Data preparation

E0420

Week 2

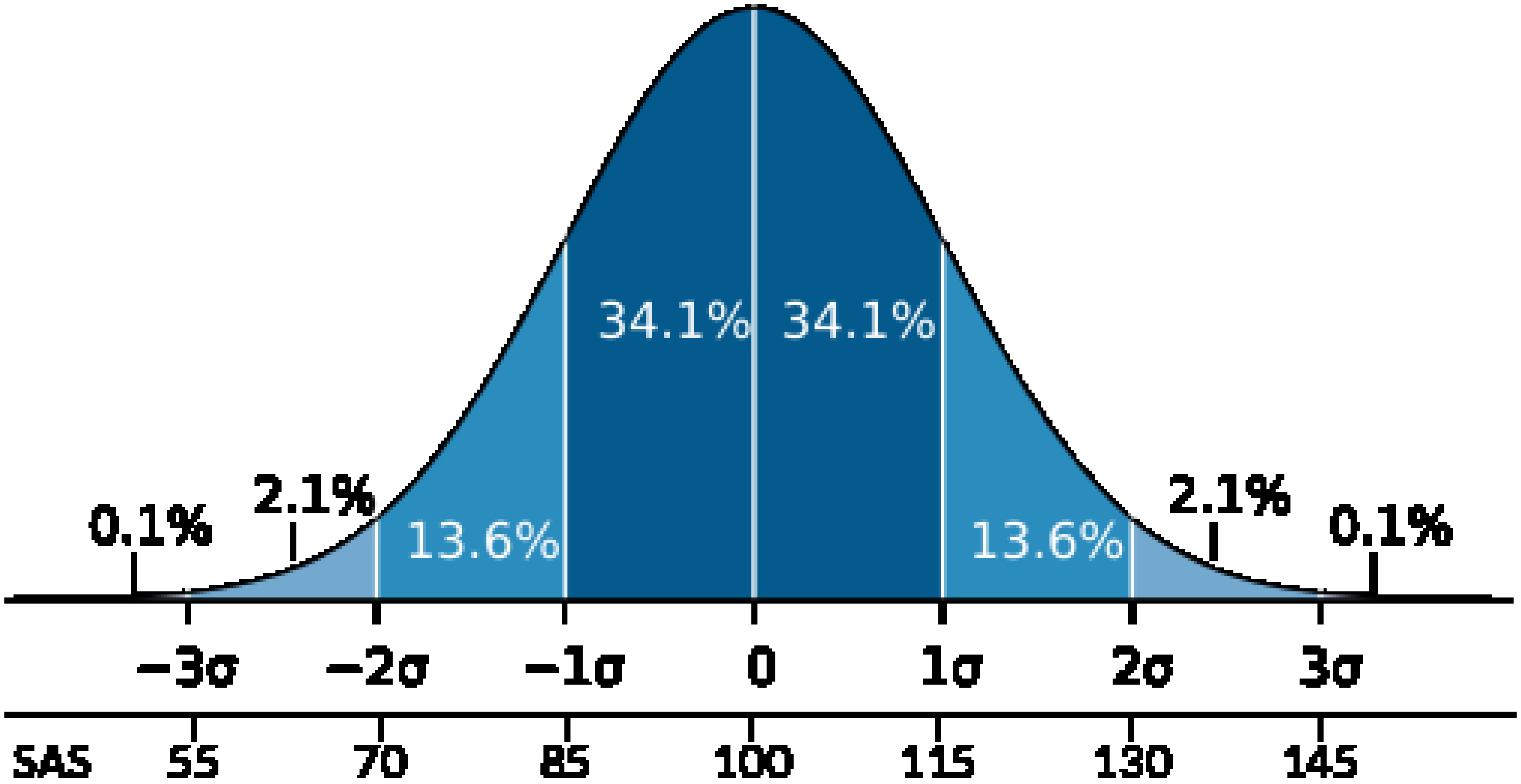
Let me analyze already!

- Different types of variables
- Basic diagnostics of variables in dataset are necessary
- Without it, findings can be meaningless/spurious/null!

Distribution

- How are values distributed within a sample
- The shape of the distribution determines how we can analyze the data
- Fortunately, majority of values in a sample **conform** to a well-known distribution

Normal/Gaussian distribution



https://commons.wikimedia.org/wiki/File:Wechsler.svg

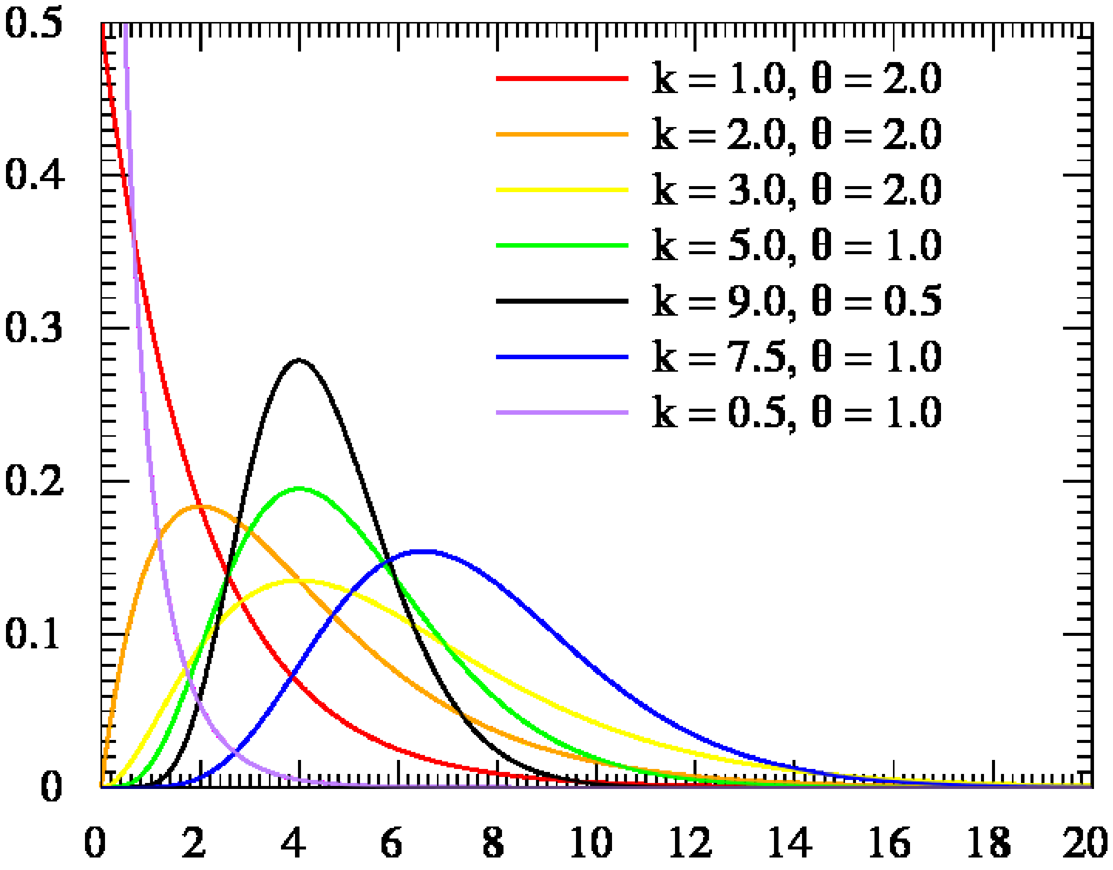
Galton board and the laws of nature

Central limit theorem

- The distribution of sums of random variables will resemble normal distribution

Specific types of distributions

- Binomial
- Beta distribution
- Gamma distribution
- and other…



https://commons.wikimedia.org/wiki/File:Gamma_distribution_pdf.svg

Basic descriptive terms

- **Sum** – adding values together
- **Mean** (M) – sum of values divided by their count
- **Mode** – most frequently occurring value
- **Median** – value at the 50% (“in the middle”)
- **Standard deviation** (SD) – distance of a value from a sample mean
- **Variance** – squared SD
- **Quantile** – cut point dividing the range of the distribution into intervals with equal probabilities
- **Minimum** – the smallest value
- **Maximum** – the largest value

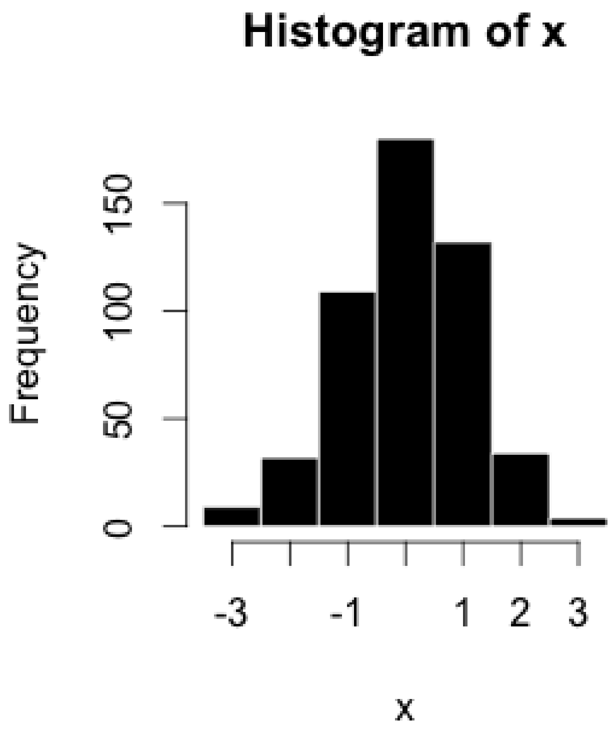
Plotting data

One variable

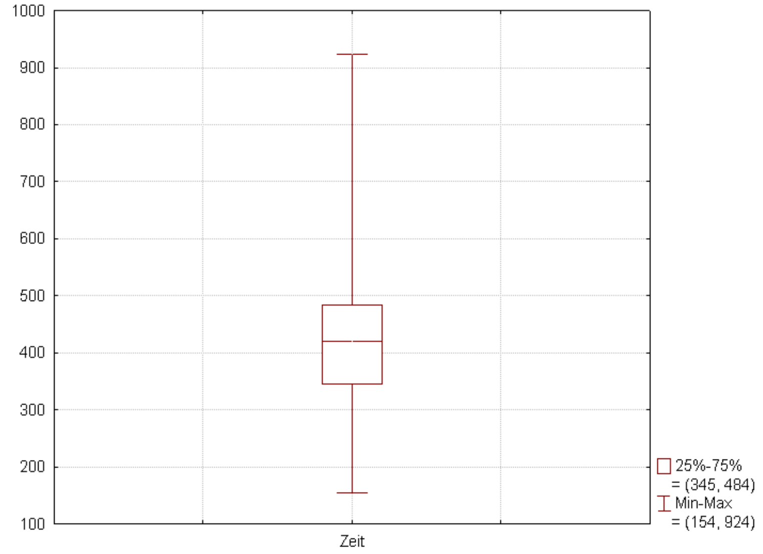
- Histogram – bars represent meaningful groups of data
- Box plot – box-and-whisker-plot
  - Represents minimum, maximum, median, and interquartile range (IQR)
  - Box is IQR (25%-75%), whiskers are min/max or 1.5 IQR

Two variables

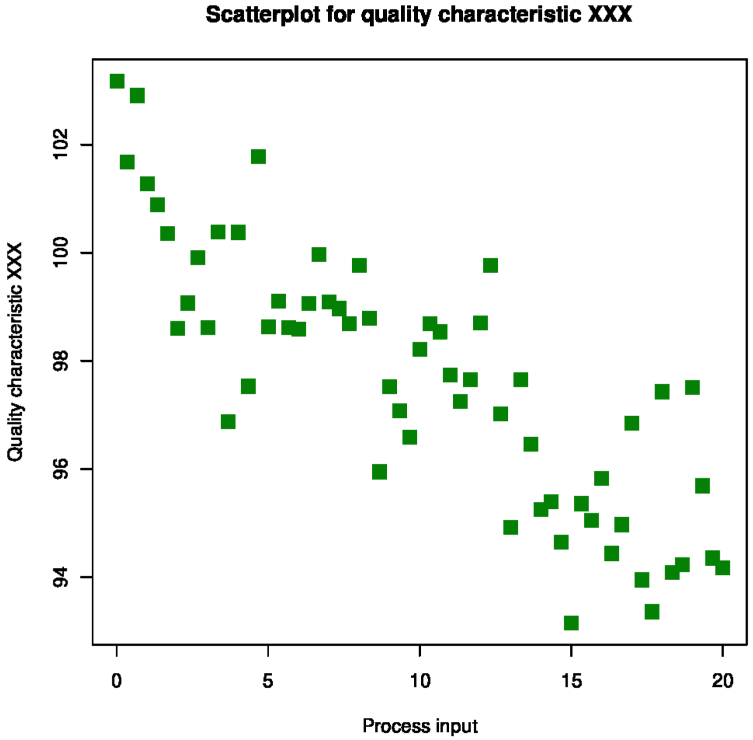
- Scatterplot
  - Represents data points as related to two variables



Box plot



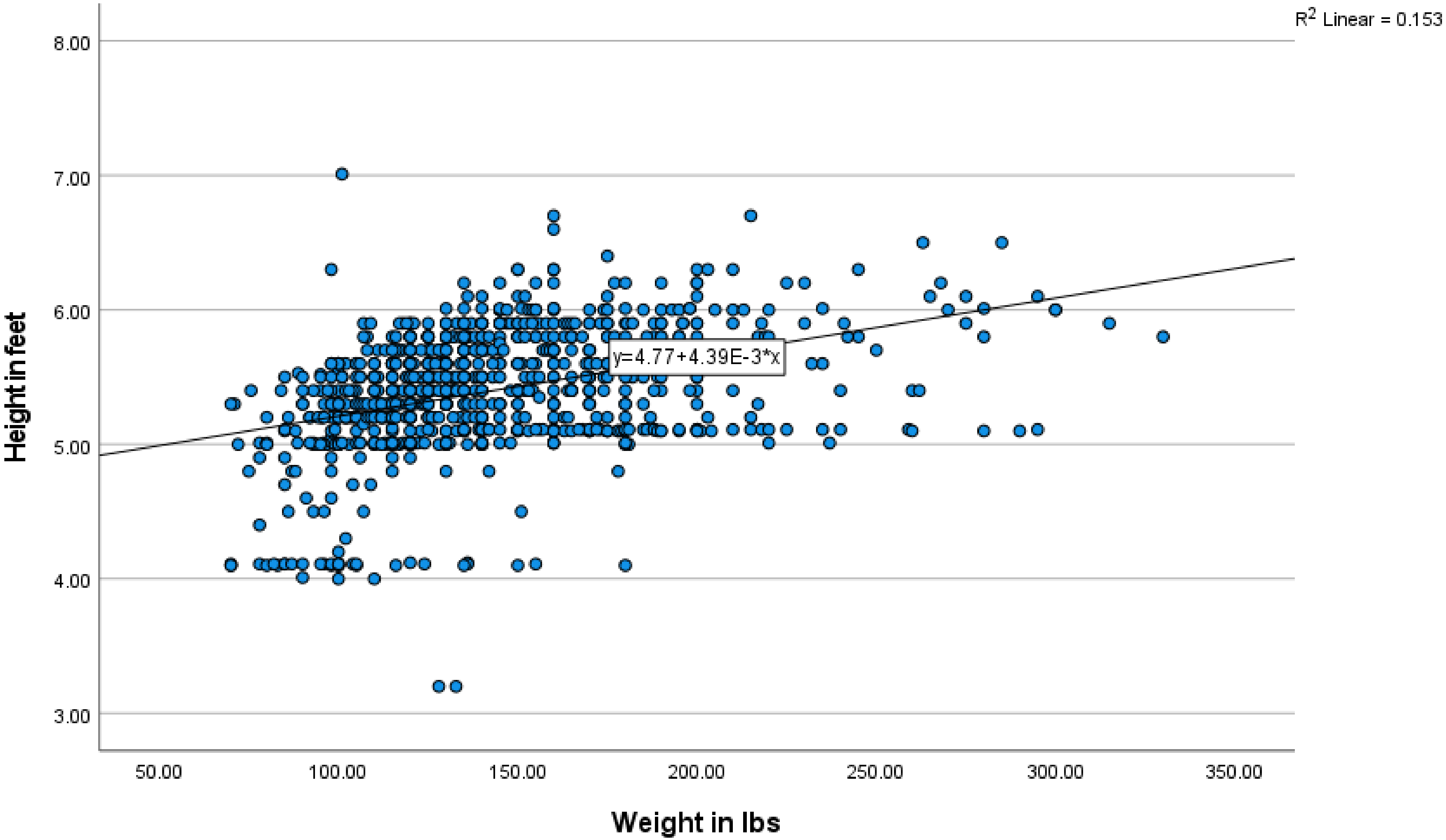
Scatterplot



https://commons.wikimedia.org/wiki/File:Example_histogram.png
https://commons.wikimedia.org/wiki/File:Box-Plot_mit_Min-Max_Abstand.png
https://commons.wikimedia.org/wiki/File:Scatter_diagram_for_quality_characteristic_XXX.svg

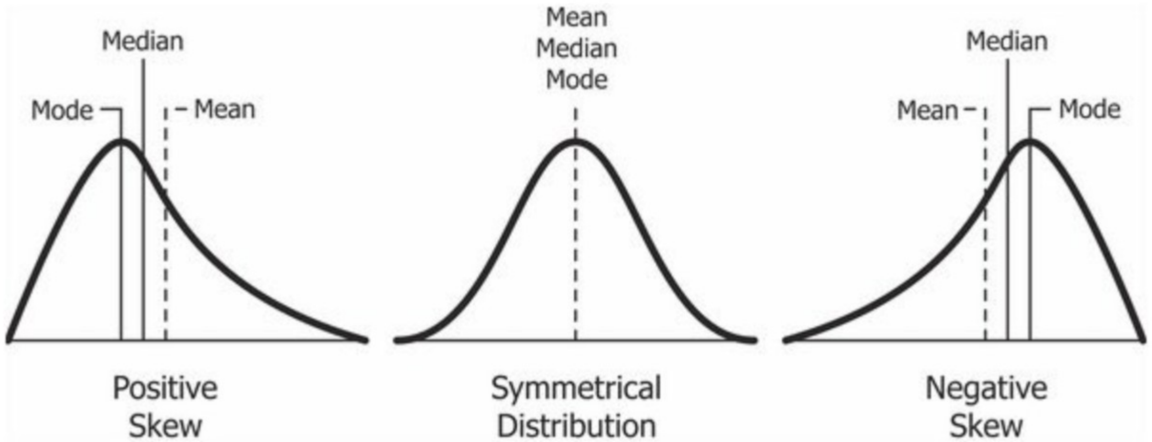
Scatterplot

- Graphical representation of association between two variables (correlation)
- can add a trendline (a line of best fit) – linear regression

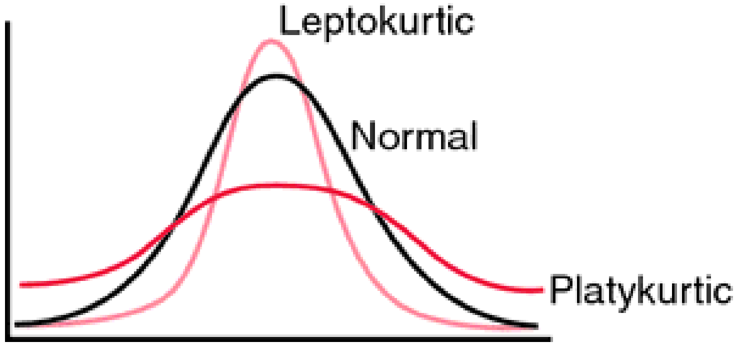

R² Linear = 0.153
y=4.77+4.39E-3*x
Height in feet
8.00
7.00
6.00
5.00
4.00
3.00
50.00
100.00
150.00
200.00
250.00
300.00
350.00
Weight in lbs

Non-normal distributions

Skewness

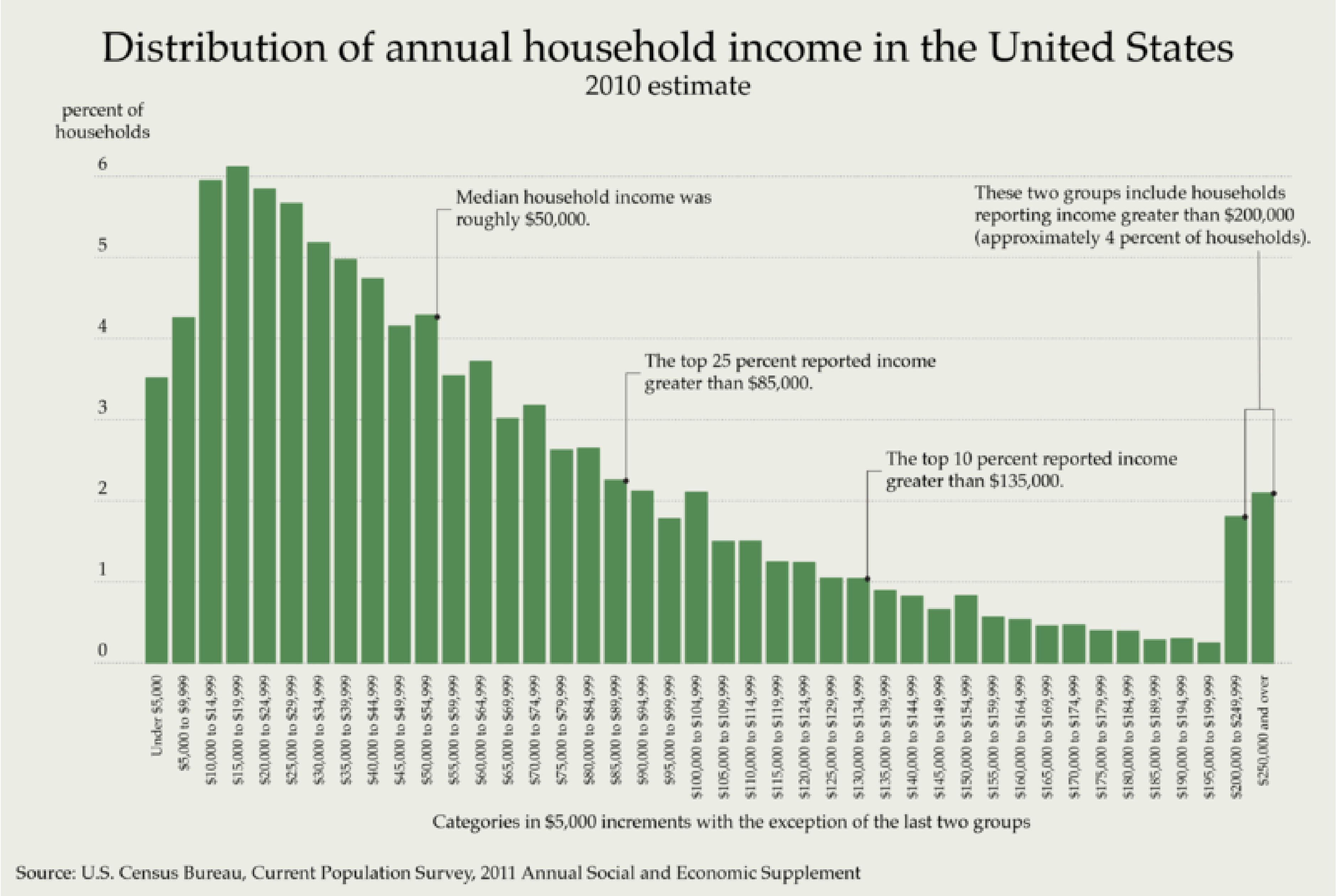


Kurtosis



https://commons.wikimedia.org/wiki/File:Relationship_between_mean_and_median_under_different_skewness.png

https://medical-dictionary.thefreedictionary.com/kurtosis



Distribution of annual household income in the United States

2010 estimate

Median household income was roughly $50,000.

The top 25 percent reported income greater than $85,000.

The top 10 percent reported income greater than $135,000.

These two groups include households reporting income greater than $200,000 (approximately 4 percent of households).

Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

https://commons.wikimedia.org/wiki/File:Distribution_of_Annual_Household_Income_in_the_United_States_2010.png

Outliers

- Atypical data points (with regards to the sample values)
- Could be due to:
  - Contamination (for bio samples)
  - Error in data entry
  - Just a really atypical case

Outliers – why do we care?

- Outliers can have a huge impact on the characteristics of the sample

- Example
  - Erasmus students in class – 10 students

  - **With outlier:**
    - M = 25.8
    - SD = 15.9
    - Median = 21

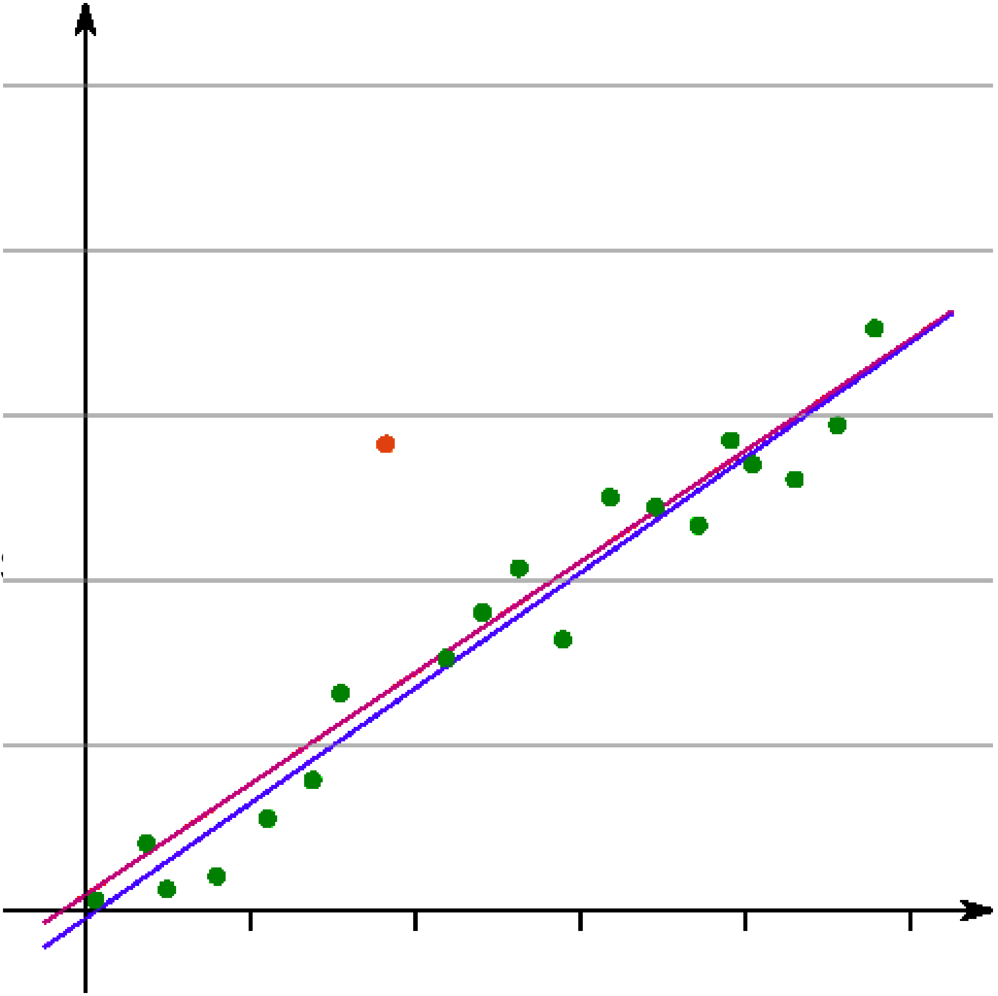
  - **Without outlier:**
    - M = 20.8
    - SD = 0.83
    - Median = 21

#	**age**
1	20
2	21
3	20
4	22
5	21
6	20
7	22
8	20
9	71
10	21

Identifying outliers

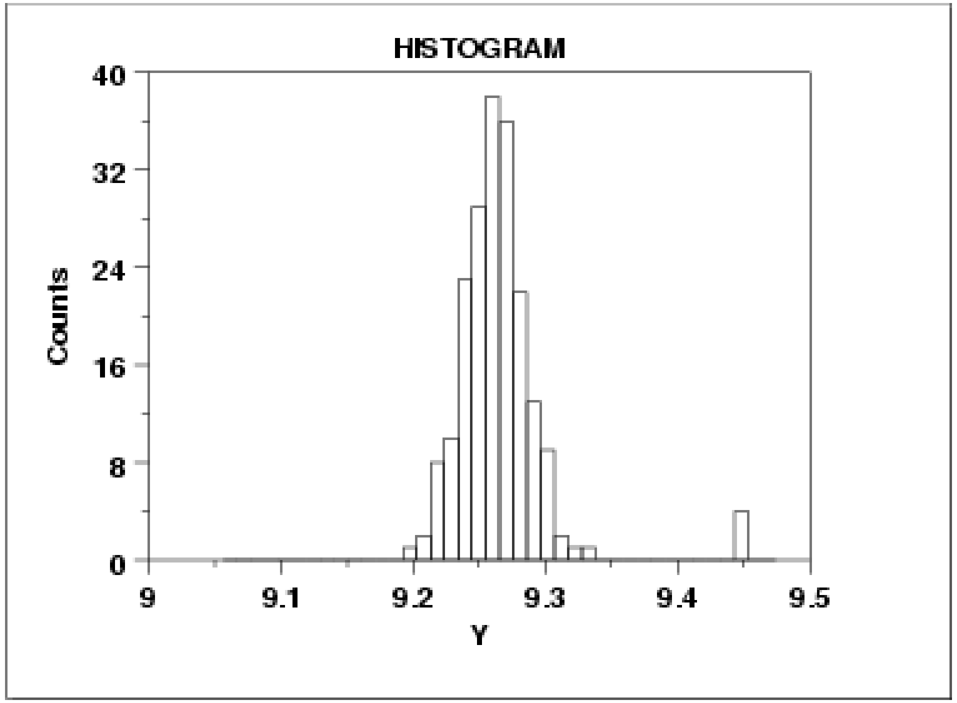
1. “Eyeballing it” - scatterplot
2. Using box plot or histogram
3. Using some cut-off
   - +- 2 SDs or 1.5*IQR -Q1

4. Using indices for multivariate outliers

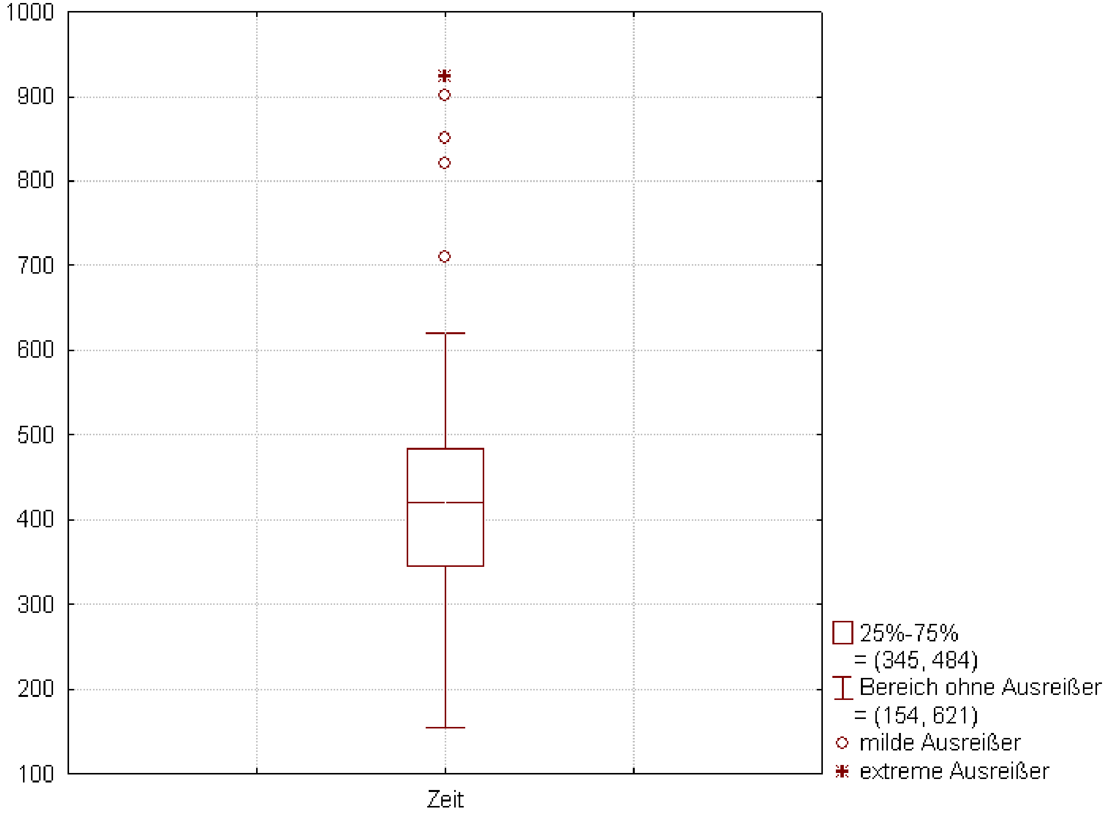


https://commons.wikimedia.org/wiki/File:Outlier_statistics.svg

Identifying outliers – graphs



Box plot with 1.5 IQR = everything beyond that is outlier

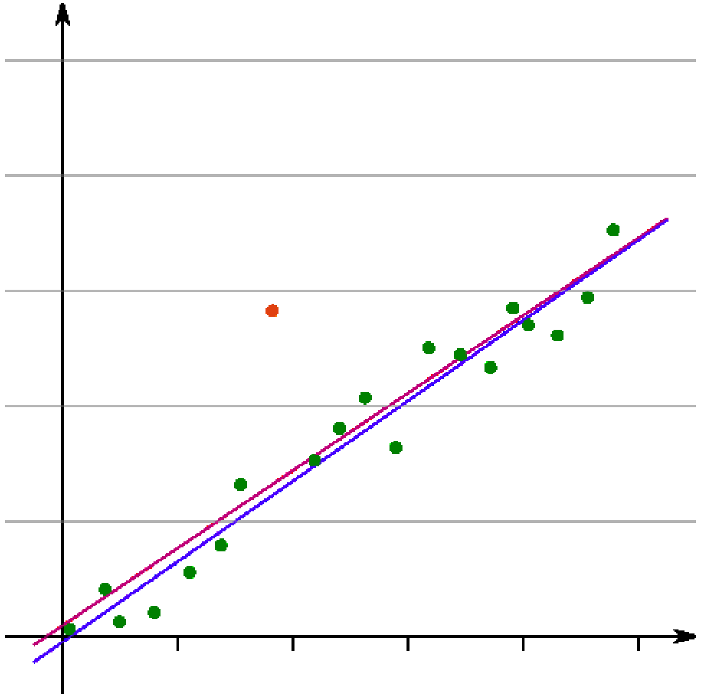


https://en.wikipedia.org/wiki/Box_plot#/media/File:Box-Plot_mit_Interquartilsabstand.png

https://www.itl.nist.gov/div898/handbook/eda/section3/eda33e8.htm

Mahalanobis distance

- Identifying multivariate outliers – outliers that are distant from a combination of scores
- A point can be a multivariate outlier even if it is not a univariate outlier



https://commons.wikimedia.org/wiki/File:Outlier_statistics.svg

Outliers – what should we do?

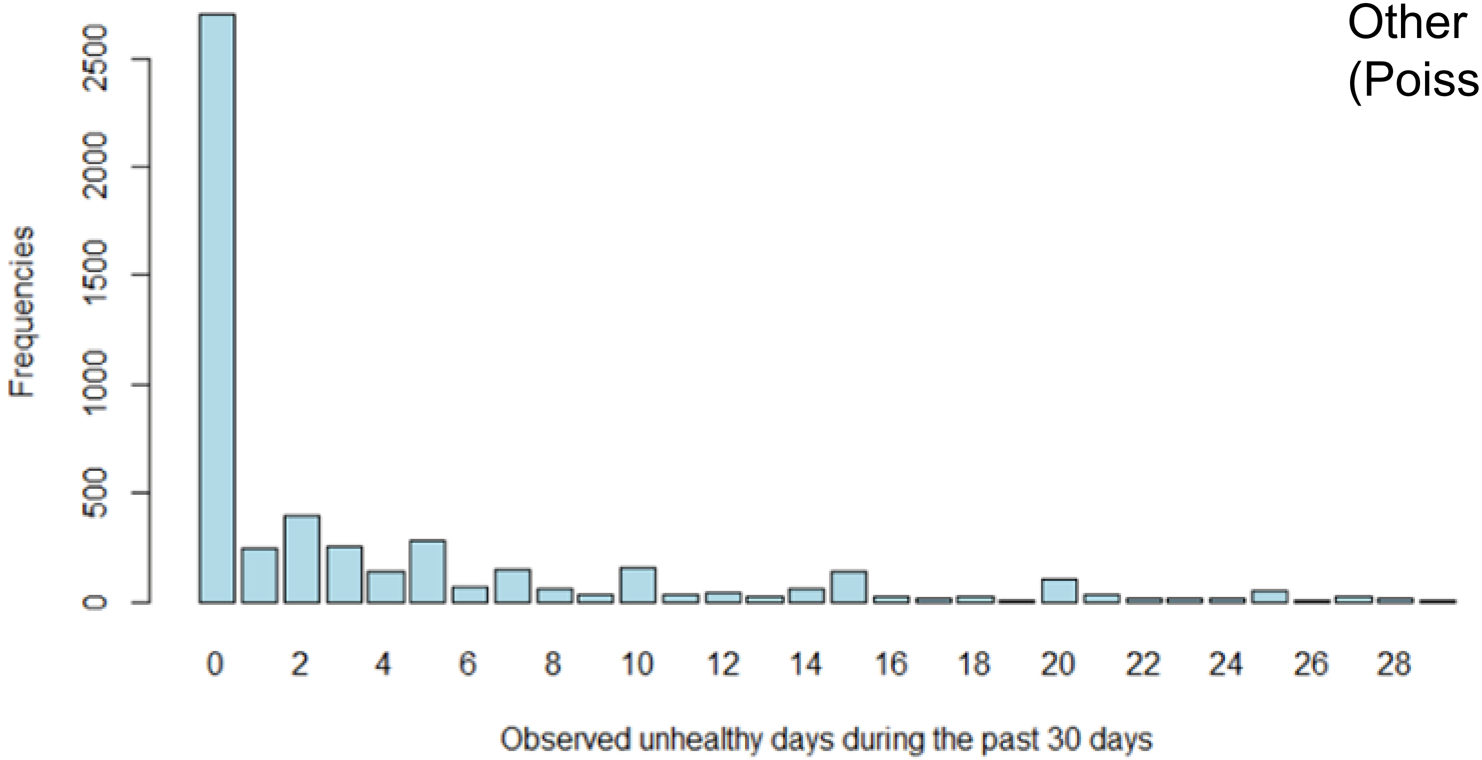
- Errors in data entry – need to fix
- Extreme values
  - Remove?
  - Keep in?
  - Substitute?
- Depends on the type of data

Outliers?

Remove only unlikely values

Variables can be transformed

Other analytic techniques can be used (Poisson regression)



Yang, S., Puggioni, G., Harlow, L. L., & Redding, C. A. (2017). A Comparison of Different Methods of Zero-Inflated Data Analysis and an Application in Health Surveys. *JMASM Editors*, *16*(1), 518-543.