Beyond a Reasonable Doubt

in statistics...
Recap: Meaningful v.s. Significant

How meaningful: does the measured difference imply the desired differences in individuals.

How significant: could the measured difference be attributed to random chance.
Start with a simple problem: Means

Population 1
Mean Age = 45.6

Population 2
Mean Age = 32.1

What does this tell us about both populations?
Rule of thumb

Population 1

Mean Age = 45.6

Population 2

Mean Age = 32.1

Reasonable comparison between two populations when the spread is similar and the expected difference on the order of magnitude of the spread.
Always use medians unless you are interested in measuring a “rate” or an “expected effect”.

If someone uses “average” or mean when they are comparing populations, they are probably trying to mislead you.
"Meaningful"

Is the quantity that we calculate meaningful?

Is the population division meaningful?
# Simpson’s Paradox

Aggregates of heterogeneous populations can be misleading.

<table>
<thead>
<tr>
<th></th>
<th>Major 1</th>
<th>Major 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>West Coast Students</strong></td>
<td>7% (N=100)</td>
<td>2% (N=200)</td>
</tr>
<tr>
<td><strong>East Cost Students</strong></td>
<td>11% (N=12)</td>
<td>3% (N=1000)</td>
</tr>
</tbody>
</table>

**Principle of similar confidence!** Compare and aggregate things with similar spread and similar size.
Recap: Meaningful v.s. Significant

Population 1

stat1

Population 2

stat2

How meaningful: does the measured difference imply the desired differences in individuals.

How significant: could the measured difference be attributed to random chance.
“Significance”: Example

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Weight Loss</td>
<td></td>
</tr>
<tr>
<td>[-1.3, -4.5, 10.5]</td>
<td></td>
</tr>
<tr>
<td>[1.7, 5.2, 10.5]</td>
<td></td>
</tr>
</tbody>
</table>

Uncertainty about who got the treatment
“Significance”: Example

Interface 1
Click-Through-Rate

Interface 2
Click-Through-Rate

? Randomly assign users to interfaces
Randomized Controlled Trial

Randomly assign individuals to treatment and control

Why do you need randomness?

Same as opinion polling, you want to determine the full population (or expected) effect of an

Same as opinion polling, randomness must be independent.
How significant: could the measured difference be attributed to random chance.
Can think of it as a generalization of SRS....
Start with the assumption that there is **no meaningful** difference between the populations.
Null Hypothesis: Example

Treatment (No Effect)

Avg. Weight Loss

Control

Avg. Weight Loss

Avg Weight Loss for Treatment and Control are the same
Null Hypothesis: Example

Interface 1

Click-Through-Rate

Interface 2

Click-Through-Rate

Click-Through-Rate on both interfaces is the same
Null Hypotheses are Subjective

Think of it as an intellectual baseline...

No single “correct” choice (we’ll see this on Friday!)
Significance (or P-Value)

Probability that the difference between two statistics is at least as big assuming the null hypothesis is true.
Two-Sample Z-Test: Comparing Means

A simple p-value calculation that gives us intuition

Sample 1

Mean
\( \bar{\mu}_1 \approx \mu_1 \)

Sample 2

Mean
\( \bar{\mu}_2 \approx \mu_2 \)

Apply ideas from sampling stats!

\[ \epsilon \approx N(0, \frac{\sigma_{\text{pop}}^2}{K}) \]

Variance of the population

Size of the sample
Step 1. Assume the Null Hypothesis

Assume: $\mu_1 = \mu_2$

Define: $Z = \bar{\mu}_1 - \bar{\mu}_2$

No expected difference

$Z \sim \mathcal{N}(0, \text{var}(\bar{\mu}_1) + \text{var}(\bar{\mu}_2))$
Step 1. Assume the Null Hypothesis

Assume: $\mu_1 = \mu_2$

Define: $Z = \bar{\mu}_1 - \bar{\mu}_2$

$Z \sim \mathcal{N}(0, \text{var}(\bar{\mu}_1) + \text{var}(\bar{\mu}_2))$

$Z \sim \mathcal{N}(0, \frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2})$

- No expected difference
- Variances add up
- Variance of the sample
- Number of samples
Step 1. Assume the Null Hypothesis

Assume: \( \mu_1 = \mu_2 \)

\[
Z \sim \mathcal{N}(0, \frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2})
\]

“Null Hypothesis Model”: A model for the world assuming the null hypothesis is true
Step 2. Observe the Actual Difference

\[ \bar{\mu}_1 \approx \mu_1 \]

\[ \bar{\mu}_2 \approx \mu_2 \]

\[ \Delta = \bar{\mu}_1 - \bar{\mu}_2 \]
Step 3. Calculate P-Value

Assume:

\[ Z \sim \mathcal{N}(0, \frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2}) \]

Calculate:

\[ p = \Pr[Z > \mu] \]
Rules of Thumb Again
Step 3. Calculate P-Value

Assume:

\[ Z \sim \mathcal{N}(0, \frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2}) \]

Calculate:

\[ p = \Pr[Z > \Delta] \]

\[ z = \frac{\Delta}{\sqrt{\frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2}}} \]

# standard deviations of difference
Can be turned into a p-value using a table
Step 3. Calculate P-Value

\[ z = \frac{\Delta}{\sqrt{\frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2}}} \]

Normal Curve, mean = 0, SD = 1
Shaded Area = 0.95

```python
>>> import scipy.stats as st
>>> st.norm.ppf(.95)
1.6448536269514722
>>> st.norm.cdf(1.64)
0.94949741652589625
(1 - p-value)!
```
Two-Sample Z-Test: Comparing Means

A simple p-value calculation that gives us intuition

Sample 1

\[ \bar{\mu}_1, \sigma^2_1, K_1 \]

Calculate mean, variance, and size of sample

Sample 2

\[ \bar{\mu}_2, \sigma^2_2, K_2 \]

Calculate mean, variance, and size of sample

\[
Z = \frac{\Delta}{\sqrt{\frac{\sigma^2_1}{K_1} + \frac{\sigma^2_2}{K_2}}} \quad \text{Z-statistic}
\]
Randomized Controlled Trial

Randomly assign individuals to treatment and control

Apply the desired intervention

Observe results and calculate a p-value

Accept conclusion when p-value is below some confidence threshold (e.g., 0.05 or 0.01)
What are p-values?

Control the probability of “accidentally” accepting a null-hypothesis

\[ p < \alpha \]

**Calculated p-value** | **Acceptance thresh**

**False Discovery Rate:** The chance of accepting a null hypothesis in a RCT procedure
Multiple Hypothesis Testing

**Scenario:** Test 20 different drugs report results for the most significant treatment.

**Why is this bad?** 20 trials each with a small FDR means the overall FDR is much higher.

Test more hypotheses, you need to lower the acceptance threshold.
Multiple Hypothesis Testing

\[ p \leq \frac{\alpha}{R} \]

Calculated p-value \quad Acceptance thresh

R is the number of tests you run, divide acceptance threshold by number of tests
RCT Assumptions?

Randomly assign individuals to treatment and control
- Samples are drawn uniformly and randomly

Apply the desired intervention
- Treatment effects are independent

Observe results and calculate a p-value
- There is a reasonable null hypothesis

Accept conclusion when p-value is below some confidence threshold (e.g., 0.05 or 0.01)
- The threshold is statistically meaningful
Comparing Estimated Quantities

Sample 1

stat1

Sample 2

stat2