



**VisWeek 08**

VIS • INFOVIS • VAST

# **Experimental Design and Analysis for Human-Subject Visualization Experiments**

IEEE Visualization 2008 Tutorial

**J. Edward Swan II, Ph.D.**

Department of Computer Science and Engineering  
Department of Psychology (Adjunct)

Mississippi State University



# Schedule: Tutorial 5



<b>10:30 am – 12:10 pm</b>	<b>100 min</b>	<b>Perceptual Issues for Vis and Eval (Part I)</b>
<b>12:10 pm – 2:00 pm</b>	<b>110 min</b>	<b>Lunch Break</b>
<b>2:00 pm – 2:50 pm</b>	<b>50 min</b>	<b>Perceptual Issues for Vis and Eval (Part II)</b>
<b>2:50 pm – 3:40 pm</b>	<b>50 min</b>	<b>Experimental Design and Analysis (Part I)</b>
<b>3:40 pm – 4:15 pm</b>	<b>35 min</b>	<b>Coffee Break</b>
<b>4:15 pm – 5:55 pm</b>	<b>100 min</b>	<b>Experimental Design and Analysis (Part II)</b>

# Motivation and Goals

- **Course attendee backgrounds?**
- **Studying experimental design and analysis at Mississippi State University:**
  - PSY 3103 Introduction to Psychological Statistics
  - PSY 3314 Experimental Psychology
  - PSY 6103 Psychometrics
  - PSY 8214 Quantitative Methods In Psychology II
  - PSY 8803 Advanced Quantitative Methods
  - IE 6613 Engineering Statistics I
  - IE 6623 Engineering Statistics II
  - ST 8114 Statistical Methods
  - ST 8214 Design & Analysis Of Experiments
  - ST 8853 Advanced Design of Experiments I
  - ST 8863 Advanced Design of Experiments II
- **7 undergrad hours; 30 grad hours; 3 departments!**

# Motivation and Goals

- **What can we accomplish in 150 minutes?**
- **Study subset of basic techniques**
  - I have found these to be the most applicable to visualization evaluation
- **Focus on **intuition** behind basic techniques**
- **Become familiar with basic concepts and terms**
  - Facilitate working with collaborators from psychology, industrial engineering, statistics, etc.

# Why Human Subject (HS) Experiments?

- **Graphics hardware / software more mature**
- **Sophisticated interactive techniques possible**
- **Focus of field:**
  - **Implementing technology → using technology**
  - **Trend at IEEE Virtual Reality, SIGGRAPH**
  - **Called for in *NIH-NSF Visualization Research Challenges Report* [Johnson et al 06]**
- **Increasingly running HS experiments:**
  - **How do humans perceive, manipulate, cognate with CG-mediated information?**
  - **Measure utility of visualizations for application domains**

# Conducting Human-Subject Experiments

- Human subject experiments at IEEE Visualization:

Year	Vis	%	Info Vis	%	VAST	%
2006	8 / 63	13%	2 / 24	8%		
2007	3 / 56	4%	12 / 27	44%		
2008	9 / 50	18%	14 / 27	52%	7 / 21	33%

- Human subject experiments at IEEE Virtual Reality:

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%
2006	12 / 27	44%	2 / 10	20%	1 / 10	10%
2007	9 / 26	35%	3 / 15	20%	5 / 18	28%
2008	21 / 37	68%			9 / 29	31%

# Outline

- *Experimental Validity*
- **Experimental Design**
- **Describing Data**
  - Graphing Data
  - Descriptive Statistics
- **Inferential Statistics**
  - Hypothesis Testing
  - Analysis of Variance
  - Power
- **Graphical Data Analysis**

# The Empirical Method

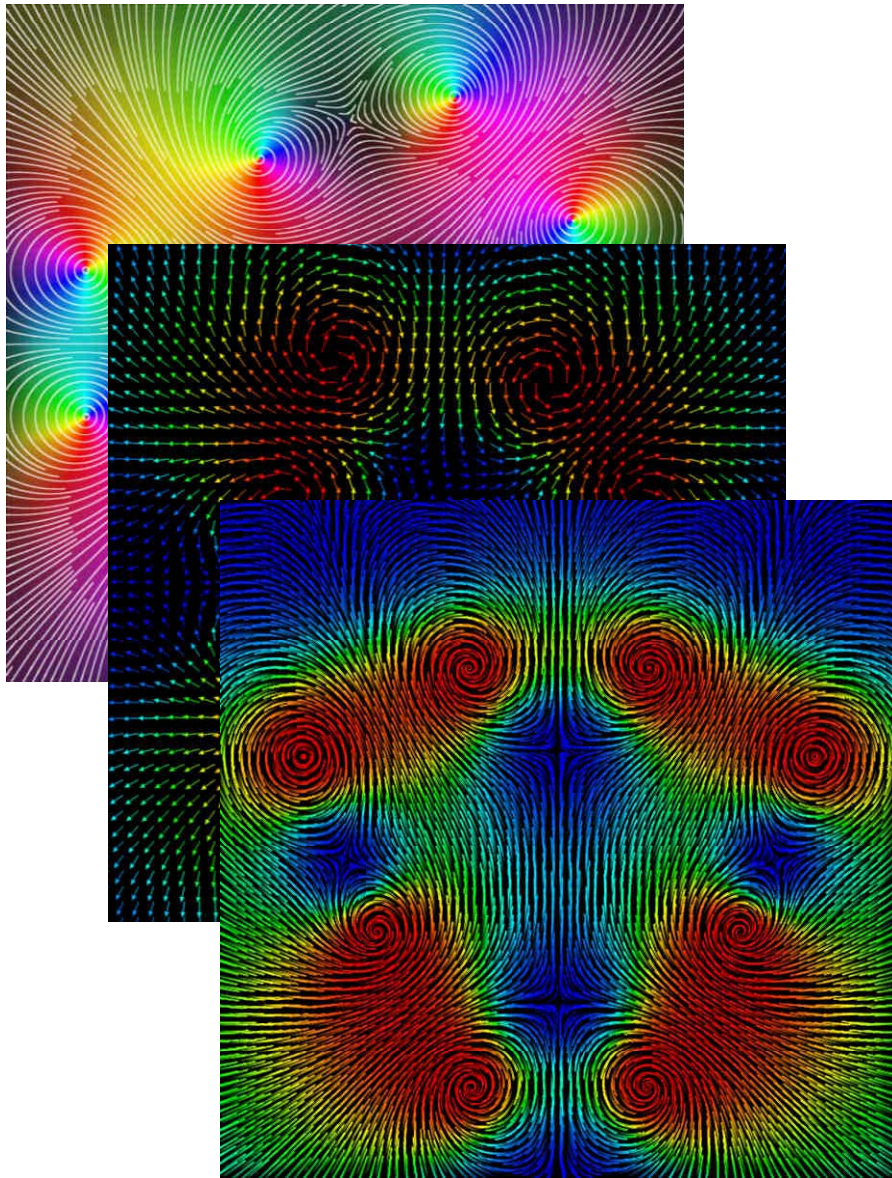
- The *Empirical Method*:
  - Develop a **hypothesis**, perhaps based on a theory
  - Make the hypothesis **testable**
  - Develop an empirical **experiment**
  - Collect and analyze data
  - Accept or refute the hypothesis
  - Relate the results back to the theory
  - If worthy, communicate the results to scientific community
- **Statistics**:
  - Foundation for empirical work; necessary but not sufficient
  - Often not useful for managing problems of **gathering**, **interpreting**, and **communicating** empirical information.



# 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**

# Example of a Validity Issue



- 2D Flow Visualization Experiment
- Tested different visualization methods
- Measured subjects' ability to locate critical points
  - error, response time
- **Validity Issue:**
  - Interested in which visualization method is **most effective**
  - How well does what we measured relate to “**effectiveness?**”

# 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:  $\frac{1}{2}$  of subjects see stereo,  $\frac{1}{2}$  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

# Experimental Design

- **Experimental Validity**
- *Experimental Design*
- **Describing Data**
  - Graphing Data
  - Descriptive Statistics
- **Inferential Statistics**
  - Hypothesis Testing
  - Analysis of Variance
  - Power
- **Graphical Data Analysis**

# Experimental Designs

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

Variable
level 1
level 2

Stereopsis
stereo
mono

- 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

Subjects	1 <sup>st</sup> condition	2 <sup>nd</sup> condition
1, 3, 5, 7	stereo	mono
2, 4, 6, 8	mono	stereo

# Factorial Designs

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

VE terrain type
flat
hilly
mountainous

- **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:

	Stereopsis	
VE terrain type	stereo	mono
flat		
hilly		
mountainous		

# Cells and Repetitions

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

	Stereopsis	
VE terrain type	stereo	mono
flat		
hilly		
mountainous		

cell

- **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

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \\ 3 & 1 & 4 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}$$

$2 \times 2$                        $3 \times 3$                        $4 \times 4$

- 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 that has all 3 properties)

# 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

Subject	Presentation Order
1	1A, 1B, 2A, 2B, 3A, 3B
2	1B, 1A, 2B, 2A, 3B, 3A
3	2A, 2B, 3A, 3B, 1A, 1B
4	2B, 2A, 3B, 3A, 1B, 1A
5	3A, 3B, 1A, 1B, 2A, 2B
6	3B, 3A, 1B, 1A, 2B, 2A
7	1A, 1B, 3A, 3B, 2A, 2B
8	1B, 1A, 3B, 3A, 2B, 2A
9	2A, 2B, 1A, 1B, 3A, 3B
10	2B, 2A, 1B, 1A, 3B, 3A
11	3A, 3B, 2A, 2B, 1A, 1B
12	3B, 3A, 2B, 2A, 1B, 1A

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 & 2 \\ 2 & 1 & 3 \\ 3 & 2 & 1 \end{bmatrix}$$

$$\begin{bmatrix} A & B \\ B & A \end{bmatrix}$$

# Experimental Design Example #1

trial number		1 ..... 216						217 ..... 432					
sv <sup>1</sup>	ground plane	on						off					
	stereo	on			off			on			off		
rp <sup>2</sup>	drawing style	wire				fill				wire+fill			
	alpha	const		decr		const		decr		const		decr	
	intensity	const	decr	const	decr	const	decr	const	decr	const	decr	const	decr
rp <sup>2</sup>	target position	close			middle			far					
	repetition	1	2	3	1	2	3	1	2	3			

<sup>1</sup> sv = systemically varied, <sup>2</sup> rp = randomly permuted

- All variables within-subject

From [Living Swan et al 03]

# Experimental Design Example #2

<b>Between Subject</b>	Stereo Viewing		<i>on</i>				<i>off</i>			
	Control Movement		<i>rate</i>		<i>position</i>		<i>rate</i>		<i>position</i>	
	Frame of Reference		<i>ego</i>	<i>exo</i>	<i>ego</i>	<i>exo</i>	<i>ego</i>	<i>exo</i>	<i>ego</i>	<i>exo</i>
<b>Within Subject</b>	<b>Computer Platform</b>	<i>cave</i>	<i>subjects 1 – 4</i>	<i>subjects 5 – 8</i>	<i>subjects 9 – 12</i>	<i>subjects 13 – 16</i>	<i>subjects 17 – 20</i>	<i>subjects 21 – 24</i>	<i>subjects 25 – 28</i>	<i>subjects 29 – 32</i>
		<i>wall</i>								
		<i>workbench</i>								
		<i>desktop</i>								

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

# Types of Statistics

- **Descriptive Statistics:**
  - Describe and explore data
  - Summary statistics:  
many numbers → 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

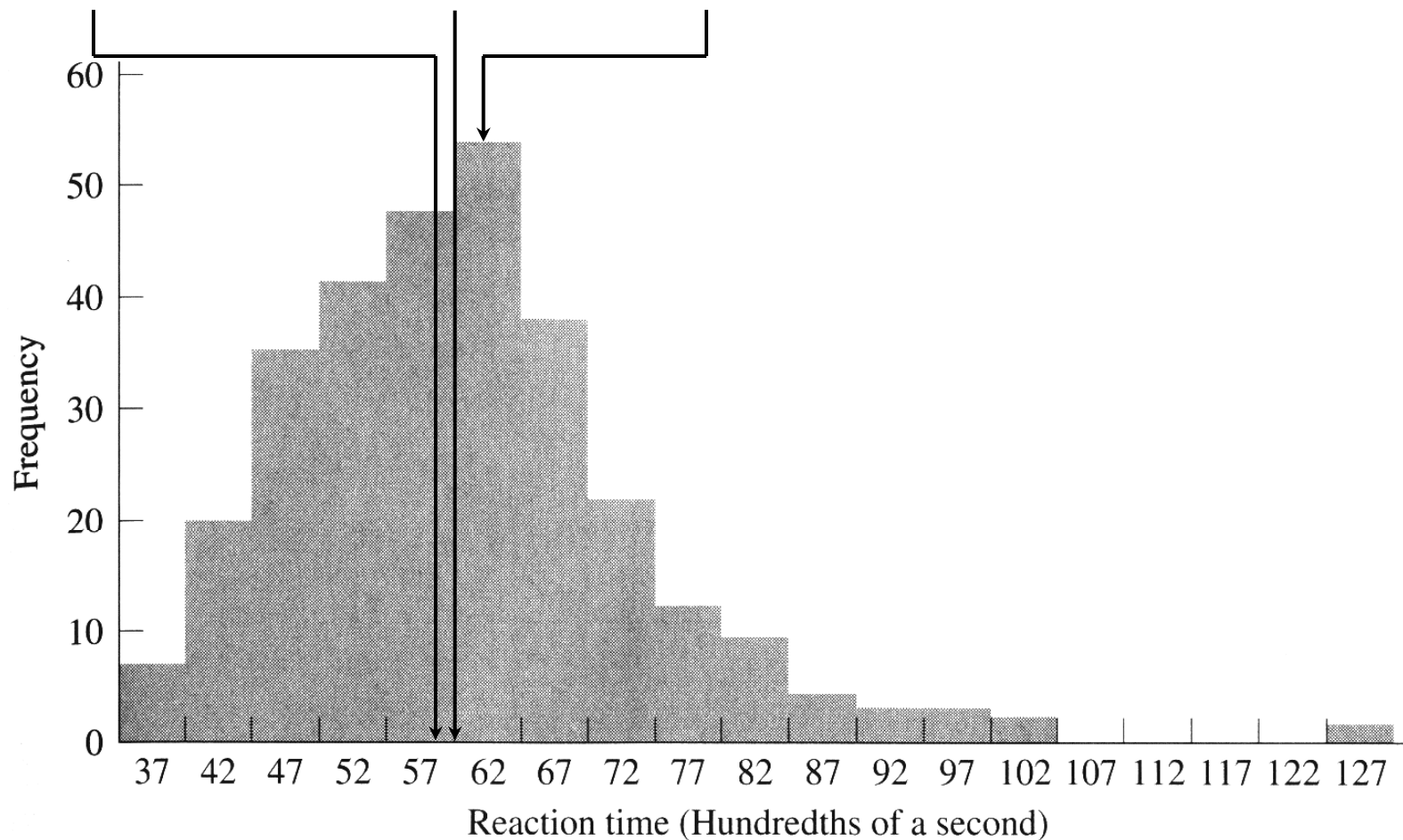
# Graphing Data

- **Experimental Validity**
- **Experimental Design**
- *Describing Data*
  - *Graphing Data*
  - **Descriptive Statistics**
- **Inferential Statistics**
  - **Hypothesis Testing**
  - **Analysis of Variance**
  - **Power**
- **Graphical Data Analysis**

# Exploring Data with Graphs

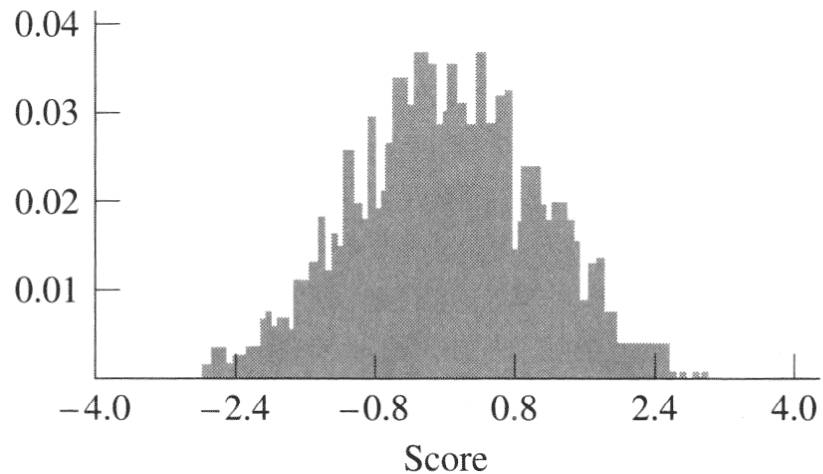
- Histogram common data overview method

median = 59.5    mean = 60.26    mode = 62

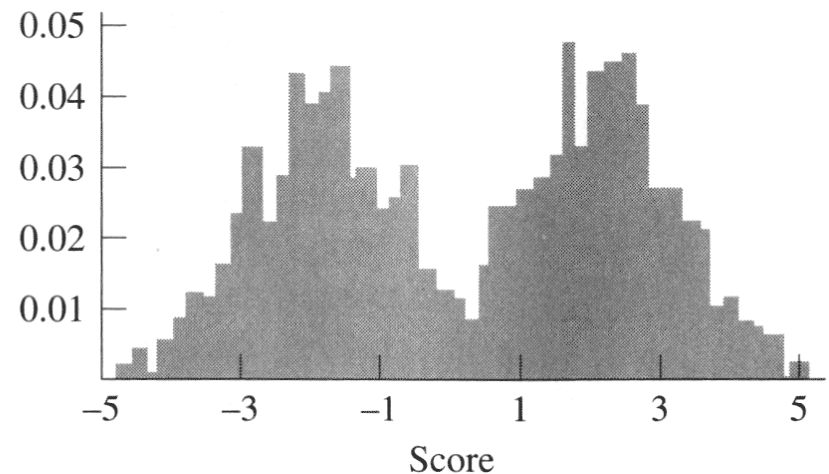


From [Howell 02] p 21

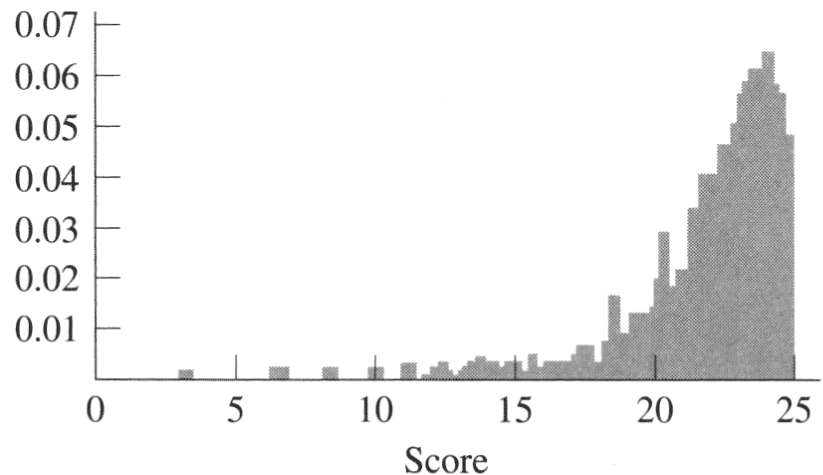
# Classifying Data with Histograms



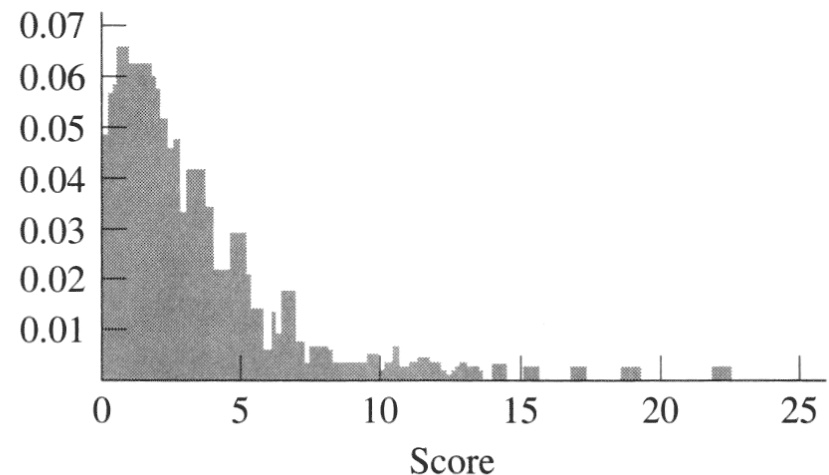
(a) Normal



(b) Bimodal



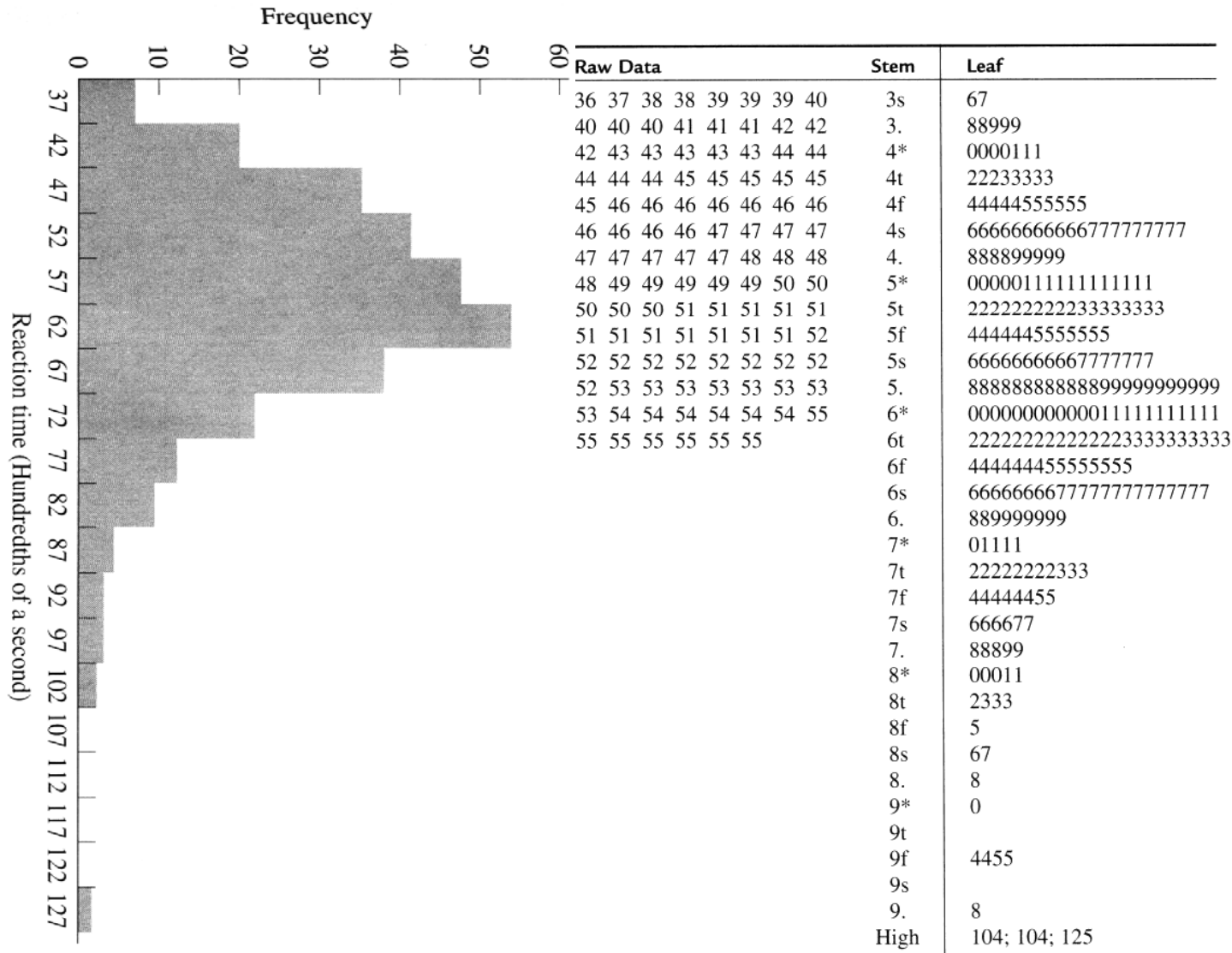
(c) Negatively skewed



(d) Positively skewed



# Stem-and-Leaf: Histogram From Actual Data



From [Howell 02] p 21, 23

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

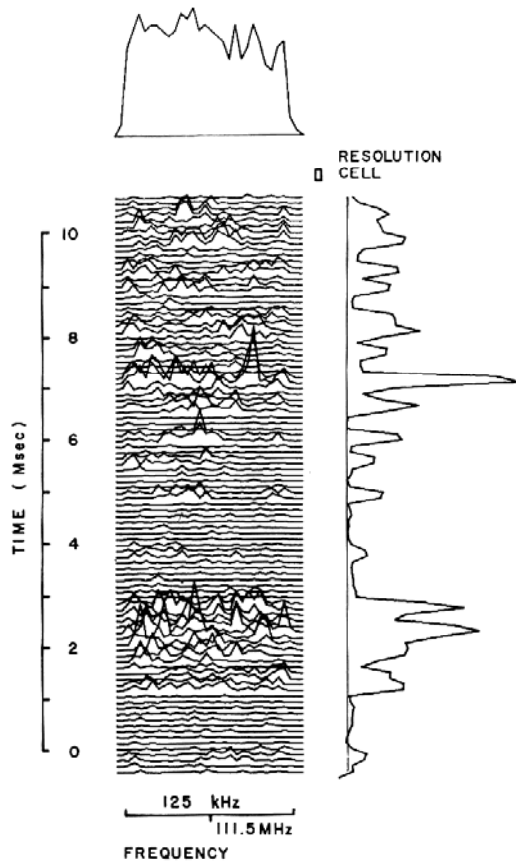
# Stem-and-Leaf: Histogram From Actual Data

## Midterm #1 Grades

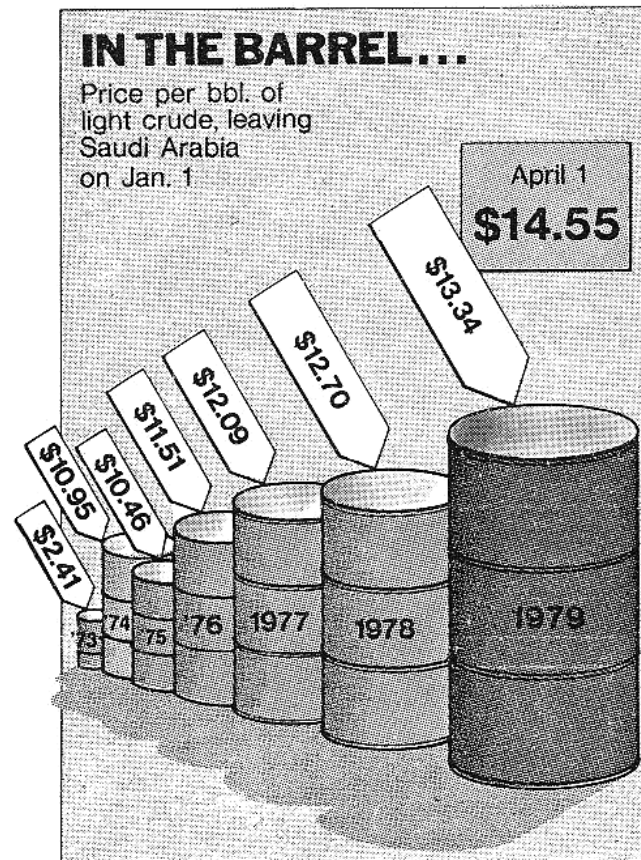
0	0% F	0
0	0% F	1
0	0% F	2
0	0% F	3
1	3% F	4 7
0	0% F	5
4	12% D	6 1789
3	9% C	7 024
10	30% B	8 0014458889
14	42% A	9 01112333455579
0	0% A	10
32		

# 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 02]).



Lots of good examples...



And plenty of bad examples!

From [Tufte 83], p 134, 62

# Descriptive Statistics

- **Experimental Validity**
- **Experimental Design**
- **Describing Data**
  - Graphing Data
  - *Descriptive Statistics*
- **Inferential Statistics**
  - Hypothesis Testing
  - Analysis of Variance
  - Power
- **Graphical Data Analysis**

# Summary Statistics

- **Many numbers → few numbers**
- **Measures of central tendency:**
  - Mean: average
  - Median: middle data value
  - Mode: most common data value
- **Measures of variability / dispersion:**
  - Mean absolute deviation
  - Variance
  - Standard Deviation

# 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:**

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

**Mean absolute deviation:**

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

**Variance:**

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

**Standard deviation:**

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

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

- **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:**

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

- **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



# Example: Degrees of Freedom

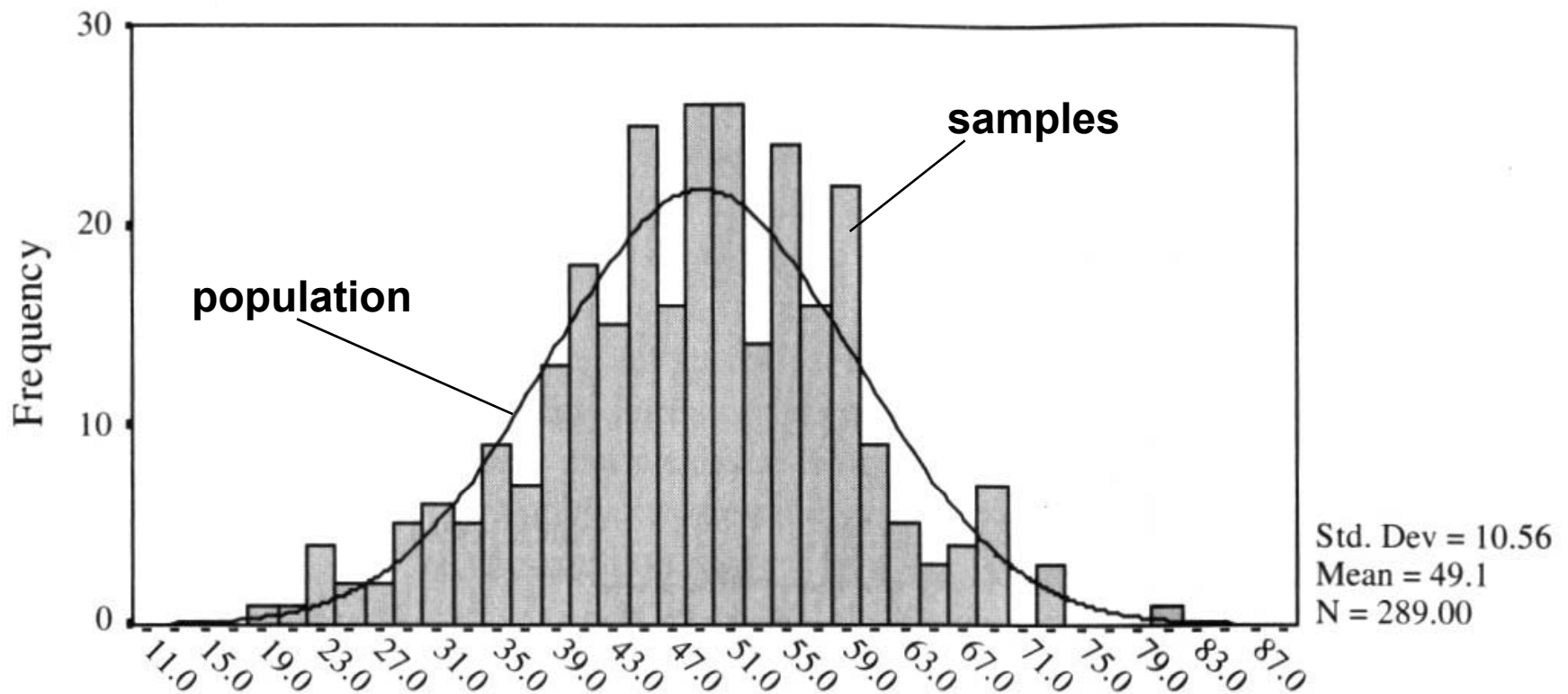
- **Samples:  $\{6, 8, 10\}$ ;  $N = 3$ ;  $X = 8$**
- **If mean must remain  $X = 8$ ;  
how many numbers may vary?**
- **Answer: 2 may vary**
  - **Example: let  $6 \rightarrow 4$ ,  $8 \rightarrow 14$ , then  $(4 + 14 + a)/3 = 8$**
  - **$a = 6$  if  $X = 8$ ; value of  $a$  is constrained**
- **We say that set  $\{6, 8, 10\}$  has  
 $N - 1 = 2$  degrees of freedom (*dof, df*)**
  - **Generally equal to  $N$  minus 1 per mean calculated**

# Hypothesis Testing

- **Experimental Validity**
- **Experimental Design**
- **Describing Data**
  - Graphing Data
  - Descriptive Statistics
- ***Inferential Statistics***
  - *Hypothesis Testing*
  - Analysis of Variance
  - Power
- **Graphical Data Analysis**

# Hypothesis Testing

- Goal is to infer population characteristics from sample characteristics



# Testable Hypothesis

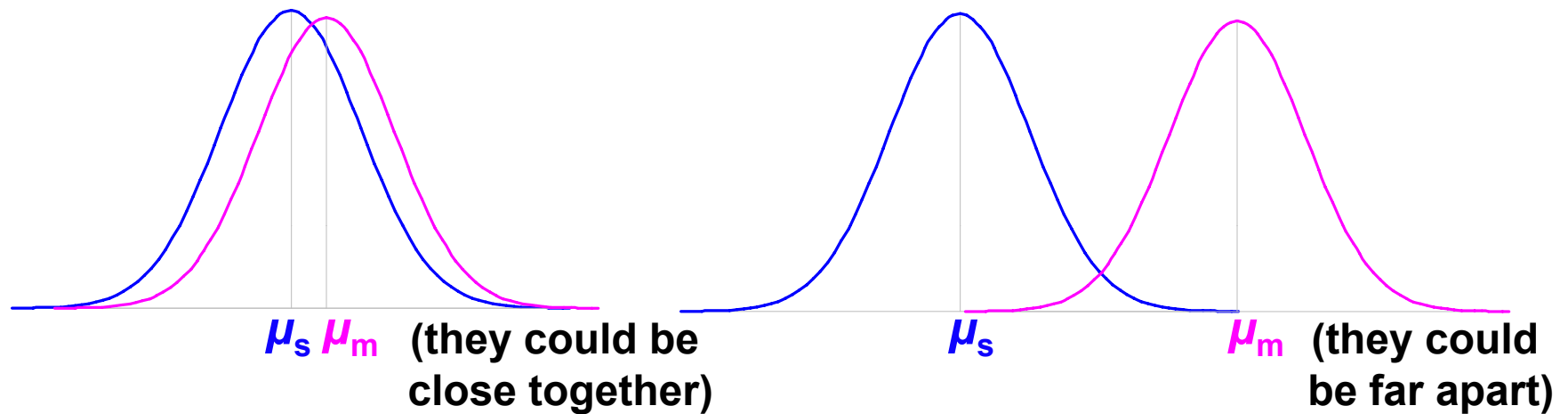
- **General hypothesis:** The research question that motivates the experiment.
- **Testable hypothesis:** The research question expressed in a way that can be measured and studied.
- **Generating a good testable hypothesis is a real skill of experimental design.**
  - By **good**, we mean contributes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

# Testable Hypothesis Example

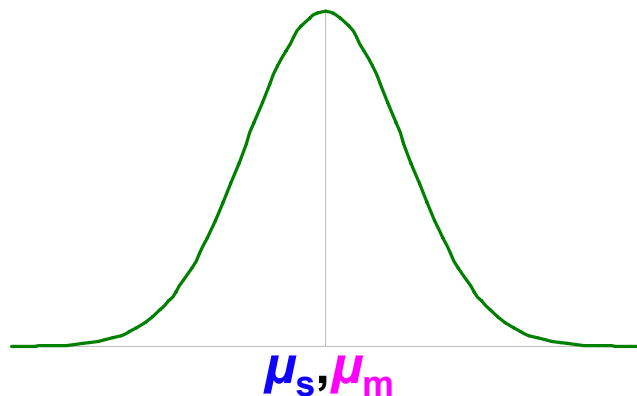
- **General hypothesis:** Stereo will make people more effective when navigating through a virtual environment (VE).
- **Testable hypothesis:** We measure time it takes for subjects to navigate through a particular VE, under conditions of stereo and mono viewing. We hypothesis subjects will be faster 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  $\mu_s$ : 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  $H_0: p(X | H_0)$ .
  - 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  $H_0: p(1.983 \text{ sec} | H_0) = 0.445$ .
  - $p = 0.445$  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.



## 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  $H_0: p(252.5 \text{ msec} | H_0) = 0.008$ .
  - $p = 0.008$  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.

# Some Considerations...

- The conditional probability  $p( X | H_0 )$ 
  - 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  $p( X | H_0 )$  is calculated
  - If you can calculate  $p( X | H_0 )$ , you can test a hypothesis
- The null hypothesis  $H_0$ 
  - $H_0$  usually in form  $f(\mu_1, \mu_2, \dots) = 0$
  - 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  $f(\mu_1, \mu_2, \dots) \neq 0$ ; 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

# Probabilistic Reasoning

- If hypothesis testing was absolute:
  - If  $H_0$  is true, then  $X$  **cannot occur**...however,  $X$  has occurred...therefore  $H_0$  is **false**.
  - e.g.: If a person is a Martian, then they are not a member of Congress (**true**)...this person is a member of Congress...therefore they are not a Martian. (**correct result**)
  - e.g.: If a person is an American, then they are not a member of Congress (**false**)...this person is a member of Congress...therefore they are not an American. (**if-then false, so correct result**)
- However, hypothesis testing is probabilistic:
  - If  $H_0$  is true, then  $X$  is **highly unlikely**...however,  $X$  has occurred...therefore  $H_0$  is **highly unlikely**.
  - e.g.: If a person is an American, then they are probably not a member of Congress (**true, right?**)...this person is a member of Congress...therefore they are probably not an American. (**incorrect result, but correct hypothesis testing reasoning**)

# Hypothesis Testing Outcomes

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

- $p(X | H_0)$  compared to  $\alpha$ , so hypothesis testing involves setting  $\alpha$  (typically 0.05 or 0.01)
- 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

# Analysis of Variance

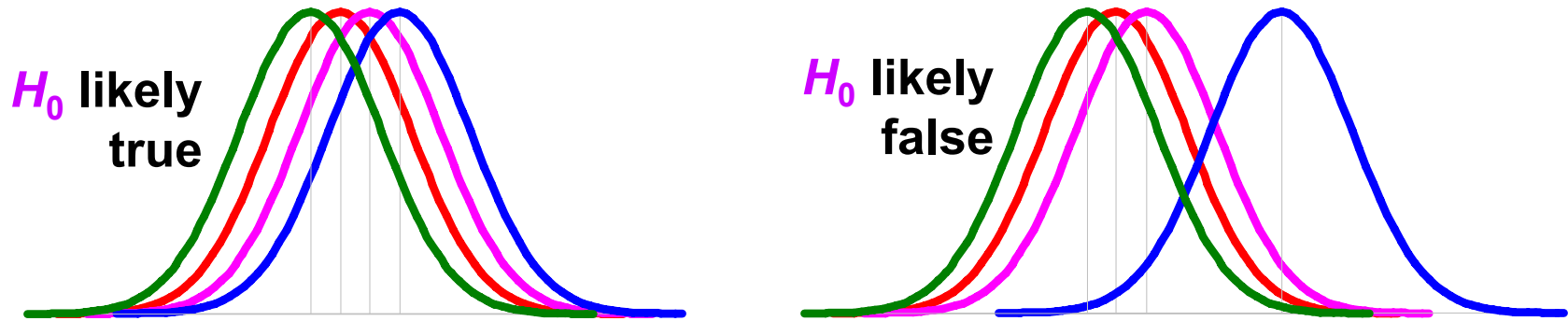
- **Experimental Validity**
- **Experimental Design**
- **Describing Data**
  - Graphing Data
  - Descriptive Statistics
- **Inferential Statistics**
  - Hypothesis Testing
  - *Analysis of Variance*
  - Power
- **Graphical Data Analysis**



# ANOVA: Analysis of Variance

- ***t*-test used for comparing two means**
  - (2 x 1 designs)
- **ANOVA used for factorial designs**
  - Comparing multiple levels ( $n \times 1$  designs)
  - Comparing multiple independent variables ( $n \times m$ ,  $n \times m \times p$ ), 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**

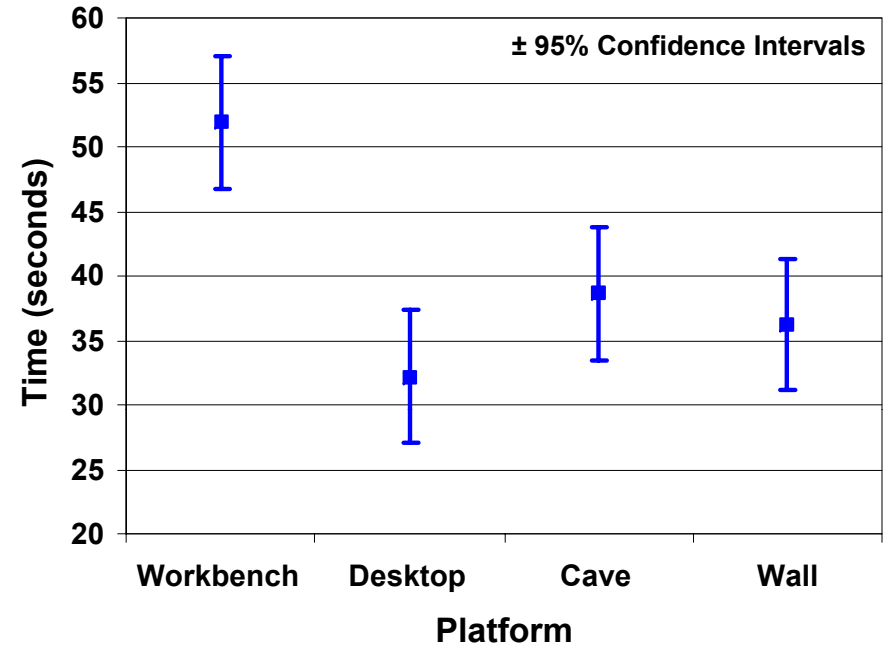
# How ANOVA Works



- 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  $F \gg 1$ , 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.



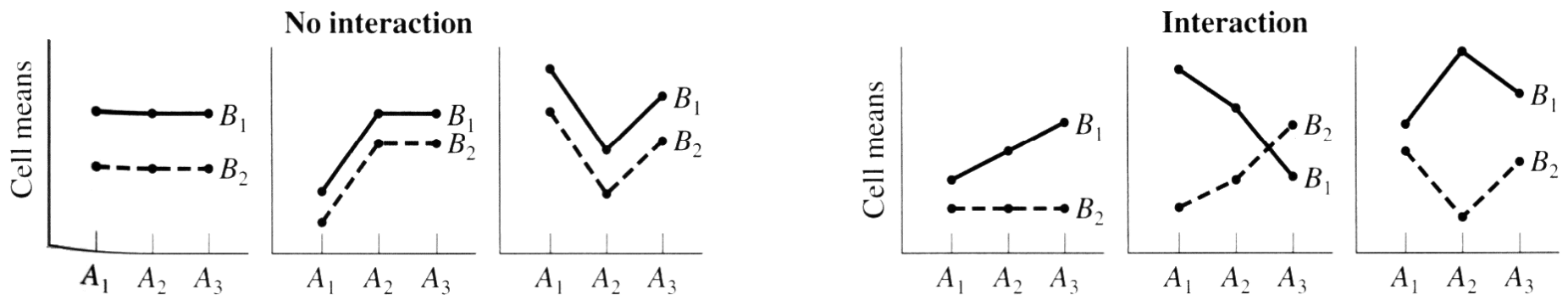
Source	SS	df	MS	F	p
Between (platform)	1205.8876	3	401.9625	3.100*	0.031
Within (P x S)	12059.0950	93	129.6677		

\* $p < .05$

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

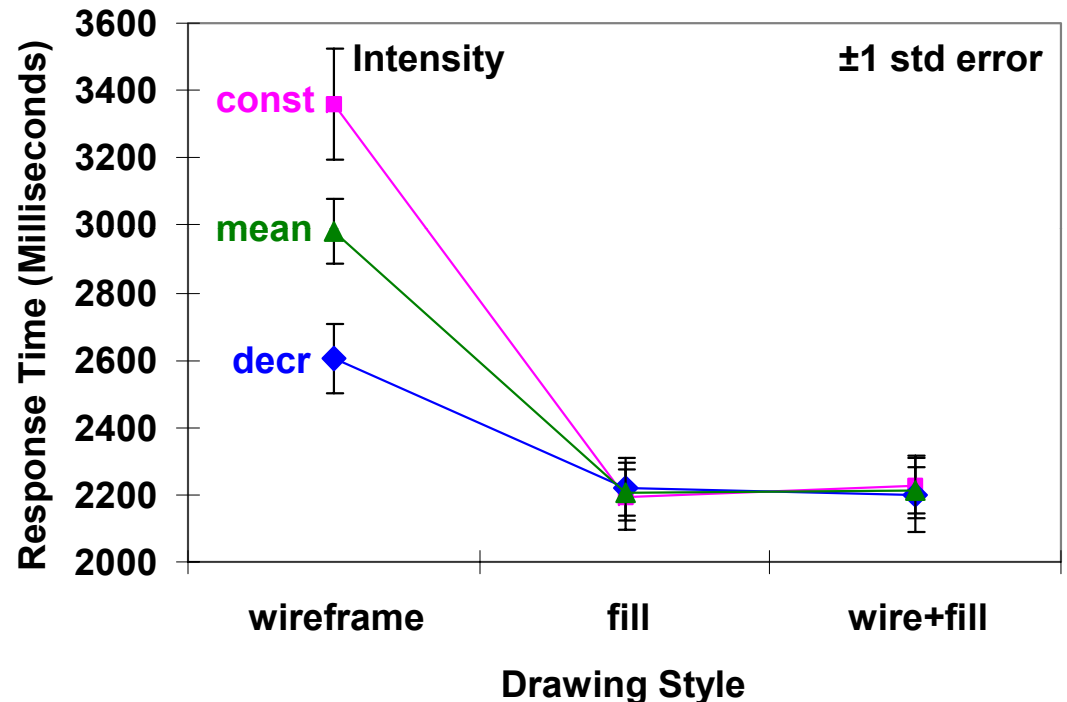
# Main Effects and Interactions

- **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



# 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



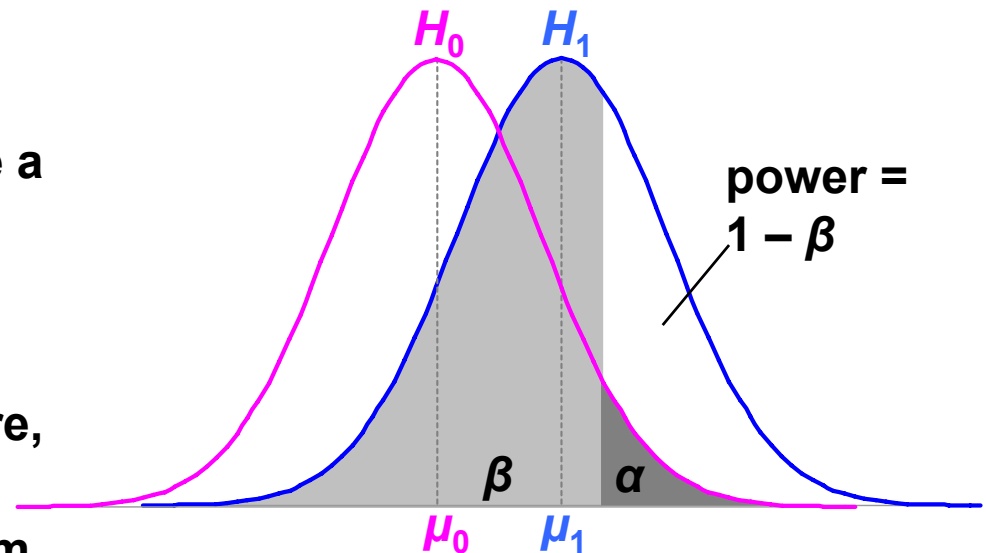
# Power

- **Experimental Validity**
- **Experimental Design**
- **Describing Data**
  - Graphing Data
  - Descriptive Statistics
- **Inferential Statistics**
  - Hypothesis Testing
  - Analysis of Variance
  - *Power*
- **Graphical Data Analysis**

# Interpreting $\alpha$ , $\beta$ , and Power

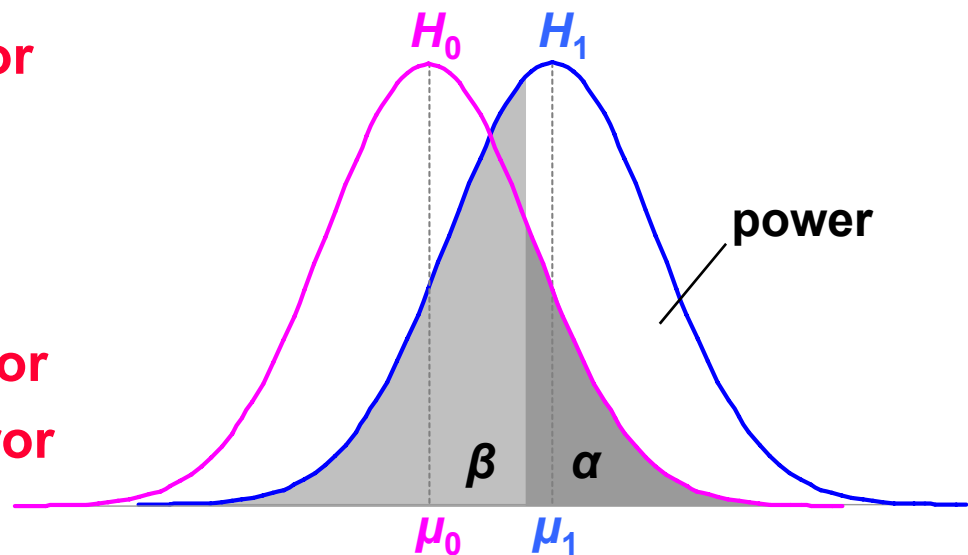
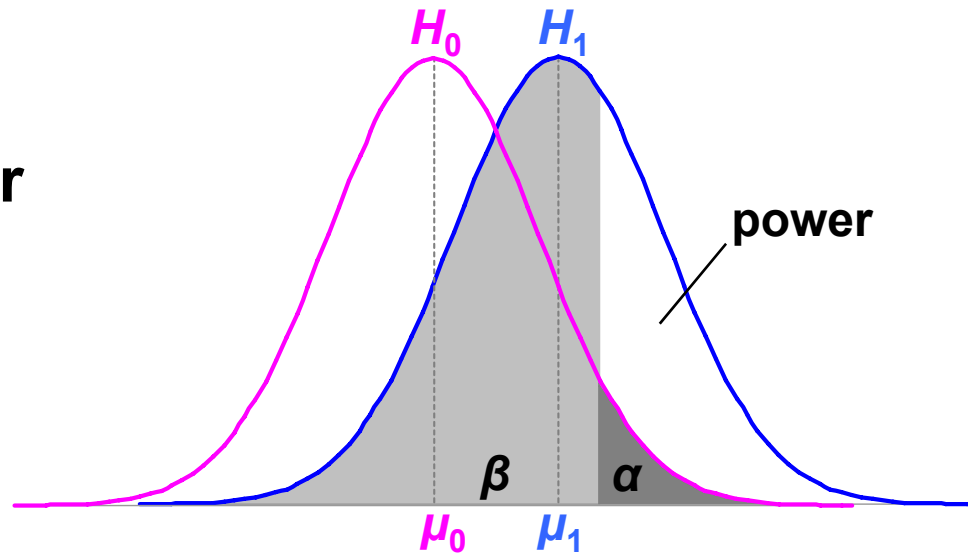
		Decision	
		Reject $H_0$	Don't reject $H_0$
True state of the world	$H_0$ false	a result! $p = 1 - \beta = \text{power}$	type II error $p = \beta$
	$H_0$ true	type I error $p = \alpha$	wasted time $p = 1 - \alpha$

- 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  $\beta$



# Increasing Power by Increasing $\alpha$

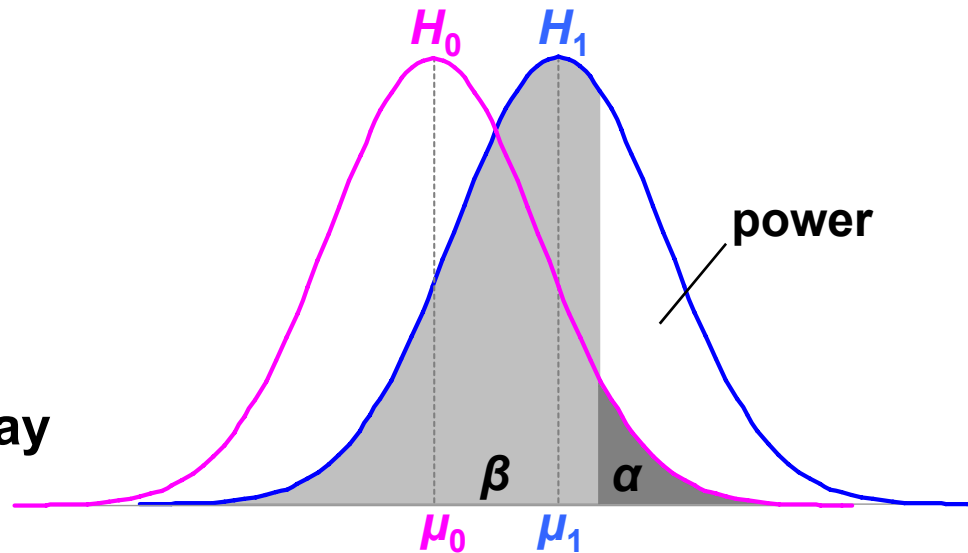
- Illustrates  $\alpha$  / power tradeoff
- Increasing  $\alpha$ :
  - Increases power
  - Decreases **type II error**
  - Increases **type I error**
- Decreasing  $\alpha$ :
  - Decreases power
  - Increases **type II error**
  - Decreases **type I error**



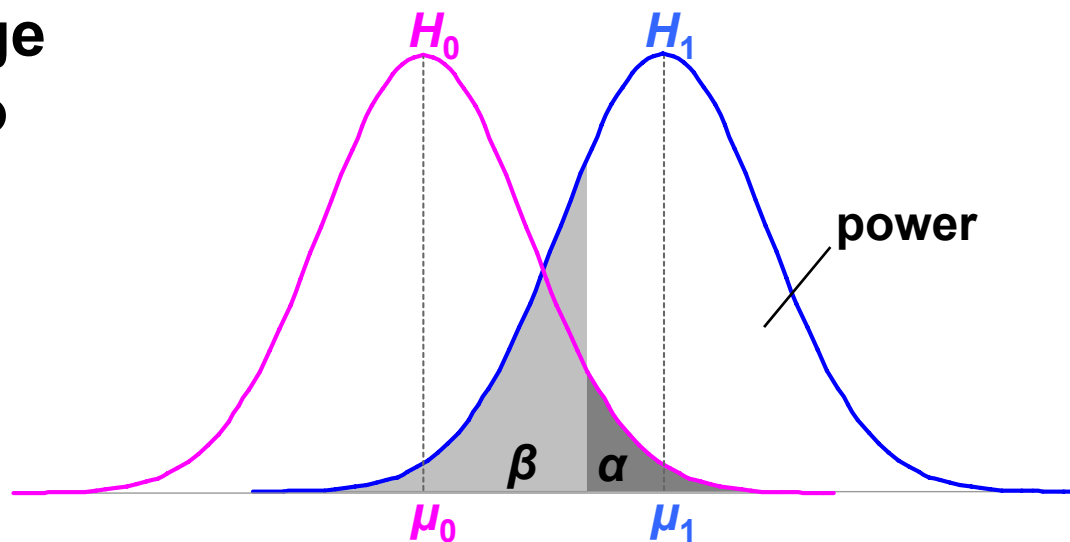


# Increasing Power by Measuring a Bigger Effect

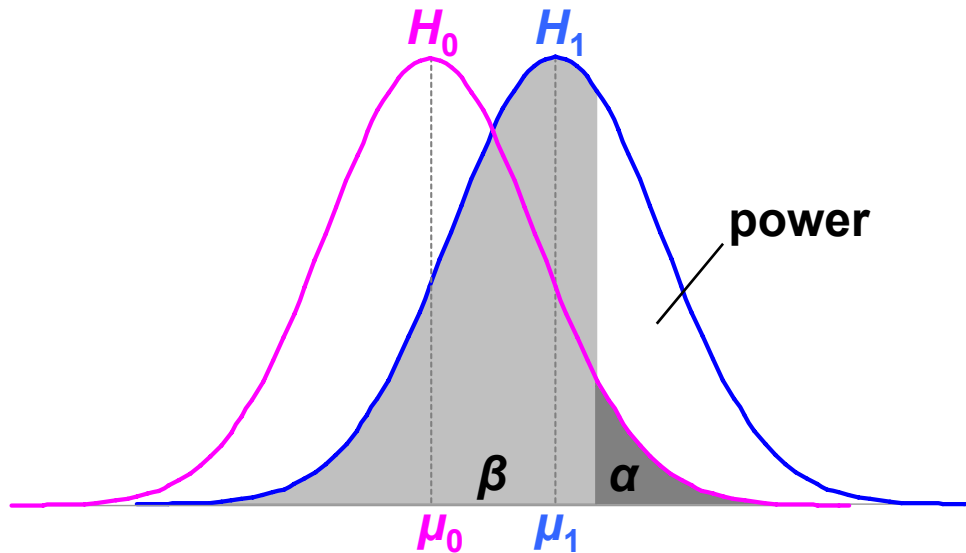
- If the effect size is large:
  - Power increases
  - **Type II error** decreases
  - $\alpha$  and **type I error** stay the same



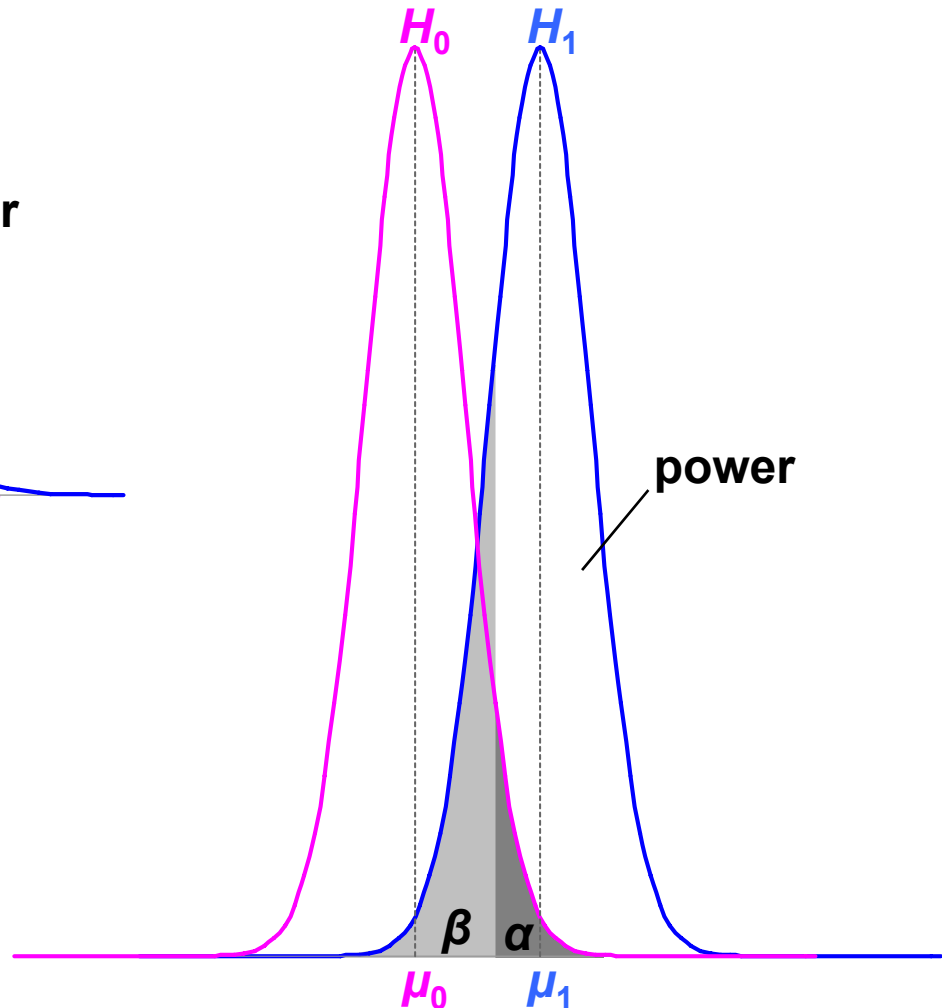
- Unsurprisingly, large effects are easier to detect than small effects



# Increasing Power by Collecting More Data

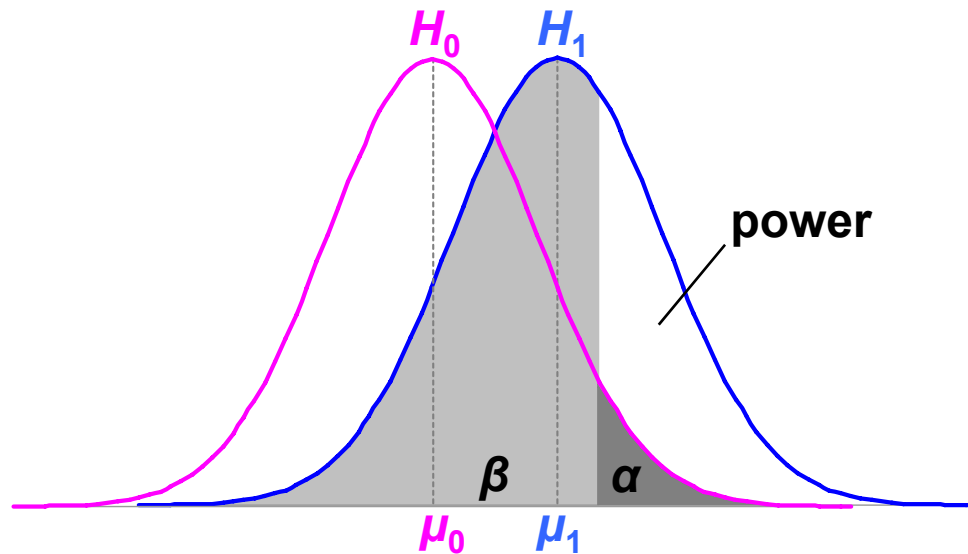


- Increasing sample size ( $N$ ):
  - Decreases variance
  - Increases power
  - Decreases **type II error**
  - $\alpha$  and **type I error** stay the same
- There are techniques that give the value of  $N$  required for a certain power level.

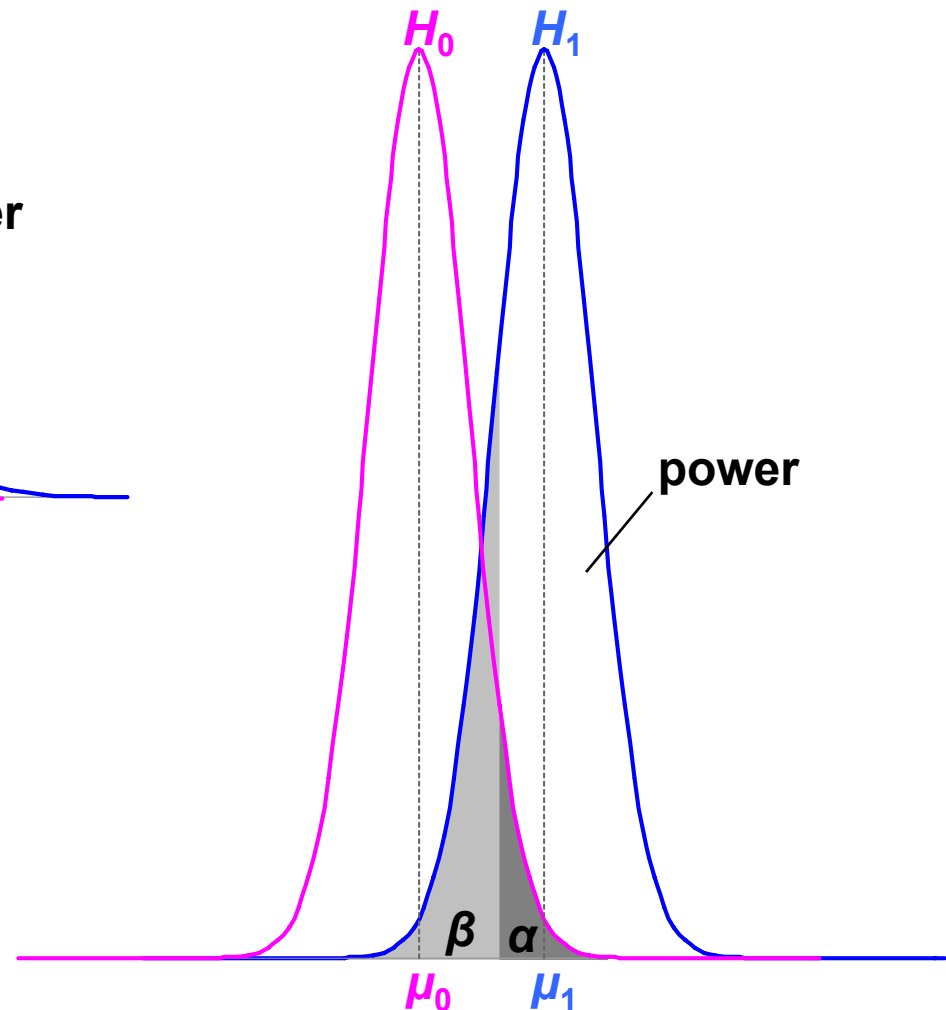


- Here, effect size remains the same, but variance drops by half.

# Increasing Power by Decreasing Noise



- Decreasing experimental noise:
  - Decreases variance
  - Increases power
  - Decreases **type II error**
  - $\alpha$  and **type I error** stay the same
- More careful experimental results give lower noise.



- Here, effect size remains the same, but variance drops by half.

# Using Power

- Need  $\alpha$ , effect size, and sample size for power:

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

- Problem for Visualization:

- 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:

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

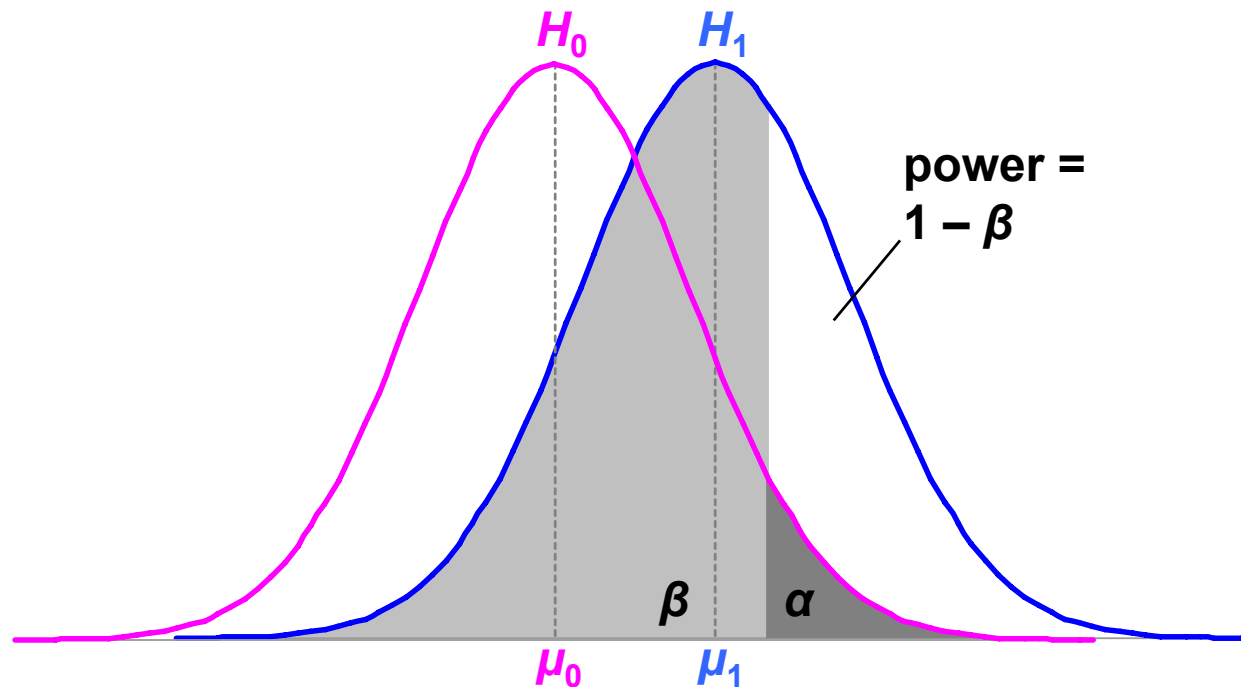
## 3. Significance level needed:

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

(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.



# Graphical Data Analysis

- **Experimental Validity**
- **Experimental Design**
- **Describing Data**
  - Graphing Data
  - Descriptive Statistics
- **Inferential Statistics**
  - Hypothesis Testing
  - Analysis of Variance and Factorial Experiments
  - Power
- *Graphical Data Analysis*

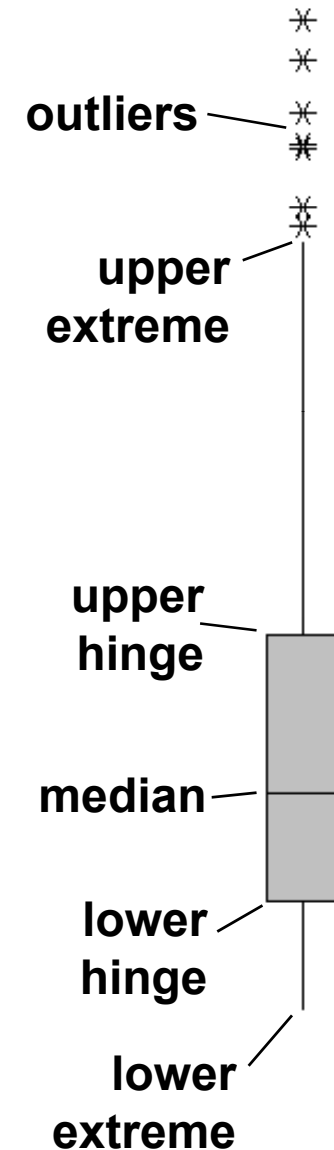
# Exploratory Data Analysis (EDA)

- EDA is:
  - A set of **data analysis tools** and **techniques**
  - A **philosophy** of how to investigate data
- EDA philosophy: data should be **explored**, with an open mind
  - Contrary to then-popular view: statistical tests should be planned before data collected
  - Data may reveal **more** than anticipated, **other** than anticipated
  - Emphasizes **images** that yield rapid insight
  - Greatest value “when it forces us to notice what we never expected to see.” [Tukey 77]
- EDA workflow:
  - **1<sup>st</sup>**: explore the data (descriptive statistics)
  - **2<sup>nd</sup>**: confirm the findings (hypothesis testing)
- EDA is visualization philosophy applied to data analysis

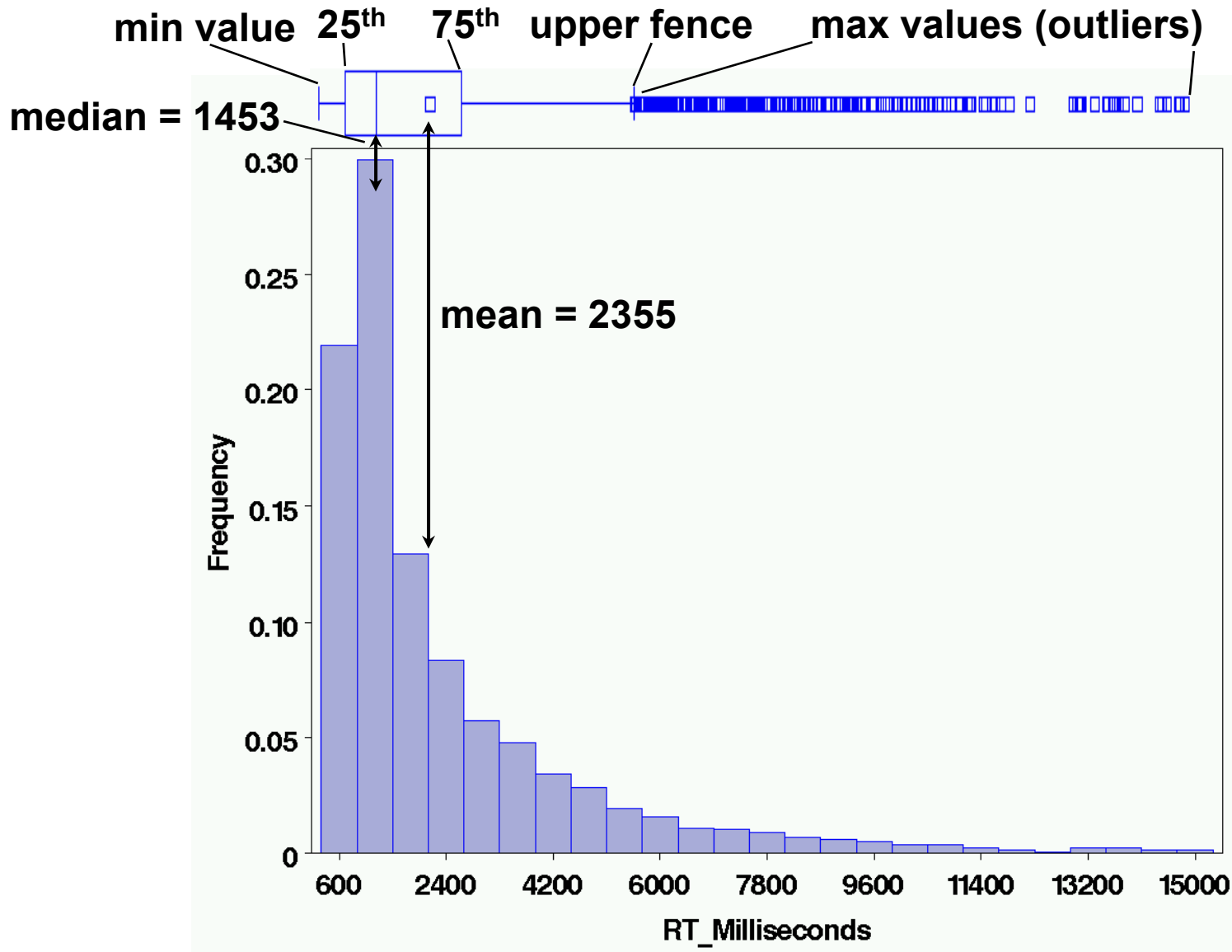


# EDA and Median Statistics

- **EDA** emphasizes **median statistics**:
  - median
  - upper hinge, lower hinge
  - upper extreme, lower extreme
- 5 values often drawn as a **boxplot**:
- Calculation of hinges and extremes depends on software
- Median statistics insensitive to
  - Data distribution
  - Outliers
- Use mean statistics once distribution is established and outliers removed

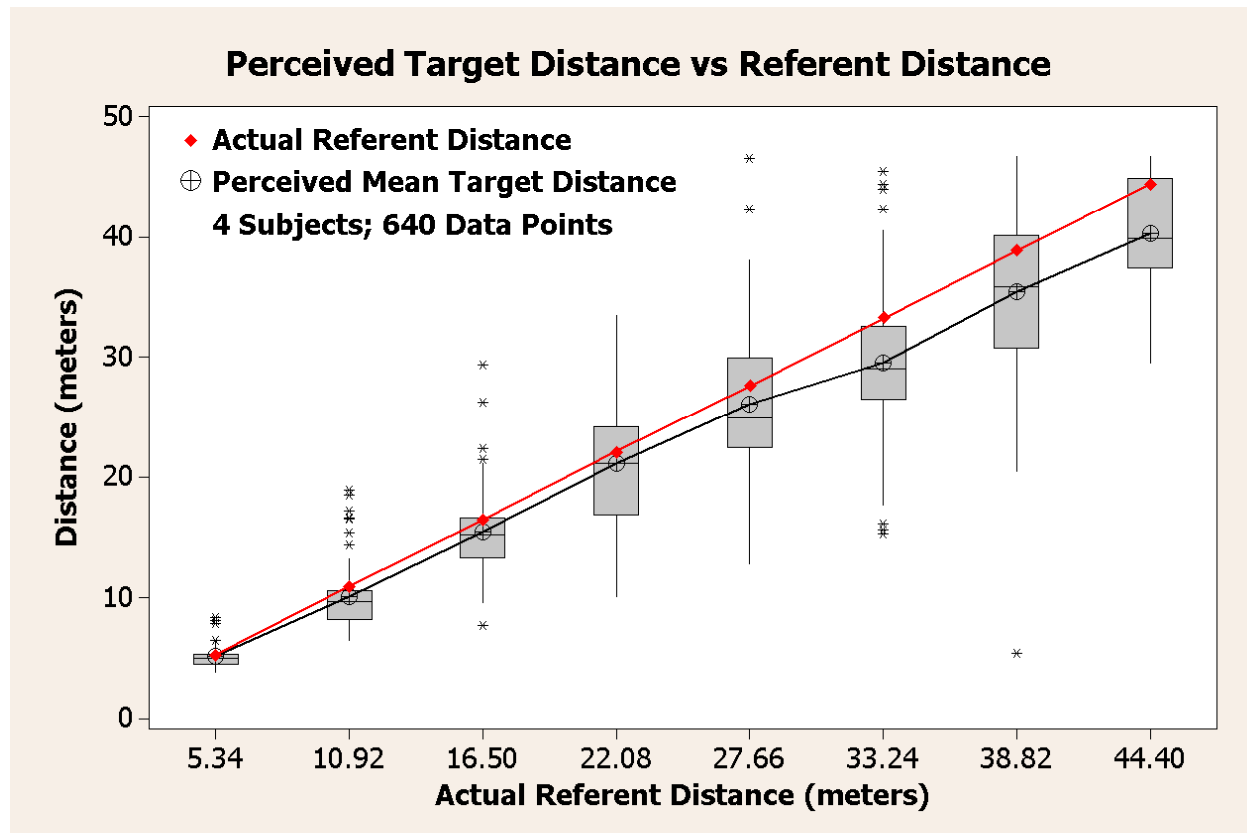


# Example Histogram and Boxplot from Real Data



Data from [Living Swan et al 03]

# Boxplots Displaying Groups

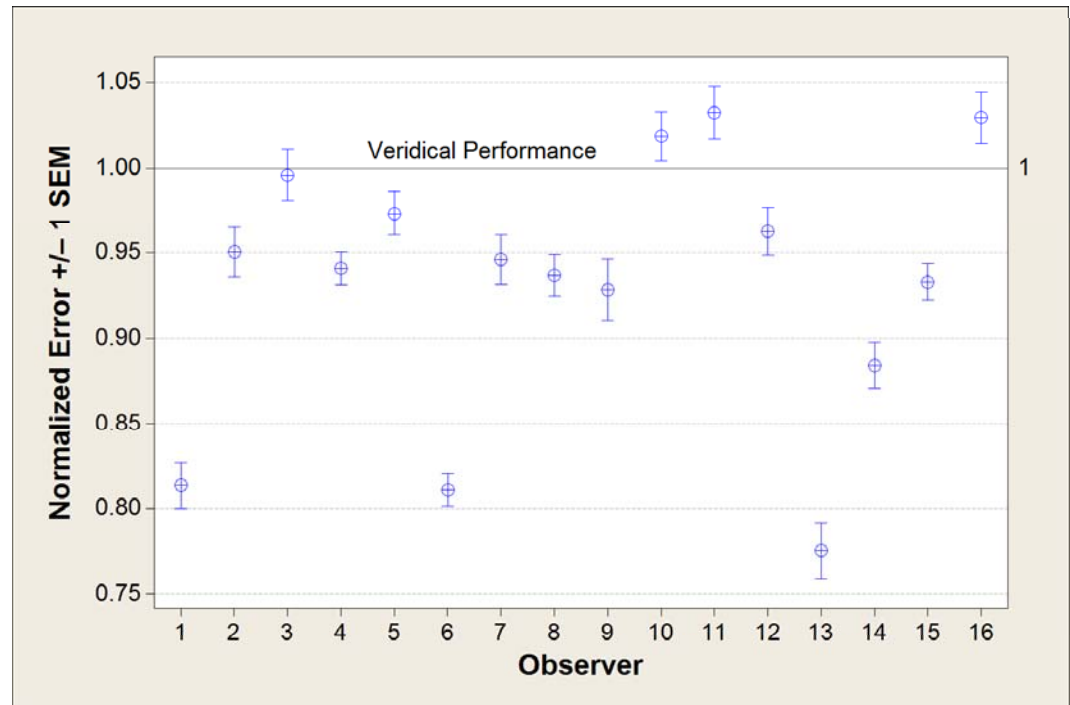


- Emphasizes variation and relationship to mean
- Because narrow, can be used to display side-by-side groups
- EDA includes many other innovative graphical techniques...

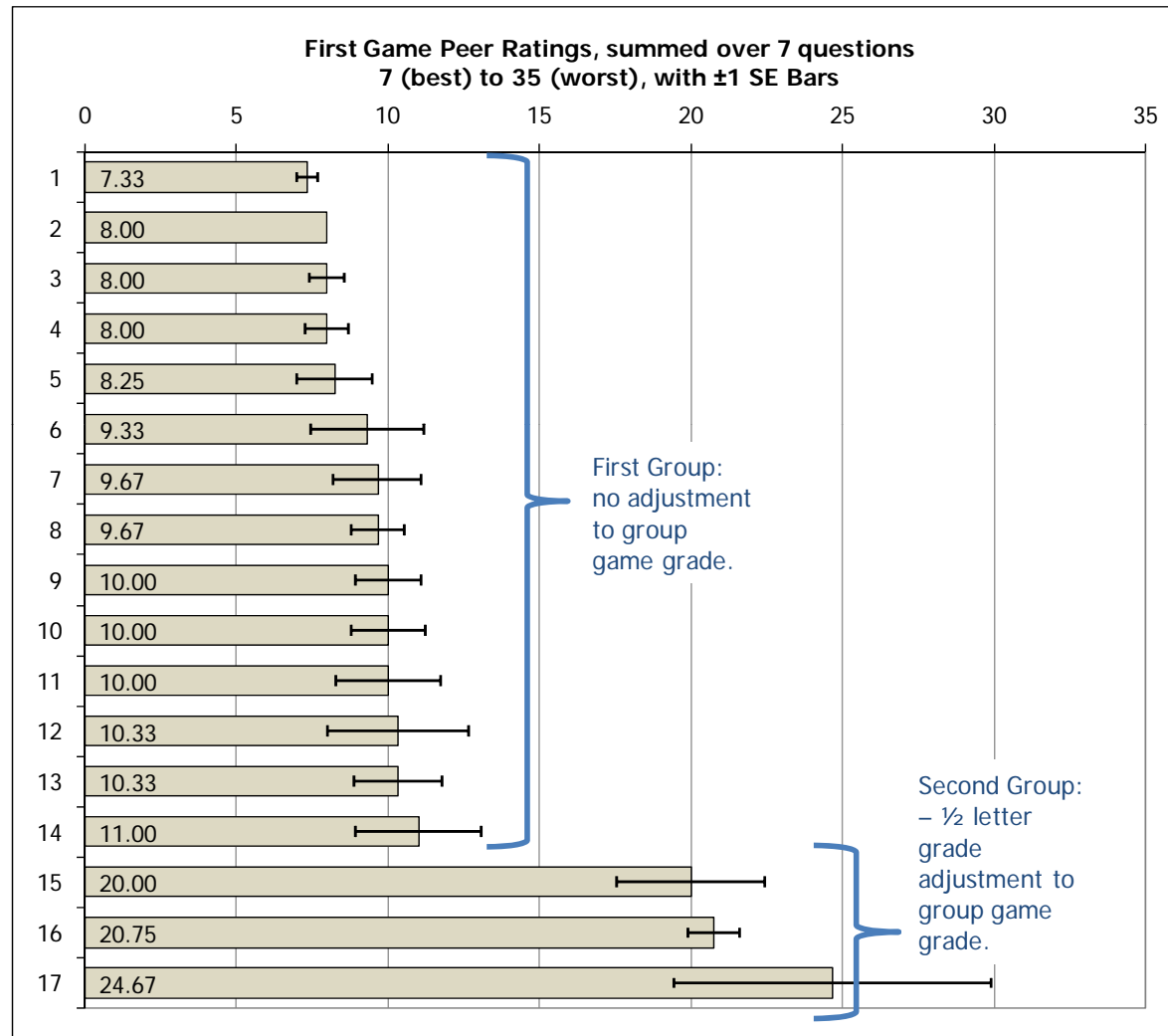
Data from [Swan et al 06]

# Mean $\pm$ SEM Plots

- **Most important considerations:**
  - Size of difference between means
  - Distance between error bars (separation / overlap)
    - Graphical indication of power
  - Size of smallest meaningful interval on *y*-axis
- **Note that considerations are all graphical**

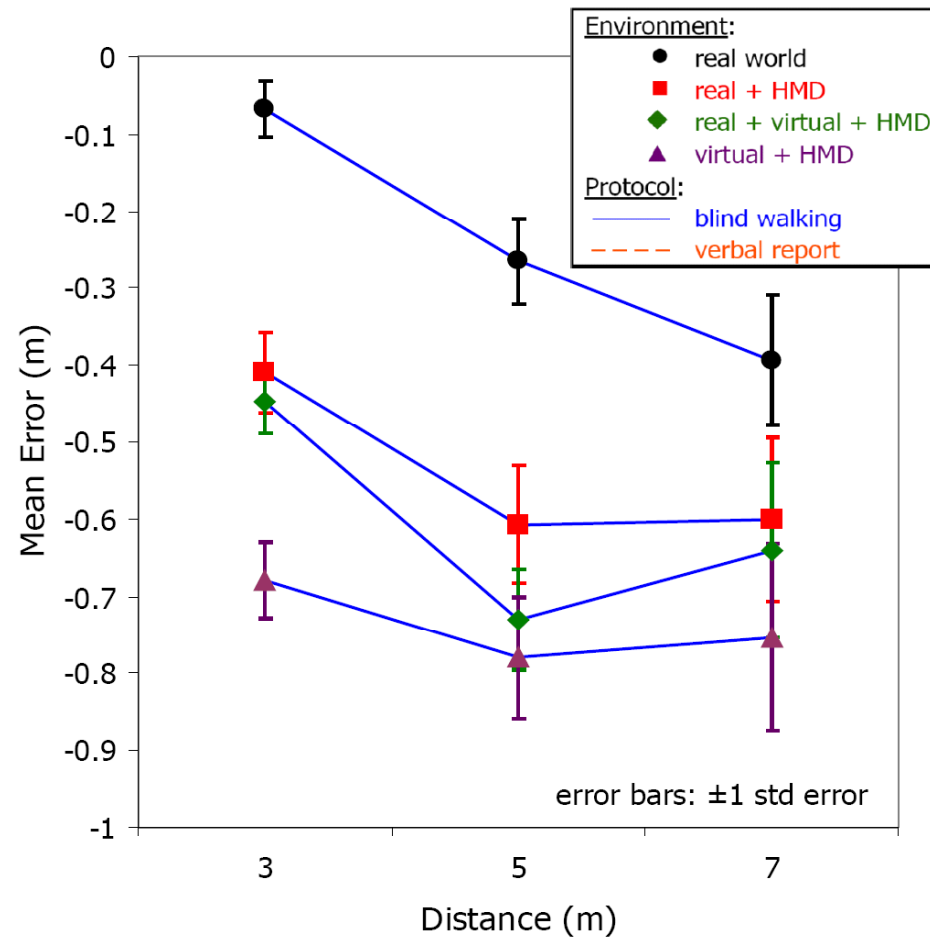


# Mean $\pm$ SEM Bar Graphs



- Error bars give much more context to the results
- Always use error bars if bars showing means

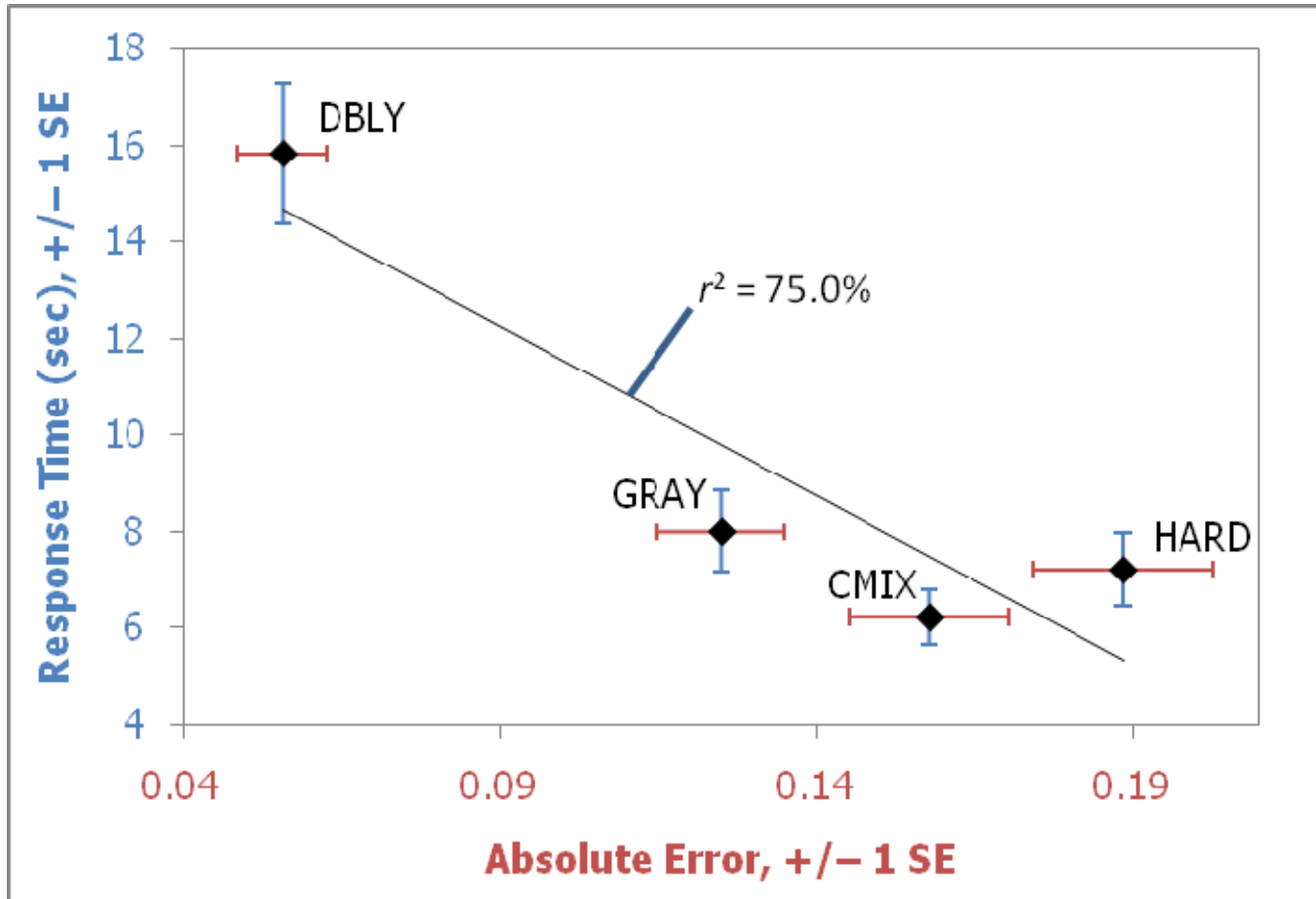
# Mean $\pm$ SEM Interaction Plots



Data from [Swan et al 07]

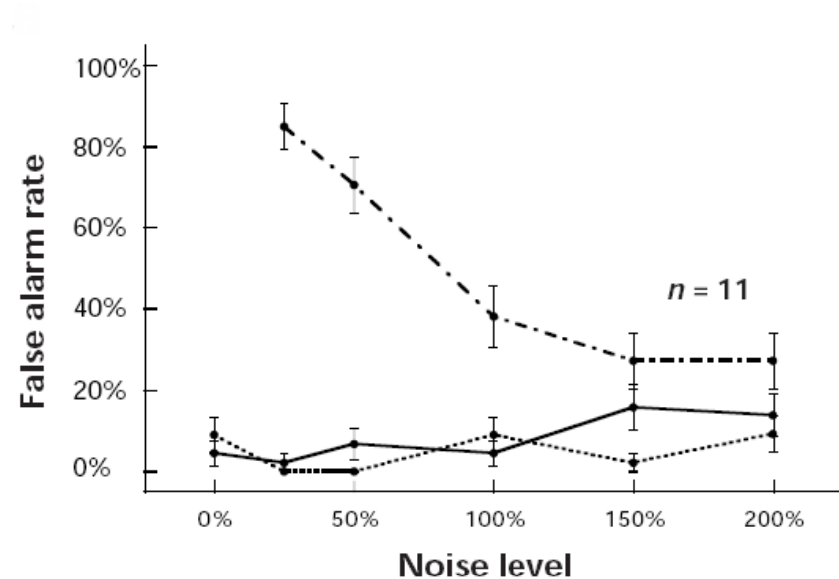
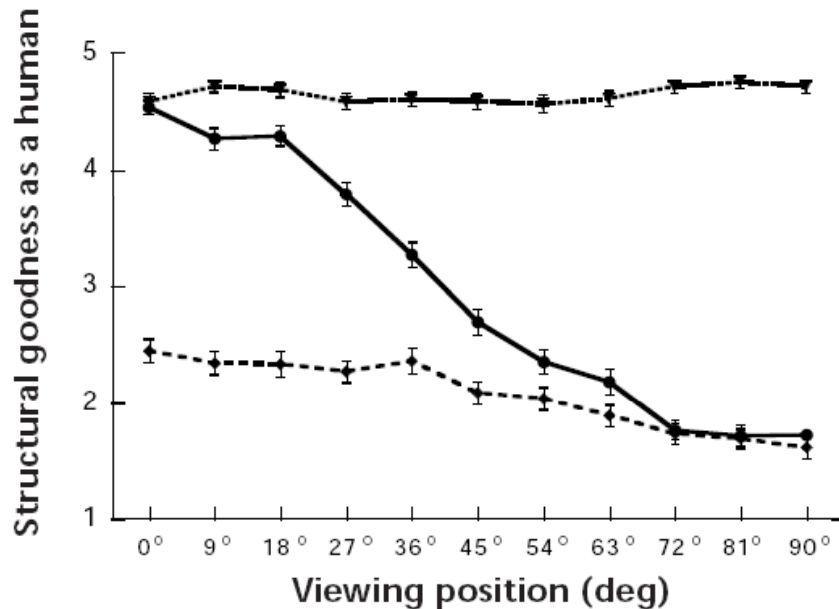
- Again, error bars give much more context to the results
- Here, error bars suggest where to group and separate the means

# XY Mean +/- SEM Plots



- Error bars are against both axes
- Suggests a clear speed / accuracy tradeoff

# Are Plots All You Need?



- Two plots from [Bülthoff et al 98], *Nature Neuroscience*
- Small error bars relative to
  - (1) effect sizes,
  - (2) smallest meaningful interval → large amount of power
- Paper contains no hypothesis testing!
- In some fields (e.g., psychophysics) hypothesis testing culturally unnecessary if plots convincingly show enough power



# My Data Analysis Work Flow

- **Create MS Word data analysis file**
  - Can throw in text and graphics
  - Can organize using headings and outliner
- **In a very non-linear fashion:**
  - Draw histograms and boxplots; understand distributions
  - Remove outliers
  - Draw mean  $\pm$  SEM plots
  - Explain dependent measures calculations
  - Hypothesize as to what we (might have) found and why
  - Perform hypothesis testing on interesting results
  - Perhaps collect more data if results look promising but are not yet powerful
- **Eventually determine what is the overall story of the data; what graphs to show**

# Example of My Analysis Document

## 1 Dependent Measures

We have calculated 4 dependent measures:

(1) *judged distance*, (meters)

(2)  $error = judged\ distance - correct\ distance$ , (meters)

$error = 0$ : a veridical answer (no error)

$error > 0$ : increasing overestimation

$error < 0$ : increasing underestimation

(3)  $absolute\ error = |judged\ distance - correct\ distance|$ , (meters)

$absolute\ error = 0$ : a veridical answer (no error)

$absolute\ error > 0$ : increasing overestimation / underestimation; folds the direction of the error together

(4)  $normalized\ error = judged\ distance / correct\ distance$ , (no units)

$normalized\ error = 1$ : a veridical answer (no error)

$normalized\ error > 1$ : increasing overestimation (normalized to units of *correct distance*)

$0 < normalized\ error < 1$ : increasing underestimation (normalized to units of *correct distance*)

Often *normalized error* is considered as a percentage.

## 2 Ideas

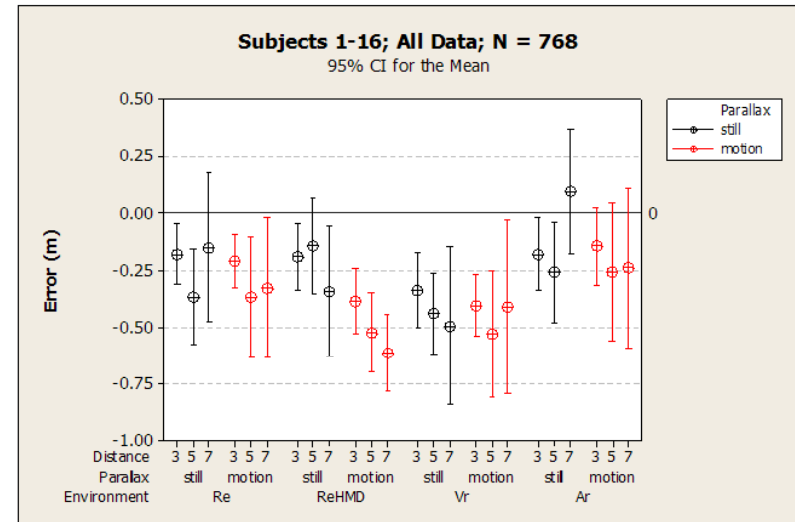
### 2.1 Analysis Tasks To Do

- Do discriminate analysis to justify splitting out subjects 1, 6, and 13.
- Make and consider “learning” graph.
- Calculate between-results pooled confidence intervals as Laidlaw does. But, we have to us Bonferroni corrections, which reduces power. Howell [1] indicates that Bonferroni corrections loose too much power when there are too many multiple comparisons, and recommends either Ryan REGWQ or Tukey HSD post-hoc tests. Perhaps the better approach is to just use standard error bars, and indicate the a-priori groupings using another method.
- Try removing the .1 meter “correction”, just to see what happens.
- Redraw the big graph in Excel 2007.
- Normalize per subject.

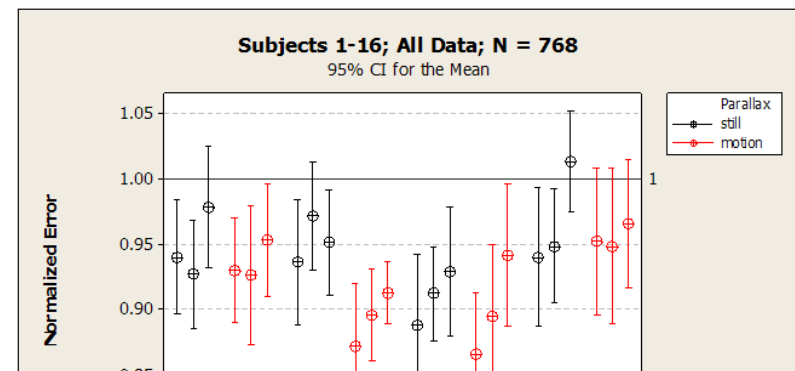
### 2.2 Overall Findings

- The degree of underestimation for all conditions is low compared to many previous studies.
- There does not appear to be large interactions with increasing distance (the only exceptions: Re / still, ReHMD / still, Ar / still); we could do a power analysis to see how non-existent the distance interaction really is.

## 4 Analysis (Original Data, with 0.1 meter subtracted)



This graph indicates that the variability of error increases with increasing distance (the confidence intervals tend to increase with increasing distance). This means that the assumption of homogeneity of variance over distance is not met for error, and hence it is not appropriate to perform an ANOVA over distance for error. I believe this also means it is more appropriate to sum over distance for normalized error as well. [Run an omnibus ANOVA here.](#)



# Final Thoughts on Experimental Design and Data Analysis

- In the end, what matters are:
  - (1) **the results**, and
  - (2) **how they relate to what's being studied**.
- ...not hypothesis testing (e.g., [Bülthoff et al 98])
- Paraphrased quote from many applied statistics texts:  
    **“Data analysis is an art, not a science”**
- When applying data analysis to results:
  - There is no one way to be right
  - There is no one way to be wrong
- The best way to learn data analysis and experimental design: read and critique existing papers, both in visualization and in other fields.  
    **“A month in the lab will save you a day in the library”**

# References

- [Cohen 88] J Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2<sup>nd</sup> edition, Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.
- [Bülthoff et al 98] I Bülthoff, H Bülthoff, P Sinha, “*Top-Down Influences on Stereoscopic Depth-Perception*”, *Nature Neuroscience*, 1(3), July 1998, pages 254–257.
- [Cohen 94] J Cohen, “*The Earth is Round ( $p < .05$ )*”, *American Psychologist*, 49(12), pages 997–1003.
- [Johnson et al 06] CR Johnson, R Moorhead, T Munzner, H Pfister, P Rheingans, TS Yoo (Eds), *NIH-NSF Visualization Research Challenges Report*, IEEE Press, 2006.
- [Jones Swan et al 08] JA Jones, JE Swan II, G Singh, E Kolstad, SR Ellis, “*The Effects of Virtual Reality, Augmented Reality, and Motion Parallax on Egocentric Depth Perception*”, *Proceedings of the Symposium on Applied Perception in Graphics and Visualization*, Los Angeles, California, USA, August 9–10, 2008, pages 9–14.
- [Living Swan 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.
- [Howell 02] DC Howell, *Statistical Methods for Psychology*, 5<sup>th</sup> edition, Duxbury, Pacific Grove, CA, 2002.
- [Smith Prentice 93] AF Smith, DA Prentice, “*Exploratory Data Analysis*”, in G Keren, C Lewis (eds), *A Handbook for Data Analysis in the Behavioral Sciences*, Lawrence Erlbaum, Hillsdale, NJ, 1993.
- [Swan et al 07] JE Swan II, A Jones, E Kolstad, MA Livingston, HS Smallman, “*Egocentric Depth Judgments in Optical, See-Through Augmented Reality*”, *IEEE Transactions on Visualization and Computer Graphics*, Volume 13, Number 3, May/June 2007, pages 429–442.
- [Swan et al 06] JE Swan II, MA Livingston, HS Smallman, D Brown, Y Baillot, JL Gabbard, D Hix, “*A Perceptual Matching Technique for Depth Judgments in Optical, See-Through Augmented Reality*”, *Technical Papers, IEEE Virtual Reality 2006*, March 25–29, 2006.
- [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 90] ER Tufte, *Envisioning Information*, Graphics Press, Cheshire, Connecticut, 1990.
- [Tufte 83] ER Tufte, *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut, 1983.
- [Tukey 77] JW Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.

# Contact Information



## J. Edward Swan II, Ph.D.

Associate Professor

Department of Computer Science and Engineering

Department of Psychology (Adjunct)

Mississippi State University

swan@acm.org

(662) 325-7507

## Slide Location:

<http://www.cse.msstate.edu/~swan/teaching/tutorials/Swan-Vis2008-Tutorial.pdf>

