read FPP ch. 27
Significance testing

Case Study 12 continued
* put long run SD in model from May 10th
  estimated SD = \( \hat{\sigma}_{\text{long}}(\bar{Y}) = \frac{3.9}{\sqrt{100}} = 0.3 \) days

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

\[ \begin{align*}
\hat{\sigma} &= \frac{3.9}{\sqrt{100}} \\
&= 0.3 \\
&= (CT) \\
&= \text{raw units} \\
&= \text{standard units (z)} \\
&= \frac{5.4 - 0.3}{0.3} \\
&= 17 \\
&= \text{z score}
\end{align*} \]

\( z = -3 \), so what?
\( \rightarrow \) how unusual is \( z = -3 \)?

Answer:

\( p\text{-value} = P = \text{chance, if } H_0 \text{ is true, of getting data as extreme as, or more extreme than, what we got (find by calculating area to the left of 5.4 days under the curve because this is more extreme).} \)

\( \rightarrow \) how do you know where "more extreme" is?

Answer:

Look at the form of the alternative hypothesis; here alt. is \( \mu < 0.3 \) so we look only at \( \bar{Y} < 5.4 \)

here we look only at one tail (left) of the normal curve to get \( P \): one-tailed test; here \( P = 0.15\% \)

Final Step:
if \( P \) is small \( \rightarrow \) favor alt. hypothesis
if \( P \) is large \( \rightarrow \) favor null hypothesis
→ How small is small enough for \( P \)?

No general answer (depends on real world consequences of choosing wrong hypothesis)

**Conventional answer:** (stat. sig. = statistically significant)

\[
P \leq 5\% \iff \text{"stat. sig."} \\
P \leq 1\% \iff \text{"highly stat. sig."}
\]

So here, result is highly stat. sig. \((P = 0.15\%\) \(\rightarrow \)

favor alt. hypothesis (the mean really has gone down).

**But**

you can't tell if flex time caused this decline (might have been due to some other change over time).

**Better design:**

Compare 2 groups at the same time, one on flex time (treatment), the other not (control).

**Case Study B**

*Under null model \( H_0 : \theta = M \)

\[
\begin{align*}
\text{Pop:} & \quad \text{all students at VCB in 1977} \\
\text{Sample:} & \quad \text{the observed students} \\
\text{Imaginary data set:} & \quad \text{possible } \hat{\theta} \text{'s}
\end{align*}
\]

\[
\begin{align*}
N = 2 \\
\text{Gender:} & \quad \left[ \begin{array}{c} 1's \\ 3's \\ \emptyset \end{array} \right] \\
\text{Mean } \hat{p} = 46\% \\
\text{SD } \sigma = \sqrt{\frac{p(1-p)}{n}} \\
\text{Hypothetical } \hat{p} = \frac{1}{3} (\text{ex. 32%}) \\
\text{Pop. histogram:} & \quad \left[ \begin{array}{c} 1's \\ 3's \\ \emptyset \end{array} \right]
\end{align*}
\]

\[
\begin{align*}
\text{Long run mean } E_{\text{IID}}(\hat{p}) = p = 33\% \\
\text{Long run SD } SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \\
& = 4.7\% \\
\end{align*}
\]

\[
\begin{align*}
\text{Long run histogram of } \hat{p} \text{ if } H_0 \text{ true}
\end{align*}
\]

12 May 2005
**H₀ (null hypothesis)** (method of gathering) (data is like SRS) (expected to be = 33%)  
**Hₐ (alt. hypothesis)** (not SRS) (p might be either above or below 33%)  
= 2 sided alternative  

*we say H₀ is method like SRS because we have to be able to try out null.*

\[ Z = \frac{(\text{obs. p}) - (\text{expected p if H₀ true})}{(\text{SE of p if H₀ true})} = \frac{\text{signal}}{\text{noise}} = \frac{46\% - 33\%}{4.7\%} = 2.75 \]

2-tailed p-value (for a 2-sided alternative) = 0.04% (using z table)  
if 1-tailed p = 0.3% so arrive at same conclusion

Stat. Sig.? → Yes, (p ≤ 1%)  
Pract. Sig.? → Yes, 46% is quite different in a real-world sense from 33% in terms of gender

**Pitfalls of Sig. testing**

   (pract. sig. = practical significance)

ex. new drug is given to lower blood pressure  
**H₀** : (drug doesn't work) (μ = 0)  
**Hₐ** : (drug makes a change) (μ ≠ 0)

**Blood Pressure (systolic)**  

<table>
<thead>
<tr>
<th>after (a)</th>
<th>before (b)</th>
<th>diff (a-b)</th>
<th>sample data: pop. mean difference</th>
<th>(a-b) = μ</th>
</tr>
</thead>
</table>
| ≥         | ≥          | ex. (-3 mmHg) | ↑ n=8000 | ex. (+1 mmHg) | ↓  

Mean \( \bar{y} = -1 \text{ mmHg} \)  

\[ Z = \frac{\text{signal}}{\text{noise}} = \frac{(\text{obs. diff.)} - (\text{expected diff. if H₀ true})}{(\text{SE of diff. if H₀ true})} = \frac{-1 \text{ mmHg} - 0 \text{ mmHg}}{a/2 \text{ mmHg}} = -4.5 \]

*SE of diff = \( \sqrt{\frac{\text{SE}(\bar{y})}{n}} = \frac{20 \text{ mmHg}}{\sqrt{8000}} = 0.22 \text{ mmHg} \)
long run histogram of $\bar{y}$ if $H_0$ true:

$p \approx 0\%$ so **WAY** stat. sig. but not pract. sig.

$\Rightarrow$ this happened because $n$ was too big!

ex: same situation but pilot study

$\begin{bmatrix} A & B \end{bmatrix}$

$n = 8$

this difference is large in clinical terms

so it is pract. sig., but stat. sig.?

$\mu = -10 \text{ mmHg}$

$s = 20 \text{ mmHg}$

$z = \frac{-10 \text{ mmHg}}{10 \text{ mmHg}} = -1.4 \Rightarrow P \approx 16\% \Rightarrow \Rightarrow \Rightarrow$ so not stat. sig.

$*SE(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{20 \text{ mmHg}}{\sqrt{8}} \approx 7 \text{ mmHg}$

$\Rightarrow$ this happened because $n$

was too small!

Design idea: choose $n$ so that stat. sig. $=$ pract. sig.