Comparing 2 samples (pg. 22 VI. A) 17 May 2005

CONTINUING Pitfalls of significance testing

2. estimates + give-or-takes (interval estimation) are far more informative than p-values

Ex. Flex Time Case Study 12 p. 89

$H_0: \mu = 6.3 \text{ days}$

$H_a: \mu < 6.3 \text{ days}$

$p = 0.15\%$ so highly stat. sig., so null certainly looks wrong, $\mu$ is less than 6.3, but by how much?

Q. Does p-value by itself tell us anything about practical significance?

histogram of $z$ if $H_0$ true

$p = 0.15\% \rightarrow z = -3 = \frac{\text{signal}}{\text{noise}}$

but what you can infer stops there.

A: No, because you can't work out signal from \frac{\text{signal}}{\text{noise}}

by contrast: 95% C.I. for $\mu$

- estimate of $\mu$: $\bar{y} = 5.4$ days
- give-or-take: $SE(\bar{y}) = 0.3$ days
- 95% C.I.: $\bar{y} \pm 2SE(\bar{y}) = 5.4 \pm 2(0.3) \text{ days} = 5.4 \pm (0.6) \text{ day}$

→ shows pratic. sig.: $5.4 \text{ vs. } 6.3$ (a decline of 0.9 days per year per employee is large in real-world terms)

→ the 95% C.I. for $\mu$ "rejects null" of 6.3 because it is not inside the 95% interval

Intervals answer both questions: stat. and pract. sig., whereas sig. tests only answer stat. sig.
Case Study 14

treatment (Supposedly causal) variable: 
\( T \) = discount strategy vs. \( C \) = standard strategy

Outcome (effect/response) variable: Sales (# of cases of product sold in 6-week period)

design 1: basic design: controlled experiment
n = 120 stores (60 to treatment, 60 to control) which is a randomized controlled trial (RCT):
→ good design:
  1) validity (on average, if this design is repeated, you would get right answer) ; RCT is valid.
  2) efficiency (how accurately can basic question be answered)

design 2: matched pairs

[remember PCFs; ex. Location (overall sales volume)]
divide 120 stores into 60 pairs matched within each pair on PCF, overall sales volume and comparable location; within each pair assign to treatment and control at random

<table>
<thead>
<tr>
<th>Pair #</th>
<th>Discount Strategy</th>
<th>Standard Strategy</th>
<th>Overall Sales Volumes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>851</td>
<td>916</td>
<td>$2.0 M</td>
</tr>
<tr>
<td>2</td>
<td>903</td>
<td>1004</td>
<td>$2.5 M</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>60</td>
<td>787</td>
<td>699</td>
<td>$1.0 M</td>
</tr>
</tbody>
</table>

mean: 864  
SD: 58

Is this difference large in practical terms? \( \frac{T - C}{C} \) 
\( = \frac{(854 - 923)}{923} = -7.5\% \) so the average sales under treatment are 7.5% lower than under C (this is a pretty big difference)
→ IS design 2 valid? Yes
→ IS design 2 more efficient than design 1? Yes, if the PCF is strongly associated with the outcome which is true in this case.

<table>
<thead>
<tr>
<th>pair #</th>
<th>Disc.</th>
<th>Stan.</th>
<th>(D-S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>851</td>
<td>910</td>
<td>-65</td>
</tr>
<tr>
<td>2</td>
<td>903</td>
<td>1004</td>
<td>-101</td>
</tr>
<tr>
<td>3</td>
<td>33</td>
<td>12</td>
<td>98</td>
</tr>
<tr>
<td>40</td>
<td>787</td>
<td>1099</td>
<td>312</td>
</tr>
</tbody>
</table>

2 samples (D) vs. (S) but they are dependent samples because of pairing % with 2 dependent samples, focus on differences in each pair (which reduces problem to 1 sample)

\[ \text{Imaginary data set} \]
\[ \{ [-69, -40, \ldots] \} \]
\[ N = 100 \]
\[ \text{long run mean: } E_{iid} (\bar{y}) = \mu \]
\[ \text{long run SD: } \hat{\sigma}_{iid} (\bar{y}) = \frac{s}{\sqrt{n}} \approx \frac{150}{\sqrt{100}} = 15 \]
\[ \text{long run Histogram of } \bar{y} \]
\[ \hat{\sigma} = 10 \]
\[ \text{CLT} \]
\[ \mu \pm 2\hat{\sigma} = 10 \pm 2(10) = 10 \pm 20 \]
\[ \mu \pm 2\hat{\sigma} = 10 \pm 2(10) = 10 \pm 20 \]

**Inferential summary**

<table>
<thead>
<tr>
<th>Quantity of Interest</th>
<th>( \mu = \text{mean difference (D-S) in sales of overall possible pairs} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>( \bar{y} = -69 \text{ cases} )</td>
</tr>
<tr>
<td>Give or take</td>
<td>( \hat{\sigma} (\bar{y}) = \frac{s}{\sqrt{n}} = 19 \text{ cases} )</td>
</tr>
<tr>
<td>95% CI for ( \mu )</td>
<td>( \bar{y} \pm 2\hat{\sigma} = -69 \pm 2(19) = -69 \pm 38 = (-107, -31) \text{ cases} )</td>
</tr>
</tbody>
</table>
Stat. Sig. ?

H₀: \( \mu = 0 \) (obs. diff. is due to unlucky sampling)

\( \rightarrow \) By looking at the 95\% C.I., \( \mu = 0 \) is nowhere near the interval so null looks wrong. This diff. is both stat. and pract. sig.

Therefore, this evidence supports the psychologist’s theory and the use of pairing is for accuracy \((\text{smaller SE for diff.})\).

* Another similar design:

- before–after design (person \( \leftrightarrow \) pair)

\[
\begin{array}{ccc}
\text{model:} & \text{outcome} & \text{difference} \\
\text{before T} & \text{after T} & \text{after-before}
\end{array}
\]

**CASE STUDY 15**

Data:

<table>
<thead>
<tr>
<th>Tribe 1</th>
<th>Tribe 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>height</td>
<td>height</td>
</tr>
</tbody>
</table>
| \[
\begin{array}{cc}
\{ \} & n_1 = 25 \\
\{ \} & n_2 = 27
\end{array}
\] | \[
\begin{array}{cc}
\{ \} & \bar{y}_1 = 59.4 \text{ in} \\
\{ \} & \bar{y}_2 = 61.3 \text{ in}
\end{array}
\] |
| mean=\( \bar{y}_1 = 59.4 \text{ in} \) | mean=\( \bar{y}_2 = 61.3 \text{ in} \) |
| SD=\( s_1 = 1.8 \text{ in} \) | SD=\( s_2 = 2.4 \text{ in} \) |

2 independent samples: Know this because 2 sample sizes are different here and there is no connection between the two samples (very different from matched pairs or before–after which are 2 dependent samples).