

Stat 3355

Statistical Methods for Statisticians and Actuaries

The notes and scripts included here are copyrighted by their author, Larry P. Ammann, and are intended for the use of students currently registered for Stat 3355. They may not be copied or used for any other purpose without permission of the author.

Syllabus

Stat 3355 Course Information

Instructor: Dr. Larry P. Ammann
Office hours: Thurs. 3-4pm, others by appt.
Email: ammann@utdallas.edu
Office: FO 2.604D
Phone: (972) 883-2164
Text: Introduction to Statistics & Data Analysis, 3rd Ed.
Authors: R. Peck, C. Olsen, and J. Devore

Topics	Chapters
Graphical summaries	3
Numerical summaries	4
Bivariate summaries	5
Probability, random variables, and simulation	6,7
Sampling distributions	8
One sample estimation and hypothesis tests	9,10
Two sample estimation and hypothesis tests	11
Regression and ANOVA	13, 15

Exam Schedule

To be determined

Grading Policy

Course grade will be based on exams and homework projects.

Exam 1: 20%

Exam 2: 20%

Exam 3: 30%

Homework: 30%

Note: the complete syllabus is available here:

http://www.utdallas.edu/~ammann/stat3355_syllabus.pdf

Class Notes

Graphical tools

A picture is worth a thousand words...

A thousand words spoken with no one listening are worth exactly nothing.

The computer tools that we have available today give us access to a wide array of graphical techniques and tools that can be used for effective presentation of complex data. However, we must first understand what type of data we wish to present, since the presentation tool that should be used for a set of data depends on the questions we wish to answer and the type of data we are using to answer those questions.

Note: graphics for this section are generated by the script file <http://www.UTDallas.edu/~ammann/stat3355scripts/graphex1.r>

Categorical (qualitative) data

Categorical data is derived from populations that consist of some number of subpopulations and we record only the subpopulation membership of selected individuals. In such cases the basic data summary is a frequency table that counts the number of individuals within each category. If there is more than one set of categories, then we can summarize the data using a multi-dimensional frequency table. For example, here is part of a dataset that records the hair color, eye color, and sex of a group of 592 students.

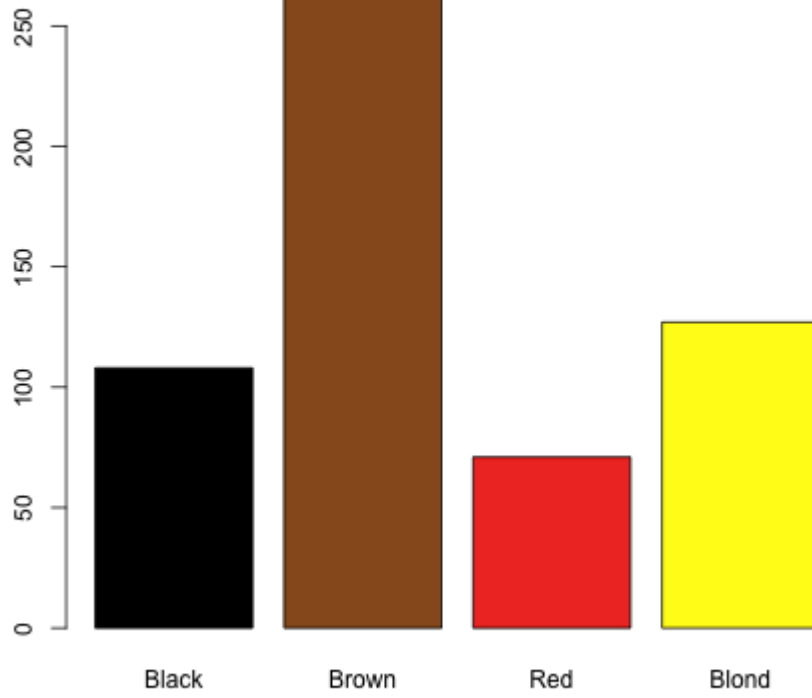
Hair	Eye	Sex
Black	Brown	Female
Red	Green	Male
Blond	Blue	Male
Brown	Hazel	Female
...		

Sometimes numerical codes are used in place of names, but it is important to remember that these codes are not quantitative values, just labels. The frequency table for hair color in this dataset is:

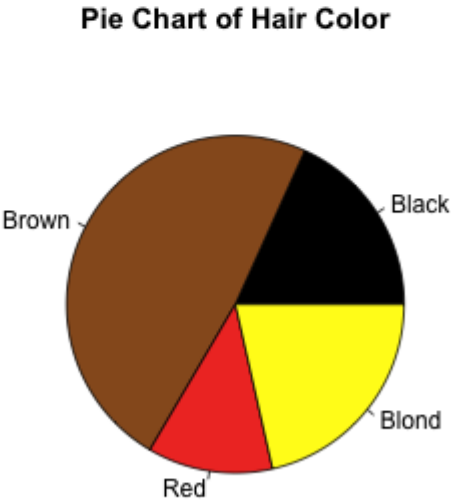
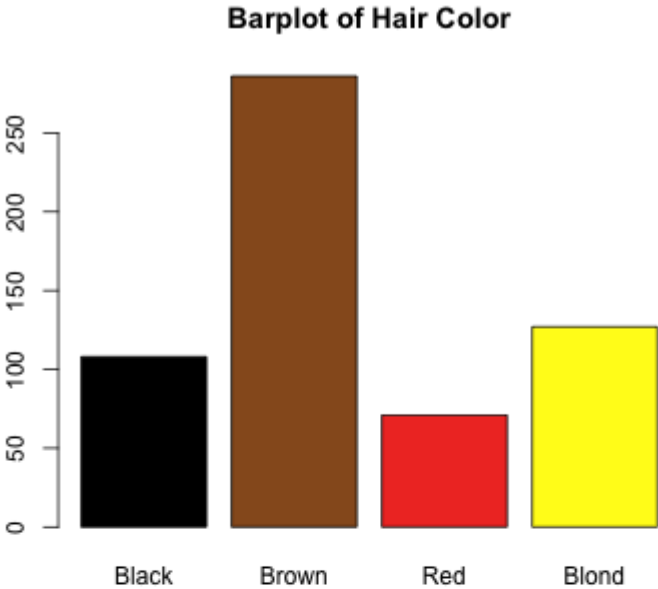
Black	Brown	Red	Blond
108	286	71	127

The basic graphical tool for categorical data is the barplot. This plots bars for each category, the height of which is the frequency or relative frequency of that category.

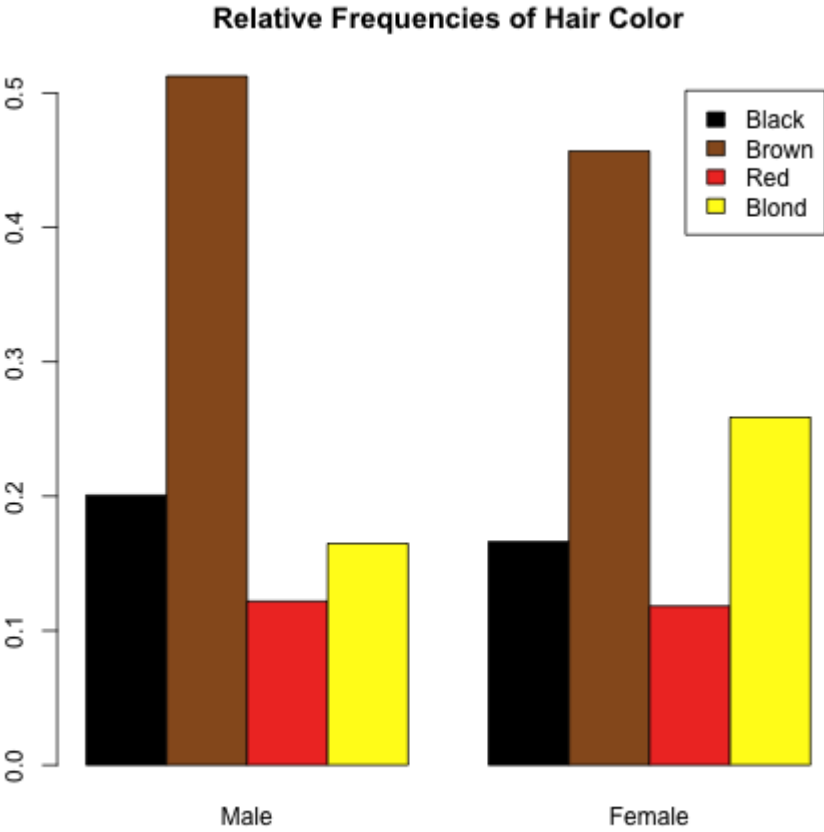
Barplot of Hair Color



Barplots are more effective than pie charts because we can more readily make a visual comparison of rectangles that have the same base width than a visual comparison of wedges in a pie.

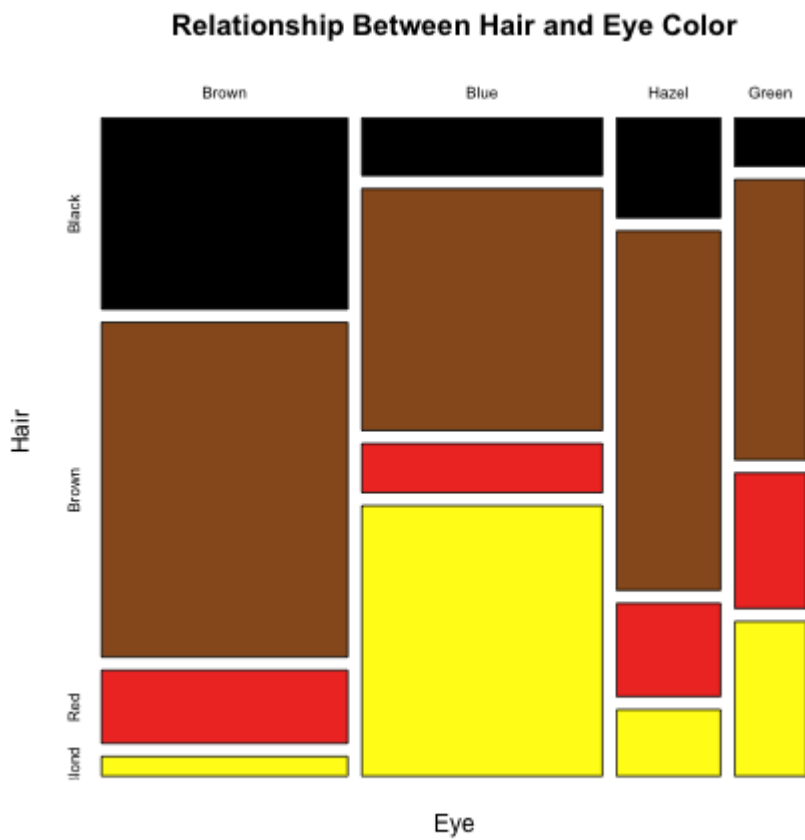


If a second categorical variable also is observed, for example hair color and gender, a barplot with side-by-side bars for each level of the first variable plotted contiguously, and each such group plotted with space between groups, is most effective to compare each level of the first variable across levels of the second. For example, the following plot shows how hair color is distributed for a sample of males and females. A comparison of the relative frequencies for males and females shows that a relatively higher proportion of females have blond hair and somewhat lower proportion of females have black or brown hair.



We can also display the relationship between hair and eye color using a 2-dimensional frequency table and barplot. The areas of the rectangles in this plot represent the relative frequency of the corresponding category combination.

Hair	Eye			
	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

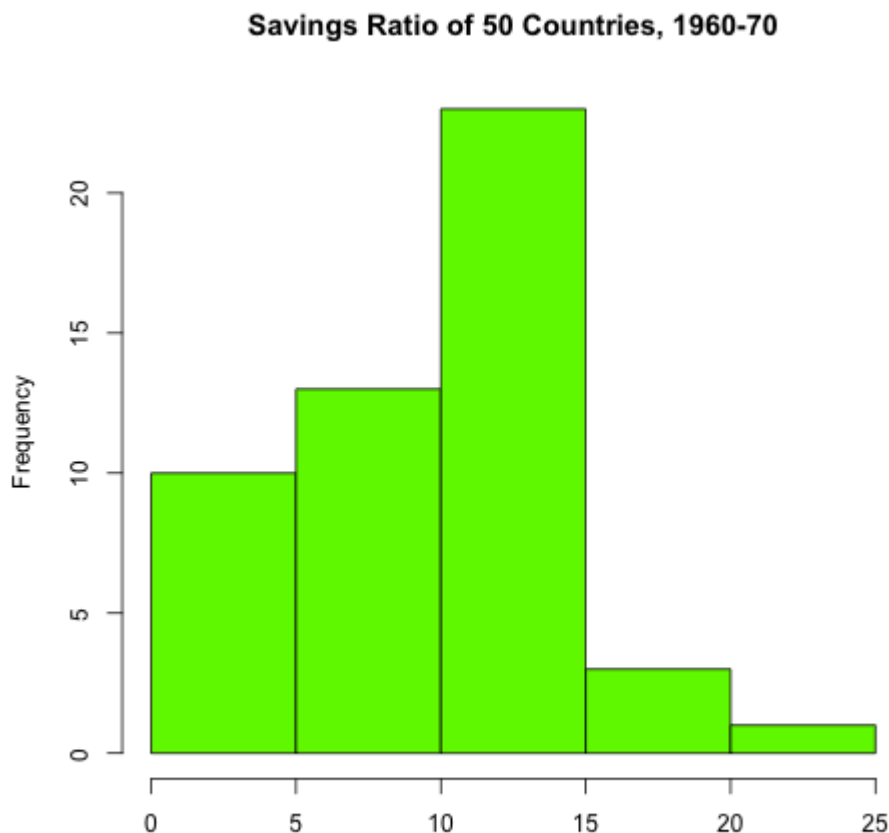


Quantitative data

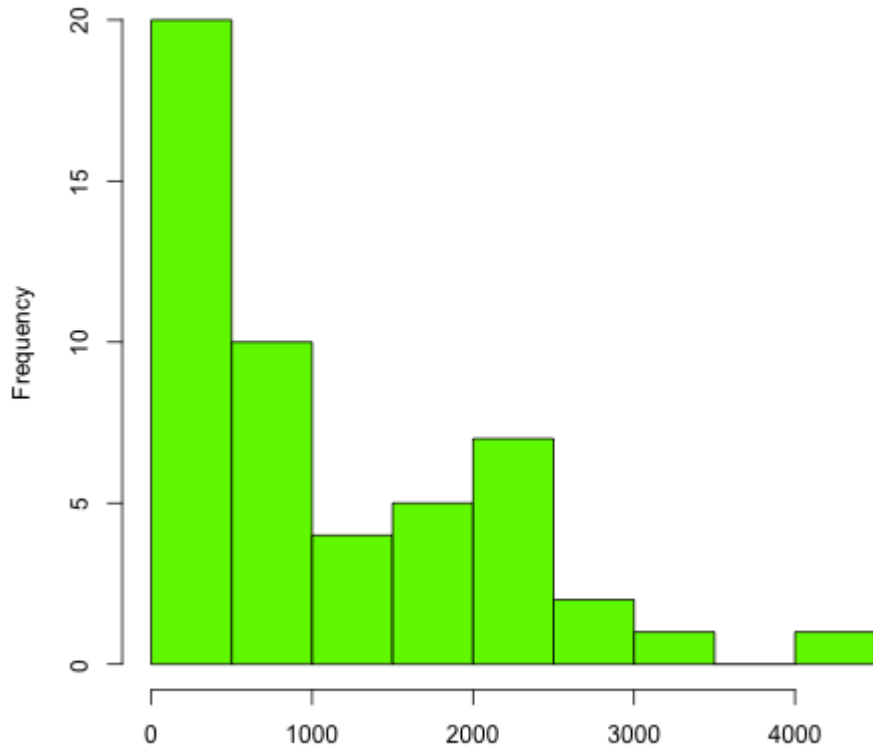
Data in which the values represent some numerical quantity are referred to as quantitative data. For example, here is a portion of a dataset that contains savings rates along with other demographic variables for 50 countries during 1960-70.

	sr	pop15	pop75	dpi	ddpi
Australia	11.43	29.35	2.87	2329.68	2.87
Austria	12.07	23.32	4.41	1507.99	3.93
Belgium	13.17	23.80	4.43	2108.47	3.82
...					

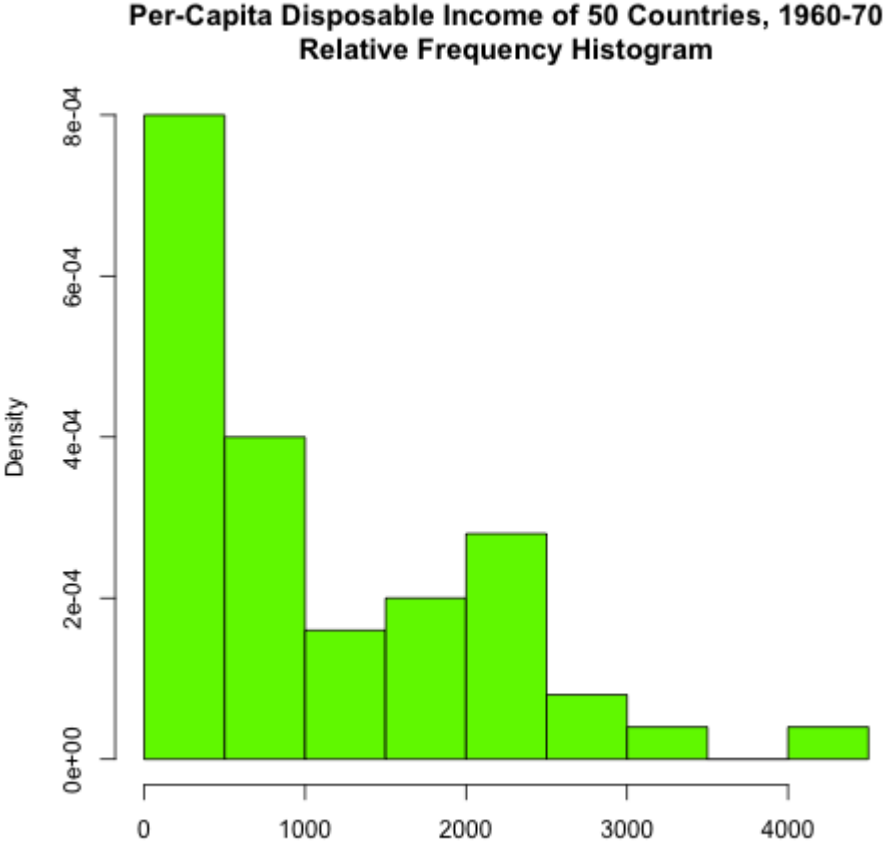
In this dataset *sr* represents savings ratio, *pop15* represents the percent of population under age 15, *pop75* is the percent of population over age 75, *dpi* is the real per-capita disposable income, and *ddpi* is the percent growth rate of *dpi*. The most commonly used graphical method for summarizing quantitative data is the **histogram**. To construct a histogram, we first partition the data values into a set of non-overlapping intervals and then obtain a frequency table. A histogram is the barplot of the corresponding frequency data. Here are histograms for savings ratio and disposable income.



Per-Capita Disposable Income of 50 Countries, 1960-70

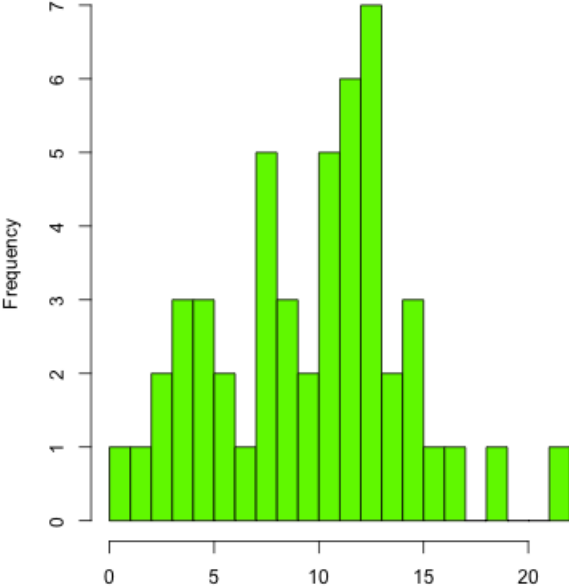
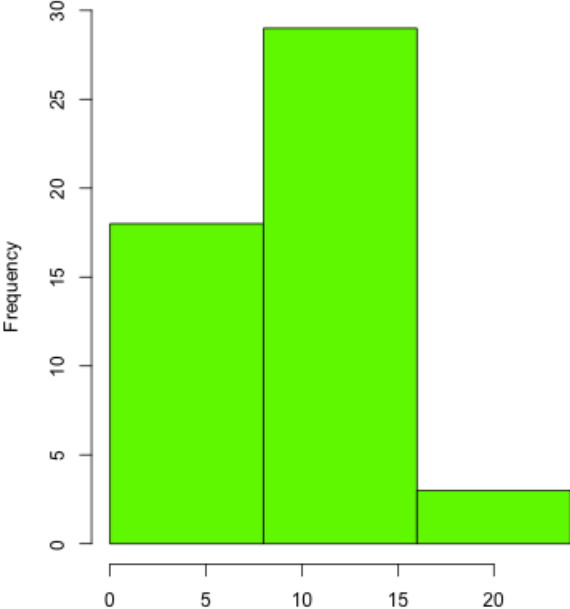


In some applications, the proportions within the sub-intervals are of greater interest than the frequencies. In these cases a relative frequency histogram can be used instead. In this case the vertical axis is re-scaled by dividing the frequencies by the total number of observations. The shape of a relative frequency histogram is unchanged; the only quantity that changes is the scale of the vertical axis.

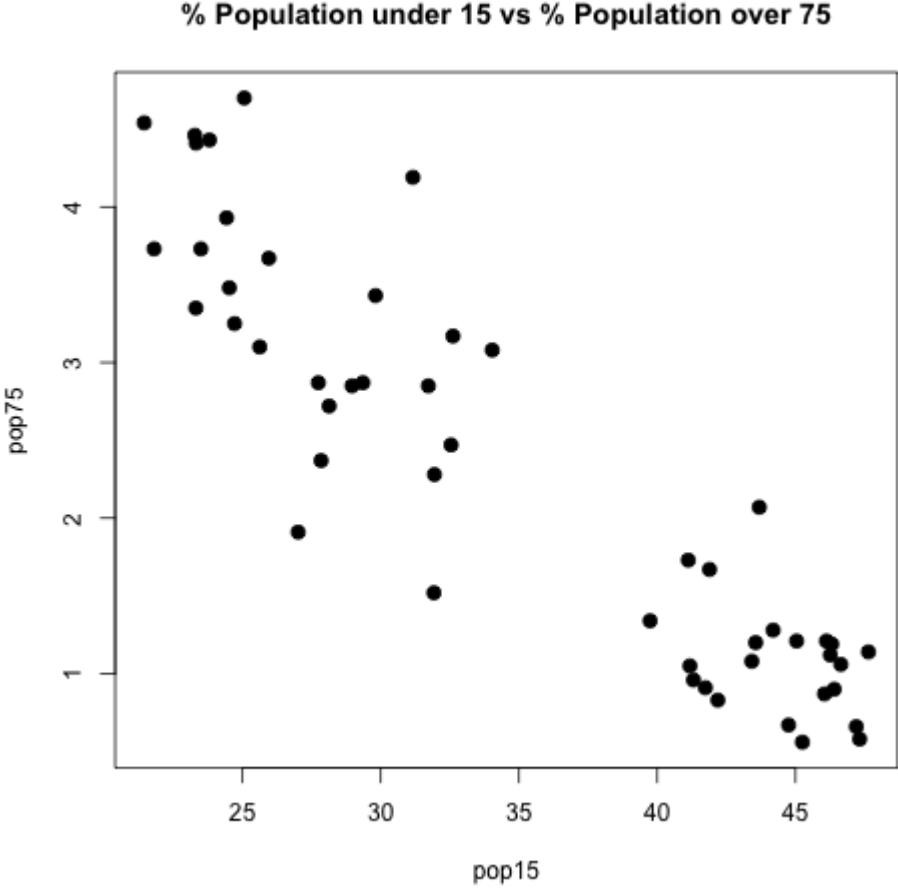


There is no fixed number of sub-intervals that should be used. A large number of sub-intervals corresponds to less summarization of the data, and a small number of sub-intervals corresponds to more summarization.

Savings Ratio of 50 Countries, 1960-70

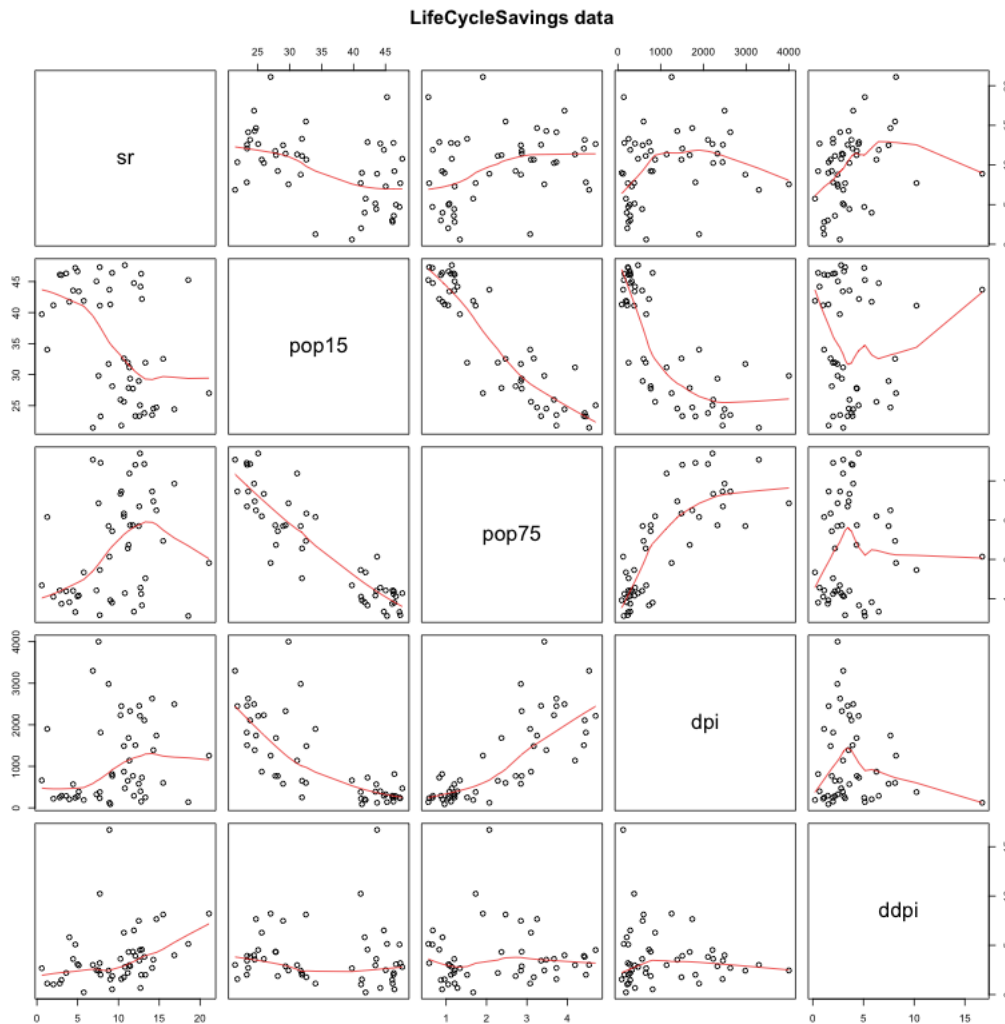


When two or more variables are measured for each individual in the dataset, then we may be interested in the relationship between these variables. The type of graphical display we use depends on the types of the variables. We have already seen an example of a 2-dimensional barplot for the case in which both variables are categorical. If both variables are quantitative, then the basic graphical tool is the **scatterplot**. For example, here is a scatterplot of *pop15* versus *pop75*.



The relationships among all 5 of the variables in this dataset can be displayed simultaneously by constructing pairwise scatterplots on the same graphic.

Note: we will defer until later in the course a discussion of numerical descriptions of these relationships.



Numerical summaries of data

Although graphical techniques are useful visualization tools, they are not very good for making decisions or inferences based on data. For those situations we need to consider numerical measures. Numerical measures describe various attributes of a dataset, the most common of which are measures of location and measures of dispersion.

Note: graphics for this section are generated by the script file <http://www.UTDallas.edu/~ammann/stat3355scripts/numeric.r>

Measures of Location

We used a histogram to describe the distribution of savings rate and per capita disposable income. Now suppose instead we would like to know where the middle of the savings rate and disposable income is located. This requires that we first define what we mean by the **middle** of a dataset. There are three such measures in common use: the **mean**, **median**, and **mode**.

The **mean** usually refers to the arithmetic mean or average. This is just the sum of the measurements divided by the number of measurements. We make a notational distinction between the mean of a population and the mean of a sample. The general rule is that greek letters are used for population characteristics and latin letters are used for sample characteristics. Therefore,

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i,$$

denotes the (arithmetic) mean of a population of N observations, and

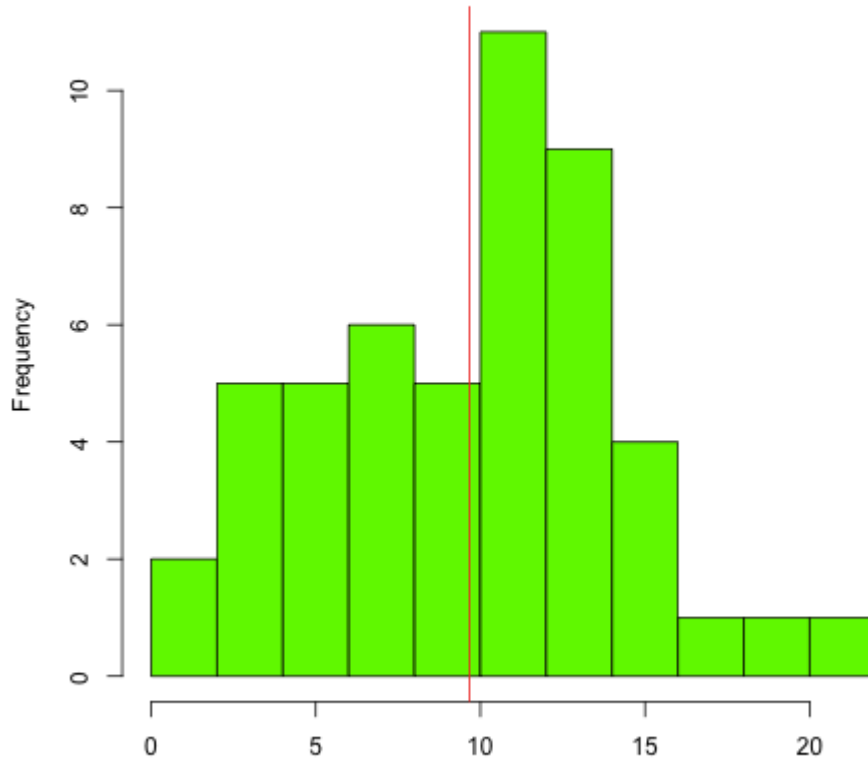
$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$

denotes the mean of a sample of size n selected from a population. The mean can be thought of as a center of gravity of the data values. That is, the histogram of the data would balance at the location defined by the mean. We can express this property mathematically by noting that the mean is the solution to the equation,

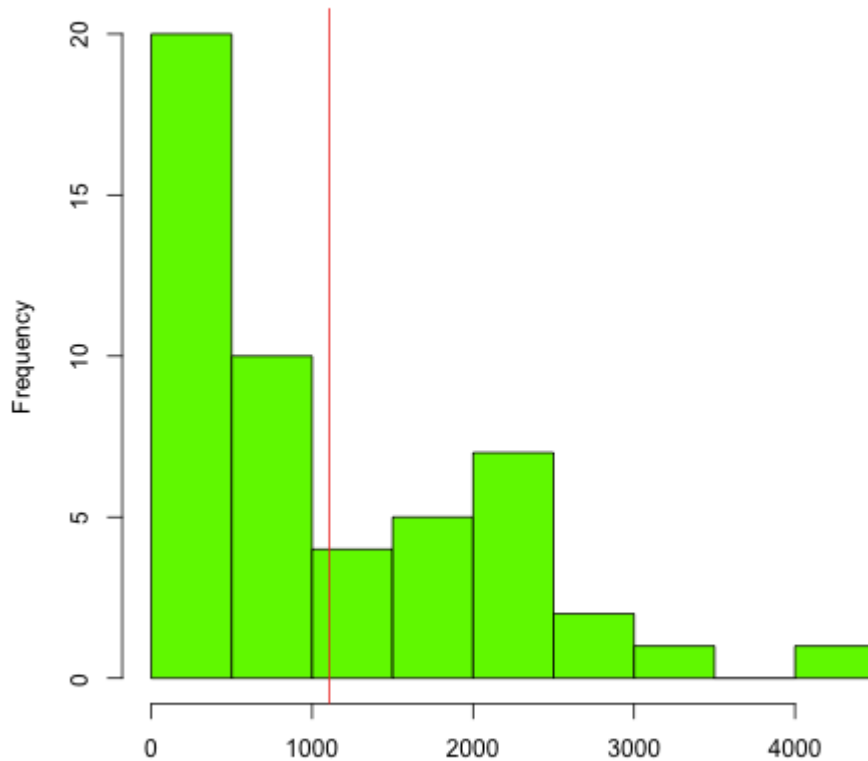
$$\sum_{i=1}^n (X_i - c) = 0.$$

This property of the mean has advantages and disadvantages. The mean is a natural measure of location for data that have a well-defined middle of high concentration with the frequency decreasing more or less evenly as we move away from the middle in either direction. The mean is not as useful when the data is heavily skewed. This is illustrated in the following two histograms. The first is the histogram of savings ratio with its mean superimposed, and the second is the histogram of disposable income.

Savings Ratio of 50 Countries, 1960-70



Per Capita Disposable Income of 50 Countries, 1960-70



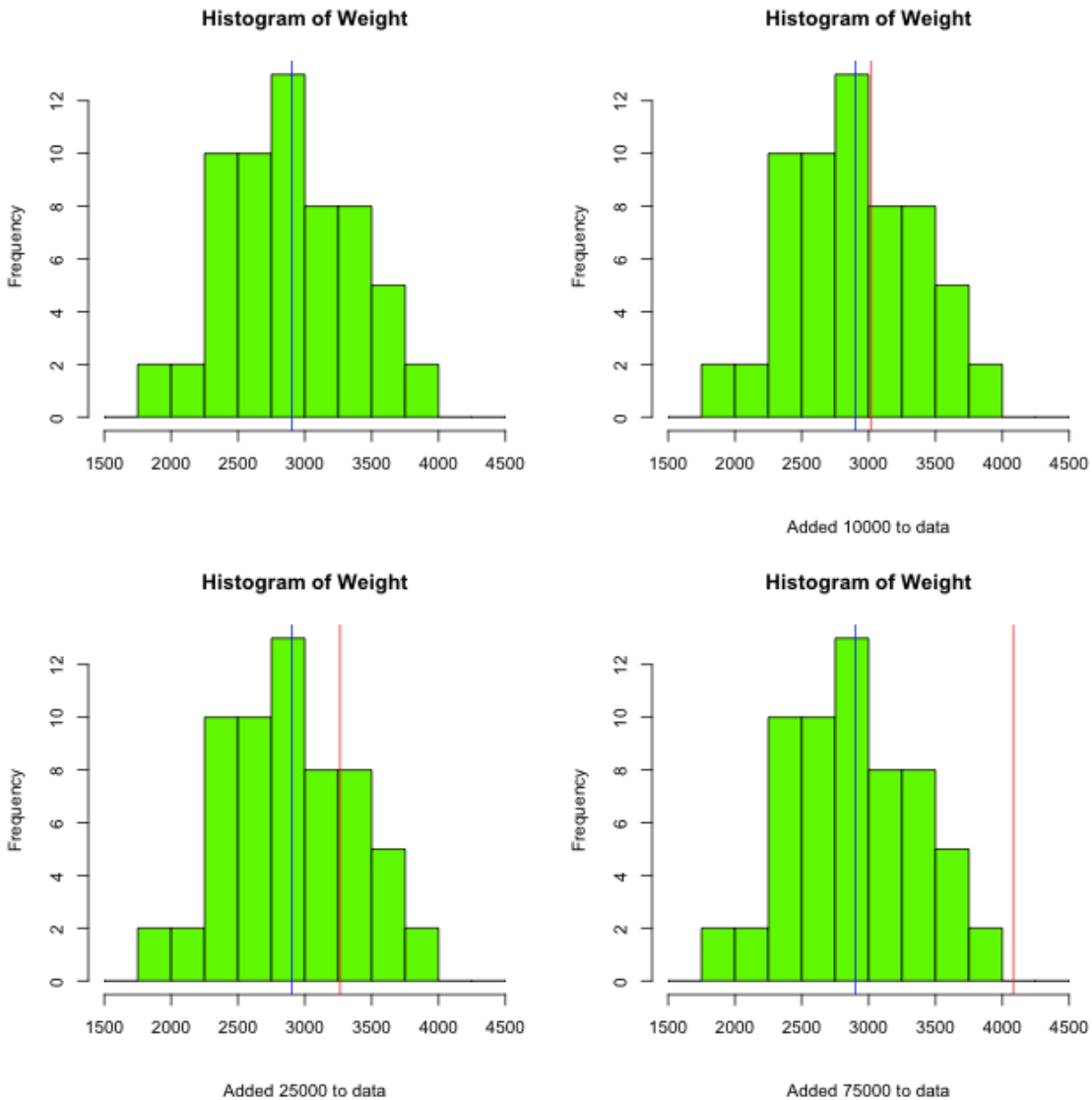
Another disadvantage of this measure is that it is very sensitive to the presence of a relatively few extreme observations. For example, the following data gives some quantities associated with 60 automobiles.

	Weight	Disp.	Mileage	Fuel	Type
Eagle Summit 4	2560	97	33	3.030303	Small
Ford Escort 4	2345	114	33	3.030303	Small
Ford Festiva 4	1845	81	37	2.702703	Small
Honda Civic 4	2260	91	32	3.125000	Small
Mazda Protege 4	2440	113	32	3.125000	Small
Mercury Tracer 4	2285	97	26	3.846154	Small
Nissan Sentra 4	2275	97	33	3.030303	Small
Pontiac LeMans 4	2350	98	28	3.571429	Small
Subaru Loyale 4	2295	109	25	4.000000	Small
Subaru Justy 3	1900	73	34	2.941176	Small
Toyota Corolla 4	2390	97	29	3.448276	Small
Toyota Tercel 4	2075	89	35	2.857143	Small
Volkswagen Jetta 4	2330	109	26	3.846154	Small

Chevrolet Camaro V8	3320	305	20	5.000000	Sporty
Dodge Daytona	2885	153	27	3.703704	Sporty
Ford Mustang V8	3310	302	19	5.263158	Sporty
Ford Probe	2695	133	30	3.333333	Sporty
Honda Civic CRX Si 4	2170	97	33	3.030303	Sporty
Honda Prelude Si 4WS 4	2710	125	27	3.703704	Sporty
Nissan 240SX 4	2775	146	24	4.166667	Sporty
Plymouth Laser	2840	107	26	3.846154	Sporty
Subaru XT 4	2485	109	28	3.571429	Sporty
Audi 80 4	2670	121	27	3.703704	Compact
Buick Skylark 4	2640	151	23	4.347826	Compact
Chevrolet Beretta 4	2655	133	26	3.846154	Compact
Chrysler Le Baron V6	3065	181	25	4.000000	Compact
Ford Tempo 4	2750	141	24	4.166667	Compact
Honda Accord 4	2920	132	26	3.846154	Compact
Mazda 626 4	2780	133	24	4.166667	Compact
Mitsubishi Galant 4	2745	122	25	4.000000	Compact
Mitsubishi Sigma V6	3110	181	21	4.761905	Compact
Nissan Stanza 4	2920	146	21	4.761905	Compact
Oldsmobile Calais 4	2645	151	23	4.347826	Compact
Peugeot 405 4	2575	116	24	4.166667	Compact
Subaru Legacy 4	2935	135	23	4.347826	Compact
Toyota Camry 4	2920	122	27	3.703704	Compact
Volvo 240 4	2985	141	23	4.347826	Compact
Acura Legend V6	3265	163	20	5.000000	Medium
Buick Century 4	2880	151	21	4.761905	Medium
Chrysler Le Baron Coupe	2975	153	22	4.545455	Medium
Chrysler New Yorker V6	3450	202	22	4.545455	Medium
Eagle Premier V6	3145	180	22	4.545455	Medium
Ford Taurus V6	3190	182	22	4.545455	Medium
Ford Thunderbird V6	3610	232	23	4.347826	Medium
Hyundai Sonata 4	2885	143	23	4.347826	Medium
Mazda 929 V6	3480	180	21	4.761905	Medium
Nissan Maxima V6	3200	180	22	4.545455	Medium
Oldsmobile Cutlass Ciera 4	2765	151	21	4.761905	Medium
Oldsmobile Cutlass Supreme V6	3220	189	21	4.761905	Medium
Toyota Cressida 6	3480	180	23	4.347826	Medium
Buick Le Sabre V6	3325	231	23	4.347826	Large
Chevrolet Caprice V8	3855	305	18	5.555556	Large
Ford LTD Crown Victoria V8	3850	302	20	5.000000	Large
Chevrolet Lumina APV V6	3195	151	18	5.555556	Van
Dodge Grand Caravan V6	3735	202	18	5.555556	Van
Ford Aerostar V6	3665	182	18	5.555556	Van
Mazda MPV V6	3735	181	19	5.263158	Van
Mitsubishi Wagon 4	3415	143	20	5.000000	Van

Nissan Axxess 4	3185	146	20	5.000000	Van
Nissan Van 4	3690	146	19	5.263158	Van

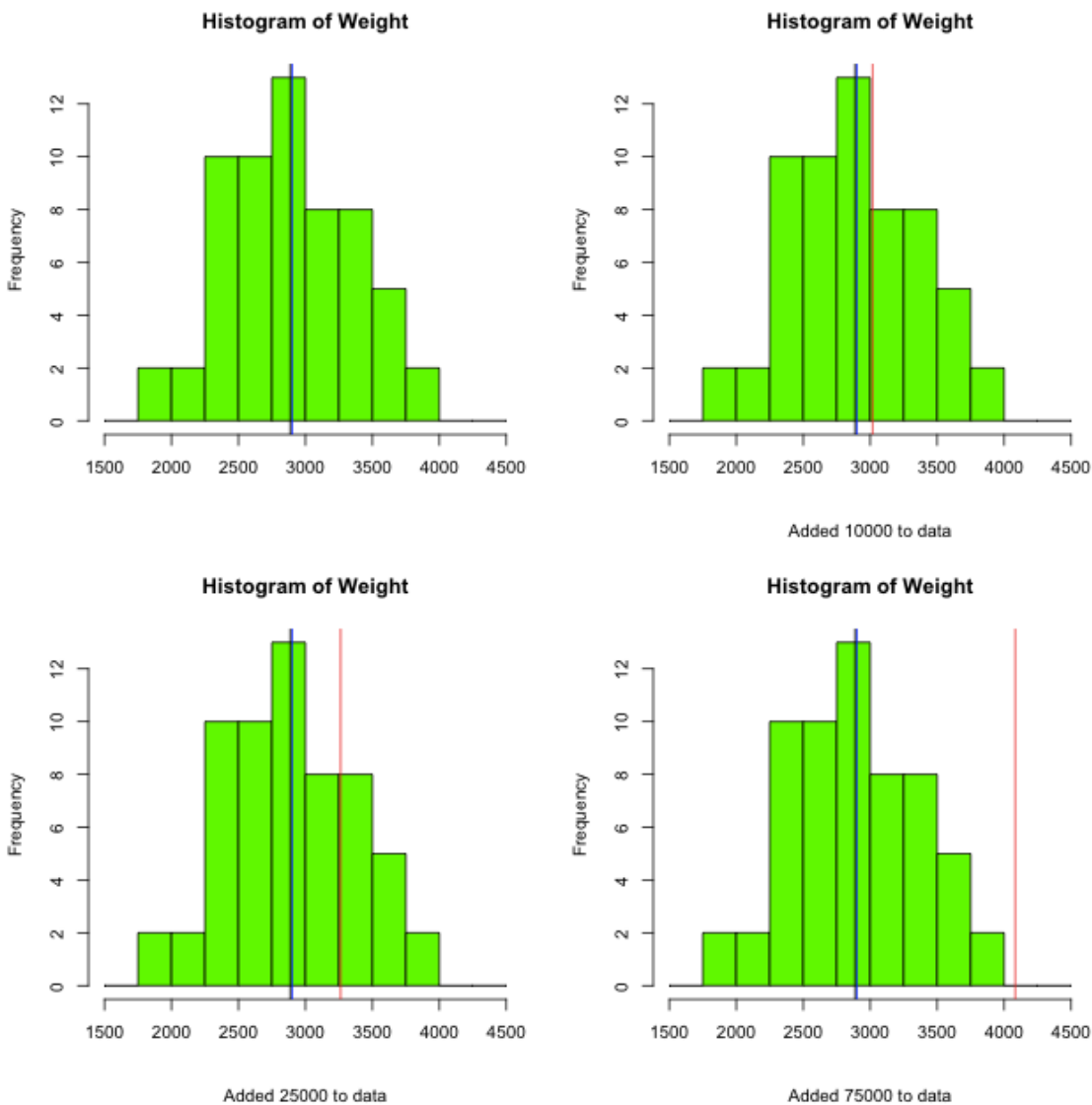
The 4 plots given below represent histograms of **Weight** with the mean of Weight superimposed. The second, third, and fourth plots are histograms of Weight with the values 10000, 25000, and 70000, respectively, added to the dataset. The *blue* line is the original mean and the red lines are the means of the modified data.



An alternative measure of location is the **median**. This measure is defined to be a number such that half of the measurements are below this number and half are above. The advantage of this measure is that it is not sensitive to the presence of a few outliers. Also, it gives an intuitive description of location regardless of the shape of the histogram. The median is obtained by first ordering the data values from smallest to largest. it the number

of observations n is odd, then the median is the ordered value in position $(n+1)/2$. If n is even, then the median is half-way between the $n/2$ and $n/2 + 1$ ordered values.

The plots below are identical to the previous plots except that the median is superimposed in black on each histogram. Note that the location of the median is much more stable than the mean. For that reason the median is used to describe the middle of data such as real estate prices and wages.



The **mode** is simply the most frequently occurring measurement or category. It is not used much except for some very specialized applications.

R notes

There is a dataset named *state.x77* in **R** that is a matrix with 50 rows and 8 columns. We can obtain the means for each column using the function *colMeans*:

```
state.means = colMeans(state.x77)
```

There also is a vector named *state.region* giving the geographic region (Northeast, South, North Central, West) for each state. We can use this to extract data for states belonging to a particular region as follows.

```
NorthEast.x77 = state.x77[state.region == "Northeast",]  
South.x77 = state.x77[state.region == "South",]  
NorthCentral.x77 = state.x77[state.region == "North Central",]  
West.x77 = state.x77[state.region == "West",]
```

Suppose we wanted to build a matrix that contains the means for each variable within each region so that rows correspond to region and columns correspond to variables. We could accomplish that as follows.

```
#construct blank matrix with dimnames  
state.means = matrix(0,4,dim(state.x77)[2],  
                    dimnames=list(levels(state.region),dimnames(state.x77)[[2]]))  
state.means["Northeast",] = colMeans(NorthEast.x77)  
state.means["South",] = colMeans(South.x77)  
state.means["North Central",] = colMeans(NorthCentral.x77)  
state.means["West",] = colMeans(West.x77)  
state.means  
round(state.means,2)
```

Measures of Dispersion

It is possible to have two very different datasets with the same means and medians. For that reason, measures of the middle are useful but limited. Another important attribute of a dataset is its dispersion or variability about its middle. The most useful measures of dispersion are the **range**, **percentiles**, and the **standard deviation**. The **range** is the difference between the largest and the smallest data values. Therefore, the more spread out the data values are, the larger the range will be. However, if a few observations are relatively far from the middle but the rest are relatively close to the middle, the range can give a distorted measure of dispersion.

Percentiles are positional measures for a dataset that enable one to determine the relative standing of a single measurement within the dataset. In particular, the p^{th} *%ile* is defined to be a number such that $p\%$ of the observations are less than or equal to that number and $(100 - p)\%$ are greater than that number. So, for example, an observation that is at the 75th *%ile* is less than only 25% of the data. In practice, we often cannot satisfy

the definition exactly. However, the steps outlined below at least satisfies the spirit of the definition.

1. Order the data values from smallest to largest; include ties.
2. Determine the position,

$$k.ddd = 1 + \frac{p(n-1)}{100}.$$

3. The p^{th} %ile is located between the k^{th} and the $(k+1)^{th}$ ordered value. Use the fractional part of the position, $.ddd$ as an interpolation factor between these values. If $k = 0$, then take the smallest observation as the percentile and if $k = n$, then take the largest observation as the percentile. For example, if $n = 75$ and we wish to find the 35th percentile, then the position is $1 + 35 * 74/100 = 26.9$. The percentile is then located between the 26th and 27th ordered values. Suppose that these are 57.8 and 61.3, respectively. Then the percentile would be

$$57.8 + .9 * (61.3 - 57.8) = 60.95.$$

Note. Quantiles are equivalent to percentiles with the percentile expressed as a proportion (70th %ile is the .70 quantile).

The 50th percentile is the median and partitions the data into a lower half (below median) and upper half (above median). The 25th, 50th, 75th percentiles are referred to as *quartiles*. They partition the data into 4 groups with 25% of the values below the 25th percentile (lower quartile), 25% between the lower quartile and the median, 25% between the median and the 75th percentile (upper quartile), and 25% above the upper quartile. The difference between the upper and lower quartiles is referred to as the *inter-quartile range*. This is the range of the middle 50% of the data.

The third measure of dispersion we will consider here is associated with the concept of distance between a number and a set of data. Suppose we are interested in a particular dataset and would like to summarize the information in that data with a single value that represents the *closest* number to the data. To accomplish this requires that we first define a measure of distance between a number and a dataset. One such measure can be defined as the *total distance between the number and the values in the dataset*. That is, the distance between a number c and a set of data values, X_i , $1 \leq i \leq n$, would be

$$D(c) = \sum_{i=1}^n |X_i - c|.$$

It can be shown that the value that minimizes $D(c)$ is the median. However, this measure of distance is not widely used for several reasons, one of which is that this minimization problem does not always have a unique solution.

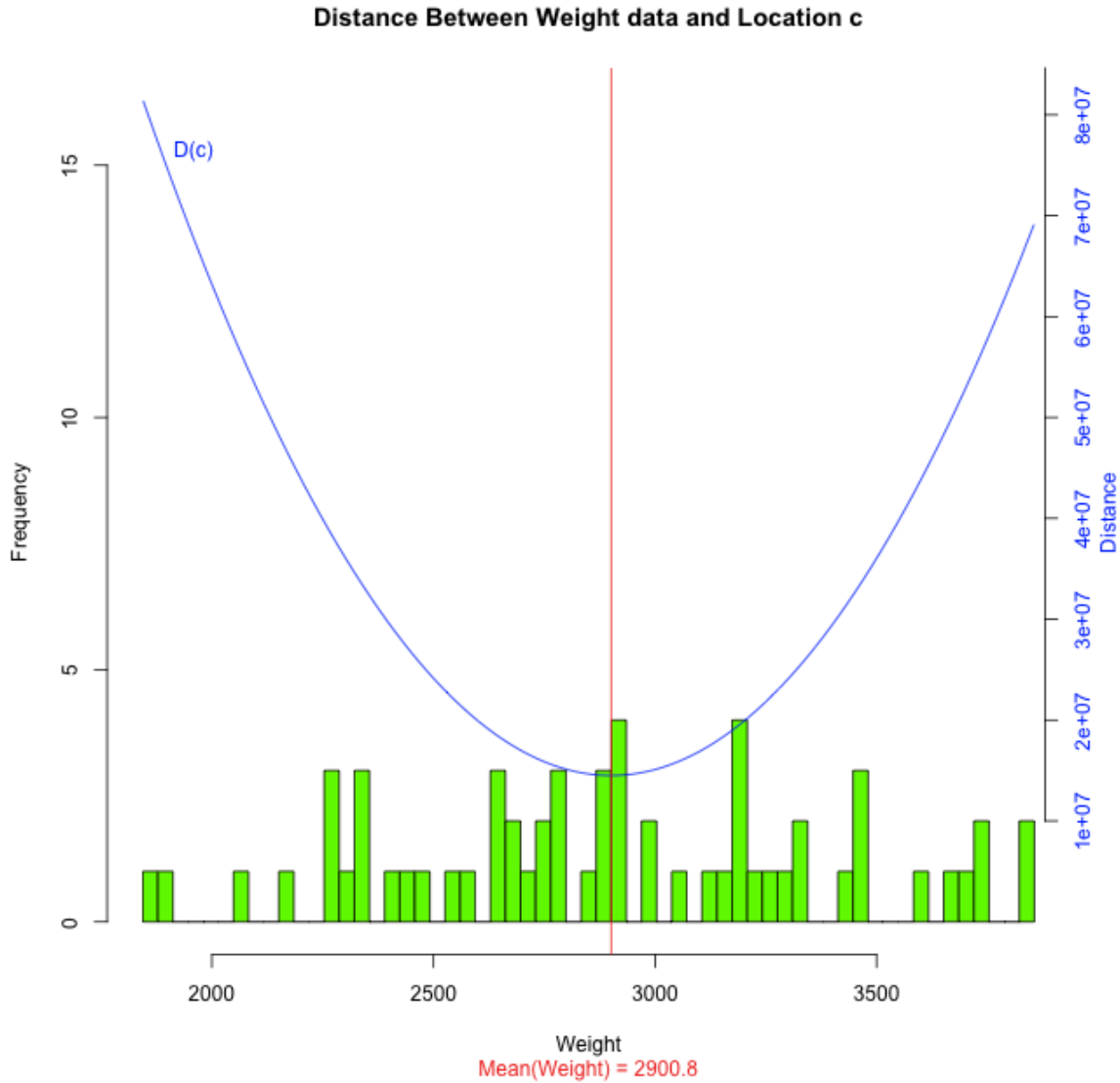
An alternative measure of distance between a number and a set of data that is widely used and does have a unique solution is defined by,

$$D(c) = \sum_{i=1}^n (X_i - c)^2.$$

That is, the distance between a number c and the data is the sum of the *squared* distances between c and each data value. We can take as our single number summary the value of c that is closest to the dataset, i.e., the value of c which minimizes $D(c)$. It can be shown that the value that minimizes this distance is $c = \bar{X}$. This is accomplished by differentiating $D(c)$ with respect to c and setting the derivative equal to 0.

$$0 = \frac{\partial}{\partial c} D(c) = \sum_{i=1}^n -2(X_i - c) = -2 \sum_{i=1}^n (X_i - c).$$

As we have already seen, the solution to this equation is $c = \bar{X}$. The graphic below gives a histogram of the Weight data with the distance function $D(c)$ superimposed. This graph shows that the minimum distance occurs at the mean of Weight.



The mean is the closest single number to the data when we define distance by the square of the deviation between the number and a data value. The *average distance* between the data and the mean is referred to as the **variance** of the data. We make a notational distinction and a minor arithmetic distinction between variance defined for populations and variance defined for samples. We use

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2,$$

for population variances, and

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

for sample variances. Note that the unit of measure for the variance is the square of the unit of measure for the data. For that reason (and others), the square root of the variance, called the **standard deviation**, is more commonly used as a measure of dispersion,

$$\sigma = \sqrt{\sum_{i=1}^N (X_i - \mu)^2 / N},$$

$$s = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)}.$$

Note that datasets in which the values tend to be far away from the middle have a large variance (and hence large standard deviation), and datasets in which the values cluster closely around the middle have small variance. Unfortunately, it is also the case that a dataset with one value very far from the middle and the rest very close to the middle also will have a large variance. **See sections 4.1, 4.2 in the textbook for details and examples.**

The standard deviation of a dataset can be interpreted by **Chebychev's Theorem**:

for any $k > 1$, the proportion of observations within the interval $\mu \pm k\sigma$ is at least $(1 - 1/k^2)$.

For example, the mean of the *Mileage* data is 24.583 and the standard deviation is 4.79. Therefore, at least 75% of the cars in this dataset have weights between $24.583 - 2 * 4.79 = 15.003$ and $24.583 + 2 * 4.79 = 34.163$. Chebychev's theorem is very conservative since it is applicable to every dataset. The actual number of cars whose Mileage falls in the interval (15.003,34.163) is 58, corresponding to 96.7%. Nevertheless, knowing just the mean and standard deviation of a dataset allows us to obtain a rough picture of the distribution of the data values. Note that the smaller the standard deviation, the smaller is the interval that is guaranteed to contain at least 75% of the observations. Conversely, the larger the standard deviation, the more likely it is that an observation will not be close to the mean. From the point of view of a manufacturer, reduction in variability of some product characteristic would correspond to an increase of consistency of the product. From the point of view of a financial manager, variability of a portfolio's return is referred to as volatility.

Note that **Chebychev's Theorem** applies to all data and therefore must be conservative. In many situations the actual percentages contained within these intervals are much higher than the minimums specified by this theorem. If the shape of the data histogram is known, then better results can be given. In particular, if it is known that the data histogram is approximately bell-shaped, then we can say

$\mu \pm \sigma$ contains approximately 68%,
 $\mu \pm 2\sigma$ contains approximately 95%,
 $\mu \pm 3\sigma$ contains essentially all

of the data values. This set of results is called the **empirical rule**. Later in the course we will study the bell-shaped curve (known as the normal distribution) in more detail.

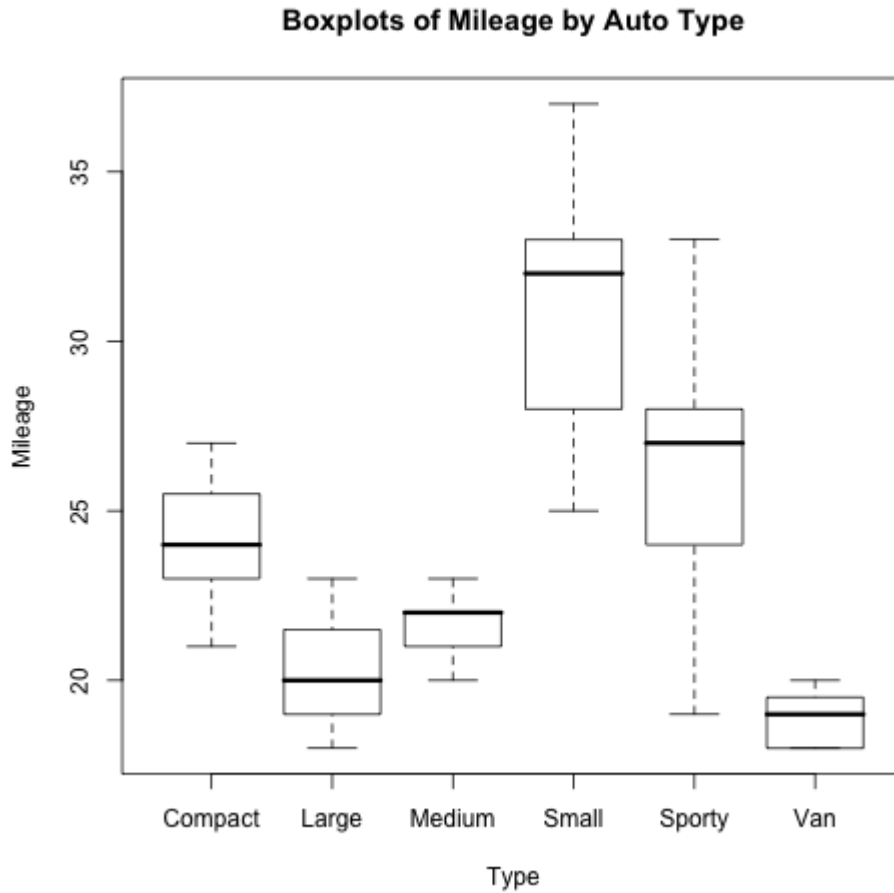
The relative position of an observation in a data set can be represented by its distance from the mean expressed in terms of the s.d. That is,

$$z = \frac{x - \mu}{\sigma},$$

and is referred to as the z-score of the observation. Positive z-scores are above the mean, negative z-scores are below the mean. Z-scores greater than 2 are more than 2 s.d.'s above the mean. From Chebychev's theorem, at least 75% of observations in any dataset will have z-scores between -2 and 2

Since z-scores are dimension-less, then we can compare the relative positions of observations from different populations or samples by comparing their respective z-scores. For example, directly comparing the heights of a husband and wife would not be appropriate since males tend to be taller than females. However, if we knew the means and s.d.'s of males and females, then we could compare their z-scores. This comparison would be more meaningful than a direct comparison of their heights.

If the data histogram is approximately bell-shaped, then essentially all values should be within 3 s.d.'s of the mean, which is an interval of width 6 s.d.'s. A small number of observations that are unusually large or small can greatly inflate the s.d. Such observations are referred to as outliers. Identification of outliers is important, but this can be difficult since they will distort the mean and the s.d. For that reason, we can't simply use $\bar{X} \pm 2s$ or $\bar{X} \pm 3s$ for this purpose. We instead make use of some relationships between quartiles and the s.d. of bell-shaped data. In particular, if the data histogram is approximately bell-shaped, then $IQR \approx 1.35s$. This relationship can be used to define a robust estimate of the s.d. which is then used to identify outliers. Observations that are more than 1.5(IQR) from the nearest quartile are considered to be outliers. Boxplots in **R** are constructed so that the box edges are at the quartiles, the median is marked by a line within the box, and this the box is extended by whiskers indicating the range of observations that are no more than 1.5(IQR) from the nearest quartile. Any observations falling outside this range are plotted with a circle. For example, the following plot shows boxplots of mileage for each automobile type.

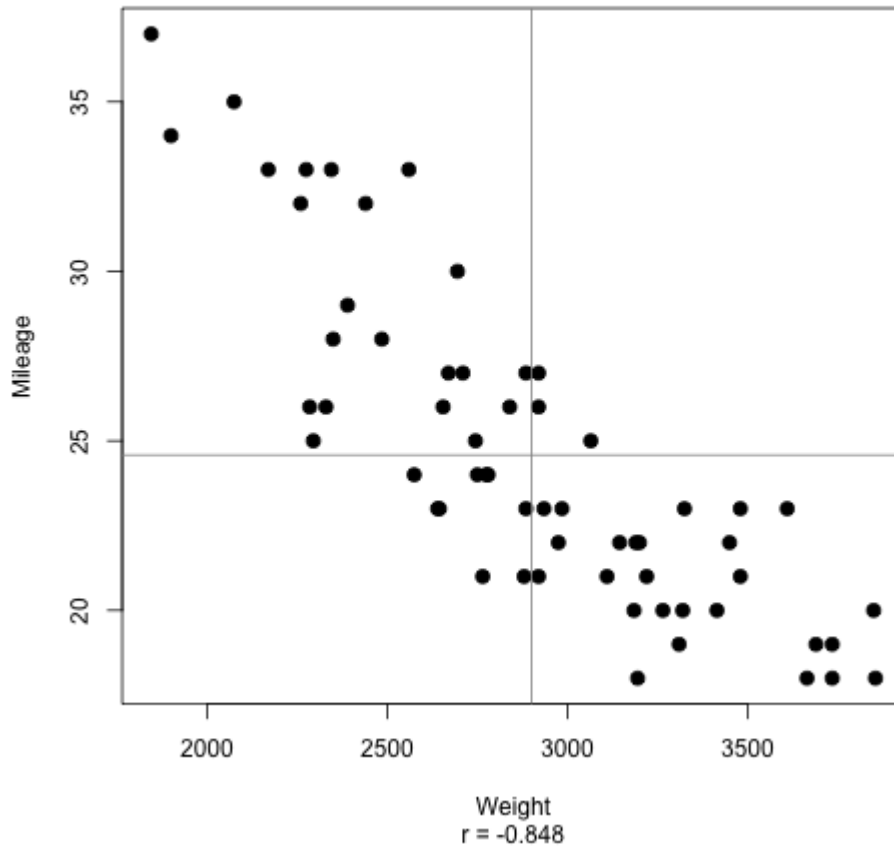


See sections 4.3, 4.4 in the textbook for details and additional examples.

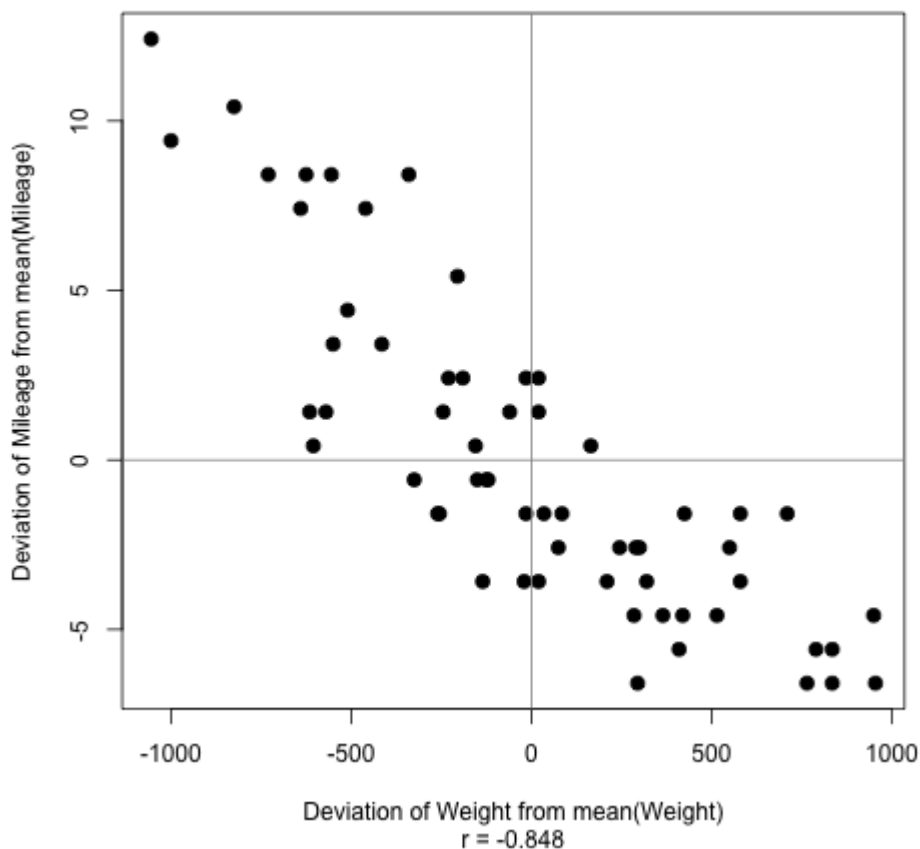
Measures of Association

The automobile dataset given above includes both Weight and Mileage of 60 automobiles. In addition to describing location and dispersion for each variable separately, we also may be interested in what kind of relationship exists between these variables. The following figure represents a scatterplot of these variables with the respective means superimposed. This shows that for a high percentage of cars, those with above average Weight tend to have below average Mileage, and those with below average Weight have above average Mileage. This is an example of a decreasing relationship, and most of the data points in the plot fall in quadrants 2,4. In an increasing relationship, most of the points will fall in quadrants 1,3.

Scatterplot of Weight vs Mileage



Deviations from the Means: Weight vs Mileage



We can derive a measure of association for two variables by considering the deviations of the data values from their respective means. Note that the product of deviations for a data point in quadrants 2,3 is positive and the product of deviations for a data point in quadrants 1,3 is negative. Therefore, most of these products for variables with a strong increasing relationship will be positive, and most of these products for variables with a strong decreasing relationship will be negative. This implies that the sum of these products will be a large positive number for variables that have a strong increasing relationship, and the sum will be a large negative number for variables that have a strong decreasing relationship. This is the motivation for using

$$r = \frac{\frac{1}{N} \sum_{i=1}^N (X_i - \mu_x)(Y_i - \mu_y)}{\sigma_x \sigma_y}.$$

as a measure of association between two variables. This quantity is called the **correlation coefficient**. The denominator of r is a scale factor that makes the correlation coefficient dimension-less and scales so that $0 \leq |r| \leq 1$. Note that this can be expressed equivalently

as

$$r = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_x s_y}.$$

If the correlation coefficient is close to 1, then the variables have a strong increasing relationship and if the correlation coefficient is close to -1, then the variables have a strong decreasing relationship. If the correlation is exactly 1 or -1, then the data must fall exactly on a straight line. The correlation coefficient is limited in that it is only valid for *linear* relationships. A correlation coefficient close to 0 indicates that there is no *linear* relationship. There may be a strong relationship in this case, just not linear. Furthermore, the correlation may understate the strength of the relationship even when r is large, if the relationship is non-linear.

The correlation coefficient between Weight and Mileage is -0.848. This is a fairly large negative number, and so there is a fairly strong linear, decreasing relationship between Weight and Mileage. This is confirmed by the scatterplot. Since these variables are so strongly related, we can ask how well can we predict Mileage just by knowing the Weight of a vehicle. To answer this question, we first define a measure of distance between a dataset and a line.

Suppose we have measured two variables for each individual in a sample, denoted by $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, and we wish to predict the value of Y given the value of X for a particular individual using a straight line for the prediction. A reasonable approach would be to use the line that comes closest to the data for this prediction. Let $Y=a+bX$ denote the equation of a prediction line, and let $\hat{Y}_i = a + bX_i$ denote the predicted value of Y for X_i . The difference between an actual and predicted Y -value represents the error of prediction for that data point. We define the *distance* between a prediction line and a point in the dataset to be the square of the prediction error for that observation. The total distance between the actual and predicted Y -values is then the sum of the squared errors, which is the variance of the prediction errors multiplied by n . Since the predicted values, and hence the errors, depend on the slope and intercept of the prediction line, we can express this total distance by

$$D(a, b) = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - a - bX_i)^2.$$

Our goal now is to find the line that is closest to the data using this definition of distance. This line has slope and intercept that minimize $D(a, b)$. We can use differential calculus to find the minimum.

$$\frac{\partial}{\partial a} D(a, b) = -2 \sum_{i=1}^n (Y_i - a - bX_i),$$

$$\frac{\partial}{\partial b} D(a, b) = -2 \sum_{i=1}^n X_i (Y_i - a - bX_i).$$

Setting these equal to 0 gives the system of equations

$$0 = \sum_{i=1}^n (Y_i - a - bX_i) = n(\bar{Y} - b\bar{X} - a),$$

$$0 = \sum_{i=1}^n X_i Y_i - na\bar{X} - b \sum_{i=1}^n X_i^2.$$

Therefore,

$$a = \bar{Y} - b\bar{X},$$

and, after substituting for a in the second equation and solving for b ,

$$b = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}.$$

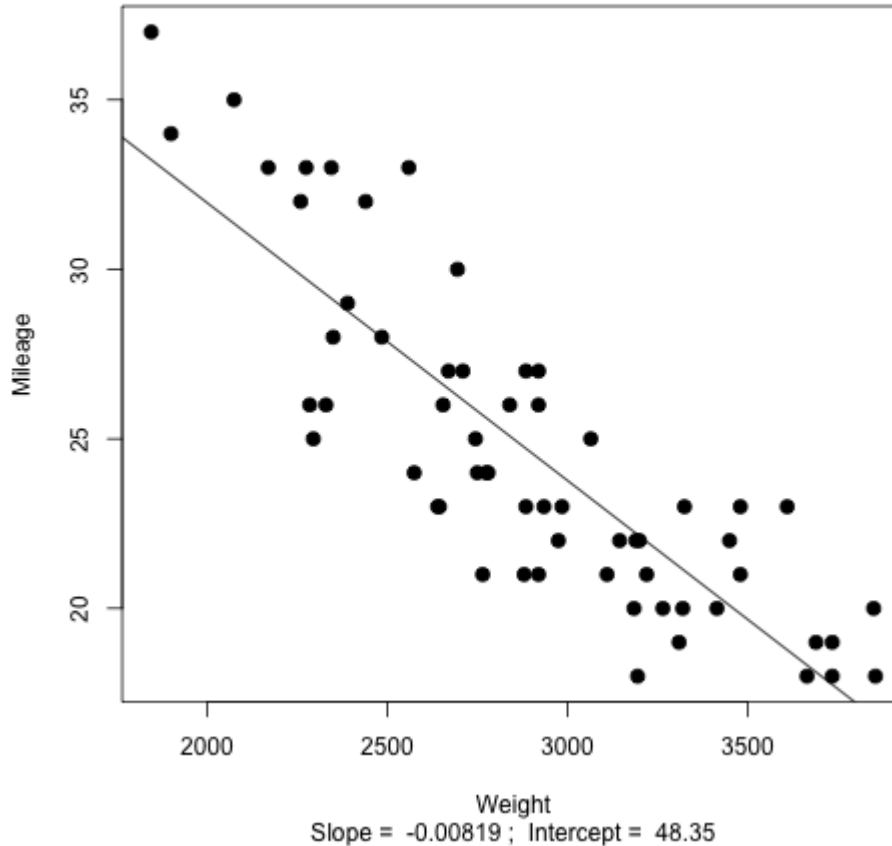
It can be shown that the numerator equals $(n-1)r s_x s_y$ and the denominator equals $(n-1)s_x^2$. Hence,

$$b = r \frac{s_y}{s_x}, \quad a = \bar{Y} - b\bar{X}.$$

The prediction line, referred to as the **least squares regression line**, is then

$$\hat{Y} = a + bX.$$

Scatterplot of Weight vs Mileage



The next question that can be asked related to this prediction problem is how well does the prediction line predict? We can't answer that question completely yet because the full answer requires inference tools that we have not yet covered, but we can give a descriptive answer to this question. The distance measure, $D(a,b)$, represents the variance of the prediction errors. One way of describing how well the prediction line performs is to compare it to the best prediction we could obtain without using the X values to predict. In that case, our predictor would be a single number. We have already seen that the closest single number to a dataset is the mean of the data, so in this case, the best predictor based only on the Y values is \bar{Y} . This corresponds to a horizontal line with intercept \bar{Y} , and so the distance between this line and the data is $D(\bar{Y}, 0)$. This quantity represents the error variance for the best predictor that does not make use of the X values, and so the difference,

$$D(\bar{Y}, 0) - D(a, b),$$

represents the reduction in error variance (improvement in prediction) that results from use

of the X values to predict. If we express this as a percent,

$$100 \frac{D(\bar{Y}, 0) - D(a, b)}{D(\bar{Y}, 0)},$$

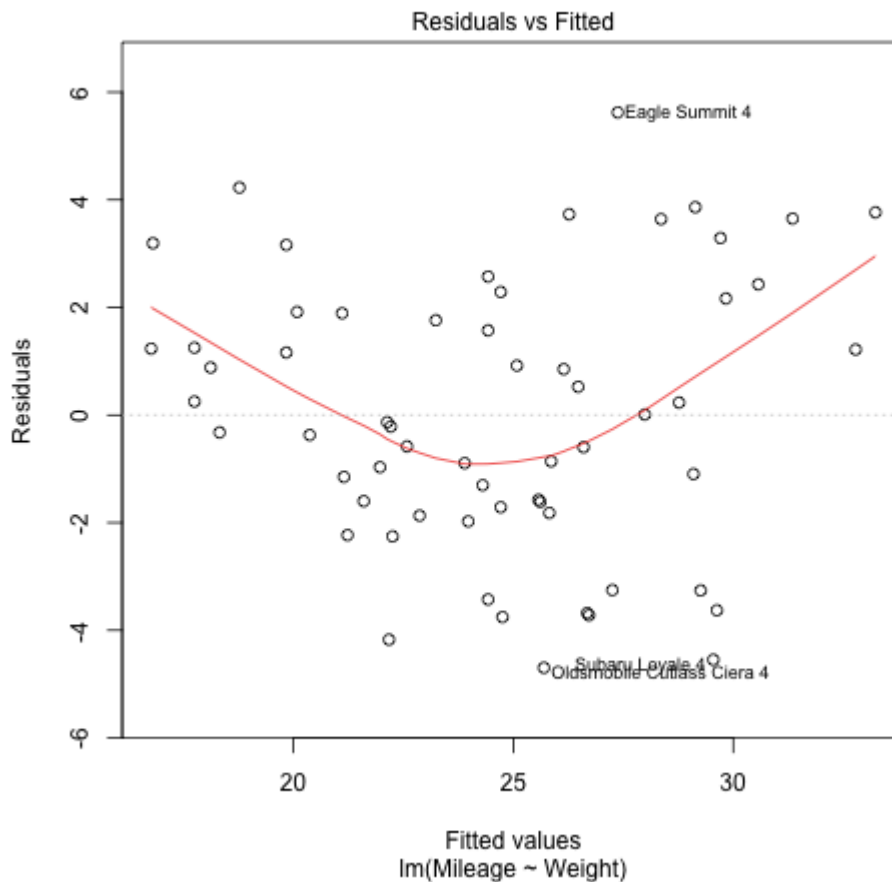
then this is the percent of the error variance that can be removed if we use the least squares regression line to predict as opposed to simply using the mean of the Y 's. It can be shown that this quantity is equal to the square of the correlation coefficient,

$$r^2 = \frac{D(\bar{Y}, 0) - D(a, b)}{D(\bar{Y}, 0)}.$$

R-squared also can be interpreted as the proportion of variability in the Y -variable that can be explained by the presence of a linear relationship between X and Y .

In the automobile example, the correlation between Weight and Mileage was $r = -0.848$, and so $r^2 = 0.719$. If we use the regression line to predict Mileage based on Weight, we can remove 71.9% of the variance of the Mileage data by using Weight to predict Mileage. Another way of expressing this is to ask: Why don't all cars have the same mileage. Part of the answer to that question is that cars don't all weigh the same and there is a fairly strong linear relationship between weight and mileage that accounts for 71.9% of the variability in mileage. This leaves 28.1% of this variability that is related to other factors, including the possibility of a non-linear relationship between Mileage and Weight.

To help judge the adequacy of a linear regression fit, we can plot the residuals vs the predictor variable X . The residuals are the prediction errors, $e_i = Y_i - \hat{Y}_i$, $1 \leq i \leq n$. If a linear fit is reasonable, then the residuals should have no discernable relationship with X and should be essentially noise. This plot for a linear fit to predict Mileage based on Weight is shown below.



This shows that the residuals are still related to Weight, so a linear fit is not adequate. Note that removal of the linear component of the relationship between weight and mileage, as represented by the residuals from a linear fit, does a better job of revealing this non-linearity than a scatterplot of these variables. This will be discussed in greater detail later.

It is important to remember that correlation is a mathematical concept that says nothing about causation. The presence of a strong correlation between two variables indicates that there *may* be a causal relationship, but does not prove that one exists, nor does it indicate the direction of any causality. **Read pages 266-7 in the textbook for a more thorough discussion of this and related issues.**

The **R** code to generate the graphics in this section can be found at:
<http://www.utdallas.edu/~ammann/stat3355scripts/NumericGraphics.r>

Probability

Probability is a mathematical description of a process whose outcome is uncertain. We call such a process an **experiment**. This could be something as simple as tossing a coin or as

complicated as a large-scale clinical trial consisting of three phases involving hundreds of patients and a variety of treatments. The **sample space** of an experiment is the set of all possible outcomes of the experiment, and an **event** is a set of possible outcomes, that is, a subset of the sample space.

For example, the sample space of an experiment in which three coins are tossed consists of the outcomes

$$\{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

while the sample space of an experiment in which a disk drive is selected and the time until first failure is observed would consist of the positive real numbers. In the first case, the event that exactly one head is observed is the set $\{HHT, HTH, THH\}$. In the second case, the event that the time to first failure of the drive exceeds 1000 hours is the interval $(1000, \infty)$.

Probability arose originally from descriptions of games of chance – gambling – that have their origins far back in human history. It is usually interpreted as the proportion or percentage of times a particular outcome is observed if the experiment is repeated a large number of times. We can think of this proportion as representing the *likelihood* of that outcome occurring whenever the experiment is performed. **Probability** is formally defined to be a function that assigns a real number to each event associated with an experiment according to a set of basic rules. These rules are designed to coincide with our intuitive notions of likelihood, but they must also be mathematically consistent.

This mathematical representation is simplest when the sample space contains a finite or countably infinite number of elements. However, our mathematics and our intuition collide when working with an experiment that has an uncountable sample space, for example an interval of real numbers. Consider for example the following experiment. You purchase a spring driven clock, set it at 12:00 (ignore AM and PM), wind the clock and let it run until it stops. We can represent the sample space of this experiment as the interval, $[0, 12)$, and we can ask questions such as

1. What is the probability the clock stops between 1:00 and 2:00?
2. What is the probability the clock stops between 4:00 and 4:30?
3. What is the probability the clock stops between 7:05 and 7:06?

We can answer each of these questions using our intuitive ideas of likelihood. For the first question, since we know nothing about the clock, we can assume that there is no preference of one interval of time over any other interval of time for the clock to stop. Therefore, we would expect that each of the 12 intervals of length one hour are equally likely to contain the stopping time of the clock, and so the likelihood that it stops between 1:00 and 2:00 would be $1/12$. Similarly, the likelihood that it stops between 4:00 and 4:30 would be $1/24$ since there are 24 intervals of length $1/2$ hour, and the likelihood that it stops between 7:05 and 7:06 would be $1/720$ since there are 720 intervals of length one minute. In each case our intuition tells us that the likelihood of an event for this experiment is the reciprocal of the number

of non-overlapping intervals of the same length, since each such interval is assumed to be equally likely to contain the stopping point of the clock. Note also that the interval $[1, 2)$, corresponding to the times between 1:00 and 2:00, contains the non-overlapping intervals, $[1, 1.5)$ and $[1.5, 2)$. Each of these intervals would have likelihoods $1/24$ and the sum of these two likelihoods equals the likelihood of the entire interval. This illustrates the additive nature of likelihood that we have for this concept.

A problem occurs if we ask a question such as what is the probability that the clock stops at precisely $\sqrt{2}$ minutes past 1? In this case there is an uncountably infinite number of such times in the interval $[0, 12)$, so that the likelihood we would assign to such an event would be $1/\infty = 0$. However, the sum of the likelihoods for all such events between 1:00 and 2:00 would be 0, not $1/12$ as we have derived above. This inconsistency requires that we modify the rules somewhat. In the case of uncountably infinite sample spaces, we only require that probability be defined for an *interesting* set of events. In the case of the clock experiment, this *interesting* set of events would consist of all interval subsets of the sample space with positive length along with events that can be formed from countable unions and intersections of such intervals. This collection of events is referred to as the **probability space** for the experiment. In the case of finite or countably infinite sample spaces, the probability space can be the set of all possible subsets of the sample space. Unless specified otherwise, all events used here are assumed to be in the probability space.

The basic rules or axioms of probability are then:

1. Probability is a function $P : \mathcal{F} \rightarrow [0, 1]$, where \mathcal{F} is the probability space. That is, the probability function assigns a number between 0 and 1 to each event in the probability space.
2. $P(S) = 1$, where S is the sample space. That is, the probability that an outcome in the sample space occurs is 1.
3. For any countable collection of mutually exclusive events in \mathcal{F} , A_i , $i \geq 1$, we have

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

That is, the probability of the union of non-overlapping events equals the sum of the individual probabilities.

All other properties of probability derive from these basic axioms along with any additional definitions we construct.

As noted previously, when working with experiments that have equally likely outcomes, it is only necessary to count the number of outcomes contained in events to determine their probabilities. Events in many such experiments involve the selection of objects from a population. There are different methods used for counting outcomes for such situations depending on whether or not the selection order of the selected objects is recognized and whether a selected object is returned (**selection with replacement**) to the population before the next

selection is made or not returned (**selection without replacement**). We use the term **permutation** to refer to selection of objects in which selection order is distinguished and use the term **combinations** to refer to the case in which selection order is not distinguished. We will consider here three of these methods, permutations with and without replacement, and combinations without replacement.

Permutations without replacement

If the object selected is not returned to the population before the next object is selected, then an object can appear in the selected subset no more than once. There are n choices in the population to fill the first position, but then that leaves $n-1$ choices in the population to fill the second position. Therefore, there are $n(n-1)$ ways to fill the first two positions. Continuing this argument, we can see that there are $n(n-1)\dots(n+1-k)$ ways to select k objects without replacement from a population of n objects when selection order is distinguished. This number is commonly expressed using factorial notation,

$$n(n-1)\dots(n+1-k) = \frac{n!}{(n-k)!}$$

Permutations with replacement

This case occurs when we wish to select k objects with replacement from a population of n objects and selection order is distinguished. Replacement implies that the same object could be selected multiple times. What is required is to count the number of distinct sets of k objects could be selected in this way. We can view this selection process by considering the ways in which each of the positions, $1, \dots, n$, of the set are filled. Note that there are n choices in the population to fill the first position, and since the object selected for this first position is then returned to the population, there are n choices available for the second selection as well. Therefore, there are n^2 ways to fill the first two positions. Continuing this argument, we can see that there are n^k ways to select k objects with replacement from a population of n distinguishable objects.

Combinations without replacement

The only difference between this case and the case of permutations without replacement is that the selection order of the k selected objects is not distinguished here. This implies that a different arrangement of the same objects is not counted for this case and so this case involves simply selecting a subset of size k from the population. Therefore, we can view the number of permutations without replacement as a two-stage process: first select a subset (combinations without replacement) and then generate every possible rearrangement of each of these subsets. Note that the number of ways to generate every possible rearrangement of k objects is equivalent to counting the number of permutations without replacement of

k objects selected from a population of size k , and so is equal to $k!$. Denote by $C(n,k)$ the number of combinations without replacement. Then we have,

$$\frac{n!}{(n-k)!} = C(n,k)k!.$$

Hence,

$$C(n,k) = \frac{n!}{k!(n-k)!}.$$

This quantity is usually denoted by

$$\binom{n}{k}$$

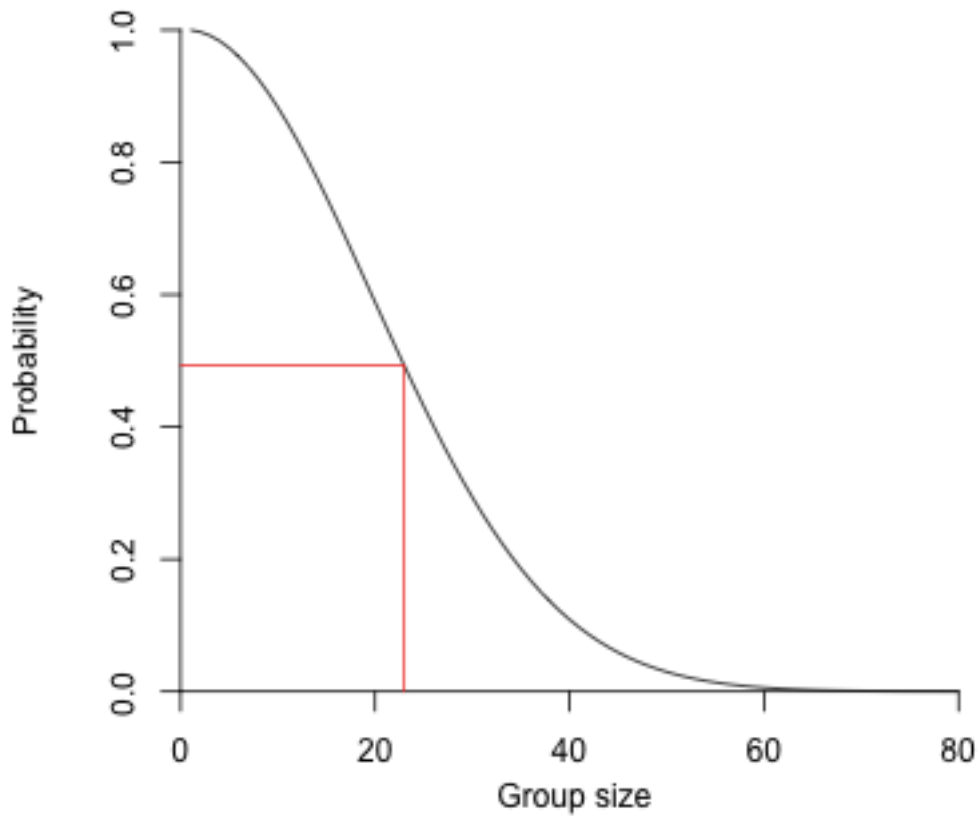
Examples

Birthday Problem. This is a classic example of how probability in some applications does not coincide with our intuition. Suppose we have a class of n individuals and want to determine the probability that there is at least one pair of individuals who have the same birthday. To simplify this problem, we will ignore birthdays that occurred on Feb. 29 during a leap year and count those as occurring on March 1. The model we will assume for this problem treats an individual's birthdate as if it was randomly selected from the set of 365 possible birthdays. Therefore, the experiment in which each individual selects a birthdate is an experiment with equally likely outcomes. Therefore, we must count the number of ways to select n birthdays from the population of 365 possible birthdays, and then count the number of ways to select n birthdays with at least one matching pair. It turns out to be easier to count the number of ways to select n birthdays with no matches. This is equivalent to counting the number of permutations without replacement of k objects selected from a population of 365 objects. This number is therefore $365!/(365-n)!$. The total number of possible birthdates for this group is equivalent to the number of permutations with replacement of n birthdates from the population of 365 possible birthdates. This gives,

$$P(\text{no birthdate matches}) = \frac{365!/(365-n)!}{365^n}.$$

A plot of this probability as a function of n is given below. Note that when $n=23$, there is about a 50% probability of no matches in the group, and when $n=50$, there is about a 3% chance of no matches in the group.

Probability of No Birthdate Match



$p = 0.4927$ when group size = 23

Binomial coefficients. Note that the number of combinations without replacement occurs in the binomial series,

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}.$$

Now consider an experiment in which two gamblers play a series of 10 games, the results of which are independent. That is, the event that the first gambler wins (or loses) on game r is independent of the event that he wins (or loses) any other game. Suppose that the probability that the first gambler wins a particular game is p , and his probability of winning any other game is the same. Find the probability that the first gambler wins exactly 4 games. To solve this problem, first note that an arbitrary outcome of this experiment can be represented by a string of 10 characters, each of which is either W or L , denoting the outcomes of each game. The event that the first gambler wins 4 games consists of all possible strings in which W occurs 4 times and L occurs 6 times. Each such string can be specified by the 4 positions of W in this string. For example, the outcome $WWWWLLLLLL$ could be specified by the positions, 1234 of W . Since the games are independent, the probability of observing this outcome would be

$$P(WWWWWLLLLLL) = pppp(1-p)(1-p)(1-p)(1-p)(1-p)(1-p) = p^4(1-p)^6.$$

Any other outcome with exactly 4 wins would just be a rearrangement of the 10 characters in the string, and so would have the same probability. Therefore, the probability that the first gambler wins exactly 4 games is this probability times the number of such outcomes. We can obtain this number by counting the number of combinations of 4 positions taken from the possible 10 positions for W in the string. Hence,

$$P(4 \text{ wins}) = \binom{10}{4} p^4 (1-p)^6.$$

Using the same arguments, we can see that

$$P(k \text{ wins in } 10 \text{ games}) = \binom{10}{k} p^k (1-p)^{10-k}, \quad 0 \leq k \leq 10.$$

We can easily extend this to n games to obtain

$$P(k \text{ wins in } n \text{ games}) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leq k \leq n.$$

Finally, note that these probabilities are terms in a binomial series, and that

$$\sum_{k=0}^n P(k \text{ wins in } n \text{ games}) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = (p + 1 - p)^n = 1.$$

Subpopulation selection. Suppose a committee consists of 40 males and 20 females and must select a subcommittee of 5 members. It decides to make this selection randomly.

What is the probability that all 5 members of the subcommittee will be female? What is the probability that at least 2 member of the subcommittee will be male? First note that an outcome of this experiment is a set of 5 members selected without replacement from the committee, and this experiment has equally likely outcomes. To answer the questions, we will first obtain the probability that exactly k members of the subcommittee will be female for $0 \leq k \leq 5$. Note that if k members are female, then $5-k$ members will be male. Hence, the number of outcomes contained in the event that exactly k members are female can be obtained by counting the number of ways to select a subset of size k from the 20 females and multiplying that times the number of ways to select a subset of size $5-k$ from the 40 males. Since order of selection does not count, this number is then

$$\binom{20}{k} \binom{40}{5-k}.$$

The number of outcomes in the sample space is the total number of ways to select a subset of size 5 from the 60 committee members, so the probability that exactly k are female is,

$$P(k) = \frac{\binom{20}{k} \binom{40}{5-k}}{\binom{60}{5}}.$$

We can now answer the questions.

$$\begin{aligned} P(5 \text{ females}) &= P(5) = \frac{\binom{20}{5} \binom{40}{0}}{\binom{60}{5}} \\ &= \frac{20!5!55!}{5!15!60!} \\ &= \frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16}{60 \cdot 59 \cdot 58 \cdot 57 \cdot 56} \\ &= 0.0028. \end{aligned}$$

$$\begin{aligned} P(\text{at least 2 males}) &= P(\text{no more than 3 females}) = P(0) + P(1) + P(2) + P(3) \\ &= 1 - P(4) - P(5). \end{aligned}$$

$$\begin{aligned} P(4) &= \frac{\binom{20}{4} \binom{40}{1}}{\binom{60}{5}} \\ &= \frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 40 \cdot 5}{60 \cdot 59 \cdot 58 \cdot 57 \cdot 56} \\ &= 0.0355. \end{aligned}$$

So, $P(\text{at least 2 males}) = 1 - 0.0355 - 0.0028 = 1 - 0.0383 = .9617$.

Additional Properties of Probability

The **complement** of an event is defined to be the set of all outcomes contained in the sample space that are not contained in the event. It is denoted by A^c . Note that the complement of the sample space is defined to be the empty set, \emptyset , the set with no elements. Also, $A \cup \emptyset = A$ and $A \cap \emptyset = \emptyset$ for any event A . Therefore, if we set $A_1 = A$, $A_i = \emptyset$, $i \geq 2$, then $\{A_i\}$ is a countable collection of mutually exclusive events. Hence, from axiom 3 we have,

$$P(A) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A) + \sum_{i=2}^{\infty} P(\emptyset).$$

Since $P(\emptyset) \geq 0$ from Axiom 1, then this equation implies that $P(\emptyset) = 0$.

Note: mathematical equations are sentences with the same syntax as English and can be read as such. The set operations, *intersection*, *union*, and *complement* are often read as the English equivalents, *and*, *or*, and *not*, respectively. Also, the word *or* used in this context is assumed to mean the *inclusive or*.

Now let A_i , $1 \leq i \leq n$ be a finite collection of mutually exclusive events and set $A_i = \emptyset$ for $i > n$. Then from Axiom 3, we have

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &= P\left(\bigcup_{i=1}^{\infty} A_i\right) \\ &= \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^n P(A_i) + \sum_{i=n+1}^{\infty} P(\emptyset) \\ &= \sum_{i=1}^n P(A_i). \end{aligned}$$

That is, the probability of a finite union of mutually exclusive events equals the sum of the probabilities.

Suppose we are interested in an experiment in which the sample space consists of a finite collection of n outcomes, O_i , $1 \leq i \leq n$, and that each outcome is equally likely with probability p . Then the previous result implies that

$$1 = \sum_{i=1}^n P(O_i) = np.$$

Therefore, we must have $p = 1/n$. Furthermore, since an event for such an experiment may be written as the union of the individual outcomes contained in the event, then

$$P(A) = \frac{\#\{A\}}{n},$$

where $\#\{A\}$ represents the number of elements in the set A . For experiments with equally likely outcomes, the probability of an event is just the number of outcomes in the event divided by the total number of outcomes.

Next note that A and A^c are mutually exclusive and $A \cup A^c = S$. Therefore, from the previous result we have,

$$1 = P(A \cup A^c) = P(A) + P(A^c).$$

So, the probability of the complement of an event is one minus the probability of the event. This result is useful for situations in which an event of interest is very complicated and its probability is difficult to obtain directly, but the complement of the event is simple with an easily obtainable probability.

The axioms of probability tell us how to find the probability of the union of mutually exclusive events, but not how to find the probability of the union of arbitrary, not necessarily mutually exclusive, events. We can use the results derived thus far to solve this problem. Suppose we are interested in two events, A and B . We need to write the union of these two events as the union of two mutually exclusive events. This can be done by noting that $A \cup B = A \cup \{B \cap A^c\}$. Since A and $B \cap A^c$ are mutually exclusive, then

$$P(A \cup B) = P(A) + P(B \cap A^c).$$

Next note that $B = \{A \cap B\} \cup \{B \cap A^c\}$, which is a disjoint union. Therefore,

$$P(B) = P(A \cap B) + P(B \cap A^c)$$

and so,

$$P(B \cap A^c) = P(B) - P(A \cap B).$$

Combining this with the previous result gives,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

the probability of A union B equals the sum of the probabilities minus the probability of the intersection.

In a similar way, we can show that probability is a monotone function. Suppose that $A \subset B$. Then we may express B as a disjoint union, $B = A \cup (B \cap A^c)$ and apply the additivity property of probability,

$$\begin{aligned} P(B) &= P(A \cup (B \cap A^c)) \\ &= P(A) + P(B \cap A^c) \\ &\geq P(A), \end{aligned}$$

since $P(B \cap A^c) \geq 0$. Hence, if $A \subset B$, then $P(A) \leq P(B)$.

Another extension that can be derived directly from the axioms is an extremely useful result called the **Law of Total Probability**. A **partition** of the sample space is defined to be a collection, finite or countably infinite, of mutually exclusive events in the probability space whose union is the sample space. Suppose that $\{B_i\}$ is partition and A is an arbitrary

event. Then $A = \cup\{A \cap B_i\}$, and the events, $A \cap B_i$ are mutually exclusive. The Law of Total Probability is just the application of Axiom 3 to this expression,

$$P(A) = \sum P(A \cap B_i).$$

This property allows us to breakdown a complicated event A into more manageable pieces, $A \cap B_i$.

Example. Suppose a standard card deck (13 denominations in 4 suits) is well-shuffled and then the top card is discarded. What is the probability that the 2nd card (the new top card) is an ace? Let A enote the event that the 2nd card is an ace. The partitioning events we will use are the events

$$B_1 = \{1^{st} \text{ card is Ace}\}, \quad B_2 = \{1^{st} \text{ card is not Ace}\}$$

Then,

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) \\ &= P(1^{st} \text{ card is Ace} \cap 2^{nd} \text{ card is Ace}) + P(1^{st} \text{ card is not Ace} \cap 2^{nd} \text{ card is Ace}). \end{aligned}$$

The first term has numerator which is the number of ways the first card is an ace and the second card is an ace, and has denominator which is the total number of different outcomes for the first two cards. We can use permutations to count the number of outcomes for both numerator and denominator. The numerator is $4 \cdot 3$ and the denominator is $52 \cdot 51$. Hence,

$$P(1^{st} \text{ card is Ace} \cap 2^{nd} \text{ card is Ace}) = \frac{(4)(3)}{(52)(51)}.$$

Similarly, the second term is

$$P(1^{st} \text{ card is not Ace} \cap 2^{nd} \text{ card is Ace}) = \frac{(48)(4)}{(52)(51)}.$$

These give

$$\begin{aligned} P(A) &= \frac{(4)(3)}{(52)(51)} + \frac{(48)(4)}{(52)(51)} \\ &= \frac{(4)(3 + 48)}{(52)(51)} \\ &= \frac{(4)(51)}{(52)(51)} \\ &= \frac{4}{52} = \frac{1}{13}. \end{aligned}$$

Note that this probability is the same as the probability that the first card is an ace.

Conditional Probability

The additional properties derived above illustrate how the basic probability axioms can be extended to include concepts that coincide with our intuitive notions of probability. However, our mathematical representation of probability can lead us to some results that are not necessarily intuitive, but are true nonetheless. One example of this involves the concept of conditional probability.

Consider the following table of frequencies,

	Hired	Not Hired	Total
M	8	57	65
F	2	33	35
Total	10	90	100

where these represent qualified applicants for a position. Suppose we model this situation as an experiment in which a single applicant is randomly selected – selected in such a way that each applicant has the same chance of being selected. Then this is an experiment with equally likely outcomes. We can then use the counting rule to obtain the following probabilities:

$$P(\text{Hired}) = 10/100 = .10$$

$$P(\text{Not Hired}) = 90/100 = .9$$

$$P(\text{M}) = 65/100 = .65$$

$$P(\text{F}) = 35/100 = .35$$

$$P(\text{M and Hired}) = 8/100 = .08$$

$$P(\text{F and Hired}) = 2/100 = .02$$

The English translation of the first probability is “10% of all qualified applicants were hired”, and the translation of the last probability is “2% of all qualified applicants were female and were hired”.

Now suppose we perform this experiment, and discover that the person selected was male. We don’t know yet who was selected, but we do have some partial information about the outcome of the experiment. This partial information causes us to modify the probabilities of other events, since now the experiment can be viewed as randomly selecting one of the males. Therefore, the probabilities we would assign to each male after receiving this information would be $1/65$, and so, the probability that the selected applicant was hired, *given the information that a male was selected*, would then be $8/65 = .123$. We could then say that 12.3% of all qualified *male* applicants were hired. This modified probability is referred to as a conditional probability, and is expressed as

$$P(\text{Hired}|\text{M}) = \frac{8}{65}.$$

In general, the conditional probability of an event A , given the partial information about the experiment that an outcome in event B has occurred, is defined to be

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

In English: the conditional probability of A given B equals the probability of A and B divided by the probability of B .

Some other examples:

What proportion of females were hired?

$$P(Hired|F) = \frac{2}{35} = .057.$$

What proportion of those hired were males?

$$P(M|Hired) = \frac{8}{10}.$$

What proportion of those not hired were female?

$$P(F|Not Hired) = \frac{33}{90} = .367.$$

Note that in describing conditional probabilities, the reference group has changed from the sample space to the given event and that event is what appears after the $|$ in the conditional probability expression.

We can rearrange the definition of conditional probability to show how the probability of an intersection can be derived from a conditional probability,

$$P(A \cap B) = P(A|B)P(B).$$

This result also can be applied to the law of total probability by replacing $P(A \cap B_i)$ with $P(A|B_i)P(B_i)$,

$$P(A) = \sum P(A|B_i)P(B_i)$$

Example. An on-line computer system has four incoming communication lines with the properties described in the table below. What is the probability that a randomly chosen message has been received without error?

Line	Fraction of traffic	Fraction of messages without error
1	0.4	0.998
2	0.3	0.999
3	0.1	0.997
4	0.2	0.992

Solution. First note that the probabilities in column 3 are conditional probabilities because they do not refer to all messages, just those coming from the respective lines. The line numbers represent a partition of the sample space of all incoming messages, so we can separate messages by their line number and apply this extension of the Theorem of Total Probability.

$$\begin{aligned}
 P(w/o\ error) &= P(w/o\ error|Line\ 1)P(Line\ 1) + P(w/o\ error|Line\ 2)P(Line\ 2) \\
 &\quad + P(w/o\ error|Line\ 3)P(Line\ 3) + P(w/o\ error|Line\ 4)P(Line\ 4) \\
 &= (.998)(.4) + (.999)(.3) + (.997)(.1) + (.992)(.2) \\
 &= .997
 \end{aligned}$$

Note that this probability is essentially a weighted average of the individual line probabilities in which the weights are the proportions of incoming traffic carried by the lines.

A further extension of this result is known as **Bayes' Theorem**: if B_i is a partition and A is an arbitrary event, then

$$\begin{aligned}
 P(B_i|A) &= \frac{P(A \cap B_i)}{P(A)} \\
 &= \frac{P(A \cap B_i)}{\sum P(A|B_i)P(B_i)}
 \end{aligned}$$

In this context, $P(B_i)$ is referred to as a *prior* probability since it is given prior to any observations associated with the experiment. The result of Bayes Theorem, $P(B_i|A)$, is referred to as a *posterior* probability. It represents an updated assessment of the likelihood of event B_i after observing some partial information about the experiment, A .

There are many important applications of Bayes Theorem. One such application is automated fault diagnosis. Here the partition, $\{B_i\}$, would represent the potential causes of failure for a system. The prior probabilities, $P(B_i)$, could be obtained from historical records and/or expert opinion regarding the likelihoods of each possible cause of failure. The event A would represent a particular symptom or set of symptoms that was observed during a failure episode. Then the posterior probabilities, $P(B_i|A)$, would represent updated likelihoods for each cause of failure after observing that symptom, and a search for the actual cause of failure would first look at the cause with the highest posterior probability. If the most likely cause of failure is not the actual cause, then the search would look at the second highest posterior probability, etc.

Example. Suppose that a pharmaceutical company has developed a new screening test for a particular form of cancer that is very inexpensive to administer and is non-invasive. It is tested by applying it to a very large group that is known to have this form of cancer and to a very large group that is known to not have cancer. Suppose that these tests give the following results: 95% of those known to have this form of cancer receive a positive result from the screening test, and 97% of those known to not have cancer receive a negative result from the test. Finally, suppose that this screening test is designed to be given only to a target population that exhibits a set of symptoms that are possible indicators for this form

of cancer and that 10% of the target population have this form of cancer. Do you think this is a good test? Now answer this question after considering the following scenerio: a patient is identified by his doctor as being a member of the target population and is given this screening test. The test comes back positive. What then does the doctor tell the patient?

Example. Adult employees of a certain county in Texas are required periodically to take a drug test. It has been reported that the test they take has the following properties: 97% of those who have used illegal drugs within the previous week will get a positive test result and 94% of those who have not used illegal drugs within the past week will get a negative test result. Also, it has been reported that 1% of adults in this county use illegal drugs. Does this test give useful information about drug use by these employees?

Independence

We have seen that conditional probability represents a modification of the probability function to adjust for partial information that has been obtained about the outcome of the experiment. In some situations, however, this partial information results in no change in the probabilities. This would occur if $P(A|B) = P(A)$. That is, the conditional probability of A given B is the same as the original probability of A . If this occurs, we say that the events A and B are **independent**. Note that from the definition of conditional probability, this is equivalent to $P(A \cap B) = P(A)P(B)$. That is, if two events are independent, then the probability of the intersection of the events equals the product of the probabilities. This definition can be extended to include an arbitrary number of events. A collection of events, $\{A_i\}$, is said to be **independent** if the probability of the intersection of any subcollection of these events equals the product of the respective probabilities.

Example: redundancy improves reliability. Suppose a product must pass 5 different quality tests before it is approved for delivery. If the probability that a defective product passes a test is 0.1 for each of these tests, and if these tests are independent, what is the probability that a defective product will pass all 5 tests?

$$P(\text{Pass all tests}) = P\left(\bigcap_{i=1}^5 \{\text{Pass test } i\}\right) = \prod_{i=1}^5 P(\text{Pass test } i) = .1^5 = .00001.$$

Random Variables

For many of the experiments that we model, we are not interested in the outcomes themselves, but instead in some numerical attribute associated with the outcome. In the example above that modelled the games played by two gamblers, we were not interested in the particular outcome, $WWWWLLLLLL$, but instead all that was important was the number of wins, in this case 4, associated with the outcome. This numerical attribute, the number of wins, is an attribute possessed by each possible outcome of the experiment. We call such numerical attributes **random variables**.

Formally, a **random variable** is a function, $X : \Omega \rightarrow \mathfrak{R}$, that assigns a real number to each outcome of an experiment. In the case of the gambler's problem, if X denotes the

number of wins in 10 games, and if $\omega = \{WWWWLLLLLL\}$, then $X(\omega) = 4$. If all that we care to observe in an experiment is the value of X , then the only events we need to work with are events defined in terms of the random variable. For example,

$$\{\omega \in \Omega : X(\omega) \leq 5\}$$

is an event defined in terms of X , and its probability,

$$P(\omega \in \Omega : X(\omega) \leq 5)$$

is ordinarily expressed as $P(X \leq 5)$. Note that this notation suppresses the fact that X is really a function, not a number.

One requirement that we have regarding random variables is that we must be able to obtain the probability of events in which the random variable belongs to an interval of real numbers, along with all the sets one can obtain by finite unions, intersections, and complements of such events. The probabilities of events of this form can be obtained from the **distribution function** (d.f.) of the random variable, defined by

$$F(x) = P(X \leq x), \quad -\infty < x < \infty.$$

For example,

$$\begin{aligned} P(a < X \leq b) &= P(\{X \leq b\} \cap \{X \leq a\}^c) \\ &= P(X \leq b) - P(X \leq a) \\ &= F(b) - F(a), \quad -\infty < a < b < \infty. \end{aligned}$$

Another example:

$$P(X > a) = P(\{X \leq a\}^c) = 1 - P(X \leq a) = 1 - F(a).$$

Note: we make a notational distinction between the random variable and a value of the random variable by using upper case letters to denote random variables and lower case letters to denote values of a random variable.

The **sample space** of a random variable is the set of all possible values of the random variable. If the sample space is finite or countably infinite, we say that the random variable is **discrete**. If the distribution function is differentiable, we say that the random variable is **continuous**. If X is a discrete random variable, it is usually easier to obtain probabilities of events in terms of its **probability mass function** (pmf), defined by $p(x) = P(X = x)$. The correspondence between the d.f. of a random variable and its pmf is given by the following relationships:

$$F(x) = \sum_{a \leq x} p(a), \quad p(a) = F(a) - F(a-),$$

where $F(a-)$ denotes the limit from below of the d.f. This implies that the d.f. of a discrete random variable is a step-function. In the case of an integer-valued random variable, we have,

$$F(n) = \sum_{i \leq n} p(i), \quad p(n) = F(n) - F(n-1).$$

Likewise,

$$P(X \in A) = \sum_{n \in A} p(n).$$

The basic axioms of probability can be used to show that distribution functions of random variables satisfy the following properties:

1. $F(x)$ is a monotone non-decreasing function.
2. $\lim_{x \rightarrow -\infty} F(x) = 0.$
3. $\lim_{x \rightarrow \infty} F(x) = 1.$

Any function that satisfies these properties is the distribution function of some random variable.

Probability mass functions satisfy the properties,

1. $p(x) > 0$ for only a finite or countably infinite number of values of x , and is 0 for all other values of x .
2. $\sum_x p(x) = 1.$

A function that satisfies these properties is the p.m.f. of some random variable.

In the previous section, we obtained the probability of randomly selecting k females for a subcommittee of size 5 from a committee that has 40 males and 20 females. Let N denote the number of females selected for the subcommittee. Then the p.m.f. of this random variable is given by,

$$p(k) = P(N = k) = \frac{\binom{20}{k} \binom{40}{5-k}}{\binom{60}{5}}, \quad 0 \leq k \leq 5.$$

Note that $p(x) = 0$ for x not equal to one of the values, 0, 1, 2, 3, 4, 5.

Expectation of Discrete Random Variables

Suppose you are really bored one afternoon and decide to toss two coins a large number of times, say 10,000 times, recording the number of heads after each pair of coins is tossed. We can treat the tosses as binomial experiments with $n = 2$ and $p = 0.5$. Let X_i denote the number of heads obtained on the i^{th} toss. Then, we can expect after performing these experiments that around 2,500 of the X_i 's will be 0, around 5,000 of the X_i 's will be 1, and around 2,500 of the X_i 's will be 2. Now suppose we wish to find the average of the X_i 's, that

is, the average number of heads per experiment. Based on the number of times we expect to observe each of the possible values, 0,1,2, the average we can expect to see would be,

$$\begin{aligned}\bar{X} &= \frac{\sum X_i}{10000} \approx \frac{0 * 2500 + 1 * 5000 + 2 * 2500}{10000} \\ &= 0 \left(\frac{2500}{10000} \right) + 1 \left(\frac{5000}{10000} \right) + 2 \left(\frac{2500}{10000} \right) \\ &= 0(.25) + 1(.50) + 2(.25) = 1.\end{aligned}$$

This quantity is referred to as the **expected value** of the random variable X , the number of heads when two coins are tossed. Note that the expected value represents the average we would expect to observe if an experiment is repeated a large number of times, just as the probability of an event is the proportion of times we would expect the event to occur if we repeated the experiment a large number of times. It is no coincidence that the expected value in this case coincides with

$$\sum_x xp(X = x) = \sum_x xp(x),$$

where $p(x)$ is the p.m.f of X .

Definition. The **expected value** of a discrete random variable with p.m.f. p is defined to be

$$E(X) = \sum_x xp(x).$$

The expected value is also commonly referred to as the **mean** of the random variable. Properties of summation lead to the properties of expected values listed below. In this list, X, Y represent random variables and a, b, c represent constants.

1. If $Y = a + bX$, then

$$E(Y) = a + bE(X).$$

2. More generally, if

$$Y = a_0 + \sum_{i=1}^n a_i X_i,$$

then

$$E(Y) = a_0 + \sum_{i=1}^n a_i E(X_i).$$

3. Let $I_A(x)$ denote the indicator function for the set A . That is, $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ if $x \notin A$. Then

$$E[I_A(X)] = P(X \in A).$$

4. If X and Y are independent, then

$$E[XY] = E[X]E[Y].$$

The expected value of a r.v. describes the center of the possible values of the r.v. Note that it is a weighted average of those values in which the weights are given by the p.m.f. A related quantity, called the *variance*, describes the variability of those values about this center. Let X be a discrete r.v. with expected value $E(X) = \mu$. Then the *variance* of X is defined by:

$$\begin{aligned} \text{Var}(X) &= E[(X - \mu)^2] \\ &= \sum_x (x - \mu)^2 p(x). \end{aligned}$$

The variance of a r.v. is commonly represented by σ^2 . The square is because the unit of measure of the variance is the square of the unit of measure of the r.v.

Note that the properties of expectation show:

$$\begin{aligned} \text{Var}(X) &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E(X) + \mu^2 \\ &= E[X^2] - \mu^2. \end{aligned}$$

This is usually the easiest way to obtain the variance.

The square root of the variance, called the *standard deviation* and denoted by σ , represents a measure of distance between a r.v. and its mean. In particular, since $E[cX] = cE[X]$, then

$$\text{Var}(cX) = E[c^2 X^2] - c^2 \mu^2 = c^2 \text{Var}(X),$$

and so,

$$SD(cX) = |c| \sigma_X.$$

Also, $\sigma_X = 0$ if and only if $X = \mu$ with probability 1. If the s.d. of a r.v. is small, then the r.v. tends to be close to its mean, but if the s.d. is large, then the r.v. tends to be farther from its mean. This is made more precise by Chebychev's inequality.

Chebychev's inequality: if X is a random variable with mean μ and variance σ^2 , then for any positive constant ϵ ,

$$P(|X - \mu| > \epsilon) \leq \left(\frac{\sigma}{\epsilon}\right)^2.$$

In particular,

$$\begin{aligned} P(|X - \mu| > 2\sigma) &\leq \frac{1}{4}, \\ P(|X - \mu| > 3\sigma) &\leq \frac{1}{9}. \end{aligned}$$

Example. Suppose that a company receives a shipment of 50 new PC's, 4 of which are defective. Suppose that your cost center is given 5 of these PC's, assumed to be randomly selected from the shipment. Find the expected value and s.d. of the number of defective PC's received by the cost center.

solution. Let N denote the number of defective PC's. We must first obtain the p.m.f. of N . Note that the sample space of N is 0, 1, 2, 3, 4.

$$p(0) = \frac{\binom{46}{5}}{\binom{50}{5}} = 0.64696$$

$$p(1) = \frac{\binom{4}{1}\binom{46}{4}}{\binom{50}{5}} = 0.30808$$

$$p(2) = \frac{\binom{4}{2}\binom{46}{3}}{\binom{50}{5}} = 0.04299$$

$$p(3) = \frac{\binom{4}{3}\binom{46}{2}}{\binom{50}{5}} = 0.00195$$

$$p(4) = \frac{\binom{46}{1}}{\binom{50}{5}} = 0.00002$$

Note the these probabilities sum to 1. The expected value and variance can be obtained most easily from the following table.

x	$p(x)$	$xp(x)$	x^2	$x^2p(x)$
0	0.64696	0	0	0
1	0.30808	0.30808	1	0.30808
2	0.04299	0.08598	4	0.17196
3	0.00195	0.00585	9	0.01755
4	0.00002	0.00008	16	0.00032
sum	1	0.4		0.49791

So, $E[N] = 0.4$, $E[N^2] = 0.49791$, $Var(N) = 0.49791 - 0.4^2 = 0.33791$, $SD(X) = \sqrt{0.33791} = 0.5813$.

Special Discrete Distributions

Some random variables and distribution functions occur frequently enough in applications that we will study them individually.

Bernoulli and Related Distributions

The simplest experiment we can model is one in which there are just two possible outcomes. By convention, one of these is labelled S or *Success* and the other is labelled F or *Failure*.

Such experiments are referred to as Bernoulli trials. The corresponding Bernoulli random variable assigns the value 1 to the outcome S and 0 to F . If p denotes the probability of observing the outcome labelled S , then the p.m.f. of a Bernoulli random variable N is given by

$$p(x) = \begin{cases} 1 - p, & x = 0, \\ p, & x = 1, \\ 0, & \text{otherwise.} \end{cases}$$

The corresponding distribution function is,

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - p, & 0 \leq x < 1, \\ 1, & x \geq 1. \end{cases}$$

Note that this distribution is characterized by the single parameter, p .

The expected value and variance of a Bernoulli r.v. can be obtained easily:

$$E(X) = (0)(1 - p) + (1)p = p,$$

$$E(X^2) = (0^2)(1 - p) + (1^2)p = p,$$

and so,

$$Var(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1 - p).$$

Note that the s.d. of the Bernoulli is just

$$SD(X) = \sqrt{p(1 - p)}.$$

A direct extension of this experiment is an experiment in which a series of n independent Bernoulli trials are performed, each with the same probability p of success. Let X_i denote the i^{th} Bernoulli random variable, and let N denote the total number of successes among the n trials. Then,

$$N = \sum_{i=1}^n X_i.$$

Note that the gambler's problem described in the previous section is an example of this type of experiment in which S represents the event that the first gambler wins a game and each game is a Bernoulli trial. Using the same arguments here as we did for the gambler's problem, we can see that the p.m.f. of N is given by,

$$p(k) = P(N = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad 0 \leq k \leq n.$$

As noted earlier, the p.m.f of this random variable involves terms of the binomial series, and so this random variable is referred to as the binomial random variable and its distribution is

referred to as the binomial distribution. This distribution is characterized by two parameters, the number of trials, n and the success probability, p .

The binomial distribution can be used to model a sampling experiment in which a sample of n objects is to be randomly selected with replacement from a population that consists of two types of objects. Let p denote the proportion of the population that are type 1 objects. The sampling can be viewed as a sequence of Bernoulli trials, with S denoting the event that the first type is selected on a trial. Since the sampling is done with replacement, the second selection trial is identical to the first and the outcome of the second trial does not depend on the outcome of the first trial, since the object selected on the first trial is returned to the population. Hence, the number of type 1 objects selected in such an experiment would have a binomial distribution with n trials and success probability p .

If the sampling is performed without replacement, then the trials will no longer be independent and the probability of success for each trial will no longer be the same. However, if the sample size is small compared to the population size, then the binomial distribution is a reasonable approximation to the actual probabilities.

The expected value and variance can be derived directly from the Bernoulli distribution.

$$E(N) = E\left[\sum_{k=1}^n X_k\right] = \sum_{k=1}^n E(X_k) = np.$$

Since N can be expressed as a sum of independent Bernoulli r.v.'s, then the variance of this sum is the sum of the individual variances,

$$Var(N) = \sum_{k=1}^n Var(X_k) = np(1-p),$$

and so,

$$SD(N) = \sqrt{np(1-p)}.$$

Example. Suppose that we are interested in the proportion of defectives within a large population, and so randomly select a sample of size n from this population. If n is small compared to the population size, then we can use the binomial distribution as an approximation to the distribution of the number of defectives that will be found in the sample. Let N denote this number and note that

$$\hat{p}_n = \frac{N}{n}$$

represents the proportion of defectives in the sample. We would expect that the sample proportion should be fairly close to the population proportion p . This can be expressed probabilistically by applying Chebychev's inequality. Let c denote an arbitrarily small positive real number. Then,

$$\begin{aligned} P(|\hat{p}_n - p| > c) &= P(|N - np| > cn) \\ &\leq \frac{np(1-p)}{c^2n^2} \\ &= \frac{p(1-p)}{nc^2}. \end{aligned}$$

This means that if n is large, then the probability that the sample proportion \hat{p}_n is more than c from the population proportion p is very small. In fact, as $n \rightarrow \infty$, then this probability goes to 0.

A different extension of Bernoulli trials is to continue performing the trials until the first success is observed. Let G denote the number of trials required to obtain the first success. Under the assumption that the trials are independent with the same probability of success, we have,

$$P(G = k) = (1 - p)^{k-1}p, \quad k \geq 1.$$

This follows from the fact that the first success occurs on trial k if and only if the first $k - 1$ trials are failures and trial k is a success. Since the p.m.f. of this random variable involves terms of the geometric series, this random variable is referred to as the geometric random variable and its distribution is referred to as the geometric distribution.

A related random variable that is sometimes more convenient to work with is Y , defined to be the number of failures observed before the first success occurs. It can be seen that Y is related to G by $Y = G - 1$. Hence, its p.m.f. is given by,

$$p_Y(k) = P(Y = k) = (1 - p)^k p, \quad k \geq 0.$$

The expected value and variance of the geometric distribution are:

$$\begin{aligned} E[G] &= \frac{1}{p}, \\ \text{Var}(G) &= \frac{q}{p^2} \\ E[Y] &= E[G - 1] = \frac{q}{p} \\ \text{Var}(Y) &= \text{Var}(G) = \frac{q}{p^2}. \end{aligned}$$

Continuous Random Variables

Continuous random variables are variables that take values that could be any real number within some interval. One common example of such variables is *time*, for example, the time to failure of a system or the time to complete some task. Other examples include physical measurements such as length or diameter. As will be seen, continuous random variables also can be used to approximate discrete random variables.

To develop probability models for continuous r.v.'s, it is necessary to make one important restriction: we only consider events associated with these r.v.'s that are defined in terms of intervals of real numbers, including intersections and unions of intervals. Probability models are constructed by representing the probability that a r.v. is contained within an interval as the area under a curve over that interval. That curve is called the *density function* of the r.v. To satisfy the laws of probability, density functions must satisfy the following two conditions:

1. $f(t) \geq 0, \forall t,$
2. $\int_{-\infty}^{\infty} f(t)dt = 1.$

The second condition corresponds to the requirement that the probability of the entire sample space must be 1. Any function that satisfies these two conditions is the density function of some r.v.

The probability that the r.v. is contained within an interval is then

$$P(a < X \leq b) = \int_a^b f(t)dt.$$

Note that in the case of continuous r.v.'s,

$$P(a < X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a \leq X \leq b),$$

since the area under a curve at a point is 0. The distribution function of a continuous r.v. is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt.$$

Note that the Fundamental Theorem of Calculus implies that

$$f(x) = \frac{d}{dx}F(x).$$

Also note that the value of a density function is not a probability; nor is a density necessarily bounded by 1. It can be thought of as the concentration of likelihood at a point.

The expected value of a continuous r.v. is defined analogously to the expected value of a discrete r.v. with the p.m.f. replaced by the density function and the sum replaced by an integral:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

Also, the variance of a continuous r.v. is defined by

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx,$$

where $\mu = E(X)$. Note that the additive property of integrals gives

$$\begin{aligned} Var(X) &= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f(x)dx \\ &= \int_{-\infty}^{\infty} x^2 f(x)dx - 2\mu \int_{-\infty}^{\infty} x f(x)dx + \mu^2 \\ &= \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 \\ &= E(X^2) - \mu^2, \end{aligned}$$

where $\mu = E(X)$.

To construct probability models for continuous r.v.'s, it is only necessary to find a density function that models appropriately the concentration of likelihood.

Normal Distribution

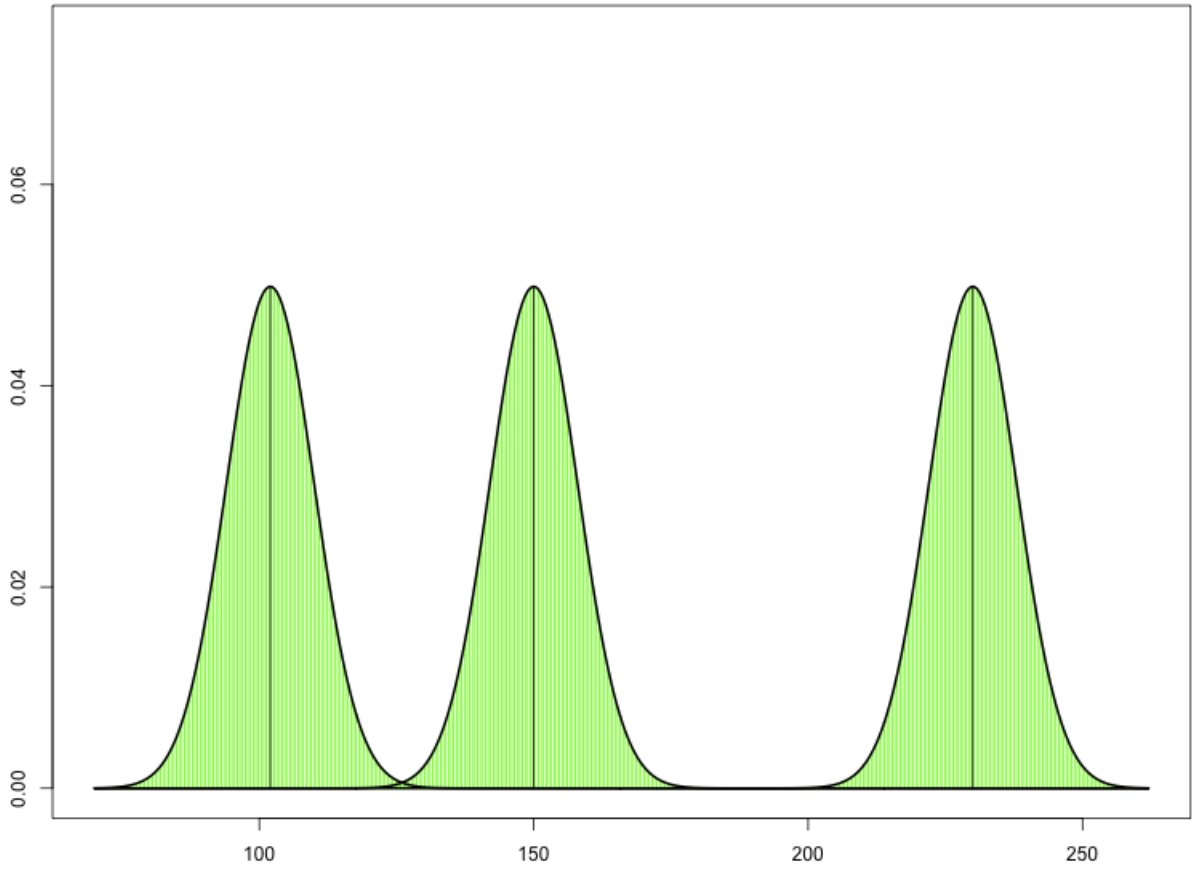
The **normal distribution**, also known as the *Bell Curve*, has been used (and abused) as a model for a wide variety of phenomena to the point that some have the impression that any data that does not fit this model is in some way *abnormal*. That is not the case. The name *normal distribution* comes from the title of the paper Carl Friedrich Gauss wrote that first described the mathematical properties of the bell curve, “On the Normal Distribution of Errors”. For this reason, the distribution is sometimes referred to as the **gaussian distribution**. Perhaps that name would be less misleading. The main importance of this model comes from the central role it plays in the behavior of many statistics that are derived from large samples.

The normal distribution represents a family of distribution functions, parametrized by the mean and standard deviation, denoted by $N(\mu, \sigma)$. The density function for this distribution is

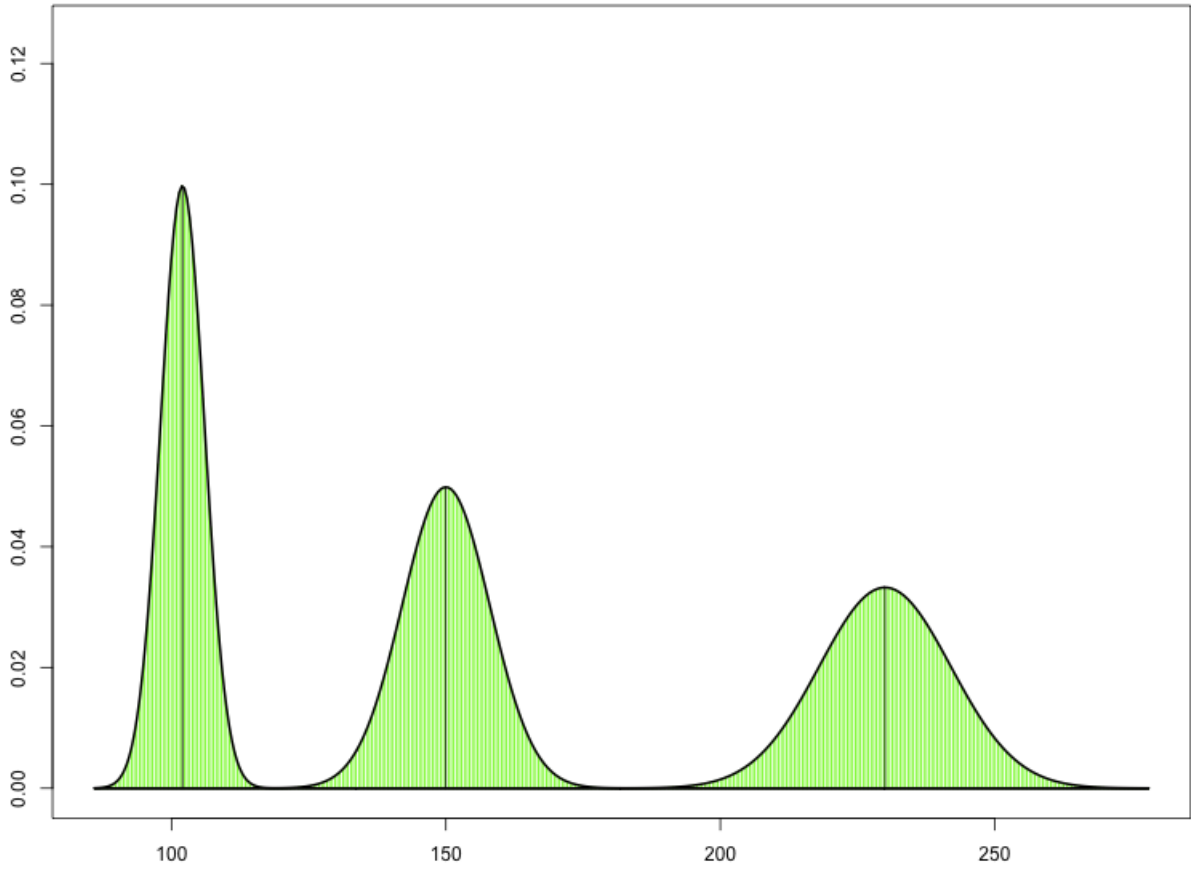
$$f(x; \mu, \sigma) = (2\pi\sigma^2)^{-1/2} \exp\{-(x - \mu)^2/(2\sigma^2)\}.$$

The mean is referred to as a location parameter since it determines the location of the peak of the curve. The standard deviation is referred to as a scale parameter since it determines how spread out or concentrated the curve is. The plots below illustrate these properties. In the first plot, the means differ but the standard deviations are all the same. In the second plot, both the means and the standard deviations differ.

Normal distributions: $\mu = 102, 150, 230$; $\sigma = 8$



Normal distributions: $\mu = 102, 150, 230$; $\sigma = 4, 8, 12$



Probability that a continuous random variable is contained within an interval is modeled by the area under the curve corresponding to the interval. Suppose for example we have a random variable that has a $N(50, 5)$ distribution and we are interested in the probability that this r.v. takes a value between 45 and 60. The problem now is to determine this area. Unfortunately (or perhaps fortunately from the point of view of students) the normal density function does not have an explicit integral. This implies that we must either use a set of tabulated values to obtain areas under the curve or use a computer routine to determine the areas. One property satisfied by the family of normal distributions is *closure under linear transformations*. That is, if $X \sim N(\mu, \sigma)$, and if $Y = a + bX$, then $Y \sim N(a + b\mu, |b|\sigma)$. We can make use of this property by noting that

$$Z = \frac{X - \mu}{\sigma} = -\frac{\mu}{\sigma} + \frac{1}{\sigma}X$$

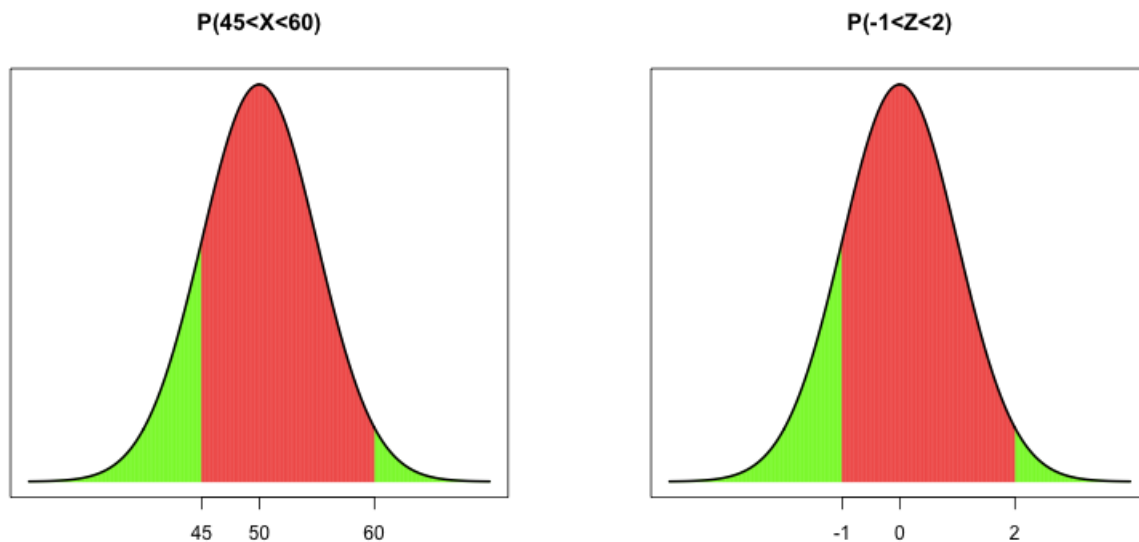
has a $N(0, 1)$ distribution. This distribution is referred to as the **standard normal distribution**, and the value of Z corresponding to X is referred to as the **standardized score** or **Z-score** for X . This property implies that the probability of any interval can be transformed into a probability involving the standard normal distribution. The interpretation of the *Z-score* can be seen by expressing X in terms of Z ,

$$X = \mu + Z\sigma.$$

This shows that the *z-score* represents the number of standard deviations X is from its mean.

For example, if $X \sim N(50, 5)$, then

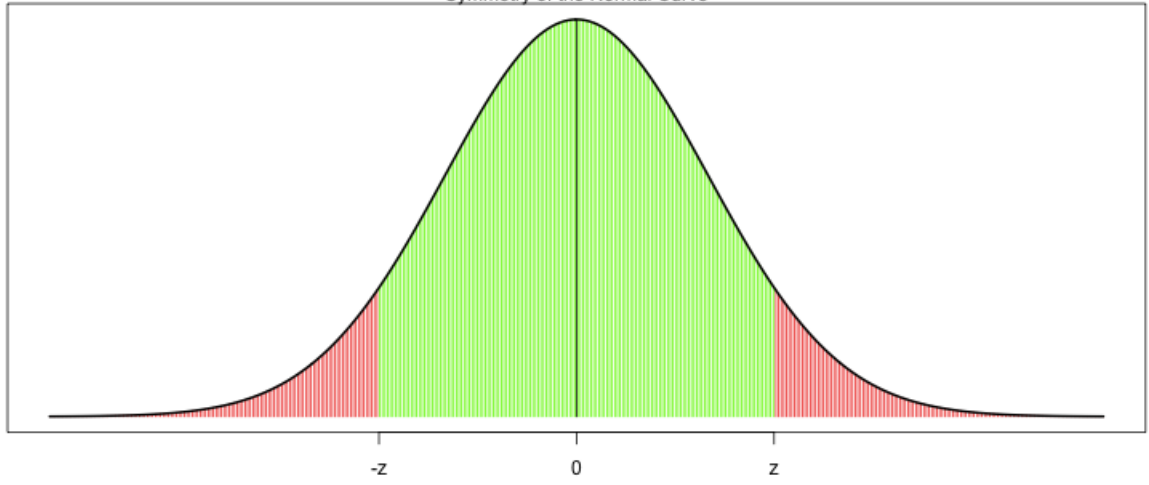
$$\begin{aligned} P(45 < X < 60) &= P\left(\frac{45 - 50}{5} < \frac{X - 50}{5} < \frac{60 - 50}{5}\right) \\ &= P(-1 < Z < 2). \end{aligned}$$



As can be seen by comparing these two plots, the areas for $P(45 < X < 60)$ and $P(-1 < Z < 2)$ are the same. Therefore, it is only necessary to tabulate areas for the standard normal distribution. The textbook contains such a table on page 789. This table gives areas under the standard normal curve below z for $z > 0$. This table requires an additional property of normal distributions called symmetry:

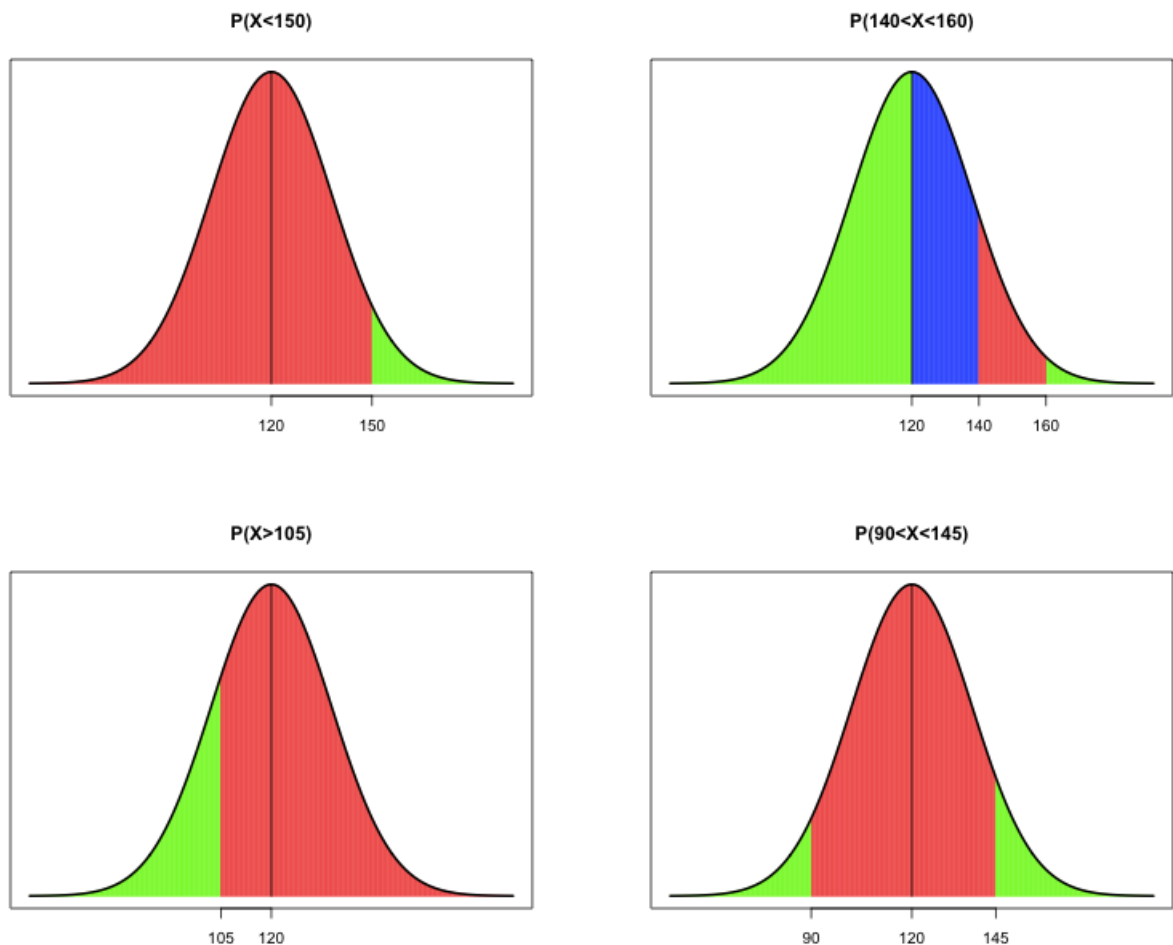
$$P(Z < -z) = P(Z > z), \quad P(0 < Z < z) = P(-z < Z < 0).$$

Symmetry of the Normal Curve



Example. Suppose a questionnaire designed to assess employee satisfaction with working conditions is given to the employees of a large corporation, and that the scores on this questionnaire are approximately normally distributed with mean 120 and standard deviation 18.

- a) Find the proportion of employees who scored below 150.
 - b) Find the proportion of employees who scored between 140 and 160.
 - c) What proportion scored above 105?
 - d) What proportion scored between 90 and 145?
- These areas are represented in the plots given below.
- e) 15% of employees scored below what value?



Solutions

a) First transform to $N(0, 1)$.

$$z = \frac{150 - 120}{18} = 1.67,$$

$$P(X < 150) = P(Z < 1.67).$$

From the table on the inside back cover of the text, the area below 1.67 is 0.9525. Therefore,

$$P(X < 150) = P(Z < 1.67) = 0.9525.$$

b) Transform to $N(0, 1)$.

$$z_1 = \frac{140 - 120}{18} = 1.11$$

$$z_2 = \frac{160 - 120}{18} = 2.22.$$

In this case we must subtract the area below 1.11 from the area below 2.22. From the table these areas are, respectively, .8665 and .9868. This gives

$$P(140 < X < 160) = P(1.11 < Z < 2.22) = 0.9868 - 0.8665 = 0.1203.$$

c) Transform to $N(0, 1)$.

$$z = \frac{105 - 120}{18} = -0.83.$$

The symmetry property of the normal distribution implies that the area above -0.83 is the same as the area below 0.83, which we get from the table.

$$P(X > 105) = P(Z > -0.83) = P(Z < 0.83) = 0.7967.$$

d) Transform to $N(0, 1)$.

$$z_1 = \frac{90 - 120}{18} = -1.67$$

$$z_2 = \frac{145 - 120}{18} = 1.39$$

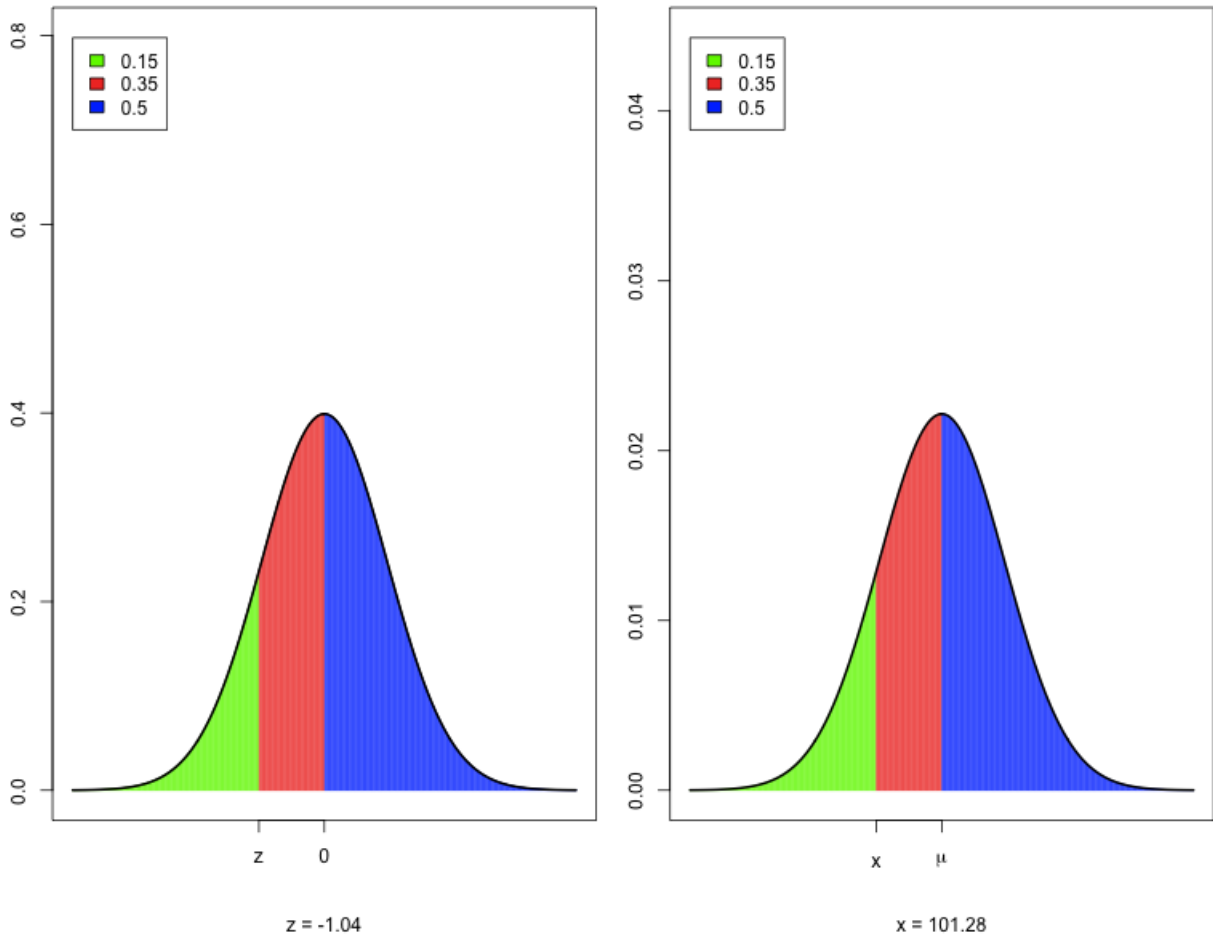
The area we require is the difference between the area below 1.39 and the area below -1.67. By symmetry, the area below -1.67 is the same as the area above 1.67.

$$\begin{aligned} P(90 < X < 145) &= P(Z < 1.39) - P(Z < -1.67) \\ &= 0.9177 - [1 - P(Z < 1.67)] \\ &= 0.9177 - [1 - 0.9525] \\ &= 0.8702. \end{aligned}$$

e) This problem is different than the others because we are given an area and must use this to determine the appropriate value. The first step is to determine on which side of the mean the required value is located. This is determined by two quantities: whether the area is less than 0.5 or greater than 0.5, and the direction relative to the required value occupied by the specified area. In this case, the area (**15%=0.15**) is less than 0.5 and the direction is specified by *scored below what value*. These imply that the required value must be less than the mean. A picture of this area is given below. To answer this question, we first answer the corresponding question for the standard normal distribution. What *z-value* has an area of **0.15 below** it? This *z-value* must be negative since the area is less than **0.15** and the direction is **below** (or to the left of) the required value. Since the table gives areas below **z**, the area we must find in the table is $1 - 0.15 = 0.85$. The closest area in the table to 0.85 is 0.8508 which corresponds to a z-score of 1.04. Since the z-score for this problem is negative, then the answer to this question for the standard normal distribution is $z = -1.04$. Finally, we must convert this z-score to the x-value,

$$x = \mu + z\sigma = 120 + (-1.04)(18) = 101.28.$$

If you check this answer by finding the area below 101.28, you will see that the steps we just followed are the same steps we used to find areas but applied in reverse order. Also note that the value of 101.28 represents the 15th percentile of this normal distribution. Other percentiles can be obtained similarly.



Since z-scores represent the number of standard deviations from the mean, and since they are directly associated with percentiles, they can be used to determine the relative standing of an observation from a normally distributed population. In particular, consider the following three intervals: $\mu \pm \sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$. After converting these intervals to z-scores, they become, respectively, $(-1,1)$, $(-2,2)$, and $(-3,3)$. Because of the symmetry property, the probabilities for these intervals are,

$$\begin{aligned}
 P(\mu - \sigma < X < \mu + \sigma) &= P(-1 < Z < 1) = 2P(0 < Z < 1) = 2(.3413) = .6826 \\
 P(\mu - 2\sigma < X < \mu + 2\sigma) &= P(-2 < Z < 2) = 2P(0 < Z < 2) = 2(.4772) = .9544 \\
 P(\mu - 3\sigma < X < \mu + 3\sigma) &= P(-3 < Z < 3) = 2P(0 < Z < 3) = 2(.4987) = .9974
 \end{aligned}$$

This is the basis for the **empirical rule**: if a set of data has a histogram that is approximately bell-shaped, then approximately 68% of the measurements are within 1 standard deviation of the mean, approximately 95% are within 2 standard deviations of the mean, and essentially

all (makes more sense than approximately 99.74%) are within 3 standard deviations of the mean.

Suppose that in the previous example an employee scored 82 on the employee satisfaction survey. The z-score for 82 is $(82-120)/18 = -2.11$. So this score is more than 2 standard deviations below the mean. Since 95% of the scores are within 2 standard deviations of the mean, this is a relatively low score. We could be more specific by determining the percentile rank for this score. From the table of normal curve areas, the area below 2.11 is 0.9826, so the area below $z = -2.11$ is $1 - 0.9826 = 0.0174$. That is, only 1.74% of those who took this questionnaire scored this low or lower.

Large Sample Approximations

The main importance of the normal distribution is associated with the **Central Limit Theorem**. This theorem was originally derived as a large sample approximation for the binomial distribution when n is large and p is not extreme. In this case we may approximate the binomial distribution function by the normal distribution with mean np and standard deviation $\sqrt{np(1-p)}$.

Suppose for example that in a very large population of voters, 48% favor Candidate A for president, and that a sample of 500 is randomly selected from this population. What is the probability that more than 250 in the sample will favor Candidate A? We can model the number in the sample who favor Candidate A with a binomial distribution with $n=500$ and $p=0.48$. Since n is large, we can approximate this distribution with a normal distribution with mean $\mu = 500(.48) = 240$ and standard deviation $\sigma = \sqrt{500(.48)(.52)} = 11.2$. Since the binomial is a discrete distribution, we can improve this approximation slightly by extending the interval of values whose probability we wish obtain by 0.5 at each end of the interval. For example, if we want to find $P(N = 230)$, then we approximate it by $P(229.5 < X < 230.5)$, where X has the appropriate approximate normal distribution. Similarly,

$$\begin{aligned}
 P(N < a) &\approx P(X < a - .5) \\
 P(N \leq a) &\approx P(X < a + .5) \\
 P(N > a) &\approx P(X > a + .5) \\
 P(N \geq a) &\approx P(X > a - .5) \\
 P(a < N < b) &\approx P(a + .5 < X < b - .5) \\
 P(a \leq N \leq b) &\approx P(a - .5 < X < b + .5)
 \end{aligned}$$

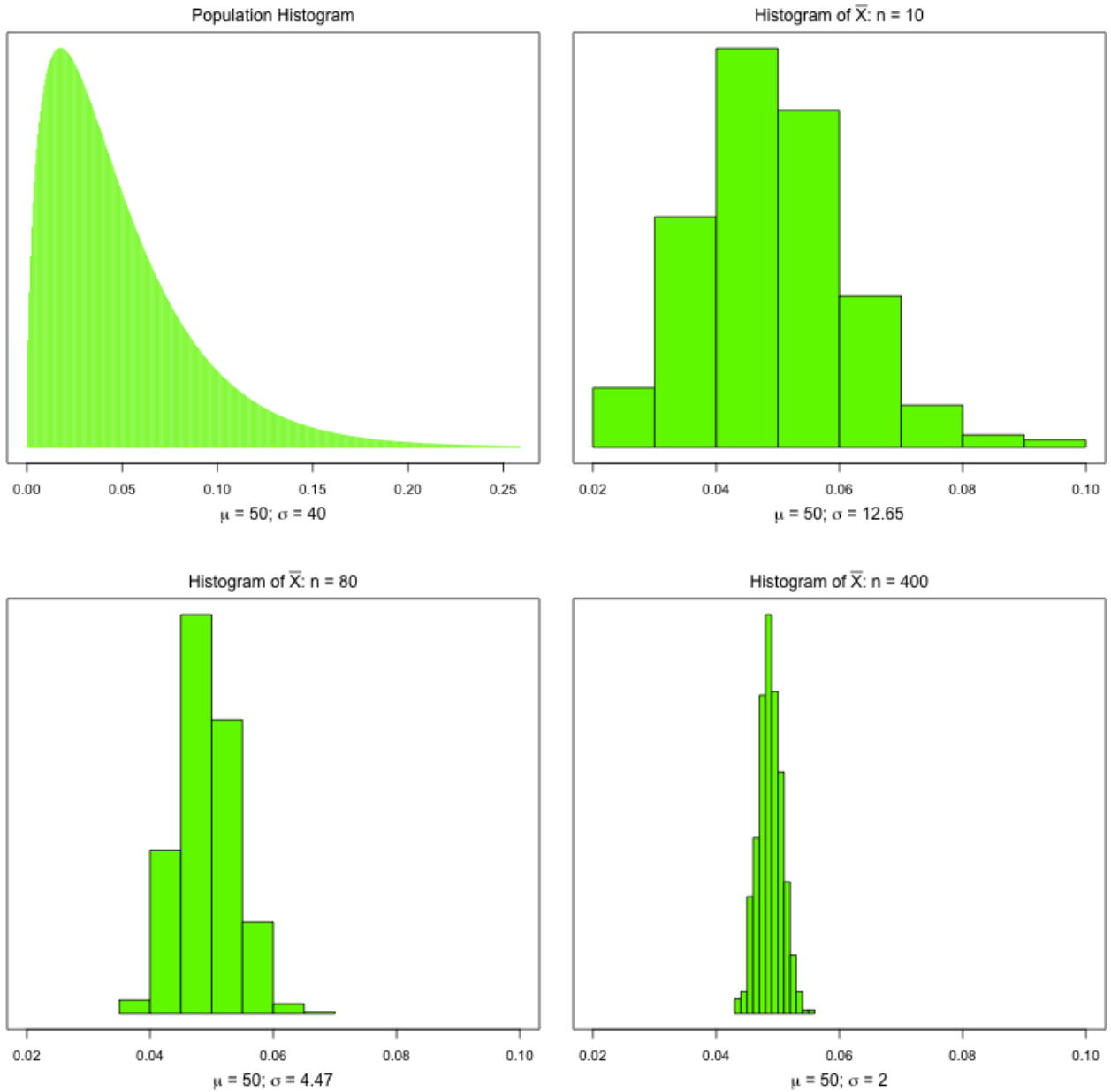
Therefore, from the table of areas under the normal curve, we obtain

$$\begin{aligned}
 P(N > 250) &\approx P(X > 250.5) \\
 &= P(Z > (250.5 - 240)/11.2) \\
 &= P(Z > 0.94) \\
 &= 1 - 0.8264 = 0.1736.
 \end{aligned}$$

Note that we could also express this event in terms of the sample proportion who favor Candidate A. Let $\hat{p} = N/500$ denote the sample proportion. Then the probability we obtained above could be expressed as $P(\hat{p} > 0.5)$. Since \hat{p} is a linear function of N , then we can use the normal distribution with mean $\mu = 240/500 = 0.48$ and standard deviation $\sigma = 11.2/500 = 0.022$ to approximate the distribution of \hat{p} . Note that the standard deviation can be obtained directly as $\sigma = \sqrt{p(1-p)/n} = \sqrt{(.48)(.52)/500} = 0.022$.

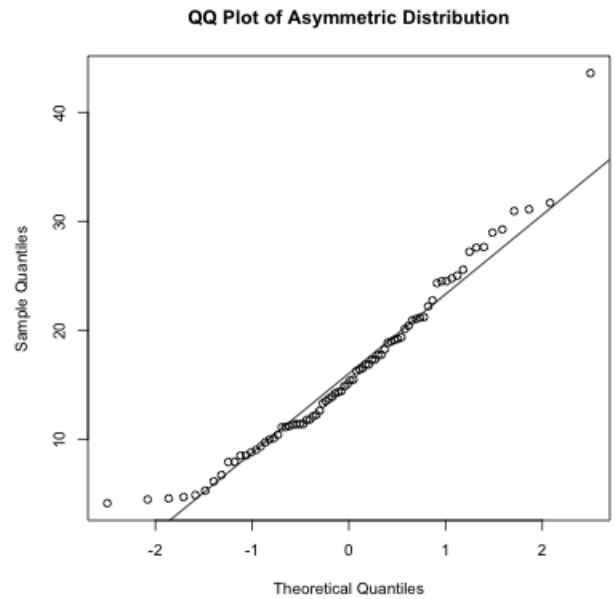
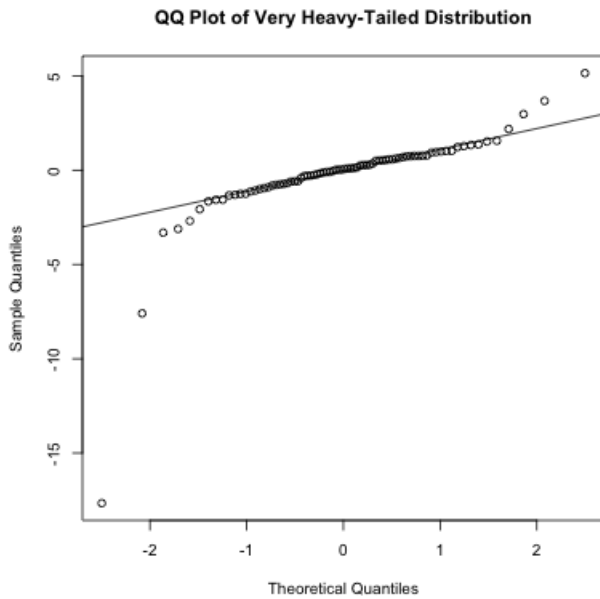
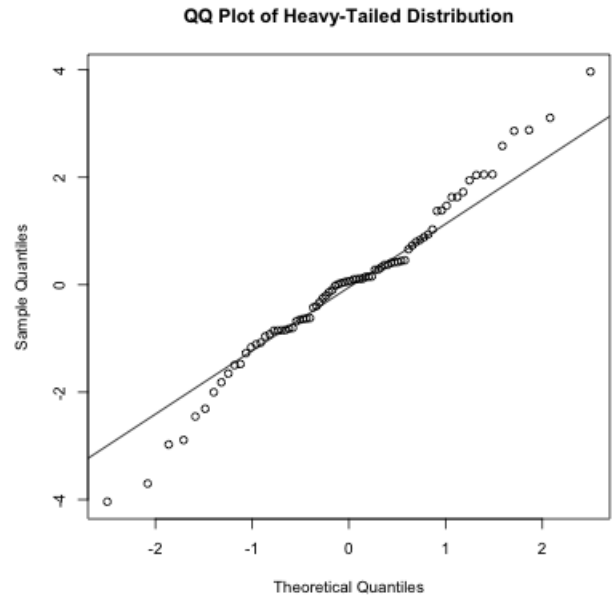
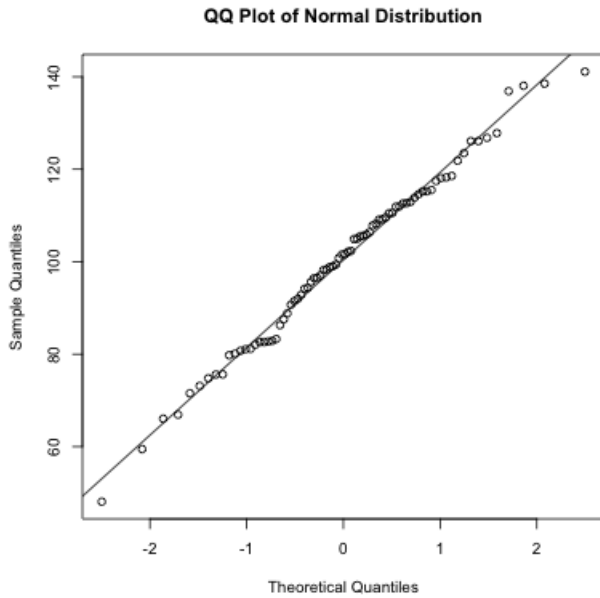
The **Central Limit Theorem** extends this result to a sampling situation in which a sample of size n is randomly selected from a very large population with mean μ and standard deviation σ . Let \bar{X} denote the mean of this sample. We can treat the sample mean as a random variable that is the numerical value associated with the particular sample we obtain when we perform the sampling experiment. The *Central Limit Theorem* states that the distribution of this random variable is approximately a normal distribution with mean μ and standard deviation σ/\sqrt{n} . Suppose we looked at every possible sample of size n that could be obtained from the population, and we computed the sample mean for each of these samples. What the **CLT** implies is that the histogram of all these sample means would be approximately a normal curve with mean μ and standard deviation σ/\sqrt{n} . The following plots illustrate this.

Histogram of \bar{X} based on 500 samples



Note that there is less asymmetry in the histogram of \bar{X} with $n = 10$ than in the population histogram, but some asymmetry still remains. However, that asymmetry is not present in the histograms corresponding to the larger sample sizes. Note also that the variability decreases with increasing sample size. This theorem holds for any distribution, but the more *non-normal* the distribution, the larger n must be for the distribution of \bar{X} to be close to the normal distribution. However, if the population distribution is itself a normal distribution, then the Central Limit Theorem holds for all $n \geq 1$.

One remaining question that will also be applicable to methods discussed later is the problem of determining how far a data set is from normality. This is accomplished most commonly by a *Quantile-Quantile* plot. Let n denote the sample size and let $y = ((1 : n) - .5)/n$. Then y represents the quantiles of the ordered values of the data. That is, $y[i]$ represents, up to a correction factor, the proportion of the sample that is at or below the i^{th} ordered value of the sample. Now let $x[i] = z_{y[i]}$. Then $x[i]$ represents the z-score such that the area below it equals the proportion of the sample that is at or below the data value corresponding to the i^{th} ordered value. If the data has a normal distribution, then these points should fall on a line with slope equal to the s.d. and intercept equal to the mean. The following plots show quantile-quantile plots for four distributions: normal, heavy-tailed, very heavy-tailed, and asymmetric.



Estimation

Many important statistical problems can be expressed as the problem of determining some characteristic of a population when it is not possible or feasible to measure every individual in the population. For example, political candidates may wish to determine the proportion of voters in a state who intend to vote for them; an advertising agency may wish to determine the proportion of a target population who react favorably to an ad campaign; a manufacturer

may wish to determine the mean cost per unit associated with warranty costs of a product. Since it is not possible or feasible to contact every individual in the respective populations, the only reasonable alternative is to select in some way a sample from the population and use the information contained within the sample to **estimate** the population characteristic of interest.

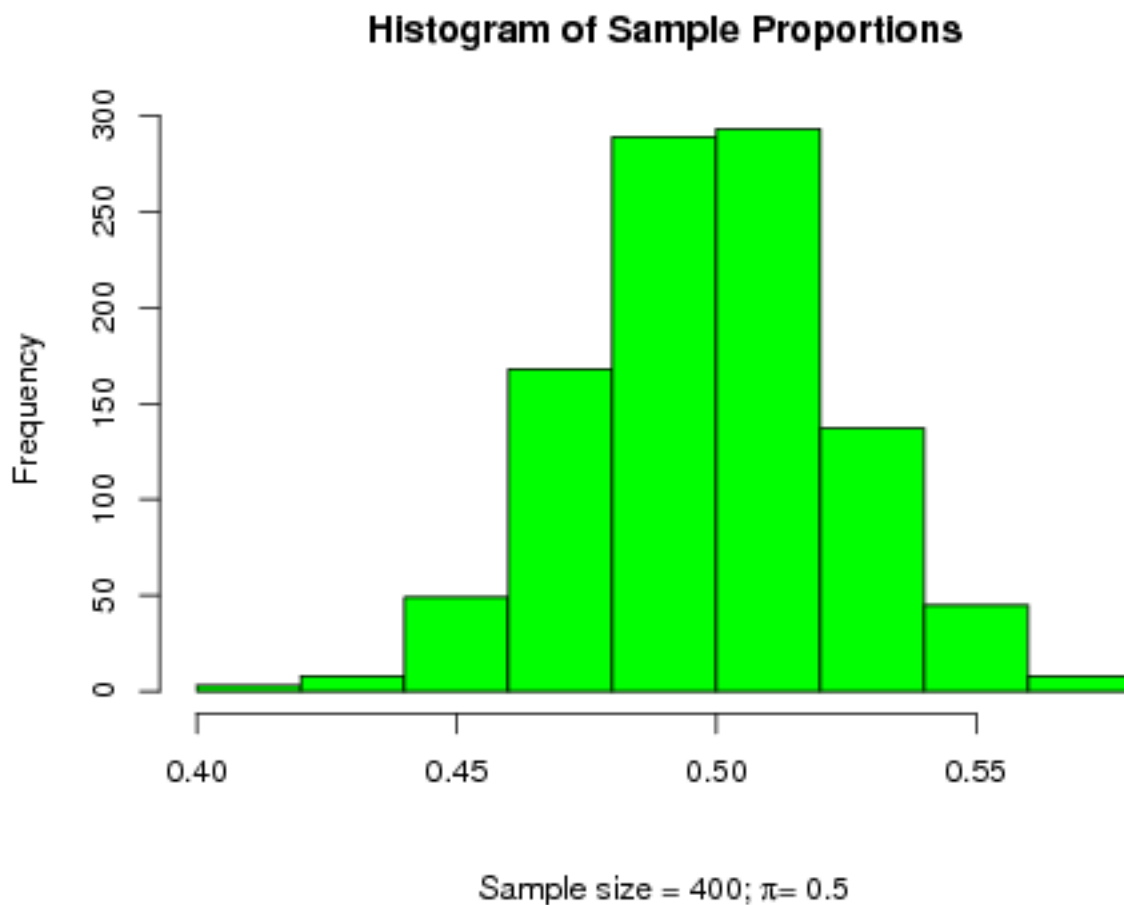
At first thought, it would seem that what should be done here is to select a *representative* sample from the population, since such a sample would mirror the properties of the population. Suppose, for example, that we would like to determine the proportion of voters in a state who intend to vote for a particular candidate for governor. Let π denote this proportion. A representative sample selected from this population should have a sample proportion that is *close* to π . The problem though is how to select such a sample. In fact, it is not possible to do this, for even if the proportion in the sample were close to π , we would not know it because we don't know the value of π .

Furthermore, an estimate derived from a sample has no value unless we can make some statement about its accuracy. Suppose that \hat{p} is the proportion in the sample that favor that candidate. Then the error of prediction would be $\epsilon = \hat{p} - \pi$. Obviously we cannot make an exact statement about this error since we do not know π . However, if the sample is selected randomly so that each individual in the population has the same chance of being selected, then it is possible to make a probability statement about the estimation error. **Random sampling is the only type of sampling with which we can make reasonable statements about the prediction error.**

Large Sample Estimation of a Population Proportion

Suppose we randomly select n individuals from the population of voters and let N denote the proportion in the sample who favor a particular candidate. Then $\hat{p} = N/n$ is our estimate of π . The value of this estimate depends on the individuals who are selected for the sample. To understand how we can make use of this fact to make a statement about estimation error, consider the following *thought experiment* (an experiment that we don't actually perform, but can think about). Suppose we select every possible sample of size n from the population and for each sample we obtain the sample proportion who favor this candidate. These estimates will vary from 0 to 1 and the actual sampling experiment we perform, selecting a random sample of size n and obtaining its sample proportion, is equivalent to randomly selecting one proportion from the population of proportions obtained from all possible samples of size n . Although we could not perform this experiment in reality, we can perform it mathematically. If we can determine the distribution of the population of all possible sample proportions, then we can use this distribution to make a **probability** statement about the estimation error. The Central Limit theorem states that if n is large, then the distribution of N is approximately a normal distribution with mean $n\pi$ and variance $n\pi(1 - \pi)$. Therefore, the distribution of $\hat{p} = N/n$ has approximately a normal distribution with mean π and variance $\pi(1 - \pi)/n$ (see the plot below). This distribution is called the **sampling distribution** of \hat{p} . One of the properties of normal curves is that approximately 95% of a normally

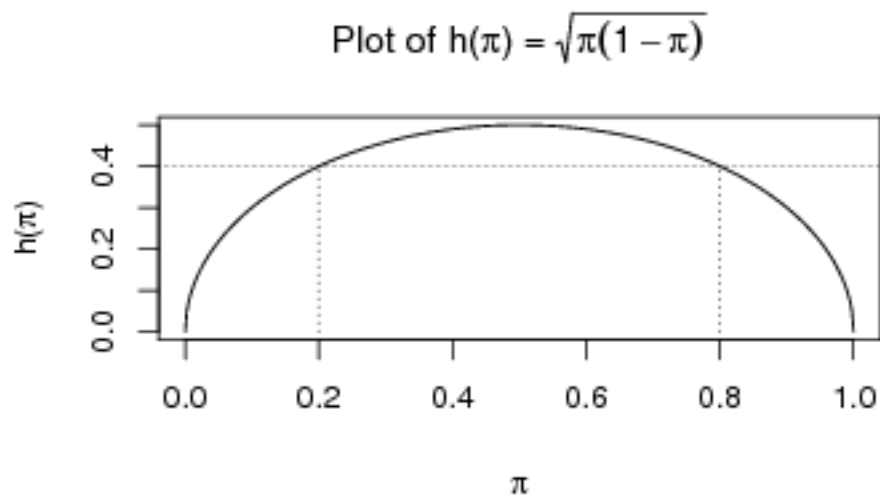
distributed population lies within 2 standard deviations of the mean. In this case that means that approximately 95% of all possible samples of size n have sample proportions that are within 2 standard deviations of their mean π . Therefore, when we randomly select our sample proportion from the population of all possible sample proportions, the **probability** is approximately 0.95 that the error of estimation, the difference between the estimate and the actual proportion, will be no more than 2 standard deviations, $2\sqrt{\pi(1-\pi)/n}$. This represents a bound on the error of estimation. It is not an absolute bound, but is a reasonable bound in the sense that there is only a 5% chance that the error of estimation will exceed this bound.



For example, suppose we randomly select 500 voters and find that 260 of these voters favor this candidate. Then our estimate of the population proportion is $\hat{p} = 260/500 = 0.520$. We are about 95% certain that the error of this estimate is no more than $2\sqrt{\pi(1-\pi)/n}$. The problem that remains to be solved is that this error bound depends on the value of π , which is unknown. There are two approaches we can take to solve this problem. The first approach is to note that the function

$$h(\pi) = \sqrt{\pi(1-\pi)}, \quad 0 \leq \pi \leq 1,$$

is a bounded function of π with upper bound $h(\pi) \leq 0.5$. The plot below shows how this function depends on π .



This implies that the bound on the error of estimation is at most $2(.5)/\sqrt{n} = 1/\sqrt{n}$. Therefore, we can make the following statement about the proportion of voters who favor our candidate based on the information contained in our sample: *the estimated proportion who favor our candidate is 0.520 and we are about 95% certain that this estimate is no more than $1/\sqrt{500} = 0.045$ from the actual population proportion.* Another way of stating this is that we are about 95% certain that the population proportion is within the interval 0.520 ± 0.045 , that is, between 0.475 and 0.565.

This bound on the error of estimation of a population proportion is conservative in the sense that it does not depend on the actual population proportion. However, if π is close to 0 or 1, then it will be too conservative because in this case, the value of $h(\pi)$ would be much smaller than the upper bound. It can be seen from the plot that if $.2 \leq \pi \leq .8$, then $.4 \leq h(\pi) \leq .5$, so the upper bound becomes too conservative when the population proportion is below .2 or above .8. In some situations, we may have prior information in the form of a bound on π that allows us to place a bound on $h(\pi)$. Suppose, for example, that we wish to estimate the proportion of memory chips that do not meet specifications, and we know from past history that this proportion has never exceeded 0.15. In that case, we can say that $h(\pi) \leq \sqrt{(.15)(.85)} = .357$. If a sample of 400 memory chips is randomly selected from a production run, and it is found that 32 fail to meet specifications, then the estimated population proportion is $\hat{p} = .080$, and a bound on the error of estimation would be $2(.357)/\sqrt{400} = 0.036$. We could present these results as follows: The estimated proportion of memory chips that do not meet specifications is 0.080. With 95% certainty, this proportion could be as low as 0.044 or as high as 0.116.

If we do not have available any prior bounds on the population proportion, then we could use \hat{p} in place of π in the error bound. That is, the estimated bound on the error of estimation would be

$$2\sqrt{\hat{p}(1 - \hat{p})/n}.$$

One of the interpretations of the estimate of the proportion of voters who favor our candidate is that we are 95% confident that this proportion is between 0.475 and 0.565. This interval represents a range of reasonable values for the population proportion. The confidence level of 95% is determined by the use of 2 standard deviations for the error bound and the property of normal curves that approximately 95% of a population falls within 2 standard deviations from the mean. However, this also implies that there is a 5% chance that the estimation error is greater than the stated bound, or that there is a 5% chance that the interval does not contain the population proportion. If there are very serious consequences of reporting an error bound that turns out to be too small, then we should decide what is an acceptable risk that the error bound is too small. We can then use the appropriate number of standard deviations so that the risk is acceptably small. Suppose for example that we are willing to accept a risk of 1% that the error bound is too small or that the resulting interval of reasonable values does not include the population proportion. To accomplish this, we must find the z-score such that the area between $-z$ and z is 0.99. To find this z-score, we must look for the area of $0.99/2 = 0.495$. The z-score that is closest to that area is $z=2.58$.

The resulting interval is

$$(0.520 \pm (2.58)\sqrt{(.52)(.48)/500}) \longleftrightarrow (0.520 \pm 0.058) \longleftrightarrow (0.462, 0.578).$$

In this case we are 99% confident that the proportion of voters who favor our candidate is somewhere within this interval. Such intervals are called **confidence intervals**. To summarize the discussion above, a confidence interval for a population proportion based on a random sample of size n is

$$\hat{p} \pm z\hat{\sigma},$$

where z is selected so that the area between $-z$ and z is the required level of confidence, and $\hat{\sigma}$ is

$$\hat{\sigma} = \begin{cases} \sqrt{p_0(1-p_0)/n}, & \text{if prior bound } p_0 \text{ is given} \\ \sqrt{\hat{p}(1-\hat{p})/n}, & \text{if no prior bound is given} \end{cases}$$

Confidence intervals have two related properties: the level of confidence and the precision as measured by the width of the confidence interval. These properties are inversely related. If the confidence level is increased, then the width is increased and so its precision is decreased. The only way to increase the confidence level while maintaining or increasing precision is to use a larger sample size. The sample size can be determined by specifying the confidence level and the required precision. Suppose for example that we would like to estimate the proportion who favor our candidate to within 0.025 with 95% confidence. To attain these goals requires that the confidence interval have the form $\hat{p} \pm e$, where e denotes the required precision, 0.025. Since there is no prior bound available for the population proportion, we must use the conservative standard deviation for the confidence interval, $\hat{p} \pm z/(2\sqrt{n})$. Therefore, to attain these goals we must have

$$\frac{z}{2\sqrt{n}} = e \iff n = \left(\frac{z}{2e}\right)^2,$$

where z is chosen so that the area between $-z$ and z is 0.95 and $e=0.025$. From the table of normal curve areas, the required z-score is 1.96, so $n = 39.2^2 = 1537$.

In situations where we have a prior bound on the population proportion, we can incorporate that bound into the sample size determination. If we would like to estimate the proportion of memory chips that do not meet specifications and we have a prior bound, $p \leq p_0$ for the proportion, then the confidence interval will have the form,

$$\hat{p} \pm z\sqrt{p_0(1-p_0)/n}.$$

This gives

$$\frac{z\sqrt{p_0(1-p_0)}}{\sqrt{n}} = e \iff n = \left(\frac{z}{e}\right)^2 p_0(1-p_0).$$

If we require that the estimate of this proportion be within .02 of the population proportion with 90% confidence, and we have a prior bound on the population of $p \leq 0.15$, then $z = 1.65$, $p_0 = 0.15$, and so the sample size would be

$$n = \left(\frac{1.65}{.02}\right)^2 (.15)(.85) = 868.$$

Estimation of a Population Mean

The results of the previous section are derived from the Central Limit Theorem. We can use similar methods to estimate the mean of a population. We will first consider this estimation problem when the population has a normal distribution, and then we will examine the extension of these methods to populations that are not necessarily normally distributed.

Recall that if the population has a normal distribution with mean μ and standard deviation σ , then the distribution of \bar{X} is $N(\mu, \sigma/\sqrt{n})$. This implies that we can use \bar{X} as an estimate of μ . The error of estimation is then $\bar{X} - \mu$, and we can make the following probability statement about this error,

$$P(|\bar{X} - \mu| > \frac{z_{\alpha/2}\sigma}{\sqrt{n}}) = P(|Z| > z_{\alpha/2}) = \alpha,$$

where $z_{\alpha/2}$ is the z-score such that the area to the right of $z_{\alpha/2}$ under the normal curve is $\alpha/2$. We use $\alpha/2$ so that the total area in both extremes is α . Therefore, the probability that the error of estimation exceeds $z_{\alpha/2}\sigma/\sqrt{n}$ is α and so, a $1 - \alpha$ confidence interval for the population mean is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

The problem here is that this confidence interval depends on σ , the population standard deviation. In most situations, σ is unknown as well as μ . Sometimes we have prior information available that gives an upper bound for σ , $\sigma \leq \sigma_0$, which can be incorporated into the confidence interval,

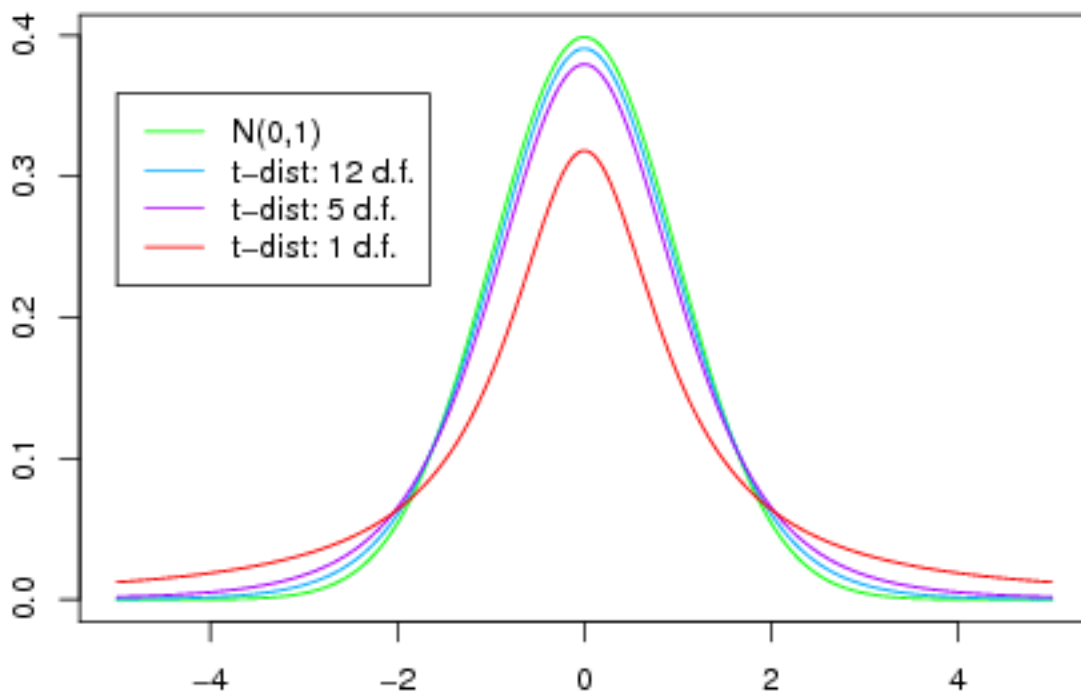
$$\bar{X} \pm z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}}.$$

Situations where no such upper bound is available require that we estimate σ with the sample standard deviation. However, using s in place of σ changes the sampling distribution of \bar{X} . What is required is to determine the distribution of

$$t_n = \frac{\bar{X} - \mu}{s/\sqrt{n}}.$$

This problem was solved around 100 years ago by a statistician named William Gossett, who solved it while working for Guinness brewery. Because of non-disclosure agreements in his employment contract, Gossett had to publish his work under the pseudonym *Student*.

For this reason, the distribution of t_n when X_1, \dots, X_n is a random sample from a normal distribution is called **Student's t distribution**. This distribution is similar to the standard normal distribution and represents an adjustment to the sampling distribution of \bar{X} caused by replacing the constant σ with a random variable s . As the sample size increases, s becomes a better estimate of σ , and so less adjustment is required. Therefore, the t-distribution depends on the sample size. This dependence is expressed by a function of the sample size called **degrees of freedom**, which for this problem is $n - 1$. That is, the sampling distribution of t_n is a **t-distribution with n-1 degrees of freedom**. A plot that compares several t-distributions with the standard normal distribution is given below. Note that the t-distribution is symmetric and has relatively more area in the extremes and less area in the central region compared to the standard normal distribution. Also, as the degrees of freedom increases, the t-distribution converges to the standard normal distribution.



We can now make use of Gossett's result to obtain a confidence interval for μ ,

$$\bar{X} \pm t_{n-1, \alpha/2} \frac{s}{\sqrt{n}},$$

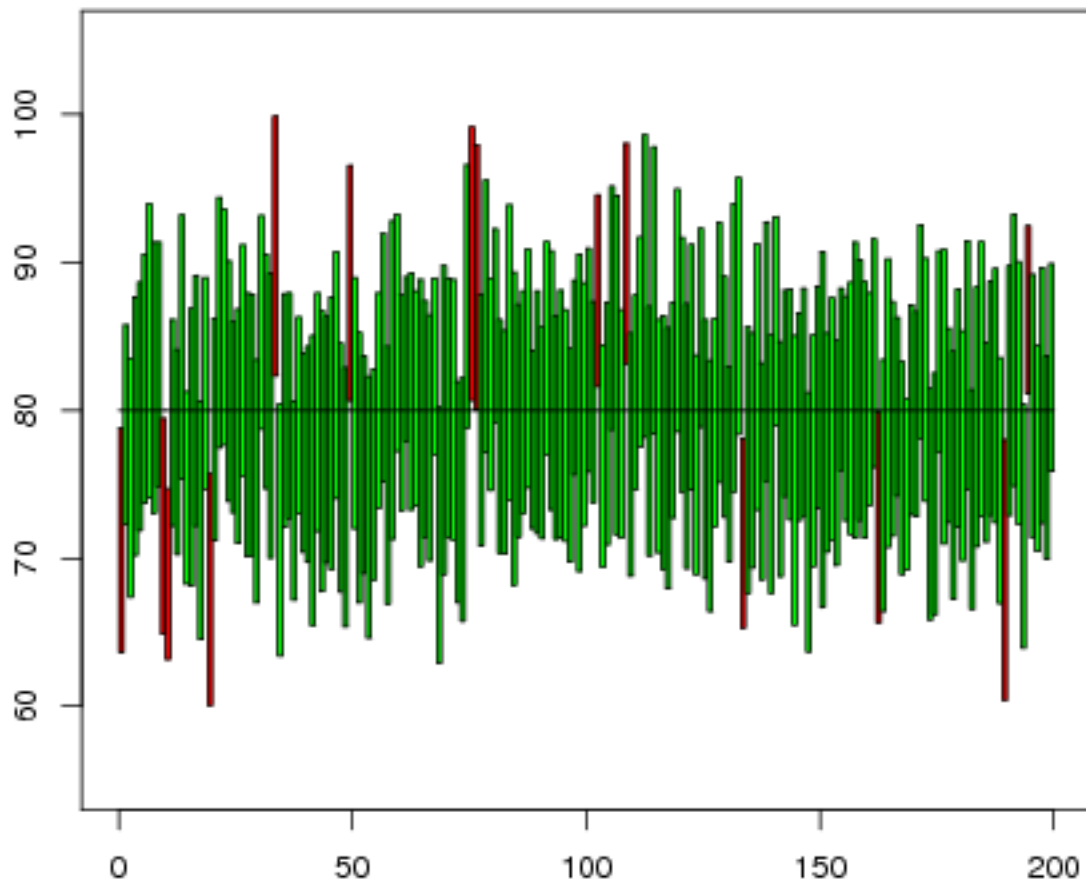
where $t_{n-1, \alpha/2}$ is the value from the t-distribution with $n-1$ degrees of freedom such that the area to the right of this value is $\alpha/2$. The interpretation of this interval is that it contains reasonable values for the population mean, reasonable in the sense that the probability that the interval does not contain the mean is α .

The probability statement associated with this confidence interval,

$$P \left(\bar{X} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \right) = 1 - \alpha,$$

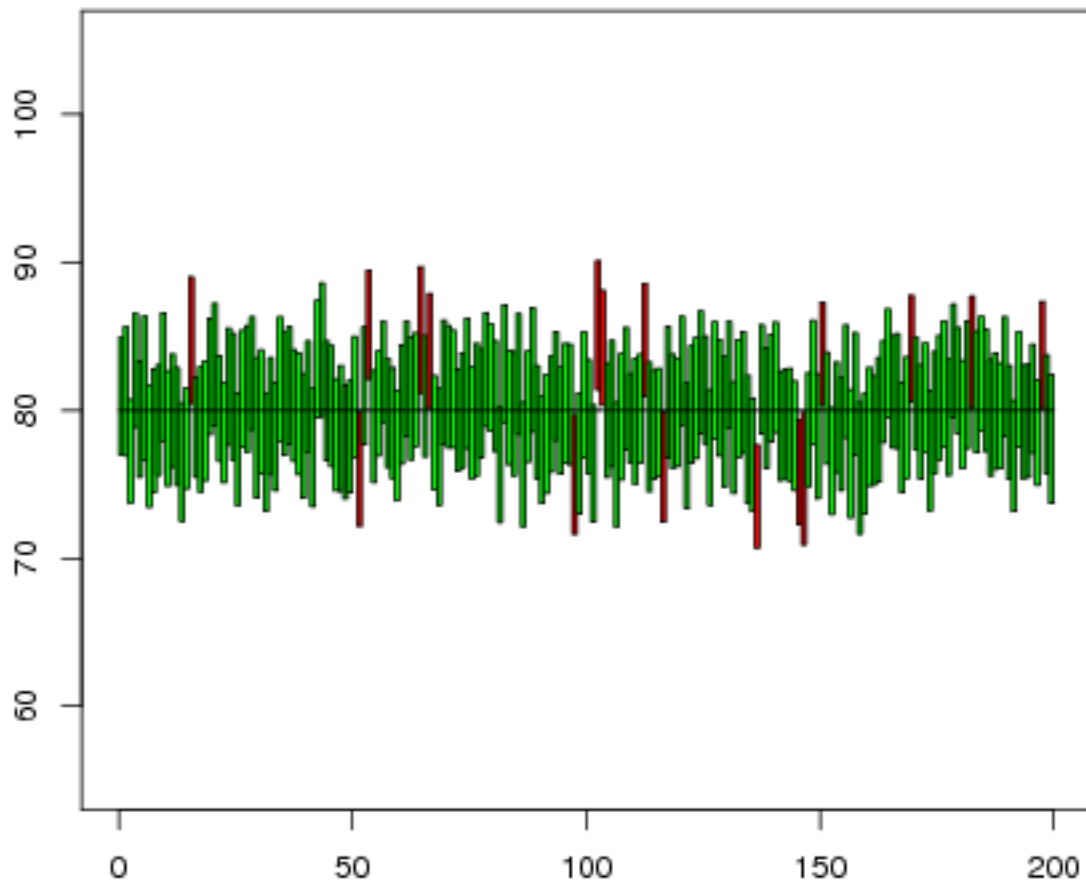
appears to imply that the mean μ is the random element of this statement. However, that is incorrect; what is random is the interval itself. This is illustrated by the following graphics. The first simulates the selection of 200 random samples each of size 25 from a population that has a normal distribution and the second performs the simulation with samples of size 100. Each vertical bar represents the confidence interval associated with one of these random samples. Green bars contain the actual mean and red bars do not. Note that the increased sample size does not change the probability that an interval contains the mean. Instead, what is different about the second graphic is that the confidence intervals are shorter than the intervals based on samples of size 25.

Simulation of 95% Confidence Intervals from $N(80,20)$
Sample size = 25



7% of these confidence intervals do not contain 80

Simulation of 95% Confidence Intervals from $N(80,20)$
Sample size = 100



8.5% of these confidence intervals do not contain 80

Sample size determination. If our estimate must satisfy requirements both for the level of confidence and for the precision of the estimate, then it is necessary to have some prior information that gives a bound on σ or an estimate of σ . Let σ_0 denote this bound or estimate, and let e denote the required precision. Then the confidence interval must have the form, $\bar{X} \pm e$, which implies that

$$\frac{z_{\alpha/2}\sigma_0}{\sqrt{n}} = e \iff n = \left(\frac{z_{\alpha/2}\sigma_0}{e}\right)^2.$$

Example. A random sample of 22 existing home sales during the last month showed that the mean difference between list price and sales price was \$4580 with a standard deviation of \$1150. Assume that the differences between list and sales prices have approximately a normal distribution and construct a 95% confidence interval for the mean difference for all existing home sales. What would you say if the mean difference between list and sales prices for the same month last year had been \$5500? Suppose you wish to estimate this mean to within \$250 with 99% confidence. What sample size would be required if you use the standard deviation of this sample as an estimate of σ ?

Solution. The confidence interval has the form

$$4580 \pm t_{21,.025} \frac{1150}{\sqrt{22}}.$$

A table of t-values is given in the text on page B10. It is formatted differently than the table of normal curve areas. In the t-table, degrees of freedom are in the left-hand margin and tail areas are in the top margin. This table gives $t_{21,.025} = 2.080$, and so the confidence interval is

$$4580 \pm (2.080)(1150)/\sqrt{22} \iff 4580 \pm 510 \iff [4070, 5090].$$

The interpretation of this interval is that it contains reasonable values for the population mean, reasonable in the sense that we are risking a 5% chance that the actual population mean is not one of these values. If the mean difference between list and sales prices for the same month last year had been \$5500, then we could say that the difference between list and sales price this year is less than last year since all of the reasonable values for this year's mean difference are lower than last year's mean. There is a risk of 5% that this conclusion is wrong. Note that the precision of this confidence interval is 510 with 95% confidence. If we require a precision of 250 with 99% confidence, then we must use a larger sample size. If the sample standard deviation, $s = 1150$, is used as an estimate of the σ for the purpose of sample size determination, then

$$n = \left(\frac{z_{\alpha/2}\sigma_0}{e}\right)^2,$$

where $\alpha = .01/2 = .005$, $\sigma_0 = 1150$, and $e = 250$. Note that the last row of the t-table with degrees of freedom equal to infinity corresponds to the standard normal distribution. Therefore, we can use this row to find the required z-score. This gives

$$n = \left(\frac{(2.576)(1150)}{250}\right)^2 = 11.85^2 = 141.$$

So a sample of size 141 would be required to meet these specifications. The actual precision attained by a confidence interval based on a sample of this size may not have a precision that is very close to 250 if the sample standard deviation in our preliminary sample of size 22 is not a good estimate of the actual population standard deviation.

Since the results discussed above are based on the Central Limit Theorem, we can apply them in the same way to the problem of estimating the mean of a population that does not necessarily have a normal distribution. This would lead to the same confidence interval for μ ,

$$\bar{X} \pm t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}.$$

The only difference is that such an interval would only be valid if the sample size is sufficiently large for the Central Limit Theorem to be applicable. Some caution must be used here, since the definition of sufficiently large depends on the distribution of the population.

Software for Statistical Analysis

Examples presented in class are obtained using the statistical programming language and environment **R**. This is freely available software with binaries for Linux, MacIntosh, Windows that can be obtained from

<http://cran.r-project.org>

For those who have never used *R* or who need a refresher on its use, there is an excellent introduction that can be downloaded from

<http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf>

This document is the basis for Homework Assignment 0 given in the next section.

R Notes

1.3 A Simple Example: the `c()` Function and the Assignment Operator There are now two assignment operators, `=` and `<-`. The second operator is available for compatibility with old versions of **R**. The `=` now is most commonly used to assign a value to a name.

1.4 The Workspace The *Workspace* only exists in the computer's memory, not on the physical hard drive, and so **R** offers a choice to the user when exiting: save the workspace or do not save it. If the Workspace is not saved, all objects created during the session will be lost. That's no problem if you are using it only as a mathematical or statistical calculator. If you are performing an analysis, but must exit before completing it, then you don't want to lose what you have already done. There is an alternative that I recommend instead of saving the workspace: write the commands you wish to enter into a text file and then copy/paste from the edit window into the **R** console. Even though this may seem like extra work, it has three advantages: 1) any mistakes can be corrected immediately with the editor; 2) you won't have to remember what the objects in a workspace represent since the file contains the

commands that created those objects; 3) if you need to perform a similar analysis at a later time, you can just copy the original file to a new name and modify/extend the commands in the file to complete the later analysis. **You must use a plain text editor to edit command files, not a document editor like Word.**

2.3 Matrix Operations. Matrix-matrix multiplication can be performed only when the two matrices are conformable, that is, their inner dimensions are the same. For example, if A is $n \times r$ and B is $r \times m$, then matrix-matrix multiplication of A and B is defined and results in a matrix C whose dimensions are $n \times m$. Elementwise multiplication of two matrices can be performed when both dimensions of the two matrices are the same. That is, $c_{i,j} = a_{i,j}b_{i,j}$, $1 \leq i \leq n$, $1 \leq j \leq m$. These two different types of multiplication operations must be differentiated by using different symbols, since both types would be possible if the matrices have the same dimensions. Matrix-matrix multiplication is denoted by $A\% * \%B$ and elementwise multiplication is denoted by $A * B$. The same situation occurs if A and B are both vectors that have the same length n . In that case, $A\% * \%B$ represents the *dot product* of these vectors,

$$A\% * \%B = \sum_{i=1}^n A_i B_i.$$

Note that this result is a scalar. $A * B$ represents elementwise multiplication. The result is a vector C with $c_i = a_i b_i$.

Sequences. Another useful function for creating vectors is *seq*. This function creates a sequence of numbers and has optional arguments *by* and *length*. The operator `:` is a shortcut for creating a sequence of consecutive integers.

```
x = 3:10
y = seq(0,10,length=101)
n = seq(1,20,by=3)
```

3.2 Reading Data from files. The two main functions to read data that is contained in a file are *scan()* and *read.table()*.

scan(Fname) reads a file whose name is the value of **Fname** as a vector. All values in the file must be the same type (numeric, string, logical). By default, *scan()* reads numeric data. If the values in this file are not numeric, than the optional argument **what=** must be included. For example, if the file contains strings, then

```
x = scan(Fname,what=character(0))
```

will read this data. Note that **Fname** as used here is an **R** object whose value is the name of the file that contains the data.

Note: if the file is not located in the working directory, then full path names must be used to specify the file. **R** uses unix conventions for path names regardless of the operating system. So, for example, in Windows a file located on the C-drive in folder Stat3355data named Data1.txt would be scanned by

```
x = scan("c:/Stat3355data/Data1.txt")
```

3.3 Data Frames and read.table(). Tabular data contained in a file can be read by **R** using the `read.table()` function. Each column in the table is treated as a separate variable and variables can be numeric, logical, or character (strings). That is, different columns can be different types, but each column must be the same type. An example of such a file is <http://www.utdallas.edu/~ammann/stat3355scripts/Temperature.data>. Note that the first few lines begin with the character `#`. This is the comment character. **R** ignores that character and the remainder of the line. The first non-comment line contains names for the columns. In that case we must include the optional argument `header=TRUE` as follows:

```
Temp = read.table("http://www.utdallas.edu/~ammann/stat3355scripts/Temperature.data",
  header=TRUE)
```

The first column in this file is not really data, but just gives the name of each city in the data set. These can be used as row names:

```
Temp = read.table("http://www.utdallas.edu/~ammann/stat3355scripts/Temperature.data",
  header=TRUE,
  row.names=1)
```

The value returned by `read.table()` is a *data.frame*. This type of object can be thought of as an enhanced matrix. It has a *dimension* just like a matrix, the value of which is a vector containing the number of rows and number of columns. However, a data frame is intended to represent a data set in which each row is the set of variables obtained for each subject in the sample and each column contains the observations for each variable being measured. In the case of the Temperature data, these variables are:

JanTemp, *Lat*, *Long*. Unlike a matrix, a data frame can have different types of variables, but each variable (column) must contain the same type.

Individual variables in a data frame can be accessed several ways.

1. `$`

```
Latitude = Temp$Lat
```

2. Name:

```
Latitude = Temp[["Lat"]]
```

3. Number:

```
Latitude = Temp[[2]]
```

Note that the object named `Latitude` is a vector. If you want to extract a subset of the variables with all rows included, then use `[]`. The result is a data frame. If the original data frame has names, these are carried over to the new data frame. If you only want some of the rows, then specify these the way it is done with matrices:

```
LatLong = Temp[2:3] #extract variables 2 through 3
LatLong = Temp[c("Lat","Long")] #extract Lat and Long
LatLong1 = Temp[1:20,c("Lat","Long")] #extract first 20 rows for Lat and Long
```

Although it may seem like more work to use names, the advantage is that one does not need to know the index of the desired column, just its name.

Additional variables can be added to a data frame as follows.

```
#create new variable named Region with same length as other variables in Temp
Region = rep("NE",dim(Temp)[1])
# NE is defined to be Lat >= 39.75 and Long < 90
# SE is defined to be Lat < 39.75 and Long < 90
# SW is defined to be Lat < 39.75 and Long >= 90
# NW is defined to be Lat >= 39.75 and Long >= 90
Region[Temp$Lat < 39.75 & Temp$Long < 90] = "SE"
Region[Temp$Lat < 39.75 & Temp$Long >= 90] = "SW"
Region[Temp$Lat >= 39.75 & Temp$Long >= 90] = "NW"
#give Region the same row names as Temp
names(Region) = dimnames(Temp)[[1]]
#make Region a factor
Region = factor(Region)
#add Region to Temp
Temp1 = data.frame(Temp,Region)
#plot January Temperature vs Region
#since Region is a factor, this gives a boxplot
plot(JanTemp ~ Region,data=Temp1)
```

R scripts. The most effective way to learn and use **R** is to use a plain text editor to write the commands and then copy/paste those lines into the **R** command line. Then any errors can be corrected immediately. Furthermore, if you have forgotten details of what was done in a previous session, the file is there to show the commands. If, in addition, you include comment lines that explain what is being done, then the commands will be easier to follow. An example of this is the file

<http://www.utdallas.edu/~ammann/stat3355scripts/dose.r>

Comment lines describe each command in this file. Once the commands are correct, the entire file can be executed within **R** by the `source()` function:

```
source("http://www.utdallas.edu/~ammann/stat3355scripts/dose.r")
```

The argument to *source()* is the name of a file either on the local system or, as is the case here, on a remote system. All commands in the file are executed line-by-line until the end-of-file is reached.

Saving graphics. By default **R** uses a separate graphical window for the display of graphic commands. A graphic can be saved to a file using any of several different graphical file types. The most commonly used are *pdf()* and *png()* since these types can be imported into documents created by **Word** or \LaTeX . The first argument for these functions is the filename. Arguments *width=*, *height=* give the dimensions of the graphic. For *pdf()* the dimension units are inches, for *png()* the units are pixels. *pdf()* supports multi-page graphics, but *png()* only allows one page per file unless the file name has the form *Myplot%d.png*. For example,

```
pdf("TempPlot.pdf",width=6,height=6)
plot(JanTemp ~ Lat,data=Temp)
plot(JanTemp ~ Region,data=Temp1)
graphics.off()
#creates a 2-page pdf document
png("TempPlot%d.png",width=480,height=480)
plot(JanTemp ~ Lat,data=Temp)
plot(JanTemp ~ Region,data=Temp1)
graphics.off()
#creates two files: TempPlot1.png and TempPlot2.png
```

The function *graphics.off()* writes any closing material required by the graphic file type and then closes the graphics file.

Homework and Project Assignments

Assignment 0

There is no due date for this assignment, and the assigned exercises do not need to be turned in for grading. However, the sooner you complete this assignment, the sooner you will be able to understand the scripts that are used for examples presented in class, and the sooner you will be able to begin work on the graded projects that will be assigned. This assignment refers to the introductory guide to **R**, **Owen-TheRGuide.pdf**. Additional information to supplement this guide is given in the **Notes** section below. Other chapters from this guide will be added as we progress through the course.

1. Read Chapters 1, 2 of **Owen-TheRGuide.pdf**. Complete the exercises in *Section 2.4*.
2. Read Chapter 3. Complete the exercises in *Section 3.4*.
3. Read Chapter 4. Complete the exercises in *Section 4.5*.
4. Read Chapter 5. Complete the exercises in *Section 5.3*.
5. Read Section 7.3.

Homework 1

1. Use the data contained in the file

<http://www.utdallas.edu/~ammann/stat3355scripts/Smoking.txt>

- Find the means and standard deviations for each variable.
- Which states are more than 2 sd's above the mean for cigarette consumption? for bladder cancer? for lung cancer?
- Which states are in the top 10% of cigarette consumption? of bladder cancer? of lung cancer? (see documentation for **R** function *quantile()*)
- Plot cigarette consumption versus lung cancer and add an informative title.
- Repeat for bladder cancer.

2. Use the data contained in the file

<http://www.utdallas.edu/~ammann/stat3355scripts/Sleep.data>

A description of this data is given in

<http://www.utdallas.edu/~ammann/stat3355scripts/Sleep.txt>

The `Species` column should be used as row names.

- Construct histograms of each variable.
- The strong asymmetry for all variables except `Sleep` indicates that a *log* transformation is appropriate for those variables. Construct a new data frame that contains `Sleep`, replaces *BodyWgt*, *BrainWgt*, *LifeSpan* by their log-transformed values, and then construct histograms of each variable in this new data frame.
- Plot `LifeSpan` vs `BrainWgt` with `LifeSpan` on the y-axis. Repeat using these variables after applying a log-transformation to both variables. Superimpose lines corresponding to the respective means of the variables for each plot.
- What proportion of each of these two variables are within 2 s.d.'s of their means? Answer this question for the original variables and for the log-transformed variables.
- Obtain the correlation between *LifeSpan* and *BrainWgt*. Repeat for *Log(LifeSpan)* and *log(BrainWgt)*. Interpret these correlations.
- Obtain the least squares regression line to predict *LifeSpan* based on *BrainWgt*. Repeat to predict *log(LifeSpan)* based on *log(BrainWgt)*. Predict *LifeSpan* of *Homo sapiens* based on each of these regression lines. Which would you expect to have the best overall accuracy? Which prediction is closest to the actual *LifeSpan* of *Homo sapiens*?

Homework 2

Problems from textbook, p. 349-351: 6.85, 6.89, 6.93.

Additional problems.

1. A researcher has developed a performance score for high schools that she believes gives a reasonable indication of how well a district prepares its students for college. She examines a group of 100 high schools from a variety of school districts, obtains her performance score for each high school, and also obtains the average freshman year GPA for a group of students from each high school. The results are given below. Interpret this information. Now suppose that a particular high school has a performance score of 220 and the average GPA of freshmen from this high school is 2.80. What can you say about the relative standing of this high school? What conclusions can you draw about the actual GPA of students from this high school compared to the GPA predicted by its performance score? Explain what possible problems might exist if performance scores are used to predict GPA based on this data.

mean performance = 156, s.d. = 30

mean GPA = 2.65, s.d. = .4

r = 0.72

2. In a study of prisoners in the state of Texas who have been convicted of crimes against persons, it was found that the average sentence length was 18.5 years with a standard deviation of 8. Also, the average years of school for these prisoners was 9.5 with a standard deviation of 2.5, and the correlation between years of school and the length of sentence was -0.82. Interpret this information. What would you say is an unusually long sentence among this group? What can you say about the relative standing of a prisoner who has received a sentence of 15 years and who has had 16 years of school? Predict the years of school of a prisoner who has received a sentence of 15 years. Predict the length of sentence for a prisoner who has 16 years of school.

Solutions for Homework 2

1. **Problem 6.85.** $P(A_1) = 0.4$, $P(L|A_1) = 0.02$, so by multiplicative law,

$$P(A_1 \cap L) = P(L|A_1)P(A_1) = (0.4)(0.02) = 0.008.$$

2. **Problem 6.89.** The event that she must examine at least two disks is equivalent to the event that the first disk is not blank. So that probability is $10/25 = 0.4$.

3. **Problem 6.93.**

- a. There are 2517 viewers total. The number who saw a PG movie was $179 + 87 = 266$, so

$$P(\text{PG}) = \frac{266}{2517} = 0.1057.$$

- b. $P(\text{PG} \cup \text{PG} - 13) = \frac{420 + 323 + 179 + 114 + 87}{2517} = 0.4462.$

- c. $P(R^c) = 1 - \frac{600 + 196 + 205 + 139}{2517} = 1 - 0.4529 = .5471.$

4. *Interpret this information:* For mean and s.d., use Chebychev's theorem or the empirical rule, whichever is appropriate. In this case, we do not know that these variables have bell-shaped histograms, so use Chebychev's theorem.

Performance score: at least 75% of schools had performance score in the interval $156 \pm 2(30)$ ([96, 216]).

Mean GPA: at least 75% of schools had mean GPA in the interval $2.65 \pm 2(.4)$ ([1.85, 3.45]).

$r = 0.72$: positive correlation indicates an increasing relationship, r-squared is $0.72^2 = 0.5184$, so 51.84% of variability in mean GPA is due to linear relationship with performance score.

Relative standing: this school has performance score that is 2.13 s.d.'s above the mean performance score, so this is relatively high. Its mean GPA is 0.375 s.d.'s above the mean for GPA, so that is only slightly above average.

GPA predicted by performance score: GPA is the Y-variable and performance score is the X-variable, so

$$b = .72 \frac{.4}{30} = 0.0096, \quad a = 2.65 - 0.0096(156) = 1.1524.$$

Predicted GPA for this school is

$$\hat{Y} = 1.1524 + 0.0096(220) = 3.2644.$$

This is much higher than its actual mean GPA of 2.80.

Problems: the relationship may not be linear; there may be a few schools with relatively extreme values which would distort the means, s.d.'s, and the correlation, and so would distort the prediction equation as well.

5. *Interpret* means and s.d.'s using Chebychev's theorem.

Correlation is negative, so relationship is decreasing – prisoners with above average years in school tended to have below average sentence length. R-squared is $(-0.82)^2 = 0.6724$. So 67.24% of variability in sentence length is due to a linear relationship with years in school.

Unusually long sentence length: use 2 s.d.'s above the mean for this (higher values than 2 also could have been used). This is $18.5 + 2(8) = 34.5$.

Relative standing: Sentence is 15 years which is 0.44 s.d.'s below the mean, so this is just a little below average. Years in school is 2.6 s.d.'s above the mean, so this is much higher than average.

Predict years of school. Y-variable is years in school and X-variable is sentence length.

$$b = (-.82) \frac{2.5}{8} = -0.256, \quad a = 9.5 - (-0.256)18.5 = 14.24,$$

$$\hat{Y} = 14.24 + (-0.256)(15) = 10.4.$$

Predict sentence length. Y-variable is sentence length and X-variable is years in school.

$$b = (-.82) \frac{8}{2.5} = -2.624, \quad a = 18.5 - (-2.624)9.5 = 43.428,$$

$$\hat{Y} = 43.428 + (-2.624)(16) = 1.444.$$

Note that the second prediction equation is *NOT* equivalent to inverting the first equation.

Homework 3

1. Textbook, p. 310, problem 6.33
2. Textbook, p. 311, problem 6.36
3. Textbook, p. 334, problem 6.70
4. Textbook, p. 334, problem 6.71
5. A small brewery has three bottling machines. Machine A produces 40% of all the bottles, machines B and C produce 30% each. Five percent of bottles filled by A, four percent of bottles filled by B, and three percent of bottles filled by C are rejected for some reason.
 - a. If a bottle is filled by A or B, what is the probability that it is rejected?
 - b. If a bottle is rejected, what is the probability that it was filled by A?
6. Determine whether or not each of the following statements satisfies the laws of probability. For those that do not, state how the laws are violated.
 - (a) $P(A) = 0.9, P(A|B) = 0.5, P(B|A) = 0.6$.
 - (b) A and B are independent, $P(A \cup B) = 0.8, P(A) = 0.6$.
 - (c) $P(A^c|B) = 0.3, P(A|B^c) = 0.7, P(B) = 0.9, P(A) = 0.6$.

Solutions to Homework 3

1. Problem 6.33.

- $P(F|\text{Predicted F}) = 432/(432 + 130) = 0.769$.
- $P(M|\text{Predicted M}) = 390/(48 + 390) = 0.890$.
- It appears that prediction for males are more reliable.

2. Problem 6.36.

- $P(\text{Former smoker}) = 99/294 = 0.3367$.
- $P(\text{Very harmful}) = 224/294 = 0.7619$.
- $P(\text{Very harmful}|\text{Current}) = 60/96 = 0.625$.
- $P(\text{Very harmful}|\text{Former}) = 78/99 = 0.7879$.
- $P(\text{Very harmful}|\text{Never}) = 86/99 = 0.8687$.

3. Problem 6.70.

- These percentages have different reference groups. For 8%, reference group is adult full-time workers; reference group for 70% is adult drug users. Even though the numerators are the same, the denominators are different, so these percentages could be correct.
- $P(D|E) = 0.08$, $P(E|D) = 0.70$.
- Not possible. We would need $P(E)$, the proportion of adults who are employed full-time, or $P(D)$, the proportion of adults who are drug users.

4. Problem 6.71. We are given,

$$P(\text{Basic}) = 0.40, P(\text{Extended}|\text{Basic}) = 0.30, P(\text{Extended}|\text{Deluxe}) = 0.50.$$

We are asked to find

$$P(\text{Basic}|\text{Extended}) = \frac{P(\text{Basic} \cap \text{Extended})}{P(\text{Extended})}.$$

Numerator is $P(\text{Basic} \cap \text{Extended}) = P(\text{Extended}|\text{Basic})P(\text{Basic}) = 0.12$. To obtain denominator, we need to use the theorem of total probability,

$$\begin{aligned} P(\text{Extended}) &= P(\text{Basic} \cap \text{Extended}) + P(\text{Deluxe} \cap \text{Extended}) \\ &= 0.12 + P(\text{Extended}|\text{Deluxe})P(\text{Deluxe}) \\ &= 0.12 + (0.5)(0.6) = 0.42. \end{aligned}$$

This gives, $P(\text{Basic}|\text{Extended}) = .12/.42 = 0.2857$.

5. We are given,

$$P(A) = 0.40, P(B) = P(C) = 0.30, P(\text{Rejected}|A) = 0.05, P(\text{Rejected}|B) = 0.04, P(\text{Rejected}|C) = 0.03$$

We can convert these conditional probabilities to probabilities of the respective intersections by the multiplicative rule.

$$P(\text{Rejected} \cap A) = (.05)(.4) = .020$$

$$P(\text{Rejected} \cap B) = (.04)(.3) = .012$$

$$P(\text{Rejected} \cap C) = (.03)(.3) = .009$$

a. We must find $P(\text{Rejected}|A \cup B)$. The denominator for this conditional probability is $P(A) + P(B) = 0.70$. The numerator is obtained by

$$\begin{aligned} P(\text{Rejected} \cap (A \cup B)) &= P((\text{Rejected} \cap A) \cup (\text{Rejected} \cap B)) \\ &= P(\text{Rejected} \cap A) + P(\text{Rejected} \cap B) \\ &= .020 + .012 = 0.032. \end{aligned}$$

Therefore, $P(\text{Rejected}|A \cup B) = .032/.70 = 0.0457$.

b. By Theorem of Total Probability,

$$P(\text{rejected}) = 0.020 + 0.012 + 0.009 = 0.041,$$

and so

$$P(A|\text{rejected}) = \frac{P(\text{rejected} \cap A)}{P(\text{rejected})} = \frac{0.020}{0.041} = 0.488.$$

6. a. $P(A) = 0.9, P(A|B) = 0.5, P(B|A) = 0.6$.

First note that

$$P(A \cap B) = P(B|A)P(A) = (0.6)(0.9) = 0.54,$$

Next look at $P(A|B)$ and how it is related to what we just found, $P(A \cap B)$.

$$0.54 = P(A \cap B) = P(A|B)P(B) = 0.5 * P(B).$$

This implies that $P(B) = 0.54/0.5 = 1.08$, which violates the probability axioms.

b. A and B are independent, $P(A \cup B) = 0.8, P(A) = 0.6$.

Note that independence implies

$$P(A \cap B) = P(A)P(B) = 0.6 * P(B).$$

The additivity law implies that

$$0.8 = P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + P(B) - 0.6 * P(B) = 0.6 + 0.4 * P(B)$$

and so

$$P(B) = \frac{0.8 - 0.6}{0.4} = 0.5.$$

So everything is OK.

c. $P(A^c|B) = 0.3$, $P(A|B^c) = 0.7$, $P(B) = 0.9$, $P(A) = 0.6$.

$$P(A^c \cap B) = P(A^c|B)P(B) = (0.3)(0.9) = 0.27,$$

$$P(A \cap B^c) = P(A|B^c)P(B^c) = (0.7)(0.1) = 0.07$$

$$0.6 = P(A) = P(A \cap B) + P(A \cap B^c) = P(A \cap B) + 0.07$$

$$\rightarrow P(A \cap B) = 0.6 - 0.07 = 0.53$$

$$P(B) = P(A \cap B) + P(A^c \cap B) = 0.53 + 0.27 = 0.80,$$

but we are told that $P(B) = 0.9$, so the law of total probability is violated.

Homework 4

1. Textbook, p. 395, problem 7.48
2. Textbook, p. 395, problem 7.51
3. Textbook, p. 395, problem 7.52
4. Textbook, p. 413, problem 7.73
5. Textbook, p. 413, problem 7.74
6. Textbook, p. 414, problem 7.80
7. Textbook, p. 432, problem 7.109
8. Textbook, p. 433, problem 7.123
9. Textbook, p. 433, problem 7.124

Additional problems

1. The Human Resources Department of a large corporation wanted to determine if a majority of its employees were satisfied with their treatment by the corporation's health care provider. A random sample of 400 employees was selected, and 251 indicated that they were satisfied with their treatment. Construct a 95% confidence interval for the proportion of all employees who are satisfied. What sample size would be required to estimate this proportion to within $\pm 2\%$ with 90% confidence if no prior bounds are placed on the population proportion?

2. A researcher is interested in the long-term results of individuals who use the weight-loss center at a particular private hospital. In particular, she would like to know if there has been any change in the mean weight of these individuals 1 year after finishing the program. Suppose that she randomly selects 25 of these individuals and finds that the mean weight loss after 1 year is 5.4 with a s.d. of 8.2 (a negative loss represents a weight gain). Construct a 90% confidence interval for the mean weight loss after 1 year.
3. Suppose a research article reported that a large population of 12 year olds with a particular learning disability had a mean score of 72 on a math skills test. You wish to determine if a new protocol for educating these students can improve their math skills. A random sample of 24 such children receive this new protocol and are then tested 1 month after completion. Suppose the mean score for these 24 students is 78 with a standard deviation of 10. Construct a 95% confidence interval for the mean score of all students with this learning disability who receive the protocol. You may assume that the test scores have approximately a normal distribution.
4. A random sample of 43 5 year olds is given a puzzle to solve, and it is found that the mean completion time is 200 seconds with a standard deviation of 28. Construct a 99% confidence interval for the mean completion time of all 5 year olds. Use the sample mean and sample standard deviation to approximate the proportion of these children who take more than 250 seconds to solve the puzzle, and the proportion who take between 180 and 240 seconds, assuming that the histogram of the completion times is approximately a normal distribution.

Solutions to Homework 4

1. Textbook, p. 395, problem 7.48. Binomial distribution where S denotes the event that a passenger rests or sleeps, N represents the number of passengers among the 10 who rest or sleep, and $P(S) = 0.80$. Use Table 9 with $n = 10, \pi = 0.8$.
 - a. $P(N = 8) = 0.302$.
 - b. $P(N \leq 7) = 1 - P(N \geq 8) = 1 - p(8) - p(9) - p(10) = 1 - .677 = .323$.
 - c. More than half corresponds to more than 5.

$$P(N > 5) = 1 - P(N \leq 5) = 1 - .001 - .006 - .026 = 0.967.$$

Note: adding the table values from 6 to 10 gives 0.966. The difference is due to round-off errors.

2. Textbook, p. 395, problem 7.51. Binomial($n = 10, \pi = .15$), where S denotes the event that an individual fails the polygraph test and N denotes the number among the 10 who fail the polygraph.

a. All pass is the event that $N = 0$. $P(N = 0) = .85^{10} = 0.1969$.

b.

$$P(N > 2) = 1 - P(N \leq 2) = 1 - p(0) - p(1) - p(2) = 1 - .85^{10} - 10(.15)(.85^9) - 45(.15^2)(.85^8)$$

c. $E(N) = 500(.15) = 75$, $sd(N) = \sqrt{500(.15)(.85)} = 7.98$.

d. Use normal approximation to binomial.

$$P(N < 25) \approx P(Z < (25 - 75)/7.98) = P(Z < 6.27) \approx 0.$$

3. Textbook, p. 395, problem 7.52. Use Binomial($n = 20, \pi$) for $\pi = 0.05, 0.10, 0.20$ to find $P(N \leq 1)$. a. $\pi = .05$: $P(N \leq 1) = 0.735$.

b. $\pi = .10$: $P(N \leq 1) = 0.392$.

a. $\pi = .05$: $P(N \leq 1) = 0.070$.

4. Textbook, p. 413, problem 7.73. Use Normal(3432, 482).

a.

$$P(X > 4000) = P(Z > (4000 - 3432)/482) = P(Z > 1.18) = 1 - .8810 = .119.$$

$$\begin{aligned} P(3000 < X < 4000) &= P((3000 - 3432)/482 < Z < (4000 - 3432)/482) \\ &= P(-.90 < Z < 1.18) \\ &= .8810 - (1 - .8159) = .6969. \end{aligned}$$

b. $P(X < 2000) = P(Z < -2.97) = 1 - .9985 = .0015$.

$P(X > 5000) = P(Z > 3.25) = 1 - .9994 = .0006$. So

$$P((X < 2000) \cup (X > 5000)) = .0021.$$

c. $P(X > 7(453.59)) = P(X > 3175.13) = P(Z > (3175.13 - 3431)/482) = P(Z > -.53) = 0.7019$.

d. Most extreme 0.1% are weights below the 0.0005 quantile and weights above the 0.9995 quantile. Z-score is 3.27, so the 0.0005 quantile is $3432 - 3.27 * 482 = 1856$ and the 0.9995 quantile is $3432 + 3.27 * 482 = 5008$.

e. Distribution in pounds is normal with mean $3432/453.59 = 7.57$ and s.d. $482/453.59 = 1.06$. Answers will be the same.

5. Textbook, p. 413, problem 7.74.

$$P(\text{Good}) = P(2.9 < X < 3.1) = P(-1 < Z < 1) = 2(.8413 - .5) = 0.6826.$$

Proportion of defective corks is $1 - 0.6826 = 0.3174$.

6. Textbook, p. 414, problem 7.80.
- $P(X > 50) = P(Z > 1) = 1 - .8413 = .1587$.
 - Use $z = 1.28$ (from normal table or bottom row of t-table). Then time is $45 + 1.28 * 5 = 51.4$.
 - $z = -0.67$, so fastest 25% is $X = 45 - .67(5) = 41.65$.

7. Textbook, p. 432, problem 7.109.
- $P(X > 45) = P(Z > (45 - 60)/10) = P(Z > -1.5) = 0.9332$.
 - $Z = 1.28$, $X = 60 + 1.28(10) = 72.8$.
 - $Y = 10 + 50X$, $E(Y) = 10 + 50E(X) = 10 + 50(1) = 60$.

8. Textbook, p. 433, problem 7.123.
- $E(N) = n\pi = 200(.16) = 32$.

$$sd(N) = \sqrt{200(.16)(.84)} = 5.1846.$$

b.

$$P(25 \leq X \leq 40) \approx P((25 - 32)/5.1846 \leq Z \leq (40 - 32)/5.1846) = P(-1.35 \leq Z \leq 1.54) = .9$$

c. $50/200 = .25$ which is higher than 16%. The probability of getting a sample proportion this high or higher is approximately

$$P(\hat{p} > .25) \approx P(Z > (.25 - .16)/\sqrt{(.16)(.84)/200}) = P(Z > 3.47) = 0.0003.$$

Not likely.

9. Textbook, p. 433, problem 7.124.

Homework 5

Note: all hypothesis testing problems should include a p-value as part of your answer.

1. The Human Resources Department of a large corporation wanted to determine if a majority of its employees were satisfied with their treatment by the corporation's health care provider. A random sample of 400 employees was selected, and 251 indicated that they were satisfied with their treatment. What conclusion should you make at the 5% level of significance? What type error might you make with this decision? What is the probability that the null hypothesis will be rejected if 60% of employees are satisfied with their treatment?
2. Experience has shown that the mean number of sick leave days taken per year by employees in a large corporation is 10. Suppose that a random sample of 62 employees who smoke is selected, and suppose that their sick leave records for the last year show that the average number of sick leave days taken is 15 with a s.d. of 12. Does this data show at the 1% level of significance that the mean number of sick leave days for all employees who smoke is greater than the overall mean? Which type error might you make with this decision?
3. A researcher is interested in the long-term results of individuals who use the weight-loss center at a particular private hospital. In particular, she would like to know if there has been any change in the mean weight of these individuals 1 year after finishing the program. Suppose that she randomly selects 25 of these individuals and finds that the mean weight loss after 1 year is 5.4 with a s.d. of 8.2 (a negative loss represents a weight gain). Does this data show at the 10% level of significance that the mean weight loss after 1 year is greater than 2?
4. Suppose a research article reported that a large population of 12 year olds with a particular learning disability had a mean score of 72 on a math skills test. You wish to determine if a new protocol for educating these students can improve their math skills. A random sample of 24 such children receive this new protocol and are then tested 1 month after completion. Suppose the mean score for these 24 students is 78 with a standard deviation of 10. Does this data show at the 5% level of significance that the mean score of all students with this learning disability who receive the protocol has improved? You may assume that the test scores have approximately a normal distribution.
5. A random sample of adult smokers was asked the age at which they started to smoke. The results are summarized in the table below. Does this data show that gender and age when smoking began are related? Use 5% level of significance.

Age	Male	Female
< 16	25	10
16-17	24	17
18-20	28	32
≥ 21	19	34

Solutions for Homework 5

1. $H_0 : \pi \leq .5$, $H_1 : \pi > .5$, $\hat{p} = 251/400 = .6275$. P-value is

$$P(Z \geq \frac{.6275 - .5}{\sqrt{(.5)(.5)/400}} = P(Z > 5.1) = 0.$$

Reject H_0 . We might be making a Type 1 error with this decision. Rejection region for $\alpha = .05$ is

$$\hat{p} \geq .5 + 1.645\sqrt{(.5)(.5)/400} = 0.54.$$

If $\pi = 0.60$, then

$$P_6(\text{reject } H_0) = P(Z \geq (.54 - .6)/\sqrt{(.6)(.4)/400}) = P(Z \geq -2.45) = 0.9929.$$

2. $H_0 : \mu \leq 10$, $H_1 : \mu > 10$. P-value is obtained from t-distribution with 61 d.f. (use d.f.=60 from table).

$$P(t \geq \frac{15 - 10}{12/\sqrt{62}}) = P(t \geq 3.28).$$

Using Table 4 and 60 d.f., $P(t \geq 3.28) = 0.001$, so we reject H_0 at 1% level of significance. We might be making a Type 1 error with this decision.

3. $H_0 : \mu \leq 2$, $H_1 : \mu > 2$. P-value comes from t-distribution with 24 d.f.

$$P(t \geq \frac{5.4 - 2}{8.2/\sqrt{25}}) = p(t \geq 2.07).$$

From Table 4, $.023 < p\text{-value} < .028$, so reject H_0 . Mean weight loss is greater than 2.

4. $H_0 : \mu \leq 72$, $H_1 : \mu > 72$. P-value comes from t-distribution with 23 d.f.

$$P(t \geq \frac{78 - 72}{10/\sqrt{24}}) = 2.94.$$

From Table 4, $.003 < p\text{-value} < .004$ so reject H_0 . Mean score is higher than 72.

5. Expected frequencies under hypothesis of independence are

Age	Male	Female
< 16	17.778	17.222
16-17	20.825	20.175
18-20	30.476	29.524
≥ 21	26.921	26.079

Table of $(O - E)^2/E$ is

Age	Male	Female
< 16	2.934	3.029
16-17	0.484	0.500
18-20	0.201	0.208
≥ 21	2.330	2.406

$\sum(O - E)^2/E = 12.09$. P-value comes from Chisquare distribution with 3 d.f. From Table 8, $0.005 < p\text{-value} < 0.010$, so reject H_0 . Gender and age first smoked are not independent. Note that the greatest discrepancies occurred in < 16 and ≥ 21 age groups. More males than expected started smoking before age 16 and more females than expected started smoking at age 21 or older.