Housekeeping

Announcements

- PS5 due Tuesday
- You might hear from me about team meetings this weekend and next week
- Project proposal feedback

Feedback from midterm evaluation

Varied opinions on flipped class / teams, consensus on clicker questions and real world examples

45/62 think the pace is about right, 60/62 think stats is worth learning

- Length of midterm - I hear ya!
- Problem set submission - start with typing, don’t transcribe after
- PA due on day after class - feel free to complete day of class
- Time commitment - expected from any class: 10 hrs/week
- Peer evaluation - be honest and fair
- OH - https://stat.duke.edu/courses/Fall14/sta101.001/info/
- Labs / TAs
- RA grading weight - If your Final score is higher than your overall RA score, that score will be replaced with the average of Final and RA score

Main ideas

Comparing means of two groups

When comparing the means of two groups, we must first think about whether the data are independent or dependent across the groups:

- dependent (paired) groups (e.g. pre/post weights of subjects in a weight loss study, twin studies, etc.)
- independent groups (e.g. grades of students across two sections)
(2) Using the T distribution

When the sample size is low we use the T distribution for inference on means.

Example 1: Zinc in water

Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water at 10 randomly sampled locations.

<table>
<thead>
<tr>
<th>Location</th>
<th>Bottom</th>
<th>Surface</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.43</td>
<td>0.415</td>
<td>0.015</td>
</tr>
<tr>
<td>2</td>
<td>0.266</td>
<td>0.238</td>
<td>0.028</td>
</tr>
<tr>
<td>3</td>
<td>0.567</td>
<td>0.39</td>
<td>0.177</td>
</tr>
<tr>
<td>4</td>
<td>0.531</td>
<td>0.41</td>
<td>0.121</td>
</tr>
<tr>
<td>5</td>
<td>0.707</td>
<td>0.605</td>
<td>0.102</td>
</tr>
<tr>
<td>6</td>
<td>0.716</td>
<td>0.609</td>
<td>0.107</td>
</tr>
<tr>
<td>7</td>
<td>0.651</td>
<td>0.632</td>
<td>0.019</td>
</tr>
<tr>
<td>8</td>
<td>0.589</td>
<td>0.523</td>
<td>0.066</td>
</tr>
<tr>
<td>9</td>
<td>0.469</td>
<td>0.411</td>
<td>0.058</td>
</tr>
<tr>
<td>10</td>
<td>0.723</td>
<td>0.612</td>
<td>0.111</td>
</tr>
</tbody>
</table>

Water samples collected at the same location, on the surface and in the bottom, cannot be assumed to be independent of each other, hence we need to use a *paired* analysis.

Source: https://onlinecourses.science.psu.edu/stat500/node/51

Comparing means of two groups

Analyzing paired data

Suppose we want to compare the average zinc concentration levels in the bottom and surface:

- When two sets of observations have this special correspondence (not independent), they are said to be *paired*.
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations.
- It is important that we always subtract using a consistent order.

Parameter and point estimate for paired data

For comparing average zinc concentration levels in the bottom and surface when the data are paired:

- **Parameter of interest**: Average difference between the bottom and surface zinc measurements of *all* drinking water.

\[ \mu_{\text{diff}} \]

- **Point estimate**: Average difference between the bottom and surface zinc measurements of drinking water from the *sampled* locations.

\[ \bar{x}_{\text{diff}} \]
Example 2: Gender gap in salaries

Since 2005, the American Community Survey polls ∼3.5 million households yearly. The following summarizes distribution of salaries of males and females from a random sample of individuals who responded to the 2012 ACS:

<table>
<thead>
<tr>
<th></th>
<th>$\bar{x}$</th>
<th>$s$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>55,890</td>
<td>68,767.88</td>
<td>470</td>
</tr>
<tr>
<td>female</td>
<td>29,240</td>
<td>32,025.98</td>
<td>373</td>
</tr>
</tbody>
</table>

ACS: Surge of media attention in spring 2012 when the House of Representatives voted to eliminate the survey. Daniel Webster, Republican congressman from Florida: “in the end this is not a scientific survey. It’s a random survey.”

Parameter and point estimate

- **Parameter of interest**: Average difference between the salaries of all males and females in the US.
  $$\mu_{\text{male}} - \mu_{\text{female}}$$

- **Point estimate**: Average difference between the salaries of sampled males and females in the US.
  $$\bar{x}_{\text{male}} - \bar{x}_{\text{female}}$$
### Clicker question

**Ex 1:** Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water at 10 randomly sampled locations. What are the hypotheses for evaluating whether the true average concentration in the bottom water exceeds that of surface water? (Note: diff = bottom - surface)

<table>
<thead>
<tr>
<th>Location</th>
<th>bottom</th>
<th>surface</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.43</td>
<td>0.415</td>
</tr>
<tr>
<td>2</td>
<td>0.266</td>
<td>0.238</td>
</tr>
<tr>
<td>3</td>
<td>0.567</td>
<td>0.39</td>
</tr>
<tr>
<td>4</td>
<td>0.531</td>
<td>0.41</td>
</tr>
<tr>
<td>5</td>
<td>0.707</td>
<td>0.605</td>
</tr>
<tr>
<td>6</td>
<td>0.716</td>
<td>0.609</td>
</tr>
<tr>
<td>7</td>
<td>0.651</td>
<td>0.632</td>
</tr>
<tr>
<td>8</td>
<td>0.589</td>
<td>0.523</td>
</tr>
<tr>
<td>9</td>
<td>0.469</td>
<td>0.411</td>
</tr>
<tr>
<td>10</td>
<td>0.723</td>
<td>0.612</td>
</tr>
</tbody>
</table>

- (a) $H_0: \mu_{\text{bottom}} = \mu_{\text{surface}}$
- (b) $H_0: \mu_{\text{bottom}} = \mu_{\text{surface}}$
- (c) $H_0: \mu_{\text{diff}} = 0$
- (d) $H_0: \mu_{\text{diff}} = 0$

- $H_A: \mu_{\text{bottom}} \neq \mu_{\text{surface}}$
- $H_A: \mu_{\text{bottom}} > \mu_{\text{surface}}$
- $H_A: \mu_{\text{diff}} 
eq 0$
- $H_A: \mu_{\text{diff}} > 0$

### Conditions for paired analysis

- **Independence:**
  - **Within groups:** both samples should be random and less than 10% of their respective populations
  - If these hold, we can assume that each observation within a sample is independent of another in that sample.
  - **Between groups:** Between groups, the observations should be dependent on each other, i.e. paired.

- **Sample size / skew:**
  - Each distribution should not be extremely skewed. If either sample size is small (less than 30 observations), use a T distribution. If both samples are large (more than 30 observations), can use a T or Z distribution.

### Clicker question

**Ex 2:** Using data from the 2012 ACS, what are the hypotheses for evaluating whether salaries of men and women are different, on average. (Note: diff = male - female)

<table>
<thead>
<tr>
<th></th>
<th>$\bar{x}$</th>
<th>$s$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>55,890</td>
<td>68,767.88</td>
<td>470</td>
</tr>
<tr>
<td>female</td>
<td>29,240</td>
<td>32,025.98</td>
<td>373</td>
</tr>
</tbody>
</table>

- (a) $H_0: \mu_{\text{male}} = \mu_{\text{female}}$
- (b) $H_0: \mu_{\text{male}} = \mu_{\text{female}}$
- (c) $H_0: \mu_{\text{diff}} = 0$
- (d) $H_0: \mu_{\text{diff}} = 0$

- $H_A: \mu_{\text{male}} \neq \mu_{\text{female}}$
- $H_A: \mu_{\text{male}} > \mu_{\text{female}}$
- $H_A: \mu_{\text{diff}} 
eq 0$
- $H_A: \mu_{\text{diff}} > 0$

### Conditions for comparing two means

- **Independence:**
  - Within groups: both samples should be random
  - Between groups: both samples should be less than 10% of their respective populations
  - If these hold, we can assume that each observation within a sample is independent of another in that sample.

- **Sample size / skew:**
  - Each distribution should not be extremely skewed. If either sample size is small (less than 30 observations), use a T distribution. If both samples are large (more than 30 observations), can use a T or Z distribution.
Comparing means of two groups

Conditions for comparing two means

1. Independence:
   - Within groups:
     - both samples should be random
     - both samples should be less than 10% of their respective populations
   
   If these hold, we can assume that each observation within a sample is independent of another in that sample.
   
   - Between groups: Between groups, the observations should be independent on each other, i.e. not paired.

2. Sample size / skew:
   
   Each distribution should not be extremely skewed. If either sample size is small (less than 30 observations), use a T distribution. If both samples are large (more than 30 observations), can use a T or Z distribution.

Mechanics for comparing two means analysis

- If groups are paired, first create a new variable, the differences. Then, proceed as if doing inference for a single variable:

  \[
  \text{test statistic} = \frac{\bar{x}_{\text{diff}} - \mu_{\text{diff}}}{\text{SE}}
  \]

  where \( SE = \frac{s_{\text{diff}}}{\sqrt{n_{\text{diff}}}} \).

- If groups are independent, use a new formula for SE:

  \[
  \text{test statistic} = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{SE}
  \]

  where \( SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \).

  In either situation, if the sample size is small, make sure to use the T distribution.

Review: what purpose does a large sample serve?

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that...

- the sampling distribution of the mean is nearly normal
- the estimate of the standard error, as \( \frac{s}{\sqrt{n}} \), is reliable

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds true for any sample size as long as the population distribution is nearly normal.
- While this is a helpful special case, it's inherently difficult to verify normality in small data sets.
- We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from.

  - For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?
The T distribution

- When $n$ is small, and the population standard deviation ($\sigma$) is unknown (almost always), the uncertainty of the standard error estimate is addressed by using the T distribution.
- This distribution also has a bell shape, but its tails are thicker than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- These extra thick tails are helpful for mitigating the effect of a less reliable estimate for the standard error of the sampling distribution (since $n$ is small).

Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: degrees of freedom ($df$).

What happens to shape of the T distribution as $df$ increases?

Degrees of freedom for inference

- One sample: $df = n - 1$ (applies to paired analysis as well)
- Two samples: $df = \min(n_1 - 1, n_2 - 1)$

Back to comparing two means

Calculating test statistic

Ex 1 - Zinc in water: Calculate the test statistic.

\[ \bar{x}_{diff} = 0.0804, s_{diff} = 0.0523, n_{diff} = 10 \]

\[ T = \frac{0.0804 - 0}{0.0523} = 4.86 \]
Finding p-values using the T distribution

Ex 1 - Zinc in water: Find the p-value for this hypothesis test.

\[ H_0 : \mu_{\text{diff}} = 0 \quad H_A : \mu_{\text{diff}} > 0 \quad n = 10 \quad T = 4.86 \]

\[ df = 10 - 1 = 9 \]

- Using R:

```r
> pt(4.86, df = 9, lower.tail = FALSE)
[1] 0.00044
```

Constructing a confidence interval

Ex 1 - Zinc in water: Construct a 95% confidence interval for the average difference between the zinc concentration in bottom and surface. Hint: Remember that the confidence interval always has the form: point estimate ± margin of error, and the margin of error is always a critical value multiplied by the standard error.

\[ x_{\text{diff}} = 0.0804, s_{\text{diff}} = 0.0523, n_{\text{diff}} = 10 \]

Recap

Critical value / test statistic

- If the sample size is large enough (both \( n > 30 \)) can use the Z distribution
- If the sample size is not large enough (either \( n < 30 \)), use the T distribution with degrees of freedom \( df = \min(n_1 - 1, n_2 - 1) \)
- If the sample size is large enough, it actually doesn’t matter whether you use Z or T (since the T distribution converges to the Z distribution as \( n \) increases)