TMATH 390
Probability and Statics for Engineers
NOTES

Ben Foster*
benf94@uw.edu

March 31, 2015

Abstract

This document contains the notes from the course TMATH 390 and does not necessarily contain all the information provided by the instructor.

*Institute of Technolgy, University of Washington Tacoma

Contents

1 Data and Distributions

The discipline of statistics teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation.

1.1 Populations, Samples, and Processes

An investigation will typically focus on a well-defined collection of objects constituting a **population** of interest. In one study, the population might consist of all gelatin capsules of a particular type produced during a specified period. When desired information is available for all objects in the population, we have what is called a **census**. A subset of the population – a **sample**– is usually selected in some prescribed manner to get feedback on the population.

1.2 Visual Displays for Univariate Data

Stem-and-Leaf Displays These can be an effective way to organize numerical data without expending much effort. It is based on separating each observation into two parts: a **stem**, consisting of one or more leading digits, and a **leaf**, consisting of the remaining or trailing digit(s). Example:

0	4
1	1345678889
2	1223456666777889999
3	01122333445556666777777888899999
4	11122222334444556666667778888899 9
5	00111222233455666667777888899
6	01111244455666778

Table 1: Percentage binge drinkers at each of 140 colleges

1.2.1 Dotplots

A dotplot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values. Each observation is presented by a dot above the corresponding location on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically. Example: Here is data on state-by-state appropriations for higher education as a percentage of state and local tax revenue for fiscal year 2009-2010:

14.0	3.1	8.6	9.6	7.4	4.0	4.5	6.5	6.1	8.8
8.2	8.6	6.4	6.7	8.0	8.5	9.4	9.5	4.6	6.8
3.9	6.9	6.3	11.9	5.8	5.8	9.9	5.9	2.7	4.2
14.9	4.0	12.1	8.0	5.2	9.2	6.8	4.3	3.9	9.6
8.0	8.6	8.6	8.7	3.1	5.8	6.2	8.7	6.8	8.9

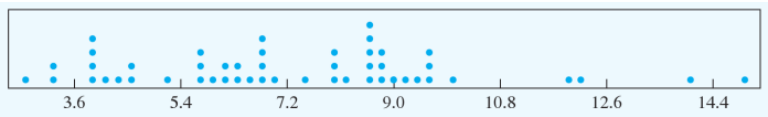


Figure 1: example caption

1.2.2 Histograms

Some numerical data is obtained by counting to determine the value of a variable, whereas other data is obtained by taking measurements. The prescription for drawing a histogram is different for these two cases.

A variable is **Discrete** if its set of possible values either is finite or else can be listed in an infinite sequence. A variable is **continuous** if its possible values consist of an entire interval on the number line.
Consider data consisting of observations on a discrete variable x. The **Frequency** of any particular x value is the number of times that value occurs in the data set. The **Relative Frequency** of a value is the fraction or proportion of time the value occurs.

$$\text{relative frequency of a value} = \frac{\text{number of times the value occurs}}{\text{number of observations in the data set}}$$

Here is an example of a histogram:

Board Size	Frequency	Relative Frequency	Board Size	Frequency	Relative Frequency
4	3	0.0147	19	0	0.0000
5	12	0.0588	20	0	0.0000
6	13	0.0637	21	1	0.0049
7	25	0.1225	22	0	0.0000
8	24	0.1176	23	0	0.0000
9	42	0.2059	24	1	0.0049
10	23	0.1127	25	0	0.0000
11	19	0.0931	26	0	0.0000
12	16	0.0784	27	0	0.0000
13	11	0.0539	28	0	0.0000
14	5	0.0245	29	0	0.0000
15	4	0.0196	30	0	0.0000
16	1	0.0049	31	0	0.0000
17	3	0.0147	32	1	0.0049
18	0	0.0000		204	0.9997

Table 2: Table of board members for hospitals

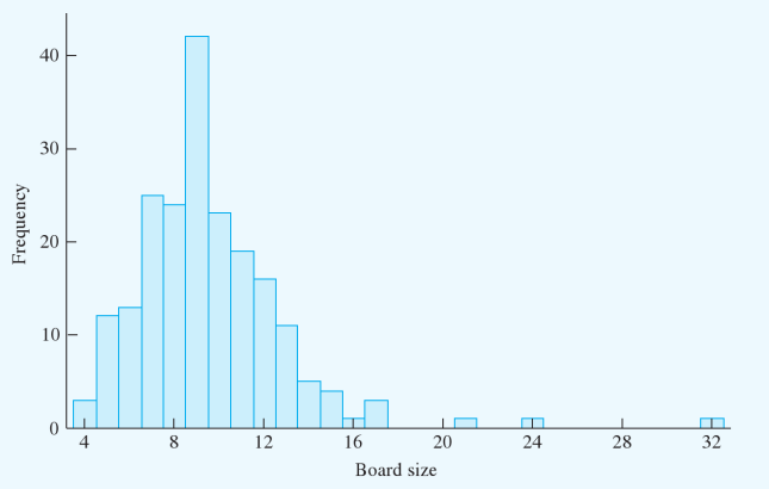


Figure 2: Example of a Histogram

Constructing a Histogram for Continuous Data: Equal Class Widths Determine the frequency and relative frequency for each class. Mark the class boundaries on a horizontal measurement axis. Above each class interval, draw a rectangle whose height is the corresponding relative frequency (or frequency).

Constructing a Histogram for Continuous Data: Unequal Class Widths After determining the frequencies and relative frequencies, calculate the height of each rectangle using the formula:

$$\text{rectangle height} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

The resulting rectangle heights are usually called *densities*, and the vertical scale is the **density scale**. This prescription will also work when the class widths are equal.

1.2.3 Histogram Shapes

Histograms come in a variety of shapes. A **unimodal** histogram is one that rises to a single peak and then declines. A **bimodal** histogram has two different peaks. Bimodality occurs when the data set consists of observations on two quite different kinds of individuals or objects. A histogram with more than two peaks is said to be **multimodal**. Of course, the number of peaks may well depend on the choice of class intervals, particularly with a small number of observations.

A histogram is **symmetric** if the left half is a mirror image of the right half. A unimodal histogram is **positively skewed** if the right or upper tail is stretched out compared with the left or lower tail, and **negatively skewed** if the longer tail extends to the left.

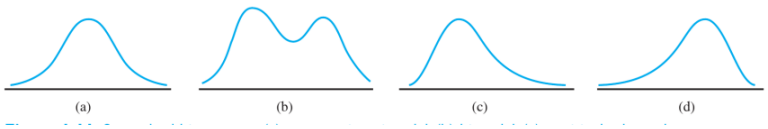


Figure 3: Smoothed Histograms: (a) symmetric unimodal; (b) bimodal; (c) positively skewed; (d) negatively skewed

1.2.4 Categorical Data

A histogram for categorical data is often called a **bar chart**. In come cases, there will be a natural ordering of classes (for example, freshman, sophomore, junior, senior), whereas in other cases, the order will be arbitrary (Honda, Yamaha, Ford, etc.). A **Pareto diagram** is a bar chart resulting from a quality control study in which category data represents a different type of product nonconformity or production problem.

1.3 Describing Distributions

In Section 1.2, we saw that a histogram could be used to describe how values of a variable x are distributed in a data set. In practice, a histogram is virtually always constructed from sample data.

1.3.1 Continuous Distributions

1.4 The Normal Distribution

1.5 Other Continuous Distributions

1.6 Several Useful Discrete Distributions

2 Numerical Summary Measures

2.1 Measures of Center

2.2 Measures of Variability

2.3 More Detailed Summary Quantities

2.4 Quantile Plots

Constructing a Quantile plot can take a little more work than constructing a regular distribution. For example, when making a Normal Quantile Plot, you would use the following definition for a sample quantile: Let $x_{(1)}$ denote the smallest sample observation, $x_{(2)}$ the second smallest observation,..., and $x_{(n)}$ the largest sample observation. For $i = 1, ..., n, x_{(i)}$ is the $[(i - 0.5/n]$th sample quantile.
Therefore, to make the Normal Quantile Plot, you would use the coordinates:

$$\left(\left(\frac{0.5}{n}\right) \text{th quantile}, x_{(1)}\right), ..., \left(\left(\frac{i-0.5}{n}\right) \text{th quantile}, x_{(n)}\right)$$

The plot of this, if a true normal distribution, should fall close to a 45° angle or a line with a slope of 1 passing through the point (0,0).

If you get a normal distribution that isn't standard, then you would use this:

$$\text{quantile for normal } (\mu, \sigma)\text{distribution} = \mu + (\text{corresponding z quantile})\sigma$$

3 Bivariate and Multivariate Data and Distributions

Bivariate and Multivariate Data and Distributions:

A multivariate data set consists of observations made simultaneously on two or more variables. One important special case is that of bivariate data, in which observations on only two variables, x and y are available. We will also discuss the correlation coefficient which is a measure of how strongly two variables are related.

3.1 Scatterplots

Scatterplots are the best way to graphically describe Bivariate data sets. In R, typing in the command `plot(A,B)` will print out the scatter plot with A being x and B being Y where as the command `plot(A ~ B)` treats A as y and B as x. Here is an example of a scatterplot with histograms related to both variables attached to it:

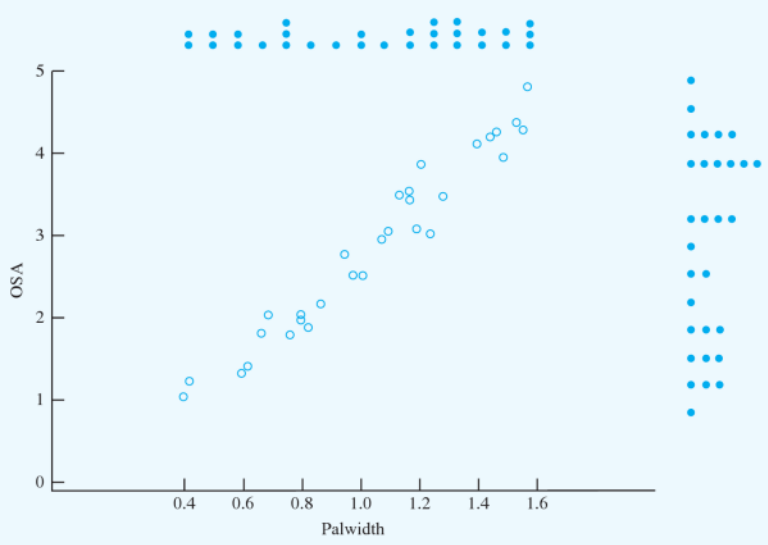


Figure 4: Example of a scatterplot

3.2 Correlation

In order to make precise statements and draw reliable conclusions from the data, we must go beyond pictures and find **Correlation Coefficient** which is a quantitative assessment of the strength of relationship between x and y values in a set of pairs.

A *positive* relationship is one where both x and y tend to increase together. A *negative* relationship is one where y tends to decrease as x increases. A strong positive or negative relationship can be linear or curved in appearance so long as x and y tend to be clustered together in particular pattern.

Pearson's sample correlation This sample correlation is r given by the following equation:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$
$$= \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Computing formulas for the three summation quantities are:

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$
$$S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

Properties of r

- The value of r does not depend on the unit of measurement for either variable, meaning the correlation coefficient measures the inherent strength of relationship between two numerical variables.
- The value of r does not depend on which of the two variables is labeled x.
- The value of r is between -1 and +1. A value near the upper limit is indicative of a substantial positive relationship, whereas an r close to the lower limit suggests a prominent negative relationship.
- r = 1 only when all the points in a scatterplot of the data lie exactly on a straight line that slopes upward. Similarly, r = -1 only when all the points lie exactly on a downward sloping line.
- The value of r is a measure of the extent to which x and y are **linearly** related. The extent to which the points in the scatterplot fall close to a straight line.

The Population Correlation Coefficient Pearson's r measures how strongly the x and y values in a *sample* of pairs are related to one another. There is an analogous measure of how strongly x and y are related in the entire population of pairs from which the sample was obtained. This is called the **population correlation coefficient** and is denoted by ρ. ρ satisfies properties paralleling those of r:

- ρ is a number between -1 and +1 that does not depend on the unit of measurement for either x or y, or on which variable is labeled x and which is labeled y.
- $\rho = +1$ or -1 if and only if all pairs in the population lie exactly on a straight line.

Correlation and Causation Be sure to remember that just because a value of r is close to 1 does not mean that relatively large values of one variable *cause* relatively large values of the other variable.

3.3 Fitting a Line to Bivariate Data

Given two numerical values of x and y, the general objective of *regression analysis* i s to use the information about x to draw some type of conclusion concerning y. The different roles played by the two variables are reflected in standard terminology: y is called the **dependent** or **response variable**, and x is referred to as the **independent, predictor, or explanatory variable**.
A scatterplot of y vs x frequently exhibits a linear pattern. In such cases, it is natural to summarize the relationship between the variables by finding a line that is as close as possible to the points in the plot.

Fitting a Straight Line Often when making a scatterplot, you can place a line that generally summarizes the scatterplot and then each point will have a deviation from that line. In order to find the best fitting line, we need to find the **Least Squares Line**. To find that we have to minimize the following sum:

$$\sum [y_i - (a + bx_i)]^2 = [y_1 - (a + bx_1)]^2 + ... + [y_n - (a + bx_n)]^2$$

To find this equation, let $g(\tilde{a}, \tilde{b}) = \sum [y_i - (\tilde{a} + \tilde{b}x_i)]^2$. Then the intercept a and the slope b of the least squares line are the values of $\tilde{a}$ and $\tilde{b}$ that minimize $g(\tilde{a}, \tilde{b})$. These minimizing values are obtained by taking the partial derivative of the g function first with respect to $\tilde{a}$ and then with respect to $\tilde{b}$, and equating these two partial derivatives to zero.

The slope b of the least squares line is given by

$$b = \frac{\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n} = \frac{S_{xy}}{S_{xx}}$$

The vertical intercept a of the least squares line is given by

$$a = \bar{y} - b\bar{x}$$

The equation of the least squares line is often written as $\hat{y} = a + bx$, where the hat above the y emphasizes that $\hat{y}$ is a prediction of y that results from the substitution of any particular x value into the equation.
The least squares line should not be used to make a prediction for an x value much beyond the range of the data. The danger of extrapolation is that the fitted relationship may not be valid for such x values.

Regression The term comes from the relationship between the least squares line and the sample correlation coefficient. A little Algebraic manipulation yields:

$$b = r\left(\frac{s_y}{s_x}\right)\hat{y} = \bar{y} + r\left(\frac{s_y}{s_x}\right)(x - \bar{x}$$

When $-1 < r < 1$, for *any* x value, the corresponding predicted value $\hat{y}$ will be closer in terms of standard deviations to $\bar{y}$ than is x to $\bar{x}$; that is, $\hat{y}$ is pulled toward (regressed toward) the mean y value. This **regression effect** was first noticed by Sir Francis Galton in the late 1800s when he studied the relation between father's height and son's height; The predicted height of a son was always closer to the mean height than was his father's height.

Assessing the Fit of the Least Squares Line How much of the observed variation in y can be attributed to the approximate linear relationship and the fact that x is varying? A quantitative assessment is based on the vertical deviations from the least squares line.
Variation in y can be effectively be explained by an approximate straight-line relationship when the points in the scatterplot fall close to the least squares line – that is, when the residuals are small in magnitude. A natural measure of variation about the least squared line is the sum of the squared residuals.
Residual sum of squares, denoted by **SSResid**, is given by:

$$\text{SSResid} = \sum(y_i - \hat{y})^2 = (y_1 - \hat{y}_1)^2 + ... + (y_n - \hat{y}_n)^2$$

(Alternatively called *error sum of squares* and denoted by SSE).
Total sum of squares, denoted by **SSTo**, is defined as

$$\text{SSTo} = \sum(y_i - \bar{y})^2$$

Alternatie notation for SSTo is S_{yy}, and a computing formula is

$$\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

A computing formula for residual sum of squares makes it unnecessary to calculate the residuals:

$$\text{SSResid} = \text{SSTo} - bS_{xy}$$

because b and S_{xy} have the same sign, bS_{xy} is a positive quantity unless $b = 0$, so the computing formula shows that SSResid = SSTo if $b = 0$ and SSResid ¡ SSTo otherwise.
SSResid is the amount of variation in y that cannot be attributed to the linear relationship between x and y.

The **coefficient of determination**, denoted by r^2, is given by

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

It is the proportion of variation in the observed y values that can be explained by a linear relationship between x and y in the sample.

Standard Deviation About the Least Squares Line The **standard deviation about the least squares line** is given by

$$s_e = \sqrt{\frac{\text{SSResid}}{n-2}}$$

Roughly speaking, s_e is the typical amount by which an observation deviates from the least squares line.

Plotting the Residuals (Optional A desirable plot exhibits no particular pattern, such as curvature or much greater spread in one part of the plot than in another part. Looking at a residual plot after fitting a line amounts to examining y after removing any linear dependence on x. This can sometimes more clearly show the existence of a nonlinear relationship.

A point that is far off typically means that there is some unusual behavior such a recording error, non-standard experimental condition or atypical experimental subject.

3.4 Nonlinear Relationships

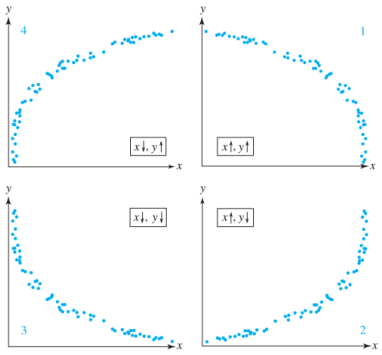
A scatterplot of bivariate data frequently shows curvature rather than a linear pattern. In this section, we discuss several different ways to fit a curve to such data.

Power Transformations Suppose that the general pattern is curved and monotonic. In this case, it is often possible to find a **power transformation** for x and y so that there is a linear pattern in a scatterplot of the transformed data. By a power transformation, we mean the use of exponents p and q such that the transformed values are $x\prime = x^p$ and/ or $y\prime = y^q$; the relevant scatterplot is of the $(x\prime, y\prime)$ pairs.

Transformed value $= (originalvalue)^{POWER}$

Power Transformation ladder:

Power	Transformed value	Name
3	$(Originalvalue)^3$	Cube
2	$(Originalvalue)^2$	Square
1	Original value	No transformation
$\frac{1}{2}$	$\sqrt{Originalvalue}$	Square root
$\frac{1}{3}$	$\sqrt[3]{Originalvalue}$	Cube root
0	Log(original value)	Logarithm
-1	$\frac{1}{originalvalue}$	Reciprocal



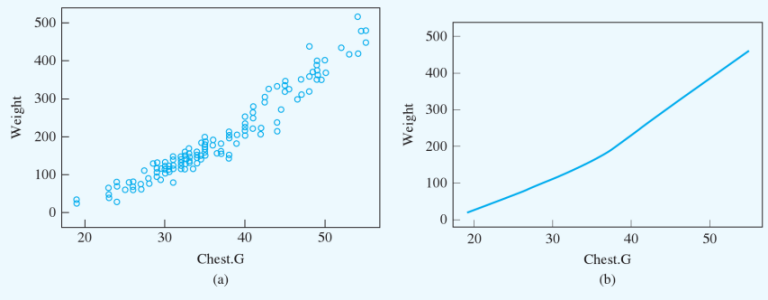
Fitting a Polynomial Function Sometimes the general patter of curvature in a scatterplot is not monotonic. In such instances, it is reasonable to fit a quadratic function $a + b_1x + b_2x^2$, whose graph is a parabola, to the data. If the quadratic coefficient b_2 is positive, the parabola turns upward, whereas it turns downward if b_2 is negative. Just as in fitting a straight line, the principle of least squares can be employed to find the best-fit quadratic. The **least squares coefficients** $a, b_1,$ **and** b_2 are the values of $\tilde{a}, \tilde{b_1},$ and$\tilde{b_2}$ that minimize

$$g(\tilde{a}, \tilde{b_1}, \tilde{b_2}) = \sum_i [y_i - (\tilde{a} + \tilde{b_1}x_i + \tilde{b_2}x_i^2)]^2$$

Which is the sum of squared vertical deviations from the points in the scatterplot to the parabola determined by the quadratic with coefficients $\tilde{a}, \tilde{b_1}, \tilde{b_2}$. Taking the partial derivative of the g function first with respect to $\tilde{a}$, then with respect to $\tilde{b_1}$, and finally with respect to $\tilde{b_2}$, and equating these three expressions to zero gives three equations in three unknowns. These *normal equations* are again linear in the unknowns, but because there are three rather than just two, there is no explicit elementary expression for their solution. Instead, matrix algebra must be used to solve the system numerically for each different data set.

The methodology employed to fit a quadratic is easily extended to fit a higher-order polynomial. For example, using the principle of least squares to fit a cubic equation gives a system of normal equations consisting of four equations in four unknowns. The arithmetic is best left to a statistical computer package. In practice, a cubic equation is rarely fit to data, and iti s virtually never appropriate to fit anything of higher order than this.

Smoothing a Scatterplot Sometimes the pattern in a scatterplot is too complex for a line or curve of a particular type (e.g., exponential or parabolic) to give a good fit. Statisticians have recently developed some more flexible methods that permit a wide variety of patterns to be modeled using the same fitting procedure. One such method is LOWESS, short for *locally weighted scatterplot smoother*. Let $(x*, y*)$ denote a particular one of the n (x, y) pairs in the sample. The $\hat{y}$ value corresponding to $(x*, y*)$ is obtained by fitting a straight line using only a specified percentage of the data (e.g, 25%) whose x values are closest to $x*$. Furthermore, rather than use "ordinary" least squares, which gives equal weight to all points, those with x values closer to $x*$ are more heavily weighted than those whose x values are farther away. The height of the resulting line above $x*$ is the fitted value $\hat{y}*$. This process is repeated for each of the n points, so n different lines are fit (you surely wouldn't want to do all this by hand). Finally, the fitted points are connected to produce a LOWESS curve.



3.5 using More Than One Predictor

In many situations, predictions of y values can be improved and more observed y variation can be explained by utilizing information in two or more explanatory variables. Notation is a bit more complex than in the case of a single predictor. Let

$$k = \text{number of explanatory variables or predictors} \quad (1)$$

$$n = \text{sample size} \quad (2)$$

and $x_1, x_2, ..., x_k$ denote the k predictors, so that each observation will consist of $k + 1$ numbers: the value of x_1, the value of $x_2, ...$, the value of x_k, and the value of y. Also let

$$x_{ij} = \text{value of the predictor } x_i \text{ in the } i\text{th observation}$$

so

$$firstobservation = (x_{11}, x_{21}, ..., x_{k1}, y_1)$$

$$...$$

$$n\text{th observation} = (x_{1n}, x_{2n}, ..., x_{kn}, y_n)$$

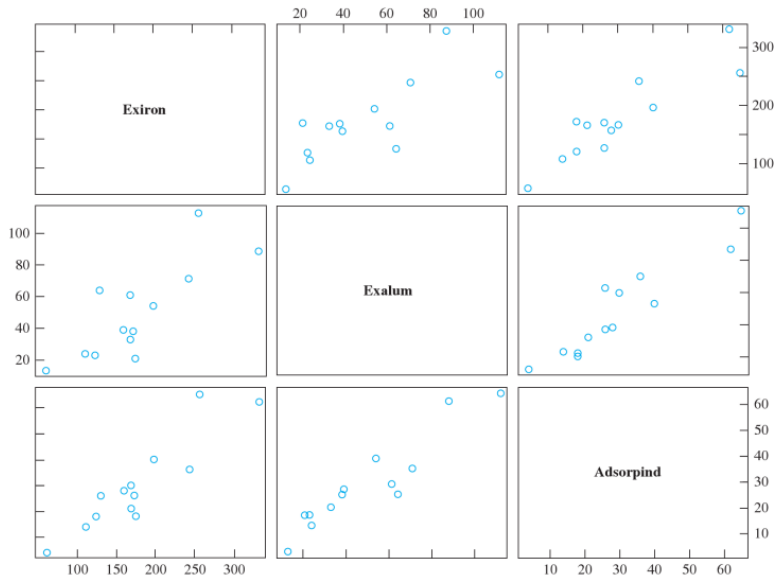
Example

Soil and sediment adsorption, the extent to which chemicals collect in a condensed form on the surface, is an important characteristic because it influences the effectiveness of pesticides and various agricultural chemicals. The article "Adsorption of Phosphate, Arsenate, Methanearsonate, and Cacodylate by Lake and Stream Sediments: Comparison with Soils" (*J. of Environ. Qual.*, 1984: 499–504) gave the following data on y = phosphate adsorption index, x_1 = amount of extractable iron, and x_2 = amount of extractable aluminum

Observation	x_1	x_2	y
1	61	13	4
2	175	21	18
3	111	24	14
4	124	23	18
5	130	64	26
6	173	38	26
7	169	33	21
8	169	61	30
9	160	39	28
10	244	71	36
11	257	112	65
12	333	88	62
13	199	54	40

Thus the first observation is the triple $(x_{11}, x_{21}, y_1) = (61, 13, 4), \ldots$, and the last observation is $(x_{1,13}, x_{2,13}, y_{13}) = (199, 54, 40)$.

A scatterplot of this data would represent each observation as a point in a three-dimensional coordinate system, which is obviously difficult to construct or visualize. Partial information about the relationship between the variables can be obtained by forming a **scatterplot matrix**. This is just a collection of two-dimensional scatterplots, arranged in a square array, in which each variable is plotted against every other variable. Here is a scatterplot matrix of the data in the previous example:



Fitting a Linear Function We now consider fitting a relation of the form:

$$y \approx a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

The reasonableness of this approximation depends on patterns in the scatterplot matrix and other characteristics of the data to be considered shortly. As with bivariate data, the values of $a, b_1, \ldots, b_k$ should be selected to give

the best fit.
The **least squares coefficients** $a, b_1, b_2, ..., b_k$ are the values of $\tilde{a}, \tilde{b_1}, ..., \tilde{b_k}$ that minimize

$$g(\tilde{a}, \tilde{b_1}, ..., \tilde{b_k}) = \sum_{j-1}^{n} [y_j - (\tilde{a} + \tilde{b_1} x_{1j} + ... + \tilde{b_k} x_{kj})]^2$$

The g() function is the sum of squared deviations between observed y values and what would be predicted by $\tilde{a} + \tilde{b_1} x_1 + ... + \tilde{b_k} x_k$.

3.6 Joint Distributions

4 Obtaining Data

4.1 Operational Definitions

4.2 Data from Sampling

4.3 Data from Experiments

4.4 Measurement Systems

5 Probability and Sampling Distributions

5.1 Chance Experiments

A **chance experiment**, also called a **random experiment**, is simply an activity or situation whose outcomes, to some degree, depend on chance. To decide whether a given activity qualifies as a chance experiment, ask yourself the question, will I get exactly the same result if I repeat the experiment more than once? If no, then the experiment is a chance experiment.

Events
Underlying the computations of probability is an organized system for describing and working with the outcomes of chance experiments. These outcomes can be divided into two types: **simple events**, which are the individual outcomes of an experiment and, more generally, **events**, which consist of collections of simple events. For instance, the chance experiment of conducting a series of stress tests on three metal parts has eight possible outcomes. Each of these eight outcomes is a simple event, which, taken together, form the **sample space** of the experiment.

Depicting Events
aasd ASDASDFADSF

5.2 Probability Concepts

5.3 Conditional Probability and Independence

5.4 Random Variables

5.5 Sampling Distributions

5.6 Describing Sampling Distributions

6 Quality and Reliability

6.1 Terminology

6.2 How Control Charts Work

6.3 Control Charts for Mean and Variation

6.4 Process Capability Analysis

6.5 Control Charts for Attributes Data

6.6 Reliability

7 Estimation and Statistical Intervals

The general objective of a statistical inference is to use sample information as a basis for drawing various types of conclusions. In an estimation problem, we want to make an educated guess about the value of some population characteristic or parameter, such as the population mean battery lifetime μ, the proportion π of all components of a certain type that need service while under warranty, or the difference $\mu_1 - \mu_2$ between the population mean lifetimes for two different types of batteries. The simplest type of estimate is a *point estimate*, a single number that represents our best guess for the value of the parameter.

7.1 Point Estimation

A **point estimate** of some parameter θ is a single number, calculated from sample data, that can be regarded as an educated guess for the value of θ.
A point estimate is usually obtained by selecting a suitable statistic and calculating its value for the given sample data. For example, a natural statistic to use for estimating a population mean μ is the sample mean $\bar{x}$, and a sensible way to estimate a population variance σ^2 is to compute the value of the sample variance s^2. The statistic used to calculate an estimate is sometimes called an **estimator**, and the symbol $\hat{\theta}$ is frequently used to denote either the estimator or the resulting estimate. Thus the statement

$$\hat{\mu} = \bar{x} = 32.5$$

says that the point estimate of the population mean μ is 32.5 and that this estimate was calculated using the sample mean $\bar{x}$ as the estimator.

A commonly used method of estimating the size of a wildlife population is to perform a capture/recapture experiment. Suppose a biologist wishes to estimate the number of fish in a certain lake; that is, the parameter to be estimated is the population size N. An initial sample of 100 fish is selected, each one is tagged, and the tagged fish are returned to the lake. After a time period sufficient to allow the tagged fish to mix with the other fish in the lake, a second sample of 250 fish is selected. If 25 of the fish in the recapture sample are tagged, what is a sensible estimate for N? Because 10% of the fish in the recapture sample are tagged, it is reasonable to estimate that 10% of all fish in the lake are tagged. Since we know that a total of 100 fish were initially tagged, this suggests that we use 1000 as a point estimate of N.

More generally, if M denotes the number of fish initially tagged, n the size of the recapture sample, and x the number of tagged fish in the recapture sample (so x is a random variable), the proposed estimator of N is $\hat{N} = [Mn/x]$. (The square bracket notation $[c]$ denotes the largest whole number that is at most c; this takes care of cases where Mn/x is not a whole number.)

Figure 5: Example 7.1

Frequently, there is more than one estimator that can sensibly be used to calculate an estimate, as the following example shows.

Properties of Estimators One desirable property that a good estimator should possess is that it be **unbiased**. An estimator is unbiased if, in repeated random samples, the numerical values of the estimator stack up around the population parameter that we are trying to estimate. An often used analogy is to think of each value of an estimator as a shot fired at a target, the target being the population parameter of interest. As long as all the shots fall in a pattern with the target value in the *middle*, we say that the shots are unbiased. Notice that we do not require that any of the individual shots actually hit the target; we require only that they be centered around the target value. If the majority of the shots are centered somewhere else, then we say that they exhibit a certain amount of **bias**.

Definition
Denote a population parameter generically by the letter θ and denote any estimator of this parameter by $\hat{\theta}$. Then $\hat{\theta}$

Consider a population of N = 5000 invoices. Associated with each invoice is its "book value," the recorded amount of that invoice. Let T = \$1,761,300 denote the known total book value. Unfortunately, some of the book values are erroneous. An audit will be carried out by randomly selecting n invoices and determining the audited (i.e., correct) value for each one. Suppose the sample gives the following results:

Invoice:	1	2	3	4	5
Book value:	300	720	526	200	127
Audited value:	300	520	526	200	157
Error:	0	200	0	0	−30

Let $\bar{y}$ = sample mean book value = \$374.60, $\bar{x}$ = sample mean audited value = \$340.60, and $\bar{e}$ = sample mean error = \$34.00. Each of the following estimators for the total audited (i.e., correct) value and resulting estimates is sensible:

mean per unit statistic = $N\bar{x}$; estimate = 5000(340.60) = \$1,703,000

difference statistic = $T - N\bar{e}$; estimate = 1,761,300 − (5000)(34)
= \$1,591,300

ratio statistic = $T(\bar{x}/\bar{y})$; estimate = (1,761,300)(340.6/374.6) = \$1,601,438

The choice among these estimates is not clear-cut. In fact, all three of the estimators have been advocated by those employing statistical methodology in auditing.

Figure 6: Example 7.2

is an **unbiased** estimator if $\mu_{\hat{\theta}} = \theta$. Otherwise, $\hat{\theta}$ is said to be biased, and the quantity $\mu_{\hat{\theta}} - \theta$ is called the **bias** of $\hat{\theta}$.

Unbiasedness does not imply that the estimate computed from any particular sample will coincide with the value of the parameter being estimated. Consider, for example, using the sample proportion p to estimate the population proportion π based on a sample of size $n = 25$, and suppose that $\pi = 0.7$. Then $\mu_p = 0.7$, so the sampling distribution of p is centered at 0.7. However, with x denoting the number of "successes" in the sample, $p = \frac{x}{25} \neq 0.77$ for any possible value of x. That is, even though p is unbiased for estimating π, the value of the estimate calculated from any particular sample will inevitably differ from π.

A second desirable property that estimators often possess is consistency. If $\hat{\theta}$ denotes an estimator of some population parameter θ, then $\hat{\theta}$ is said to be **consistent** if the probability that it lies close to θ increases to 1 as the sample size increases. That is, as you increase n, it becomes more and more likely that such estimators will be very close to the parameter they are intended to estimate. The most common method for showing that an estimator is consistent is to show that its standard error decreases as the sample size increases.

Definition
If the probability that an estimator $\hat{\theta}$ falls close to a population parameter θ can be made as near to 1 as desired by increasing the sample size n, then $\hat{\theta}$ is said to be a **consistent** estimator of θ.

7.2 Large-Sample Confidence Intervals for a Population Mean

7.3 More Large-Sample Confidence Intervals

7.4 Small-Sample Intervals Based on a Norm. Pop. Distr.

7.5 Intervals for $\mu_1 - \mu_2$ based on Norm. Pop. Distr.

7.6 Other Topics in Estimation (Optional)

8 Testing Statistical Hypotheses

8.1 Hypotheses and Test Procedures

8.2 Tests Concerning Hypotheses About Means

8.3 Tests Concerning Hypotheses About a Categorical Population

8.4 Testing the Form of a Distribution

8.5 Further Aspects of Hypothesis Testing

9 The Analysis of Variance

9.1 Terminology and Concepts

9.2 Single-Factor ANOVA

9.3 Interpreting ANOVA Results

9.4 Randomized Block Experiments

10 Experimental Design

10.1 Terminology and Concepts

10.2 Two-Factor Designs

10.3 Multifactor Designs

10.4 2^k Designs

10.5 Fractional Factorial Designs

11 Inferential Methods in Regression and Correlation

11.1 Regression Models Involving a Single Independent Variable

11.2 Inferences About the Slope Coefficient

11.3 Inferences Based on the Estimated Regression Line

11.4 Multiple Regression Models

11.5 Inferences in Multiple Regression

11.6 Further Aspects of Regression Analysis

12 Appendix Tables