Data Analysis and Introductory Statistical Inference with Statistical Formulae and Tables with SAS implementations

June 2000

Calvin L. Williams, Ph.D.
calvinw@math.clemson.edu
http://www.math.clemson.edu/∼calvinw/

Contents

Course Syllabus

Mathematical Sciences 405/605
Statistical Methods II

Instructor: Calvin L. Williams, Ph.D.	**Class Location:** M-103 Martin Hall
Office: 0-323 Martin Hall	**Class Time:**9:45-11:15 M-F
Telephone: 656-5241	**Office Hours:** M-F:2:00-3:00 or By Appointment
E-mail: calvinw@math.clemson.edu	**WWW:** http://www.math.clemson.edu/∼calvinw/MthSc405

I. **Text:** Linear Statistical Models : An Applied Approach by Bowerman and O'Connell.

Prerequisites: MthSc 301-302 or equivalent

II. **Course Description:** This course is designed to continue with intermediate probability and statistics at an intermediate level. Emphasis is placed on the understanding of the concepts of techniques in inferential statistics, data analysis, and regression analysis along with its appropriate application. This should prepare you for the practical application of regression and other modeling techniques in more general areas such as engineering, the sciences, education, and management. The intent is to cover the prescribed text omitting those sections that are unneeded with additional information given in the form of handouts and take home projects. Although there will be no requirement of a specific statistical computing package, it would be in your best interest to be familiar with a package. Examples shown in class will be done using Statistix and SAS. Notes and example code are available for students wishing to use SAS.

III. Short Course Itinerary

- Brief review of techniques in statistical inference
- Least Squares and Simple Linear Regression
- Polynomial Regression
- Multiple Linear Regression
- Diagnostics and Model Building
- Indicator variables and the Analysis of Variance
- Nonlinear Regression, just a little.

IV. **Attendance Policy:** All classes should be attended. If not, legitimate excuses must be offered with respect to the day(s) missed. Attendance will be monitored. It is to the instructors discretion whether an excuse is legitimate or not. Accordingly, the university's policy on religious holidays will be acknowledged and honored. IF YOU ARE ILL STAY HOME. You may call me or e-mail me in advance of class if you are ailing. **Note that this does not exempt you from examinations, homework or project due dates.**

V. **Tardy Professor Policy:** If the instructor is more than 15 minutes late for any class you may leave.

VI. **Examination Policy:** There will be weekly fifty minutes in class closed book quizzes and a **final** examination, also closed book. Students should bring a calculator, two clean regulation size($8\frac{1}{2}$") sheets for scratch work to be turned in with exams, and of course something with which to write, preferably pencil. There will be no sharing of calculators, scratch sheets, or writing utensils during the exams. **No makeup examinations will be given**. Any student who misses an examination without a **legitimate excuse**,*e.g. a documented medical excuse,* will receive a score of **zero** for that exam. A student with a **legitimate excuse**, will receive a final score based on all other class work. More than one missed exam with require withdrawal from the course and/or the receipt of a failing final grade.

VII. **Homework Assignments, Case Studies, and Class Project:**

There will also be several homework sets and case studies from different application areas assigned from the text as well as from material covered during class. Although it is imperative that each student be completely comfortable with these assigned problems and projects, group study is encouraged. There will also be a class project as described below.

VIII. **Requirements for Homework Assignments**

A. **Homework:**

(a) Problems will generally be due the **next** class session after the class session in which they were assigned unless stated otherwise at the time of the assignment.

(b) Solutions should be written out clearly and completely in the context of the problem posed.

B. **Case Studies:**

(a) Case studies will generally be due the **second** class session after the class session in which they were assigned.

(b) The analysis should include a *description* of the problem. I will generally include this with the assignment.

(c) The analysis should include *summary* statistics written in a **narrative** form. Tables can be included for centrally locating these results.

(d) The analysis should also include any graphical descriptions, along with a **narrative** describing the graphs, plots, etc.

(e) Complete computer printouts, command line results, or any other precursory results are not necessary and should not be turned in unless requested.

C. **Class Project:**

(a) The data set must not be taken from any text book, although data from journal articles are satisfactory. You may even consider collecting your own data. In other words, the internet or the course web page will be your best source.

(b) Data must have at least 40 cases and at least three measured characteristics. You can reduce this for you presentations, but must justify your reasons for doing so.

(c) Write ups should include all of those items required for regular class homework, ie, summaries, graphics, exposition, etc.

IX. **Grading Policy:** The weekly regular quizzes will count as 60% of the final grade, homework sets and projects 20%, and final exam 20%. The final exam will cover the more important topics covered during the semester.

X. **Grading Scale:** **A**$\Leftarrow$ 100 - 90, **B**$\Leftarrow$ 89 - 80, **C**$\Leftarrow$ 79 - 70, **D**$\Leftarrow$ 69 - 60, and **F**$\Leftarrow$ 59 - 0

XI. **Academic Dishonesty:** Academic dishonesty will not be tolerated. For information regarding the definition of acts of academic dishonesty and the subsequent penalties, you are referred to the 1999-2000 Student Handbook.

1 Class Project

Mathematical Sciences 405
Statistical Theory and Methods II
Project Description, Summer 2000

This project is an opportunity to use the statistical techniques we have learned in class, to answer real-life questions. Projects may be done individually, or as a team of 1 or 2, preferably 2. Each team must:

- Choose a question that is of interest to them, and that can be answered via a designed experiment or an observational study.
- Design and perform an experiment, gathering data to answer the question. Published data are not acceptable. Data that were gathered for a project in another class are acceptable, provided the guidelines for *this* project are met.
- Analyze the data in whatever way is appropriate.
- Report the findings.

You will have about 2 weeks to perform your experiment and analyze and report your findings. Plan your time accordingly.

The team grade will be based on the final report, which should contain the following items.

- A description of the question, and the team's reasons for wanting to know the answer,
- A description of the techniques used for gathering the data, including how randomization was performed and how the sample size was chosen,
- Analysis and illustration of the findings and conclusions.
- A listing of all the data, and example of a data-collection form (if used) and the details of any unusual calculations.

Reports should be neatly typed, well-organized and attractive. Graphical displays (either computer-generated or hand-drawn) are encouraged. Generally, graphs are more effective if they are incorporated into the text, rather than hidden at the end of the report. You may also use a computer package to aid in the data analysis. If you do so, the results should be discussed in the text of your report, and the computer output itself may be included in an appendix.

A rough draft of the final report will be due approximately 2 weeks before the final report is due. The critique and rough draft will be given back to the original group, who can change or add finishing touches before turning in the final report.

The project is worth 100 points. Grades will be based on:

Appropriate and correct procedures	50 pts
Well-written and attractive presentation	20 pts
Grammar, spelling and punctuation	20 pts
Complexity	10 pts

All members of the team will receive the same grade. It is the team's responsibility to see that all members make a fair contribution. A project proposal (not graded) must be approved before the project is started. An approved proposal must be turned in with the final report. The proposal should state:

- The question and its motivation

- Plan for collecting data, details of how randomness will be achieved, planned sample size and reason for it.
- Proposed analysis.

Due dates:

2 Homework and Case Studies Due Dates

- Reading Assignment Chapter 1
- Reading Assignment Chapter 2
- Homework Assignment: Due: 2.7, 2.10, 2.17, 2.19, 2.22
- Reading Assignment Chapter 3
- Homework Assignment: Due: 3.1, 3.10, 3.21, 3.24, 3.30
- Case Study: Medicine: Due: TBA
- Reading Assignment Chapter 4
- Homework Assignment: Due: 4.7, 4.10, 4.14, 4.17, 4.23
- Reading Assignment Chapter 5
- Homework Assignment: Due:5.6, 5.18, 5.25, 5.34, 5.37
- Homework Assignment: Due:5.46,5.47, 5.50, 5.52
- Reading Assignment Chapter 6
- Homework Assignment: Due: 6.9, 6.11, 6.14, 6.15, 6.25
- Reading Assignment Chapter 7 Section 8.
- Homework Assignment: Due: 7.12, 7.13, 7.14, 7.15
- Reading Assignment Chapter 8
- Case Study: Climatology: Due: TBA
- Homework Assignment: Due: 8.4, 8.10, 8.17
- Reading Assignment Chapter 9
- Homework Assignment: Due: 9.3, 9.6, 9.9
- Reading Assignment Chapter 10 (10.1,10.4-10.5) and Multicollinearity Handout
- Case Study: Inducer: Due: TBA
- Homework Assignment: Due: 10.3
- Reading Assignment Chapter 11
- Homework Assignment: Due: 11.5, 11.8, 11.9
- Reading Assignment Chapter 12
- Homework Assignment: Due: TBA
- Reading Assignment Chapter 13
- Homework Assignment: Due: TBA
- Reading Assignment Chapter 14
- Homework Assignment: Due: TBA
- Reading Assignment Chapter 15
- Homework Assignment: Due: TBA

3 Introduction

At its basic study, Statistics can be partitioned into two or three major foundational areas: data exploration, data categorization and analysis (eg. modeling), and statistical inference. Data exploration begins in a exploratory form and becomes more practical and provocative as data and the constraints placed on modeling data becomes more complex.

4 Data Types

- ***Quantitative data***
  - **Continuous data**
  - **Discrete data**
- ***Qualitative(categorical) data***
  - **Nominal data**
  - **Ordinal data**

5 Descriptive Statistics-Informal data definitions

5.1 Main terms and concepts

Population, population distribution, population parameters, sample, sample statistics, sampling distribution, point estimator, interval estimator, confidence interval.

- **Population:** A population is the totality of units under study. That is, units that are unmeasured as well as those measured. One or more characteristics or attributes are measured and analyzed.
- **Cumulative distribution:** A population can be described in terms of its cumulative distribution function which gives the proportion of the population less than each possible value, usually denoted, $Pr(X \leq x)$.

  A *discrete* population can be described by a probability function giving the proportion of the population equal to each possible value.
- **Density function:** A continuous population can be described by a density function, which is the derivative of the cumulative distribution function. A density function can be approximated by a histogram giving the proportion of the population lying within each of a series of intervals of values. A probability density function is like a histogram with an infinite number of infinitely small intervals.
- **Sample:** A sample is a part of the population from which the characteristic under study is measured and analyzed in order to make inferences back to the population.
- ***Sample* statistic:** A sample statistic is a mathematical function of the sample values. A statistic is to a sample what a parameter is to a population. It is customary to denote sample statistics in arabic, such as the sample mean X, and to denote population parameters in greek, such as the population mean μ.
- **Estimate:** Often we wish to estimate or guess what a population characteristic's value is under certain circumstances. We can get an estimate of the characteristic's value based on the characteristic's value for a simple random sample. There could be several ways to estimate the population parameter. For different characteristics there could be different estimates.

- ***Simple random* sample:** A simple random sample is a sample taken from the population where every unit in the population has the same probability of being selected.

- ***Parameter* estimation:** Assume a simple random sample of size *n* is taken from a population with mean μ and variance σ^2. If these population parameters are unknown, they must be estimated from the sample data.

5.2 Measures of Location or Central tendency

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size *n* drawn from some population

- ***Mean:*** The sample mean X is defined by:

  $$X = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

  It is the "best" point estimate of the population mean μ = E(X), when it is unknown.

- ***Median:*** The population median is the central value, lying above and below half of the population values. The sample median is determined similarly, that is, it is the middle value when the sample values are ordered in ascending or descending order. If *n* is odd, the median is just the $\frac{n}{2}$ + 1*st* value. If *n* is *even* it is the average of the middle two points, the ($\frac{n}{2}$) + ($\frac{n}{2}$ + 1)*st* values divided by two.

- ***Mode:*** The mode is the value at which the density of the population is at a maximum. Some densities have more than one maximum point and are said to be multimodal. The sample mode is the value that occurs most often in the sample. If there is a tie for the most often occurring sample value, the sample is said not to have a mode. If the population is continuous, then all sample values occur only once and the sample mode has very little use.

- ***Weighted* Mean** Given that the weight associated with x_i is $w_i > 0$, positive and non-zero for all *i*: $\overline{x}$ = $\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ = $\frac{x_1+x_2+\cdots+x_n}{n}$

- ***Geometric* Mean** Given that $x_i > 0$, positive and non-zero for all *i*: GM = $\sqrt[n]{x_1 \cdot x_2 \cdots x_n}$

- ***Harmonic* Mean** Given that $x_i > 0$, positive and non-zero for all *i*:
  HM = $\frac{n}{\displaystyle\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$. Given equal observations GM≤HM≤ $\overline{x}$.

- ***Percentile Trimmed (p%)* Mean** Delete the *p*% smallest and the largest *p*% of a sample. $\overline{x}_{tr(p)}$ is the arithmetic mean of the remaining data.

5.3 Quantiles

Quantiles, including percentiles, quartiles, and the median, are useful for a detailed study of a distribution.

- [**Quantiles]** For a data set consisting of *n* values that when ordered are $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$,
  1. for any number **p** of the form $\frac{i-0.5}{n}$, where *i* is an integer from 1 to *n*, the **p** quantile of the data set will be taken to be x_i. (The *i*th smallest data point will be called the $\frac{i-0.5}{n}$ quantile.)
  2. for any number **p** between $\frac{0.5}{n}$ and $\frac{n-0.5}{n}$ that is not of the form$\frac{i-0.5}{n}$, the **p** quantile of the data set will be obtained by linear interpolation between the quantiles corresponding to the two values of $\frac{i-0.5}{n}$ that bracket **p**.

     In both cases, the notation Q(**p**) will be used to symbolize the **p** quantile.

- ***Percentile:*** For a set of measurements arranged in order of magnitude, the *p*th percentile is the value that has *p*% of the measurements below it and *(100-p)*% above it.

- **Quartiles** Q_1, Q_2, Q_3**:**

$$\text{If the number } n \text{ is even} \quad : \begin{cases} Q_1 & \text{is the median of the smallest } n/2 \text{ observations} \\ Q_3 & \text{is the median of the largest } n/2 \text{ observations} \end{cases}$$

$$\text{If the number } n \text{ is odd} \quad : \begin{cases} Q_1 & \text{is the median of the smallest } (n-1)/2 \text{ observations} \\ Q_3 & \text{is the median of the largest } (n-1)/2 \text{ observations} \end{cases}$$

  The *1*st quartile is *25*th% *tile*. The *2*nd quartile is *50*th% *tile* and Median. And, the *3*rd quartile is *75*th% *tile*. Obviously the *4*th quartile is *100*th% *tile*.

- **Quintiles** $\boldsymbol{P_{20}}$,$\boldsymbol{P_{40}}$,$\boldsymbol{P_{60}}$, and $\boldsymbol{P_{80}}$ percentiles:
  P_{20} = $\frac{20}{100}$ (n+1)*st* = $\frac{1}{5}$ (n+1)*st* observation.
  P_{40} = $\frac{40}{100}$ (n+1)*st* = $\frac{2}{5}$ (n+1)*st* observation.
  P_{60} = $\frac{60}{100}$(n+1)*st* = $\frac{3}{5}$ (n+1)*st* observation.
  P_{80} = $\frac{80}{100}$ (n+1)*st* = $\frac{4}{5}$ (n+1)*st* observation.
  The common thought is to round up on all non-integer values for measures of location.

5.4 Measures of Variability or Spread

This group of measures are also important in giving a detailed study of a distribution. It is important to note that with measures of variability or spread if the entire set of observations are changed by adding or subtracting a fixed(constant) amount then the sample statistics are unchanged, but if the are multiplied by a fixed constant, they sample statistics are changed.

- ***Range:*** The sample range is the difference between the largest and the smallest values in the sample. For many populations, at least in statistical theory, the range is infinite, so the sample range may not tell you much about the population. The sample range is finite and tends to increase as the sample size increases. If all the sample values are multiplied by a constant, the sample range is multiplied by the same constant.

- **Interquartile range:** The interquartile range is the difference between the the *3*rd quartile and the *1*st quartile. If the sample values are multiplied by a constant, the sample interquartile range is multiplied by a constant.

- **Variance:** The population variance, usually denoted σ^2 when it is clear what population is being considered, is the expected value of the squared difference of the values from the population mean:

$$\sigma^2 = E(X - \mu)^2$$

The sample variance, s^2, is defined by:

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1} \\ &= \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n-1} \\ &= \frac{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}{n(n-1)}. \end{aligned}$$

The difference between a value and the mean is called a deviation from the mean. Thus the variance is the sum of the squared deviations from the mean divided by *n* in the case of the population, and *n*-1 in the case of the sample. When all of the values lie close to the mean, the variance is small but never zero. If the sample values are multiplied by a constant, the sample variance is multiplied by the square of the constant.

- **Standard Deviation:** the standard deviation is the square root of the variance, or root-mean-square deviation from the mean, in either the population or the sample. The sample standard deviation is expressed in the same units as the values in the sample, not squared units like the variance. If all sample values are multiplies by a constant, the sample standard deviation is multiplied by the same constant.
- **Standard Deviation:**
  s= $\sqrt{s^2}$(unbiased variance) or $\tilde{s}$= $\sqrt{\tilde{s}^2}$(biased variance)
- **Standard Error of $\overline{X}$ as an estimate of the population mean:**
  s.e $(\overline{x})$= $s_{\overline{x}}$ = s/$\sqrt{n}$
- **Coefficient of variation:** The coefficient of variation is a unitless measure of relative variability. It is defined as the ratio of the standard deviation to the mean expressed as a percentage. The coefficient of variation is meaningful only if the variable is measured on the ration scale. If the sample values are multiplied by a constant, the sample coefficient of variation remains unchanged.
- **Variability:(sample), unbiased** s^2= $\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2-\overline{x}\right)^2$
- **Variability:(sample), biased** $\tilde{s}^2$= $\frac{1}{n}\sum_{i=1}^{n}\left(x_i^2-\overline{x}\right)^2$
- **Range:** *R* = $\max\{x_1, x_2, \ldots, x_n\}$ - $\min\{x_1, x_2, \ldots, x_n\}$ = $x_{(n)}$ - $x_{(1)}$
- **Interquartile Range:** *IQR* = Q_3 - Q_1
- **Useful for Box Plots:**
  Inner Fences: Q_1-1.5 IQR, Q_3+1.5 IQR
  Outer Fences: Q_1-3 IQR, Q_3+3 IQR
- **Linear Transformations:** Let y_i = ax_i+ b, then $\overline{y}$ = a$\overline{x}$ + b, s_y^2 = a$^2 s_x^2$, s_y = | a | s_x
  Important notes: Linear transformations do not change the shape of the data (distribution).

5.5 Measures of Shape

- ***Skewness:*** The variance is a measure of the overall size of the deviations form the mean. Since the formula for the variance squares the deviations, both positive and negative deviations contribute to the variance in the same way. In may distributions, positive deviations may tend to be larger in magnitude than negative deviations, or vice versa. *Skewness* is a measure of the tendency of deviations to be larger in one direction than the other. The population skewness is defined by

$$\frac{E(X-\mu)^3}{\sigma^2}$$

Since the deviations are cubed, rather than squared, the signs of the deviations are maintained. Cubing the deviations also emphasizes the effects of large deviations. The formula includes a divisor of σ^3 to remove the effect of scale, so multiplying all values by a constant does not change the skewness. Skewness can thus be interpreted as a tendency for one tail of the population to be heavier than the other. The sample skewness can be calculated by:

$$\frac{m_3}{(m_2)^{\frac{3}{2}}}$$

where

$$m_j = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^j$$

- ***Kurtosis:*** the heaviness of the tails of the population affects the behavior of many statistics. Hence it is useful to have a measure of tail heaviness. One such measure is *kurtosis*. The population kurtosis is usually defined as:

  $$\frac{E(X-\mu)^4}{\sigma^4} - 3,$$

  although some statisticians omit the subtraction of 3. Since deviations are raised to the fourth power, positive and negative deviations make the same contribution, while large deviations contribute strongly. Because of the divisor σ^4, multiplying each value by a constant has no effect on kurtosis.

  Population kurtosis must lie between -2 and positive infinity, inclusive. If m_3 represents population skewness and m_4 represents population kurtosis, the $m_4 \geq (m_3)^2 - 2$.

  There is a myth in the literature that kurtosis measures the peakedness of a density.

  Sample skewness and kurtosis are rather unreliable estimators of the corresponding parameters in small samples. Trust them only if you have a very large sample. However, large values of skewness or kurtosis may merit attention even in small samples because such values indicate that statistical methods based on normality assumptions may be inappropriate.

6 Graphical Descriptions of Data

6.1 Boxplots

A boxplot or box-and-whisker plot is a graphical representation of data in which a rectangle is used to summarize the data distribution. The top and the bottom, sometimes the left and right, of the rectangle represent the third and first quartiles, respectively. The line inside the rectangle represents the median. The lines extending from the top and bottom of the rectangle represent either the actual limits of the data, or the limits of the bulk of the data (with unusual observations, sometimes referred to as **outliers** see below, being represented by individual symbols [“flagged”] if they are further out, **modified box plot**). The boxplot is particularly useful for comparing the location and variability of several batches of data, as boxes can be plotted side-by-side on one plot.

6.2 Dotplots

A dotplot is a preliminary remedial graphical representation of the data that groups the data into many small classes or intervals.

6.3 Histograms

A histogram is another graphical representation of the distribution of a batch of data. The data values are usually grouped into mutually exclusive and exhaustive intervals of equal width, and the number of observations in each interval is determined and represented by a vertical bar. In some variations the widths of the intervals are varied, resulting in potentially different appearances in the plot.

6.4 Stem and leaf

A stem-and-leaf display is another graphical representation of the distribution of a batch of data. Very similar to a histogram, it is often accompanied by additional information about the data, such as cumulative frequencies and the position of the median. The plot represents the data values by their numerical values, providing additional information over the histogram, but the grouping intervals are usually chosen based on using round numbers, rather than in an attempt to provide the most effective plot.

6.5 Scatter plot

A scatterplot a is a graphical method that can be used to study the joint variation of two variables graphically. Each observation is represented by a point (x,y) on the plot, indexed by the values on the axes. Each axis is used for a different variable. Besides showing how (and whether) two variables are related to each other, scatter plots also can indicate the existence of distinct subgroups in the data. Scatterplots can only be used if there are data pairs (x_i, y_i).

6.6 Density Curves

Density curves are functional and or graphical representations of data that are usually continuous in nature.

6.7 Q-Q plots

Quantile-Quantile Plots are useful for comparing distributions. They are generally used to determine if data in a sample follow a particular distribution. In statistics in order to make inferences, it is often assumed that data follow the normal distribution. In which case the quantiles of the sample are compared to the quantiles of the normal distribution. These are generally referred to as normal probability plots.

7 Sampling Distributions derived from the normal

7.1 The Normal distribution

The normal distribution is probably the most important probability distribution in statistics! It is a probability distribution of a continuous random variable, yet it is often used to model the distribution of other continuous random variables and discrete random variables. The reason for the versatility in using the normal distribution as a probability distribution model is indicated in the figure below. The basic form of the normal distribution is that of a bell it has a single mode and is symmetric about its central value. The flexibility in using the normal distribution is due to the fact that the "bell" may be centered over any number on the real line and it may be made flat or peaked to correspond to the amount of dispersion that the values of a random variable may assume. Examples of random variables that have been successfully modeled by the normal distribution are the height and weight of persons, the diameter of bolts of a specified size produced on a machine, the IQ of persons, and the lifetime in hours of batteries or light bulbs. Typically, in the type of experiment that produces a random variable that can be successfully approximated by a normal random variable, the values of the random variable are produced by a measuring process, where it is known that the measurements tend to cluster symmetrically about a central value. **A random variable that is an average or a sum of values of another random variable is, under very general conditions, almost always distributed approximately as a normal random variable, regardless of the form of the distribution of the random variable with values that are summed or averaged.** An example of such a random variable is the average grade point average of a group of students selected at random from the population of students at your university or college. The notion that a random variable that is an average is distributed as a normal random variable is discussed when we describe the central limit theorem. For a random variable to be normally distributed, the mathematical expression delineating the form of the bell must be of a specific type as described in the following definition:

$$f(y:\mu,\sigma^2) \;=\; \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}} \; -\infty \leq y \leq \infty$$

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}}dy \;=\; 1$$

$$E[Y] \;=\; \mu; \qquad E\left[Y^2\right] \;=\; \sigma^2+\mu^2; \qquad Var\left[Y\right] \;=\; \sigma^2$$

7.2 t-distribution

Given the sample statistics $\overline{X}$, the sample mean, $S^2$the sample variance, we now derive distributions based on the normal distribution. Let X$_1$, …, X$_n$ be a random sample from a N(μ,σ^2) distribution. The quantity $\frac{(\overline{X}-\mu)}{S/\sqrt{n}}$ has a Student's t distribution with *n* - 1 degrees of freedom. The moments of which are 0 and $\frac{n}{n-2}$. The **density** function of the *t*-distribution is given by

$$f(y,\nu) \;=\; \frac{1}{\sqrt{\pi\nu}}\frac{\Gamma(\frac{(\nu+1)}{2})}{\Gamma(\frac{\nu}{2})}\left(1+\frac{y^2}{\nu}\right)^{\frac{-(\nu+1)}{2}} \qquad -\infty \leq y \leq \infty$$

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{\pi\nu}}\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(1+\frac{y^2}{\nu}\right)^{\frac{-(\nu+1)}{2}}dy \;=\; 1.$$

$$E[Y] \;=\; 0, \;\; \nu \geq 2; \qquad E\left[Y^2\right] \;=\; \tfrac{\nu}{\nu-2}; \qquad Var\left[Y\right] \;=\; \tfrac{\nu}{\nu-2}, \;\; \nu \geq 3$$

7.3 χ^2-distribution

Let X$_1$, …, X$_n$ be a random sample from a N(μ, σ^2) distribution. The quantity $\frac{(n-1)S^2}{\sigma^2}$ has a chi-squared distribution. The **density** function of the χ^2-distribution is given by

$$f(y,\nu) \;=\; \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} y^{\frac{\nu}{2}-1} e^{-\frac{y}{2}} \qquad 0 \le y \le \infty$$

$$\int_0^\infty \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} y^{\frac{\nu}{2}-1} e^{-\frac{y}{2}} dy \;=\; 1.$$

$$E[Y] \;=\; \nu; \qquad E[Y^2] \;=\; 2\nu + \nu^2; \qquad Var[Y] \;=\; 2\nu$$

7.4 F-distribution

Let X$_1$, …, X$_n$ be a random sample from a N(μ_x, σ_x^2) population, and let Y$_1$, …, Y$_m$ be a random sample from an independent N$\left(\mu_y, \sigma_y^2\right)$ population. If we were interested in comparing the variability of the populations, one quantity of interest would be the ratio $\frac{\sigma_x^2}{\sigma_y^2}$. Information about this ratio is contained in $S_X^2/S_Y^2$the ratio of sample variances. Recall from our previous discussion that $\frac{(n-1)S^2}{\sigma^2}$ has a chi-squared distribution. Then the ratio of two chi-squares, divided by their respective degrees of freedom has an *f*-distribution with $n-1$=p numerator and $m-1$ = q denominator degrees of freedom. Note that if the null hypothesis is true then this ratio is the same as f, given before. The **density** function of the *F*-distribution is given by

$$f(y,\nu) \;=\; \left(\frac{p}{q}\right)^{p/2} \frac{\Gamma(\frac{(p+q)}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})} y^{p/2-1} \left(1 + \left(\frac{p}{q}\right) y\right)^{\frac{-(p+q)}{2}} \qquad -\infty \le y \le \infty$$

$$\int_{-\infty}^\infty \left(\frac{p}{q}\right)^{p/2} \frac{\Gamma(\frac{(p+q)}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})} y^{p/2-1} \left(1 + \left(\frac{p}{q}\right) y\right)^{\frac{-(p+q)}{2}} dy \;=\; 1.$$

The cumulative distribution is given obviously by Pr$[f_{n_1-1,n_2-1} \le f]$
Examples of using the *f*-distribution :
For instance,

(a) Pr$[f_{10,15} \le 2.54]$ = 0.95

(b) Pr$[f_{10,15} > 3.06]$ = 0.025

(c) $F_{10,15,0.0.975}$ = 3.06

(d) $F_{10,15,0.95}$ = 2.54

(e) $F_{10,15,0.0.025}$ = $\frac{1}{F_{15,10,0.0.975}}$ = $\frac{1}{3.52}$ = 0.28

(f) $F_{10,15,0.05}$ = $\frac{1}{F_{15,10,0.0.95}}$ = $\frac{1}{2.85}$ = 0.35.

The F distribution can be derived in a more general setting than is done here. A variance ratio may have an F distribution even if the parent populations are not normal. Kelker (1970) has shown that as long as the parent populations have a certain type of symmetry (spherical symmetry), then the variance ratio will have an F distribution.

8 Analyses of Univariate Data

Univariate data can be broadly classified into one of two types:

1. Cross-sectional data, measurements from a random sample, data in which the ordering is not important in the analysis.
2. Logitudinal data are measurements (observations) of the same quantity on the same subject ate different time points.

We will concentrate for the time being on cross-sectional data, considering graphical and numerical descriptions of these data, and finally making inference by determining appropriate models, estimates of model parameters,and making any inferences warranted by the analysis.

For these data types the theory of the normal distribution plays an important role.

1. For theoretical reasons real data are usually considered normally distributed.
2. Once we have determined normality, the data is usually easier to work with.
3. For descriptive reasons as well as reasons corresponding to making inferences on the data. The standard deviation and mean can be determined readily if the distribution is normal. Other distribution can create some difficulty in terms of parameter estimation.

It becomes a very important task to determine whether a distribution is normal or nonnormnal. In terms of numerical measures, this determination can get clouded. For instance, measures of location(mean, median, mode) can be similar for several distributions.

9 Assessing Normality

9.1 Probability plots

Probability plots are an extremely useful graphical tool for qualitatively assessing the fit of data to a theoretical distribution. Consider a sample of size *n* from a uniform distribution on *[0,1]*. Denote the ordered sample values by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. These are called the order statistics. It can be shown that $E(X_{(j)}) = \frac{j}{n+1}$. This suggests plotting the ordered observations $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, against the points $\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}$. This should be recognized as being the cumulative distribution function of the uniform distribution.

This technique can be extended to other continuous probability laws (distributions). Let's say that *X* is a continuous random variable with a strictly increasing cumulative distribution function F_x, and if $Y = F_x(x)$ then *Y* has a uniform distribution on $[0,1]$. $Y = F_x(x)$ is known as the probability integral transform. Hence, the following procedure is suggested. Suppose that it is hypothesized that *X* follows a certain distribution, *F*. Given a sample $x_1, x_2, \ldots, x_n$ we plot

$$\begin{aligned} F\left(X_{(k)}\right) \quad vs \quad & \frac{k}{n+1} \;\; \text{Uniform : } \mathrm{E}\left(X_{(k)}\right) \\ = \quad & \frac{k}{n+1} \end{aligned}$$

or equivalently

$$X_{(k)} \quad vs \quad F^{-1}\left(\frac{k}{n+1}\right)$$

In some cases, F is of the form

$$F(x) \quad = \quad G\left(\frac{x-\mu}{\sigma}\right)$$

where μ and σ are called location and scale paramaters, respectively. The normal distribution is of this form. We could plot

$$\frac{X_{(k)} - \mu}{\sigma} \quad vs. \quad G^{-1}\left(\frac{k}{n+1}\right)$$

or, if we plotted

$$X_{(k)} \quad vs. \quad G^{-1}\left(\frac{k}{n+1}\right)$$

the result would be approximately a straight line, if the model were correct

$$X_{(k)} \quad \approx \quad \sigma G^{-1}\left(\frac{k}{n+1}\right) + \mu$$

Slight modifications of the procedure are sometimes used. For example, rather than $G^{-1}(\frac{k}{n+1})$, $E(X_{(k)})$, the expected value of the k^{th} smallest observation can be used. But it can be argued that

$$E\left(X_{(k)}\right) \approx F^{-1}\left(\frac{k}{n+1}\right) = \sigma G^{-1}\left(\frac{k}{n+1}\right) + \mu$$

So, this modification yields very similar results to the original procedure.

The procedure can be viewed from another perspective. Given that $F^{-1}\left[\frac{k}{n+1}\right]$ is the $\frac{k}{n+1}$ st quantile of the distribution F, that is the point such that the probability that a random variable with distribution function F is less than it is $\frac{k}{n+1}$. We are thus plotting the ordered observation (which may be viewed as the observed or empirical quantile) versus the quantile of the theoretical distribution. An example set of observations. We have tensile strengths from 4 different types of die sets in which we have taken 10 observations each.

Table 1: Tensile Strengths from 4 Die Sets

Observation	Die 1	Die 2	Die 3	Die 4
1	18.9	16.9	19.9	15.9
2	19.3	17.5	20.2	16.0
3	19.5	17.8	21.3	16.8
4	20.0	18.0	21.5	17.2
5	20.5	18.3	21.7	17.4
6	20.6	18.4	21.8	17.5
7	20.7	18.6	21.9	17.7
8	20.8	18.8	21.9	17.9
9	21.0	19.2	22.5	18.1
10	22.1	19.9	23.0	19.0

We can get summary statistics for the tensile strengths of these different types of dies. They are:

```
                    DIE1        DIE2        DIE3        DIE4
N                     10          10          10          10
SUM               203.40      183.40      215.70      173.50
MEAN              20.340      18.340      21.570      17.350
SD                0.9395      0.8592      0.9393      0.9419
VARIANCE          0.8827      0.7382      0.8823      0.8872
SE MEAN           0.2971      0.2717      0.2970      0.2979
C.V.              4.6190      4.6848      4.3548      5.4290
MINIMUM           18.900      16.900      19.900      15.900
1ST QUARTILE      19.450      17.725      21.025      16.600
MEDIAN            20.550      18.350      21.750      17.450
3RD QUARTILE      20.850      18.900      22.050      17.950
MAXIMUM           22.100      19.900      23.000      19.000
MAD               0.5000      0.5000      0.3500      0.5500
BIASED VAR        0.7944      0.6644      0.7941      0.7985
SKEW              0.1641      0.1426     -0.4853     -0.0774
KURTOSIS         -0.4912     -0.4141     -0.4211     -0.5424
```

9.2 Quantile-plots

Let's construct a quantile plot for die types 2 and 3. First, reconsider the table of values. A quantile plot is simply a scatterplot of the observation versus it's quantile. So the quantile plot for the third die type is given in figure 1.

Table 2: Tensile Strengths from 4 Precision Die Sets

Observation	$p=\frac{i-0.5}{n}$	Die 1	Die 2	Die 3	Die 4
1	0.05	18.9	16.9	19.9	15.9
2	0.15	19.3	17.5	20.2	16.0
3	0.25	19.5	17.8	21.3	16.8
4	0.35	20.0	18.0	21.5	17.2
5	0.45	20.5	18.3	21.7	17.4
6	0.55	20.6	18.4	21.8	17.5
7	0.65	20.7	18.6	21.9	17.7
8	0.75	20.8	18.8	21.9	17.9
9	0.85	21.0	19.2	22.5	18.1
10	0.95	22.1	19.9	23.0	19.0

9.3 Quantile-Quantile Plots

Quantile-Quantile plots are useful for comparing distributions. They are generally used to determine if data in a sample follow a particular distribution. In statistics in order to make inferences, it is often assumed that data follow the normal distribution. In which case the quantiles of the sample are compared to the quantiles of the normal distribution. These are generally referred to as normal probability plots.

Let's construct a Q-Q plot for die types 2 and 3. This is given in figure 2. The fact that a significant amount of points do not fall on the line superimposed signifies that the two sets of observation are different in terms of there distribution.

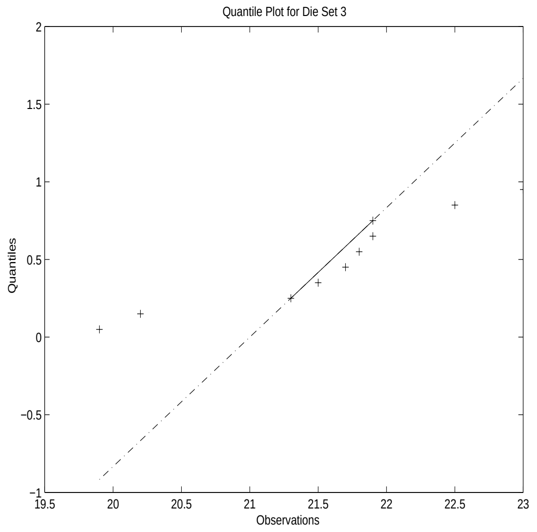


Figure 1: Quantile plot for Die Set

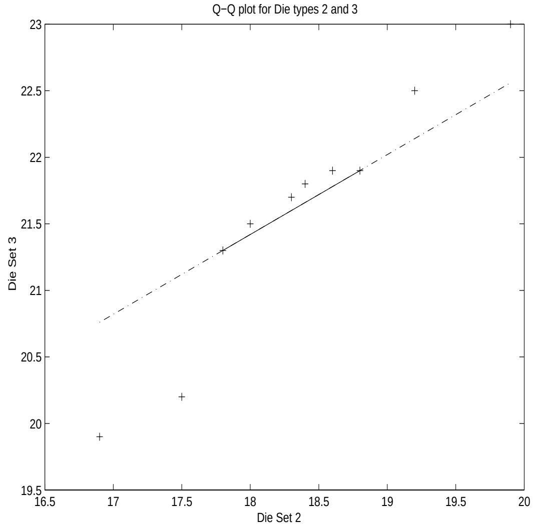


Figure 2: Comparative Quantile-Quantile plot for Die Sets 2 and 3

9.4 Comparable Normal probability plots

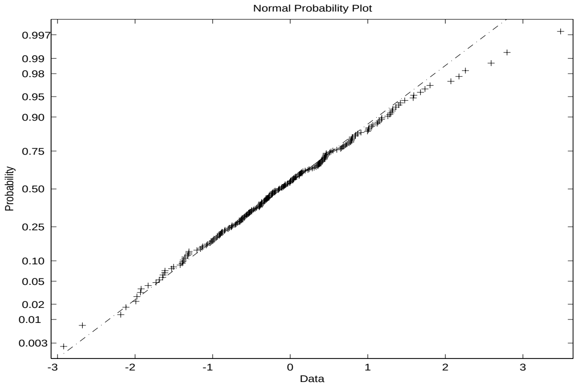


Figure 3: Ideal normal plot - signifying data is normally distributed

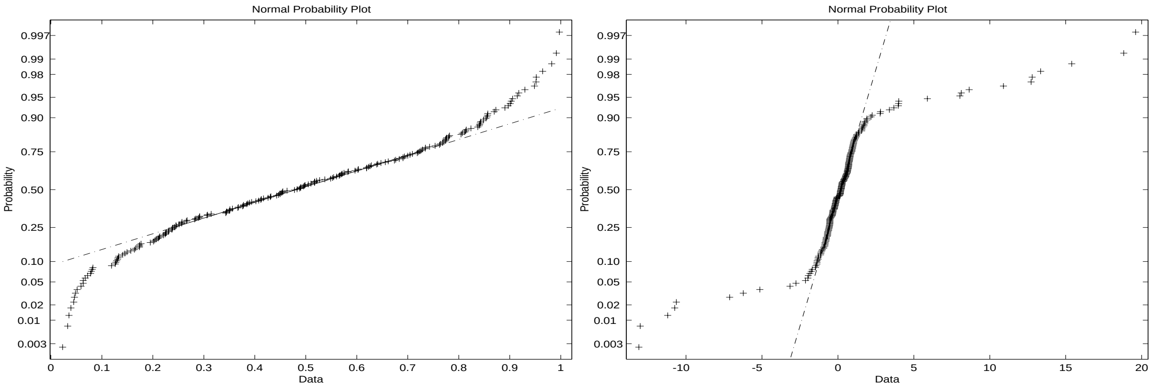


Figure 4: Figure on left - heavy tailed distribution, figure on right light - tailed

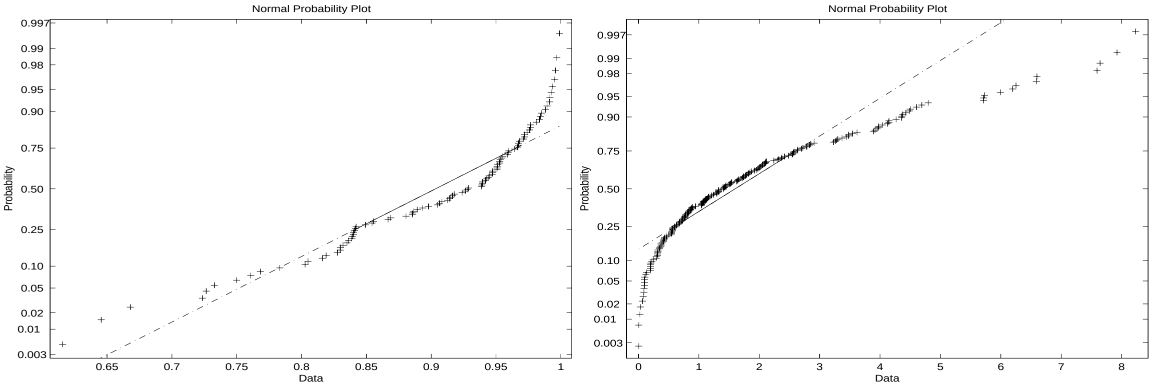


Figure 5: Figure on left - positive skew (skewed right) , figure on right negative skew (skewed left)

10 Statistical Inference Tests and Confidence Intervals (Chapt. 3 LSM)

10.1 Confidence Intervals

The general form of a confidence interval for some unknown parameter is given by:

$$\widehat{\theta} \quad \pm \quad SD_{\widehat{\theta}\frac{\alpha}{2}} SE_{\widehat{\theta}}$$

where

- $\widehat{\theta}$ is an estimator of the parameter,
- SD$_{\widehat{\theta}}$ is the sampling distribution of the estimator, and
- SE$_{\widehat{\theta}}$ is the standard error of the estimator.

That is, if we were to sample the population say, a large but finite number of times, $(1-\alpha)100\%$ of the intervals generated from the samples will contain the true population parameter.

There is a duality between confidence intervals and hypotheses tests. Consider the following example. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution having unknown mean μ and known variance σ^2. We consider testing the following hypothesis:

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_0 : \mu &\neq \mu_0 \end{aligned}$$

Consider a test at a specific level α that rejects for $|\overline{x}-\mu_0| > C$, where C is determined so that $Pr\{|\overline{x}-\mu_0| > C\}$ if H_0 is true: $C = \sigma_{\overline{x}} \mathcal{Z}_{\frac{\alpha}{2}}$. The test thus does not reject when:

$$|\overline{x} - \mu_0| < \sigma_{\overline{x}} \mathcal{Z}_{\frac{\alpha}{2}}$$

or

$$-\sigma_{\overline{x}}\mathcal{Z}_{\frac{\alpha}{2}} < \overline{x} - \mu_0 < \sigma_{\overline{x}}\mathcal{Z}_{\frac{\alpha}{2}}$$

or

$$\overline{x} - \sigma_{\overline{x}}\mathcal{Z}_{\frac{\alpha}{2}} < \mu_0 < \overline{x} + \sigma_{\overline{x}}\mathcal{Z}_{\frac{\alpha}{2}}$$

A $(1-\alpha)100\%$ confidence interval for μ_0 is

$$[\overline{x} - \sigma_{\overline{x}}\mathcal{Z}_{\frac{\alpha}{2}}, \overline{x} + \sigma_{\overline{x}}\mathcal{Z}_{\frac{\alpha}{2}}]$$

Comparing the acceptance region of the test to the confidence interval, we see that μ_0 lies in the confidence interval if and only if the hypothesis tests does not reject. In other words, the confidence interval consists precisely of all those values of μ_0 for which the null hypothesis $H_0 : \mu = \mu_0$ is not rejected.

10.2 Tables of Confidence Intervals-Single Sample(Chapt. 3 LSM)

Parameter	Assumptions	$100(1-\alpha)$% **Confidence** interval
μ	n large, σ known, or normality, σ^2 known	$\overline{x} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
μ	n large, σ unknown	$\overline{x} \pm z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$
μ	normality, σ^2 unknown	$\overline{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$
p	binomial experiment, large n	$\widehat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{pq}}{n}}$
σ^2	normality	$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},df=n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},df=n-1}}\right)$

11 Normal Models - Single Sample Tests of Hypothesis (Chapt. 3 LSM)

Table 3: Single Sample Tests of Hypothesis-Normal Models

Null Hypothesis	Assumptions	Test Statistic	Alternative Hypothesis	Rejection Region
$\mu = \mu_0$	n large, σ known, or normality, σ^2 known	$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \neq \mu_0$	$Z \geq z_\alpha$ $Z \leq -z_\alpha$ $\mid Z \mid \geq z_{\alpha/2}$
$\mu = \mu_0$	n large, σ unknown	$Z = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \neq \mu_0$	$Z \geq z_\alpha$ $Z \leq -z_\alpha$ $\mid Z \mid \geq z_{\alpha/2}$
$\mu = \mu_0$	normality, n small, σ unknown	$T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \neq \mu_0$	$T \geq t_{n-1,\alpha}$ $T \leq -t_{n-1,\alpha}$ $\mid T \mid \geq t_{n-1,\alpha/2}$
$\sigma^2 = \sigma_0^2$	normality	$X^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$\sigma^2 > \sigma_0^2$ $\sigma^2 < \sigma_0^2$ $\sigma^2 \neq \sigma_0^2$	$X^2 \geq \chi^2_{n-1,\alpha}$ $X^2 \leq \chi^2_{n-1,(1-\alpha)}$ $X^2 \geq \chi^2_{n-1,\alpha/2}$ or $X^2 \leq \chi^2_{n-1,(1-\alpha/2)}$
$p = p_0$	binomial experiment, n large	$Z = \frac{\widehat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$	$p > p_0$ $p < p_0$ $p \neq p_0$	$Z \geq z_\alpha$ $Z \leq -z_\alpha$ $\mid Z \mid \geq z_{\alpha/2}$

Case Study: Education

Mathematical Sciences 405/605
Case Study: Education

Intelligence tests are routinely administered by school guidance counselors and psychcologists as screening devices for their students. However, are all of these tests really accurate indicators of a student's IQ? In a study [†] to compare two such tests, the Slosson Intelligence Test (SIT) the Wechsler Intelligence Scale for Children-Revised (WISC-R), the tests were administered to a sample of 72 children in a large urban school district in central Ohio. The mean age of the children was 8.5 years with a standard deviation of 16.6 months. Scores on the two tests for the 72 children were as follows:

Test	Mean	Standard Deviation
WISC-R Full Scale	86.11	15.65
SIT IQ	90.47	14.77

i) Assume that the scores of the 72 students represent a random sample from the population of scores for all students who might take the test. Find a point estimate for the average grade on the WISC-R for the population. What is the margin of error for this estimate ?

ii) Find a 98% confidence interval for the mean grade on the SIT test.

iii) In fact, the sample taken by the experimenters was limited to students who were not making adequate academic progress in the regular classroom. What impact does this have on the inferences you can make in parts (b) and (c) ?

[†] Source: Prewett, Peter N., and D. B. Fowler. "Predictive Validity of the Slosson Intelligence Test with the WISC-R and WRAT-R Level 1." *Psychology in the Schools* **29** (January 1992), p. 17.

Case Study: General

Mathematical Sciences 405/605
Case Study: General
Will Your Bill for College Textbooks Continue to Rise?

The number of new U.S. book titles increased from almost 47,000 in 1990 to over 48,000 in 1991. However, this was still below the historic high of about 56,000 titles attained in 1987 (Grannis, 1992). Can we expect an increase or decrease in the price of books, especially hardbacks, if there are more competitors on the market? The following table gives the number of titles and the average price of hardback books classified according to 23 standard subject groups representing one or more specific Dewey Decimal Classification numbers. Consider the number of volumes and average price per volume in 1990 and 1991 as paired samples for two randomly chosen years for each of the 23 categories of books. Although there was an increase in the total number of books in 1991, the number of hardbacks seems relatively unchanged and the average price per volume seems to have increased over the average 1990 price.

	1990		**1991**	
Category	Volumes	Average Price	Volumes	Average Price
Agriculture	359	$54.24	371	$57.73
Art	759	42.18	717	44.99
Biography	1,337	28.58	1,416	27.52
Business	748	45.48	790	43.38
Education	562	38.72	556	41.26
Fiction	1,962	19.83	2,062	21.30
General works	1,035	54.77	1,071	51.74
History	1,450	36.43	1,442	39.87
Home economics	357	23.80	341	24.23
Juveniles	3,675	13.01	3,705	16.64
Language	312	42.98	240	51.71
Law	596	60.78	240	63.89
Literature	1,312	35.80	1,265	35.76
Medicine	2,215	72.24	2,078	71.44
Music	184	41.86	173	41.04
Philosophy/Psychology	963	40.58	945	42.74
Poetry/drama	486	32.19	511	33.29
Religion	977	31.31	958	32.33
Science	2,028	74.39	958	80.14
Sociology/Economics	4,504	42.10	4,306	48.83
Sports/recreation	403	30.52	440	30.68
Technology	1,521	76.48	1,620	76.40
Travel	181	30.41	156	33.50
Total	27,926	$42.12	26,361	$43.93

i) Determine whether the difference in the average number of volumes per category for 1991 differs significantly from the 1990 average, using a significance level of 5%.

ii) Determine whether the change in the average price of a hardback book per category in 1991 differs significantly from that in 1990 at the 5% level of significance.

iii) Summarize your results concerning the difference in the number and price of books per category in 1991 compared with 1990.

11.1 Two Sample Confidence intervals(Chapt. 3 LSM)

Parameter	Assumptions	$100(1-\alpha)\%$ **Confidence** interval
$p_1 - p_2$	binomial experiment, n$_1$, n$_2$ large	$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p_1}\widehat{q_1}}{n_1} + \frac{\widehat{p_2}\widehat{q_2}}{n_2}}$
$\mu_1 - \mu_2$	independence, n$_1$, n$_2$ large, σ_1^2, σ_2^2 known, or normality, independence, σ_1^2, σ_2^2 (un)known	$(\overline{x}_1 - \overline{x}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ $(\overline{x}_1 - \overline{x}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$\mu_1 - \mu_2$	independence, n$_1$, n$_2$ small, normality, σ_1^2, σ_2^2 unknown, but equal	$(\overline{x}_1 - \overline{x}_2) \pm t_{\frac{\alpha}{2},n_1+n_2-2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where S$_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)}$
$\mu_1 - \mu_2$	independence, n$_1$, n$_2$ small, normality, σ_1^2, σ_2^2 unknown, but unequal	$(\overline{x}_1 - \overline{x}_2) \pm t_{\frac{\alpha}{2},\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ where $\nu = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1}+\frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$
$\mu_1 - \mu_2 = \mu_D$	dependence, normality, n pairs,	$\overline{d} \pm t_{\frac{\alpha}{2},n-1}\frac{s_D}{\sqrt{n}}$
$\frac{\sigma_1^2}{\sigma_2^2}$	independence normality,	$\left(\frac{s_1^2}{s_2^2}\frac{1}{F_{n_1-1,n_2-1,\frac{\alpha}{2}}}, \frac{s_1^2}{s_2^2}\frac{1}{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}}\right)$

12 Two Sample Hypothesis Tests of Hypothesis(Chapt. 3 LSM)

12.1 Binomial Models

Table 4: Two Sample Hypothesis Tests- Binomial Models

Null Hypothesis	Assumptions	Test Statistic	Alternative Hypothesis	Rejection Region
$p_1 - p_2 = 0$	binomial experiment n$_1$, n$_2$ large	Z = $\frac{\left(\widehat{p}_1 - \widehat{p}_2\right)}{\sqrt{\widehat{p}\widehat{q}(1/n_1+1/n_2)}}$ $\widehat{p} = \frac{X_1+X_2}{n_1+n_2}$	$p_1 - p_2 > 0$ $p_1 - p_2 < 0$ $p_1 - p_2 \neq 0$	$Z \geq z_\alpha$ $Z \leq -z_\alpha$ $\mid Z \mid \geq z_{\alpha/2}$
$p_1 - p_2 = \triangle_0$	binomial experiment n$_1$, n$_2$ large	Z = $\frac{\left(\widehat{p}_1 - \widehat{p}_2\right) - \triangle_0}{\sqrt{\left(\widehat{p}_1\widehat{q}_1/n_1 + \widehat{p}_2\widehat{q}_2/n_2\right)}}$	$p_1 - p_2 > \triangle_0$ $p_1 - p_2 < \triangle_0$ $p_1 - p_2 \neq \triangle_0$	$Z \geq z_\alpha$ $Z \leq -z_\alpha$ $\mid Z \mid \geq z_{\alpha/2}$

12.2 Normal Models

Table 5: Two Sample Hypothesis Tests- Normal Models

Null Hypothesis	Assumptions	Test Statistic	Alternative Hypothesis	Rejection Region
$\mu_1 - \mu_2 = \triangle_0$	independence n$_1$, n$_2$ large, σ_1^2, σ_2^2 known, or independence normality, σ_1^2, σ_2^2 known	Z = $\frac{(\overline{X}_1-\overline{X}_2)-\triangle_0}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$	$\mu_1 - \mu_2 > \triangle_0$ $\mu_1 - \mu_2 < \triangle_0$ $\mu_1 - \mu_2 \neq \triangle_0$	$Z \geq z_\alpha$ $Z \leq -z_\alpha$ $\mid Z \mid \geq z_{\alpha/2}$
$\mu_1 - \mu_2 = \triangle_0$	independence, n$_1$, n$_2$ large, σ_1^2, σ_2^2 unknown	Z = $\frac{(\overline{X}_1-\overline{X}_2)-\triangle_0}{\sqrt{S_1^2/n_1+S_2^2/n_2}}$	$\mu_1 - \mu_2 > \triangle_0$ $\mu_1 - \mu_2 < \triangle_0$ $\mu_1 - \mu_2 \neq \triangle_0$	$Z \geq z_\alpha$ $Z \leq -z_\alpha$ $\mid Z \mid \geq z_{\alpha/2}$
$\mu_1 - \mu_2 = \triangle_0$	independence n$_1$, n$_2$ small, normality, σ_1^2, σ_2^2 unknown, $\sigma_1^2 = \sigma_2^2$	T = $\frac{(\overline{X}_1-\overline{X}_2)-\triangle_0}{S_p\sqrt{1/n_1+1/n_2}}$ S$_p^2$ $= \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)}$	$\mu_1 - \mu_2 > \triangle_0$ $\mu_1 - \mu_2 < \triangle_0$ $\mu_1 - \mu_2 \neq \triangle_0$	$T \geq t_{n_1+n_2-2,\alpha}$ $T \leq -t_{n_1+n_2-2,\alpha}$ $\mid T \mid \geq t_{n_1+n_2-2,\alpha/2}$
$\mu_1 - \mu_2 = \triangle_0$	independence n$_1$, n$_2$ small, normality, σ_1^2, σ_2^2 unknown, $\sigma_1^2 \neq \sigma_2^2$	T = $\frac{(\overline{X}_1-\overline{X}_2)-\triangle_0}{\sqrt{S_1^2/n_1+S_2^2/n_2}}$ $\nu = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}$	$\mu_1 - \mu_2 > \triangle_0$ $\mu_1 - \mu_2 < \triangle_0$ $\mu_1 - \mu_2 \neq \triangle_0$	$T \geq t_{\nu,\alpha}$ $T \leq -t_{\nu,\alpha}$ $\mid T \mid \geq t_{\nu,\alpha/2}$
$\mu_D = \triangle_0$	normality, n pairs n small dependence	T = $\frac{\overline{D}-\triangle_0}{S_D/\sqrt{n}}$	$\mu_D > \triangle_0$ $\mu_D < \triangle_0$ $\mu_D \neq \triangle_0$	$T \geq t_{n-1,\alpha}$ $T \leq -t_{n-1,\alpha}$ $\mid T \mid \geq t_{n-1,\alpha/2}$
$\sigma_1^2 = \sigma_2^2$	normality, independence	F^*=S_1^2/S_2^2	$\sigma_1^2 > \sigma_2^2$ $\sigma_1^2 < \sigma_2^2$ $\sigma_1^2 \neq \sigma_2^2$	$F^* \geq F_{n_1,n_2,\alpha}$ $F^* \leq 1/F_{n_2,n_1,\alpha}$ $F^* \geq F_{n_1,n_2,\alpha/2}$ or $F^* \leq 1/F_{n_2,n_1,\alpha/2}$

Case Study: Medicine

Mathematical Sciences 405/605
Case Study: Medicine

We consider data from Rikkers [†] et al. (1978), who report results of a prospective randomized surgical trial allocating cirrhotic patients who had bled from varices to either a nonselective shunt (standard operation) or to a selective shunt (new operation). The dependent variable is the maximal rate of urea synthesis (MRUS), which is a quantitative test of liver function. Poor liver function is associated with a low MRUS value. MRUS was measured preoperatively and early postoperatively in eight selective shunt patients and thirteen nonselective shunt patients. The purposes of the study were to assess preoperatively the comparability of the selective and the nonselective groups and to longitudinally evaluate the change in liver function of the two groups. Table 1 reports MRUS values for each patient for the preoperative and postoperative periods.

Table 6: Pre and Post Maximal Rate of Urea Synthesis Level (mg urea N/hr/kg BW$^{3/4}$) and Sample Cell Means, by Group

Group	Subject	Pre	Post
Selective Shunt	1	51	48
(new operation)	2	35	55
	3	66	60
	4	40	35
	5	39	36
	6	46	43
	7	52	46
	8	42	54
Mean		$\overline{x}_{11}$= 46.375	$\overline{x}_{12}$ = 47.125
Nonselective Shunt	1	34	16
(standard operation)	2	40	36
	3	34	16
	4	36	18
	5	38	32
	6	32	14
	7	44	20
	8	50	43
	9	60	45
	10	63	67
	11	50	36
	12	42	34
	13	43	32
Mean		$\overline{x}_{21}$= 43.538	$\overline{x}_{22}$= 31.462

[†] Rikkers, Layton F., Rudman, Daniel, Galambos, John T., Fulenwider, J. Timothy, Milliken, William J., Kutner, Michael H., Smith, Robert B., Salam, Atef A., Sones, Peter J., and Warren, W. Dean (1978), ”A Randomized, Controlled Trial of the Distal Spenorenal Shunt,” *Annals of Surgery*, **188**, 271-282.

Case Study: Medicine
Review Questions

i) Based on the pre-operative results, can we say that there is a significant difference in the MRUS for the two groups of individuals selected to compare these two procedures ? In other words are the groups comparable ? Give a *p*-value to support your conclusions.

ii) Determine a 95% confidence interval for the difference in the MRUS for the two procedures' pre-operative results ?

iii) What if we consider the post-operative results, are they significantly different ? Give a *p*-value to support your conclusions.

iv) Determine a 99% confidence interval for the difference in the MRUS for the two procedures' post-operative results ?

v) Can you suggest a reason for considering a larger confidence level for the post-operative results ?

vi) If there is a significant difference in the post-operative results, what procedure would you suggest is the more beneficial to the patients ? Explain.

13 Goodness of Fit Tests

13.1 First a binomial experiment

The following general description has wide applications. Suppose we can describe a situation as a sequence of trials, each of which has two possible outcomes commonly referred to as ’success’ or ’failure’. If the probability of a success at each trial is constant then the number of successes has a binomial distribution. We can summarize this formally as follows:

1. There is a fixed number of trials (n).
2. There are two possible outcomes for each trial (’success’ or ‘failure’).
3. There is a constant probability of success (p). This implies that the outcomes of trials are independent.

Binomial Distribution Example: Over a long period of time it has been observed that a given rifleman can hit a target on a single trial with probability equal to 0.8. Suppose he fires four shots at the target.

1. What is the probability that he will hit the target exactly two times ?
2. What is the probability that he will hit the target at least two times ?
3. What is the probability that he will hit the target exactly four times ?

Assume that the trials are independent and that the probability *p* of hitting the target remains constant from trial to trial, *n* = 4 and *p* = .8. Let *x* denote the number of shots that hit the target. Then, for *x* = 0, 1, 2, 3, 4, we have

$$p(x) = \begin{pmatrix} 4 \\ x \end{pmatrix} (0.8)^x \, (0.2)^{4-x}$$

$$\begin{aligned} p(2) &= \begin{pmatrix} 4 \\ 2 \end{pmatrix} (0.8)^2 \, (0.2)^{4-2} \\ &= \frac{4!}{2!2!}(0.64)(0.04) \\ &= \frac{(4)(3)(2)(1)}{(2)(2)}(0.64)(0.04) \\ &= 0.1536. \end{aligned}$$

The probability is .1536 that he will hit the target exactly two times.

$$\begin{aligned} P(\text{at least two}) &= p(2) + p(3) + p(4) \\ &= 1 - p(0) - p(1) \\ &= 1 \begin{pmatrix} 4 \\ 0 \end{pmatrix} (0.8)^0 \, (0.2)^4 \begin{pmatrix} 4 \\ 1 \end{pmatrix} (0.8)^1 \, (0.2)^3 \\ &= 1 - 0.0016 - 0.0256 \\ &= .9728 \end{aligned}$$

The probability is 0.9728 that he will hit the target at least two times.

$$\begin{aligned} p(4) &= \begin{pmatrix} 4 \\ 4 \end{pmatrix} (0.8)^4 \, (0.2)^{4-4} \\ &= \frac{4!}{4!0!}(0.4096)(1) \\ &= 0.4096. \end{aligned}$$

The probability is .4096 that he will hit the target exactly four times.

Note that these probabilities would be incorrect if the rifleman could observe the location of each hit on the target and thereby adjust his aim. In that case, the trials would be dependent and *p* would likely increase from trial to trial.

Binomial Distribution Example 2: A student has no knowledge whatsoever of the material to be tested on a true-false examination, and so the student flips a fair coin in order to determine the response to each question. What is the probability that the student scores at least 60% on a ten-item examination?

Here the binomial variable, X, the number of correct responses, has *n* = 10, and *p* = *q* = $\frac{1}{2}$. We need

$$P(X \geq 6) = \sum_{x=6}^{10} \binom{10}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x}$$

Now we find that P$(X \geq 6)$ = $\frac{193}{512}$= 0.376953.

These calculations can easily be done with a pocket computer. If we want to investigate the probability that at least 60% of the questions are answered correctly as the number of items on the examination increases, then use of a computer algebra system is recommended for aiding in the calculation. Many computer algebra systems contain the binomial probability distribution as a defined probability distribution; for other systems. the probability distribution function may be entered directly. The following results can be found where *n* is the number of trials and *p* is the probability of at least 60% correct:

n	10	40	80	100
p	0.376953	0.134094	0.0464559	0.028444

Clearly, guessing is not a sensible strategy on a test with a large number of items.

13.2 A Multinomial Experiment

We can extend the binomial model to the case where instead of there being only 2 possible outcomes there are *k* possible outcomes, each with it's own probability of occurring.

1. The experiment consists of *n* identical trials.
2. The outcome of each trial falls into one of *k* classes or cells.
3. The probability that the outcome of a single trial will fall in a particular cell, say, cell *i*, is π_i (*i* = 1, 2,. . . , *k*) and remains the same from trial to trial. Note that 0 $<$ π_i $<$ 1 for all i, and π_1 + π_2 + π_3 + . . . + π_k = 1.
4. The trials are independent.
5. The experimenter is interested in $n_1, n_2, \ldots, n_k$, where n_i (*i* = 1,2,. . . ,*k*) is equal to the number of trials in which the outcome falls in cell *i*. Note that $n_1 + n_2 + \cdots + n_k$ = n.

Definition: (Multinomial random variable). Let an experiment consist of *n* independent and identical multinomial trials with parameters π_1, π_2, . . . , π_k. Let n_i denote the number of trials that result in outcome *i* for *i*=1,2,. . . ,*k*. The *k*-tuple (n_1, n_2,. . . , n_k) is called a multinomial random variable with parameters *n*, π_1, π_2,. . . ,π_k. The purpose of the chi-squared goodness of fit test is to test the null hypothesis that a given set of observations is drawn from, or "fits", a specified probability distribution. We consider two distinct situations:

1. The hypothesized distribution is completely specified before the sampling is done.
2. The hypothesized distribution is completely specified only after the sampling is done.

Case 1 is useful, but case 2 is particularly interesting because it provides an alternative to the usual procedures for testing normality, ie, normal probability plots, Shapiro-Wilks, and Lilliefors tests.

13.3 Goodness of fit

Generally a researcher would be interested in testing the following hypothesis:

$$H_0: \quad \text{data follows a specified model}$$

$$H_A: \quad \text{data does not follow the specified model}$$

or

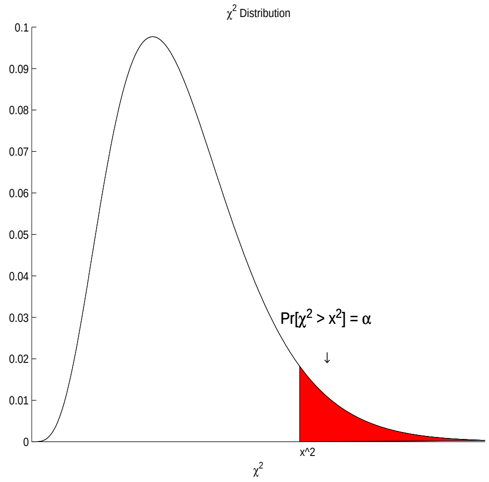
$$H_0: \quad \pi_1 = \pi_{10}, \pi_2 = \pi_{20}, \cdots \pi_k = \pi_{k0}, \quad i = 1 \ldots k$$

$$H_A: \quad \pi_i \neq \pi_{i0} \qquad\qquad\qquad \forall i = 1 \ldots k$$

Let (n_1, n_2, $\ldots$, n_k) be a multinomial random variable with parameters *n*, π_1, π_2, $\ldots$, π_k. Since a function of random variables is also a random variable, for large *n* the random variable, under the null hypothesis, that is if the null hypothesis is true,

$$X^2 \quad = \quad \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}$$

follows an approximate chi-squared distribution with *k* - 1 degrees of freedom, given that $n\pi_i$ >5 for all *i*. In practice, we would reject the null hypothesis in favor of the alternate hypothesis if $X^2 > \chi^2_{\nu,\alpha}$ as shown in the figure below, where ν = (k-1).



13.4 Exercises

I. Suppose that a response can fall in one of k = 5 categories with probabilities π_1, π_1, $\ldots$, π_5, respectively, and that n= 300 responses produced the following category counts: Conduct a test to

Category	1	2	3	4	5
Observed count	47	63	74	51	65

determine if there is a difference the proportion of counts that fall in each of the categories. Use both an hypothesis testing approach (with α = 0.01) and a significance testing approach ie, determine a p-value to make your decision.

II. Gregor Mendel was the first to describe a theory of genetics used is determining genotypes of offspring. The Mendelian theory states that the number of peas of a certain type falling into the classifications i) round and yellow, ii) wrinkled and yellow, iii) round and green, and iv) wrinkled and green should be in the ratio 9:3:3:1. Suppose that 100 such peas revealed 56, 19, 17, and 8 in the respective classes. Do these data disagree with the Mendelian theory ? Use both an hypothesis testing approach (with α = 0.05) and a significance testing approach ie, determine a p-value to make your decision.

III. Medical statistics show that deaths due to four major diseases - call them disease A, disease B, disease C, and disease D, account for 15, 21, 18, and 14 percent, respectively, of all non-accidental deaths. A study of the cases of 308 non-accidental deaths at a hospital gave the following counts of patients dying of disease A, disease B, disease C, and disease D:

Disease	**Number of Deaths**
A	43
B	76
C	85
D	21
Others	83

Do these data provide sufficient evidence to indicate that the proportion of people dying of diseases A, B, C, and D at this hospital differ from the proportions accumulated for the population at large ? Use both an hypothesis testing approach (with α = 0.025) and a significance testing approach ie, determine a p-value to make your decision.

IV. Computer systems crash for a number of different reasons, among them are software failures, hardware failures, operator errors, and system overloads. It is believed that 10% of all crashes are due to software failure, 5% to hardware failure, 25% to operator error, and 40% to system overloading. Over an extended period of time 150 computer crashes were monitored with the following results: 13 crashes due to software failures, 10 to hardware failures, 42 to operator errors, 65 to system overloading, and the rest to other causes. Do these data lead us to suspect the accuracy of the stated percentages ? Use both an hypothesis testing approach (with α = 0.05) and a significance testing approach ie, determine a p-value to make your decision.

V. Although white has long been the most popular car color, recent trends in fashion and home design have signaled the emergence of green as the new color of the 1990s. The growth in the popularity of green hues stems partially from an increased interest in the environment and increased feelings of uncertainty. According to an article in the Press-Enterprise ("White Cars Still Favored," 1993),"green symbolizes harmony and counteracts emotional stress." The article cites the top five colors and the percentage of the market share for four different classes of cars. These data are given below for the truck-van category:

		Medium/Dark			
Color	**White**	**Red**	**Green**	**Red**	**Black**
Percentage	29.72	11.00	9.24	9.08	9.01

In an attempt to verify the accuracy of these figures, we take a random sample of 250 trucks and vans and record their color. Suppose that the number of vehicles falling in each of the five categories above were 82, 22, 27, 21, and 20, respectively.

(a) Is there any category that is missing in the above classification? How many cars and trucks fell in that category?

(b) Is there sufficient evidence to indicate that the percentages of trucks and vans differ from those given above? Find the approximate *p*-value for the test.

13.5 r × c Tests for homogeneity - or “likeness”

The goodness of fit tests can be extended to those cases where there are two variables under study, and the main interest being that of determining if there is an homogeneity or “likeness” between two variables. In this case, one of the marginal totals is fixed. Assumptions:

1. Two variables - one of which is studied at *r* levels and the other at *c* levels.
2. One of the marginal totals is fixed by the researcher or resources

	Variable B						**Row Totals**
	n_{11}	n_{12}	n_{13}	$\cdots$	$n_{1(c-1)}$	n_{1c}	$n_{1.}$
	n_{21}	n_{21}	n_{23}	$\cdots$	$n_{2(c-1)}$	n_{2c}	$n_{2.}$
Variable A	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
	$n_{(r-1)1}$	$n_{(r-1)2}$	$n_{(r-1)3}$	$\cdots$	$n_{(r-1)(c-1)}$	$n_{(r-1)c}$	$n_{(r-1).}$
	n_{r1}	n_{r2}	n_{r3}	$\cdots$	$n_{r(c-1)}$	n_{rc}	$n_{r.}$
Column Totals	$n_{.1}$	$n_{.2}$	$n_{.3}$	$\cdots$	$n_{.(c-1)}$	$n_{.c}$	$n_{..}$

	Variable B						**Row Totals**
	π_{11}	π_{12}	π_{13}	$\cdots$	$\pi_{1(c-1)}$	π_{1c}	1
	π_{21}	π_{21}	π_{23}	$\cdots$	$\pi_{2(c-1)}$	π_{2c}	1
Variable A	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
	$\pi_{(r-1)1}$	$\pi_{(r-1)2}$	$\pi_{(r-1)3}$	$\cdots$	$\pi_{(r-1)(c-1)}$	$\pi_{(r-1)c}$	1
	π_{r1}	π_{r2}	π_{r3}	$\cdots$	$\pi_{r(c-1)}$	π_{rc}	1

Generally a researcher would be interested in testing the following hypothesis:

H_0 : The proportions are equal

H_A : At least one pair of proportions are unequal

or

$$H_0: \quad \pi_{1j} = \pi_{2j} = \cdots = \pi_{(r-1)c} = \pi_{rc}, \quad j = 1 \ldots c$$

$$H_A: \quad \pi_{ij} \neq \pi_{i'j} \qquad \text{for any } i' = 1 \ldots r \text{ and } j = 1 \ldots c$$

Under the null hypothesis the proportion are the same, the expected values for each of these cells are given by:

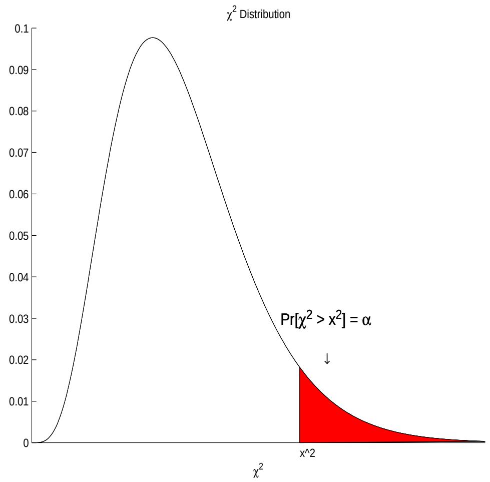
$$e_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Assumption: $e_{ij} > 5$. Under the null hypothesis, that is if the null hypothesis is true,

$$X^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(n_{ij} - e_{ij})^2}{e_{ij}}$$

follows an approximate chi-squared distribution with *k* - 1 degrees of freedom, given that $e_{ij} > 5$ for all *i* and *j*.

In practice, we would reject the null hypothesis in favor of the alternate hypothesis if $X^2 > \chi^2_{\nu,\alpha}$ as shown in the figure below, where ν = (r-1)×(c-1).



13.6 Exercises

I. A commercial nuclear power plant contains one or more nuclear power units, the term *nuclear power plant* usually refers to a single nuclear unit. The Oconee nuclear cluster would have three units and hence be considered three plants. In a study of the amount of failures in plants similar to the Oconee plant over the failure history of the plant (*time since first failure*), the number of failures reported for 7 plants were considered with the following results:

		Number of Failures in
Plant	**Failure History Period**	**Failure History Period**
A	12/82-12/88	35
B	1/78-2/86	16
C	5/76-7/86	18
D	2/83-1/87	9
E	8/83-10/86	13
G	11/78-6/84	8
H	4/84-2/91	11

Ignoring the failure history period, does there appear to be sufficient evidence that the number of failures is different across all plants ? Use both an hypothesis testing approach (with α = 0.10) and a significance testing approach ie, determine a *p*-value to make your decision. Conduct the test again, after removing plant A.

II. A study of the purchase decisions for three stock portfolio managers A, B, and C was conducted to compare the rates of stock purchases that resulted in profits over a time period that was less than or equal to 1 year. One hundred randomly selected purchases obtained for each of the managers gave the following results:

	Manager		
	A	**B**	**C**
Purchases that resulted in a profit	63	71	55
Purchases that resulted in no profit	37	29	45
Total	100	100	100

Do the data provide evidence of differences among the rates of successful purchases for the three managers? Use both an hypothesis testing approach (with α = 0.05) and a significance testing approach ie, determine a *p*-value to make your decision.

III. A study is conducted to test for independence between air quality and air temperature. These data are obtained from records on 200 randomly selected days over the last few years. Do these data indicate an association between these variables ? Use both an hypothesis testing approach (with α = 0.10) and a significance testing approach ie, determine a *p*-value to make your decision.

	Air quality		
Temperature	**Poor**	**Fair**	**Good**
Below average	1	3	24
Average	12	28	76
Above Average	12	14	30

IV. A new method for etching semiconductors is being studied. The quality of the etch is to be compared to that obtained using two older techniques. The results of the study are given in the table below. State the null hypothesis of homogeneity mathematically. Use both an hypothesis testing approach (with α = 0.10) and a significance testing approach ie, determine a *p*-value to make your decision.

	Quality				
Method	**Excellent**	**Good**	**Fair**	**Poor**	
High Pressure (old)	113	34	21	32	200
Reactive ion(old)	117	31	25	27	200
Magnetron(new)	130	40	20	10	200
					600

V. Are baby-boomers more likely to increase their investing now that they are reaching middle age? A poll was conducted by Hal Riney & Partners (Los Angeles Times, June 11, 1990). in which 400 investors were classified according to their age group and their likely investment pattern over the next 5 years versus the last 5 years. The data are shown below. Notice that there were 200 investors included from each age group, ie., a fixed marginal. Do these data provide sufficient evidence to conclude that the investing patterns of the baby-boomers age group differs from that of that of the older age group ? Use both an hypothesis testing approach (with α = 0.01) and a significance testing approach ie, determine a *p*-value to make your decision.

Age Group	**More**	**Less**	**Same**
35-54	90	18	92
55+	40	60	100

13.7 Tests for Independence

In the case of testing for independence

$$H_0: \quad \text{Independence}$$

$$H_A: \quad \text{Dependence}$$

or

$$H_0: \quad \pi_{ij} = \pi_{i.}\pi_{.j}, \quad \forall i \text{ and } j$$

$$H_A: \quad \pi_{ij} \neq \pi_{i.}\pi_{.j} \quad \text{for some } i \text{ and } j$$

	Variable B						**Row Totals**
	n_{11}	n_{12}	n_{13}	$\cdots$	$n_{1(c-1)}$	n_{1c}	$n_{1.}$
	n_{21}	n_{21}	n_{23}	$\cdots$	$n_{2(c-1)}$	n_{2c}	$n_{2.}$
Variable A	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$	$\vdots$
	$n_{(r-1)1}$	$n_{(r-1)2}$	$n_{(r-1)3}$	$\cdots$	$n_{(r-1)(c-1)}$	$n_{(r-1)c}$	$n_{(r-1).}$
	n_{r1}	n_{r2}	n_{r3}	$\cdots$	$n_{r(c-1)}$	n_{rc}	$n_{r.}$
Column Totals	$n_{.1}$	$n_{.2}$	$n_{.3}$	$\cdots$	$n_{.(c-1)}$	$n_{.c}$	$n_{..}$

	Variable B						**Row Totals**
	π_{11}	π_{12}	π_{13}	$\cdots$	$\pi_{1(c-1)}$	π_{1c}	$\pi_{1.}$
	π_{21}	π_{21}	π_{23}	$\cdots$	$\pi_{2(c-1)}$	π_{2c}	$\pi_{2.}$
Variable A	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$	$\vdots$
	$\pi_{(r-1)1}$	$\pi_{(r-1)2}$	$\pi_{(r-1)3}$	$\cdots$	$\pi_{(r-1)(c-1)}$	$\pi_{(r-1)c}$	$\pi_{(r-1).}$
	π_{r1}	π_{r2}	π_{r3}	$\cdots$	$\pi_{r(c-1)}$	π_{rc}	$\pi_{r.}$
Column Totals	$\pi_{.1}$	$\pi_{.2}$	$\pi_{.3}$	$\cdots$	$\pi_{.(c-1)}$	$\pi_{.c}$	1

Under the null hypothesis, the expected values for each of these cells are given by:

$$e_{ij} = \frac{n_{i.}n_{.j}}{n}$$

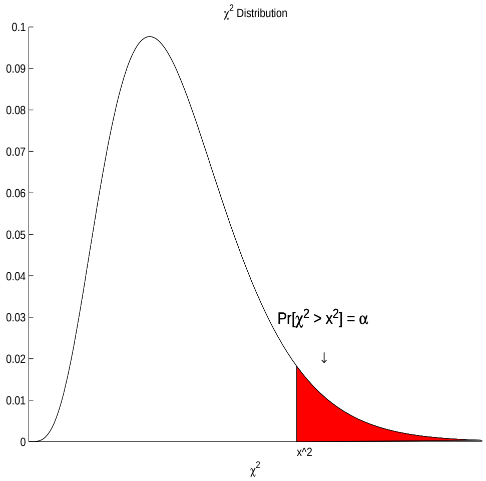
Assumption: $e_{ij} > 5$.

Under the null hypothesis, that is if the null hypothesis is true,

$$X^2 = \frac{\sum_{i=1}^{r} \sum_{j=1}^{c} (n_{ij} - e_{ij})^2}{e_{ij}}$$

follows an approximate chi-squared distribution with *k* - 1 degrees of freedom, given that $e_{ij} > 5$ for all *i* and *j*.

In practice, we would reject the null hypothesis in favor of the alternate hypothesis if $X^2 > \chi^2_{\nu,\alpha}$ as shown in the figure below, where ν = (r-1)×(c-1).



13.8 Exercises

I. A cancer researcher performs what is called a prospective be selecting a large group of individuals at random and following their progress for a long period of time. At the end of the study period each individual is classified according to whether or not lung cancer was present and according to whether the individual has been exposed to an identifiable source of airborne asbestos. The result of this classification yielded the following table:

		Exposure Status		**Totals**
		Exposed	**Unexposed**	
Cancer	**Yes**	10	40	50
	No	490	4460	4950
Totals		500	4500	5000

Do these data suggest an association of exposure to airborne asbestos and cancer development ? Use both an hypothesis testing approach (with α = 0.10) and a significance testing approach ie, determine a *p*-value to make your decision.

II. A problem that sometimes occurs during surgical operations is the occurrence of infections during blood transfusions. An experiment was conducted to determine whether the injection of antibodies reduced the probability of infection. An examination of the records of 138 patients produced the data shown in the accompanying table. Do the data provide sufficient evidence to indicate that injections of antibodies affect the likelihood of transfusional infections? Use both an hypothesis testing approach (with α = 0.01) and a significance testing approach ie, determine a *p*-value to make your decision.

	Infection	**No Infection**
Antibody	4	78
No antibody	11	45

III. A recent study claims that an increasing proportion of engineering firms are purchasing liability insurance. This claim is based on a survey of 753 engineering firms. The status of each firm is recorded for the current and for the previous year. The data upon which the claim is based are shown in the table below. Do the data support the claim? Explain, based on the *p*-value of McNemar's test.

	This year		
Last year	**Insured**	**Uninsured**	
Insured	650	5	655
Uninsured	28	70	98
	678	75	753

IV. The following table shows the categorization of 204 men awaiting bypass heart surgery according to the relative degree of each man's coronary artery obstruction and according to his perceived level of discomfort due to angina pectoris (Jenkins et al., 1983). Do the data present sufficient evidence to indicate that the level of angina is dependent on the level of coroners artery obstruction? The authors report the *p*-value for a chi-square test to be $p = 0.01$.

(a) Compute the value of χ^2 for the data.

(b) Find the *p*-value for the test and compare with the authors' value $p = 0.01$.

(c) What conclusions would you reach based on your analysis ?

	Arteries Obstructed 75% or More			
Level of Angina	**0 or 1**	**2**	**3 to 6**	**Total**
None	3	21	20	44
Mild	2	12	9	23
Moderate	26	20	31	77
Moderate/severe	13	10	18	41
Severe	7	5	7	19

Case Study: Marketing the Library

Case Study: Marketing the Library Can a Marketing Approach Improve Library Services?

Carol Day and Del Loewenthal (1992) studied the responses of young adults in their evaluation of library services. Of the n = 200 young adults involved in the study, n_1 = 152 were students, and n_2 = 48 were non-students. The following table presents the number of favorable responses for each group to seven questions in which the atmosphere, staff, and the design of the library were examined. The entry in the last column labeled P(χ^2) is the *p*-value for testing the hypothesis of no difference in the proportion of students and nonstudents answering each questionfavorably. Hence. each question gives rise to a 2 $\times$ 2 contingency table.

Table 7: Favoroble Responses to Attitude Questions for Students and Nonstudents

		% Student		% Nonstudent		
Question	Question	Favorable	n_1 = 152	Favorable	n_2 = 48	$P\left(\chi^2\right)$
3	Libraries are friendly	79.6	121	56.2	27	$<$ .01
4	Libraries are dull	77	117	58.3	28	$<$ .05
5	Library staff are helpful	91.4	139	87.5	42	N. S.
6	Library staff are less helpful to teenagers	60.5	92	45.8	22	$<$ .01
7	Libraries are so quiet they feel uncomfortable	75.6	115	52.05	25	$<$ .01
11	Libraries should be more brightly decorated	29	44	18.8	9	N.S.
13	Libraries are badly signposted	45.4	69	43.8	21	N. S.

1. Perform a test of homogeneity for each question and verify the reported *p*-value of the rest.
2. Questions 3, 4, and 7, are concerned with the atmosphere of the library; questions 5 and 6 are concerned with the library staff; and questions 11 and 13 are concerned with the library design. How would you summarize the results of your analyses regarding the seven questions concerning the image of the library ?
3. With the information given. is it possible to do any further testing concerning the proportion favorable versus unfavorable responses for two or more questions simultaneously?

14 Important Formulas and Concepts in Regression Analysis(Chapt. 4 LSM)

14.1 Strategy

Data **investigation**, model **specification**, parameter **estimation**, model **assessment**, variable **selection**.

14.2 Simple Linear regression and the principle of Least Squares

Linear model

$$y_i \quad = \quad \beta_0 + \beta_1 x_i + \epsilon_i$$

where (x_i, y_i) is the i^{th} data point and

1. x_i is a realization of the "independent" or predictor random variable X.
2. y_i is a realization of the "dependent" or prediction random variable Y.
3. β_0 is the y-intercept parameter, generally unknown.
4. β_1 is the slope or *rate of change*, generally unknown.
5. $\epsilon_i \; i = 1, 2, \ldots, n$ are unobservable "noise" or "error" random variables with mean zero and constant variance σ^2.
6. β_0 + β_1 x is called the true unknown "regression function" of y on X. ie, E[Y]=β_0 + β_1x.

One can fit any number of models using least squares

1. linear models

$$y_i \quad = \quad \beta_0 + \beta_1 x_i$$

2. polynomials

$$y_i \quad = \quad \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p$$

3. Other functions that are linear in the parameters to be estimated

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^{\frac{3}{2}} + \frac{\beta_3}{ln \mid x_{1i} \mid} \\ y_i &= \beta_0 + \beta_1 e^{-2x_{1i}} + \beta_2 sin\left(x_{1i} x_{2i}\right) + \beta_3 x_{1i} ln\left(x_{2i}^2\right) tan\left(x_{3i}\right) \end{aligned}$$

4. nonlinear models that have been "linearized".

Suppose we let

$$\begin{aligned} \epsilon_i &= y_i - (\beta_0 + \beta_1 x_i) \\ \text{then} & \\ \epsilon_i^2 &= [y_i - (\beta_0 + \beta_1 x_i)]^2 . \end{aligned}$$

Let Q be the sum of these squared differences:

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 .$$

We wish to find estimates of β_0 and β_1, call them $\widehat{\beta_0}$ and $\widehat{\beta_1}$, that would minimize Q.

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ and & \\ \frac{\partial Q}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) . \end{aligned}$$

After taking derivatives and setting equal to zero, and passing through the summation operator, we can solve the following equations:

$$\begin{aligned} \sum_{i=1}^{n} y_i &= n\widehat{\beta_0} + \widehat{\beta_1} \sum_{i=1}^{n} x_i \\ \text{and} & \\ \sum_{i=1}^{n} x_i y_i &= \widehat{\beta_0} \sum_{i=1}^{n} x_i + \widehat{\beta_1} \sum_{i=1}^{n} x_i^2 . \end{aligned}$$

Solving first for $\widehat{\beta_1}$ and them for $\widehat{\beta_0}$ we have the following estimates:

$$\begin{aligned} \widehat{\beta_1} &= \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right) \left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ \text{and} & \\ \widehat{\beta_0} &= \frac{\sum_{i=1}^{n} y_i - \widehat{\beta_1} \sum_{i=1}^{n} x_i}{n} = \overline{y} - \widehat{\beta_1} \overline{x} . \end{aligned}$$

We can simplify the expression for β_1 to:

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

where

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n} (x_i - \overline{x}) (y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right) \left(\sum_{i=1}^{n} y_i\right)}{n} = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y} \\ S_{xx} &= \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \\ \text{and also} & \\ S_{yy} &= \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 . \end{aligned}$$

14.3 The fitted regression line(Chapt. 4 LSM)

The fitted regression line is given by after finding estimates for β_0 and β_1

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i.$$

Define a residual as:

$$e_i = \widehat{\epsilon_i} = y_i - \widehat{y_i}$$

that is the actual (original) value of the response minus the fitted (predicted) value. After getting $\widehat{\beta_0}$ and $\widehat{\beta_1}$ we can substitute these values back into Q to find this minimum.

$$\widehat{Q} = \sum_{i=1}^{n} \left[y_i - \left(\widehat{\beta_0} + \widehat{\beta_1} x_i \right) \right]^2 = \sum_{i=1}^{n} [y_i - \widehat{y_i}]^2 = \sum_{i=1}^{n} \widehat{\epsilon_i}^2.$$

This sum which can be re-written as

$$S_{yy} - \widehat{\beta_1} S_{xy} = SS_E$$

is called the residual or error sum of squares (SS$_E$). An unbiased estimate of σ^2 is given by

$$\begin{aligned} \widehat{\sigma^2} &= \frac{SS_E}{n-2} = MS_E \\ &= S^2 \end{aligned}$$

which is referred to as the mean square error (or literally the mean of the squared errors).

15 Parsimony in Modeling(Chapt. 4 LSM)

The simplest useful model we can fit to data is a constant function. With this model, the dependent variable *y* does not change as the independent variable *x* changes. While different constants could be chosen (for example, any measure of center), the mean $\overline{y}$ is the most commonly chosen constant. If all the data actually has the same y coordinate, then the data has no variation and a constant function explains the data completely. If, however, the y values are not constant, then, clearly, there is some variation in the data about the mean. One way to measure this variation is called the total sum of squares, or SS_{Tot}, which is defined

$$SS_{Tot} = \sum_{i=1}^{N} (y_i - \overline{y})$$

Notice that SS_{Tot}/(n-1) can be called the sample variance, reinforcing the idea that SS_{Tot} is a measure of the variance in the data.

If we are building a model to explain the variation seen in the data, we need to use a model more complicated than a constant function; we call it $y = f(x)$. If the data lie on this function exactly, then it explains all the variation. Usually, however, there is some noise to the data causing it to lie about a model function. Sometimes this noise is due to randomness. Sometimes there is curve which is a better model. Using a line, for instance, to model perfectly parabolic data, results in data points not lying on the model curve even in the absence of any random effects. In any case, there are two types of variations in the data: variation explained by the model and variation not explained by the model. The deviations between the predicted and actual *y* values $(y_i - \widehat{y})$ are called *residuals*. The variation not explained by the model is called the residual sum of squares or SS_{Res}. This is formally defined as

$$SS_{Res} = \sum_{i=1}^{N} (y_i - \widehat{y})$$

One common form of data fitting is called regression. Because of this, the variation explained by the model (regression function) is called the regression sum of squares or SS_{Reg}. This is defined by

$$SS_{Reg} = \sum_{i=1}^{N} (\widehat{y}_i - \overline{y})$$

Again this is the deviation from the mean explained by the model.

15.1 Understanding Variation(Chapt. 5 LSM)

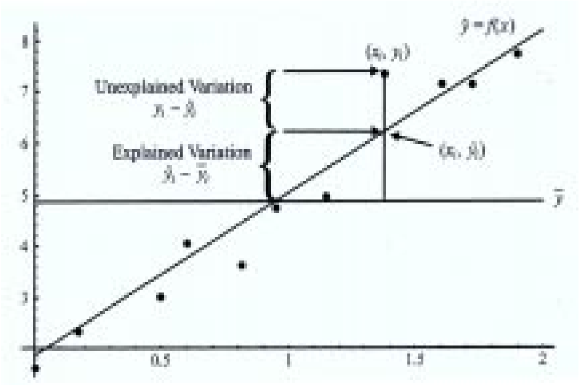


Figure 6: Total Variation: Explained and Unexplained

It is easy to show algebraically that

$$SS_{Tot} = SS_{Reg} + SS_{Res}$$

Intuitively this means that the total variation in the data is the sum of the variation explained by the model as well as the variation not explained by the model. Thus the ratio

$$\frac{SS_{Reg}}{SS_{Tot}}$$

is the fraction (or percentage if multiplied by 100%) of the variation in the data explained by the model. This ratio is called R^2. Thus

$$R^2 = \frac{SS_{Reg}}{SS_{Tot}}$$

Some books call R^2 the *coefficient of determination*. If the data is almost completely random, then almost none of the variance in the data is explained by the model. In this case, $SS_{Reg} \approx 0$ and hence $R^2 \approx 0$. On the other hand if the data has almost no noise and lies very nearly on the model or regression curve, then $SS_{Res} \approx 0$, so $SS_{Tot} \approx SS_{Reg}$ and hence $R^2 \approx 1$.

The observant reader may notice that there are some similarities between this discussion and the discussion of correlation. Indeed the notation R^2 comes from the fact that if we fit a line to data which minimizes the unexplained variance, the statistic R^2 is exactly the correlation *r* squared. For this reason, in simple linear regression, R^2 is denoted by r^2.

15.2 Correlation and Coefficient of Determination(Chapt. 5 LSM)

We can define the Pearson product moment correlation coefficient (*r*) as an estimate of the true population coefficient ρ.

$$
\begin{aligned}
r &= \frac{\left[\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}\right]}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}} \\
&= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \\
&= \widehat{\beta_1}\sqrt{\frac{S_{xx}}{S_{yy}}}.
\end{aligned}
$$

The correlation is a measure of the *strength and direction* of the **linear** relationship between *x* and *y*. Note that it does not imply that *x* causes *y* or influences *y*. It just measures the strength of the relationship. The *coefficient of determination*

$$
\begin{aligned}
R^2 &= \frac{\left[\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}\right]^2}{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)} \\
&= \frac{SS_R}{S_{yy}} \\
&= 1 - \frac{SS_E}{S_{yy}}
\end{aligned}
$$

can be defined as the proportion of variability in the responses *y* that can be explained by or accounted for by the predictor X.

15.3 Inferences on the population correlation coefficient ρ(Chapt. 5 LSM)

Null Hypothesis	Assumptions	Test Statistic	Alternative Hypothesis	Rejection Region
ρ = 0	Error assumptions (n $>$ 2)		$\rho > 0$	$T \geq t_{n-2,\alpha}$
		T = $\frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$ = $\frac{\widehat{\beta}_1 - 0}{\sqrt{\frac{MS_E}{S_{xx}}}}$	$\rho \neq 0$	$\mid T \mid \geq t_{n-2,\alpha/2}$
			$\rho < 0$	$T \leq -t_{n-2,\alpha}$

When X and Y have a bivariate normal distribution:

Null Hypothesis	Assumptions	Test Statistic	Alternative Hypothesis	Rejection Region
$\rho = \rho_0$	Error assumpt. (n $>$ 2)		$\rho > \rho_0$	$Z \geq z_\alpha$
		Z = $\frac{\sqrt{n-3}}{2} ln \left[\frac{(1+R)(1-\rho_0)}{(1-R)(1+\rho_0)}\right]$	$\rho \neq \rho_0$	$\mid Z \mid \geq z_{\alpha/2}$
			$\rho < \rho_0$	$Z \leq -z_\alpha$

15.4 Is the regression significant ?(Chapt. 4 LSM)

One of the first question we need to answer is “Is the regression significant”. In the linear case we are basically asking do we have a linear relationship. Is there statistical evidence to conclude that the slope of the true regression line is different from zero ?

Null Hypothesis	Assumptions	Test Statistic	Alternative Hypothesis	Rejection Region
β_1 = 0	Error assumptions (n $>$ 2)	T = $\frac{\widehat{\beta}_1 - 0}{S_{\widehat{\beta}_1}}$	$\beta_1 > 0$	$T \geq t_{n-2,\alpha}$
		= $\frac{\widehat{\beta}_1 - 0}{s\sqrt{\frac{1}{S_{xx}}}}$	$\beta_1 \neq 0$	$\mid T \mid \geq t_{n-2,\alpha/2}$
			$\beta_1 < 0$	$T \leq -t_{n-2,\alpha}$

If the null hypothesis is rejected then the regression is generally considered “significant”

Table 8: Analysis of Variance for Regression

Source	df	Sum of Squares	Mean Square	F-test
Regression	*1*	SS_R=$\hat{\beta}_1 S_{xy}$	MS_R=$\frac{\hat{\beta}_1 S_{xy}}{1}$	$\frac{MS_R}{MS_E} = F^* \sim F_{1,n-2}$
Error	*n-2*	SS_E=$S_{yy} - \hat{\beta}_1 S_{xy}$	MS_E =$\frac{SS_E}{n-2}$	
Total	n-1	S_{yy}		

Of lesser significance is the test for the intercept term β_0 = 0. It should be noted that the linear model can be re-written so that the intercept term is zero.

Null Hypothesis	Assumptions	Test Statistic	Alternative Hypothesis	Rejection Region
β_0 = 0	Error assumptions (n $>$ 2)	T = $\frac{\widehat{\beta_0}-0}{S_{\widehat{\beta_0}}}$ = $\frac{\widehat{\beta_0}-0}{s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}}}$	$\beta_0 > 0$ $\beta_0 \neq 0$ $\beta_0 < 0$	$T \geq t_{n-2,\alpha}$ $\mid T \mid \geq t_{n-2,\alpha/2}$ $T \leq -t_{n-2,\alpha}$

15.5 Distributional properties of $\widehat{\beta_0}$ and $\widehat{\beta_1}$(Chapt. 5 LSM)

The method of least squares along with the Gauss-Markov theorem that the estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the "best linear unbiased estimators" for β_0 and β_1. But what is the distribution of these estimates? Without going into the derivation of the variances of these estimates and following the assumptions made on the error and response random variables, we have that

$$\widehat{\beta_0} \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}\right).$$

Similarly, it can be shown for slope parameter estimate, $\widehat{\beta_1}$,

$$\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

15.6 Inferences on the parameters $\widehat{\beta_0}$ and $\widehat{\beta_1}$(Chapt. 5 LSM)

$$\begin{aligned} \frac{\widehat{\beta_0} - \beta_0}{S_{\widehat{\beta_0}}} &\sim t_{n-2} \\ &\text{and} \\ \frac{\widehat{\beta_1} - \beta_1}{S_{\widehat{\beta_1}}} &\sim t_{n-2}. \end{aligned}$$

15.6.1 Confidence Interval for β_0 and β_1.(Chapt. 5 LSM)

Using the above a (1-α) (100 %) confidence interval for β_0 is given by

$$\begin{aligned} \widehat{\beta_0} &\pm t_{n-2,\alpha/2} S_{\widehat{\beta_0}} \\ &\pm t_{n-2,\alpha/2} s \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}}. \end{aligned}$$

Using the above a (1-α) (100 %) confidence interval for β_1 is given by

$$\begin{aligned} \widehat{\beta_1} &\pm t_{n-2,\alpha/2} S_{\widehat{\beta_1}} \\ &\pm t_{n-2,\alpha/2} s \sqrt{\frac{1}{S_{xx}}}. \end{aligned}$$

15.7 Prediction and Estimation(Chapt. 5 LSM)

15.7.1 Inferences about the estimated regression, E(y) (μ_0)

Suppose we wish to use our regression function to find the mean response, E(y), your text uses (μ_0), for a single measurement at a point x_0. Then we can use the fitted regression function as an estimate of the mean response,

$$\widehat{E(y)} = \widehat{y}_{x_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

to get this value. Note that it is important that this new point x_0 be within the range of the current x values. If we wish to get a confidence interval for the mean reponse, E(y), at the "new" value x_0 we would have:

$$\widehat{y}_{x_0} \pm t_{\alpha/2,n-2}\sqrt{MS_E\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)}.$$

Predicting a response at a point x$_0$, y_0

Suppose we wish to use our regression to "predict" a response for some future point, say x_0, of the current values then the best predictor is obviously,

$$\widehat{y}_{x_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

If we wish to get a confidence interval for the future value we would have:

$$\widehat{y}_{x_p} \pm t_{\alpha/2,n-2}\sqrt{MS_E\left(1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{S_{xx}}\right)}.$$

Predicting a mean response of m future measurements at a point x$_p$

Suppose we wish to use our regression to "predict" m responses at some future point, say x_0. Then best estimate of the the mean of the responses $\overline{y}$ of the current values then the best predictor is obviously,

$$\widehat{y}_{x_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

If we wish to get a confidence interval for the future value we would have:

$$\widehat{y}_{x_0} \pm t_{\alpha/2,n-2}\sqrt{MS_E\left(\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)}.$$

15.8 Introductory Residual Analysis(Chapt. 6 LSM)

Residuals

$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ &= y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \end{aligned}$$

Residual Mean

$$\begin{aligned} \overline{e}_i &= \frac{\sum\limits_{i=1}^{n} e_i}{n} \\ &= 0 \end{aligned}$$

Residual Variance

$$\begin{aligned} MS_E &= \frac{\sum\limits_{i=1}^{n} \left(e_i - \overline{e}_i\right)^2}{n-2} \\ &= \frac{SS_E}{n-2} \\ &= \widehat{\sigma}^2 \end{aligned}$$

Standardized Residuals(semi-studentized)

$$d_i = \frac{e_i}{\sqrt{MS_E}}$$

Studentized Residuals

$$d_i^* = \frac{e_i}{\sqrt{MS_E\left[1 - \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}\right)\right]}}$$

15.9 Departures from Model to Be Studied by Residuals(Chapt. 6 LSM)

We shall consider the use of residuals for examining six important types of departures from the simple linear regression model with normal errors:

1. The regression function is not linear.
2. The error terms do not have constant variance.
3. The error terms are not independent.
4. The model fits all but one or a few outlier observations.
5. The error terms are not normally distributed.
6. One or several important predictor variables have been omitted from the model.

15.10 Diagnostics for Residuals

We take up now some informal diagnostic plots of residuals to provide information on whether any of the six types of departures from the simple linear regression model just mentioned are present. The following plots of residuals (or semistudentized residuals) will be utilized here for this purpose:

1. Plot of residuals against predictor variable.
2. Plot of absolute or squared residuals against predictor variable.
3. Plot of residuals against fitted values.
4. Plot of residuals against time or other sequence.
5. Plots of residuals against omitted predictor variables.
6. Box plot of residuals.
7. Normal probability plot of residuals.

15.11 Comparable Normal probability plots(Chapt. 6 LSM)

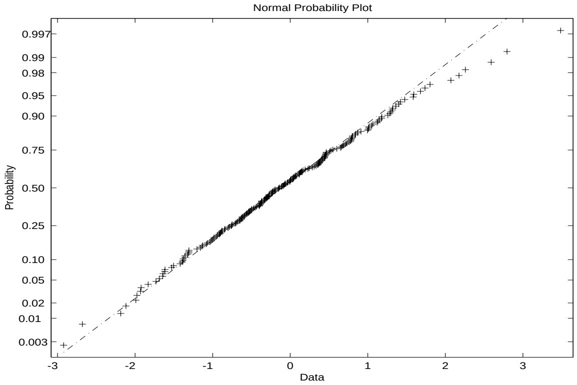


Figure 7: Ideal normal plot - signifying data is normally distributed

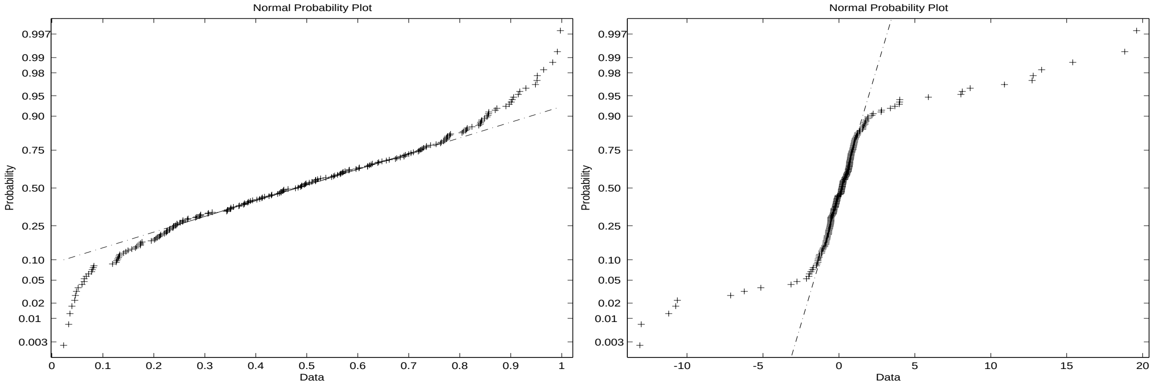


Figure 8: Figure on left - heavy tailed distribution, figure on right light - tailed

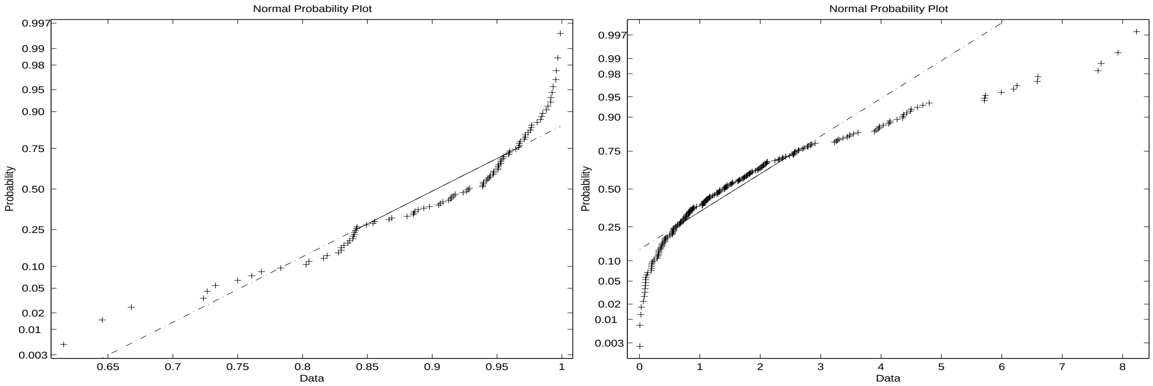


Figure 9: Figure on left - positive skew (skewed right) , figure on right negative skew (skewed left)

16 Transformations to Linearity(Chapt. 6 (13) LSM)

In certain situations, a transformation on X or Y (or both) might "straighten out" the plot so that a linear relationship would be appropriate for the transformed variables. Polynomial regression may also be employed.

But first consider the following transformations:

Relationship of σ^2 to E(y)	Transformation
$\sigma^2 \propto$ constant	y' = y (no transformation)
$\sigma^2 \propto E(y)$	y' = $\sqrt{y}$ (square root; Poisson data)
$\sigma^2 \propto E(y)[1 - E(y)]$	y' = sin^{-1} $(\sqrt{y})$ (arcsin; binomial proportions 0 $\leq$ y_i $\leq$ 1
$\sigma^2 \propto [E(y)]^2$	y' = ln (y) (log)
$\sigma^2 \propto [E(y)]^3$	y' = $y^{1/2}$ (reciprocal square root)
$\sigma^2 \propto [E(y)]^4$	y' = y^{-1} (reciprocal)

Table 9: Linearizable functions and corresponding linear form

	Linearizable	**Linear**	
Figure	**Function**	**Transformation**	**Form**
3.13a,b	$y = \beta_0 x^{\beta_1}$	$y' = log\, y\, ,\; x' = log x$	$y' = log \beta_0 + \beta_1 x'$
3.13c,d	$y = \beta_0 e^{\beta_1 x}$	$y' = lny$	$y' = ln\beta_0 + \beta_1 x$
3.13e,f	$y = \beta_0 + \beta_1 log\, x$	$x' = logx$	$y' = \beta_0 + \beta_1 x'$
313g,h	$y = \frac{x}{\beta_0 x - \beta_1}$	$y\prime = \frac{1}{y},\; x\prime = \frac{1}{x}$	$y = \beta_0 - \beta_1 x'$

16.1 Lack of Fit and Introduction to Polynomial Regression(Chapt. 8 LSM)

Assume n_i observations at each x_i, $i = 1, 2, \ldots, k$ and $n_i > 1$ at for at least one value of *i*. Let n = $\sum_{i=1}^{k} n_i$ denote the total number of observations. Then

$$SS_{yy} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2$$

and

$$SSR = \sum_{i=1}^{k} n_i(\hat{Y}_i - \overline{Y}_i)^2$$

where SSE = SS_{yy} - SSR thus

$$SS_{PE} = \sum_{i=1}^{k} S_i^2$$

where

$$S_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$$

$SS_{LF} = SSE - SS_{PE}.$

Table 10: Analysis of Variance for Regression Including Partition for Lack of Fit

Source	*df*	*Sum of Squares*	*Mean Square*	*F-test*
Regression	*1*	*SSR=*$\hat{\beta}_1 S_{xy}$	*MSR=*$\frac{\hat{\beta}_1 S_{xy}}{1}$	$\frac{MSR}{MSE} - - - - - F_{1,n-2}$
Error	*n-2*	*SSE=*$S_{yy} - \hat{\beta}_1 S_{xy}$	*MSE=*$\frac{SSE}{n-2}$	
lack of fit	*k-2*	SS_{LF}	*MSLF=*$\frac{SS_{LF}}{k-2}$	F^*=$\frac{MSLF}{MSPE} - - - - - F_{k-2,n-k}$
pure error	*n-k*	SS_{PE}	*MSPE=*$\frac{SS_{PE}}{n-k}$	
Total	*n-1*	SS_{yy}		

Reject H_0: Regression is linear, if $F^* > F_{\alpha,k-2,n-k}$. If the null hypothesis is rejected assumption of a linear fit is inappropriate. In this situation, a transformation on X or Y (or both) might “straighten out” the plot so that a linear relationship would be appropriate for the transformed variables. Polynomial regression may also be employed.

The following are the breaking strengths of six bolts at each of five different diameters. Also see exercise

Table 11: Example of Testing for Lack of Fit

Diameter				
.1	.2	.3	.4	.5
1.62	1.71	1.86	2.14	2.45
1.73	1.78	1.86	2.07	2.42
1.70	1.79	1.90	2.11	2.33
1.66	1.86	1.95	2.18	2.36
1.74	1.70	1.96	2.17	2.38
1.72	1.84	2.00	2.07	2.31

4.10 from text.

17 Polynomial models(Chapt. 8 LSM)

We begin with a simple quadratic model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i.$$

Using the technique of least squares:

$$Q = \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 \right) \right]^2 .$$

We wish to find estimates of β_0, β_1 , and β_2, call them $\widehat{\beta_0}$, $\widehat{\beta_1}$ and $\widehat{\beta_2}$, that would minimize Q.

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2\sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 \right) \\ \frac{\partial Q}{\partial \beta_1} &= -2\sum_{i=1}^{n} x_i \left(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 \right) \\ \text{and} & \\ \frac{\partial Q}{\partial \beta_2} &= -2\sum_{i=1}^{n} x_i^2 \left(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2 \right) . \end{aligned}$$

After taking derivatives and setting equal to zero, and passing through the summation operator, we can solve the following equations:

$$\begin{aligned} \sum_{i=1}^{n} y_i &= n\widehat{\beta_0} + \widehat{\beta_1} \sum_{i=1}^{n} x_i + \widehat{\beta_2} \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i y_i &= \widehat{\beta_0} \sum_{i=1}^{n} x_i + \widehat{\beta_1} \sum_{i=1}^{n} x_i^2 + \widehat{\beta_2} \sum_{i=1}^{n} x_i^3 \\ \text{and} & \\ \sum_{i=1}^{n} x_i^2 y_i &= \widehat{\beta_0} \sum_{i=1}^{n} x_i^2 + \widehat{\beta_1} \sum_{i=1}^{n} x_i^3 + \widehat{\beta_2} \sum_{i=1}^{n} x_i^4 . \end{aligned}$$

Case Study: Climatology

Mathematical Sciences 405/605
Case Study: Climatology

Table 1 gives the values of normal average January minimum temperature (y) in degree fahrenheit, latitude (x_1), and longitude (x_2) for 56 cities in the United States. It may be of interest to investigate how January temperature relates to latitude and longitude. For this purpose, it is reasonable to assume an approximate linear relationship between January temperature and latitude. In addition, some studies have found that, after adjusting for latitude, a cubic polynomial in longitude accurately predicts normal average January temperatures in the contiguous United States.

Table 12: Normal Average January Minimum Temperature (y), Latitude (x_1), and Longitude (x_2) for 56 Locations in the Contiguous United States

Location	y	x_1	x_2	Location	y	x_1	x_2
Mobile, AL	44	31.2	88.5	Omaha, NB	13	41.9	96.1
Montgomery, AL	38	32.9	86.8	Concord, NH	11	43.5	71.9
Phoenix, AZ	35	33.6	112.5	Atlantic City, NJ	27	39.8	75.3
Little Rock, AR	31	35.4	92.8	Albuquerque, NM	24	35.1	106.7
Los Angeles, CA	47	34.3	118.7	Albany, NY	14	42.6	73.7
San Francisco, CA	42	38.4	123.0	New York, NY	27	40.8	74.6
Denver, CO	15	40.7	105.3	Charlotte, NC	34	35.9	81.5
New Haven, CT	22	41.7	73.4	Raleigh, NC	31	36.4	78.9
Wilmington, DE	26	40.5	76.3	Bismarck, ND	0	47.1	101.0
Washington, DC	30	39.7	77.5	Cincinnati, OH	26	39.2	85.0
Jacksonville, FL	45	31.0	82.3	Cleveland, OH	21	42.3	82.5
Key West, FL	65	25.0	82.0	Oklahoma City, OK	28	35.9	97.5
Miami, FL	58	26.3	80.7	Portland, OR	33	45.6	123.2
Atlanta, GA	37	33.9	85.0	Harrisburg, PA	24	40.9	77.8
Boise, ID	22	43.7	117.1	Philadelphia, PA	24	40.9	75.5
Chicago, IL	19	42.3	88.0	Charleston, SC	38	33.3	80.8
Indianapolis, IN	21	39.8	86.9	Nashville, TN	31	36.7	87.6
Des Moines, IA	11	41.8	93.6	Amarillo, TX	24	35.6	101.9
Wichita KS	22	38.1	97.6	Galveston, TX	49	29.4	95.5
Louisvilie, KY	27	39.0	86.5	Houston, TX	44	30.1	95.9
New Orleans, LA	45	30.8	90.2	Salt Lake City, UT	18	41.1	112.3
Portland, ME	12	44.2	70.5	Burlington, VT	7	45.0	73.9
Baltimore, MD	25	39.7	77.3	Norfolk, VA	32	37.0	76.6
Boston, MA	23	42.7	71.4	Seattle, WA	33	48.1	122.5
Detroit, Ml	21	43.1	83.9	Spokane, WA	19	48.1	117.9
Minneapolis, MN	2	45.9	93.9	Madison, Wl	9	43.4	90.2
St. Louis, MO	24	39.3	90.5	Milwaukee, Wl	13	43.3	88.1
Helena, MT	8	47.1	112.4	Cheyenne, WY	14	41.2	104.9

NOTE: The average minimum temperature for any month is obtained by adding the daily minimum temperatures during that month and dividing by the number of days in that month. The normal average January minimum temperature (y) was obtained by adding the average minimums for January 1931, January 1932, and so on, through January 1960 and dividing the total by 30. The variables x_1 and x_2 are latitude and longitude in degrees. Source: Long (1972)[†].

[†] Long, L. H. (ed.) (1972), The 1972 World Almanac and Book of Facts New York: Newspaper Enterprise Association.

18 Multiple Linear Regression(Chapt. 8 LSM)

Generally when we wish to determine whether or not certain relationships exist between a response and certain "condition(s)", it is not unusual to have two or more variables that can influence a specific outcome. These variables or "conditions" act synergistically on predicting or estimating an outcome.

Here we will assume on a linear relationship between these variables, but others could also exist.

The multiple linear regression model expresses the response as a function of *k* distinct independent predictor variables.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i .$$

Using the technique of least squares:

$$Q = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}) \right]^2 .$$

We wish to find estimates of β_0, β_1, β_2, . . ., β_k call them $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}$, . . ., $\widehat{\beta_k}$ that would minimize Q.

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki}) \\
\frac{\partial Q}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_{1i} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki}) \\
\frac{\partial Q}{\partial \beta_2} &= -2 \sum_{i=1}^{n} x_{2i} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki}) \\
\vdots &= \vdots \\
\text{and} & \\
\frac{\partial Q}{\partial \beta_k} &= -2 \sum_{i=1}^{n} x_{ki} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki}) .
\end{aligned}$$

After taking derivatives and setting equal to zero, and passing through the summation operator, we can solve the following equations:

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= n\widehat{\beta_0} + \widehat{\beta_1} \sum_{i=1}^{n} x_{1i} + \widehat{\beta_2} \sum_{i=1}^{n} x_{2i} + \cdots + \widehat{\beta_k} \sum_{i=1}^{n} x_{ki} \\
\sum_{i=1}^{n} x_{1i} y_i &= \widehat{\beta_0} \sum_{i=1}^{n} x_{1i} + \widehat{\beta_1} \sum_{i=1}^{n} x_{1i}^2 + \widehat{\beta_2} \sum_{i=1}^{n} x_{1i} x_{2i} + \cdots + \widehat{\beta_k} \sum_{i=1}^{n} x_{1i} x_{ki} \\
\vdots &= \vdots \\
\text{and} & \\
\sum_{i=1}^{n} x_{ki} y_i &= \widehat{\beta_0} \sum_{i=1}^{n} x_{ki} + \widehat{\beta_1} \sum_{i=1}^{n} x_{ki} x_{1i} + \widehat{\beta_2} \sum_{i=1}^{n} x_{ki} x_{2i} + \cdots + \widehat{\beta_k} \sum_{i=1}^{n} x_{ki}^2 .
\end{aligned}$$

18.1 The fitted regression line(Chapt. 8 LSM)

(See Chapter 8 of LSM)

18.2 Is the multiple regression significant ?(Chapt. 8 LSM)

(See Chapter 8 of LSM)

18.3 Inferences on the fitted partial regression coefficients, the $\widehat{\beta_i}$'s.(Chapt. 8 LSM)

(See Chapter 8 of LSM)

Case Study: Computer Science

Mathematical Sciences 405/605
Case Study: Computer Science

The waiting time *y* that elapses between the time a computing job is submitted to a large computer and the time at which the job is initiated (computing commences) is a function of many variables, including the priority assigned to the job, the number and sizes of the jobs already on the computer, the size of the job being submitted, and so on. A study was initiated to investigate the relationship between waiting time y (in hours) for a job and x_1, the estimated CPU time (in seconds) for the job, and x_2, the CPU utilization factor. The estimated CPU time x_1 is an estimate of the amount of time that a job will occupy a portion of the computer's central processing unit's memory. The CPU utilization factor x_2 is the percentage of the memory bank of the central processing unit that is occupied at the time the job is submitted. We would expect the waiting time *y* to increase as the size of the job x, increases and as the CPU utilization factor x_2 increases. To conduct the study, *15* jobs of varying sizes were submitted to the computer at randomly assigned times throughout the day. The job waiting time *y*, estimated CPU time x_1 and CPU utilization factor x_2 were recorded for each job. The data[†] are shown below.

Job	x_1	x_2	y
1	2.0000	45.0000	0.0010
2	9.3000	80.0000	1.1400
3	5.6000	23.0000	0.0300
4	3.7000	25.0000	0.0010
5	12.4000	67.0000	0.7800
6	18.1000	30.0000	0.3000
7	13.5000	55.0000	0.6000
8	26.6000	21.0000	0.2000
9	34.2000	79.0000	2.2400
10	38.8000	40.0000	0.4400
11	56.1000	22.0000	0.0010
12	60.3000	37.0000	0.3200
13	4.4000	50.0000	0.1600
14	2.6000	66.0000	0.2900
15	20.9000	42.0000	0.4900

A second-order model, E(y) = $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$ was selected to model mean waiting time E(y).

i) Find the values of SSE and $\widehat{\sigma^2}$.

ii) Find the prediction equation.

iii) Find R^2 and interpret its value.

iv) Do the data provide sufficient evidence to indicate that the model contributes information for the prediction of *y* ? Test using α = 0.10.

[†] Waiting time data frequently violate the assumptions required for significance tests and confidence intervals in a regression analysis. The probability distribution for waiting times is often skewed, and its variance increases as the mean waiting time increases. Methods are available for coping with this problem, but we will ignore it for the purposes of this introductory discussion.

MODELS WITH TWO INDEPENDENT VARIABLES

No interaction model

$$E(y) \;=\; \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Comments on model parameters:

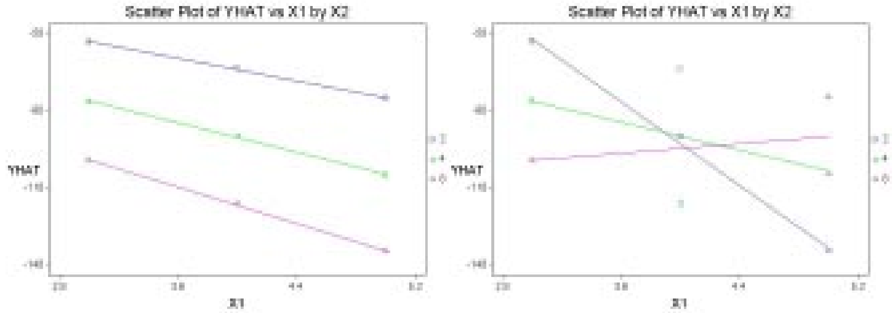
- β_0: y-intercept, the value of E(y) when $x_1 = x_2 = 0$
- β_1: Change in E(y) for a 1 unit increase in x_1, when x_2 is held fixed.
- β_2: Change in E(y) for a 1 unit increase in x_2, when x_1 is held fixed.

General comments: In particular, a first-order model relating E(y) to two independent quantitative variables, x_1 and x_2, graphs as a plane in three-dimensional space. The plane traces the value of E(y) for every combination of values (x_1, x_2) that correspond to points in the x_1, x_2 plane. Most response surfaces in the real world are well behaved (smooth), and they have curvature. Consequently, a first-order model is appropriate only if the response surface is fairly flat over the x_1, x_2 region that is of interest to you.

Interaction model

$$E(y) \;=\; \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

The assumption that a first-order model will adequately characterize the relationship between E(y) and the variables x_1 and x_2 is equivalent to assuming that x_1 and x_2 do not interact; that is, you assume that the effect on E(y) of a change in x_1 (for a fixed value of x_2) is the same regardless of the value of x_2 (and vice-versa). Thus, no interaction implies that the effect of changes in one variable (say x_1) on E(y) is independent of the value of the second variable (say x_2).



Two variables **interact** if the change in E(y) for 1-unit increase (decrease) in x_1(when x_2 is held fixed) is dependent on the values of x_2. In which case the lines in the previous plot would cross.

Interaction terms clearly allow more opportunity for individual predictor variables to exhibit joint effects with other predictor variables. Several interaction terms involving two or more predictor variables can be included in regression models but they should not be inserted routinely for several reasons. First the number of possible interaction terms can be large for regression models with several predictor variables. With only 5 predictor variables there are 10 possible two-variable interaction terms, 10 three-variable interaction terms, 5 four-variable interaction terms, and 1 five-variable interaction term. Use of all predictor variables and their interactions could result in a complicated model with 32 terms in it.

18.4 Indicator and Dummy Variables in Multiple regression(Chapt. 12 LSM)

Case Study: Industrial

Mathematical Sciences 405/605
Case Study: The Petroleum Industry

In the oil industry, water mixes with crude oil during production and transportation. The organic properties of oil prevent it from dissolving in an inorganic medium; rather, tiny oil particles are suspended within the water. This water and oil (w/o) suspension is called an emulsion.

Chemists have found that the oil can be extracted from the w/o emulsion electrically. In a high electric field, the (lighter) emulsified droplets are enlarged while the (heavier) water settles out of the mix gravitationally. Researchers at the University of Bergen (Norway) conducted a series of experiments to study the factors that influence the voltage required to separate the water from the oil in w/o emulsions (*Journal of Colloid and Interface Science. Aug. 1995*). The seven independent variables investigated in the study are described below. Each variable was measured at two levels a “low” level and a “high” level.

- x_1: Volume fraction of disperse phase (as a percentage of weight); Low = 40%, High = 80%
- x_2: Salinity of emulsion (as a percentage of weight); Low = 1%, High = 4%
- x_3: Temperature of emulsion (in C); Low = 40°, High = 23°
- x_4: Time delay after emulsification (in hours); Low = 0.25 hour(15minutes), High = 24 hours
- x_5: Concentration of surface-active agent, or ”surfactant” (as a percentage of weight); Low = 2%, High = 4%
- x_6: Ratio of two chemicals (Span and Triton) used as surfactants; Low = .25, High = .75
- x_7: Amount of solid particles added (as a percentage of weight); Low = .5%, High = 2%

Sixteen w/o emulsions were prepared using different combinations of the independent variables listed above; then each emulsion was exposed to a high electric field. In addition, three w/o emulsions were tested when all independent variables were set to 0. For all 19 emulsions, the amount of voltage (kilovolts per centimeter) where the first sign of macroscopic activity is observed was measured; this value represents the dependent variable, *y*. The data for the study are given in Table 1.

1. Propose a model for y as a function of all seven independent variables. Assume that a linear relationship exists between *y* and x_i, i = 1, 2, . . . , 7.

2. Use a statistical software package to fit the model to the data in Table 1.

3. Fully interpret the results of the regression. Part of the analysis should include an interpretation of the β estimates.

4. According to the researchers, the model predicts a negative value for the voltage *y* for experiment #14. Verify this result.

5. The researchers state that the result, part 4, “is physically not acceptable, and a model with interaction terms must be proposed.” The model the researchers selected is
E(y) = β_0 + $\beta_1 x_1$ + $\beta_2 x_2$ + $\beta_3 x_5$ +$\beta_4 x_1 x_2$ + $\beta_5 x_1 x_5$. Note that the model includes interact)¡ between disperse phase volume (x_1) and salinity (x_2) as well as interaction between disperse phase volume(x_5) and surfactant concentration (x_5). Discuss how these interaction terms affect the hypothetical relationship between *y* and x_1. Draw a sketch to support your answer.

6. Fit the interaction model, part 5, to the data. Do the model appear to fit the data better than the model in part 1? Explain.

7. Interpret the β estimates of the interaction model from part 5.

8. The researchers concluded that “in order to break; an emulsion with the lowest possible voltage, the volume fraction of the disperse phase (x_1) should high, while the salinity(x_2) and the amount of surfactant (x_5) should be low.” Use this information and the interaction model to find a 95% prediction interaction for this ”low” voltage *y*. Interpret the interval.

		Disperse Phase			**Time**	**Surfactant**	**S:T**	**Solid**
Experiment	**Voltage**	**Volume**	**Salinity**	**Temperature**	**Delay**	**Concentration**	**Ratio**	**Particles**
	(y)	(x_1)	(x_2)	(x_3)	(x_4)	(x_5)	(x_6)	(x_7)
1	.64	40	1	4	.25	2	.25	.5
2	.80	80	1	4	.25	4	.25	2
3	3.20	40	4	4	.25	4	.75	.5
4	.48	80	4	4	.25	2	.75	2
5	1.72	40	1	23	.25	4	.75	2
6	.32	80	1	23	.25	2	.75	.5
7	.64	40	4	23	.25	2	.25	2
8	.68	80	4	23	.25	4	.25	.5
9	.12	40	1	4	24	2	.75	2
10	.88	80	1	4	24	4	.75	.5
11	2.32	40	4	4	24	4	.25	2
12	.40	80	4	4	24	2	.25	.5
13	1.04	40	1	23	24	4	.25	.5
14	.12	80	1	23	24	2	.25	2
15	1.28	40	4	23	24	2	.75	.5
16	.72	80	4	23	24	4	.75	2
17	1.08	0	0	0	0	0	0	0
18	1.08	0	0	0	0	0	0	0
19	1.04	0	0	0	0	0	0	0

Collinearity Diagnostics (Chapt. 10 LSM)

When a regressor, x_i, is nearly a linear combination of other regressors in the model the affected estimates are unstable and have high standard errors. This problem is called **collinearity** or **multicollinearity**. It is a good idea to find out which variables are nearly collinear with which other variables. The approach in **PROC REG** follows that of Belsley, Kuh, and Welsch (1980). **REG** provides several methods for detecting collinearity with the **COLLIN**, **COLLINOINT**, **TOL**, and **VIF** options

The **COLLIN** option in the MODEL statement requests that a collinearity analysis be done. First, $\mathbf{X'X}$ is scaled to have *1*s on the diagonal. If **COLLINOINT** is specified, the intercept variable is adjusted out first. Then the eigenvalues and eigenvectors are extracted. The analysis in **REG** is reported with eigenvalues of $\mathbf{X'X}$ rather than values from the singular decomposition of **X**. The singular values of **X** are the square roots of the eigenvalues of $\mathbf{X'X}$.

The condition indices are the square roots of the ratio of the largest eigenvalue to each individual eigenvalue. The largest condition index is the condition number of the scaled **X** matrix. When this number is large, the data are said to be ill-conditioned. When this number is extremely large, the estimates may have a fair amount of numerical error (although the statistical standard error almost always is much greater than the numerical error).

For each variable, **REG** prints the proportion of the accounted for by each principal component. A collinearity problem occurs when a component associated with a high condition index contributes strongly to the variance of two or more variables.

The **VIF** option in the **MODEL** statement provides the variance inflation factors. These factors measure the inflation in the variances of the parameter estimates due to collinearities that exist among the regressor (dependent) variables. There are no formal criteria for deciding if a **VIF** is large enough to affect the predicted values. But, there are informal criteria that work quite well in practice. A liberal criteria is any VIFs over 10 suggest multicollinearity. A more conservative criteria would suggest multicollinearity if there are any VIFs greater than the number of parameters in the model.

The **TOL** option requests the tolerance values for the parameter estimates.

For a complete discussion of the methods discussed above, see Belsley, Kuh, and Welsch (1980). For a more detailed explanation of using the methods with **PROC REG**, see Freund and Littell (1986).

Here is an example using the **COLLIN** option on the oxidation data given in a class handout and reproduced at the end of this handout.

```
proc reg data=oxidata;
model oxidant=windspd temp humid insolate / tol vif collin;
run;
```

Influence Diagnostics

The **INFLUENCE** option requests the statistics proposed by Belsley, Kuh, and Welsch (1980) to measure the *influence* of each observation on the estimates. Influential observations are those that, according to various criteria, appear to have a large *influence* on the parameter estimates. Let $\beta_{-(i)}$ be the parameter estimates after deleting the *i*th observation; let s$^2_{-(i)}$ be the variance estimate after deleting the *i*th observation; let $\mathbf{X}_{-(i)}$ be the **X** matrix without the *i*th observation(case); let $\widehat{y}_{-(i)}$ be the *i*th value predicted without using the *i*th observation; let r$_i$ = y$_i$ - $\widehat{y}$ be the *i*th residual; and let h$_i$ be the *i*th diagonal of the projection matrix for the predictor space, also called the hat matrix:

$$h_i \quad = \quad \mathbf{x}_i \left(\mathbf{X'X}\right)^{-1} \mathbf{x}_i'$$

Belsley, Kuh, and Welsch propose a cutoff of *2*p/n*, where *n* is the number of observations used to fit the model, and *p* is the number of parameters in the model. Observations with h$_i$ values above this cutoff should be investigated.

For each observation, **REG** first prints the residual, the studentized residual, and the h$_i$. The studentized residual differs slightly from that in the previous section since the error variance is estimated by s$^2_{-(i)}$ without the *i*th observation, not by s^2, for example,

$$\textbf{RSTUDENT} = \frac{r_i}{\sqrt{s^2_{-(i)}\,(1-h_i)}}$$

Observations with **RSTUDENT** larger than 2 in absolute value may need some attention.

The **COVRATIO** statistic measures the change in the determinant of the covariance matrix of the estimates by deleting the *i*th observation:

$$\textbf{COVRATIO} = \frac{|\ s^2_{-(i)}\mathbf{X'}_{-(i)}\mathbf{X}_{-(i)}\ |}{|\ s^2(\mathbf{X'X})^{-1}\ |}$$

Belsley, Kuh, and Welsch suggest observations with

$$|\textbf{COVRATIO} - 1| \geq \frac{3p}{n}$$

where *p* is the number of parameters in the model, and *n* is the number of observations used to fit the model, are worth investigation.

The **DFFITS** statistic is a scaled measure of the change in the predicted value for the *i*th observation and is calculated by deleting the *i*th observation. A large value indicates that the observation is very influential in its neighborhood of the **X** space.

$$\textbf{DFFITS} = \frac{\widehat{y}_i - \widehat{y}_{-(i)}}{\sqrt{s^2_{-(i)}h_i}}$$

Large values of **DFFITS** indicate influential observations. A general cutoff to consider is 2; a size-adjusted cutoff recommended by Belsley, Kuh, and Welsch is $2\sqrt{\frac{p}{n}}$, where *n* and *p* are as defined above. **DFFITS** is very similar to **Cook's Distance**.

Cook's D, for short, is also a scaled measure. Cases for which D_i is large have substantial influence on $\widehat{\beta}$ and on the fitted values, and deletion of them may result in important changes in conclusions. Typically the case with the largest D_i, or in large data setsthe cases with the largest few D_i, will be of interest. A proposed cut-off, see Weisberg (1985), is if D_i is substantially less than 1, deletion of a case will not change the estimate β by much. To investigate the influence of a case more closely, the analyst should delete the large D_i case and recompute the analysis to see exactly what aspects of it have changed.

The simplest form for D_i is

$$D_i = \frac{1}{p}\textbf{RSTUDENT}^2_i\left(\frac{h_i}{1-h_i}\right)$$

If *p* is fixed, the size of D_i will be determined by two different sources: the size of $\textbf{RSTUDENT}_i$, a random variable reflecting lack of fit of the model at the *i* th case, and the potential h_i, reflecting the location of $\mathbf{x}_i$ relative to $\overline{x}$. A large value of D_i may be due to large $\textbf{RSTUDENT}_i$, large h_i, or both.

DFBETAS are the scaled measures of the change in each parameter estimate and are calculated by deleting the *i*th observation:

$$\textbf{DFBETAS}_j = \frac{\widehat{\beta}_j - \widehat{\beta}_{j_{-(i)}}}{\sqrt{s^2_{-(i)}(\mathbf{X'X})^{-1}_{jj}}}$$

where

$(\mathbf{X'X})^{-1}_{jj}$ is the *(j,j)*th element of $(\mathbf{X'X})^{-1}$.

In general, large values of **DFBETAS** indicate observations that are influential in estimating a given parameter. Belsley, Kuh, and Welsch recommend 2 as a general cutoff value to indicate influential observations and $2/\sqrt{n}$ as a size-adjusted cutoff.

The output below shows the portion of output produced by the **INFLUENCE** option for the oxidation example. See the subsequent output for the fitted regression equation.

```
proc reg data=oxidata;
model oxidant=windspd temp humid insolate / influence;
run;
```

Table 13: Influence Diagnostics for Ozone Data

Criteria	*Criteria Cutoff*	*Size Adjusted*			*Suspect Cases*		
h (Hat)	0.33				22,23,30		
RSTUDENT	2				11		
COVRATIO	0.5				8,11,24		
					25,29,30		
DFFITS	2	0.82			4,22,23		
DFBETAS	2	0.37			**Parameters**		
			Intercept	**WINDSPD**	**TEMP**	**HUMID**	**INSOLATE**
			4,22	4,22,23	4,21	4,28	22,23

The **PARTIAL** option produces **PARTIAL** regression leverage plots. One plot is printed for each regressor in the full, current model. For example, plots are produced for regressors included by using ADD statements; plots are not produced for interim models in the various model-selection methods but only for the full model. If you use a model-selection method and the final model contains only a subset of the original regressors, the **PARTIAL** option still produces plots for all regressors in the full model.

For a given regressor, the **PARTIAL** regression leverage plot is the plot of the dependent variable and the regressor after they have been made orthogonal to the other regressors in the model. These can be obtained by plotting the residuals for the dependent variable against the residuals for the selected regressor, where the residuals for the dependent variable are calculated with the selected regressor omitted, and the residuals for the selected regressor are calculated from a model where the selected regressor is regressed on the remaining regressors. A line fit to the points has a slope equal to the parameter estimate in the full model.

On the plot, points are marked by the number of replicates appearing at one print position. The symbol ’*’ is used if there are ten or more replicates. if an ID statement is specified, the left-most nonblank character in the value of the ID variable is used as the plotting symbol.

The following statements use the oxidation data.

```
proc reg data=oxidata;
model oxidant=windspd temp humid insolate / partial;
run;
```

References

[1] Belsley, Kuh, and Welsch (1980). “Regression Diagnostics”, *John Wiley & Sons*

[2] Freund and Littell (1986). “SAS System for Regression”, *SAS Institute*

Case Study: Toxicity

Mathematical Sciences 405/605
Case Study:Toxicity
100 points

I. **The Data:** It is known that in mammals the toxicity of various drugs, pesticides, and chemical carcinogens can be altered by inducing liver enzyme activity. A study to investigate this sort of phenomena in a vertabrate model similar to that of humans is reported in an article in the *American Journal of Veterinary Research*. Regression analysis was used to study the relationship between induced enzyme activity and detoxification on the insecticide malathion. Butylated hydroxytaluene(BHT) and 3-methylcholanthrene (MC) were used to induce enzyme activity. Each observation represents the percentage of activity relative to a control, an untreated lab animal. The response variable is the percentage of detoxification of malathion. Five enzyme activities were measured and serve as the predictor variables.

Table 14: Detoxification Data

Inducer	**Detoxification**	**Enzyme 1**	**Enzyme 2**	**Enzyme 3**	**Enzyme 4**	**Enzyme 5**
x_1	y	x_2	x_3	x_4	x_5	x_5
BHT	146.1040	348.4750	337.5000	108.1220	106.6670	107.6920
BHT	152.5970	233.2200	260.4170	82.2340	80.0000	88.8890
BHT	168.8310	287.4580	273.9580	74.6190	66.6670	87.1790
BHT	178.5710	152.5420	310.4170	86.8020	73.3330	96.5810
BHT	191.5580	276.2710	818.7500	122.8430	86.6670	97.4360
BHT	113.6360	78.6440	156.2500	112.6900	93.3330	94.8720
BHT	188.3120	196.9490	260.4170	79.1880	80.0000	106.8380
BHT	94.1560	101.6950	112.5000	127.9190	93.3330	80.3420
BHT	159.0910	194.5760	280.2080	239.5940	106.6670	91.4530
BHT	142.8570	325.4240	326.0420	173.0960	113.3330	100.0000
MC	56.2500	106.3290	90.7560	94.6500	162.7910	114.7370
MC	75.0000	144.7260	203.3610	131.6870	255.8140	112.6320
MC	115.6250	136.2870	672.2690	123.4570	191.8600	153.6840
MC	68.7500	154.4300	183.1930	113.1690	133.7210	116.8420
MC	96.8750	385.2320	140.3360	117.2840	174.4190	87.3680
MC	168.7500	583.5440	146.2180	152.2630	273.7560	94.7370
MC	84.3750	489.4510	184.8740	121.3990	255.8140	95.7890
MC	171.8750	445.9920	537.8150	150.2060	552.3260	113.6840
MC	109.3750	270.8860	309.2440	185.1850	534.8840	108.4210
MC	103.1250	163.2910	190.7560	139.9180	360.4650	106.3160

(i) Fit a multiple linear regression that models detoxification as a function of the five enzymes for each of the two types of inducers.

(ii) Using the methods described in your handout on collinearity and influence, determine if multicollinearity exists for each of the models and whether or not there are any influential cases.

(iii) Using the variable selection methods described in class determine the variables that are "best" in predicting the percentage of detoxification of malathion when looking at BHT as the inducer and the variables that are "best" in predicting the percentage of detoxification of malathion when considering MC as the inducer.

19 Model-Selection Methods(Chapt. 11 LSM)

Introduction

The nine methods of model selection implemented in **PROC REG** are specified with the **SELECTION=** option in the **MODEL** statement. Each method is discussed below.

Full Model Fitted (NONE)

This method is the default and provides no model selection capability. The complete model specified in the **MODEL** statement is used to fit the model. For many regression analyses, this may be the only method you need.

Forward Selection (FORWARD)

The forward-selection technique begins with no variables in the model. For each of the independent variables, **FORWARD** calculates ***F*** statistics that reflect the variable's contribution to the model if it is included. The *p*-values for these ***F*** statistics are compared to the **SLENTRY=** value that is specified in the **MODEL** statement (or to 0.50 if the **SLENTRY=** option is omitted). If no ***F*** statistic has a significance level greater than the **SLENTRY=** value, **FORWARD** stops. Otherwise, **FORWARD** adds the variable that has the largest ***F*** statistic to the model. **FORWARD** then calculates ***F*** statistics again for the variables still remaining out side the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant ***F*** statistic. Once a variable is in the model, it stays.

Backward Elimination (BACKWARD)

The backward-elimination technique begins by calculating statistics for a model, including all of the independent variables. Then the variables are deleted from the model one by one until all the variables remaining in the model produce ***F*** statistics significant at the **SLSTAY=** level specified in the **MODEL** statement (or at the 0.10 level if the **SLSTAY=** option is omitted). At each step, the variable showing the smallest contribution to the model is deleted.

Stepwise (STEPWISE)

The stepwise method is a modification of the forward-selection technique and differs in that variables already in the model do not necessarily stay there. As in the forward-selection method, variables are added one by one to the model, and the ***F*** statistic for a variable to be added must be significant at the **SLENTRY=** level. After a variable is added, however, the stepwise method looks at all the variables already included in the model and deletes any variable that does not produce an ***F*** statistic significant at the **SLSTAY=** level. Only after this check is made and the necessary deletions accomplished can another variable be added to the model. The stepwise process ends when none of the variables outside the model has an ***F*** statistic significant at the **SLENTRY=** level and every variable in the model is significant at the **SLSTAY=** level, or when the variable to be added to the model is the one just deleted from it.

Maximum $\boldsymbol{R}^2$ Improvement (MAXR)

The maximum $\boldsymbol{R}^2$ improvement technique does not settle on a single model Instead, it tries to find the ”best” one-variable model, the ”best” two-variable model, and so forth, although it is not guaranteed to find the model with the largest $\boldsymbol{R}^2$ for each size. The **MAXR** method begins by finding the one-variable model producing the highest $\boldsymbol{R}^2$. Then another variable, the one that yields the greatest increase in $\boldsymbol{R}^2$, is added. Once the two-variable model is obtained, each of the variables in the model is compared to each variable not in the model. For each comparison, **MAXR** deter mines if removing one variable and replacing it with the other variable increases $\boldsymbol{R}^2$. After comparing all possible switches, **MAXR** makes the switch that produces the largest increase in $\boldsymbol{R}^2$. Comparisons begin again, and the process continues until **MAXR** finds that no switch could increase $\boldsymbol{R}^2$. Thus, the two-variable model achieved is considered the ”best” two-variable model the technique can find. Another variable is then added to the model, and the comparing-and-switching process is repeated to find the ”best” three-variable model, and so forth. The difference between the STEPWISE method and the **MAXR** method is that all switches are evaluated before any switch is made in **MAXR**. In the STEPWISE method, the “worst” variable can be removed without considering what adding the best” remaining variable might accomplish. **MAXR** may require much more computer time than STEPWISE.

Minimum $\boldsymbol{R}^2$ Improvement (MINR)

The **MINR** method closely resembles **MAXR**, but the switch chosen is the one that produces the smallest increase in $\boldsymbol{R}^2$. For a given number of variables in the model, **MAXR** and **MINR** usually produce the same “best” model, but **MINR** considers more models of each size.

$\boldsymbol{R}^2$ Selection (RSQUARE)

The **RSQUARE** method finds subsets of independent variables that best predict a dependent variable by linear regression in the given sample. You can specify the largest and smallest number of independent variables to appear in a subset and the number of subsets of each size to be selected. The **RSQUARE** method can efficiently perform all possible subset regressions and print the models in decreasing order of $\boldsymbol{R}^2$ magnitude within each subset size. Other statistics are available for comparing subsets of different sizes. These statistics, as well as estimated regression coefficients, can be printed or output to a SAS data set. The subset models selected by **RSQUARE** are optimal in terms of $\boldsymbol{R}^2$ for the given sample, but they are not necessarily optimal for the population from which the sample was drawn or for any other sample for which you may want to make predictions. If a subset model is selected on the basis of a large $\boldsymbol{R}^2$ value or any other criterion commonly used for model selection, then all regression statistics computed for that model under the assumption that the model is given a priori, including all statistics computed by **REG**, are biased. While the **RSQUARE** method is a useful tool for exploratory model building, no statistical method can be relied on to identify the ”true” model. Effective model building requires substantive theory to suggest relevant predictors and plausible functional forms for the model. The **RSQUARE** method differs from the other selection methods in that **RSQUARE** always identifies the model with the largest $\boldsymbol{R}^2$ for each number of variables considered . The other selection methods are not guaranteed to find the model with the largest $\boldsymbol{R}^2$. **RSQUARE** requires much more computer time than the other selection methods, so a different selection method such as **STEPWISE** is a good choice when there are many independent variables to consider.

Adjusted $\boldsymbol{R}^2$ Selection (ADJRSQ)

This method is similar to **RSQUARE**, except that the adjusted $\boldsymbol{R}^2$ statistic is used as the criterion for selecting models, and the method finds the models with the highest adjusted $\boldsymbol{R}^2$ within the range of sizes.

Mallows' C$_p$ Selection (C$_p$)

This method is similar to **ADJRSQ**, except that Mallow's C$_p$ statistic is used as the criterion for model selection.

Additional Information on Model-Selection Methods

If the **RSQUARE** or **STEPWISE** procedure (as documented in SAS User's Guide: Statistics, Version 5 Edition) is requested, **PROC REG** with the appropriate model-selection method is actually used. Reviews of model-selection methods by Hocking (1976) and Judge et al. (1980) describe these and other variable-selection methods.

20 The Analysis of Variance(Chapt. 14 LSM)

20.1 The Completely Randomized Design

Given the common hypothesis:

$$\mathrm{H}_0\text{: } \mu_1 = \mu_2 = \ldots \mu_r$$

$$\mathrm{H}_a\text{: } \mu_i \;\neq\; \mu_j \quad \{i \neq j\} \quad \leq r,$$

there are a number of techniques for comparing means from several populations or processes. One particularly interesting method is called the *Analysis of Variance*. Although it seems a misnomer, we will see how the analysis of variance is used for testing the inequality of means from several populations. For experiments involving *r* means a model of the form:

$$y_{ij} \;=\; \mu_i + \epsilon_{ij} \quad i = 1, 2, \ldots, r \quad j = 1, 2, \ldots, n,$$

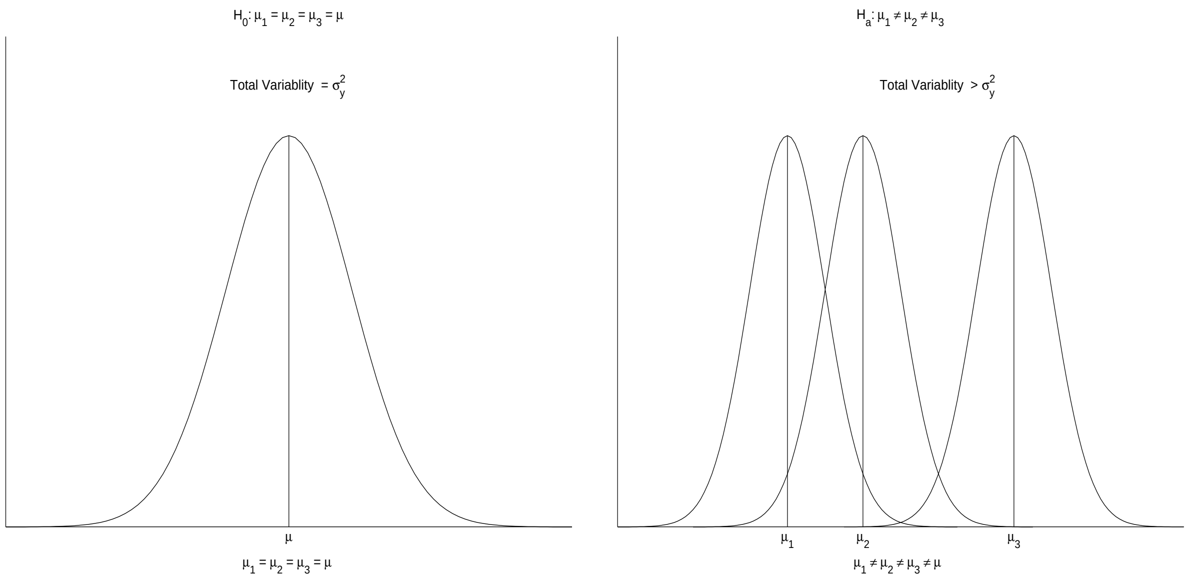
can be used. This model is equivalent to the one-factor model used to analyze data resulting from designed experiments. The simplest of these is the completely randomized design. The analysis of variance (ANOVA) is especially suited for comparing means of populations when it can be assumed that the population variances are equal.

Let's assume that the observations from the *r* different processes are independent. The sample sizes n_i need not be the same but we will keep things simple by making that assumption here. Assume further that the errors are normally distributed with mean 0 and common variance σ^2, (recall we must make this assumption as stated above).

The purpose of the ANOVA is to assess whether the means in the model given above are significantly different from each other. Note that this is different from the analysis of means(ANOM), which compares the means to an overall mean. Given that H_0 is true, the *r* sample means, $\overline{y}_i$ provide an unbiased estimate of the population mean μ and each of the sample variances, s_i^2 provides an unbiased estimate of the population variance σ_y^2. Thus, we are taking, in effect, *r* repeated random samples, each of size *n*, from the same population. Recall from Chapter 3, that the variance of the sample mean, denoted by $\sigma_{\overline{y}}^2$ is equal to the population variance σ_y^2 divided by the sample size *n*:

$$\sigma_{\overline{y}}^2 \;=\; \frac{\sigma_y^2}{n}.$$

Thus, if the null hypothesis is true, the population variance σ_y^2 should be equal to *n* times the variance of the sample means, $\sigma_{\overline{y}}^2$, ie., σ_y^2 = *n* $\sigma_{\overline{y}}^2$. If the null hypothesis is not true, then the equation σ_y^2 = *n* $\sigma_{\overline{y}}^2$ will not hold; indeed, $n\sigma_{\overline{y}}^2$ will be greater than σ_y^2 due to the fact that the population means corresponding to the *r* populations are different. This relationship may be seen in the figure below.



If the *r* treatment effects are equal, we are drawing the random samples from one distribution with variance σ_y^2. If the *r* treatments effects are not equal, then the total amount of variability in the *Y* population must be greater than σ_y^2 as illustrated on the right side of the figure above. The repeated samples of size *n* would then be drawn from a "composite" population, indicated in the figure by the shaded area. Then, *n* times the variability of the sample average statistic $\overline{Y}$ in repeated samples must be larger than σ_y^2, because these samples are being drawn from a population in which the variability is greater than the one in which the population means are equal. Therefore, the hypotheses:

$$\text{H}_0\text{: } \mu_1 = \mu_2 = \ldots \mu_r$$

$$\text{H}_a\text{: } \mu_i \; \neq \; \mu_j \quad \{i \neq j\} \quad \leq r$$

are equivalent to the hypotheses:

$$\text{H}_0\text{: } n\ \sigma_{\overline{y}}^2 = \sigma_y^2$$

$$\text{H}_a\text{: } n\ \sigma_{\overline{y}}^2 > \sigma_y^2.$$

That is, we can test the equality of population mean effects by comparing estimates of σ_y^2, the population variance, and $\sigma_{\overline{y}}^2$, the variance of the sample mean statistic. The analysis of variance procedure does, in fact, analyze variances to compare means.

The analysis of variance procedure compares an estimate of σ_y^2,denoted by $\widehat{\sigma}_y^2$, with an estimate of $\sigma_{\overline{y}}$ denoted by $\widehat{\sigma}_{\overline{y}}$. If $\widehat{\sigma}_y^2$ is "much less" than *n* $\sigma_{\overline{y}}^2$ then there is reason to suspect that the null hypothesis is not true. To determine whether $\widehat{\sigma}_y^2$ is significantly less than *n* $\widehat{\sigma}_{\overline{y}}^2$, we compute, the probability that the difference *n* $\widehat{\sigma}_{\overline{y}}^2$ - $\widehat{\sigma}_y^2$ could arise by chance, (sampling error) if $\widehat{\sigma}_y^2 = \widehat{\sigma}_{\overline{y}}^2$. To illustrate the testing process, consider data from a polish cannons exercise:

[**Example:].** Bauer. Dirks, Palkovic and Wittmer fired tennis balls out of "Polish cannons" inclined at an angle of 45° using three different Propellants and two different Charge Sizes of propellant. They observed the distances traveled in the air by the tennis balls. Their data are given in the accompanying table. (Five trials were made for each Propellant/Charge Size combination and the values given are in feet.)

		Propellant		
		Lighter Fluid	Gasoline	Carburetor Fluid
Charge Size	2.5 ml	58	76	90
		50	79	86
		53	84	79
		49	73	82
		59	71	86
	5.0 ml	65	96	107
		59	101	102
		61	94	91
		68	91	95
		67	87	97

For the moment let's just consider the three types of propellant. And we want to determine if the population means for the three types of propellant differ. The means model:

$$y_{ij} \quad = \quad \mu_i + \epsilon_{ij} \quad i = 1, 2, \ldots, 3 \quad j = 1, 2, \ldots, 10$$

Propellant

	Lighter Fluid	Gasoline	Carburetor Fluid	Overall Mean
Propellant means	$\overline{y}_{1.}$=58.9	$\overline{y}_{2.}$=85.2	$\overline{y}_{3.}$=91.5	$\overline{y}_{..}$=78.53

Since the three sample means are quite dissimilar ($\overline{y}_{1.} = 58.9, \overline{y}_{2.} = 85.2,$ and $\overline{y}_{3.} = 91.5$) we might expect the analysis of variance procedure to suggest that the null hypothesis should be rejected.

20.2 Estimate of the population variance σ_y^2

To calculate a sample estimate of σ_y^2, we will use the sample variances s_1^2, s_2^2, and s_3^2. The variance for the *i*th sample is given by:

$$\begin{aligned} s_i^2 &= \frac{\sum_{i=1}^{n=10} (y_{ij} - \overline{y}_{i.})^2}{n-1} \\ &= \frac{\sum_{i=1}^{n=10} y_{ij}^2 - \frac{\left(\sum_{i=1}^{n=10} y_{ij}\right)^2}{n}}{n-1}. \end{aligned}$$

Thus,

$$\begin{aligned} s_i^2 &= \frac{(58)^2 + (50)^2 + (53)^2 + \cdots (67)^2 - \frac{(58+50+53+\cdots+67)^2}{10}}{9} \\ &= \frac{35095 - \frac{(589)^2}{10}}{9} \quad = \quad 44.7667. \end{aligned}$$

Similarly, s_2^2 = 106.1778 and s_3^2 = 78.0556. A critical assumption of analysis of variance is that the population variances corresponding to the *r=3* treatments are equal, regardless of whether or not the null-hypothesis is true (that is, whether or not the *r* population means are equal). Thus, any differences in the sample variances must always be attributable to sampling error. The three sample variances appear to be reasonably similar in this example. But let's apply the F$_{max}$ test.

$$\begin{aligned} F &= \frac{max\{s_i^2\}}{min\{s_i^2\}} \\ &= \frac{106.1778}{44.667} \\ &= 2.3718, \end{aligned}$$

which is less than 5.34, α=0.05 and obviously less than 8.5, with α=0.01. So the assumption is appropriate.

Since s_1^2, s_2^2, and s_3^2 each estimate the polulation variance σ_y^2, we can produce an improved estimate of σ_y^2 over each sample variance taken individually by pooling the three estimates (recall the pooled estimate of the population variance in the denominator of the *t*-statistic used to compare two population means).

The pooled estimate of σ_y^2 is given by:

$$\begin{aligned} s_p^2 &= \sum_{i=1}^{r} \frac{s_i^2}{r} \\ &= \frac{44.7667 + 106.1778 + 78.0556}{3} \\ &= 76.3333. \end{aligned}$$

20.3 Estimate of the variance the sample mean $\sigma_{\overline{y}}^2$

The three sample means are ($\overline{y}_{1.}$ = 58.9, $\overline{y}_{2.}$ = 85.2, and $\overline{y}_{3.}$ = 91.5). To compute the estimate $\widehat{\sigma}_{\overline{y}}^2$ of the three sample means we used the variance formula:

$$s_{\overline{y}}^2 = \frac{\sum_{i=1}^{r=3} (\overline{y}_{i.} - \overline{y}_{..})^2}{r-1}$$

In our example, $\overline{y}_{..}$=78.53. Then

$$\begin{aligned} s_{\overline{y}}^2 &= \frac{\sum_{i=1}^{r=3} (\overline{y}_{i.} - \overline{y}_{..})^2}{r-1} \\ &= \frac{(58.9 - 78.53)^2 + (85.2 - 78.53)^2 + (91.5 - 78.53)^2}{2} \\ &= 299.0234. \end{aligned}$$

If the null hypothesis is true, *n* $\widehat{\sigma}_{\overline{y}}^2$ should be an unbiased estimate of the population variance,σ_y^2. Its value is:

$$n\widehat{\sigma}_{\overline{y}}^2 = (10)(299.0234) = 2990.234.$$

Since *n* $\widehat{\sigma}_{\overline{y}}^2 > \widehat{\sigma}_y^2$ (2990.234 > 76.333), the sample means appear to be much too variable to have been drawn from the same common population with mean μ. But is the difference between *n* $\widehat{\sigma}_{\overline{y}}^2$ and $\widehat{\sigma}_y^2$ statistically significant ? We have developed two procedures for testing the equivalence of two population variances. The common F-test and Hartley's F$_{max}$ which we recently used. For the usual F-test,

$$\begin{aligned} f &= \frac{2990.234}{76.3333} \\ &= 39.1734. \end{aligned}$$

To determine whether the calculated value of *f* is significantly different from 1, we look for the critical value of *f* based on the numerator degrees of freedom ν_1 = r - 1 = 2, the denominator degrees of freedom ν_2 = *r(n-1)* = 3(9) = 27, and a selected value of α, say α=0.05. From the table in the appendix, we get $F_{2,24,0.05}$=3.40 and $F_{2,30,0.05}$= 3.32. Since the calculated value of the statistic **F**(*f* = 39.1734) is greater that the critical value(s) (39.1734 $>$ 3.40), we reject the null hypothesis H_0: μ_1 = μ_2 = μ_3 at the α=0.05 significance level. Thus we conclude that there appears to be strong evidence that the three propellants do produce different mean distances. It is possible to calculate the observed level of significance, but the closest tabled value yields a *p*-value $<$ 0.0005. Thus it is very likely that the three propellants produce different distances.

20.4 The Analysis of Variance table

A convenient computational format for calculating the statistics necessary to determine whether the null hypothesis should be rejected is provided by the analysis of variance table. Its form is presented in the table below.

Source of Variation	*Degrees of freedom*	*Sum of squares*	*Mean Square*	*F-ratio*
Among treatments	r - 1	SS_{tr}	MS_{tr}	$\frac{MS_{tr}}{MS_E}$
Experimental error	r(n - 1)	SS_E	MS_E	
Total	rn - 1	SS_T		

The first row of the table, “among treatments,” produces the estimate of the variance*n* $\sigma^2_{\overline{y}}$ it is denoted by MS_{tr} (mean square for factor levels) in the table. The second row of the table, “experimental error” produces the pooled estimate of the population variance σ^2_y it is denoted by MS_E (mean square for error) in the table. The *F*-statistic is the ratio of the mean square for treatments, MS_{tr} , and the mean square for error, MS_E .

The formula for MS_{tr} is:

$$MSTr \;=\; n\widehat{\sigma}^2_{\overline{y}} \;=\; \frac{n\sum_{i=1}^{r}\left(\overline{y}_{i.}-\overline{y}_{..}\right)^2}{r-1}$$

The numerator of MS_{tr} is called the sum of squares among treatments and is denoted by SS_{tr} in the table. The formula for MS_E is:

$$\begin{aligned} MS_{tr} \;&=\; \widehat{\sigma}^2_y \;=\; \frac{(n-1)\sum_{i=1}^{r} s_i^2}{r(n-1)} \\ &=\; \frac{(n-1)\sum_{i=1}^{r}\sum_{j=1}^{n}\left(\overline{y}_{ij}-\overline{y}_{i.}\right)^2}{r(n-1)} \end{aligned}$$

The numerator of MS_E is called the error sum of squares and is denoted by SS_E in the table. The “Degree of freedom” column in the table gives the appropriate divisors of the sums of squares to produce the mean squares.

The last row in the table gives the “Total degrees of freedom” *(tn - 1) = (t - 1) + t(n - 1)* - and the total sum of squares given by:

$$SS_T \;=\; \sum_{i=1}^{r}\sum_{j=1}^{n}\left(\overline{y}_{ij}-\overline{y}_{..}\right)^2$$

Hence, the analysis of variance table can also be given as:

Source of Variation	*Degrees of freedom*	*Sum of squares*	*Mean Square*	*F-ratio*
Among treatments	r - 1	$n\sum_{i=1}^{r}\left(\overline{y}_{i.}-\overline{y}_{..}\right)^2$	$MS_{tr}=\frac{(n-1)\sum_{i=1}^{r}s_i^2}{r(n-1)}$	$\frac{MS_{tr}}{MS_E}$
Experimental error	r(n - 1)	SS_E	MS_E	
Total	rn - 1	SS_T		

Notice that SS$_T$ gives the sum of the squared deviations of each observation y_{ij} about the grand mean of the data $y_{..}$ Thus, this quantity is a measure of the total variability in the dependent variable.

An important relationship given in the analysis of variance table is:

$$SST = SST_r + SS_E$$

That is, the total sum of squares can be partitioned into the sum of squares due to treatments plus the error sum of squares. So, for the polish cannon data we have the following analysis of variance table:

	Degrees of	*Sum of*		
Source of Variation	*freedom*	*squares*	*Mean Square*	*F-ratio*
Among propellants	3 - 1 [2]	5980.467	2990.233	39.173
Experimental error	3(10 - 1) [27]	2061.000	76.333	
Total	3(10)-1 [29]	8041.467		

Or in Statistix we would get the following:

```
STUDENT EDITION OF STATISTIX
ONE-WAY AOV FOR DISTANCE BY PROP

SOURCE    DF       SS         MS        F       P
-------  ----  ---------  ---------  ------  ------
BETWEEN    2     5980.47    2990.23   39.17  0.0000
WITHIN    27     2061.00    76.3333
TOTAL     29     8041.47

                    CHI-SQ    DF       P
BARTLETT’S TEST OF  ------  ------  ------
   EQUAL VARIANCES    1.56     2    0.4595

COCHRAN’S Q                    0.4637
LARGEST VAR / SMALLEST VAR     2.3718

COMPONENT OF VARIANCE FOR BETWEEN GROUPS    291.390
EFFECTIVE CELL SIZE                            10.0

                        SAMPLE     GROUP
   PROP         MEAN     SIZE     STD DEV
---------   ----------  ------  ----------
    1          58.900     10       6.6908
    2          85.200     10       10.304
    3          91.500     10       8.8349
TOTAL          78.533     30       8.7369
CASES INCLUDED 30    MISSING CASES 0
```

Suppose we wanted to do the same thing for the Charge size in our example. Then we can let a similar model as that posed before for the propellant, represent the mean charge.

$$y_{ij} \;=\; \mu_i + \epsilon_{ij} \quad i = 1, 2. \;\; j = 1, 2, \ldots, 15.$$

Similarly for Charge type alone we would get:

```
STUDENT EDITION OF STATISTIX
ONE-WAY AOV FOR DISTANCE BY CHARGE

SOURCE    DF       SS         MS         F       P
-------  ----  ---------  ---------  ------  ------
BETWEEN    1     1414.53    1414.53    5.98  0.0210
WITHIN    28     6626.93    236.676
TOTAL     29     8041.47

                    CHI-SQ    DF       P
BARTLETT'S TEST OF  ------  ------  ------
   EQUAL VARIANCES    0.31     1    0.5767

COCHRAN'S Q                       0.5755
LARGEST VAR / SMALLEST VAR        1.3555

COMPONENT OF VARIANCE FOR BETWEEN GROUPS    78.5238
EFFECTIVE CELL SIZE                            15.0

                        SAMPLE     GROUP
  CHARGE        MEAN     SIZE     STD DEV
---------    ----------  ------  ----------
    1          71.667      15       14.176
    2          85.400      15       16.505
TOTAL          78.533      30       15.384

CASES INCLUDED 30    MISSING CASES 0
```

Here's a third example of comparing several means and the completely randomized design. The production manager of a company which manufactures filters for liquids, for use in the pharmaceutical and food industries, wishes to compare the burst strength of four types of membrane. The first (A) is the company's own standard membrane material, the second (B) is a new material the company has developed, and C and D are membrane materials from other manufacturers. The manager has tested five filter cartridges from ten different batches of each material. The mean burst strengths for each set of five cartridges are given in above. The data can be analysed by setting up a multiple regression model. We let *Y* be the average burst strength for each set of five cartridges and x_1,x_2,x_3 be indicator variables coded as: The coefficients then represent the differences between the company's standard membrane and the others. This sets up four of the six possible comparisons. If we fit the model

Table 15: Burst strength of filter membranes (kPa)

Type A	95.5	103.2	93.1	89.3	90.4	92.1	93.1	91.9	95.3
Type B	90 5	98.1	97.8	97.0	98.0	95.2	95.3	97.1	90.5
Type C	86.3	84.0	86.2	80.2	83.7	93.4	77.1	86.8	83.7
Type D	89.5	93.4	87.5	89.4	87.9	86.2	89.9	89.5	90.0

$$y = \beta_0 + \beta_1 x_1 = +\beta_2 x_2 + \beta_3 x_3$$

	x_1	x_2	x_3
Type A	0	0	0
Type B	1	0	0
Type C	0	1	0
Type D	0	0	1

we obtain the following results

$$y = 92.84 + 3.24x_1 - 8.21x_2 - 2.95x_3$$

with *s*= 3.901 and the table of coefficients below:

Predictor	Coef	Stdev	t-ratio	p-value
Constant	92.84	1.234	75.27	0.000
Type B	3.24	1.744	1.86	0.071
Type C	8.21	1.744	-4.71	0.000
Type D	2.95	1.744	-1.69	0.099

20.5 Paired Comparisons for Analysis of Variance

I. Bonferroni's Procedure:

i) Equal Sample Sizes: Let *n* = n_i, *i* = 1,. . . ,*k*, the set of confidence intervals with endpoints:

$$(\bar{y}_{i.} - \bar{y}_{j.}) \pm t_{(N-k),\alpha'}\sqrt{\frac{2MSE}{n}}$$

Each confidence interval that does not include zero suggests $\mu_i \neq \mu_j$ at α.

ii) Unequal Sample Sizes:

$$(\bar{y}_{i.} - \bar{y}_{j.}) \pm t_{(N-k),\alpha'}\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}$$

Each confidence interval that does not include zero suggests $\mu_i \neq \mu_j$ at α. Notice here that $\alpha' = \frac{\alpha}{k(k-1)}$.

II. Tukey's Procedure:

i) Equal Sample Sizes: Let n_i = n, i = 1,. . . ,k and let Q_{α,ν_1,ν_2} be a critical value of the Studentized Range Distribution. The set of cofidence intervals with end points

$$(\bar{y}_{i.} - \bar{y}_{j.}) \pm Q_{\alpha,k,N-k}\sqrt{\frac{MSE}{n}}\text{for all i and j, i} \neq \text{j}$$

is a collection of simultaneous $100(1-\alpha)$% confidence intervals for the differences between the true treatment means, $\mu_i - \mu_j$. Each confidence interval that does not include zero suggests $\mu_i \neq \mu_j$ at α.

ii) Unequal Sample sizes: The set of confidence intervals with endpoints

$$(\bar{y}_{i.} - \bar{y}_{j.}) \pm \frac{1}{\sqrt{2}}Q_{\alpha,k,N-k}\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}\text{for all i and j, i} \neq \text{j}$$

is a collection of simultaneous $100(1-\alpha)$% confidence intervals for the differences between the true treatment means, $\mu_i - \mu_j$. Similarly, each confidence interval that does not include zero suggests $\mu_i \neq \mu_j$ at α.

III. Duncan's Multiple Range Procedure:
Let *n*=n_i, i

i) Linearly order the k sample means (smallest to largest).

ii) Find the value of the least significant studentized range r_p,for p = 2,3,. . . ,k. Table XI,γ denotes the number of degrees of freedom associated with the MSE.

iii) For each $p = 2, 3, \ldots, k$ find the shortest significant range, SSR_p. This value is given by

(a) Equal sample sizes:

$$SSR_p = r_p\sqrt{\frac{MSE}{n}}$$

(b) Unequal sample sizes:

$$SSR_p == r_p\sqrt{MSE}$$

iv) Consider any subset of *p* adjacent sample means. Let $|\bar{y}_{i.} - \bar{y}_{j.}|$ denote the range of the means in this subgroup. Hence $\mu_i \neq \mu_j$ if

(a) Equal sample sizes:

$$|\bar{y}_{i.} - \bar{y}_{j.}| > SSR_p$$

(b) Unequal sample sizes:

$$|\bar{y}_{i.} - \bar{y}_{j.}|\sqrt{\frac{2n_in_j}{n_i + n_j}} > SSR_p$$

v) Summaring your results by underlining any subset of adjacent samples means that are not considered significantly different.

IV. Dunnett's Procedure:

Let *n* = n_i, *i* = 1,. . . ,*k* and let d_{α,ν_1,ν_2} be a critical value for Dunnett's procedure and let treatment O be the control group. Dunnett's procedure for determining significant differences between each treatment and control at the joint significance level α is given by:

H_0: $\mu_0 = \mu_i$ *i* = 1,. . . ,*k*

H_a : $\mu_0 > \mu_i$

$\mu_0 < \mu_i$ *i* = 1,. . . ,*k*

$\mu_0 \neq \mu_i$

Test statistic: D_i = $\frac{\bar{Y}_{i\cdot} - \bar{Y}_{0\cdot}}{\sqrt{2\hat{\sigma}^2/n}}$ *i* = 1,. . . ,*k*

Rejection Region:

$D_i \geq d_{\alpha,k,k(n-1)}$

$D_i \leq -d_{\alpha,k,k(n-1)}$ *i* = 1,. . . ,*k*

$\mid D_i \mid\geq d_{\alpha,k,k(n-1)}$

20.6 The Randomized Complete Block Design

20.7 Factorial Designs

And finally, we can combine the two factors and create a two-factor model

$$y_{ijk} \;=\; \mu_i + \mu_j + \epsilon_{ijk} \quad i = 1, 2, \ldots, r_1, \quad j = 1, 2, \ldots, r_2, \quad k = 1, 2, \ldots, n$$

More specifically, for our example this would be:

$$y_{ijk} \;=\; \mu_i + \mu_j + \epsilon_{ijk} \quad i = 1, 2, \ldots, 3, \quad j = 1, 2, \quad k = 1, 2, \ldots, 5$$

Notice that $2\times3\times5$ represents the 30 observations. Finally, we can form the interaction model:

$$y_{ijk} \;=\; \mu_i + \mu_j + \mu_{ij} + \epsilon_{ijk} \quad i = 1, 2, \ldots, 3, \quad j = 1, 2, \quad k = 1, 2, \ldots, 5$$

For our example recall interactions were of interest. The usual form of the interaction model is:

$$y_{ijk} \;=\; \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \quad i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, b, \quad k = 1, 2, \ldots, n$$

where the μ corresponds to an overall mean, α corresponds to the “effect” of the first factor, call it Factor A, measured at *a* levels, β is the “effect” of the second factor, factor B, measured at *b* levels, and $\alpha\beta$ corresponds to the interaction term. The general form the analysis of variance table in this two-factor interaction model is given below.

Source of Variation	*Degrees of freedom*	*Sum of squares*	*Mean Square*	*F-ratio*
Factor A	a - 1	SS$_A$=$bn\sum_{i=1}^{a}\left(\overline{y}_{i..}-\overline{y}_{...}\right)^2$	MS$_A$=$\frac{SS_A}{a-1}$	$f \;=\; \frac{MS_A}{MS_E}$
Factor B	b - 1	SS$_B$=$an\sum_{j=1}^{b}\left(\overline{y}_{.j.}-\overline{y}_{...}\right)^2$	MS$_B$=$\frac{SS_B}{b-1}$	$f \;=\; \frac{MS_B}{MS_E}$
AB interaction	(a-1)(b-1)	SS$_{AB}$=$n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{y}_{ij.}-\overline{y}_{i..}-\overline{y}_{.j.}+\overline{y}_{...}\right)^2$	MS$_{AB}$=$\frac{SS_{AB}}{(a-1)(b-1)}$	$f \;=\; \frac{MS_{AB}}{MS_E}$
Experimental Error	ab(n-1)	SS$_E$=$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\overline{y}_{ijk}-\overline{y}_{ij.}\right)^2$	MS$_E$=$\frac{SS_E}{ab(n-1)}$	
Total	rn - 1	SS$_T$=$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\overline{y}_{ijk}-\overline{y}_{...}\right)^2$		

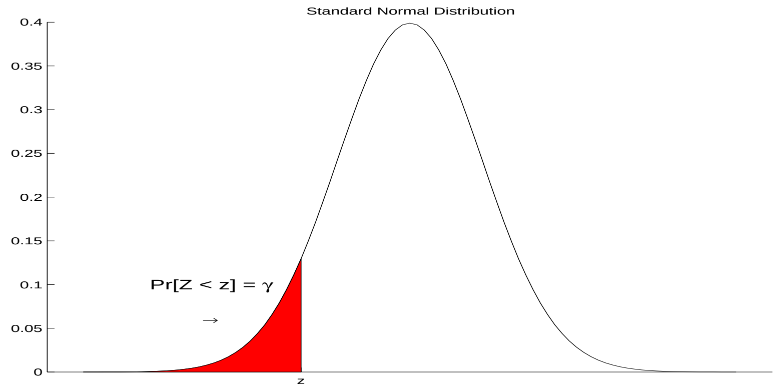
```
STUDENT EDITION OF STATISTIX

ANALYSIS OF VARIANCE TABLE FOR DISTANCE

SOURCE          DF       SS          MS          F        P
------------   ----  ----------  ----------  -------  ------
PROP (A)          2    5980.47     2990.23    122.63  0.0000
CHARGE (B)        1    1414.53     1414.53     58.01  0.0000
A*B               2    61.2667     30.6333      1.26  0.3028
RESIDUAL         24    585.200     24.3833
------------   ----  ----------
TOTAL            29    8041.47
```

21 Tables

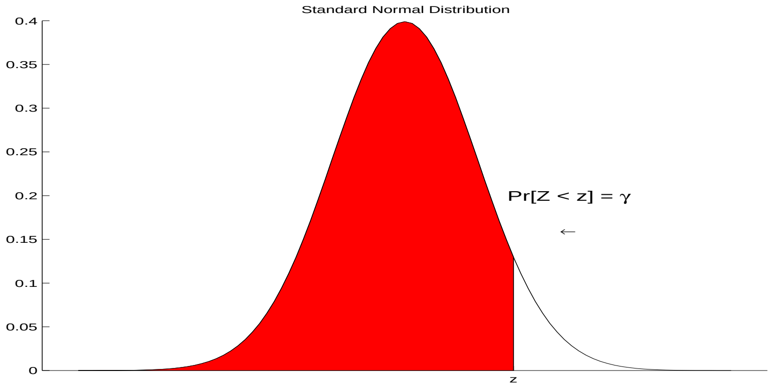
21.1 Cumulative Standard Normal Distribution tables



Pr$\{$Z $\leq$ z$\}$
Table entry for z is the probability γ
lying below z (ie. cumulative probabilities)

Table 16: Cumulative Standard Normal distribution probabilities

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641



Pr{Z ≤ z}

Table entry for z is the probability γ
lying below z (ie. cumulative probabilities)

Table 16: Standard Normal distribution probabilities continued

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

$\mathbf{Pr}\{\mathbf{Z} \leq \mathbf{z}\}$ **-** Table entry is the critical value z below which probability p lies under the curve†

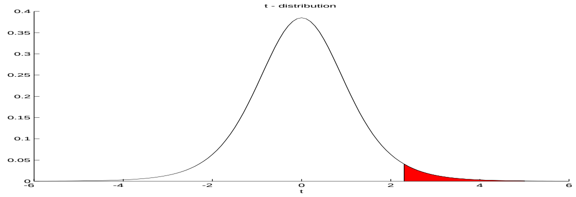
Table 17: Standard Normal Distribution Quantiles

p	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.50	0.000	0.003	0.005	0.008	0.010	0.013	0.015	0.018	0.020	0.023
0.51	0.025	0.028	0.030	0.033	0.035	0.038	0.040	0.043	0.045	0.048
0.52	0.050	0.053	0.055	0.058	0.060	0.063	0.065	0.068	0.070	0.073
0.53	0.075	0.078	0.080	0.083	0.085	0.088	0.090	0.093	0.095	0.098
0.54	0.100	0.103	0.105	0.108	0.111	0.113	0.116	0.118	0.121	0.123
0.55	0.126	0.128	0.131	0.133	0.136	0.138	0.141	0.143	0.146	0.148
0.56	0.151	0.154	0.156	0.159	0.161	0.164	0.166	0.169	0.171	0.174
0.57	0.176	0.179	0.181	0.184	0.187	0.189	0.192	0.194	0.197	0.199
0.58	0.202	0.204	0.207	0.210	0.212	0.215	0.217	0.220	0.222	0.225
0.59	0.228	0.230	0.233	0.235	0.238	0.240	0.243	0.246	0.248	0.251
0.60	0.253	0.256	0.259	0.261	0.264	0.266	0.269	0.272	0.274	0.277
0.61	0.279	0.282	0.285	0.287	0.290	0.292	0.295	0.298	0.300	0.303
0.62	0.305	0.308	0.311	0.313	0.316	0.319	0.321	0.324	0.327	0.329
0.63	0.332	0.335	0.337	0.340	0.342	0.345	0.348	0.350	0.353	0.356
0.64	0.358	0.361	0.364	0.366	0.369	0.372	0.375	0.377	0.380	0.383
0.65	0.385	0.388	0.391	0.393	0.396	0.399	0.402	0.404	0.407	0.410
0.66	0.412	0.415	0.418	0.421	0.423	0.426	0.429	0.432	0.434	0.437
0.67	0.440	0.443	0.445	0.448	0.451	0.454	0.457	0.459	0.462	0.465
0.68	0.468	0.470	0.473	0.476	0.479	0.482	0.485	0.487	0.490	0.493
0.69	0.496	0.499	0.502	0.504	0.507	0.510	0.513	0.516	0.519	0.522
0.70	0.524	0.527	0.530	0.533	0.536	0.539	0.542	0.545	0.548	0.550
0.71	0.553	0.556	0.559	0.562	0.565	0.568	0.571	0.574	0.577	0.580
0.72	0.583	0.586	0.589	0.592	0.595	0.598	0.601	0.604	0.607	0.610
0.73	0.613	0.616	0.619	0.622	0.625	0.628	0.631	0.634	0.637	0.640
0.74	0.643	0.646	0.650	0.653	0.656	0.659	0.662	0.665	0.668	0.671
0.75	0.674	0.678	0.681	0.684	0.687	0.690	0.693	0.697	0.700	0.703
0.76	0.706	0.710	0.713	0.716	0.719	0.722	0.726	0.729	0.732	0.736
0.77	0.739	0.742	0.745	0.749	0.752	0.755	0.759	0.762	0.765	0.769
0.78	0.772	0.776	0.779	0.782	0.786	0.789	0.793	0.796	0.800	0.803
0.79	0.806	0.810	0.813	0.817	0.820	0.824	0.827	0.831	0.834	0.838
0.80	0.842	0.845	0.849	0.852	0.856	0.860	0.863	0.867	0.871	0.874
0.81	0.878	0.882	0.885	0.889	0.893	0.896	0.900	0.904	0.908	0.912
0.82	0.915	0.919	0.923	0.927	0.931	0.935	0.938	0.942	0.946	0.950
0.83	0.954	0.958	0.962	0.966	0.970	0.974	0.978	0.982	0.986	0.990
0.84	0.994	0.999	1.003	1.007	1.011	1.015	1.019	1.024	1.028	1.032
0.85	1.036	1.041	1.045	1.049	1.054	1.058	1.063	1.067	1.071	1.076
0.86	1.080	1.085	1.089	1.094	1.098	1.103	1.108	1.112	1.117	1.122
0.87	1.126	1.131	1.136	1.141	1.146	1.150	1.155	1.160	1.165	1.170
0.88	1.175	1.180	1.185	1.190	1.195	1.200	1.206	1.211	1.216	1.221
0.89	1.227	1.232	1.237	1.243	1.248	1.254	1.259	1.265	1.270	1.276
0.90	1.282	1.287	1.293	1.299	1.305	1.311	1.317	1.323	1.329	1.335
0.91	1.341	1.347	1.353	1.359	1.366	1.372	1.379	1.385	1.392	1.398
0.92	1.405	1.412	1.419	1.426	1.433	1.440	1.447	1.454	1.461	1.468
0.93	1.476	1.483	1.491	1.499	1.506	1.514	1.522	1.530	1.538	1.546
0.94	1.555	1.563	1.572	1.580	1.589	1.598	1.607	1.616	1.626	1.635
0.95	1.645	1.655	1.665	1.675	1.685	1.695	1.706	1.717	1.728	1.739
0.96	1.751	1.762	1.774	1.787	1.799	1.812	1.825	1.838	1.852	1.866
0.97	1.881	1.896	1.911	1.927	1.943	1.960	1.977	1.995	2.014	2.034
0.98	2.054	2.075	2.097	2.120	2.144	2.170	2.197	2.226	2.257	2.290
0.99	2.326	2.366	2.409	2.457	2.512	2.576	2.652	2.748	2.878	3.090

Table 18: Standard Normal Critical Values

α	Z_α
0.1	1.2816
0.05	1.6449
0.025	1.96
0.01	2.3263
0.005	2.5758
0.001	3.0902
0.0005	3.2905
0.0001	3.719
0.00009	3.7455
0.00008	3.775
0.00007	3.8082
0.00006	3.8461
0.00005	3.8906
0.00004	3.9444
0.00003	4.0128
0.00002	4.1075
0.00001	4.2649

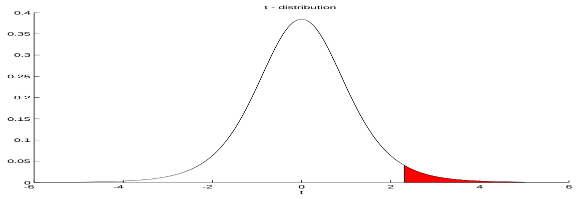
21.2 t-distribution tables



Pr$\{\mathbf{T} \geq t^*\}$
Table entry is the critical value t^*
above which probability α
lies under the curve

Table 19: t-distribution critical values

	α							
df	0.4	0.3	0.25	0.2	0.15	0.1	0.05	0.025
1	0.325	0.727	1.000	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	0.816	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.765	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.741	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.727	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.718	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.711	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.706	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.703	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.700	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.697	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.695	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.694	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.692	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.691	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.690	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.689	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.688	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.688	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.687	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.686	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.686	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.685	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.685	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.684	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.684	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.684	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.683	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.683	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.683	0.854	1.055	1.310	1.697	2.042
35	0.255	0.529	0.682	0.852	1.052	1.306	1.690	2.030
40	0.255	0.529	0.681	0.851	1.050	1.303	1.684	2.021
45	0.255	0.528	0.680	0.850	1.049	1.301	1.679	2.014
50	0.255	0.528	0.679	0.849	1.047	1.299	1.676	2.009
55	0.255	0.527	0.679	0.848	1.046	1.297	1.673	2.004
60	0.254	0.527	0.679	0.848	1.046	1.296	1.671	2.000
65	0.254	0.527	0.678	0.847	1.045	1.295	1.669	1.997
70	0.254	0.527	0.678	0.847	1.044	1.294	1.667	1.994
75	0.254	0.527	0.678	0.846	1.044	1.293	1.665	1.992
80	0.254	0.526	0.678	0.846	1.043	1.292	1.664	1.990
85	0.254	0.526	0.677	0.846	1.043	1.292	1.663	1.988
90	0.254	0.526	0.677	0.846	1.042	1.291	1.662	1.987
95	0.254	0.526	0.677	0.845	1.042	1.291	1.661	1.985
100	0.254	0.526	0.677	0.845	1.042	1.290	1.660	1.984
105	0.254	0.526	0.677	0.845	1.042	1.290	1.659	1.983
110	0.254	0.526	0.677	0.845	1.041	1.289	1.659	1.982
115	0.254	0.526	0.677	0.845	1.041	1.289	1.658	1.981
120	0.254	0.526	0.677	0.845	1.041	1.289	1.658	1.980
1000	0.253	0.525	0.675	0.842	1.037	1.282	1.646	1.960



Pr$\{$**T** $\geq t^*\}$
Table entry is the critical value t^*
above which probability α
lies under the curve

Table 19: t-distribution critical values continued

	α						
df	0.02	0.01	0.005	0.0025	0.001	0.0005	0.0001
1	15.890	31.821	63.657	127.300	318.309	636.600	3183.099
2	4.849	6.965	9.925	14.090	22.327	31.600	70.700
3	3.482	4.541	5.841	7.453	10.215	12.920	22.204
4	2.999	3.747	4.604	5.598	7.173	8.610	13.034
5	2.757	3.365	4.032	4.773	5.893	6.869	9.678
6	2.612	3.143	3.707	4.317	5.208	5.959	8.025
7	2.517	2.998	3.499	4.029	4.785	5.408	7.063
8	2.449	2.896	3.355	3.833	4.501	5.041	6.442
9	2.398	2.821	3.250	3.690	4.297	4.781	6.010
10	2.359	2.764	3.169	3.581	4.144	4.587	5.694
11	2.328	2.718	3.106	3.497	4.025	4.437	5.453
12	2.303	2.681	3.055	3.428	3.930	4.318	5.263
13	2.282	2.650	3.012	3.372	3.852	4.221	5.111
14	2.264	2.624	2.977	3.326	3.787	4.140	4.985
15	2.249	2.602	2.947	3.286	3.733	4.073	4.880
16	2.235	2.583	2.921	3.252	3.686	4.015	4.791
17	2.224	2.567	2.898	3.222	3.646	3.965	4.714
18	2.214	2.552	2.878	3.197	3.610	3.922	4.648
19	2.205	2.539	2.861	3.174	3.579	3.883	4.590
20	2.197	2.528	2.845	3.153	3.552	3.850	4.539
21	2.189	2.518	2.831	3.135	3.527	3.819	4.493
22	2.183	2.508	2.819	3.119	3.505	3.792	4.452
23	2.177	2.500	2.807	3.104	3.485	3.768	4.415
24	2.172	2.492	2.797	3.091	3.467	3.745	4.382
25	2.167	2.485	2.787	3.078	3.450	3.725	4.352
26	2.162	2.479	2.779	3.067	3.435	3.707	4.324
27	2.158	2.473	2.771	3.057	3.421	3.690	4.299
28	2.154	2.467	2.763	3.047	3.408	3.674	4.275
29	2.150	2.462	2.756	3.038	3.396	3.659	4.254
30	2.147	2.457	2.750	3.030	3.385	3.646	4.234
35	2.133	2.438	2.724	2.996	3.340	3.591	4.153
40	2.123	2.423	2.704	2.971	3.307	3.551	4.094
45	2.115	2.412	2.690	2.952	3.281	3.520	4.049
50	2.109	2.403	2.678	2.937	3.261	3.496	4.014
55	2.104	2.396	2.668	2.925	3.245	3.476	3.986
60	2.099	2.390	2.660	2.915	3.232	3.460	3.962
65	2.096	2.385	2.654	2.906	3.220	3.447	3.942
70	2.093	2.381	2.648	2.899	3.211	3.435	3.926
75	2.090	2.377	2.643	2.892	3.202	3.425	3.911
80	2.088	2.374	2.639	2.887	3.195	3.416	3.899
85	2.086	2.371	2.635	2.882	3.189	3.409	3.888
90	2.084	2.368	2.632	2.878	3.183	3.402	3.878
95	2.082	2.366	2.629	2.874	3.178	3.396	3.869
100	2.081	2.364	2.626	2.871	3.174	3.391	3.862
105	2.080	2.362	2.623	2.868	3.170	3.386	3.855
110	2.078	2.361	2.621	2.865	3.166	3.381	3.848
115	2.077	2.359	2.619	2.862	3.163	3.377	3.843
120	2.076	2.358	2.617	2.860	3.160	3.374	3.837
1000	2.056	2.330	2.581	2.813	3.098	3.300	3.733

21.3 Tables of the χ^2-distribution

Pr$\{\chi^2 \geq x^2\}$ - Table entry is the critical value x^2 above which probability α lies under the curve (upper tail probabilities)

Table 20: Quantiles of the χ^2_ν- distribution

	α							
ν	.9999	.9995	.999	.995	.99	.975	.95	.90
1	1.57E-08	3.93E-07	1.57E-06	3.93E-05	0.0002	0.0010	0.0039	0.0158
2	0.0002	0.0010	0.0020	0.0100	0.0201	0.0506	0.1026	0.2107
3	0.0052	0.0153	0.0243	0.0717	0.1148	0.2158	0.3518	0.5844
4	0.0284	0.0639	0.0908	0.2070	0.2971	0.4844	0.7107	1.0636
5	0.0822	0.1581	0.2102	0.4117	0.5543	0.8312	1.1455	1.6103
6	0.1724	0.2994	0.3811	0.6757	0.8721	1.2373	1.6354	2.2041
7	0.3000	0.4849	0.5985	0.9893	1.2390	1.6899	2.1673	2.8331
8	0.4636	0.7104	0.8571	1.3444	1.6465	2.1797	2.7326	3.4895
9	0.6608	0.9717	1.1519	1.7349	2.0879	2.7004	3.3251	4.1682
10	0.8889	1.2650	1.4787	2.1559	2.5582	3.2470	3.9403	4.8652
11	1.1453	1.5868	1.8339	2.6032	3.0535	3.8157	4.5748	5.5778
12	1.4275	1.9344	2.2142	3.0738	3.5706	4.4038	5.2260	6.3038
13	1.7333	2.3051	2.6172	3.5650	4.1069	5.0088	5.8919	7.0415
14	2.0608	2.6967	3.0407	4.0747	4.6604	5.6287	6.5706	7.7895
15	2.4082	3.1075	3.4827	4.6009	5.2293	6.2621	7.2609	8.5468
16	2.7739	3.5358	3.9416	5.1422	5.8122	6.9077	7.9616	9.3122
17	3.1567	3.9802	4.4161	5.6972	6.4078	7.5642	8.6718	10.0852
18	3.5552	4.4394	4.9048	6.2648	7.0149	8.2307	9.3905	10.8649
19	3.9683	4.9123	5.4068	6.8440	7.6327	8.9065	10.1170	11.6509
20	4.3952	5.3981	5.9210	7.4338	8.2604	9.5908	10.8508	12.4426
21	4.8348	5.8957	6.4467	8.0337	8.8972	10.2829	11.5913	13.2396
22	5.2865	6.4045	6.9830	8.6427	9.5425	10.9823	12.3380	14.0415
23	5.7494	6.9237	7.5292	9.2604	10.1957	11.6886	13.0905	14.8480
24	6.2230	7.4527	8.0849	9.8862	10.8564	12.4012	13.8484	15.6587
25	6.7066	7.9910	8.6493	10.5197	11.5240	13.1197	14.6114	16.4734
26	7.1998	8.5379	9.2221	11.1602	12.1981	13.8439	15.3792	17.2919
27	7.7019	9.0932	9.8028	11.8076	12.8785	14.5734	16.1514	18.1139
28	8.2126	9.6563	10.3909	12.4613	13.5647	15.3079	16.9279	18.9392
29	8.7315	10.2268	10.9861	13.1211	14.2565	16.0471	17.7084	19.7677
30	9.2581	10.8044	11.5880	13.7867	14.9535	16.7908	18.4927	20.5992
35	11.9957	13.7875	14.6878	17.1918	18.5089	20.5694	22.4650	24.7967
40	14.8831	16.9062	17.9164	20.7065	22.1643	24.4330	26.5093	29.0505
45	17.8940	20.1366	21.2507	24.3110	25.9013	28.3662	30.6123	33.3504
50	21.0093	23.4610	24.6739	27.9907	29.7067	32.3574	34.7643	37.6886
55	24.2141	26.8658	28.1731	31.7348	33.5705	36.3981	38.9580	42.0596
60	27.4969	30.3405	31.7383	35.5345	37.4849	40.4817	43.1880	46.4589
65	30.8483	33.8767	35.3616	39.3831	41.4436	44.6030	47.4496	50.8829
70	34.2607	37.4674	39.0364	43.2752	45.4417	48.7576	51.7393	55.3289
75	37.7279	41.1072	42.7573	47.2060	49.4750	52.9419	56.0541	59.7946
80	41.2445	44.7910	46.5199	51.1719	53.5401	57.1532	60.3915	64.2778
85	44.8060	48.5151	50.3203	55.1696	57.6339	61.3888	64.7494	68.7772
90	48.4087	52.2758	54.1552	59.1963	61.7541	65.6466	69.1260	73.2911
95	52.0492	56.0702	58.0220	63.2496	65.8984	69.9249	73.5198	77.8184
100	55.7246	59.8957	61.9179	67.3276	70.0649	74.2219	77.9295	82.3581
105	59.4323	63.7499	65.8411	71.4282	74.2520	78.5364	82.3537	86.9093
110	63.1701	67.6310	69.7894	75.5500	78.4583	82.8671	86.7916	91.4710
115	66.9360	71.5371	73.7613	79.6916	82.6824	87.2128	91.2422	96.0427
120	70.7281	75.4665	77.7551	83.8516	86.9233	91.5726	95.7046	100.6236

Pr$\{\chi^2 \geq x^2\}$ - Table entry is the critical value x^2 above which probability α lies under the curve (upper tail probabilities)

Table 20: Quantiles of the χ^2_ν- distribution continued

	α							
ν	.10	.05	.025	.01	.005	.001	.0005	.0001
1	2.7055	3.8415	5.0239	6.6349	7.8794	10.8276	12.1157	15.1367
2	4.6052	5.9915	7.3778	9.2103	10.5966	13.8155	15.2018	18.4207
3	6.2514	7.8147	9.3484	11.3449	12.8382	16.2662	17.7300	21.1075
4	7.7794	9.4877	11.1433	13.2767	14.8603	18.4668	19.9974	23.5127
5	9.2364	11.0705	12.8325	15.0863	16.7496	20.5150	22.1053	25.7448
6	10.6446	12.5916	14.4494	16.8119	18.5476	22.4577	24.1028	27.8563
7	12.0170	14.0671	16.0128	18.4753	20.2777	24.3219	26.0178	29.8775
8	13.3616	15.5073	17.5345	20.0902	21.9550	26.1245	27.8680	31.8276
9	14.6837	16.9190	19.0228	21.6660	23.5894	27.8772	29.6658	33.7199
10	15.9872	18.3070	20.4832	23.2093	25.1882	29.5883	31.4198	35.5640
11	17.2750	19.6751	21.9200	24.7250	26.7568	31.2641	33.1366	37.3670
12	18.5493	21.0261	23.3367	26.2170	28.2995	32.9095	34.8213	39.1344
13	19.8119	22.3620	24.7356	27.6882	29.8195	34.5282	36.4778	40.8707
14	21.0641	23.6848	26.1189	29.1412	31.3193	36.1233	38.1094	42.5793
15	22.3071	24.9958	27.4884	30.5779	32.8013	37.6973	39.7188	44.2632
16	23.5418	26.2962	28.8454	31.9999	34.2672	39.2524	41.3081	45.9249
17	24.7690	27.5871	30.1910	33.4087	35.7185	40.7902	42.8792	47.5664
18	25.9894	28.8693	31.5264	34.8053	37.1565	42.3124	44.4338	49.1894
19	27.2036	30.1435	32.8523	36.1909	38.5823	43.8202	45.9731	50.7955
20	28.4120	31.4104	34.1696	37.5662	39.9968	45.3147	47.4985	52.3860
21	29.6151	32.6706	35.4789	38.9322	41.4011	46.7970	49.0108	53.9620
22	30.8133	33.9244	36.7807	40.2894	42.7957	48.2679	50.5111	55.5246
23	32.0069	35.1725	38.0756	41.6384	44.1813	49.7282	52.0002	57.0746
24	33.1962	36.4150	39.3641	42.9798	45.5585	51.1786	53.4788	58.6130
25	34.3816	37.6525	40.6465	44.3141	46.9279	52.6197	54.9475	60.1403
26	35.5632	38.8851	41.9232	45.6417	48.2899	54.0520	56.4069	61.6573
27	36.7412	40.1133	43.1945	46.9629	49.6449	55.4760	57.8576	63.1645
28	37.9159	41.3371	44.4608	48.2782	50.9934	56.8923	59.3000	64.6624
29	39.0875	42.5570	45.7223	49.5879	52.3356	58.3012	60.7346	66.1517
30	40.2560	43.7730	46.9792	50.8922	53.6720	59.7031	62.1619	67.6326
35	46.0588	49.8018	53.2033	57.3421	60.2748	66.6188	69.1986	74.9262
40	51.8051	55.7585	59.3417	63.6907	66.7660	73.4020	76.0946	82.0623
45	57.5053	61.6562	65.4102	69.9568	73.1661	80.0767	82.8757	89.0695
50	63.1671	67.5048	71.4202	76.1539	79.4900	86.6608	89.5605	95.9687
55	68.7962	73.3115	77.3805	82.2921	85.7490	93.1675	96.1632	102.7758
60	74.3970	79.0819	83.2977	88.3794	91.9517	99.6072	102.6948	109.5029
65	79.9730	84.8206	89.1771	94.4221	98.1051	105.9881	109.1639	116.1599
70	85.5270	90.5312	95.0232	100.4252	104.2149	112.3169	115.5776	122.7547
75	91.0615	96.2167	100.8393	106.3929	110.2856	118.5991	121.9418	129.2937
80	96.5782	101.8795	106.6286	112.3288	116.3211	124.8392	128.2613	135.7825
85	102.0789	107.5217	112.3934	118.2357	122.3246	131.0412	134.5403	142.2257
90	107.5650	113.1453	118.1359	124.1163	128.2989	137.2084	140.7823	148.6273
95	113.0377	118.7516	123.8580	129.9727	134.2465	143.3435	146.9903	154.9906
100	118.4980	124.3421	129.5612	135.8067	140.1695	149.4493	153.1670	161.3187
105	123.9469	129.9180	135.2470	141.6201	146.0696	155.5277	159.3146	167.6140
110	129.3851	135.4802	140.9166	147.4143	151.9485	161.5807	165.4353	173.8791
115	134.8135	141.0297	146.5711	153.1906	157.8076	167.6102	171.5309	180.1158
120	140.2326	146.5674	152.2114	158.9502	163.6482	173.6174	177.6029	186.3260

21.4 F-distribution tables

Table E: F-critical values

		Degrees of freedom in the numerator								
DFD	p	1	2	3	4	5	6	7	8	9
	0.1	39.86	49.5	53.59	55.83	57.24	58.2	58.91	59.44	59.86
	0.05	161.45	199.5	215.71	224.58	230.16	233.99	236.77	238.88	240.54
1	0.025	647.79	799.5	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	0.01	4052.2	4999.5	5403.4	5624.6	5763.6	5859	5928.4	5981.1	6022.5
	0.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
	0.1	8.53	9	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	0.05	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38
2	0.025	38.51	39	39.17	39.25	39.3	39.33	39.36	39.37	39.39
	0.01	98.5	99	99.17	99.25	99.3	99.33	99.36	99.37	99.39
	0.001	998.5	999	999.17	999.25	999.3	999.33	999.36	999.37	999.39
	0.1	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
3	0.025	17.44	16.04	15.44	15.1	14.88	14.73	14.62	14.54	14.47
	0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	0.001	167.03	148.5	141.11	137.1	134.58	132.85	131.58	130.62	129.86
	0.1	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
4	0.025	12.22	10.65	9.98	9.6	9.36	9.2	9.07	8.98	8.90
	0.01	21.2	18	16.69	15.98	15.52	15.21	14.98	14.8	14.66
	0.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49	48.47
	0.1	4.06	3.78	3.62	3.52	3.45	3.4	3.37	3.34	3.32
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
5	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	0.001	47.18	37.12	33.2	31.09	29.75	28.83	28.16	27.65	27.24
	0.1	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
6	0.025	8.81	7.26	6.6	6.23	5.99	5.82	5.7	5.6	5.52
	0.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.1	7.98
	0.001	35.51	27	23.7	21.92	20.8	20.03	19.46	19.03	18.69
	0.1	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
7	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.9	4.82
	0.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	0.001	29.25	21.69	18.77	17.2	16.21	15.52	15.02	14.63	14.33
	0.1	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39
8	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	0.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	0.001	25.41	18.49	15.83	14.39	13.48	12.86	12.4	12.05	11.77
	0.1	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
9	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.2	4.1	4.03
	0.01	10.56	8.02	6.99	6.42	6.06	5.8	5.61	5.47	5.35
	0.001	22.86	16.39	13.9	12.56	11.71	11.13	10.7	10.37	10.11
	0.1	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	0.05	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02
10	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	0.01	10.04	7.56	6.55	5.99	5.64	5.39	5.2	5.06	4.94
	0.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.2	8.96

Table E: F-critical values

		Degrees of freedom in the numerator										
DFD	p	10	12	15	20	25	30	40	50	60	120	1000
	0.1	60.19	60.71	61.22	61.74	62.05	62.26	62.53	62.69	62.79	63.06	63.30
	0.05	241.88	243.91	245.95	248.01	249.26	250.1	251.14	251.77	252.2	253.25	254.19
1	0.025	968.63	976.71	984.87	993.1	998.08	1001.4	1005.6	1008.1	1009.8	1014	1017.7
	0.01	6055.8	6106.3	6157.3	6208.7	6239.8	6260.6	6286.8	6302.5	6313	6339.4	6362.7
	0.001	605621	610668	615764	620908	624017	626099	628712	630285	631337	633972	636301
	0.1	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.48	9.49
	0.05	19.4	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.48	19.49	19.49
2	0.025	39.4	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.48	39.49	39.50
	0.01	99.4	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.48	99.49	99.50
	0.001	999.4	999.42	999.43	999.45	999.46	999.47	999.47	999.48	999.48	999.49	999.50
	0.1	5.23	5.22	5.2	5.18	5.17	5.17	5.16	5.15	5.15	5.14	5.13
	0.05	8.79	8.74	8.7	8.66	8.63	8.62	8.59	8.58	8.57	8.55	8.53
3	0.025	14.42	14.34	14.25	14.17	14.12	14.08	14.04	14.01	13.99	13.95	13.91
	0.01	27.23	27.05	26.87	26.69	26.58	26.5	26.41	26.35	26.32	26.22	26.14
	0.001	129.25	128.32	127.37	126.42	125.84	125.45	124.96	124.66	124.47	123.97	123.53
	0.1	3.92	3.9	3.87	3.84	3.83	3.82	3.8	3.8	3.79	3.78	3.76
	0.05	5.96	5.91	5.86	5.8	5.77	5.75	5.72	5.7	5.69	5.66	5.63
4	0.025	8.84	8.75	8.66	8.56	8.5	8.46	8.41	8.38	8.36	8.31	8.26
	0.01	14.55	14.37	14.2	14.02	13.91	13.84	13.75	13.69	13.65	13.56	13.47
	0.001	48.05	47.41	46.76	46.1	45.7	45.43	45.09	44.88	44.75	44.4	44.09
	0.1	3.3	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.12	3.11
	0.05	4.74	4.68	4.62	4.56	4.52	4.5	4.46	4.44	4.43	4.4	4.37
5	0.025	6.62	6.52	6.43	6.33	6.27	6.23	6.18	6.14	6.12	6.07	6.02
	0.01	10.05	9.89	9.72	9.55	9.45	9.38	9.29	9.24	9.2	9.11	9.03
	0.001	26.92	26.42	25.91	25.39	25.08	24.87	24.6	24.44	24.33	24.06	23.82
	0.1	2.94	2.9	2.87	2.84	2.81	2.8	2.78	2.77	2.76	2.74	2.72
	0.05	4.06	4	3.94	3.87	3.83	3.81	3.77	3.75	3.74	3.7	3.67
6	0.025	5.46	5.37	5.27	5.17	5.11	5.07	5.01	4.98	4.96	4.9	4.86
	0.01	7.87	7.72	7.56	7.4	7.3	7.23	7.14	7.09	7.06	6.97	6.89
	0.001	18.41	17.99	17.56	17.12	16.85	16.67	16.44	16.31	16.21	15.98	15.77
	0.1	2.7	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.51	2.49	2.47
	0.05	3.64	3.57	3.51	3.44	3.4	3.38	3.34	3.32	3.3	3.27	3.23
7	0.025	4.76	4.67	4.57	4.47	4.4	4.36	4.31	4.28	4.25	4.2	4.15
	0.01	6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.82	5.74	5.66
	0.001	14.08	13.71	13.32	12.93	12.69	12.53	12.33	12.2	12.12	11.91	11.72
	0.1	2.54	2.5	2.46	2.42	2.4	2.38	2.36	2.35	2.34	2.32	2.30
	0.05	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.02	3.01	2.97	2.93
8	0.025	4.3	4.2	4.1	4	3.94	3.89	3.84	3.81	3.78	3.73	3.68
	0.01	5.81	5.67	5.52	5.36	5.26	5.2	5.12	5.07	5.03	4.95	4.87
	0.001	11.54	11.19	10.84	10.48	10.26	10.11	9.92	9.8	9.73	9.53	9.36
	0.1	2.42	2.38	2.34	2.3	2.27	2.25	2.23	2.22	2.21	2.18	2.16
	0.05	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.8	2.79	2.75	2.71
9	0.025	3.96	3.87	3.77	3.67	3.6	3.56	3.51	3.47	3.45	3.39	3.34
	0.01	5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.48	4.4	4.32
	0.001	9.89	9.57	9.24	8.9	8.69	8.55	8.37	8.26	8.19	8	7.84
	0.1	2.32	2.28	2.24	2.2	2.17	2.16	2.13	2.12	2.11	2.08	2.06
	0.05	2.98	2.91	2.85	2.77	2.73	2.7	2.66	2.64	2.62	2.58	2.54
10	0.025	3.72	3.62	3.52	3.42	3.35	3.31	3.26	3.22	3.2	3.14	3.09
	0.01	4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.08	4	3.92
	0.001	8.75	8.45	8.13	7.8	7.6	7.47	7.3	7.19	7.12	6.94	6.78

Table E: F-critical values

		Degrees of freedom in the numerator								
DFD	p	1	2	3	4	5	6	7	8	9
11	0.1	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.3	2.27
	0.05	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.90
	0.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	0.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	0.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12
12	0.1	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	0.05	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.80
	0.025	6.55	5.1	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.5	4.39
	0.001	18.64	12.97	10.8	9.63	8.89	8.38	8	7.71	7.48
13	0.1	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.2	2.16
	0.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	0.025	6.41	4.97	4.35	4	3.77	3.6	3.48	3.39	3.31
	0.01	9.07	6.7	5.74	5.21	4.86	4.62	4.44	4.3	4.19
	0.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98
14	0.1	3.1	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	0.05	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65
	0.025	6.3	4.86	4.24	3.89	3.66	3.5	3.38	3.29	3.21
	0.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	0.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.8	6.58
15	0.1	3.07	2.7	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	0.05	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59
	0.025	6.2	4.77	4.15	3.8	3.58	3.41	3.29	3.2	3.12
	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4	3.89
	0.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26
16	0.1	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	0.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	0.025	6.12	4.69	4.08	3.73	3.5	3.34	3.22	3.12	3.05
	0.01	8.53	6.23	5.29	4.77	4.44	4.2	4.03	3.89	3.78
	0.001	16.12	10.97	9.01	7.94	7.27	6.8	6.46	6.19	5.98
17	0.1	3.03	2.64	2.44	2.31	2.22	2.15	2.1	2.06	2.03
	0.05	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49
	0.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
	0.01	8.4	6.11	5.19	4.67	4.34	4.1	3.93	3.79	3.68
	0.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75
18	0.1	3.01	2.62	2.42	2.29	2.2	2.13	2.08	2.04	2.00
	0.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	0.025	5.98	4.56	3.95	3.61	3.38	3.22	3.1	3.01	2.93
	0.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	0.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
19	0.1	2.99	2.61	2.4	2.27	2.18	2.11	2.06	2.02	1.98
	0.05	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42
	0.025	5.92	4.51	3.9	3.56	3.33	3.17	3.05	2.96	2.88
	0.01	8.18	5.93	5.01	4.5	4.17	3.94	3.77	3.63	3.52
	0.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39
20	0.1	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2	1.96
	0.05	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	0.01	8.1	5.85	4.94	4.43	4.1	3.87	3.7	3.56	3.46
	0.001	14.82	9.95	8.1	7.1	6.46	6.02	5.69	5.44	5.24

Table E: F-critical values

		Degrees of freedom in the numerator										
DFD	p	10	12	15	20	25	30	40	50	60	120	1000
11	0.1	2.25	2.21	2.17	2.12	2.1	2.08	2.05	2.04	2.03	2	1.98
	0.05	2.85	2.79	2.72	2.65	2.6	2.57	2.53	2.51	2.49	2.45	2.41
	0.025	3.53	3.43	3.33	3.23	3.16	3.12	3.06	3.03	3	2.94	2.89
	0.01	4.54	4.4	4.25	4.1	4.01	3.94	3.86	3.81	3.78	3.69	3.61
	0.001	7.92	7.63	7.32	7.01	6.81	6.68	6.52	6.42	6.35	6.18	6.02
12	0.1	2.19	2.15	2.1	2.06	2.03	2.01	1.99	1.97	1.96	1.93	1.91
	0.05	2.75	2.69	2.62	2.54	2.5	2.47	2.43	2.4	2.38	2.34	2.30
	0.025	3.37	3.28	3.18	3.07	3.01	2.96	2.91	2.87	2.85	2.79	2.73
	0.01	4.3	4.16	4.01	3.86	3.76	3.7	3.62	3.57	3.54	3.45	3.37
	0.001	7.29	7	6.71	6.4	6.22	6.09	5.93	5.83	5.76	5.59	5.44
13	0.1	2.14	2.1	2.05	2.01	1.98	1.96	1.93	1.92	1.9	1.88	1.85
	0.05	2.67	2.6	2.53	2.46	2.41	2.38	2.34	2.31	2.3	2.25	2.21
	0.025	3.25	3.15	3.05	2.95	2.88	2.84	2.78	2.74	2.72	2.66	2.60
	0.01	4.1	3.96	3.82	3.66	3.57	3.51	3.43	3.38	3.34	3.25	3.18
	0.001	6.8	6.52	6.23	5.93	5.75	5.63	5.47	5.37	5.3	5.14	4.99
14	0.1	2.1	2.05	2.01	1.96	1.93	1.91	1.89	1.87	1.86	1.83	1.80
	0.05	2.6	2.53	2.46	2.39	2.34	2.31	2.27	2.24	2.22	2.18	2.14
	0.025	3.15	3.05	2.95	2.84	2.78	2.73	2.67	2.64	2.61	2.55	2.50
	0.01	3.94	3.8	3.66	3.51	3.41	3.35	3.27	3.22	3.18	3.09	3.02
	0.001	6.4	6.13	5.85	5.56	5.38	5.25	5.1	5	4.94	4.77	4.62
15	0.1	2.06	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.82	1.79	1.76
	0.05	2.54	2.48	2.4	2.33	2.28	2.25	2.2	2.18	2.16	2.11	2.07
	0.025	3.06	2.96	2.86	2.76	2.69	2.64	2.59	2.55	2.52	2.46	2.40
	0.01	3.8	3.67	3.52	3.37	3.28	3.21	3.13	3.08	3.05	2.96	2.88
	0.001	6.08	5.81	5.54	5.25	5.07	4.95	4.8	4.7	4.64	4.47	4.33
16	0.1	2.03	1.99	1.94	1.89	1.86	1.84	1.81	1.79	1.78	1.75	1.72
	0.05	2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.12	2.11	2.06	2.02
	0.025	2.99	2.89	2.79	2.68	2.61	2.57	2.51	2.47	2.45	2.38	2.32
	0.01	3.69	3.55	3.41	3.26	3.16	3.1	3.02	2.97	2.93	2.84	2.76
	0.001	5.81	5.55	5.27	4.99	4.82	4.7	4.54	4.45	4.39	4.23	4.08
17	0.1	2	1.96	1.91	1.86	1.83	1.81	1.78	1.76	1.75	1.72	1.69
	0.05	2.45	2.38	2.31	2.23	2.18	2.15	2.1	2.08	2.06	2.01	1.97
	0.025	2.92	2.82	2.72	2.62	2.55	2.5	2.44	2.41	2.38	2.32	2.26
	0.01	3.59	3.46	3.31	3.16	3.07	3	2.92	2.87	2.83	2.75	2.66
	0.001	5.58	5.32	5.05	4.78	4.6	4.48	4.33	4.24	4.18	4.02	3.87
18	0.1	1.98	1.93	1.89	1.84	1.8	1.78	1.75	1.74	1.72	1.69	1.66
	0.05	2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.04	2.02	1.97	1.92
	0.025	2.87	2.77	2.67	2.56	2.49	2.44	2.38	2.35	2.32	2.26	2.20
	0.01	3.51	3.37	3.23	3.08	2.98	2.92	2.84	2.78	2.75	2.66	2.58
	0.001	5.39	5.13	4.87	4.59	4.42	4.3	4.15	4.06	4	3.84	3.69
19	0.1	1.96	1.91	1.86	1.81	1.78	1.76	1.73	1.71	1.7	1.67	1.64
	0.05	2.38	2.31	2.23	2.16	2.11	2.07	2.03	2	1.98	1.93	1.88
	0.025	2.82	2.72	2.62	2.51	2.44	2.39	2.33	2.3	2.27	2.2	2.14
	0.01	3.43	3.3	3.15	3	2.91	2.84	2.76	2.71	2.67	2.58	2.50
	0.001	5.22	4.97	4.7	4.43	4.26	4.14	3.99	3.9	3.84	3.68	3.53
20	0.1	1.94	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.68	1.64	1.61
	0.05	2.35	2.28	2.2	2.12	2.07	2.04	1.99	1.97	1.95	1.9	1.85
	0.025	2.77	2.68	2.57	2.46	2.4	2.35	2.29	2.25	2.22	2.16	2.09
	0.01	3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.61	2.52	2.43
	0.001	5.08	4.82	4.56	4.29	4.12	4	3.86	3.77	3.7	3.54	3.40

Table E: F-critical values

		Degrees of freedom in the numerator								
DFD	p	1	2	3	4	5	6	7	8	9
21	0.1	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	0.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	0.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	0.01	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	0.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11
22	0.1	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	0.05	4.3	3.44	3.05	2.82	2.66	2.55	2.46	2.4	2.34
	0.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	0.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	0.001	14.38	9.61	7.8	6.81	6.19	5.76	5.44	5.19	4.99
23	0.1	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	0.05	4.28	3.42	3.03	2.8	2.64	2.53	2.44	2.37	2.32
	0.025	5.75	4.35	3.75	3.41	3.18	3.02	2.9	2.81	2.73
	0.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	0.001	14.2	9.47	7.67	6.7	6.08	5.65	5.33	5.09	4.89
24	0.1	2.93	2.54	2.33	2.19	2.1	2.04	1.98	1.94	1.91
	0.05	4.26	3.4	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	0.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	0.01	7.82	5.61	4.72	4.22	3.9	3.67	3.5	3.36	3.26
	0.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80
25	0.1	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	0.05	4.24	3.39	2.99	2.76	2.6	2.49	2.4	2.34	2.28
	0.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	0.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	0.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71
26	0.1	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	0.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
	0.025	5.66	4.27	3.67	3.33	3.1	2.94	2.82	2.73	2.65
	0.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	0.001	13.74	9.12	7.36	6.41	5.8	5.38	5.07	4.83	4.64
27	0.1	2.9	2.51	2.3	2.17	2.07	2	1.95	1.91	1.87
	0.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	0.025	5.63	4.24	3.65	3.31	3.08	2.92	2.8	2.71	2.63
	0.01	7.68	5.49	4.6	4.11	3.78	3.56	3.39	3.26	3.15
	0.001	13.61	9.02	7.27	6.33	5.73	5.31	5	4.76	4.57
28	0.1	2.89	2.5	2.29	2.16	2.06	2	1.94	1.9	1.87
	0.05	4.2	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
	0.025	5.61	4.22	3.63	3.29	3.06	2.9	2.78	2.69	2.61
	0.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	0.001	13.5	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50
29	0.1	2.89	2.5	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	0.05	4.18	3.33	2.93	2.7	2.55	2.43	2.35	2.28	2.22
	0.025	5.59	4.2	3.61	3.27	3.04	2.88	2.76	2.67	2.59
	0.01	7.6	5.42	4.54	4.04	3.73	3.5	3.33	3.2	3.09
	0.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45
30	0.1	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	0.01	7.56	5.39	4.51	4.02	3.7	3.47	3.3	3.17	3.07
	0.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39

Table E: F-critical values

		Degrees of freedom in the numerator										
DFD	p	10	12	15	20	25	30	40	50	60	120	1000
21	0.1	1.92	1.87	1.83	1.78	1.74	1.72	1.69	1.67	1.66	1.62	1.59
	0.05	2.32	2.25	2.18	2.1	2.05	2.01	1.96	1.94	1.92	1.87	1.82
	0.025	2.73	2.64	2.53	2.42	2.36	2.31	2.25	2.21	2.18	2.11	2.05
	0.01	3.31	3.17	3.03	2.88	2.79	2.72	2.64	2.58	2.55	2.46	2.37
	0.001	4.95	4.7	4.44	4.17	4	3.88	3.74	3.64	3.58	3.42	3.28
22	0.1	1.9	1.86	1.81	1.76	1.73	1.7	1.67	1.65	1.64	1.6	1.57
	0.05	2.3	2.23	2.15	2.07	2.02	1.98	1.94	1.91	1.89	1.84	1.79
	0.025	2.7	2.6	2.5	2.39	2.32	2.27	2.21	2.17	2.14	2.08	2.01
	0.01	3.26	3.12	2.98	2.83	2.73	2.67	2.58	2.53	2.5	2.4	2.32
	0.001	4.83	4.58	4.33	4.06	3.89	3.78	3.63	3.54	3.48	3.32	3.17
23	0.1	1.89	1.84	1.8	1.74	1.71	1.69	1.66	1.64	1.62	1.59	1.55
	0.05	2.27	2.2	2.13	2.05	2	1.96	1.91	1.88	1.86	1.81	1.76
	0.025	2.67	2.57	2.47	2.36	2.29	2.24	2.18	2.14	2.11	2.04	1.98
	0.01	3.21	3.07	2.93	2.78	2.69	2.62	2.54	2.48	2.45	2.35	2.27
	0.001	4.73	4.48	4.23	3.96	3.79	3.68	3.53	3.44	3.38	3.22	3.08
24	0.1	1.88	1.83	1.78	1.73	1.7	1.67	1.64	1.62	1.61	1.57	1.54
	0.05	2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.86	1.84	1.79	1.74
	0.025	2.64	2.54	2.44	2.33	2.26	2.21	2.15	2.11	2.08	2.01	1.94
	0.01	3.17	3.03	2.89	2.74	2.64	2.58	2.49	2.44	2.4	2.31	2.22
	0.001	4.64	4.39	4.14	3.87	3.71	3.59	3.45	3.36	3.29	3.14	2.99
25	0.1	1.87	1.82	1.77	1.72	1.68	1.66	1.63	1.61	1.59	1.56	1.52
	0.05	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.84	1.82	1.77	1.72
	0.025	2.61	2.51	2.41	2.3	2.23	2.18	2.12	2.08	2.05	1.98	1.91
	0.01	3.13	2.99	2.85	2.7	2.6	2.54	2.45	2.4	2.36	2.27	2.18
	0.001	4.56	4.31	4.06	3.79	3.63	3.52	3.37	3.28	3.22	3.06	2.91
26	0.1	1.86	1.81	1.76	1.71	1.67	1.65	1.61	1.59	1.58	1.54	1.51
	0.05	2.22	2.15	2.07	1.99	1.94	1.9	1.85	1.82	1.8	1.75	1.70
	0.025	2.59	2.49	2.39	2.28	2.21	2.16	2.09	2.05	2.03	1.95	1.89
	0.01	3.09	2.96	2.81	2.66	2.57	2.5	2.42	2.36	2.33	2.23	2.14
	0.001	4.48	4.24	3.99	3.72	3.56	3.44	3.3	3.21	3.15	2.99	2.84
27	0.1	1.85	1.8	1.75	1.7	1.66	1.64	1.6	1.58	1.57	1.53	1.50
	0.05	2.2	2.13	2.06	1.97	1.92	1.88	1.84	1.81	1.79	1.73	1.68
	0.025	2.57	2.47	2.36	2.25	2.18	2.13	2.07	2.03	2	1.93	1.86
	0.01	3.06	2.93	2.78	2.63	2.54	2.47	2.38	2.33	2.29	2.2	2.11
	0.001	4.41	4.17	3.92	3.66	3.49	3.38	3.23	3.14	3.08	2.92	2.78
28	0.1	1.84	1.79	1.74	1.69	1.65	1.63	1.59	1.57	1.56	1.52	1.48
	0.05	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.79	1.77	1.71	1.66
	0.025	2.55	2.45	2.34	2.23	2.16	2.11	2.05	2.01	1.98	1.91	1.84
	0.01	3.03	2.9	2.75	2.6	2.51	2.44	2.35	2.3	2.26	2.17	2.08
	0.001	4.35	4.11	3.86	3.6	3.43	3.32	3.18	3.09	3.02	2.86	2.72
29	0.1	1.83	1.78	1.73	1.68	1.64	1.62	1.58	1.56	1.55	1.51	1.47
	0.05	2.18	2.1	2.03	1.94	1.89	1.85	1.81	1.77	1.75	1.7	1.65
	0.025	2.53	2.43	2.32	2.21	2.14	2.09	2.03	1.99	1.96	1.89	1.82
	0.01	3	2.87	2.73	2.57	2.48	2.41	2.33	2.27	2.23	2.14	2.05
	0.001	4.29	4.05	3.8	3.54	3.38	3.27	3.12	3.03	2.97	2.81	2.66
30	0.1	1.82	1.77	1.72	1.67	1.63	1.61	1.57	1.55	1.54	1.5	1.46
	0.05	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.76	1.74	1.68	1.63
	0.025	2.51	2.41	2.31	2.2	2.12	2.07	2.01	1.97	1.94	1.87	1.80
	0.01	2.98	2.84	2.7	2.55	2.45	2.39	2.3	2.25	2.21	2.11	2.02
	0.001	4.24	4	3.75	3.49	3.33	3.22	3.07	2.98	2.92	2.76	2.61

Table E: F-critical values

		Degrees of freedom in the numerator								
DFD	p	1	2	3	4	5	6	7	8	9
40	0.1	2.84	2.44	2.23	2.09	2	1.93	1.87	1.83	1.79
	0.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
	0.025	5.42	4.05	3.46	3.13	2.9	2.74	2.62	2.53	2.45
	0.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	0.001	12.61	8.25	6.59	5.7	5.13	4.73	4.44	4.21	4.02
50	0.1	2.81	2.41	2.2	2.06	1.97	1.9	1.84	1.8	1.76
	0.05	4.03	3.18	2.79	2.56	2.4	2.29	2.2	2.13	2.07
	0.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38
	0.01	7.17	5.06	4.2	3.72	3.41	3.19	3.02	2.89	2.78
	0.001	12.22	7.96	6.34	5.46	4.9	4.51	4.22	4	3.82
60	0.1	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	0.05	4	3.15	2.76	2.53	2.37	2.25	2.17	2.1	2.04
	0.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	0.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	0.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
100	0.1	2.76	2.36	2.14	2	1.91	1.83	1.78	1.73	1.69
	0.05	3.94	3.09	2.7	2.46	2.31	2.19	2.1	2.03	1.97
	0.025	5.18	3.83	3.25	2.92	2.7	2.54	2.42	2.32	2.24
	0.01	6.9	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
	0.001	11.5	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44
200	0.1	2.73	2.33	2.11	1.97	1.88	1.8	1.75	1.7	1.66
	0.05	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
	0.025	5.1	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18
	0.01	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.6	2.50
	0.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26
1000	0.1	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64
	0.05	3.85	3	2.61	2.38	2.22	2.11	2.02	1.95	1.89
	0.025	5.04	3.7	3.13	2.8	2.58	2.42	2.3	2.2	2.13
	0.01	6.66	4.63	3.8	3.34	3.04	2.82	2.66	2.53	2.43
	0.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.3	3.13

Table E: F-critical values

		Degrees of freedom in the numerator										
DFD	p	10	12	15	20	25	30	40	50	60	120	1000
40	0.1	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.42	1.38
	0.05	2.08	2	1.92	1.84	1.78	1.74	1.69	1.66	1.64	1.58	1.52
	0.025	2.39	2.29	2.18	2.07	1.99	1.94	1.88	1.83	1.8	1.72	1.65
	0.01	2.8	2.66	2.52	2.37	2.27	2.2	2.11	2.06	2.02	1.92	1.82
	0.001	3.87	3.64	3.4	3.14	2.98	2.87	2.73	2.64	2.57	2.41	2.25
50	0.1	1.73	1.68	1.63	1.57	1.53	1.5	1.46	1.44	1.42	1.38	1.33
	0.05	2.03	1.95	1.87	1.78	1.73	1.69	1.63	1.6	1.58	1.51	1.45
	0.025	2.32	2.22	2.11	1.99	1.92	1.87	1.8	1.75	1.72	1.64	1.56
	0.01	2.7	2.56	2.42	2.27	2.17	2.1	2.01	1.95	1.91	1.8	1.70
	0.001	3.67	3.44	3.2	2.95	2.79	2.68	2.53	2.44	2.38	2.21	2.05
60	0.1	1.71	1.66	1.6	1.54	1.5	1.48	1.44	1.41	1.4	1.35	1.30
	0.05	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.56	1.53	1.47	1.40
	0.025	2.27	2.17	2.06	1.94	1.87	1.82	1.74	1.7	1.67	1.58	1.49
	0.01	2.63	2.5	2.35	2.2	2.1	2.03	1.94	1.88	1.84	1.73	1.62
	0.001	3.54	3.32	3.08	2.83	2.67	2.55	2.41	2.32	2.25	2.08	1.92
100	0.1	1.66	1.61	1.56	1.49	1.45	1.42	1.38	1.35	1.34	1.28	1.22
	0.05	1.93	1.85	1.77	1.68	1.62	1.57	1.52	1.48	1.45	1.38	1.30
	0.025	2.18	2.08	1.97	1.85	1.77	1.71	1.64	1.59	1.56	1.46	1.36
	0.01	2.5	2.37	2.22	2.07	1.97	1.89	1.8	1.74	1.69	1.57	1.45
	0.001	3.3	3.07	2.84	2.59	2.43	2.32	2.17	2.08	2.01	1.83	1.64
200	0.1	1.63	1.58	1.52	1.46	1.41	1.38	1.34	1.31	1.29	1.23	1.16
	0.05	1.88	1.8	1.72	1.62	1.56	1.52	1.46	1.41	1.39	1.3	1.21
	0.025	2.11	2.01	1.9	1.78	1.7	1.64	1.56	1.51	1.47	1.37	1.25
	0.01	2.41	2.27	2.13	1.97	1.87	1.79	1.69	1.63	1.58	1.45	1.30
	0.001	3.12	2.9	2.67	2.42	2.26	2.15	2	1.9	1.83	1.64	1.43
1000	0.1	1.61	1.55	1.49	1.43	1.38	1.35	1.3	1.27	1.25	1.18	1.08
	0.05	1.84	1.76	1.68	1.58	1.52	1.47	1.41	1.36	1.33	1.24	1.11
	0.025	2.06	1.96	1.85	1.72	1.64	1.58	1.5	1.45	1.41	1.29	1.13
	0.01	2.34	2.2	2.06	1.9	1.79	1.72	1.61	1.54	1.5	1.35	1.16
	0.001	2.99	2.77	2.54	2.3	2.14	2.02	1.87	1.77	1.69	1.49	1.22

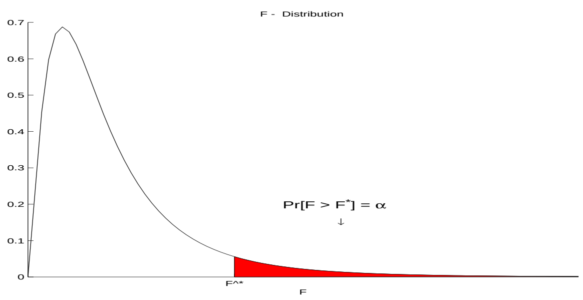
Table 21: Percentage Points of the Maximum F-Ratio

α=0.05

	r										
ν	2	3	4	5	6	7	8	9	10	11	12
2	39.0	87.5	142	202	266	333	403	475	550	626	704
3	15.4	27.8	39.2	50.7	62.0	72.9	83.5	93.9	104	114	124
4	9.60	15.5	20.6	25.2	29.5	33.6	37.5	41.1	44.6	48.0	51.4
5	7.15	10.8	13.7	16.3	18.7	20.8	22.9	24.7	26.5	28.2	29.9
6	5.82	8.38	10.4	12.1	13.7	15.0	16.3	17.5	18.6	19.7	20.7
7	4.99	6.94	8.44	9.70	10.8	11.8	12.7	13.5	14.3	15.1	15.8
8	4.43	6.00	7.18	8.12	9.03	9.78	10.5	11.1	11.7	12.2	12.7
9	4.03	5.34	6.31	7.11	7.80	8.41	8.95	9.45	9.91	10.3	10.7
10	3.72	4.85	5.67	6.34	6.92	7.42	7.87	8.28	8.66	9.01	9.34
12	3.28	4.16	4.79	5.30	5.72	6.09	6.42	6.72	7.00	7.25	7.48
15	2.86	3.54	4.01	4.37	4.68	4.95	5.19	5.40	5.59	5.77	5.93
20	2.46	2.95	3.29	3.54	3.76	3.94	4.10	4.24	4.37	4.49	4.59
30	2.07	2.40	2.61	2.78	2.91	3.02	3.12	3.21	3.29	3.36	3.39
60	1.67	1.85	1.96	2.04	2.11	2.17	2.22	2.26	2.30	2.33	2.36
∞	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.0	1.0

α=0.01

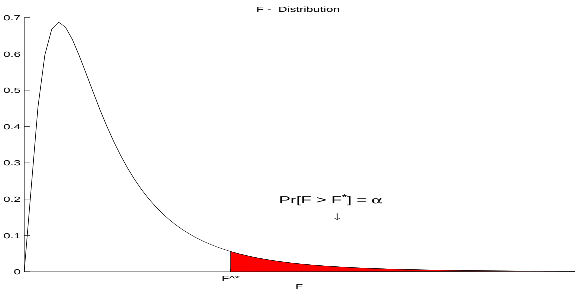
	r										
ν	2	3	4	5	6	7	8	9	10	11	12
2	199	448	729	1036	1362	1705	2063	2432	2813	3204	3605
3	47.5	85	120	151	184	21(6)	24(9)	28(1)	31(0)	33(7)	36(1)
4	23.2	37	49	59	69	79	89	97	106	113	120
5	14.9	22	28	33	38	42	46	50	54	57	60
6	11.1	15.5	19.1	22	25	27	30	32	34	36	37
7	8.89	12.1	14.5	16.5	18.4	20	22	23	24	26	27
8	7.50	9.9	11.7	13.2	14.5	15.8	16.9	17.9	18.9	19.8	21
9	6.54	8.5	9.9	11.1	12.1	13.1	13.9	14.7	15.3	16.0	16.6
10	5.85	7.4	8.6	9.6	10.4	11.1	11.8	12.4	12.9	13.4	13.9
12	4.91	6.1	6.9	7.6	8.2	8.7	9.1	9.5	9.9	10.2	10.6
15	4.07	4.9	5.5	6.0	6.4	6.7	7.1	7.3	7.5	7.8	8.0
20	3.32	3.8	4.3	4.6	4.9	5.1	5.3	5.5	5.6	5.8	5.9
30	2.63	3.0	3.3	3.4	3.6	3.7	3.8	3.9	4.0	4.1	4.2
60	1.96	2.2	2.3	2.4	2.4	2.5	2.5	2.6	2.6	2.7	2.7
∞	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0



Pr{**F** $\geq F^*$}
Table entry for F is the probability α lying above F^* (ie. tail probabilities)

Table 22: Critical values of the F - distribution - α = 0.05

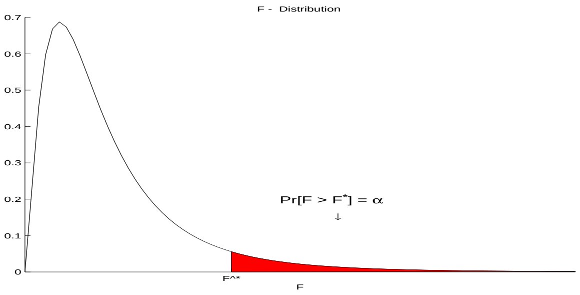
	ν_1																
ν_2	1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	248.01	250.10	251.14	252.20	253.25	254.25
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.57	2.53	2.49	2.45	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04	1.99	1.95	1.90	1.85
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.18	2.10	2.01	1.96	1.92	1.87	1.82
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	1.98	1.94	1.89	1.84	1.79
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.13	2.05	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.94	1.89	1.84	1.79	1.74
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09	2.01	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.84	1.79	1.74	1.68	1.63
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.74	1.69	1.64	1.58	1.51
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.69	1.63	1.58	1.51	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.75	1.66	1.55	1.50	1.43	1.35	1.26
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.01	1.94	1.88	1.84	1.67	1.58	1.46	1.40	1.32	1.23	1.00



Pr$\{$**F** $\geq F^*\}$
Table entry for F is the probability α lying above F^* (ie. tail probabilities)

Table 22: Critical values of the F - distribution (continued) - α = 0.01

	ν_1																
ν_2	1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.43	99.45	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87	26.69	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20	14.02	13.84	13.75	13.65	13.56	13.47
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.55	9.38	9.29	9.20	9.11	9.03
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.23	7.14	7.06	6.97	6.89
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.65	4.57	4.48	4.40	4.32
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	3.94	3.86	3.78	3.69	3.61
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.70	3.62	3.54	3.45	3.37
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.35	3.27	3.18	3.09	3.01
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.10	3.02	2.93	2.84	2.76
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31	3.16	3.00	2.92	2.83	2.75	2.66
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15	3.00	2.84	2.76	2.67	2.58	2.50
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.78	2.69	2.61	2.52	2.43
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.03	2.88	2.72	2.64	2.55	2.46	2.37
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98	2.83	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	2.93	2.78	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.89	2.74	2.58	2.49	2.40	2.31	2.22
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.85	2.70	2.54	2.45	2.36	2.27	2.18
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.20	2.11	2.02	1.92	1.81
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.42	2.27	2.10	2.01	1.91	1.80	1.69
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.03	1.94	1.84	1.73	1.61
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.19	2.03	1.86	1.76	1.66	1.53	1.39
1000	6.65	4.62	3.79	3.33	3.03	2.81	2.65	2.52	2.42	2.33	2.05	1.89	1.71	1.60	1.48	1.34	1.00



Pr$\{\mathbf{F} \geq F^*\}$
Table entry for F is the probability α lying above F^* (ie. tail probabilities)

Table 22: Critical values of the F - distribution(continued) - α = 0.001

	ν_1																
ν_2	1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
2	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40	999.43	999.45	999.47	999.47	999.48	999.49	999.50
3	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86	129.25	127.37	126.42	125.45	124.96	124.47	123.97	123.50
4	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05	46.76	46.10	45.43	45.09	44.75	44.40	44.07
5	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92	25.91	25.39	24.87	24.60	24.33	24.06	23.80
6	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41	17.56	17.12	16.67	16.44	16.21	15.98	15.76
7	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08	13.32	12.93	12.53	12.33	12.12	11.91	11.71
8	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54	10.84	10.48	10.11	9.92	9.73	9.53	9.35
9	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	9.24	8.90	8.55	8.37	8.19	8.00	7.82
10	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75	8.13	7.80	7.47	7.30	7.12	6.94	6.77
11	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92	7.32	7.01	6.68	6.52	6.35	6.18	6.01
12	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	6.71	6.40	6.09	5.93	5.76	5.59	5.43
13	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.23	5.93	5.63	5.47	5.30	5.14	4.98
14	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	5.85	5.56	5.25	5.10	4.94	4.77	4.61
15	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.54	5.25	4.95	4.80	4.64	4.47	4.32
16	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81	5.27	4.99	4.70	4.54	4.39	4.23	4.07
17	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58	5.05	4.78	4.48	4.33	4.18	4.02	3.86
18	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39	4.87	4.59	4.30	4.15	4.00	3.84	3.68
19	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22	4.70	4.43	4.14	3.99	3.84	3.68	3.52
20	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.56	4.29	4.00	3.86	3.70	3.54	3.39
21	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95	4.44	4.17	3.88	3.74	3.58	3.42	3.27
22	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.33	4.06	3.78	3.63	3.48	3.32	3.16
23	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73	4.23	3.96	3.68	3.53	3.38	3.22	3.07
24	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64	4.14	3.87	3.59	3.45	3.29	3.14	2.98
25	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56	4.06	3.79	3.52	3.37	3.22	3.06	2.90
30	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	3.75	3.49	3.22	3.07	2.92	2.76	2.60
40	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.40	3.14	2.87	2.73	2.57	2.41	2.24
50	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67	3.20	2.95	2.68	2.53	2.38	2.21	2.04
60	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54	3.08	2.83	2.55	2.41	2.25	2.08	1.90
120	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24	2.78	2.53	2.26	2.11	1.95	1.77	1.56
1000	10.86	6.93	5.44	4.64	4.12	3.76	3.49	3.28	3.11	2.97	2.53	2.28	2.01	1.85	1.68	1.47	1.00

22 Table credits

1. Table 21.4 Reprinted from: Handbook of Tables for Probability and Statistics, Second Edition. Edited by William H. Beyer, © The Chemical Rubber Co., 1968. Used by permission of CRC Press Inc., Boca Raton, FL.