Chapter 6

Key Ideas

Density Curve – Uniform Distribution, Standard Normal Distribution, Z-Score, Z-table (finding areas above and below values using them), Sampling Distributions (of the mean, of a proportion), The Central Limit Theorem

Section 6-1: Overview

In Chapter 5, we considered discrete probability distributions. We paid special attention to a particularly useful distribution called the Binomial Distribution. Now, we will turn our attention on continuous probability distributions. Here, the possible outcomes fall in a range of values without gaps (note that this means there are an infinite number of values the random variable could take).

Section 6-2: The Standard Normal Distribution

Before we explore the more complicated Standard Normal Distribution, we must examine how the concept of a probability distribution changes when the random variable is continuous.

Recall that in Chapter 5, a probability distribution gave a value P(x) = P(X = x) to each possible outcome x. For the values to make a probability distribution, we needed two things to happen:

1. $\sum_{x} \mathrm{P(x)} = 1$
2. $0 \le \mathrm{P(x)} \le 1$

For a *continuous* random variable, a probability distribution must be what is called a **density curve**. This means:

1. The area under the curve is 1.
2. $0 \le \mathrm{P(x)} \le 1$ for all outcomes x.

Notice that the first condition is similar to the first condition for discrete distributions. The second condition, however, does not require probabilities to be less than 1 anymore.

Note: Whether we use < or ≤ makes no difference in the case of a continuous random variable. This is because the probability that X is *exactly* a certain value (i.e. exactly 7, and not 7.00001, 6.999998, etc.) is infinitesimally small. In other words, the probability that X is exactly a certain value is always 0.

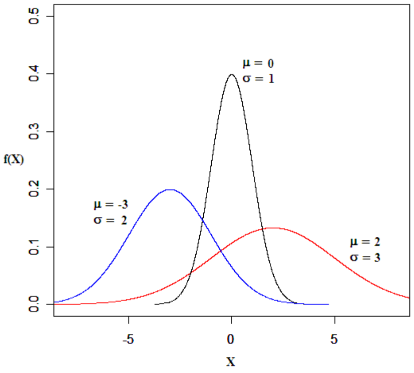
Normal Distributions

Another example of a continuous random variable (and the most commonly used in statistics) is called a *normal random variable*. A variable with a normal distribution displays what we call a bell-shaped distribution. It is called "normal" because it is a good model for random error (i.e. "noise") around a particular value. For instance, a person's height will not be exactly equal to the average height for all people in the world. However, we would expect a lot of people to be fairly close to the average, and less people to be much taller or shorter. Additionally, we might expect that there will be as many people below average as there are people above average. For this reason, we might use a bell-shaped distribution to signify more people in the middle and fewer in the extremes.

A normal distribution has two parameters. The mean (average) is called μ, and the standard deviation is σ.

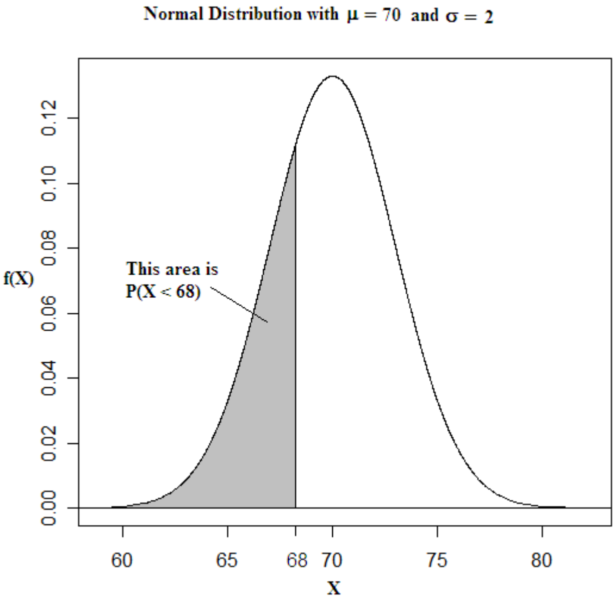
The formula for the density curve of the normal distribution is: $f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ (this determines the bell shape)

While the formula is too complicated for the scope of this class, by observation you can see that the formula is entirely determined given values of μ and σ. For this reason, the mean and standard deviation completely define a normal distribution. Here is a plot of normal distributions with different values of μ and σ:



Note: It may not be immediately obvious, but the area under each curve is actually 1. Therefore, these are all density curves.

Example: Consider a normal distribution with mean 70 and standard deviation 2. What is P(X < 68)? We may be able to draw the area that needs to be found, but how can we compute it?



The Standard Normal Distribution

Fortunately, there is a way to easily compute areas under any normal density curve. To do this, we first examine a special type of normal distribution.

Definition: The **standard normal distribution** is the normal distribution with mean 0 and standard deviation 1. It is often denoted Z to differentiate it from a regular random variable X or Y.

To find areas under this special distribution, statisticians used numerical integration methods to find areas below certain values of Z, (these are called z-scores). These areas have been compiled into a table (Table A-2, back cover of the book). The organization of the table is as follows:

- The left column gives a value z to one decimal place.
- The top row gives the 2nd decimal place for z.
- The numbers in the body of the table give the area *below* the z-score given by the corresponding numbers in the row/column. i.e. the numbers in the body of the table are P(Z < z), where Z is the standard normal random variable.

Note that the table only gives the area *below* particular z-values, but not the area *above* or *between* values. To find these areas, you must use the fact that the area under the entire curve is 1.

Examples (Finding areas given Z-scores)

Let Z be a standard normal random variable, and find the following probabilities.

1. Probability that Z is less than 0.63.
2. Probability that Z is less than –1.5.
3. Probability that Z is greater than 1.34.
4. Probability that Z is greater than –0.78.
5. Probability that Z is between 0.1 and 0.5.
6. Probability that Z is between –0.84 and 0.93.

Solutions

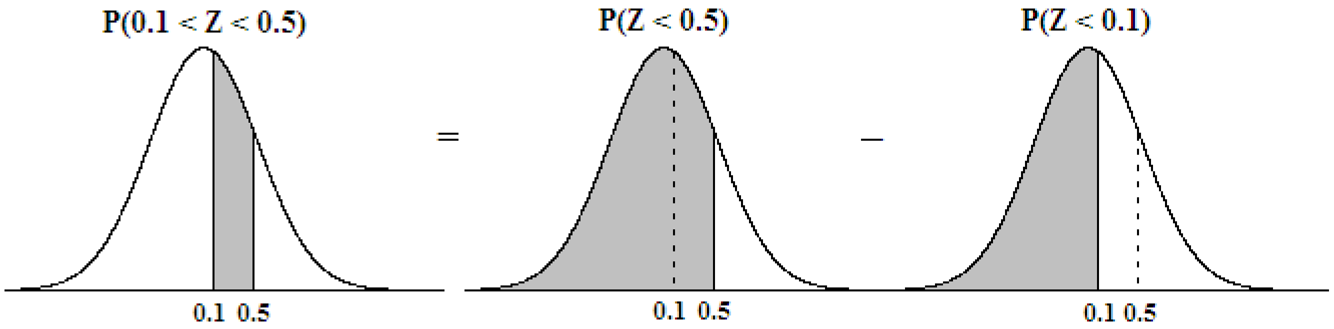
1. To find P(Z < 0.63), we note that z = 0.63. So we look down the left column to find 0.6, then along the top row to find 0.03. The number in the body of the table will be the area below z = 0.63. According to the table, then, P(Z < 0.63) = **0.7357**.

2. To find P(Z < -1.5), we use the same technique. Now, in the left column we find –1.5. In the top row, we find 0.00. Then the number in the body of the table is the area below z = -1.50, which gives us P(Z < -1.5) = **0.0668**.

3. Now we want P(Z > 1.34). Looking for 1.3 in the left column and 0.04 in the top row tells us the area *below* 1.34 is 0.9099. However, this is not the area we want. We want the area *above* 1.34. Since the total area under the curve is 1, this means the area below and the area above must add to 1. In other words, P(Z > 1.34) = 1 – 0.9099 = **0.0901**.

4. Using the same technique as in #3, we find that the area *below* –0.78 is 0.2177. Therefore, P(Z > -0.78) = 1 – 0.2177 = **0.7823**.

5. To do these last 2 problems, look at the following picture:



Finding the area between 0.1 and 0.5 is the same as taking the area below 0.5 and subtracting away the area below 0.1. Thus, we can look up $P(Z < 0.5)$ and $P(Z < 0.1)$ in the table and find the difference to get the answer. From the table, we see that:
$P(Z < 0.5) = 0.6915$
$P(Z < 0.1) = 0.5398$
Therefore, $P(0.1 < Z < 0.5) = P(Z < 0.5) - P(Z < 0.1) = 0.6915 - 0.5398 = \mathbf{0.1517}$.

6. $P(-0.84 < Z < 0.93) = P(Z < 0.93) - P(Z < -0.84) = 0.8238 - 0.2005 = \mathbf{0.6233}$.

Examples (Finding Z-scores given areas)
Let Z be a standard normal random variable.
1. Find the value z for which $P(Z < z) = 0.3594$.
2. Find the value z for which $P(Z < z) = 0.65$.
3. Find the value z for which $P(Z > z) = 0.0139$.
4. Find the value z for which $P(Z > z) = 0.10$.
5. Find the 70th percentile of Z.

Solutions
1. Now, we know the *area* below z, and we want to find z. To do this, we look through the table to find the desired area 0.3594 in the *body* of the table (where areas are located). It appears in the bottom right corner on the left page. Now, we find out which z-score this area corresponds to. The value on the left column in this row is –0.3. The value at the top of this column is 0.06. This means the z-score with an area of 0.3594 below is **z = –0.36**.

2. Again, we are given the area below z: it is 0.65. This time, however, when we look at the z-table, there is no z-score with exactly 0.65 as the area below. The closest ones are 0.6480 (z = 0.38) and 0.6517 (z = 0.39). A good rule of thumb when trying to approximate z is to choose the z-score with the area closer to the target value. If both areas are roughly an equal distance away (as in this case), the convention is to take the average of the two. So here, we would use **z = 0.385**.
Note: For the purposes of this class, 0.38, 0.39, or 0.385 would have all been acceptable answers.

3. We now want the z-score where the area *above* is 0.0139. However, values in the table only give the area *below* particular z-scores. Therefore, we have to find the area below the desired value. If the area above z is 0.0139, that means the area below z is 1 – 0.0139 = 0.9861. Now we can look in the table for 0.9861 (as in #1) and find **z = 2.20**.

4. Again, an area above z of 0.10 translates into an area below of 0.90. Looking in the table for 0.90, we find two areas that are close: 0.8997 (z = 1.28) and 0.9015 (z = 1.29). Since 0.8997 is a lot closer, we just use **z = 1.28**.
Note: 1.28, 1.29, or 1.285 would have all been acceptable for the purposes of this class.

5. The 70th percentile of Z is the z-score for which the area *below* is 0.70. Looking in the table for 0.70, we find two areas that are close: 0.6985 (z = 0.52) and 0.7019 (z = 0.53). These are roughly the same distance away, so we go with the average: **z = 0.525**.
Note: 0.52, 0.53, or 0.525 would have all been acceptable for the purposes of this class.

Section 6-3: Applications of Normal Distributions

In application, many processes around the world follow normal distributions. However, the mean and standard deviation are almost always *not* 0 and 1, as is the case with the standard normal distribution. How can we find probabilities involving these general normal distributions? The answer is surprisingly simple.

In Section 3-4, we discussed how to find a z-score to obtain standardized values to be used in comparing relative standing.

Recall: The *z-score* of an observation x is $z = \frac{x-\mu}{\sigma}$, where μ and σ are the mean and standard deviation of the population.

It turns out that if X is a normal random variable with mean μ and standard deviation σ, then $Z = \frac{X-\mu}{\sigma}$ is a standard normal distribution! Furthermore, any area under the curve for X is the same as the corresponding area under Z.

This means that in order to find areas under a general normal curve, we just compute the z-score for the cutoff value and find the area using the tables as in 6-2.

Examples (Finding the probabilities given X)
Suppose X is the cost per gallon of gas at a pump anywhere in the U.S., and X is normally distributed with mean 2.25 and standard deviation 0.2. If you fill up at a random gas pump, what is:

1. The probability that the gas is less than $2.36?
2. The probability that the gas is less than $1.90?
3. The probability that the gas is more than $2.00?
4. The probability that the gas is more than $1.75?
5. The probability that the gas is between $2.00 and $2.50?

Solutions

1. We want P(X < 2.36), but we only have a table for the standard normal random variable Z. Thus we have to compute the z-score for 2.36, then find the area below that value. $z = \frac{x-\mu}{\sigma} = \frac{2.36-2.25}{0.2} = 0.55$. Now we look up z = 0.55 in the table and find that the area below is 0.7088. Thus, P(X < 2.36) = **0.7088**.

2. Again, we compute the z-score. This time, x = 1.90. Using the conversion formula, we find $z = \frac{x-\mu}{\sigma} = \frac{1.90-2.25}{0.2} = -1.75$. The area below z = -1.75 is 0.0401 from the table, so P(X < 1.90) = **0.0401**.

3. Now we want P(X > 2). As in #1 and #2, we compute the z-score. This time, x = 2.00, which gives a z-score of –1.25. Now, however, we want the area *above* this value. From the table, the area below z = -1.25 is 0.1056.
Therefore, P(X > 2) = 1 – 0.1056 = **0.8944**.

4. This is just like #3, only now x = 1.75. The associated z-score is –2.5. The area below z = -2.5 is 0.0062, which means the area above is 1 – 0.0062 = **0.9938**, which is therefore P(X > 1.75).

5. To find P(2.00 < X < 2.50), we have to find the z-score for x = 2.00 and x = 2.50. These values are z = -1.25 and z = 1.25, respectively. Thus, P(2.00 < X < 2.50) = P(-1.25 < Z < 1.25). Using the same technique as in 6-2, we say that:
P(2.00 < X < 2.50) = P(-1.25 < Z < 1.25) = P(Z < 1.25) – P(Z < -1.25) = 0.8944 – 0.1056 = **0.7888**.

Examples (Finding Xs given areas)
In the same situation as the previous examples, let X = price per gallon of gas at a random pump. Mean = 2.25, SD = 0.20.

1. Find the value x where P(X < x) = 0.2483.
2. Find the gas price where 58.71% of gas prices fall *below* the value.
3. Find the gas price where 80% of gas prices fall *below* the value.
4. Find the gas price where 17.88% of gas prices fall *above* the value.
5. Find the cutoff price for the top 10% of gas prices.
6. Find the 45th percentile of the gas prices.

Solutions

1. Now, we are given an area. Since the only given areas are in the Z-table, we can use that table to find the z-score associated with this area. The value is z = –0.68. Now, we need to convert the z into an x:
$$z = \frac{x-\mu}{\sigma} \Rightarrow -0.68 = \frac{x-2.25}{0.2} \Rightarrow -0.136 = x-2.25 \Rightarrow x = 2.114.$$
Therefore, the value we want is x = 2.114 (roughly **$2.11**/gallon)

2. As in #1, we are given an area (this time in percent form). The probability below x needs to be 0.5871. Looking in the table, the associated z-score is z = 0.22. As in #1, we convert this z into an x:
$$z = \frac{x-\mu}{\sigma} \Rightarrow 0.22 = \frac{x-2.25}{0.2} \Rightarrow 0.044 = x-2.25 \Rightarrow x = 2.294.$$
So the price with 58.72% gas prices below it is roughly **$2.29**.

3. As in #1 and #2, we want an area below of 0.80. This value does not fall exactly in the table, but the two closest values are 0.7995 (z = 0.84) and 0.8023 (z = 0.85). The first one is a lot closer, so we use z = 0.84. Doing the same conversion from z to x, we find that x = 2.418. So the desired value is roughly **$2.42**.

4. Now, we want 17.88% of prices *above* the value, which means P(X > x) = P(Z > z) = 0.1788. Unfortunately, the table only has areas *below* values, so we look for an area below of 1 – 0.1788 = 0.8212. The associated z-score is 0.92. Converting from z to x as in the first 3 examples gives x = 2.434. Thus the desired value is approximately **$2.43**.

5. The cutoff price for the top 10% of gas prices means that 90% of prices are below it. Thus the area below is 0.90. Looking in the table, the closest values are 0.8997 (z = 1.28) and 0.9015 (z = 1.29). We use z = 1.28 since it is closer. Converting from z to x yields a cutoff of x = 2.506, or roughly **$2.51** per gallon.

6. The 45th percentile has an area below of 0.45. In the table, the closest two areas are 0.4483 (z = -0.13) and 0.4522 (z = -0.12). Since they are roughly the same distance away, we use the average: z = -0.125. Converting z to x using the formula, we obtain the value x = 2.225. Thus, the 45th percentile is about **$2.23**.

Section 6-4: Sampling Distributions and Estimators

Suppose that you were trying to estimate the average number of children (people less than age 12) in a U.S. home. It is impossible to calculate this number directly, so you would need to take a representative sample of the population and find the mean of the sample. Generally speaking, the mean should be close to what the true population average is if the sample minimized bias. If you repeated this procedure over and over (i.e. draw a new sample and calculate the mean), would you get the exact same mean every time? Most likely not. Each time, the mean should be close to the true population average, but the means themselves will be *varying* from sample to sample.

In this section, we will explore estimators (like the mean) as *random variables*. In the example above, the mean is a random variable because it changes from sample to sample. We can even ask questions about the values it takes on. For instance, around what values will the mean usually be? How often will it be bigger than 10? How often will it be 0? Certainly, values like 10 and 0 can occur if you select an “unlucky” sample, but it will not happen as often as a mean around 2 or 3. Thus, at least for the mean, we can guess that most samples will give means very close to the true population average, and less of them will be further out. (sort of like a normal distribution, perhaps?)

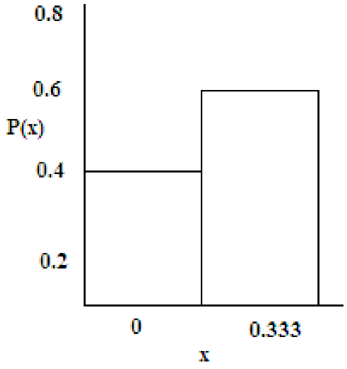
First, we look at a few definitions:
Definition: The **sampling distribution of a statistic** is the distribution of all values of the statistic when all possible samples of size *n* are taken from the population. It is usually displayed as a table, probability histogram, or formula. Here, a *statistic* is the sample mean, proportion, median, standard deviation, etc. (anything calculated from the collected data).

Definition: The **sampling distribution of the proportion** is the distribution of all values of the sample proportions when all possible samples of size *n* are taken from the population.
Definition: The **sampling distribution of the mean** is the distribution of all values of the sample mean when all possible samples of size *n* are taken from the population.

Example (Sampling Distribution of the Proportion)
Suppose you are trying to estimate the % of marbles in a jar that are yellow. For simplicity, let’s suppose there are only 5 marbles in the jar (1 yellow, 1 red, 1 orange, 1 green and 1 blue) and you are estimating this proportion using a sample of size 3. There are 10 different samples of size 3 that could be picked, and they give the following sample proportions (note that the true proportion is 0.2):

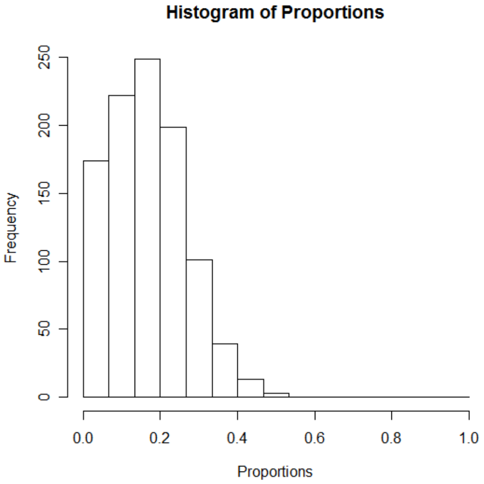
Sample	Sample Proportion
YRO	0.333
YRG	0.333
YRB	0.333
YOG	0.333
YOB	0.333
YGB	0.333
ROG	0
ROB	0
RGB	0
OGB	0

The probability histogram for *p*:



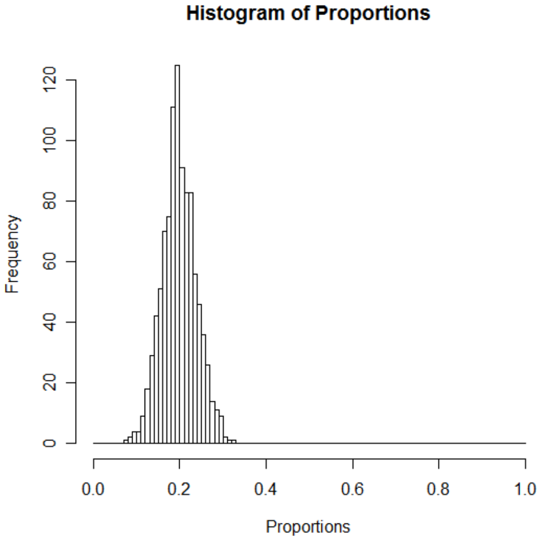
Example

Now suppose there are 200 marbles in the jar, and we take samples of size 15. There are 1.46 x 10^{22} possible samples that could be taken. Suppose again that 20% of the marbles are yellow (40 of them). The following plot is a frequency histogram of the sample proportions for 1000 different randomly chosen samples.



We see that most sample proportions are close to the true proportion of 0.2, and they trail off as they get further from the true value. Still, it appears that a few times we got a sample that had as many as half the marbles being yellow, even though only 20% of the total population was yellow.

If we increase the sample size to 100 and rerun the 1000 samples, the histogram changes a lot:

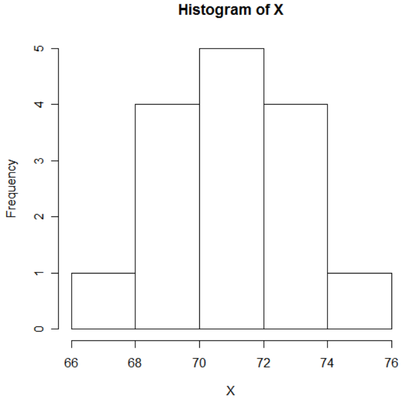


Now, the variation in proportions is much smaller, due to the fact that the sample is so large. Also, you may notice that the shape of this distribution is familiar: it is close to a normal distribution. This illustrates an important fact: under certain circumstances, the sampling distribution of a proportion can be approximated by a *normal distribution*.

Another Example (Sampling Distribution of the Mean)

Now suppose you have a group of 6 people whose heights (in inches) are 62, 64, 69, 72, 79, and 80. The true population height, then, is the average of these numbers: 71 inches. Suppose you wanted to estimate this value using a sample of size 4. There are 15 possible samples you could take, and the histogram of these means is to the right:

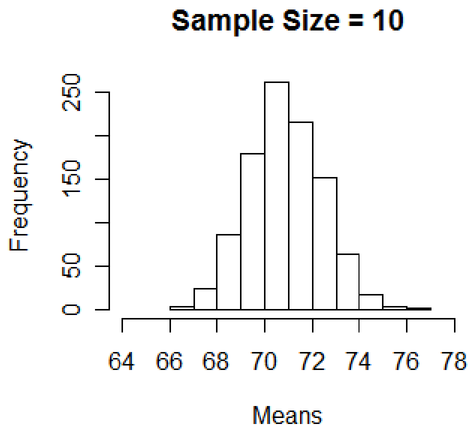
Sample	Sample Mean
62, 64, 69, 72	66.75
62, 64, 69, 79	68.5
62, 64, 69, 80	68.75
62, 64, 72, 79	69.25
62, 64, 72, 80	69.5
62, 64, 79, 80	71.25
62, 69, 72, 79	70.5
62, 69, 72, 80	70.75
62, 69, 79, 80	72.5
62, 72, 79, 80	73.25
64, 69, 72, 79	71

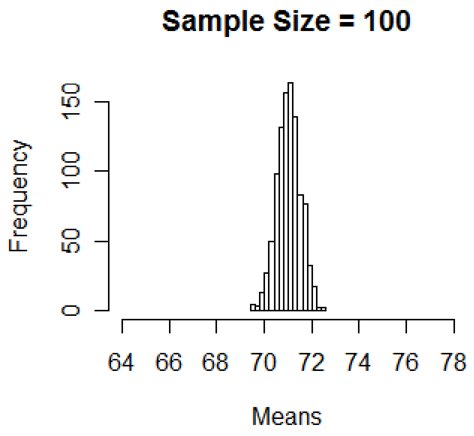


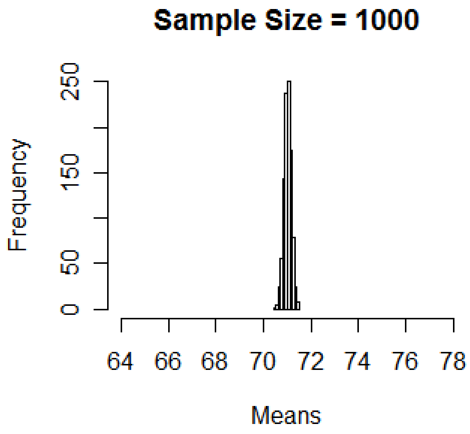
64, 69, 72, 80	71.25
64, 69, 79, 80	73
64, 72, 79, 80	73.75
69, 72, 79, 80	75

We see again that most means are close to the true mean of 71, while larger/smaller values are less common.

To illustrate this in a large sample case, suppose that the population of the entire U.S. has an average height of 71 inches, with a standard deviation of 5 inches. 1000 simulated samples are taken from this population, and the results shown in a histogram. The three figures show histograms for the means of these 1000 samples when the sample size is 10, 100, and 1000.







Again, we see the normal shape, and the variation goes down as the sample size increases.

Which Estimators Target the Population Parameter?
For a discussion of this topic, see p. 276 in the book. It turns out that only some of the sample statistics target the true value. The two most common are:

1. The sample mean targets the population mean.
2. The sample variance targets the population variance.

It turns out, however, that:

1. The sample median does *not* target the population median.
2. The sample standard deviation does *not* target the population standard deviation.
3. The sample range does *not* target the population range.

Section 6-5: The Central Limit Theorem

All of the examples in the previous section form the basis for an intuitive understanding of one of the most important results in statistics: the Central Limit Theorem. First, let's summarize what we saw in the previous section.

1. Sample statistics vary from sample to sample, and thus are random variables which have probability distributions.
2. The bigger the sample size, the more on target the sample mean seems to be (there is less variability).
3. The bigger the sample size, the more bell-shaped (normal) the distribution of the sample mean seems to be.

The theorem itself is as follows.

The Central Limit Theorem:
Given these conditions –

- The random variable X has a distribution with a mean of μ and standard deviation of σ.
- All samples taken are of size *n*.

We can conclude –

- If X was *normally distributed*, and $n > 0$, then:

  The distribution of the sample mean $\overline{X}$ will be *exactly* normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

- If X was *not* normally distributed, and *n* is large*, then:

The distribution of the sample mean $\overline{\mathrm{X}}$ will be *approximately* normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

- If X was *not* normally distributed, and *n* is small*, then:

  No conclusions can be drawn about the distribution of $\overline{\mathrm{X}}$.

*For convention, "*n* is large" usually means $n \geq 30$, and "*n* is small" means $n < 30$.

This theorem is incredibly useful, because it helps us estimate the distribution of the sample mean when the distribution of X is *unknown*, as well as when it is not normal.

Example (X is normal)
Let X represent the height of a U.S. resident. X is normally distributed, with a population mean of 71 inches and standard deviation of 5 inches. Suppose we take a sample of size $n = 64$.

1. Find the probability that the sample mean will be larger than 72.
2. Find the 90th percentile for means.

Solutions

From the CLT, the distribution of the sample mean is *exactly* normal, with mean $\mu = 71$ and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{5}{8} = 0.625$.

Now, we can use the techniques from Section 6-3 to answer the questions.

1. The z-score for 72 is $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{72 - 71}{0.625} = 1.60$

   (Note that the form of the z-score is the same, but we use the mean and standard deviation of the mean, not X)
   Using the table, the area above z = 1.60 is 1 – 0.9452 = **0.0548**.

2. Since we are given an area, we first go to the table to find the associated z-score. In this case, we get z = 1.28 (see previous examples in 6-3). So we use the formula in reverse to get $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \Rightarrow 1.28 = \frac{\bar{x} - 71}{0.625} \Rightarrow 0.8 = \bar{x} - 71 \Rightarrow \bar{x} = 71.8$.

   Thus the 90th percentile is **71.8** inches.

Example (X is not normal)
In an assembly line, the chance of an item being defective is 0.05. Suppose we take a sample of size 100.

1. What is the probability that the average number of defective items in the sample is less than 5.5?
2. What is the probability that the average number of defective items in the sample is more than 4.8?
3. Find the 90th percentile for the average number of defective items in the sample.

Solutions
Now, you may recognize X as being binomial, with $n = 100$ and $p = 0.05$. From the previous chapter, we saw that:
$\mu = np = 100(0.05) = 5$
$\sigma^2 = np(1\text{-}p) = 100(0.05)(0.95) = 4.75$, which means $\sigma = 2.18$.

Therefore, since $n > 30$, the mean will be approximately normal with mean 5 and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{2.18}{10} = 0.218$.

At this point, the solutions become the same as the previous example, only with different numbers.

1. The z-score for a mean of 5.5 will be z = 2.29. From the table, the area below this value is **0.9890**.
2. The z-score for 4.8 is z = -0.92. From the table, the area above this value is 1 – 0.1788 = **0.8212**.
3. The z-score for the 90th percentile is z = 1.28, which gives a mean of **5.2790**.

A Technical Note:

The standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ assumes an infinite population size. Often, it may be useful to consider a *finite population correction factor* when the sample size *n* is more than 5% of the population size *N*.

In this case, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ is used for the standard deviation of the mean.

We will always assume the sample is less than 5% of the population size, and therefore will not be using this correction in class.