Wrap up from Section 1: Decreasing our margin of error.

How many monkeys?

Determining sample size from margin of error

Researchers would like to estimate the mean cholesterol level of a particular variety of monkey that is often used in laboratory experiments. They would like their estimate to be within 1 milligram per deciliter of blood of the true value of the mean at a 95% confidence level. A previous study involving this variety of monkey suggests that the standard deviation of the cholesterol level is about 5 mg/dl. Obtaining the monkeys is time-consuming and expensive, so the researchers want to know the minimum number of monkeys they will need to generate a satisfactory estimate.

Sample Size for a Desired Margin of Error

To determine the sample size $n$ that will yield a confidence interval for a population mean with a specified margin of error $m$, set the expression for the margin of error to be less than or equal to $m$ and solve for $n$:

$$z^* \frac{\sigma}{\sqrt{n}} \leq m$$

Pg. 642 shows how to get a confidence interval on your calculator.
Cautions from section 1:

Notice that it is the size of the SAMPLE that determines the margin of error. The size of the population does not influence the sample size that we need.

The data must be an SRS from the population.

Different methods for confidence intervals are needed for other sampling methods.

There is no correct method for inference from data haphazardly collected with bias of unknown size. Fancy formulas cannot rescue badly produced data.

Outliers can distort results.

The shape of the population distribution matters.

You must know the standard deviation of the population in order to use the method we have learned in section 1.

So, what do we do if we do not have the population standard deviation?

In practicality, we never know the population standard deviation.

When we do inference in practice, verifying the conditions is often a bit more complicated.

Still we need to verify three important conditions before we estimate a population mean.

**Conditions for Inference about a Population Mean**

- **SRS** Our data are a simple random sample (SRS) of size \( n \) from the population of interest or come from a randomized experiment. This condition is very important.

- **Normality** Observations from the population have a Normal distribution with mean \( \mu \) and standard deviation \( \sigma \). In practice, it is enough that the distribution be symmetric and single-peaked unless the sample is very small. Both \( \mu \) and \( \sigma \) are unknown parameters.

- **Independence** The method for calculating a confidence interval assumes that individual observations are independent. To keep the calculations reasonably accurate when we sample without replacement from a finite population, we should verify that the population size is at least 10 times the sample size (\( N \geq 10n \)).

In this setting, the sample mean has the Normal distribution with mean \( \mu \) and standard deviation \( \sigma \). Because we don’t know \( \sigma \), we estimate it by the sample standard deviation of \( s \). We then estimate the standard deviation of our sample mean by the following formula:

**Standard Error**

When the standard deviation of a statistic is estimated from the data, the result is called the standard error of the statistic. The standard error of the sample mean is \( s/\sqrt{n} \).

When we know the value of our population standard deviation, we base our confidence interval on the sampling distribution of our sample mean, which has a Normal distribution of the Normality condition is satisfied.

When we don’t know the population standard deviation, I can no longer use the normal curve. I must use a \( t \)-distribution.
t-distributions:

Unlike the standard normal distribution, there is a different t-distribution for each sample size n. We specify a particular t-distribution by giving its degrees of freedom, (df).

When we perform inference about the mean using a t-distribution, our degrees of freedom is

\[ df = n - 1 \]

This is because we are using our sample standard deviation in our calculations.

The standard normal curve we will refer to as a Z-distribution.

The t-distribution we call a t-distribution with k degrees of freedom, or \( t(k) \) for short.

The t-distribution looks approximately normal.

There is more spread in a t-distribution. They also have more area in the tails and less in the center than does the Normal Curve. This happens because we substitute the estimate \( s \) for the population standard deviation which gives more variation.

The are symmetric about zero and single peaked, and bell-shaped.

The larger our degrees of freedom, the closer our t-distribution looks to the normal curve. This is because it is based on sample size, and the larger the sample size the better our approximation of what really happens.

When using t-distributions, we will be using Table C in the back of the book.
Finding \( t \)

Using table C

Suppose you want to construct a 95% confidence interval for the mean of the population based on an SRS of size \( n = 12 \). What critical value \( t \) should we use?

Our degrees of freedom our \( n - 1 \).

Or \( 12 - 1 = 11 \).

In table C, we consult the row corresponding to 11 degrees of freedom. We move across the row to the entry that is directly above the 95% confidence level at the bottom of the chart. The desired critical value is \( t = 2.201 \).

The corresponding critical value from the \( z \)-distribution \( z = 1.96 \).

### Upper-tail probability \( p \)

<table>
<thead>
<tr>
<th>df</th>
<th>.05</th>
<th>.025</th>
<th>.02</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.812</td>
<td>2.228</td>
<td>2.359</td>
<td>2.764</td>
</tr>
<tr>
<td>11</td>
<td>1.796</td>
<td>2.201</td>
<td>2.328</td>
<td>2.718</td>
</tr>
<tr>
<td>12</td>
<td>1.782</td>
<td>2.179</td>
<td>2.303</td>
<td>2.681</td>
</tr>
</tbody>
</table>

\( z^* \)

<table>
<thead>
<tr>
<th>90%</th>
<th>95%</th>
<th>96%</th>
<th>98%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.645</td>
<td>1.960</td>
<td>2.054</td>
<td>2.326</td>
</tr>
</tbody>
</table>

Confidence level \( C \)

Auto Pollution

Constructing a one-sample \( t \)-interval

Environmentalists, government officials, and vehicle manufacturers are all interested in studying the auto exhaust emissions produced by motor vehicles. The major pollutants in auto exhaust from the gasoline engines are hydrocarbons, monoxide, and nitrogen oxides (NOX). The data below are given in NOX levels (in grams per mile) for a random sample of light-duty engines of the same type.

| 1.28 | 1.17 | 1.16 | 1.08 | 0.60 | 1.32 | 1.24 | 0.71 | 0.49 | 1.38 | 1.20 | 0.78 |
| 1.06 | 2.20 | 1.75 | 1.63 | 1.26 | 1.73 | 1.31 | 1.96 | 0.97 | 1.12 | 0.72 |
| 1.31 | 1.45 | 1.22 | 1.32 | 1.42 | 1.44 | 0.51 | 1.49 | 1.33 | 0.88 | 0.57 |
| 2.27 | 1.87 | 2.34 | 1.16 | 1.45 | 1.51 | 1.47 | 1.06 | 2.01 | 1.39 |

Construct and interpret a 95% confidence interval for the mean amount of NOX emitted by light-duty engines of this type.

What are the steps for solving an inference problem?
Step 1: Parameter

Step 2: Conditions
- SRS
- Normality
- Independence

Step 3: Calculations

Step 4: Interpretation

Normality:

Computation:

<table>
<thead>
<tr>
<th>Upper-tail probability $p$</th>
<th>df</th>
<th>.05</th>
<th>.025</th>
<th>.02</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>1.697</td>
<td>2.042</td>
<td>2.147</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>1.684</td>
<td>2.021</td>
<td>2.123</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>1.676</td>
<td>2.009</td>
<td>2.109</td>
<td></td>
</tr>
<tr>
<td>90% 95% 99%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Interpretation: (1.185, 1.473)

We are 95% confident that the true mean level of nitrogen oxides emitted by this type of light-duty engine is between 1.185 and 1.473 grams/mile.

Estimate + or - t(SE)estimate

SE stands for Standard Error

Homework: 10.27-10.32 pg. 648