**Study Design**

**What is involved (generally)**

- Formulating a hypothesis
- Designing a test plan
  - Picking a method
  - Selecting users
  - Writing out procedure
- Getting IRB permission
- Dealing with users
- Dealing with data
Hypothesis Testing

What is a Hypothesis?

Why do hypothesis testing?

• What kinds of hypotheses do we usually test?
  – X is better/larger/faster than Y
  – X improved more than Y
Goals of Good Study Design

• Testing a hypothesis
  – Clearly state a research question
  – Identify a method for measurement of a single dependent variable as a number of independent variables are manipulated
  – Control for all other extraneous variables if possible

• Issues to deal with:
  – Selectivity of study method
  – Fidelity of research method
  – Quantitative vs. Qualitative data
  – Control vs Ecological validity
  – Cost vs. Relevance

Hypothesis Testing

1. Specify null hypothesis (H0) and the alternative hypothesis (H1). For example H0: \( \mu_1 - \mu_2 = 0 \) and H1: \( \mu_1 \neq \mu_2 \).

   Remember, define so that H1=true iff H0=false

2. Select a significance level. Typically P=0.05 or P=0.01
3. Sample population and calculate statistics.
4. Calculate the probability (p-value) of obtaining a statistic as different or more from the norm.
5. Compared the p-value with the chosen significance level. If P<=Significance level, reject null hypothesis

   That P is greater does not mean H0 is true! Just means we can't tell that it is false.

   When the null hypothesis is rejected, the result is "statistically significant". When we can't reject null hypothesis, result is not statistically significant
Hypothesis Testing

1. Specify null hypothesis (H0) and the alternative hypothesis (H1). For example H0: \( \mu_1 - \mu_2 = 0 \) and H1: \( \mu_1 \neq \mu_2 \).
   
   Remember, define so that H1=true iff H0=false

2. Select a significance level. Typically \( P=0.05 \) or \( P=0.01 \)

3. Sample population and calculate statistics.

4. Calculate the probability (p-value) of obtaining a statistic as different or more from the norm.

5. Compared the p-value with the chosen significance level. If \( P \leq \) significance level, reject null hypothesis
   
   That P is greater does not mean H0 is true! Just means we can't tell that it is false.

When the null hypothesis is rejected, the result is "statistically significant". When we can't reject null hypothesis, result is not statistically significant.

Determining significance

- Significance = threshold at which occurrence was seen as due to some force rather than random chance

- Different thresholds for different fields, depends on stakes
  - CS & many others: \( P=0.05 \) is significant, \( P=0.1 \) is sometimes referred to as "marginal significance" (CS only)
  - Medicine, psychology, etc: Anything over 0.01 or 0.005 is not generally accepted for any treatment related experiment

- Be Aware: Most of our statistical methods assume normally distributed populations and random samples
**Determining Significance**

- Standard deviation ($\sigma$) = $\sqrt{\text{variance}}$
  - Variance = average squared deviation from mean

- In a normal distribution, 68% of scores are within one standard deviation and 95% within two standard deviations
  - This is why CS and others arrived at $p=0.05$, easy to calculate without computers; we used to have to look things up in tables

**Disclaimers**

- Difference between HCI and Psychology / Medicine / Statistics
  - Probabilities
  - Sample size
  - Effect size
  - Study duration
  - Replication
Statistical Significance

• Statistical significance means
  – Two populations differ along some variable, not necessarily all, to a degree that it is likely not caused by chance.
Statistical Significance

• Statistical significance means
  – Two populations differ to a significant extent along some variable, not necessarily all.

• Typically in either the rate of occurrence, or the avg. value of some result. Can have one without the other…?
  – Group A can be twice as likely to do well on tests than Group B (statistically significant), yet the difference in scores may not be large enough to be significant

• Statistical significance does NOT necessarily mean noteworthy
  – It may be the case that CS grad students are better at math than the average grad student (unlikely), but that this difference is so small as to have no practical significance in real life

Significance: Type I and II Errors

• What does significance mean?
  – Type I error – False positive
  – Type II error – False negative
  – Type I error rate set by significance threshold
  – Type II error rate grows as Type I shrinks

• Set significance to balance risks of type I or type II error
  – When might low Type I and high Type II be preferable?
  – When might high Type I and low Type II be preferable?

<table>
<thead>
<tr>
<th>Action</th>
<th>True State</th>
<th>H0 False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject H0</td>
<td>Type I error</td>
<td>Correct</td>
</tr>
<tr>
<td>Not reject H0</td>
<td>Correct</td>
<td>Type II Error</td>
</tr>
</tbody>
</table>
What does it mean when you do not get significance?

Testing Assumptions and Sampling

Most of our statistical tests make a number of assumptions:

– Normally distributed population
– Random sampling
– Single test

What happens if these assumptions are violated?
Sample sizes

- Two factors determine how big a sample you need:
  - Power:
    - The likelihood we’ll have enough data to reliably reject H0, given a difference
    - Power = (1 - Probability of Type II error)
  - Central Limit Theorem:
    - Central Limit Theory: A large sample of a non-normal population will tend to approach normal distribution
    - Most statistical tests assume a normally distributed sample
    - The more the overall populations departs from normality, the larger the sample size needed
  - Advice: 10-20 per condition, the more the better

Within-subject vs. Between-subject Design

- Repeated measures vs. single sample (or low number of samples)

- Are we testing whether two groups are different (between subjects), or whether a treatment had an effect (within subject)?
  - Between subjects we typically look at population averages
  - Within subjects we typically look at the average change in subjects (analysis of variance)
**Within-subject or Between-subject Design**

- Within-subject design
  - Cheap, fewer subjects, more data
  - Removes individual differences
  - Introduces learning and carryover effects
  - Can’t use the same tests as on between subjects because the observations are no longer independent

**Hypothesis Testing**

1. Specify null hypothesis (H0) and the alternative hypothesis (H1). For example H0: \( \mu_1 - \mu_2 = 0 \) and H1: \( \mu_1 \neq \mu_2 \).
   
   Remember, define so that H1=true iff H0=false

2. Select a significance level. Typically \( P=0.05 \) or \( P=0.01 \)
3. Sample population and calculate statistics.
4. Calculate the probability (p-value) of obtaining a statistic as different or more from the norm.
5. Compared the p-value with the chosen significance level. If \( P<\text{Significance level} \), reject null hypothesis

That \( P \) is greater does not mean H0 is true! Just means we can’t tell that it is false.

When the null hypothesis is rejected, the result is “**statistically significant**” When we can’t reject null hypothesis, result is **not statistically significant**
Basic Stats Concepts

• The variance is a measure of how spread out a distribution is. It is the average squared deviation of each number from its mean.
• Standard deviation = $\sqrt{\text{variance}}$
• $N =$ degrees of freedom = Observations-conditions
• Standard Error = Standard dev / $\sqrt{N}$

Error Bars

![Error Bars Diagram]
**Student t-test**

- Most basic test used (very handy)
- For single group
  \[ t = \frac{\text{statistic} - \text{hypothesized value}}{\text{Std error}} \]
- For test of proportions
  \[
  z = \frac{p_1 - p_2}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}}
  \]

Where \( p_1 - p_2 \) is the difference between sample proportions, \( p \) is a weighted average of \( p_1 \) and \( p_2 \) and the \( n \)’s are the # of samples

---

**What does a “t” or a “z” mean?**

Look it up in table!

Or use program to calculate probability value

Analysis of variance:
ANOVA

• t-test checks for difference between two means

• Fishing bad, don’t test every single possible combination of factors in a large experiment

• Remember Type I error, something is bound to slip by sooner or later

• Break out your stats program
  – It will calculate your p’s for you, but you still need to figure out what it all means

Dealing with data

• Academic honesty key
  – Falsifiability of theory/hypothesis/results
  – Need for meticulous records
    • Need to keep records unmodified
    • Giving others your data…?
  – Objectivity
  – Peer review
  – Pilots
  – Replication
**Pitfalls (general biases)**

- Biased testing
  - Tests that cannot disprove your hypothesis

- Biased selection
  - Exclude subjects which may not fit your model

- Biased subjects
  - Want to help
  - You may tell them what you want
  - Hawthorn effect

- Biased interpretation
  - “Read” your expectations into data

**Resources**

- “Design and Analysis: A Researcher’s Handbook” by Geoffrey Keppel

- HyperStat Online (http://davidmlane.com/hyperstat/)

- The Little Handbook of Statistical Practice (http://www.tufts.edu/~gdallal/LHSP.HTM)