Human-Centered Fidelity Metrics for Virtual Environment Simulations

Three Numbers from Standard Experimental Design and Analysis: *α, power, effect magnitude*

VR 2005 Tutorial

J. Edward Swan II, Mississippi State University

Outline

- Introduction and Motivation
- Alpha (α):
  - The Logic of Hypothesis Testing
  - Interpreting α; accepting and rejecting H_0
  - VR and AR examples
- Power:
  - Power and hypothesis testing
  - Ways to use power
  - VR and AR examples
- Effect Magnitude:
  - The Logic of ANOVA
  - Calculating η^2 and ω^2
  - VR and AR examples

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 VR / AR for applications
- HS experiments at VR:

VR year	papers	%	sketches	%	posters	%
2003	10 / 29	35%			5 / 14	36%
2004	9 / 26	35%			5 / 23	22%
2005	13 / 29	45%	1 / 8	13%	8 / 15	53%

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

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

Alpha (α)

- Introduction and Motivation
- Alpha (α):
  - The Logic of Hypothesis Testing
  - Interpreting α; accepting and rejecting H_0
  - VR and AR examples
- ***Power***:
  - Power and hypothesis testing
  - Ways to use power
  - VR and AR examples
- ***Effect Magnitude***:
  - The Logic of ANOVA
  - Calculating η^2 and ω^2
  - VR and AR examples

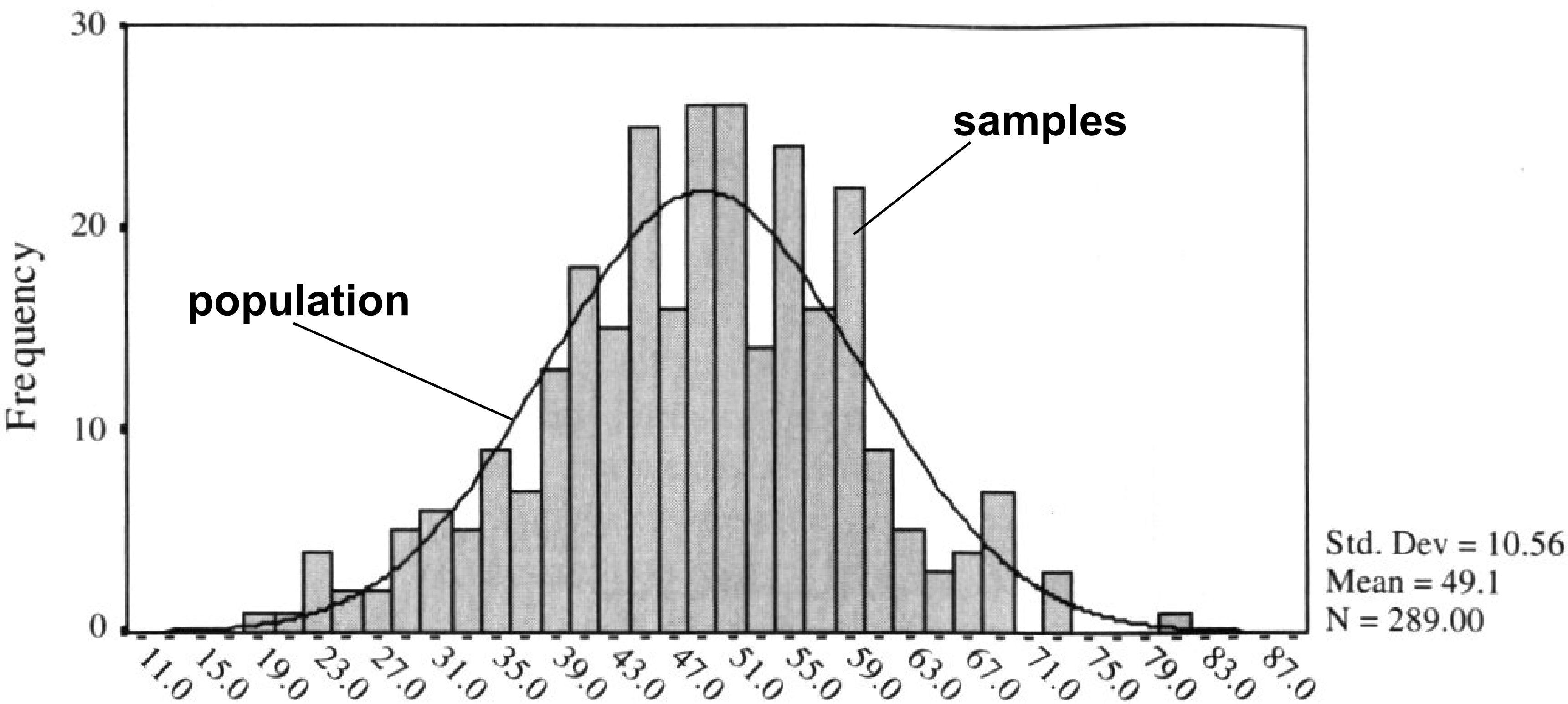
Populations and Samples

- **Population:**
  - Set containing every possible element that we want to measure
  - Usually a Platonic, theoretical construct
  - Mean: μ Variance: σ^2 Standard deviation: σ

- **Sample:**
  - Set containing the elements we actually measure (our subjects)
  - Subset of related population
  - Mean: $\overline{X}$ Variance: s^2 Standard deviation: s Number of samples: N

Hypothesis Testing

- **Goal is to infer population characteristics from sample characteristics**



From [Howell 02], p 78

Testable Hypothesis

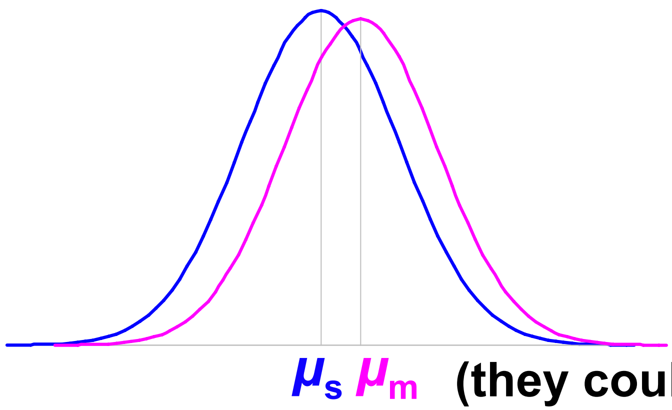
- **General hypothesis: The research question that motivates the experiment.**
- **Testable hypothesis: The research question expressed in a way that can be measured and studied.**
- **Generating a good testable hypothesis is a real skill of experimental design.**
  - **By *good*, we mean contributes to experimental validity.**
  - **Skill best learned by studying and critiquing previous experiments.**

Testable Hypothesis Example

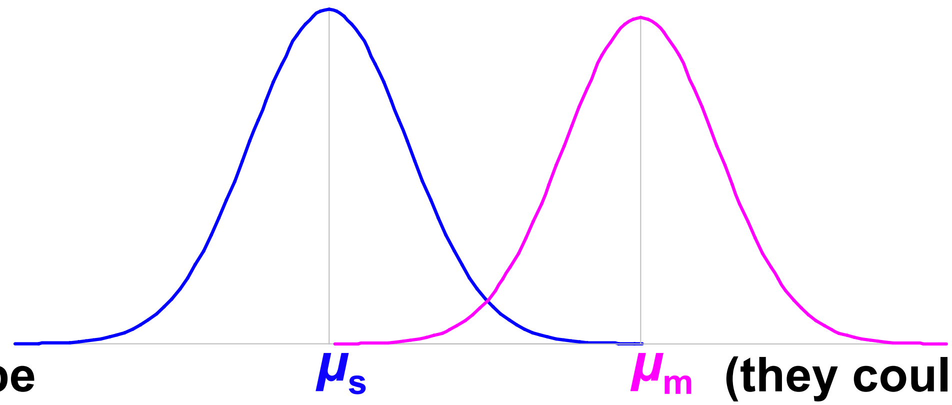
- General hypothesis: Stereo will make people more effective when navigating through a virtual environment (VE).
- Testable hypothesis: We measure time it takes for subjects to navigate through a particular VE, under conditions of stereo and mono viewing. We hypothesis subjects will be faster under stereo 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 μ_s: stereo time; μ_m: mono time
  - Perhaps there are two populations: $\mu_s - \mu_m = d$

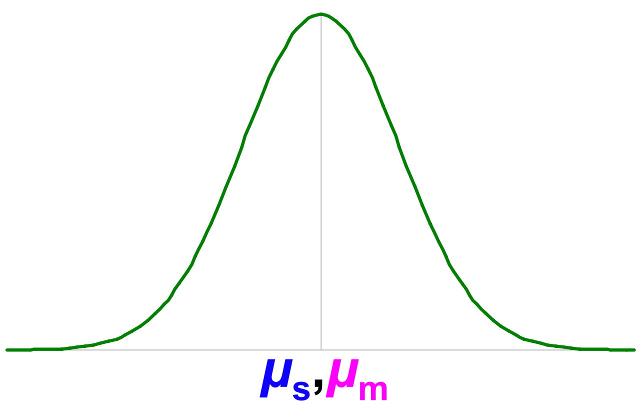


(they could be close together)



(they could be far apart)

  - 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: $\alpha = p(X \mid H_0)$.**
   - **If probability is low ($\alpha \leq 0.05$, $\alpha \leq 0.01$), we are unlikely to see X when H_0 is true. We reject H_0, and embrace H_1.**
   - **If probability is not low ($\alpha > 0.05$), 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: $\alpha = p(\text{ 1.983 sec } | H_0) = 0.445$.**
   - **α = 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: $\alpha = p($ 252.5 msec $| H_0) = 0.008$.
   - $\alpha = 0.008$ is low ($\alpha \leq 0.01$); 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 $\alpha = p(X \mid H_0)$
  - 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 $\alpha = p(X \mid H_0)$ is calculated
  - If you can calculate $\alpha = p(X \mid H_0)$, 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 can never 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 $\alpha = p(X \mid H_0)$, 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 $p(H_0 \mid X)$, 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. (But see the section on Power).**

- **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

		Decision	
		Reject H_0	Don't reject H_0
True state of the world	H_0 false	correct a result! $p = 1 - \beta$ = power	wrong type II error $p = \beta$
	H_0 true	wrong type I error $p = \alpha$	correct (but wasted time) $p = 1 - \alpha$

- $\alpha = p(\, X \mid H_0 \,)$, so hypothesis testing involves calculating α
- 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
    - α level: ($p < .05$, $p < .01$, $p < .001$)
    - power, effect magnitude
  - 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

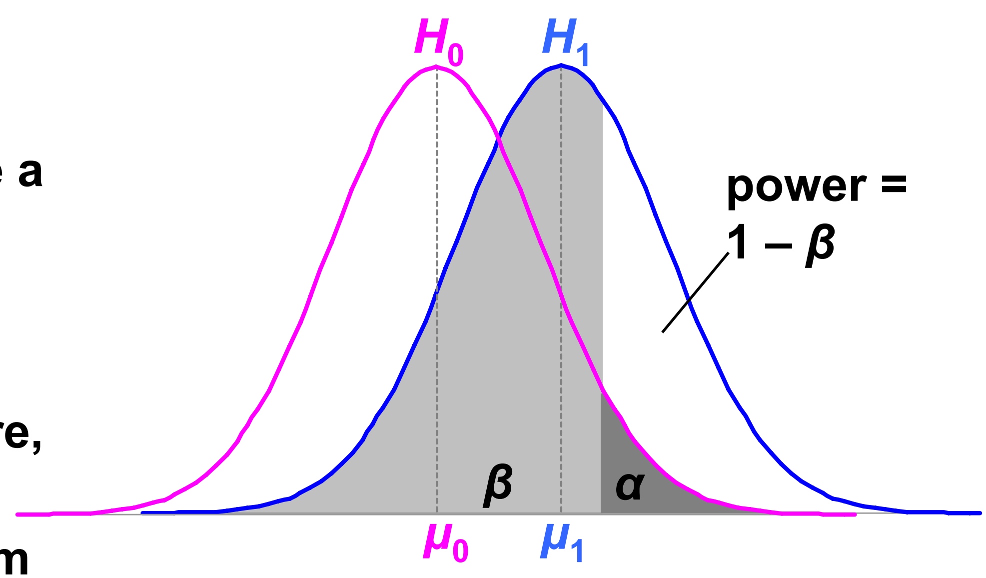
Power

- **Introduction and Motivation**
- **Alpha (α):**
  - The Logic of Hypothesis Testing
  - Interpreting α; accepting and rejecting H_0
  - VR and AR examples
- ***Power*:**
  - Power and hypothesis testing
  - Ways to use power
  - VR and AR examples
- ***Effect Magnitude*:**
  - The Logic of ANOVA
  - Calculating η^2 and ω^2
  - VR and AR examples

Interpreting α, β, and Power

		Decision	
		Reject H_0	Don't reject H_0
True state of the world	H_0 false	a result! $p = 1 - \beta$ = power	type II error $p = \beta$
	H_0 true	type I error $p = \alpha$	wasted time $p = 1 - \alpha$

- If H_0 is true:
  - α is probability we make a type I error: we think we have a result, but we are wrong
- If H_1 is true:
  - β is probability we make a type II error: a result was there, but we missed it
  - Power is a more common term than β

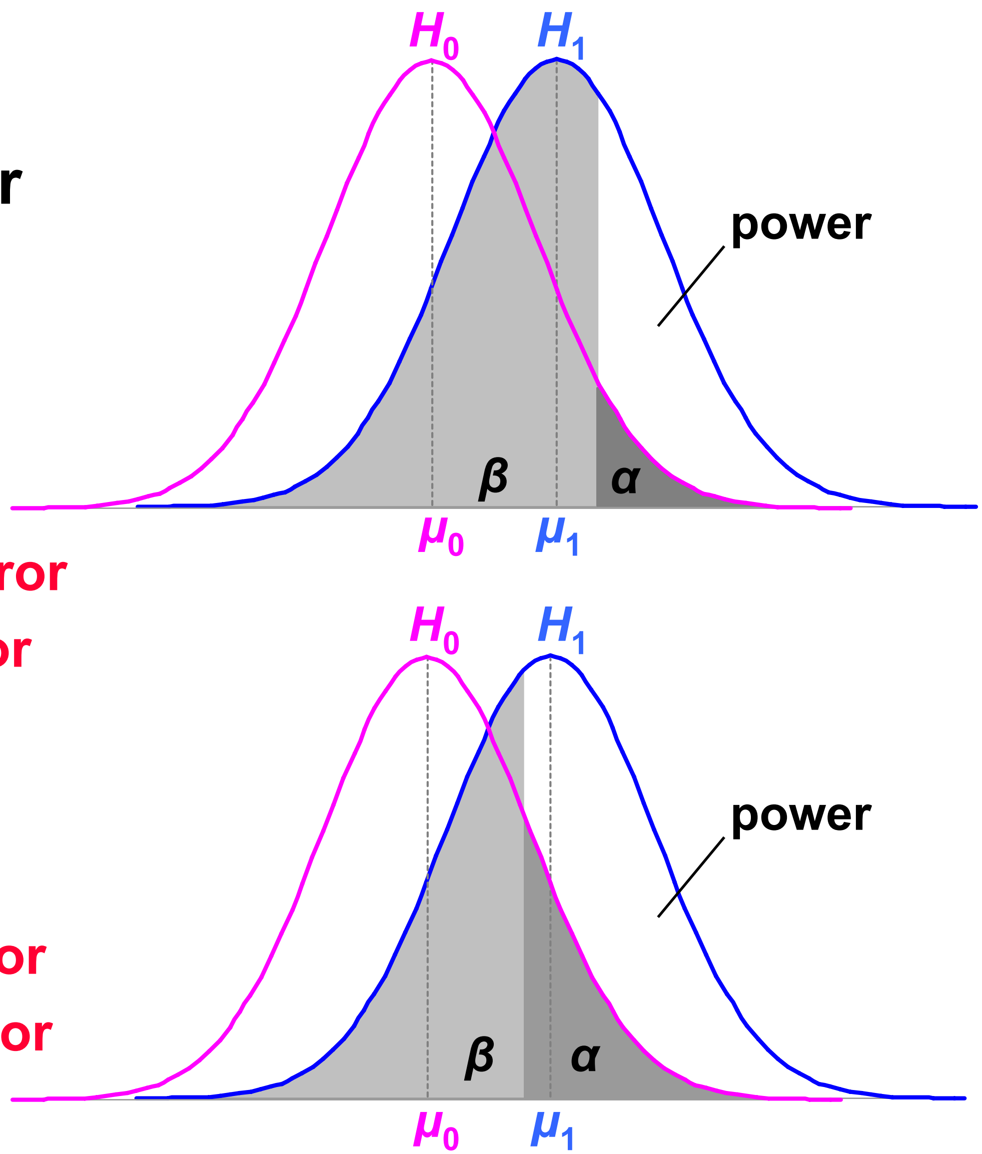


Increasing Power by Increasing α

- **Illustrates α / power tradeoff**

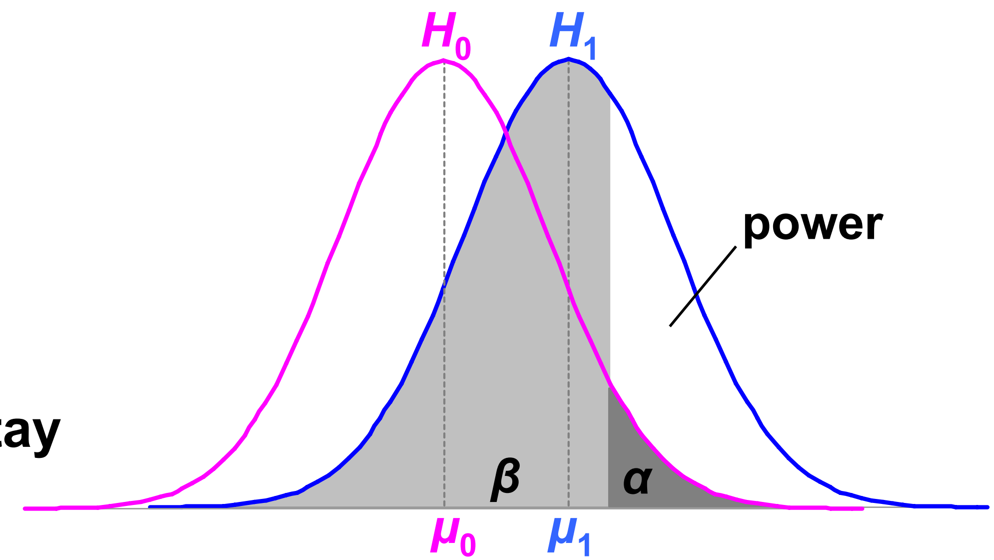
- **Increasing α:**
  - Increases power
  - Decreases type II error
  - Increases type I error

- **Decreasing α:**
  - Decreases power
  - Increases type II error
  - Decreases type I error

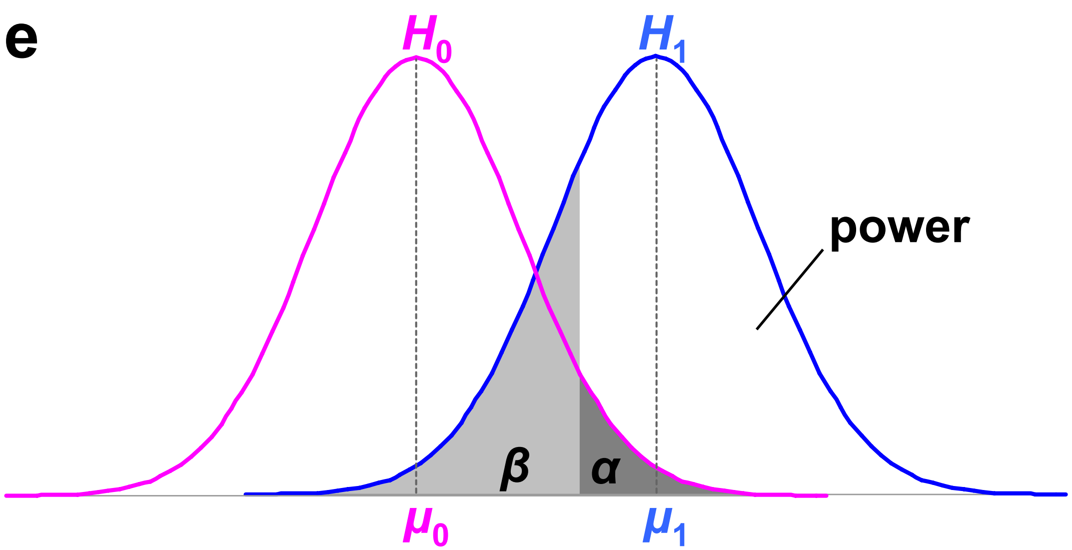


Increasing Power by Measuring a Bigger Effect

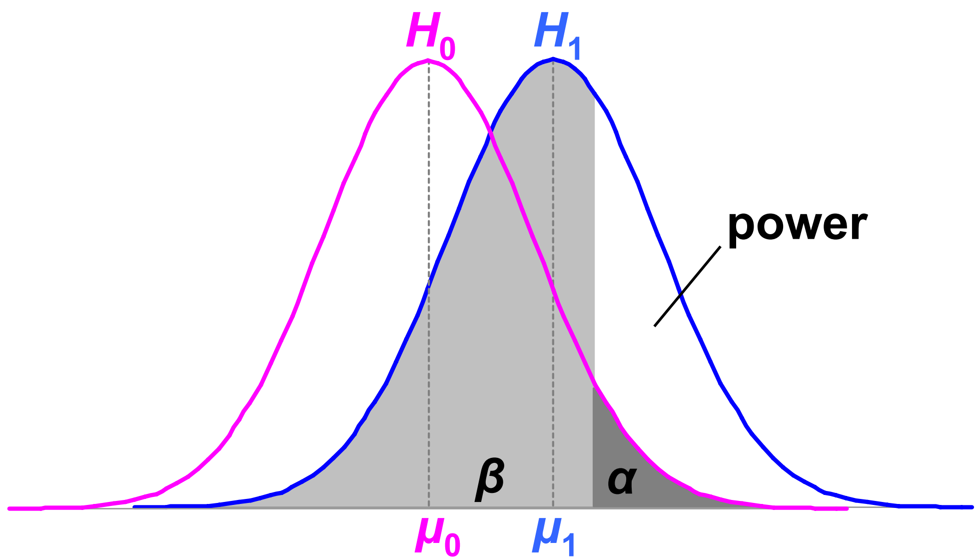
- **If the effect size is large:**
  - Power increases
  - Type II error decreases
  - α and type I error stay the same



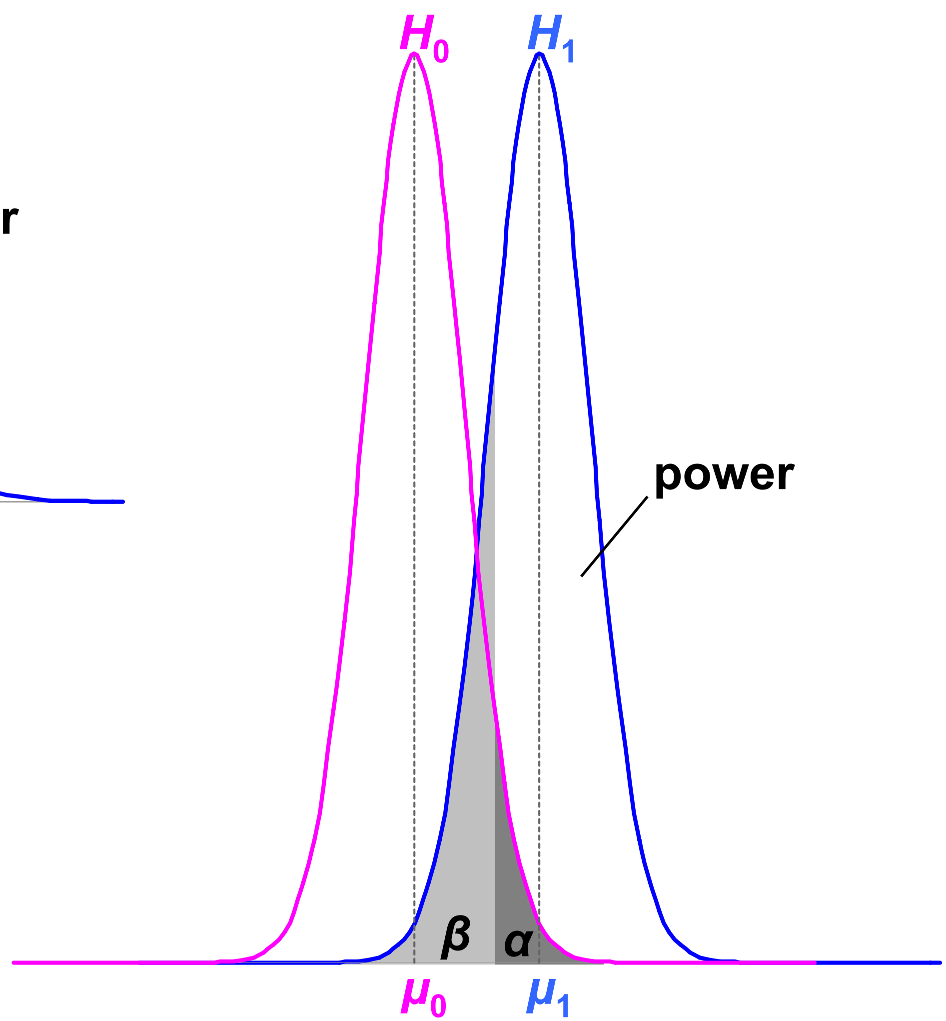
- **Unsurprisingly, large effects are easier to detect than small effects**



Increasing Power by Collecting More Data



- **Increasing sample size (*N*):**
  - **Decreases variance**
  - **Increases power**
  - **Decreases type II error**
  - ***α* 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.**

Power and VR / AR Fidelity Metrics

- **Need *α*, effect size, and sample size for power:**

$$\text{power} = f(\alpha, |\mu_0 - \mu_1|, N)$$

- **Problem for VR / AR:**
  - Effect size $|\mu_0 - \mu_1|$ hard to know in our field
    - Population parameters estimated from prior studies
    - But our field is so new, not many prior studies
  - Can find effect sizes in more mature fields

- **Post-hoc power analysis:**

$$\text{effect size} = |X_0 - X_1|$$

  - Estimate from sample statistics
  - But this makes statisticians grumble (e.g. [Howell 02] [Cohen 88])

Other Uses for Power

1. Number samples needed for certain power level:

   $$N = f(\text{ power}, \alpha, |\mu_0 - \mu_1| \text{ or } |X_0 - X_1|)$$

   – Number extra samples needed for more powerful result
   – Gives "rational basis" for deciding *N* [Cohen 88]

2. Effect size that will be detectable:

   $$|\mu_0 - \mu_1| = f(N, \text{ power}, \alpha)$$

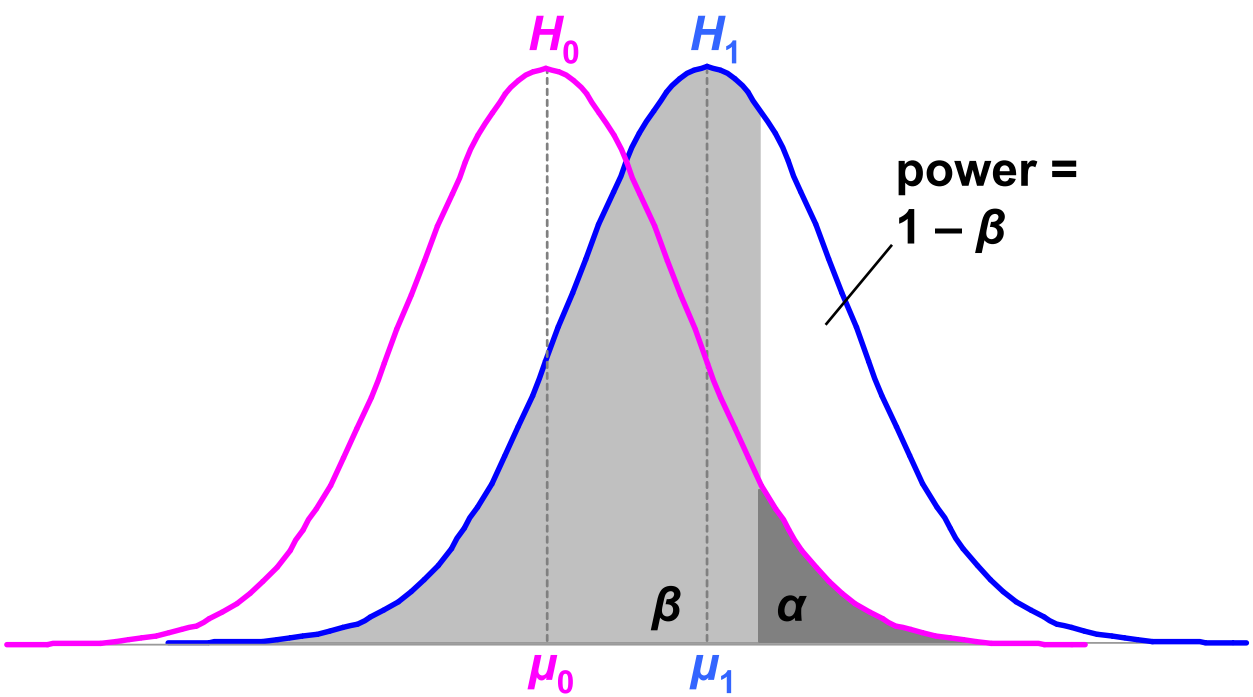
3. Significance level needed:

   $$\alpha = f(|\mu_0 - \mu_1| \text{ or } |X_0 - X_1|, N, \text{ power})$$

(1) is the most common power usage

Arguing the Null Hypothesis

- Cannot directly argue H_0: $\mu_s - \mu_m = 0$. But we can argue that $|\mu_0 - \mu_1| < d$.
  - Thus, we have bound our effect size by *d*.
  - If *d* is *small*, effectively argued null hypothesis.



From [Cohen 88], p 16

Example of Arguing H_0

- We know GP is effective depth cue,
  but can we get close with other graphical cues?

ground plane	drawing style	opacity	intensity	mean error*
on	all levels	both levels	both levels	0.144
off	wire+fill	decreasing	decreasing	0.111

*$F(1,1870) = 1.002$, $p = .317$

- Our effect size is $d = .087$ standard deviations

  power($\alpha = .05$, $d = .087$, $N = 265$) = .17

- Not very powerful. Where can our experiment bound d?

  d($N = 265$, power = .95, $\alpha = .05$) = .31 standard deviations

- This bound is significant at $\alpha = .05$, $\beta = .05$, using same logic as hypothesis testing.
  But how meaningful is $d < .31$? Other significant d's:

  .37, .12, .093, .19

- Not very meaningful. If we ran an experiment to bound $d < .1$, how much data would we need?

  N(power = .95, $\alpha = .05$, $d = .1$) = 2600

- Original study collected $N = 3456$, so $N = 2600$ reasonable

Data from [Living et al. 03]

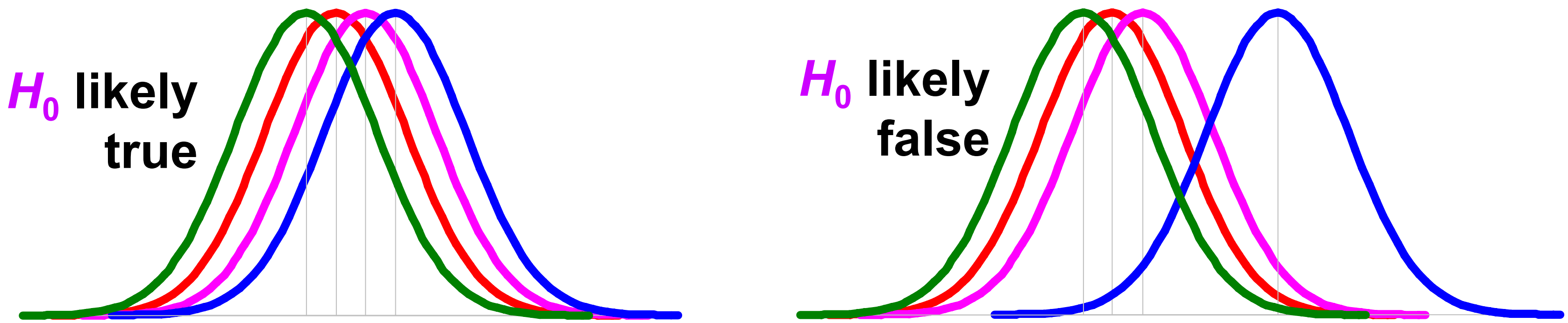
Effect Magnitude

- **Introduction and Motivation**
- **Alpha (α):**
  - The Logic of Hypothesis Testing
  - Interpreting α; accepting and rejecting H_0
  - VR and AR examples
- ***Power*:**
  - Power and hypothesis testing
  - Ways to use power
  - VR and AR examples
- ***Effect Magnitude*:**
  - The Logic of ANOVA
  - Calculating η^2 and ω^2
  - VR and AR examples

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

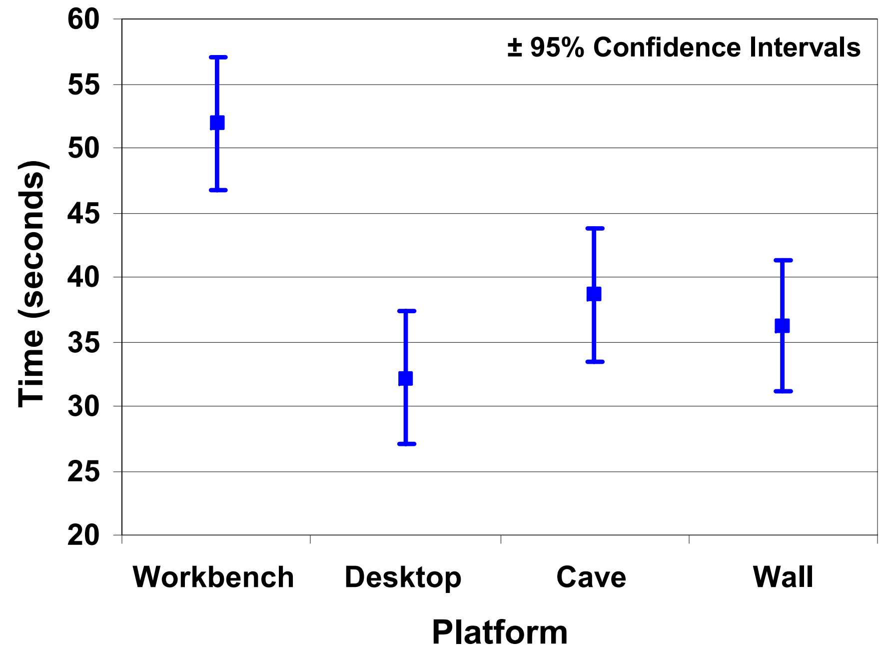
How ANOVA Works



- Null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; H_1: at least one mean differs
- Estimate variance between each group: $MS_{between}$
  - 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_{within}
  - Treats each group separately
  - Accurate estimation whether H_0 is true or false
- Calculate F critical value from ratio: $F = MS_{between} / MS_{within}$
  - If $F \approx 1$, then accept H_0
  - If $F >> 1$, then reject H_0

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.



Source	SS	*df*	MS	*F*	*p*
Between (platform)	1205.8876	3	401.9625	3.100*	0.031
Within (P x S)	12059.0950	93	129.6677		

*$p < .05$

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

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

Measures of Effect Magnitude

- **Hypothesis Testing with ANOVA gives us:**
  - ***α*: measures *effect significance***
- **From ANOVA table, can calculate measures of *effect magnitude***
  - **Related to effect size *d* from power analysis**
- **Many calls for reporting *effect magnitude* in addition to *α*:**
  - **Current statistics textbooks**
  - **American Psychological Association**
  - **Many journals and other venues**
- **Related to considering / controlling both:**
  - **Probability of type I error (*α*)**
  - **Probability of type II error (*β*)**

Calculating η^2

- η^2 (eta-squared):
  - Percentage of variance accounted for by an effect
  - Ratio of $SS_{between}$ / SS_{within} :

Source	SS	*df*	MS	*F*	*p*
Between (platform)	1205.8876	3	401.9625	3.100*	0.031
Within (P x S)	12059.0950	93	129.6677		

- η^2 = .100
  - Platform accounts for 10% of observed variance

- Calculate by putting ANOVA table in spreadsheet
  - η^2 not given by Minitab
  - η^2 not given by SPSS (but it gives *partial-* η^2 and calls it η^2!)

Data from [Swan et al. 03]

Calculating ω^2

- ω^2 (omega-squared):
  - Percentage of variance accounted for by an effect
  - Better than η^2: η^2 is biased; ω^2 is less biased

    $\omega^2 = f($ various MS measures, various *df* measures $)$
  - *f* depends on ANOVA design (fixed, random, mixed)
- Generally ω^2 preferred over η^2
- *However*:
  - ω^2 not computable for within-subject, repeated-measures designs
    - Each subject sees multiple levels of independent variables
  - This describes most low-level, perceptual, psychophysical studies
    - E.g., fidelity metrics
  - Therefore η^2 still very useful

From [Howell 02], p 446

Example of using η^2

- When deciding what effects are important:
  - Consider α (e.g., $\alpha \leq .05$), and consider η^2 (e.g., $\eta^2 \geq 1\%$)

- In repeated-measures experiments, factorial designs can give "*spurious*" *n*-way interactions
  - Arise because large *df* in denominator of *F* ratio
  - These effects *significant*, but not *important*

- Example: 3-way interaction below is not in [Gabbard et al. 05], because of low η^2 value

Source	SS	*df*	MS	*F*	*p*	η^2
Distance x Background x Drawing Style	22423249	50	448465	1.5*	0.016	.83%
Within (D x B x DS x Subject)	254047337	850	298879			

Data from [Gabbard et al. 05]

Human-Centered Fidelity Metrics for Virtual Environment Simulations

Three Numbers from Standard Experimental Design and Analysis: *α, power, effect magnitude*

VR 2005 Tutorial

J. Edward Swan II, Mississippi State University

References

[Cohen 88] J Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition, Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[Devore Peck 86] J Devore, R Peck, *Statistics: The Exploration and Analysis of Data*, West Publishing Co., St. Paul, MN, 1986.

[Gabbard et al. 05] JL Gabbard, JE Swan II, D Hix, RS Schulman, J Lucas, D Gupta, "*An Empirical User-Based Study of Text Drawing Styles and Outdoor Background Textures for Augmented Reality*", Technical Papers, IEEE Virtual Reality 2005, March 12-16, Bonn, Germany.

[Howell 02] DC Howell, *Statistical Methods for Psychology*, 5th edition, Duxbury, Pacific Grove, CA, 2002.

[Living et al. 03] MA Livingston, JE Swan II, JL Gabbard, TH Höllerer, D Hix, SJ Julier, Y Baillot, D Brown, "*Resolving Multiple Occluded Layers in Augmented Reality*", The 2nd International Symposium on Mixed and Augmented Reality (ISMAR '03), October 7–10, 2003, Tokyo, Japan, pages 56–65.

[Saville Wood 91] DJ Saville, GR Wood, *Statistical Methods: The Geometric Approach*, Springer-Verlag, New York, NY, 1991.

[Swan et al. 03] JE Swan II, JL Gabbard, D Hix, RS Schulman, KP Kim, "*A Comparative Study of User Performance in a Map-Based Virtual Environment*", Technical Papers, IEEE Virtual Reality 2003, March 22–26, Los Angeles, California: IEEE Computer Society, 2003, pages 259–266.

[Tufte 83] ER Tufte, *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut, 1983.

Contact Information

J. Edward Swan II, Ph.D.

Associate Professor

Department of Computer Science and Engineering

swan@acm.org

(662)325-7507