May 12, 2005 Lecture Notes

Significance testing

Case Study 12 continued:

estimated long run SD: \( \hat{SE}_{\bar{y}} = \frac{s}{\sqrt{n}} = \frac{5}{\sqrt{100}} = \frac{3 \text{ days}}{10} = 0.3 \text{ days} \)

long run histogram of \( \bar{y} \) if null is true:

\[ \frac{5.4 - 6.3}{0.3} = -3 = \text{z} \]

Even more extreme than -3 standard units

\[ 5.4 \]

\( z \)-test = -3 = \text{ z ; how unusual is this? } \)

\( P \text{ value} = P(\text{chance, if null is true, of getting data as extreme as or more extreme than what we got}) \)

go back to the alternative

where is more extreme than what we got?

if you look at the alternative, \( \mu < 6.3 \), so we look at \( \bar{y} < 5.4 \)

because that is the most extreme

(here we are only looking at one tail to get the \( P \)-value: one tail test)

\( P = 0.15\% \)

final step: if \( P \) is small, favor the alternative

if \( P \) is large, favor the null

How small is small enough? No general answer—depends on the real world consequences of choosing the wrong hypothesis

Conventional Answers:

\( P \leq 5\% \) \( \Rightarrow \) "statistically significant" (stat.slg)

\( P \leq 1\% \) \( \Rightarrow \) "highly stat. sig."

Here, the result is "highly statsig" (\( P = 0.15\% \)) \( \Rightarrow \) favor the alternative (the mean really has gone down)
However, you can’t tell if flextime caused this decline, might have been some other change over time

better design:

(Treatment) Compare 2 groups at the same time - one on flextime and the other not on flextime (control)

(End of Case Study 12)

Case Study 13: 1 = Female, 0 = Male

population

all students at UCB in 1977

sample

observed students

imaginary data set

[possible $\hat{p}$'s]

$\begin{align*}
N &= \text{big} \left[ \begin{array}{c} 15+ \\text{like SRS} \\ 0\text{'s} \end{array} \right] \\
\text{like IID} \\
\text{mean } p &= 33\% \\
\text{SD } \sigma &= \sqrt{\frac{p(1-p)}{n}} \\
\end{align*}$

population

histogram:

$\begin{array}{c}
67\% \\
33\% \\
\end{array}$

$\begin{align*}
\text{gender} &\xrightarrow{\text{like SRS}} \text{gender} \\
\left[ \begin{array}{c} 15+ \\
0\text{'s} \end{array} \right] &\xrightarrow{\text{n=100}} \text{mean } \hat{p} = 46\% \\
\left[ \begin{array}{c} 15+ \\
0\text{'s} \end{array} \right] &\xrightarrow{\text{n=100}} \text{mean } \hat{p} = ? (\text{ex: } 32\%) \\
\end{align*}$

null hypothesis: his method is like SRS

$\hat{p}$ should be expected to be about 33%

(we have to be able to try null on for size)

alternative hypothesis: his method is not like SRS

$\hat{p}$ might be < 33% or > 33% (2 sided alternative)
Long Run mean: $\hat{p}_{100} = p = 33\%$

Long Run SD: 
\[ SE_{100}(p) = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.33)(0.67)}{100}} = 0.047 = 4.7\% \]

Long Run histogram of $p$ if null is true:
- $SE = 4.7\%$
- Follows normal curve because of CLT
- Would be equally surprising as getting $40\%$
- Because $Z = 2.75$, inside $97.4\%$. $100\% - 97.4\% = 2.6\% = P$

2-sided $p$-test (because of the 2-sided alternative)
- $P < 0.6\%$: highly statistically significant because $p < 1\%$
- Is also practically significant because $40\%$ is quite different from $33\%$ as far as gender is concerned.
- If it was a one-tailed $p$-test, we would have gotten $0.3\%$ and we would have arrived at the same conclusion.

Thus, this is not an SRS - he probably likes talking to women.

(End of Case Study 13)

Pitfalls of significance testing:
- Statsig is not at all the same as practically sig
  - Example: new drug to lower systolic blood pressure
    - Null: (drug does not work) $\mu = 0$
    - Alt: (drug does make a change) $\mu = 0$
difference
(A - B)

\[
\begin{bmatrix}
-33 \text{ mm Hg} \\
+1 \text{ mm Hg}
\end{bmatrix}
\] \[n = 8000\]
pop. mean difference (A - B) = \(\mu\)

mean \(\bar{y} = -1\) mm Hg

(190 to 189 - not big enough of a difference in practical terms)

SD \(s = 20\) mm Hg

\[
\frac{s}{\sqrt{n}} = \frac{20 \text{ mm Hg}}{\sqrt{8000}} = 0.22 \text{ mm Hg}
\]

SE of difference: \(\hat{SE}(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{20 \text{ mm Hg}}{\sqrt{8000}} = 0.22 \text{ mm Hg} = 22\%\)

Long Run histogram of \(\bar{y}\) if null true:

\(\hat{SE} = 0.22\)

[Diagram of normal curve b/c CLT]

\(-1\) \(0\) \(1\)

\(-1.5\) \(0\) \(1.5\)

\(P = 0\%\) so way stat sig but not practically sig because too much data

ex: same situation but pilot study (much less people)

(A - B)

\[
\begin{bmatrix}
\cdot \\
\cdot \\
\cdot
\end{bmatrix}
\] \[n = 8\]

mean \(\bar{y} = -10\) mm Hg

SD = 20 mm Hg

\[
\frac{s}{\sqrt{n}} = \frac{7 \text{ mm Hg}}{\sqrt{8}} = 7 \text{ mm Hg}
\]

\[
\hat{SE}(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{20 \text{ mm Hg}}{\sqrt{8}} = 7 \text{ mm Hg}
\]

this difference is large in clinical terms, so is practically sig, but is it stat sig?
P = 16% "not stat sig" : insufficient evidence to reject null hyp.
- this happened because we had too little data.

Thus, choose n so that stat sig = prac. sig.