STAT*2060DE
Summary Notes for Unit #3

In this unit we encounter a few very important concepts in statistics, namely *the normal distribution*, *sampling distributions*, and *the Central Limit Theorem*. Although there are many questions regarding probability, they are a little simpler than those of Unit 2.

Many variables we encounter have *approximately normal distributions*, and a great many statistical inference techniques rely on the assumption of a normally distributed population. Some of our inference techniques will require us to find probabilities and percentiles based on the normal distribution.

The concept of a *sampling distribution* is fundamental to all of statistical inference. The idea seems quite basic to some, but many students struggle with this concept. The sampling distribution of a statistic is the probability distribution of that statistic. One way to visualize this is through the repeated sampling argument: If we were to repeatedly draw samples of size n from the population and calculate the mean, say, of each sample, then that sample mean would vary from sample to sample. If we took a great many samples, and plotted out the frequency histogram of the means of the samples, then that histogram would look like the *sampling distribution of the sample mean* for that scenario. In practice we only take one sample for a given problem, but the concept of a sampling distribution is important. The value of the statistic that we obtain in our one sample is simply a randomly selected value from the sampling distribution of that statistic. We usually obtain the sampling distribution of a statistic via mathematical arguments, but in more complex situations we can use simulations to approximate it.

The *Central Limit Theorem* is extremely important to statistical inference. There are formal, mathematical definitions, but the gist of it is that sums and means of random variables have distributions that tend toward the normal distribution as the sample size increases. One implication of this is that the sample mean is approximately normally distributed for large sample sizes, regardless of the distribution of the original variable. For example, consider a strongly right-skewed distribution such as salaries of workers in a company. The distribution of the salary of a single randomly selected person is strongly skewed to the right, but the distribution of the sample mean of 200 randomly selected salaries will be approximately normal, due to the Central Limit Theorem. The Central Limit Theorem is a big reason why we often see normally distributed variables in many practical situations.

In this unit you will be expected to:

- Understand what is meant by a *continuous* random variable.
- Know the basic properties of a continuous random variable.
- Understand and carry out calculations for a continuous uniform random variable.
- Understand and carry out probability calculations for a normally distributed random variable.
- Know what is meant by the term *sampling distribution* of a statistic.
- Understand and carry out probability calculations based on the mean of a sample.
- Know the gist of the Central Limit Theorem, and why it is important to us.
- Be able to properly interpret a normal quantile plot.

On to the more detailed notes:

Continuous Probability Distributions

Continuous random variables take on an infinite number of possible values, corresponding to every value in an interval. Recall that discrete random variables took on a *countable* number of possible values. Continuous random variables will not have a countable number of values.

Common examples of continuous variables include heights, weights, time to an event. Note that for a variable like time there are an infinite number of possible values between 1 second and 2 seconds, an infinite number of possible values between 1.4 seconds and 1.5 seconds, an infinite number of possible values between 1.67824 and 1.67825 seconds, etc. We cannot possibly count up the possible values of time.

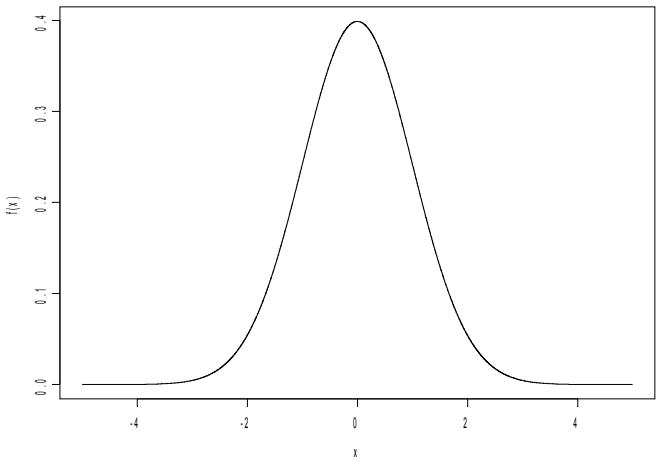
As an example of a continuous random variable, let the random variable X represent the time to the next nuclear weapon detonation on earth. Then X is a continuous random variable, with possible values of $X > 0$.

Let Y represent the amount of water in a randomly selected 500 mL bottle of water. Then Y is a random variable. Then Y is a continuous random variable with possible values of $0 \leq Y \leq$ Maximum capacity.

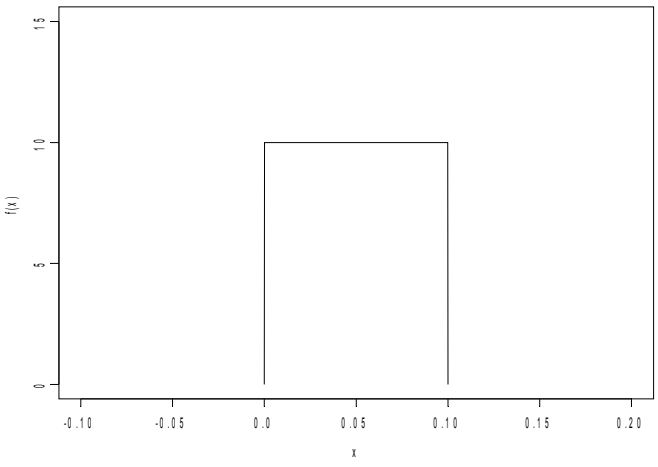
We cannot model continuous random variables with the same methods we used for discrete random variables. Many of the ideas will be similar, but we will have to make a few adjustments. We model a continuous random variable with a curve $f(x)$, called a **probability distribution** or **probability density function** (pdf).

Examples of common continuous probability distributions:

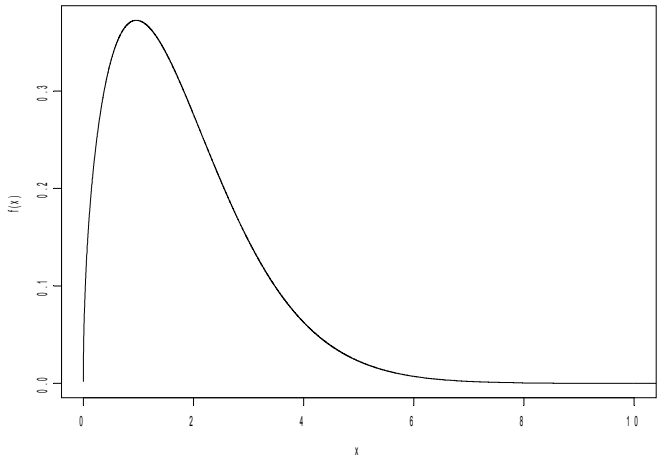
The Normal Distribution:



The Continuous Uniform Distribution:



The Weibull Distribution:

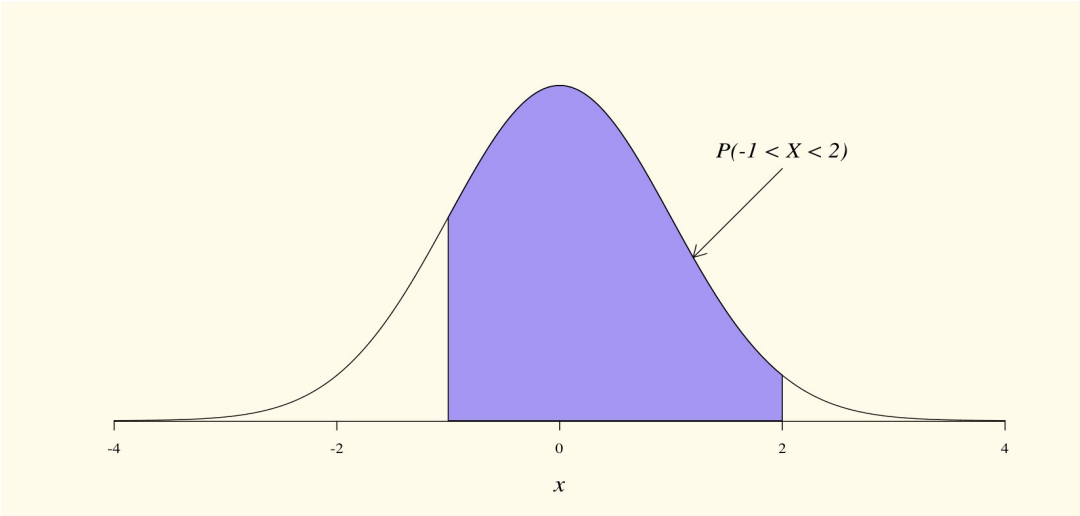


The Weibull distribution is commonly used to model time-to-event data, but we won't discuss it further in this course. The normal distribution is the most important distribution for us in this course, and we will discuss it in detail below. We will also discuss the uniform distribution in a little more detail. But first we will discuss some general properties of continuous distributions.

Properties of Continuous Probability Distributions

- *f(x)* represents the height of the curve at point *x*. The curve cannot dip below the *x* axis. In other words, $f(x) \geq 0$.
- *f(x)* is the height of the curve. It is not a probability. However, the height of the curve helps us find probabilities, since for continuous random variables probabilities are **areas under the curve**. The probability that the random variable *X* lies between two points *a* and *b* is the area under the curve between *a* and *b*.
- The area under the entire curve is equal to one. ($\int_{-\infty}^{\infty} f(x)dx = 1$) This is the continuous analog of the discrete case, in which the probabilities must **sum** to 1.

It bears repeating: For continuous random variables, probabilities correspond to **areas under the curve:**



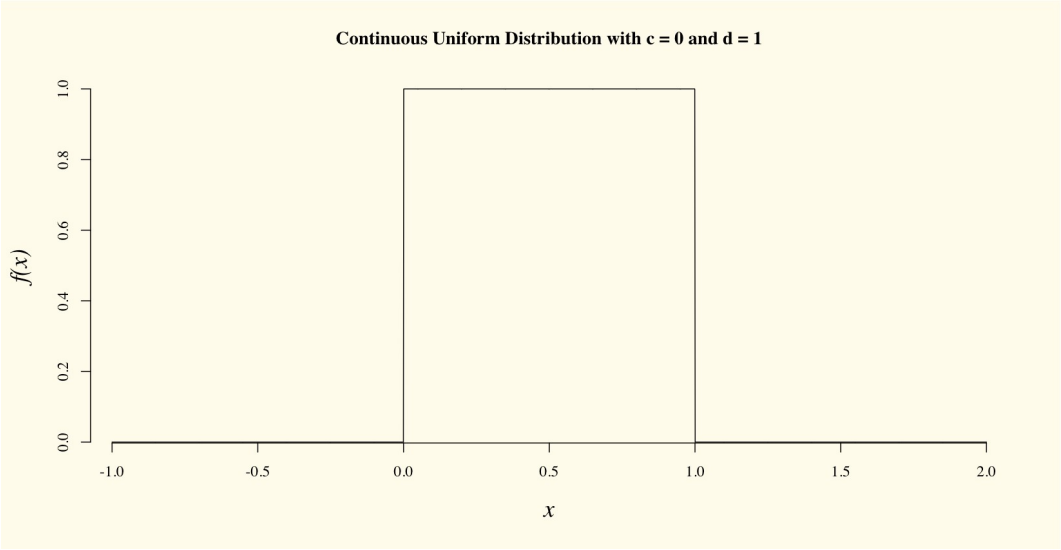
One more (possibly subtle) point: The probability that X is equal to any constant is zero ($P(X = a) = 0$ for all a. Since probabilities are areas under a curve, and any constant a is just a point, it has an infinitesimally small area above it, and thus $P(X = a) = 0$. Note that this is distinctly different from the discrete random variable case. If I shoot 20 free throws, say, then there is some positive probability that I make exactly 3 ($P(X = 3) > 0$). But the probability my kettle takes exactly 287 seconds to boil water is 0. ($P(X = 287) = 0$). Note that here 287 means *exactly* 287, 287 with infinite trailing zeros, 287.000000000000... This can be an important point, and we may have to realize in some cases that for continuous random variables $P(X \geq 5) = P(X > 5)$ since $P(X = 5) = 0$. Note that this holds only for *continuous* random variables and not *discrete* random variables.

The Continuous Uniform Distribution

The simplest continuous distribution is the uniform distribution. For both theoretical and practical reasons it is an important distribution, but we look at it here merely as a simple introduction to continuous probability distribution.

In the continuous uniform distribution, the random variable X takes on values between a lower bound c and an upper bound d, and all intervals of equal length are equally likely to occur.

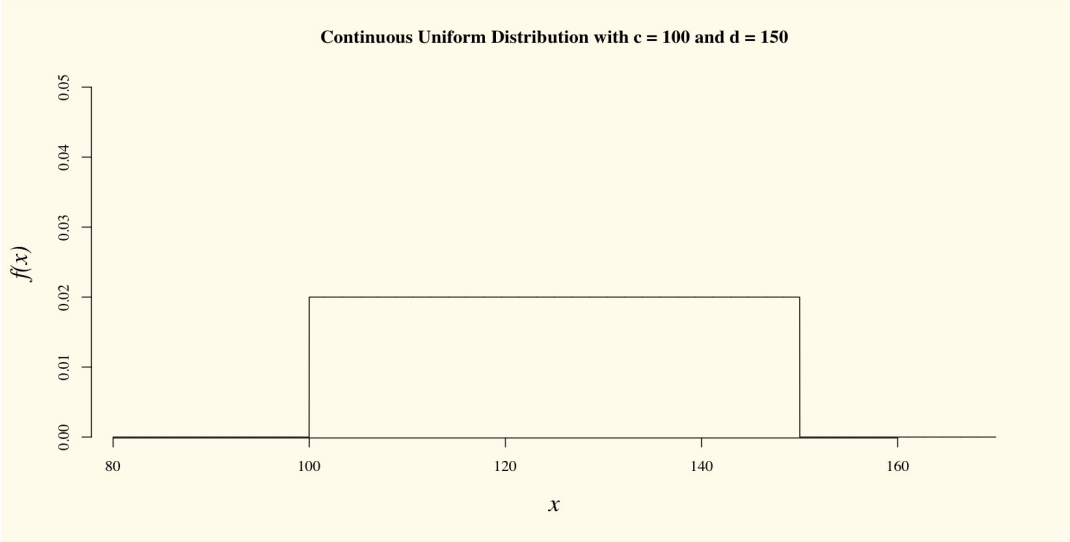
Plot of the most common continuous uniform distribution ($c = 0$, $d = 1$):
(This is sometimes called the **standard** uniform distribution, but that is not overly important to know)



Note that since the area under the entire curve must equal one, in this case the height of the curve at its peak must equal 1. The shape of the continuous uniform distribution will be a simple rectangle (square, in this case), which will make it simple for us to calculate probabilities and discuss some of the basic ideas of continuous probability distributions.

Let's look at a continuous uniform distribution that has bounds that are different from 0 and 1, as that is best to illustrate a few points.

Example. Suppose a continuous random variable X has the uniform distribution, with a minimum of 100 and a maximum of 150:

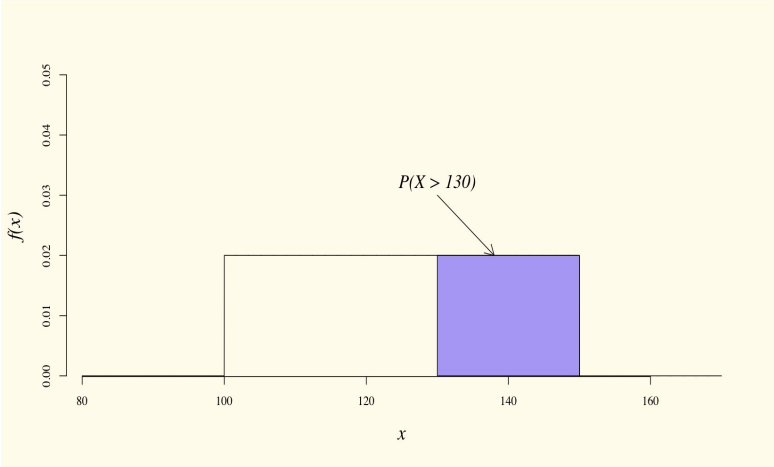


Note that the entire under the entire curve is a rectangle with a base of (150-100) = 50, and a height of $f(x)$. We know that the area under the entire curve must equal 1, and we know that the area of a rectangle is base*height. We can easily find out what the height of the curve must be: $50f(x) = 1$ implies that $f(x) = 1/50 = .02$. More formally, for this distribution,

$f(x) = .02$ for $100 \leq x \leq 150$
$f(x) = 0$ elsewhere.

The random variable X can only take on values between 100 and 150.

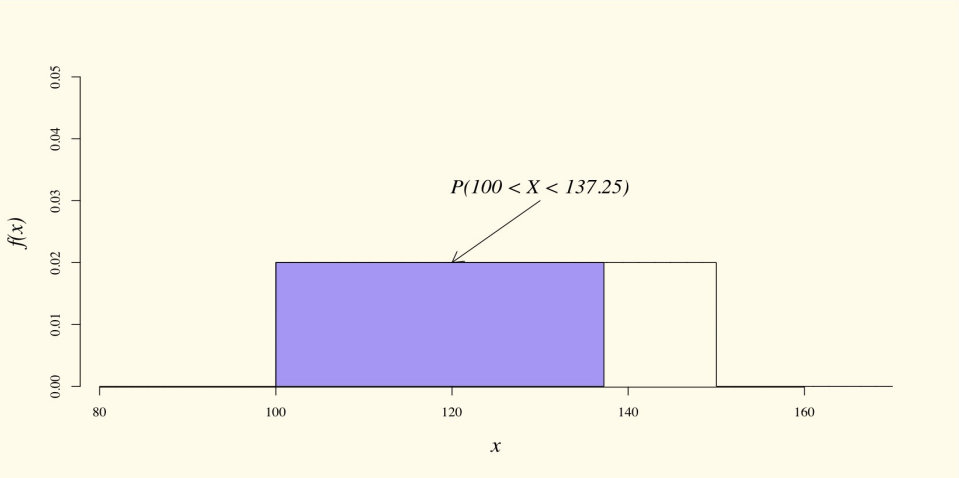
What is P($X > 130$)? Since probabilities are areas under the curve, P($X > 130$) is simply the area to the right of 130:



This area is a rectangle, with a base of 150-130 = 20, and a height of $f(x) = .02$.
Area to the right of 130 = P($X > 130$) = 20(.02) = 0.40.

What is P(85 < X < 137.25)?

P(85 < X < 137.25) = P(100 < X < 137.25), since P(85 < X < 100) = 0 (there is no area under the curve below X = 100).

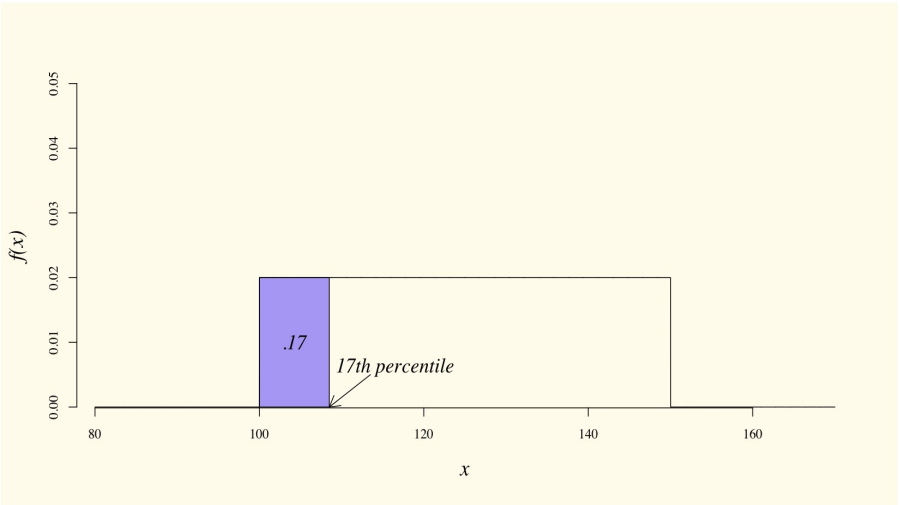


P(100 < X < 137.25) is the area under the curve between 100 and 137.25 = (137.25 – 100) (.02) = 37.25(.02) = 0.745.

What is the median of this probability distribution? Equivalently, what is the mean of the random variable X? The median is the value of the random variable that splits the distribution in half (50% of the area to the left, 50% of the area to the right). For this distribution, that point is 125.

What is the mean? Finding the mean of a continuous probability distribution can be a little tricky, as it often requires integration. However, we can find it out here rather easily, as the distribution is symmetric. For symmetric distributions the mean and median are equal, and thus for this distribution, mean = median = 125.

What is the 17th percentile of the distribution? The 17th percentile is the *value of* x such that the area to the left of x is .17:



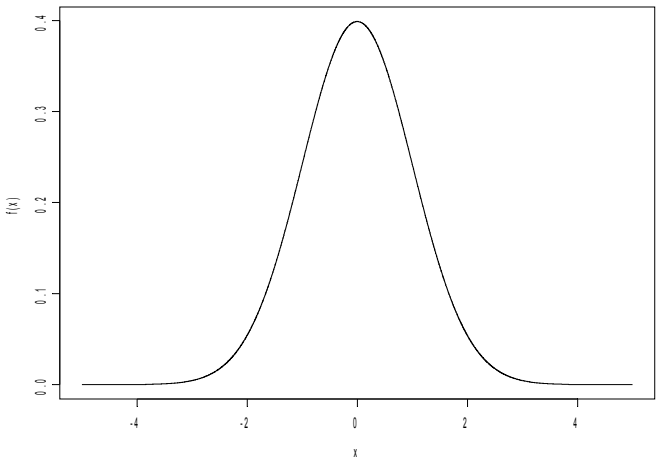
If we let a represent the 17th percentile, then we know the area to the left of a is .17. We also know that the area to the left of a must equal base*height = $(a - 100)(.02)$. $(a - 100)(.02) = .17$ implies that $a = 108.5$. The 17th percentile of this distribution is 108.5.

The calculations for the uniform distribution are very straightforward. The other continuous distributions we use in the course have more complicated density functions. Finding the area under these curves means numerically integrating the density function. Fortunately for us, others have done for that us and put the values in tables. So for our other continuous probability distributions we will have to learn how to look up the appropriate areas in the tables.

The Normal Distribution

The normal distribution is an extremely important continuous probability distribution. Many of our statistical inference techniques are based on the normal distribution. The normal distribution arises very frequently. One reason for this is the **Central Limit Theorem**, but more on this later.

The normal curve is sometimes called "bell-shaped", and it looks like:



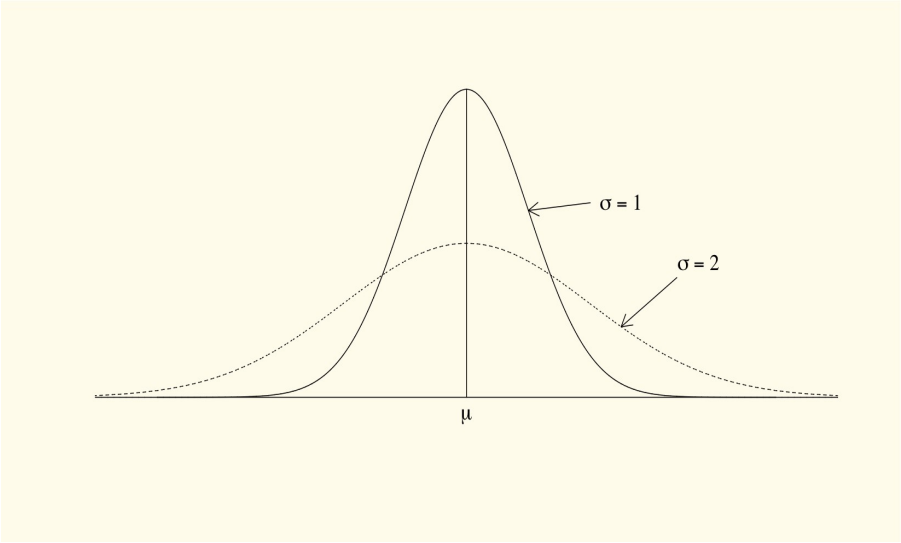
The equation of the curve is $f(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{1}{2\sigma^2}(x-\mu)^2}$. This is the **probability density function** (pdf) of the normal distribution. It is not overly important for you to know this formula; we will be looking up necessary values in a table. The table values are based on this curve.

If a random variable X has the normal distribution, then it can take on any finite value: $-\infty < x < \infty$. Although this may seem to be a problem for modelling variables that only take on positive values like height and weight, it is not a problem in practice. Although there will always be some positive probability that $X < 0$ for a normally distributed random variable, depending on the parameters of the distribution it can be very, very, very close to 0.

The normal distribution has two parameters, μ and σ. μ (the Greek letter mu, usually pronounced

myoo) represents the mean of this probability distribution. Since the distribution is symmetric, the mean and median of this distribution are equal. Mean = Median = μ.

The Greek letter σ (sigma) represents the standard deviation of the probability distribution. The greater the standard deviation, the more spread out the distribution and the lower the lower peak. The following plot shows two different normal distributions. The two distributions have equal means (μ), but different standard deviations.



The normal distribution is symmetric about the mean μ.
The mean μ of a normal distribution can be any value: $-\infty < \mu < \infty$
The variance of the normal distribution must be greater than 0: $\sigma^2 > 0$. The variance of *any* distribution has to be at least 0. But if we have a variance of 0, that means the distribution can only take on one possible value, and we would not have a normal distribution. This leaves us with $\sigma^2 > 0$

There is an infinite number of possible normal distributions, corresponding to the different possible values of μ and σ. But our textbook only contains one normal distribution table, the **standard normal distribution**. The **standard normal distribution** is, by definition, a normal distribution with a mean of μ = 0 and standard deviation of σ = 1. We will soon see that if *X* has a normal distribution with any mean and any variance, we can easily convert it into a random having a standard normal distribution. The standard normal distribution is a very important distribution for us. We will use the letter *Z* to represent random variables that have a standard normal distribution.

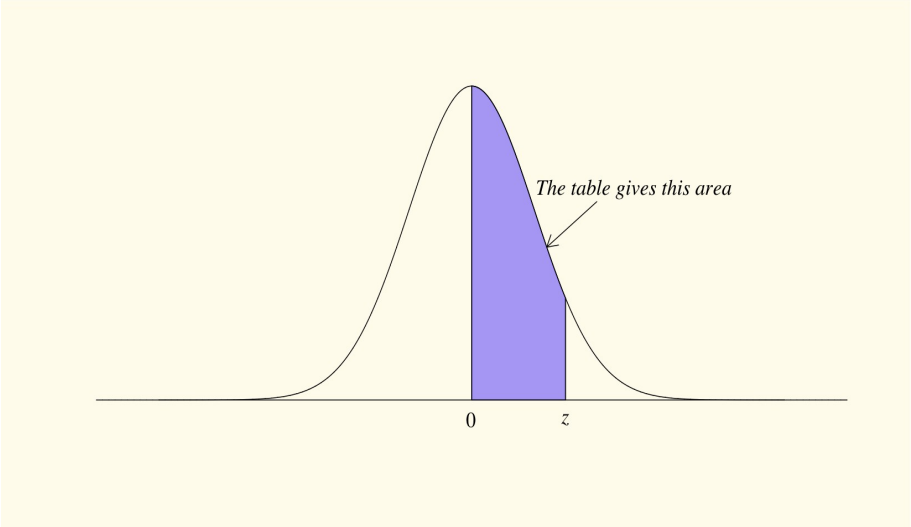
Finding Areas Under the Standard Normal Curve

It is very important that you fully understand the methods of this section! We will be using these techniques for the remainder of the course, and I will be assuming you can understand all of this perfectly.

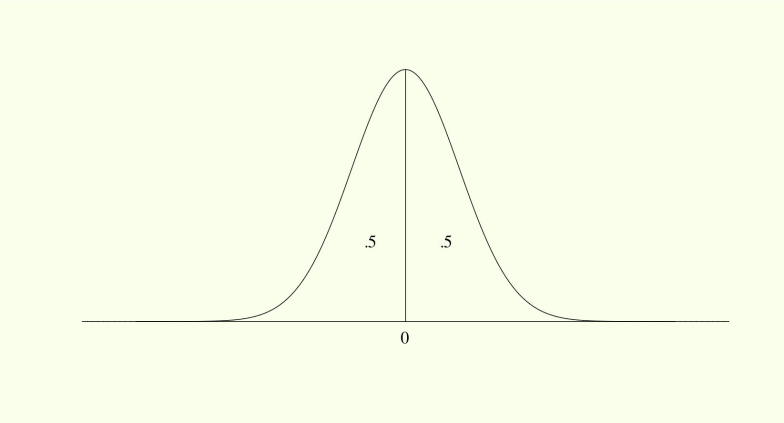
Many of our statistical inference procedures will require us to find areas under the normal curve, as well as find percentiles of normally distributed random variables. In this section we'll learn to find

these values.

Areas under the standard normal curve are found by integrating the curve. The curve must be integrated numerically, so this is not something we can easily do on our own. Fortunately, this has been done for us and areas under the standard normal curve are found in Table IV of our text. If we look up a value z in the table, the table entry gives the area between 0 and z:

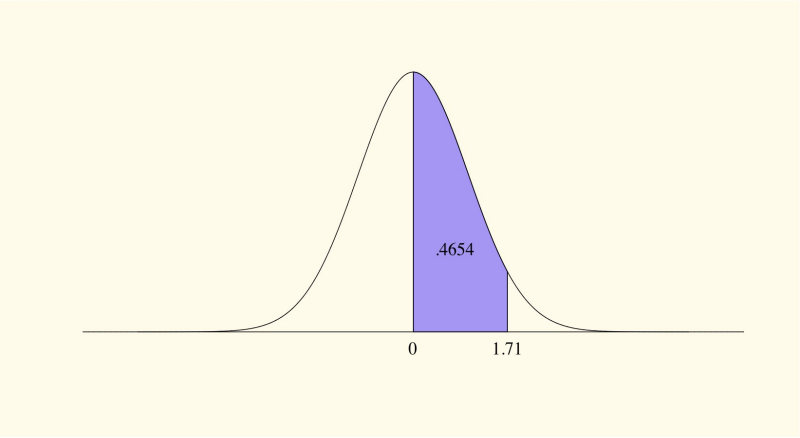


This is not my favourite table format, but it is used by many texts. We can use this area to find any area that we require. We need need only the important information that the **standard normal distribution is symmetric about 0**. Since the area under the entire curve equals 1, we know that the area to the left of 0 is 0.5, and the area to the right of 0 is 0.5:



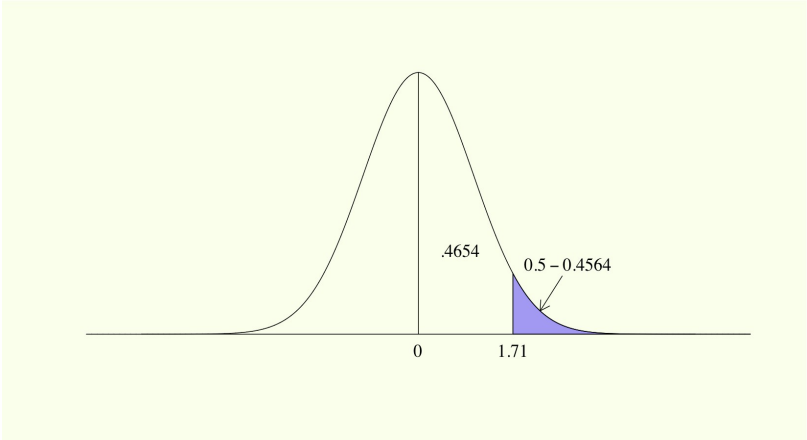
Let's do a few examples. It's important that you get these ideas down cold, as we will be looking up values in tables for the remainder of the course. The ideas are relatively straightforward, and you should get to the point where these questions are very easy for you to do.

What is $P(0 < Z < 1.71)$?
Until you are 100% confident in your abilities to get the proper probabilities, it is best to draw a picture. This will help to ensure you are looking up the proper area.



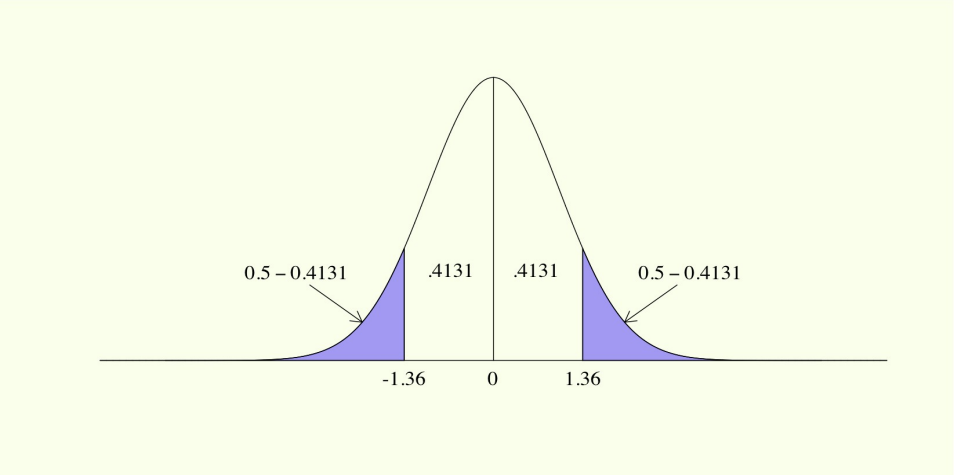
Since our textbook table gives us the area between 0 and the *z* value that we look up, nothing else needs to be done here. The value in the table is our final answer. $P(0 < Z < 1.71) = .4564$.

What is $P(Z > 1.71)$?
If we draw a picture, shading in the area that we require:



Here we need the area to the *right* of 1.71. The table gives the area between 0 and 1.71. Since the area to the right of 0 is 0.5, $P(Z > 1.71) = 0.5 - .4564 = .0436$.

The table contains only positive values of *z*. *Z* can take on negative values, but since the distribution is symmetric about 0, we can find areas corresponding to negative values of *z* easily. For example, suppose we wish to find $P(Z < -1.36)$. The area to the *left* of -1.36 is equal to the area to the *right* of 1.36:



When we look up 1.36 in the table, we get the value .4131, which is the area between 0 and 1.36. It is also the area between 0 and -1.36. P($Z < -1.36$) = P($Z > 1.36$) = .5 - .4131 = .0869.

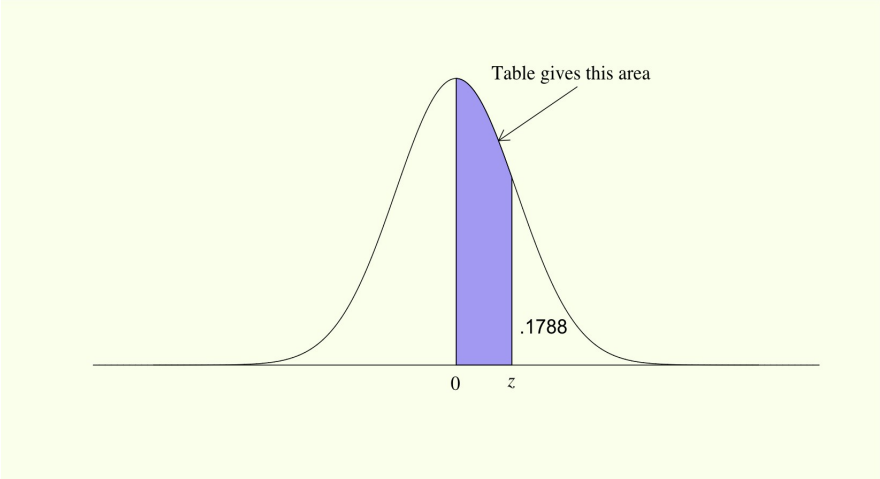
Using similar logic to the above, find the following probabilities.

P($1.23 < Z < 2.27$) = .4884 - .3907 = .0977.

P($-2.00 < Z < 2.00$) = P($-2 < z < 0$) + P($0 < z < 2$) = .4772 + .4772 = .9544.
Note that for the normal distribution, approximately 95% of the area lies within two standard deviations of the mean. The normal distribution is what the **empirical rule** is based on. We learned of the empirical rule earlier in the course.

In many situations we will need to find percentiles of the standard normal distribution. For example, we may need to find the value of z such that the area to the left is .99. We may wish to be 99% sure that we have enough product on hand to meet demand. Or 99.999% sure, or whatever is appropriate for the given scenario. In these situations to find the appropriate amount that is required, we would need the appropriate percentile. The following questions are inverse problems to those above. We are given a probability, and need to find the value z that makes the statement true. Students usually find these questions a little harder.

Find the value z such that the area to the right is .1788.
It is best to illustrate the given scenario with a quick plot:

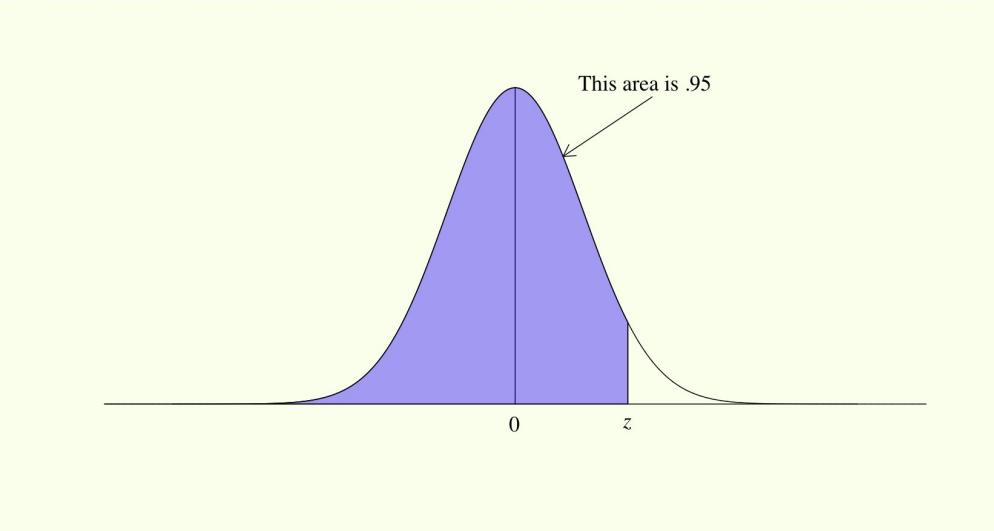


We cannot simply go running off to our table, blindly looking up .1788. This is not how our table is set up. Our table gives us areas between 0 and z. Since the table always gives us areas between 0 and the value z we look up, we need to convert our area into the area the table gives us. If the area to the right of z is .1788, this implies the area between 0 and z is .5 - .1788 = .3212.

If we go into the **body** of the standard normal table and find .3212, we see that the corresponding value of z is 0.92. $P(z > 0.92) = .1788$. N.B. We **do not** look up .32 as a value of z in the table. .32 is not a value of z, it is an area. Areas are found in the body (the middle part) of the table.

Example. Find the 95th percentile of the standard normal distribution.

By the definition of a percentile, this is equivalent to asking to find the value of z such that the area to the left is 0.95.



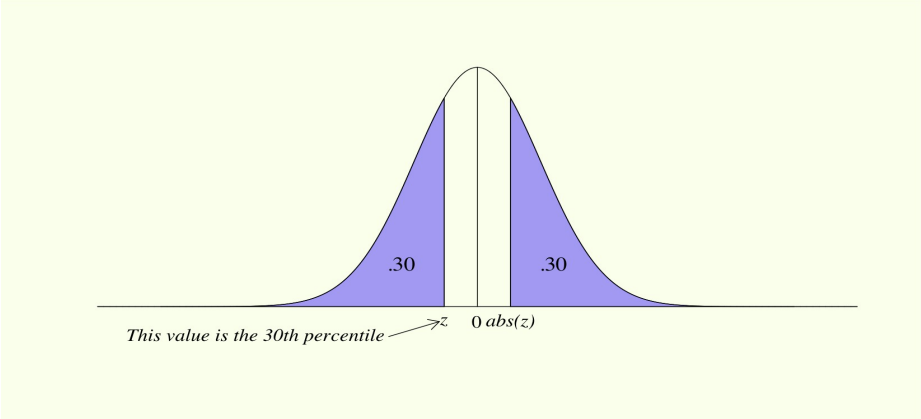
But our textbook gives us areas between 0 and z. Finding the 95th percentile of the standard normal distribution is equivalent to finding the value z such that $P(0 < Z < z) = 0.45$. In other words, we need

to find z such that the area between 0 and z is .45. If we look up .45 in the **body** of the table, we can see that the two closest values are .4495 and .4505. These correspond to z values of 1.64 and 1.65. Since the desired value is halfway in between .4495 and .4505, we can take the average of the two z value, and call the 95th percentile of the standard normal distribution 1.645. N.B. Under most circumstances, if you pick the closest value in the table that is good enough for me. There is no need to average or linearly interpolate. But the 95th percentile is an important number and comes up frequently, so we may as well get it right.

Find the 5th percentile of the standard normal distribution.

By symmetry, the 5th percentile of the standard normal distribution is -1.645.

N.B. Many students have trouble with finding z values that are negative. To do these questions, you must use the symmetry argument, and find the absolute value of the z value, then include the minus sign. For example, if we wish to find the 30th percentile of the standard normal distribution,



The 30th percentile of the standard normal distribution must be **negative** (it is to the left of 0). To find this value, we must recognize that the absolute value of the 30th percentile is equal to the 70th percentile. If we look up .5 - .3 = .2 in the body of the table, we find that the corresponding z value is approximately .52. Thus the 70th percentile of the standard normal distribution is approximately .52. Since we know the 30th percentile is a negative value, the 30th percentile is approximately -.52.

Some notation that will come up:
z_α is the value of z such that the area to the right is α.
$z_{.05} = 1.645$ (The area to the right of 1.645 under the standard normal curve is .05)
$z_{.95} = -1.645$ (The area to the right of -1.645 is .95)

Work through the above examples until you understand them and can do them perfectly.

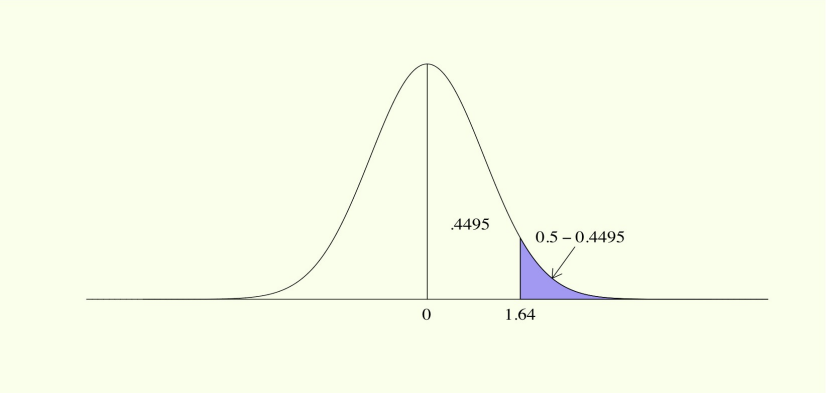
Suppose a random variable X is normally distributed with mean μ and and standard deviation σ. We can convert X into something having the standard normal distribution, so that we can use our standard normal table for our probability calculations.

If we let $Z = \frac{X - \mu}{\sigma}$, , then Z is a random variable having the standard normal distribution. Using this basic linear transformation, we can convert any normally distributed random variable into a variable that has the standard normal distribution.

Example #1. A soft drink dispensing machine fills cups with an average of 11.10 ounces of pop, with a standard deviation of 0.55 ounces. If the amount of the fill is approximately normal, what is the probability that a 12.00 oz cup will be filled to overflowing?

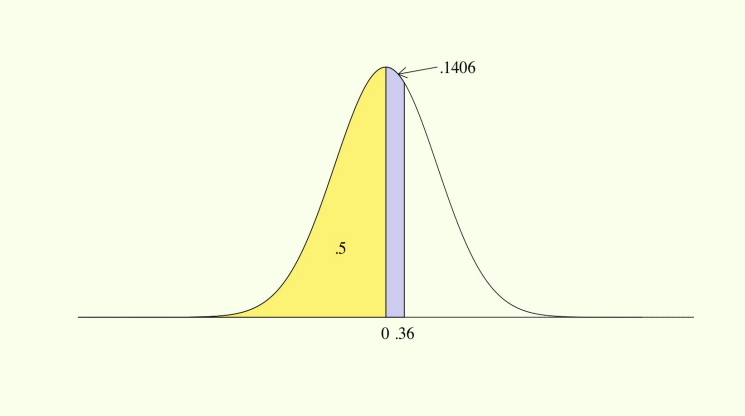
Here we are looking for *P(X > 12.00)*. If we had a table in the textbook for a normal distribution with $\mu = 12.00$ and $\sigma = 0.55$, then we could simply look up the appropriate area. We don't have such a table, but that doesn't pose much of an issue. We can simply convert X to a standard normal random variable:

$$P(X > 12.00) = P\left(\frac{X - \mu}{\sigma} > \frac{12 - \mu}{\sigma}\right) = P\left(Z > \frac{12.00 - 11.10}{0.55}\right) = P(Z > 1.636364) \approx 0.5 - .4495 = 0.0505.$$



What is the probability that the cup will contain less than 11.30 ounces?

$$P(X < 11.30) = P\left(Z < \frac{11.30 - 11.10}{0.55}\right) = P(Z < 0.3636364) \approx 0.5 + .1406 = .6406$$



What is the 32nd percentile of the amount of pop the machine puts in a cup?

Students often find this type of question more difficult than finding an area. Solving this type of problem involves going in the reverse order of what we did above. Steps:

1) Find the 32nd percentile of the standard normal distribution.

2) Convert from z to x. $z=\frac{x-\mu}{\sigma}$, which implies $x=\mu+\sigma z$

The 32nd percentile of the standard normal distribution is the value z that has an area of .32 to the left. This value is -0.47. Converting back to x: $x=\mu+\sigma z$, and thus the 32nd percentile of the amount of pop the machine puts in a cup is 11.10 + 0.55(-.47) = 10.8415.

Example #2.
Scores on a certain IQ test are approximately normally distributed with a mean of 100.0, and a standard deviation of 14.5. If a randomly picked person takes the test, what is the probability that:

They score less than 130?

X is approximately normally distributed with a mean of 100, and a standard deviation of 14.5.

$$P(X<130)=P(Z<\frac{130-100}{14.5})=P(Z<2.068966)=0.5+0.4808=0.9808$$

(At least approximately, as there is some rounding error there)

What is the lowest score that would put a person in the top 2% of the population?

This is equivalent to asking for the 98th percentile of the scores on the test. To answer this, find the 98th percentile of the standard normal, then convert back to X.

The 98th percentile of the standard normal is (approximately) 2.05. The 98th percentile of the IQ scores is 100 + 14.5(2.05) = 129.725.

N.B. Choosing the closest value in the table is good enough for me in these spots. Some profs may take issue with this, as the 98th percentile needs to have **at least** 98% of the area to the left. But I don't feel this is an overly important point, and it can be distracting. Choose the closest value in the table. If you can understand it from that perspective, that's good enough for me.

Checking for Normality: Normal Quantile-Quantile Plots

In our inference procedures in the course, at times we will need our sample data to come from a normally distributed population in order for the techniques to be valid. To check whether our data is normally distributed, we could plot a frequency histogram, but this is not typically as informative as a normal quantile plot. There are several different related techniques and plotting methods. The gist of all of them is that we end up with an approximately straight line if the data is normally distributed.

For this section your job will be fairly easy: I simply want you to be able to properly interpret a normal

quantile plot. You will not have to draw a plot, or even know how the calculations are done. But I illustrate this below. You should be able to properly interpret plots like those given at the end of this section.

In a normal quantile plot, we plot the ordered data against the appropriate quantiles of the standard normal. That is, we plot the *i*th ordered data value against the *z* value with *i/(n+1)* of the area to the left. (There are some slightly different methods that are used). If the data is normally distributed, this should result in an approximately straight line.

If there are systematic deviations, then our assumption of normality is violated, and any statistical inferences we make may be suspect.

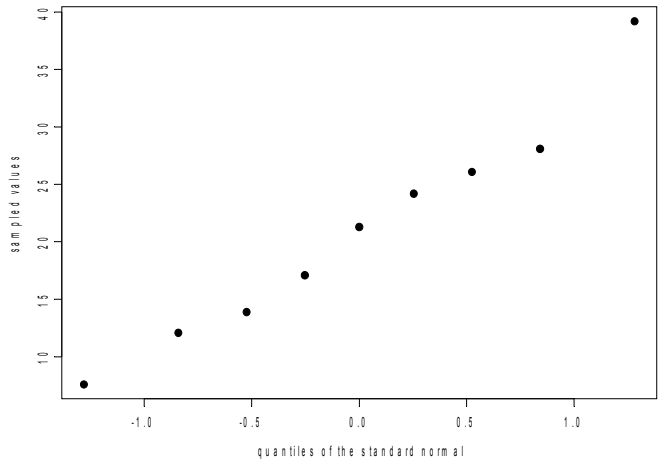
Example of the Calculations

Suppose we have a sample of $n = 9$ values, and we wish to see if there is any evidence of non-normality. The sample data: 12.1, 7.6, 39.2, 21.3, 24.2, 28.1, 26.1, 13.9, 17.1.

These data points are then sorted from smallest to largest, then the corresponding quantile of the standard normal distribution is calculated in the following manner:

Sorted values	7.6	12.1	13.9	17.1	21.3	24.2	26.1	28.1	39.2
$\frac{i}{n+1}$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
z	-1.28155	-0.84162	-0.52440	0.25335	0.00000	0.25335	0.52440	0.84162	1.28155

Where the *z* values are values of a standard normal random variable that result in *i/(n+1)* of the area to the left. The *z* values above were calculated using a computer, but we could get the values from the table in the text, with a slight loss of accuracy. Now, if we plots the sorted values against the corresponding Z-values, we have ourselves a normal quantile plot:

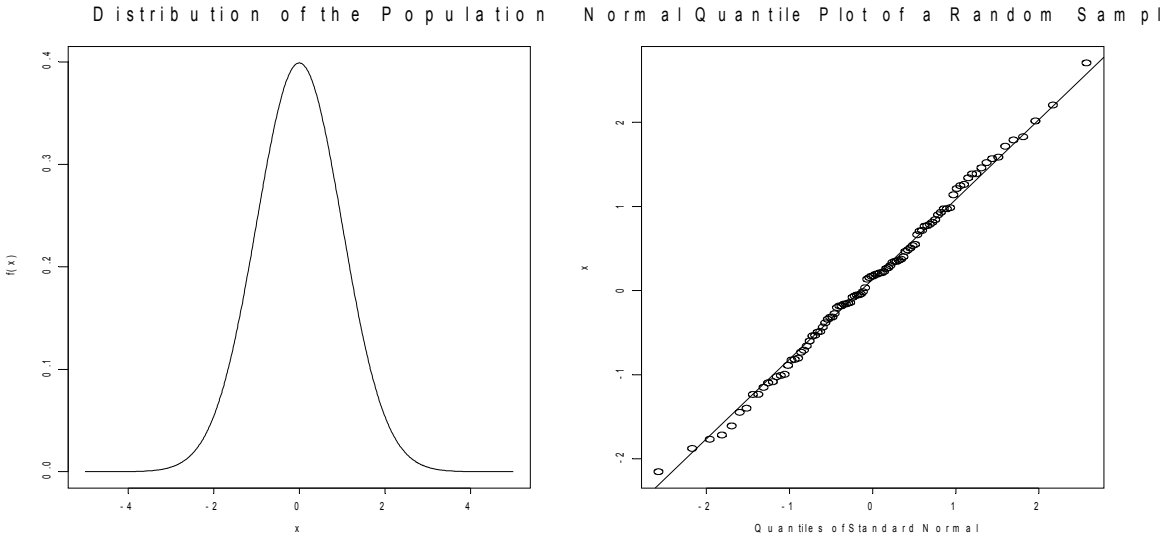


The resulting points fall close to a straight line. There is one slight outlier (the largest observation) that falls above the line, but the overall pattern is pretty close to linear. There are no obvious systematic deviations from linearity. This implies that our sample is close to normally distributed, and it is not unreasonable to assume our population is at least approximately normal.

In practice we do not have to do the calculations above, as statistical computing packages have built-in functions to plot normal quantile plots. Often we draw in a line to give some perspective and to make it easier for our eyes to detect non-linearity.

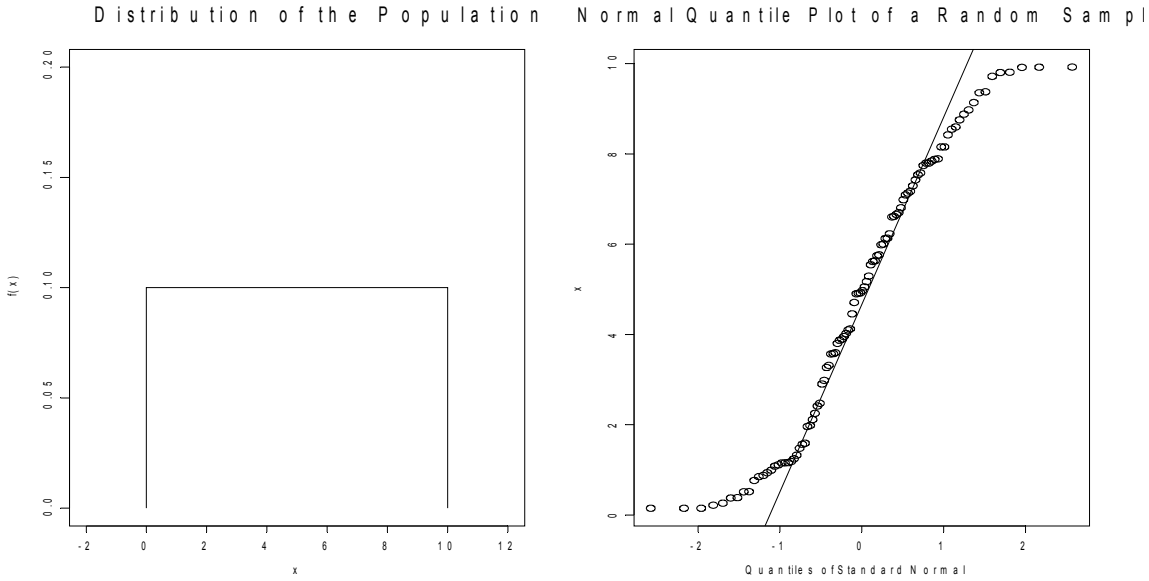
The following plots give several different population distributions, and the corresponding normal quantile plot for a random sample from the distribution.

Example 1: Population is normally distributed.



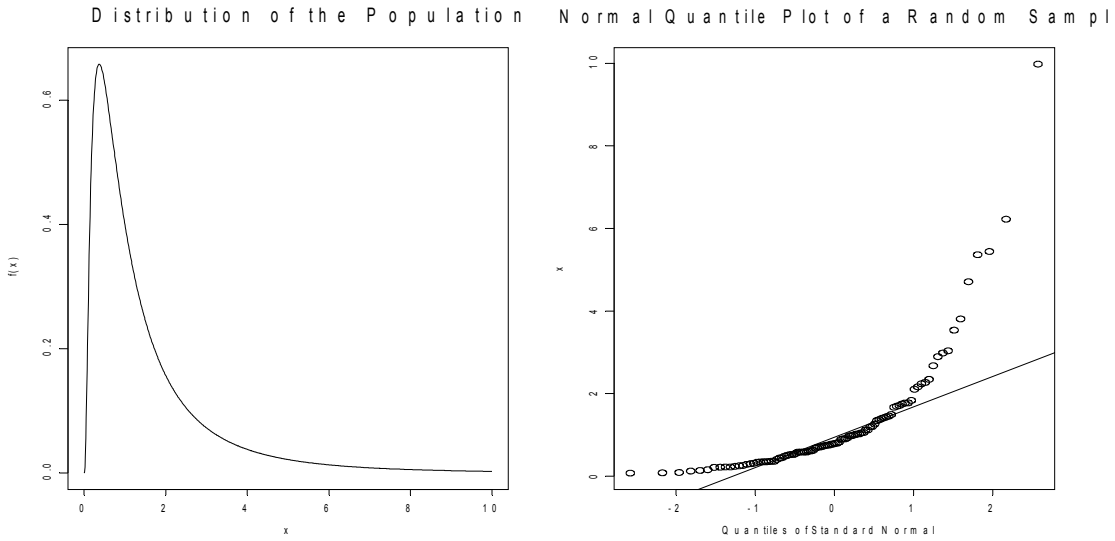
Here the true distribution is normal. The normal quantile plot of a random sample from this distribution results in a very straight line. (This line is even a little bit straighter than would be expected – there is usually a little random variability)

Example 2: Population is uniform on the interval (0,10)



Note that the middle part of the normal quantile plot seems quite linear, but there is systematic deviation from linearity in the tails. The uniform distribution does not have the tails that the normal distribution has, and is "chopped off". This results in the normal quantile plot curving away from linearity at the extremes. The values in the right tail of the distribution are not as large as would be expected from the normal distribution. Similarly for the left tail.

Example 3: A Right-Skewed Distribution.



This population is right-skewed, and has a heavier right tail than the normal distribution. The largest

sampled values are much larger than would be expected from a normal distribution, and this results in the normal quantile plot curving upwards on the right side. The smaller values in the sample are not as small as would be expected if the distribution was normal, and this results in the normal quantile plot curving upwards on the left side.

The purpose of these plots are to check if the normality assumption is justified. If we have evidence of non-normality, and we do not have a large sample size, we are going to have to go about the analysis another way. The procedures based on the normal distribution would not be reasonable, but perhaps we could use a procedure that doesn't require the normality assumption (a nonparametric, or distribution-free procedure). Alternatively, we may be able to transform the data, making it more normally distributed. More on this later.

Sampling Distributions

Fundamental to statistical inference is the concept of the **sampling distribution** of a statistic.

Recall: The **population** is the entire group of individuals or items we want information about.
The **sample** is the subset of the population that we actually examine.

Soon we will use sample *statistics* to estimate and make inferences about population *parameters*.

We do not typically know the value of parameters, so how will we know how good our estimates are? We use arguments based on the **sampling distribution** of a statistic. In short, the **sampling distribution** of a statistic is the **probability distribution** of that statistic. Let's look at the concept of a sampling distribution in a little more detail below.

In a simple random sample (SRS), each possible sample of size n has the same chance of being selected. (n/N = sample size/population size)

To conduct a SRS, we usually let a computer program randomly pick a sample for us. Humans are not very good at picking a random sequence of numbers, so we usually rely on computer programs, random number tables, or possibly internet sites that supply "true" random numbers.

Side note: It is actually fairly difficult to generate truly random numbers. Computer programs generate what we call *pseudo-random* numbers. They are not truly random, but for almost all purposes they are close enough to be considered random. Truly random numbers are available in some places. These places typically get their numbers by observing and measuring physical processes that are random, such as radioactive decay. But this is **way** more of a pain that generating pseudo-random numbers, so we go with the simple method pretty much always.

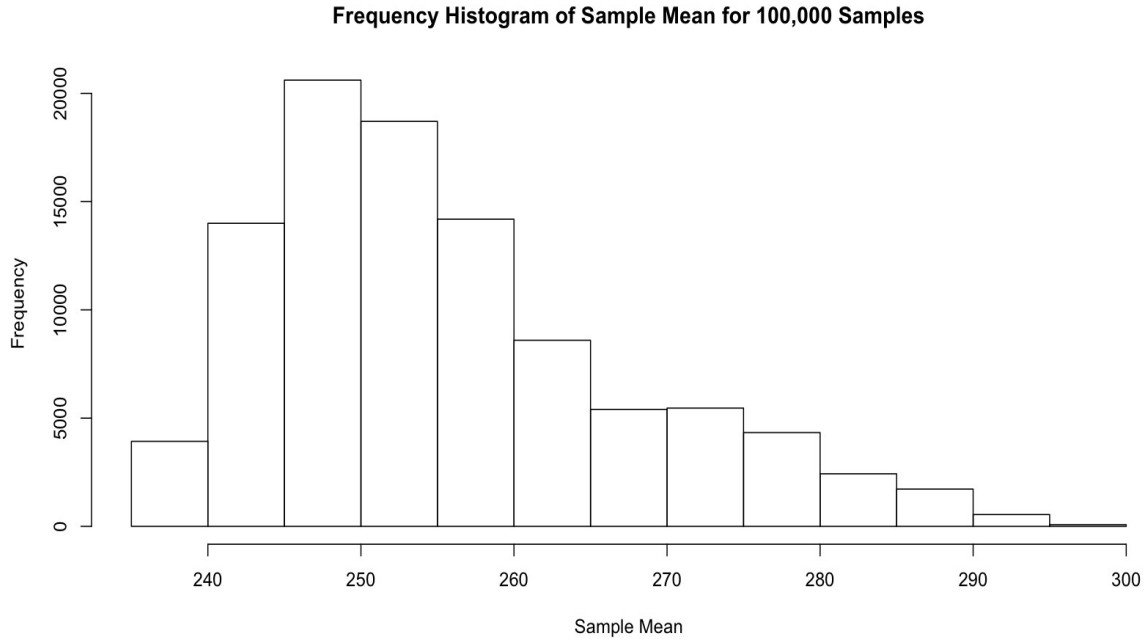
Example. Suppose I have a class of 20 students (the population) and wish to estimate the average age of my class. Suppose that I do not have this information available to me and do not have time to check into the ages of every student. If I have enough time to find out the ages of 3 of the students, then I should draw a SRS of size 3.

Student: Age in months (Unknown before the sample is drawn)

01 Adamson	228
02 Anderson	253
03 Bond	244
04 Crandell	249
05 Daley	263
06 Gates	288
07 Jackson	329
08 Krisp	255
09 McCullough	242
10 McKinnon	241
11 Peoples	267
12 Pringle	243
13 Rankin	276
14 Robertson	265
15 Rogers	239
16 Smith	248
17 Spry	251
18 Thomson	255
19 Watts	238
20 Wood	241
	Unknown to us, the population mean $\mu = 255.75$

Suppose we take a simple random sample of size $n = 3$, and it happens to be Watts, Daley, Rankin. Our sample mean age is (238+263+276)/3 = 259. If this sample is the only information we had about the population, then our best estimate of the mean age of the entire class (the population) is 259. But it shouldn't be too hard to see that if we take another SRS, then we will usually end up with a different sample mean age. For example, if our SRS ended up being Adamson, Bond, Robertson, then the sample mean age would be (228+244+265)/3 = 245.6667.

We could work out the exact sampling distribution of the sample mean in the above scenario, but we can also estimate it by repeatedly sampling samples of size 3 from the population, calculating the sample mean for each sample, and plotting out the resulting histogram:



This illustrates the **sampling distribution of the sample mean** in this scenario.

In practical situations we will not see the sampling distribution, as we will be **taking only one sample**. But the ***concept*** of a sampling distribution underlies everything we do in statistical inference. We will take one sample and get the value of a statistic from that sample. But we should recognize that the **value of the statistic we see is a single value sampled from that statistic's sampling distribution.** We will then use our mathematical knowledge of the sampling distribution to make statements like:

"We are 95% confident that the population mean age of students in this class lies between 245 and 260 months"

Much more on this type of statement later in the course.

The Sampling Distribution of the Sample Mean

It turns out that in most situations we do not have to repeatedly sample from a population to know something about the sampling distribution. We can often work it out using mathematical arguments. In this section we'll look at some simple but important properties we know about the sampling distribution of the sample mean.

We will assume here that the population is infinite (or at least very large compared to the sample size). Slight adjustments need to be made if that is not the case, but we won't worry about that for now.

Example.

Let X be the the number of heads when a fair coin is tossed once. What is the probability distribution of X?

x	0	1
$p(x)$	0.5	0.5

(We'll get heads half the time). Note that we can use our formulas for the mean and variance of a discrete probability distribution to find: $\mu_X = 0.50, \sigma_X^2 = 0.25$.

Let $\bar{X}$ be the mean number of heads per toss when a fair coin is tossed **twice**.
What is the probability distribution of $\bar{X}$? The possible outcomes are HH, HT, TH, TT. These lead to $\bar{X}$ values of 1, 0.5, 0.5, and 0, respectively. The probability distribution of the sample mean $\bar{X}$ is:

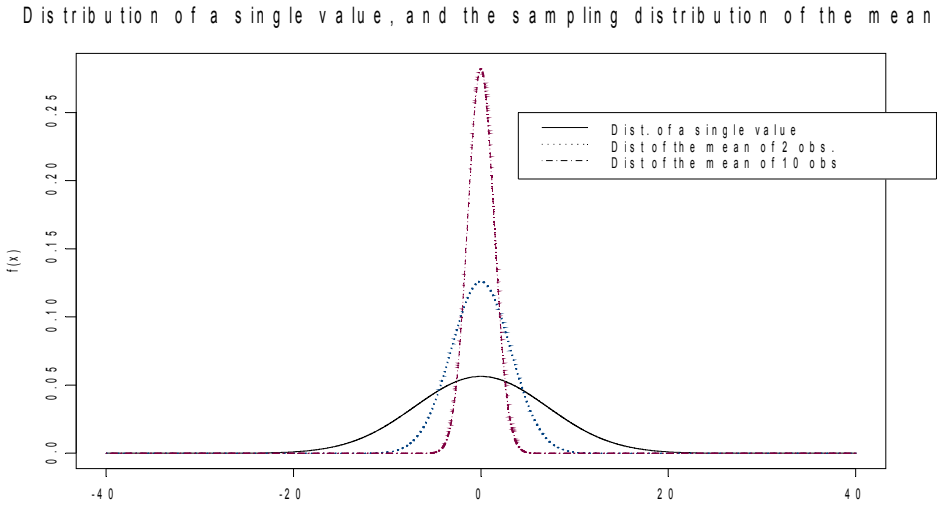
$\bar{x}$	0	0.5	1
$p(\bar{x})$	0.25	0.5	0.25

Note that we can use our formulas for the mean and variance of a discrete probability distribution to find: $\mu_{\bar{X}} = 0.50, \sigma_{\bar{X}}^2 = 0.125$.

Note that the mean of the probability distribution of the sample mean is equal to our original population mean, but the variance has decreased by a factor of 2. This is a specific example of the following important rules.

If the random variable X has mean μ and standard deviation σ, and $\bar{X}$ is the sample mean of n independent observations, then:

- $\mu_{\bar{X}} = \mu$. The mean of the sample mean is equal to the population mean.
- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- If the random variable X is normally distributed, then $\bar{X}$ is also normally distributed



To calculate a probability based on the *mean* of *n* observations: $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

Recall from above that we use: $Z = \frac{\bar{X} - \mu}{\sigma}$ for a *single* observation.

Example A manufacturer of potato chips claims their bags contain 170 grams of chips. In reality, the amount in each bag is approximately normally distributed, with a mean of 170.0 grams, and a standard deviation of 3.5 grams.

If a single bag is randomly selected, what is the probability that it contains more than 175.0 grams of chips? Since we are looking for a probability based on a **single** observation (one bag of chips), we will need to use $Z = \frac{\bar{X} - \mu}{\sigma}$:

$$P(X > 175) = P(Z > \frac{175 - 170}{3.5}) = P(Z > 1.429) \approx 0.5 - .4236 = 0.0764$$

If 3 bags are randomly selected, what is the probability the average weight is more than 175 grams? This question is fundamentally different from the previous one. Here we need to calculate a probability based on the **mean of 3** observations. We thus need to use $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$:

$$P(\bar{X} > 175) = P(Z > \frac{175 - 170}{3.5/\sqrt{3}}) \;=\; P(Z > 2.474) = 0.5 - .4932 = .0068$$

Central Limit Theorem

We learned above that if the population is normally distributed, then the sampling distribution of the sample mean will be normal, regardless of the sample size. But in this section we learn a much more interesting fact: for large sample sizes the sampling distribution of the sample mean will be approximately normal, **regardless of the distribution of the population**.

The C**entral Limit Theorem** tells us that regardless of the distribution of the random variable *X*, the distribution of the sample mean $\bar{X}$ is approximately normal, provided *n* is sufficiently large.

This is a rather loose definition of the Central Limit Theorem, but this is the gist of it, and the part that is important for us.

The distribution of $\bar{X}$ tends toward the normal distribution as *n* increases, and can be considered approximately normal under most conditions for $n > 30$.

If *X* is exactly normally distributed, $\bar{X}$ will be exactly normally distributed, regardless of the sample size. But the central limit theorem tells us that we do not need the distribution of our data to be normal – the distribution of the sample mean will be approximately normal provided we have a large sample

size.

$n > 30$ is only a ROUGH GUIDELINE. I hesitate to even put it down. In reality, it depends a great deal on the shape of the original population. If the original distribution is close to normal, then the sample mean will be approximately normally distributed, even for small sample sizes. If the distribution of the population is very different from normal, strongly skewed to the right, say, then one may have to take a sample size of that is much larger than 30 in order for the approximation to be reasonable. But under most circumstances, $n > 30$ is a reasonable rough guideline.

Example.
Suppose Guelph housing prices are known to have a mean of $389,000 and standard deviation 120,000.

What is the probability that a randomly picked house costs more than 400,000?

It may be tempting to try: $P(X > 400{,}000) = P(Z > \frac{400\,000 - 389\,000}{120\,000})$, but at this point we cannot reasonably look up this value of Z in the standard normal table. There is nothing in this question that states that housing prices are approximately normal. We have in fact learned previously in the course that housing prices are not normally distributed, and are skewed to the right. Unless we know the true distribution, this question can not be answered.

What is the probability that the mean of 100 randomly selected houses is more than 400,000?

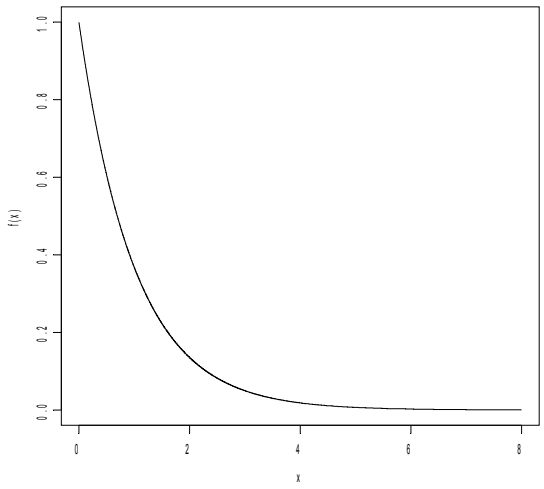
Although we could not answer this question for a **single** house, the Central Limit Theorem allows us to answer this question based on the **mean of 100** houses (at least approximately). Since the sample size is large, the distribution of the sample mean is approximately normal. We can thus use our usual techniques:

$$P(\bar{X} > 400{,}000) = P(Z > \frac{400\,000 - 389\,000}{120\,000/\sqrt{100}}) = P(Z > 0.9166667) \approx 1 - .8212 = .1788$$

Because of the Central Limit Theorem we are able to use methods based on the normal distribution in a wide variety of situations.

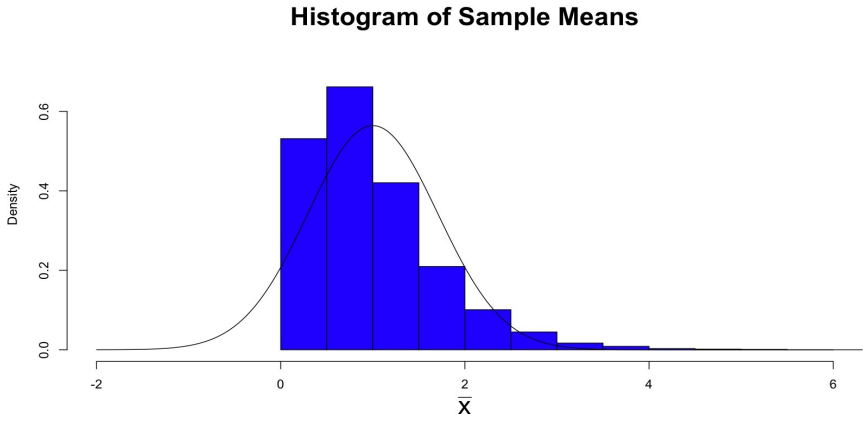
Illustration of the Central Limit Theorem

Suppose the distribution of X is:



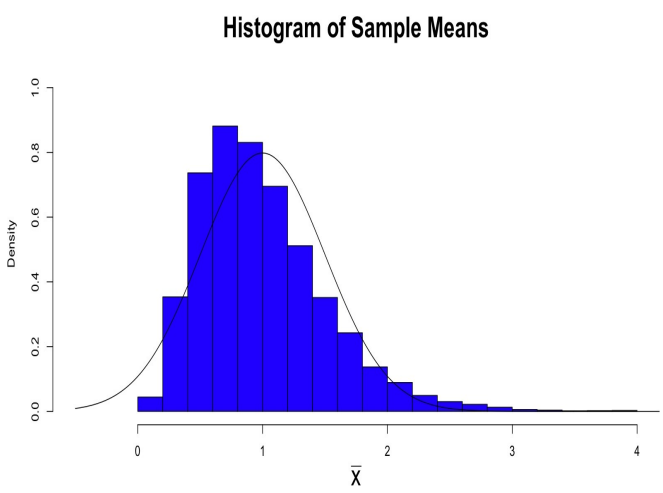
The original distribution above is strongly skewed to the right. But what is the distribution of the sample mean if we draw a sample from this distribution? It can be worked out mathematically in this spot, but let's look at it here through simulation. What happens when we draw random samples from this distribution and look at the distribution of the sample mean?

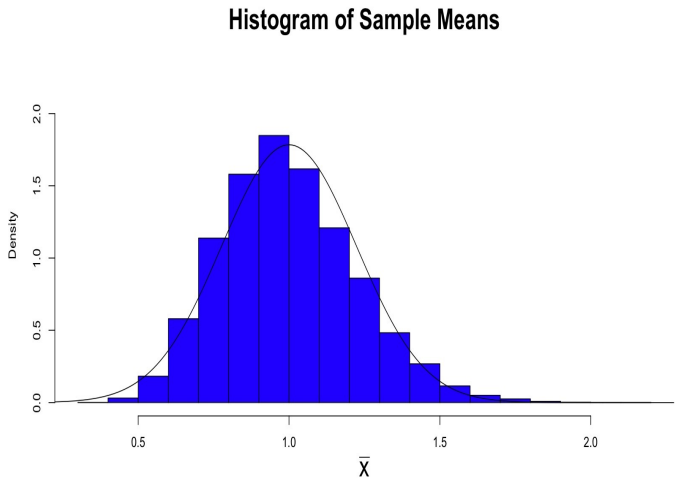
Suppose we draw a sample of size 2, calculate the mean, and repeat 10,000 times. The histogram of sampled means:



This distribution is still very skewed, but not quite as skewed as the distribution of a single value.

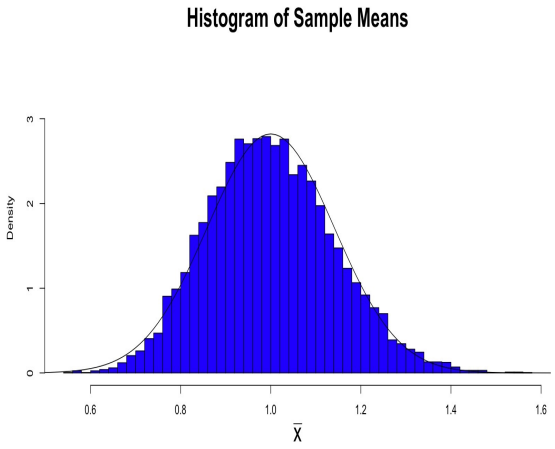
Note that the scale is not the same on these plots! These plots merely illustrate the shape of the distribution.
As the sample size increases, the distribution of the sample mean becomes less and less skewed. By a sample of size 20 or so, it's looking very normal.
Samples of size 4: Samples of size 20:

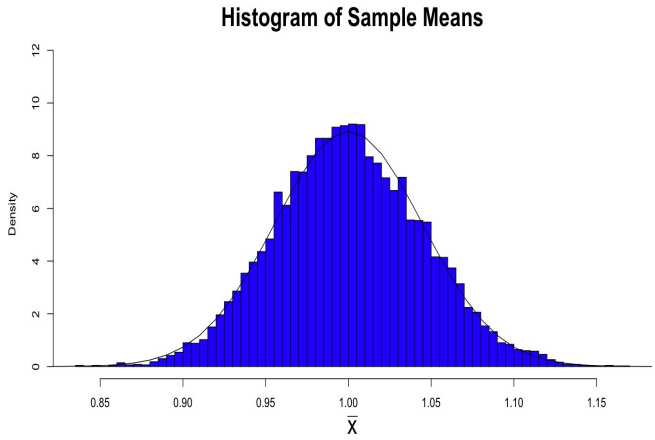




Samples of size 50:

Samples of size 500:





When the sample size gets very large (500), pretty much all of the skewness is gone and we have a very normal-looking distribution.

The important point to note is that the sampling distribution of the sample mean is **tending toward the normal distribution** as the **sample size increases**. This relationship has nothing to do with the number of samples being draw; it depends on the sample size.

As stated above, this is very important for our statistical inference techniques. We can often use inference techniques that are based on the assumption of a normally distributed population, even if the population is NOT normally distributed, provided we have a large enough sample size.

Unbiased Estimators

A statistic is *unbiased* if its average value is equal to the parameter it estimates.

We have previously learned that: $\mu_{\bar{X}} = \mu$. Since the average value of the sample mean is equal to the population mean, the sample mean is an **unbiased** estimator of μ.

The most useful statistics are unbiased estimators that have low variability. The sample mean $\bar{X}$ is the best estimator of the population mean μ. The sample mean is an unbiased estimator, which is good, but it also has the lowest possible variance of any unbiased estimator of μ (this was not shown above, but is an established mathematical fact).

We will be omitting the Normal Approximation to the Binomial section at this time. We will use some of the ideas later on in the course, but we will look at it from a slightly different perspective.

Suggested Textbook Questions:

Continuous uniform distribution: 4.141, 4.147, 4.151

Normal distribution: 4.79, 4.81, 4.85, 4.87, 4.89, 4.91, 4.93, 4.97, 4.99, 4.101, 4.103

Sampling distribution of the sample mean: 4.156, 4.163, 4.165, 4.167, 4.169, 4.171, 4.173, 4.175, 4.177

Supplementary: 4.183, 4.185, 4.189, 4.191, 4.195, 4.203, 4.207, 4.211, 4.213

If you are looking for some extra questions to do:

Course Manual: 3.1, 3.2. Omit the "normal approximation to the binomial" questions of 3.3.