Statistics 101

Mine Çetinkaya-Rundel

February 16, 2012
Announcements

- **Due:** Online quiz over the weekend
- **OH next week:**
  - Monday 2pm - 4pm
  - Tuesday 2:30pm - 5:00pm
  - Wednesday 2pm - 4pm
- **Optional review session:** Wednesday Tuesday 5:30pm - 7pm. Room TBA.
- **HW 4** will be posted right after class today, is due **by noon next Wednesday** Feb 22.
Review question

Which of the following is false?

(a) A parameter is a measure from a population, a point estimate is a measure from a sample.
(b) A parameter is a mean, a point estimate is a proportion.
(c) Parameters are rarely known, point estimates can be calculated from sample data.
(d) Point estimates are used to estimate parameters.
1 Confidence intervals
- Assumptions & conditions for inference
- A more accurate confidence interval
- Changing the confidence level
- Interpreting confidence intervals
- Hypothesis testing using confidence intervals

2 Hypothesis testing
This semester we asked Duke stat students how many hours of sleep they get per night. A sample of 217 respondents yielded an average of 6.7 hours of sleep with a standard deviation of 2.03 hours. Assuming that this sample is random and representative of all Duke students (*might be leap of faith!*), construct a 95% confidence interval for the average amount of sleep Duke students get per night.

CLT states that sample means will be nearly normally distributed, and the standard error of the sampling distribution can be estimated by \( \frac{s}{\sqrt{n}} \). However there are certain assumptions and conditions that must be verified in order for the CLT to apply.
Assumptions & conditions for inference

1. **Independence Assumption:**
   - **Random sampling condition:** We are assuming that this sample is random.
   - **10% Condition:** \( 217 < 10\% \) of all Duke students.

   We can assume that how much sleep one student in this sample gets is independent of another.

2. **Nearly Normal Condition:** The sample data has a symmetric distribution, so we can assume that it comes from a nearly normal population. In addition, \( n > 50 \), so we can assume that the sampling distribution will be approximately normal as well.
An approximate interval for sleep

An approximate confidence interval for the average amount of sleep Duke students get can be calculated as

$$6.7 \pm \left(2 \times \frac{2.03}{\sqrt{217}}\right) = 6.7 \pm (2 \times 0.14) = (6.42, 6.98)$$

But we can actually obtain a confidence interval that’s a little more accurate.
Clicker question

Which of the below Z scores mark the cutoff for the middle 95% of a normal distribution?

(a) $Z = -3.49$ and $Z = 1.65$
(b) $Z = -2.58$ and $Z = 2.58$
(c) $Z = -1.96$ and $Z = 1.96$
(d) $Z = -1.65$ and $Z = 1.65$
A more accurate 95% confidence interval

Calculate an exact 95% confidence interval for the average sleep Duke students get per night.

\[
\bar{x} = 6.7, \; s = 2.03, \; n = 217
\]

\[
\bar{x} \pm z^* \times \frac{s}{\sqrt{n}} = 6.7 \pm \left(1.96 \times \frac{2.03}{\sqrt{217}}\right)
\]

\[
= 6.7 \pm 0.27
\]

\[
= (6.43, 6.97)
\]

Note: We used the approximate confidence interval to introduce this concept and to illustrate how it relates to the 68-95-99.7% rule. When asked for a confidence interval you should calculate it using this more accurate approach.
In order to change the confidence level all we need to do is adjust $z^*$ in the above formula.

Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.

However, using the $z$ table it is possible to find the appropriate $z^*$ for any confidence level.
A 98% confidence interval

Calculate a 98% confidence interval for the average sleep Duke students get per night.

\[ \bar{x} = 6.7, \ s = 2.03, \ n = 217 \]

\[ = \bar{x} \pm z^* \times \]

\[ = 6.7 \pm \left( 2.33 \times \frac{2.03}{\sqrt{217}} \right) \]

\[ = (6.7 - 0.32, 6.7 + 0.32) \]

\[ = (6.38, 7.02) \]
Clicker question

Which of the following is correct?

(a) 98% of Duke students sleep between 6.38 and 7.02 hours per night, on average.

(b) We are 98% confident that Duke students on average sleep 6.38 to 7.02 hours per night.

(c) 98% of the time Duke students sleep 6.38 hours to 7.02 hours per night.

(d) We are 98% confident that the average sleep the 217 students in this sample get is between 6.38 and 7.02 hours per night.

(e) The standard error is 0.32 hours.
The 95% confidence interval for the average hours of sleep Duke students get was (6.43, 6.97). Does this provide convincing evidence that Duke students do not get the 8 hours of sleep recommended by the CDC?

(a) Yes
(b) No
Testing claims based on a confidence interval (cont.)

- Using a confidence interval for hypothesis testing might be insufficient in some cases since it gives a yes/no (reject/don’t reject) answer, as opposed to quantifying our decision with a probability.
- Formal hypothesis testing allows us to report a probability along with our decision.
1. Confidence intervals

2. Hypothesis testing
   - Hypothesis testing framework
   - Assumptions & conditions for inference
   - Formal testing using p-values
   - Two-sided hypothesis testing with p-values
Remember when...

MythBusters yawning experiment:

<table>
<thead>
<tr>
<th></th>
<th>Seeded</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yawn</td>
<td>10</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>Not Yawn</td>
<td>24</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>16</td>
<td>50</td>
</tr>
</tbody>
</table>

\[
\hat{p}_{seeded} = \frac{10}{34} = 0.29
\]

\[
\hat{p}_{control} = \frac{4}{16} = 0.25
\]
Remember when...

MythBusters yawning experiment:

<table>
<thead>
<tr>
<th></th>
<th>Seeded</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yawn</td>
<td>10</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>Not Yawn</td>
<td>24</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>16</td>
<td>50</td>
</tr>
</tbody>
</table>

\[ \hat{p}_{\text{seeded}} = \frac{10}{34} = 0.29 \]

\[ \hat{p}_{\text{control}} = \frac{4}{16} = 0.25 \]

Possible explanations:

- Yawning is *independent* of seeing someone else yawn; therefore, the difference between the proportions of yawners in the control and seeded groups is *due to chance*.  → *nothing is going on*

- Yawning is *dependent* on seeing someone else yawn; therefore, the difference between the proportions of yawners in the control and seeded groups is *real*.  → *something is going on*
Hypothesis testing framework

With a slightly different terminology

- We started with the assumption that yawning is independent of seeing someone else yawn. → null hypothesis
- We then investigated how the results would look if we simulated the experiment many times assuming the null hypothesis is true. → testing

Since the simulation results were similar to the actual data (on average roughly 10 people yawning in the seeded group), we decided not to reject the null hypothesis in favor of the alternative hypothesis.
A trial as a hypothesis test

- Hypothesis testing is very much like a court trial.
  - In a trial, the burden of proof is on the prosecution.
  - In a hypothesis test, the burden of proof is on the unusual claim.

- $H_0$: Defendant is innocent
- $H_A$: Defendant is guilty

- Collect data - The null hypothesis is the ordinary state of affairs (the status quo), so it’s the alternative hypothesis that we consider unusual (and for which we must gather evidence).

- Then we judge the evidence - “Could these data plausibly have happened by chance if the null hypothesis were true?”
  - If they were very unlikely to have occurred, then the evidence raises more than a reasonable doubt in our minds about the null hypothesis.

- Ultimately we must make a decision. How unlikely is unlikely?
A trial as a hypothesis test (cont.)

- If the evidence is not strong enough to reject the assumption of innocence, the jury returns with a verdict of “not guilty”.
  - The jury does not say that the defendant is innocent, just that there is not enough evidence to convict.
  - The defendant may, in fact, be innocent, but the jury has no way of being sure.

- Said statistically, we fail to reject the null hypothesis.
  - We never declare the null hypothesis to be true, because we simply do not know whether it’s true or not.
  - Therefore we never “accept the null hypothesis”.

Statistics 101 (Mine Çetinkaya-Rundel)
Grade inflation?

In 2001 the average GPA of students at Duke University was 3.37. This semester we asked Duke stats students their GPA. A sample of 203 respondents yielded an average GPA of 3.59 with a standard deviation of 0.28. Assuming that this sample is random and representative of all Duke students *(another leap of faith!)*, do these data provide convincing evidence that the average GPA of Duke students has changed over the last decade?
Setting the hypotheses

- The parameter of interest is the average GPA of current Duke students.
- There may be two explanations why our sample mean is higher than the average GPA from 2001.
  - The true population mean has changed.
  - The true population mean remained at 3.37, the difference between the true population mean and the sample mean is simply due to natural sampling variability.
- We start with the assumption that nothing has changed.

\[ H_0 : \mu = 3.37 \]

- We test the claim that average GPA has changed.

\[ H_A : \mu \neq 3.37 \]
Before doing inference using this data set, we must make sure that the assumptions & conditions necessary for inference are satisfied:

1. **Independence Assumption:**
   - *Random sampling condition:* Assuming this sample is random.
   - *10% Condition:* \(203 < 10\%\) of all current Duke students.
   
   We can assume that GPA of one student in this sample is independent of another.

2. **Nearly Normal Condition:** The distribution appears to be slightly skewed (but not extremely) and \(n > 50\) so we can assume that the distribution of the sample means is nearly normal.
The same survey asked how many colleges students applied to, and 206 students responded to this question. This sample yielded an average of 9.7 college applications with a standard deviation of 7. College Board website states that counselors recommend students apply to roughly 8 colleges. Which of the following are the correct set of hypotheses to test if these data provide convincing evidence that the average number of colleges all Duke students apply to is higher than recommended.

(a) $H_0 : \mu = 9.7$
   $H_A : \mu > 9.7$

(b) $H_0 : \mu = 8$
   $H_A : \mu > 8$

(c) $H_0 : \bar{x} = 8$
   $H_A : \bar{x} > 8$

(d) $H_0 : \mu = 8$
   $H_A : \mu > 9.7$
Number of college applications - assumptions & conditions

1. **Independence Assumption:**
   - *Random sampling condition:* Assuming this sample is random.
   - *10% Condition:* $206 < 10\%$ of all current Duke students.

   We can assume that how many colleges of one student in this sample is applied to independent of another.

2. **Nearly Normal Condition:** We are not provided a plot of the distribution of the data, however as long as the data aren’t extremely skewed we can assume that the sampling distribution of the means will be nearly normal since $n > 50$. 
Hypothesis testing

Formal testing using p-values

Number of college applications - test statistic

The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually (significantly) high?

\[
\bar{x} \sim N\left(\mu = 8, \frac{SE}{\sqrt{206}} = 0.5\right)
\]

\[
Z = \frac{9.7 - 8}{0.5} = 3.4
\]
p-values

- The *p-value* is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.

- If the p-value is *low* (lower than the significance level, $\alpha$, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* $H_0$.

- If the p-value is *high* (higher than $\alpha$) we say that it is likely to observe the data even if the null hypothesis were true, and hence *do not reject* $H_0$.

- We never accept $H_0$ since we’re not in the business of trying to prove it. We simply want to know if the data provide convincing evidence to support $H_A$. 
**p-value:** probability of observing data at least as favorable to $H_A$ as our current data set (a sample mean greater than 9.7), if in fact $H_0$ was true (the true population mean was 8).

\[
P(\bar{x} > 9.7 \mid \mu = 8) = P(Z > 3.4) = 0.0003
\]
Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is low (lower than 5%) we reject $H_0$.
- The data provide convincing evidence that Duke students average apply to more than 8 schools.
- The difference between the null value of 8 schools and observed sample mean of 9.7 schools is not due to chance or sampling variability.
A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 217 Duke students yielded an average of 6.7 hours, with a standard deviation of 2.03 hours. Assuming that this is a random sample representative of all college students (bit of a leap of faith?), do these data provide convincing evidence that Duke students on average sleep less than 7 hours per night?
Hypothesis testing

Clicker question

Which of the following conditions do not need to be satisfied in order to answer this question using statistical inference techniques that rely on the central limit theorem?

(a) The sample should be random.
(b) How much one student in the sample sleeps should be independent of another.
(c) The distribution of sleep should not be extremely skewed.
(d) There should be at least 10 expected successes and 10 expected failures.
Setting the hypotheses

\[ H_0 : \mu = 7 \] (Duke students sleep 7 hours per night on average)
\[ H_A : \mu < 7 \] (Duke students sleep less than 7 hours per night on average)

- If in fact the null hypothesis is true, \( \bar{x} \) is distributed nearly normally with mean \( \mu = 7 \) and standard error
  \[ SE = \frac{s}{\sqrt{n}} = \frac{2.03}{\sqrt{217}} = 0.14. \]
- We would like to find out how likely it is to observe a sample mean at least as far from the data as our current sample mean (6.7), if in fact the null hypothesis is true.
Calculating the p-value

\[ \bar{x} \sim N \left( \mu = 7, SE = \frac{2.03}{\sqrt{217}} = 0.14 \right) \]

\[ Z = \frac{6.7 - 7}{0.14} = -2.14 \]

\[ p - \text{value} = P(\bar{x} < 6.7 \mid \mu = 7) = P(Z < -2.14) = 0.0162 \]
Clicker question

Based on a p-value of 0.0162, which of the following is true?

- $H_0 : \mu = 7$ (Duke students sleep 7 hours per night on average)
- $H_A : \mu < 7$ (Duke students sleep less than 7 hours per night on average)

(a) Fail to reject $H_0$, the data provide convincing evidence that Duke students sleep less than 7 hours on average.
(b) Reject $H_0$, the data provide convincing evidence that Duke students sleep less than 7 hours on average.
(c) Reject $H_0$, the data prove that Duke students sleep more than 7 hours on average.
(d) Fail to reject $H_0$, the data do not provide convincing evidence that Duke students sleep less than 7 hours on average.
(e) Reject $H_0$, the data provide convincing evidence that Duke students in this sample sleep less than 7 hours on average.
If the research question was “Do the data provide convincing evidence that the average amount of sleep Duke students get per night is different than the national average?”, the alternative hypothesis would be different.

\[ H_0 : \mu = 7 \]
\[ H_A : \mu \neq 7 \]

Hence the p-value would change as well:

\[ p-value = 0.0162 \times 2 \]
\[ = 0.0324 \]
the next two slides are provided as a brief summary of hypothesis testing...
Recap: Hypothesis testing framework

1. Set the hypotheses.
2. Check assumptions and conditions.
3. Calculate a *test statistic* and a p-value.
4. Make a decision, and interpret it in context of the research question.
Recap: Hypothesis testing for a population mean

1. Set the hypotheses
   - $H_0 : \mu = \text{null value}$
   - $H_A : \mu < \text{or} > \text{or} \neq \text{null value}$

2. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Normality: nearly normal population or $n \geq 50$, no extreme skew

3. Calculate a test statistic and a p-value (draw a picture!)
   - $Z = \frac{\bar{x} - \mu}{SE}$, where $SE = \frac{s}{\sqrt{n}}$

4. Make a decision, and interpret it in context of the research question
   - If p-value $< \alpha$, reject $H_0$, data provide evidence for $H_A$
   - If p-value $> \alpha$, do not reject $H_0$, data do not provide evidence for $H_A$