Pitfalls of Significance Testing (continued)

2. Estimates + give/ take s. (interval estimation) are typically far more informative than P-values.

Example: Case Study 12 - Flextime

- \( H_0 \) (null): \( \mu_0 = 6.3 \) days
- \( H_a \) (alt.): \( \mu < 6.3 \) day

- P-value = 0.15%, so highly statistically significant.
- By the logic of significance testing, the null looks wrong.
- But... can you work backwards from the P-value to look for practical significance?
- \( \mu \) is < 6.3, but by how much?

Q: Does the P-value say anything about practical significance? (How far backwards can we go?)

hist of \( Z \) under \( H_0 \)

\[ P = 0.15\% \rightarrow Z = -3 = \frac{\text{signal}}{\text{noise}} \]

\[ \text{signal over noise} = \frac{\text{obs. } \bar{y}}{\text{exp. } \bar{y} \text{ under } H_0} \]

\[ = \frac{?}{(0.3)} \]

We know the ratio, but not the numerator and denominator of \( Z \).

A: No, because you cannot work out the "signal" of the signal/noise ratio.

By contrast...

95% CI for \( \mu \)

- Estimate of \( \mu = \bar{y} = 5.4 \) days
- give/take: estimated SE (\( \bar{y} \)) = 0.3 days

- 95% CI for \( \mu = \bar{y} \pm 2 \times \text{SE} (\bar{y}) = (5.4 \pm 2 \times 0.3) \) days

\[ = (5.4 \pm 0.6) \text{ days} \]
95% CI for μ:
4.8, 5.4, 6.0

Practical significance: 5.4 vs 6.0
- A decline of 0.9 days per yr. per employee
is large in real-world terms.
- The null value of 6.0 is not inside
the 95% CI, so we reject it.
- Intervals answer both questions:
  - Statistical significance?
  - Practical significance?
- Whereas significance tests only answer:
  - Statistical significance?

**Case Study**

**Comparing 2 samples**

Discount Pricing

Treatment (supposedly causal): discount strategy
Control: standard strategy

Outcome (effect/response): variable: sales (# of cases of
product sold over a 1-week period)

Basic design: controlled experiment

Design 1: Randomize 120 stores: 60 to T, 60 to C
- Randomized controlled trial (RCT)
- Good design: (but what is "good")
  1. Validity: on avg., if the design were
    repeated, you would get the right answer
    - RCT is valid
  2. Efficiency: how accurately the basic
    question can be answered
What if... let store owners choose whether they use the discount or standard strategy: observational study

PCFs: if location (overall sales volume)
- to control for this PCF j match up stores w/ equivalent locations.

matched pairs design: match up equivalent stores & randomize which store gets discount and which is standard.
- Divide 120 stores into 60 pairs matched within each pair on PCF (overall sales volume & comparable location), within each pair assign to treatment and control at random.
- Even better than RCC because it is more accurate.

<table>
<thead>
<tr>
<th>Pair #</th>
<th>Discount Strategy</th>
<th>Standard Strategy</th>
<th>Overall Sales Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>851</td>
<td>916</td>
<td>$2 M</td>
</tr>
<tr>
<td>2</td>
<td>908</td>
<td>1024</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>679</td>
<td>$1.0 M</td>
</tr>
</tbody>
</table>

mean: 854
SD: 58

Difference is practically large. About a 7.5% increase from discount strategy to the standard strategy.

\[
\frac{T - C}{C} = \frac{854 - 923}{923} \approx -7.5\%
\]

So avg. sales under treatment are 7.5% lower than under the control.
- This is a big difference in the world of business.
Is design 2 valid? - Yes: its results are repeatable.

Is design 2 more efficient than design 1? - Yes, if the PCF is strongly associated with the outcome.

True in this case: product sales are associated with the overall sales of each store.

Pair # | Discount | Standard | Difference (d-s) |
--- | --- | --- | --- |
1 | 85 | 91.6 | -65
2 | 70.8 | 125 | -54.2
3 | 68 | 78.7 | +10.7
4 | 70 | 89 | +19

2 Samples (D vs S) but 2 dependent samples

- Because of pairing.
- Look at the differences inside each pair.
- With 2 dependent samples, focus on differences within each pair.
- Reduces problem to 1 sample.

Pop: all possible D vs pairs

Sample: the observed pairs

Discount - standard

\[
\begin{align*}
N & = 50 \\
\bar{d} & = -69 cases \\
S & = 150 cases
\end{align*}
\]

\[
\frac{\bar{d}}{S}\]

Long run hist?

\[
\hat{SE} = 19
\]

95% CI = -69 ± 2(19)

-69 ± 38

-107 cases
Inferential Summary

<table>
<thead>
<tr>
<th>Quantity of Interest</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M = \text{mean difference (d-s)} ) in sales overall</td>
<td>( \bar{y} = -0.9 )</td>
<td>( \text{SE}_{\bar{y}} = 0.9 )</td>
<td>( -0.9 \pm 2.14(0.9) = (-1.97, -0.03) )</td>
</tr>
</tbody>
</table>

Stat.

- \( H_0: \mu = 0 \) (no difference between strategies); obs. difference is due to unlucky sampling.

- 0 is not in Confidence Interval (nowhere near this CI, so Null looks wrong).

\[
\text{CI: } [-1.97, -0.03]
\]

- This difference is both statistically and practically significant.
- Supports the (alt. Hypothesis) psychologist's theory.
- The point of pairing stores was to increase accuracy: smaller SE for the mean difference.

To summarize:

Another similar design:

- hold person constant; before + after design
- measure outcome before + treatment + after treatment, and find difference (after-before)

Case Study 15 Skeletons

<table>
<thead>
<tr>
<th>Tribe 1</th>
<th>Tribe 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{x} = 59.4'' )</td>
<td>( \bar{x} = 61.3'' )</td>
</tr>
<tr>
<td>SD = 2.9''</td>
<td>( S = 2.4'' )</td>
</tr>
</tbody>
</table>

2 independent samples: Dead giveaway =
the two sample sizes (n) are different (so they can’t be paired). 2) no connection between the 2 samples.

Different from analysis of matched pairs (dependent samples) or before/after.

Notation:

\[ n_1 = 25 \] (use subscripts to keep track of the Tribes/samples)

\[ n_2 = 27 \]

\[ y_1, \text{ etc.} \]

\[ y_2 \]