Confidence Intervals and Hypothesis Testing

• Readings: Howell, Ch. 4, 7

The Sampling Distribution of the Mean (SDM)

• Derivation - See Thorne & Giesen (T&G), pp. 169-171 or online Chapter Overview for Ch. 9 http://www.mhhe.com/thorne4
• See also Howell, Chapter 4.2, 7.1
Getting to Know You

• Name, address, and phone number?
• With distributions, it’s
  – Central tendency
    • The mean of the sampling distribution of $M$
  – Variability
    • The variance of the sampling distribution of $M$
    • The standard deviation of the sampling distribution of $M$
  – Shape

Properties of the Sampling Distribution of the Mean (SDM):
The Sampling Distribution of the Mean: How is it constructed?

• The sampling distribution of the mean (SDM) is constructed by drawing samples of some fixed size (say $N = 10$), calculating the mean of the samples, repeating the process for a large number of samples, and plotting the resulting means in a frequency polygon. The resulting distribution is the sampling distribution of the mean.

Property 1

• The population mean of the SDM equals the mean of the Parent Population.

• In symbols:
Property 2

- The larger the size of the sample ($N$) taken when constructing the SDM, the more closely the SDM will approximate a normal curve.

- If the parent population is normal, the SDM will be exactly normal.

Property 3

- The larger the size of the sample ($N$), the smaller the standard deviation of the SDM.

- The standard deviation of the sampling distribution of the mean is called the standard error of the mean.

- In symbols:
AKA The Central Limit Theorem

- Howell says, “This is one of the most important theorems in statistics.” (7.1) His version, p. 180:

Given a population with mean $\mu$ and variance $\sigma^2$, the sampling distribution of the mean (the distribution of sample means) will have a mean equal to $\mu$ (i.e., $\mu_{\bar{X}} = \mu$), a variance $\left(\sigma_{\bar{X}}^2\right)$ equal to $\sigma^2/n$, and a standard deviation $\left(\sigma_{\bar{X}}\right)$ equal to $\sigma/\sqrt{n}$. The distribution will approach the normal distribution as $n$, the **sample size**, increases.$^1$

- Also see this section for his illustration of the “shape” property.

- Rice Virtual [Statistics](http://www.rdlabs.com/Statistics) Lab Simulation

Forming Any $z$ Score

- $z = \frac{\text{score} - \text{mean of scores}}{\text{st dev of scores}}$

- $z \left( X \right) =$

- $z \left( M \right) =$

- Finding $M$ for a $z$ - Formula 9-3 (T&G); Howell, 7.3 in context of Conf. Intervals.
Estimation and Degrees of Freedom

- Estimation for \( \mu \)
- Estimation of sigma-squared or sigma
  - \( N-1 \) in that case was the degrees of freedom
- General definition for degrees of freedom
  - the number of values that are free to vary after certain restrictions have been placed on the data.
  - Given that 5 scores must sum to zero, let’s try it.

- See Howell, p. 187 for a more specific rationale.

Estimating the Standard Error of the Mean

- Standard Error of the Mean formula:
- Estimated Standard Error of the Mean formula:
- Formula for z score:
- Formula when we estimate the standard error of the mean:
Summary

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Symbol</th>
<th>Equivalence</th>
<th>Sample Estimate Name</th>
<th>Symbol</th>
<th>Equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Mean of the SDM</td>
<td></td>
<td></td>
<td>Sample Mean of the SDM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population Variance of the SDM</td>
<td></td>
<td></td>
<td>Sample Variance of the SDM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population Standard Deviation of the SDM*</td>
<td></td>
<td></td>
<td>Sample standard error of the mean</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The t distribution- A little history

- Ireland, the Guinness brewing company, and William Sealy Gosset
- t distribution developed to be used with small samples and unknown population variances
- Student’s t -- rather than Gosset’s t-- so that the company wouldn’t have to admit to a bad batch. More
t and z -- What’s the difference?

• Sketch
• Role of the df
• Formula

• See Howell, Figure 7.5, p. 187.

The Confidence Interval

• Finding “deviant” mean scores
• Sketch: Note
  – The horizontal axis is the X axis
  – The deviant scores are now means
• Formula (where M = X-bar)
• 95% CI = +/- t(.05/2) s(M) + M
• 99% CI = +/- t(.01/2) s(M) + M
• See also, Howell, 7.3, Confidence Interval on µ
Example: Checking Progress Solution

Page 184

Digit span: \( N = 31, \sum X = 220.41, \sum X^2 = 1,705.72 \)

\[
\bar{X} = \frac{\sum X}{N} = \frac{220.41}{31} = 7.11
\]

\[
s = \sqrt{\frac{\sum X^2 - (\sum X)^2}{N - 1}} = \sqrt{\frac{1,705.72 - (220.41)^2}{31 - 1}}
\]

\[
s_\bar{X} = \frac{s}{\sqrt{N}} = \frac{2.15}{\sqrt{30}} = 0.386
df = N - 1 = 31 - 1 = 30
\]

95% CI = \( \bar{X} \pm t_{.05}s_\bar{X} \)

\[
= 7.11 \pm 2.0423(0.386) + 7.11
= 7.11 \pm 0.788 + 7.11
= 6.32 \text{ to } 7.90
\]

99% CI = \( \bar{X} \pm t_{.01}s_\bar{X} \)

\[
= 7.11 \pm 2.7500(0.386) + 7.11
= 7.11 \pm 1.062 + 7.11
= 6.05 \text{ to } 8.17
\]

Interpretations

- We can be 95% confident that the average digit span of the adult population is at least 6.32 and at most 7.90 digits.

- We can be 99% confident that the average digit span of the adult population is at least 6.05 and at most 8.17 digits.

- Sketches
What the Confidence Interval Really Means

• Confidence -- not probability
  – The sample has already been drawn
• For a 95% confidence interval, if the experiment were repeated 100 times, 95 of the intervals would contain (capture) the true population mean $\mu$.
  – The “rub” is, for any given one, we don’t know for sure if it captures $\mu$ or not, hence the confidence language.

Confidence vs. Probabililty

• “We place our confidence in the method rather than the interval.”
• You can say your confidence is .95 that Mu is found in the interval.
• You’re at risk if you say that the probability is .95 that Mu lies in the interval.
• Many will pounce if you do.
• See Howell, p. 193.
Hypothesis Testing: One-Sample t Test

- Example Handout – Ratings of Speech
  - [http://www2.msstate.edu/~jmg1/3103/one_sample_ttest.pdf](http://www2.msstate.edu/~jmg1/3103/one_sample_ttest.pdf)

---

**Example**

A new college president has been hired at a large state university. The president has just given their first speech to the general faculty. As the faculty members are leaving, 30 are randomly selected and asked to rate the speech on a scale from 1 to 7 with 1 being "horrible," 4 being "average," and 7 "fantastic." A computer program indicated the mean rating to be 4.4 with a = .05. These values were verified by using computational formulas on a pocket calculator. Test the null hypothesis that the faculty as a whole was neutral towards the speech.

1. **H₀:** μ = 4 (The faculty as a whole was neutral toward the speech. The faculty rating was average = 4.0)
2. **H₁:** μ ≠ 4 (The faculty as a whole was not neutral (rating not average) toward the speech)
3. Let α = .05
4. **Rejection Rule:** Reject H₀ if |tₙ₋₁| > tₙ₋₁(α/2) = 2.042
5. **Calculations and Test Statistic:**
   
   \[ t = \frac{\bar{X} - \mu_0}{S_x} = \frac{4.4 - 4.0}{1.1417565} = 0.34 = 3.0625 \]
6. **Decision:** Reject H₀ since |tₙ₋₁| = 3.0625 > 2.042 = tₙ₋₁, α = .05.
7. **Conclusion:** There is sufficient evidence to conclude that the faculty rating of the speech was not neutral (one significantly different from "average"). The faculty rating of the speech was (significantly) better than average, (4.4 - 2.5), \( p < .05 \).

\[ 95\% \text{ C.I. for } \mu: \bar{X} \pm t_{0.025, n-1} \cdot S_x = \text{CI}_{0.95} \]

\[ = 4.4 \pm (2.042)(1.1417565) \]

\[ = 4.4 \pm 0.3258 \]

\[ \text{CI}_{0.95} = 4.0744 \leq \mu \leq 4.7256 \]

**Interpretation:**

We can be 95% confident that the faculty rating of the speech was at least 4.074 and at most, 4.7258, on the average.
Hypothesis Testing: One-Sample t Test

- How deviant or unlikely is a particular sample mean?
  - Analogous to finding the deviant probability of a score with the z curve
- Start from knowledge or assumption of the parent population.
  - We know from much previous intelligence research that the population mean adult digit span is 7.0 (μ = 7).

Another Example
Research Context

- Does amphetamine (a stimulant) affect recent memory?
  - 25 volunteers take a small dose then take the digit span test
    - M = 7.53 and s = 0.97
- What is the probability of a mean this deviant?
- Outline: M --> t --> Area (or probability)
Solution

• Set up 7-step procedure
  • We assumed our sample came from the adult population with average digit span of 7.

• Our sample mean was 7.53, and this was 2.73 standard error units away from the mean.
  – \( S(M) = s/\sqrt{0.97} = 0.194 \)
  – \( T_{cal} = (7.53 - 7.00)/0.194 = 2.73 \)
  – \( T_{critical} = t(25, .025) = 2.0639 \) (look-up)
  – \( t(calculated) = 2.73 \), and the probability of a score this or more deviant was less than .05.
  – A sample mean this deviant is very unlikely to have come from a population where \( \mu = 7 \).

Conclusion

• A one-sample t test was performed to determine if amphetamine affects adult digit span. The digit span for the amphetamine sample (\( M = 7.53 \)) was significantly different from the normal adult digit span of 7, \( t(24) = 2.73, p < .05 (p < .02) \). Amphetamine significantly enhances recent memory.
The Seven-Step Procedure

• An easy, organized way to test hypotheses and come to a conclusion.
• See T&G, pp. 190-191.
• mhhe.com/thorne4 Chapter 9 Review

Gunsmoke: Errors in Hypothesis Testing

<table>
<thead>
<tr>
<th>Reality</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₀: BB Honest</td>
<td>Cheating</td>
</tr>
<tr>
<td></td>
<td>Error I: Shot</td>
</tr>
<tr>
<td></td>
<td>Innocent Man</td>
</tr>
<tr>
<td></td>
<td>Consequences?</td>
</tr>
<tr>
<td></td>
<td>Correct</td>
</tr>
<tr>
<td>Hₐ: BB Cheating</td>
<td>Not Cheating</td>
</tr>
<tr>
<td></td>
<td>Error II:</td>
</tr>
<tr>
<td></td>
<td>Continue to</td>
</tr>
<tr>
<td></td>
<td>Play</td>
</tr>
<tr>
<td></td>
<td>Consequences?</td>
</tr>
</tbody>
</table>
## Errors in Hypothesis Testing

### Decision

<table>
<thead>
<tr>
<th>Reality</th>
<th>Reject (Claim an Effect)</th>
<th>Fail to Reject (No Claim)</th>
</tr>
</thead>
</table>
| H₀: True | Type I Error
Alpha Error
 falsel claim
Consequences? | Correct |
| H₀: False | Correct | Type II Error
Beta Error
(failure of detection)
Consequences? |

### Consequences?
- Correct
- Type I Error
Alpha Error
(false claim)

### Errors in Hypothesis Testing

## Signal Detection Theory Version

### Is Signal Present? Decision

| Reality | “Yes” (Reject)
( Claim an Sig.
Present) | “No” Signal
Fail to Reject
(No Claim) |
|---------|--------------------------|---------------------------|
| H₀: True (No Sign.) | Type I, Alpha Error
Alpha Error
(false positive –
false alarm
Consequences?) | Correct
(Correct
Rejection) |
| H₀: False (Yes Signal present) | Correct
“Hit” | Type II, β Error
(False negative–
“Miss”)
(failure of detection)
Consequences? |
Consequences of Errors

• False Positive (false claim, Type I error)
  • Gunsmoke: If you shoot an innocent man, you’ll hang
  • Death penalty trials – avoid conviction of innocents. Evidence must by very very good.
  • Hunters in the woods must not shoot everything that moves in the bushes. It could be a person or dog. Threshold for pulling the trigger must be raised “until the signal is clearly present.”

• In these situations, really want to avoid false claims.

• False negatives (failures of detection) – not that bad
  – Gunsmoke: Continue to play & loose $.
  – Some criminals get lesser sentence or go free
  – The hunter misses their game once in a while.

False positives less harmful than False negatives

• Blood bank screening for AIDS (HIV) virus
  – Test very sensitive; tends to “detect” HIV sometimes when it is not there (false positives; Type I errors) but “always” detect when it “is there.” Thus we eliminate failures of detection (“Misses”).
  – “Shoot and ask questions later.”
  – Cost of false positives – some clean blood samples are wasted; a small price to pay for knowing you have no infected blood in YOUR transfusion!
The Almighty $d$ prime
A bias-free measure of test sensitivity
(under construction)

- A bias for claiming an effect ("shoot anything that moves") will result in more false positives
- A bias for "holding back" on claiming an effect ("be sure before you shoot") will result in more false negatives (Misses).
- $d'$ is the hit rate $H$ – the false alarm rate $F$ when $H$ and $F$ are $z$-transformed:
  - $d' = z(H) - z(F)$,
  - where $H = P("yes" | YES)$ and $F = P("yes" | NO)$
  - Colin Wilson has provided his Excel formula: $d' = \text{NORMINV(hit-rate,0,1)} - \text{NORMINV(false-alarm-rate,0,1)}$
    where Excel's NORMINV "Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation", 0 being the specified mean and 1 being the specified SD. See this link:
    - [http://wise.cgu.edu/sdtmod/signal_applet.asp](http://wise.cgu.edu/sdtmod/signal_applet.asp)

Power-Definitions

- Beta = Probability of a Type II error
  - Probability of failure of detection
- Power = 1 - Beta
- Power = 1 - Probability of a failure of detection
- Power = Probability of detection of an effect
Why Power

• How hard would you work if the chance of getting paid = .20, .50, .80?

• In behavioral science research we want a “good” chance of finding an effect, if it is present
  • (e.g., a cure for the common cold or
  • a form of cancer,
  • confirming YOUR thesis/dissertation hypothesis).

• Statistical power tells us our chance of finding an effect
  – the Probability of detection of an effect

Factors Affecting Power

• Alpha level
  – Increasing the alpha level _________ the level of power.

• Sample Size
  – Increasing the sample size _________ the level of power.

• Size of the effect (mu0 - mu1)
  – The larger the effect, the _________ the level of power.
Power Diagrams

• Text illustrations - Howell, 4.7
• Note best shown in two curves:
  – one curve for $H_0$ shows $\alpha$ and $1-\alpha$; when $H0$ is true
  – Another curve for $H_1$ true shows $\beta$ and $1-\beta$
  – See Howell’s Figure 4.3
• Spreadsheet illustration
  – Power.xls

What Should We Be Looking For?

• What statistical significance tells us
• What effect size tells us
• Practical significance
  – Statistical significance = statistical reliability
    • How sure we are of the presence of some effect
  – Practical significance = effect size = importance!
• More Later.
Using SPSS

- One sample t test
- Confidence Interval
- Example from Ratings of Speech

(enter 4.40 x 33 times, 0.51, 8.29. This simulates data for example.)

- Enter data; Use this syntax (or point-click)

T-TEST
  /TESTVAL = 4
  /VARIABLES = Ratings.
* Comment: Set Testval = 0 to get CI on the mean.

T-TEST
  /TESTVAL = 0
  /VARIABLES = Ratings.
Lab Assignment 1

• Review questions and formulae
• Formulae for #2.
• Help on setup for #5.
• SPSS startup