Conducting Human-Subject Experiments with Virtual and Augmented Reality

VR 2004 Tutorial

J. Edward Swan II, Naval Research Laboratory (organizer)

Joseph L. Gabbard, Virginia Tech

Deborah Hix, Virginia Tech

Steve Ellis, NASA Ames

Bernard D. Adelstein, NASA Ames
<table>
<thead>
<tr>
<th>Time</th>
<th>Duration</th>
<th>Event</th>
<th>Instructor(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>800 AM</td>
<td>0.4 hours</td>
<td>Introduction and Overview</td>
<td>Ed</td>
</tr>
<tr>
<td>830 AM</td>
<td>1.6 hours</td>
<td>Usability Engineering</td>
<td>Joe, Debby</td>
</tr>
<tr>
<td>1000 AM</td>
<td>0.5 hours</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>1030 AM</td>
<td>1.5 hours</td>
<td>Experimental Design and Analysis</td>
<td>Ed</td>
</tr>
<tr>
<td>1200 PM</td>
<td>1.5 hours</td>
<td>Lunch Break</td>
<td></td>
</tr>
<tr>
<td>130 PM</td>
<td>0.5 hours</td>
<td>Experimental Design and Analysis</td>
<td>Ed</td>
</tr>
<tr>
<td>200 PM</td>
<td>1.0 hours</td>
<td>Human Performance Studies in Virtual Environments</td>
<td>Steve</td>
</tr>
<tr>
<td>300 PM</td>
<td>0.5 hours</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>330 PM</td>
<td>1.0 hours</td>
<td>Psychophysics: Classical Methods</td>
<td>Dov</td>
</tr>
<tr>
<td>430 PM</td>
<td>0.5 hours</td>
<td>Final Questions and Discussion</td>
<td>Ed</td>
</tr>
</tbody>
</table>
Outline

• **Empiricism**
• Experimental Validity
• Usability Engineering
• Experimental Design
• Gathering Data
• Describing Data
  – Graphing Data
  – Descriptive Statistics
• Inferential Statistics
  – Hypothesis Testing
  – Hypothesis Testing Means
  – Power
  – Analysis of Variance and Factorial Experiments
Why Human Subject (HS) Experiments?

• VR and AR hardware/software more mature
• Focus of field:
  – Implementing technology → using technology
• Increasingly running HS experiments:
  – How do humans perceive, manipulate, cognate with VR, AR-mediated information?
  – Measure utility of AR / VR for applications

• HS experiments at VR 2003:
  – 10/29 papers (35%)
  – 5/14 posters (36%)
Logical Deduction vs. Empiricism

• **Logical Deduction**
  – Analytic solutions in closed form
  – Amenable to proof techniques
  – Much of computer science fits here
  – Examples:
    • Computability (what can be calculated?)
    • Complexity theory (how efficient is this algorithm?)

• **Empirical Inquiry**
  – Answers questions that cannot be proved analytically
  – Much of science falls into this area
  – Antithetical to mathematics, computer science
What is Empiricism?

• The Empirical Technique
  – Develop a hypothesis, perhaps based on a theory
  – Make the hypothesis testable
  – Develop an empirical experiment
  – Collect and analyze data
  – Accept or refute the hypothesis
  – Relate the results back to the theory
  – If worthy, communicate the results to your community

• Statistics:
  – Foundation for empirical work; necessary but not sufficient
  – Often not useful for managing problems of gathering, interpreting, and communicating empirical information.
Where is Empiricism Used?

• Humans are very non-analytic
• Fields that study humans:
  – Psychology / social sciences
  – Industrial engineering
  – Ergonomics
  – Business / management
  – Medicine
• Fields that don’t study humans:
  – Agriculture, natural sciences, etc.
• Computer Science:
  – HCI
  – Software engineering
Experimental Validity

• Empiricism
• *Experimental Validity*
• Usability Engineering
• Experimental Design
• Gathering Data
• Describing Data
  – Graphing Data
  – Descriptive Statistics
• Inferential Statistics
  – Hypothesis Testing
  – Hypothesis Testing Means
  – Power
  – Analysis of Variance and Factorial Experiments
Designing Valid Empirical Experiments

• **Experimental Validity**
  – Does experiment really measure what we want it to measure?
  – Do our results really mean what we think (and hope) they mean?
  – Are our results **reliable**?
    • If we run the experiment again, will we get the same results?
    • Will others get the same results?

• **Validity is a large topic in empirical inquiry**
  – **Usability Engineering** can greatly enhance validity of VR / AR experiments
Experimental Variables

• Independent Variables
  – What the experiment is studying
  – Occur at different levels
    • Example: stereopsis, at the levels of stereo, mono
  – Systematically varied by experiment

• Dependent Variables
  – What the experiment measures
  – Assume dependent variables will be effected by independent variables
  – Must be measurable quantities
    • Time, task completion counts, error counts, survey answers, scores, etc.
    • Example: VR navigation performance, in total time
Experimental Variables

• Independent variables can vary in two ways
  – **Between-subjects**: each subject sees a different level of the variable
    • Example: ½ of subjects see stereo, ½ see mono
  – **Within-subjects**: each subject sees all levels of the variable
    • Example: each subject sees both stereo and mono

• **Confounding factors** (or confounding variables)
  – Factors that are not being studied, but will still affect experiment
    • Example: stereo condition less bright than mono condition
  – Important to **predict and control confounding factors**, or experimental validity will suffer
Usability Engineering

- Empiricism
- Experimental Validity
- *Usability Engineering*
- Experimental Design
- Gathering Data
- Describing Data
  - Graphing Data
  - Descriptive Statistics
- Inferential Statistics
  - Hypothesis Testing
  - Hypothesis Testing Means
  - Power
  - Analysis of Variance and Factorial Experiments
Experimental Design

• Empiricism
• Experimental Validity
• Usability Engineering
• *Experimental Design*
• Gathering Data
• Describing Data
  – Graphing Data
  – Descriptive Statistics
• Inferential Statistics
  – Hypothesis Testing
  – Hypothesis Testing Means
  – Power
  – Analysis of Variance and Factorial Experiments
Experimental Designs

• 2 x 1 is simplest possible design, with one independent variable at two levels:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Stereopsis</th>
</tr>
</thead>
<tbody>
<tr>
<td>level 1</td>
<td>stereo</td>
</tr>
<tr>
<td>level 2</td>
<td>mono</td>
</tr>
</tbody>
</table>

• Important confounding factors for within subject variables:
  – Learning effects
  – Fatigue effects

• Control these by **counterbalancing** the design
  – Ensure no systematic variation between levels and the order they are presented to subjects

<table>
<thead>
<tr>
<th>Subjects</th>
<th>1st condition</th>
<th>2nd condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 3, 5, 7</td>
<td>stereo</td>
<td>mono</td>
</tr>
<tr>
<td>2, 4, 6, 8</td>
<td>mono</td>
<td>stereo</td>
</tr>
</tbody>
</table>
Factorial Designs

- $n \times 1$ designs generalize the number of levels:

<table>
<thead>
<tr>
<th>VE terrain type</th>
<th>stereo</th>
<th>mono</th>
</tr>
</thead>
<tbody>
<tr>
<td>flat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hilly</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mountainous</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Factorial designs generalize number of independent variables and the number of levels of each variable
- Examples: $n \times m$ design, $n \times m \times p$ design, etc.
- Must watch for factorial explosion of design size!

3 x 2 design:

<table>
<thead>
<tr>
<th>VE terrain type</th>
<th>Stereopsis</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>stereo</td>
<td>mono</td>
</tr>
<tr>
<td>flat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hilly</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mountainous</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cells and Levels

- **Cell**: each combination of levels
- **Repetitions**: typically, the combination of levels at each cell is repeated a number of times

Example of how this design might be described:
- “A 3 (VE terrain type) by 2 (stereopsis) within-subjects design, with 4 repetitions of each cell.”
- This means each subject would see $3 \times 2 \times 4 = 24$ total conditions
- The presentation order would be counterbalanced
Counterbalancing

• Addresses time-based confounding factors:
  – Within-subjects variables: control learning and fatigue effects
  – Between-subjects variables: control calibration drift, weather, other factors that vary with time

• There are two counterbalancing methods:
  – Random permutations
  – Systematic variation
  • Latin squares are a very useful and popular technique

\[
\begin{bmatrix}
1 & 2 \\
2 & 1 \\
\end{bmatrix}
\begin{bmatrix}
1 & 2 & 3 \\
1 & 3 & 2 \\
2 & 1 & 3 \\
2 & 3 & 1 \\
3 & 1 & 2 \\
3 & 2 & 1 \\
\end{bmatrix}
\begin{bmatrix}
1 & 2 & 3 & 4 \\
2 & 4 & 1 & 3 \\
3 & 1 & 4 & 2 \\
4 & 3 & 2 & 1 \\
\end{bmatrix}
\]

• Latin square properties:
  – Every level appears in every position the same number of times
  – Every level is followed by every other level
  – Every level is preceded by every other level

6 x 3 (there is no 3 x 3)

2 x 2

6 x 3 (there is no 3 x 3)
Counterbalancing Example

- “A 3 (VE terrain type) by 2 (stereopsis) within-subjects design, with 4 repetitions of each cell.”
- Form Cartesian product of Latin squares
  \( \{6 \times 3\} \) (VE Terrain Type) \( \otimes \) \( \{2 \times 2\} \) (Stereopsis)
- Perfectly counterbalances groups of 12 subjects

<table>
<thead>
<tr>
<th>Subject</th>
<th>Presentation Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1A, 1B, 2A, 2B, 3A, 3B</td>
</tr>
<tr>
<td>2</td>
<td>1B, 1A, 2B, 2A, 3B, 3A</td>
</tr>
<tr>
<td>3</td>
<td>1A, 1B, 3A, 3B, 2A, 2B</td>
</tr>
<tr>
<td>4</td>
<td>1B, 1A, 3B, 3A, 2B, 2A</td>
</tr>
<tr>
<td>5</td>
<td>2A, 2B, 1A, 1B, 3A, 3B</td>
</tr>
<tr>
<td>6</td>
<td>2B, 2A, 1B, 1A, 3B, 3A</td>
</tr>
<tr>
<td>7</td>
<td>2A, 2B, 3A, 3B, 1A, 1B</td>
</tr>
<tr>
<td>8</td>
<td>2B, 2A, 3B, 3A, 1B, 1A</td>
</tr>
<tr>
<td>9</td>
<td>3A, 3B, 1A, 1B, 2A, 2B</td>
</tr>
<tr>
<td>10</td>
<td>3B, 3A, 1B, 1A, 2B, 2A</td>
</tr>
<tr>
<td>11</td>
<td>3A, 3B, 2A, 2B, 1A, 1B</td>
</tr>
<tr>
<td>12</td>
<td>3B, 3A, 2B, 1A, 1B, 1A</td>
</tr>
</tbody>
</table>

\[
\begin{bmatrix}
1 & 2 & 3 \\
1 & 3 & 2 \\
2 & 1 & 3 \\
2 & 3 & 1 \\
3 & 1 & 2 \\
3 & 2 & 1 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
A & B \\
B & A \\
\end{bmatrix}
\]
Experimental Design Example #1

- All variables within-subject

From [Living et al. 03]
# Experimental Design Example #2

<table>
<thead>
<tr>
<th>Stereo Viewing</th>
<th>on</th>
<th>off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Movement</td>
<td>rate</td>
<td>position</td>
</tr>
<tr>
<td>Frame of Reference</td>
<td>ego</td>
<td>exo</td>
</tr>
<tr>
<td>Between Subject</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer Platform</td>
<td>Within Subject</td>
<td></td>
</tr>
<tr>
<td>wall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>workbench</td>
<td></td>
<td></td>
</tr>
<tr>
<td>desktop</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Mixed design: some variables between-subject, others within-subject.

From [Swan et al. 03]
Gathering Data

• Empiricism
• Experimental Validity
• Usability Engineering
• Experimental Design
• Gathering Data
• Describing Data
  – Graphing Data
  – Descriptive Statistics
• Inferential Statistics
  – Hypothesis Testing
  – Hypothesis Testing Means
  – Power
  – Analysis of Variance and Factorial Experiments
Gathering Data

- Some unique aspects of VR and AR
  - Can capture, log, and analyze head trajectory
  - If we log head / hand trajectory so we can play it back, must have way of logging critical incidents
  - VR / AR equipment more fragile than other UI setups

- In a CAVE:
  - Observing a subject can break their presence / immersion
  - Determining button presses when experimenter cannot see wand

- In AR, very difficult to know what user is seeing
  - Can mount separate display near user or on their back
  - Could mount lightweight camera on user’s head
Graphing Data

- Empiricism
- Experimental Validity
- Usability Engineering
- Experimental Design
- Gathering Data

- Describing Data
  - Graphing Data
  - Descriptive Statistics

- Inferential Statistics
  - Hypothesis Testing
  - Hypothesis Testing Means
  - Power
  - Analysis of Variance and Factorial Experiments
Types of Statistics

• Descriptive Statistics
  – Describe and explore data
  – Summary statistics:
    many numbers $\rightarrow$ few numbers
  – All types of graphs and visual representations
  – Data analysis begins with descriptive stats
    • Understand data distribution
    • Test assumptions of significance tests

• Inferential Statistics
  – Detect relationships in data
  – Significance tests
  – Infer population characteristics from sample characteristics
Summary Statistics

• Many numbers $\rightarrow$ few numbers

• Measures of central tendency:
  – Mean: average
  – Median: middle value
  – Mode: most common, highest point

• Measures of variability / dispersion:
  – Mean absolute deviation
  – Variance
  – Standard Deviation
Exploring Data with Graphs

• Histogram common data overview method

median = 59.5  mean = 60.26  mode = 62
Classifying Data with Histograms

From [Howell 02] p 28
Stem-and-Leaf:
Histogram From Actual Data

From [Howell 02] p 21, 23

FIGURE 2.4 Stem-and-leaf display for reaction time data
Boxplot

- Emphasizes variation and relationship to mean
- Because narrow, can be used to display side-by-side groups
Example Histogram and Boxplot from Real Data

- Mean = 2355
- Median = 1453
- Min value
- 25th
- 75th
- Upper fence
- Max values (outliers)

Data from [Living et al. 03]
We Have Only Scratched the Surface...

- There are a vary large number of graphing techniques
- Tufte’s [83, 90] works are classic, and stat books show many more examples (e.g. Howell [03]).

Lots of good examples...

And plenty of bad examples!

From [Tufte 83], p 134, 62
Descriptive Statistics

• Empiricism
• Experimental Validity
• Usability Engineering
• Experimental Design
• Gathering Data
• Describing Data
  – Graphing Data
  – Descriptive Statistics
• Inferential Statistics
  – Hypothesis Testing
  – Hypothesis Testing Means
  – Power
  – Analysis of Variance and Factorial Experiments
Populations and Samples

• Population:
  – Set containing every possible element that we want to measure
  – Usually a Platonic, theoretical construct
  – Mean: $\mu$  Variance: $\sigma^2$  Standard deviation: $\sigma$

• Sample:
  – Set containing the elements we actually measure (our subjects)
  – Subset of related population
  – Mean: $\bar{X}$  Variance: $s^2$  Standard deviation: $s$
  Number of samples: $N$
Measuring Variability / Dispersion

Mean:

$$\bar{X} = \frac{\sum X}{N}$$

Variance:

$$s^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Mean absolute deviation:

$$\text{m.a.d.} = \frac{\sum |X - \bar{X}|}{N}$$

Standard deviation:

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$$

- Standard deviation uses same units as samples and mean.
- Calculation of population variance $\sigma^2$ is theoretical, because $\mu$ almost never known and the population size $N$ would be very large (perhaps infinity).
Sums of Squares, Degrees of Freedom, Mean Squares

• Very common terms and concepts

\[ s^2 = \frac{\sum (X - \bar{X})^2}{N - 1} = \frac{SS}{df} = \frac{\text{sums of squares}}{\text{degrees of freedom}} = MS (\text{mean squares}) \]

• Sums of squares:
  – Summed squared deviations from mean

• Degrees of freedom:
  – Given a set of \( N \) observations used in a calculation, how many numbers in the set may vary
  – Equal to \( N \) minus number of means calculated

• Mean squares:
  – Sums of squares divided by degrees of freedom
  – Another term for variance, used in ANOVA
Hypothesis Testing

- Empiricism
- Experimental Validity
- Usability Engineering
- Experimental Design
- Gathering Data
- Describing Data
  - Graphing Data
  - Descriptive Statistics
- **Inferential Statistics**
  - *Hypothesis Testing*
  - Hypothesis Testing Means
  - Power
  - Analysis of Variance and Factorial Experiments
Hypothesis Testing

• Goal is to infer population characteristics from sample characteristics

From [Howell 02], p 78
Testable Hypothesis

- **General hypothesis**: The research question that motivates the experiment.

- **Testable hypothesis**: The research question expressed in a way that can be measured and studied.

- Generating a **good** testable hypothesis is a real skill of experimental design.
  - By **good**, we mean contributes to experimental validity.
  - Skill best learned by studying and critiquing previous experiments.
Testable Hypothesis Example

- **General hypothesis:** Stereo will make people more effective when navigating through a virtual environment (VE).

- **Testable hypothesis:** We measure time it takes for subjects to navigate through a particular VE, under conditions of stereo and mono viewing. We hypothesize subjects will be faster under stereo viewing.

- Testable hypothesis requires a measurable quantity:
  - Time, task completion counts, error counts, etc.

- Some factors effecting experimental validity:
  - Is VE representative of something interesting (e.g., a real-world situation)?
  - Is navigation task representative of something interesting?
  - Is there an underlying theory of human performance that can help predict the results? Could our results contribute to this theory?
What Are the Possible Alternatives?

• Let time to navigate be $\mu_s$: stereo time; $\mu_m$: mono time
  – Perhaps there are two populations: $\mu_s - \mu_m = d$
  – Perhaps there is one population: $\mu_s - \mu_m = 0$
Hypothesis Testing Procedure

1. Develop testable hypothesis $H_1: \mu_s - \mu_m = d$
   - (E.g., subjects faster under stereo viewing)

2. Develop null hypothesis $H_0: \mu_s - \mu_m = 0$
   - Logical opposite of testable hypothesis

3. Construct sampling distribution assuming $H_0$ is true.

4. Run an experiment and collect samples; yielding sampling statistic $X$.
   - (E.g., measure subjects under stereo and mono conditions)

5. Referring to sampling distribution, calculate conditional probability of seeing $X$ given $H_0$: $p( X \mid H_0 )$.
   - If probability is low ($p \leq 0.05$, $p \leq 0.01$), we are unlikely to see $X$ when $H_0$ is true. We reject $H_0$, and embrace $H_1$.
   - If probability is not low ($p > 0.05$), we are likely to see $X$ when $H_0$ is true. We do not reject $H_0$. 
Example 1: VE Navigation with Stereo Viewing

1. Hypothesis $H_1$: $\mu_s - \mu_m = d$
   - Subjects faster under stereo viewing.

2. Null hypothesis $H_0$: $\mu_s - \mu_m = 0$
   - Subjects same speed whether stereo or mono viewing.

3. Constructed sampling distribution assuming $H_0$ is true.

4. Ran an experiment and collected samples:
   - 32 subjects, collected 128 samples
   - $X_s = 36.431$ sec; $X_m = 34.449$ sec; $X_s - X_m = 1.983$ sec

5. Calculated conditional probability of seeing 1.983 sec given $H_0$: $p(1.983$ sec $| H_0) = 0.445$.
   - $p = 0.445$ not low, we are likely to see 1.983 sec when $H_0$ is true. We do not reject $H_0$.
   - This experiment did not tell us that subjects were faster under stereo viewing.

Data from [Swan et al. 03]
Example 2: Effect of Intensity on AR Occluded Layer Perception

1. Hypothesis \( H_1: \mu_c - \mu_d = d \)
   - Tested constant and decreasing intensity. Subjects faster under decreasing intensity.

2. Null hypothesis \( H_0: \mu_c - \mu_d = 0 \)
   - Subjects same speed whether constant or decreasing intensity.

3. Constructed sampling distribution assuming \( H_0 \) is true.

4. Ran an experiment and collected samples:
   - 8 subjects, collected 1728 samples
   - \( X_c = 2592.4 \) msec; \( X_d = 2339.9 \) msec; \( X_c - X_d = 252.5 \) msec

5. Calculated conditional probability of seeing 252.5 msec given \( H_0: p(252.5 \text{ msec} \mid H_0) = 0.008. \)
   - \( p = 0.008 \) is low \( (p \leq 0.01) \); we are unlikely to see 252.5 msec when \( H_0 \) is true. We reject \( H_0 \), and embrace \( H_1 \).
   - This experiment suggests that subjects are faster under decreasing intensity.

Data from [Living et al. 03]
Some Considerations…

• The conditional probability \( p(X \mid H_0) \)
  – Much of statistics involves how to calculate this probability; source of most of statistic’s complexity
  – Logic of hypothesis testing the same regardless of how \( p(X \mid H_0) \) is calculated
  – If you can calculate \( p(X \mid H_0) \), you can test a hypothesis

• The null hypothesis \( H_0 \)
  – \( H_0 \) usually in form \( f(\mu_1, \mu_2, \ldots) = 0 \)
  – Gives hypothesis testing a double-negative logic: assume \( H_0 \) as the opposite of \( H_1 \), then reject \( H_0 \)
  – Philosophy is that cannot prove something true, but can prove it false
  – \( H_1 \) usually in form \( f(\mu_1, \mu_2, \ldots) \neq 0 \); we don’t know what value it will take, but main interest is that it is not 0
When We Reject $H_0$

• Calculate $\alpha = p( X | H_0 )$, when do we reject $H_0$?
  – In psychology, two levels: $\alpha \leq 0.05$; $\alpha \leq 0.01$
  – Other fields have different values

• What can we say when we reject $H_0$ at $\alpha = 0.008$?
  – “If $H_0$ is true, there is only an 0.008 probability of getting our results, and this is unlikely.”
    • Correct!
  
    – “There is only a 0.008 probability that our result is in error.”
    • Wrong, this statement refers to $p( H_0 )$, but that’s not what we calculated.
  
    – “There is only a 0.008 probability that $H_0$ could have been true in this experiment.”
    • Wrong, this statement refers to $p( H_0 | X )$, but that’s not what we calculated.
When We Don’t Reject $H_0$

• What can we say when we don’t reject $H_0$ at $\alpha = 0.445$?
  – “We have proved that $H_0$ is true.”
  – “Our experiment indicates that $H_0$ is true.”
    • Wrong, statisticians agree that hypothesis testing cannot prove $H_0$ is true.

• Statisticians do not agree on what failing to reject $H_0$ means.
  – Conservative viewpoint (Fisher):
    • We must suspend judgment, and cannot say anything about the truth of $H_0$.
  – Alternative viewpoint (Neyman & Pearson):
    • We “accept” $H_0$, and act as if it’s true for now…
    • But future data may cause us to change our mind

From [Howell 02], p 99
Hypothesis Testing Outcomes

<table>
<thead>
<tr>
<th>Decision</th>
<th>Reject $H_0$</th>
<th>Don’t reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ false</td>
<td>correct a result! $p = 1 - \beta = \text{power}$</td>
<td>wrong type II error $p = \beta$</td>
</tr>
<tr>
<td>$H_0$ true</td>
<td>wrong type I error $p = \alpha$</td>
<td>correct (but wasted time) $p = 1 - \alpha$</td>
</tr>
</tbody>
</table>

- $\alpha = \Pr( X | H_0 )$, so hypothesis testing involves calculating $\alpha$
- Two ways to be right:
  - Find a result
  - Fail to find a result and waste time running an experiment
- Two ways to be wrong:
  - Type I error: we think we have a result, but we are wrong
  - Type II error: a result was there, but we missed it
When Do We *Really* Believe a Result?

• When we reject $H_0$, we have a result, but:
  – It’s possible we made a type I error
  – It’s possible our finding is not reliable
    • Just an artifact of our particular experiment

• So when do we *really* believe a result?
  – Statistical evidence
    • $\alpha$ level: ($p < .05$, $p < .01$, $p < .001$)
    • Power
  – Meta-statistical evidence
    • Plausible explanation of observed phenomena
      – Based on theories of human behavior: perceptual, cognitive psychology; control theory, etc.
    • Repeated results
      – Especially by others
Hypothesis Testing Means

- Empiricism
- Experimental Validity
- Usability Engineering
- Experimental Design
- Gathering Data
- Describing Data
  - Graphing Data
  - Descriptive Statistics
- Inferential Statistics
  - Hypothesis Testing
  - Hypothesis Testing Means
  - Power
  - Analysis of Variance and Factorial Experiments
Hypothesis Testing Means

• How do we calculate $\alpha = p( X \mid H_0 )$, when $X$ is a mean?
  – Calculation possible for other statistics, but most common for means

• Answer: we refer to a sampling distribution

• We have two conceptual functions:
  – Population: unknowable property of the universe
  – Distribution: analytically defined function, has been found to match certain population statistics
Calculating $\alpha = p( X | H_0 )$ with A Sampling Distribution

- Sampling distributions are analytic functions with area 1
- To calculate $\alpha = p( X | H_0 )$ given a distribution, we first calculate the value $D$, which comes from an equation of the form:

$$D = \frac{\left(\text{size of effect} : f(\bar{X})\right)}{\left(\text{variability of effect} : f(s^2, N)\right)}$$

- $\alpha = p( X | H_0 )$ is equal to:
  - Probability of seeing a value $\geq |D|$
  - $2 \times (\text{area of the distribution to the right of } |D|)$
- If $H_0$ true, we expect $D$ to be near central peak of distribution
- If $D$ far from central peak, we have reason to reject the idea that $H_0$ is true

Represents assumption that $H_0$ true
A Distribution for Hypothesis Testing Means

- The Standard Normal Distribution \((\mu = 0, \sigma = 1)\)

\[
N(X; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}
\]
The Central Limit Theorem

• Full Statement:
  – Given population with \((\mu, \sigma^2)\), the sampling distribution of means drawn from this population is distributed \((\mu, \sigma^2/n)\), where \(n\) is the sample size. As \(n\) increases, the sampling distribution of means approaches the normal distribution.

• Implication:
  – As \(n\) increases, distribution of means becomes normal, regardless of how “non-normal” the population looks.

• How big does \(n\) have to be before means look normally distributed?
  – For very “non-normal” data, \(n \approx 30\).
Central Limit Theorem in Action

Response time data set $A$; $N = 3436$ data points. Data from [Living et al. 03].

Plotting 100 means drawn from $A$ at random without replacement, where $n$ is number of samples used to calculate mean.

• This demonstrates:
  – As number of samples increases, distribution of means approaches normal distribution;
  – Regardless of how “non-normal” the source distribution is!
The $t$ Distribution

- In practice, when $H_0: \mu_c - \mu_d = 0$ (two means come from same population), we calculate $\alpha = p(X | H_0)$ from $t$ distribution, not $Z$ distribution.

- Why? $Z$ requires the population parameter $\sigma^2$, but $\sigma^2$ almost never known. We estimate $\sigma^2$ with $s^2$, but $s^2$ biased to underestimate $\sigma^2$. Thus, $t$ more spread out than $Z$ distribution.

- $t$ distribution parametric: parameter is $df$ (degrees of freedom).

From [Howell 02], p 185
**t-Test Example**

- **Null hypothesis** $H_0$: $\mu_s - \mu_m = 0$
  - Subjects same speed whether stereo or mono viewing.

- Ran an experiment and collected samples:
  - 32 subjects, collected 128 samples
  - $n_s = 64$, $X_s = 36.431$ sec, $s_s = 15.954$ sec
  - $n_m = 64$, $X_m = 34.449$ sec, $s_m = 13.175$ sec

\[
t(126) = \frac{f\left(\overline{X}\right)}{f\left(s^2, N\right)} = \frac{\overline{X}_s - \overline{X}_m}{\sqrt{s^2_{p}\left(\frac{1}{n_s} + \frac{1}{n_m}\right)}} = 0.766, \quad s^2_p = \frac{(n_s - 1)s^2_s + (n_m - 1)s^2_m}{n_s + n_m - 2}
\]

- Look up $t(126) = 0.766$ in a $t$-distribution table: 0.445

- Thus, $\alpha = p(1.983 \text{ sec} \mid H_0) = 0.445$, and we do not reject $H_0$.

Calculation described by [Howell 02], p 202
One- and Two-Tailed Tests

• *t*-Test example is a two-tailed test.
  – Testing whether two means differ, no preferred direction of difference: $H_1: \mu_s - \mu_m = d$, either $\mu_s > \mu_m$ or $\mu_s < \mu_m$
  – E.g. comparing stereo or mono in VE: either might be faster
  – Most stat packages return two-tailed results by default

• One-tailed test is performed when preferred direction of difference: $H_1: \mu_s > \mu_m$
  – E.g. in [Meehan et al. 03], hypothesis is that heart rate & skin conductance will rise in stressful virtual environment
Power

• Empiricism
• Experimental Validity
• Usability Engineering
• Experimental Design
• Gathering Data
• Describing Data
  – Graphing Data
  – Descriptive Statistics

• Inferential Statistics
  – Hypothesis Testing
  – Hypothesis Testing Means
  – Power
  – Analysis of Variance and Factorial Experiments
**Interpreting $\alpha$, $\beta$, and Power**

<table>
<thead>
<tr>
<th>Decision</th>
<th>Reject $H_0$</th>
<th>Don’t reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ false</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$ true</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- If $H_0$ is true:
  - $\alpha$ is probability we make a **type I error**: we think we have a result, but we are wrong

- If $H_1$ is true:
  - $\beta$ is probability we make a **type II error**: a result was there, but we missed it
  - **Power** is a more common term than $\beta$

$$p = 1 - \beta = \text{power}$$

$$p = \alpha$$

$$p = \beta$$

$$p = 1 - \alpha$$

*Don't reject $H_0$*  
*Reject $H_0$*

$H_0$  
$H_1$

$\mu_0$  
$\mu_1$

$\beta$  
$\alpha$

Power $= 1 - \beta$
Increasing Power by Increasing $\alpha$

- Illustrates $\alpha$ / power tradeoff

- Increasing $\alpha$:
  - Increases power
  - Decreases type II error
  - Increases type I error

- Decreasing $\alpha$:
  - Decreases power
  - Increases type II error
  - Decreases type I error
Increasing Power by Measuring a Bigger Effect

• If the effect size is large:
  – Power increases
  – Type II error decreases
  – $\alpha$ and type I error stay the same

• Unsurprisingly, large effects are easier to detect than small effects
Increasing Power by Collecting More Data

- **Increasing sample size (N):**
  - Decreases variance
  - Increases power
  - Decreases type II error
  - \( \alpha \) and type I error stay the same

- There are techniques that give the value of \( N \) required for a certain power level.

- Here, effect size remains the same, but variance drops by half.
Analysis of Variance and Factorial Experiments

• Empiricism
• Experimental Validity
• Usability Engineering
• Experimental Design
• Gathering Data
• Describing Data
  – Graphing Data
  – Descriptive Statistics
• Inferential Statistics
  – Hypothesis Testing
  – Hypothesis Testing Means
  – Power
  – Analysis of Variance and Factorial Experiments
ANOVA: Analysis of Variance

• *t*-test used for comparing two means
  – *(2 x 1 designs)*

• ANOVA used for factorial designs
  – Comparing multiple levels *(n x 1 designs)*
  – Comparing multiple independent variables *(n x m, n x m x p)*, etc.
  – Can also compare two levels *(2 x 1 designs)*; ANOVA can be considered a generalization of a *t*-Test

• No limit to experimental design size or complexity

• Most widely used statistical test in psychological research

• ANOVA based on the *F* Distribution; also called an *F*-Test
How ANOVA Works

- Null hypothesis $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; $H_1$: at least one mean differs.
- Estimate variance between each group: $MS_{\text{between}}$
  - Based on the difference between group means.
  - If $H_0$ is true, accurate estimation.
  - If $H_0$ is false, biased estimation: overestimates variance.
- Estimate variance within each group: $MS_{\text{within}}$
  - Treats each group separately.
  - Accurate estimation whether $H_0$ is true or false.
- Calculate $F$ critical value from ratio: $F = MS_{\text{between}} / MS_{\text{within}}$
  - If $F \approx 1$, then accept $H_0$.
  - If $F >> 1$, then reject $H_0$.
ANOVA Uses The $F$ Distribution

- Calculate $\alpha = p( X | H_0 )$ by looking up $F$ critical value in $F$-distribution table
- $F$-distribution parametric: $F$ (numerator $df$, denominator $df$)
- $\alpha$ is area to right of $F$ critical value (one-tailed test)
- $F$ and $t$ are distributions are related: $F(1, q) = t(q)$

From [Saville Wood 91], p 52, and [Devore Peck 86], p 563
ANOVA Example

• Hypothesis $H_1$:
  – Platform (Workbench, Desktop, Cave, or Wall) will affect user navigation time in a virtual environment.

• Null hypothesis $H_0$: $\mu_b = \mu_d = \mu_c = \mu_w$.
  – Platform will have no effect on user navigation time.

• Ran 32 subjects, each subject used each platform, collected 128 data points.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between (platform)</td>
<td>1205.8876</td>
<td>3</td>
<td>401.9625</td>
<td>3.100*</td>
<td>0.031</td>
</tr>
<tr>
<td>Within (P x S)</td>
<td>12059.0950</td>
<td>93</td>
<td>129.6677</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\*$p < .05$

• Reporting in a paper: $F(3, 93) = 3.1$, $p < .05$

Data from [Swan et al. 03], calculations shown in [Howell 02], p 471
Main Effects and Interactions

- **Main Effect**
  - The effect of a single independent variable
  - In previous example, a *main effect* of platform on user navigation time: users were slower on the Workbench, relative to other platforms

- **Interaction**
  - Two or more variables interact
  - Often, a 2-way interaction can describe main effects

From [Howell 02], p 431
Example of an Interaction

• Main effect of drawing style:
  – $F(2,14) = 8.84, p < .01$
  – Subjects slower with wireframe style

• Main effect of intensity:
  – $F(1,7) = 13.16, p < .01$
  – Subjects faster with decreasing intensity

• Interaction between drawing style and intensity:
  – $F(2,14) = 9.38, p < .01$
  – The effect of decreasing intensity occurs only for the wireframe drawing style; for fill and wire+fill, intensity had no effect
  – This completely describes the main effects discussed above

Data from [Living et al. 03]
Reporting Statistical Results

• For parametric tests, give degrees of freedom, critical value, \( p \) value:
  - \( F(2,14) = 8.84^*, \ p < .01 \) (report pre-planned significance value)
  - \( t(8) = 4.11, \ p = .0034 \) (report exact \( p \) value)
  - \( F(8,12) = 5.826403, \ p = 3.4778689e10^{-3} \)
    \( \text{(too many insignificant digits)} \)

• Give primary trends and findings in graphs
  – Best guide is [Tufte 83]

• Use graphs / tables to give data, and use text to discuss what the data means
  – Avoid giving too much data in running text
References


Contact Information

J. Edward Swan II, Ph.D.
The Naval Research Laboratory
Washington, DC
swan@acm.org
202/404-4984