These are the notes for Professor Miller's sections of BUS 350 at Emory University. Professor Miller created these notes and is solely responsible for their content including any errors or omissions. These notes are just that, notes, they are not meant to be a text book. You cannot just read these notes as a substitute for attending class. There are unanswered questions in the notes (these will for the most part be answered in class) and in some places information is missing (missing information will, to the extent possible, be filled in during class.) You are free to use these notes as you wish, one idea might be to use them as a guide during class adding your own comments and notes in the margins. We will roughly follow the notes in order during the class sessions but we might skip some sections or pass over a section and then return to it. In particular, these notes do not necessarily represent what is covered in other sections of BUS 350 and students should not use these notes to replace or supplement materials provided by instructors in those sections.

An Example

Imagine you are the COO at Chipotle Mexican Grill. You oversee all the individual stores. You have data on the revenue and costs of each location and you see that some locations are more profitable than others. The question is why? Knowing this would be useful in a number of ways:

1) To the extent that profitability was driven by factors within the control of the manager, you could coach the managers to take actions that would improve profitability
2) To the extent that profitability was driven by factors outside the control of the manager (e.g., competition, population density, etc.) you could make sure you did not reward or punish managers for things they didn't control
3) You could predict what might be a good location to put a new store

To determine why some locations are more profitable you would have to collect data. You would start with a long list of everything that might be a factor, e.g.,

- Median income in the local area
- Age of the local population
- Minimum wage in the area
- Distance to the closest competing business
- Foot traffic passing by the store
- Ease of parking near the store
- Advertising spend and effectiveness
- Measures of employee and customer satisfaction
- Years of experience of managers and employees, etc.

There would be many items on your list. For each item on the list you would have to define precisely what you were going to measure (i.e., how do you define "local area"? what constitutes a "competing business"? how is satisfaction measured? Etc.) You would not think that all of these factors necessarily were important but you would want to cast a wide net at first. You can always remove items later.

Once you have a list of candidate factors you would need to collect data on them. Here is where you run into a problem. Chipotle has over 1600 stores. It would be a monumental task to collect data on all of them, in fact, it would likely be infeasible due to the cost and amount of time it would take.

We refer to all 1600+ stores as the <u>population</u>. We want to know things about the population values – these will be <u>parameters</u>, but It's not feasible to collect data on the entire population. You would take a <u>sample</u> of stores (a sample is any subset of the population) and you would collect data on the sample. You would then calculate things from the sample (this will be exactly what we mean by a <u>statistic</u> – something calculated from a sample) and you would use the sample values as proxies or stand-ins for the (unknown) population values.

If you understand that example you understand what this course is about. This leads us to state the basic idea of the course.

The Basic Idea

All of statistics essentially comes down to the following scenario:

1. We want to know something about a population. For example, what factors drive profitability? What percentage of our customers were satisfied with their most recent service interaction? How effective is our firm's advertising? Does providing health insurance to workers reduce absenteeism?
2. It is infeasible (too much time, money, or both) to collect data on the entire population.
3. Instead of collecting data on the population we take a sample and calculate values from the sample, these are statistics.
4. We use the statistics to estimate the population values, the parameters. This is estimation.
5. We use techniques from statistical inference (confidence intervals, hypothesis tests, regression etc.) to make further statements about the population values.

So we want to know something about a population. We take a sample and calculate a value from the sample but the sample we take is one of a very large number of possible samples[1]. The value we get for our statistic then is one of a very large number of possible values we could have gotten. This means that the statistic is random – in the parlance of probability we say that statistics are random variables.

Once we talk about probability that statement will make more sense but I think it is the most critical insight in the study of statistics so it bears repeating – statistics are random variables.

Random variables are variables that take on a value randomly and to make sense of a random variable we need to identify its probability distribution – the probability distribution will tell us the probability the random variable takes on each of its possible values. Since statistics are random variables they have distributions and if we know the distribution we know the probability that the statistic takes on each of its possible values which we can then use to make inferences.

To recap:

- We want to know something about a population but it is infeasible to collect data on the population so we collect data on a sample and calculate statistics from the sample.
- We use the statistics to estimate the population values.
- Statistics are random variables, random variables have distributions and if we can figure out the distributions of the statistics we can then make inferences about the population values.

So we need to understand random variables and to do that we need to cover probability so that is where we will begin.

[1] As an example, if there were 1600 Chipotle stores, there would be 2×10^{260} different samples of size 200.

Probability

Probability forms the basis for statistical analysis and also allows us to characterize uncertainty in decision problems and in simulation. Some of what we will cover here is not specifically required for the study of statistics but you will find it useful in other pursuits.

We will discuss three ways of thinking about probability: a priori, empirical or relative frequency and subjective.

A Priori

These are situations where the process lets us calculate the probability. If we roll a fair six sided die, for example, we would say the probability of each of the 6 numbers (1 through 6) resulting would be 1/6. Often these are games of chance: roll of a die, flip a coin, number on a roulette wheel.

Example: there are 38 numbers on a roulette wheel, 0, 00 and the numbers 1 through 36. (This applies to roulette as played in the US, roulette wheels in Europe are different.) The probability any one of these comes up is 1/38 as they are all equally likely (we assume that all games are fair, i.e., not rigged, in this class.)

The 36 numbers on a roulette wheel that are greater than 0 are split between red and black. The 0 and 00 are green. The probability that the number that comes up is black is 18/38, the probability it is green is 2/38, etc.

A priori probabilities are usually very easy to work with but they are not that common in actual problems (other than say gambling problems).

Say you work for a Toyota dealership and you are going to place an order for Camry cars. You need to order by color so you want to know the probability customers will want cars in blue, red, black, etc. Do you think these colors are equally likely? Or is there some other process you could imagine that would allow you to specify the probabilities? Unlikely, but you can collect data that would help you.

Empirical or Relative Frequency

These probabilities are based on data.

Say you wanted to know the probability a random customer wants each color of Camry. You would collect data on cars sold in the past. The probability a car was red would be the number of red cars divided by the total, the probability a car was blue would be the number of blue cars divided by the total, etc. What you would be doing is calculating the relative frequency of each color of car in the data and using those as the probabilities.

Another example, last month a particular store had 530 unique visitors. Two were caught shoplifting. Probability a visitor to this store was caught shoplifting would be 2/530.

Two concerns:

- We are using historical data (cars sold in the past, previous shoplifters, etc.) to determine the probability of something happening in the future (next car sold is a particular color; probability future customer is a shoplifter, etc.) This is reasonable as long as things don't change but sometimes they do. For example, the popularity of car colors changes due to changing tastes, or the prevalence of shoplifting might increase or decrease due to changes in the economy. This can invalidate our approach so we need to consider whether it is an issue or not.
- Our relative frequencies are based on samples when they really should be based on the population. For example, we only used data for cars purchased from our dealership but this is a subset of all potential purchases. As long as we have a reasonable amount of data (i.e., our sample is not that small) this is usually not a problem.

Subjective

These are individual beliefs – we can each have our own.

What is the probability that the S&P 500 ends higher this year than last?

You could go back and look at how many years the market finished higher than the year before and use that to calculate your probability but most of us probably would not do that. The reason has to do with the first concern mentioned above; as the mutual fund companies famously state: "past performance is no indication of future results".

Instead we would consider what we thought the economy was going to do over the remainder of the year, our expectations for Federal Reserve or other policy maker actions, geopolitical events we might imagine, etc. and come up with a subjective probability.

Frequentist Vs Subjective Approach

There is something of a divide between the frequentists and the subjective probability types.

The frequentists argue that all probabilities are objective, that is, based on relative frequency.

The subjective types think they are all just subjective beliefs.

I am more in the camp of the subjectivists.

I would say that all probabilities are subjective (that is, they represent my belief in the likelihood of something happening) but sometimes my subjective belief is informed by a relative frequency or a priori approach.

Some Terminology

An experiment is a process with more than one outcome and which outcome occurs is uncertain. So a customer coming into our dealership and choosing a particular color of car is an experiment.

The sample space (denoted S) is the set of all possible outcomes of an experiment. These should be "finest grain" – i.e., the most basic outcomes we can imagine. In our car example if dark blue and light blue were each choices we would want to identify those two colors separately in sample space rather than calling them both blue.

An event is a collection of one or more items in sample space. We typically use capital letters (e.g., A, B, C) to denote events. Again, if we sold dark blue and light blue cars, A = car sold was blue would be an event made up of two items in sample space, dark blue and light blue.

Experiment: ask n customers, where n > 0, if they prefer our product or the competition's product and count how many prefer ours. Is this an experiment? Why or why not? What is the sample space?

Let A = more than half prefer our product and let B = fewer than half prefer our product. A and B will be events.

Two or more events are mutually exclusive (abbreviated ME) if they do not overlap at all – that is each outcome appears in at most one of the events.

Two or more events are collectively exhaustive (abbreviated CE) if they include all outcomes – that is each outcome appears in at least one of the events.

If we ask 10 customers if they prefer our product or the competition's product, the possible outcomes are 0, 1, 2, …, 10 prefer our product.

A = no more than 2 prefer our product and B = 7 or more prefer our product are ME, but not CE.

A = five or less prefer our product and B = 4 or more prefer our product are CE, but not ME.

A = no one prefers our product and B = at least one person prefers our product are both ME and CE.

The set S is made up of the outcomes 0, 1, 2, 3, …, 10 prefer our product and these outcomes are both ME and CE. Since the outcomes in sample space are "finest grained" and since each appears only once, they always will be ME and CE.

Typically we are interested in the probabilities of events.

For an event A, P(A) will denote the probability that the event A occurs.

Rules of Probability

Let S be the sample space of an experiment.

Let O_1, O_2, …, O_k be the outcomes in S.

Let A be a collection of one or more of the outcomes in S (so A is an event).

For any event A, $0 \le P(A) \le 1$

If P(A) = 0, then ?

If P(A) = 1, then ?

P(S) = ?

P(A) will be the sum of the probabilities of the outcomes in A. More formally, $P(A) = \sum_{O_i \in A} P(O_i)$

So in the experiment where we ask 10 people if they prefer our product, if the event A = more than half prefer our product, we could calculate P(A) by summing the probability that 6, 7, 8, 9 and 10 people prefer our product.

Complements

A' (read A prime) is the complement of A (you will also see the notation $\bar{A}$). The complement of A is everything not in A. So if A = more than half prefer our product then A' = half or less prefer our product

A and A' will be ME and CE

We have to have P(A) + P(A')= 1 so P(A) = 1 – P(A'). This is the rule of complements.

We can visualize this using a Venn diagram – we will do this in class.

Sometimes it is easier to calculate things using complements.

Say we ask 10 people if they prefer our product and let A = at least one prefers our product.

Now A is the event 1,2,3, …, 10 people prefer our product.

The complement of A is that no one prefers our product.

It is much easier to calculate P(A') and then calculate P(A) = 1- P(A') then to calculate and sum the probabilities of the 10 outcomes in A to get P(A) directly.

Probability of A or B

If A and B are ME then

P(A or B) = P(A) + P(B)

If A and B are not mutually exclusive, then P(A or B) = P(A) + P(B) - ?

Again, a Venn diagram can help us visualize this – we will do this in class.

Solving Probability Problems

The following steps are the **sure-fire method for solving probability problems**:

1) Identify the events of interest
2) Identify what you know in terms of the events of interest
3) Identify what you want to know in terms of the events of interest
4) Use the rules of probability and what you know (in terms of the events of interest) to find what you want to know

Joint, Marginal and Conditional Probabilities

We will often represent probabilities using contingency tables and often we start with raw data from which we calculate probabilities. The following data are from the Ireland Central Statistics Office:

Persons Aged 15 and Over (thousands)	Male	Female
At work	1014.7	844.8
Unemployed	191.4	103.4
Student	204.2	196.8
Home duties	9.6	460.7
Retired	257.9	158.8
Other	84.8	68.4

The items in the cells represent joint events, that is there are two things going on. The intersection of the Male column and the Unemployed row tell us that there are 191.4 thousand people who are male and unemployed; similarly there are 158.8 thousand retired females, etc. We can add a column and a row that represent the totals as in the next table.

Persons Aged 15 and Over (thousands)	Male	Female	Totals
At work	1014.7	844.8	1859.5
Unemployed	191.4	103.4	294.8
Student	204.2	196.8	401
Home duties	9.6	460.7	470.3
Retired	257.9	158.8	416.7
Other	84.8	68.4	153.2
Totals	1762.6	1832.9	3595.5

The numbers in this new row and column for the totals represent individual or simple events. The total in the Student rows tells us that there are 401 thousand students, the total in the Female column tells us there are 1832.9 thousand females, etc.

If we want to know, for example, the likelihood someone in the population was an unemployed female we would take the number of unemployed females and divide it by the total number in the population. We would get 103.4/3595.5 = .029 so 2.9% of the population are unemployed females.

If we chose a random person, the probability the person we chose was an unemployed female would be .029. (Probabilities should always be decimals, i.e., numbers between 0 and 1 but for convenience we will often refer to them as percentages, e.g., 2.9% chance of choosing a person who was an unemployed female.)

If we divide each element in the table by 3595.5 (the total) we get the following table where the capital letters have been added to identify events:

Persons Aged 15 and Over (thousands)	Male (A)	Female (B)	Totals
At work (C)	0.282	0.235	0.517
Unemployed (D)	0.053	0.029	0.082
Student (E)	0.057	0.055	0.112
Home duties (F)	0.003	0.128	0.131
Retired (G)	0.072	0.044	0.116
Other (H)	0.024	0.019	0.043
Totals	0.490	0.510	1

The values in the cells in the matrix are joint probabilities – the probability of two events happening.

Values in the row and column totals (in the "margins") are called marginal probabilities; they are the probabilities of individual events.

Examples of joint probabilities:

P(A and G) =

P(B and E) =

Note that the order of the events does not matter: P(B and E) = P(E and B)

Examples of marginal probabilities:

P(B) =

P(F) =

Notice where the marginal probabilities come from, for example, P(B) was calculated by adding

P(B and C) + P(B and D) + P(B and E) + P(B and F) + P(B and G) + P(B and H).

This works because C, D, E, F, G and H are ME and CE.

In general for an event A and a set of events B_1, B_2, …, B_k that are ME and CE,

P(A) = P(A and B_1) + P(A and B_2) + … + P(A and B_k)

This is known as the law of total probability.

Again, a Venn diagram can help us visualize this – we will do this in class.

We can calculate pretty much any probability we want by adding up the appropriate cells.

For example, people who are working or unemployed are considered to be in the labor force (the other categories, students, retired, etc., are not in the labor force). We can calculate:

P(someone in labor force) = P(C or D) = .282 + .235 + .053 + .029 = .599

The percentage of people in the labor force is known as the labor force participation rate.

Now what if I told you I chose a random person from the population and it turned out the person was unemployed. What would you calculate as the probability the person is female?

This will be a conditional probability.

We want the probability of one event ("female") knowing that another event ("unemployed") occurs. The event that we know occurs is referred to as the conditioning event.

CAUTION: this is not the same as the joint event female and unemployed! We calculated this earlier as .029 but now we know that one event (unemployed) occurs. This event (the conditioning event) is not random. We want the probability of the other event (female) knowing unemployed has occurred.

The notation for this conditional probability is P(B|D) read "probability of B given D"

This can be calculated in various ways.

We can go back to the original data.

If we know the person is unemployed we know they are one of the 294.8 thousand unemployed people. Of these, 103.4 thousand are female so $P(B|D) = \frac{103.4}{294.8} = .351$.

We also have a formula we can use. In general

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Using this formula,

$$P(B|D) = \frac{P(B \text{ and } D)}{P(D)} = \frac{.029}{.082} = .351$$

We often will have two (or maybe more) valid ways of calculating something. It goes without saying that if the methods are both valid then they need to give us the same answer.

Let's think for a minute about what is going on here. In conditional probability we know that one event occurs. Knowing that event occurs changes the sample space and what we are doing is calculating the new probability in this new sample space. In the previous example once we knew the person was unemployed we knew the sample space was the single row represented by event D (unemployed) and we calculated the probability of the person being female knowing they were unemployed as the percentage of unemployed people that were female.

As another example, say we chose a random person and we know the person is male. The sample space now consists of only the Male column:

Persons Aged 15 and Over (thousands)	Male (A)
At work (C)	0.282
Unemployed (D)	0.053
Student (E)	0.057
Home duties (F)	0.003
Retired (G)	0.072
Other (H)	0.024
Totals	0.490

The problem is that these probabilities do not add up to one and they need to. We know that the person is either working, unemployed, a student, etc. so conditional on the person we chose being male the sum of the probabilities of the categories the person can be in still has to add to one. The solution is to divide each of the probabilities in the Male column by the probability the person is male (.490). This will scale all the probabilities up so they now add to 1:

Persons Aged 15 and Over (thousands)	Male
At work	0.576
Unemployed	0.109
Student	0.116
Home duties (F)	0.005
Retired (G)	0.146
Other	0.048
Totals	1

We could have gotten the same probabilities by taking the original data and dividing the values in the male column by the total number of males. You should try this to verify that it works and also think about why this works.

Some notes:

1) **Conditional probabilities have to follow all the rules of probability**.
2) We can only add probabilities all conditional on the same event (in this case conditional on the person being male). **It does not make sense to add up probabilities conditional on different events**.
3) The denominator in the formula for conditional probability is the probability of the conditioning event. You should now see that the reason this is the denominator is to appropriately scale up the probability of the joint event.

Independence of Events

Did knowing the person was unemployed change our belief about whether they were male or female?

Before we knew the person was unemployed, what was the likelihood they were female?

How about after we knew the person was unemployed?

If knowing an event B occurs does not change your belief in the likelihood that the event A occurs then we say that A and B are independent.

A more formal way of stating this is that **A and B are independent if and only if P(A|B) = P(A)**.

In order to show that two events are independent **you have to check the probabilities**, that is, you have to calculate whether knowing one event occurs changed your belief in the other. Sometimes you may think that you can come up with a plausible argument for why two events are independent or not but this does not prove anything.

CAUTION: a very common mistake that students make is to think they can prove that two events are independent without calculating the values – **you cannot do this**.

Calculating Joint Probabilities – Multiplication Rules

We showed how to calculate joint probabilities from a data table like the one above but what about a formula? Look again at the formula for conditional probability:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

If we multiply both sides by P(B) we get that P(A and B) = P(A|B) * P(B) which is the general multiplication rule.

CAUTION: another very common mistake is to think that P(A and B) = P(A)P(B). This is only the case **if A and B are independent**. Can you see why?

Probability Trees

Your book calls these decision trees but I prefer to refer to them as probability trees (decision trees typically have both decisions and random events while probability trees only have random events).

Probability trees are another way to represent random events.

Example: a survey taken for the Urban Land Institute[2] reports on, among other things, marital and home ownership status of survey respondents.

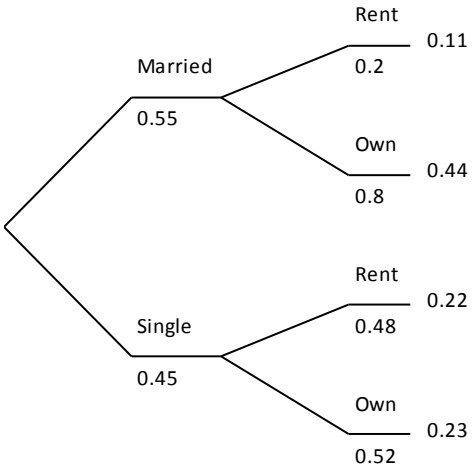
Let M be the event "survey respondent is married" and let R be the event "survey respondent rents their principle residence". Then M' will be the event "survey respondent is not married" (i.e., single) and R' will be the event "survey respondent does not rent" (i.e., owns their home).

The survey reports that 20% of married people rent their principle residence. What is this in terms of the events M and R? (Hint: given that someone is married there is a 20% chance they are a renter.)

The survey also reported that 33% of renters are married. What is this in terms of the events M and R? How does this differ from the 20% in the previous statement?

The report also says that 55% of all the survey respondents were married and 48% of single respondents rent their principle residence.

We can represent these data in a probability tree:



[2] "Americans' Views on their Communities, Housing, and Transportation", March 2013, Belden Russonello Strategists, LLC. The report has been posted to Blackboard.

Some notes:

1) The probabilities at the ends of the paths in the tree are the probabilities of the joint events on the branches leading to that end point (e.g., .11 is the probability that a survey respondent is married and rents).
2) The end of path probabilities are calculated by taking the product of the probabilities on the branches leading to the end point (e.g., .55*.2 = .11). What are the calculations in terms of the events M and R?
3) Probabilities in probability trees (and in decision trees) **are always conditional on where you are in the tree**.

Now, what about the 33% of renters who are married? What is this in terms of our events? Can we calculate this from the tree?

Try applying the sure-fire method of solving probability problems.

1) We already have our events of interest (M and R, and by extension M' and R')
2) What do we know in terms of these events? (These are the probabilities in the tree.)
3) What do we want to know in terms of these events? (This will be the probability someone is married given they are a renter.)
4) Use the definition of conditional probability to solve for the probability you want.

We will do this in class.

If you work through this you will see that it can be calculated very easily from the tree. It can also be calculated using Bayes' Theorem. We will not cover Bayes' Theorem directly as probability trees accomplish the same thing but it is covered in your book and is useful in more advanced study.

Random Variables and Probability Distributions

Random variables are variables that take on a value as a result of an experiment.

The outcome of an experiment might be a number (e.g., quarterly profit of Google) in which case we just let a variable take on the value of the outcome (e.g., let X = Google's profit next quarter.)

Sometimes the outcome is not directly a number (e.g., test an item, the item will either fail or pass) in which case we create a variable that takes on a value depending on the outcome (e.g., let the variable X be 1 if the item fails and 0 if it passes – now X will be a random variable.)

Random variables can be discrete or continuous.

Let X = the number of survey respondents out of 10 that prefer our product.

X is a discrete random variable as it only takes on specific discrete values.

Let Y = change in value of the DJIA the next day the market is open.

Y will be continuous as it takes on any value in a range. (I realize that when they report the market results on the news they will say the Dow was up by 43 points or down by 87 points but they are rounding off or discretizing a continuous quantity.)

We use capital letters for random variables and the corresponding lowercase letter to represent the values it takes on. For a discrete random variable, X, we use the notation P(X=x) to mean the probability that the random variable X takes on the value x.

Random variables have distributions.

The distribution of a random variable is a table, graph or formula that specifies the probability that the random variable takes on each of its possible values

Random variables also have expected values.

The expected value of a random variable (also called the mean of the random variable) is the average value that the random variable takes on.

More formally the expected value of a random variable is the weighted average of the values that the random variable takes on where the weights are the probabilities that the random variable takes on that value.

Even more formally, we define the expected value, written E(X) or μ, for a discrete random variable as

$E(X) = \mu = \sum xP(X=x)$ where the sum is taken over all values x that the random variable takes on.

Simple example: say at a roulette table you bet \$1 that the number that comes up is black. There are two possible outcomes, you win or you lose. If you win you get your \$1 plus \$1 in winnings so you are

plus $1, if you lose you lose your $1. Let X = 1 if black comes up and let X = -1 if black does not come up. Recall that 18 of the 38 numbers are black and 20 are not (18 are red and 2 are green).

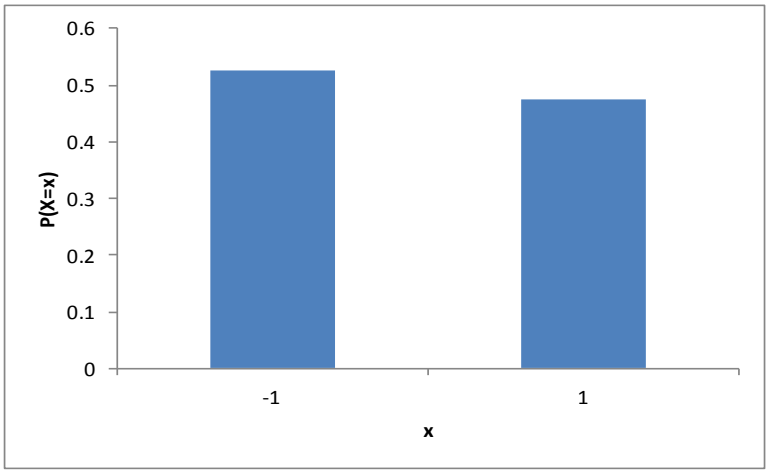
So X represents your return (gains or losses) from one play of the game.

X is a random variable as it takes on its value based on the outcome of an experiment (spin of the roulette wheel.)

The distribution of X can be represented as a table:

x	P(X=x)
-1	20/38
1	18/38

The distribution can also be represented as a graph:



The expected value of X will be E(X) = μ = (-$1)(20/38) + ($1)(18/38) = -2/38 = -$.05263 or -5.263¢

The variable X represented the amount won on one play of the roulette wheel. E(X) will be the average value of X over many plays so this is saying that on average you will lose just over a nickel, or just over 5% of the amount you are betting, every time you play. (Do you see why you will not win in the long run at the roulette table?)

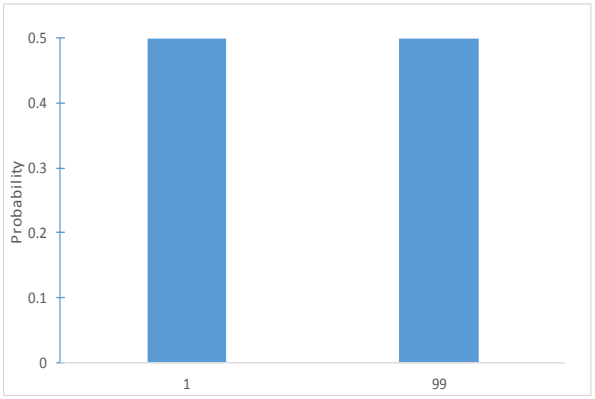
The expected value will be the center of the distribution above. If you thought of the two bars above as weights, the line would balance at the mean or at -.05263 which is slightly to the left of the center tick mark (which represents zero). In general this is another way of thinking about the expected value for a discrete random variable – if you thought of the bars as weights the expected value is where the line would balance.

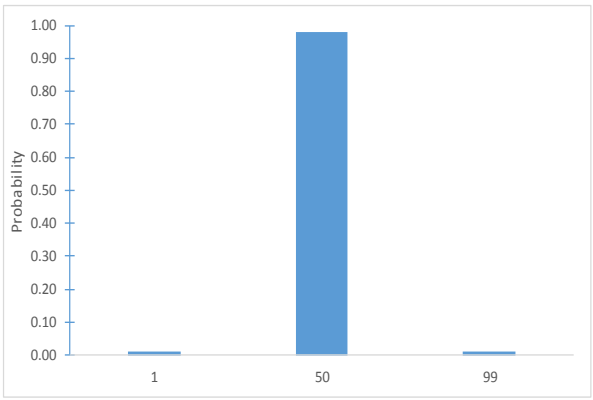
We can also talk about measures of the spread or variation in the distribution.

One we might think of is the range – the difference between the largest value taken on by the random variable and the smallest value.

The problem with this is that it does not consider the likelihood the random variable takes on these values. For example, say you were considering buying an item at auction. It there was an equally likely chance it would sell for $1 or $99 the range would be $98. If, on the other hand, there was a .01 chance it would sell for $1 and a .01 chance it would sell for $99 and a .98 chance it would sell for $50, the range would still be $98.

These two distributions are graphed below.





In both cases the center of the distribution (the expected value) is 50 but the first one is always relatively far from the center while the second one is at the center almost all the time and rarely far from the center. This would lead us to say that the first one has more variation (or is more dispersed) but the range is the same in both cases. Clearly the range is not giving us a good measure of the variation in this case.

One thought might be to look at how far, on average, the values are from the center. We could calculate $\sum[x - E(X)]P(X = x)$ where the sum is taken over all values x that the random variable X takes on.

This will be the weighted average distance to the expected value (notice that $[x - E(X)]$ is the distance to the expected value for a given value x) where the weights are the probabilities.

There is a problem with this. Can you see what it is? (Hint: this formula will always result in the same value.)

To avoid the problem inherent in the previous formula we take the weighted average squared distance to the mean where again the weights are the probabilities. This is called the variance, denoted σ^2, and defined for a discrete random variable:

$\sigma^2 = \sum[x - E(X)]^2P(X = x)$ where again the sum is taken over all values x that the random variable takes on.

Notice that the formula for the variance is calculating $E[X - E(X)]^2$, that is, it is the weighted average squared distance to the mean where again the weights are the probabilities.

For the random variable X where X = -1 with probability 20/38 and X = 1 with probability 18/38,

$\sigma^2 = [-1 - (-.05263)]^2*20/38 + [1 - (-.05263)]^2*18/38 = .997$

The variance is a commonly used measure of the spread or variation in a distribution but it has a quirky feature – the units of the variance are the square of the units of the random variable. If the random variable is measured in dollars (as in the roulette example), the variance will be measured in square dollars.

The standard deviation is the square root of the variance and it will be measured in the same units as the random variable.

The symbol σ is used for the standard deviation and it is defined $\sigma = \sqrt{\sigma^2}$

The file DiscreteDistributions.xlsx, which has been posted to Blackboard, has examples of discrete probability distributions.

We will look at the worksheet Empirical Distribution in this spreadsheet for an example of an empirical distribution (i.e., one based on data) and for calculations of the mean, variance and standard deviation.

There are a number of standard random variables that come up often in practice. We will discuss five of these, the Bernoulli, the binomial, the Poisson, the exponential and the normal.

Bernoulli Random Variable

A Bernoulli experiment is one with two possible outcomes commonly referred to as success and failure.

We define a random variable to take on the value 1 if the outcome is success and 0 otherwise. We use π to denote the probability that the outcome is success.

Say a consulting firm submits a proposal that will either be accepted or rejected. If the proposal is accepted we will call that a success and if it isn't we will call that a failure.

Define the random variable X to be 1 if the outcome is success and 0 if it is failure.

So $P(X = 1) = \pi$ and $P(X = 0) = 1-\pi$.

In this case, X is a Bernoulli random variable. The expected value of X will be $(1)(\pi) + (0)(1-\pi) = \pi$.

The variance of X will be $(0 - \pi)^2(1-\pi) + (1-\pi)^2\pi = \pi^2 - \pi^3 + \pi + \pi^3 - 2\pi^2 = \pi - \pi^2 = \pi(1-\pi)$

Let's say this firm thinks that $\pi = .3$, that is they have a 30% chance of winning the job.

Then $E[X] = .3$ and $Var[X] = (.3)(.7) = .21$

It is more interesting if we assume that the firm has submitted a number of proposals and they have a probability of .3 of winning each one, independent of the others.

This gives rise to a binomial random variable.

Binomial Random Variable

Say we have n replications of a Bernoulli experiment (we call each replication a trial) with probability π of success on each.

Further assume that the trials are independent of each other. (Independence means that knowing something about the outcome of one trial does not change our belief in the likelihood of the outcomes on another trial.)

Let X = the total number of successes on all the n trials.

X will be a binomial random variable.

We have a formula for the distribution of X:

$$\mathrm{P}(\mathrm{X}=\mathrm{x}) = \binom{n}{x}\pi^{x}(1-\pi)^{n-x}$$

$\binom{n}{x}$ is read "n choose x" and is equal to $\frac{n!}{x!(n-x)!}$ where n! is n factorial. Your book uses the notation ${}_{n}C_{x}$ to denote n choose x.

In Excel the formula is =binom.dist(x,n,p,0)

We don't usually graph the binomial but an example is in the Binomial worksheet in DiscreteDistributions.xlsx.

Say the consulting firm submits 100 proposals with π = .3. What is the probability they are successful on exactly 30 of the proposals?

We can plug the values into the formula above to get:

$$\mathrm{P}(\mathrm{X}=30) = \binom{100}{30}.3^{30}(1-.3)^{100-30} = \frac{100!}{30!\,70!}.3^{30}(.7)^{70} = .0868$$

Your calculator will likely not calculate 100! as it is too large (100! is approximately 9.3 X 10^{157}). You may have a key that calculates the choose term, otherwise you would have to do the appropriate cancellation first. When the value of n and x are this large it is much easier to use Excel. In Excel:

P(X=30) = binom.dist(30,100,.3,0) = .0868

What if we want the probability that they are successful on 30 or fewer of the proposals? We again use the formula but this would be highly impractical (we would have to calculate the probability of 0, 1, 2, …, 30 successes which would require us to apply the formula as above 31 times). In Excel, we calculate:

P(X≤30) = binom.dist(30,100,.3,1) = .549

Note the change to the last argument from 0 to 1 – this last argument is referred to as the "cumulative" argument and it tells Excel to calculate the probability that X is less than or equal to x.

What about P(X > 20)? We can calculate this as follows:

P(X > 20) = 1 – P(X ≤ 20) = 1 – binomdist(20,100,.3,1) = 1 - .0165 = .9835

What about the expected value of X? Well if the firm submits 100 proposals and the probability is .3 that they win each one, on average how many will they win?

Imagine they submitted 100 proposals each week, each with .3 chance of winning. On average how many would they win per week?

Can you write down a formula in terms of n and π for this? In general this will be the expected value of a binomial random variable.

There is another way to derive the expected value of a binomial. The binomial random variable is the sum of n Bernoulli random variables, one for each trial. It turns out that the expected value of a sum is the sum of the expected values. How can we use this to derive an expression for the expected value of a binomial random variable?

The variance of a binomial random variable will be $n\pi(1-\pi)$. What was the variance of the Bernoulli random variable? Do you see a relationship here? Careful, though, the variance of a sum of random variables is the sum of the variances of the random variables **only if the random variables are independent of each other**. Do we have independence in this case? How do you know?

If we have a binomial random variable and we know n and π we can write down the distribution, expected value and variance.

We refer to n and π as the <u>parameters</u> of the distribution.

If you know the name of the distribution and the values of the parameters you know everything you need to know to calculate any probability you want.

<u>Poisson Random Variable</u>

The Poisson is another standard random variable used frequently in operations especially in queuing problems.

It describes the number of arrivals to a process (say calls to a call center, jobs to a manufacturing work center, customers entering a fast food restaurant) in unit time.

It is also used to describe the number of items in some area of space, say the number of defects in 1 square meter of carpet or the number of violent crimes in a 10 square mile policing district.

Just like the binomial, the Poisson has a set of conditions that need to be met. (For the binomial, we needed n independent trials with success of π on each and then X defined as the total number of successes on the n trials was a binomial random variable.)

The conditions for the Poisson are not quite as straightforward but we will state them for completeness.

For a Poisson we need three conditions:

- In each very small time interval we have either 0 or 1 arrivals.
- The number of arrivals in a time interval is independent of what happened before
- The probability of a particular number of arrivals in a given time interval depends only on the length of the interval.

So let's think about the call center example.

Say you are managing a large call center.

There are people all over the country, maybe the world, that at any instant of time either call or don't call your call center.

If we make the time interval small enough we will get either zero or one calls in a given time interval (that is, if we measure it precisely enough no two calls will arrive at exactly the same instant.) This meets the first condition.

Also, since these people calling you are all doing so independently (that is, these people are not coordinating their decisions to call) we would not expect past history to matter so the second condition will be met.

The third condition essentially says that the decisions these people are making to call or not call don't depend on the time of day.

This is more problematic as we might very well have busy (or not busy) periods when people are more (or less) likely to call.

That is OK as all we need is for the busy periods (or the not busy periods) to be consistent.

So if we know that Monday mornings are busy periods what we require is that during Monday mornings the probability of a particular number of arrivals to only depend on the length of the interval. Again since these are all independent decisions being made this is a reasonable condition.

Now let X = the total number arrivals in a unit time interval and

Let λ = the average number of arrivals in a unit time interval

(By unit time interval we mean a specific time interval that we pick, say ½ hour or 1 day, 1 year, etc.)

If the three conditions above are met then X is a Poisson random variable. The distribution of X will be:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad X = 0, 1, 2, 3, \ldots$$

This distribution has one parameter, λ.

The expected value of X will be λ (if you look at how λ is defined that has to be the case) and also the variance of X will be λ (this is just the way it works out, the variance and expected value are not usually the same).

We can use the formula directly, as with the binomial formula, or the Excel function poisson.dist can be used to calculate Poisson probabilities.

Say we have a call center that on average receives a call every 5 minutes. Notice how λ was defined – it is the average number per time period so it needs to be a rate. In this case we would say that $\lambda = 12$ calls per hour as if we get a call every 5 minutes then we get 60/5 = 12 per hour.

Now assume we wanted to know the probability of exactly 10 calls in an hour. This would be

$$P(X = 10) = \frac{e^{-12}12^{10}}{10!} = .1048$$

And in Excel, we would calculate =poisson.dist(10,12,0) = .1048 or just over 10%.

If we wanted the probability that we received 10 or fewer calls in an hour, we would calculate in Excel:

=poisson.dist(10,12,1) = .3473.

Again, this would be much more difficult with the formula above; we would have to use the formula 11 times to calculate the probability of 0 through 10 calls and then sum these values.

It might make sense to look at a graph of this distribution. The Poisson worksheet in DiscreteDistributions.xlsx has an example.

The Bernoulli, binomial and Poisson are all examples of discrete distributions.

There are many other discrete distributions and there are also many continuous distributions.

With a continuous distribution we can no longer talk about the probability that the random variable X takes on a specific value x. Continuous random variables take on any value in a range and there are too many possible values of the random variable to assign a non-zero probability to each value and have the probabilities all add up to one. The probability that a continuous random variable takes on any exact value is zero. This makes sense when you consider that when we say the random variable takes on an exact value, say 4.875, what we mean is that it takes on that value to infinite precision, i.e., it takes on the value 4.87500000000000000000000 … where the zeroes continue indefinitely. The probability of this happening is zero.

For a continuous random variable, rather than talking about the probability it takes on some specific value, we calculate the probability that the random variable takes on a value in a range.

Probability distributions for continuous random variables will be represented for the most part by continuous curves and the area under the curve between two points will give us the probability the

random variable takes on a value between those two points. The total area under the curve will have to equal 1, which is analogous to the probabilities all adding up to 1 for a discrete random variable.

The ideas of expected value and variance for discrete random variables carry over directly to continuous random variables, the only difference is that instead of a summation sign in the formula we would have an integral.

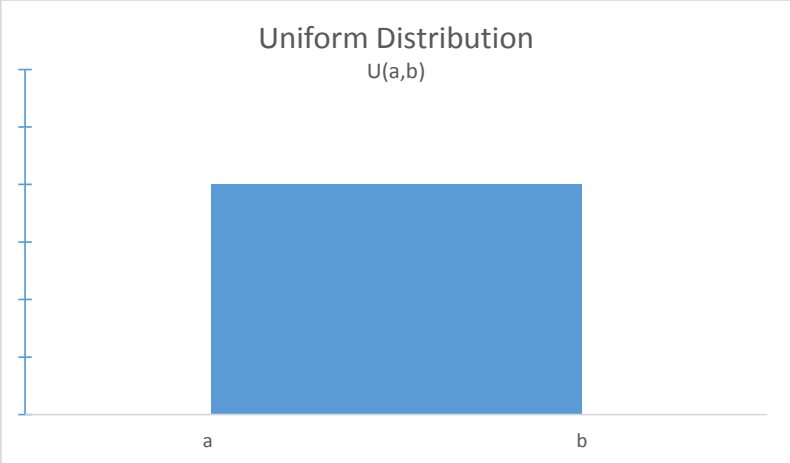
We will discuss three continuous distributions, the uniform, exponential and normal.

Uniform Distribution

A uniformly distributed random variable is equally likely to take on any value in a range from a to b where a and b are two real numbers.

a and b are the parameters of the distribution and the notation U(a,b) is used to represent the distribution where values are filled in for a and b. For example, if X is a random variable equally likely to take on any value between 0 and 100, we would write X ~ U(0,100) where the tilde symbol, ~ , stands for "is distributed as".

We can represent a U(a,b) distribution graphically:



We can also use a formula to represent this distribution:

$$f(X) = \frac{1}{b-a} \; for \; a \leq X \; \leq b \; and \; 0 \; otherwise$$

The distribution of a continuous random variable is sometimes referred to as a density function.

The mean of a U(a,b) random variable is $\mu = \frac{a+b}{2}$ which is of course the midpoint between a and b.

The variance of a U(a,b) random variable is $\sigma^2 = \frac{(b-a)^2}{12}$ and the standard deviation is the square root of this.

It is easy to calculate probabilities for a uniform. For example, if X ~ U(0,100), the probability that X takes on a value between 25 and 50 will be the area under the curve between 25 and 50. The curve is a straight line at a height of 1/100 so this will be the area of a rectangle that is 25 wide and 1/100 high. This area is (25)(1/100) = .25 so there is a 25% chance the random variable takes on a value between 25 and 50.

The uniform is instrumental in choosing random numbers either in sampling, simulation or other uses. Specifically a U(0,1) random variable is used and there is an excel function =rand() that generates random numbers[3] that come from this distribution. Notice that the rand function has no arguments (there is nothing between the parentheses) and the function also will recalculate every time the spreadsheet recalculates.

Exponential Distribution

The exponential distribution is closely related to the Poisson and is also used a lot in operations.

Say we have a Poisson process. This describes the number of arrivals in a time interval.

The number of arrivals in a given time interval varies (that is, one day we might get 180 customers at our store, the next day we might get 195, etc.) so this is a random variable.

Just like the total will vary, the time between two successive arrivals will also vary.

That is we might get a customer and then wait 12 minutes before the next customer arrives, then another customer after only 5 more minutes, etc.

So the time between arrivals is also a random variable.

If the arrival process is Poisson then the time between arrivals has an exponential distribution.

The exponential distribution is a continuous distribution as time is a continuous quantity.

The exponential is usually represented by a function:

$$f(X) = \lambda e^{-\lambda x} \; for \; X \geq 0$$

λ is the only parameter of an exponential distribution and it is the mean number of arrivals per unit time interval. *e* is a mathematical constant similar to π. *e* is the base of the natural logarithms and it is approximately 2.71828.

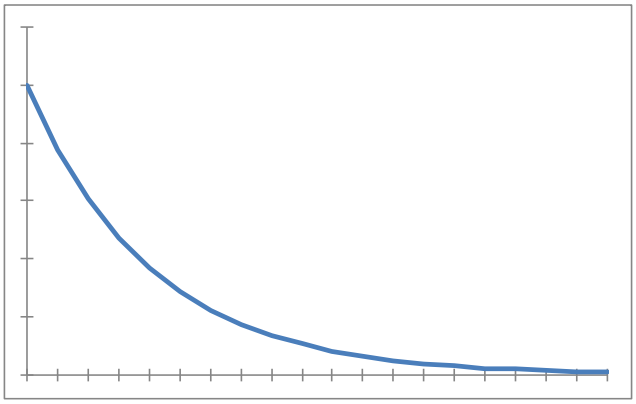
If X has an exponential distribution with parameter λ we write X ~ exp(λ).

Notice that just like the Poisson, λ has to be a rate, that is arrivals per unit time.

[3] These random numbers are generated by an algorithm so are technically pseudorandom numbers (eventually the values will repeat) but we will ignore this distinction.

Since λ is the mean number of arrivals per unit time, the mean time between arrivals (the mean of the exponential random variable), will be $\mu = \frac{1}{\lambda}$. The variance of the exponential will be $\sigma^2 = \frac{1}{\lambda^2}$

A graph that shows the general shape of the exponential distribution is shown below.



It may appear that the curve hits the X axis but in fact the curve is asymptotic to the axis (that is, it keeps getting closer and closer but never hits it.)

Calculating probabilities for an exponential is a little harder than for a uniform. We still want to calculate the area under the curve but to do this we need to integrate the function to find the area under the curve.

We can solve for the cumulative distribution function and use that to find probabilities.

The cumulative tells us the probability that X ≤ x. To find the function we integrate the exponential distribution from 0 to x:

$$P(X \leq x) = \int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda t}|_0^x = -e^{-\lambda x} - -e^0 = 1 - e^{-\lambda x}$$

So we can calculate the probability that the random variable takes on any value less than or equal to x. What if we want to calculate the probability that the random variable takes on a value between say x_1 and x_2? We can do this using the same cumulative function and the following relationship:

$$P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1)$$

That is, the area under the curve between x_1 and x_2 is the area from 0 to x_2 minus the area from 0 to x_1.

As an example, say we again had a call to a call center on average every 5 minutes and we wanted to know the probability that there were less than 6 minutes between calls. So again we would have λ = 12. Now we have to be careful here. λ is measured in calls per hour so we have to put the value of x (6 minutes in this case) into hours also. Our units have to be consistent. So what we want is P(X ≤ .1) because 6 minutes is .1 hours. We would now calculate

$$P(X \leq .1) = 1 - e^{-12*.1} = 1 - e^{-1.2} = .6988$$

and in Excel this would be = expon.dist(.1,12,1) = .6988 or just about 70%.

Two notes:

1) If the last argument (the cumulative argument) in expon.dist is set equal to 0, the formula will calculate the height of the curve at the value of x. This is not usually of any value so this is not commonly done.
2) The units have to be consistent but they are somewhat arbitrary. We could have put everything in terms of minutes (i.e., λ = .2 calls per minute and x = 6 minutes) and we would get the same answer. Do you see why this is? (Hint: does the value of λx depend on the time units, as long as we are consistent?)

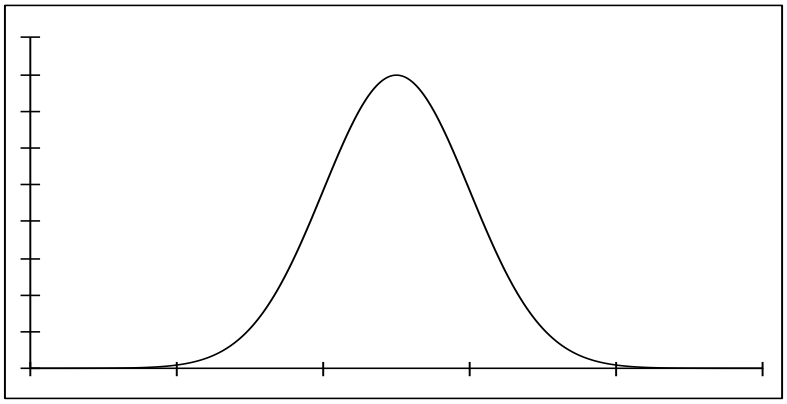
The spreadsheet ContinuousDistributions.xlsx has been posted to Blackboard and the Exponential worksheet in this spreadsheet gives an example of an exponential distribution. The worksheet shows both the probability distribution and the cumulative distribution.

Normal Distribution

The normal distribution is the standard bell curve that I am sure you are familiar with. The normal is probably the most used distribution in statistics.

There is a formula for the normal distribution and it is in your book. We will not do much with this formula so you do not need to know it. You do need to know what a normal distribution is and how to work with one.

A graphical representation of a normal random variable is shown below.



Some notes about the normal:

1) The normal has two parameters, the mean μ and the standard deviation σ. We need to know both of these to uniquely define the distribution.
2) The distribution will be centered at the mean, that is, the high point of the peak will be above μ.
3) The mean will also be the median (this is a feature of any symmetric distribution, not just the normal).

4) As with the exponential, the tails are asymptotic to the axes, they only appear to hit the axes because of the thickness of the line.
5) The standard deviation determines how spread out the distribution is.
6) Since the area under the curve has to equal 1, if the standard deviation is greater (so the curve is more spread out) then the peak of the curve has to be lower. Similarly, the smaller the standard deviation the higher and narrower the peak will be.

If X has a normal distribution with mean μ and standard deviation σ we write $X \sim N(\mu,\sigma)$

In order to calculate probabilities for the normal distribution we need to integrate the distribution function. Unfortunately there is no closed form solution to this integral. (This is why it is not terribly useful to know the formula.) That means that we cannot find the cumulative distribution function like we could with the exponential distribution.

We can, however, perform the integration numerically. That still presents a problem in that the function is different for every different combination of μ and σ (the two parameters).

We can use a trick to get around this. It turns out that we can take any statement about a random variable $X \sim N(\mu,\sigma)$ and transform it into a statement about a random variable $Z \sim N(0,1)$. Z is called a standard normal random variable (your book refers to is as a "standardized normal" but this is not common) and it is normal with a mean of 0 and a standard deviation of 1. We have tables that have been pre-calculated for the standard normal random variable (not surprisingly called Z tables) and we can then use these tables to get the probabilities we want. This table is on the back inside cover of your textbook.

We define Z as follows:

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

We are subtracting μ from X and then dividing by σ which gives us a random variable that has mean 0 and standard deviation 1.

So why does this work? Well first, if we take a random variable X and subtract a constant from it (and μ is of course just a number) we are just shifting all the values over by that constant. If X was normal before it will still be normal but the mean will now be 0. (Since all the values are now μ less than they were before it stands to reason that the mean will now be μ less but $\mu - \mu = 0$.)

It is not as intuitive but it turns out that dividing by a constant σ divides the variance by the square of that constant.

Let X have mean μ and standard deviation σ. We already saw that $X - \mu$ will have mean 0 so $Z = \frac{X-\mu}{\sigma}$ will also have mean 0. (The mean will actually be 0 divided by σ which is of course still 0.) Recall that the variance is defined as the average squared distance to the mean which is $E[X - \mu]^2$.

We have $Var(Z) = E\left[\frac{X-\mu}{\sigma} - 0\right]^2 = E\left[\frac{(X-\mu)^2}{\sigma^2}\right] = \frac{1}{\sigma^2}E[X-\mu]^2 = \frac{1}{\sigma^2}Var(X) = \frac{1}{\sigma^2}\sigma^2 = 1$

and since the variance is 1 then the standard deviation (the square root of the variance) is also 1.

Hopefully all of that is straight forward. The only part that you may question is why we can take the σ^2 term out of the expected value calculation. We can do this as it is just a constant. The expected value is defined as either a summation (for a discrete random variable) or an integral (for a continuous random variable), in either case we can take constants out of summations or integrals.

So we have Z ~ N(0,1) and we have tables that give us probabilities for Z. How do we use these?

Say X ~ N(5, 20) and we want to know P(X < 10). We do the following:

$$P(X < 10) = P\left(\frac{X-\mu}{\sigma} < \frac{10-5}{20}\right) = P(Z < .25)$$

Notice in the second step that we are taking the inequality X < 10 and doing the same thing to both sides; we are subtracting the mean from both sides and then dividing both sides by the standard deviation. We just do it in terms of symbols on the left side of the inequality X < 10 and in terms of the numbers on the right side of the inequality. Since we are doing the same things to both sides of the inequality we are not changing the nature of the inequality.

So we took a statement about X and turned it into a statement about Z. We can then look in the table to find that P(Z < .25) = .5987 which will be the probability X is less than 10 which is what we wanted to find in the first place.

Examples of normal distributions are in the Normal worksheet in ContinuousDistributions.xlsx on Blackboard.

To calculate probabilities in Excel just like with expon.dist we need to use the cumulative function which means we want to set the last argument in norm.dist equal to 1.

P(X ≤ x) = norm.dist(x,μ,σ,1)

P(X ≥ x) = 1 – norm.dist(x,μ,σ,1)

Say we are filling 1 liter (1000 ml) bottles. The process fills the bottles on average with 1000 ml with a standard deviation of 2 ml normally distributed.

What is the probability a bottle is filled with less than 998 ml?

Let X = amount in a bottle. What is the distribution of X? (Note that you have to specify the name of the distribution as well as the values of the parameters of the distribution.)

P(X ≤ 998) = norm.dist(998,1000,2,1) = .1587

Alternatively, using the Z table approach we would calculate

$$P(X \leq 998) = P\left(\frac{X-\mu}{\sigma} \leq \frac{998-1000}{2}\right) = P(Z \leq -1)$$

Which we would look up in the table and get a value of .1587.

What is the probability the bottle is filled with between 998 and 1002 ml?

P(998 ≤ x ≤ 1002) = norm.dist(1002,1000,2,1) – norm.dist(998,1000,2,1) = .8413 - .1587 = .6827

And using the Z table approach this would be

$$P(998 \leq X \leq 1002) = P\left(\frac{998-1000}{2} \leq \frac{X-\mu}{\sigma} \leq \frac{1002-1000}{2}\right) = P(-1 \leq Z \leq 1)$$
$$= P(Z \leq 1) - P(Z \leq -1) = .8413 - .1587 = .6827$$

The normal distribution has some nice properties.

For any normal distribution, the probability that the random variable takes on a value within

± 1 σ of its mean will be .6827

± 2 σ of its mean will be .9545

± 3 σ of its mean will be .9973

This is sometimes referred to as “the empirical rule”.

In particular, note that pretty much all values of a normally distributed random variable will be within 3 standard deviations of its mean.

Also, look at the calculation we just did for P(998 ≤ X ≤ 1002). We got .6827 which we could have seen from the empirical rule since what we were calculating was the probability that X was within 1 standard deviation either side of its mean.

Sampling and Sampling Distributions

Populations, Samples and Sampling

The population is the set of all things we are interested in. We can talk about a population of people (e.g., the population of Georgia, or the population of Emory students) or a population of things (e.g., the population of all Volt automobiles or all ice cream cones sold in Ocean City, NJ this year.)

We will use $X_1, X_2, \ldots, X_N$ to represent the elements of a population (note there are capital N elements).

A sample is any subset of the population. So the students in this class are a sample of all Emory students. Note that the students in this class are not a representative sample but they do make up a sample.

We will use $X_1, X_2, \ldots, X_n$ to represent the elements of a sample (note there are now lowercase n elements).

As we said before, we typically are interested in the population but only have samples to work with.

The process of obtaining a sample is known as sampling.

There are a number of sampling issues that, while important, are not really quantitative in nature:

- Primary vs secondary data
- Data collection method for primary data – focus groups, survey, experiments, etc.
- Survey administration – mail, phone, internet, in-person
- Choosing and posing questions, questionnaire design

We are more interested in the quantitative sampling issues and we will focus more on the sample design.

Sampling Plan

- Target population – population of interest
- Population units – individuals, households, companies, departments within companies, transactions, etc.
- Population frame – the method (usually a list) used to identify every member of the population. If not a list, usually use a counting method, e.g., 5^{th} house in each block, every 3^{rd} shopper

Sampling Procedure

- Non-probability sampling
    - judgment sampling, e.g., test market selection
    - convenience sampling, e.g., faculty using students from their classes
    - does not lead to valid statistical conclusions

- Probability sampling
    - Simple random sampling
        - each member of the population has an equal chance of being selected
        - given a list, generate random numbers and use the matching elements or choose a starting point and choose every n^{th} element
    - Stratified sampling
        - divide the population into strata and sample from each strata
        - usually strata sampled at different rates
    - Cluster sampling
        - Divide population into clusters
        - Randomly choose clusters to sample
        - Sample within chosen cluster

Types of errors

- Non-sampling error
    - usually bad practice
    - can be reduced or eliminated by better technique
- Sampling error
    - introduced by fact we are sampling
    - results from variation in population
    - can be reduced. How?

We need to use a good sampling technique – for the most part this will be probability sampling, and we need to eliminate or reduce as much as possible the non-sampling error.

<u>Sampling Distributions</u>

Let's go back to the basic idea that we started the course with.

We said that all of statistics essentially comes down to the following scenario:

1. We want to know something about a population. For example, what factors drive profitability? What percentage of our customers were satisfied with their most recent service interaction? How effective is our firm's advertising? Does providing health insurance to workers reduce absenteeism?
2. It is infeasible (too much time, money, or both) to collect data on the entire population.
3. Instead of collecting data on the population we take a sample and calculate values from the sample, these are statistics.
4. We use the statistics to estimate the population values. This is estimation.
5. We use techniques from statistical inference (confidence intervals, hypothesis tests, regression etc.) to make further statements about the population values.

Let's tie this back to probability.

Recall that an experiment is a process with more than one outcome and which outcome occurs is uncertain.

We will take one sample but there are many different samples we could have taken so the process of taking a sample is an experiment.

Recall also that a random variable is a variable that takes on a value based on the outcome of an experiment.

Notice that the statistic we calculate is a number that is generated based on the outcome of an experiment, so by definition, the statistic is a random variable.

Again, as we pointed out at the beginning of the course this is so important it bears repeating yet again: s**tatistics are random variables**.

That means that **the statistic has a distribution** so if we can figure out the distribution of the statistic we can use it to make inferences.

The distribution of the statistic is referred to as a sampling distribution because it is the distribution we would get if we took all the possible samples, calculated our statistic for each sample, and then drew the histogram for the values we calculated.

If we were sampling from some population and we knew the distribution of the population we might be able to figure out (derive) the sampling distribution. It turns out that even if we don't know the distribution of the underlying population from which we are sampling, we can often identify the form of the sampling distribution.

One of the things we are most interested in is the mean of the population. If we had all the elements of the population we would calculate the mean (which we refer to as μ) by adding up all the N elements and dividing by N. Specifically,

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

We can rewrite this as

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

If you think about it, each element of the population is equally likely, so has probability $\frac{1}{N}$, so what we are doing here is calculating a weighted average of the data where the weights are the probabilities. But this is what we did to calculate the expected value of a random variable so you see why we use μ in both cases – they really are the same thing.

It turns out that we know the sampling distribution of the sample mean due to the following theorem.

The Central Limit Theorem (CLT)

Let $x_1, x_2, \ldots, x_n$ be a sample from a population with mean μ and standard deviation σ. Let $\bar{X} = \frac{\sum x_i}{n}$.

$\bar{X}$ (read X bar) is the sample mean.

The expected value or mean of $\bar{X}$ will be μ and the standard deviation of $\bar{X}$ will be $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

So the expected value of the sample mean is the same as the mean of the population we are sampling from (again, we call this the underlying population) and the standard deviation of the sample mean (also referred to as the standard error of the mean) will be the standard deviation of the underlying population divided by the square root of the sample size.

If the original population is normal then $\bar{X}$ will be normal.

So if we are sampling from a population that fits the bell curve, we can say that the sample mean will be normally distributed. But what if we can't say that the population fits a bell curve?

The important result is the following, which tells us the distribution of $\bar{X}$ even if we don't know the distribution of the original population:

If the original population is not normal and if n is large enough then $\bar{X}$ will be approximately normal and the approximation will get better as n gets bigger.

The remaining question is how large n has to be for this to be a reasonable approximation. It turns out surprisingly small. If the original population is not too skewed then $n \geq 30$ or so will give a reasonable approximation.

This result allows us to characterize our uncertainty in $\bar{X}$.

Example

An electric car manufacturer claims that the car can be driven, on average, for 100 miles on a fully charged battery.

It is impossible to test this for every car so we take a sample.

We calculate the average number of miles for the cars in the sample, say this turns out to be 99.8 hours.

This is a sample value that depends on the sample we took. (Since a statistic is a random variable.)

If we took another sample we might get 100.5 or 99.6 or whatever.

What we want to know is whether a sample value of 99.8 gives us evidence that allows us to conclude that the population average is not really at least 100 hours, in other words, that the manufacturer's claim is not true.

Certainly the fact that it is less than 100 makes us wonder if the claim is true (on average the value of our statistic should be equal to the mean) but our sample value is a random variable and even if the population mean was 100 there would be a 50% chance that we would get a value less than 100. What we want to know is how likely it is that we would get a value this small. In other words, if the mean really was 100 is it reasonable to get 99.8 for the sample mean?

Knowing the distribution of the sample mean we can calculate the probability of getting the value we got. If it turns out that assuming the mean was 100 it was then unlikely that we would get the value we got we would conclude that it probably did not happen by chance alone and we would think the manufacturer's claim was not true. If it turns out that getting a value of 99.8 was relatively likely we would not be able to draw that conclusion. Note that in this case we are not saying the claim is true, just that we can't show it false. We will come back to this idea; what we just described is essentially a hypothesis test.

In any case, knowing the sampling distribution is what allows us to calculate the probability we want.

As an example of specific calculations, recall the previous example where we were filling 1 liter (1000 ml) bottles. We said the process filled the bottles on average with 1000 ml with a standard deviation of 2 ml but let's assume now that we don't know the distribution of the amount each bottle is filled with.

Say we take a sample of 100 bottles and calculate the sample mean. What is the probability that the sample mean is less than 999.95 ml?

By the CLT, the sample mean will be approximately normal with a mean of 1000 and a standard deviation of $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = \frac{2}{10} = .2$

$$P(\bar{X} < 999.95) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{999.95 - 1000}{.2}\right) = P\left(Z < \frac{-.05}{.2}\right) = P(Z < -.25) = .4013$$

We can also use the CLT to identify the sampling distribution for the sample proportion.

The population proportion is the percentage of elements of the population that have some attribute of interest. The attribute could be pretty much anything, for example

- Prefer our product to competitor
- Earn over $400,000 per year
- McDonald's orders that include fries

Again, it is not feasible to determine the value of the population proportion so we take a sample of size n. If X elements of the sample have the attribute of interest (e.g., we sample n McDonald's orders and let X = the number of orders that include fries), we define the sample proportion as X/n.

There is another way of calculating the sample proportion. Take a sample of size n and let $x_i = 1$ if the i^{th} element has the attribute of interest and let $x_i = 0$ otherwise for $i = 1, \ldots, n$

What is X in terms of the x_i?

If you consider this new definition for the sample proportion it turns out we can again use the CLT. Do you see why? (Hint: compare this way of defining the sample proportion to the formula for the sample mean.)

So it turns out that the sampling distribution of the sample proportion will also be approximately normal.

For the sample proportion we have the following result:

Say we have a population in which the proportion of items in the population with some attribute of interest is π. Take a sample of n items and let X = number of items in the sample that have the attribute of interest. Let $p = \frac{X}{n}$

So π is the population proportion and p is the sample proportion. It turns out the standard error of the proportion (the standard deviation of p) will be $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$ so the CLT tells us that **p will be approximately normal with mean of π and standard deviation of $\boldsymbol{\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}}$**

As an example say we know that exactly 60% of all customers prefer our product and 40% prefer our competitor's product. Take a sample of 36 customers and let p be the proportion of these customers that prefer our product. What is the probability that p ≥ .62?

$$P(p \geq .62) = P\left(\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \geq \frac{.62-.6}{\sqrt{\frac{.6(1-.6)}{36}}}\right) = P\left(Z \geq \frac{.02}{.0816}\right) = P(Z \geq .245)$$

Notice that the table gives us probabilities that Z is less than specific values so we need to put this in that form. We use the rule of complements to say that P(Z ≥.245) = 1 – P(Z<.245). If you now look at the table you will find a value for .24 and one for .25 but not one for .245. In this case we interpolate[4]; since we want a value midway between .24 and .25 we split the difference to get P(Z < .245) = .5968 so P(Z ≥.245) = 1 – .5968 = .4032

[4] This is linear interpolation which is not exactly correct as the function is not linear, but it is a reasonable approximation and better than not interpolating at all.

Classifying Data

We can classify data as categorical or numerical.

Categorical data are things that can only be put into a category: do you own or rent your home, are you employed, a student, retired, etc.

Numerical data are things that represent quantities. How many people own their own home, how many rent?

We can also classify data as continuous or discrete. Similar to random variables, discrete variables take on specific (discrete) values while continuous variables take on any value in a range.

Measurement Scales

A nominal scale puts data into categories but there is no ranking of the categories. Bus route numbers are nominal data – the number on the bus is just a designator telling us the route (the category) but nothing more.

An ordinal scale puts data into categories but now the categories are ordered in some way. Student's grades are an ordinal scale as are small, medium and large drink sizes at a fast food restaurant.

Ordinal data is more useful than nominal data as it provides a ranking but the scale does not tell us anything about the differences between levels on the scale. How much larger is a large soft drink than a medium?

An interval scale is an ordinal scale where the difference between values has meaning but which has no absolute zero. The Celsius and Farhenheit scales are interval scales – a 1 degree difference has the same meaning everywhere on the scale but values below 0 are possible.

A ratio scale is an interval scale that has an absolute zero. Most scales we use to measure quantitative variables are ratio scales, e.g., weight, height, age, salary, speed, etc.

Descriptive Statistics

We will use *DescriptiveStatistics.xlsx* which has been posted to Blackboard for data and calculations.

We will also use the data analysis tools. The data analysis tool pak comes with Excel but it is not part of the basic install. We will go over in class how to install it.

You can also use PHStat2 which you can get from the companion web site for the textbook (see the book for instructions). If you have questions about PHStat2 I will try and answer them but be aware that I do not use it a lot so am not totally fluent. In the language of IT shops, PHStat2 is not a supported product in this class but you are welcome to use it.

Given a set of data we typically do two things: first we draw pictures (graphs, charts, etc.) to get a sense of what is going on in the data and then second, we calculate measures of the data to quantify the relationships.

Categorical Data

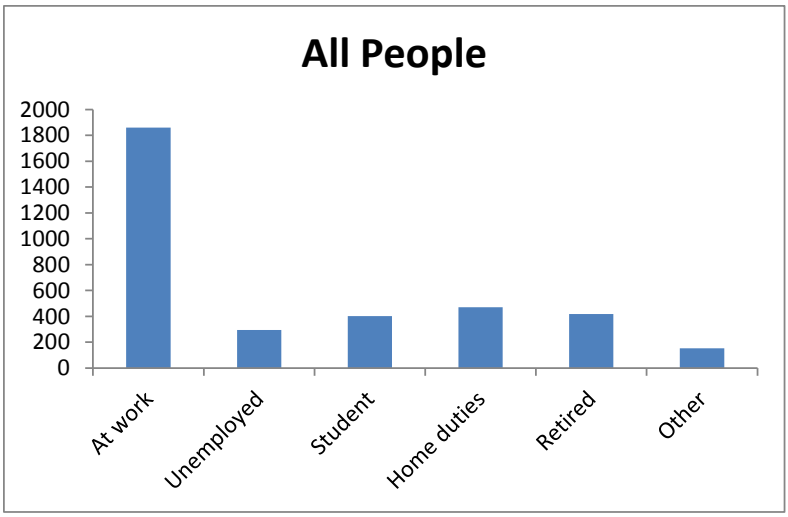
We will use the data from the Ireland Central Statistics Office that we used before. We had the following table:

Persons Aged 15 and Over (thousands)	Male	Female	Totals
At work	1014.7	844.8	1859.5
Unemployed	191.4	103.4	294.8
Student	204.2	196.8	401
Home duties	9.6	460.7	470.3
Retired	257.9	158.8	416.7
Other	84.8	68.4	153.2
Totals	1762.6	1832.9	3595.5

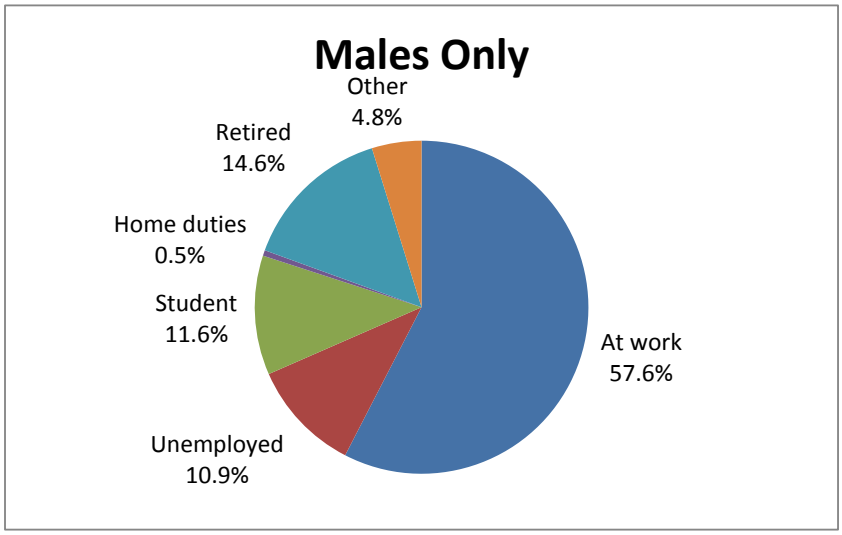
This is categorical data and with these type of data we often create bar charts[5], pie charts and Pareto charts. These examples are in the Ireland CSO worksheet in DescriptiveStatistics.xlsx

A bar chart shows a bar for each data point. A bar chart for the totals column is on the next page.

[5] Excel, as does your book, refers to the chart we will create as a column chart. In Excel's usage a bar chart represents the data horizontally while a column chart represents it vertically. We will represent the data vertically but call it a bar chart as that is common industry usage.



A pie chart uses parts of a circle (slices of a pie) to represent the number in each category. This is a pie chart for the Males only. We might want to put the counts for each category either instead of or in addition to the percentages shown. It is a matter of personal preference but showing both does get a little busy. We also could add a legend if we wanted to but that cuts down on the space available for the chart itself.

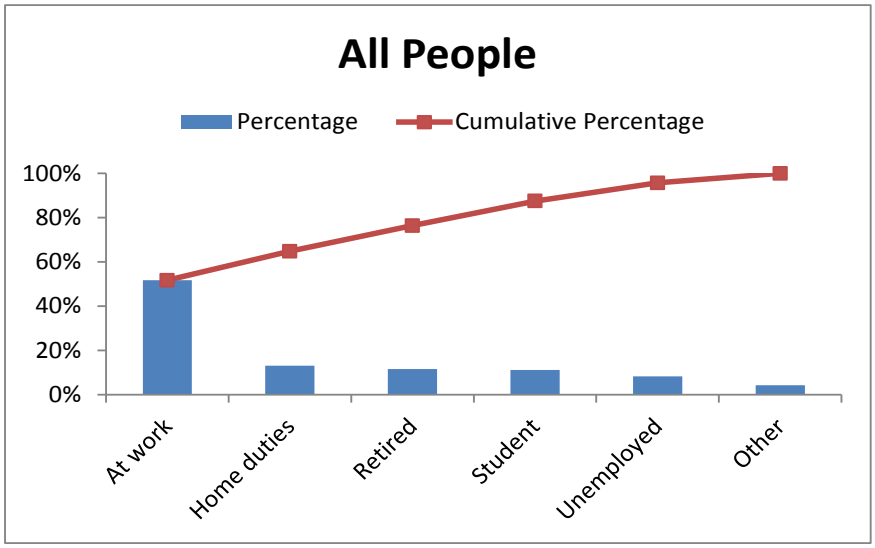


Another type of chart is a Pareto chart.

You may have heard of the 80-20 rule. The idea is that often 20% of some set of categories accounts for 80% of the total. Often these are not exactly 80-20 but they mean to convey the idea that there are a "vital few" things we should focus on (the 20%) and a "trivial many" (the rest) that do not warrant our attention. During the heat of the debate over the healthcare law President Obama claimed that 20% of hospital patients accounted for 80% of hospital costs. Whether that is precisely true or not, it is an example of the 80-20 rule. The more formal name is the Pareto Principle.

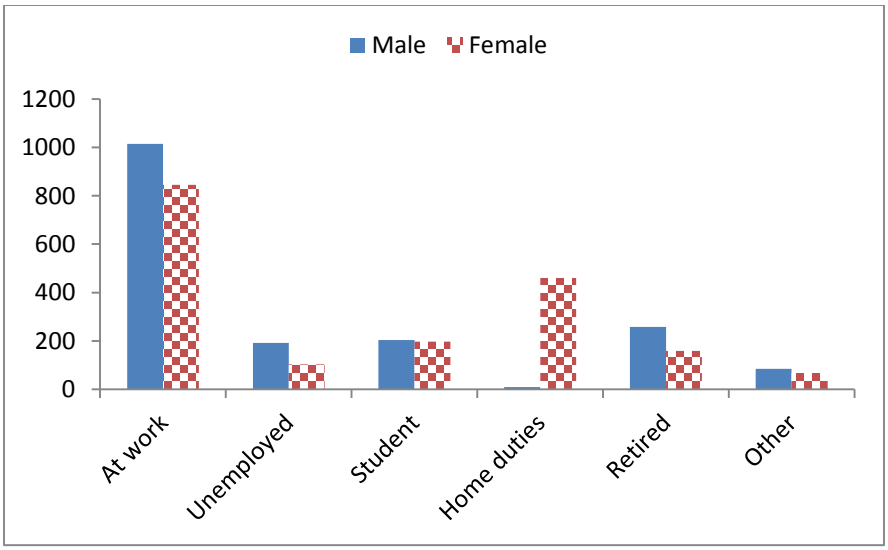
A Pareto chart is an ordered bar chart (ordered so that the bars get smaller as you go left to right) where the bars represent the percentages each category is of the total. A cumulative percentage line (called an ogive) is also displayed on the chart. This lets us easily see the percent that each bar is by itself and cumulatively what percentage the largest bars (or any subset of the bars) account for.

This is the Pareto chart for All People.



A fourth type of chart we might create (and there are many others as you can see in Excel) is a side-by-side bar chart. This lets us see how two data series compare to each other.

This is the side-by-side bar chart for Males and Females.



We will demonstrate how to create all of these charts in class. Instructions are also in your textbook.

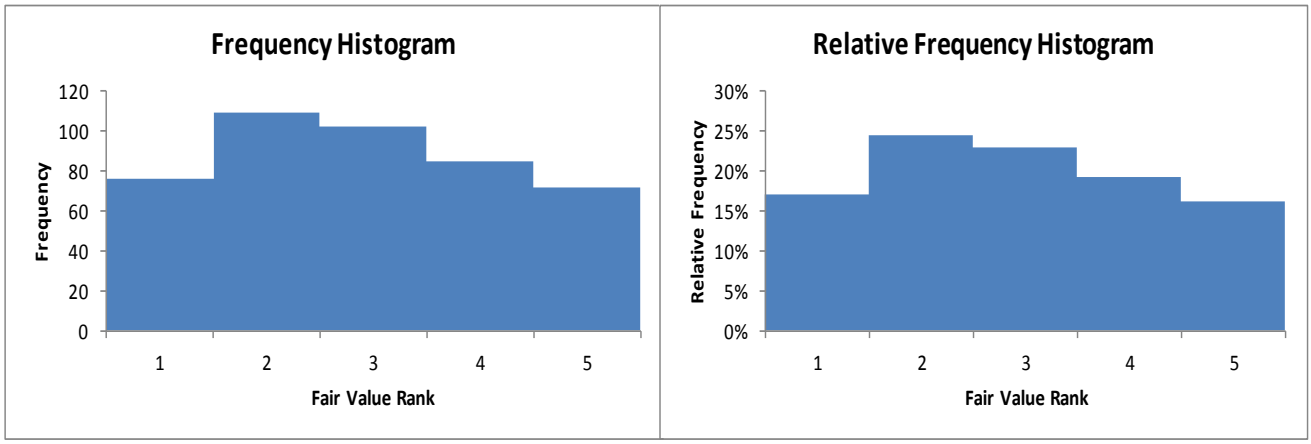
Numerical Data

Often we can use the categorical data charts for numerical data but there are also some new kinds of charts that would not make sense for categorical data but do make sense for numerical data.

We will look at histograms, both with and without cumulative curves (the cumulative curve is the ogive we had in a Pareto chart), scatterplots when we want to compare two data series and line charts to represent time series.

For histograms and scatterplots we will use the data in the S&P500 Data worksheet in DescriptiveStatistics.xlsx. This worksheet has data on the companies in the S&P 500, a commonly used market index.

One of the data series is the S&P Fair Value Rank. This is a rank applied by S&P to the stock of various companies. There are 5 possible values and the frequency and relative frequency histograms are below.



The frequency histogram gives counts of the data while the relative frequency histogram gives the percentages of the data. Note that these two histograms look exactly the same except for the left scale.

Notice that the relative frequency histogram is essentially an empirical probability distribution. A rank of 1 was given to 17% of the companies so we would say that the probability a randomly selected company had a rank of 1 was .17 or 17%. We can make similar statements about the probability a randomly selected company has each of the other ranks, but this is exactly what a probability distribution would tell us – the probability the random variable took on each of its possible values.

Since there were 5 possible rank values this was a pretty straightforward histogram – we just made each category or bin (in Excel talk) equal to one of the values.

Usually things aren't that simple. For another example consider the Average Daily Volume column in this same worksheet.

The average daily volume for these companies ranges from around 225,000 to around 80,000,000. The first thing we need to consider is how to set the bins for a histogram of these data. If you don't specify

your own bins Excel will pick some for you but their choice will not be as good as if you picked them yourself.

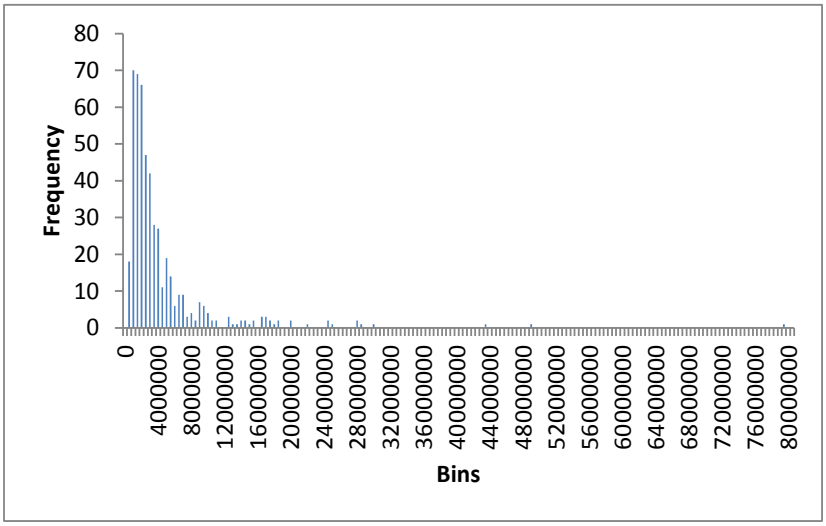
Your book says that you should have 5 to 15 bins but this is not a great approach, especially with the kind of data we have here. If you notice even though the range is from 225 thousand to 80 million, 80% of the values are less than 4.5 million. Even if we went to 15 bins the bin width would be just over 5 million so the great majority of the data would be in the first bin. This is not good.

There are various more sophisticated ways of picking the bin widths but they can be somewhat complicated. A simple approach is to look at how much data you have and pick bins so you have at least 5 to 10 data points on average per bin. We have around 500 data points so this would be 50 to 100 bins.

I would probably start with something like this (since 100 bins would still give them a width of 800,000 or so) and see whether it seemed to give a good representation of the data.

The key is that you are trying to graph the data to get an idea of how the data is distributed. Too few bins will not do this and too many will not either. It is OK to start with one set of bins and then modify them – as long as you are doing this to give a more fair representation and not to obfuscate.

You also want bin sizes that make sense so rather than 800,000, I would probably go with 500,000 or 1 million. Using a bin width of 500,000 we get the histogram on the next page.

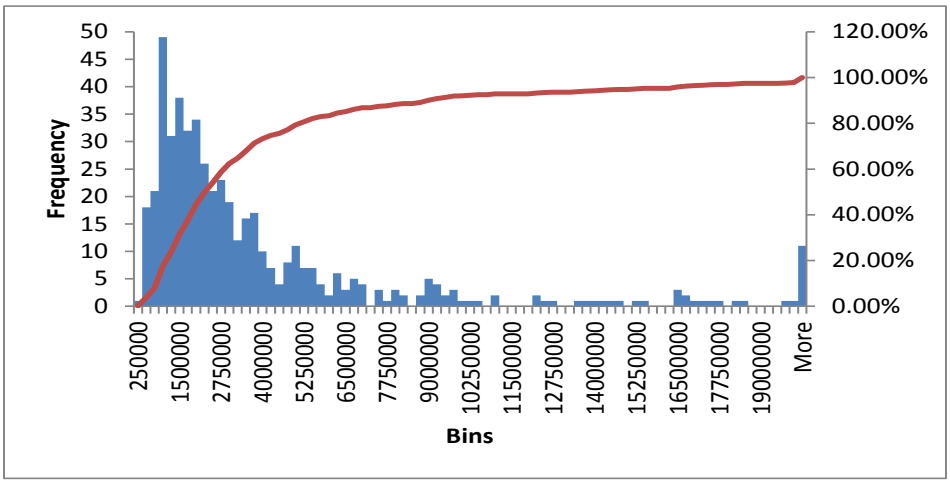


We can see what is going on. The data are left skewed with a very long tail; there are very few data points once we get past 2 million or so.

We have a few choices now:

1) Use the histogram above.
2) Use bins up to a certain level and put everything above the largest bin into a separate category (Excel will do this using a “More” category).
3) Use bins up to a certain level and report how many are not being plotted.

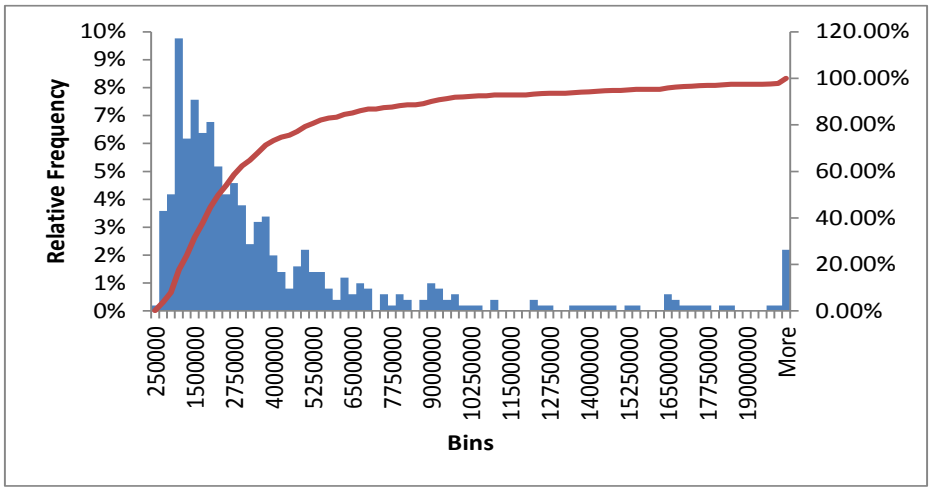
Using the More category with bins up to 2 million with a width of 250,000 and adding an ogive we get:



This gives a pretty good picture of what is going on. We have good definition on the left side of the distribution and we can see the long tail nature of the distribution. There are only about 10 stocks that had volume over 2 million shares (these are in the "More" category) so we have captured the great majority of the data.

The ogive is the red line and it shows us the cumulative probability. Notice that it slopes up pretty quickly at the start suggesting that most of the stocks have relatively small daily volume.

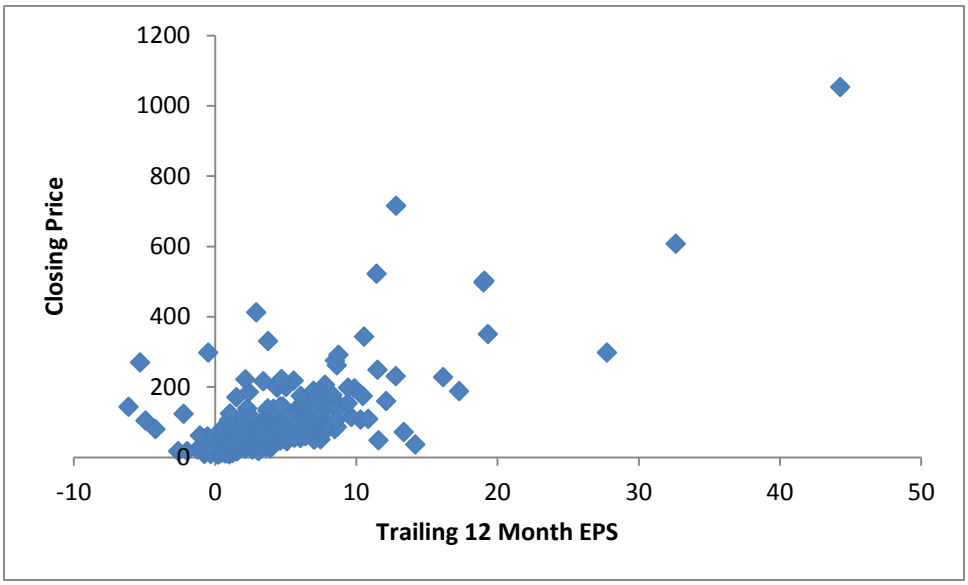
We can also do a relative frequency histogram:



Notice that the frequency histogram and relative frequency histogram look exactly the same except for the scale on the left axis.

If we want to see how two data sets relate to each other we can do a scatterplot. In a scatter plot we plot the data as pairs of values where the X value is for one variable and the Y variable is for the other.

We have data on closing price and on trailing 12 month earnings per share. We might expect these data to be related as companies with greater earnings typically sell for higher prices. This is the scatterplot:

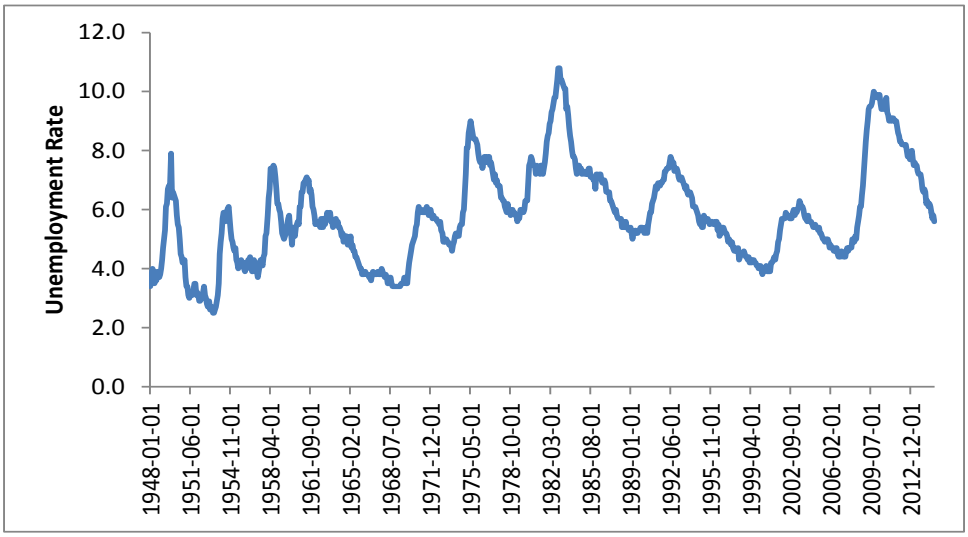


There is an increasing relationship that looks roughly linear or maybe slightly quadratic.

Often data will be in the form of a time series which is any data set referenced by time. So the monthly or quarterly sales of a company is a time series as are weekly initial claims data or annual automobile production.

The Unemployment worksheet has the civilian unemployment rate both seasonally adjusted and not seasonally adjusted from 1948 through December of 2014. With a time series we want to see how the data varies over time so we use a line chart with time on the X axis.

This is the seasonally adjusted unemployment rate as a line chart:



In general, you want to choose the chart type and chart that best (and by best I mean most accurately and fairly) represents your data.

You certainly want to present the data in a way that makes your case but you don't want to misrepresent the data to make your point.

We said that we would also calculate measures of data so we will do that next. There are many things we could calculate. For a single data series we will calculate two types of measures: measures of central tendency and measures of dispersion.

In some of the measures we will make a distinction between the measure for a sample and for a population. We have already introduced some of the measures but we will repeat them here for completeness.

In fact you should see many parallels between what we do with data and what we previously did with random variables.

Measures of Central Tendency

These are measures of the center of the data set. There are three common measures: mean, median and mode.

The mode is the most common value. This is typically of least interest to us. First, there does not have to be a mode. If the data are all unique there will be no mode. Even if there is a mode it might not tell us much about the center of the data. Say we measured the weight of everyone in the class to the nearest tenth of a pound. It is very likely that no two people weigh the same amount to a tenth of a pound (probably even to the pound) so there may very well be no mode. If two people did weigh exactly the same to the tenth of a pound we would probably think it was just a coincidence.

Mode can be calculated using the *mode* function in Excel.

Mode does have interest when we look at the distribution of the data. We talk about unimodal (one "peak" in the distribution), bimodal (two "peaks") and multimodal (more than two "peaks") data sets.

The median is the middle value. If there are an even number of values you take the average of the middle two values. There is a *median* function in Excel or you can sort the data and find the middle value.

The mean is the average value. The *average* function in Excel calculates the mean. As discussed previously the mean of a data set is analogous to the mean for a random variable.

We already defined the population mean as

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

and the sample mean as

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

There are two differences between these formulas.

First we use N for the population and n for the sample, consistent with our notation above.

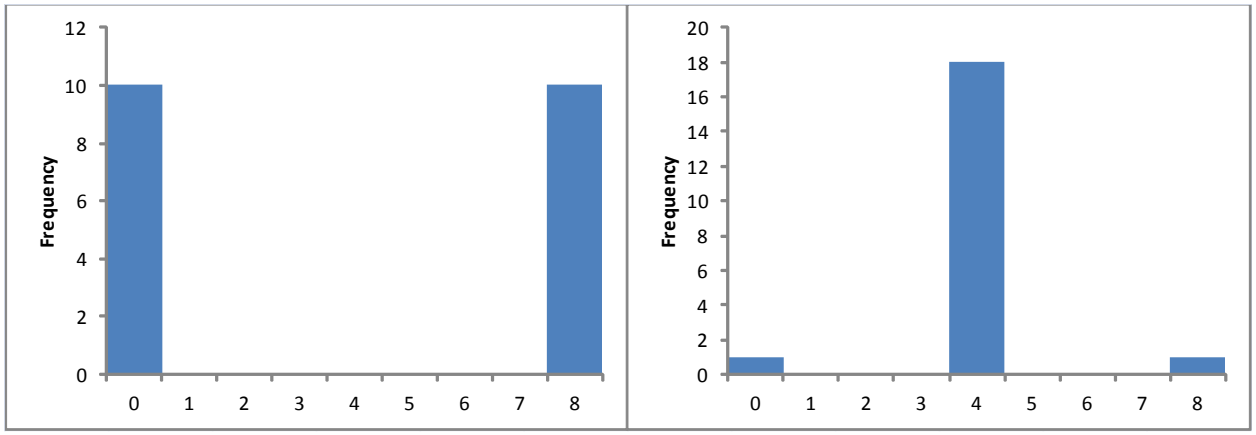
Second we refer to the population mean as μ, a Greek letter and to the sample mean as $\bar{X}$, a Latin alphabet letter. In general **we use Greek letters to refer to population values and Latin letters to refer to sample values**.

Measures of Dispersion

We already talked about these in terms of random variables and the discussion carries over directly to data for the most part.

We can look at the range for a data set but we still have the problem that it is not a great indicator of dispersion.

In terms of data (note the left scale is now frequency not probability) we could compare the two histograms below:



The range is the same but clearly the one on the left is more disperse than the one on the right. The average of each of these data sets will be the same (4). In the histogram on the left everything is far from the center while in the one on the right most values are in the center and only two are at the extremes (this is why we would say the one on the right is less spread out.)

We could try to measure the average distance to the mean. For a population this would be

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

but this won't work for the same reason it did not work for random variables (again we will always get the same value).

Instead, just like for random variables, we calculate the average of the squared distance to the mean which is the variance.

For a population this is defined:

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

And just like for random variables we define the standard deviation as the square root of the variance.

We can rewrite the formula for σ^2 (as we did for the population mean) as:

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

Again, realizing that each data point appears once in the population (hence has probability $\frac{1}{N}$) we can interpret this as the weighted average squared distance to the mean where the weights are the probabilities so it is directly analogous to the variance of a random variable.

For a sample we define:

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

And we define the sample standard deviation, S, as the square root of the sample variance.

Note that we divide by n-1 instead of n in the sample variance. It turns out that this gives us a better value than dividing by n would. We will explain why this is when we get to estimation.

Another measure of variation is the coefficient of variation. This measures the variation relative to the mean. It is defined:

$$CV = \left(\frac{S}{\bar{X}}\right)100\%$$

It is useful if we want to compare two different data sets that are measured in different units.

We can also use the Z score to identify values far from the mean. We calculate:

$$Z = \frac{X - \bar{X}}{S}$$

Some people would conclude that anything with a Z score less than -3 or greater than +3 is an outlier.

You have to be careful with outliers. Often people in industry will identify a data point as an outlier and decide to exclude it from analysis on that basis. Before you do that you have to consider whether the outlier is legitimate or not. If the data are normally distributed we would expect that almost all the values would be within plus or minus three standard deviations (so Z scores outside the range -3 to 3 might be a cause for concern) but for other distributions like the long tailed distribution above, something more than 3 standard deviations might not be unexpected.

If you have some reason to think the outlier is a mistake (e.g., possibly data was captured or recorded incorrectly) you should probably exclude it but otherwise you should consider whether the outlier came from the process you are sampling. If it did then it should probably not be eliminated.

Quartiles

The quartiles split the data into fourths. The first quartile is the value such that 25% of all values are smaller. The second quartile is the median, the third is the value such that 75% of the values are less.

Covariance and Correlation

These are measures of how two variables relate to each other.

The covariance is a measure of how two variables vary together, that is, when one is above its mean does the other tend to be above its mean? Or is the other below its mean? Or maybe sometimes when one is above its mean the other is above and sometimes it is below.

If the two variables tend to both be above their means at the same time they will have a positive covariance, if when one is above the other tends to be below they will have a negative covariance and if there is no pattern the covariance will tend to be close to zero.

For example we might expect height and weight to have a positive covariance (people taller than average probably in general also are heavier than average and vice versa.)

The covariance is difficult to interpret, however, as it depends on the size of the data, that is, how big the values of the data tend to be.

The covariance of height and weight for people will tend to be greater than it is for miniature parakeets just because the values are bigger, irrespective of which relationship is stronger.

The covariance is defined as

$$\sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_x)(Y_i - \mu_y)}{N}$$

The *covar* function in Excel or the data analysis tools can be used to calculate covariance.

The covariance is used extensively in finance.

One application is to calculate the standard deviation of a portfolio of stocks. The standard deviation is a common measure of the risk of a stock.

Say we have a portfolio of two stocks. The standard deviation of the portfolio can be calculated as

$$\sigma_p = \sqrt{W_X\sigma_X^2 + W_Y\sigma_Y^2 + 2W_XW_Y\sigma_{XY}}$$

Where W_X is the fraction of the portfolio in stock X and W_Y is the fraction of the portfolio in stock Y.

Notice the covariance term in the radical.

If the covariance of the two stocks was zero, that is if there was no relationship between them, this last term would drop out. If the covariance is negative, that is if the two stocks vary in opposite directions, the standard deviation will be less then if the covariance is zero.

This says that the risk of the portfolio decreases by combining stocks with a negative covariance.

By the same token, the risk would increase if we combined stocks with a positive covariance.

As we said above, it is difficult to interpret the covariance as it depends on the size of the data.

Google and Apple (which both trade in the hundreds) will probably have a greater covariance then Intel and Texas Instruments (which both trade in the 20s or 30s or so) simply because the values are bigger.

Another measure of the relationship between two variables that we can easily interpret is the coefficient of correlation (usually referred to as just the correlation). The correlation is the standardized covariance:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Since we used a Greek letter for this (rho), it is a population value.

We use r to denote the sample correlation.

Both the population and the sample correlation will always be between -1 and 1.

The correlation measures the strength of the linear relationship between the two variables.

The closer the correlation is to 1, the closer the points in a scatter plot of the two variables would fall to an increasing straight line. The closer the correlation is to -1, the closer the points in a scatter plot of the two variables would fall to a decreasing straight line.

It is critical to remember that the correlation measures the strength of the linear relationship.

Two variables that have a strong non-linear relationship can have a small correlation. A simple example of this is Y = |X| (the absolute value of X) for X = -2, -1, 0, 1, 2. Y is uniquely determined by X so the relationship could not be stronger and yet the correlation will be zero.

The *correl* function in Excel or the data analysis tools can be used to calculate the correlation.

Example

Below are the mean, median, mode, population standard deviation, sample standard deviation and coefficient of variation for the Daily Average Volume data in the S&P500 Data worksheet.

Mean: function *average*
4,140,829
Median: function *median*
2,258,266
Mode: function *mode*
#N/A
Population St Dev: function *stdevp*
6,255,892
Sample St Dev: function *stdev*
6,262,132
Coefficient of Variation
151.1%

Some notes:

1) The Excel functions are given in the table except for the coefficient of variation as this is just the standard deviation divided by the mean.
2) The argument for each of the Excel functions is the array that contains the data.
3) The mean is much larger than the median. This is because the mean is much more sensitive to extreme values and we already saw that we have a long tail with some very extreme values.
4) The mode function in Excel returns #N/A as there is no mode (that is, no value appears more than once) which is not surprising. If there were more than one mode, Excel would only return the first of them.
5) There are two formulas for standard deviation: stdevp for a population and stdev for a sample. The difference of course is the denominator.
6) The table shows both the population and sample standard deviations. Usually we have either a sample or a population – in this case we can think of the data as the population of S&P 500 stocks or as a sample of all publicly traded stocks so we calculated both.

We can also calculate the quartiles. The Excel function =quartile.inc(array, quart) returns the quartiles. Array is the data and quart = 0 for the minimum value, 1 for the first quartile, 2 for the second quartile or median, 3 for the third quartile and 4 for the maximum.

These are the values returned by the quartile.inc function

Quart	Average Daily Volume
0	225,490
1	1,324,902
2	2,258,266
3	4,329,627
4	79,467,172

The quartiles show us the long tailed nature of the distribution. Note that the third quartile is just over 4.3 million so 75% of all values are less than this value but the maximum is almost 80 million. That means that 25% of the values are between 4.3 million and 80 million.

The skewness (Excel function: *skew*) is a measure of the skew in the distribution and the kurtosis (Excel function: *kurt*) is a measure of how peaked the distribution is. We will not cover either of these more than these simple definitions.

If we have a pair of variables we can calculate the covariance and the correlation. These are the values of the covariance and correlation for the trailing 12 month EPS and the Closing Price.

Trailing 12 Month EPS and Closing Price	
Population Covariance: function *covariance.p*	243.3
Sample Covariance: function *covariance.s*	243.8
Correlation: function *correl*	0.74

The Excel functions are identified in the table. The arguments of each of these are the two arrays of data. The correl function is used for both the population and sample correlation.

Recall that the scatter plot showed a reasonable linear relationship and this is confirmed by the value of .74 for the correlation.

Estimation

We've been saying all along that we will use sample values (statistics) to make statements about the population values (parameters).

One of the most basic ways we do this is estimation.

We will discuss two types of estimators: point estimators and interval estimators.

Point Estimators

Point estimators are single values used to estimate a parameter. The following table shows a number of parameters and the corresponding statistic we use as a point estimator.

Parameter	Statistic
μ	$\bar{X}$
σ^2	S^2
σ	S
π	p
ρ	r

The statistics in this table are the ones used to estimate the corresponding parameter.

But why are these chosen? They are not the only things we could use.

For example, we could use the sample median (i.e., take a sample, calculate the median of the sample) to estimate the population mean.

Also, we defined the sample variance using the formula:

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

What if we defined a new variable:

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

where we divide by n instead of n-1. Why don't we use this variable to estimate the population variance?

Well we could. We can use whatever we want to estimate a parameter. What we want is the "best" estimator in each case so we need to consider what we mean by "best".

First, we would want our estimator, on average, to equal the thing we are estimating. If an estimator has this property we say it is unbiased.

More formally:

Estimators are statistics and statistics are random variables so estimators are random variables. That means that estimators have a distribution, expected value and standard deviation.

For an estimator to be unbiased we want the mean, or expected value, of the statistic to equal the parameter we are estimating.

So to be unbiased means that the expected value of the estimator is equal to the parameter being estimated.

We will not show it here but it turns out that $E[S^2] = \sigma^2$ so S^2 is unbiased.

$E[S_n^2]$ in the other hand is biased.

It turns out that $E[S_n^2] = \frac{n-1}{n}\sigma^2$ so S_n^2 is biased, its expected value is a small factor off from the parameter we are estimating.

Can you see why dividing by n-1 instead of n makes the estimator unbiased?

But why is S_n^2 biased? We defined the population variance as

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

which is an average of N things so we would think it would make sense to define S^2 as the average of n things (as we did with S_n^2) rather than as the average of n-1 things (as we did with our actual definition of S^2).

The reason S^2 is unbiased while S_n^2 is biased is that while we are averaging n terms, only n-1 of them are independent. We are actually averaging n-1 independent terms so dividing by n-1 works better.

To see why only n-1 of the terms are independent, take a simple example.

Say n=4 and the data are $x_1 = 7$, $x_2 = 3$, $x_3 = 5$, $x_4 = 9$

So the sample mean is (7+3+5+9)/4 = 6

Now imagine we don't know x_4. It turns out we can figure out what it has to be if we know the first three of the $(x_i - \bar{X})$ values.

These values are

7 – 6 = 1

3 – 6 = - 3

5 – 6 = - 1

It isn't hard to show that the sum of the $(x_i - \bar{X})$ values has to be zero. (We actually spoke about something similar when we looked at the average distance to the mean – recall that had to be zero and this is the same thing.)

This means that x_4 -6 = 3 (to make the sum of the $(x_i - \bar{X})$ values equal zero), which means that x_4 =9.

In other words, if we know the first n-1 of the $(x_i - \bar{X})$ values we can figure out what the last one has to be so the last one is not independent (that is, it is not free to be anything); the last one is determined by the first n-1.

We refer to these as degrees of freedom. The number of degrees of freedom is the number of things that are free to take on any value. Here we have one less degree of freedom than we have data points.

We say that we lost one degree of freedom when we used $\bar{X}$ in the formula.

So we defined S^2 the way we did to make it unbiased, which is a good thing.

But being unbiased is not enough.

It turns out that the sample median is an unbiased estimator of the population mean.

If we take all the possible samples and calculated the sample median for each of them, and averaged together all those sample medians, we would get the population mean.

Put another way, the expected value of the sample median is the population mean.

Another property that we would want is for the estimator to have a smaller variance than any other possible estimator.

If we looked at the sampling distribution of each possible estimator, we would want to choose the one with the least dispersion.

We call this the minimum variance estimator.

It turns out the variance of the sample median is about 1.57 times the variance of the sample mean so, while both of them are unbiased estimators, the sample mean has less variance so is a better estimator of the population mean.

What we would like are minimum variance unbiased estimators (MVUE).

The sample mean and sample variance are unbiased estimators of the population mean and variance respectively. If the data are normal then the sample mean and sample variance are the MVUE for the respective parameters.

The sample proportion is the MVUE for the population proportion.

The sample correlation is a biased estimator of the population correlation. The sample correlation can be multiplied by a factor to remove the bias but this is not commonly done. (With a sufficient amount of data the bias is relatively small.)

Also, the sample standard deviation is a biased estimator of the population standard deviation and the adjustment to make it unbiased is not simple.

All of the preceding are point estimators, they are one value that we pick as our "best guess" of the value of the parameter.

They don't, however, capture our uncertainty about our estimate. For that we use confidence intervals.

Confidence Intervals

Confidence intervals (CIs) are a type of interval estimator. Rather than pick a specific point estimate we pick a range of values that we have some confidence cover or include the parameter value.

We can create confidence intervals for a range of parameters, we will discuss only CIs for the mean and the population proportion.

To do this we will rely on the Central Limit Theorem.

We are going to create an interval that we think includes the true value of the population parameter.

α is called the significance level, it is the chance that we are willing to accept that we are wrong, essentially the chance that our interval does not include the population parameter.

α will be a decimal, typically .01, .05 or .10 and we create a 100(1-α)% CI.

So for example, if α is .05 a 100(1-.05)% CI would be a 95% CI.

A 95% CI is an interval that we think has a 95% chance of including the true parameter value.

(The last statement is somewhat imprecise and a purist would quibble with it but it is a good working definition for our purposes.)

We will create CIs for the population mean assuming we know the population standard deviation, σ, then assuming we don't know σ and we will also create CIs for the population proportion.

CI for the Mean with known σ

Assume we take a sample of size n from a population with mean μ and standard deviation σ and we calculate the sample mean for this sample.

Let the significance level be α.

We want a general formula but for now, assume that α = .05

If n is large enough, by the CLT the sample mean will be approximately normal with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$. From the Z tables we can see that

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) = .95$$

(Recall that we talked about this before; 95% of all values of a normal random variable are within 1.96 standard deviations of its mean.)

We can multiply all three terms inside the parentheses by $\frac{\sigma}{\sqrt{n}}$, subtract $\bar{X}$ from each term, multiply each term by -1 and rearrange to get

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95$$

So this gives us our 95% CI. The interval is $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ and we are 95% confident that μ is in this interval.

We want a more general formula though and this is just for α = .05. The thing that made it unique to this value of α was the use of 1.96. We chose 1.96 because that is the value from the Z table such that .025 of the probability is above this value and .025 is below the negative of this value. Notice that .025 is α/2 and we refer to the value from the Z table such that α/2 probability is above the value and α/2 probability is below the negative of the value as $Z_{\alpha/2}$

So if we want a general formula we should replace 1.96 with $Z_{\alpha/2}$

With this change a 100(1-α)% CI for μ with σ known will be

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

Example:

Simple Cuts operates a chain of hair cutting salons. They have found that the standard deviation of the time for a haircut is 5 minutes. They measure the time it takes to cut the hair of 100 customers at a new location and find that the sample mean for these 100 customers is 20 minutes.

Construct a 90% confidence interval for the mean haircut time.

What is α?

What is $Z_{\alpha/2}$? (You should look this up in the table.)

What are $\bar{X}$ and σ?

If you were able to answer the three questions just asked you should be able to plug into the formula and get the interval.

If you constructed a 95% CI for these same data would it be wider or narrower?

What if the standard deviation (σ) were larger? Would the interval be wider or narrower?

Confidence Interval for the Mean with Unknown σ

When we constructed the CI for the mean with known σ we used the fact that $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ has a standard normal distribution. If we don't know σ we have to do something else.

What would you use instead of σ (since we don't know this) in the formula?

The problem is that whatever we use is going to introduce additional uncertainty since it will not necessarily be as good as using σ in the formula.

We can define $t = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ which will have a t distribution with n-1 degrees of freedom if the underlying population is normally distributed. We talked about degrees of freedom earlier when we explained why dividing by n-1 in the formula for the sample variance is better than dividing by n.

Values for the t distribution are also in a table in your book.

Again, let α = .05 and let n = 100 (we have to know the sample size now as the distribution of t will depend on the sample size). Also assume the population we are sampling from is normally distributed.

We can say the following (the value of 1.9842 is found in the table "Critical Values of t" in your book),

$$P\left(-1.9842 < \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < 1.9842\right) = .95$$

and then doing algebra similar to before we get,

$$P\left(\bar{X} - 1.9842\frac{S}{\sqrt{n}} < \mu < \bar{X} + 1.9842\frac{S}{\sqrt{n}}\right) = .95$$

So this gives us our 95% CI. The interval is $\bar{X} \pm 1.9842\frac{S}{\sqrt{n}}$ and we are 95% confident that μ is in this interval.

Notice the difference, before we had a factor of 1.96, now we have a slightly larger factor of 1.9842 so our interval is slightly larger. This is because using S instead of σ introduced additional uncertainty.

As the sample size gets larger, the factors from the t table will get smaller and will eventually converge to the same factors as the Z table. When the sample size is greater than 1000 the factors are probably

close enough to use the Z table values. For example, at n = 1000 the t-table factor would be 1.962 as opposed to a Z-table factor of 1.96.

The formula we just derived was for α = .05 and just like before we want a more general formula.

We now use $t_{\alpha/2}$ which is the value from the t table such that $\alpha/2$ probability is above this value and $\alpha/2$ probability is below the negative of this value.

Our more general formula for a 100(1-α)% CI for μ with σ unknown is

$$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

In order to use the t distribution we had to assume that the underlying population was normally distributed. It turns out that if the data are not normal, as long as they are not terribly skewed and our sample size is large, we should be OK using this formula.

Example:

A survey of 600 graduating BBA students reported an average starting salary of $58,769 with a sample standard deviation of $10,780. Construct a 90% CI for the population mean starting salary. Assume that starting salaries of BBA students are roughly normally distributed.

We have everything we need except the value of $t_{\alpha/2}$. We need the critical value for the t distribution with 599 degrees of freedom. The table in your book, however, only has values up to 120 degrees of freedom. If the value we want is not in the table but there is something close, for example if we had the value for 600 degrees of freedom, we would just use that. Since we don't have anything close we will use Excel to find the value from the t distribution. There are a few formulas in Excel that will give you the appropriate value, two are:

=TINV(α, d.f)

=T.INV.2T(α, d.f)

In each of these notice that the first argument is α, not $\alpha/2$, and the second argument is the number of degrees of freedom.

For our values, TINV(.1,599) = 1.6474 and the 90% CI will be

$$\$58{,}769 \pm (1.6474)\left(\frac{\$10{,}780}{\sqrt{600}}\right) = \$58{,}769 \pm \$725$$

This says that we are 90% confident that the true mean is between $58,044 and $59,494.

CI for the Population Proportion

Let p be the proportion in the population with some attribute:

- prefers our product to that of our competitor
- supports candidate X
- has individual income over $400,000

Take a sample and calculate p the sample proportion. (Recall that p is a sample mean.)

The standard error of p will be $\sqrt{\frac{\pi(1-\pi)}{n}}$ and a 100(1-α)% CI for π will be

$$p \pm Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$$

But of course we don't know π so we use p instead and our 100(1-α)% CI is

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

To use this formula we need np > 5and n(1-p) > 5. Another way to state this (which is the way your text states it) is that we need at least 5 items with the attribute of interest and at least 5 items without the attribute of interest.

We are not doing the derivation but this formula is similar to the two we did previously and if you compare the formulas you should see the similarities.

Example:

In a sample of 1250 college students, 49.2% reported that they ate breakfast less than three days per week[6]. Construct a 99% CI for the population proportion of college students that eat breakfast less than three days per week.

What is α?

What is $Z_{\alpha/2}$?

What is the 99% CI?

What are the implications for a new breakfast restaurant opening near a college campus?

[6] Irregular Breakfast Eating and Associated Health Behaviors: A Pilot Study Among College Students by Thiagarajah and Torabi, posted to Blackboard.

Sample Size

Sometimes we want to know what sample size we would need to have a confidence interval with a specific accuracy.

The term in the CI after the ± sign is the margin of error of the CI. It is also the half-width of the interval.

Your book refers to this quantity as the sampling error.

So if we decide the margin of error we are willing to accept we can then figure out how big a sample we need to have at most this margin of error.

We do this differently for confidence intervals for μ and for π but the idea is the same.

For the population proportion, the margin of error is

$$Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Say we want this to be less than or equal to some value *e*.

$$e \geq Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Then some algebra lets us solve for n:

$$e^2 \geq Z_{\alpha/2}^2 \frac{p(1-p)}{n}$$

$$n \geq Z_{\alpha/2}^2 \frac{p(1-p)}{e^2}$$

So we need n at least this large.

We specify α so we can find $Z_{\alpha/2}$, and we know the value of *e* as we choose *e*. The one problem is that we don't know p. We will know p after we take the sample but that will be too late.

We want an upper bound on n, that is we want a value of n that will guarantee our margin of error is at most *e*, so if we pick p to get the biggest possible n we should be OK. It turns out that p(1-p) is maximized[7] when p = .5 so that is what we will use.

Example:

If we want the margin of error to be at most .02 at a 95% CI we would need a sample size of at least $(1.96)^2(.5)(.5)/(.02^2) = 2401$.

[7] This is not difficult to show and there is an intuitive explanation for this that we will discuss in class.

Notice that if we were willing to accept a margin of error twice as large (i.e., .04 versus .02), the required sample size would drop to 600 (600.25 to be exact) which is ¼ the sample size calculated above.

When it comes to political polling typically a confidence interval is reported. For example, when they report on the news that candidate X has 45% support with a margin of error of 4% they are telling you that a CI for the population proportion π is .45 ± .04

What they don't tell you is the α they used but typically they use .05 so they are giving you a 95% CI.

You can figure out from the information they gave you how big their sample must have been.

If a 95% CI for π is .45 ± .04, then we must have $n = (1.96)^2(.45)(.55)/(.05^2) \approx 594$. (Of course the .45 and the .04 were presumably also rounded off so this calculation is only approximate, but it gets you in the ballpark.)

We can do a similar derivation to figure out the sample size we need to have a margin of error of at most *e* for a CI for μ and we would get:

$$n \geq Z_{\alpha/2}^2 \frac{\sigma^2}{e^2}$$

Again, we have a problem in that we don't know σ (we will have the other factors on the right side of this inequality). This is a bigger problem as we also can't specify a worst case as we could when we did not know p.

If we had done previous studies we might have an idea of what σ was and we would use that. Otherwise we might do a pilot study to get an idea of σ or we might take an educated guess.

Of course once you take the sample and calculate the confidence interval you will see what the margin of error is for your specific results, what these calculations are trying to do is give you an upper bound on what n has to be to get at most some specific margin of error.

Hypothesis Tests

An article in the Wall Street Journal from 1/21/15 reported on a study[8] conducted by the consultancy McKinsey & Co. The article said that the study "found a statistically significant relationship between companies with women and minorities in their upper ranks and better financial performance".

The important phrase here is statistically significant. What does this mean?

Sometimes we see a relationship in a set of data.

Consider the Chipotle example we started the course with.

Maybe if we collected some data we would find that the stores with more parking had greater profit.

Does that mean that more parking leads to greater profit?

There actually are two concerns here:

1) All we can tell from the data is that there is a correlation and correlation does not imply causation. The correlation could be due to a mistake in the data or even if the data were correct, it might be that the two factors were both correlated with something else and it was this other factor that was causal. For example, maybe stores with more parking happen to be larger and larger stores are more profitable. Or possibly stores in less urban environments (e.g., suburbs, shopping centers, etc.), which would tend to have more parking, are more profitable.
2) We are working with sample results which are of course random. It might be that the results we got were due to chance alone – that is, we just happened to get these results and most other samples we could have taken would not have given us these same results.

Statistical significance has to do with the second of these concerns. To say that a result is statistically significant means that we are reasonably certain that the result was not due to chance alone.

We say we are reasonably certain as there is no way of being absolutely certain that the results did not happen by chance alone, what we want is to have only a small probability that the results happened by chance.

We use hypothesis tests to determine if our results are statistically significant.

Just like with CIs there are many hypothesis tests. We will only cover a few of them.

Again, as with CIs we will start with a test of the mean with known standard deviation and we will use an example to illuminate the concepts.

[8] Diversity Matters, 24 November 2014, McKinsey & Company. The report has been posted to Blackboard.

Example

Wooden bats are lathed from wood blanks known as billets. A company that manufactures maple billets claims their billets are on average 37 inches long with a population standard deviation of .2 inches. If the billets are too long or too short they will not fit on the lathe properly and the bat will not be made to specifications so you want to check if the billets really are 37 inches on average.

You take a sample of 36 maple billets from this company and calculate that the average length for this sample is 36.92 inches. Do you have evidence that the population mean is not 37 inches as the company claims?

So on the one hand, you got an average different from 37 inches. That would suggest that the manufacturer's claim is not correct. On the other hand, the average you got is pretty close to 37, maybe the small difference is due to chance alone. That is, there is variation in the sizes (the manufacturer admits to a standard deviation of .2 inches) so maybe you just happened to get a sample with an average different from 37. Maybe if you measured every billet in the population you would get an average of 37 inches.

If you think about it, if you got 37 for your average you would not have any reason to believe that the population mean was not 37. But what about something different from 37? At some point, that is far enough from 37, you are going to have to conclude that getting that value for the sample mean would be very unlikely if the population mean really were 37. This will be evidence that the population mean really is not 37.

OK, so what about 36.92? Is this far enough from 37 to make us believe the population mean is not 37?

It will depend on two things: the sample size and how much variation there is in the population.

We can do the following hypothesis test:

$$H_0: \mu = 37$$

$$H_1: \mu \neq 37$$

Some notes:

1. H_0 is called the null hypothesis
2. H_1 is called the alternative hypothesis
3. Both H_0 and H_1 are **statements about parameters, not statistics**
4. H_0 will always include an equals sign (it might be an inequality but if it is it will be either "greater than or equal" or "less than or equal" it will never be strictly greater than or strictly less than). When we run the test we will take the case where we have equality.
5. H_1 is everything else not covered by the null so it will be that the parameter is not equal to the specific value in the null hypothesis or that the parameter is greater than the value in the null or that the parameter is less than the value in the null hypothesis.

6. If our alternative is that the parameter is not equal to some value, we call that a two tailed test and the alternative is a two sided alternative.
7. If our alternative is that the parameter is greater than (or less than) some value, we call that a one tailed test and the alternative is a one sided alternative.

This is the general approach we take (specifics for the billet example will follow):

1. We assume the null is true and in particular we assume that we have equality (that is why the null has to include the possibility of equality and cannot be strictly greater or strictly less than). Assuming the null to be true has to give us something to work with.
2. We then calculate a test statistic. The test statistic will vary depending on the test but it usually will be something that should be close to zero if the null is really true (as we assumed).
3. Since the test statistic should be close to zero it will be very unlikely to be far from zero.
4. If it is in fact far from zero, since that would be very unlikely if the null were true, we conclude that the null must in fact be false.
5. In this case we say that we reject the null.
6. If we reject the null than we accept the alternative (conclude that it is true).
7. If we fail to reject the null (that is, if the test statistic value is not unlikely) then we can't actually say the null is true. All we can say is that we cannot show the null to be false.

The question we still have to answer is how unlikely does the test statistic value have to be before we can reject the null?

Or another way of asking this is for what values of the test statistic will we reject the null? These values will form the rejection region – we will reject if we get a value in this region and fail to reject if we get a value outside this region (sometimes called the non-rejection region).

The test statistic is of course a random variable and so has a distribution. We will know what this distribution is, assuming the null is true.

So if the null is true the test statistic came from that particular distribution.

Recall the significance level, α, from a confidence interval. α was the chance we were willing to accept that we were wrong.

With a CI it is the chance that the interval does not include the parameter value.

In a hypothesis test, α will be the chance we are making a mistake if we reject the null. That is, the chance that the null is actually true and we reject it. This is called a Type I error.

A typical value of α is .05. With $\alpha = .05$ we reject the null if there is a 5% chance or less of getting the test statistic value we got or something even more unlikely.

So we will reject the null for some values of the test statistic and we fail to reject for others

We look at the distribution of the test statistic to find the rejection region. If the null is true then the value we got came from this distribution.

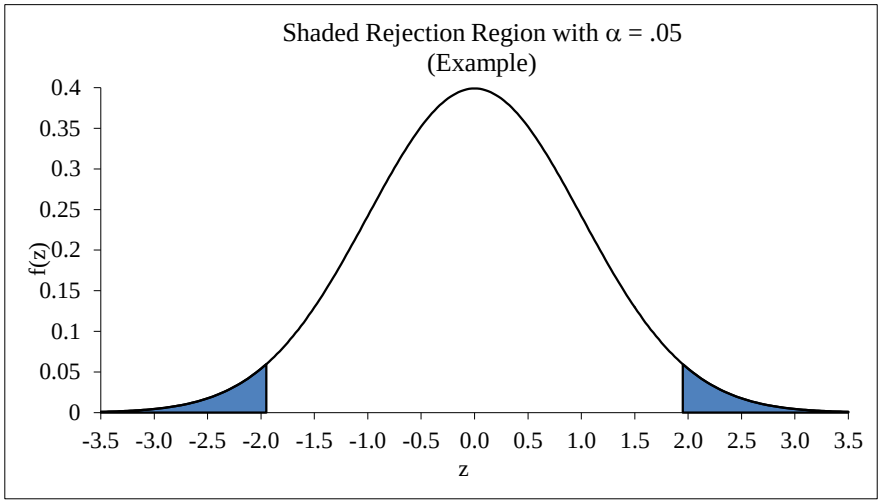
For the first test we are doing which is a two sided test of the mean assuming we know the value of σ, the test statistic will have a standard normal distribution and we would not expect the value we got to be in the tails. Rather we would expect to get a value in the center as this is where all the probability is.

If we get a value in the tails we could not be sure that it did not come from this distribution but we would think it unlikely.

Of course if it did come from this distribution and we rejected the null we would be committing a Type I error. We already said that we were willing to accept α probability of committing a Type I error so we put α probability in the rejection region.

Since the rejection region for our first example will be in the two tails we split α between the two tails. If α is .05 we put .025 probability in each tail.

The graph below shows an example of the rejection region for this test:



Recall that 95% of all values are between -1.96 and 1.96 for a standard normal random variable. That is why we reject the null if we get a value less than -1.96 or greater than 1.96 (i.e., a value in the shaded regions above).

This is called the critical value approach as we come up with critical values (-1.96 and 1.96 in the example above) and reject the null if the value we got for our test statistic is less than the smaller of these or greater than the larger of these. (Note that this is for a two sided test. For a one-sided test we would only have one critical value.)

Another way to look at it is to calculate the p-value.

The p-value for a hypothesis test is the actual chance you are wrong if you reject the null. If the p-value is greater than α you reject the null, if it is less than α you fail to reject.

The p-value will be the probability that would be in the rejection region if the border of the rejection region was at the actual value we got for our test statistic.

Using the p-value is preferable in some ways as it does not require you to specify a value of α. It calculates the value of α that would have made you indifferent between accepting and rejecting the null and then allows any observer to draw their own conclusion.

Back to our test.

H_0: $\mu = 37$

H_1: $\mu \neq 37$

For a test of the mean with σ known, the test statistic will be

$$Z_{STAT} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

By the CLT this will be approximately N(0,1).

Now we have a value for the sample mean (36.92 in this example), we know σ by assumption (.2 in this example) and we have n (the sample size of 36).

All we need is a value for μ in this formula.

But we said that we would assume the null hypothesis was true. If this is true then $\mu = 37$ so we have everything we need to calculate the value of the test statistic.

(Recall the discussion of why the null had to include equality and that we would assume equality in running the test – now you can see why.)

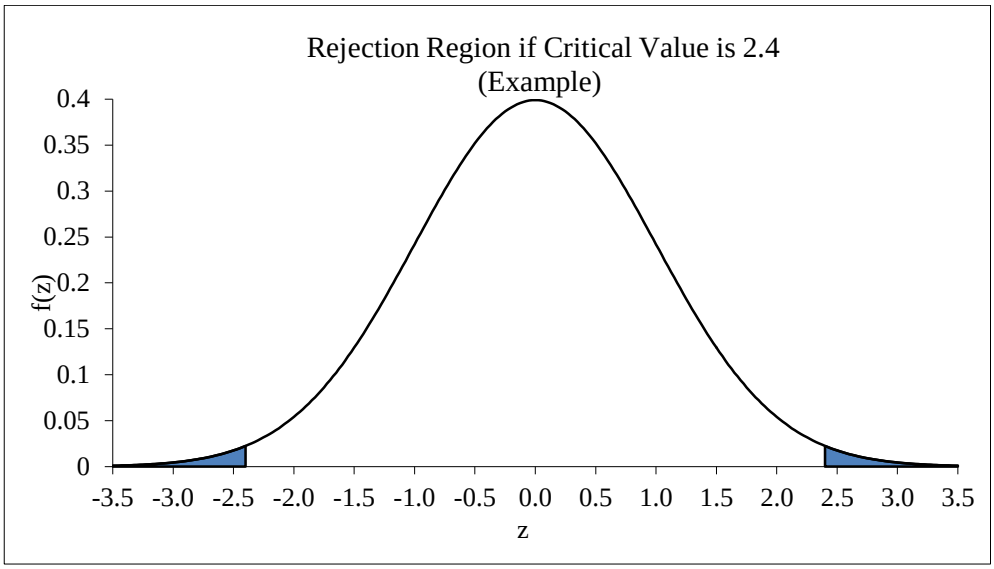
The value of the test statistic will be

$$Z_{STAT} = \frac{36.92 - 37}{.2/\sqrt{36}} = \frac{-.08}{.2/6} = -2.4$$

With $\alpha = .05$ we would reject if the value of our test statistic is greater than 1.96 or less than -1.96. Here the test statistic value is less than -1.96 so we reject the null and conclude that $\mu \neq 37$.

To calculate the p-value we look at what the rejection region would have been if the border of the region had been at -2.4 and 2.4. The picture on the next page shows this scenario.

The p-value will be the probability that would be in this rejection region. From the Z-table we can calculate this as (2)(.0082) = .0164 so we would reject the null for any α greater than .0164 and we would fail to reject for any α less than .0164.



This was a two sided test as we would reject if our test statistic value was in either of the tails of the distribution.

We also have one-sided tests where we reject only in one of the tails.

Let's imagine another scenario. Say the bat manufacturer can live with billets that are too long (there will just be extra material that can be trimmed off) but cannot work with billets that are short.

The risk now is in the billets not being 37 inches on average and the manufacturer would want to know if there was evidence that the billets were less than 37 inches on average.

The appropriate test would be

$$H_0: \mu \geq 37$$

$$H_1: \mu < 37$$

The test statistic will be the same as the previous test but the rejection region will only be in one tail. Do you see which tail it will be? Think about what values would give you evidence that the null was false and the alternative was true – if you got something in the right tail would you think the alternative was true? What about the left tail?

Since the rejection region will be in only one tail the critical value will be different as we will put all of α in the one tail rather than splitting it between the two tails.

If we got the same sample mean (36.92) for a sample of the same size (n=36) the test statistic value would again be -2.4.

With $\alpha = .05$ we would reject if our test statistic value were less than -1.645, which it is, so we would reject the null and conclude that $\mu < 37$.

The p-value would now be .0082 as we would again have a rejection region only in one tail. The p-value for a one-sided test will be half what the p-value would be for the equivalent two-tailed test.

We could also have another one-tailed test where our alternative hypothesis was that $\mu > 37$:

$$H_0: \mu \leq 37$$

$$H_1: \mu > 37$$

Again, the test statistic will be the same but the rejection region will only be in one tail. Can you see which tail it will be?

If we were running this test and we got a sample mean of 36.92 we would never reject the null. Do you see why?

Caution: you can see that we might have different hypotheses for the same data. **You need to decide what your hypotheses will be before you look at the data**. You can't look at the data, see what you can prove, and then specify your hypotheses to match what you can prove. That is considered cheating. The hypotheses should be driven by what you want to show not by what you can show.

In all of the above we controlled for the probability of a Type I error. Type I error is committed when you reject the null and the null is true. It is sometimes referred to as false rejection.

We controlled for Type I error by specifying a value of α which was the probability of a Type I error we were willing to accept. The test we ran will falsely reject the null (100)(α) percent of the time

There is another kind of error we could make, we could fail to reject the null when it is false. This is called Type II error. The symbol β is used for the probability of a Type II error

Type II error is difficult to evaluate. The problem is that it depends on the actual alternative value. If you notice, the alternative hypothesis is always a range of values. Which one is it really? In the one-sided test above where the alternative is that $\mu > 37$, if the alternative is really that $\mu = 45$ there will not be much of a chance of that we commit a Type II error. This is because it will be unlikely that we get a value from the alternative distribution and think it came from the null distribution. If on the other hand the alternative is that $\mu = 37.01$ we will be relatively likely to commit a Type II error as if we get a value from the alternative distribution it will be difficult to distinguish it from a value from the null distribution.

What we can do is determine the value of β for each possible value that the alternative can be. This gives rise to the power function of the test which is a statement of β as a function of specific alternative values.

What you need to keep in mind about α and β is that they tend to move in opposite directions. For a given sample size, if we reduce the possibility of a Type I error we will necessarily increase the possibility of a Type II error and vice versa.

So far we looked at only one test, the test of the mean with σ known. We can do other tests but they follow the same approach. What will be different is the specific test statistic and the distribution of the test statistic and then because these two are different the specific rejection region will be different. Everything else is pretty much the same.

Relationship of CIs to Hypothesis Tests

If you notice, the test statistic we use for the hypothesis test of the mean with σ known is what we started with when we derived the CI for the mean with σ known. Also, the idea of the significance level, α, is the chance we are willing to accept we are wrong in both; for a CI this would be that the interval did not include μ, for a hypothesis test it would be Type I error.

It turns out that if you would reject the null at a level of α in a two-tailed test of the mean with known σ then a 100(1-α)% CI of the mean with known σ would not include that same hypothesized value of μ.

This applies to all the two-sided tests: rejecting the null in the hypothesis test is equivalent to the corresponding CI not including the hypothesized value and failing to reject the null is equivalent to the interval including the hypothesized value.

Test of the Mean with σ Unknown

If we don't know the value of σ we have to use something else in its place. We did this before with confidence intervals and we will replace σ with the sample standard deviation again here.

Our test statistic will now be

$$t_{STAT} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

which has a t-distribution with n-1 degrees of freedom if the underlying population is normally distributed.

The test now proceeds the same but with this new test statistic and test statistic distribution.

Say we have the same example where the manufacturer claims that the billets are on average 37 inches but now we don't know the standard deviation. Assume that the billet lengths are normally distributed.

We will do the one-tailed test where we were concerned about the possibility that the billets were on average less than 37 inches.

$$H_0: \mu \geq 37$$

$$H_1: \mu < 37$$

We would still take a sample and assume again that n = 36 and the sample mean is 36.92 inches. We would also now calculate the sample standard deviation, S. Assume that S = .25 inches.

With α = .05 we would reject the null if the test statistic takes on a value less than -1.6896.

Our test statistic value would be $\frac{36.92-37}{.25/\sqrt{36}} = -1.92$

We can calculate the p-value using Excel. There are a few different functions you can use, one that will work is =TDIST(x value, d.f., tails) where x value is the absolute value of the test statistic value (the function only works with positive values), d.f. is the degrees of freedom and tails is the number of tails (1 or 2) in the test.

For this example we would calculate =TDIST(1.92,35,1) = .0315.

We usually can't figure out the p-value from the tables as the value we want will usually not be in the table.

For this example the best we could do using the table would be to bound the p-value. For a t-distribution with 35 degrees of freedom, the table tells us that 5% of the probability is less than -1.6896 and 2.5% of the probability is less than -2.0301. Since our test statistic took on a value of -1.92 which is between -1.6896 and -2.0301, this tells us that the p-value will be between .025 and .05. We can't really tell more than that from the table.

Z Test of the Population Proportion

For this test our test statistic will be

$$Z_{STAT} = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Where p is the sample proportion and π is the population proportion.

Now our null hypothesis is going to be that π is equal to some specific value and our alternative hypothesis will either be that π is not equal to that value (a two-sided test) or that it is less than or greater than that value (two different one-sided tests).

Say a particular state is going to vote on whether to allow casino gambling. You are a local politician who is generally in favor of casino gambling but you want to determine if the voting public supports casino gambling before you announce your support.

You will poll a sample of voters and determine a value for p, the sample proportion, and you want to know if this provides evidence that $\pi > .5$

You would run this hypothesis test:

H_0: $\pi \leq .5$

H_1: $\pi > .5$

Say you poll 400 voters and 216 of them support casino gambling.

Does this provide evidence that $\pi > .5$?

We will have the sample proportion, p, equal to 216/400 = .54. We assume the null is true, and consistent with our earlier practice, take the value of π to be .5 when calculating the value of our test statistic.

The test statistic value will be

$$Z_{STAT} = \frac{.54 - .5}{\sqrt{\frac{.5(1-.5)}{400}}} = \frac{.04}{\frac{.5}{20}} = 1.6$$

If we use the critical value approach we have to choose a value of α. Since this is a one-sided test we would put all of α in one of the tails, in this case we would put it all in the right tail.

If α = .05 our critical value would be 1.645 as the probability a standard normal random variable is greater than 1.645 is .05. In this case we would fail to reject the null as 1.6 < 1.645.

If α = .1 our critical value would be 1.28 as the probability a standard normal random variable is greater than 1.28 is .10 (technically it is .1003, we are using the closet value from the table). In this case we would reject the null as 1.6 > 1.28.

This highlights the shortcoming of the critical value approach. With some reasonable α we would reject while with other reasonable values we would fail to reject, but the result of our test should not depend so importantly on the value of α.

The p-value will be P(Z > 1.6) = 1- P(Z < 1.6) = 1 - .9452 = .0548 so we will reject for any α < .0548 and fail to reject for α > .0548.

Notice that if we report the p-value (which is commonly done), a reader can make their own determination of whether the results are statistically significant or not.

Statistical Vs Practical Significance

The preceding all deals with statistical significance: can we show that a parameter is statistically different from some value?

If we can say this what we are saying is that yes we realize that our sample is just one picture of what is happening but the results are strong enough that we think they are not just due to random chance (that is, the randomness in sampling.)

You also have to consider practical significance. Is the difference you see of any practical importance?

It turns out that any difference, no matter how small, can be shown to be statistically significant with a large enough sample. But that does not mean it is practically significant.

Say you work for a large auto insurance company.

You want to claim that your firm is less expensive than the competition.

You calculate your average policy cost to be $754 but you can't know the competition's policy costs so you take a sample.

The sample mean for the competitor is $756.

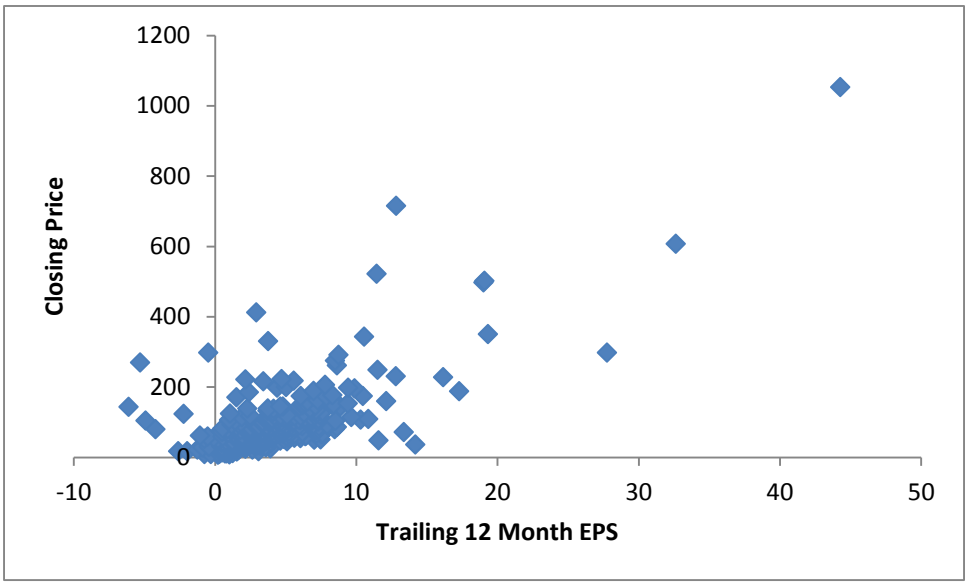
If the sample size is large enough you could show this $2 difference to be statistically significant. But so what? No one would care, everyone would think of your average costs as being the same.

Regression

One of the charts we looked at was a scatterplot.

A scatterplot gives a visual representation of the relationship between two variables.

Previously we drew the scatterplot of the closing price and the trailing 12 month earnings per share for the S&P 500 stocks.



Looking at the scatterplot we said there was an increasing relationship that looked roughly linear or maybe slightly quadratic.

We also calculated the coefficient of correlation which was .74.

Recall that the correlation gave a measure of the strength of the linear relationship between the two variables.

Now in regression we want to identify the form of the relationship. We will do linear regression, but as we will see later, linear regression can also sometimes be used to fit non-linear relationships.

Simple Linear Regression

We will start with simple regression.

We think that one variable (the Y variable) depends on another variable (the X variable.)

Y is called the dependent variable

X is called the independent variable

Simple regression refers to the scenarios when we have only one independent variable. Later we will discuss multiple regression in which we can have more than one independent variable.

We will do linear regression so we assume a linear model:

$$Y = \beta_o + \beta_1 X$$

You should recognize this as the equation of a line. We used β_0 and β_1 for the intercept and slope respectively rather than b and m because the intercept and slope are parameters and consistent with our convention we use Greek letters for parameters.

There is a problem here. This is the equation of a line and, if this model were true, it would imply that all values fell on a straight line. If you look at pretty much any data set (other than carefully constructed examples) you will see that the points do not all fall on a line.

The solution is to add an error term, ε, to the model[9]:

$$Y = \beta_o + \beta_1 X + \varepsilon$$

This is the population regression line.

We can interpret the error term as a fudge factor, essentially a factor that makes the model work, but there is a better way to think about ε.

Say we are trying to develop a relationship between crop yield and rain fall. Over some range we might think this relationship is linear. (Over some range as we might think that more rain led to better crop yield up to a point. If there was too much rain we might see yields decrease.)

So if we had data on crop yield (the Y variable) and rain fall (the X variable) we might think there was a linear relationship between the two, something like $Y = \beta_o + \beta_1 X$. Now as we pointed out above if we collected data on crop yield and rain fall we would not expect the data to fall exactly on a line.

There is a good reason they would not, though. Crop yield presumably depends on rain fall but it probably also depends on many other factors, e.g., sunlight, fertilizer, ph balance of the soil, type of seeds used, etc. None of these are in our model. The effects of these other factors will cause the data to not fall exactly on a line and we use the error term, ε, to account for all these factors not in the model.

We don't need to specify the missing factors, by the way, in any case it would be unlikely that we would know them all, we just know that ε accounts for whatever is not in the model.

This is not to say that we want to have missing factors. We certainly want to include everything that is relevant but there may be factors about which we are unaware and there may be factors for which we can't collect data even though we are aware of them.

So we have our population regression line: $Y = \beta_o + \beta_1 X + \varepsilon$

[9] The error term is the Greek letter epsilon which is commonly used to represent error or other small quantities.

Now we are right back where we have been all along in statistics: we want to know something about a population, in this case the slope and the intercept of the population regression line, but it is infeasible to collect data on the entire population so we take a sample. We then estimate the slope and intercept which define the sample regression line.

To estimate the parameters we use least squares regression. Least squares regression finds the line that minimizes the sum of the squared error terms.

Why do you think we minimize the sum of the squared error terms?

The calculations can be done in Excel or some other software package. They can in theory be done by hand or with a calculator but that is not recommended.

Least squares regression requires four assumptions:

1) The relationship is linear
2) The errors are independent
3) The error is normally distributed
4) The variance of the errors is constant (does not vary with the level of the independent variable).

These assumptions can be recalled with the mnemonic LINE (for Linear relationship, Independent errors, Normal errors, Equal variance).

Using least squares regression, we estimate β_0 and $\beta_1$and get the sample regression line:

$$\hat{Y} = b_0 + b_1 X$$

The "hat" (^) on Y denotes that this is an estimator of the independent variable Y and not Y itself. $\hat{Y}$ is read "Y hat".

So part of what we want to do is calculate the values of the estimators b_0 and b_1.

If the four assumptions above are met then the estimators b_0 and b_1 will be the minimum variance unbiased estimators (MVUE) of β_0 and β_1 respectively.

We will also want to determine whether the assumptions of least squares regression are valid (for much of this we will look at the error values called the residuals) and we want to try and determine how well the data fits our model.

When we do a regression we will get a lot of data concerning the relationship that will let us evaluate the assumptions and our model.

We will use *RegressionExamples.xlsx* for examples.

We will start with the Hotel Data worksheet. The data in this worksheet address a similar problem as the Chopotle example we started the course with.

The data is for the La Quinta chain and tells us the margin for each of 100 hotels as well as the values for a number of other variables for these 100 hotels.

We will start by regressing Margin against Office Space.

Which will be the independent variable and which will be the dependent variable?

The coefficients are in the output.

Our estimated line is

$$\hat{Y} = 34.19 + .023X$$

The slope tells us the rate of change.

It gives us the increase, on average, in the Y variable for a 1 unit increase in the X variable.

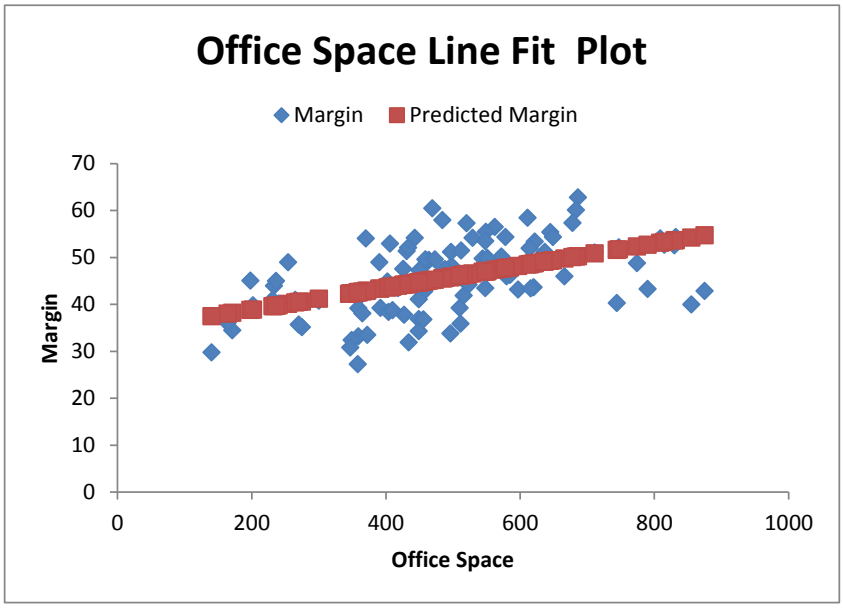
So on average, an extra 1000 square feet of office space increases margin by .023%

You have to be careful interpreting the intercept. There is a temptation to say that the intercept is the value of the Y variable you would predict if the X value took on a value of zero.

The problem with this is that the regression is only good, if it has any validity at all, within the range of the data. We don't have any X values anywhere close to zero (this is not surprising as we would probably not put a hotel in an area so remote that it had no office space) so we don't know what would happen to margin if there was no office space close by.

You can always say that the intercept tells you where the line would hit the axis (essentially it is an anchor) but unless you have data for values of X around zero you can't really say more.

The line fit plot is:



The blue diamonds are the original data and the red squares are the predicted values based on our sample regression line. This gives a visual representation of the fit of the line.

We can also quantify this relationship.

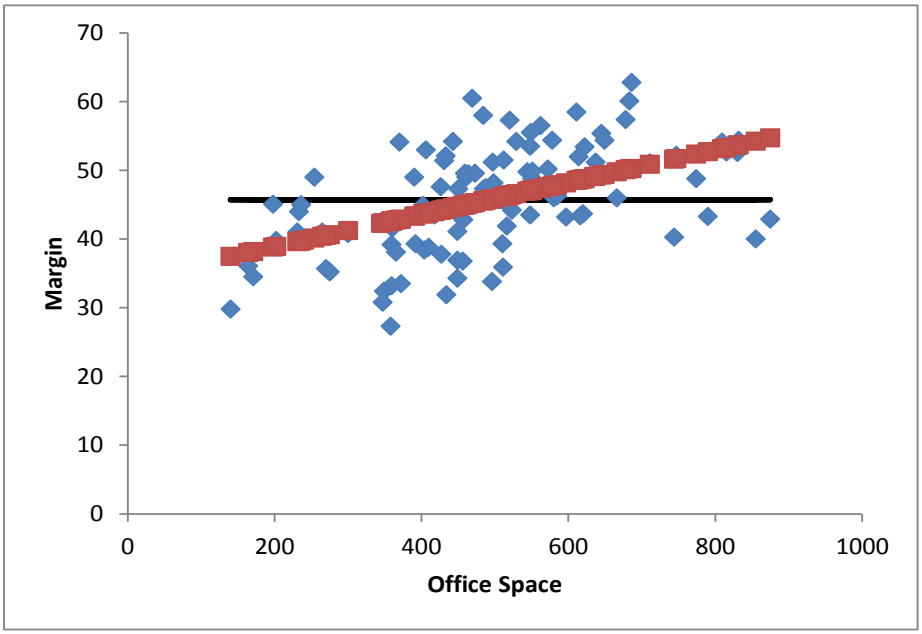
The multiple R is .501 – what Excel calls multiple R is what we called just r, the sample correlation. Recall that the correlation is a measure of the strength of the linear relationship between two variables.

The R^2 is .251. R^2 is more formally called the coefficient of determination although most people just refer to it as R-squared. R^2 happens to be the square of the correlation but it has a more important interpretation.

An R^2 of .251 tells us that 25.1% of the variation in the dependent variable is explained by the independent variable.

If you think about it, our goal is to explain variation. That is, the margin varies across the different hotels and we want to know why. Does it depend on the office space in the area, or income levels in the area, or maybe manager ability, or something else? The coefficient of determination is telling us how much of the variation we are able to explain with our current model.

In class we will use the chart below to discuss what we mean when we say variation is being explained.



The sum of the squared distances from the original data (blue diamonds) to the predicted values (red squares) is the error sum of squares or SSE. This is the unexplained variation.

The sum of the squared distances from the predicted values (red squares) to the mean margin (black line) is the regression sum of squares or SSR. This is the explained variation.

The total sum of squares or SST = SSR + SSE

The coefficient of determination is the ratio of the regression sum of squares (explained variation) to the total sum of squares (total variation).

Towards the top of the output in the Regression Statistics section we also have the Standard Error of 6.74. This is the standard error of the estimate and it is a measure of the variation around the sample regression line. The units are the same as the units of the Y variable (margin in this case).

In the rows for intercept and office space there is more useful information than just the values of the coefficients.

First, these are sample values, hence statistics, hence random variables. The value of the intercept and the slope of the sample regression line are point estimates for β_0 and β_1. Confidence interval estimates for β_0 and β_1 are given at the ends of the rows that have the point estimates. By default Excel calculates the upper and lower limits for 95% CIs and there is an option to calculate the limits for another CI.

The standard error of each of these statistics is also given (in fact the confidence intervals are based on the standard errors).

Also in the row for Office Space there is a t stat of 5.74 and a p-value of 1.07E-07

These are the test statistic value and the p-value for a test of

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Why is this test important?

It turns out that if you tested a null that $\rho = 0$ versus a two sided alternative you would get the same p-value. (In fact the tests are equivalent.) Do you see why that is?

In the line for intercept there is a t Stat of 16.10 and a p-value of 2.78E-29.

These are the test statistic value and the p-value for a test of

$$H_0: \beta_0 = 0$$

$$H_1: \beta_0 \neq 0$$

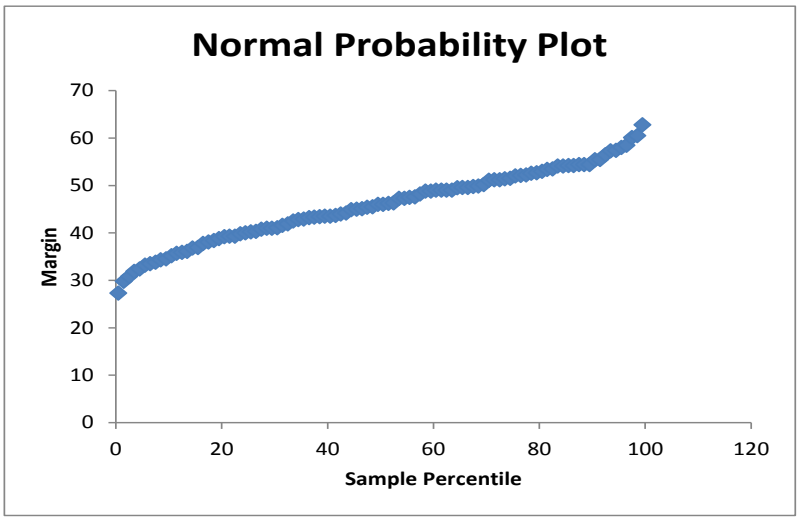
Sometimes this is important to us. Why might this be of importance?

As we said earlier we also want to determine whether the OLS assumptions hold. To check this we look at the residuals (error terms).

We assumed that the error terms were independent. This is a potential issue when the data are a time series as often there is correlation between consecutive error terms (referred to as autocorrelation). To check this you can plot the error terms in the order the data was collected (i.e., by time). If you see a

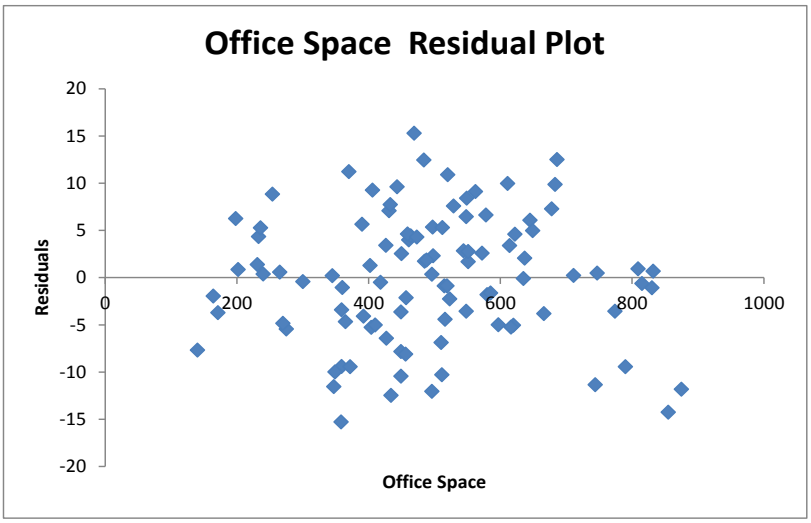
cyclical pattern there is evidence that the error is not independent. You can also calculate the Durbin-Watson statistic to test whether there is autocorrelation. Since the Hotel data was collected at the same time this is not a concern in this case.

We also assumed that the error was normally distributed. To check this you can look at a Normal Probability Plot which can be created by checking off the appropriate box in the regression tool. This is the normal probability plot for the office space regression:



This is a plot of the quantiles of the residuals versus the quantiles of a normal distribution. If the quantiles match (implying the residuals are normally distributed) the plot will be a straight line. The plot above is not exactly a straight line but it is close (the "bends at the ends of the lines imply the quantiles do not match up in the tails) but OLS is relatively robust to this assumption so this is not a great concern.

To check the assumption of equal variance we look at a plot of the residuals against the X variable (office space). This plot, which is also created by Excel, is shown below.



We are now looking for a cone shape which would suggest that the variance is not constant across the values of the X variable. If the residuals roughly sketch out a rectangle, which they do in this case, we consider this assumption to be met.

Multiple Regression

We always have only one dependent or Y variable.

We can have more than one independent or X variable. Multiple regression refers to the scenario in which we have more than one independent variable.

Now our population model with k independent variables will be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

The intercept is still β_0, the β_1, β_2, …, β_k terms are the coefficients of the independent variables and ε is still the error term which still accounts for everything not in the model.

We use Excel to estimate the sample regression line:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

Note that in Excel the independent variables have to be in contiguous (consecutive) columns.

Use the Hotel Data again. We will use the same independent variable but now we will use all the other variables as X variables.

We want to find the best model.

We could start with one possible X variable and then add additional independent variables or we could start with everything and then exclude what didn't belong.

You can do either but it turns out it is better to start with everything.

If you start with everything, at first (before you eliminate unnecessary variables), you might include variables that don't belong.

This is not good as it means you have to estimate coefficient for these variables that don't really belong in the model. The error goes up with every additional coefficient you have to estimate, which means you will have more error than necessary, but the estimators will still be unbiased.

If you start with one and build up, until you get to the right set of variables, you will be missing variables.

This is a bigger problem as the estimates you have will be biased if there are missing variables.

So we usually prefer to start with everything and eliminate variables.

Note that there is also a technique known as stepwise regression which is an automated approach different from what we will discuss here (it varies in how it decides to exclude or include variables). That technique either starts with one variable and adds variables (known as forward stepwise regression) or starts with everything and removes variables (backward stepwise) or a does combination of the two (mixed step-wise).

Using all the variables, we get a sample regression line of:

$$\hat{Y} = 38.14 - .0076(Number) + 1.65(Nearest) + .020(Office\ Space) + .41(Income) + .21(Enrollment) - .23(Distance)$$

The coefficient for each X variable tells us the change, on average, for a 1 unit increase in that X variable, holding all the other X variables in the model constant.

This is what lets us control for other variables.

If office space increases we would expect margin to increase but we might also expect Number (which measures competition) to increase also which we would think would cause margin to decrease.

We want to tease out the individual effects, what happens if just office space increases? This is what the coefficient of office space tells us in a multiple regression.

Now that we have multiple independent variables it can make sense to test whether the entire set of independent variables exhibits a linear relationship with the dependent variable. This would be a test

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_1: \text{At least one } \beta_i \neq 0,\ i = 1, 2, \ldots, k$$

So the null is that all β_i are zero and the alternative is that at least one is non-zero. Essentially this is a test that something (at least one of the independent variables) matters.

The p-value for this test is calculated by the regression tool in Excel. It is what Excel calls the significance F (the test statistic for this test has an F distribution).

The regression tool reports the significance F for a simple linear regression also. When there is only one independent variable, the significance F will equal the p-value in the β_1 (slope) output row. Do you see why this is?

For a multiple regression the p-value in the intercept row has the same interpretation as it did for the simple linear regression.

Now the p-values in each row for the X variables are tests of

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

for each specific β_i (i.e., each coefficient).

Note that some of these p-values are above .05, what does this tell us?

It means that we can't conclude that these coefficients are significantly different from zero which may imply the variables do not belong in the model. Should we remove these variables?

To answer this we need to look at the adjusted R^2

The adjusted R^2 adjusts the R^2 for the number of independent variables in the model and for the sample size used in estimating the coefficients.

The adjusted R^2 is an unbiased estimator of the population R^2 and we have to use the adjusted R^2 if we want to compare models with different numbers of independent variables.

As discussed before, R^2 tells us how much of the variation in the Y variable is explained by the X variable(s). We want a higher R^2.

The problem is that any time you add a variable R^2 will, in practice, go up. (Technically it does not have to go up, but it cannot go down and it will pretty much always go up even if just by a very small fraction.)

Since we want a higher R^2 this implies we should just keep adding variables.

As discussed above, though, every time we have to estimate a new coefficient our error goes up so we don't want to just add variables if they are not going to help very much.

The R^2 will go up with every new variable but the adjusted R^2, which tells us how much of the variation is explained by the variables in the model adjusting for the number of variables and the sample size, will not always go up.

If we add a variable and the adjusted R^2 goes up, it tells us that the additional error introduced by having to estimate one more coefficient is justified by the increase in explanatory power of the model.

If the adjusted R^2 goes down, it tells us the opposite, that we cannot justify including the variable.

Consider the Salary Data worksheet.

Use Y = Salary and X = Age

Note that the variable Age looks significant.

Now use Y = salary and X = all variables in columns B through J

Adjusted R^2 = .7198; R^2 = .7320

Now rerun but exclude Age as a variable

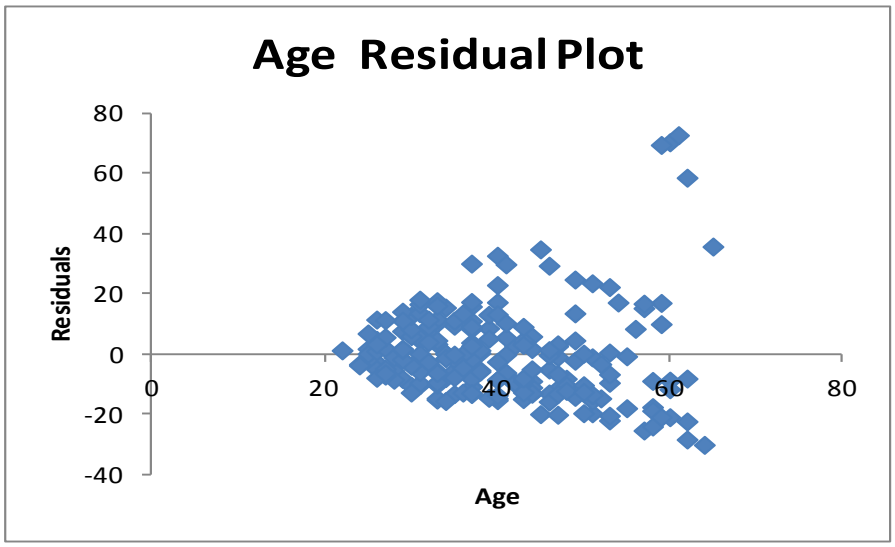
Adjusted R^2 = .7210; R^2 = .7318

So with one less variable (the second set of values) the R^2 was lower (as we would expect with fewer variables) but the adjusted R^2 was higher. This suggests that Age does not belong in the model.

So why does Age look significant by itself but not when we add other variables?

Hint: what relationship would you expect to see between Age and the other variables?

Look also at the residual plot for the regression with just Age as an independent variable. Recall that we use this plot to determine if the error terms have constant variance.



Note that it looks like a cone. This is bad and it is what we said we wanted to look out for when we evaluated the residuals.

It says that the model fits better at the low end of the Age range than it does at the high end of the range. It would be equally bad if the cone went the other way, that is, if the fit was better for older people than it was for younger ones.

This suggests that the assumption of equal variances is violated.

As we said before what we want to see is the points on the residual plot roughly sketching out a rectangle.

So why do you think the model with Age as an independent variable fits better for younger workers?

Dummy or Indicator Variables

If you look at the salary data you will see some data that is categorical. For example, one of the data series is the gender of the employees. This is male or female hence is not a number.

Since we can't use non-numeric or categorical variables in a regression we use dummy or indicator variables.

For gender we use a gender dummy variable. We assign a value of 0 to this variable if the employee is male and 1 to it if the employee is female (note that we could have used 1 for males and 0 for females if we wanted).

If we have a categorical variable that has more than two categories we need to use more than one dummy variable. We need one less dummy variable then we have categories. So male-female with two categories requires one dummy variable. If we had data on college students categorized as first years, sophomores, juniors and seniors (four categories) we would need three dummy variables.

Fitting Non-Linear Models

We can also use linear regression to fit non-linear models.

There is a thing called the Phillips curve that posits that there is a relationship between inflation and unemployment.

Let Y = inflation

If X = unemployment we can use regression to fit the simple population model we have been using:

$Y = \beta_0 + \beta_1 X + \varepsilon$

But what if we think that inflation depends on the inverse of unemployment?

In this case, still using X as unemployment, our model might be

$$Y = \beta_0 + \beta_1 \left(\frac{1}{X}\right) + \varepsilon$$

which is not linear. We can use a change of variable to make it linear, however.

Let $W = \frac{1}{X}$ so for example, in Excel, set up a new column defining W as the inverse of X.

To Excel this looks like any other column of data so can be used in a linear regression.

Now why might we want to do this?

It might give us a better model, that is, the fit could be better.

But we also might think that it is a more reasonable model.

With a linear relationship the slope is the same for all values of X. This means that the change in Y for a 1 unit increase in X will not depend on the level of X.

Sometimes this does not make sense.

If we use a non-linear model the change in Y will depend on where we are on the X axis and this might be more reasonable in a particular situation (as it probably is for inflation and unemployment).

We can use this same approach to use the techniques of linear regression to model other non-linear relationships. For example if we wanted to model a quadratic or higher order polynomial we can create new columns in Excel with X^2 or X^3 or whatever power we wanted and use these columns as independent variables.

If we are doing that we have to be careful, though, to not over fit the model. If we add powers of X we allow the regression line to curve. We can often get a much better fit using powers of X but we have to consider whether there is a reason to believe that this is the appropriate relationship. If there is no reasonable explanation why we would expect to see this relationship then we probably should not use this model even if we get a good fit.

The important point is that linear regression requires linearity in the coefficients not in the independent variables.

Two caveats both of which are demonstrated by the next example:

- Correlation is not causation (which we discussed before)
- Just because we can get a good fit to a model does not mean the model is appropriate

Another Model of Inflation

Economist David Hendry developed the following model of inflation[10]:

$Y = \beta_0 + \beta_1 X + \beta_2 x^2 + \varepsilon$

Notice by the way that he used a quadratic model which he could estimate using liner regression by transforming the variable X.

He estimated β_0, β_1 and β_2 using quarterly data from 1958 to 1975 and got

$$\hat{Y} = 10.9 - 3.2X + .39X^2$$

With all coefficients significantly different from zero and $R^2 = .982$.

What do you think X was?

Lessons:

- Models need to be plausible
- Correlation is not causation.

As a final example we will use regression to talk about something that is covered in Finance.

[10] "Econometrics – Alchemy or Science", Economica, Nov 1980, Vol 47, #188 pgs 387-406

Capital Asset Pricing Model

r = return on stock

r_f = risk free rate (typically treasuries)

r_m = market return

β = beta of stock

$r - r_f = \beta(r_m - r_f)$

The left side of this equation is the expected risk premium of the stock, that is the expected return above the risk free rate.

The right side is β times the expected risk premium of the market.

It says that the expected risk premium of the stock is proportional to the expected risk premium of the market and in fact the β of the stock tells you the ratio.

But what is the β of a stock?

We can re-write the model as:

$r = r_f + \beta(r_m - r_f)$

So the model tells you the return investors expect on an investment in a particular company.

If you have values for r_f , β, and r_m then you can calculate the expected return.

How can this be used?

Well say you are the CFO of a company considering making an investment in your firm, e.g., buying a piece of equipment or developing a new product.

What discount rate should you use?

Why?

Now how to get β?

We can define β as the ratio of the covariance of the stock and market returns to the variance of the market returns:

$$\beta = \frac{covar(r, r_m)}{var(r_m)}$$

We can estimate β using regression.

See the stock returns worksheet.

We need to decide what measure of the market we will use. Typically the S&P 500 is used.

Independent variable will be the S&P 500 returns

Dependent variable will be the stock returns

What is the model?

Should the intercept be zero?

What does the slope tell us?

What about R^2? How do we interpret this?

Careful, there is a special interpretation of the R^2 for this specific regression.