Week 3: Data types, summary statistics, & transformations

ANTH 674: Research Design & Analysis in Anthropology

Professor Andrew Du
Andrew.Du2@colostate.edu

1

Lecture outline

1. Different types of data ("**qualitative**" vs. **quantitative**)
   1. How are they described & summarized?
   2. How are they plotted (visualizing the distribution)?
2. Different types of data transformations
   1. What are they & what are they used for?
3. Plotting two data types against each other

2

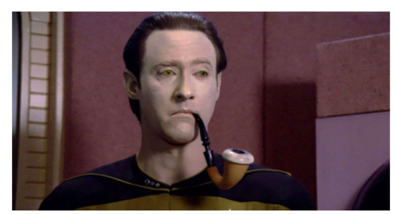
Different data types

How are they described, summarized, and plotted?

3

What is data?

- Wikipedia: **Data** are characteristics or information, usually numerical, that are collected through observation
- Want to learn something from data through analysis and/or visualization (plotting)



4

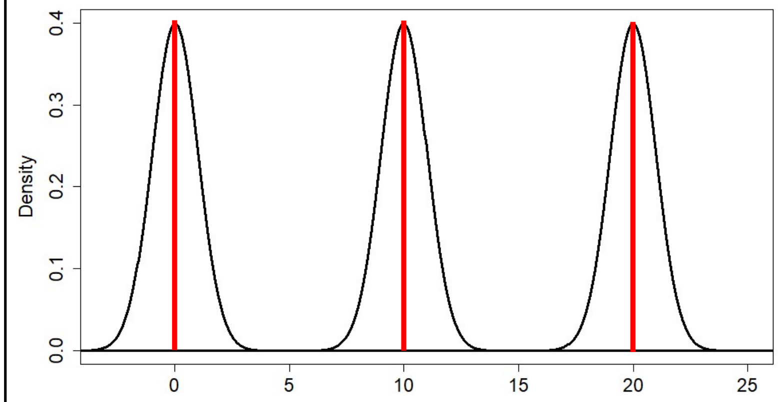
Summary statistics

- Used to summarize distribution of data w/ one number (except for multivariate distributions)

1. Location or central tendency
2. Spread or variation

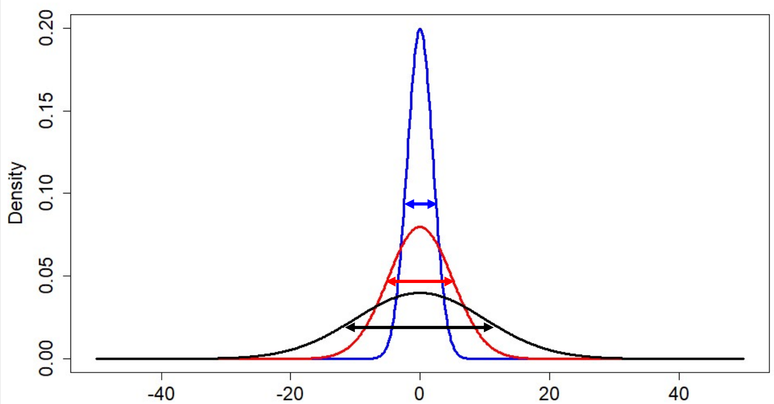
5

Location/central tendency



6

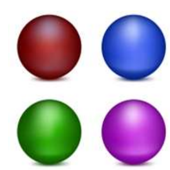
Spread/variation



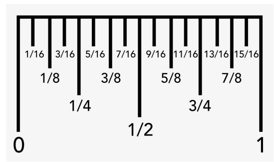
7

Different types of data

<u>"Qualitative"</u>
- Categorical/nominal
- Ordinal



<u>Quantitative</u>
- Discrete
- Continuous



- Data type tells you which summary statistics, plots, and analyses to use!

8

Questions?

9

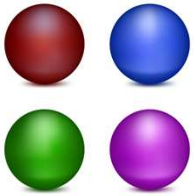
“Qualitative” data

10

What is “qualitative” data?

- Data assigned to groups, usually based on some qualitative property

1. Categorical/nominal data (unordered)
2. Ordinal data (ordered)



11

Categorical/nominal data

- Unordered qualitative data
- E.g., head/tail (binomial); basalt/chert/quartz (multinomial)
- Quantified as counts or proportions
- = factors in R

12

Plotting categorical data

black	brown	red	blond
108	286	71	127

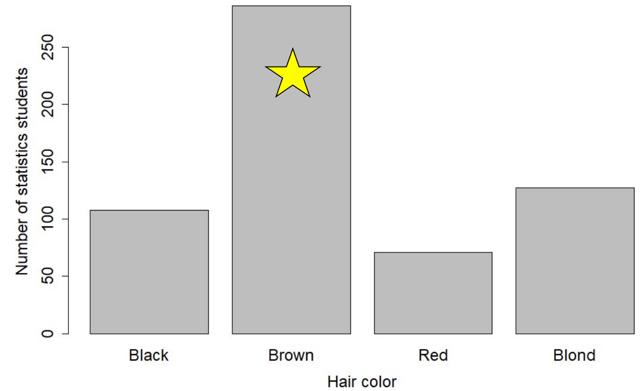
- Bar plot

*Order along x-axis is arbitrary

13

Central tendency

- Mode: most common category



14

Spread

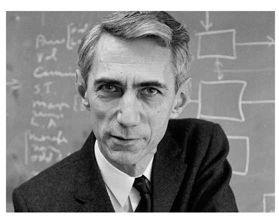
	black	brown	red	blond
#	108	286	71	127
prop.	0.18	0.48	0.12	0.21

- Information theory measures (most common is Shannon's index)
- Not commonly used

$$-\sum p_i \log(p_i)$$

p_i = proportion of i^{th} category

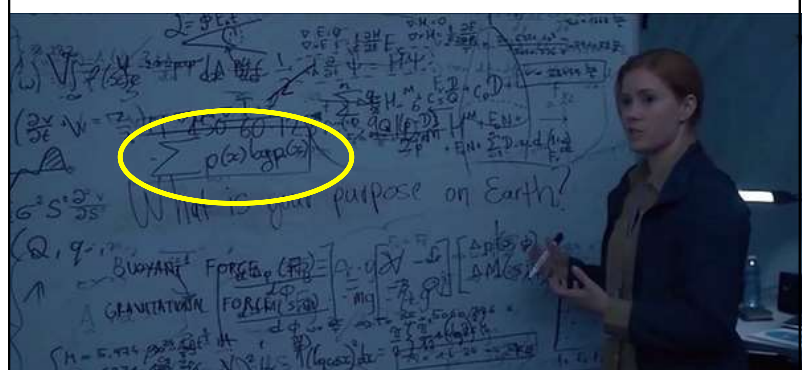
$$-[0.18 \times \log(0.18) + 0.48 \times \log(0.48) + 0.12 \times \log(0.12) + 0.21 \times \log(0.21)] = \mathbf{1.25}$$



Claude Shannon

15

From the movie *Arrival*



16

Ordinal data

- Ordered qualitative data
- Distance between categories not known
- E.g., small/medium/large; juvenile/adult
- Quantified as counts or proportions
- = ordered factor levels in R

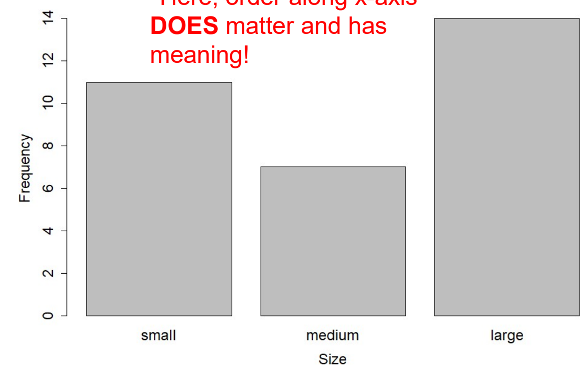
17

Plotting ordinal data

- Bar plot

small	med.	large
11	7	14

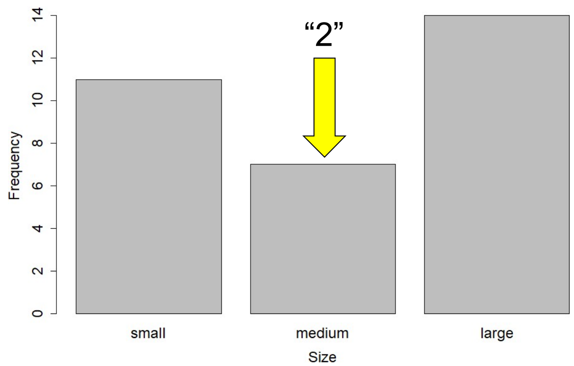
*Here, order along x-axis **DOES** matter and has meaning!



18

Central tendency

- Median: middle value in ordered data (first convert to ranks: 1, 2, 3)



19

Central tendency

20

Spread

small	med.	large
11	7	14

- Interquartile range (IQR): difference between 75th and 25th percentiles (AKA 3rd and 1st quartiles; median is the 2nd quartile)
- Middle 50% of data

After converting sizes to ranks

1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

25% of data

1

$\mathrm{IQR} = 3 - 1 = \mathbf{2}$

3

75% of data

21

Questions?

22

Quantitative data

23

What is quantitative data?

- Each data point is a number, and distances have meaning (e.g., 1 vs. 3)

Discrete

- Finite or countable values (e.g., integers)

Continuous

- Any value within a *continuous* interval (e.g., 2.4575)

- In practice, all continuous numbers are discrete due to limited precision of measurements (e.g., 1.21, 1.22, 1.23)

24

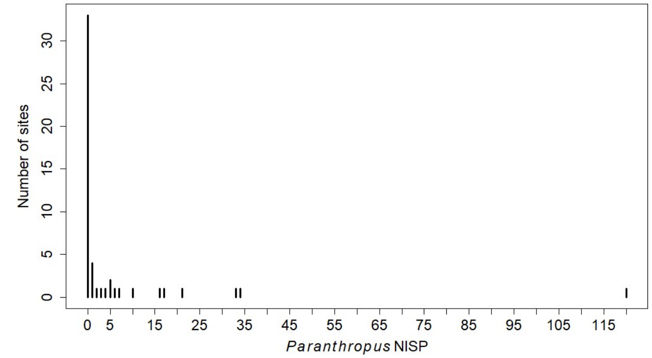
Discrete data

- Finite or countable values
- E.g., count data, anything measured in integers
- Treated as numeric class in R

25

Plotting discrete data

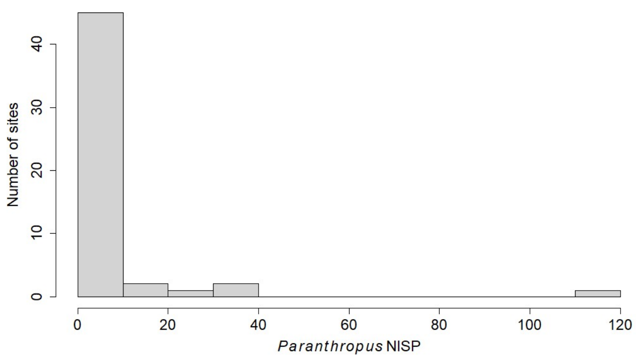
- Line plot (like a bar plot w/ more categories)



26

Plotting discrete data

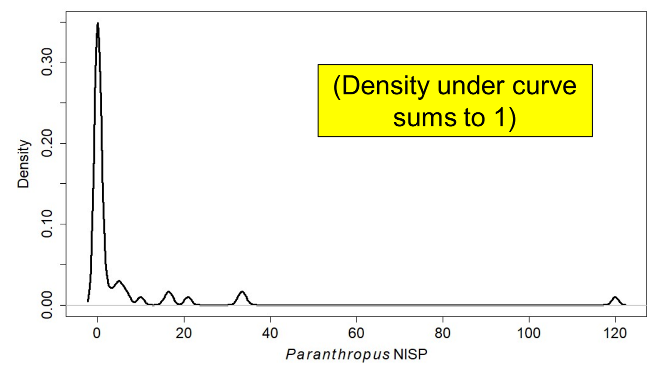
- Histogram



27

Plotting discrete data

- Density plot ("smoothed histogram")



28

Central tendency & spread

- Same as with continuous data

29

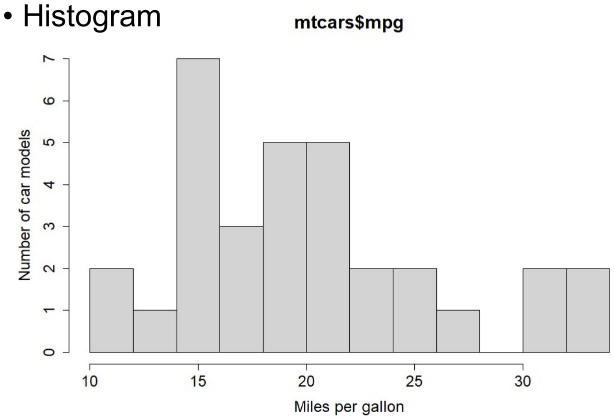
Continuous data

- Any value within a *continuous* interval (e.g., 2.4575)
- E.g., stone tool mass, hominin femur length
- Treated as numeric class in R

30

Plotting continuous data

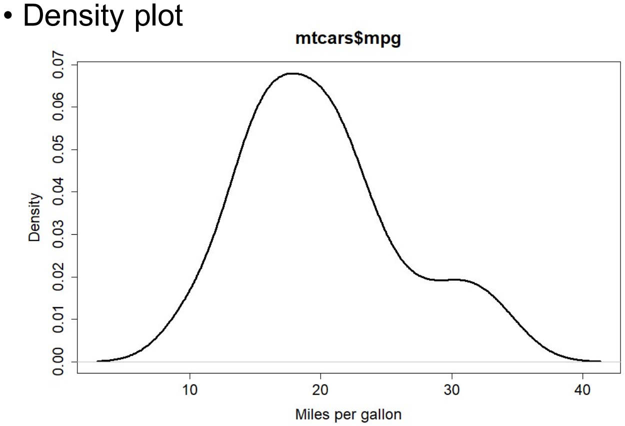
- Histogram



31

Plotting continuous data

- Density plot

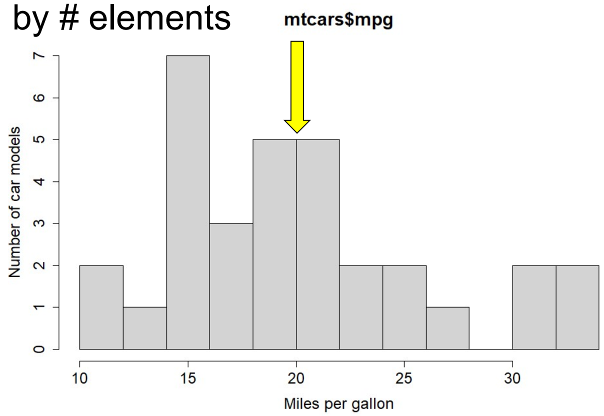


32

Central tendency

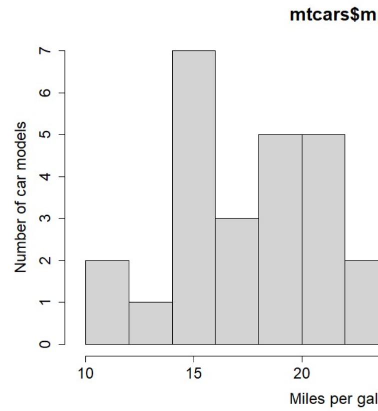
elements $\frac{\sum x_i}{n}$ ith element

- Arithmetic mean: sum elements & divide by # elements



33

Central tendency

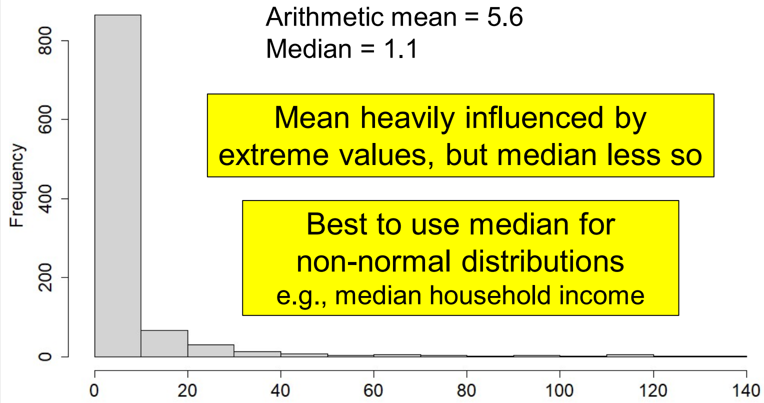


Arithmetic mean = 20.1
Median = 19.2

Arithmetic mean & median identical if distribution is symmetrical

34

What if data are non-normal?



Arithmetic mean = 5.6
Median = 1.1

Mean heavily influenced by extreme values, but median less so

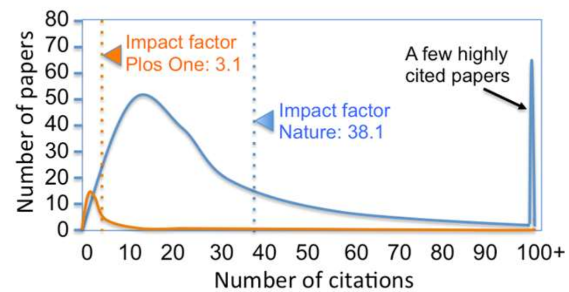
Best to use median for non-normal distributions
e.g., median household income

35

An example

90% of *Nature*'s 2004 impact factor was based on only 25% of its publications

- Journal impact factor: average # of times articles from journal published in the past two years have been cited in year of consideration



36

Central tendency

ith element

$$\left(\prod x_i\right)^{\frac{1}{n}}$$

\# elem.

- Geometric mean: nth root of elements multiplied together

  E.g., `x <- c(1, 2, 3)`

  $GM = \sqrt[3]{(1 \times 2 \times 3)}$
- Same as taking arithmetic mean of log-transformed values & calculating antilog

$$GM = \exp\left(\frac{\sum \log(x_i)}{n}\right)$$

E.g., $GM = \exp\left(\frac{\log(1)+\mathrm{l}\quad(2)+\log(3)}{3}\right)$

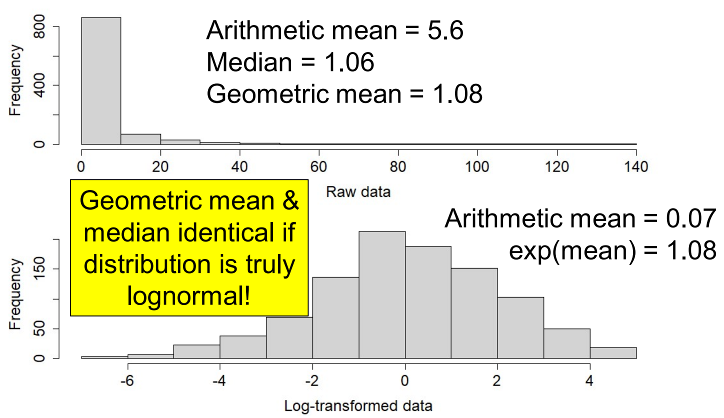
37

Central tendency (geometric mean)

- Used when dealing with data produced by multiplicative processes (e.g., % increases) → lognormal distributions
- E.g., population size, body size, household income, citation numbers
- Can't use when you have zeros or negative numbers

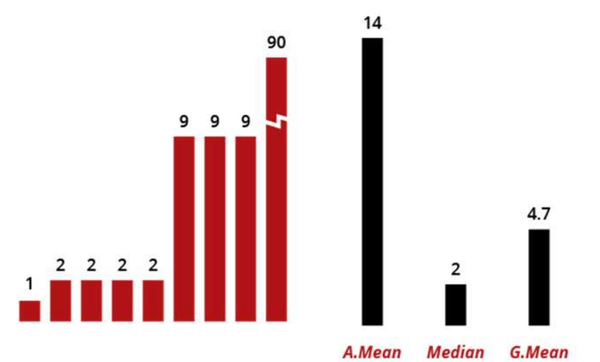
38

Geometric mean (on lognormal distribution)



39

Comparing all three



Order always the same!

40

Questions?

41

Spread

- Variance: measures how far values deviate from the arithmetic mean
- Subtract the mean from each element, square the results, add them up, and divide by number of elements minus 1

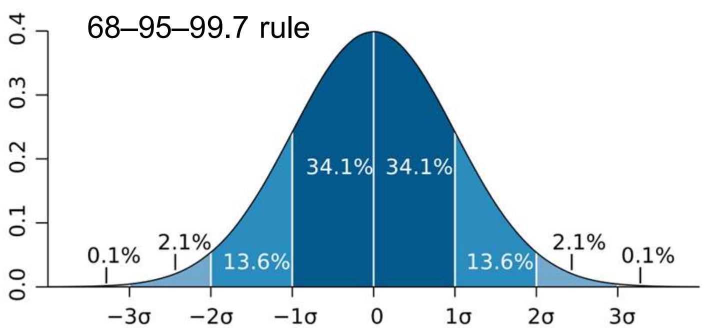
ith element, # elements, Arithmetic mean

$$\frac{\sum(x_i - \bar{x})^2}{n-1}$$

- Square-root of variance = standard deviation

42

Spread (standard deviation)

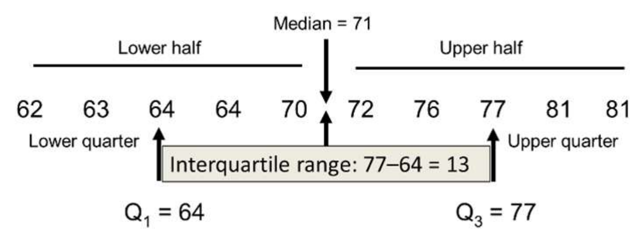


- As with the arithmetic mean, variance and SD are most interpretable for normal distributions

43

What if data are non-normal?

- Interquartile range
- Doesn't depend on arithmetic mean or type of distribution



45

Geometric standard deviation

- Used when you would use the geometric mean (e.g., lognormal data)
- Calculate std. dev. of log-transformed values and take the antilog

ith element

$$\exp\left(\sqrt{\frac{\sum\left[\log\left(\frac{x_i}{\mathrm{GM}}\right)\right]^2}{n}}\right)$$

Geometric mean

elements

46

Summary: which plot to make to visualize distribution?

Data type	**Plot**
Categorical	Bar plot
Ordinal	Bar plot
Discrete	Line plot Histogram Density plot
Continuous	Histogram Density plot

47

Summary: which statistic to use?

Data type	**Location**	**Spread**
Categorical	Mode	Information measures
Ordinal	Median	Interquartile range
Discrete/Continuous		
Normal	Arithmetic mean	- Variance - Standard deviation
Non-normal	- Median - Geometric mean	- Interquartile range - Geometric SD

48

Questions?

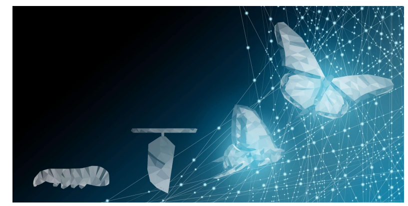
49

Data transformations

What are they & what are they used for?

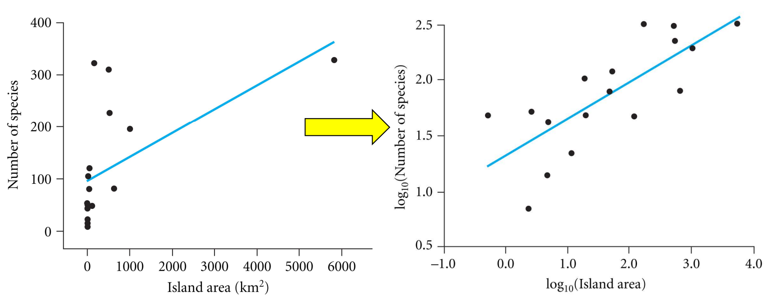
What is data transformation?

- Applying a mathematical function to data to change its distribution
- Rank order of data maintained (monotonic transformation)



Why transform data?

- To make data & results easier to understand and visualize



Why transform data?

- To make data & results easier to understand and visualize
- To make sure assumptions of statistical methods are not violated

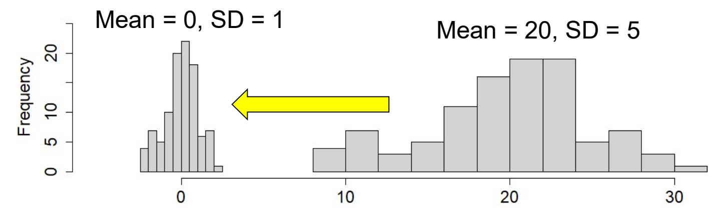
Types of data transformations

1. Centering and scaling
2. Log transformations
3. Square-root transformations
4. Arcsine & logit transformations

54

Centering and scaling

- Transforms data to have mean = 0 & standard deviation = 1 (i.e., Z-scores)
- Subtract the mean & divide by SD



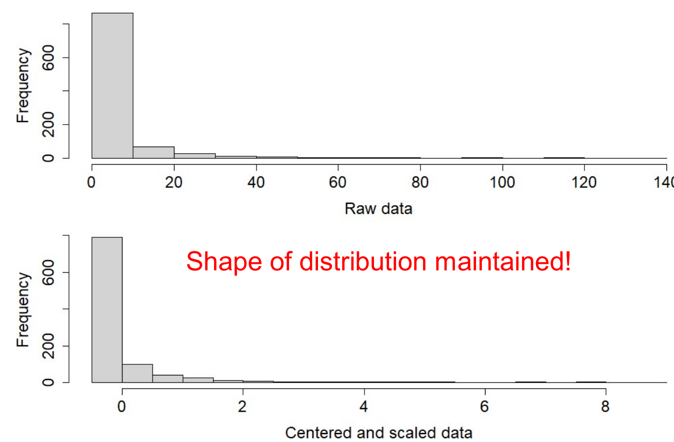
55

Centering and scaling

- Transforms data to have mean = 0 & standard deviation = 1 (i.e., Z-scores)
- Subtract the mean & divide by SD
- Converts data to units of standard deviation, so variables of different units can be compared (e.g., mass & inches)
- Can be used on non-normal data!

56

Centering and scaling



57

Log transformations

- Replaces data with their logarithms
- $b^a = x;\ \log_b x = a$
- Base *e* (analyses) and 10 (plotting) most common
- Cannot transform zeros and negative numbers

58

Log transformations

- Many statistical & plotting methods deal with additive/absolute/linear change
- E.g., linear regression: $y = a + bx$

- 1 → 2: +1
- 100 → 200: +100

Treated differently w/ linear methods

- 1 → 2: x2
- 100 → 200: x2

But multiplicative/ relative change the same!

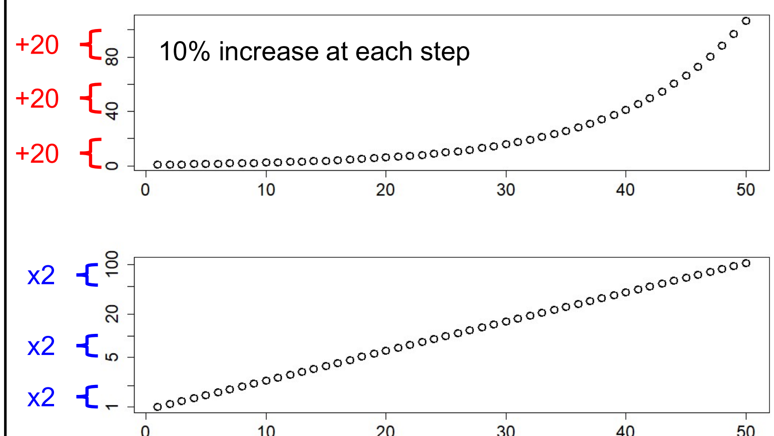
59

Log transformations

- What if you are interested in multiplicative/relative/proportional/percent change?
- Log-transformations: multiplicative → additive
- $\log\left(\frac{2}{1}\right) = \log(2) - \log(1) = 0.69$
- $\log\left(\frac{200}{100}\right) = \log(200) - \log(100) = 0.69$
- Doubling from 1 to 2 now treated the same as doubling from 100 to 200!
- Can now use linear methods to investigate multiplicative change

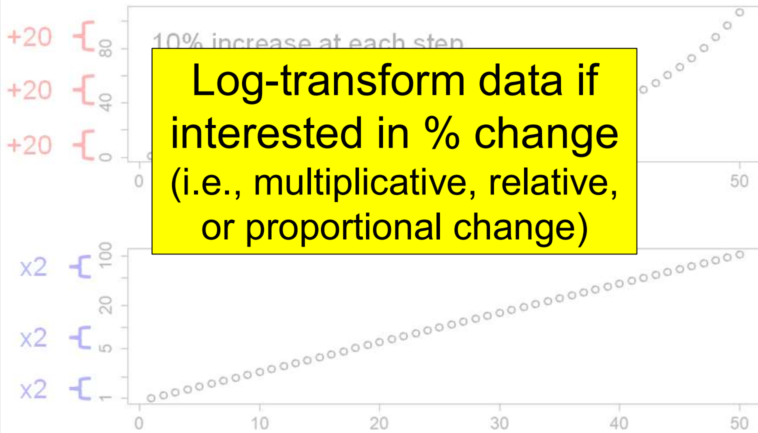
60

Log transformations



61

Log transformations

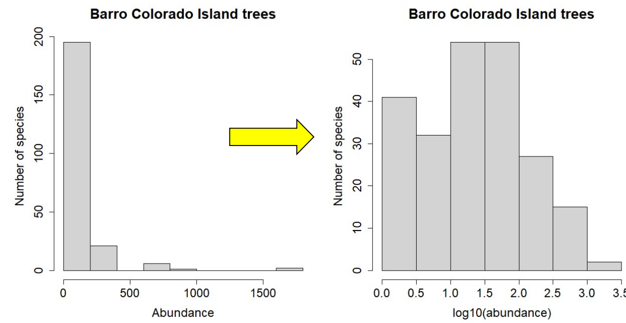


Log-transform data if interested in % change (i.e., multiplicative, relative, or proportional change)

62

Log transformations

- Can aid visualization if data vary over orders of magnitude (spreads out small numbers and squeezes large ones)



63

Other transformations

- Removes dependence between mean & variance of a variable

1. Square-root: $\sqrt{x}$
   - Used for count data
2. Arcsine: $\text{arcsine}(\sqrt{x})$
   - Used for proportions
3. Logit: $\log(\frac{x}{1-x})$
   - Used for proportions

64

Questions?

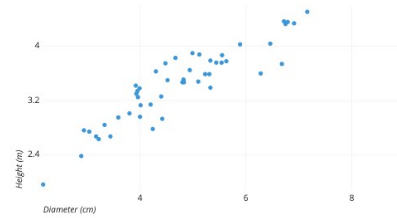
65

Plotting two data types against each other

66

Why do this?

- Want to see how two variables are related to each other
- Good way to visually describe two variables



67

First, a note on ordinal & discrete data when plotting

- Ordinal treated as categorical (maintaining order of categories)
- Discrete treated as continuous
- When ordinal data have many categories, can mimic discrete data (e.g., ranks of all countries by GDP)
- When discrete data are too few, can mimic ordinal data (e.g., 3, 4, & 5 number of forward gears in `mtcars`)

68

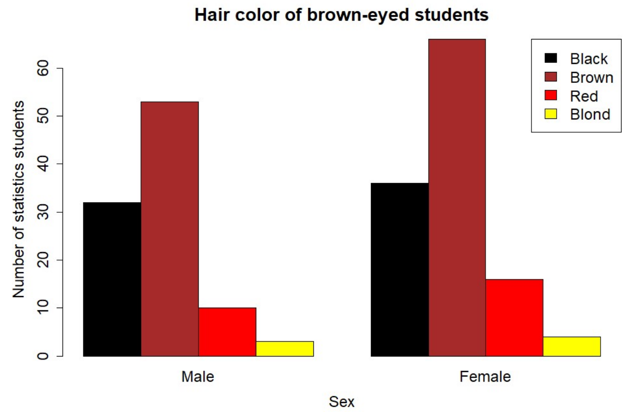
Which plot to make

X-axis

Y-axis	Categorical	Continuous
Categorical	Bar plot	Box plot Violin plot
Continuous	Box plot Violin plot	Scatter plot

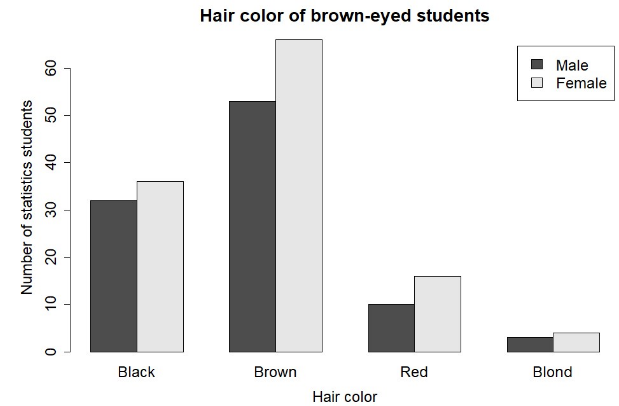
69

Bar plots
(categorical vs. categorical)



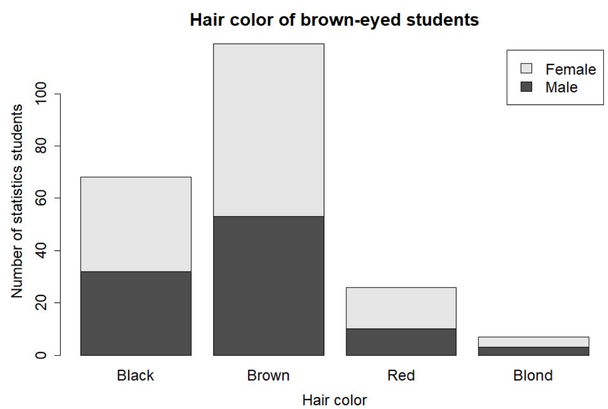
70

Bar plots
(categorical vs. categorical)



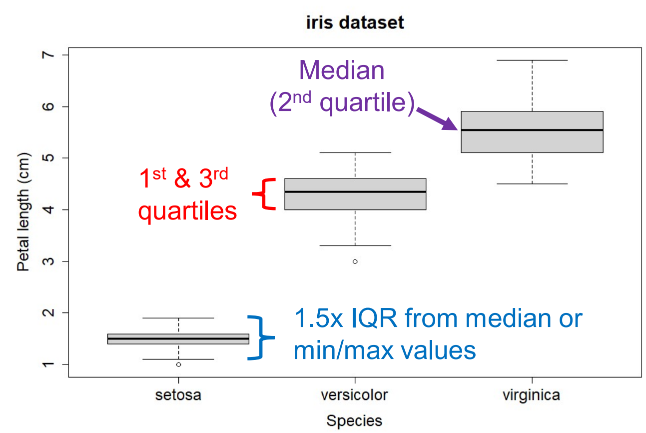
71

Bar plots (stacked)
(categorical vs. categorical)



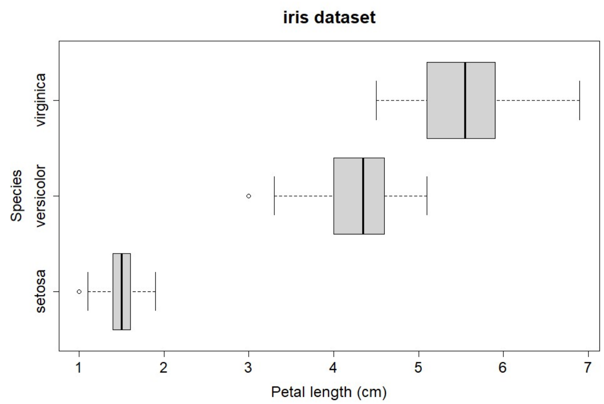
72

Box plots
(categorical vs. continuous)



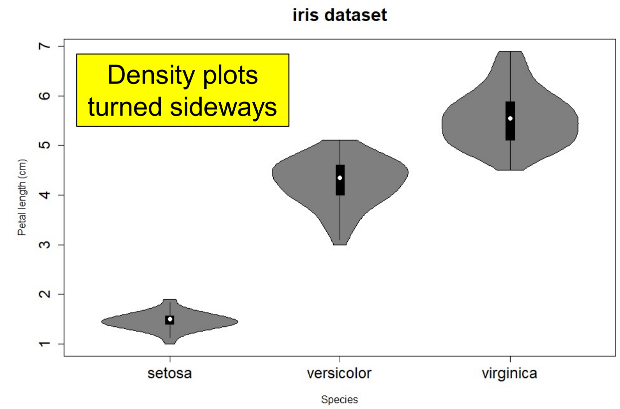
73

Box plots
(continuous vs. categorical)



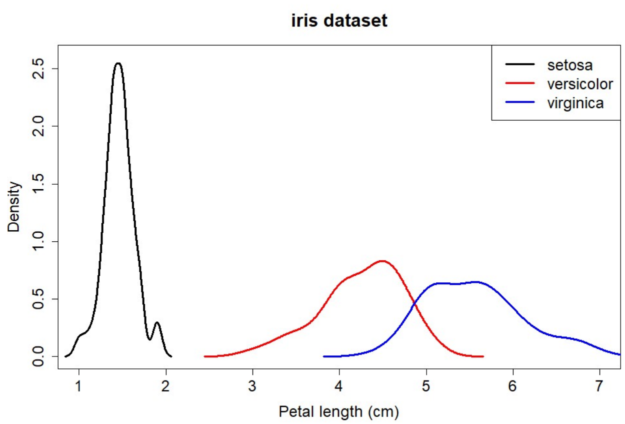
74

Violin plots
(categorical vs. continuous)



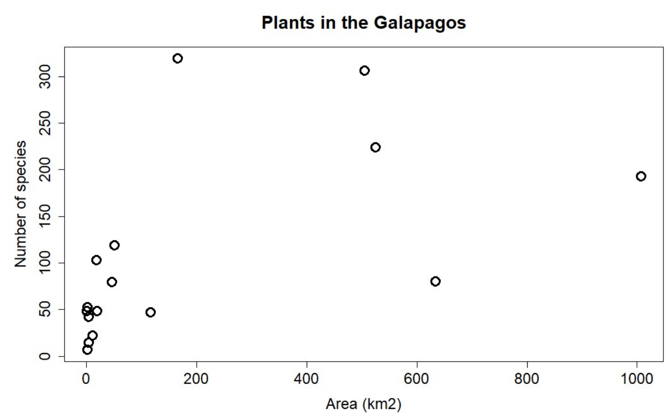
75

Density plot works too!
(categorical vs. continuous)



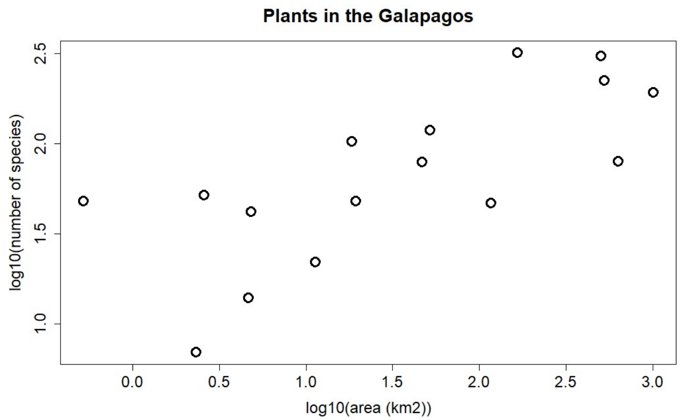
76

Scatter plots
(continuous vs. continuous)



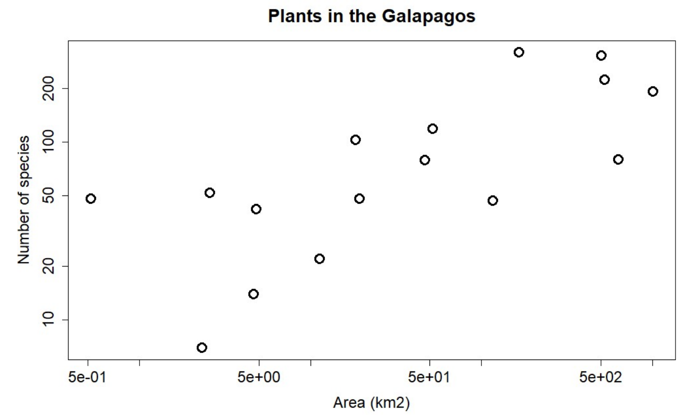
77

Scatter plots (log-transformed variables) (continuous vs. continuous)



78

Scatter plots (log-transformed axes) (continuous vs. continuous)



79

Plotting summary

- These are the general rules of plotting
  (R will actually automatically make these plots according to your variable type)
- BUT, use best judgment for showing what YOU *want* to show (<u>according to your research question</u>)
- Data visualization is very important: want to convey your data and results as clearly & effectively as possible!

80

Questions?

81

Summary

- There are four main data types: categorical, ordinal, discrete, & continuous
- Data type tells you which summary statistics and plots to use
- Data transformations aid visualization and interpretation & help data satisfy statistical assumptions
- How to plot & compare two variables depends on data type

82

Statistics vignette

- What is Euler's number & where does it come from?
- $e \approx 2.71828$
- Used as base of natural logarithm
- Fundamental to continuous growth & rate of change


Leonhard Euler

83

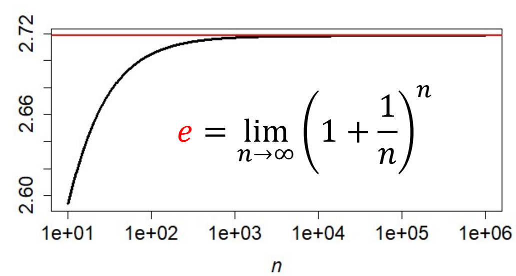
Derivation using compound interest

- You have \$1 in your bank, which offers 100% interest every year
  - After one year, \$1 → \$2
- 50% interest twice a year
  - After one year, \$1 → \$1.50 → \$2.25
- 1/12th interest every month
  - After one year, \$1 → \$1.08 → … → \$2.61
- <u>General formula</u>: $N_0\left(1+\frac{1}{n}\right)^n$

84

Derivation using compound interest

- <u>Every day</u>: $1\left(1+\frac{1}{365}\right)^{365} = 2.715$
- <u>Every hour</u>: $1\left(1+\frac{1}{8760}\right)^{8760} = 2.718$



85

A probability interpretation

- Probability every dropped chocolate is placed in wrong position = $\mathbf{1/e}$ (as # chocolates → ∞)

86

Widely considered the most beautiful equation in math

- Euler's identity

$$e^{i\pi} + 1 = 0$$

87