Empirical Research Methods in Information Science

IS 4800 / CS 6350

Lecture 5

Research Models

1

Ethnography Homework Status?

- Goal: idea to make the library more efficient and friendly
  - Pick a location and spend an hour people watching with a notebook and pencil.
  - Identify an activity you find interesting.
  - Watch several people do it.
  - Interview one or two about it.

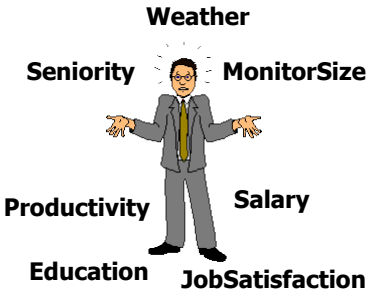
3

Research Models

4

Quantitative, Empirical Research

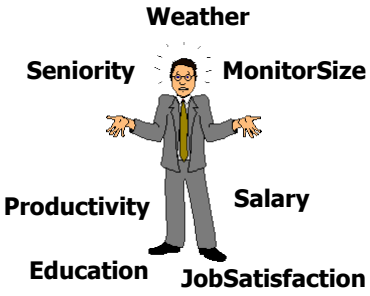
- Basic assumptions
  - The world can be decomposed into variables.
    - They can be observed and measured.
    - They can have numeric or categorical values.
    - Until proven otherwise, they are assumed to vary randomly.



5

Variables

- If we assume that all variables are independent…
- What kind of study can we do?



Descriptive

aka "Exploratory Data Collection"

6

Observation vs. Intervention

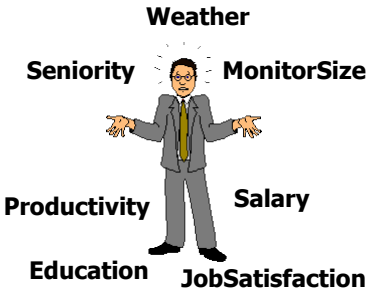
- Observation = passive recording and measurement
- Intervention (aka "manipulation") = actively changing the world to see what happens

7

Intervening on the world…

- If you can manipulate some part of the world…
- And measure variables afterwards…
- What kind of study can you do?

Demonstration

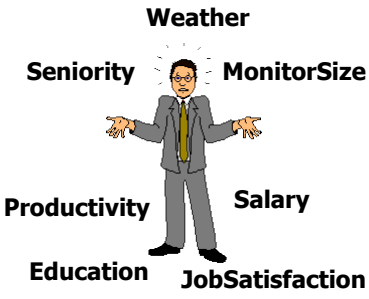


8

Associations between Variables

- You notice that some pairs of variables seem to change together in systematic ways…
- And, you're just observing…
- What kind of study can you do?

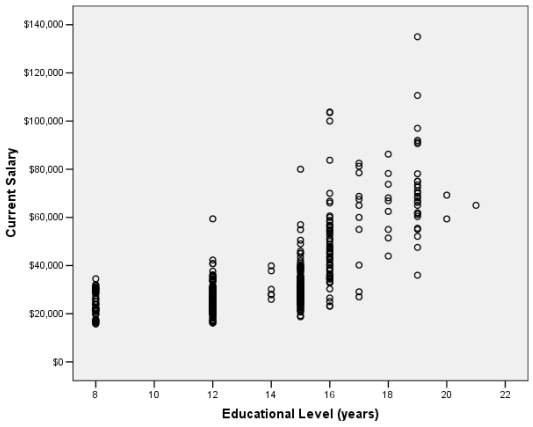
Correlational



9

Correlational study

- How to characterize the association between two variables?
- e.g. Salary & Education Level?

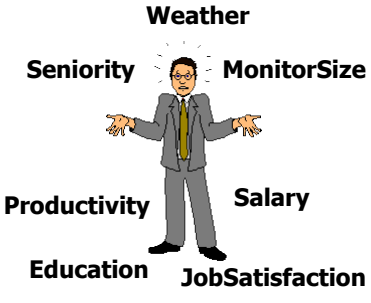


Causality between Variables

- After observing two variables covarying, you hypothesize that there is a causal link between them…

MonitorSize ⟶	Productivity
14″	15 wpd*
17″	20 wpd
21″	21 wpd

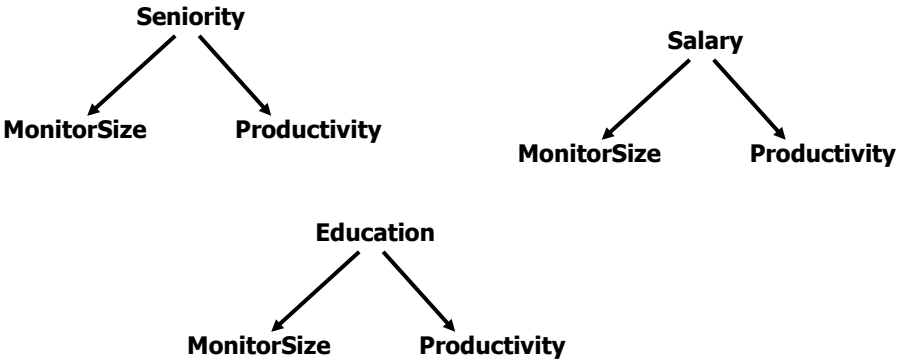
*widgets/day



11

Fundamental difference: association vs. causality

- Other explanations?



12

Experiments

- Isolate IV as the ONLY difference between treatment groups
  - Rules out possible "3rd variables"
- How?
  - Hold extraneous variables constant, OR
  - Randomize subjects between treatments

13

Sample Research Model: Descriptive Study

- aka "exploratory data collection"
- Example:
  - Characterize salaries, job satisfaction, and education level for the company

Salary

JobSatisfaction

Education

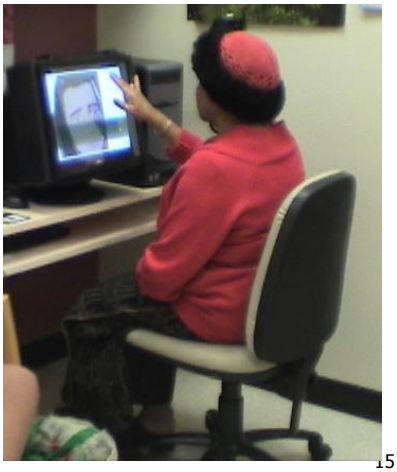
Or, more conveniently …

Salary
JobSatisfaction
Education

14

Example Descriptive Study

- For the "ECAs to Promote Health Literacy…" – what's an example of a descriptive study?



15

Sample Research Model: Demonstration

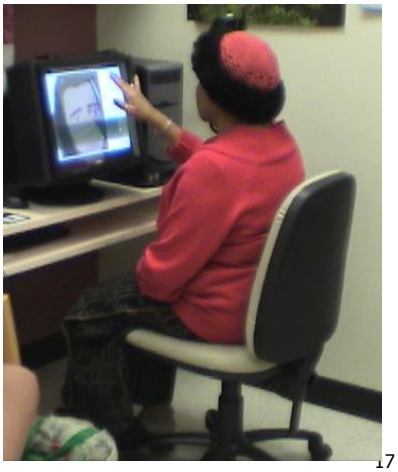
- Example:
  - Characterize JobSatisfaction and Productivity after introducing 36" monitors for all engineers

MonitorSize
JobSatisfaction
Productivity

16

Demonstrations

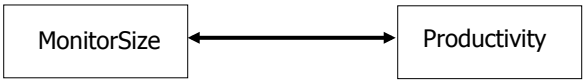
- For the "ECAs to Promote Health Literacy…" – what's an example of a demonstration?



17

Sample Research Model: Correlational Study

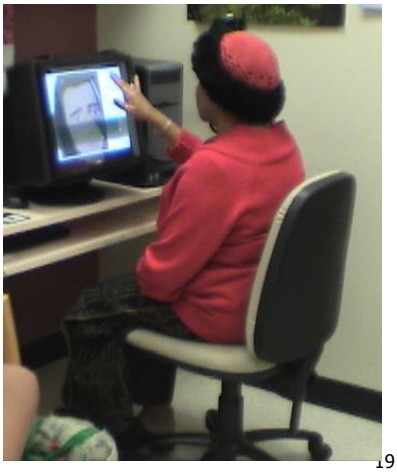
- Example:
  - Characterize the relationship between Productivity and monitors size for all engineers.



18

Correlational Studies

- For the "ECAs to Promote Health Literacy…" – what's an example of a correlational study?



19

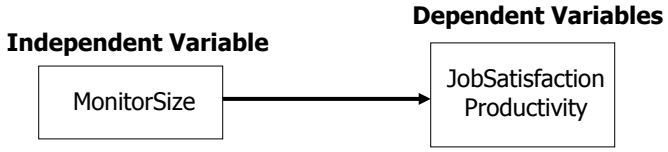
Experimental Studies

- Defining characteristics
  - Manipulation of a variable ("independent variable")
  - Comparison between two or more conditions
  - Control of extraneous variables
- Measured variable is "dependent" or "outcome" variable
- Values of IV = "treatments" or "conditions" or "arms"

20

Sample Research Model: Experimental Study

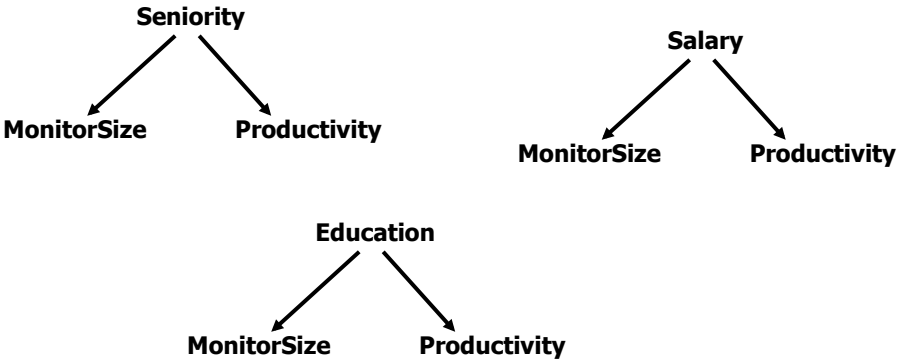
- Determine the effect of increasing MonitorSize on JobSatisfaction and Productivity.



21

Sample Research Model: Experimental Study

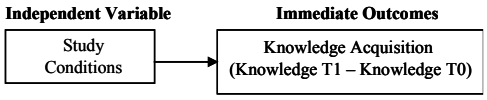
- How to eliminate these alternatives?



22

Sample Research Model: Experimental Study

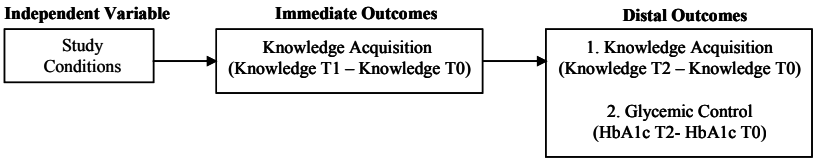
- For the "ECAs to Promote Health Literacy…" – what's an example of an experimental study?



23

Mediating Variables

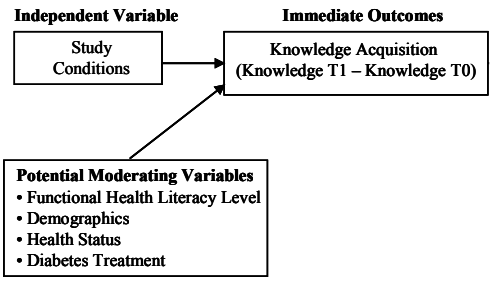
- Affect some variables and are affected by other variables.
- Change during the course of intervention, are correlated with the intervention, and have an effect on outcomes.



25

Moderating Variables

- Baseline variables that are uncorrelated with intervention and define subgroups that may respond differentially to intervention.
- Modifies the relationship between IV & DV.



26

Example: what kind of study?

- Researchers found that average temperatures were much higher in years when more people wore t-shirts, shorts, bikinis and other garments that expose lots of skin to direct sunlight.
- "These data show the important linkage between world climate change and the deterioration of the moral fiber of our society," said Melvin Ebbles, professor of sociometeorology at the University of Alberta and leader of the research team.
- "One of the first steps in arresting global warming must be to rein in the growing tendency towards public exhibitionism of the human body."
- Professor Ebbles suggested that risqué garments allow the skin to release more carbon dioxide and other greenhouse gases into the atmoshpere. Greenhouse gases prevent solar energy from escaping into outer space, effectively transforming the earth into an enormous pressure cooker.
- Professor Ebbles suggested that the United States should take steps to encourage its citizens to dress more modestly. "After all," he said, "hardly anyone wears bikinis up here in Canada and its always very cold."

27

Example: what kind of study?

- Viagra, the anti-impotence drug that has improved the sex lives of men, also works for women, Italian sex researchers reported.
- "Our results demonstrate that Viagra may directly improve female arousal disorder and thus other sexual qualitative functions such as enjoyment and orgasm," Professor Salvatore Caruso said in a report in the British Journal of Obstetrics and Gynaecology.
- In the first study on the use of Viagra on women with sexual arousal disorder, Caruso and his colleagues tested the pills on 51 women ages 22 to 38.
- The women were randomly selected to receive Viagra in 25-mg. or 50-mg. doses or a placebo over three four-week periods, with a week's interruption between each. Each month, the women rated their arousal, orgasms, enjoyment and sexual fantasies on a five-point scale. The scientists said arousal scores of the women taking Viagra rose from 1.5 to 4.2 on both drug doses. But the placebo group achieved only a 2.6 grade. Enjoyment scores and sexual fantasies all rose in the Viagra group.

28

Example: what kind of study?

- Subjects, believing that they were going to complete a personality inventory, were seated in a chemistry classroom facing a large cabinet.
- There were three signs posted on the cabinet saying: "DANGER," "KEEP OUT," and "Attention: Cabinet Contains Hazardous Chemicals Intended ONLY for Animal Research. Possible Harm to Humans if Exposed!!! DO NOT OPEN." Inside the cabinet was a sealed brown box.
- After one minute, an "authority figure" entered the room. The authority figure was dressed in a police uniform.
- The authority figure said to the subject: "I am late for a meeting with your dean. I want you to get in that cabinet and take the box to the president's office immediately."
- Results showed that 13 of 17 subjects obeyed the authority figure, despite the signs posted on the cabinet.

29

Example: What kind of study?

- Raising the price of beer leads to a reduction in cases of gonorrhoea, researchers have said.
- The research from the US suggests that raising the price of a six-pack of beer by 20 cents would cut gonorrhoea rates by almost 9%.
- Researchers at the Centers for Disease Control and Prevention looked at gonorrhoea rates between 1981 and 1995 among teenagers and young adults in US states that raised the legal drinking age or increased the state beer tax.
- Dr Kathleen Irwin, at the centre's Division of Sexually Transmitted Diseases Prevention, said: "Of the 36 beer tax increases that we reviewed, gonorrhoea rates declined among teens aged 15 to 19 in 24 instances.

30

Example: what kind of study?

- A Stanford study suggests that, for grade-school children, watching less television may be a key to limiting weight gain. Children who were involved in a one-year curriculum to reduce their TV viewing gained significantly less body fat than a control group of their peers.
- Local education officials picked two schools with similar ethnic composition, socioeconomic standing and scholastic achievement.
- At one of the schools, the third- and fourth-graders received an 18-lesson program, presented by their classroom teachers as part of the normal school curriculum, that was designed to reduce TV and videotape watching and video game playing.
- Both schools agreed to participate before learning which school would receive the curriculum, and the students at each school were found to have similar TV viewing habits and body fatness at the beginning of the school year, Robinson said.

31

Example

- Want to study efficiency of WizziWord software in BigBucks, Inc.
  - Design a descriptive study
  - Design a demonstration study
  - Design a correlational study
  - Design an experimental study

33

Example

- Want to determine cleanliness of houses cleaned with Roomba.

- Design a descriptive study
- Design a demonstration study
- Design a correlational study
- Design an experimental study

34

Correlational Research: Major Features

- No independent variables are manipulated
- Two or more dependent variables are measured, and a relationship is established
- Correlational relationships can be used for predictive purposes
  - A PREDICTOR VARIABLE can be used to predict the value of a CRITERION VARIABLE
- Correlational research cannot be used to establish causal relationships among variables
  - THIRD VARIABLE PROBLEM
  - DIRECTIONALITY PROBLEM

35

Correlational Research: When Is It Used?

- When gathering data in the early stages of research
- When manipulating an independent variable is impossible or unethical
- When you are relating two or more naturally occurring variables
- You don't have subjects or other resources to run an experimental study
- Retrospective data analysis

36

Experimental Research: Major Features

- An independent variable is manipulated (with at least two levels)
- A dependent variable is measured
- *The most basic experiment consists of an experimental and a control group*
- Control is exercised over extraneous variables either by holding them constant or by randomizing their effects across treatments
- A causal relationship between the independent and dependent variables can be established

37

Strength and Limitations of Experimental Research

- Strength
  - Identification of causal relationships among variables
    - Not possible with correlational research
- Limitations
  - Can't use experimental method if you cannot manipulate variables
  - Tight control over extraneous variables limits generality of results
    - Tradeoff exists between tight control and generality
  - Most expensive & difficult to do.

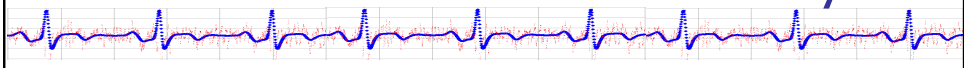
38

Types of study design

	Number of Variables	Number of IV Levels	Manipulation
Descriptive	1	NA	NA
Demonstration	≥ 2	1	√
Correlational	≥ 2	NA	NA
Experimental	≥ 2	≥ 2	√

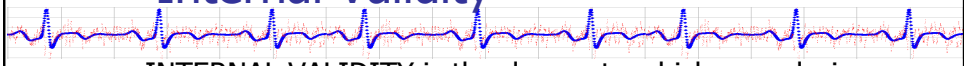
39

Internal vs. External Validity?



40

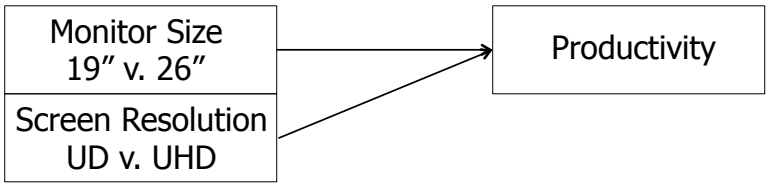
Internal Validity



- INTERNAL VALIDITY is the degree to which your design tests what it was intended to test
  - In an experiment, internal validity means showing that variation in the dependent variable is caused *only* by variation in the independent variable
- Internal validity is threatened by CONFOUNDING and EXTRANEOUS VARIABLES
- Internal validity must be considered during the design phase of research

41

Example Confound



42

External Validity

- EXTERNAL VALIDITY is the degree to which results generalize beyond your sample and research setting
- External validity is threatened by the use of a highly controlled laboratory setting, restricted populations, pretests, demand characteristics, experimenter bias, and subject selection bias
- Steps taken to increase internal validity may decrease external validity and vice versa
- Internal validity may be more important in basic research; external validity, in applied research

44

Example:

- You want to evaluate a new sensor to detect whether people are happy or not.
- You hire actors and randomly assign them to act happy or sad, and test your sensors on them.
- What kind of validity (internal/external) might be challenged?

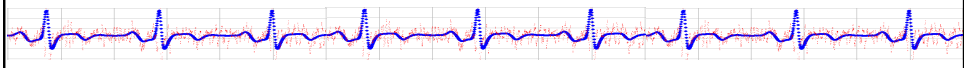
45

Example:

- You conduct the "Conversational Agents to Promote Health Literacy" study by assigning the first 30 patients who volunteer to the intervention group, and the next 30 to the control group.
- What kind of validity (internal/external) might be challenged?

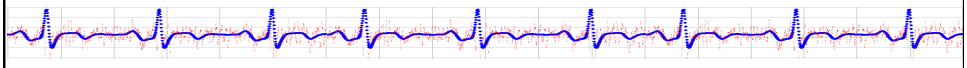
46

Extraneous vs. Confounding Variables?



48

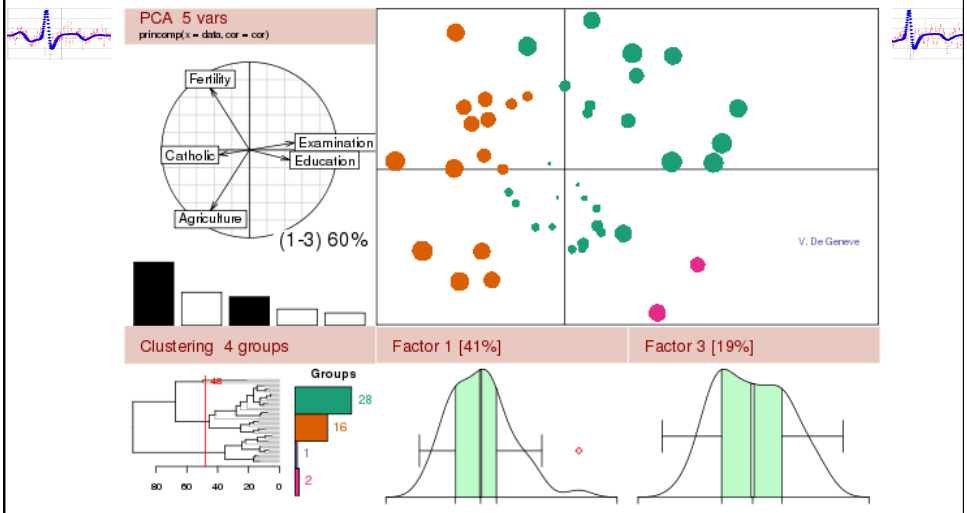
Research Settings



- The laboratory setting
  - Affords greatest control over extraneous variables
  - Simulations
    - Attempt to recreate the real world in the laboratory
    - Realism is an issue
- The field setting
  - Study conducted in a real world environment
    - Field experiment: Manipulate variables in the field
    - High degree of external validity, but internal validity may be low

49

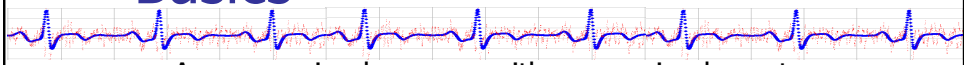
The R Project for Statistical Computing



R Introduction

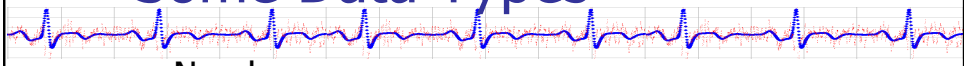
- An integrated suite of software facilities for data manipulation, calculation and graphical display.
- Features:
  - data handling and storage facility,
  - suite of operators for calculations on arrays, in particular matrices,
  - a large, coherent, integrated collection of intermediate tools for data analysis,
  - graphical facilities for data analysis and display either directly at the computer or on hardcopy,
  - a programming language (called 'S')
  - Extensible, with many contributed *packages*
- An environment within which many statistical techniques have been implemented

Basics



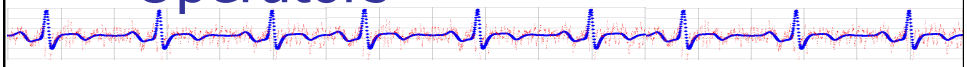
- An expression language with a very simple syntax.
- Elementary commands are either expressions or assignments.
- Commands
  - Separated by ';' or newline.
  - Grouped with '{' '}'
- Comments : '#' to end of line
- Case sensitive.
- Names
  - All alphanumeric symbols are allowed, plus '.' and '_'
  - must start with '.' or a letter, and if it starts with '.' the second character must not be a digit.

Some Data Types



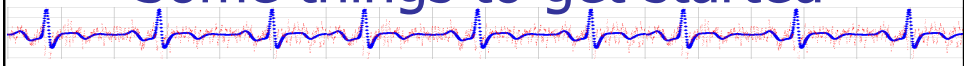
- Numbers
- Boolean (`TRUE, FALSE`)
- `NA, NaN, Inf`
- Strings
  - `"foo", 'bar', "this is a \" quote"`
- Vectors (all elements of same type, ordered)
- Matrices, Arrays, Factors, Lists
- Data Frames
- Objects
- Functions

Some Basic Numeric Operators



- `+, -, *, /, ^`
- `log, exp, sin, cos, tan, sqrt`
- `max, min, length, sum, prod`
- `range` -> vector of (min,max)
- `mean` – sample mean
- `var` – sample variance
- `order` – sort in increasing order

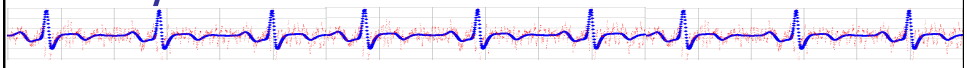
Some things to get started



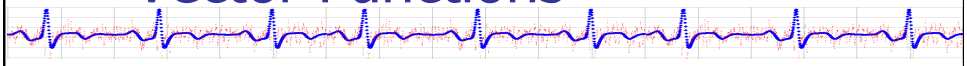
- c – concatenate – makes a vector
  - `c(2,7,3,9,4)`
- <- assignment
  - `x <- 52`
  - `x <- c(2,7,3,9,4)`
  - `y <- c("A","B","C","D")`
- `barplot(x)`
- `barplot(x,names.arg=y)`

55

R/R Studio - Demo



Vector Functions



- Creation; ‘c’ function (‘concatenate’):

```
x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

  - ‘c’ zero or more values and concatenates them.
  - A number by itself is a vector of length 1.

- Vector arithmetic

```
x/2+1
6.20 3.80 2.55 4.20 11.85
```

Vector Functions

- Selecting elements
  - V[indices] - According to index (1..length)
    - E.g., x[x>0] -- all elements greater than zero
    - x[2] -- the second element
    - x[1:10] -- the first 10 elements
  - V[neg-indices] - Elements to drop
  - By name (must assign names attribute)

```
fruit <- c(5, 10, 1, 20)
names(fruit) <- c("orange", "banana", "apple", "peach")
lunch <- fruit[c("apple","orange")]
```

  - Assigning particular vector elements

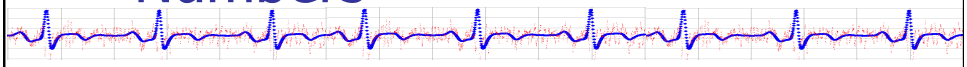
```
x[x<0] <- 0
```

--Change all negative values to zero.

Vector Functions

- Changing length of a vector
  - x[23] <- 12 extends x to 23 eles, using NA if necc
  - length(x) <- 5 truncates x to 5 eles

Creating Sequences of Numbers



- `1:5 ->  1, 2, 3, 4, 5`
- `seq(…)` - generate regular sequences with start, end, increment, and/or length specified.
- `rep(…)` -generate vector of same value repeated.

Factors

- A vector object used to specify a discrete classification (grouping) of the components of other vectors of the same length
  - E.g., the IV values for rows in a stacked frame

```
Conditions <- c("I","C","I","I","C","C")
```

- A factor is created using the factor() function:

```
arms <- factor(conditions)
```

- To find out the levels of a factor the function levels() can be used.

```
levels(arms)
```

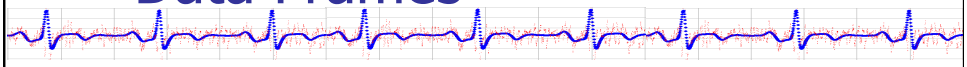
Lists

- An object consisting of an ordered collection of objects known as its components.
- The components can be of varying mode or type

Data Frames

- A special kind of List
  - Components must be vectors (numeric, character, or logical), factors, numeric matrices, lists, or other data frames.
  - Matrices, lists, and data frames provide as many variables to the new data frame as they have columns, elements, or variables, respectively.
  - Numeric vectors, logicals and factors are included as is, and character vectors are coerced to be factors, whose levels are the unique values appearing in the vector.
  - Vector structures appearing as variables of the data frame must all have the same length, and matrix structures must all have the same row size.
- Generally used to represent a dataset in statistical analyses.

Data Frames



- Creating

```
data.frame(home=statef, loot=incomes, shot=incomef)
```

- read.table() to load from a file.
  - Default: string columns are assumed to be factors.
- Index a variable (col)

```
data1$result
```

- Add a variable to a data frame:

```
data1$result <- data1$loot * 2;
```

Read Table
for command line R

- The first line of the file should have a name for each variable in the data frame.
- Each additional line of the file has as its first item a row label and the values for each variable.

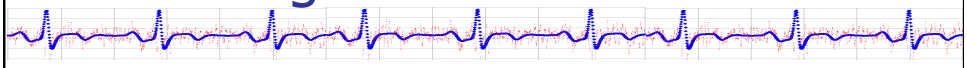
Price	Floor	Area	Rooms	Age	Cent.heat
52.00	111.0	830	5	6.2	no
54.75	128.0	710	5	7.5	no
57.50	101.0	1000	5	4.2	no
57.50	131.0	690	6	8.8	no
59.75	93.0	900	5	1.9	yes

```
data1 <- read.table("myfile.txt", header=TRUE)
```

- For excel CSV files:

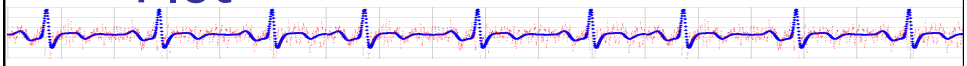
```
data1 <- read.table("myfile.txt", header=TRUE, sep=",")
```

Demo:
Loading Excel Data into R



70

Plot



- plot(x,y,…)
- plot(data,…)

- type="p" for points; "l" for lines; "h" for histograms

Boxplot

- boxplot(var1)
- title("Boxplot of var1")

- boxplot(data$result ~ data$condition)

Histogram

- hist(data)
- hist(wage, col = "yellow")

- table(vector)
  - Returns "contingency table"
  - Matrix:
    - first row = values
    - second row = freq counts

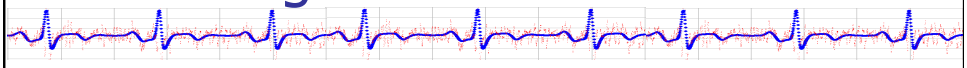
Barchart

- barplot(x)
- *If names attached to values, used as X-labels*
- barplot(table(eyecol)) # a simple histogram
- barplot(table(eyecol),col=c("blue","grey","brown","green"),main="Eye color")
- # a nicer one

Pie Chart

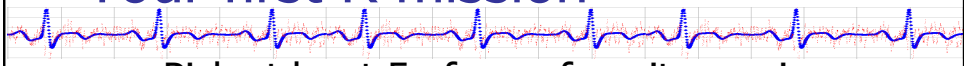
- library(MASS) # load the MASS package ???
- school = painters$School # the painter schools
- school.freq = table(school) # apply the table function
- pie(school.freq) # apply the pie function
- pie.sales <- c(0.12, 0.3, 0.26, 0.16, 0.04, 0.12)
- names(pie.sales) <- c("Blueberry", "Cherry", "Apple", "Boston Cream", "Other", "Vanilla Cream")
- pie(pie.sales) # default colours

Demo:
Plotting in R



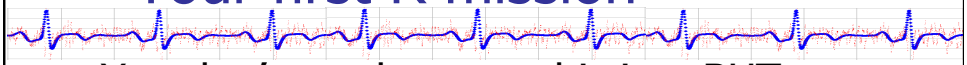
80

Your first R mission



1. Pick at least 5 of your favorite music albums. Create an Excel spreadsheet with the following columns: Album, Tracks, Cost, PctCost. Fill in the data.
2. Save a copy to csv format and import to R.
3. Create a bar chart showing the number of tracks per album.
4. Create a pie chart showing PctCost labeled by Album.
5. Paste the charts into a Word doc.

Your first R mission



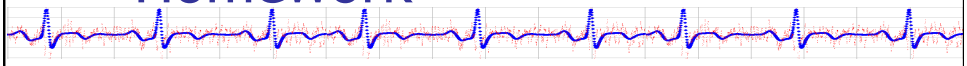
You don't need to turn this in… BUT

by 1/30 you will need to be fluent enough in R to:

- Load data from excel
- Compute a variety of descriptive statistics
- Create a variety of visualizations/plots
- Export all to Word as part of a homework

82

Homework



- Finish Homework 2a (Ethnography) & Do Homework 2b
  - Identify two measurable variables from your ethnographic study that might conceivably be associated.
  - Design separate descriptive, correlational and experimental research designs for studying these variables.
  - For each, include a diagram (such as the one in Figure 6. from the sample research plan, but with just one or two boxes) and text describing the purpose of the study, the measures you plan to use, and what the results would be useful for.
  - Identify a possible "third variable" that might invalidate predictions made with results from the correlational study and how this will be controlled in the experimental study.

83