IEEE Virtual Reality
Charlotte NC March 10-14, 2007

Conducting Human-Subject Experiments with Virtual and Augmented Reality

VR 2007 Tutorial

J. Edward Swan II, Mississippi State University (organizer)
Stephen R. Ellis, NASA Ames Research Center
Bernard D. Adelstein, NASA Ames Research Center

Schedule

8:30–9:00	0.5 hrs	Intro and Group Discussion	All
9:00–10:00	1.0 hrs	Basic Experimental Design and Analysis	Ed
10:00–10:30	*0.5 hrs*	*Coffee Break*	
10:30–12:00	1.5 hrs	Basic Experimental Design and Analysis	Ed
12:00–1:30	*1.5 hrs*	*Lunch Break*	
1:30–3:00	1.5 hrs	Classical and Other Psychophysical Methods for Virtual Environments	Dov
3:00–3:30	*0.5 hrs*	*Coffee Break*	
3:30–5:00	1.5 hrs	Human Performance and Preference Studies: Exhortations and Illustrations	Steve

Basic Experimental Design and Analysis

J. Edward Swan II, Ph.D.

Department of Computer Science and Engineering

Institute for Neurocognitive Science and Technology

Mississippi State University

Motivation and Goals

- 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!
- Course attendee backgrounds?

Motivation and Goals

- **What can we accomplish in one day?**
- **Study subset of basic techniques**
  - **Presenters have found these to be the most applicable to VR, AR systems**
- **Focus on intuition behind basic techniques**
- **Become familiar with basic concepts and terms**
  - **Facilitate working with collaborators from psychology, industrial engineering, statistics, etc.**

Outline

- *Empiricism*
- Experimental Validity
- 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 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%
2006	12 / 27	44%	2 / 10	20%	1 / 10	10%
2007	9 / 26	35%	3 / 15	20%	5 / 18	28%

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***
- **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**

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

Experimental Design

- Empiricism
- Experimental Validity
- *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:

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	1st condition	2nd condition
1, 3, 5, 7	stereo	mono
2, 4, 6, 8	mono	stereo

Factorial Designs

- *n* x 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* x *m* design, *n* x *m* x *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 Levels

- **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 x 2 x 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

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

2 x 2

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

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

6 x 3 (there is no 3 x 3 that has all 3 properties)

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

4 x 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

Counterbalancing Example

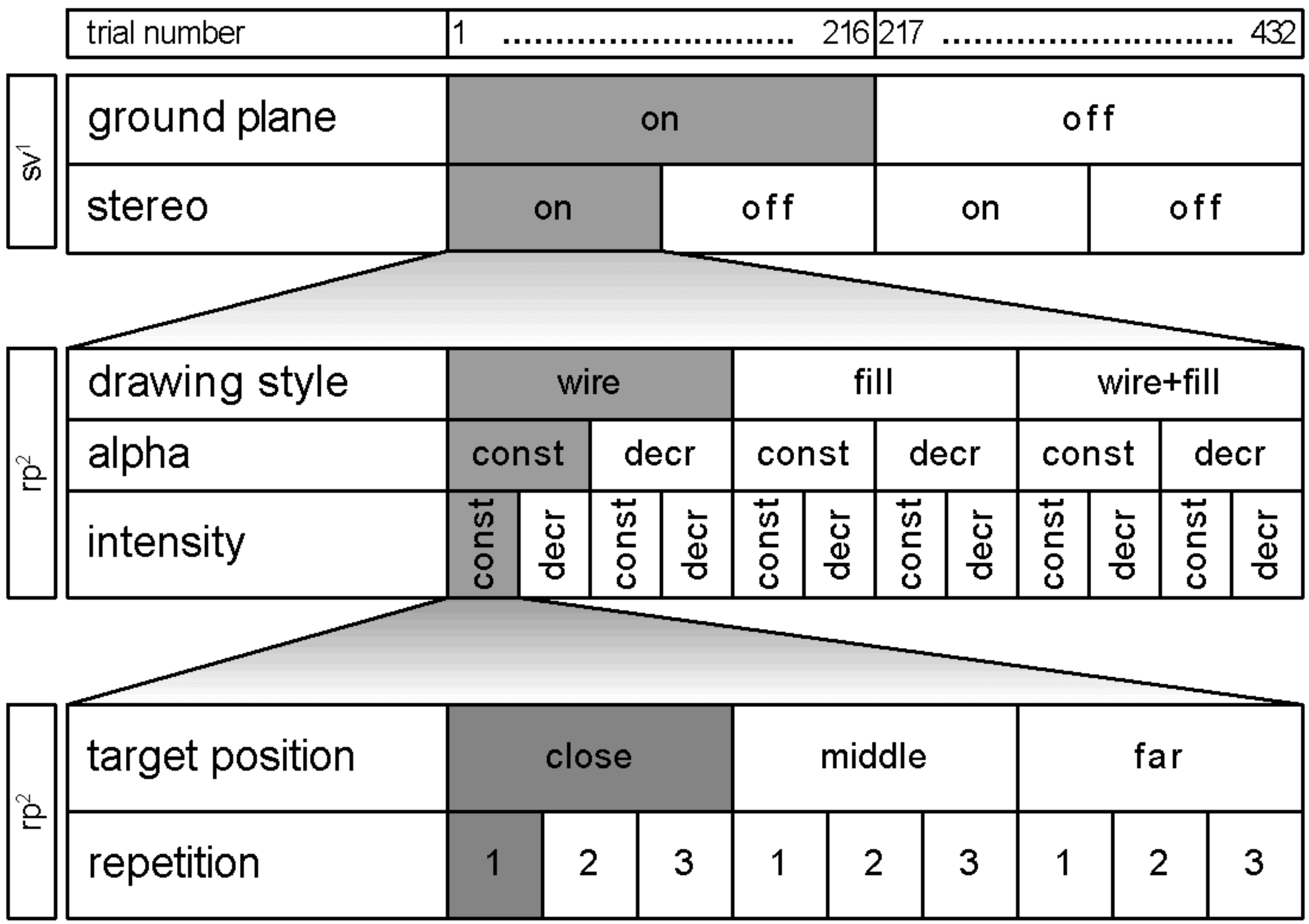
- **"A 3 (VE terrain type) by 2 (stereopsis) within-subjects design, with 4 repetitions of each cell."**
- **Form Cartesian product of Latin squares {6 x 3} (VE Terrain Type) ⊗ {2 x 2} (Stereopsis)**
- **Perfectly counterbalances groups of 12 subjects**

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

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

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

Experimental Design Example #1



[1] sv = systemically varied, [2] rp = randomly permuted

- **All variables within-subject**

From [Living et al. 03]

Experimental Design Example #2

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

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

From [Swan et al. 03]

Gathering Data

- Empiricism
- Experimental Validity
- 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 tracker 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
- Measurable phenomena:
  - Button presses, physical actions, answers

Pilot Testing a Design

- Experimental designs have to be tested and iterated (debugged)
- Typical flow:
  - 1st run: subjects are you, collaborators
  - 2nd run: small number of preliminary subjects
  - 3rd run: subset of real subjects
- With each run, problems are revealed; fix and iterate
- For later runs, perform data analysis before gathering additional data

Graphing Data

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

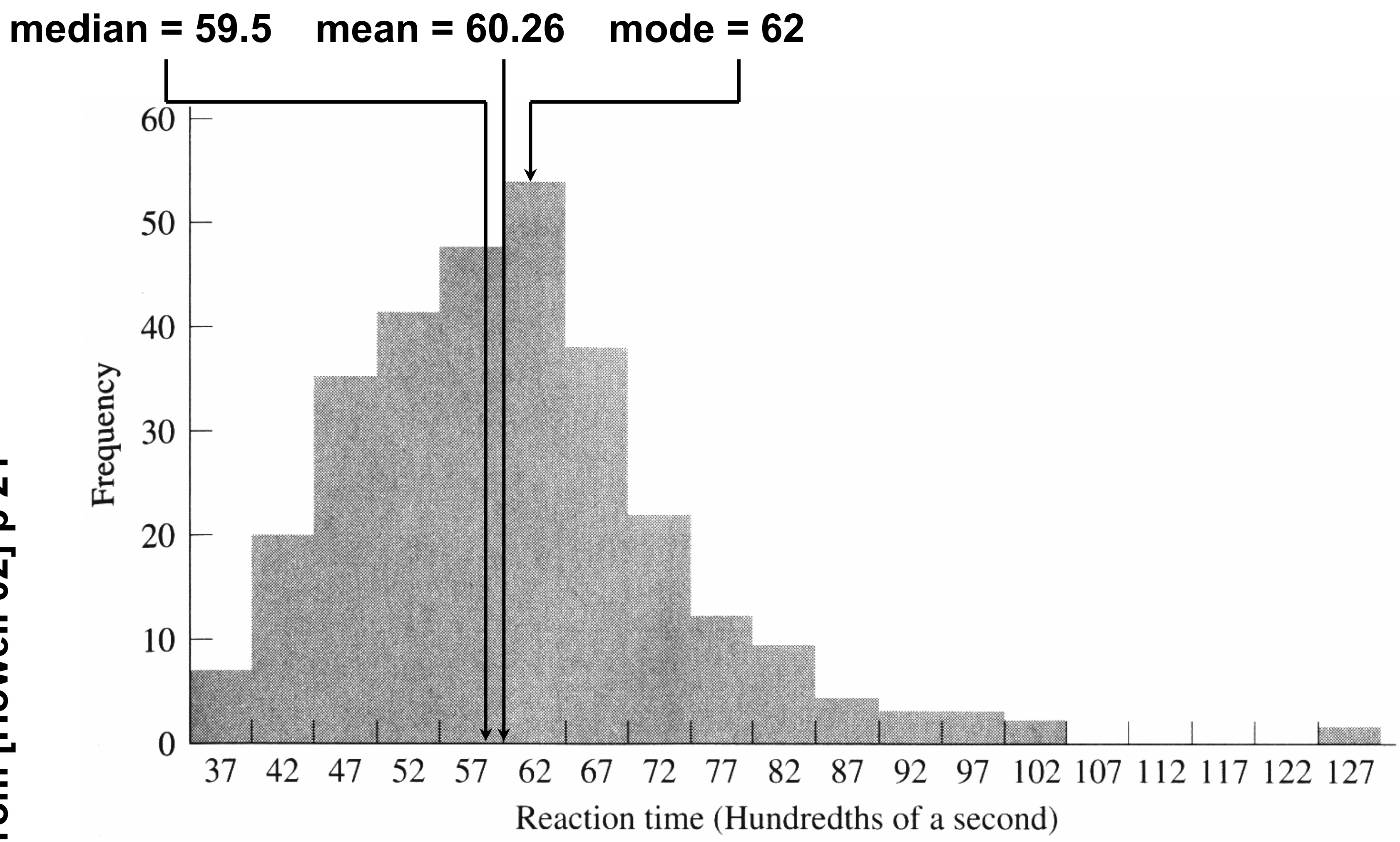
Types of Statistics

- 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

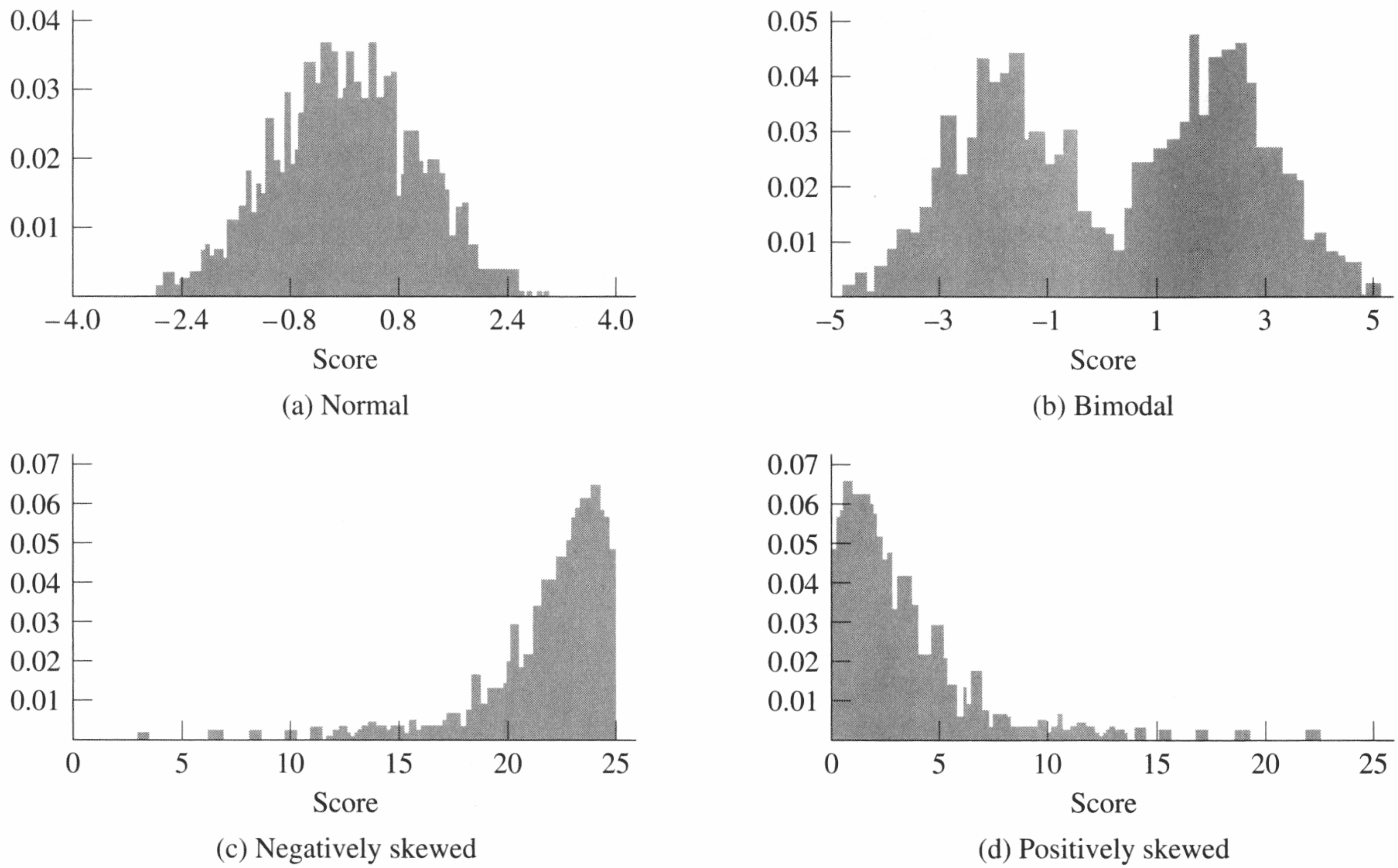
Exploring Data with Graphs

- **Histogram common data overview method**



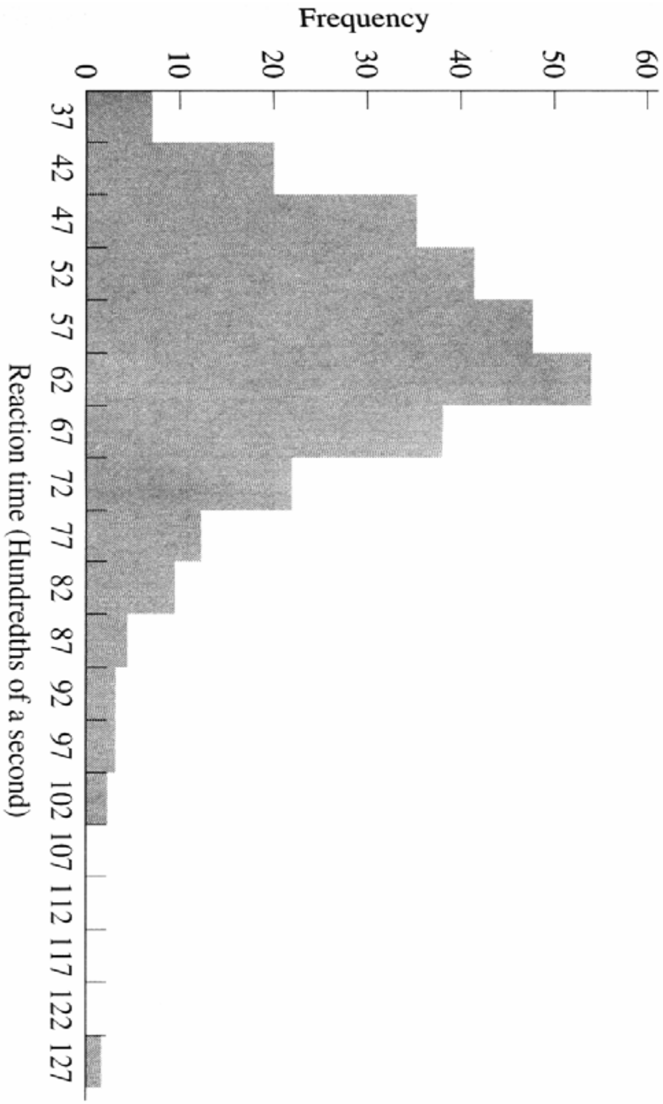
From [Howell 02] p 21

Classifying Data with Histograms



From [Howell 02] p 28

Stem-and-Leaf: Histogram From Actual Data



Raw Data	Stem	Leaf
36 37 38 38 39 39 39 40	3s	67
40 40 40 41 41 41 42 42	3.	88999
42 43 43 43 43 43 44 44	4*	0000111
44 44 44 45 45 45 45 45	4t	22233333
45 46 46 46 46 46 46 46	4f	44444555555
46 46 46 46 47 47 47 47	4s	66666666666777777777
47 47 47 47 47 48 48 48	4.	888899999
48 49 49 49 49 49 50 50	5*	0000011111111111
50 50 50 51 51 51 51 51	5t	222222222233333333
51 51 51 51 51 51 51 52	5f	4444445555555
52 52 52 52 52 52 52 52	5s	66666666667777777
52 53 53 53 53 53 53 53	5.	8888888888888999999999
53 54 54 54 54 54 54 55	6*	00000000000011111111111
55 55 55 55 55 55	6t	2222222222222223333333333
	6f	444444455555555
	6s	66666666777777777777777
	6.	889999999
	7*	01111
	7t	22222222333
	7f	44444455
	7s	666677
	7.	88899
	8*	00011
	8t	2333
	8f	5
	8s	67
	8.	8
	9*	0
	9t	
	9f	4455
	9s	
	9.	8
	High	104; 104; 125

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

From [Howell 02] p 21, 23

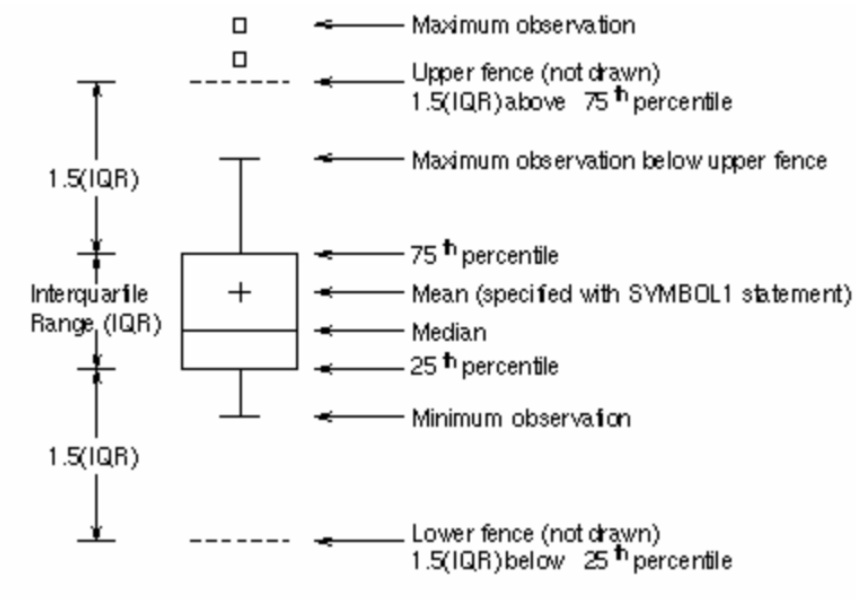
Stem-and-Leaf: Histogram From Actual Data

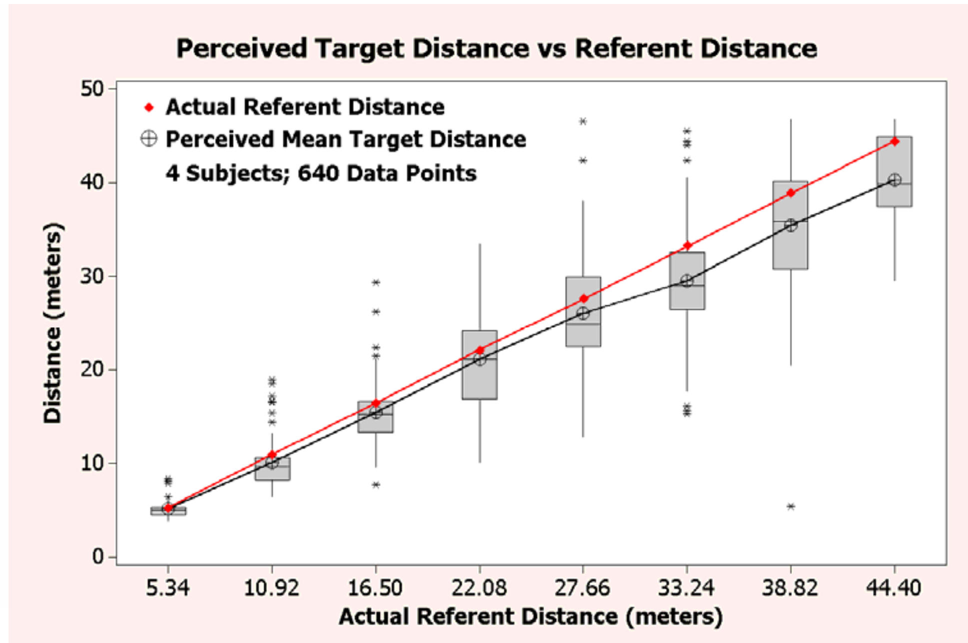
Final Recorded Grades

1	3%	F	0	0
0	0%	F	1	
0	0%	F	2	
0	0%	F	3	
0	0%	F	4	
0	0%	F	5	
5	16%	D	6	34788
8	26%	C	7	12233469
8	26%	B	8	01244699
9	29%	A	9	001123346
31				

Grades from my Autumn 2005 analysis of algorithms class

Boxplot

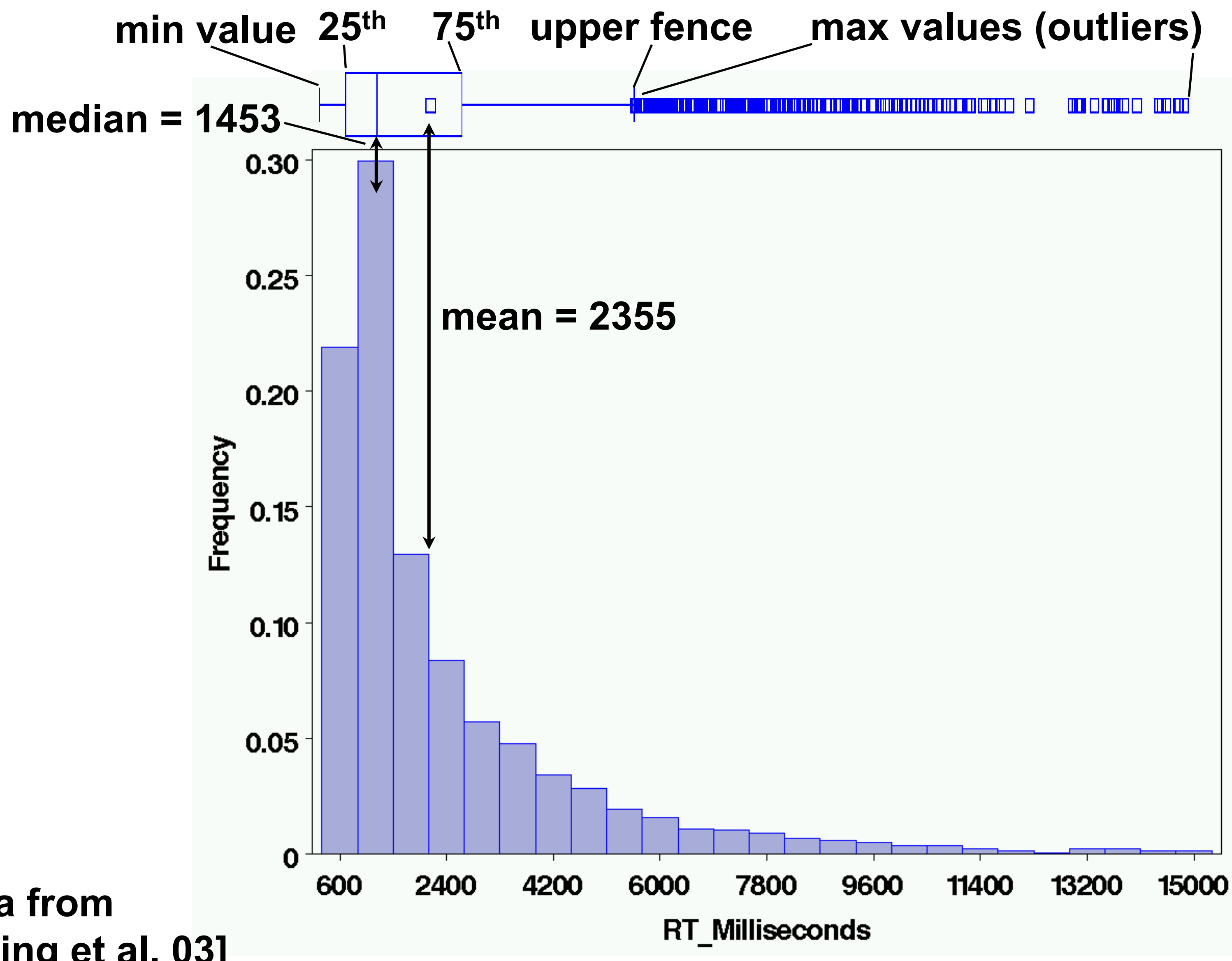




- **Emphasizes variation and relationship to mean**
- **Because narrow, can be used to display side-by-side groups**

Data from [Swan et al. 06]

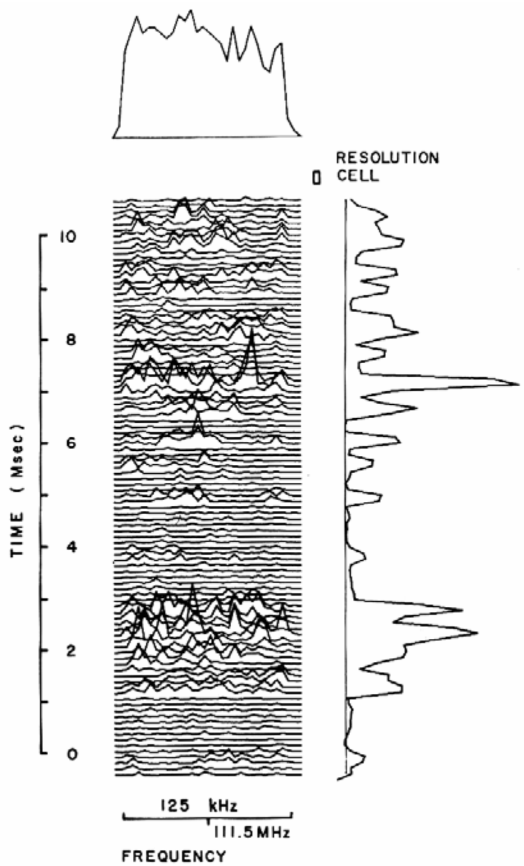
Example Histogram and Boxplot from Real Data



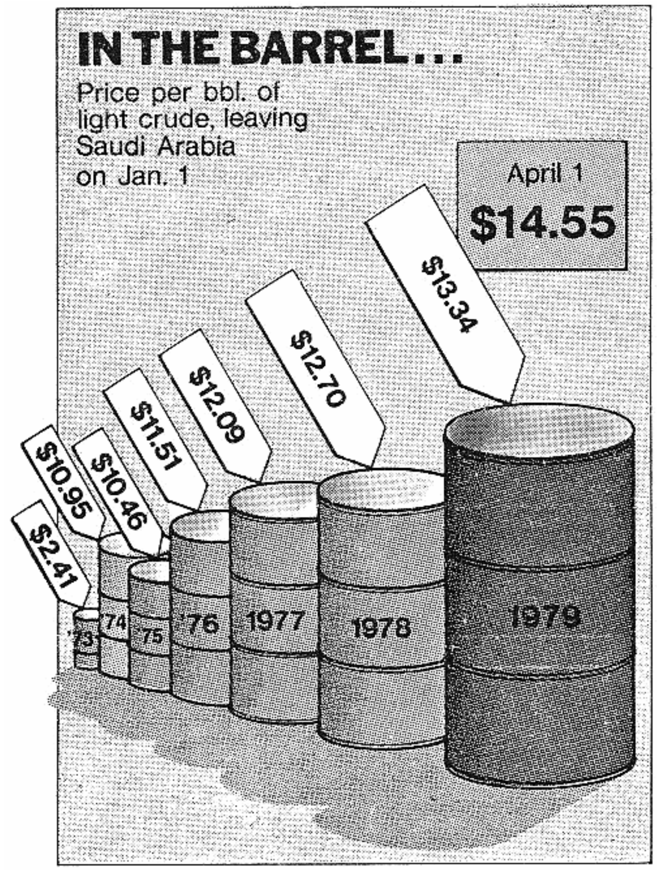
Data from
[Living et al. 03]

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

Summary Statistics

- Many numbers $\rightarrow$ 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: 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

Measuring Variability / Dispersion

Mean:

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

Mean absolute deviation:

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

Variance:

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

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

Standard deviation:

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

- **Standard deviation uses same units as samples and mean.**
- **Calculation of population variance σ^2 is theoretical, because μ 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**

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

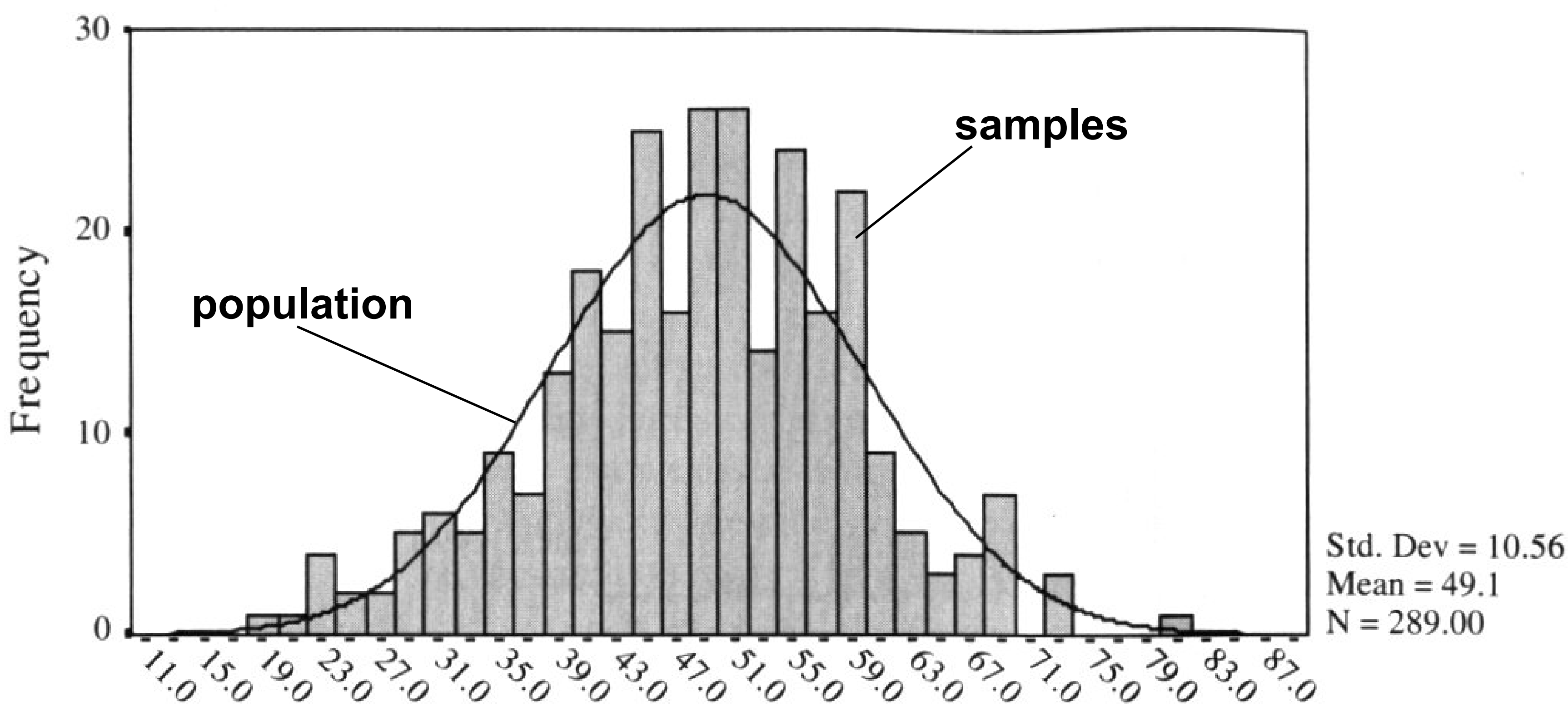
- **Sums of squares:**
  - **Summed squared deviations from mean**
- **Degrees of freedom:**
  - **Given a set of *N* observations used in a calculation, how many numbers in the set may vary**
  - **Equal to *N* minus number of means calculated**
- **Mean squares:**
  - **Sums of squares divided by degrees of freedom**
  - **Another term for variance, used in ANOVA**

Hypothesis Testing

- **Empiricism**
- **Experimental Validity**
- **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).

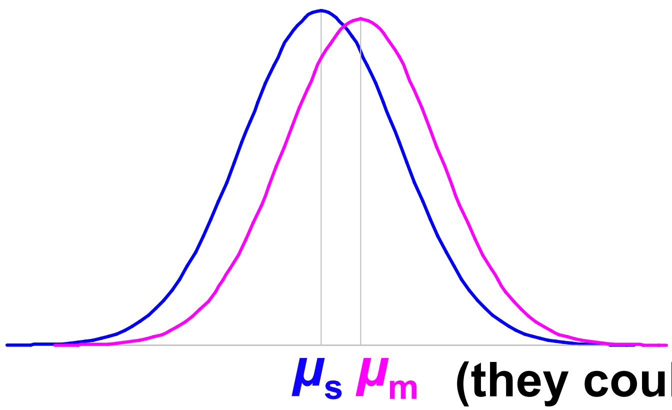
- Testable hypothesis: We measure time it takes for subjects to navigate through a particular VE, under conditions of stereo and mono viewing. We hypothesis subjects will be faster under stereo viewing.

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

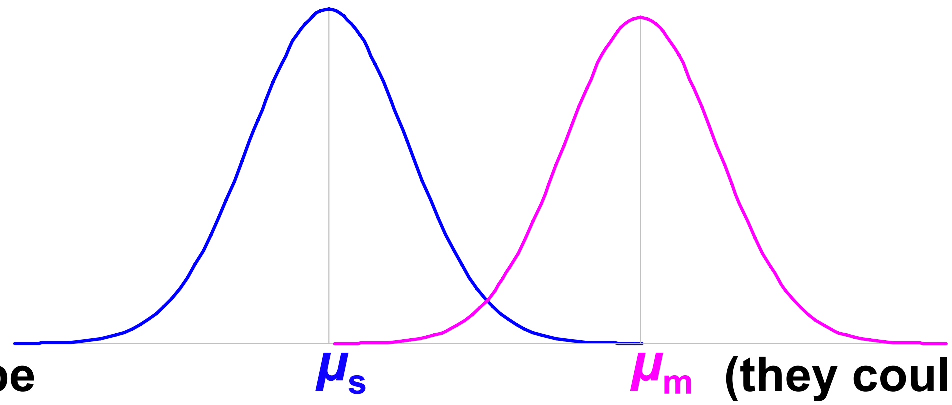
- Some factors effecting experimental validity:
  - Is VE representative of something interesting (e.g., a real-world situation)?
  - Is navigation task representative of something interesting?
  - Is there an underlying theory of human performance that can help predict the results? Could our results 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: $\mu_s - \mu_m = d$**

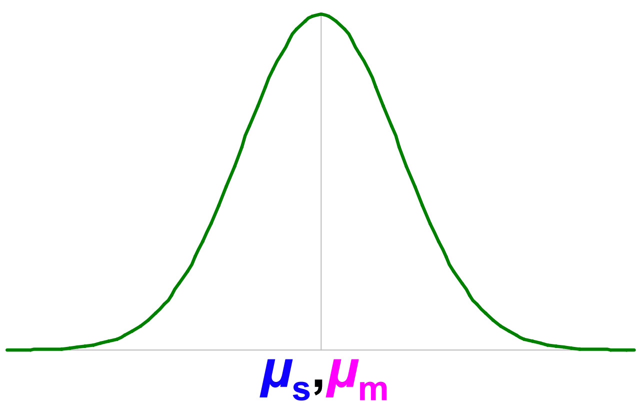


(they could be close together)



(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: $p(X \mid H_0)$.**
   - **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 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.

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 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.

- 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

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. (correct result because if-then false)**

- **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. (correct hypothesis testing reasoning, but incorrect result)**

From [Cohen 94]

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) $p = 1 - \alpha$

- $p(X \mid H_0)$ compared to α, so hypothesis testing involves setting α (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
    - α 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

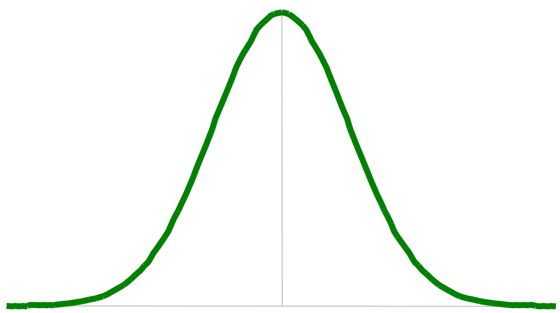
Hypothesis Testing Means

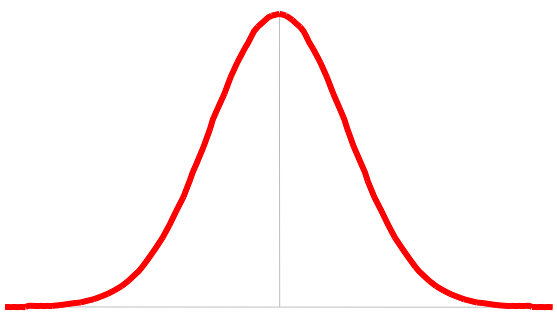
- **Empiricism**
- **Experimental Validity**
- **Experimental Design**
- **Gathering Data**
- **Describing Data**
  - **Graphing Data**
  - **Descriptive Statistics**
- **Inferential Statistics**
  - **Hypothesis Testing**
  - ***Hypothesis Testing Means***
  - **Power**
  - **Analysis of Variance and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

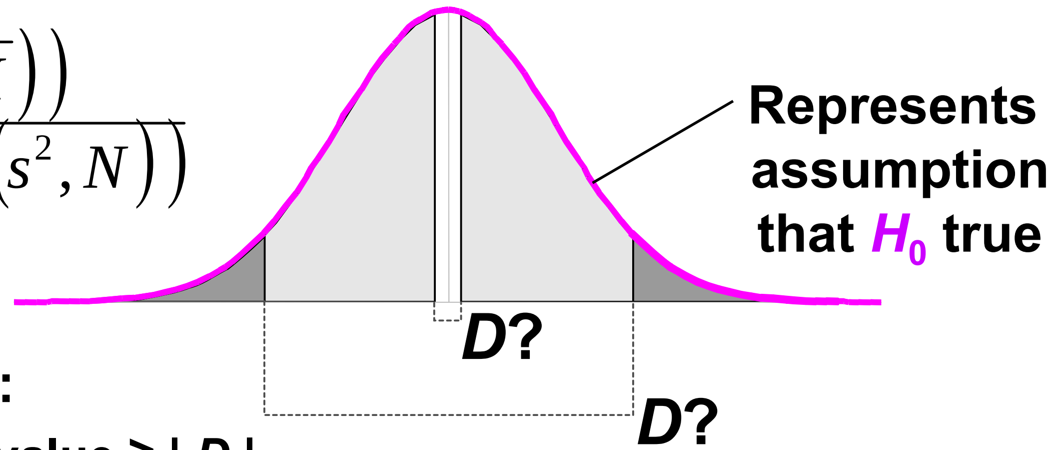




Calculating $\alpha = p(X \mid H_0)$ with A Sampling Distribution

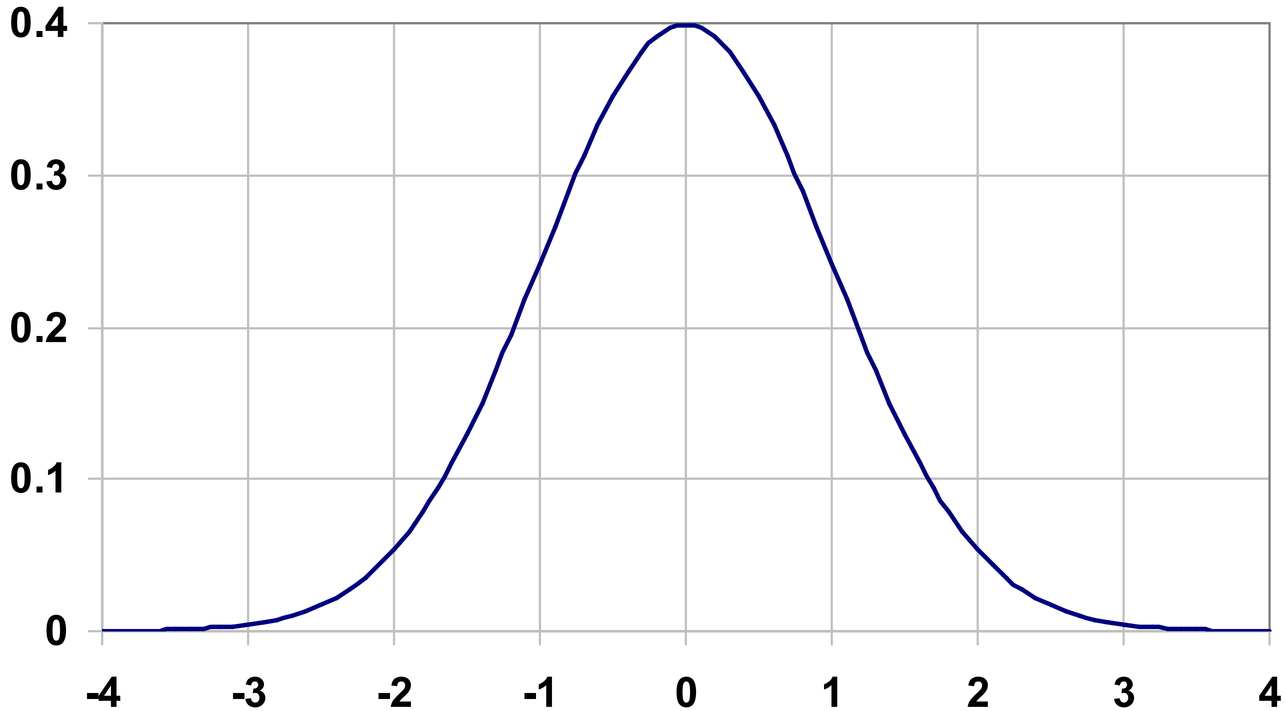
- **Sampling distributions are analytic functions with area 1**
- **To calculate $\alpha = p(X \mid H_0)$ given a distribution, we first calculate the value *D*, which comes from an equation of the form:**

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



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

A Distribution for Hypothesis Testing Means



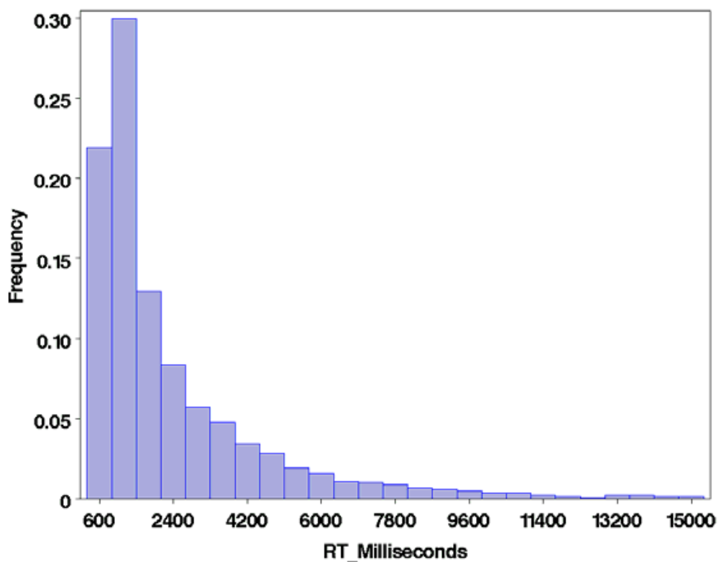
- **The Standard Normal Distribution (μ = 0, σ = 1) (also called the *Z*-distribution):**

$$N(X;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

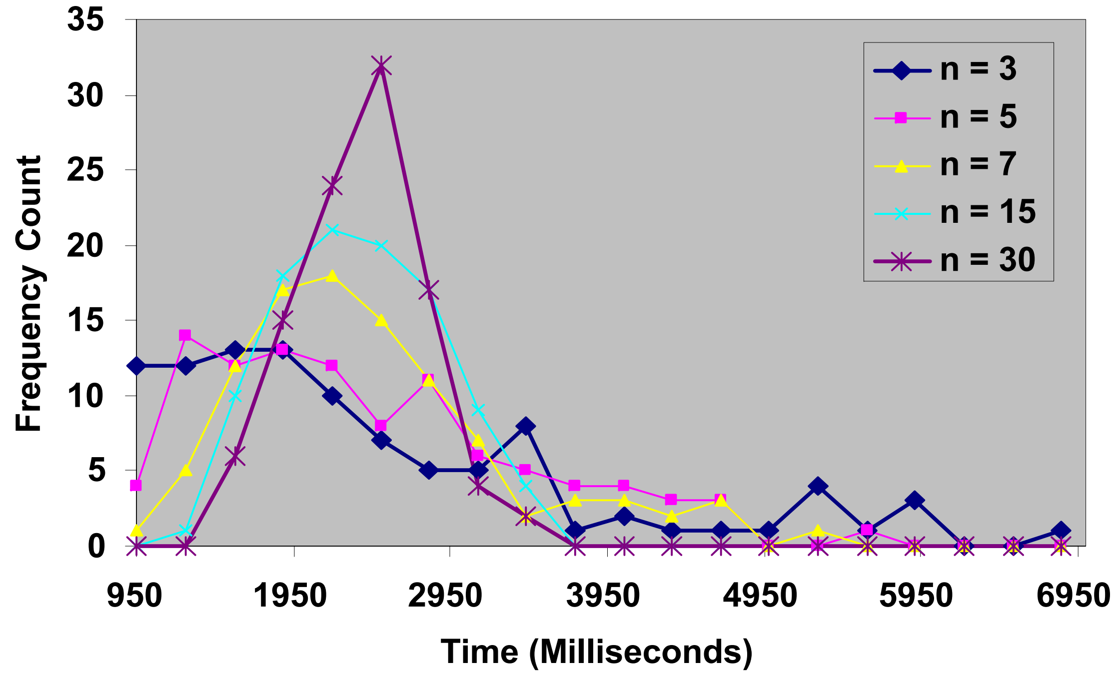
The Central Limit Theorem

- **Full Statement:**
  - Given population with (μ, σ^2), the sampling distribution of means drawn from this population is distributed (μ, σ^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].

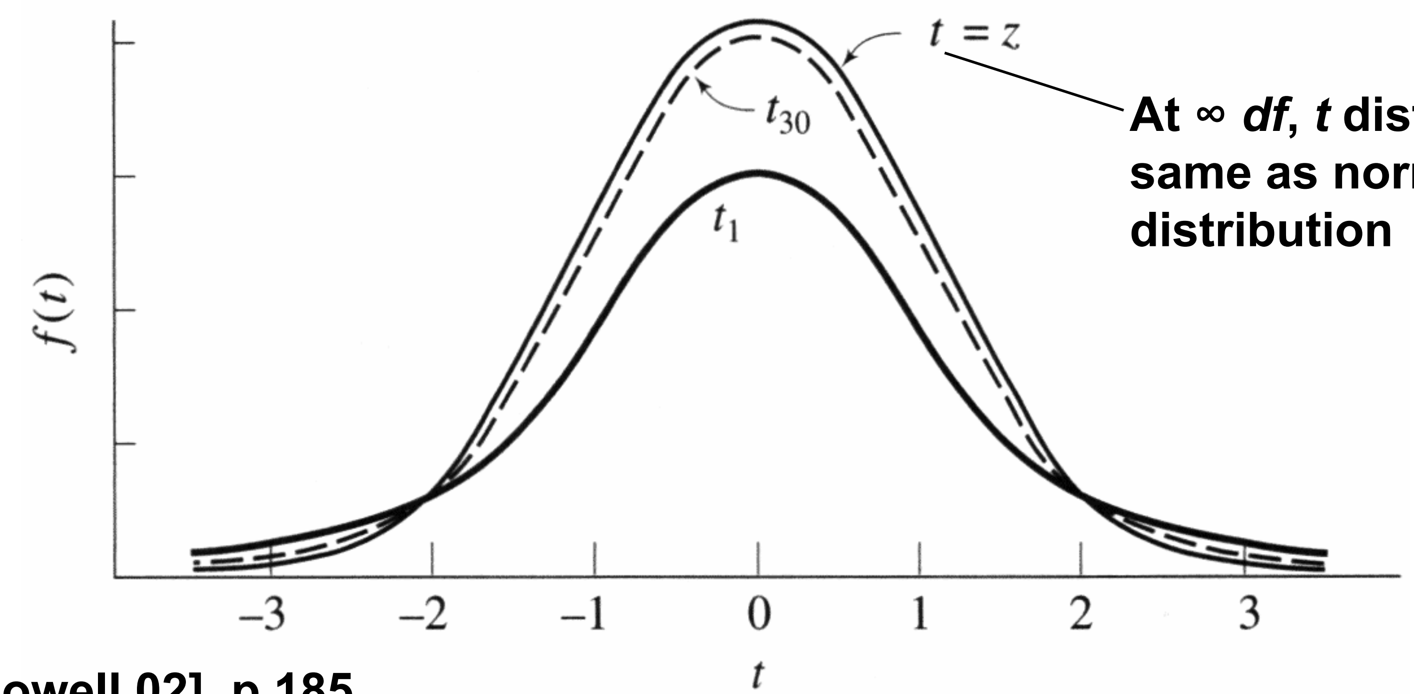


Plotting 100 means drawn from *A* at random without replacement, where *n* is number of samples used to calculate mean.

- 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 $\alpha = p(X \mid H_0)$ from *t* distribution, not *Z* distribution
- Why? *Z* requires the population parameter σ^2, but σ^2 almost never known. We estimate σ^2 with s^2, but s^2 biased to underestimate σ^2. Thus, *t* more spread out than *Z* distribution.
- *t* distribution parametric: parameter is *df* (degrees of freedom)



At ∞ *df*, *t* distribution same as normal distribution

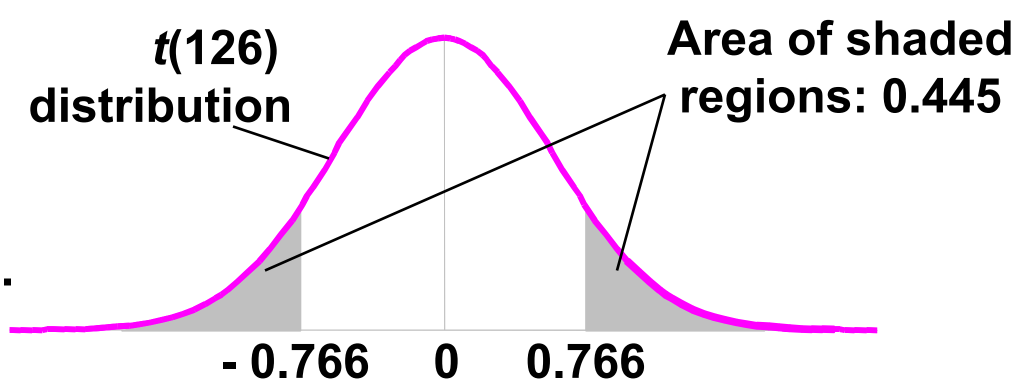
From [Howell 02], p 185

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

$$t(126) = \frac{f\left(\overline{X}\right)}{f\left(s^2, N\right)} = \frac{\overline{X}_s - \overline{X}_m}{\sqrt{s_p^2\left(\frac{1}{n_s} + \frac{1}{n_m}\right)}} = 0.766, \quad s_p^2 = \frac{(n_s - 1)s_s^2 + (n_m - 1)s_m^2}{n_s + n_m - 2}$$

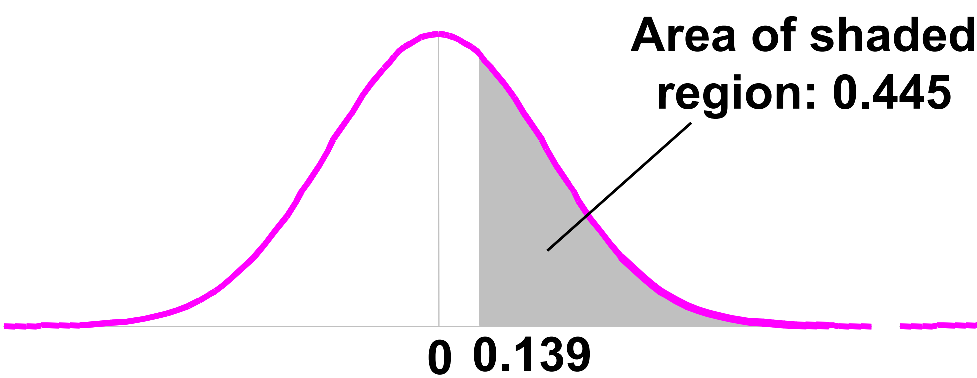
- Look up $t(126) = 0.766$ in a *t*-distribution table: 0.445
- Thus, $\alpha = p(\text{ 1.983 sec} \mid H_0) = 0.445$, and we do not reject H_0.



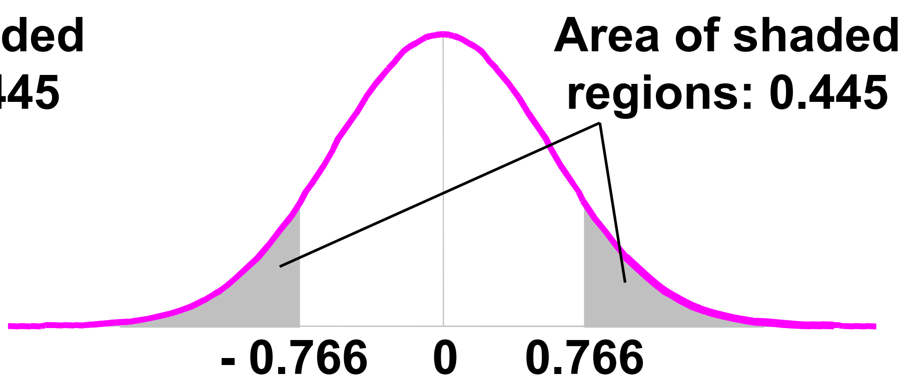
Calculation described by [Howell 02], p 202

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



one-tailed test



two-tailed test

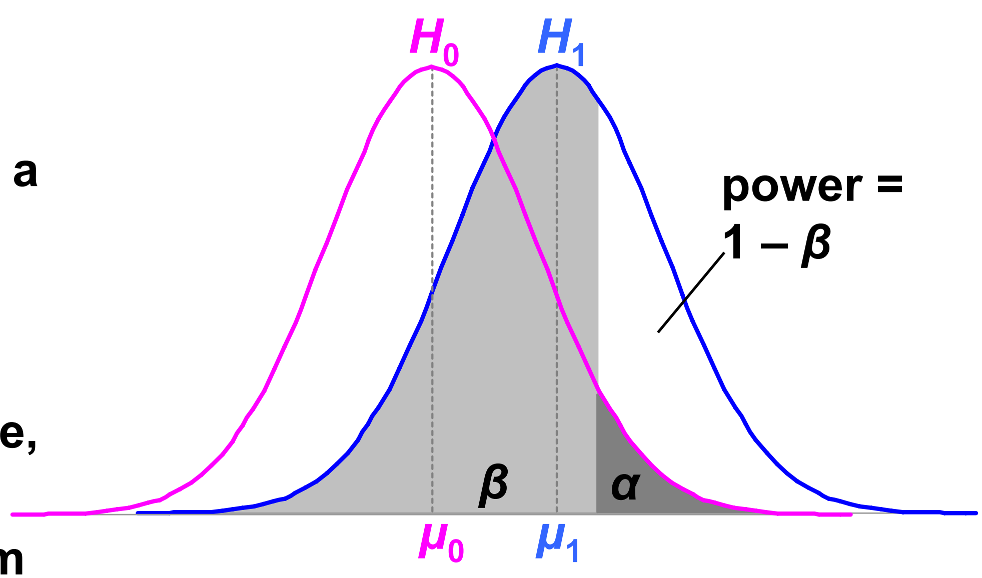
Power

- **Empiricism**
- **Experimental Validity**
- **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 α, β, and Power

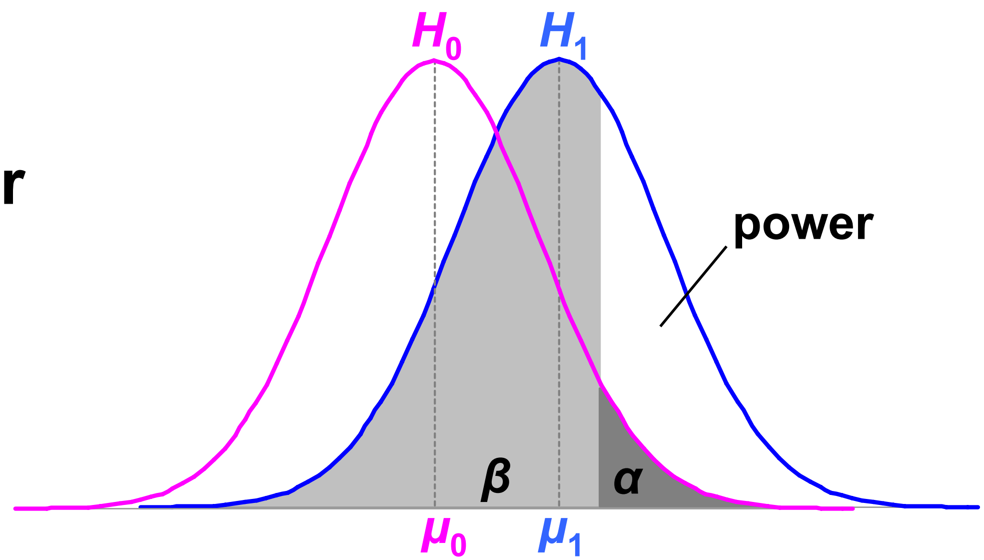
		Decision	
		Reject H_0	Don't reject H_0
True state of the world	H_0 false	a result! $p = 1 - \beta$ = power	type II error $p = \beta$
	H_0 true	type I error $p = \alpha$	wasted time $p = 1 - \alpha$

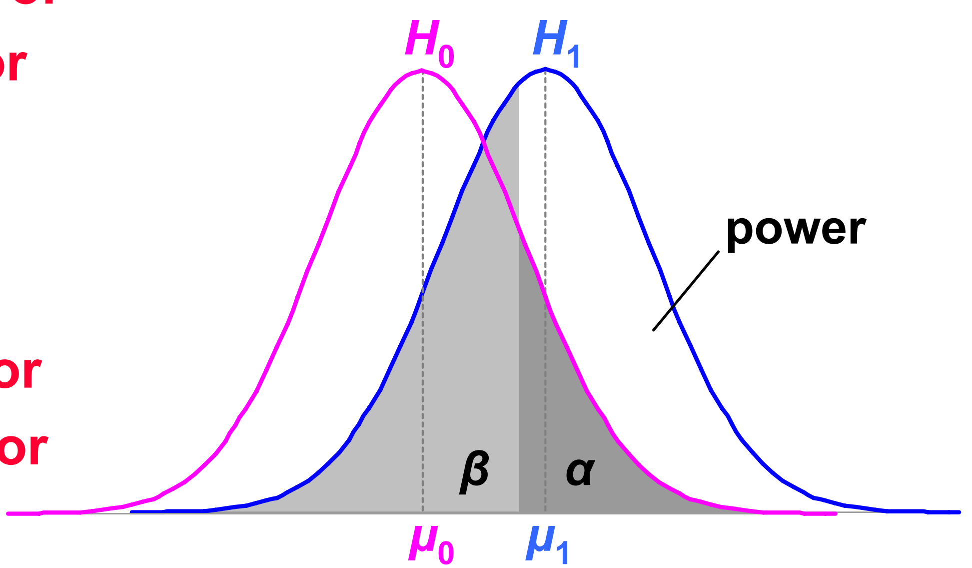
- If H_0 is true:
  - α is probability we make a type I error: we think we have a result, but we are wrong
- If H_1 is true:
  - β is probability we make a type II error: a result was there, but we missed it
  - Power is a more common term than β



Increasing Power by Increasing α

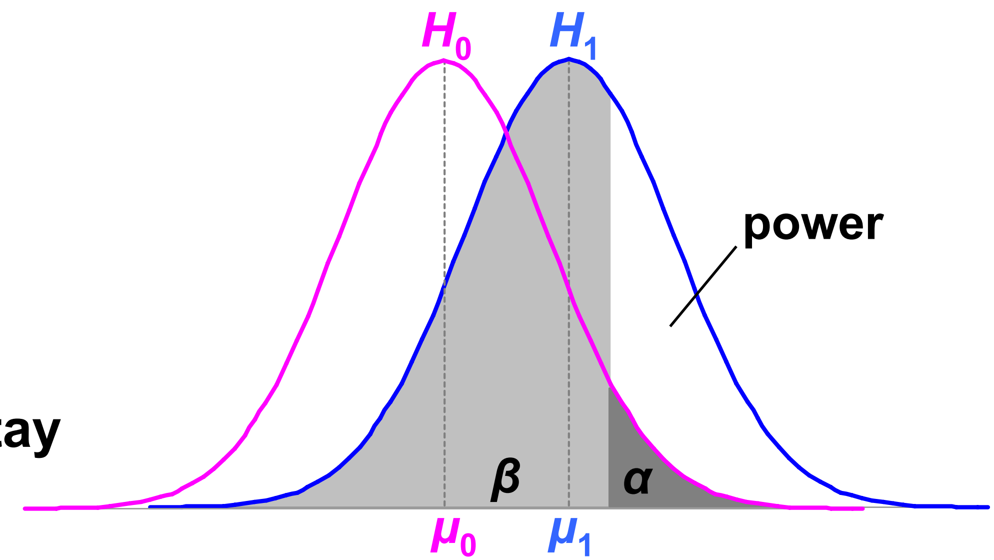
- **Illustrates α / power tradeoff**
- **Increasing α:**
  - Increases power
  - Decreases type II error
  - Increases type I error
- **Decreasing α:**
  - Decreases power
  - Increases type II error
  - Decreases type I error



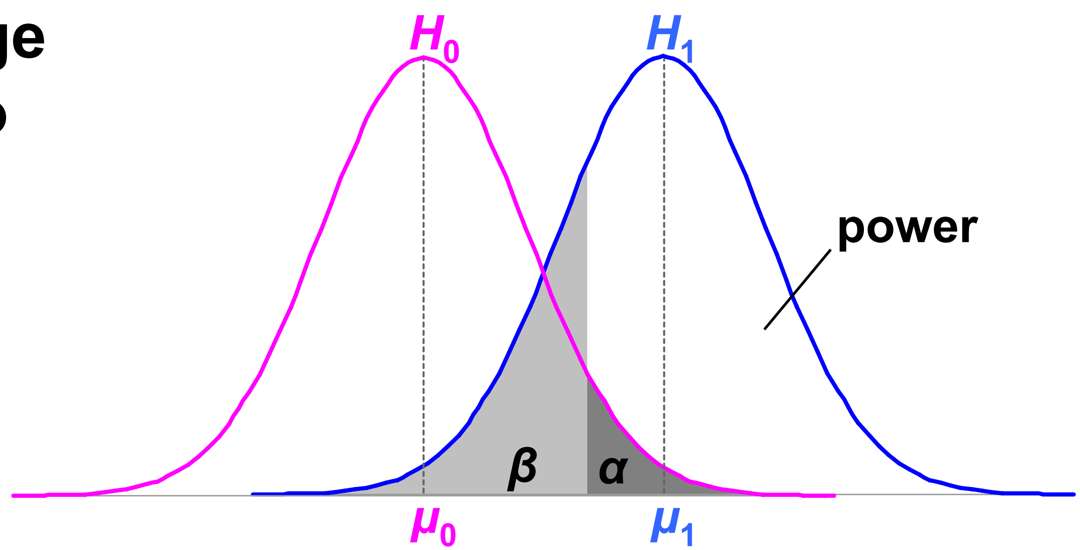


Increasing Power by Measuring a Bigger Effect

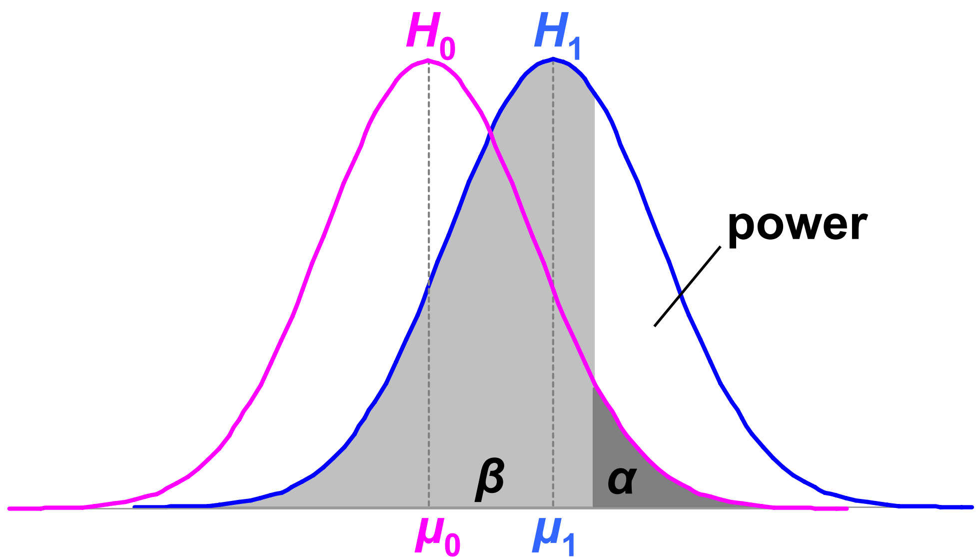
- If the effect size is large:
  - Power increases
  - Type II error decreases
  - α and type I error stay the same



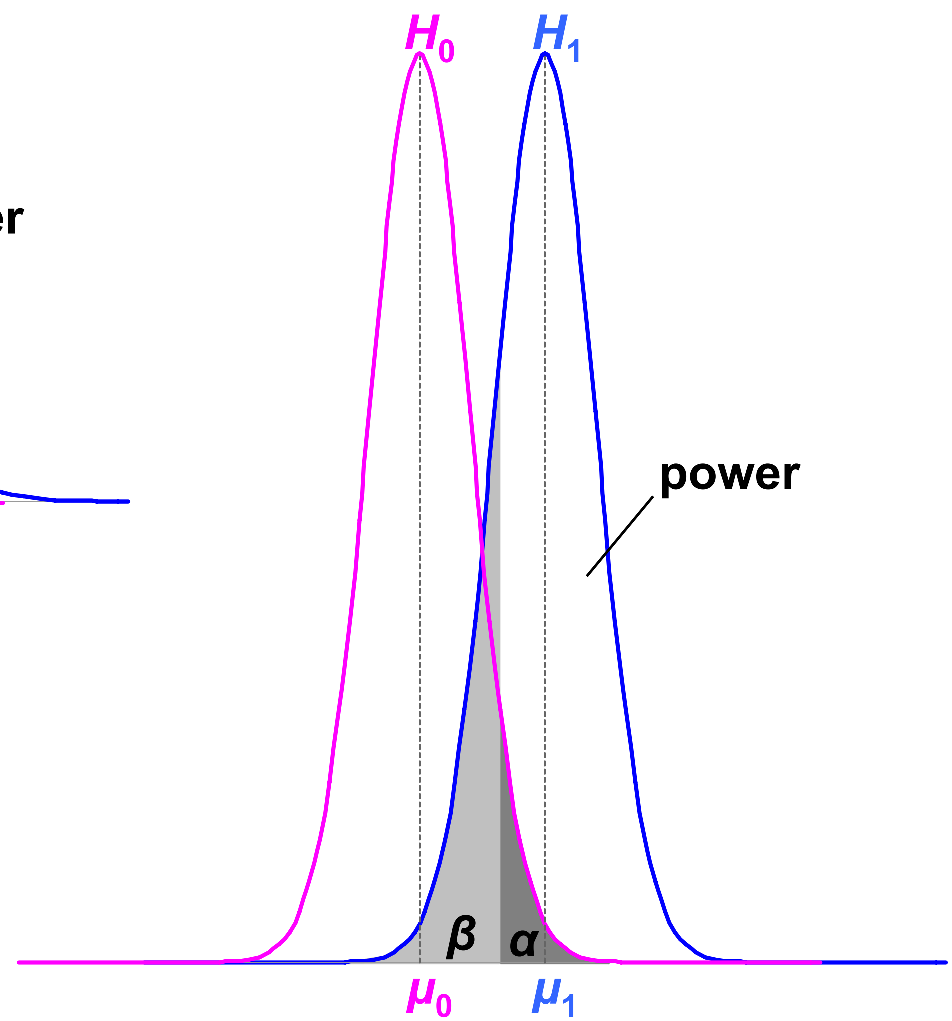
- Unsurprisingly, large effects are easier to detect than small effects



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.

Using Power

- **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:**

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

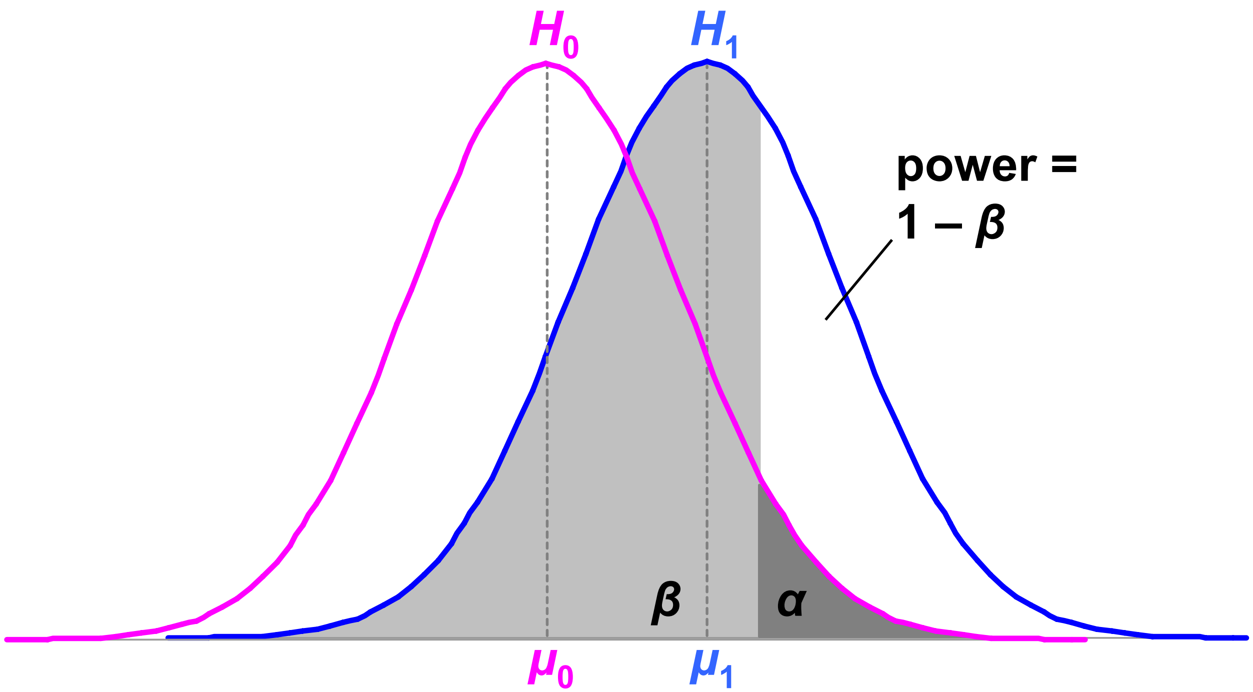
3. **Significance level needed:**

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

(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]

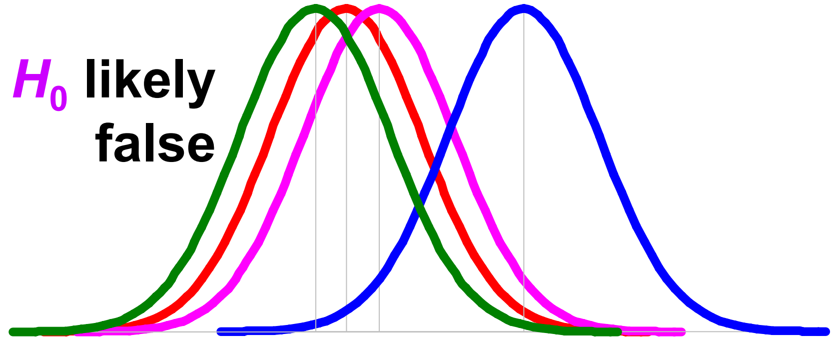
Analysis of Variance and Factorial Experiments

- **Empiricism**
- **Experimental Validity**
- **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**

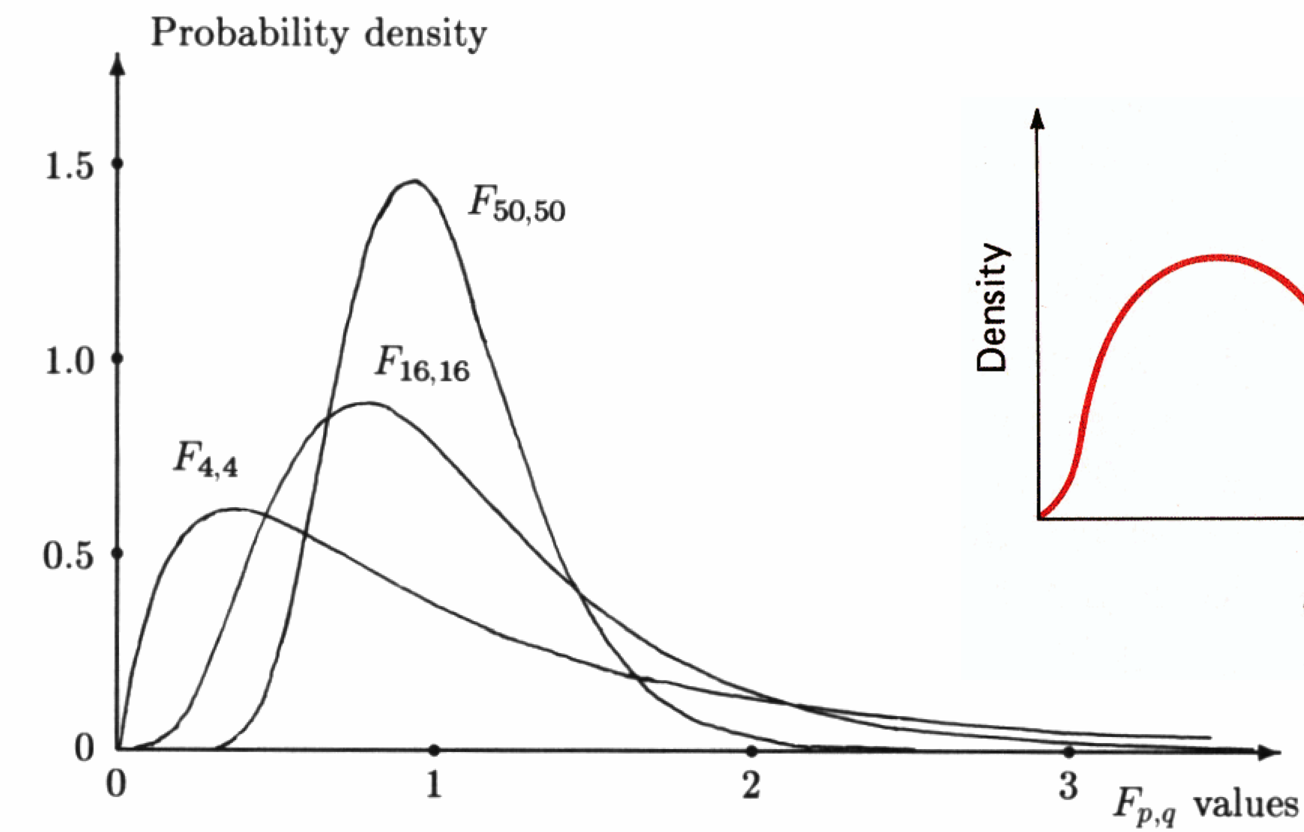
How ANOVA Works

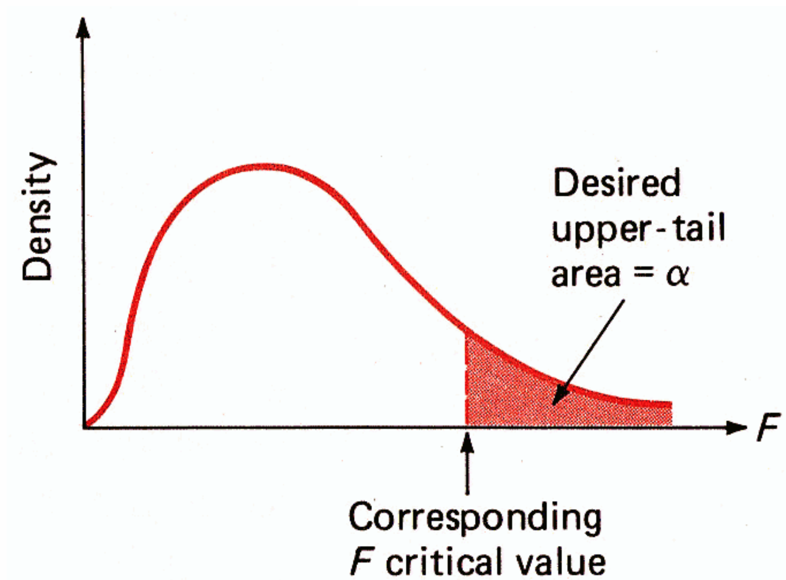


- Null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; H_1: at least one mean differs
- Estimate variance between each group: $MS_{between}$
  - Based on the difference between group means
  - If H_0 is true, accurate estimation
  - If H_0 is false, biased estimation: overestimates variance
- Estimate 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 Uses The *F* Distribution

- **Calculate $\alpha = p(X \mid H_0)$ by looking up *F* critical value in *F*-distribution table**
- ***F*-distribution parametric: *F* (numerator *df*, denominator *df*)**
- **α is area to right of *F* critical value (one-tailed test)**
- ***F* and *t* are distributions are related: $F(1, q) = t(q)^2$**

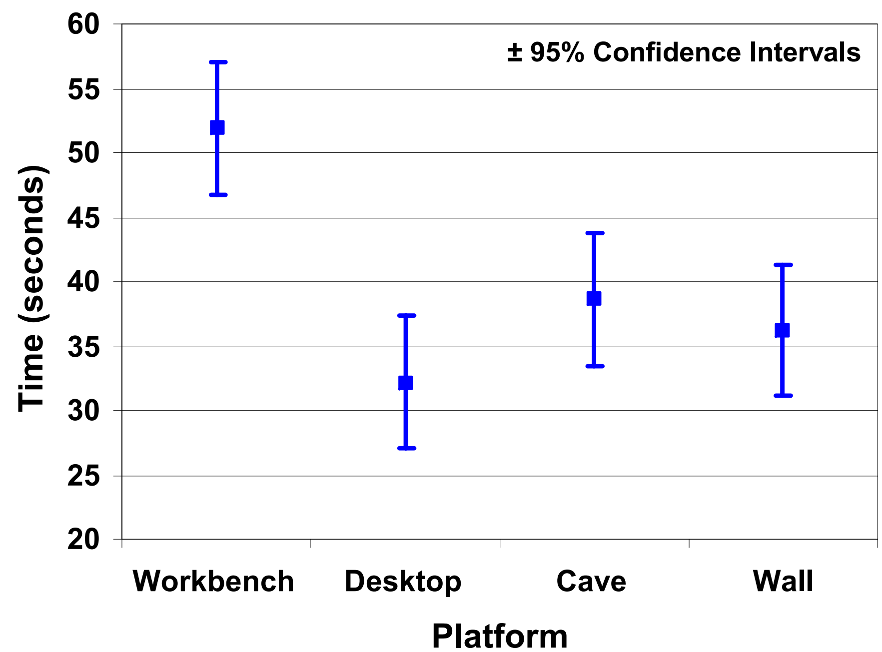




From [Saville Wood 91], p 52, and [Devore Peck 86], p 563

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.



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		

*$p < .05$

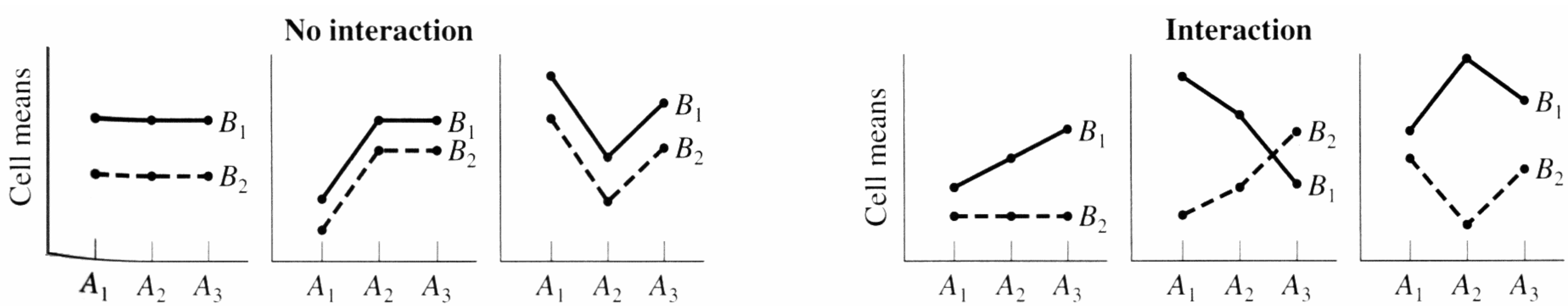
- Reporting in a paper: $F(3, 93) = 3.1$, $p < .05$

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

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**



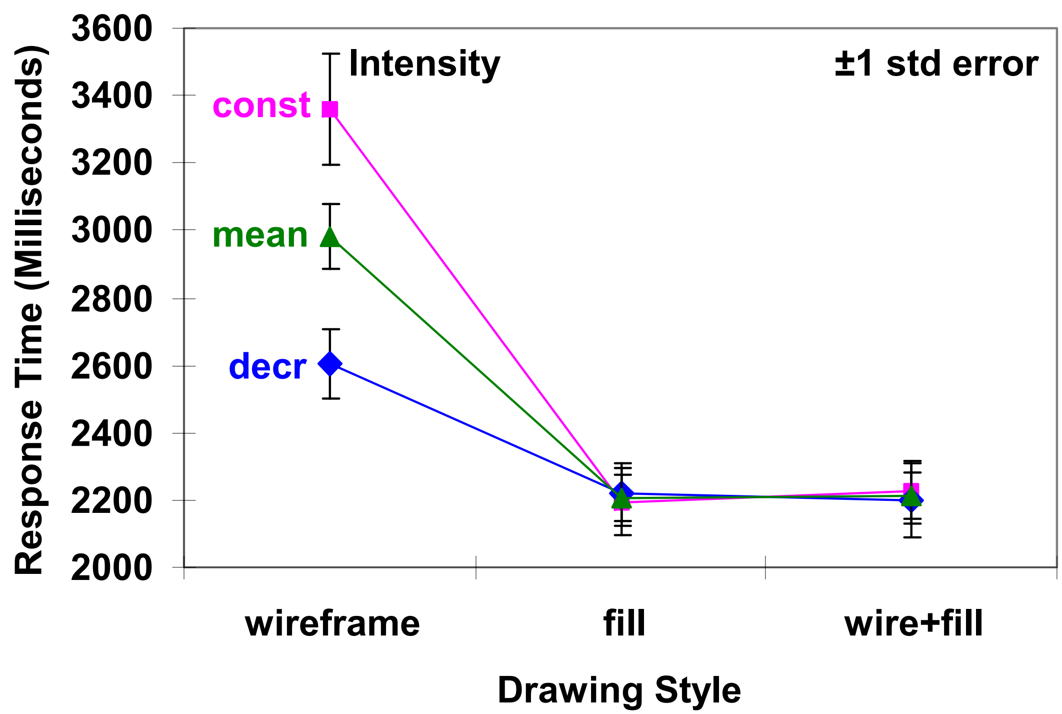
From [Howell 02], p 431

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)
  - $F(8,12) = 5.826403$, $p = 3.4778689e10\text{-}3$ (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

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

[Cohen 94] J Cohen, "*The Earth is Round (*p *< .05)*", American Psychologist, 49(12), pages 997–1003.

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

[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.

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

[Meehan et al. 03] M Meehan, S Razzaque, MC Whitton, FP Brooks, Jr., "*Effect of Latency on Presence in Stressful Virtual Environments*", Technical Papers, IEEE Virtual Reality 2003, March 22–26, Los Angeles, California: IEEE Computer Society, 2003, pages 141–148.

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

[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.

Contact Information

J. Edward Swan II, Ph.D.

Associate Professor

Department of Computer Science and Engineering

swan@acm.org

(662) 325-7507

Slide Location:

http://www.cse.msstate.edu/~swan/teaching/tutorials/Swan-VR2007-Tutorial.pdf

Mississippi State UNIVERSITY

Classical and Other Psychophysical Methods for Virtual Environments

Bernard D. (Dov) Adelstein, Ph.D.
(Bernard.D.Adelstein@nasa.gov)
Advanced Controls & Displays Group
NASA Ames Research Center

Human Systems Integration Division

Outline

- Motivation: VE Latency/Asynchrony characterization
- Psychophysics: What and why?
- Classical methods of psychophysics
  - Method of Constant Stimuli
    - Detection theory
  - Method of Limits
    - Up-Down procedures
- Adaptive methods of psychophysics
- Psychometric function

Illustrations from NASA-Ames studies

Temporal & Spatial Imperfection in (Visual) VEs

Excessive time delay and insufficient frame (update) rate

- Poor dynamic registration, dynamic instability
- "Sloshiness," jumpiness in response to observer motion
  Whole image lags in response to head motion

Systematic and random error in spatial measurement

- Poor static registration wrt external world
- VE image jitter

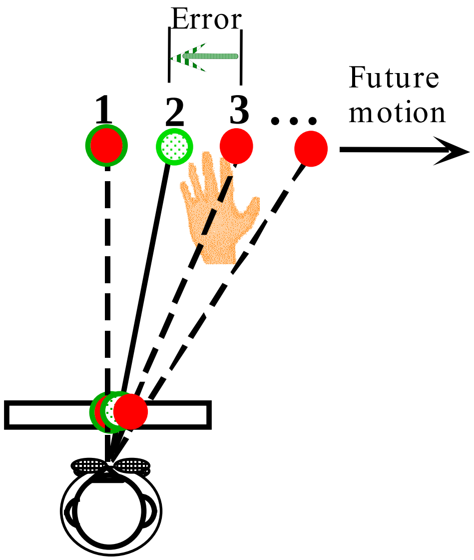
Degraded motor and perceptual performance

- Diminished interactivity, immersion, & sense of "presence"
- "Cybersickness"

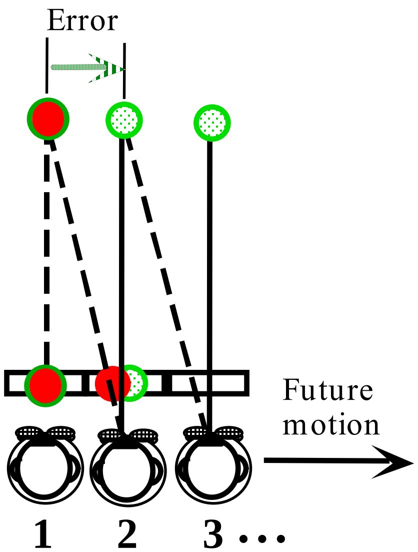
Latency Induced Rendering Errors

6 Frames Delay
20 Hz Update Rate
~ 380 ms Latency

Hand Translation



Head Translation



Latency/Asynchrony Studies
ADSP/ACD Group

•Phenomenon: Tracking & tracing performance (latency & update rate)

•First quantification of VE head and hand latency perception MoCS

•Compensation techniques: perceptually based design validation MoCS

•Latency perception mechanism: direct time vs. image "slip" MoL

•How we perceive image "slip": displacement vs velocity AS

•Generalizability of perceptual threshold quantification AS

•Why we perceive image "slip" *velocity*

•Haptic-audio asynchrony thresholds AS

Latency/Asynchrony Studies
ADSP/ACD Group

- •Phenomenon: Tracking & tracing performance (latency & update rate)
- •First quantification of VE head and hand latency perception MoCS
- •Compensation techniques: perceptually based design validation MoCS
- •Latency perception mechanism: direct time vs. image "slip" MoL
- •How we perceive image "slip": displacement vs velocity AS
- •Generalizability of perceptual threshold quantification AS
- •Why we perceive image "slip" *velocity*

- •Haptic-audio asynchrony thresholds AS

Definition

- Psychophysics:
  - Area of psychology that employs specific behavioral methods to study the relation between the physical world and subjective experience (after S. Lederman)
  - Quantitative evaluation of **perceptual** characteristics (e.g., sensitivity) as a function of physical stimulus parameters
    - ***Empirical***, *Analytical*, *Theoretical*

The Questions

- What is it?
  - A priori; qualitative
- Is it there?
  - Absolute threshold (RL)
- How different is it [than standard]?
  - Differential threshold (DL)
- How much is there?
  - Magnitude estimation

Why Psychophysics for/in VE?

- Quantify perceptual tolerances that are relevant to Virtual Environment (VE) system use
  - Establish guidelines and specifications for the design, implementation, and effective deployment of VE systems and interfaces
- Ultimately, to use appropriately implemented and well calibrated VE systems to rapidly prototype psychophysical (and other performance) studies
- We want to measure human performance, not system artifact!

(Classical) Psychophysical Methods

- Method of Adjustment
- Method of Constant Stimuli
- Method of Limits
  - Staircases
  - Up-down Staircases
  - Adaptive Staircases

Method of Adjustment

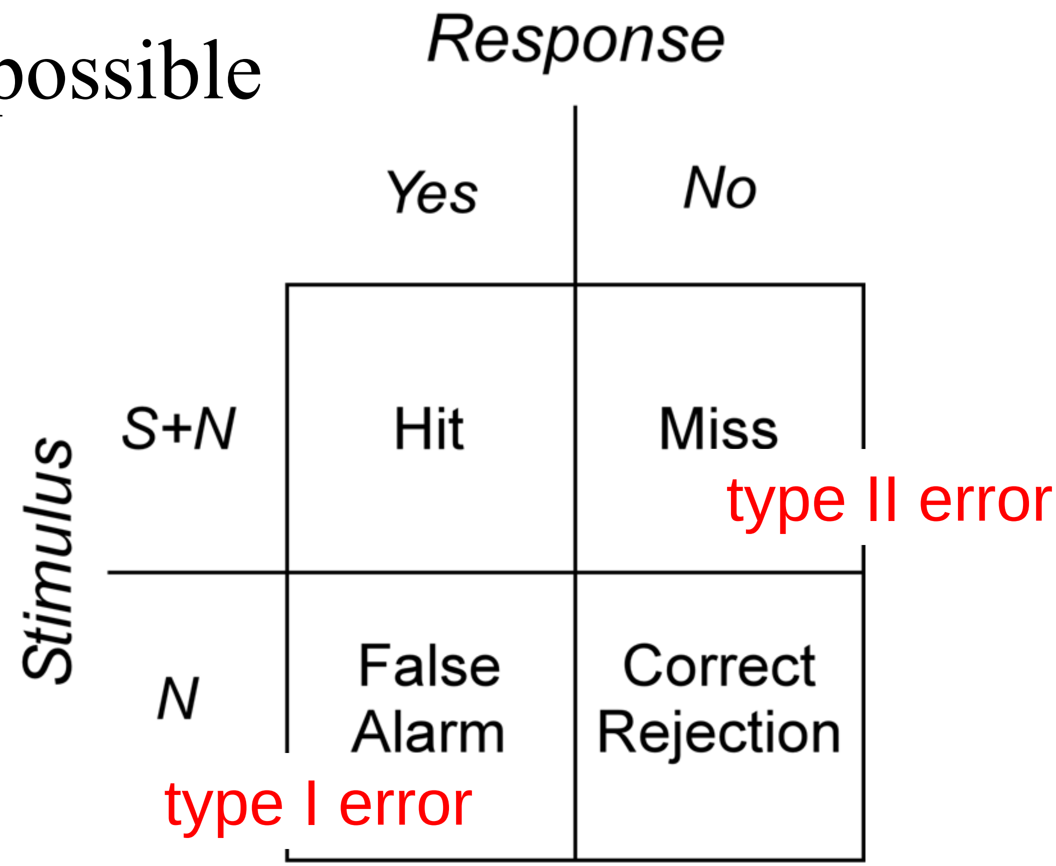
- Observer adjusts a stimulus
  - to exceed a threshold (*RL*): absolute threshold
  - to match a standard (*DL*): difference threshold
- Example
  - Manually (literally or figuratively) adjust an apparatus setting (e.g., by turning a knob) until a temporal or spatial (or other intensity) separation is (or is no longer) {heard|felt|seen} between sequentially presented stimuli

Method of Constant Stimuli

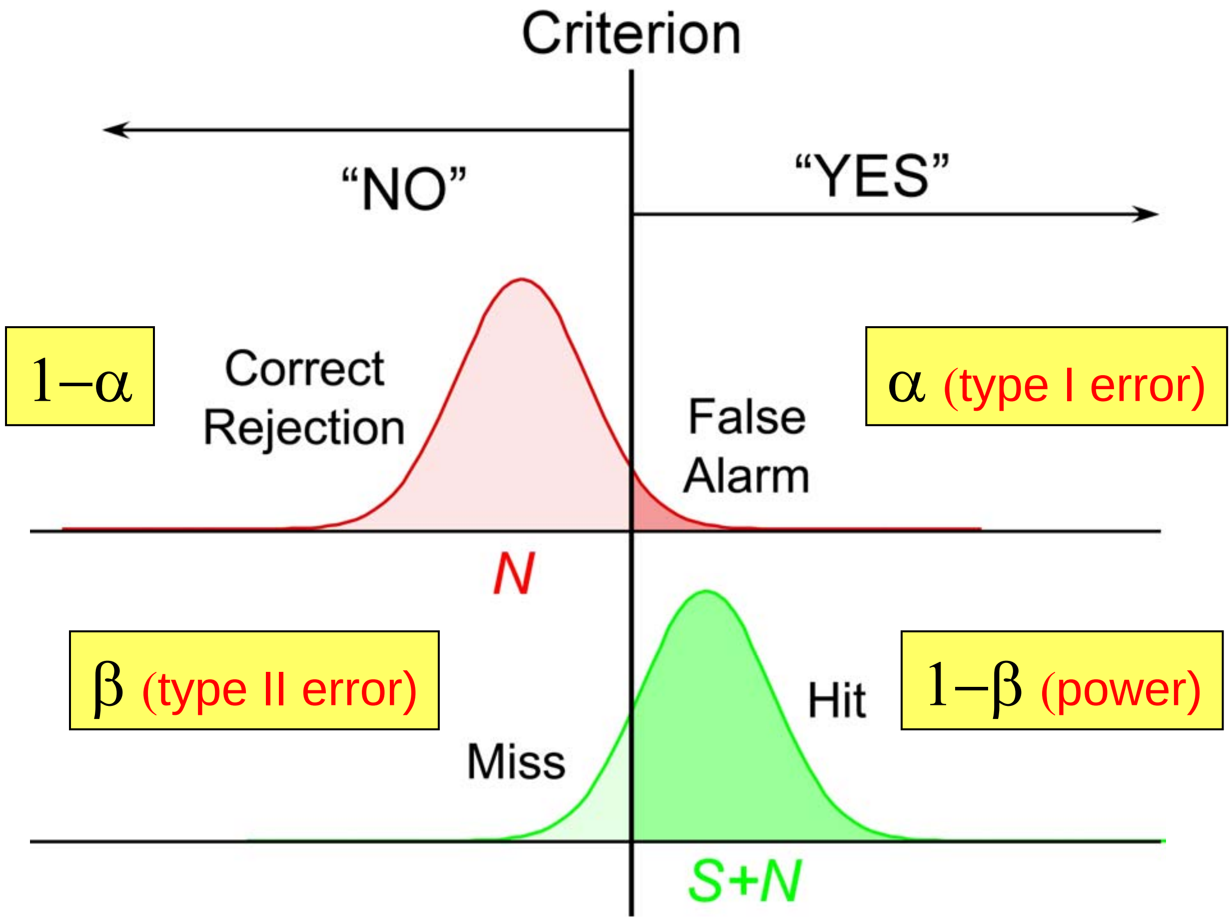
- Intervals presented
  - N (noise)
    - absence of stimulus, reference condition, standard
  - $S+N$ (signal plus noise)
    - stimulus, probe condition
- Depending on the stimulus type, intervals are presented
  - individually (single interval) for absolute threshold (RL)
    - yes|no response
  - pairs (two-interval)
    - simultaneously in adjacent locations, sequentially in same or adjacent location
    - which interval is bigger|smaller?
  - n-interval

Method of Constant Stimuli

- Response: Two Alternative Force Choice (2AFC)
- Q: Is signal (*S*) present?
- Other designs are possible



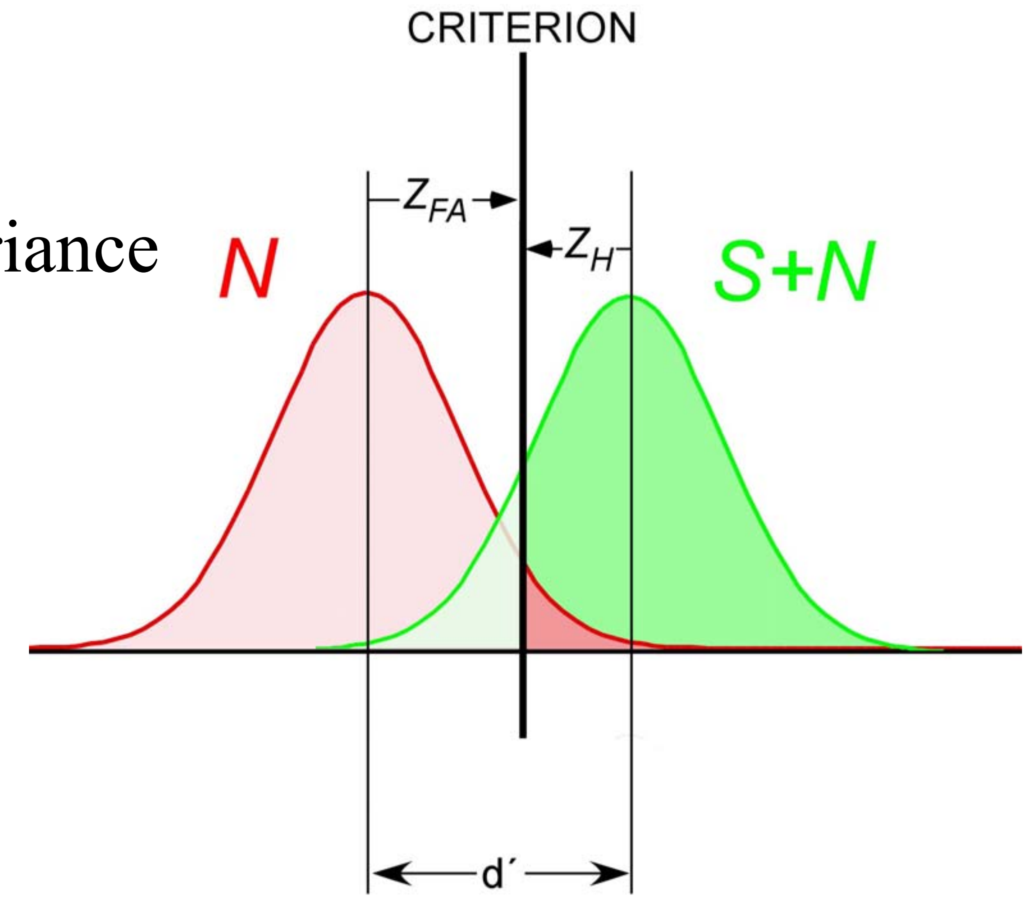
Detection Theory: Internal Response



- Criterion is individual observer's preference or bias; depends on cost/pay-off

Discriminability: *d′* (d-prime)

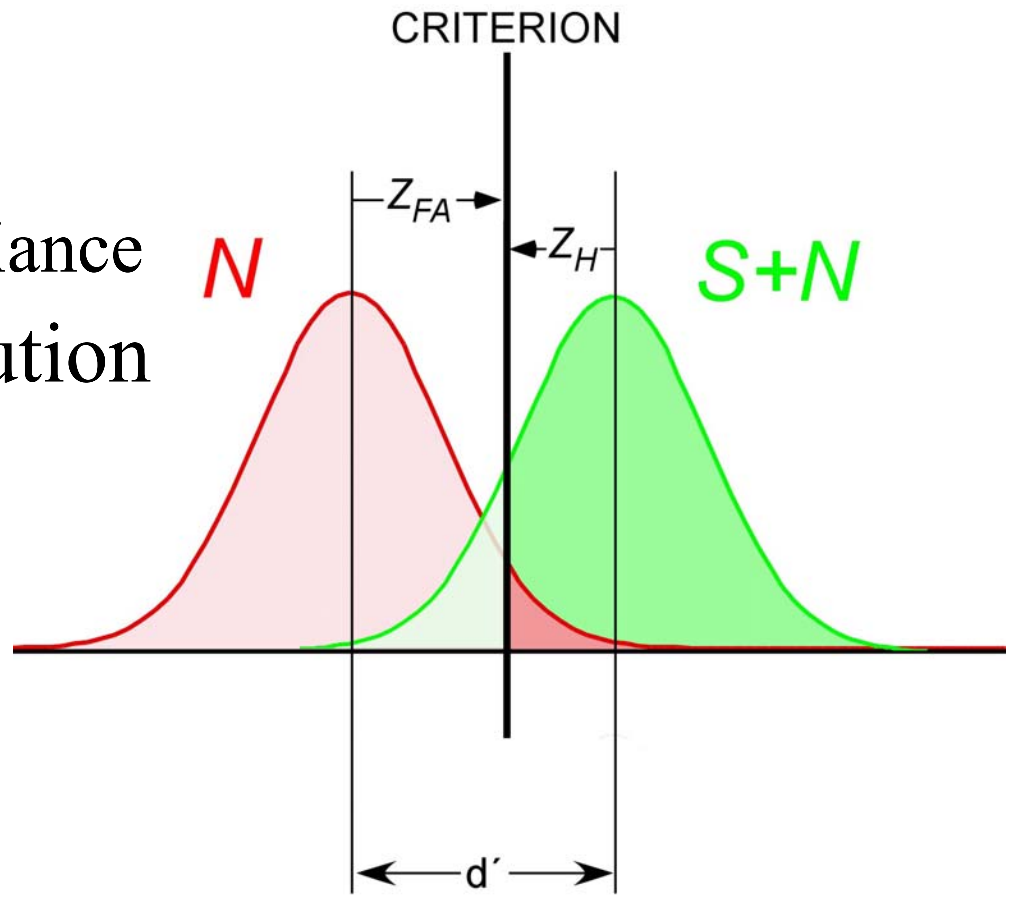
- Assumptions
  - *N*, *S*+*N* are Gaussian
  - *N*, *S*+*N* have equal variance



Discriminability: *d′* (d-prime)

- Assumptions
  - *N*, *S*+*N* are Gaussian
  - *N*, *S*+*N* have equal variance
- For Z (normal) distribution ($\sigma_n = 1$):

$$d' = Z_H - Z_{FA}$$



Discriminability: *d′* (d-prime)

- Assumptions

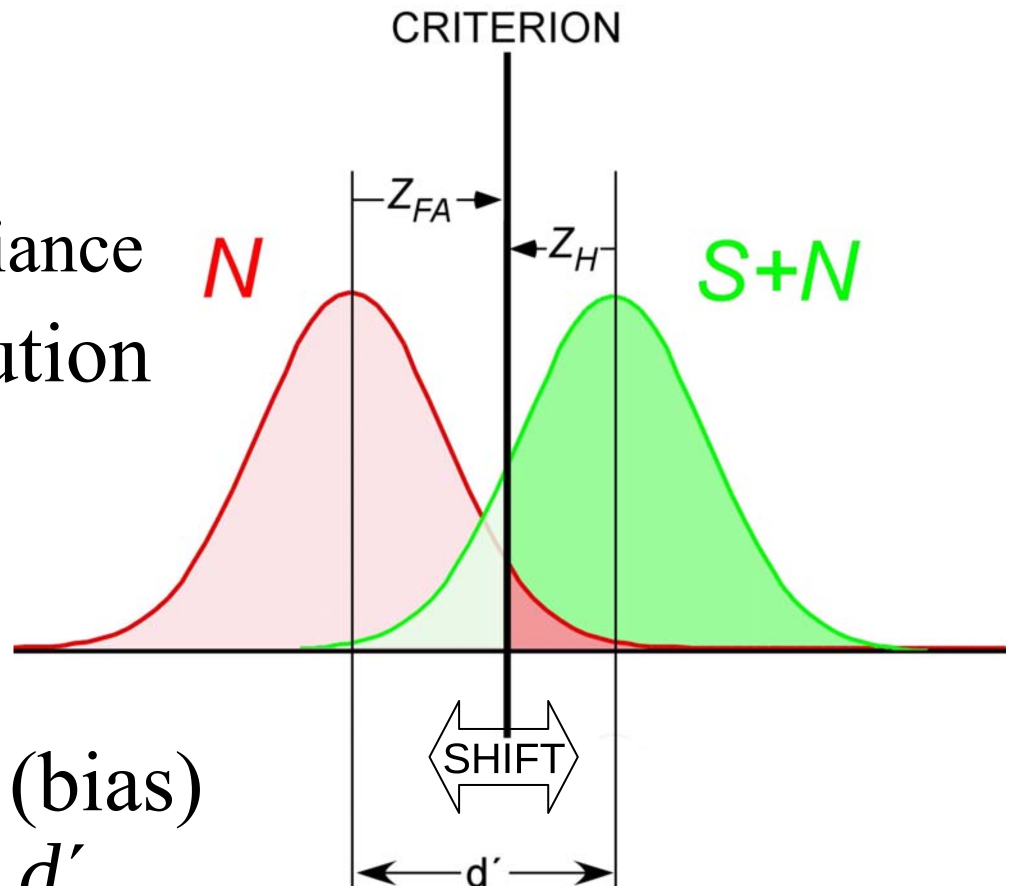
  N, *S*+*N* are Gaussian

  N, *S*+*N* have equal variance

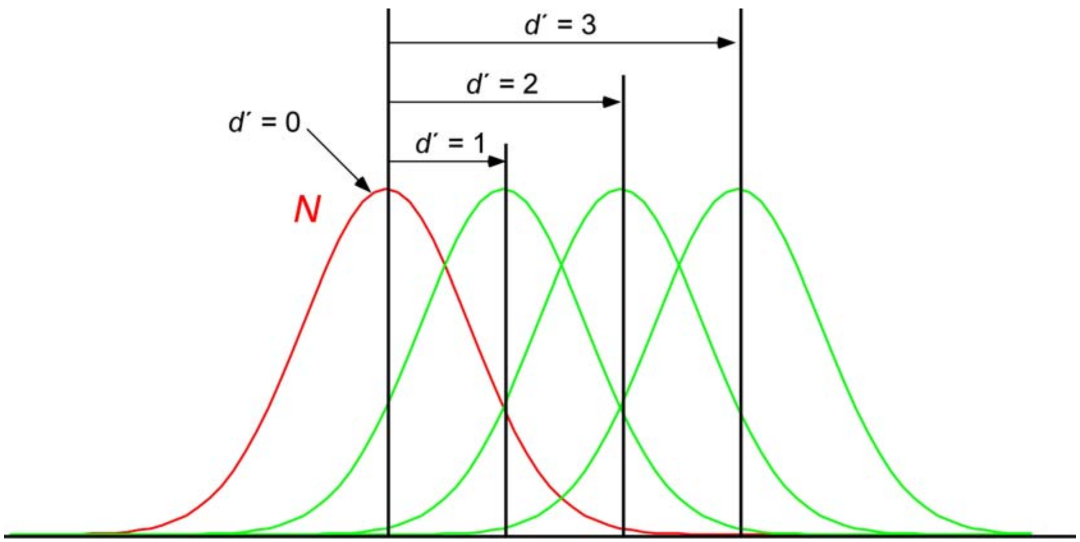
- For Z (normal) distribution ($\sigma_n = 1$):

$$d' = Z_H - Z_{FA}$$

- Possibility of criterion (bias) shift with constant *d′*

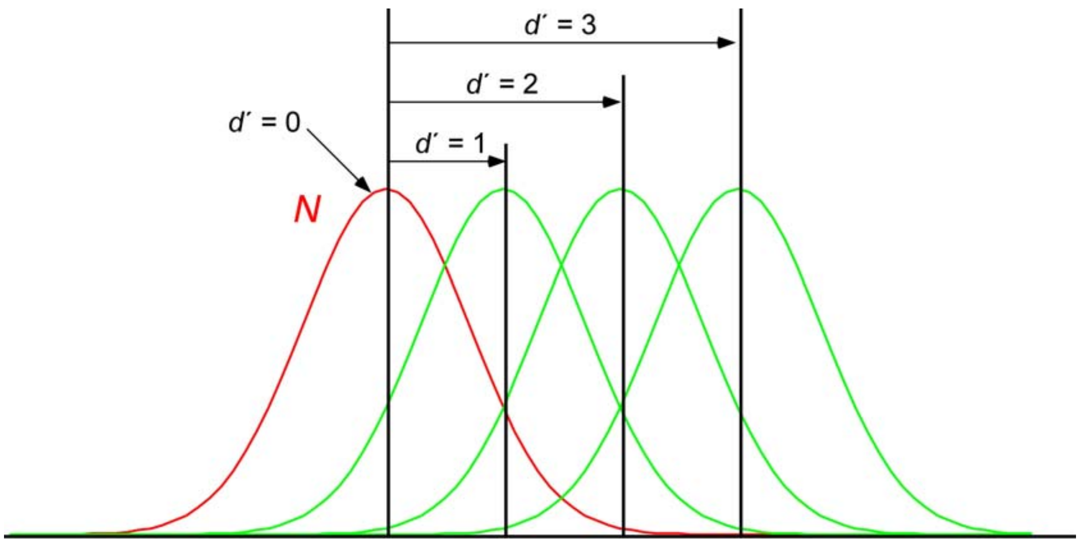


ROC: Receiver Operating Characteristic
(AKA Relative Operating Characteristic)



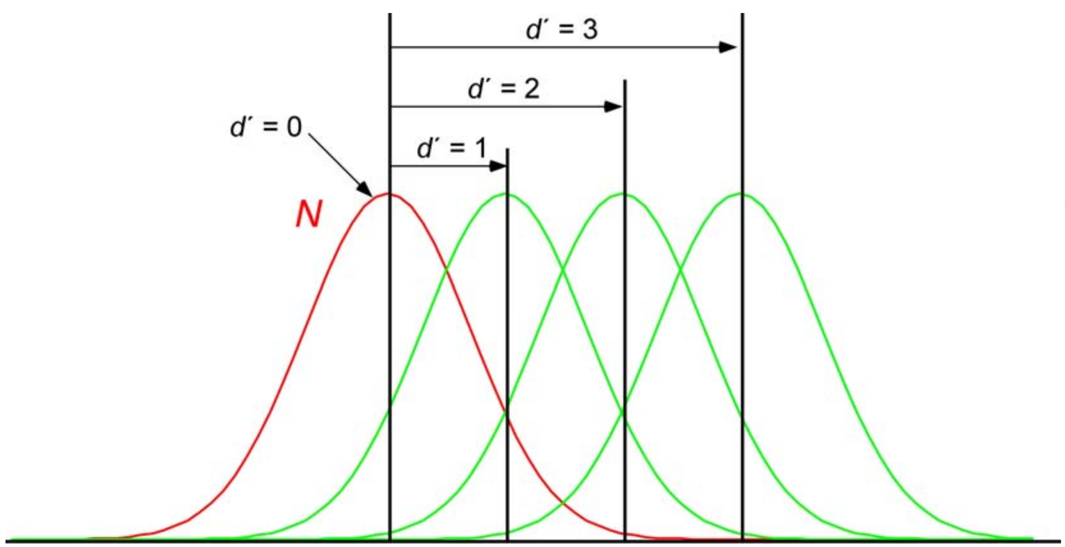
Increased d' →
improved discriminability

ROC: Receiver Operating Characteristic
(AKA Relative Operating Characteristic)

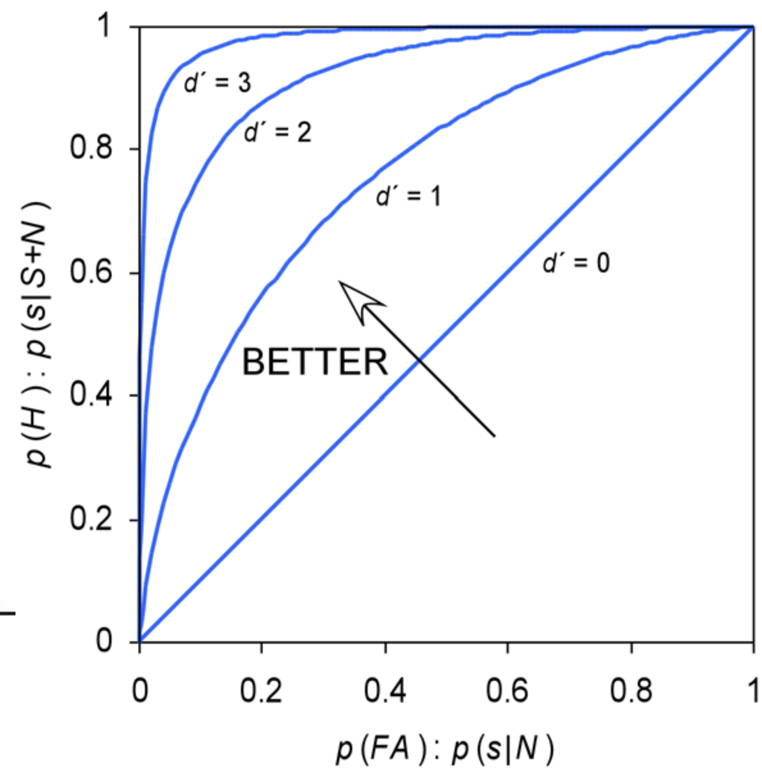


Increased d' →
improved discriminability
(threshold $d' > 1$)

ROC: Receiver Operating Characteristic
(AKA Relative Operating Characteristic)

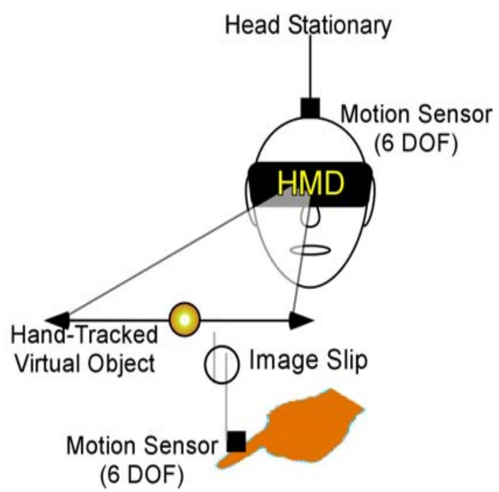


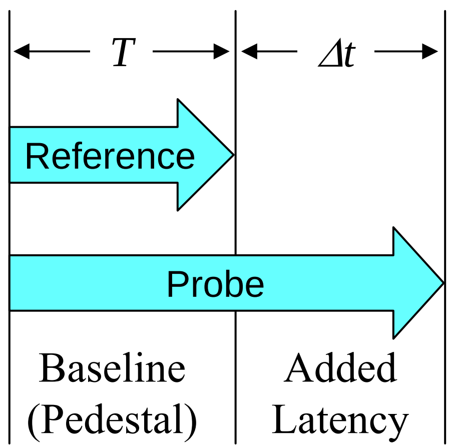
Increased d' →
improved discriminability
(threshold $d' > 1$)



ROC: $p(H)$ vs $p(FA)$
as function of d'

Example: Latency Discrimination
Constant Stimuli Experiment [1]





Q (2AFC): Same or different?

T = {33, 100, 200} ms
Δt = {16.7, 33.3,…,116.7} ms

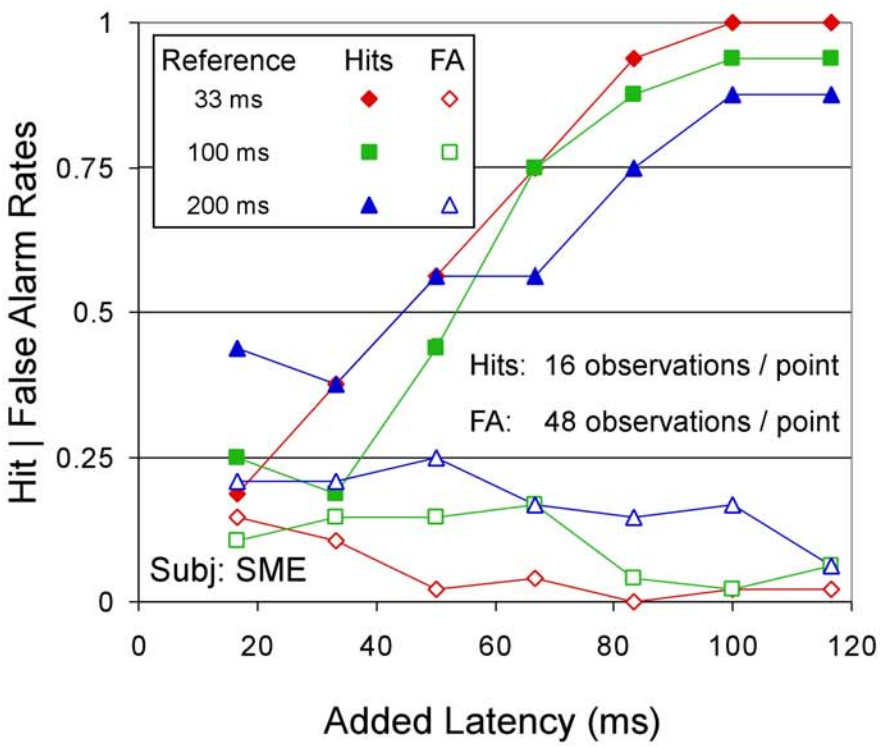
Experiment Factors (3 X 7 levels)

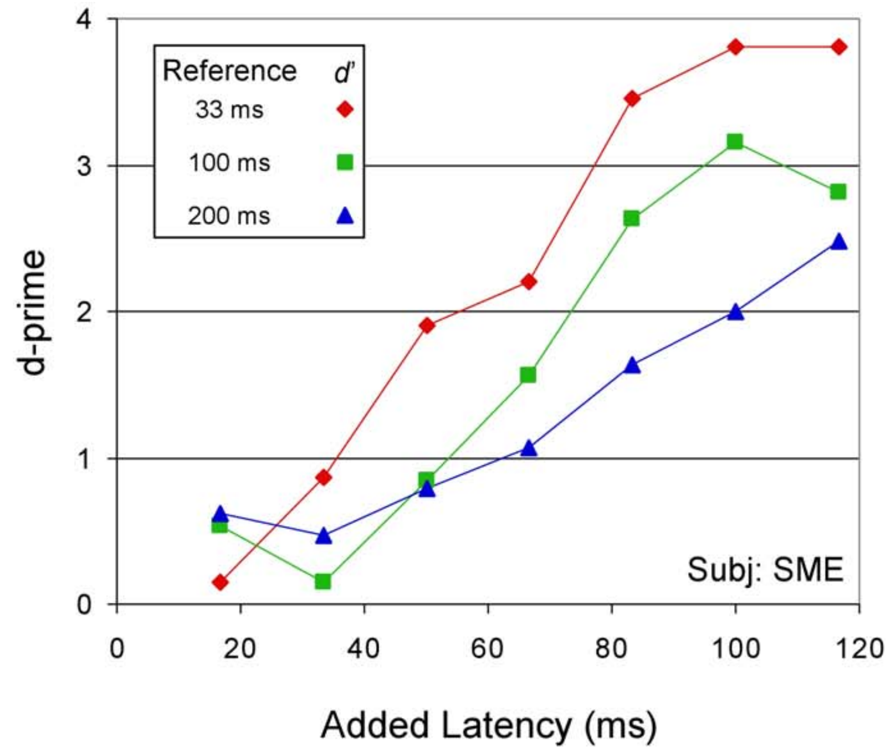
1st Condition \ 2nd Condition	T	$T+\Delta t$
$T+\Delta t$	.125 S	.375 N
T	.375 N	.125 S

Randomized Stimulus Block

Example: Latency Discrimination

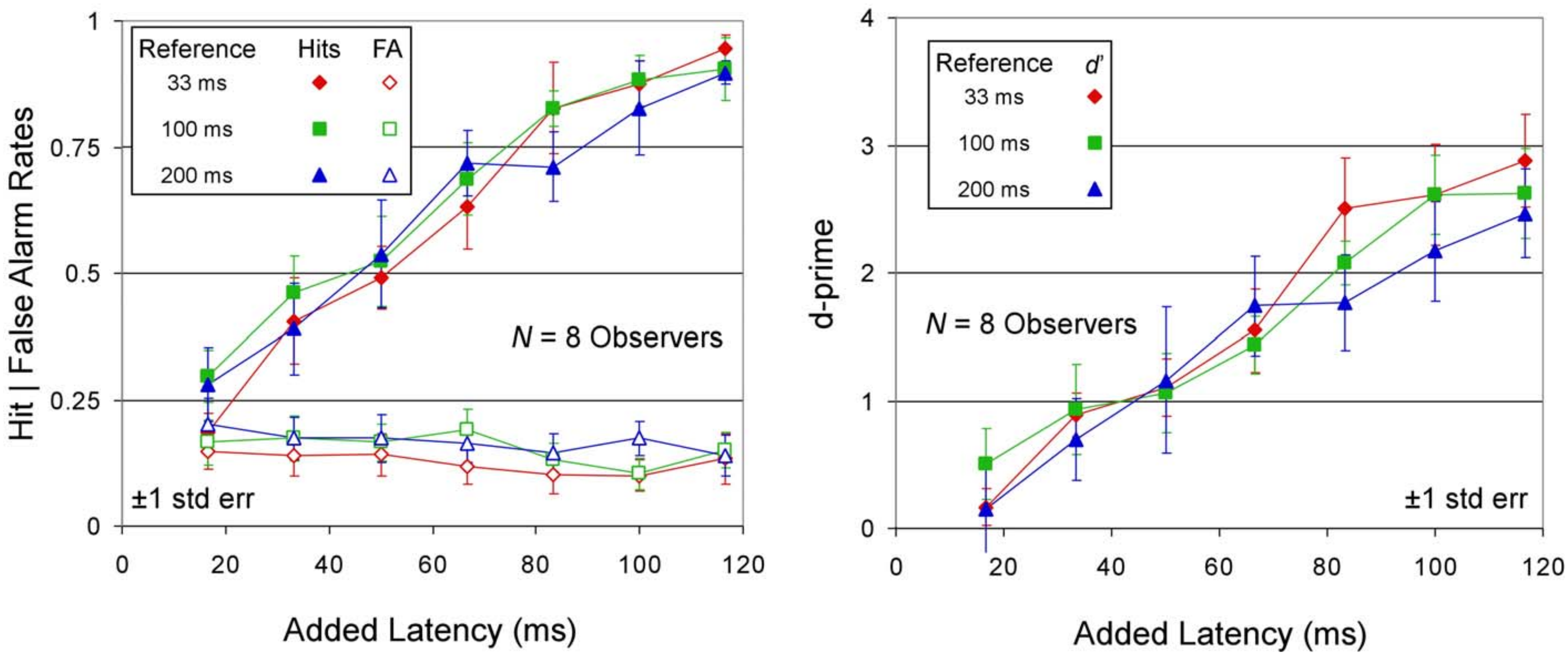
Hit/FA rates and d-prime





(One Observer)

Example: Latency Discrimination
Average Hit/FA rate and d-prime



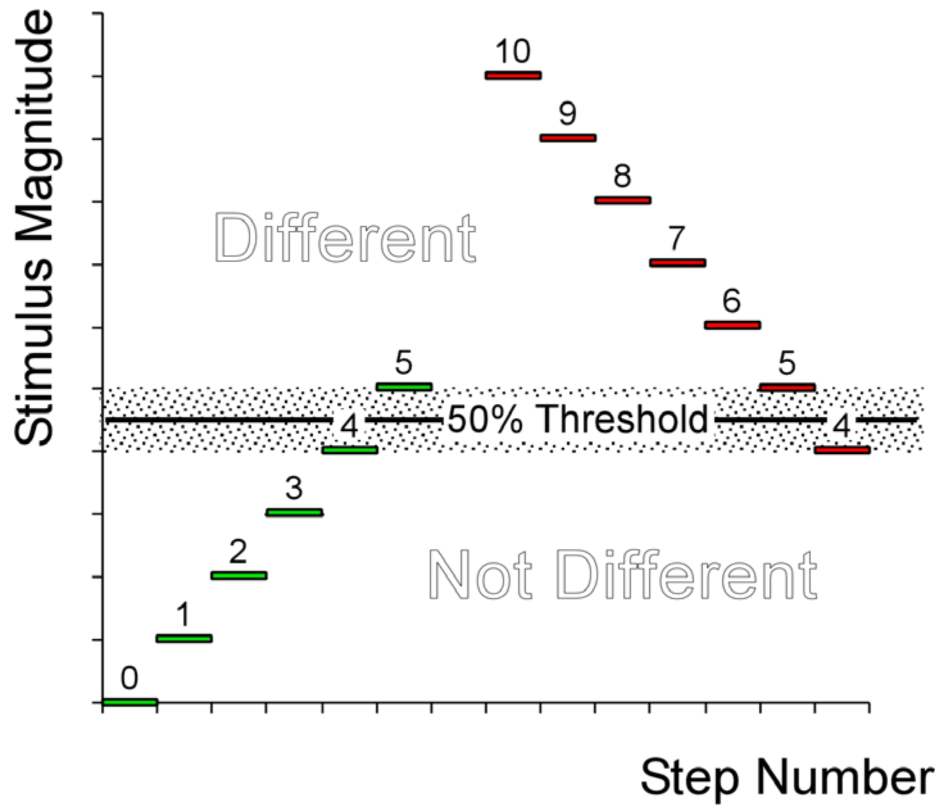
(8 Observers)

Hit/FA Rates vs. Stimulus

- Ideally we want low and uniform *p*(*FA*)
  - Reliability in performing the judgment task
  - Constant criterion; no drift in bias
  - d' depends on hit rate, *p*(*H*)
- *p*(*H*) as a function of stimulus intensity
  - Psychometric function
- Thresholds and bias
  - More later w/ Methods of Limits

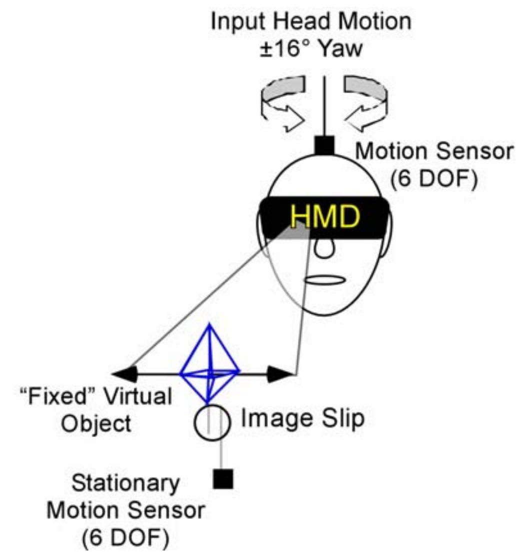
(Truncated) Method of Limits

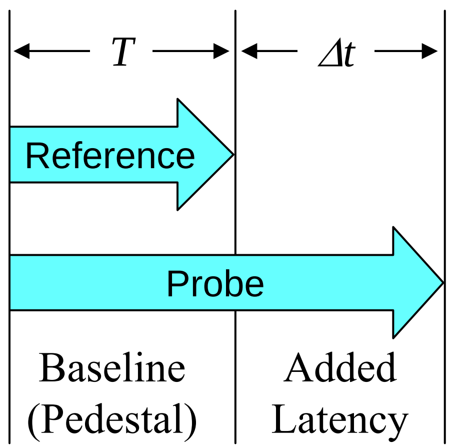
- Staircases (non-reversing) algorithm
  - Define stimulus range
  - Start high: descend
    until "Not Different"
  - Start low: ascend
    until "Different"
- 50% threshold



Example: Latency Discrimination

Staircase Experiment [2]





Q (2AFC): Same or different?

$T = \{33, 100, 200\}$ ms

Experiment Factor (3 levels)

Staircases start either

LOW

$\Delta t = 0$ ms (randomly 1 to 3 times)

and increase until "different"

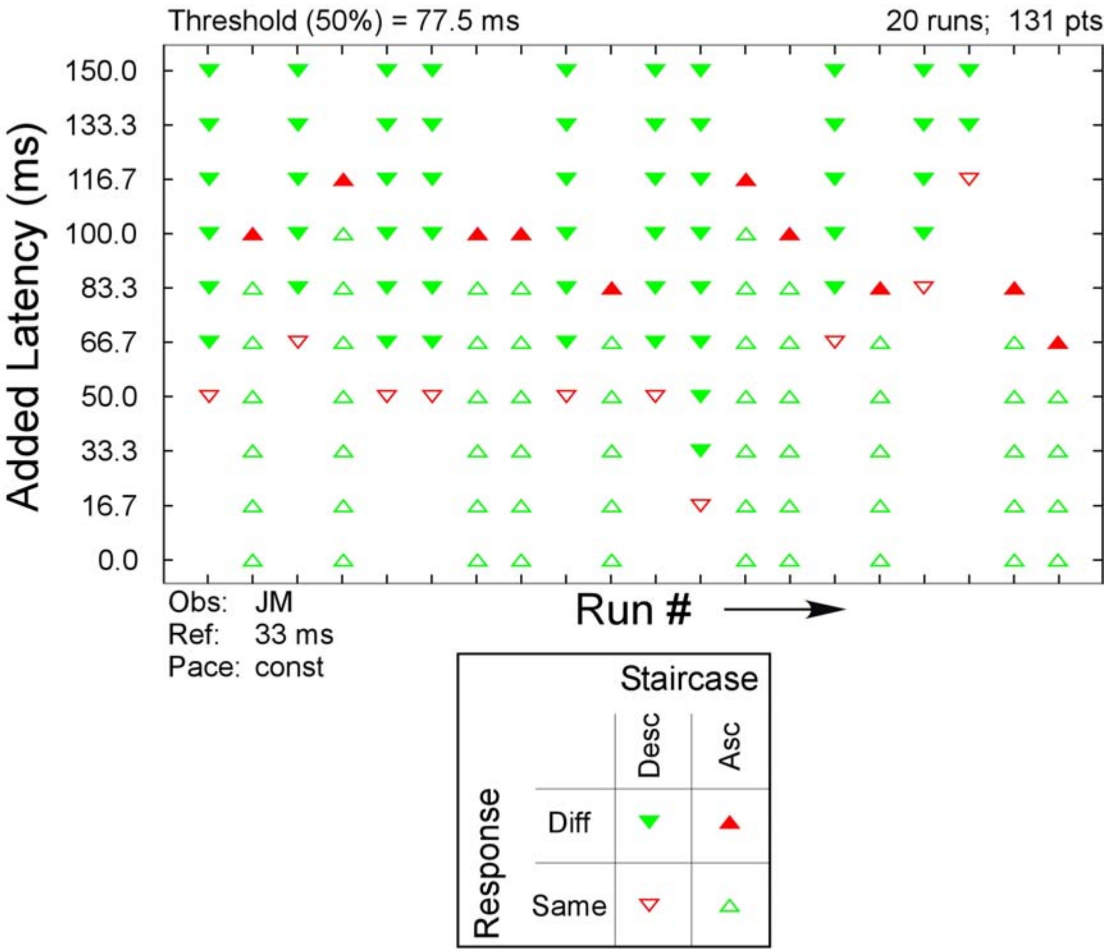
or

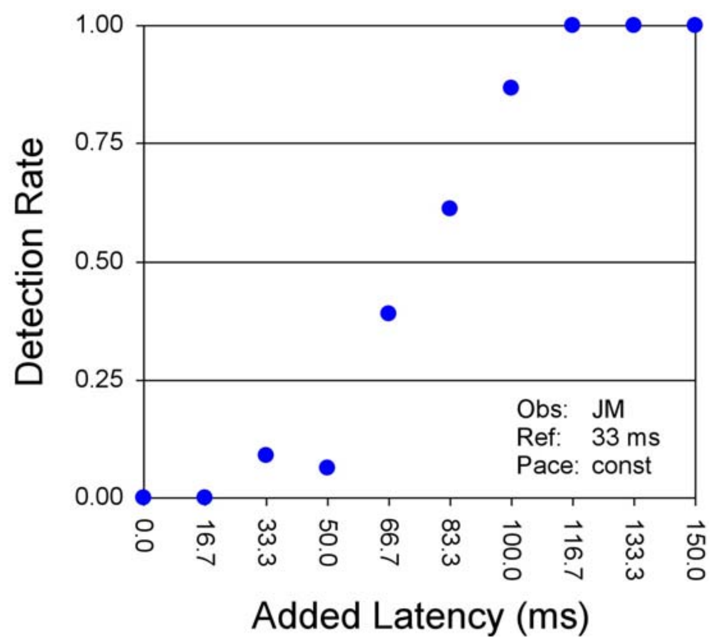
HIGH

$\Delta t = \{116.7, 133.3, 150.0\}$ ms (randomly selected)

and decrease until "not different"

Example: Latency Discrimination
Staircase Experiment





Example: Latency Discrimination
Staircase Experiment

- Staircases (non-reversing)
  - Each staircase yields a termination level
  - From which can reconstitute raw staircase data
  - Construct $p(H)$ as a function of stimulus intensity

Method of Limits

- Simple truncated method of limits up–down method or :

$$x_{n+1} = x_n - \delta(2\, z_n - 1)$$

$$\{\text{miss}|\text{hit}\}: z_n = \{0|1\}$$

- Fixed step size δ
- Avoids stimulus presentation far above and below threshold

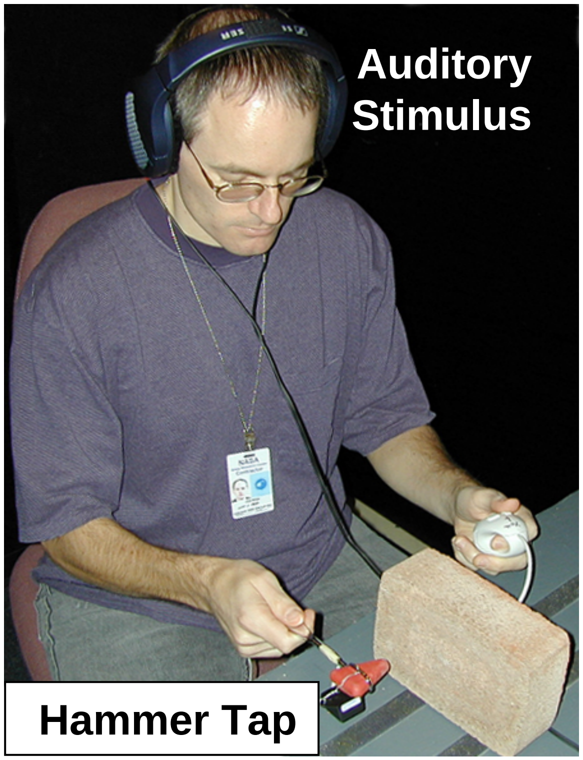
Method of Limits

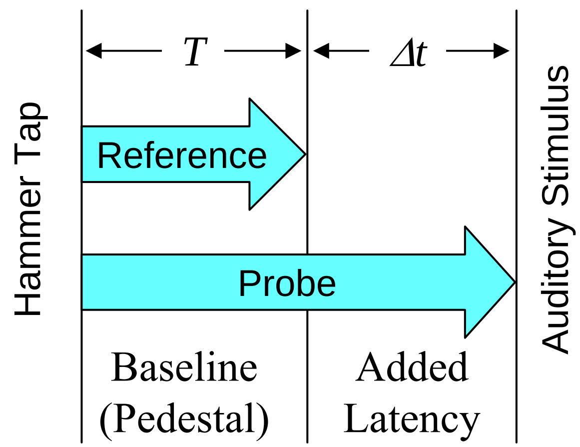
Up-Down Staircases

- Transformed Up-Down staircases w/ or w/o adaptation
- 1 Up-*N* Down staircase theoretical convergence levels ("equilibrium"): $0.5 = 1 - p(H)^N$
  - 1U-1D: 50.0%
  - 1U-2D: 70.7% *(Transformed)*
  - 1U-3D: 79.4% *(Transformed)*
- Analytic relation of d' to equilibrium for 1U-*N*D staircase and *M*-Alternative Force Choice
  - Use d' to help choose staircase method

Example: Asynchrony Discrimination

1U-2D Adaptive Staircase Experiment [3]





Q (2AFC): **Which** (1 or 2) was “Reference”?

$T = \{7.2\}$ms

Staircases start either

LOW

$\Delta t = 0$ ms

or

HIGH

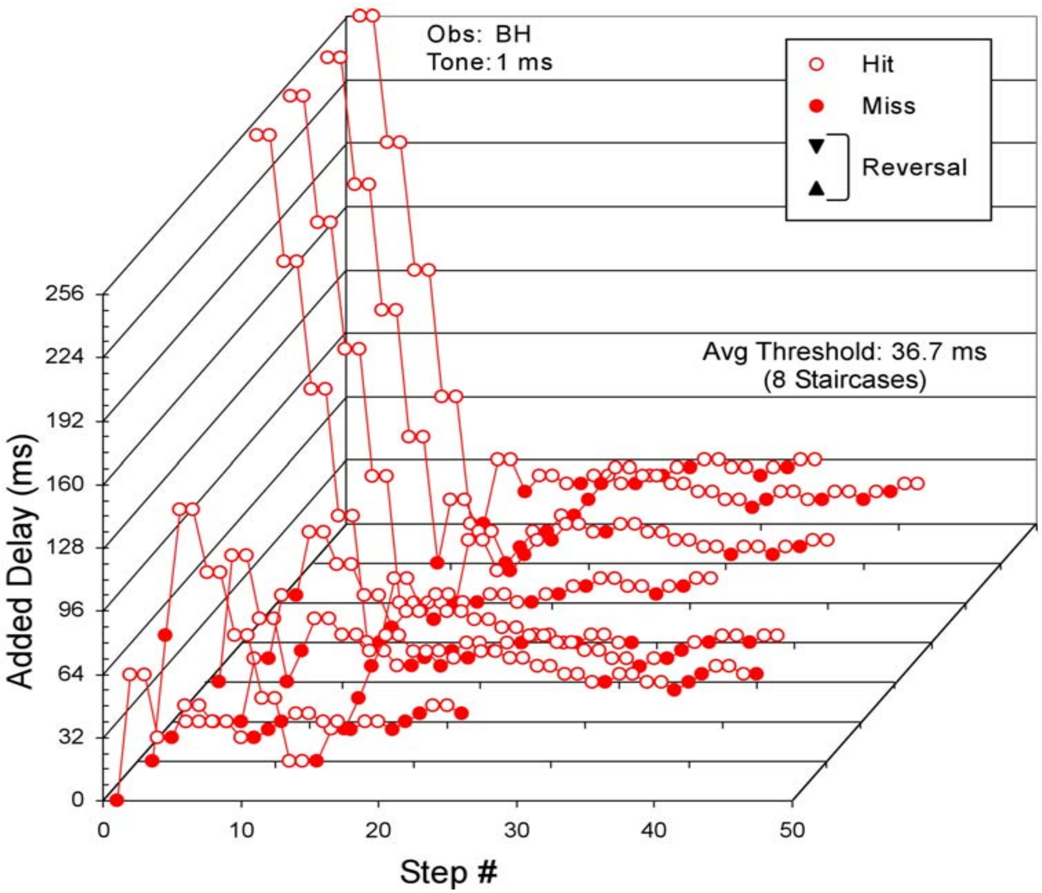
$\Delta t = 256$ ms

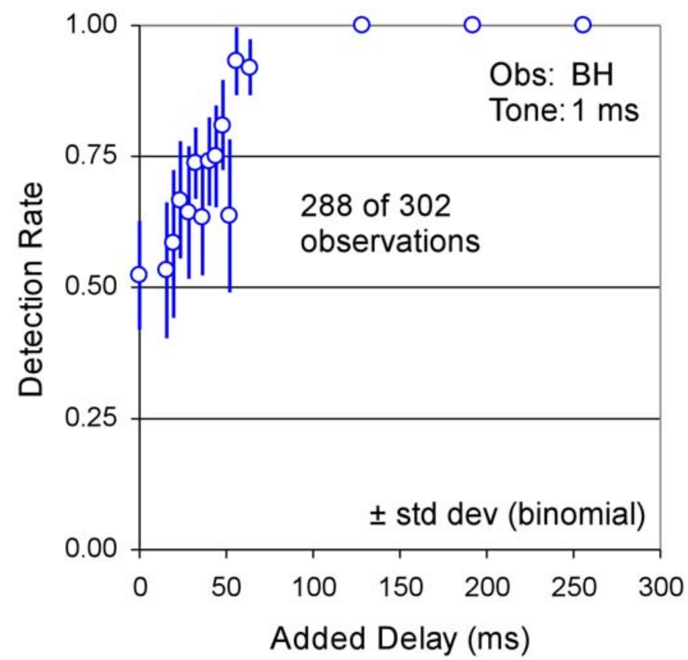
$\min(\Delta t) = 4$ ms

Example: Asynchrony Discrimination

1U-2D Adaptive Staircase Experiment

(70.7% Threshold)





Example: Asynchrony Discrimination Adaptive Staircase Experiment

- Each staircase *ideally* converges to an equilibrium level, corresponding to a theoretical threshold
  - Drift (criterion shift) with extended duration
- Construct $p(H)$ as a function of stimulus intensity
- Adaptive staircases
  - Focus most quickly on region of interest
  - More on region of interest in section on Psychometric Function
- Interleaved staircases
  - Prevent observer tracking/prediction

Psychometric Function

- Construct a model (i.e., psychometric function) describing relation between input stimulus intensity and observer's detection/discrimination rate.

- Looking for best fit of given function to experimentally measured data through optimization of model parameter space.

- In the following illustrations, Gaussian distributions (i.e., two parameter model: μ and σ) are fitted to minimization of weighted least-square (χ^2) error.

Psychometric Functions

•Some typical monotonically increasing functions

Gaussian distribution: $N(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$

Logistic distribution: $L(x;\alpha,\beta,\gamma)=\frac{1}{1+\exp\left(\frac{\alpha-x}{\beta}\right)}$

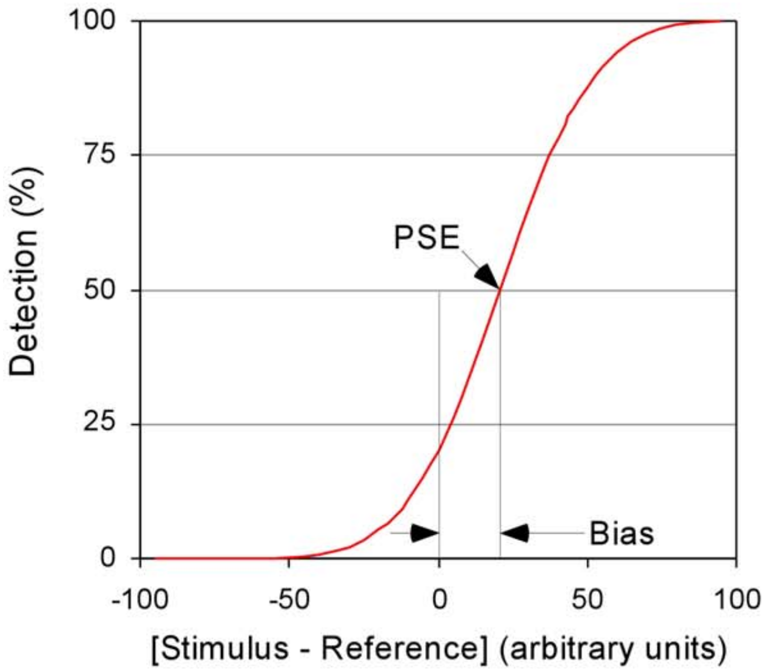
$(\beta=\sigma/1.7;\ \alpha=\mu)$

(Gaussian approx.)

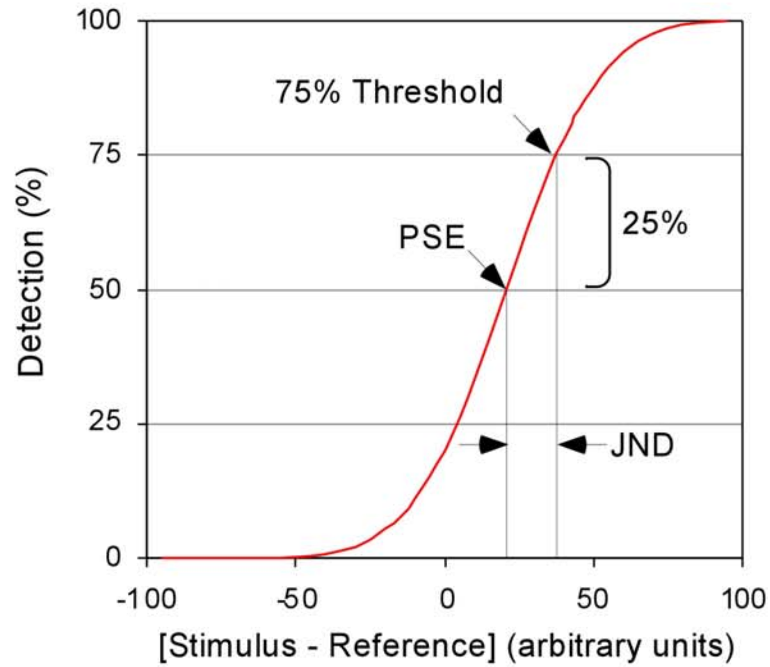
Weibull distribution: $W^{(1)}(x;\alpha,\beta,\gamma)=1-\exp\left(\frac{(x-\gamma)^{\beta}}{\alpha}\right)$

Psychometric Function

- Features of the ogive



Point of Subjective Equality (PSE)
and bias with respect to
reference stimulus



Just Noticeable Difference (JND)
for psychometric function
symmetric about PSE

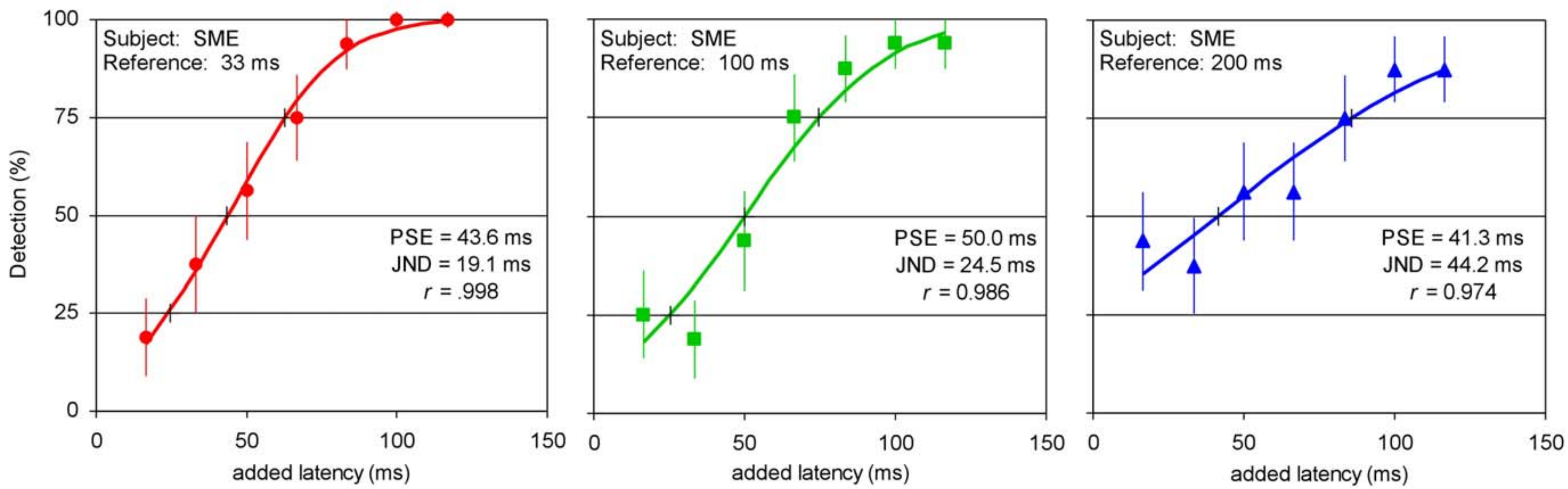
Psychometric Function

- Features of the ogive
  - Point of Subjective Equality/Equivalence (PSE)
    - Bias in observer's response
    - Criterion dependent
    - Question posed as a source of bias
  - Just Noticeable Difference (JND)
    - Generally defined by ½ of stimulus difference between 1st and 3rd detection quartiles
    - For symmetric functions, the amount of additional stimulus difference to increase detection by 25% from PSE
    - JND is related to variance and is therefore a statistical measure of detectability

Fitting a Psychometric Function

- Practical considerations (for standard normal model)
  - Transform data to standard normal (Z) coordinates and apply linear regression
  - *Probability paper (cf. semi-log paper)*
  - Functional fit minimizing weighted error of fit to data
    - Weighted by binomial standard error for fitted model (Probit with. χ^2 error/model)
  - "Finger error":
    - Rates of guessing (p_g); rates of lapsing (p_l)
    - Alleviates problem of $P = 0$ or 1, i.e., $Z \rightarrow -\infty$ or ∞

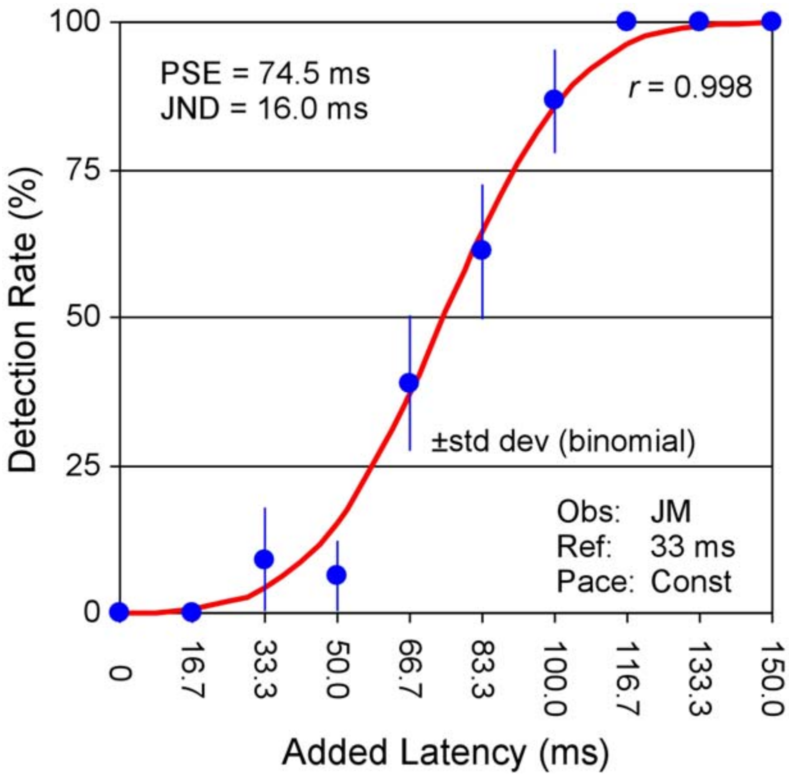
Psychometric Functions
Constant Stimuli Study [1]



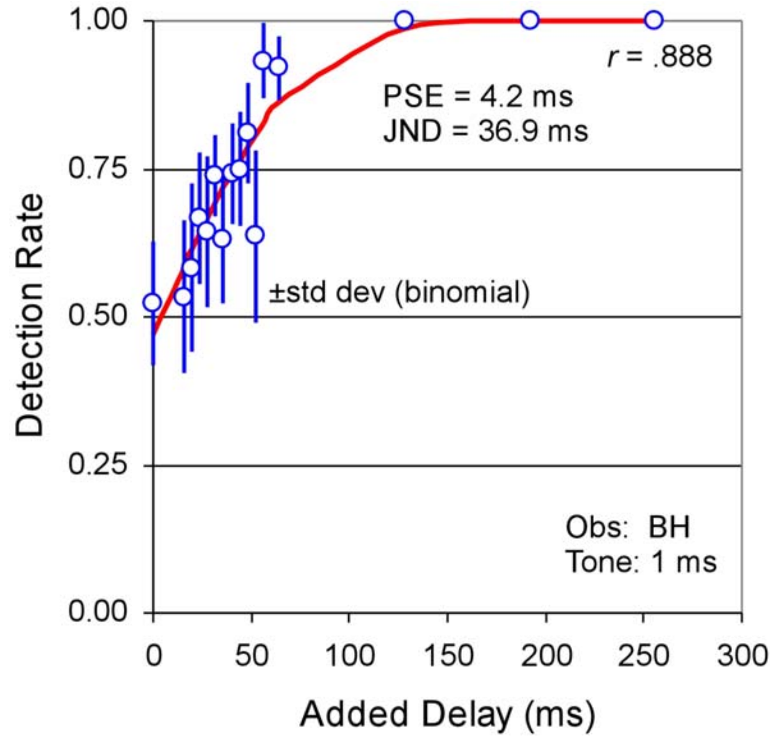
Fitted to Cumulative Gaussian Distribution

Psychometric Functions

Staircase Studies [2],[3]



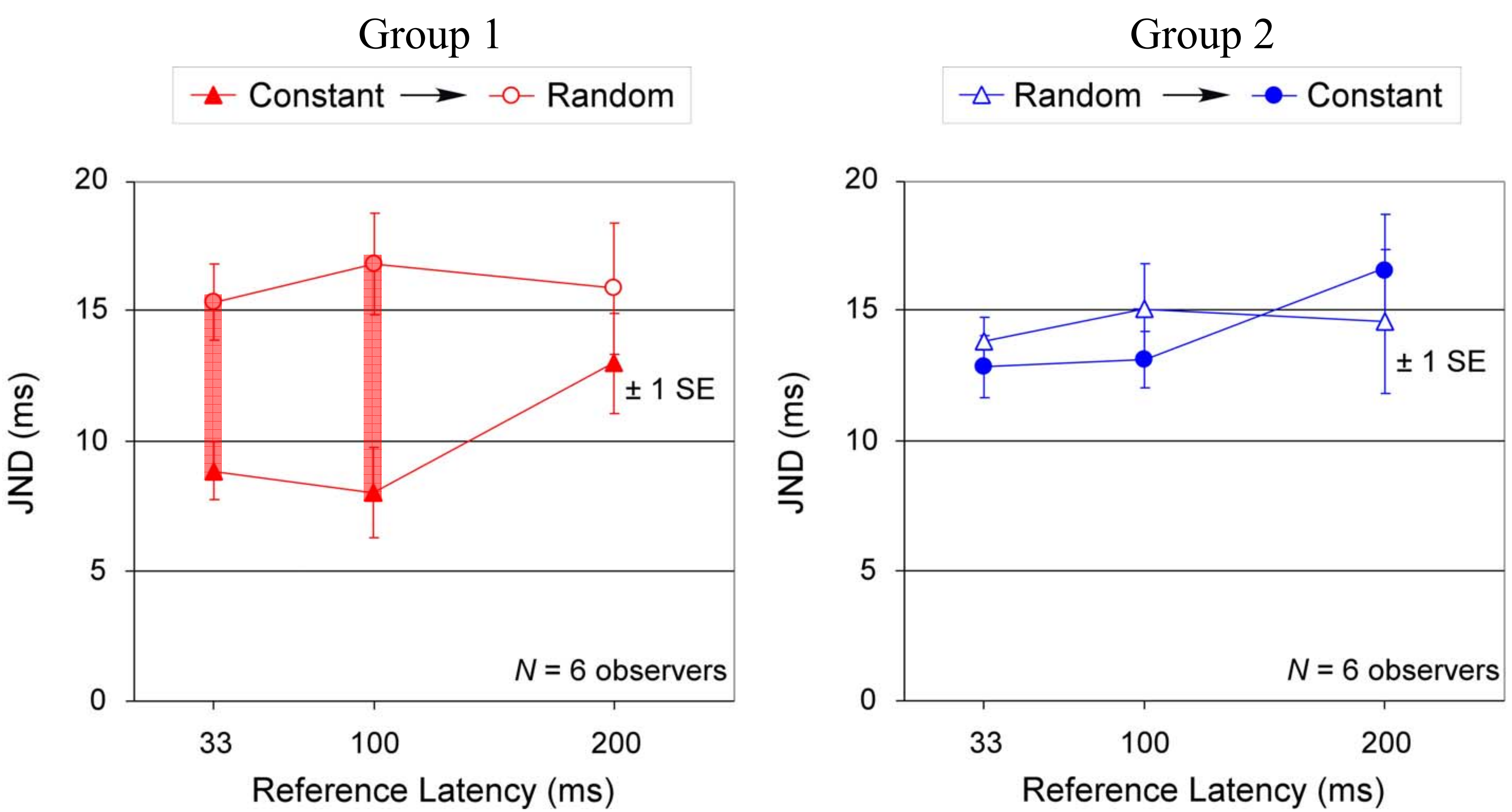
50% threshold = 77.5 ms



75% level = 41.1 ms
70.7% threshold = 36.7 ms

Just-Noticeable Differences (JND): 12 Observers

[2] HFES (2003)



Base:	$F(2,20) = 4.044$	$p < .022$
Group X Epoch X Base:	$F(2,20) = 4.866$	$p < .019$

Summary Comments on Methods

- Method choice should depend on objectives
- Use Method of Constant Stimuli first, when have insufficient knowledge of detection capacity
  - Measure *d*-prime and FA rates
  - Time-consuming (inefficient)
- Method of Limits w/ U-D Adapting Staircases
  - Can select U-D ratio to concentrate data in region of interest
    - Efficient (fewer observations than Constant Stimuli)
  - Does not measure FA rate
    - Has a prescribed d' for given *M*-alternative forced choice

Summary Comments on Methods

- Method choice should depend on objectives
- Use Method of Constant Stimuli first, when have insufficient knowledge of detection capacity
  - Measure *d*-prime and FA rates
  - Time-consuming (inefficient)
- Method of Limits w/ U-D Adapting Staircases
  - Can select U-D ratio to concentrate data in region of interest
    - Efficient (fewer observations than Constant Stimuli)
  - Does not measure FA rate
    - Has a prescribed d' for given *M*-alternative forced choice
- Caveat: Pure perception experiments may be far removed from ecological experience—i.e., detached from realistic action and task performance

References

http://human-factors.arc.nasa.gov/ihh/spatial/papers.html

[1] Ellis, S.R., Young, M.J., Ehrlich, S.M., & Adelstein, B.D. (1999). Discrimination of changes of rendering latency during voluntary hand movement. *Proceedings, 43rd Annual Meeting Human Factors and Ergonomics Society*, pp. 1182-1186.

[2] Adelstein, B.D., Lee, T.G., & Ellis, S.R. (2003). Head Tracking Latency in Virtual Environments: Psychophysics and a Model. *Proceedings, 47th Annual Meeting Human Factors and Ergonomics Society*, pp. 2083-2087.

[3] Adelstein, B.D., Begault, D.R., Anderson, M.R., & Wenzel, E.M. (2003). Sensitivity to Haptic-Audio Asynchrony. *Proceedings, 5th International Conference on Multimodal Interfaces*, ACM, pp. 73-76.

Further Reading

Falmagne, J.-C. (1985). *Elements of Psychophysical Theory*. Oxford University Press, New York.

Gescheider, G.A. (1997). *Psychophysics: The Fundamentals, 3rd Edition*. Lawrence Erlbaum Associates, Mahwah, NJ.

Green, D.M., & Swets, J.A. (1989). *Signal Detection Theory and Psychophysics*. Peninsula Publishing, Los Altos, CA.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception and Psychophysics, 63*(8), 1279-1292.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, *35*(17), 2503-2522.

Ames Research Center

Human Factors research and technology

NASA

Human Performance and Preference Studies

Stephen R. Ellis
Ames Research Center
Moffett Field, CA USA

Exhortations And Illustrations

Outline

1. Purpose and Human Performance assessment
2. Example: Interface preference data
3. Measurement
    - 3.1 Stevens's classification of measurement
    - 3.2. Critique of Stevens's classification
4. Three Illustrative Cases Studies
    - 4.1. Nominal data: Maneuver distributions
    - 4.2. Ordinal data: correlation, Friedman ANOVA
    - 4.3. Interval data: ANOVA
5. Some Heuristics for Behavioral Data Analysis

Purpose and Human Performance assessment

The purpose of a human performance assessment within a virtual environment is to determine whether the virtual environmental users are able to realize the goals and expectations they bring upon entering it without unacceptable costs and risks.

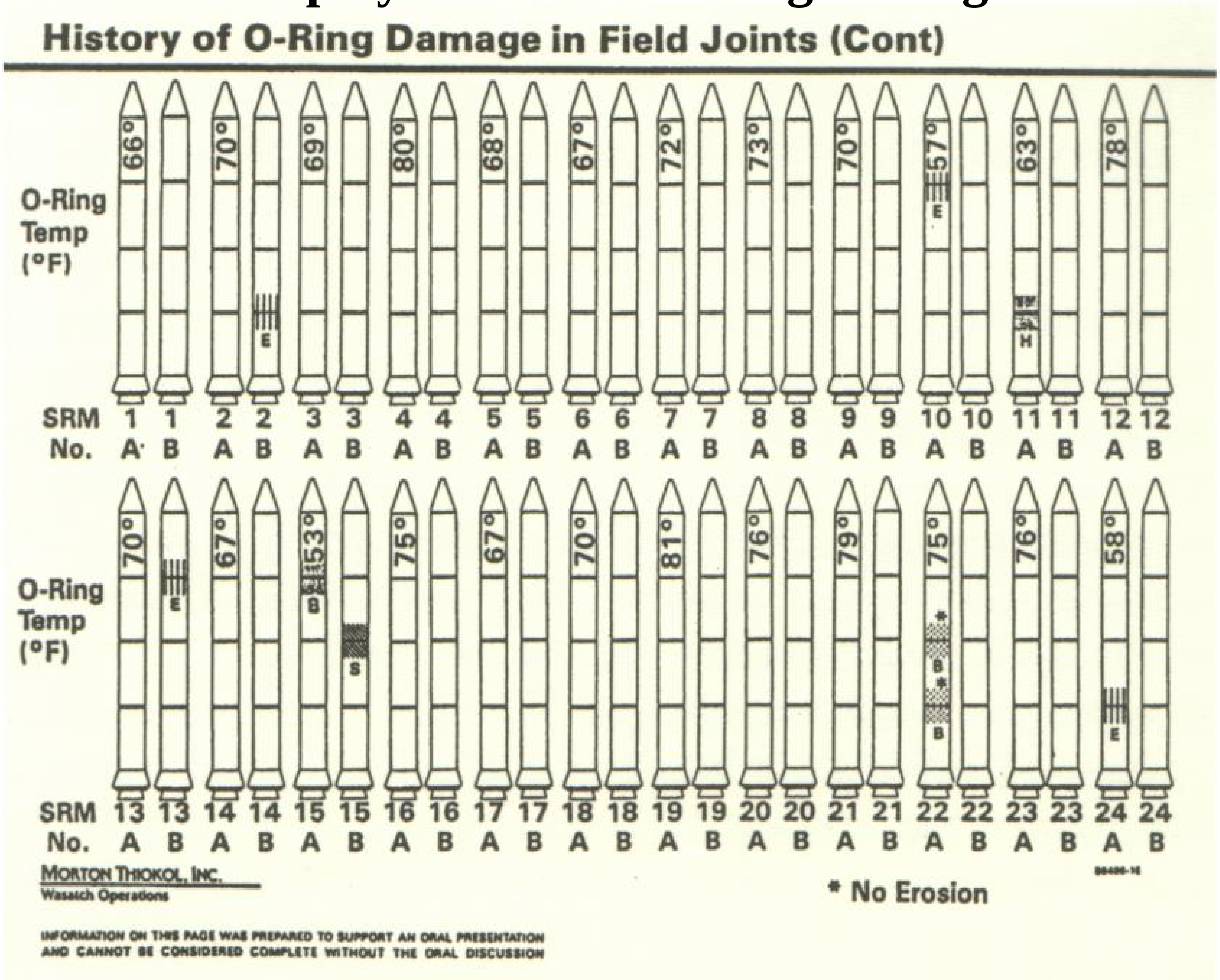
Seek information that is

1. True
2. Reliable
3. Valid
4. Knowably generalizeable
5. Task appropriate

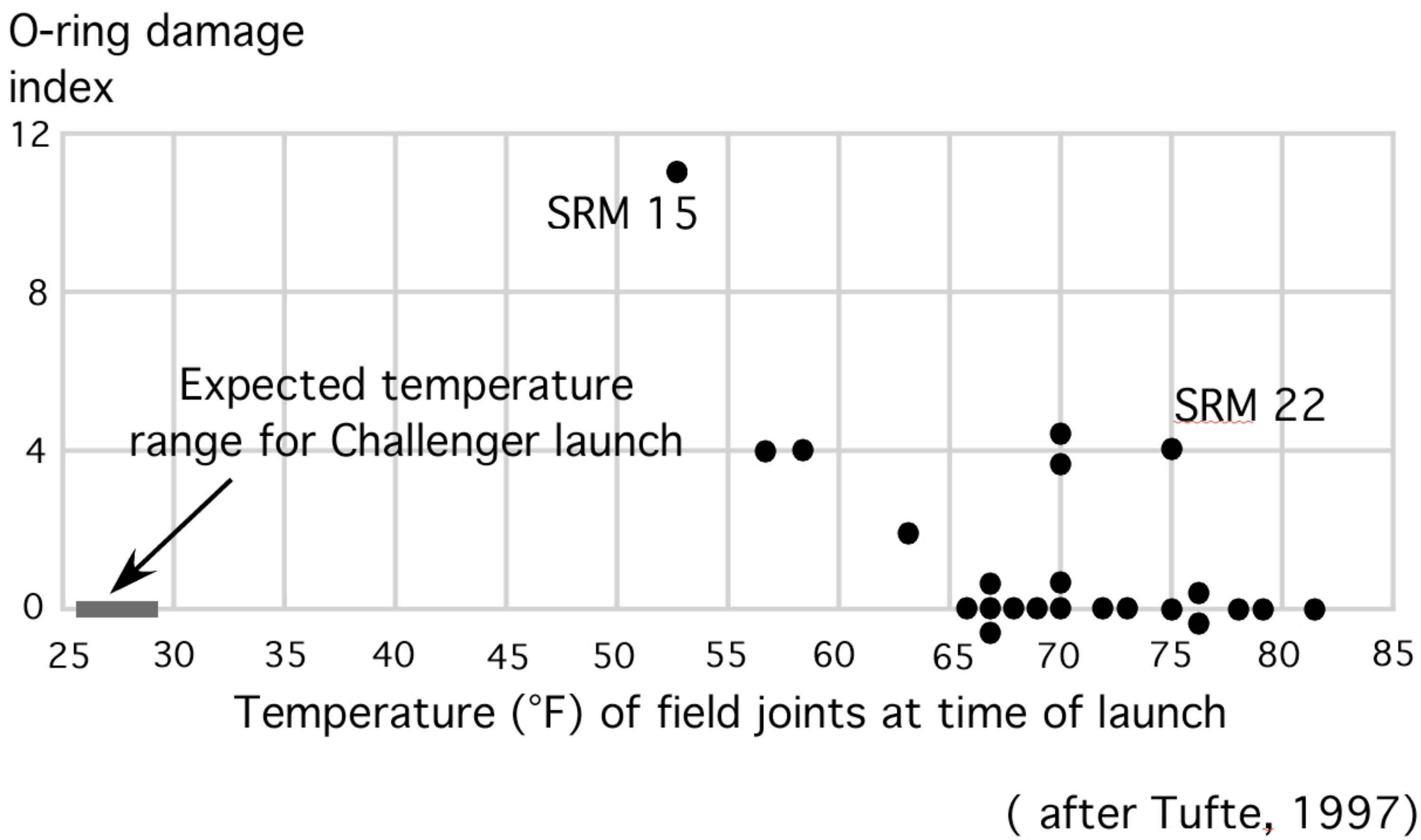
Exhortation #1

Think and Argue Causally

Historical Display of Orbiter O-ring Damage



Explanatory Display of Orbiter O-ring Damage



Exhortation #2

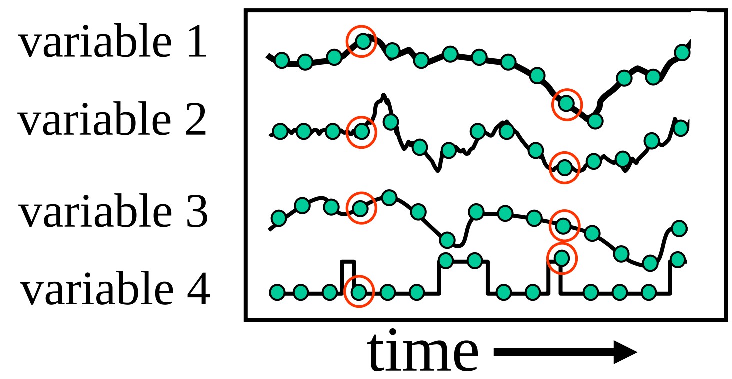
Consider Alternative Investigative Approaches

How is the evaluation done?

Investigative Techniques

Theoretical

Mathematical
Logical
Computational



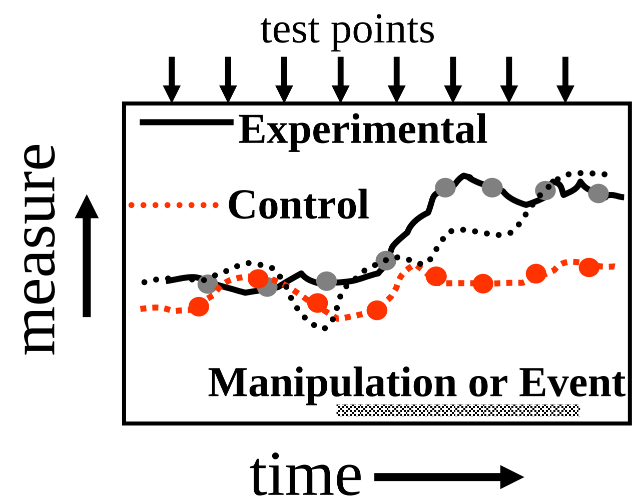
Empirical

Observational

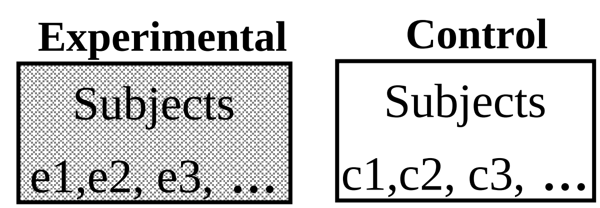
Longitudinal
Cross-sectional

Experimental

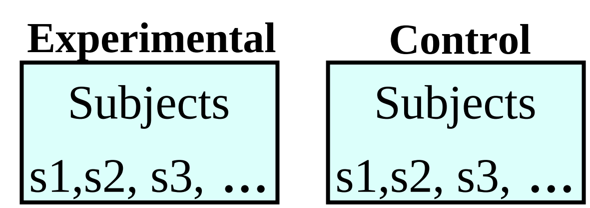
Subject is own control.



Independent groups



Repeated measures



Meta-analytic

Exhortation #3

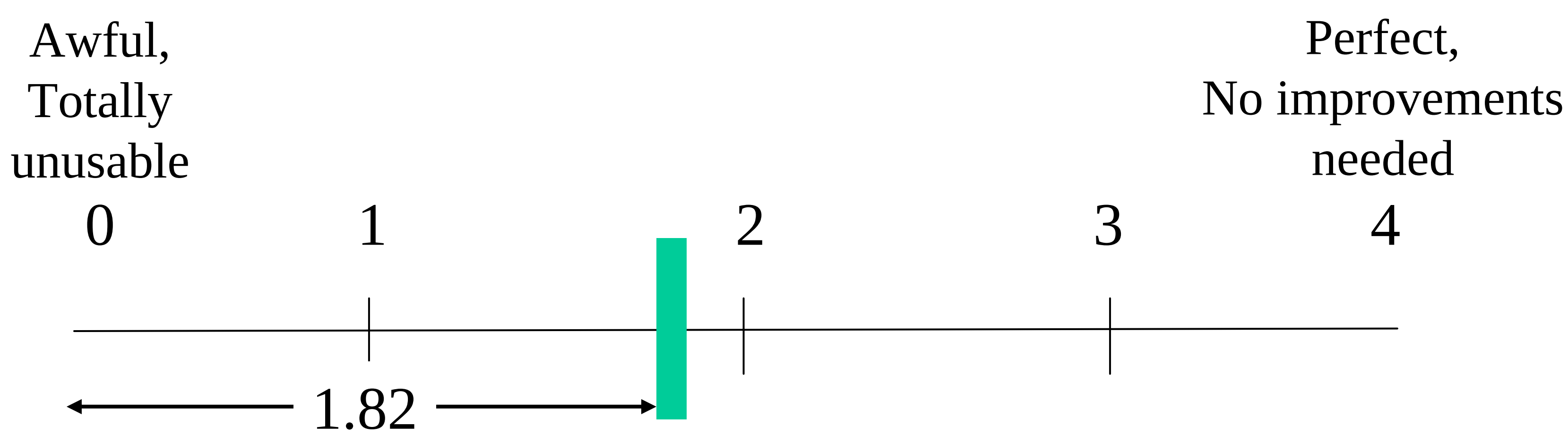
Behavioral "Measurements" Sometimes Yield Surprising Paradoxes

Lickert Scale Opinion Assessment I:questionnaire

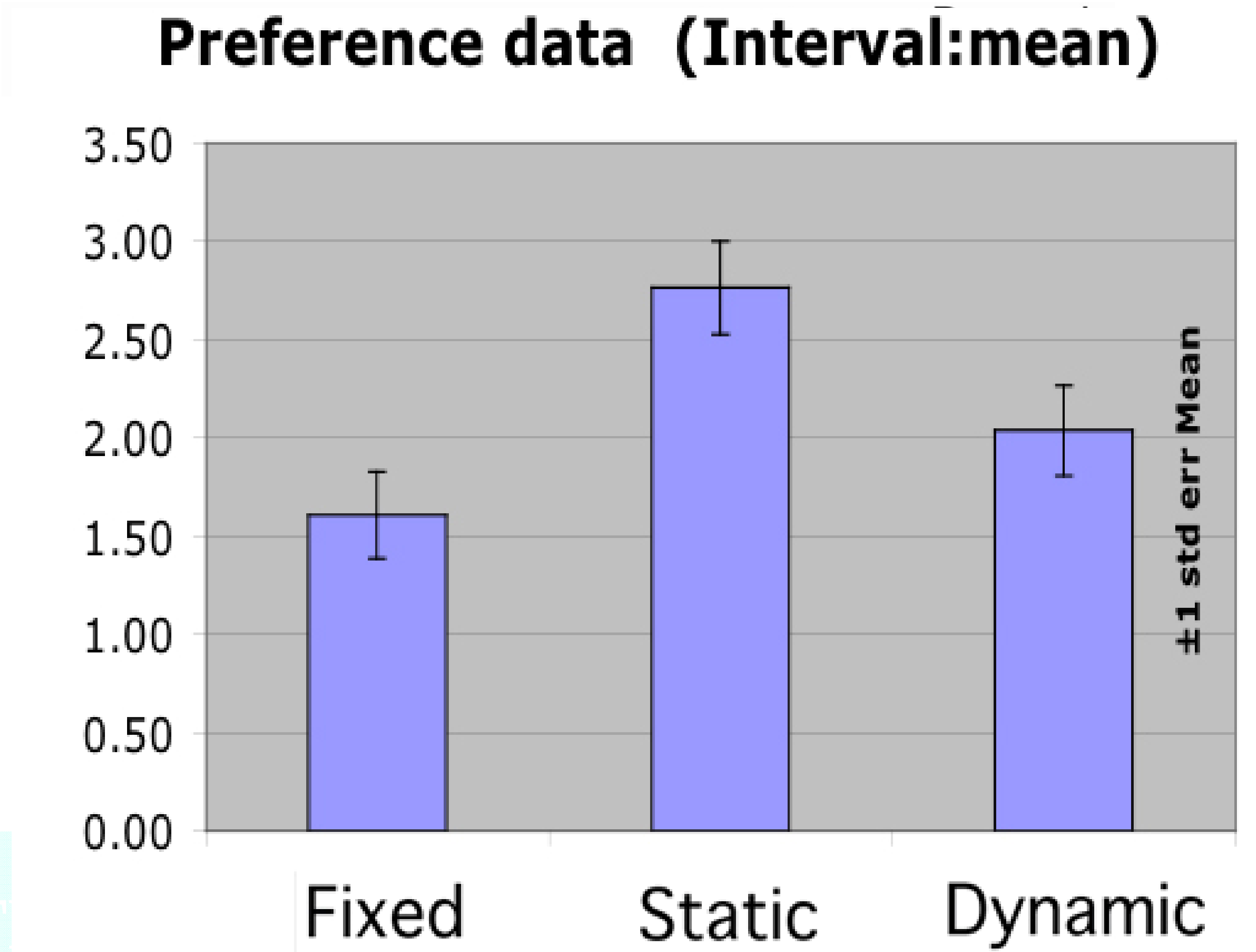
User interface options: option layout in menus

Fixed	Static Match	Dynamic Matching
Fixed menu sequence	Usage frequency adjusted menu sequence	Usage frequency menu sequence: dynamically adjusted

- written instructions, training
- laptop based data collection
- repeated measures, randomized presentation, order balancing, +



Lickert Scale Opinion Assessment II: preference scores



	Fixed	Static	Dynamic
Mean	**1.61**	**2.76**	**2.04**
SE	**0.2214**	**0.2359**	**0.2303**

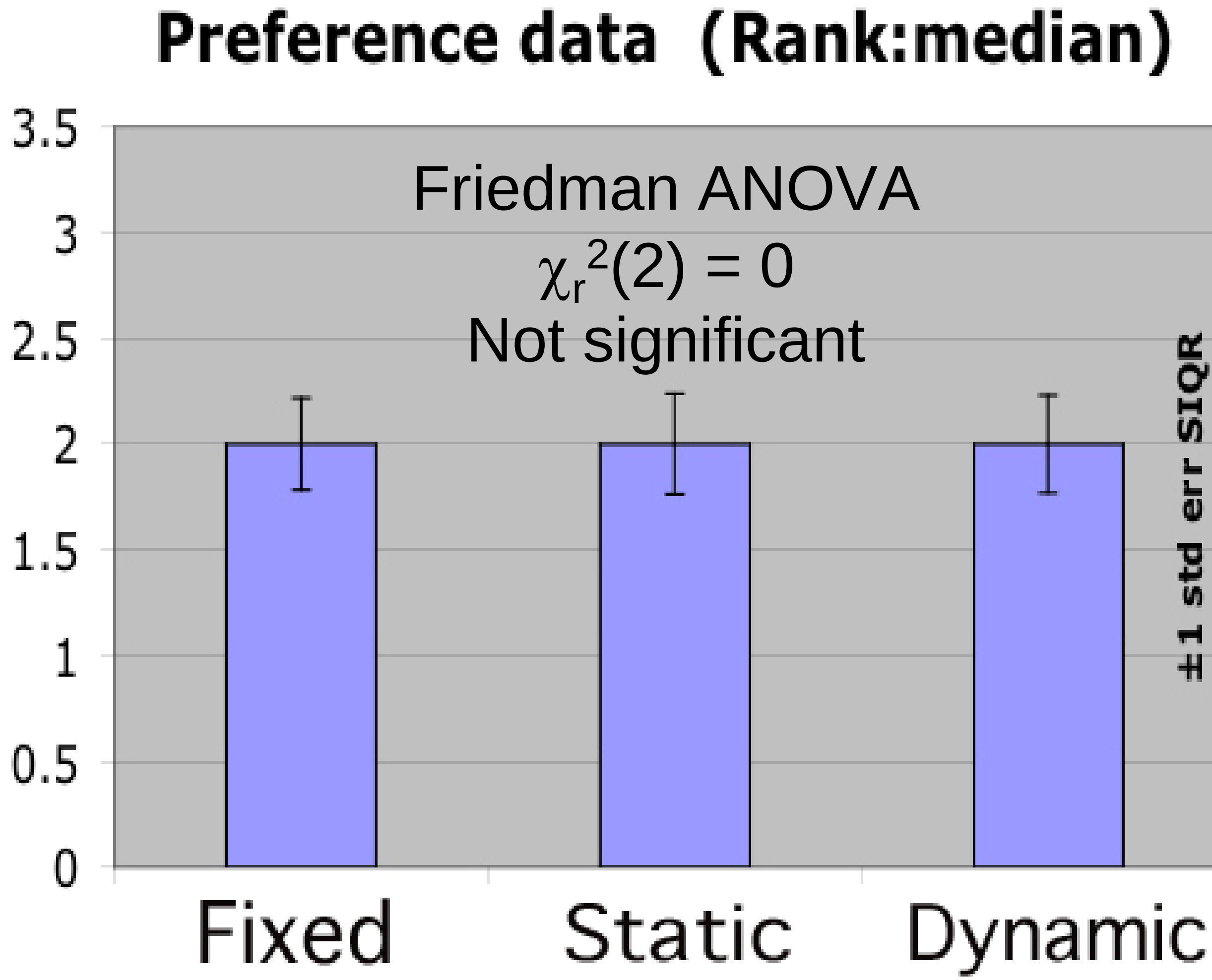
Lickert Scale Opinion Assessment III: ANOVA

Repeated measures ANOVA

Sum of Squares		df	Mean Square (variance)	
SSqr (total)	43.3185			
SSqr between	10.5103	2	MSqr between	5.255
SSqr within	32.8081	28	MSqr within(error)	1.172
			F=	4.485
			F(crit, .05)=	3.340
			F(crit, .025)=	4.221

Lickert Scale Opinion Assessment IV: rank transforms

1 ~ least preferred   3 ~ most preferred



	Fixed	Static	Dynamic
Median	**2**	**2**	**2**
sterr SIQR	**0.2031**	**0.2031**	**0.2031**

K. Arrow Voting Paradox with Heterogeneous Ordinal Preferences

1 ~ Yes “ “ ~ No

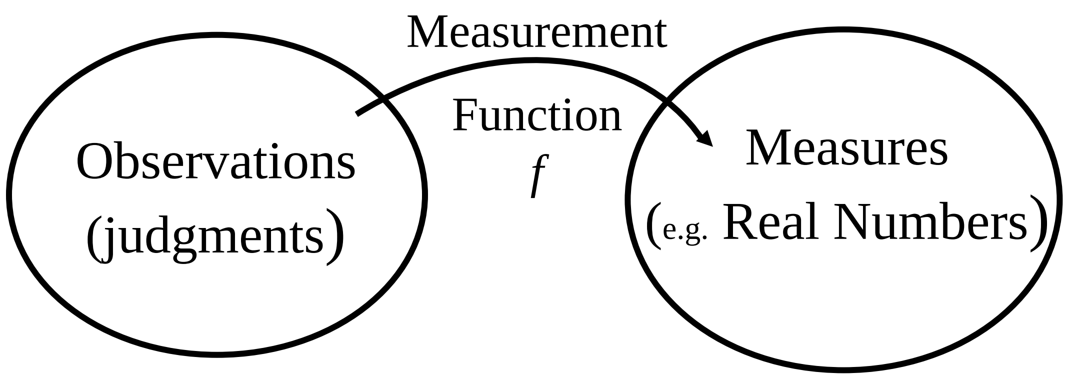
Subjects	A	B	C		a < b	b < c	?
1	3	1	2			1	
2	1	2	3		1	1	
3	2	3	1		1		
4	1	2	3		1	1	
5	3	1	2			1	
6	2	3	1		1		
7	1	2	3		1	1	
8	3	1	2			1	
9	1	2	3		1	1	
10	2	3	1		1		
11	3	1	2			1	
12	2	3	1		1		
13	3	1	2			1	
14	1	2	3		1	1	
15	2	3	1		1		
					Vote		
				Yes	**10**	**10**	5
				No	5	5	**10**

Exhortation #4

A number is not always a number!

Measurement

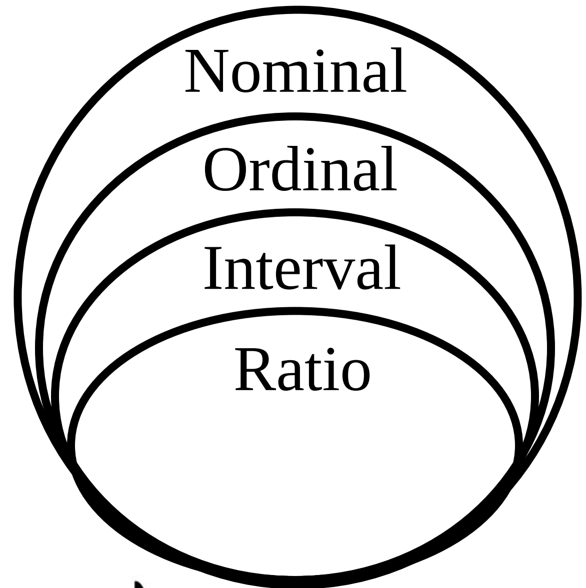
The systematic assignment of scale values, usually numbers, to observations or objects with the purpose of representing and modeling the measured entities.



Some desirable properties of measurements

1. Public
2. Unique
3. Knowably precise
4. Reliable & stable
5. Robust
6. Valid

Variety of Measurement Scales due to Stevens



Scale	Property	Symmetry Operation (inverse)
Nominal	identity	Replacement (Replacement)
Ordinal	rank	Sorting (Unsorting)
Interval	unit	Aggregation (Disaggregation)
Ratio	zero	Multiplication (Division)

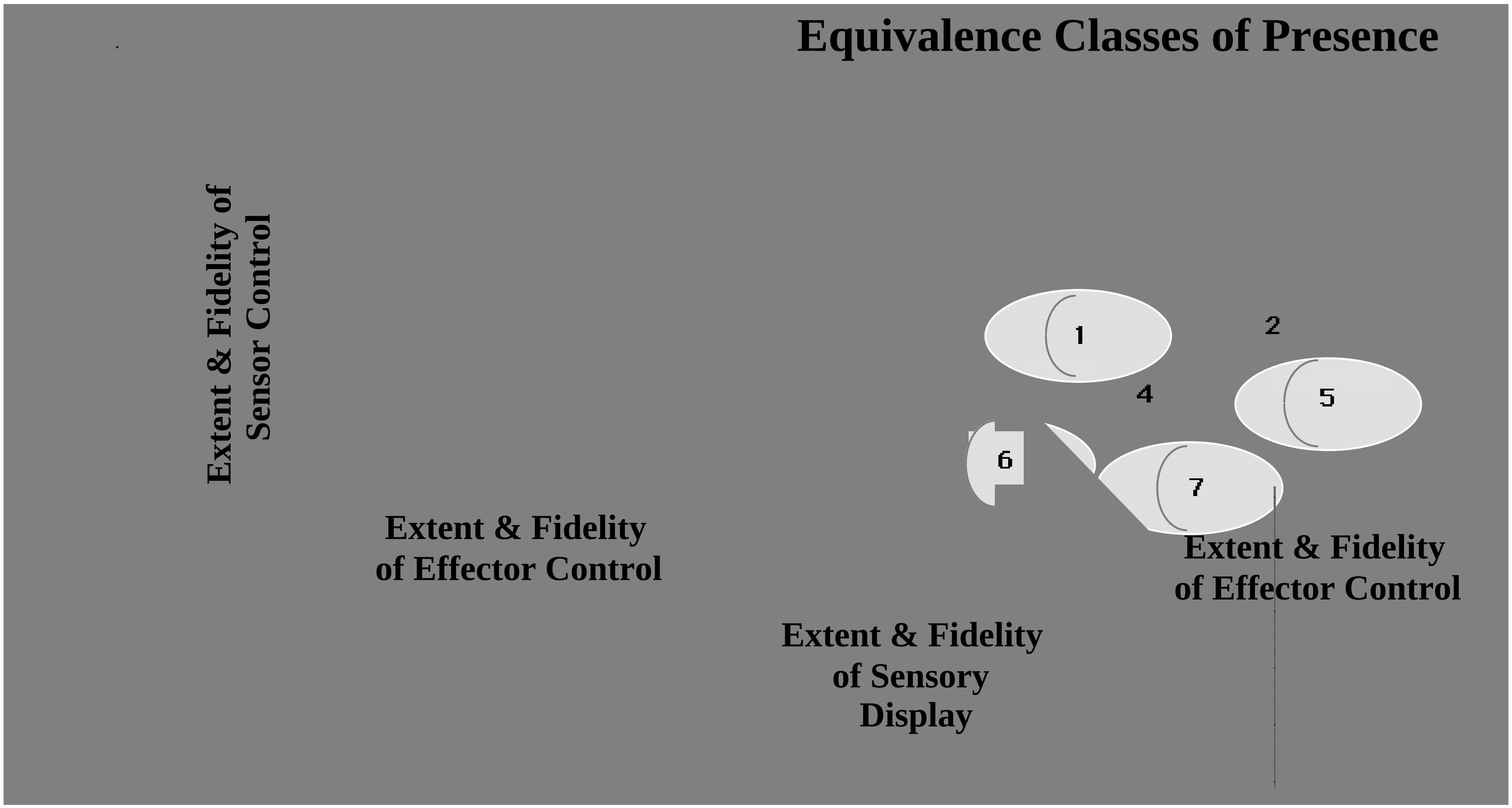
Nominal, Ordinal: Quality

Interval, Ratio: Quantity

Quality, Quantity: Equivalence class

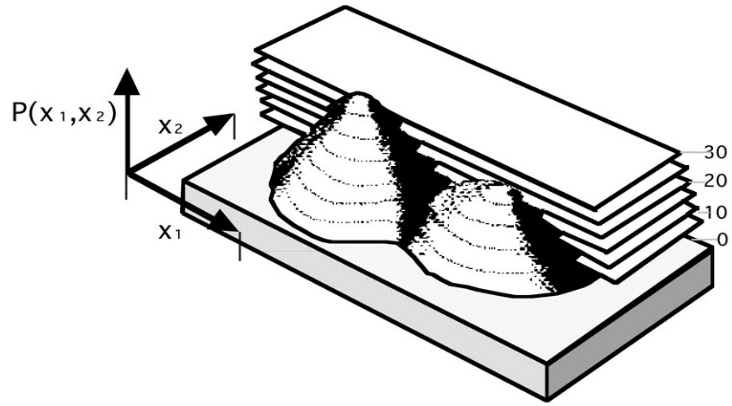
(Stevens, 1946)

The Meaning of an Equivalence Class

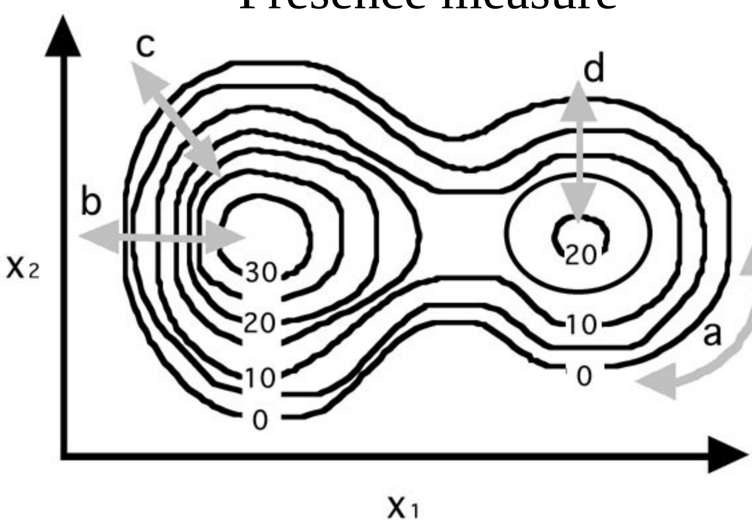


(Ellis, 1996)

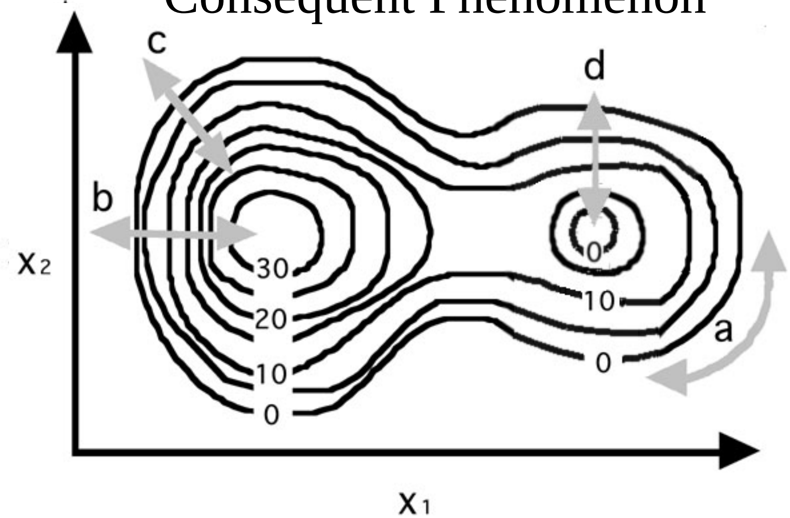
Equivalence Classes and Explanation



Presence measure



Consequent Phenomenon



Measurement Scales and Appropriate Statistics

Scale	Property	Allowable Transformations	Associated Statistics (Centeral tendency, dispersion, correlation)
Nonparametric			
Nominal	identity	Renaming	Mode = $max[frequenc(x_i)]$ Index of variety = - Σ $prob(x_I)log(prob(x_I))$ (bits) Contingency correlation $= \sqrt{\frac{X^2}{N(k-1)}}$ $X^2 = \sum \frac{o_i^2}{e_i} - N$ $\max(X^2) = N(k-1)$
Ordinal	rank	Monotonic transformation Preserving order	Median = percentile$_{50}$ Interquartile range = percentile$_{75}$ - percentile$_{25}$ Rank order Spearman correlation $= 1 - \frac{6\sum_i d_i^2}{N^2 - N}$ Friedman ANOVA
Parametric			
Interval	unit	Linear transformation preserving differences to a scale factor	Mean $= \frac{\sum_i X_i}{N}$ Standard deviation, $= \frac{\sum_i (X_i - \overline{X})^2}{N}$ Product-moment correlation $= \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2 \sum_i (Y_i - \overline{Y})^2}}$ ANOVA
Ratio	zero	Nonlinear Transforma-tions preserving ratios to a scale factor.	Mean Standard deviation Product-moment correlation ANOVA

Pros & Cons of Stevens's Measurement Classification

Pros

1. Discourage use of measurement properties implicit in numerical measurement but not necessarily supported by the measurement technique. *Tells what kinds of difference make a difference!*

2. Reinforces due consideration of the assumptions underlying conventional statistical processing, i.e. sampling, distribution, variance

3. Potential for algorithmic or heuristic control of data analysis.

4. Can be an aid for selecting appropriate statistics for analysis.

Cons

1. *A. priori* data typing may preclude serendipitous discovery.

2. Stevens's scale categorization are absolute resulting in demoting to lower scales resulting in loss of information

3. Statistics should be selected based on what kinds of questions we ask of the data not properties of the data themselves.

4. Potential for algorithmic or heuristic control of data analysis.

Nominal Data: Cockpit Traffic Display Based Avoidance Maneuvers

Distribution

	Horiz.	Vertical	Mixed	Row Sum
Counted	76	2	18	96
Ho	32	32	32	96
	108	34	50	192

Expected freq. f_e, all $f_e >> 5$

$$Cellfreq = \left(\frac{RowSum}{Total} * \frac{ColSum}{Total} * Total\right)$$

	Horiz.	Vertical	Mixed	Row Sum
Counted	54	17	25	96
Ho	54	17	25	96
	108	34	50	192

$$X^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

	Horiz.	Vertical	Mixed	Row Sum
Counted	8.963	13.235	1.96	24.158
Ho	8.963	13.235	1.96	24.158
		Xsqr(2)=	48.317	

48.317 > 13.82, critical $\chi^2(2) @ p < 0.001$

Spearman Rank Order Correlation: r_s

Measure of correlation for data that are only meaningful in terms of order, derived from the standard product moment correlation, r, i.e. $r_s = r$ of the data reduced to ranks for N pairs of correlated variables x, y, with mean ranks X and Y and rank differences d_i.

ranks

$$r_s = r = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\,\text{var}(y)}} = \frac{\sum_i (x_i - X)(y_i - Y)}{\sqrt{\sum_i (x_i - X)^2 \sum_i (y_i - Y)^2}} \le |1|$$

definition

$$r_s = 1 - \frac{6\sum_i d_i^2}{N(N^2 - 1)}$$

$$t = \frac{r\sqrt{N-2}_s}{\sqrt{1 - r_s}}, N >> 10$$

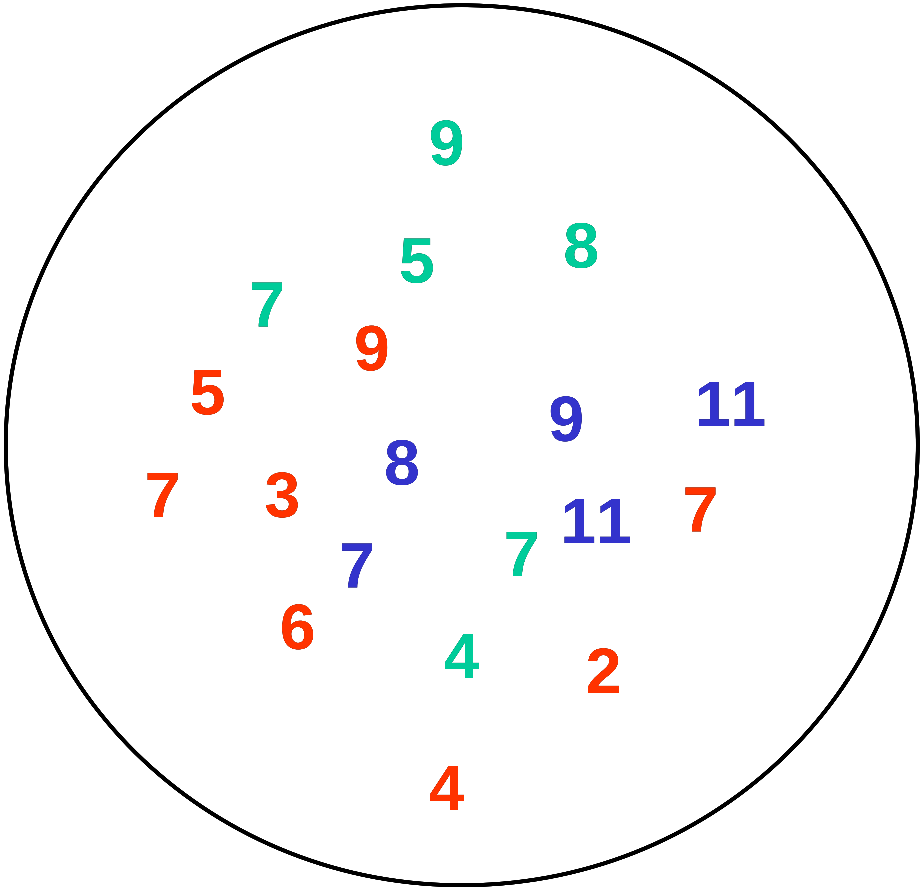
Spearman Rank Order Correlation: r_s

Subjects	Modified Cooper-Harper	Subjective Stability	Mod C/H rank	SS rank
1	4.0	3.0	4.5	7
2	4.0	1.5	4.5	2.5
3	4.2		8.5	4
4	3.5		1	2.5
5	4.0		4.5	7
6	4.0		4.5	1
7	6.0	6.0	12	12
8	4.0	3.7	4.5	10
9	5.0	3.0	10	7
10	5.2	3.0	11	7
11	4.7	3.0	8.5	7
12	4.0	4.5	4.5	11

Ties are assigned the average of ranks otherwise assigned.

r_s= 0.425ns r = 0.625*, *df* = 10, *crit(0.05)= 0.576

Single Factor ANOVA



Subdivisions of a random selection of sample statistics should provide estimates of the same population parameter if the classifi-ation into subgroups has no effect on subgroup statistics.

Example of One way Independent Groups ANOVA

Strategy: estimate a population statistic (a variance) two different different ways so that if H_o is true the ratio of these estimates will be 1. Significant deviations from 1, refute H_o, **given** assumption of random sampling, normal distribution, homogeneity of variance.

Notation

	Group 1	Group 2	…	Group k	
	$x_{1,1}$	$x_{1,2}$		$x_{1,k}$	
	$x_{2,1}$	$x_{2,2}$		$x_{2,k}$	
	$x_{3,1}$	$x_{3,2}$		$x_{3,k}$	
	.	.		.	
	$x_{n1,1}$	$x_{n2,2}$		$x_{nk,k}$	Grand Mean X
Group means	X_1	X_2		X_k	$\Sigma_j n_j = N$
	$df = n_1 -1$	$df = n_2 -1$	etc		$df = N -1$

One-way ANOVA: Partitioning Sums of Squares (SS) and Definition of Mean Squares (MS) variance estimates (R. Fisher)

For the jth group

$$(x_{ij} - X) = (x_{ij} - X_j) + (X_j - X)$$ identity

$$\Sigma_i (x_{ij} - X)^2 = \Sigma_i (x_{ij} - X_j)^2 + \Sigma_i (X_j - X)^2 + 2\,(X_j - X)\,\Sigma_i (x_{ij} - X_j)$$ sqr, sum

$$\Sigma_i (x_{ij} - X)^2 = \Sigma_i (x_{ij} - X_j)^2 + n_j (X_j - X)^2$$ summation of constant, dev sum to 0

$$\Sigma_j \Sigma_i (x_{ij} - X)^2 = \Sigma_j \Sigma_i (x_{ij} - X_j)^2 + \Sigma_j n_j (X_j - X)^2$$ sum over k groups

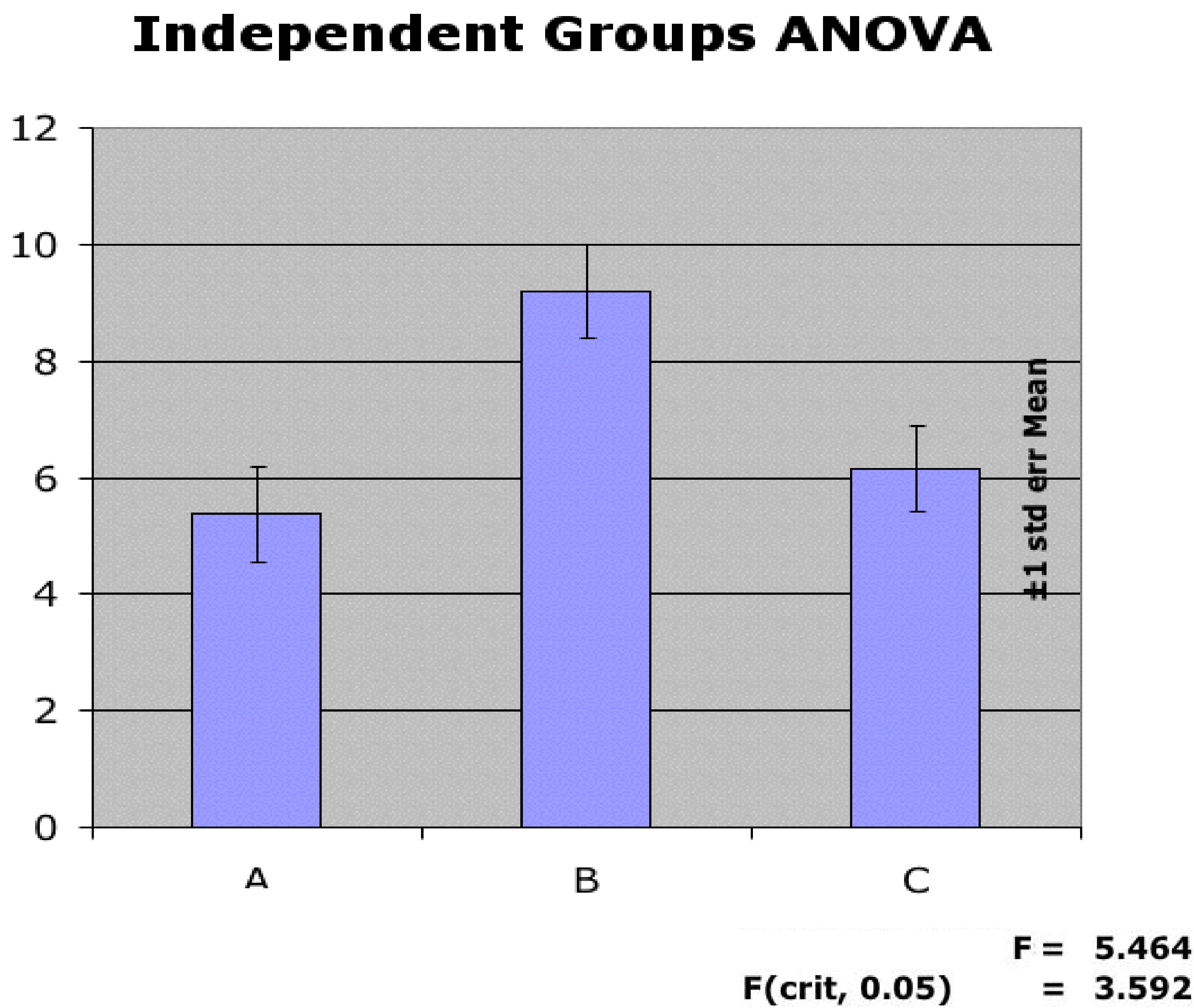
$$\text{Total SS} = SS_{\text{within groups}} + SS_{\text{between groups}}$$ definition

$$df_{within} = (n_1 - 1) + (n_2 - 1) + .. + (n_K - 1) = \Sigma_j n_j - k = N - k$$

$$df_{between} = k - 1$$

$$MS_{\text{within}} = SS_{\text{within}} / df_{\text{within,}} \quad MS_{\text{between}} = SS_{\text{beween}} / df_{\text{bewtween}}$$ definition

F statistic with k-1, N-k degrees of freedom = $MS_{\text{between}} \;/\; MS_{\text{within}}$


Independent Groups ANOVA
12
10
8
6
4
2
0
A
B
C
±1 std err Mean
F = 5.464
F(crit, 0.05) = 3.592

Example of Friedman Nonparametric ANOVA

Lickert Scale Opinion Assessment IV: rank transforms

1 ~ least preferred   3 ~ most preferred

***J* = 3**

Subjects	A	B	C
1	3	1	2
2	1	2	3
3	2	3	1
4	1	2	3
5	3	1	2
6	2	3	1
7	1	2	3
8	3	1	2
9	1	2	3
10	2	3	1
11	3	1	2
12	2	3	1
13	3	1	2
14	1	2	3
15	2	3	1
T_j	**30**	**30**	**30**

***K*=15**

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture. 0.033, ns

$$\chi_r^2 = \frac{12}{KJ(J+1)}\sum T_i^2 - 3K(J+1)$$

$0 \le \chi_r^2 \le K(J-1)$, *df* = *J*-1,

If $J > 3$ *and* $K > 9$, *use* χ^2

Otherwise use tables for χ_r^2

1

Nonparametric Statistics: Pros & Cons

Pros

1. No assumed population normality or homogeneity of variance.

2. Even data that may be higher order than ordinal may be evaluated with relaxed statistical assumptions.

3. Some nonparametric tests may be used with very small sample sizes (~5) and provide exact probabilities (e.g. binomial test)

4. Nonparametric tests may be applied to nominal data for which there are no alternatives.

5. Simpler to calculate, may aid intuition about size of effects.

Cons

1. Nonp … r
samp

Power-efficient of $test_B$ = (100) N_A/N_B

N_A = Observations for given power for $test_A$

N_B = Observations for the same power for $test_B$

2. Nonp … s
(e.g. …

3. Converting data to ranks throws away scientifically interesting ordinal or ratio information.

Some Heuristics for Behavioral Experimentation

In General

- Statistics are ideally descriptive and reinforce results evident by plots and model fits, the goals of an experiment are *data and models*, *not statistics*.
- Review handbooks/design and user performance reference material before starting.

About Methods

- *Placebo* and *Hawthorne* effects are real: consider a variety of control groups.
- Use balanced independent groups for major independent variables when possible, distribute group assignment over experimental run.
- Evaluate behavior related to closed-loop performance.
- Check statistical assumptions when possible, i.e. normality, at least unimodality, symmetry and equality of variance.

About Results and Conclusions

- Results should not be dependent upon a specific measurement scale
- Results should be robust to exclusion of outliers.
- Statistical conclusions should not depend upon a specific analytic approach.

References

Aristotle, (~350BC) Posterior Analytics, book 2. http://etext.library.adelaide.edu.au/a/aristotle/a8poa/

Adelstein, Bernard D., Johnston, Eric R., and Ellis, Stephen R. (1996) Dynamic response of electromagnetic spatial displacement trackers. *Presence* , *5*, 3, 302-318

Columbia Accident Investigation Board (CAIB) (2003) Volume 1, NASA, Government Printing Office, Washington, D.C., http://www.caib.us

Ellis, S. R. (1996) Presence of mind: a reaction to Sheridan's musings on telepresence. *Presence 5*, 2, 247-259.

Ellis, Stephen R. (2000) Collision in space. *Ergonomics in design*, *8*, 1, 4-9.

Ellis, Stephen, McGreevy, Michael W., and Hitchcock, Robert (1987) Perspective traffic display format and airline pilot traffic avoidance. *Human Factors*, 29, 371-382.

Ellis, Stephen R., Wolfram, Anthony, and Adelstein, Bernard D. (2002)Proceedings of the Human Factors and Ergonomics Society, 2002, Baltimore MD, pp. 2149-2154.

Ellis, Stephen R. (1993) What are virtual environments? *Computer Graphics and Applications*, *14*, 1, 17-22.

McCandless, Jeffrey W., Ellis, Stephen R., and Bernard D. Adelstein (2000) Localization of a time-delayed monocular virtual object superimposed on a real environment. *Presence*, *9*, 1,15-24.

Siegel, Sidney (1956) *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York.

Stevens, S.S. (1946) On the theory of scales of measurement. *Science*,*103*, 677-680.

Smith, J.D., Ellis, Stephen R., and Lee, Edward C. (1984) Perceived threat and avoidance maneuvers in response to cockpit traffic displays. *Human Factors*, *26*, 1, 33-48.

Tufte, E. (1990) Envisioning Information. Graphics Press, Cheshire, CT.

Velleman, Paul F. and Wilkinson, Leland (1993) Nominal, ordinal, interval and ratio typologies are misleading. *The American Statistician*, *47*,1,65-72.

End