Research Methods Primer

Lecture #3
What will we cover today?

Research Methodology
• Design and Timeline
• Independent and Dependent Variables
• Participant Selection and Recruitment
• Experimental Conditions and Controls, Ground Truth
• Data Collection and Management
• Statistical Power
• Correlations and Conclusions
Research Design

• What is the **Scientific Method**?
  – Falsifiable
  – Probabilities
  – Explanatory
  – Null hypothesis
    • Disprove ideas, not prove them
• Identify a specific research question or form a hypothesis
Formulating the Research Question

• Know the area and consider the purpose of experiment
• Types of research questions (1)
  – Frequency questions
    • How often do programmers take breaks?
  – Difference questions
    • Compared to artists, do programmers take longer breaks?
  – Relationship questions
    • Is there a relationship between the time when programmers take breaks and duration of the break?
Formulating the Research Question

• Types of research questions (2)
  – Descriptive
    • What proportion of programmers take breaks in the morning?
  – Causal
    • Does working remotely change the frequency and duration of breaks taken by programmers?
Research Design – Types

• Cross-sectional
  • Takes place in a single moment in time
  • Measures or observations taken at a “cross-section” or “slice” of time

• Longitudinal
  • Takes place over time
  • At least two measurements are taken at different points in time
Independent and Dependent Variables

• A variable is any entity that can take on different values
• Examples: gender, age, country of birth
• Independent Variable
  – What we manipulate
• Dependent Variable
  – What is affected by the manipulation
Research Question - Variables

• Is there a difference between male and female programmers for time taken during breaks?

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Dependent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALE</td>
<td>TIME TAKEN DURING BREAKS</td>
</tr>
<tr>
<td>FEMALE</td>
<td></td>
</tr>
</tbody>
</table>
Research Design – Variables

• Does gender or having a university degree have an effect on the time taken by programmers during breaks?

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Dependent Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALE</td>
<td>TIME TAKEN DURING BREAKS</td>
</tr>
<tr>
<td>FEMALE</td>
<td></td>
</tr>
<tr>
<td>HAS DEGREE</td>
<td></td>
</tr>
<tr>
<td>NO DEGREE</td>
<td></td>
</tr>
</tbody>
</table>
Research Question - Hypotheses

• Is there a difference between male and female programmers in time taken during breaks?

• Hypothesis: Female programmers take shorter breaks than male programmers

<table>
<thead>
<tr>
<th>MALE</th>
<th>TIME TAKEN DURING BREAKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEMALE</td>
<td></td>
</tr>
</tbody>
</table>

Independent Variables  Dependent Variable
Research Question - Hypotheses

• Null hypothesis (no difference)
  – $H_0 = \text{There is no } \textit{significant} \text{ difference between time taken during breaks and gender of programmer}$

• Alternate hypothesis (our prediction)
  – $H_1 = \text{Female programmers take } \textit{significantly} \text{ shorter breaks than male programmers}$

Null and alternate hypothesis should be mutually exclusive!
Hypotheses

- One-tailed
  - hypothesis specifies the direction
- Two-tailed
  - no direction specified
Hypothesis

• What is the two-tailed hypothesis for our programmer question?
• Null hypothesis (no difference)
  – $H_0$ = There is no *significant* difference between time taken during breaks and gender of programmer
• Alternate hypothesis (prediction)
  – $H_1$ = There is a *significant* difference between time taken during breaks and gender of programmer
Statistical Power

• The four components are:
  – **sample size**, or the number of units (e.g., people) accessible to the study
  – **effect size**, or the salience of the treatment relative to the noise in measurement
  – **alpha level** \((\alpha\text{, or significance level})\), or the odds that the observed result is due to chance
  – **power** \((\beta)\), or the odds that you will observe a treatment effect when it occurs

• It is possible to compute the value of fourth component given any three of the above
## Statistical Inference Decision Matrix

<table>
<thead>
<tr>
<th>In reality</th>
<th>$H_0$ is true, $H_1$ is false</th>
<th>$H_0$ is false, $H_1$ is true</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What we conclude</strong></td>
<td><strong>1-$\alpha$ (e.g., .95)</strong></td>
<td><strong>$\beta$ (e.g., .20)</strong></td>
</tr>
<tr>
<td><strong>We accept $H_0$, reject $H_1$</strong></td>
<td><strong>THE CONFIDENCE LEVEL</strong></td>
<td><strong>TYPE II ERROR</strong></td>
</tr>
<tr>
<td><strong>We say</strong>... &quot;There is no relationship&quot;; &quot;There is no difference, no gain&quot;; &quot;Our theory is wrong&quot;</td>
<td>The odds of saying there is no relationship, difference, gain, when in fact there is none</td>
<td>The odds of saying there is no relationship, difference, gain, when in fact there is one</td>
</tr>
<tr>
<td>95 times out of 100 when there is no effect, we'll say there is none</td>
<td>The odds of correctly not confirming our theory when it's true</td>
<td>20 times out of 100, when there is an effect, we'll say there isn't</td>
</tr>
<tr>
<td><strong>We reject $H_0$, accept $H_1$</strong></td>
<td><strong>$\alpha$ (e.g., .05)</strong></td>
<td><strong>TYPE I ERROR</strong> (SIGNIFICANCE LEVEL)</td>
</tr>
<tr>
<td><strong>We say</strong>... &quot;There is a relationship&quot;; &quot;There is a difference or gain&quot;; &quot;Our theory is correct&quot;</td>
<td>The odds of saying there is a relationship, difference, gain, when in fact there is not</td>
<td>The odds of saying that there is a relationship, difference, gain, when in fact there is one</td>
</tr>
<tr>
<td>5 times out of 100, when there is no effect, we'll say there is</td>
<td>The odds of confirming our theory correctly</td>
<td>80 times out of 100, when there is an effect, we'll say there is</td>
</tr>
<tr>
<td>We should keep this small when we can't afford/ risk wrongly concluding that our program works</td>
<td><strong>POWER</strong></td>
<td>We generally want this to be as large as possible</td>
</tr>
</tbody>
</table>
Some implications (1)

• the lower the $\alpha$, the lower the power; the higher the $\alpha$, the higher the power

• the lower the $\alpha$, the less likely it is that you will make a Type I Error (i.e., reject the null when it’s true)

• the lower the $\alpha$, the more "rigorous" the test

• an $\alpha$ of .01 (compared with .05 or .10) means the researcher is being relatively careful, s/he is only willing to risk being wrong 1 in a 100 times in rejecting the null when it’s true (i.e., saying there’s an effect when there really isn’t)
Some implications (2)

- an $\alpha$ of .01 (compared with .05 or .10) limits one’s chances of ending up in the bottom row, of concluding that the program has an effect. This means that both your statistical power and the chances of making a Type I Error are lower.

- an $\alpha$ of .01 means you have a 99% chance of saying there is no difference when there in fact is no difference (being in the upper left box)

- increasing $\alpha$ (e.g., from .01 to .05 or .10) increases the chances of making a Type I Error (i.e., saying there is a difference when there is not), decreases the chances of making a Type II Error (i.e., saying there is no difference when there is) and decreases the rigor of the test

- increasing $\alpha$ (e.g., from .01 to .05 or .10) increases power because one will be rejecting the null more often (i.e., accepting the alternative) and, consequently, when the alternative is true, there is a greater chance of accepting it (i.e., power)
Trustworthiness of findings

Reliability and validity

- Reliability: the consistency in the results of the measurement
- Validity: whether the measure is measuring what it claims to be measuring

- Low reliability and low validity
- High reliability and low validity
- High reliability and high validity
Trustworthiness of findings

• Sampling
  – Selecting units from a population of interest so that results can be generalized back to larger population

• External validity
  – Threats to external validity
    • Selection: Findings being specific to the group studies
    • Setting: Findings being specific to the particular context in which the study took place
    • History: Specific and unique historical experiences may determine or affect the findings
    • Construct effects: The particular constructs studied may be specific to the group studied
Trustworthiness of findings

• Internal Validity

The extent to which a study establishes that a factor or variable has actually caused the effect that is found (and in particular that it has not been caused by other factors)

Can be affected by:
1. History
2. Testing
3. Instrumentation
4. Regression
5. Mortality
6. Maturation
7. Selection
8. Selection by maturation interaction
9. Ambiguity about causal direction
10. Diffusion of treatment
11. Compensatory equalization of treatments
12. Compensatory rivalry
Reliability

• Can be measured using Cronbach’s alpha
• 0.7 or higher is satisfactory
• All items should measure construct in same direction

\[ \alpha = \frac{N \times \bar{r}}{1 + (N - 1) \times \bar{r}} \]
Participant Selection and Recruitment

• Ethics in Research
  – Voluntary participation
  – Informed consent
  – Risk of harm
  – Confidentiality and Anonymity

• Institutional Review Board
Survey Measures – Ground Truth

• Questionnaires and interviews
• Interviews are qualitatively analyzed
• Questionnaires can contain multiple questions that measure the same variable
• Reliability of measurement of questions estimated using Cronbach’s alpha
Constructing the survey

• Some ground rules:
  – Pose one question at a time
  – Question should be unambiguous
  – The language and expected knowledge of respondents should be taken into consideration
  – Neutral language, no leading questions
  – No unnecessary sensitive questions
  – Avoid negative or double negative questions
Experimental Conditions
Data Collection and Management

• Pre and post-test data
• Observational data
• Experimental data

• Data Logging
  – Usually in a database
  – Securely stored to ensure confidentiality and anonymity
Data Analysis

• Descriptive Statistics
  – Simply describing what the data is or shows
  – For example, the proportion of <20, 20-40 and >40 year olds in a given research study
  – Gives us a summary of what the data contains
Inferential Analysis

• We use inferential statistics to draw conclusions from the data
• Many different types of statistical methods
  – Correlations
  – T-test
  – Analysis of covariance
  – and others..
Correlation

- Describes the degree to which two variables are related
- Suppose this data from 20 male respondents

<table>
<thead>
<tr>
<th>Person</th>
<th>Height</th>
<th>Self Esteem</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>4.1</td>
</tr>
<tr>
<td>2</td>
<td>71</td>
<td>4.6</td>
</tr>
<tr>
<td>3</td>
<td>62</td>
<td>3.8</td>
</tr>
<tr>
<td>4</td>
<td>75</td>
<td>4.4</td>
</tr>
<tr>
<td>5</td>
<td>58</td>
<td>3.2</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>3.1</td>
</tr>
<tr>
<td>7</td>
<td>67</td>
<td>3.8</td>
</tr>
<tr>
<td>8</td>
<td>68</td>
<td>4.1</td>
</tr>
<tr>
<td>9</td>
<td>71</td>
<td>4.3</td>
</tr>
<tr>
<td>10</td>
<td>69</td>
<td>3.7</td>
</tr>
<tr>
<td>11</td>
<td>68</td>
<td>3.5</td>
</tr>
<tr>
<td>12</td>
<td>67</td>
<td>3.2</td>
</tr>
<tr>
<td>13</td>
<td>63</td>
<td>3.7</td>
</tr>
<tr>
<td>14</td>
<td>62</td>
<td>3.3</td>
</tr>
<tr>
<td>15</td>
<td>60</td>
<td>3.4</td>
</tr>
<tr>
<td>16</td>
<td>63</td>
<td>4.0</td>
</tr>
<tr>
<td>17</td>
<td>65</td>
<td>4.1</td>
</tr>
<tr>
<td>18</td>
<td>67</td>
<td>3.8</td>
</tr>
<tr>
<td>19</td>
<td>63</td>
<td>3.4</td>
</tr>
<tr>
<td>20</td>
<td>61</td>
<td>3.6</td>
</tr>
</tbody>
</table>
Plotting both variables

The slope of the line indicates a positive relationship.
Calculating the correlation

For our sample data,

\[ r = 0.73 \]
Is the correlation significant?

• Null hypothesis $H_0 = \text{no significant relationship between height and self-esteem}$

• Alternate hypothesis $H_1 = \text{relationship between height and self-esteem is statistically significant}$
Is the correlation significant?

• Most common significance level chosen is 0.05
  – this means odds that the correlation is a chance occurrence are no more than 5 out of 100

• Degrees of freedom
  \[ = N - 2 \]
  \[ = 20 - 2 = 18 \]
Is the correlation significant?

• One-tailed or two-tailed?
• We do not have any apriori assumptions regarding the relationship between height and self-esteem, so two-tailed
• Using the significance level ($\alpha = 0.05$), degrees of freedom ($df = 18$) and type of test (two-tailed), look up value of significance in a lookup table readily available in statistics textbooks (or online 😊!)
Is the correlation significant?

- The lookup table reveals critical value to be 0.4438
- This means, if our computed correlation is >0.4438 OR <-0.4438, then the correlation is significant
- Since our computed value is 0.73, the result is statistically significant
- Reject the null hypothesis and accept alternate hypothesis
Other statistical analyses

• T-test
  – Assesses whether the means of two groups are statistically different from one another
T-test
The t-test is used to differentiate between these three scenarios.

The two groups at the bottom are most different because there is little overlap between curves.

The t-test measures the difference between means relative to the variability in scores.
What we covered today?

Research Methodology

• Design and Timeline
• Independent and Dependent Variables
• Participant Selection and Recruitment
• Experimental Conditions and Controls, Ground Truth
• Data Collection and Management
• Statistical Power
• Correlations and Conclusions
Assigned Reading

Discriminating Gender on Twitter
  – John D. Burger and John Henderson and George Kim and Guido Zarrella
  – EMNLP 2011

• Responses due on