Chapter 5: Introduction to Inference: Estimation

Estimating Population Quantities: An Example

Estimating the Effectiveness of a Medication

A pharmaceutical company is testing the effectiveness of a new cholesterol-lowering medication. Specifically, they want to know (1) If the medication reduces low density lipoprotein (LDL) level in people with high LDL, and (2) On average, how much it reduces LDL among those people.

$\approx$

Estimating Population Quantities: An Example

To do so, the company's scientists propose the following study. They will

- Recruit 10 subjects with high levels of LDL cholesterol.
- Make sure the subjects don't take any cholesterol medication for two weeks to ensure an accurate baseline measurement.
- Take an initial baseline reading.
- Take a follow-up LDL measurement after the subject has been 30 days on the test medication.

Estimating Population Quantities: An Example

Here are the resulting data:

Subject	Baseline	Follow-up	LDL Decrease
1	160.5	168.1	−7.6
2	195.3	181.4	13.9
3	181.7	154.6	27.1
4	175.1	160.3	14.8
5	198.3	192.0	6.3
6	215.5	173.5	42.0
7	227.9	186.2	41.7
8	201.7	183.2	18.5
9	161.5	130.3	31.2
10	189.0	165.0	24.0

Estimating Population Quantities: An Example

Suppose for the sake of argument, that there are only 10 people with high LDL in the world, and the company has tested them all. Can the company now answer its two questions: (1) Does the medication reduce LDL in people with high LDL?

Subject	Baseline	Follow-up	LDL Decrease
1	160.5	168.1	−7.6
2	195.3	181.4	13.9
3	181.7	154.6	27.1
4	175.1	160.3	14.8
5	198.3	192.0	6.3
6	215.5	173.5	42.0
7	227.9	186.2	41.7
8	201.7	183.2	18.5
9	161.5	130.3	31.2
10	189.0	165.0	24.0

It did in 9 of the 10. Or, putting it more scientifically, p, the proportion of the population for whom the drug lowers LDL, is 0.9.

$\approx$

Estimating Population Quantities: An Example

(2) On average, how much does the medication change LDL among those people?

Subject	Baseline	Follow-up	LDL Decrease
1	160.5	168.1	−7.6
2	195.3	181.4	13.9
3	181.7	154.6	27.1
4	175.1	160.3	14.8
5	198.3	192.0	6.3
6	215.5	173.5	42.0
7	227.9	186.2	41.7
8	201.7	183.2	18.5
9	161.5	130.3	31.2
10	189.0	165.0	24.0

Answer: It reduced their LDL levels an average of 21.19 mg/dL. Or, putting it more scientifically, μ, the population mean decrease in LDL, equals 21.19.

$\approx$

Estimating Population Quantities: An Example

Because there were only 10 people with high LDL, the company was able to measure the change in LDL levels in all of them, and so could get exact answers to its questions.

In this case the 10 people constitute the **target population** (which we will usually shorten to just **population**): the group to which we want the conclusions to apply.

$\approx$

Estimating Population Quantities: An Example

Of course, there are many more than 10 people with high LDL levels; there are, in fact, millions, and these make up the target population for the company's study. The company still wants to answer its two questions, but it cannot measure the LDL levels of everyone with high LDL.

To get acceptable answers will take more thought and effort.

$\approx$

Estimating Population Quantities: An Example

Selecting the Sample

Since the company can only test its product on a relatively small number of people, the first thing to decide is how to select the subjects for study.

The answer, if they want to draw scientific conclusions and have the results apply to the full population, is to select the subjects using an appropriate **probability sampling method**, such as simple random sampling, stratified random sampling, or other methods discussed in Chapter 3.

Throughout the rest of this course, we will assume samples are selected by **simple random sampling**.

Estimating Population Quantities: An Example

So, suppose the company decides it can afford 10 subjects for its study. Its researchers select 10 at random from the population of all people with high LDL, and once they have their subjects they proceed as described earlier to obtain the data.

$\approx$

Estimating Population Quantities: An Example

If the data are

Subject	Baseline	Follow-up	LDL Decrease
1	160.5	168.1	−7.6
2	195.3	181.4	13.9
3	181.7	154.6	27.1
4	175.1	160.3	14.8
5	198.3	192.0	6.3
6	215.5	173.5	42.0
7	227.9	186.2	41.7
8	201.7	183.2	18.5
9	161.5	130.3	31.2
10	189.0	165.0	24.0

they would answer their questions by saying LDL decreased for 9 of 10 subjects, and the mean decrease was 21.19 mg/dL.

Estimating Population Quantities: An Example

There is one further complication, however.

Because the sample is only a subset of the population (and not the entire population, as we had assumed before), the proportion 0.9 and the mean 21.19 are not the population proportion and mean. Rather, they are called the **sample proportion** and **sample mean**.

$\approx$

Estimating Population Quantities: An Example

Because we are using these sample quantities to estimate their population counterparts, we call them **estimators**. You know that the sample mean is denoted $\overline{y}$, which is the notation we will use.

Sometimes, in order to indicate what is being estimated, the estimator is represented by putting a little hat on the quantity being estimated. For example, we will denote the sample proportion $\hat{p}$ (We could have used $\hat{\mu}$ to represent the sample mean, but since the $\overline{y}$ notation is so widely used, we chose not to).

Estimators such as $\overline{y}$ and $\hat{p}$ that give a single value as an estimate are called **point estimators**.

Estimating Population Quantities: An Example

So from the sample, we estimate the population proportion for whom the medication lowers LDL (p), by the sample proportion $\hat{p} = 0.9$, and we estimate the population mean decrease in LDL (μ) by the sample mean $\overline{y} = 21.19$.

The problem is, we don't know how close these estimates are to the true population values.

To figure this out, we need the notion of a **sampling distribution**.

$\approx$

Sampling Distributions

The sampling distribution of an estimator arises from the idea that we obtain the subjects in our study by sampling randomly from the population.

As a result, each sample is different and will give a different value of the estimator we are using (e.g., the sample mean or sample proportion).

Sampling Distributions

To make this concrete, here are the LDL decreases for 5 different samples from the population. Notice how the values of $\overline{y}$ and $\hat{p}$ vary from sample to sample.

	Sample				
	1	2	3	4	5
	8.2	10.4	30.2	11.7	−1.0
	52.5	52.0	17.2	36.3	25.4
	41.4	25.1	19.0	15.5	3.5
	7.1	3.3	12.1	−23.4	7.9
	38.9	53.8	29.3	21.9	10.5
	50.4	17.0	33.3	13.8	28.9
	30.5	43.1	29.6	35.5	-1.6
	14.4	18.0	22.4	30.4	11.2
	−6.1	7.9	−2.7	16.2	26.6
	−18.9	25.5	15.5	23.9	10.3
$\overline{y}$	21.84	25.61	20.59	18.18	12.17
$\hat{p}$	0.8	1.0	0.9	0.9	0.8

Sampling Distributions

The **sampling distribution** of an estimator is the pattern of variation shown by the values obtained when the estimator is calculated for all possible samples. Let's focus on $\overline{y}$ for the present.

In the LDL example, the sampling distribution of $\overline{y}$ would be the set of values of $\overline{y}$ computed from all possible samples of size 10.

You can think of this as creating a table like the one on the previous slide; instead of just 5 columns, though, there would be one column for each possible sample (which would be a lot of columns!). The resulting set of $\overline{y}$ values would be the sampling distribution of $\overline{y}$.

A good way to display the sampling distribution is to use a histogram. For the LDL example, a histogram of the sampling distribution of $\overline{y}$ for all samples of size 10 might look like Figure 1 on the next slide.

Sampling Distributions

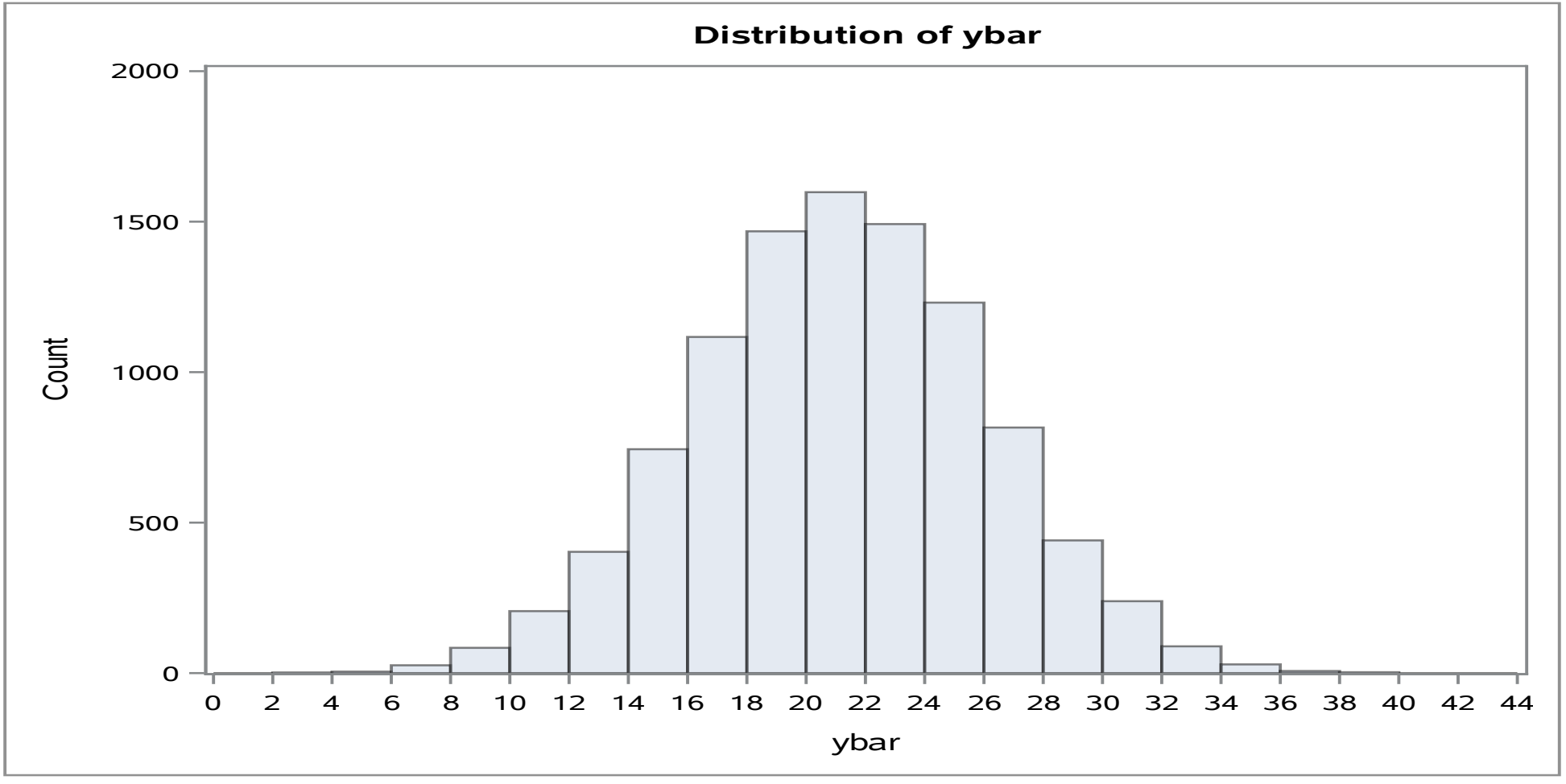

Figure: 1: Histogram showing sampling distribution of $\overline{y}$ for samples of size 10.

Sampling Distributions

Let's explore the sampling distribution of $\overline{y}$ a little more closely.

To do so, we begin with the original population of LDL decrease values. This population will have a mean μ and a standard deviation σ.

It can be shown that the sampling distribution of $\overline{y}$ based on samples of size 10 has mean μ and standard deviation $\sigma/\sqrt{10}$. This is shown in Figure 2 on the next slide, which displays a histogram of the population of LDL decrease values (upper histogram) and a histogram of the sampling distribution of $\overline{y}$ (lower histogram).

$\approx$

Sampling Distributions

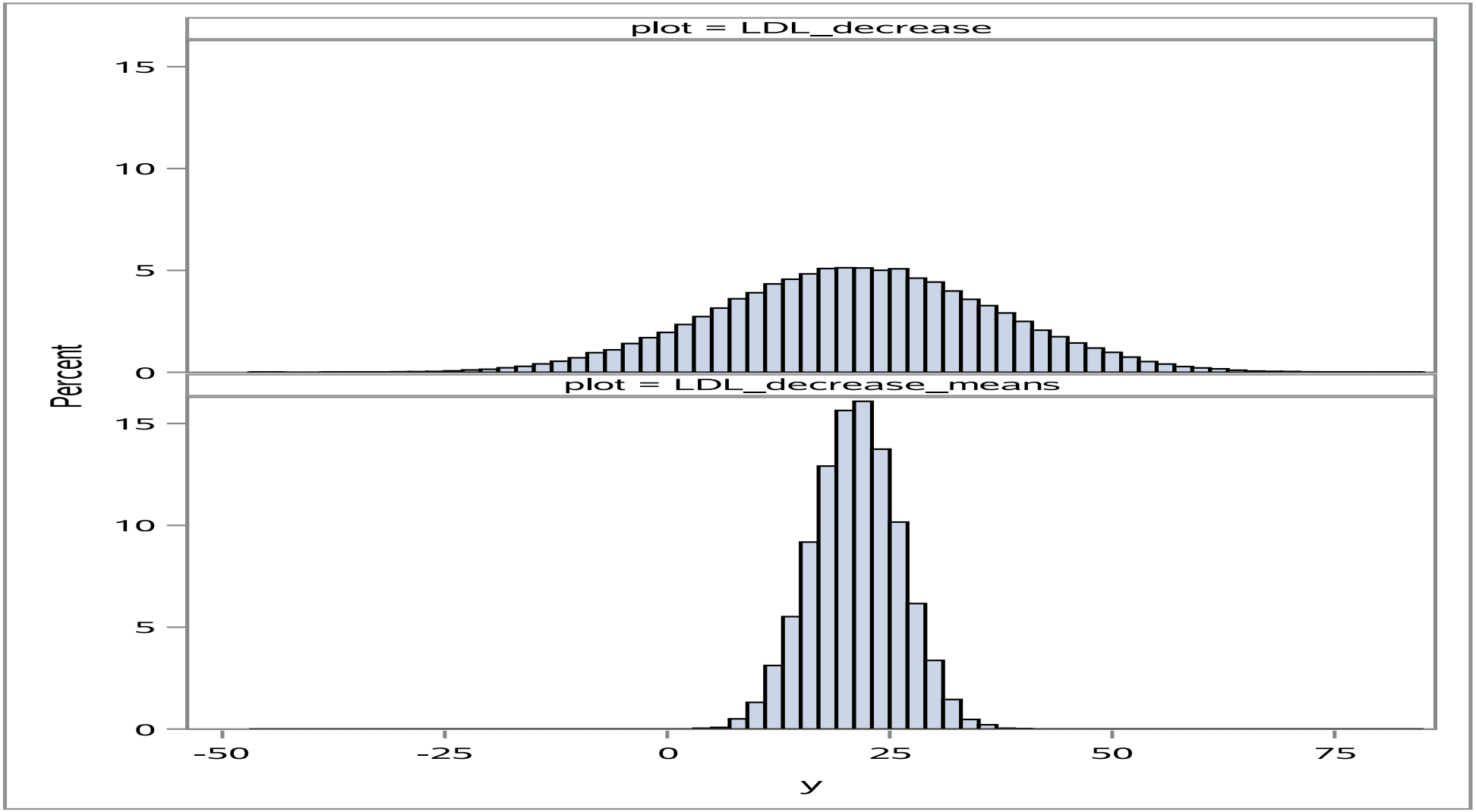

Figure: 2: Top: distribution of population values of LDL decrease; Bottom: sampling distribution of $\bar{y}$ for samples of size 10.

Sampling Distributions

The standard deviation of the sampling distribution of an estimator is called the **standard error** of the estimator. The standard error of $\overline{y}$ based on samples of size 10 is $\sigma/\sqrt{10}$.

$\approx$

Development of Confidence Intervals

Let's **standardize** all the $\overline{y}$ values in the sampling distribution by subtracting their mean μ and dividing the result by their standard error $\sigma/\sqrt{10}$. The resulting values will have a distribution that has mean 0 and standard deviation 1.

Here is the distribution of the standardized $\overline{y}$ values for samples of size 10:

$\approx$

Sampling Distributions

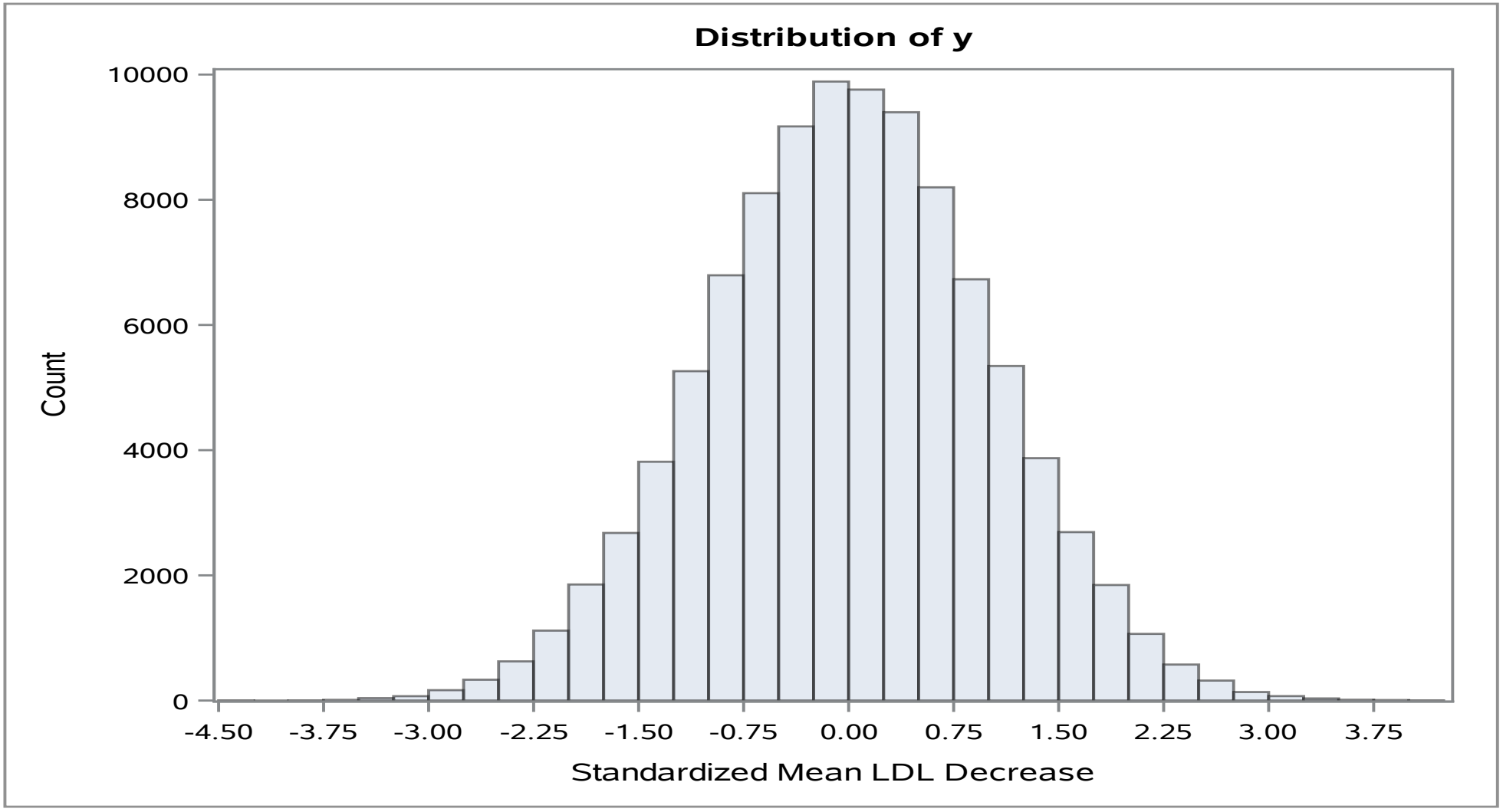

Figure: 3: Distribution of standardized $\bar{y}$ values of LDL decrease for samples of size 10.

Sampling Distributions

If we know the sampling distribution, then given any two values, $a < b$, we know the proportion of the standardized values that lie between a and b.

For example, we have reason to suppose that the proportion of standardized values of $\overline{y}$ between -1.96 and 1.96 is 0.95 (we'll give the reason later).

$\approx$

Development of Confidence Intervals

Assume the proportion of standardized values between -1.96 and 1.96 is 0.95. If we let Pr denote "proportion", we have

$$\begin{aligned} 0.95 &= Pr\left(-1.96 < \frac{\overline{y}-\mu}{\sigma/\sqrt{10}} < 1.96\right) \\ &= Pr\left(-1.96\frac{\sigma}{\sqrt{10}} < \overline{y}-\mu < 1.96\frac{\sigma}{\sqrt{10}}\right) \\ &= Pr\left(\overline{y}-1.96\frac{\sigma}{\sqrt{10}} < \mu < \overline{y}+1.96\frac{\sigma}{\sqrt{10}}\right) \end{aligned}$$

$\approx$

Development of Confidence Intervals

Look carefully at the first and last items in the chain of equalities:

$$0.95 = Pr\left(\overline{y} - 1.96\frac{\sigma}{\sqrt{10}} < \mu < \overline{y} + 1.96\frac{\sigma}{\sqrt{10}}\right).$$

What this says is that if for each possible sample we calculate the interval

$$\left(\overline{y} - 1.96\frac{\sigma}{\sqrt{10}}, \overline{y} + 1.96\frac{\sigma}{\sqrt{10}}\right),$$

then 95% of those intervals will contain the true population mean μ.

For this reason, the interval $\left(\overline{y} - 1.96\frac{\sigma}{\sqrt{10}}, \overline{y} + 1.96\frac{\sigma}{\sqrt{10}}\right)$ is called a **95% confidence interval** for μ.

Development of Confidence Intervals

Because they give a range of likely values for what is being estimated (here, the population mean μ), confidence intervals are examples of what are called **interval estimators**.

The interval estimator, $\left(\overline{y} - 1.96\frac{\sigma}{\sqrt{10}}, \overline{y} + 1.96\frac{\sigma}{\sqrt{10}}\right)$ is more informative than the point estimator $\overline{y}$, because

- If you know the interval, you can figure out what $\overline{y}$ is (it's the center of the interval).
- The interval gives a range of likely values for μ based on the variation in the sampling distribution of $\overline{y}$.

$\approx$

Interpretation of Confidence Intervals

OK, so we've just obtained a 95% confidence interval for μ:

$$\left(\overline{y} - 1.96\frac{\sigma}{\sqrt{10}}, \overline{y} + 1.96\frac{\sigma}{\sqrt{10}}\right).$$

Notice that there is one interval for every possible sample, and from the derivation, 95% of all these intervals contain the true population mean μ.

This is what we mean when we say we are "95% confident" that the interval contains μ.

Here are the results of several simulations to help illustrate this idea:

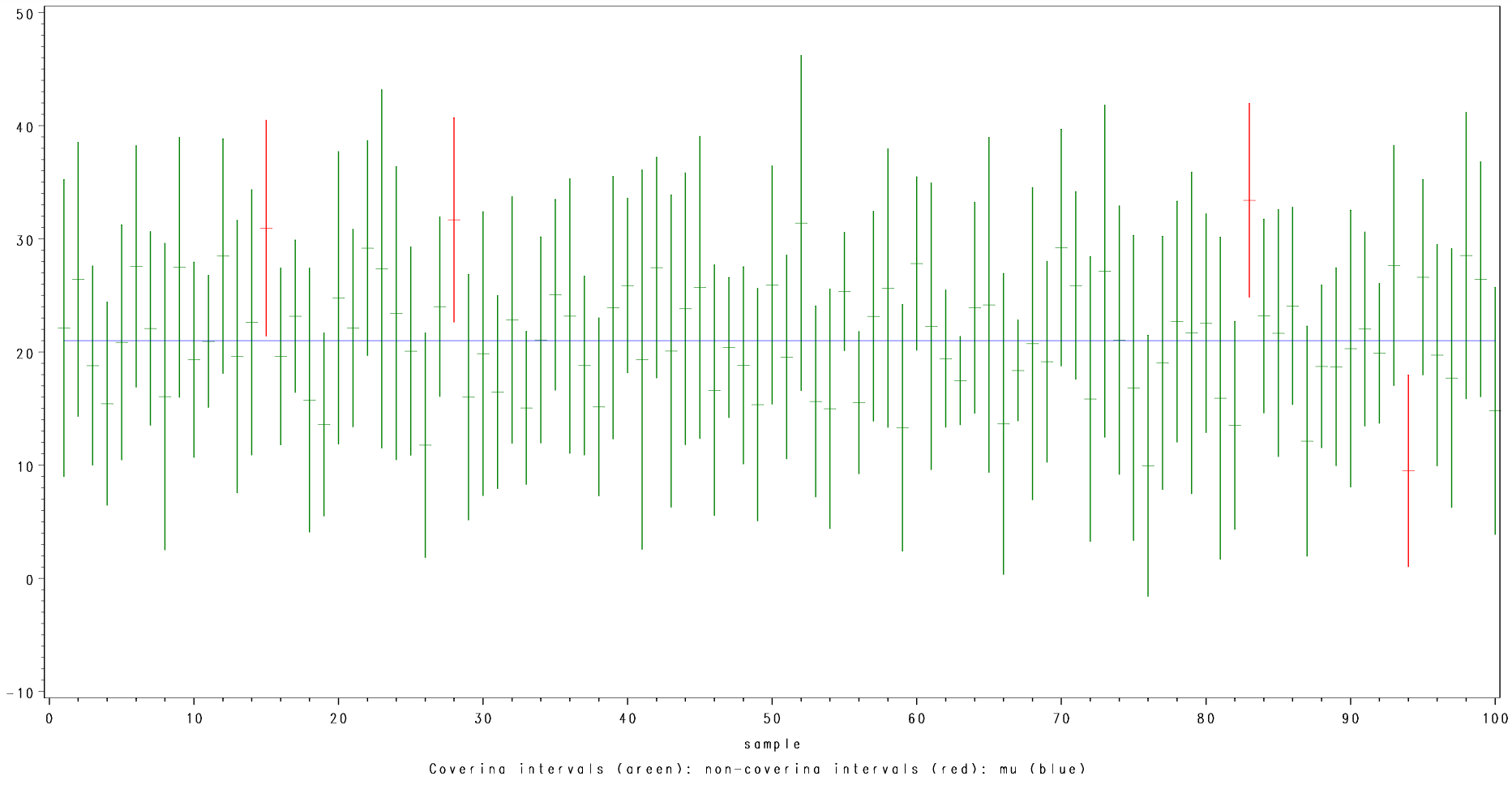

Figure: 4: Simulation of 100 95% confidence intervals for a population mean.

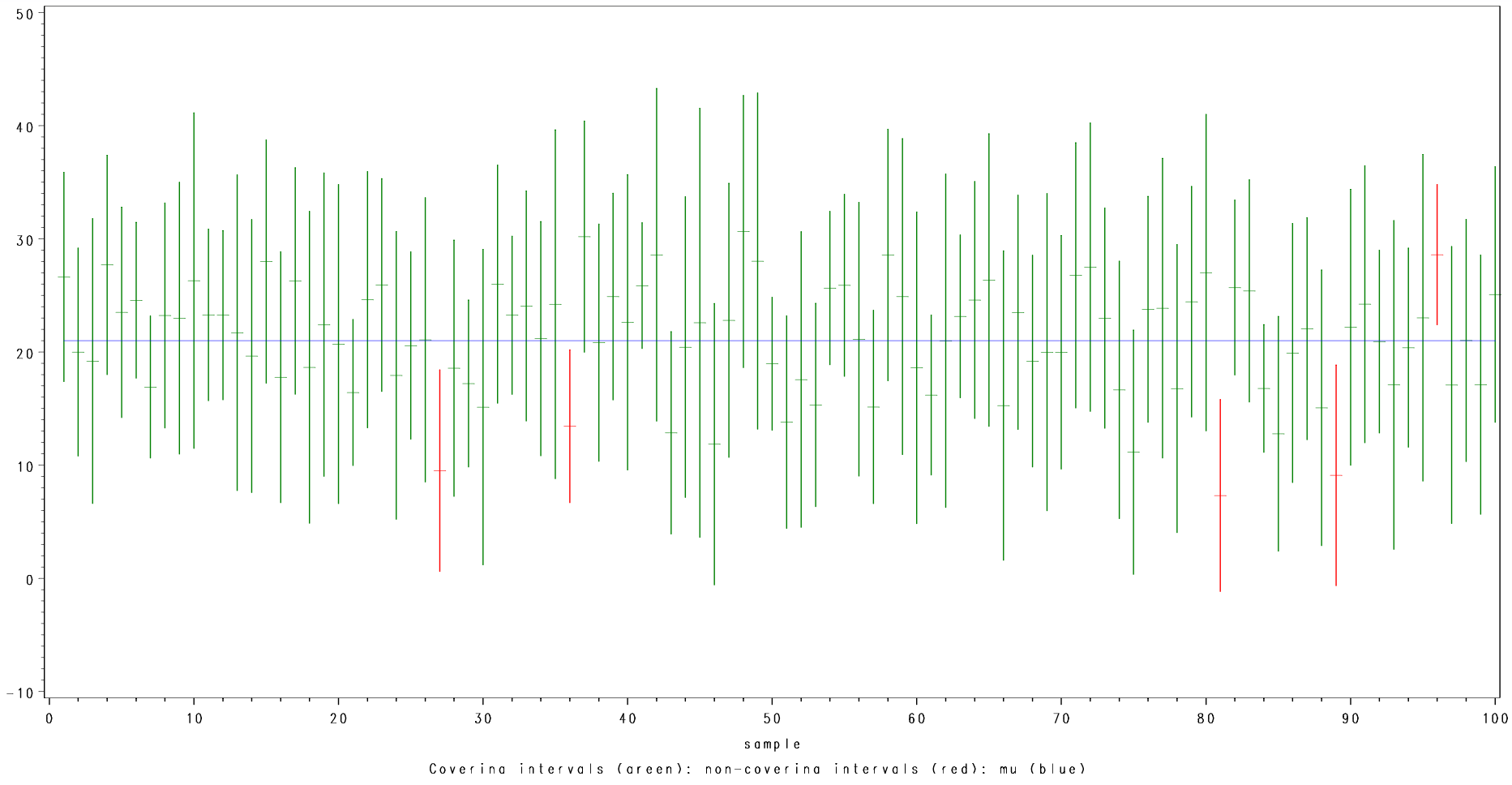

Figure: 5: *Simulation of 100 95% confidence intervals for a population mean.*

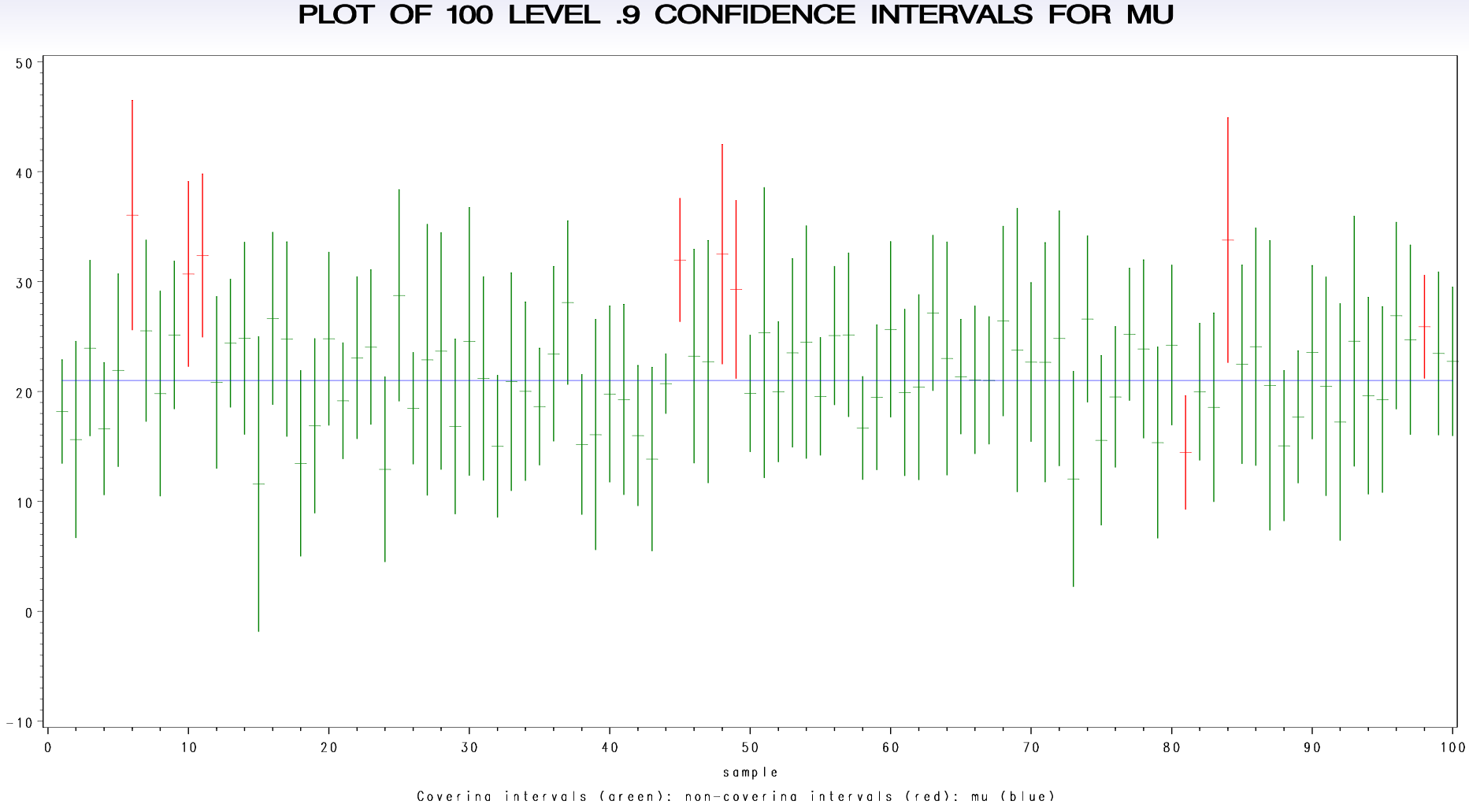

Figure: 6: *Simulation of 100 90% confidence intervals for a population mean.*

Example 1

Assume the 10 LDL decrease values presented earlier resulted from a simple random sample of subjects with high LDL levels. We have seen that the mean of the values is 21.19, which is the estimate $\overline{y}$. Assume we know the population standard deviation, σ, to be 16. Then a 95% confidence interval for the population mean μ is

$$\left(\overline{y} - 1.96\frac{\sigma}{\sqrt{10}}, \overline{y} + 1.96\frac{\sigma}{\sqrt{10}}\right) =$$

$$\left(21.19 - 1.96\frac{16}{\sqrt{10}}, 21.19 + 1.96\frac{16}{\sqrt{10}}\right) = (11.27, 31.11).$$

Thus, with 95% confidence, we estimate that the population mean decrease in LDL is between 11.27 and 31.11. [1]

$\approx$

[1] SAS code found here

Statistical Inference

Using data from a sample to estimate a population quantity, such as the mean, is an example of **statistical inference**. We are using information about a subset of the population (the sample) to **infer** (that is, to conclude) something about the population.

Now that we have some basic ideas of what estimation and confidence intervals are and how they are used, let's develop these ideas in a little more generality.

Theoretical Distributions

In developing the confidence interval, we used the idea of a population distribution: the pattern of variation in data from the entire population. Of course, we can't really know what this distribution is: if we could take data from every unit in the population, we wouldn't need to take a sample.

So what we do is use a mathematical model to give a theoretical population distribution, and derive methods and results from this model.

The Normal Distribution Model

The most frequently used model for continuous data (such as LDL decrease) is the **normal (aka Gaussian) model**. The normal model is given by the normal density curve:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right), \quad -\infty < x < \infty.$$

As you might guess, the mean of this distribution is μ and the standard deviation is σ (which means the variance is σ^2). We use the notation $N(\mu, \sigma^2)$ to denote the normal distribution with mean μ and variance σ^2.

The density curve of the normal distribution is the famous bell-shaped curve. Here are several examples:

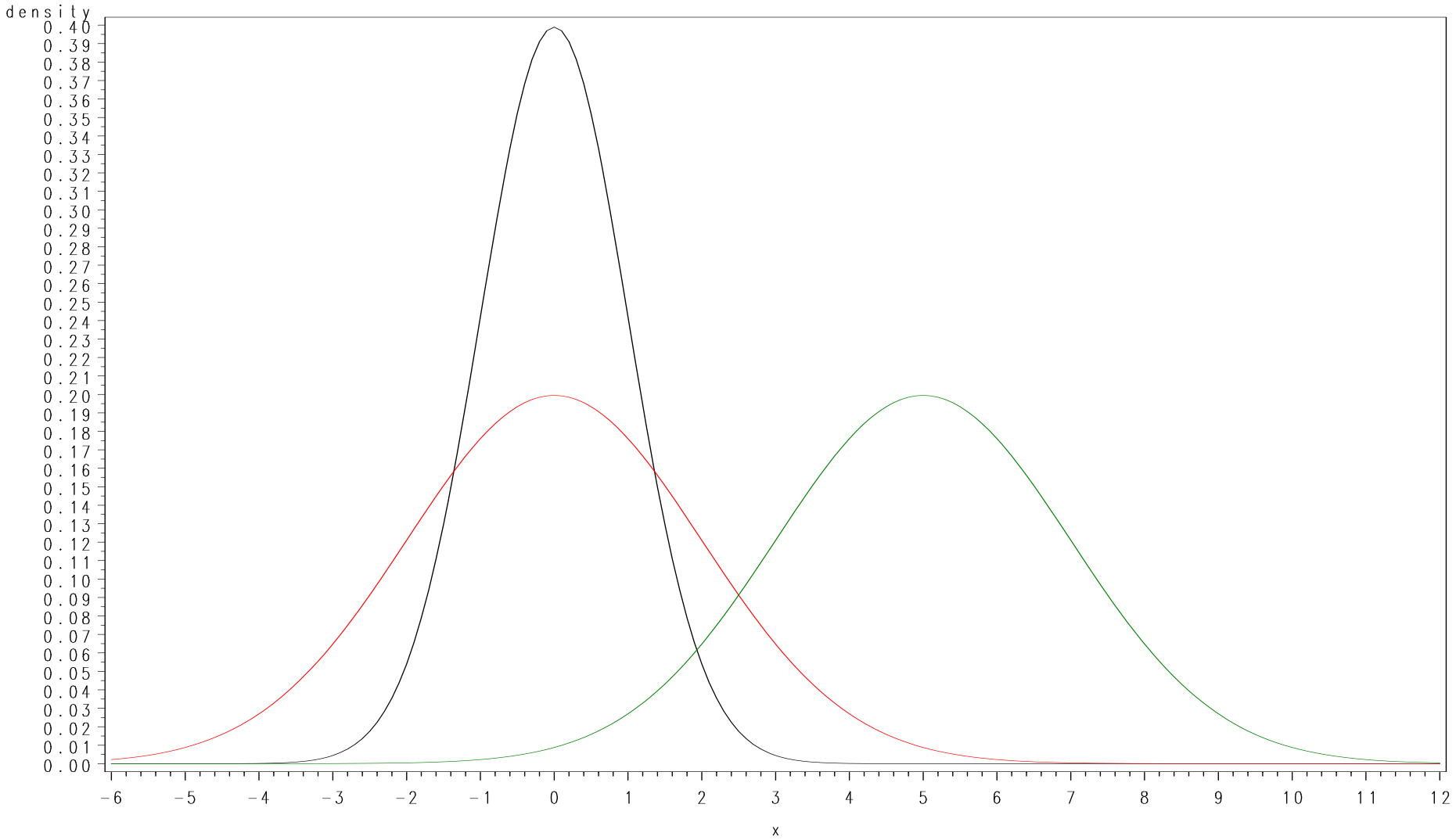

Figure: 7: Normal density curves. N(0,1) (tall curve at left), N(0,4) (short curve at left), and N(5,4).

The Normal Distribution Model

Some Important Characteristics of the Normal Density Curve

1. The curve is unimodal and symmetric about μ.
2. For any a<b, the area under the curve between a and b is the proportion of the population values falling between a and b.
3. The total area under the curve is ... 1.
4. If a population of values follows a $N(\mu, \sigma^2)$ distribution, and if we standardize each value by subtracting the mean μ and dividing the result by the standard deviation σ, the population of standardized values follows a $N(0, 1)$ (called a **standard normal**) distribution.

The Normal Distribution Model

Several slides ago, when we were developing a confidence interval for μ, we stated that "the proportion of standardized values between -1.96 and 1.96 is 0.95."

We were making the assumption that the population of estimators $\overline{y}$ followed a normal distribution. When we standardized these, the standardized values followed a standard normal distribution.

By numerical computation (since we can't generally find areas under a normal curve exactly), we can show that the area under the standard normal density between -1.96 and 1.96 is 0.95.

$\approx$

The Normal Distribution Model

Table A.3 of the text provides areas under the standard normal density curve. As shown below, the value in the table for $z = 1.96$ is 0.9750, which means that 97.5% of all $N(0, 1)$ population values lie below 1.96.

	$N(0,1)$ Probabilities								
z	.00	.01	.02	.03	.04	.05	.06	.07	.
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5
.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9

$\approx$

The Normal Distribution Model

Similarly, the value in the table for $z = -1.96$ is 0.0250, which means that 2.5% of all $N(0, 1)$ population values lie below -1.96. From this, we deduce that 95% of all $N(0, 1)$ population values lie between -1.96 and 1.96.

	$N(0,1)$ Probabilities							
z	.00	.01	.02	.03	.04	.05	.06	.07
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244

$\approx$

The Normal Distribution Model

For any number q between 0 and 1, we define the q quantile of the standard normal distribution as the value z_q below which lies area q under the standard normal curve. From what we have just seen, $z_{0.025} = -1.96$ and $z_{0.975} = 1.96$. In a similar way, we can find any quantile z_q using Table A.3. Note this and other tables are in the text's appendices (pdf posted on Canvas).

Modern technology has given other options for finding the quantile as well. Many calculators (my TI-84, for instance) have functions to do so. There are also a number of online calculators available, such as this.

$\approx$

Example 1.5

Specification limits for the diameter of a nanoglobule are 1 to 8 microns. If the distribution of nanoglobule diameters is $N(5, 4)$, what percent of nanoglobules meet specification?

If X represents the diameter of a nanoglobule, the proportion of nanoglobules in spec is (keep in mind that $\mu = 5$ and $\sigma = 2$)

$$\begin{aligned} Pr(1 < X < 8) &= Pr\left(\frac{1-5}{2} < \frac{X-5}{2} < \frac{8-5}{2}\right) \\ &= Pr(-2 < Z < 1.5) \\ &= Pr(Z < 1.5) - Pr(Z < -2) \\ &= 0.9332 - 0.0227 \\ &= 0.9105, \end{aligned}$$

so 91.05% of nanoglobules meet specification.

$\approx$

Synopsis: Confidence Intervals for the Population Mean

- We derived the formula for a 95% confidence interval for the population mean μ based on a sample of size 10 from a $N(\mu, \sigma^2)$ population: $\left(\overline{y} - 1.96\frac{\sigma}{\sqrt{10}}, \overline{y} + 1.96\frac{\sigma}{\sqrt{10}}\right)$.
- We can generalize this to samples of any size n: $\left(\overline{y} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{y} + 1.96\frac{\sigma}{\sqrt{n}}\right)$.
- 1.96 equals $z_{0.975}$, the 0.975 quantile of the $N(0, 1)$ (standard normal) distribution, so the formula is $\left(\overline{y} - z_{0.975}\frac{\sigma}{\sqrt{n}}, \overline{y} + z_{0.975}\frac{\sigma}{\sqrt{n}}\right)$.

$\approx$

Confidence Intervals for the Population Mean

Finally, for any number L between 0 and 1, we can obtain the formula for a level L confidence interval (also known as a $(100 \times L)\%$ confidence interval) for μ:

$$\left(\overline{y} - z_{\frac{1+L}{2}} \frac{\sigma}{\sqrt{n}}, \overline{y} + z_{\frac{1+L}{2}} \frac{\sigma}{\sqrt{n}} \right).$$

Some common confidence levels are

Percent	Level	Normal Quantile
90%	$L = 0.90$	$z_{\frac{1+L}{2}} = z_{\frac{1+0.90}{2}} = z_{0.95} = 1.6449$
95%	$L = 0.95$	$z_{\frac{1+L}{2}} = z_{\frac{1+0.95}{2}} = z_{0.975} = 1.9600$
99%	$L = 0.99$	$z_{\frac{1+L}{2}} = z_{\frac{1+0.99}{2}} = z_{0.995} = 2.5758$

The Central Limit Theorem

In deriving the formula for the confidence interval for the population mean, we have been assuming that the sampling distribution of the sample mean $\overline{y}$ is normal.

If the original population distribution is normal, we know that this assumption is correct: For a sample of size n from a $N(\mu, \sigma^2)$ population, the sampling distribution of $\overline{y}$ is $N(\mu, \sigma^2/n)$.

However, these same confidence interval formulas work quite well for many populations with non-normal distributions: better, in fact, than we might suppose they would.

The Central Limit Theorem

It turns out that the reason is a mathematical result called **The Central Limit Theorem**.

Basically, the Central Limit Theorem says that regardless of the population distribution of the quantity being measured, if the sample size is sufficiently large, then the sampling distribution of the sample mean is approximately normal.

$\approx$

The Central Limit Theorem

Specifically, if the population mean is μ and the population standard deviation is σ, then if the sample size n is sufficiently large, the sample mean will have approximately a $N(\mu, \sigma^2/n)$ distribution.

"Sufficiently large" varies with the population distribution, but samples of size 25 or 30 will make the CLT valid for most cases, and samples of size 100 for almost all cases found in practice.

As a result, when the Central Limit Theorem applies, the confidence interval formula we have been studying gives reliable results in a wide range of applications. Here is a picture to illustrate the Central Limit Theorem.

The Central Limit Theorem

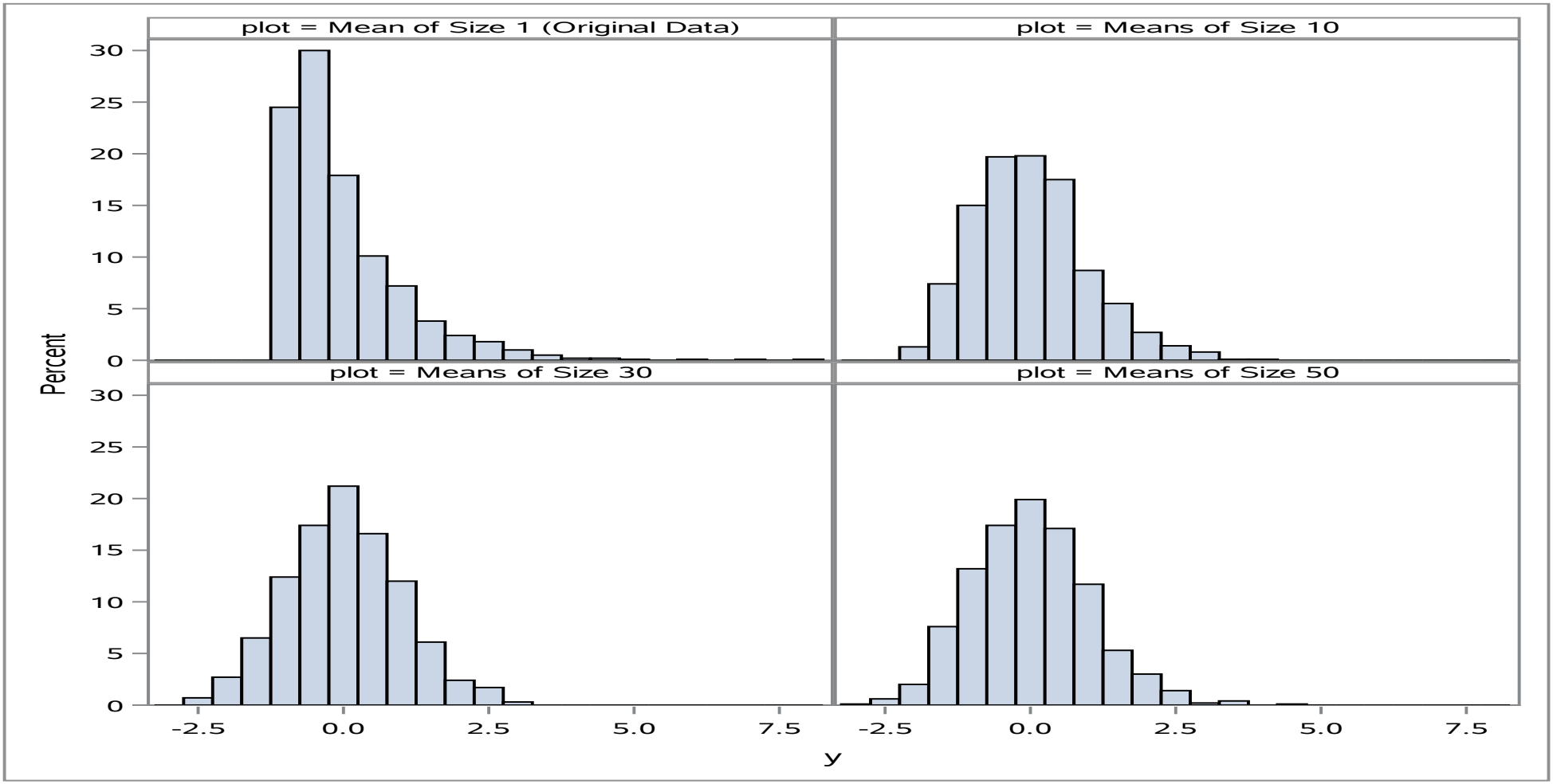

Figure: 1: The CLT in action. Distribution of standardized original data (upper left), and of standardized means of size 10 (upper right), 30 (lower left), and 50 (lower right).

Determining Sample Size

One of the first questions that has to be answered in designing any study is "What sample size is needed?". We have seen that the Central Limit Theorem provides some guidance on this.

Another consideration is the **precision** desired in estimators. Precision of an estimator is a measure of how variable that estimator is. One way of expressing precision is the width of a level L confidence interval. For a given population and confidence level, precision is a function of the size of the sample: the larger the sample, the greater the precision.

Determining Sample Size

For example, recall the confidence interval for a population mean:

$$\left(\overline{y} - z_{\frac{1+L}{2}} \frac{\sigma}{\sqrt{n}}, \overline{y} + z_{\frac{1+L}{2}} \frac{\sigma}{\sqrt{n}}\right).$$

Half the width of this interval is

$$z_{\frac{1+L}{2}} \frac{\sigma}{\sqrt{n}}.$$

If we want a precision d (so that the half length of this interval is less than or equal to d), we have

$$d \geq z_{\frac{1+L}{2}} \frac{\sigma}{\sqrt{n}},$$

which implies

$$n \geq (\sigma^2 \cdot z^2_{\frac{1+L}{2}})/d^2.$$

$\approx$

Example 2

In the LDL decrease study, the standard deviation was taken to be 16. If the researchers want a level 0.95 confidence interval to estimate the population mean decrease with a precision of 1/2 mg/dL, they should select a sample of size

$$n \geq 16^2 \cdot 1.96^2/(1/2)^2 = 3933.8.$$

Their sample should consist of 3934 subjects: a far cry from the 10 subjects used in the study!

$\approx$

The Components of a Statistical Estimation Problem

We will divide a statistical estimation problem into five steps

1. **The Scientific Goal**
2. **The Statistical Model**
3. **The Model Parameter(s) to Be Estimated**
4. **Point and Interval Estimates**
5. **Results and Interpretation**

We illustrate using the LDL reduction study

$\approx$

The Components of a Statistical Estimation Problem

1. **The Scientific Goal** The scientific goal is the reason for doing the experiment or study. In this example, there are two scientific goals: (a) Does the medication reduce LDL for a substantial proportion of people with high LDL levels? and (b) By how much does it reduce LDL among people with high LDL? Here, we will focus on (b).
2. **The Statistical Model** The statistical model is the distribution of the population of measurements that are being taken. In this case, the measurements are the LDL decreases and we will assume the population has a $N(\mu, 16^2)$ distribution.

The Components of a Statistical Estimation Problem

3. **The Model Parameter(s) to Be Estimated** At this point we examine how to achieve the scientific goal in terms of the statistical model. If we can't formulate the scientific goal in these terms, we shouldn't be doing a statistical estimation problem.

   In the LDL reduction study, we will focus on the average effect of the medication, so that the scientific goal is to find how much, on average, the medication reduces LDL among people with high LDL. This suggests that the model parameter we want to estimate is the population mean LDL reduction: μ.

The Components of a Statistical Estimation Problem

4. **Point and Interval Estimates** Often, the point estimator is the sample version of the model parameter to be estimated. This is true when we want to estimate the mean of a $N(\mu, \sigma^2)$ population: the estimator of μ is the sample mean, $\overline{y}$. For the LDL data, the point estimate is the value of $\overline{y} = 21.19$.

   We have seen that the formula for a level L confidence interval for μ when σ is known is

$$\left(\overline{y} - z_{\frac{1+L}{2}}\frac{\sigma}{\sqrt{n}}, \overline{y} + z_{\frac{1+L}{2}}\frac{\sigma}{\sqrt{n}}\right).$$

   To compute the interval, we have to choose a confidence level, L. For $L = 0.95$, we have seen that the interval for the LDL data is

$$\left(\overline{y} - 1.96\frac{16}{\sqrt{10}}, \overline{y} + 1.96\frac{16}{\sqrt{10}}\right) = (11.27, 31.11).$$

$\approx$

The Components of a Statistical Estimation Problem

5. **Results and Interpretation** We have to be a bit careful to state these correctly. For point estimation,you are always on solid ground making a statement like "The estimate of the mean LDL reduction is 21.19 mg/dL."

   However, you should always give some indication of the variation in the estimate. You can do this by giving the standard error of the estimate (here, $16/\sqrt{10} = 5.06$), or a confidence interval.

   When reporting a confidence interval, you are on safest ground making a statement such as "A 95% confidence interval for the mean LDL reduction is (11.27,31.11) mg/dL."

The Components of a Statistical Estimation Problem

Be warned, however, that on homework, quiz and test, I may ask you for a practical interpretation of "confidence" and of the interval you produce.

Here, the correct interpretation of "95% confidence" is that 95% of all possible samples will produce intervals that contain the true population mean LDL reduction.

As an example of practical interpretation, suppose it is known that the mean LDL reduction provided by the present LDL reduction medication is 5.88. The 95% interval produced by this study estimates the mean reduction under the new medication is in the range (11.27,31.11). Does this result support the contention that the new medication is more effective than the present medication?

Answer: yes, since 5.88 lies below the entire confidence interval.

Estimating the Mean when the Variance is Unknown

Up to now in our development of confidence intervals for a population mean μ, we have assumed that the population variance σ^2 is known. This is often unrealistic.

It turns out that the right thing to do in this case is to use the sample standard deviation, s, in place of the unknown population standard deviation σ.

Recall that

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

$\approx$

Estimating the Mean when the Variance is Unknown

That is, instead of assuming that the sampling distribution of the sample mean $\overline{y}$ has mean μ and standard deviation $\sigma/\sqrt{n}$, we assume it has mean μ and standard deviation $s/\sqrt{n}$

$\approx$

Estimating the Mean when the Variance is Unknown

Recall that the derivation of the confidence interval used the fact that the standardized sample mean $Z = (\overline{y} - \mu)/(\sigma/\sqrt{n})$ has a $N(0, 1)$ distribution.

Replacing σ by s gives a different standarized estimator $t = (\overline{y} - \mu)/(s/\sqrt{n})$, which not surprisingly has a different sampling distribution, called a **t distribution**, sometimes named **Student's t distribution** after the statistician who discovered it.

$\approx$

A Slight Detour:

The t distributions are a family of distributions having an integer parameter ν (and written t_ν), called the **degrees of freedom**. The t distribution density curves look like standard normal density curves, except they are lower in the center and higher in the tails.

Figure 8 compares some t densities with the standard normal density.

$\approx$

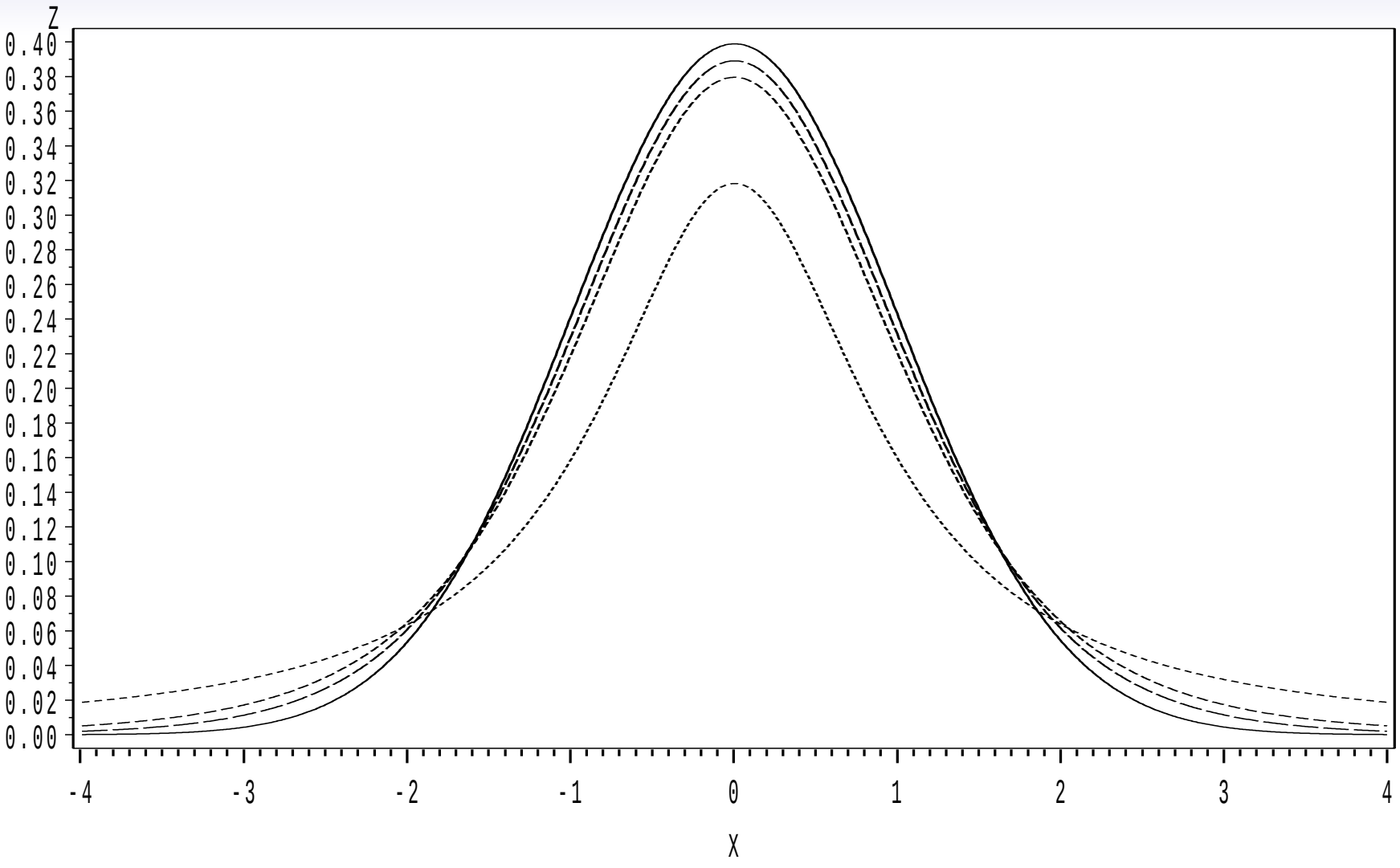

Figure: 8: In order of decreasing center height: $N(0,1)$, t_{10}, t_5 *and* t_1.

Estimating the Mean when the Variance is Unknown

If the original data are from a $N(\mu, \sigma^2)$ distribution, the standardized mean $t = (\overline{y} - \mu)/(s/\sqrt{n})$ has a t distribution with $n - 1$ degrees of freedom.

If we let $t_{n-1,q}$ denote the q^{th} quantile of the t_{n-1} distribution, and mimic the derivation of the confidence interval for the mean when σ is known, we get the formula for a level L confidence interval for the mean:

$$\left(\overline{y} - t_{n-1,\frac{1+L}{2}} \frac{s}{\sqrt{n}}, \overline{y} + t_{n-1,\frac{1+L}{2}} \frac{s}{\sqrt{n}}\right).$$

$\approx$

Example 1, Revisited

Let's go back to the LDL decrease data. Recall that the mean decrease was $\bar{y} = 21.19$. Previously, we assumed $\sigma = 16$. Let's now suppose we don't know σ, but use the sample standard deviation $s = 15.45$. To compute a 95% confidence interval for μ, we need to know $t_{9,0.975}$. The value is 2.2622, and as with normal quantiles, can be obtained in a number of ways: a calculator, an online site such as this, or from a table such as Table A.4 of the text, as the next slide shows.

$\approx$

Estimating the Mean when the Variance is Unknown

	Critical Values of the t Distribution					
Degrees of Freedom, k	$t_{k,.90}$	$t_{k,.95}$	$t_{k,.975}$	$t_{k,.99}$	$t_{k,.995}$	$t_{k,.99}$
1	3.0777	6.3137	12.7062	31.8205	63.6567	318.309
2	1.8856	2.9200	4.3027	6.9646	9.9248	22.327
3	1.6377	2.3534	3.1824	4.5407	5.8409	10.215
4	1.5332	2.1319	2.7764	3.7469	4.6041	7.173
5	1.4759	2.0150	2.5706	3.3649	4.0321	5.893
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.208
7	1.4149	1.8946	2.3646	2.9980	3.4995	4.785
8	1.3968	1.8595	2.3060	2.8965	3.3554	4.501
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.297

$\approx$

Estimating the Mean when the Variance is Unknown

Thus, a 95% confidence interval for μ is

$$\left(\overline{y} - t_{n-1,\frac{1+L}{2}}\frac{s}{\sqrt{n}}, \overline{y} + t_{n-1,\frac{1+L}{2}}\frac{s}{\sqrt{n}}\right) =$$

$$\left(21.19 - 2.2622\frac{15.45}{\sqrt{10}}, 21.19 + 2.2622\frac{15.45}{\sqrt{10}}\right) = (10.14, 32.24)^2$$

Notice that this interval is a bit wider than the earlier one ((11.27,31.11)), despite the fact that s is smaller than the value of σ we assumed before. This reflects the greater uncertainty that results from estimating σ.

$\approx$

[2] SAS code found here

Inference for a Population Proportion

Recall that one of the initial questions for the LDL reduction study was whether the tested medication reduces LDL. One way of answering this question might be to decide for what proportion of the population the medication does lower LDL.

The data obtained showed that LDL decreased for 9 of the 10 subjects. How can we use this information to answer the question?

Inference for a Population Proportion

To begin, let p denote the proportion of the population for whom the medication will lower LDL. To estimate p, we will use the proportion of the sample whose LDL decreased. Denoting this sample proportion $\hat{p}$, we have $\hat{p} = 9/10 = 0.9$.

$\hat{p}$ is the point estimate for p.

To obtain a 95% confidence interval for p, we need information about the sampling distribution of $\hat{p}$, which we now present in a little more generality.

$\approx$

Inference for a Population Proportion

Suppose we have a sample of size n (in our example $n = 10$), and that $\hat{p}$ is the proportion in the sample having the characteristic of interest (in our example, a decrease in LDL). It can be shown that the mean and variance of $\hat{p}$ are p and $p(1-p)/n$, respectively.

If n is large, the Central Limit Theorem will apply to $\hat{p}$, and will ensure that the distribution of $\hat{p}$ standardized by subtracting the mean p and dividing by the standard error $\sqrt{p(1-p)/n}$, will approximately follow a standard normal distribution.

$\approx$

Inference for a Population Proportion

So, an approximate large sample level L confidence interval for p has endpoints

$$\hat{p} \pm z_{\frac{(1+L)}{2}} \sqrt{\frac{p(1-p)}{n}}$$

This presents a problem. Can you tell me why?

That's right: we have to know p to compute it.

One solution, which works well for large samples, is to replace p with $\hat{p}$.

$\approx$

Inference for a Population Proportion

Therefore, an approximate large sample level L confidence interval for p has endpoints

$$\hat{p} \pm z_{\frac{(1+L)}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

For our example, a level 0.95 interval would be

$$0.9 \pm 1.96\sqrt{\frac{(0.9)(1-0.9)}{10}} = (0.714, 1.086) \longrightarrow (0.714, 1.0),$$

with the rounding being done since p cannot be bigger than 1.

$\approx$

Inference for a Population Proportion

While this interval works well for large samples, it does not work so well for small samples. However, a simple adjustment will make this interval work well for both small and large samples.

The adjustment consists of adding "fudge factors" to the numerator and denominator of $\hat{p}$. Here's how it goes:

$\approx$

Inference for a Population Proportion

Assume we want a level L confidence interval. If we let y denote the number of items in the sample having the characteristic of interest (so that $\hat{p} = y/n$), compute the adjusted sample proportion

$$\tilde{p} = \frac{y + 0.5z^2_{\frac{(1+L)}{2}}}{\tilde{n}},$$

where

$$\tilde{n} = n + z^2_{\frac{(1+L)}{2}}$$

The adjusted confidence interval has endpoints

$$\tilde{p} \pm z_{\frac{(1+L)}{2}} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}.$$

$\approx$

Inference for a Population Proportion

For the LDL example, $y = 9$, $n = 10$, $L = 0.95$, so $z_{\frac{(1+L)}{2}} = z_{0.975} = 1.96$. Thus,

$$\tilde{n} = 10 + 1.96^2 = 13.8416,$$

$$\tilde{p} = \frac{9 + (0.5)(1.96^2)}{13.8416} = 0.789,$$

And the interval is

$$0.789 \pm 1.96\sqrt{\frac{0.789(1 - 0.789)}{13.8416}} = (0.574, 1.003),$$

which we should report as $(0.574, 1.0)$.[3]

[3] SAS code found here

Inference for a Population Proportion

We can say that with 95% confidence we estimate that the proportion of the population for whom the medication will decrease LDL is between 0.574 and 1.

The interpretation of 95% confidence is that if we take all possible samples from the population, and for each conduct the experiment and construct a confidence interval of this type, then 95% of all those intervals will contain the true population proportion p.

$\approx$

Inference for a Population Proportion

This interval is called an **approximate score or Agresti-Coull confidence interval** for p. "Approximate score" tells the statistical idea behind the interval, and "Agresti-Coull" credits the two researchers who developed the idea. Because this is an approximate interval, the true confidence level may differ from the advertized level, though in general the approximation is good.

NOTE: In all homework and lab assignments, and the final I want you to use the approximate score interval for estimating a population proportion.

$\approx$

The Components of a Statistical Estimation Problem: Estimating a Population Proportion in the LDL Study

1. **The Scientific Goal:** As stated previously: "Does the medication reduce LDL in people with high LDL?" We will take this to be answered if we can tell what proportion of the population obtains lower LDL from the medication.

The Components of a Statistical Estimation Problem: Estimating a Population Proportion in the LDL Study

2. **The Statistical Model:** Suppose

(a) All units in a large population can be classified as having or not having a certain characteristic.

(b) The proportion of the population having the characteristic is p.

(c) A simple random sample of size n is taken from the population.

The statistical model for the number of units in the sample having the characteristic is called a **binomial model with parameters n and p** (abbreviated $b(n, p)$). [4]

$\approx$

[4] We will have more to say about the binomial model in chapter 6.

The Components of a Statistical Estimation Problem: Estimating a Population Proportion in the LDL Study

In the LDL example, the characteristic is reduction in LDL level after taking the medication. Since a simple random sample of size 10 is taken, the statistical model is $b(10, p)$, where p is proportion of the target population for whom this medication would lower LDL levels.

$\approx$

The Components of a Statistical Estimation Problem: Estimating a Population Proportion in the LDL Study

3. **The Model Parameter(s) to Be Estimated:** p

4. **Point and Interval Estimates:**

a. Point estimate: $\hat{p} = 9/10 = 0.9$.

b. 95% approximate score CI:

Adjusted estimate of p:

$$\tilde{p} = \frac{9 + (0.5)(1.96^2)}{10 + 1.96^2} = 0.789.$$

Interval:

$$0.789 \pm 1.96\sqrt{\frac{0.789(1 - 0.789)}{10 + 1.96^2}} = (0.574, 1.003) \longrightarrow (0.574, 1.0).$$

$\approx$

The Components of a Statistical Estimation Problem: Estimating a Population Proportion in the LDL Study

5. **Results and Interpretation:** Point estimate of population proportion who obtain lower LDL from the medication: 0.9. 95% confidence interval: (0.574,1.0). In particular, we estimate with 95% confidence that the medication will lower LDL for more than 57% of the population.

$\approx$

Comparing Two Population Means

Many applications of statistics involve comparisons: different products, processes, and treatments are frequently compared.

Although it was not presented as such, estimation of mean LDL reduction in the LDL study can be viewed as comparing two population means. It will be instructive to see why.

Comparing Two Population Means: Paired Data

Recall that in the study each of 10 subjects obtained in a simple random sample was measured for LDL at the outset and then after 30 days on a particular medication. Our analysis focused on estimating the mean reduction in LDL.

But viewed another way, this mean reduction is the difference of two population means: $\mu_{pre} - \mu_{post}$, where μ_{pre} is the mean of the before (or untreated) population, and μ_{post} is the mean of the after (or treated) population.

$\approx$

Comparing Two Population Means: Paired Data

The before and after measurements are said to be **paired**, because each individual provides one before-after pair. Pairing is an example of blocking, as you studied in chapter 3.

When paired data measurements are of the same quantity, as in the LDL example, analysis is often done by subtracting one paired measurement from the other and treating the resulting difference as if it were a single measurement.

This is exactly what we did by choosing to analyze the decrease in LDL.

Comparing Two Population Means: Paired Data

Recall that the data gave a point estimate of 21.19 for the population mean decrease, and a 95% confidence interval (10.14,32.24). If we choose, we can interpret the first result as estimating the difference between pre and post-treatment population mean LDL (i.e., $\mu_{pre} - \mu_{post}$) as 21.19, and the second as saying that with 95% confidence we estimate that difference in population means to be between 10.14 and 32.24.

$\approx$

Comparing Two Population Means: Independent Populations

Not all comparisons are done using paired data. Sometimes our data consist of independent random samples from two separate populations.

Suppose that we take a random sample of size n_1 from population 1, which follows a $N(\mu_1, \sigma^2)$ distribution, and independently a random sample of size n_2 from population 2, which follows a $N(\mu_2, \sigma^2)$ distribution.

Notice that the only possible difference in the population distributions is in their means.

Comparing Two Population Means: Independent Populations

We already know that the estimator of μ_1 is the sample mean of the first sample, $\overline{y}_1$, and that of μ_2 is the sample mean of the second sample, $\overline{y}_2$.

We also know that the sampling distribution of $\overline{y}_1$ is $N(\mu_1, \sigma^2/n_1)$ and that of $\overline{y}_2$ is $N(\mu_2, \sigma^2/n_2)$.

$\approx$

Comparing Two Population Means: Independent Populations

If we want to estimate $\mu_1 - \mu_2$, common sense says to use $\overline{y}_1 - \overline{y}_2$.

Although "common sense" can sometimes lead you astray, in this case, it doesn't fail: $\overline{y}_1 - \overline{y}_2$ is exactly the right point estimator to use.

We also want to obtain a confidence interval for $\mu_1 - \mu_2$. For this, we need the information that the sampling distribution of $\overline{y}_1 - \overline{y}_2$ is $N(\mu_1 - \mu_2, \sigma^2/n_1 + \sigma^2/n_2) = N(\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2))$.

$\approx$

Comparing Two Population Means: Independent Populations

Then using the fact that the standardized estimator

$$Z = \frac{\overline{y}_1 - \overline{y}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1),$$

and following the same logic as we did in developing the one sample confidence interval for the mean, we get a level L confidence interval for $\mu_1 - \mu_2$:

$$\left(\overline{y}_1 - \overline{y}_2 - z_{\frac{1+L}{2}}\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \overline{y}_1 - \overline{y}_2 + z_{\frac{1+L}{2}}\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right).$$

This interval is fine if we know the population variance, but as you know, we often do not.

$\approx$

Comparing Two Population Means: Independent Populations

If we don't know the population variance, we do the usual: we estimate it from the sample. In our case, we have two samples, and hence two sample variances to use as estimates of σ^2: s_1^2 and s_2^2.

We combine these together by a process known as pooling, which is an average of s_1^2 and s_2^2 weighted by the degrees of freedom (i.e., number of data values minus 1) in each sample.

Comparing Two Population Means: Independent Populations

The result is the **pooled variance estimate**:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Using s_p instead of σ in the standardization formula gives

$$t^{(p)} = \frac{\overline{y}_1 - \overline{y}_2 - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

which has a $t_{n_1+n_2-2}$ distribution.

$\approx$

Comparing Two Population Means: Independent Populations

From this, we get a level L confidence interval for $\mu_1 - \mu_2$ having endpoints:

$$\overline{y}_1 - \overline{y}_2 \pm t_{n_1+n_2-2,\frac{1+L}{2}} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

This interval is called a **pooled variance interval**.

Example 3

A company buys cutting blades used in its manufacturing process from two suppliers. In order to decide if there is a difference in blade life, the lifetimes of 10 blades from manufacturer 1 and 13 blades from manufacturer 2 used in the same application are compared. A summary of the data shows the following (units are hours):

Manufacturer	n	$\overline{y}$	s
1	10	108.4	26.9
2	13	134.9	18.4

Management decides a 0.90 level of confidence is sufficient for their needs, and based on previous experience, they are willing to assume the two population variances are equal.

Example 3

1. **The Scientific Goal:** Decide if there is a difference in the life of blades from the two manufacturers.
2. **The Statistical Model:** Two independent normal populations with equal variances: $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$.
3. **The Model Parameter(s) to Be Estimated:** The difference in mean blade life, $\mu_1 - \mu_2$.

$\approx$

Example 3

4. **Point and Interval Estimates:**
   a. Point estimate: $\overline{y}_1 - \overline{y}_2 = 108.4 - 134.9 = -26.5$.
   b. Confidence Interval: The pooled variance estimate is

$$s_p^2 = \frac{(10-1)(26.9)^2 + (13-1)(18.4)^2}{10+13-2} = 503.6.$$

This gives the estimate of the standard error of $\overline{y}_1 - \overline{y}_2$ as

$$\sqrt{503.6\left(\frac{1}{10}+\frac{1}{13}\right)} = 9.44.$$

Finally, $n_1 + n_2 - 2 = 10 + 13 - 2 = 21$, and $t_{21,0.95} = 1.7207$, so a level 0.90 confidence interval for $\mu_1 - \mu_2$ is

$$(-26.5 - (9.44)(1.7207),\ -26.5 + (9.44)(1.7207))$$

$$= (-42.7, -10.3)$$

(SAS code found here).

Example 3

5. **Results and Interpretation:** Management can say that with 90% confidence they estimate that the mean lifetime of the blades from manufacturer 1 is between 10.3 and 42.7 hours less than that of the blades from manufacturer 2.

   An interesting point is that if this interval contained 0, they would not be able to conclude there was a difference in population means, which, since they are assuming normal distributions with equal variances, implies they would not be able to conclude the two populations had different distributions.

Comparing Two Population Means: Independent Populations

What do we do if the population variances are not equal?

The most fundamental question is: "If the population variances are not equal, does it make sense to compare the population means?"

In the case of unequal variances, even if the means are equal, the two populations will have different distributions. So comparing the means is inappropriate if the goal is to decide if the two population distributions are the same.

Comparing Two Population Means: Independent Populations

If our interest is solely in comparing the means, and we are not willing to assume the population variances are equal, here is how we can proceed.

Suppose that we take a random sample of size n_1 from population 1, which follows a $N(\mu_1, \sigma_1^2)$ distribution, and independently a random sample of size n_2 from population 2, which follows a $N(\mu_2, \sigma_2^2)$ distribution.

$\approx$

Comparing Two Population Means: Independent Populations

We already know that the estimator of μ_1 is the sample mean of the first sample, $\overline{y}_1$, and that of μ_2 is the sample mean of the second sample, $\overline{y}_2$.

We also know that the sampling distribution of $\overline{y}_1$ is $N(\mu_1, \sigma_1^2/n_1)$ and that of $\overline{y}_2$ is $N(\mu_2, \sigma_2^2/n_2)$.

Comparing Two Population Means: Independent Populations

As in the equal variance case, the best estimator of $\mu_1 - \mu_2$ is $\overline{y}_1 - \overline{y}_2$.

To construct a confidence interval for $\mu_1 - \mu_2$ we will use the information that the sampling distribution of $\overline{y}_1 - \overline{y}_2$ is $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

$\approx$

Comparing Two Population Means: Independent Populations

Then using the fact that the standardized estimator

$$Z = \frac{\overline{y}_1 - \overline{y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1),$$

and following the same logic as we did in developing the one sample confidence interval for the mean, we get a level L confidence interval for $\mu_1 - \mu_2$:

$$\left(\overline{y}_1 - \overline{y}_2 - z_{\frac{1+L}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \overline{y}_1 - \overline{y}_2 + z_{\frac{1+L}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right).$$

This interval is fine if we know the population variances, but as you know, we often do not.

Comparing Two Population Means: Independent Populations

As before, if we do not know the population variances, we estimate them from the data using the sample variances: s_1^2 and s_2^2. This gives us the standardized estimator

$$t^{(ap)} = \frac{\overline{y}_1 - \overline{y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Based on the equal variance case, $t^{(ap)}$ should have a t distribution.

It doesn't. In fact, its distribution isn't very nice at all.

$\approx$

Comparing Two Population Means: Independent Populations

However, all is not lost. Its distribution can be approximated by a t distribution with ν degrees of freedom, where ν is the largest integer less than or equal to

$$\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}.$$

Yuck! Fortunately, the calculation is not as bad as it looks.

$\approx$

Comparing Two Population Means: Independent Populations

Once we resign ourselves to calculating ν, an approximate level L confidence interval for $\mu_1 - \mu_2$ is

$$\left(\overline{y}_1 - \overline{y}_2 - t_{\nu, \frac{1+L}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \overline{y}_1 - \overline{y}_2 + t_{\nu, \frac{1+L}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) .$$

This interval is called a **separate variance or Satterhwaite interval**.

$\approx$

Example 3, Revisited

A company buys cutting blades used in its manufacturing process from two suppliers. In order to decide if there is a difference in blade life, the lifetimes of 10 blades from manufacturer 1 and 13 blades from manufacturer 2 used in the same application are compared. A summary of the data shows the following (units are hours):

Manufacturer	n	$\overline{y}$	s
1	10	108.4	26.9
2	13	134.9	18.4

Management is not willing to assume equal population variances. Even so, the only component of the statistical estimation problem this will change is the computation of the confidence interval.

$\approx$

Example 3, Revisited

The estimated standard error of $\overline{y}_1 - \overline{y}_2$ is

$$\sqrt{\frac{(26.9)^2}{10} + \frac{(18.4)^2}{13}} = 9.92.$$

The degrees of freedom ν is computed as the greatest integer less than or equal to

$$\frac{\left(\frac{(26.9)^2}{10} + \frac{(18.4)^2}{13}\right)^2}{\frac{\left(\frac{(26.9)^2}{10}\right)^2}{10-1} + \frac{\left(\frac{(18.4)^2}{13}\right)^2}{13-1}} = 15.17,$$

so $\nu = 15$. Finally, $t_{15,0.95} = 1.7530$, so a level 0.90 confidence interval for $\mu_1 - \mu_2$ is

$$(-26.5 - (9.92)(1.753),\ -26.5 + (9.92)(1.753))$$

$$= (-43.9, -9.1)$$

(SAS code found here).

$\approx$

Example 3, Revisited

So, we have

- Pooled variance interval for $\mu_1 - \mu_2$: $(-42.7, -10.3)$.
- Separate variance interval for $\mu_1 - \mu_2$: $(-43.9, -9.1)$.

Both give similar results in this case.

$\approx$

Comparing Two Population Proportions

Suppose there are two populations: population 1, in which a proportion p_1 have a certain characteristic, and population 2, in which a proportion p_2 have a certain (possibly different) characteristic. We will use a sample of size n_1 from population 1, and n_2 from population 2 to estimate the difference $p_1 - p_2$.

$\approx$

Comparing Two Population Proportions

Specifically, if y_1 is the number having the population 1 characteristic in the n_1 items in sample 1, and if if y_2 is the number having the population 2 characteristic in the n_2 items in sample 2, then the sample proportion having the population 1 characteristic is $\hat{p}_1 = y_1/n_1$, and the sample proportion having the population 2 characteristic is $\hat{p}_2 = y_2/n_2$.

A point estimator of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$.

$\approx$

Comparing Two Population Proportions

The standard error of $\hat{p}_1 - \hat{p}_2$ is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Further, for large n_1 and n_2, the Central Limit Theorem ensures that $\hat{p}_1 - \hat{p}_2$ has approximately a normal distribution, so

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

has approximately a $N(0,1)$ distribution.

Comparing Two Population Proportions

Based on this, and on the fact that if n_1 and n_2 are large, then $\hat{p}_1$ and $\hat{p}_2$ are close to p_1 and p_2, respectively, an approximate level L confidence interval for $p_1 - p_2$ has endpoints

$$\hat{p}_1 - \hat{p}_2 \pm z_{(1+L)/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

$\approx$

Comparing Two Population Proportions

As for the one sample case, this large-sample interval does not work well when one or both sample sizes are small.

However, by "fudging" the sample proportions in much the same way as we did in the one sample case, we can get an approximate interval that works well for all sample sizes.

$\approx$

Comparing Two Population Proportions

Specifically, to compute the level L approximate score (or Agresti-Coull) interval, first compute the adjusted estimates of n_1 and n_2:

$$\tilde{n}_1 = n_1 + 0.5z^2_{(1+L)/2}, \ \tilde{n}_2 = n_2 + 0.5z^2_{(1+L)/2},$$

and then the adjusted estimates of p_1 and p_2:

$$\tilde{p}_1 = \frac{y_1 + 0.25z^2_{(1+L)/2}}{\tilde{n}_1}, \ \tilde{p}_2 = \frac{y_2 + 0.25z^2_{(1+L)/2}}{\tilde{n}_2}$$

The approximate score interval for $p_1 - p_2$ is then given by the formula:

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{(1+L)/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{\tilde{n}_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{\tilde{n}_2}}$$

$\approx$

Example 4

In a recent survey on academic dishonesty 24 of the 200 female college students surveyed and 26 of the 100 male college students surveyed agreed or strongly agreed with the statement "Under some circumstances academic dishonesty is justified." With 95% confidence estimate the difference in the proportions p_f of all female and p_m of all male college students who agree or strongly agree with this statement.

$\approx$

Example 4

1. **The Scientific Goal:** Estimate the difference in the proportions p_f of all female and p_m of all male college students who agree or strongly agree with the statement.
2. **The Statistical Model:** Two independent binomials $b(200, p_f)$, $b(100, p_m)$.
3. **The Model Parameter(s) to Be Estimated:** $p_f - p_m$

$\approx$

Example 4

4. **Point and Interval Estimates:**
   a. Point estimate: $\hat{p}_f - \hat{p}_m = \frac{24}{200} - \frac{26}{100} = -0.14$.
   b. Confidence interval: Since $z_{0.975} = 1.96$, $y_f = 24$, $n_f = 200$, $y_m = 26$, and $n_m = 100$, the adjusted estimates of n_f and n_m are

$$\tilde{n}_1 = 200 + 0.5 \cdot 1.96^2 = 201.9208, \ \tilde{n}_2 = 100 + 0.5 \cdot 1.96^2 = 101.9208$$

The adjusted estimates of p_f and p_m are then

$$\tilde{p}_f = \frac{24 + 0.25 \cdot 1.96^2}{\tilde{n}_1} = 0.1236,$$

and

$$\tilde{p}_m = \frac{26 + 0.25 \cdot 1.96^2}{\tilde{n}_2} = 0.2645.$$

$\approx$

Example 4

The approximate score interval for $p_f - p_m$ is then

$$0.1236 - 0.2645 \pm$$

$$1.96\sqrt{\frac{0.1236(1-0.1236)}{201.9208} + \frac{0.2645(1-0.2645)}{101.9208}}$$

$$= (-0.2378, -0.0440)$$

(SAS code here)

$\approx$

Example 4

5. **Results and Interpretation:** With 95% confidence we estimate that the percentage of male college students who agree or strongly agree with the statement is between 4.4 and 23.78 percent greater than the corresponding percentage of female college students.

$\approx$

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation
- Sampling Distribution
  - Normal
  - t
  - Binomial
- Interval Estimation
- The Components of a Statistical Estimation Problem
  - The Scientific Goal
  - The Statistical Model
  - The Model Parameter(s) to Be Estimated
  - Point and Interval Estimates
  - Results and Interpretation

Recap: Estimation: Our First Look at Statistical Inference

- Specific Estimation Problems:
  - 1-Sample Mean, Known Variance
  - 1 Sample Mean, Unknown Variance
  - 1-Sample Proportion, Large Sample
  - 1-Sample Proportion, All Sample (Approx. Score Interval)
  - 2-Sample Mean, Paired Observations
  - 2-Sample Mean, Known Variance
  - 2 Sample Mean, Unknown Variance
  - 2-Sample Proportion, Large Sample
  - 2-Sample Proportion, All Sample (Approx. Score Interval)