Quantitative and Qualitative Methods for Human-Subject Experiments in Virtual and Augmented Reality

VR 2012 Tutorial

Joseph L. Gabbard, Virginia Tech (organizer)
J. Edward Swan II, Mississippi State University
Stephen R. Ellis, NASA Ames Research Center

Experimental Design and Analysis

J. Edward Swan II, Ph.D.

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

Why Human Subject (HS) Experiments?

- **Graphics hardware / software more mature**
- **Sophisticated interactive techniques possible**

- **Focus of field:**
  - **Implementing technology → using technology**
  - **Trend at IEEE Visualization, 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 VR / AR for applications**

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

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

- **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	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 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 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

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

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

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

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

Counterbalancing Example

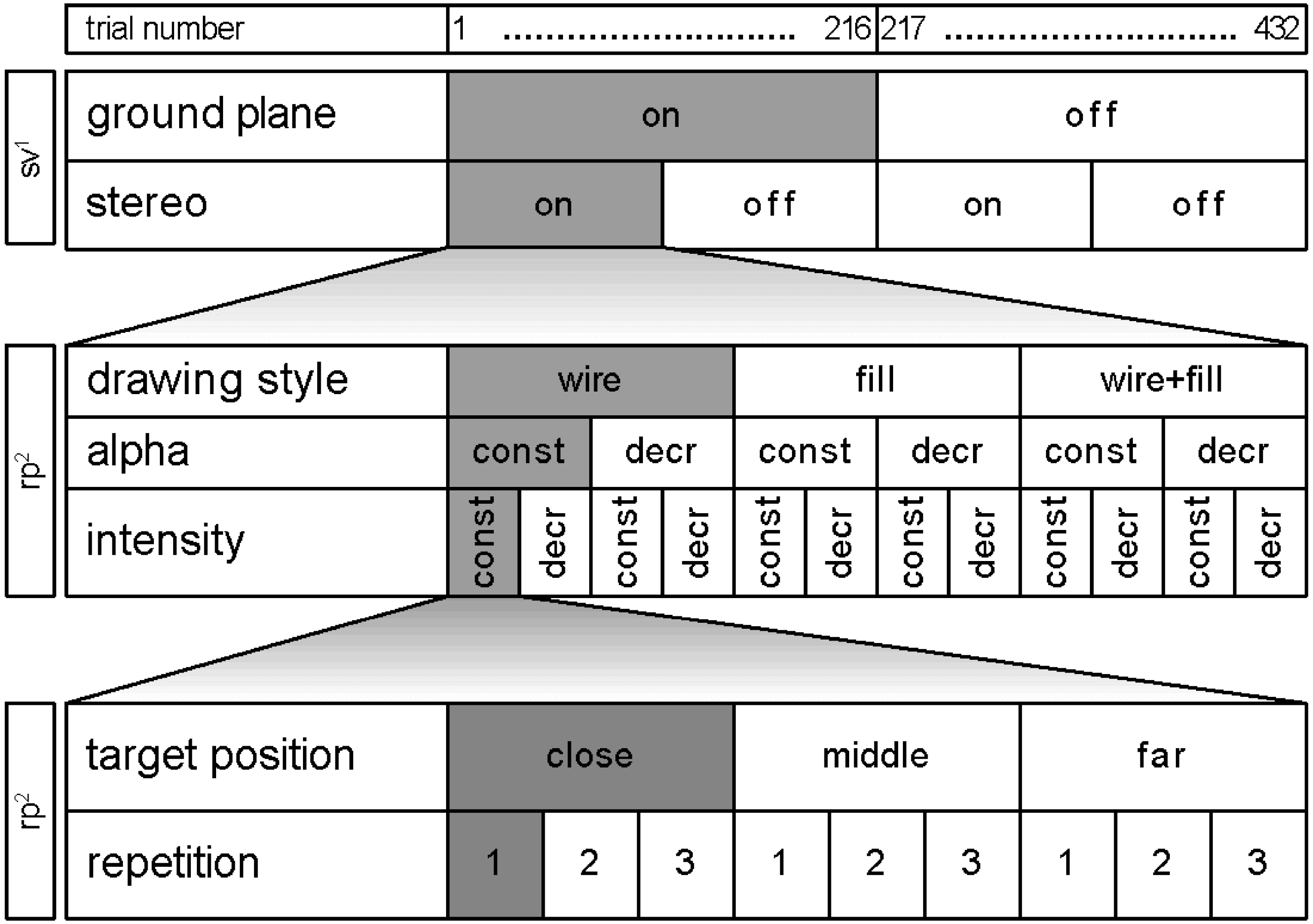
- "A 3 (VE terrain type) by 2 (stereopsis) within-subjects design, with 4 repetitions of each cell."
- Form Cartesian product of Latin squares {6 x 3} (VE Terrain Type) ⊗ {2 x 2} (Stereopsis)
- Perfectly counterbalances groups of 12 subjects

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



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

- **All variables within-subject**

From [Living Swan 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* *wall* *workbench* *desktop*	subjects 1 – 4	subjects 5 – 8	subjects 9 – 12	subjects 13 – 16	subjects 17 – 20	subjects 21 – 24	subjects 25 – 28	subjects 29 – 32

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

From [Swan et al. 03]

Types of Statistics

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

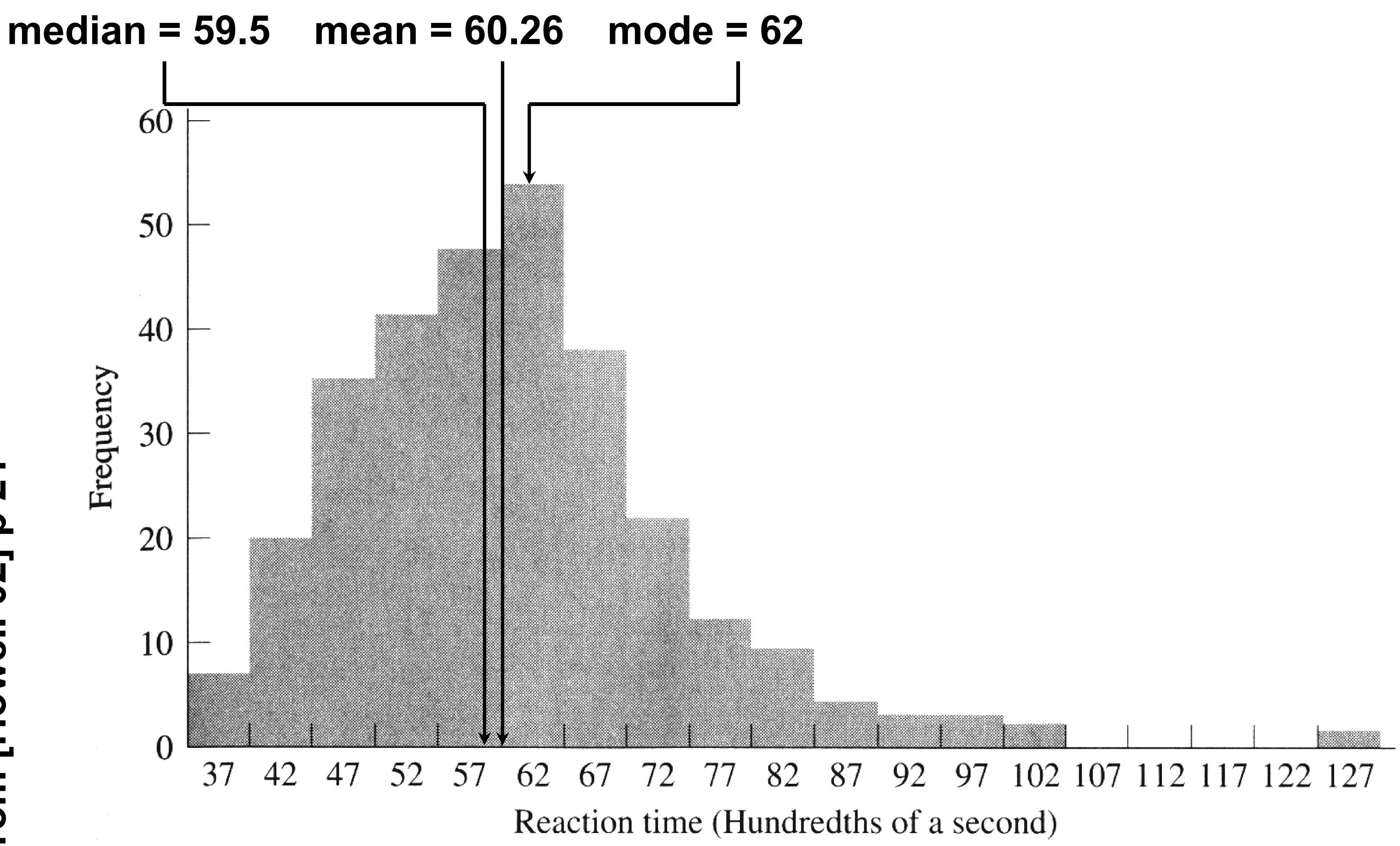
- **Inferential Statistics**:
  - Detect relationships in data
  - Significance tests
  - Infer population characteristics from sample characteristics

Graphing Data

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

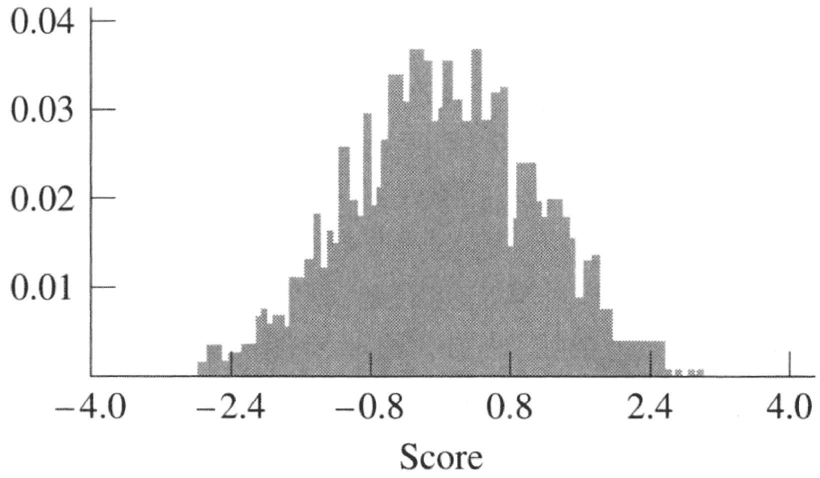
Exploring Data with Graphs

- **Histogram common data overview method**

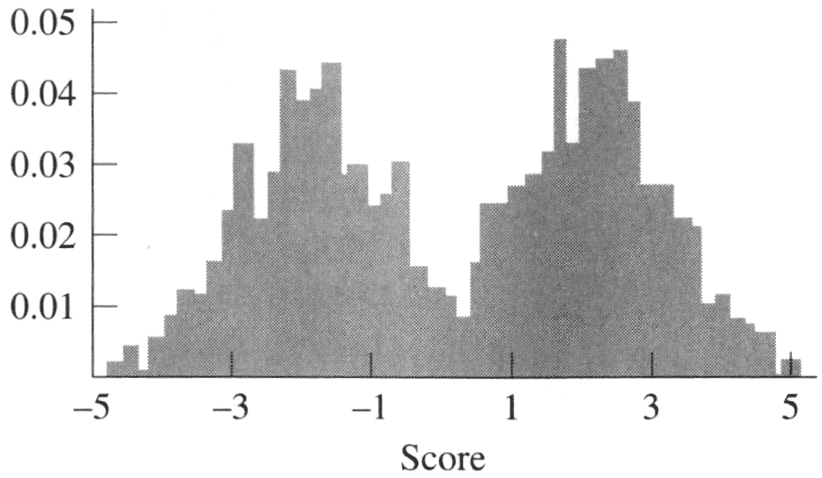


From [Howell 02] p 21

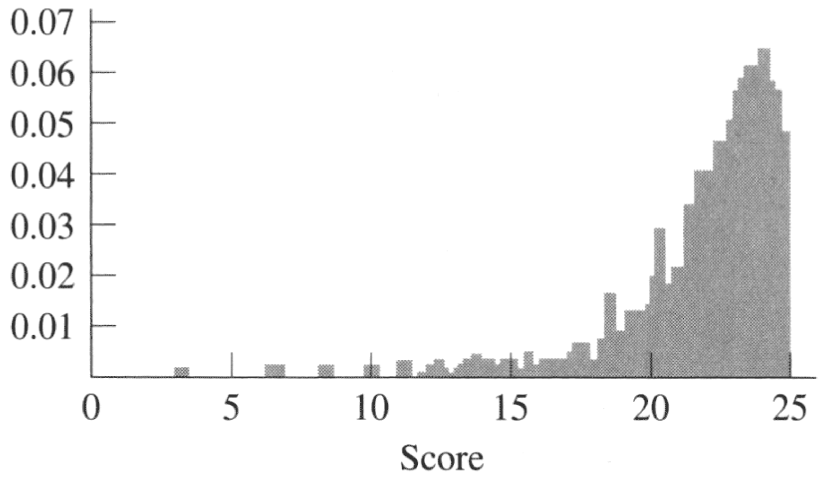
Classifying Data with Histograms



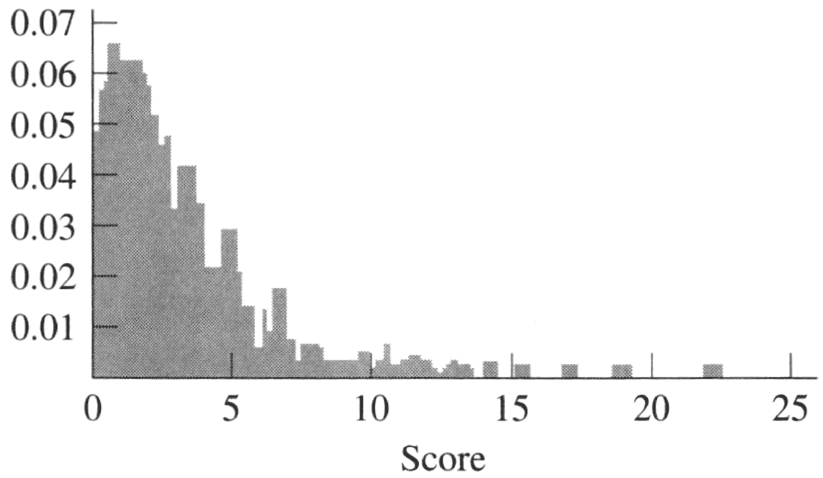
(a) Normal



(b) Bimodal



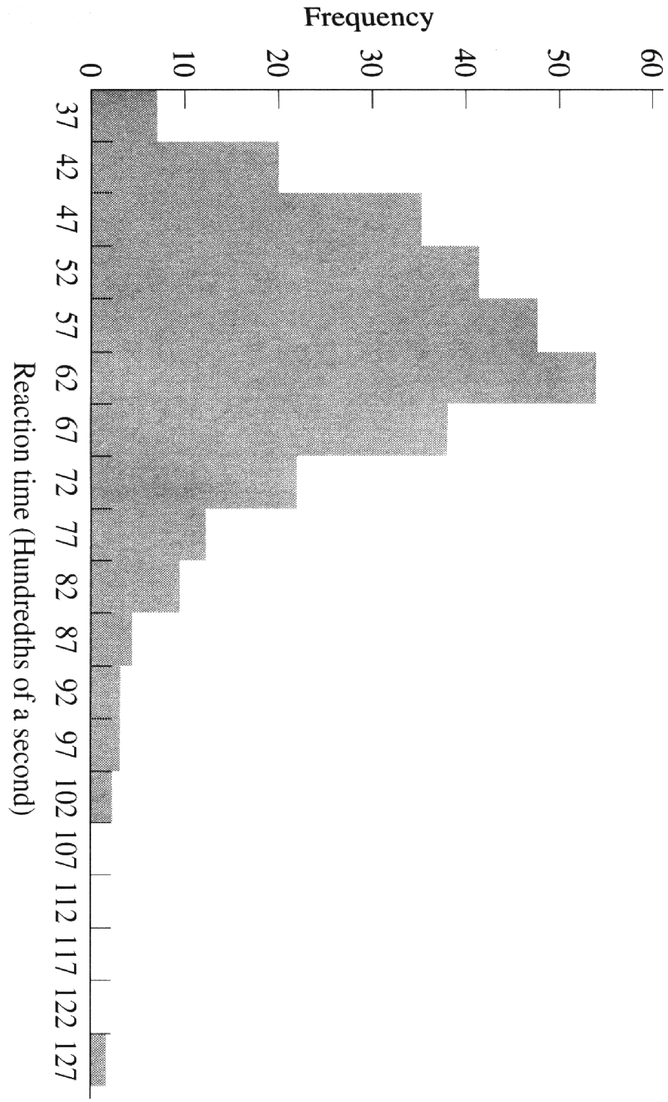
(c) Negatively skewed



(d) Positively skewed

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*	00000111111111111
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.	88888888888899999999999
53 54 54 54 54 54 54 55	6*	00000000000011111111111
55 55 55 55 55 55	6t	222222222222222333333333
	6f	444444455555555
	6s	6666666677777777777777
	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

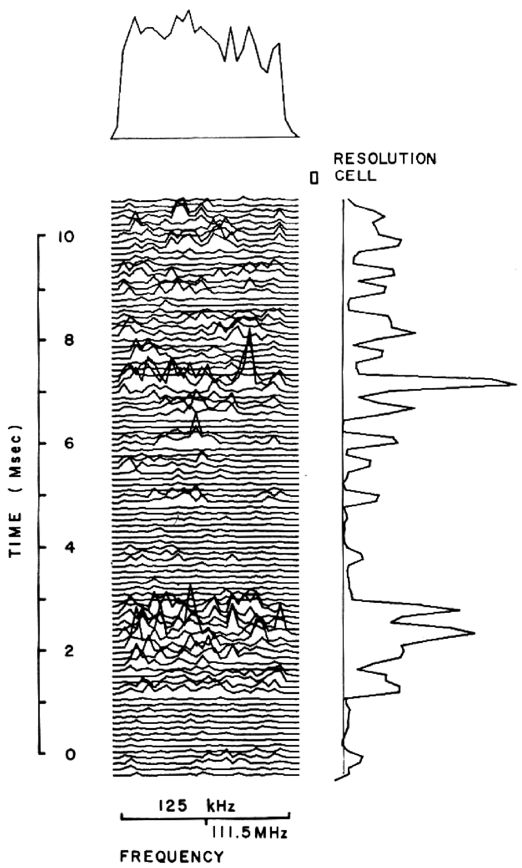
Midterm 1

%	Count				
3%	1	0	0		
0%	0	1			
0%	0	2			
0%	0	3			
0%	0	4			
13%	5	5	04689		
8%	3	6	249	F	3%
26%	10	7	0011122568	D	13%
24%	9	8	123334678	C	34%
24%	9	9	002222366	B	24%
3%	1	10	0	A	26%
sum:	38				

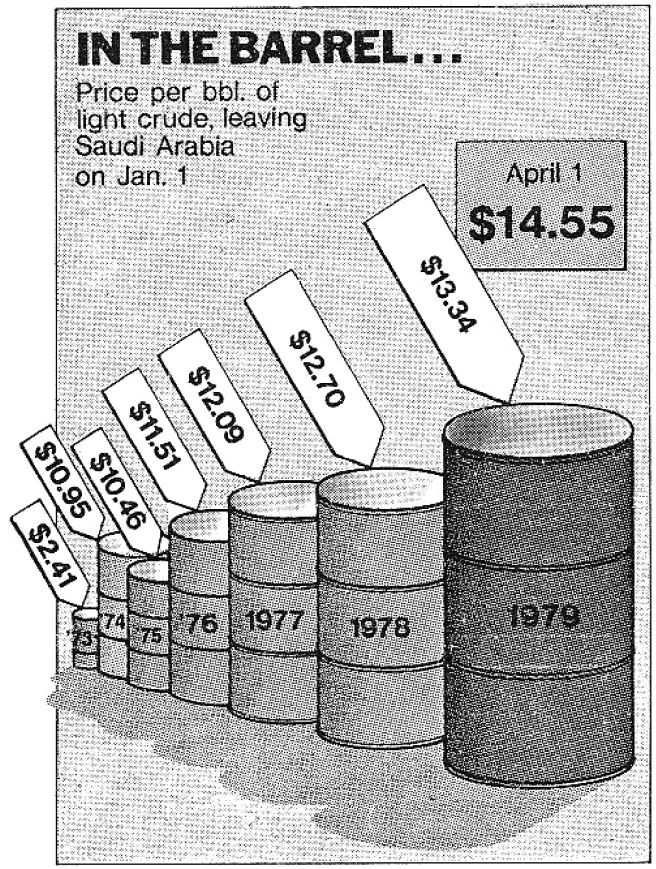
Grades from my fall 2011 Formal Languages class; first midterm

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

- **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: 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:

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

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

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**

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

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

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