Quantitative Research Review: 2-1

- The Scientific Method
- Null Hypotheses, Alternative Hypotheses
- Defining a rejection region based on hypothesis
- T-tests
- Degrees of Freedom
- Error types
Type I, Type II Errors

<table>
<thead>
<tr>
<th>Our decision</th>
<th>True state of nature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>Type I error</td>
</tr>
<tr>
<td>‘Accept’ $H_0$</td>
<td>correct decision</td>
</tr>
<tr>
<td></td>
<td>$H_0$</td>
</tr>
<tr>
<td></td>
<td>$H_A$</td>
</tr>
</tbody>
</table>

(Orloff & Bloom, 2014)
Type I, Type II Errors

<table>
<thead>
<tr>
<th>Our decision</th>
<th>True state of nature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>Type I error</td>
</tr>
<tr>
<td>‘Accept’ $H_0$</td>
<td>correct decision</td>
</tr>
</tbody>
</table>

(Orloff & Bloom, 2014)

<table>
<thead>
<tr>
<th></th>
<th>$H_0$</th>
<th>$H_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>P(Reject $H_0$</td>
<td>P(Reject $H_0$</td>
</tr>
<tr>
<td>‘Accept’ $H_0$</td>
<td>P(Fail to</td>
<td>P(Fail to</td>
</tr>
<tr>
<td></td>
<td>$H_0</td>
<td>H_0$)</td>
</tr>
</tbody>
</table>
Type I, Type II Errors

<table>
<thead>
<tr>
<th>Reject $H_0$</th>
<th>$H_0$</th>
<th>$H_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(Reject $H_0$</td>
<td>$H_0$)</td>
<td>P(Reject $H_0$</td>
</tr>
</tbody>
</table>
Power

**significance level** ("p-value") = \( P(\text{type I error}) = P(\text{Reject } H_0 | H_0) \)
(probability we are incorrect)

\( power = 1 - P(\text{type II error}) = P(\text{Reject } H_0 | H_1) \)
(probability we are correct)

\[
\begin{array}{ccc}
\text{Reject } H_0 & | & H_0 & | & H_A \\
\hline
P(\text{Reject } H_0 | H_0) & | & P(\text{Reject } H_0 | H_1)
\end{array}
\]
Power

**significance level** ("p-value") = $P(\text{type I error}) = P(\text{Reject } H_0 \mid H_0)$
(probability we are incorrect)

$power = 1 - P(\text{type II error}) = P(\text{Reject } H_0 \mid H_1)$
(probability we are correct)

Formally, a power function of a test with rejection region, $R$, is:

$$P(\text{Reject } H_0 \mid \theta)$$

where $\theta$ is the parameters of the distribution over which $R$ is defined.
(e.g. $p$, $n$ for a binomial distribution)
Multi-test Correction

If alpha = .05, and I run 40 variables through significance tests, then, by chance, how many are likely to be significant?
Multi-test Correction

If alpha = .05, and I run 40 variables through significance tests, then, by chance, how many are likely to be significant?

2 (5% any test rejects the null, by chance)
Multi-test Correction

How to fix?

2 (5% any test rejects the null, by chance)
Multi-test Correction

How to fix?

What if all tests are independent?

=> “Bonferroni Correction” ($\alpha/m$)
Multi-test Correction

What if all tests are independent?

=> “Bonferroni Correction” (α/m)

But this may over-correct.
Multi-test Correction

Benjamini-Hochberg (a.k.a. Simes) Correction Procedure

1. Let $P_{(1)} < \ldots < P_{(m)}$ denote ordered p-values
2. Define:
   \[
   C_m = \sum_{i=1}^{m} \frac{1}{i} \quad \text{otherwise}
   \]
   where $C_m = 1$ if p-values are independent,

3. Let $T = P_{(R)}$ the “rejection threshold”
4. Reject all $H_{(0)}$ for which $P_i \leq T$

(Weiss, 2005)

But this may over-correct.
The Scientific Method

Think of
Interesting
Questions

What do I see in nature?
This can be from one's
own experiences, thoughts
or reading.

Why does that
pattern occur?

Refine, Alter,
Expand or Reject
Hypotheses

Develop
General
Theories

General theories must be
consistent with most or all
available data and with other
current theories.

Gather Data to
Test Predictions

Relevant data can come from the
literature, new observations or
formal experiments. Thorough
testing requires replication to
verify results.

Formulate
Hypotheses

What are the general
causes of the
phenomenon I am
wondering about?

Develop Testable
Predictions

If my hypothesis is correct,
then I expect a, b, c, …
The Scientific Method

Which steps are most subjective?

Develop General Theories
General theories must be consistent with most or all available data and with other current theories.

Make Observations
What do I see in nature? This can be from one’s own experiences, thoughts or reading.

Think of Interesting Questions
Why does that pattern occur?

Refine, Alter, Expand or Reject Hypotheses

Formulate Hypotheses
What are the general causes of the phenomenon I am wondering about?

Gather Data to Test Predictions
Relevant data can come from the literature, new observations or formal experiments. Thorough testing requires replication to verify results.

Develop Testable Predictions
If my hypothesis is correct, then I expect a, b, c, …
The Scientific Method
Potential Effect from Big Data

Gather Data to Test Predictions
Relevant data can come from the literature, new observations or formal experiments. Thorough testing requires replication to verify results.

Refine, Alter, Expand or Reject Hypotheses

Develop Testable Predictions
If my hypothesis is correct, then I expect a, b, c, …
Resampling Techniques

“nonparametric” tests

The permutation test:

- $t_{\text{obs}} = \text{Compute observed score}$
- passes = 0
- for 1 to $B$:
  - randomly permute the data, keeping the same sizes per class
  - $t_B = \text{compute score on permuted data}$
  - if $t_B > (\text{or} <) t_{\text{obs}}$: passes+=1
- $p_{\text{value}} = \frac{\text{passes}}{B}$

Application: comparing two distributions, especially when they are unknown.
Linear Regression

Finding a linear function based on $X$ to best yield $Y$.

$X = \text{“covariate” = “feature” = “predictor” = “regressor” = “independent variable”}$

$Y = \text{“response variable” = “outcome” = “dependent variable”}$

Regression:

goal: estimate the function $r$
Linear Regression

Finding a linear function based on $X$ to best yield $Y$.

$X$ = “covariate” = “feature” = “predictor” = “regressor” = “independent variable”

$Y$ = “response variable” = “outcome” = “dependent variable”

Regression:

**goal:** estimate the function $r$

Linear Regression (univariate version): $r(x) = \beta_0 + \beta_1 x$

**goal:** find $\beta_0, \beta_1$ such that $r(x) \approx \mathbb{E}(Y|X = x)$
Linear Regression

Simple Linear Regression

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

\[ r(x) = \beta_0 + \beta_1 x \]
Linear Regression

Simple Linear Regression

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

- **intercept**
- **slope**
- **error**
- **expected variance**
Linear Regression

Simple Linear Regression

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

**Estimated intercept and slope:**

\[ \hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \]

**Residual:**