Inference for Population Means

Inference about population mean \( \mu \)

- We used sampling distribution results (Chapter 5) to create two different tools for inference

- **Confidence Intervals**: Use sample mean as the center of an interval of likely values for pop. mean \( \mu \)
  - Width of interval is a multiple of standard deviation (or standard error) of sample mean

- **Hypothesis Tests**: Compare sample mean to a hypothesized population mean \( \mu_0 \)
  - Test statistic is also a multiple of standard deviation of the sample mean

**Known vs. Unknown Variance**

- **Known population SD \( \sigma \)**
  - Sample mean is centered at \( \mu \) and has standard deviation:
    \[
    \text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}
    \]
  - Sample mean has Normal distribution
    \[
    Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ follows } N(0,1)
    \]

- **Unknown population SD \( \sigma \)**
  - Sample mean is centered at \( \mu \) and has standard error:
    \[
    \text{SE}(\bar{X}) = \frac{s}{\sqrt{n}}
    \]
  - Sample mean has \( t \) distribution with \( n-1 \) degrees of freedom
    \[
    T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \text{ follows } t_{n-1}
    \]

**t distribution**

- \( t \) distribution looks like a normal distribution, but has "thicker" tails. The tail thickness is controlled by the degrees of freedom

- The smaller the degrees of freedom, the thicker the tails of the \( t \) distribution
- If the degrees of freedom is large enough, the \( t \) distribution is pretty much identical to the normal distribution

**Tables for the \( t \) distribution**

- If we want a 100\( \cdot \)C\% confidence interval, we need to find the value so that we have a probability of \( C \) between \(-t^*\) and \( t^*\) in a \( t \) distribution with \( n-1 \) degrees of freedom
- Example: 95\% confidence interval when \( n = 14 \) means that we need a tail probability of 0.025, so \( t^* = 2.16 \)
Confidence Intervals

- Known population SD $\sigma$: confidence intervals involve standard deviation and normal critical values
  $$\left( \overline{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

- Unknown population SD $\sigma$: confidence intervals involve standard error and critical values from a $t$ distribution with $n-1$ degrees of freedom
  $$\left( \overline{X} - t_{n-1, \alpha/2} \cdot \frac{s}{\sqrt{n}}, \overline{X} + t_{n-1, \alpha/2} \cdot \frac{s}{\sqrt{n}} \right)$$

- Confidence intervals involving $t$ distribution are usually wider (more conservative)

Hypothesis Tests

- Z-test (known population SD $\sigma$): test statistic involves standard deviation
  $$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$
  - p-value calculated from a standard normal distribution

- t-test (unknown population SD $\sigma$): test statistic involves standard error
  $$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$
  - p-value calculated from a $t$ distribution with $n-1$ df

- p-values are generally larger with t-test than z-test, so t-tests are less likely to reject $H_0$ (more conservative)

Example: NCAA Requirements

- NCAA Division 1 SAT requirements:
  - Athletes must get at least 820 on combined math/verbal SAT

- Combined SAT scores from sample of 100 athletes from high-school system:
  - $\overline{X} = 1019$
  - $s = 209$
  - $n = 100$

- Are athletes from this high-school system, on average, meeting the NCAA requirements?

Test Statistic for mean SAT $\mu$

- Do high-school athletes, on average, have SAT scores equal to the NCAA requirement?
  - Null hypothesis: $\mu = 820$
  - Two-sided alternative hypothesis: $\mu \neq 820$
  - Again, we are not given a known population SD $\sigma$ so we need to use standard error in our test statistic:
    $$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{1019 - 820}{209/\sqrt{100}} = 9.52$$

- Next step is calculating the p-value: what is the probability of getting such an extreme $T$ statistic?
  - To find p-value for $T$, we look at t-table row for 99 df, but this row doesn’t exist, so look at row for 100 df

Confidence Interval for mean SAT $\mu$

- We are not given a known population SD $\sigma$ so we need to use standard error and $t$ distribution

- We need to look up $n-1 = 99$ df in $t$-table, but there is no row for 99 df, so we just use the row for 100 df

- Looking at the column for a confidence level of 95%, we get our critical value $t = 1.98$

- Critical value $t = 1.98$ is only slightly bigger than $Z = 1.96$ from normal table, so we get a slightly wider interval by correctly using $t$ distribution for unknown SD $\sigma$

p-value for mean SAT $\mu$

- Values in 100 df row of $t$-table are all less than 9.52, with the closest being 3.39, which corresponds to a probability of 0.0005

- We conclude that the probability that test statistic $T$ is greater than 9.52 must be less than 0.0005
Example: Low-Carb Diets

- 41 overweight subjects placed on low carb diet
- no limits on calorie intake but few carbs allowed
- $X = \text{weight change measured after 16 weeks}$
  \[ \bar{X} = -5.4 \text{ lbs} \]
  \[ s = 7.48 \text{ lbs} \]
  \[ n = 41 \]
- Observed an average weight drop of 5.4 lbs, but also large standard deviation. Is this a significant change from zero?

Confidence Interval for mean weight loss $\mu$

- We are not given a known population SD $\sigma$, so we need to use standard error and $t$ distribution
- We need to look up $n-1 = 40$ df in $t$-table: looking at the column for a confidence level of 95%, we get our critical value $t^* = 2.02$

  \[ \left( \bar{X} \pm t^* \cdot \frac{s}{\sqrt{n}} \right) = \left( -5.4 \pm 2.02 \cdot \frac{7.48}{\sqrt{41}} \right) = (-7.8, -3.0) \]

- Intervals only contain negative weight change values, so clear evidence of significant weight loss

Exercise: do hypothesis test for significant weight loss yourselves

$X \pm t^* \cdot \frac{s}{\sqrt{n}}$

Small Samples

- We have used the standard error and $t$ distribution to correct our assumption of known population SD $\sigma$
- However, even $t$ distribution intervals/tests not as accurate if data is skewed or has influential outliers
- Rough guidelines from your textbook:
  - Large samples ($n > 40$): $t$ distribution can be used even for strongly skewed data or with outliers
  - Intermediate samples ($n > 15$): $t$ distribution can be used except for strongly skewed data or presence of outliers
  - Small samples ($n < 15$): $t$ distribution can only be used if data does not have skewness or outliers
  - What can we do for small samples of skewed data?

Techniques for Small Samples

- One option: use log transformation on data
  - Taking logarithm of data can often make it look more normal

- Another option: non-parametric tests like the sign test
  - Not required for this course, but mentioned in textbook if you're interested

Comparing Two Samples

- Up to now, we have looked at inference for one sample of continuous data
- Our next focus in this course is comparing the data from two different samples
- For now, we will assume that our these two different samples are independent of each other and come from two distinct populations

Population 1: $\mu_1$, $\sigma_1$

Population 2: $\mu_2$, $\sigma_2$

Sample 1: $\bar{x}_1$, $s_1$

Sample 2: $\bar{x}_2$, $s_2$

Blackout Baby Boom Revisited

- Nine months (Monday, August 8th) after Nov 1965 blackout, NY Times claimed an increased birth rate
- Already looked at single two-week sample: found no significant difference from usual rate (430 births/day)
- What if we instead look at difference between weekends and weekdays?

<table>
<thead>
<tr>
<th>Sun</th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
</tr>
</thead>
<tbody>
<tr>
<td>452</td>
<td>470</td>
<td>431</td>
<td>448</td>
<td>487</td>
<td>377</td>
<td></td>
</tr>
<tr>
<td>348</td>
<td>449</td>
<td>461</td>
<td>444</td>
<td>444</td>
<td>415</td>
<td></td>
</tr>
<tr>
<td>356</td>
<td>470</td>
<td>519</td>
<td>443</td>
<td>449</td>
<td>418</td>
<td>394</td>
</tr>
<tr>
<td>399</td>
<td>451</td>
<td>468</td>
<td>432</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weekdays</th>
<th>Weekends</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi_1 = 456.4$</td>
<td>$\chi_2 = 383.4$</td>
</tr>
<tr>
<td>$s_1 = 21.7$</td>
<td>$s_2 = 24.5$</td>
</tr>
<tr>
<td>$n_1 = 23$</td>
<td>$n_2 = 8$</td>
</tr>
</tbody>
</table>
Two-Sample Z test

- We want to test the null hypothesis that the two populations have different means.
  - $H_0: \mu_1 = \mu_2$ or equivalently, $\mu_1 - \mu_2 = 0$
  - Two-sided alternative hypothesis: $\mu_1 - \mu_2 \neq 0$

  - If we assume our population SDs $\sigma_1$ and $\sigma_2$ are known, we can calculate a two-sample Z statistic:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- We can then calculate a p-value from this Z statistic using the standard normal distribution.

Two-Sample Z test for Blackout Data

- To use Z test, we need to assume that our pop. SDs are known: $\sigma_1 = 21.7$ and $\sigma_2 = 24.5$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(456.4 - 383.4) - (0)}{\sqrt{\frac{21.7^2}{24} + \frac{24.5^2}{8}}} = 7.5$$

- We can then calculate a two-sided p-value for $Z=7.5$

  - From normal table, $P(Z > 7.5)$ is less than 0.0002, so our p-value $= 2 \times P(Z > 7.5)$ is less than 0.0004

  - We reject the null hypothesis at $\alpha$-level of 0.05 and conclude there is a significant difference between birth rates on weekends and weekdays

- Next class: get rid of assumption of known $\sigma_1$ and $\sigma_2$

Next Class – Lecture 18

- More on Comparing Means between Two Samples

  - Moore and McCabe: Section 7.1-7.2