Frequency distributions, central tendency & variability

Displaying data

Software

SPSS

Excel/Numbers/Google sheets

Social Science Statistics website (socscistatistics.com),
http://www.alcula.com/calculators/statistics/dispersion/

Labor site:
https://sites.google.com/site/markojaadam/home/statisztika-gyakorlat-2017

Creating an SPSS file

- Open the Google Sheet data file
- Tidy up the data
- Open SPSS – type in new data
- Copy and paste the data into the SPSS spreadsheet
- Go to Variable View and select the correct measurement scale
- Recode string variables into numerical variables: Transform -> Recode into same variables
  - Add labels: choose Variable View and in the “Values” column add the labels

Displaying data

- Histogram (frequency polygon): shows the frequency (or probability) of each value of the dependent variable
- Classic bar chart: shows the average value of the dependent variable at each level of the independent variable
- Scatterplot: shows the relationship between two dependent variables

Frequency Distributions

- For Nominal or Ordinal data:
  - Eye colour
    Brown, brown, hazel, blue, blue, grey, blue, brown, hazel, grey, green, blue, brown
  - World cup:
    How may times each country finished 1st or 2nd place

- For scale (ratio or interval) data:
  - Height (cm)
    176, 178, 184, 172, 180, 178, 165, 160, 172, 171, 176, 166, 176, 183, 165, 158, 165, 173, 162, 185

Frequency Table: Steps

- Find Max and Min. scores
- Determine Range (Max-Min+1)
- Determine number of bins
  - More art than science, judgment call
- Decide bottom number
- From Highest to lowest, count *number of scores* that belongs in each bin.

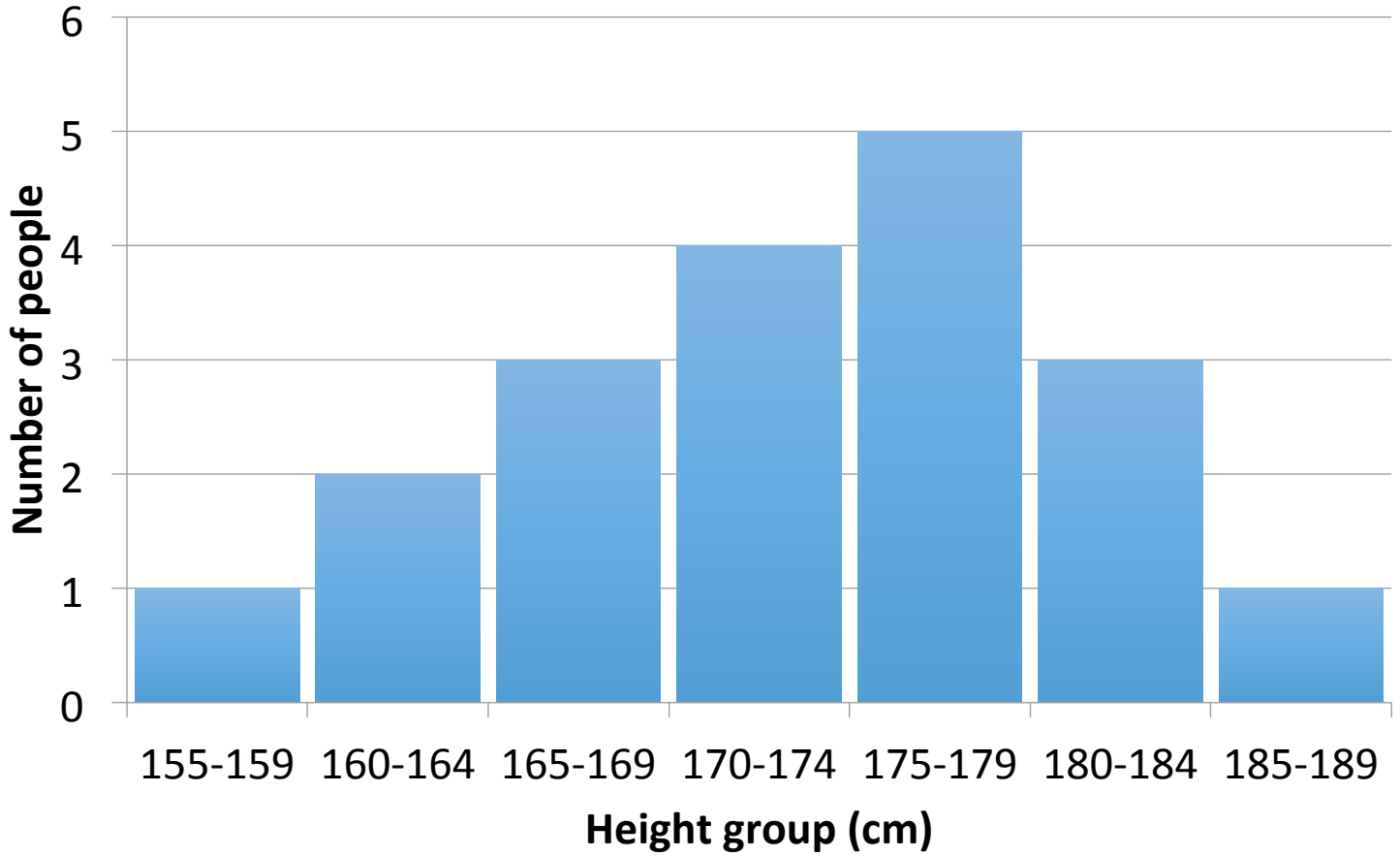
Frequency table

Height	Frequency	Percent	Cumulative percent
155-159	1	5	5
160-164	2	11	16
165-169	3	16	32
170-174	4	21	53
175-179	5	26	79
180-184	3	16	95
185-189	1	5	100

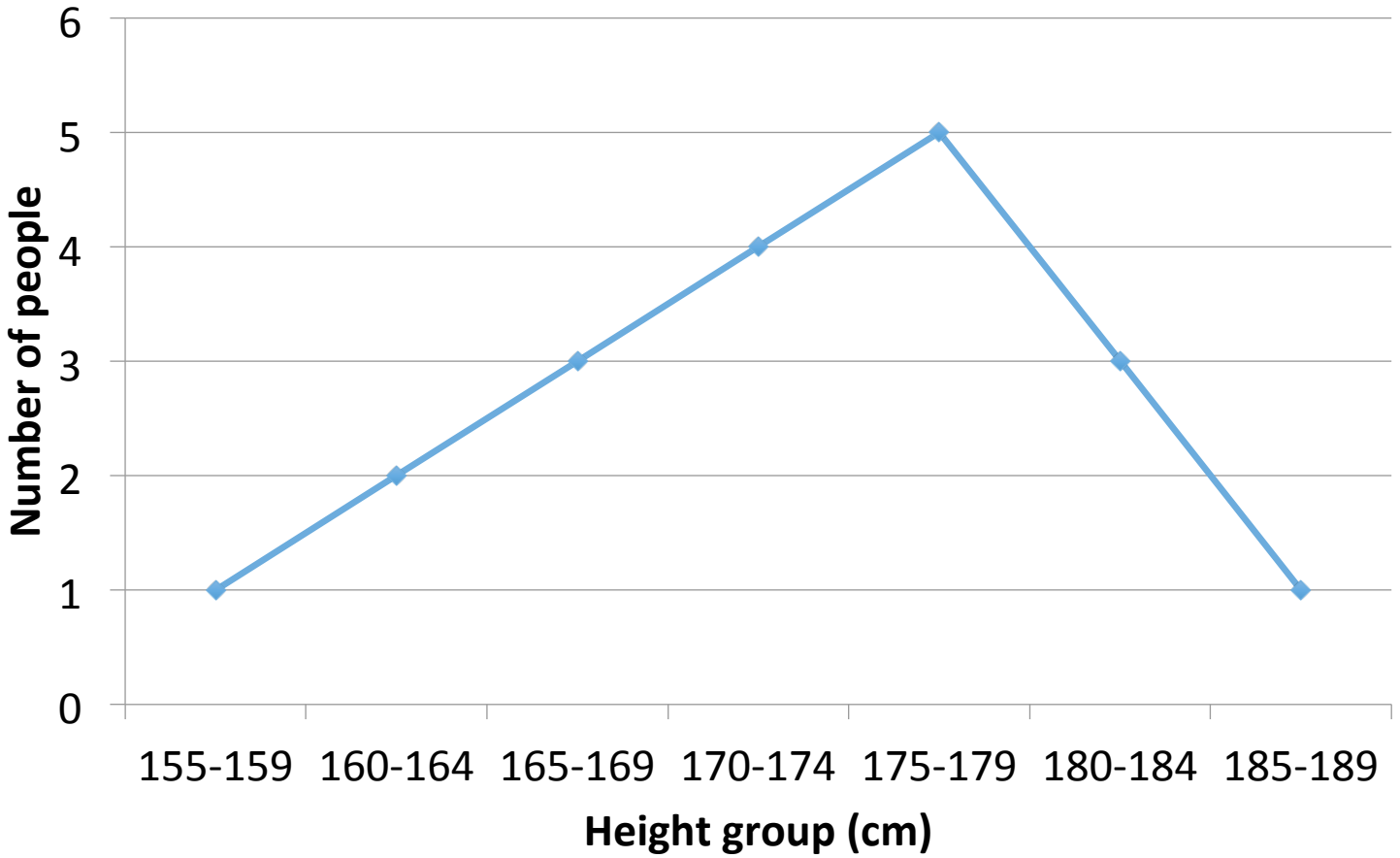
Histogram: Steps

- Define X-axis
- Define the range of X-axis variable.
- Define range of frequency on the Y-axis .
- CHOOSE the bin size (wisely).

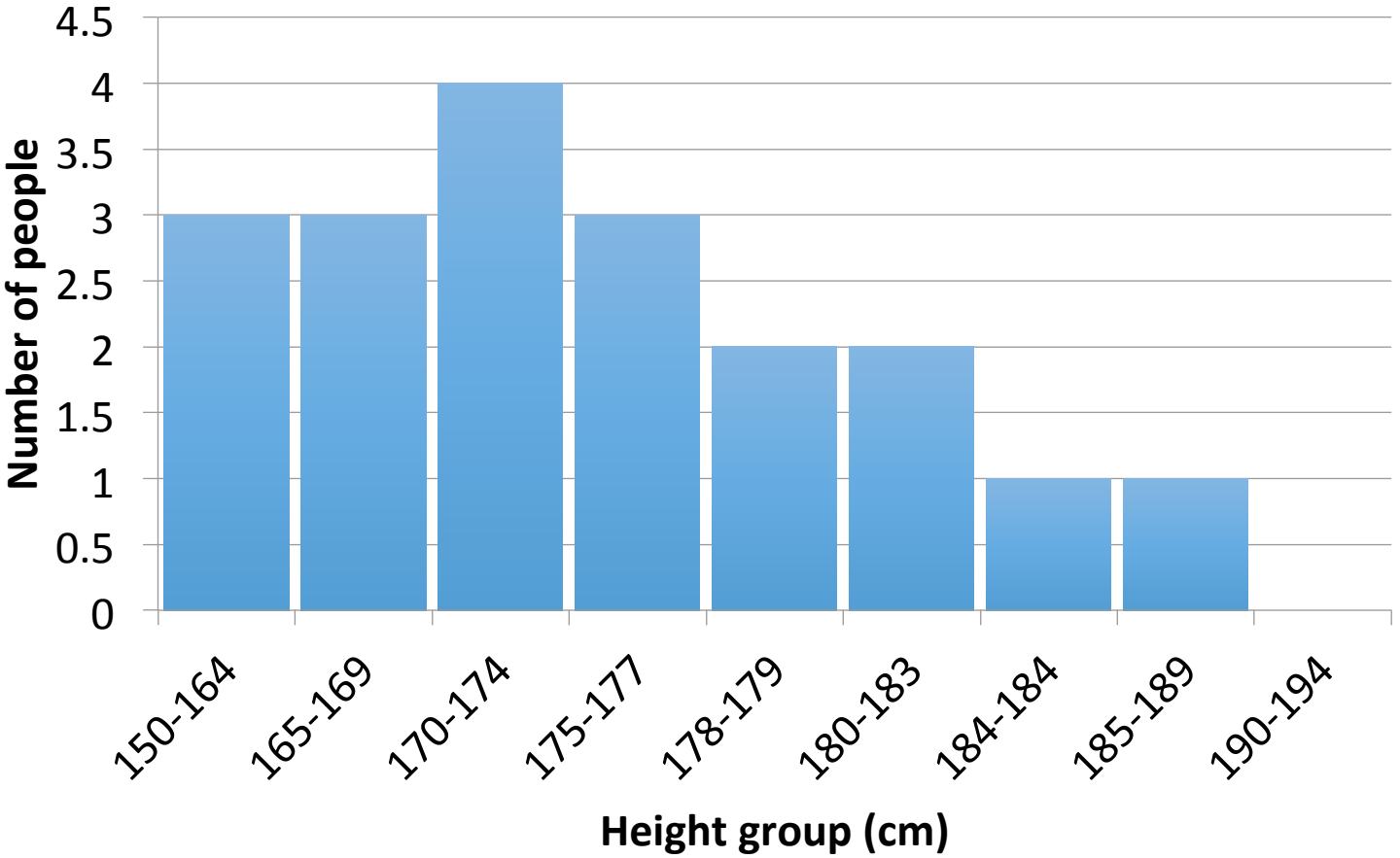
Histogram


6
5
4
3
2
1
0
Number of people
155-159
160-164
165-169
170-174
175-179
180-184
185-189
Height group (cm)

Frequency polygon



Histogram: cheating

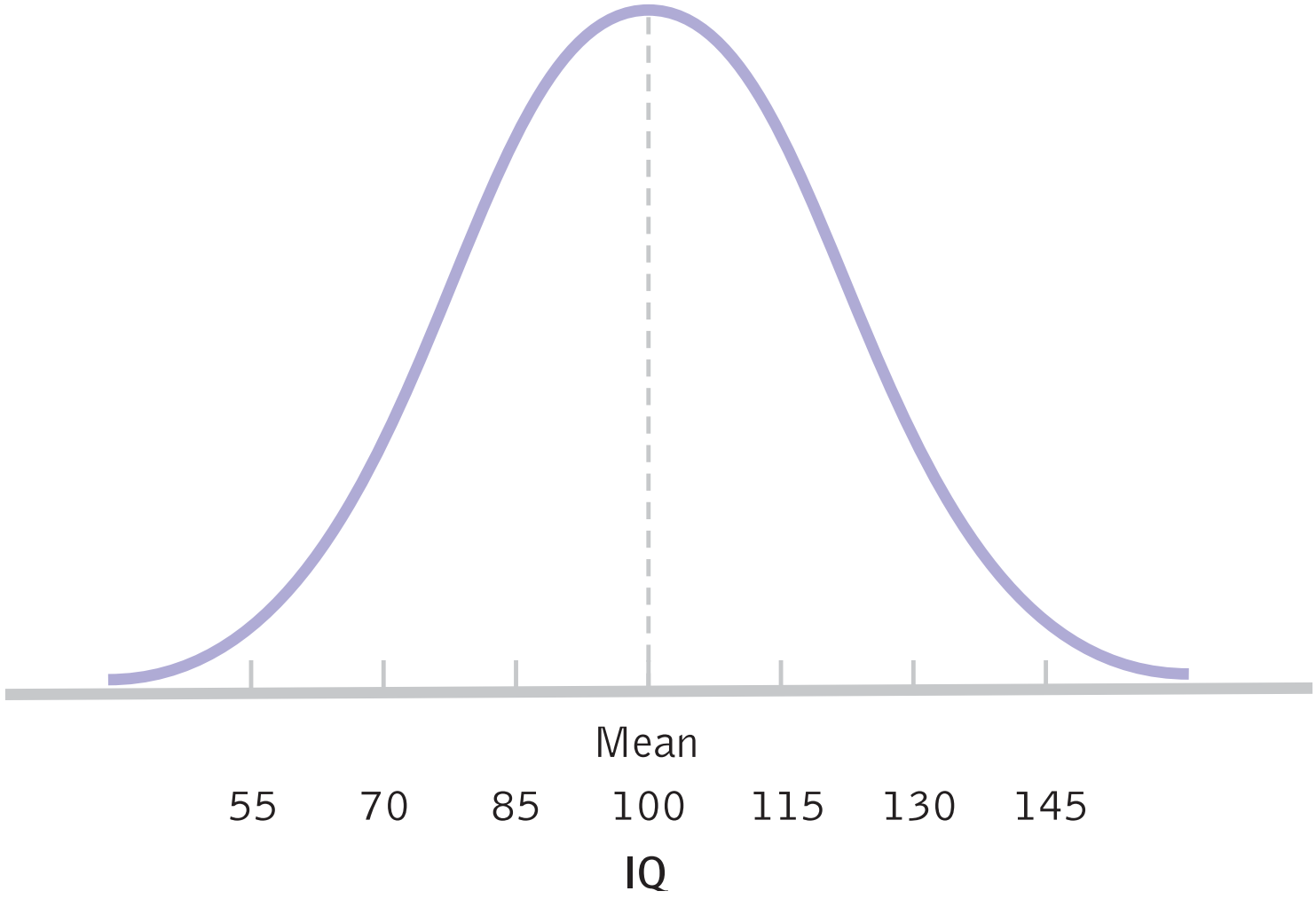


Probability distributions

- http://students.brown.edu/seeing-theory/distributions/index.html#first

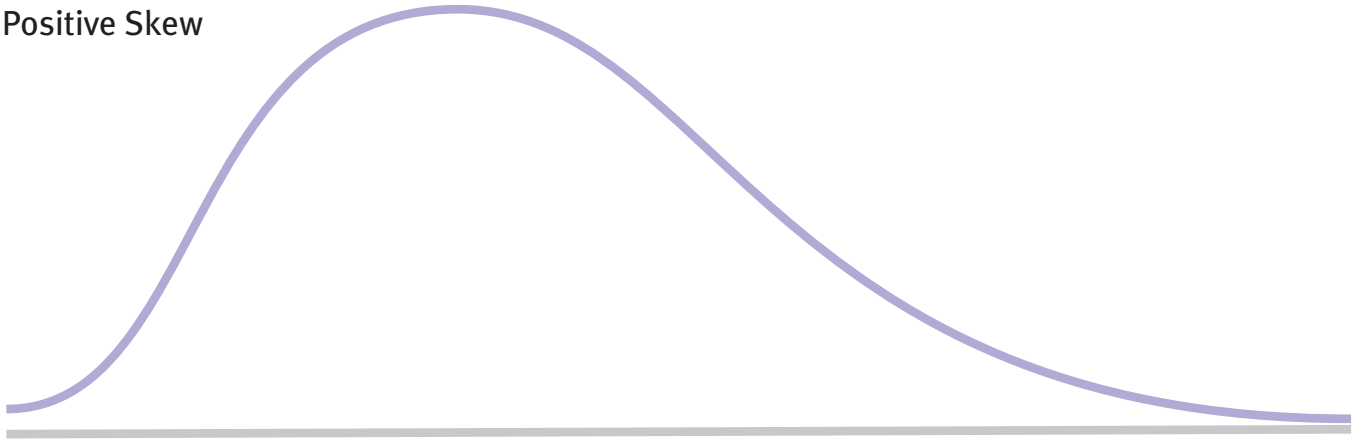
- Distribution probability calculators: http://www.distributome.org/V3/calc/index.html

Normal distribution (bell-shaped)

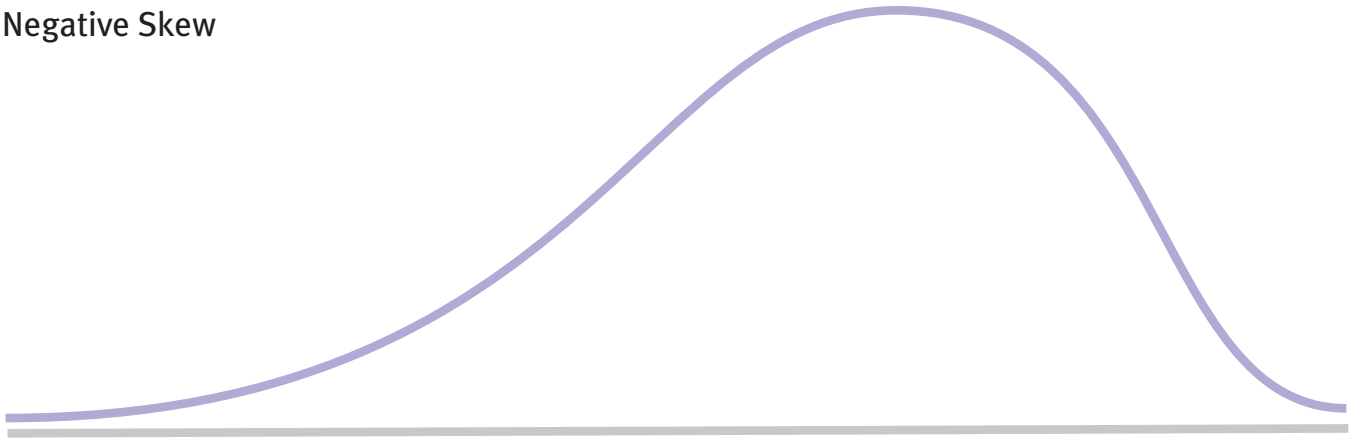


Skewed distributions

(a) Positive Skew



(b) Negative Skew



Exercises

1 What would the histograms look like for

English Test 1 (ceiling effect)

English Test 2 (floor effect)?

2 Draw a histogram for number of Facebook friends

3 Draw histograms for hours of sleep on Saturday vs. Tuesday

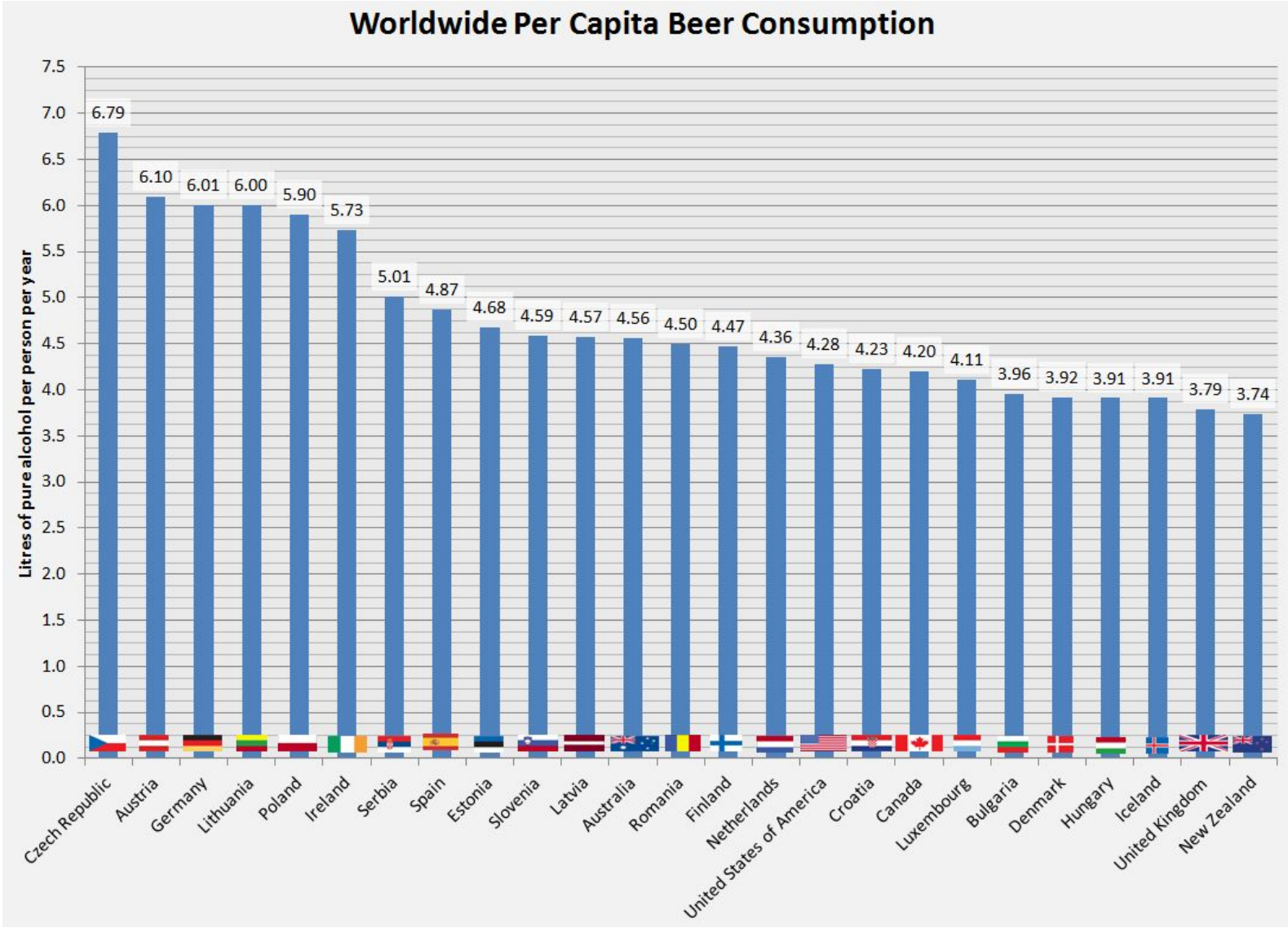
4 Get SPSS to draw a normal curve over the histogram and to display a frequency table. Analyse –> Desriptive Statistics –> Frequencies or Chart Builder > Histogram

5 Look at the various options the SPSS dialogue gives you. Try to guess what they are.

6 Did SPSS choose the right bin size? If not, change it. (Double click chart. Bin/un-bin element.)

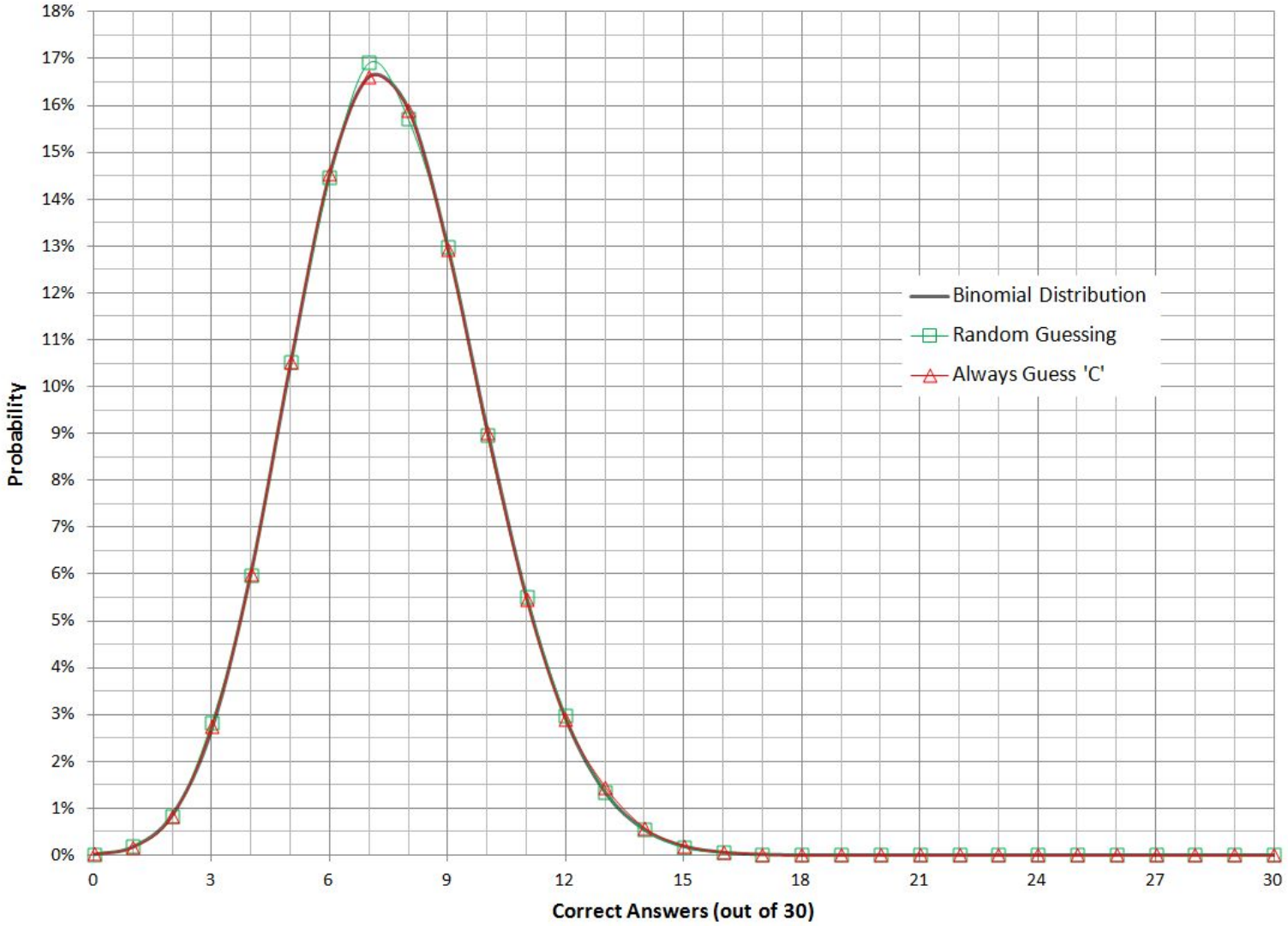
7 Draw histograms for the number of Facebook friends of those who went out last Saturday vs. those who did not. (You'll either need to split the file and then use Frequencies; use Explore; or use the coloured histogram from Chart Builder)

A histogram is NOT the same as a bar chart showing averages



Beer and labour

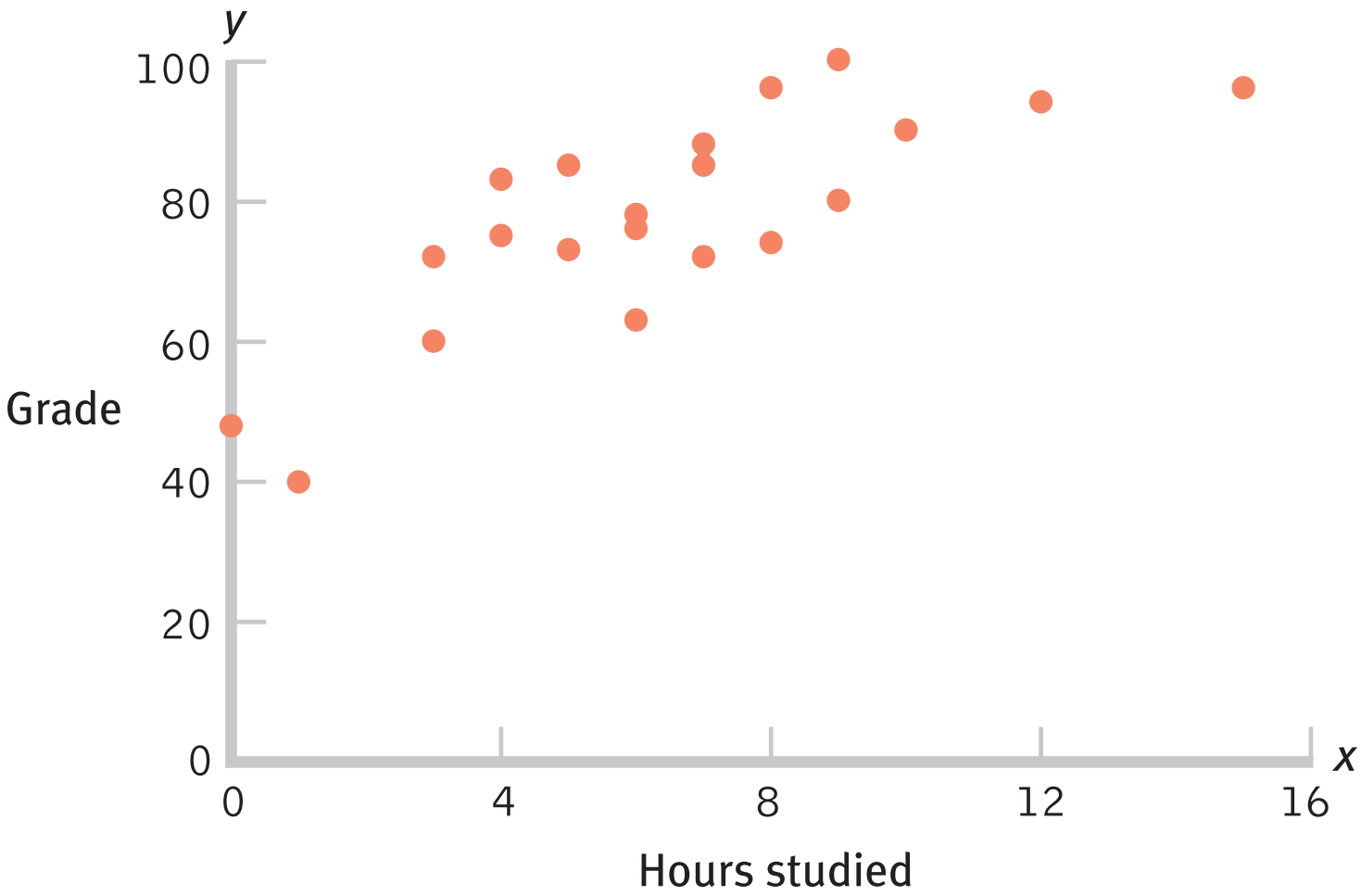
What graph is this? (Multiple choice tests)



scatterplots

- Correlation between two variables
- The variables have several levels
- The horizontal axis shows one variable and vertical axis the other variable
- The points correspond to individuals

A scatterplot
(interval, ratio or ordinal data)



Contingency table (nominal data)

	Lives in Buda	Lives in Pest	Total
By bike	8	3	
Public transport	11	14	
Other way	2	1	
Total			

Exercise

1. From your homework data
   - Which relationships would be best displayed in a bar chart?
   - Which relationships would be best displayed in a scatterplot?
   - Which relationships would be best displayed in a contingency table?

2. Create a bar chart, a scatterplot and a contingency table
   - In SPSS go to Graphs for a bar chart and a scatterplot
   - For a contingency table, go to Analyze -> Descriptive Statistics -> Crosstabs and choose the two variables

   In Google Sheets:
   - Select columns
   - Insert > Chart
   - Choose chart, customise it.
   - For a contingency table use Data > Pivot table

Characterising a data set

- Two important characteristics:

- Central tendency: the “middle” value of the set of data
  - Definition of the “middle” depends on the type of data: for a normal curve, the middle of the histogram

- Variability: to what extent scores deviate from the middle value
  - The spread of the histogram

Measures of Central Tendency

- Mean: the arithmetic average
  - Most commonly used central tendency measure
  - Used in later inferential statistics
  - For interval/ratio data
- Median: the middle score in a distribution
  - Less affected than the mean by a few extreme scores
- Mode: the most frequently occurring score
  - Easy to compute from frequency distribution
  - Can be used for any type of data including nominal

Computing the sample Mean (M or x bar)

- Compute the mean of 3, 4, 2, 5, 7, & 5
- Sum the numbers (26)
- Count the number of scores (6)
- Plug these values into the equations

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

$$M = \frac{26}{6} = 4.33$$

- Google Sheets: =AVERAGE()

symbols

- μ = population mean (“mew”)

- M or $\overline{X}$ (x bar) = sample mean
- ∑ = Sigma, summation (“add all of these”)
- N = sample size

Computing the Median

- Order the scores from smallest to largest
- Determine the middle score [(N+1)/2]
  - If 7 scores, the middle is the fourth score [(7+1)/2]=4

    Median of [11, 27, 32, 43, 46, 47, 51] = 43

  - If 8 scores, the middle score is half way between the 4th and 5th scores [(8+1)/2]=4.5

    Median of [11, 27, 32, 43, 46, 47, 51, 56] = 44.5

- Google Sheets: =Median()
- It has no symbol! (Median, Mdn)

Outliers

- Outliers are extreme values: very high or very low in comparison with the rest of the scores

- Use Number of Facebook friends, Hours of sleep or Units of alcohol

1. Create a frequency table
2. Draw a histogram – does it look normal?
3. Are there any outliers?
4. Estimate the mode, mean and the median from the histogram
5. Calculate the mean and the median – are they similar?
6. You can do all that in SPSS: Analyze -> Descriptives -> Frequencies

Measuring Dispersion

Measures of dispersion show us how well the mean (median, mode) represents the sample – and the population

- **Range**: distance between lowest and highest score
  - Simple but outliers are a big problem
- **Variance**: average squared distance from the mean
  - Used in later inferential statistics
- **Standard Deviation**: square root of variance
- (inter-quartile range)

Variance & Standard Deviation

σ = Standard Deviation = square root of variance (for population)

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

SD = Standard Deviation = square root of variance (for sample)

$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{N - 1}}$$

The Variance

- Computing the Variance
  - Compute the mean (M)
  - Compute the distance of each score from the mean (X – M)
  - Square those distances $(X – M)^2$
  - Sum those squared distances: Sum of Squares: SS
  - Divide by
    - the number of observations (n) for the set of observations (**population**)
    - the degrees of freedom (*N* - 1) for the population estimate from the **sample**
- Google Sheets: =VARP() or VAR()
- Good statistical properties, but this measure of variability is in squared units

The Standard Deviation

- Computing the Standard Deviation
  - Compute the variance
  - Take the square root of the variance

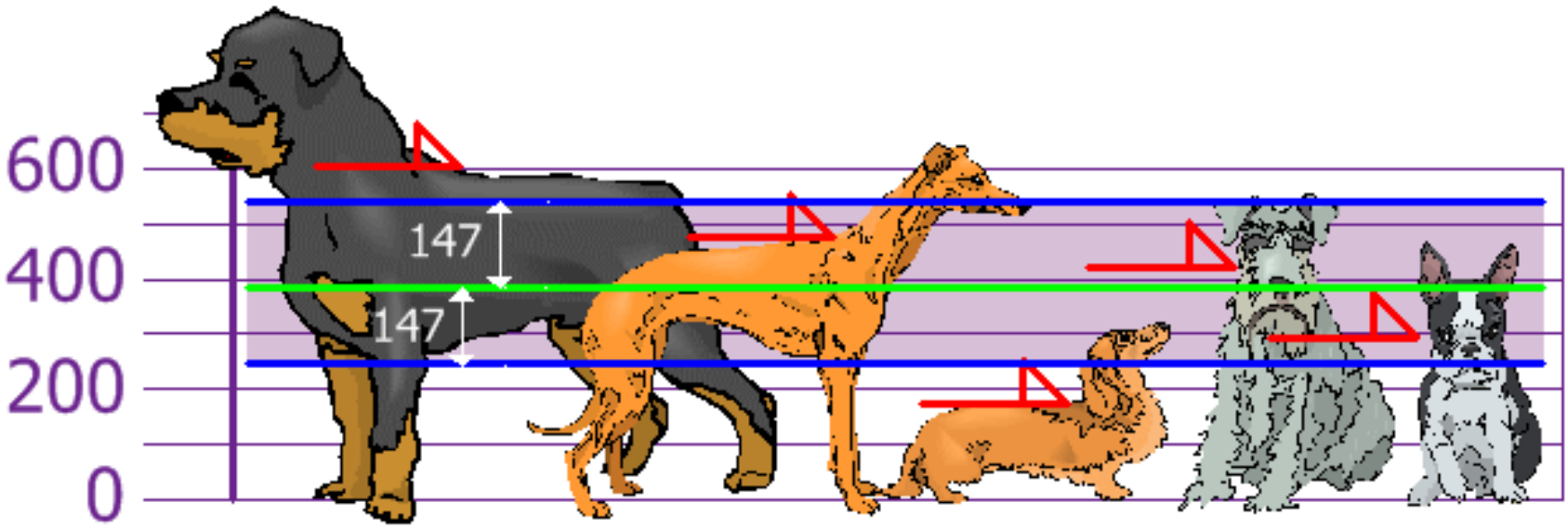
- Google Sheets: =STDEVP() or STDEV()

- This measure, like the variance, has good statistical properties and is measured in the same units as the mean

Symbols

Target	Standard Deviation	Variance	Number of observation
Sample (from which you generalise)	SD, s	SD^2, s^2	N
Population (just the observations)	σ	σ^2	n

What does the standard deviation mean?



Mathisfun.com

Dogs' heights at the shoulders (mm): 600, 470, 170, 430, 300

SD (set of observations) = 147 mm

➔ A dog up to 147 mm taller or shorter than average, is within the "standard" range (those even shorter or even taller are tiny or huge)

SD (population estimate) = 164 mm

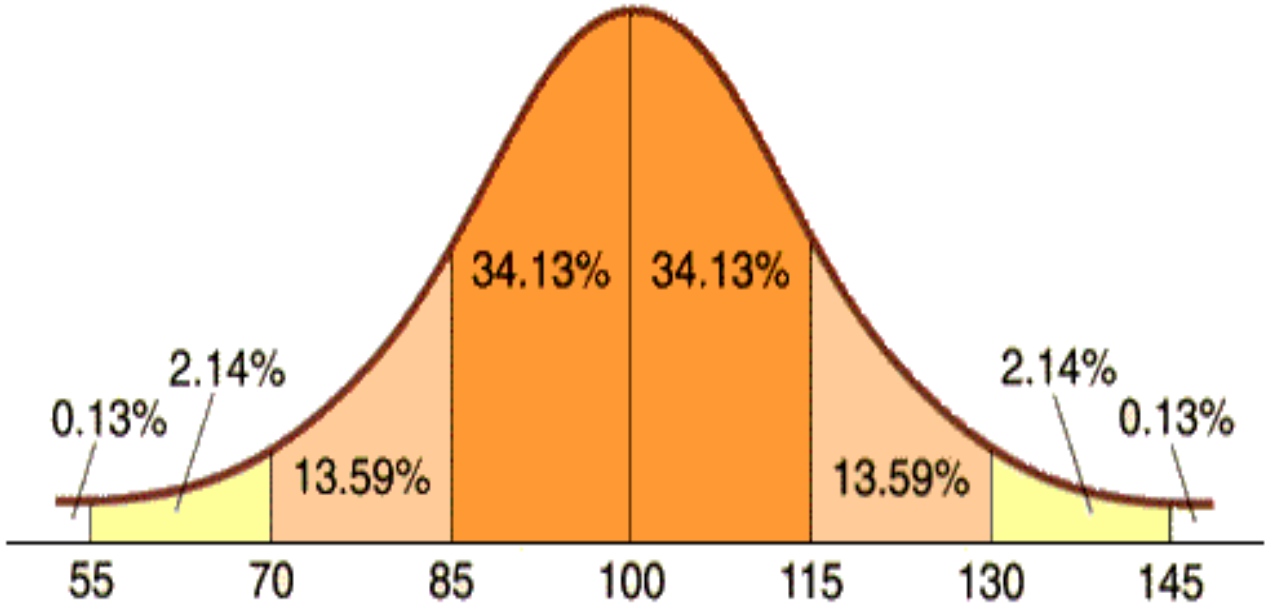
→ Allows for slightly greater variation to compensate for an imperfect sample

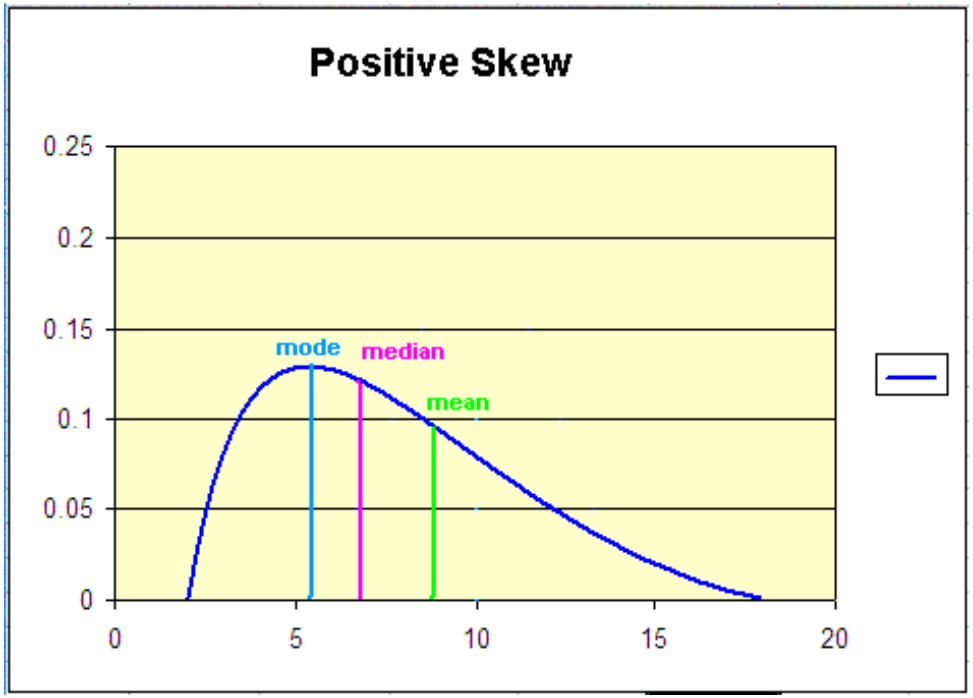
The Normal Distribution

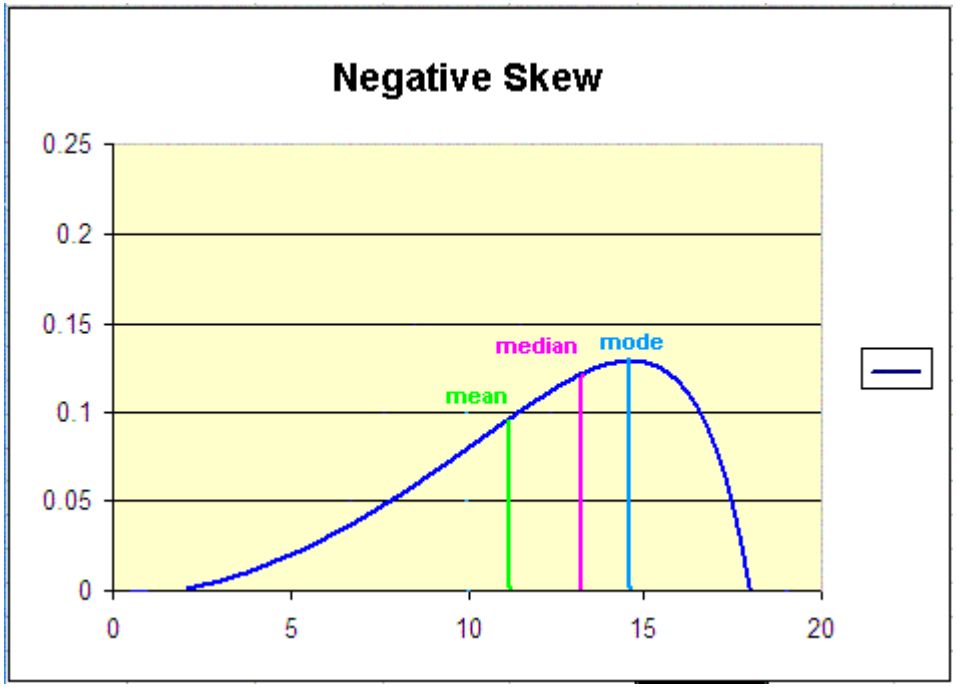
https://www.mathsisfun.com/data/standard-normal-distribution-table.html

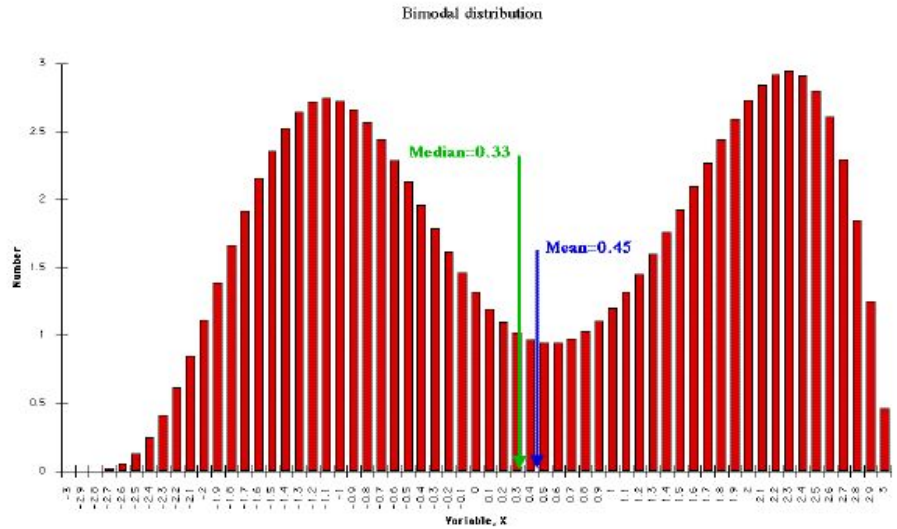
Described by:

- Shape: unimodal
- Central Tendency: mean = median = mode
- Variability:
  68% of values are within 1 SD of the mean
  95% of values are within 2 SD of the mean
  99.7% of values are within 3 SD of the mean

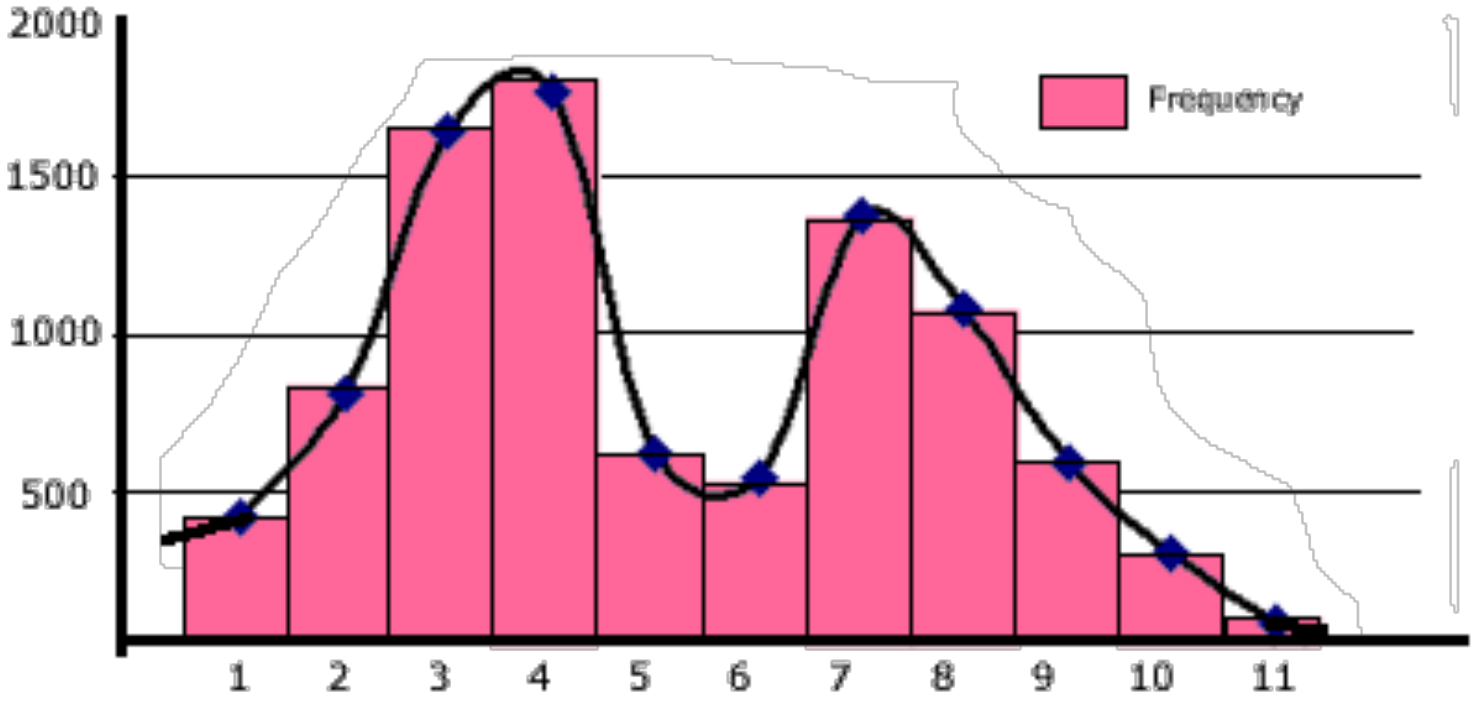



Positive Skew
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
mode
median
mean

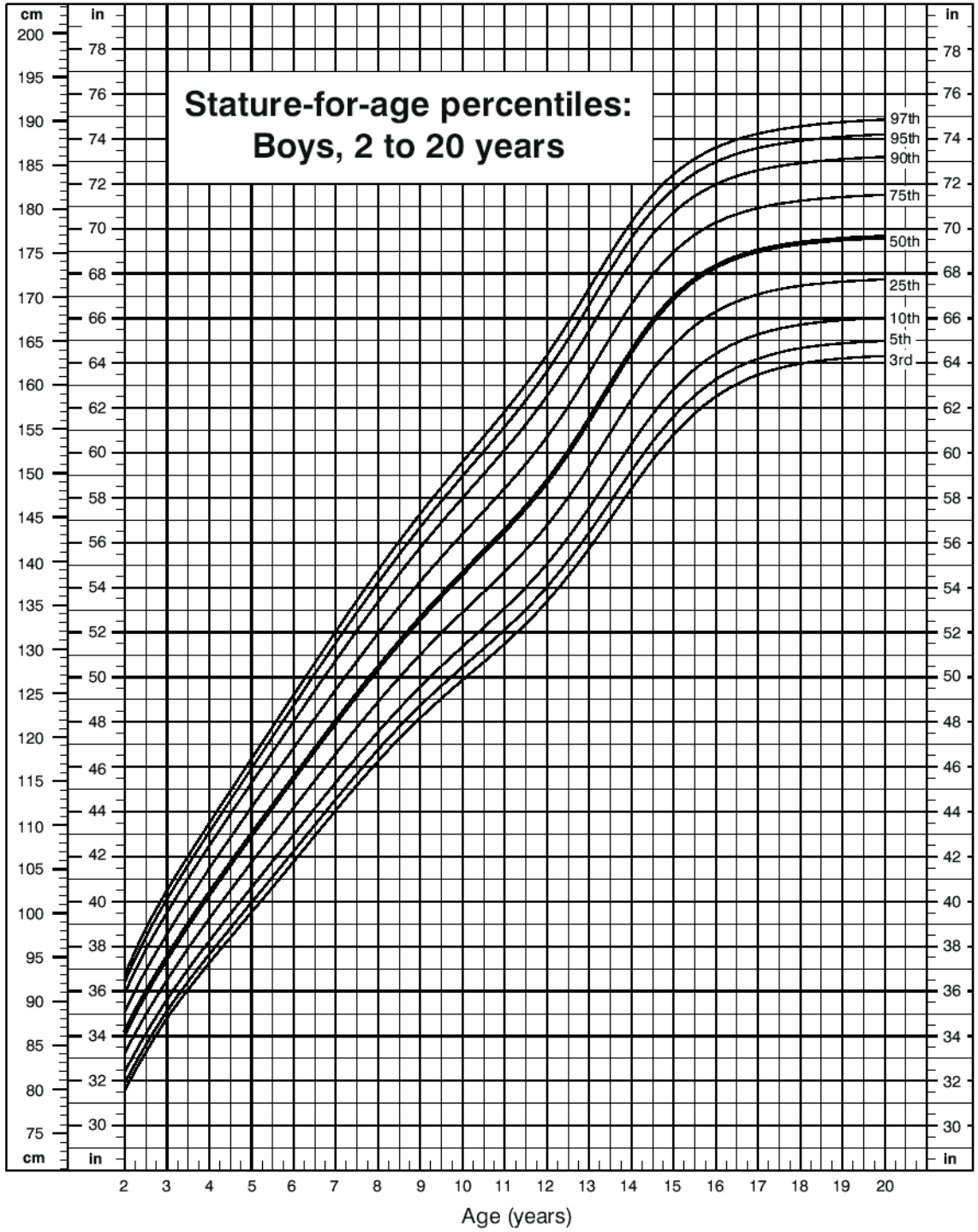

Negative Skew
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
median
mode
mean


Bimodal distribution
Median=0.33
Mean=0.45
Number
Variable, X

Multi-modal distribution



Growth chart



percentiles

- *P*th percentile: the value for which *P* percent of the scores are less than that value
- Deciles: 10^{th}, 20^{th}, etc percentiles
- Quartiles: 25^{th}, 50^{th} and 75^{th} percentiles

Q1 Q2 Q3

(In Excel: =QUARTILE(range,which))

- Another measure of variability:
- Interquartile range: Q3 – Q1

- Q2 is of course the … (which measure of central tendency?)

Approximating quartiles by hand

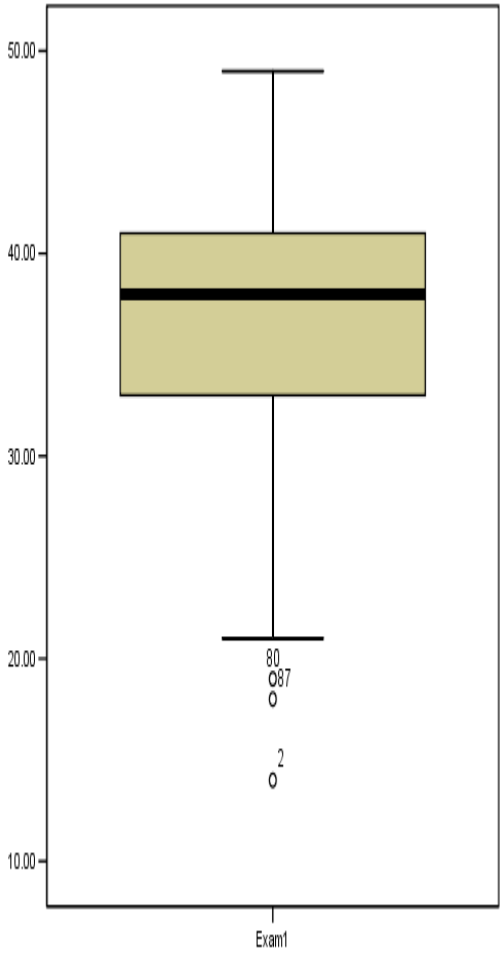
- 1 3 3 4 5 6 6 7 8 8

- Put the values in order
- Find the middle value (median)
- For 1st quartile, take the first half of the values and find the middle value
- For 3rd quartile, take the second half of the values and find the middle value

1st quartile	3
Median, 2nd quartile	5.5
3rd quartile	7

Box-Plot

- Box 'Bottom' = 1st quartile
- Box 'Midline' = 2nd quartile (Median)
- Box 'Top' = 3rd Quartile
- Whiskers = Values closest to 1.5 times the interquartile range above or below the box (or sometimes above or below the median)
- Beyond the whiskers: outliers and extreme outliers



A more accurate method

- To calculate all measures of central tendency and variability, including interquartile range and box plot:

- http://www.alcula.com/calculators/statistics/dispersion/

- or SPSS Descriptive Statistics > Explore

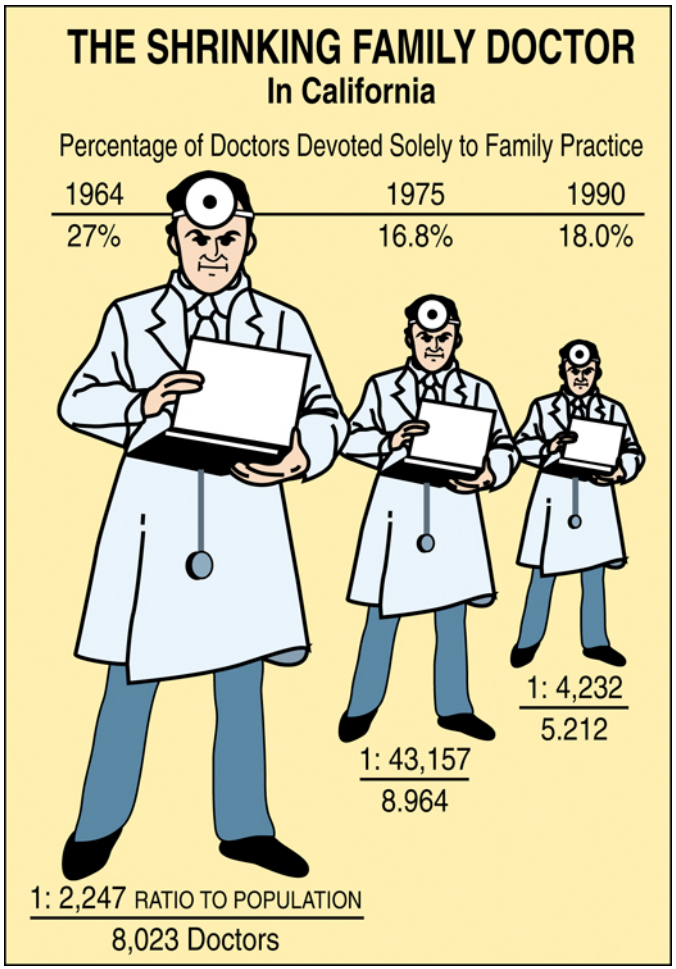
exercise

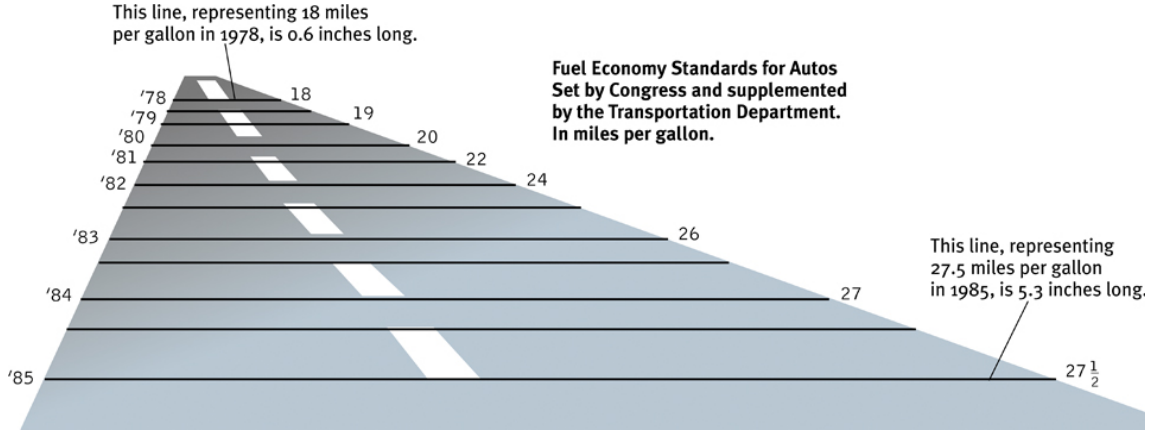
- Take the data of units of alcohol for those who went out and those who did not go out
- Draw histogram and box plot, find mode, median, mean, SD interquartile range and outliers
  - You can do this separately for the two groups via Analyze -> Descriptive statistics -> Frequencies + the Box plot via Graphs
  - Or in one go via Analyze -> Descriptive statistics -> Explore
- Calculate the percentile for the number of hours you slept on Saturday
- How much alcohol does the data predict you would have had if you had/had not gone out that night?

Misleading graphs

- http://www.statisticshowto.com/misleading-graphs/
- http://simplystatistics.org/2012/11/26/the-statisticians-at-fox-news-use-classic-and-novel-graphical-techniques-to-lead-with-data/

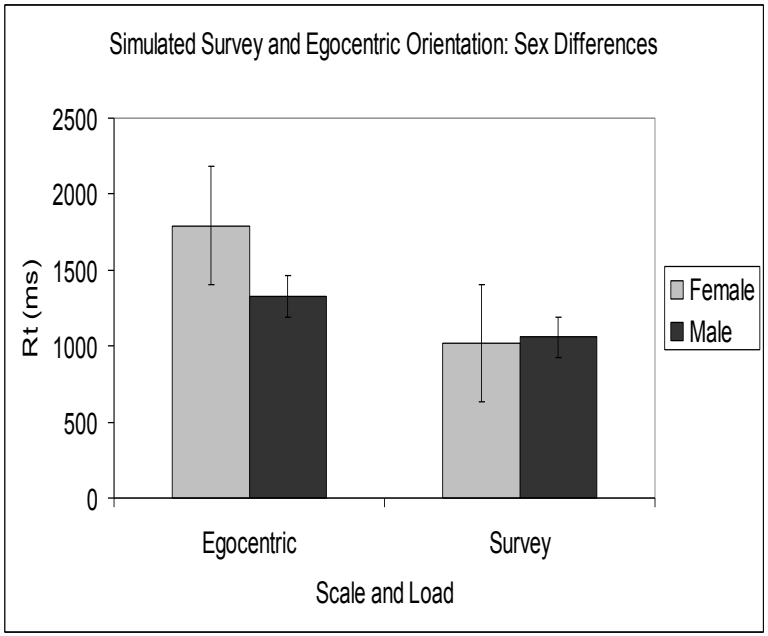
How To Lie with Graphs

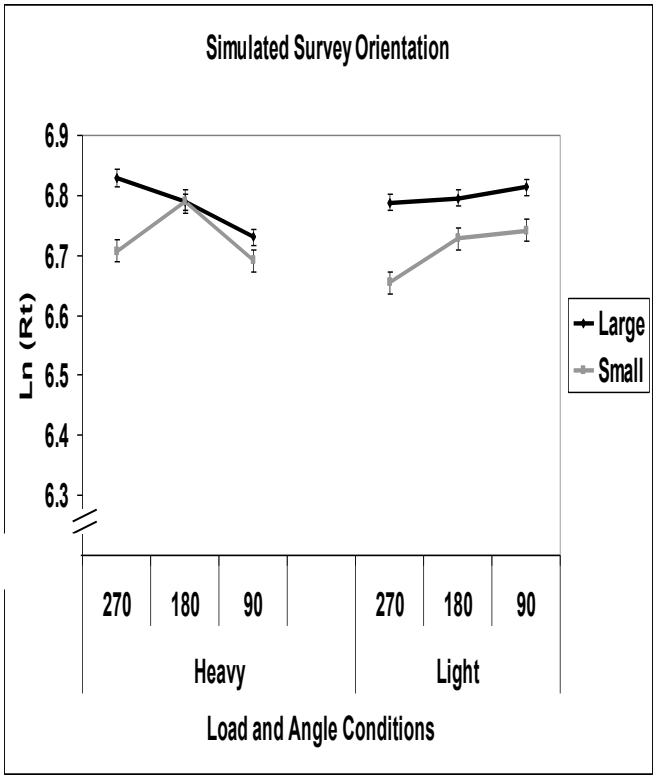




Tricks in Describing Stats

- If you omit Zero in your scale, must indicate with ellipsis

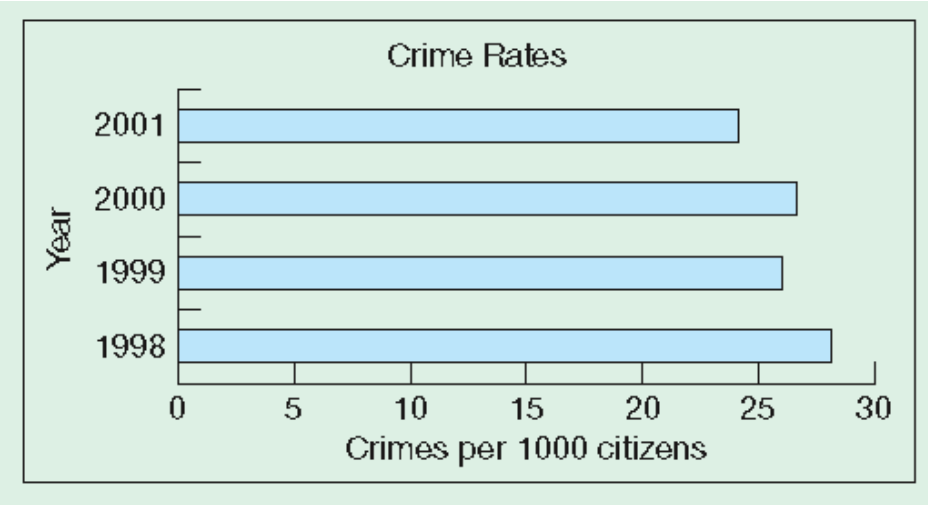


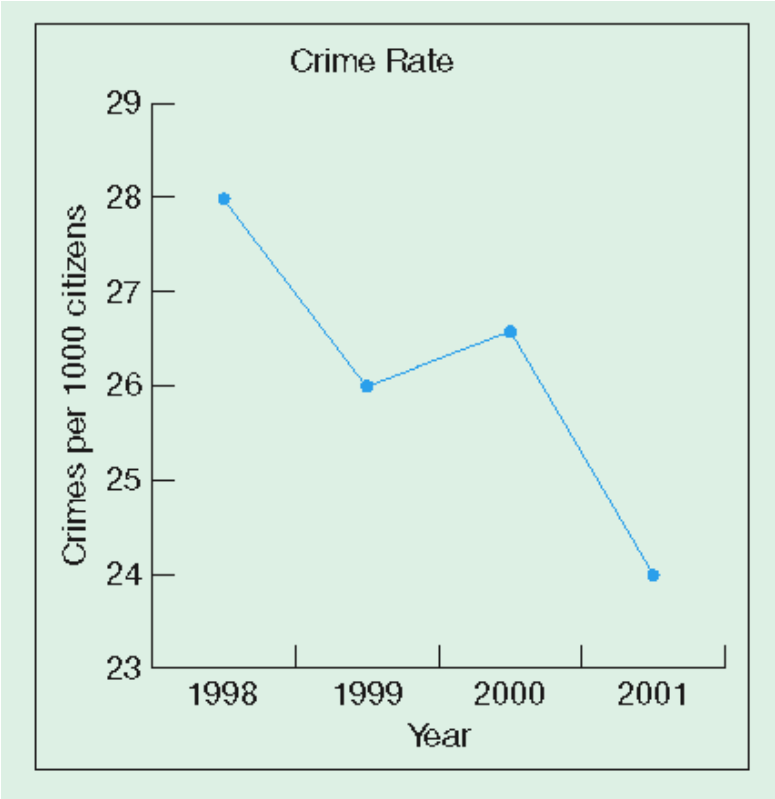


Starting Point

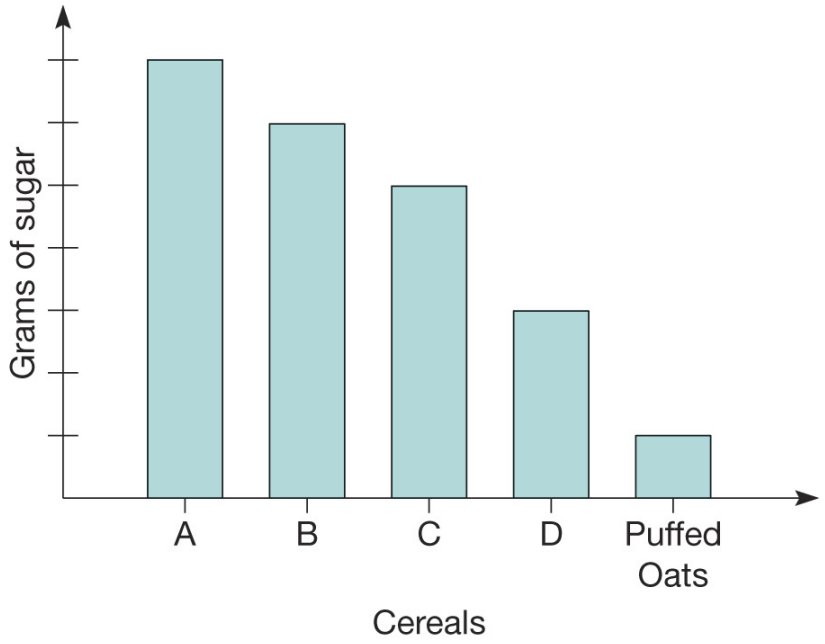
- Mayor Marcus is running for a second term against a challenger. Which graph should he send to the local journalist who is reporting

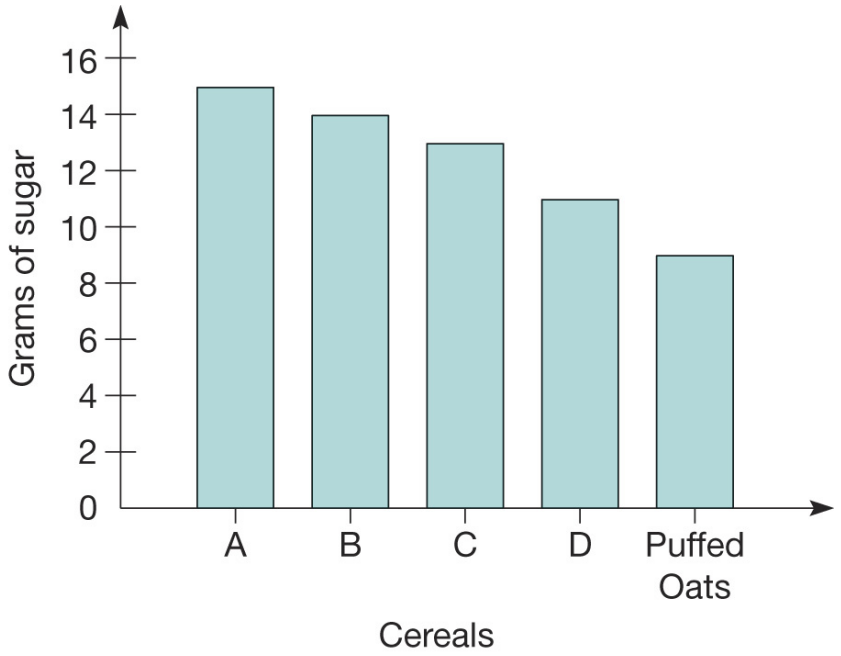
  on crime rates?



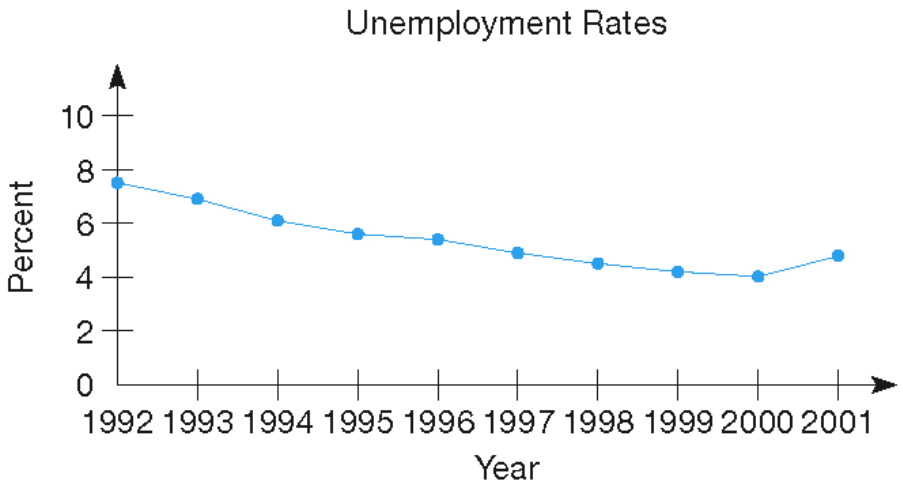


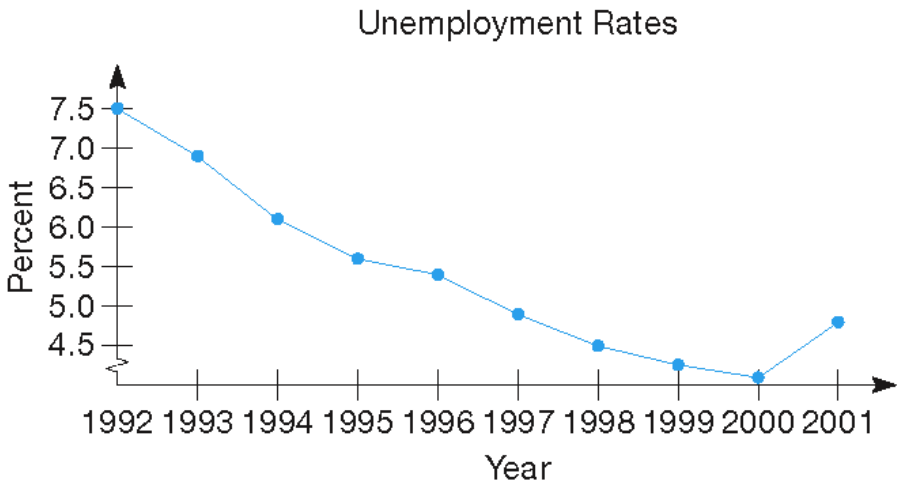
What's wrong with this picture?





What's wrong with this picture?





Homework

Mary Claire magazine's observation of shopping habits:

Put the data into an SPSS file, draw histograms, calculate means and standard deviations, draw boxplots and identify outliers, draw scatterplots, bar charts.

(Their conclusion was that shopping was good for you because of all the exercise you get.)

Men		Women	
Minutes spent shopping	Kilometres walked	Minutes spend shopping	Kilometres walked
15	0.16	22	1.40
30	0.40	140	1.81
37	1.36	160	1.96
65	1.99	183	3.02
103	3.61	245	4.82
17	0.22	145	1.88
32	0.44	163	1.90
36	1.36	189	2.93
68	2.0	235	3.40
47	1.75	222	3.17