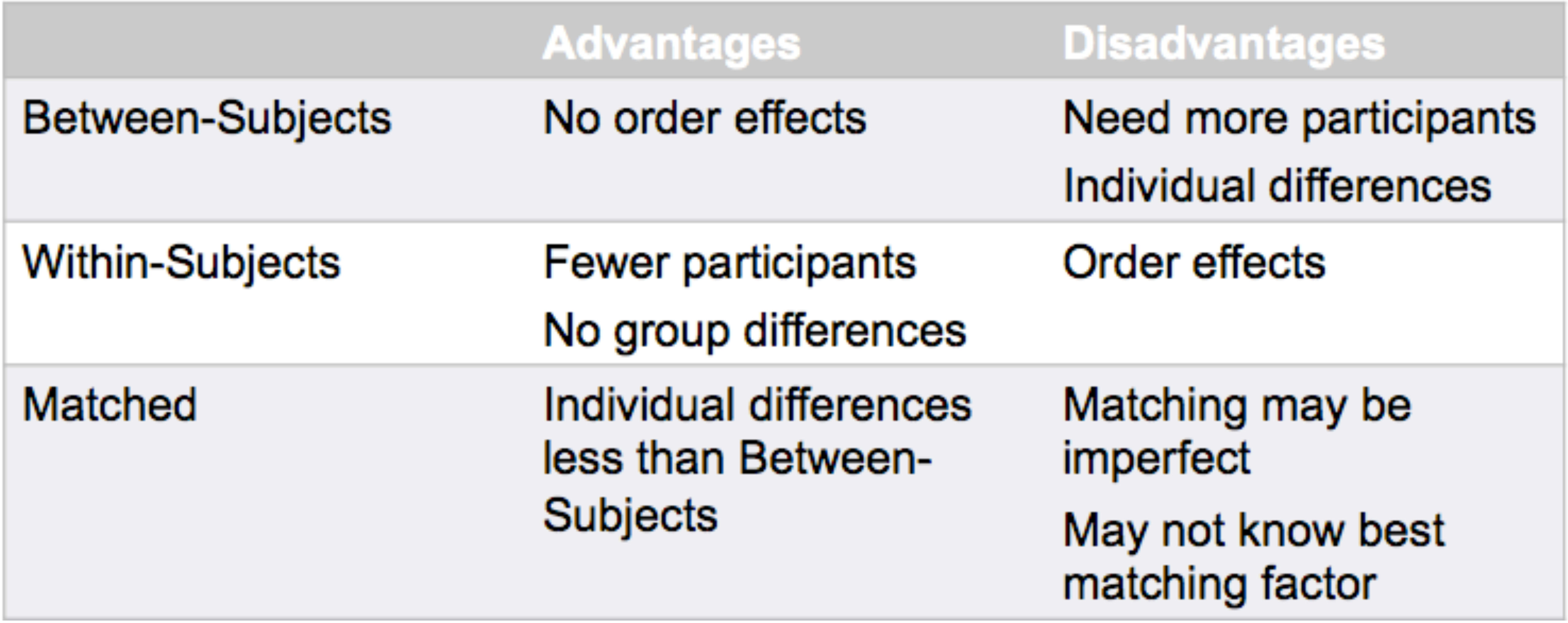
CSCI 310: Experiment Design and Analysis

Sarah Carruthers
Fall 2014

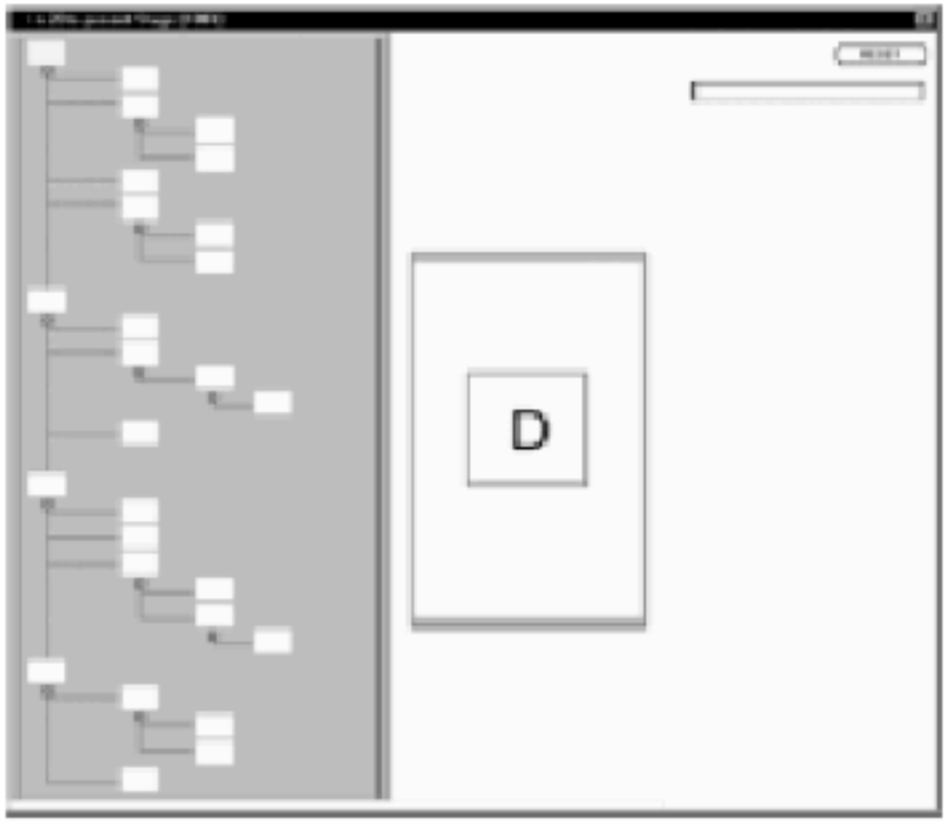
Pros/cons

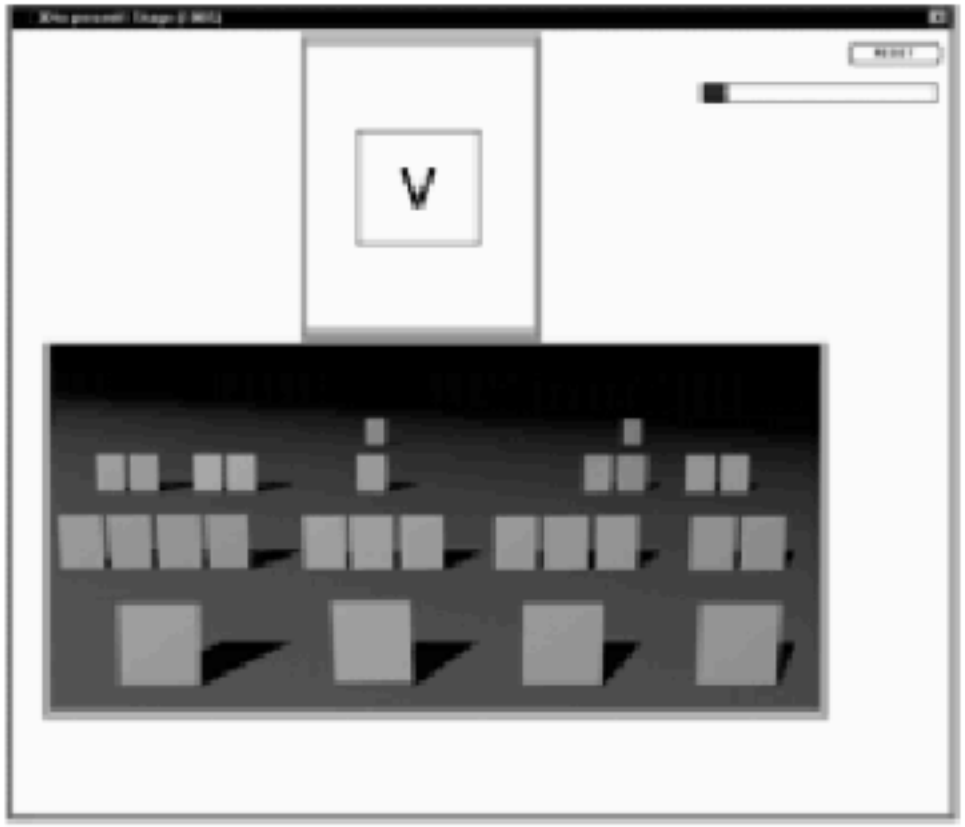


	Advantages	Disadvantages
Between-Subjects	No order effects	Need more participants Individual differences
Within-Subjects	Fewer participants No group differences	Order effects
Matched	Individual differences less than Between-Subjects	Matching may be imperfect May not know best matching factor

Exercise

- Do people remember information better with 2D displays or 3D displays? Participants were randomly assigned to 2D and 3D display groups. Instructions for experimental task were the same for both groups





Exercise

- An experiment compares
  - 1: maps with overview window to
  - 2: maps without an overview
- So that they do not repeat the same task twice, participants user a map of California with 1, and a map or Oregon with 2

Exercise: Choose a design

- Design an experiment to test the effect of the following factors on the speed and accuracy of selecting an item in a menu:
  - menu type: pop-up or pull-down
  - menu length: 3,6,9,12,15 items
  - subject type: expert or novice
- Hypothesis?
- Independent/dependent variables
- Methods?
  - Decide whether each factor will be within or between subjects
- Based on chosen methods how do you assign your task?

Example solution

- Hypotheses:
  - Experts will be faster with pull down menus, than novices, independent of menu length
  - Speed will be faster for shorter menus, but accuracy will be poorer, for both kinds of participant, and both kind of menu.
- Independent variables:
  - menu type, menu length, subject type
- Dependent variables:
  - speed and accuracy

Example solution

- Within subject factors: menu length
- Between subject factors: menu type, subject type

Novices		Experts	
Popup	Pulldown	Popup	Pulldown
3	3	3	3
6	6	6	6
9	9	9	9
12	12	12	12
15	15	15	15

Alternative and Null Hypotheses

- Alternate hypothesis is a prediction of a relationship between variables
- The null hypothesis states that there will be no relationship between variables
- It is the null hypothesis that is tested, leading us to accept it or reject it
  - if we accept it, this means we conclude that the alternative hypothesis is false
- GOAL: prove the prediction
- ACTION: disprove the null hypothesis

Example null hypothesis

- H_0: there is no difference in user performance (time and error rate) when selecting a single item from a popup or pulldown menu, regardless of the subject's previous expertise in using a mouse or using different menu types

Analysis of Data

- Before you start to do any statistical analysis:
  - look at data
  - save original data
- Choice of statistical technique depends on
  - type of data
  - information required
- Type of data
  - discrete: finite number of values
  - continuous: any value

Back to example

- What kinds of data will we be collecting?
  - discrete?
  - continuous?

Back to example

- What kinds of data will we be collecting?
  - timing data
    - ms from cue to task completion
    - continuous
  - accuracy data
    - right or wrong?
    - discrete

Data, data, data

- Which is preferred?
  - quantitative
  - qualitative
- Can you trust your data?
  - what factors contribute to validity of data?
    - human factors
    - measurement
- How do you know?
- What if you can’t?

Back to our example

- Can you trust your data?
  - timing data
  - accuracy data
- What if we added observation data:
  - how might that impact how much we can trust it?
  - how could we deal with it?

Preparing the data for analysis

- Score the data
- Determine what types of scores to analyze
  - **single item scores**: e.g. 1 for no, 2 for yes
  - **summed scores**: may want to sum responses to all questions on the instrument. Done when individual items may not completely capture a participant's perspective
  - **difference scores**: scores in a quantitative study that represent a difference or change for each individual (e.g. compare at different time intervals)

Selecting a statistical program

- Find a program with documentation that includes sample data and tutorials
  - Is there support available?
- Ease of use
- Ease of learning
- Make sure the program includes the types of statistics you need to do
- Can it analyze the *amount of data* you are going to have
- Useful to have a program that outputs graphs and charts and tables
- Costs? Academic or campus licenses?

Some examples

- Minitabl3 (http://www.minitab.com)
- StatView (http://www.statview.com)
- SYSTAT (http://www.spssscience.com)
- SAS/STAT (http://www.sas.com)
- SPSS (http://www.spss.com)
- R package (http://www.r-project.org/)

R basics

```
    > x = 1 #assignment
> x       #request output
[1] 1    #output

> n = c(2, 3, 5)    #declare a vector
> s = c("aa", "bb", "cc", "dd", "ee") #declare a vector
> b = c(TRUE, FALSE, TRUE, FALSE, FALSE) #declare a vector
> x = list(n, s, b, 3)   # x contains copies of n, s, b

> x[2] #get a slice of a list
[[1]]
[1] "aa" "bb" "cc" "dd" "ee"

#dataframes are for storing data tables
# a list of vectors of equal length
> n = c(2, 3, 5)
> s = c("aa", "bb", "cc")
> b = c(TRUE, FALSE, TRUE)
> df = data.frame(n, s, b)       # df is a data frame
```

R basics

```
    #built in data frames:
> mtcars
              mpg cyl disp  hp drat   wt ...
Mazda RX4     21.0   6  160 110 3.90 2.62 ...
Mazda RX4 Wag 21.0   6  160 110 3.90 2.88 ...
Datsun 710    22.8   4  108  93 3.85 2.32 ...

#Here is the cell value from the first row, second column of mtcars.

> mtcars[1, 2]
[1] 6

#we can use the row and column names instead of the numeric coordinates.

> mtcars["Mazda RX4", "cyl"]
[1] 6
#preview data frame
> head(mtcars)
              mpg cyl disp  hp drat   wt ...
Mazda RX4     21.0   6  160 110 3.90 2.62 ...
              ............
```

Inputting data

- Generally use sheets similar to spreadsheet programs
- Can often import from excel
- Enter the data by rows for each individual
- Assign an identification number for each individual
- Replace column headings with variable names
- be careful!

R data import

```
#importing table files
#cells separated by blank characters
#100   a1   b1
#200   a2   b2
#300   a3   b3
#400   a4   b4
> mydata = read.table("mydata.txt")  # read text file
> mydata                              # print data frame
   V1 V2 V3
1 100 a1 b1
2 200 a2 b2
3 300 a3 b3
4 400 a4 b4

#or csv file
#Col1,Col2,Col3
#100,a1,b1
#200,a2,b2
#300,a3,b3
> mydata = read.csv("mydata.csv")  # read csv file
> mydata
  Col1 Col2 Col3
1  100   a1   b1
2  200   a2   b2
3  300   a3   b3
```

Data management

- For even moderate sized studies:
    - Keep all data (or for one condition) in one sheet
    - Extract needed data for particular analyses
- May need to clean data
    - missing data
    - scores out of range
    - How to fix?
        - eliminate participants
        - replace with substitute values
            - **check with a statistician before doing this!

Levels of measurement

- Nominal measurement
  - e.g. program of study
  - assigned a number (female = 1, male = 2)
- Ordinal measurement
  - e.g. strongly disagree - strongly disagree, scale 1-5
  - intervals may not be truly equal
- Ratio measurement
  - e.g. weight, salary, number of correct answers

Descriptive statistics

- Indicate general tendencies in the data
- Central tendencies: mean, mode, median
- Variance: standard deviation

- But we’re also interested in studying relationships between variables.

Measures of central tendency

- Mean: measure that typifies a set of observations with a single value
  - used on ratio level data
  - sum the list, divide by number of values
- Median: represents the midpoint of distribution
  - can be used on ordinal level data, or in cases where the mean may be problematic
    - when atypical values are present, the median may provide a better description of the data
  - one half of cases are above it, one half below. For even number of cases, take the mean of two middle values.
- Mode: category of variables with the most cases
  - useful for nominal data: e.g. Canada in list of countries of residence, for grad students at UBC
  - Calculated as the highest frequency of any one answer

In R

```
> duration=faithful$eruptions
> mean(duration)
[1] 3.487783
>
> median(duration)
[1] 4
> max(duration)-min(duration)
[1] 3.5
>
```

Measures of dispersion

- Tells us about the variability of the data
  - are most values close to each other, or spread out?
- Range: indicates the gap between the highest and lowest values in a distribution
  - subtract lowest from highest

Measures of dispersion

- Standard deviation: reflects the variability in a set of values
  - very important to know this for most analyses
  - reflects the *average amount of deviation from the mean*

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

- Variance: The standard deviation squared
  - a numerical measure of how the data values are dispersed around the mean

Descriptive analysis example in R

- from: http://www.r-tutor.com/elementary-statistics/numerical-measures/variance
- in R: built in data for old faithful eruptions:

```
> head(faithful)
    eruptions waiting
1       3.600      79
2       1.800      54
3       3.333      74
4       2.283      62
5       4.533      85
6       2.883      55
> duration=faithful$eruptions
> sd(duration)
[1] 1.141371
> var(duration)
[1] 1.302728
>
```

R commands

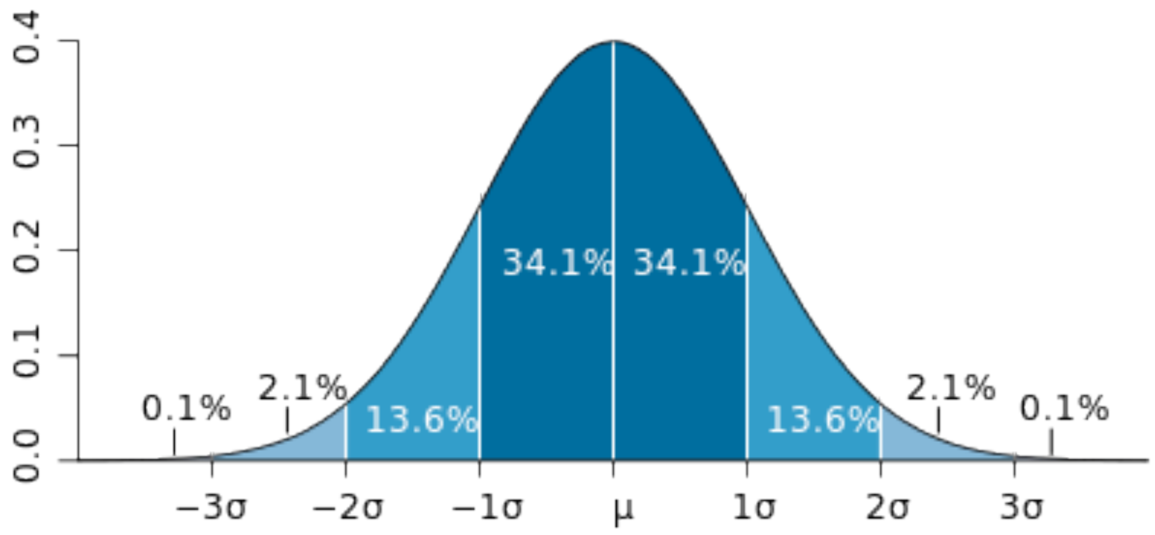
- head(): peek at the top few elements
- sd(): standard deviation
- var(): variance

Standardizing data

- Adjustments are made so that comparisons between units of different size may be made
- Data may also be standardized to create variables that have a similar variability in them (Z Scores)
- Proportions, percentages (proportion x 100)
- Percent change (e.g. between measurements)
- Rates (per standard unit, e.g.100,000)

Normal Distributions

- Bell-shaped curve:
  - few cases on the extremes, most in the middle



  - about 2/3 of cases should fall within one standard deviations of the mean
  - approx. 95% of the cases should fall within 2 standard deviations of the mean

http://en.wikipedia.org/wiki/Normal_distribution

Properties of a normal distribution

- The mean, mode and median values will be the same
- As the number of observations and number of measurement units becomes finer, the distribution curve will become smoother

Z-scores

- Measures the distance, in standard deviation units, of any value in a distribution from the mean: $z = (x-\mu)/\sigma$
  - μ is mean, σ is standard deviation, x is a raw score to be standardized
  - e.g. if someone's income has a z-score of +1.43, it would indicate that their income is 1.43 standard deviation units above the mean of the distribution

Chi-squared test

- Allows us to compare the rates of observed and expected results
- measure of goodness of fit
- As the number of trials increase:
  - the confidence of the results increase
    - $\chi^2 = \sum(O_i - E_i)^2/E_i$
    - O_i observed frequency
    - E_i expected frequency

Chi-squared test

- Calculate the chi-squared statistic
- Determine the degrees of freedom (df)
  - the number of frequencies minus the number of parameters
- Compare to the critical value from the chi-squared distribution

Example: testing fairness of dice

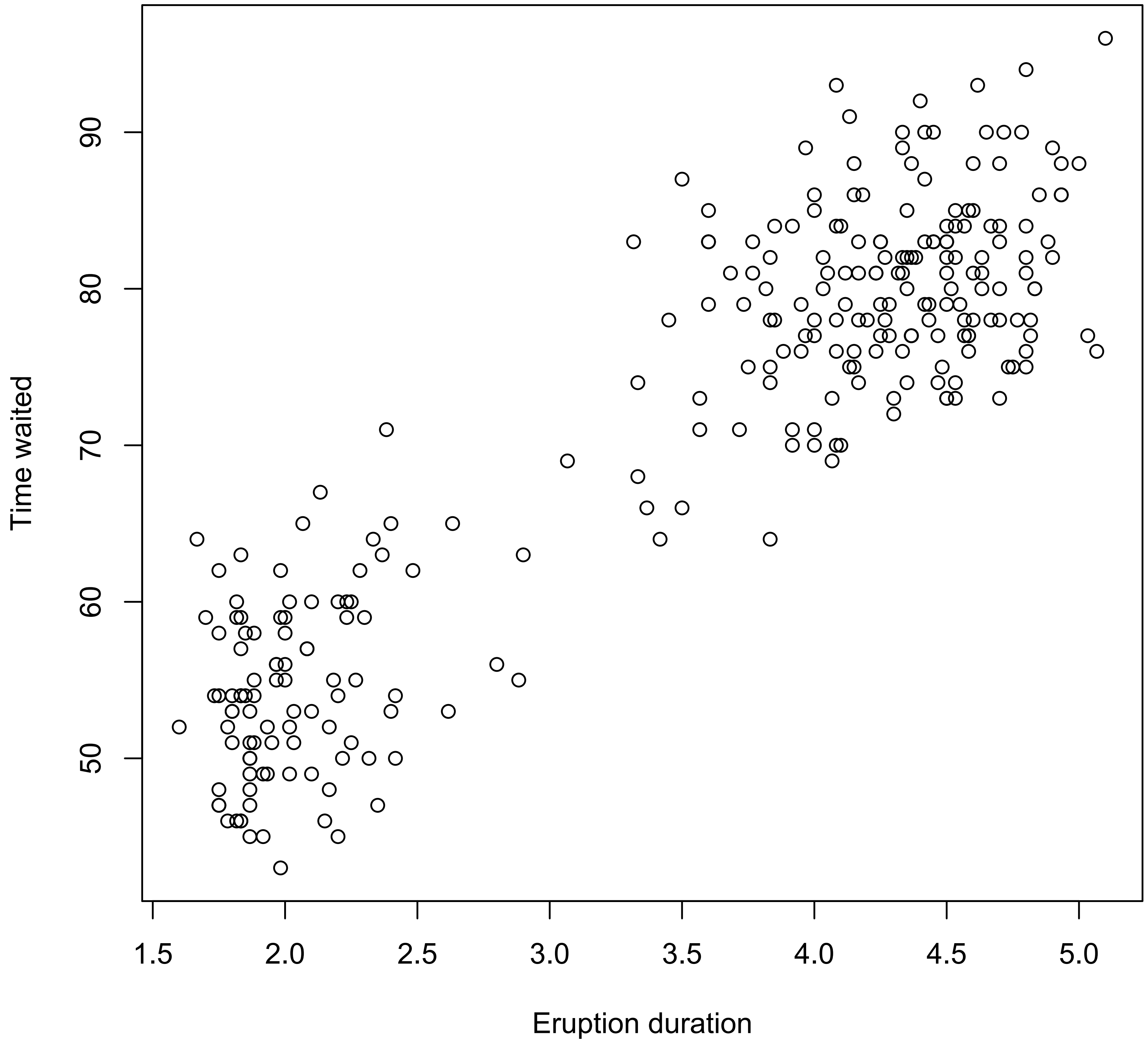
- Suppose you throw a 6-sided die 60 times
- you *expect* each face to land equally likely (10 times)
- you *observe* the following: 1: 5, 2: 8, 3: 9, 4: 8, 5: 10, 6: 20
- *df* is n-1 = 5
- $\chi^2 = 13.4$
- question: is the die biased with a significance of 95%? 99%?
  - look up in table: critical value for df = 5, at 95% is 11.07. our value exceeds this, therefore we reject null, and conclude die is biased at this level
  - look up in table: critical value for df = 5 at 99% is 15.086, therefore we cannot reject the null. There is not sufficient evidence to show that the die is biased at this level

adapted from: http://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

Graphical Visualization

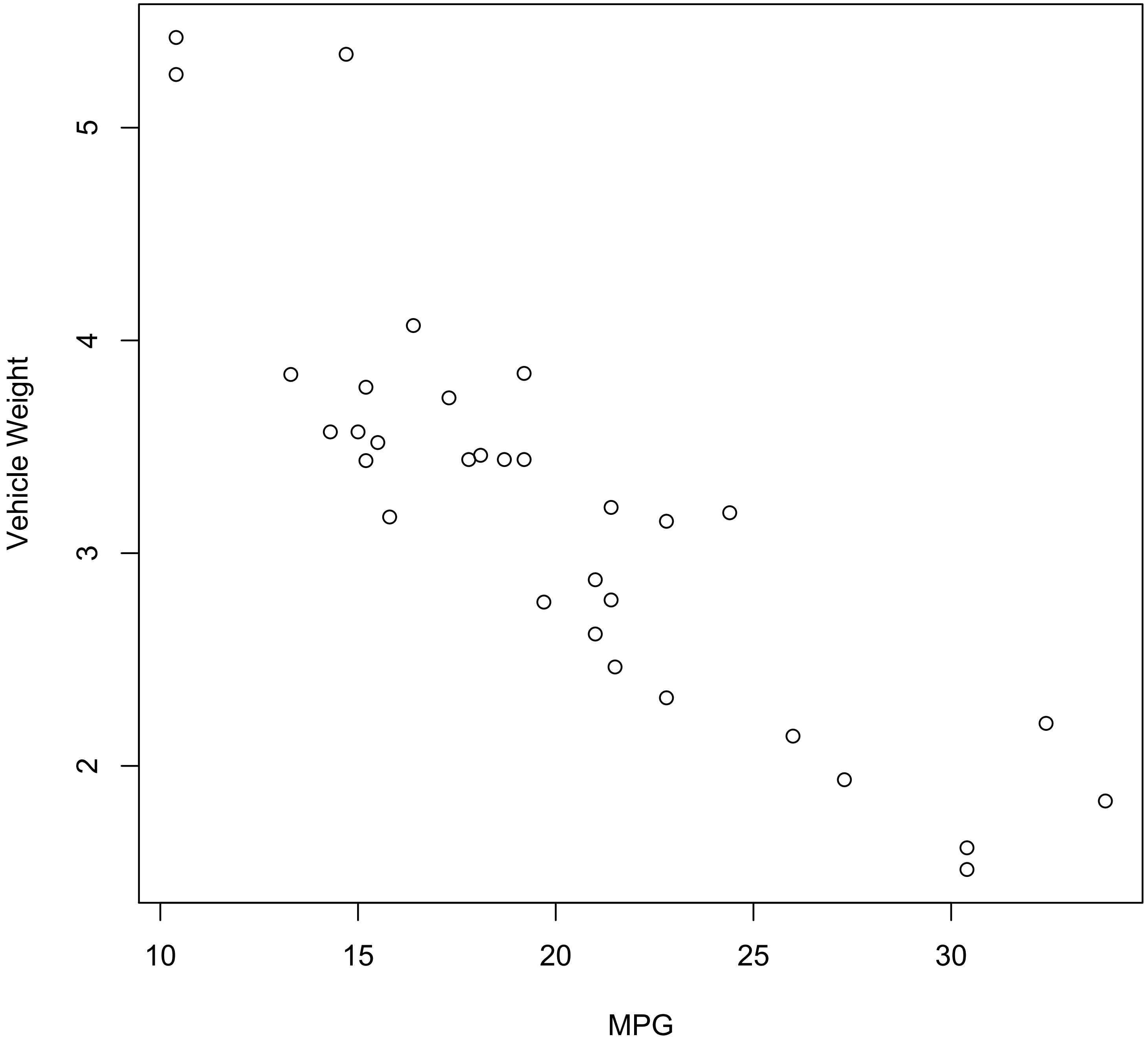
- Scatter plots (pairwise samples are plotted in two dimensions)
  - Good for seeing dependencies between variables
  - Can see linear tendency
  - Observe potential outliers
  - Can see how concentrated data points may be

Graphical visualization



```
R commands:
> waiting=faithful$waiting
> head(cbind(duration,waiting))
     duration waiting
[1,]    3.600      79
[2,]    1.800      54
[3,]    3.333      74
[4,]    2.283      62
[5,]    4.533      85
[6,]    2.883      55
> plot(duration, waiting,
xlab="Eruption duration",
ylab="Time waited")
>
```

Graphical visualization

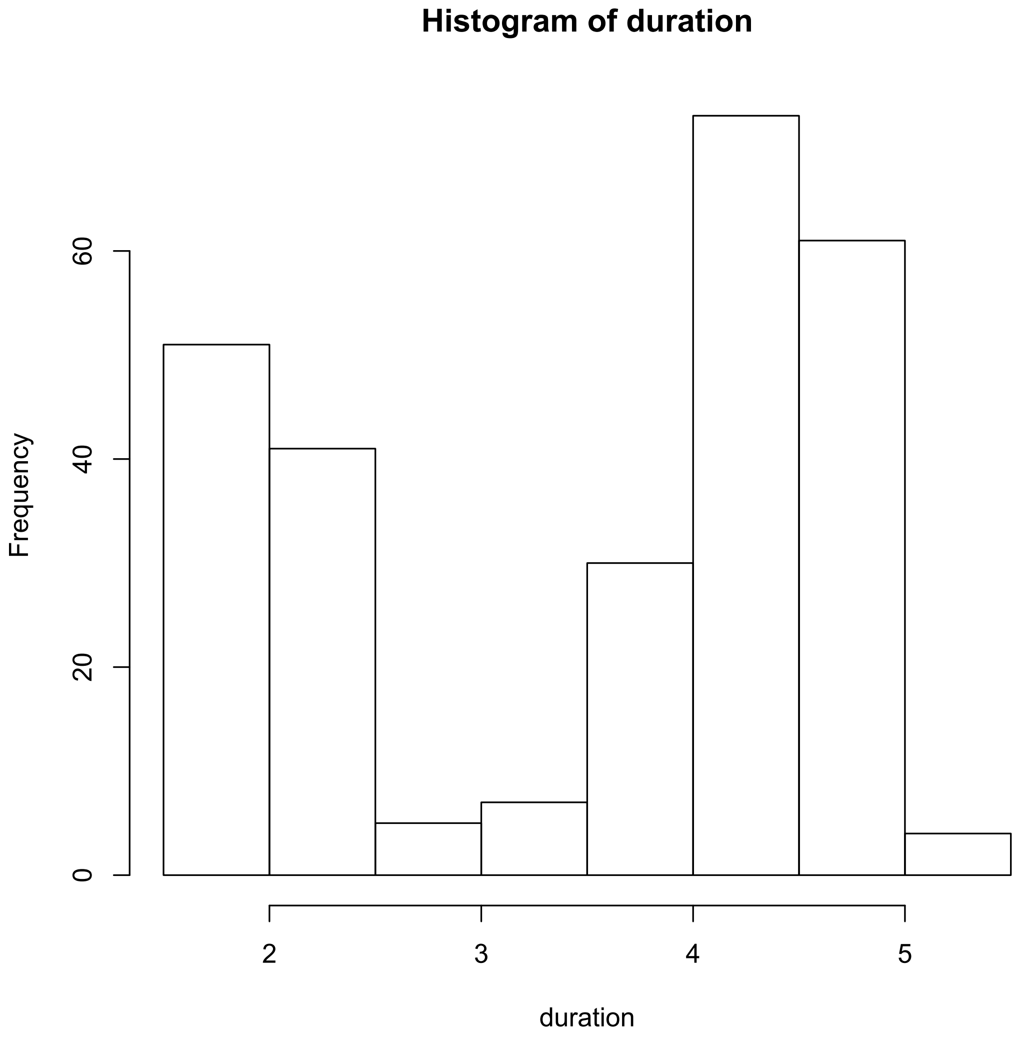


```
> economy=mtcars$mpg
> weight=mtcars$wt
> plot(economy, weight,
xlab="MPG", ylab="Vehicle
Weight")
>
```

Histograms

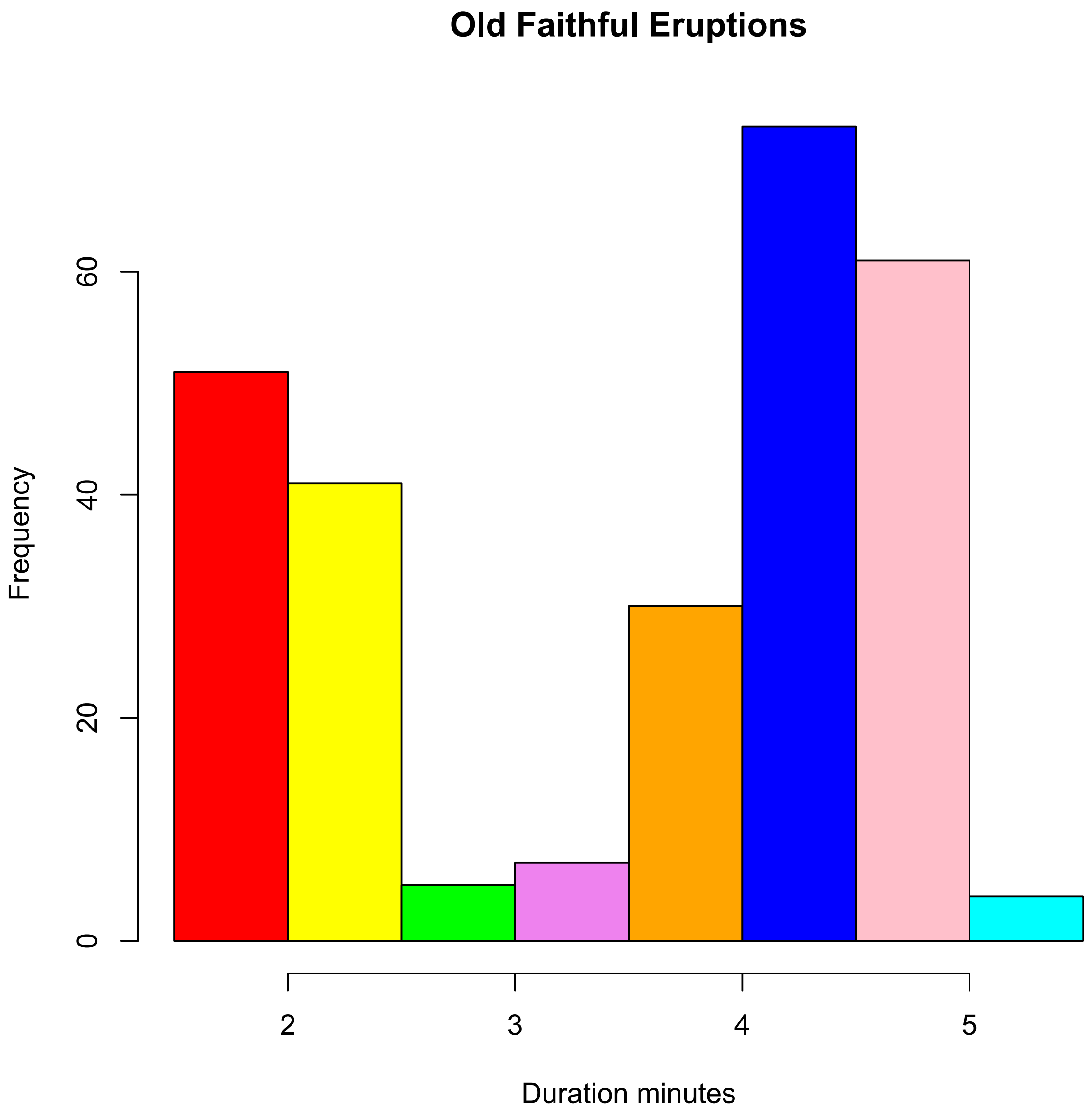
- Give an overview of the distribution density of the samples from one variable
- Histogram consists of bars with heights that represent the frequency of a value or interval of values (graphical representation of a frequency table)
- Such a plot can be useful in determining if the distribution is normal

Histogram



```
> duration = faithful$eruptions
> hist(duration, right=FALSE)    # intervals close
d on the left
```

Histogram

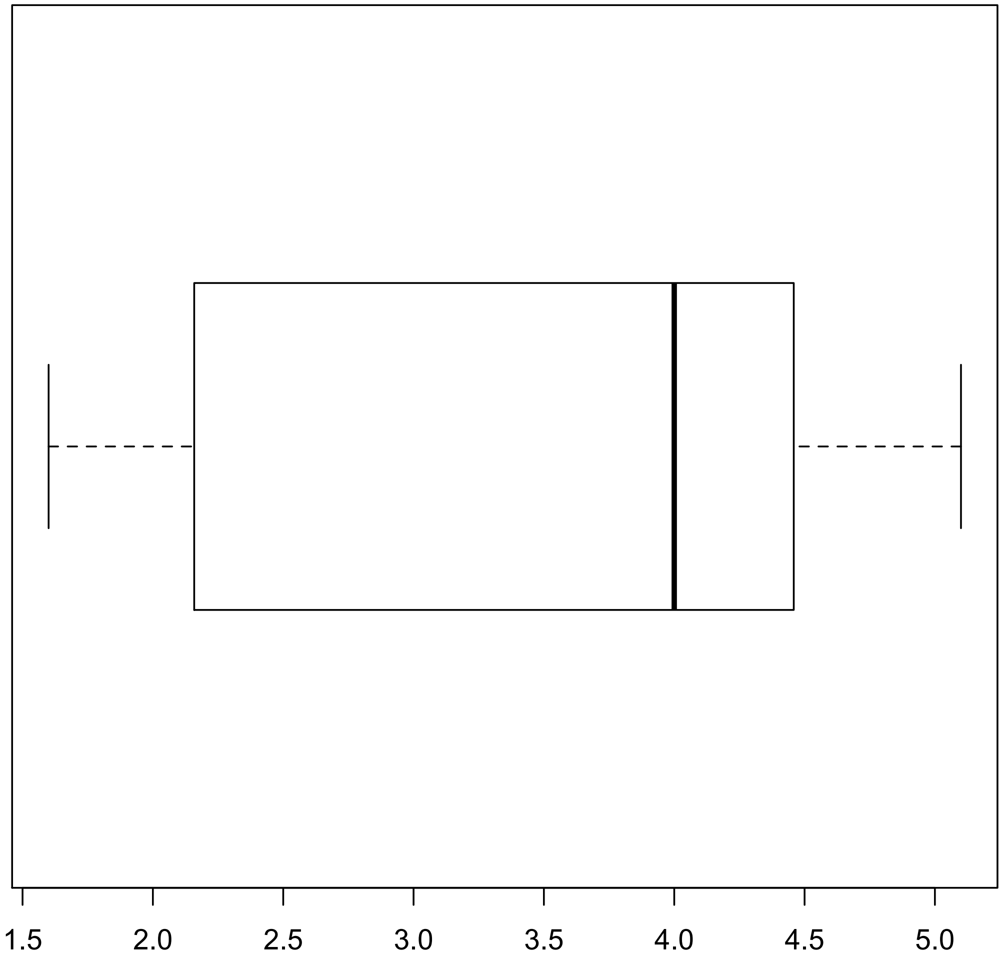


```
> duration = faithful$eruptions
>colors=c("red", "yellow", "green", "violet", "orange", "blue", "pink", "cyan")
> hist(duration,right=FALSE,col=colors,main="Old Faithful Eruptions", xlab="Duration minutes")
>
```

Graphical visualization

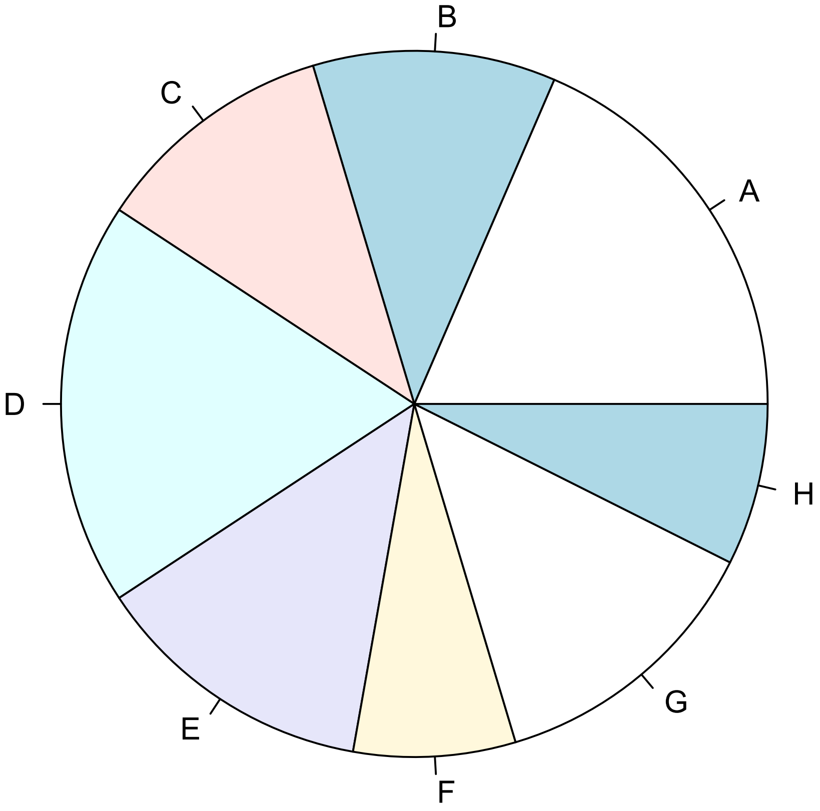
- Box plots:
    - good for visualizing the dispersion and skewedness of samples
    - Constructed by indicating different percentiles graphically
    - Ends of the box indicate the 25% adn 75% quartiles, with median in the middle
    - Tails of the box indicate upper and lower bounds if the distribution is normal
    - Values outside the lower and upper tails are extreme outliers
- Pie Charts are useful for displaying the relative frequency of data values

Box Plots



```
>duration = faithful$eruptions
>boxplot(duration,horizontal=TRUE)
```

Box Plots



```
> school=painters$School
> school.freq=table(school)
> pie(school.freq)
```

Statistical significance

- A test of significance reports the probability that an observed association of difference is the result of sampling fluctuations
- The concept of statistical significance is based on the assumption that events which occur very infrequently by chance, are “significant”
- Traditionally, events which occur less than 5% of time (or p-value < 0.05) are said to be statistically significant

Analysis of variance (ANOVA)

- Compare the variance of multiple groups of data
- Variance might be
  - due to chance
  - due to other factors
  - due to our manipulation
- Test whether the means are equal
- Look at how *significant* the difference is

ANOVA

- Assumptions:
  - The expected error values are 0
  - the variances of all errors are equal
  - the errors are independent
  - they are normally distributed

ANOVA example

```
> dat=read.table("PAODeltaAll.csv",header=T)
> head(dat)
  subjectID         PAO      delta isOpt isVC  sizeDelta       PAO.1
1       s15 0.000000000 0.00000000    NO   NO 0.00000000 0.000000000
2       s16 0.000000000 0.00000000    NO   NO 0.00000000 0.000000000
3       s62 0.000000000 0.00000000    NO   NO 0.00000000 0.000000000
4       s94 0.000000000 0.00000000    NO   NO 0.00000000 0.000000000
5       s61 0.003636364 0.04000000    NO   NO 0.04000000 0.003636364
6       s79 0.004132231 0.04545455    NO   NO 0.04545455 0.004132231
> aov.dat=aov(delta~(isVC*isOpt),dat)
> summary(aov.dat)
            Df Sum Sq Mean Sq F value Pr(>F)
isVC         1   0.17   0.174   0.095 0.7592
isOpt        1  17.54  17.539   9.510 0.0027 **
isVC:isOpt   1   0.08   0.081   0.044 0.8346
Residuals   92 169.68   1.844
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> print(model.tables(aov.dat,"means"),digits=4)
Tables of means
Grand mean

0.7642077

 isVC
isVC
    NO    YES
0.8068 0.7216

 isOpt
isOpt
    NO    YES
0.3368 1.1916

 isVC:isOpt
     isOpt
isVC  NO     YES
  NO  0.3504 1.2633
  YES 0.3232 1.1200
```

Design

- Participants assigned to 1 of 4 conditions:
  - Vertex cover (isVC) or Independent set (!isVC)
  - Search (!isOPT) or Optimization (isOPT)
- *2x2 factor design*
- Dependent measure:
  - deltaOpt = |solutionSize - opt|
- Hypotheses:
  1. deltaOpt score will not be different for either the Vertex Cover or Independent Set conditions
  2. deltaOpt score will not be different for either the Search or Optimization conditions
  3. deltaOpt score will not be different for the interaction between these conditions

ANOVA example

```
> dat=read.table("PAODeltaAll.csv",header=T)
> aov.dat=aov(delta~(isVC*isOpt),dat)
> summary(aov.dat)
            Df Sum Sq Mean Sq F value Pr(>F)
isVC         1   0.17   0.174   0.095 0.7592
isOpt        1  17.54  17.539   9.510 0.0027 **
isVC:isOpt   1   0.08   0.081   0.044 0.8346
Residuals   92 169.68   1.844
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The variance is significantly different for the isOpt factor

Therefore null hypothesis #2 can be rejected.

The other two nulls cannot be rejected

ANOVA example

```
>print(model.tables(aov.dat,"means"),digits=4)
Tables of means
Grand mean

0.7642077

 isVC
isVC
    NO    YES
0.8068 0.7216

 isOpt
isOpt
    NO    YES
0.3368 1.1916

 isVC:isOpt
     isOpt
isVC  NO     YES
  NO  0.3504 1.2633
  YES 0.3232 1.1200
```

Some other output:

Grand mean

Means for conditions

How to present ANOVA data

- Explain why analysis technique you used is appropriate
- Describe the measures, and how they were collected/calculated
- Give means for each condition (if appropriate)
- Present ANOVA results:
  - were they significant?

Example

Analysis

For normally (or quasi-normally) distributed data, both analysis of variance (ANOVA) and Bayes factor analysis were used to reveal any effect of the differing conditions. According to the Central Limit Theorem, the sum or average of a number of values approximates a normal distribution, regardless of the original distributions of the individual values (Judd & McClelland, 1997), and therefore these analysis techniques are appropriate for mean solution time measures. While ANOVA has traditionally been used for this type of analysis, Bayes factor has long been known to be a superior method for evaluating competing models. Recent work by Rouder, Morey, Speckman, & Province, (2012) provides an easy to use Bayes factor alternative to ANOVA, including a ready to use package for the R statistical analysis application (The R Project for Statistical Computing, n.d.; Morey, & Rouder, 2012). See Appendix B for an example of how to execute Bayes factor analysis for a within subject design. For Likert data collected, Wicoxon Signed-ranks test was used to determine if the population mean ranks differed, appropriate for paired and ordinal data.

Analysis of variance (ANOVA) revealed no significant effect of device ($p>>0.05$) for either PAO or DFO.

Example

Playability

Playability was assessed in three ways: measures of solution quality, solution time and playability questionnaire items. Playability is in part a balance between making a game challenging and achievable. Solution quality is a measure of achievement, and time required to find a solution is a measure of difficulty. Playability questionnaire items were gathered to assess more qualitative differences in playability between devices.

Measures of Solution Quality

Three measures of solution quality were used to assess playability: mean percent above optimal (PAO), mean difference from optimal (DFO), and total number of optimal solutions. PAO and DFO were calculated for each solution, and averaged by device (iPad or iPhone), see Tables (1, 2) and Figures (3, 4). Mean PAO for iPhone condition was 3.00%, and for iPad was 2.50%. Mean DFO for iPhone was 0.875, and for iPad was 0.708. The mean number of optimal solutions on the iPhone was 2.75, while the mean number of optimal solutions on the iPad was 3.50.

Example

Subject	iPad	iPhone
s1	2.92%	4.98%
s2	2.53%	4.25%
s3	1.21%	1.15%
s4	3.49%	2.45%

Table 1: Mean Percent Above Optimal (PAO) by Device for Each Subject

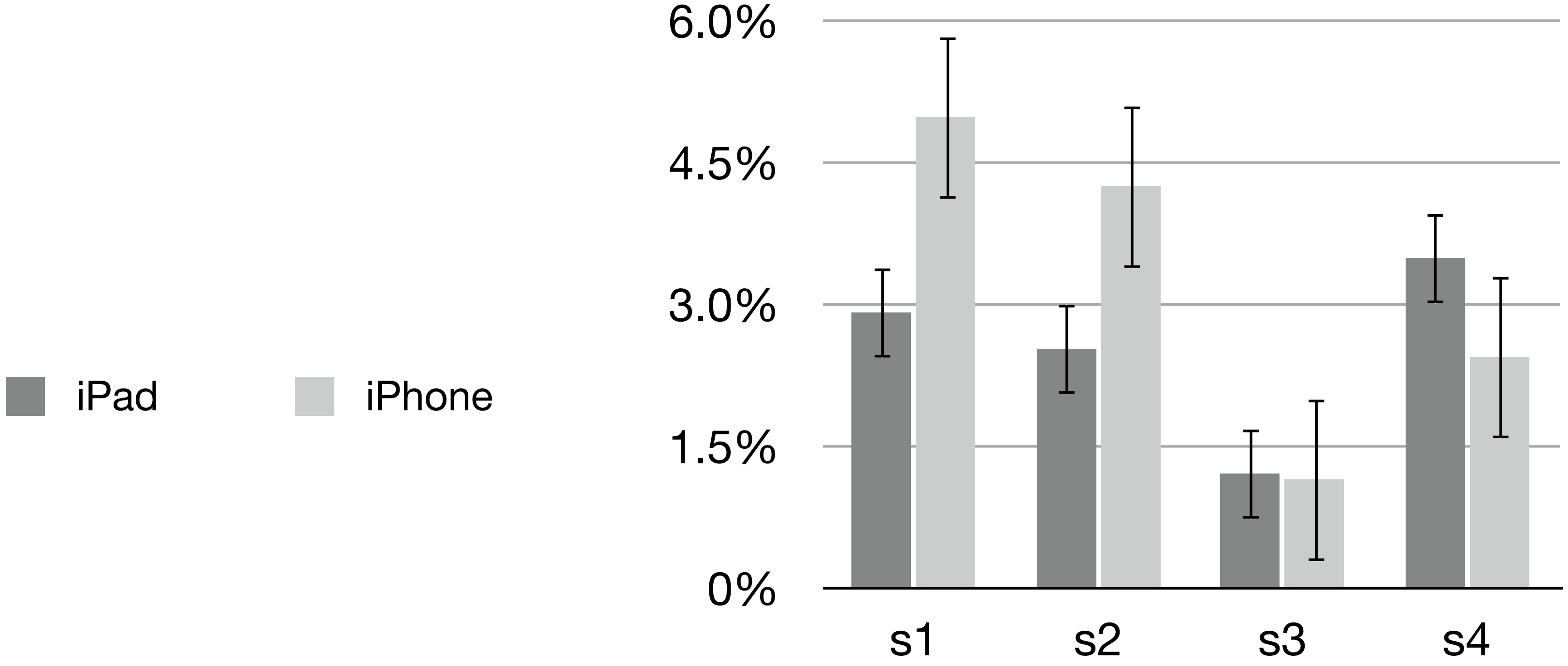


Figure 3: Mean Percent Above Optimal (PAO) by Device for Each Subject

Criticism of null-hypothesis testing

“The American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson & the APA Task Force on Statistical Inference, 1999) initially considered suggesting a ban on the use of NHST, but decided not to, stating instead, “Always provide some effect size estimate when reporting a *p* value” ”

(Wilkinson & the APA Task Force on Statistical Inference, 1999)

“In scientific inference, what we want to know is the probability that the null hypothesis (*H*0) is true given that we have obtained a set of data (*D*); that is, $p(H0|D)$. What null hypothesis significance testing tells us is the probability of obtaining these data or more extreme data if the null hypothesis is true, $p(D|H0)$.”

(Kirk, R. E. (1996)

Criticism of null-hypothesis testing

- There is no way to determine the null is more likely
- People often mis-interpret the results of ANOVA:
    - they infer that a low p-value implies a low probability that the null is based on the data
- “The value of p is the *probability of obtaining a value of a test statistic, say, D, as large as the one obtained-conditional on the null hypothesis true-p(D|H_0)*: which is not the same as the probability that the null hypothesis is true, conditional on the observed result, p(H_0|D)” (Nickerson, 2000)
- Tests the likelihood of an observation if the null is true, *not* the likelihood of the null given the observation

Alternatives to NHST

- Bayes Factor Analysis
  - a measure of evidence for 2 or more competing models
  - ratio of probabilities: $B_{01} = \Pr(D|M_0)/\Pr(D|M_1)$
  - ratio between the probability of the data conditional on model 0, and the probability of the data conditional on model 1
  - reformulated as ratio of the models given the data (posteriors) is equal to the ratio of the probability of the models (the priors

Alternatives to NHST

- Bayes Factor Analysis
  - Calculations are intensive
  - Choose the prior
    - the belief in the two models
    - can’t really be measured
    - need to estimate it
  - for ANOVA prior can be based on Cauchy and Jeffrey’s distributions (JZS priors)

Bayes Factor

- Interpreting the results
- Given two models, the one with the highest Bayes factor is preferred

B	Strength of Evidence
<1:1	Negative (supports M2)
1:1 to 3:1	Barely worth mentioning
3:1 to 10:1	Substantial
10:1 to 30:1	Strong
30:1 to 100:1	Very strong
> 100:1	Decisive

Presenting BF results

- Same as ANOVA, except presenting the actual Bayes Factor:

“Bayes factor analysis (BFA) did not give evidence either for or against the device model for PAO (BF=0.620 ±1.04%) or DFO (BF=0.606 ±1.64%).“

Bayes Factor example

```
>
anovaBF(delta~isVC*isOpt,data=dat,whichRandom="subjectID",prog
ress=FALSE)
Bayes factor analysis
--------------
[1] isOpt                     : 13.90272  ±0.02%
[2] isVC                      : 0.2231203 ±0%
[3] isOpt + isVC              : 3.056703  ±1.52%
[4] isOpt + isVC + isOpt:isVC : 0.930465  ±3.27%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

Presenting the results

- Consider the audience
- Put details of you methods in an appendix
- Consider using video snapshots to highlight critical issues
- Graphics help reader understand the data
- Be responsible and honest when reporting your data!

Key points

- Data analysis requires a lot of expertise, but there are many tools to help
- Basic measures like descriptive statistics can help you gain insight into what your data is telling you
- Statistical methods like ANOVA and Chi-square test can give deeper insight into your data
- Bayes Factor Analysis, once a computationally intense analysis method, is not usable for model comparison