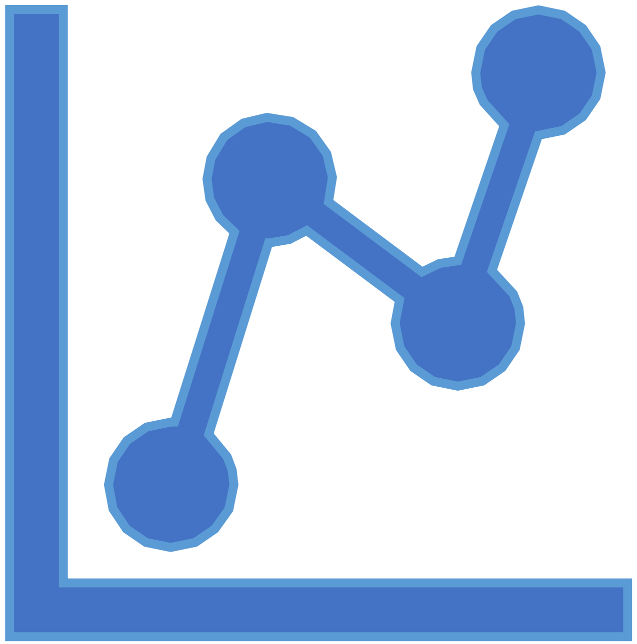
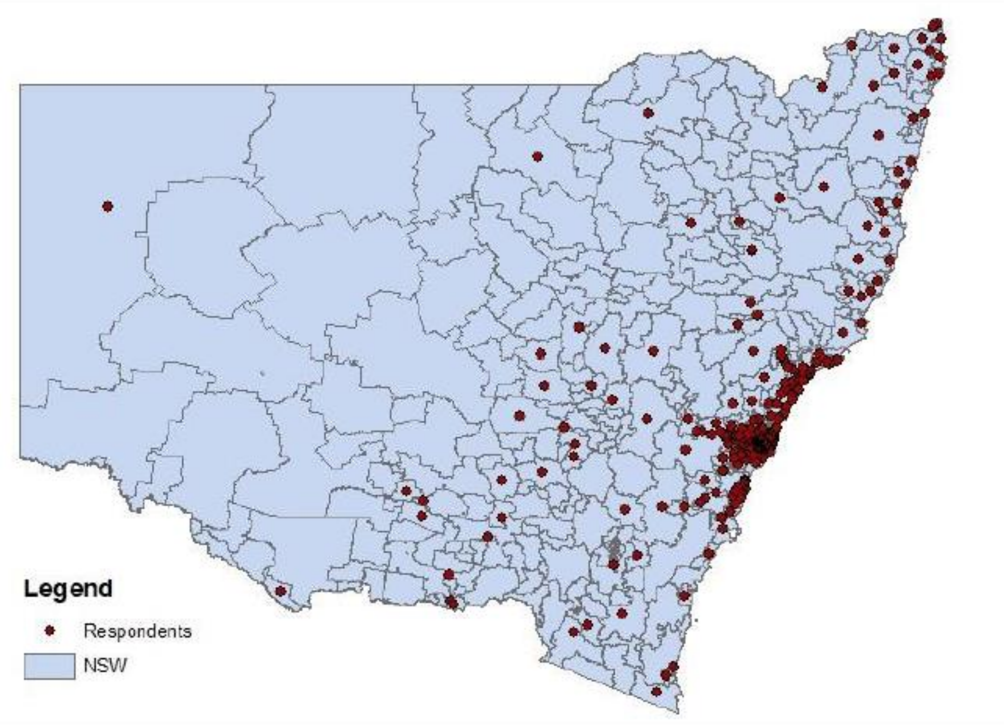


Lecture 1:
Descriptive Statistics

Introduction

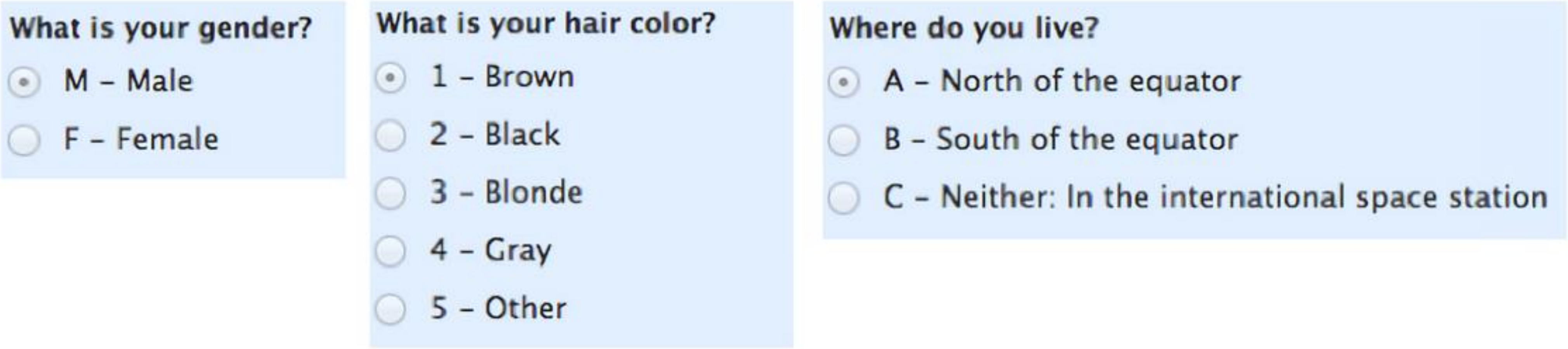
- **Descriptive statistics**: "summarises and describes the important characteristics of a set of measurements"

- **Inferential statistics**: "make inferences about **population** characteristics from information contained in a **sample** drawn from this population"



Data types

- **Nominal:** labels, mutually exclusive, no numerical significance, may or may not have orders

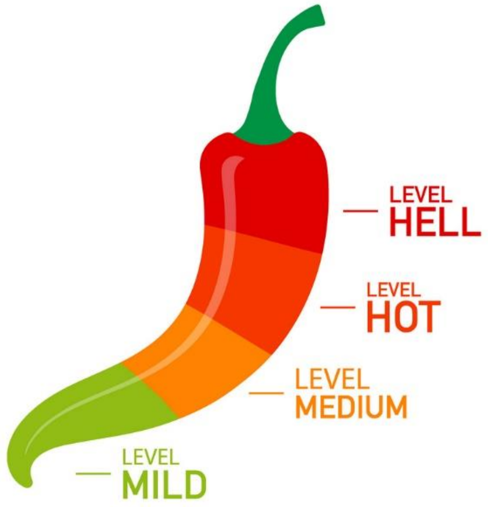


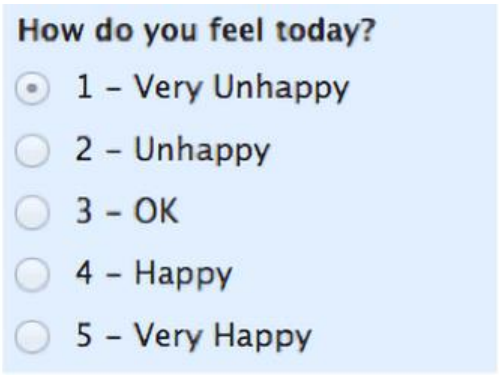
Source: https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/#targetText=Summary,the%20difference%20between%20each%20one.

Data types

- **Ordinal:** in order but the difference between variables not defined, e.g. Likert scales, time of day (morning, noon, evening), energy rating (1 star, 2 stars, 3 stars)

Likert scales – Very Happy is better (higher) than Happy. The difference between Very Happy and Happy doesn't make sense, and does not equal the difference between OK and Unhappy.





[Source: https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/#targetText=Summary,the%20difference%20between%20each%20one.](https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/#targetText=Summary,the%20difference%20between%20each%20one.)

Source: VectorStock.

Data types

- **Interval:** in order, difference between variables defined, but don't have a "true zero" and thus cannot be divided or multiplied, e.g. temperature, time on a clock, IQ score

Temperature - water from 20^{o} needs an increase of 80^{o} to 100^{o} to boil, but 0^{o} does not mean water has **no** temperature. Also, 80^{o} is not 4 times of 20^{o} because 0^{o} is not a starting/reference point.

- **Ratio**: like interval but with a "true zero", e.g. income, years of education, weight.

Source: https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/#targetText=Summary,the%20difference%20between%20each%20one.

https://www.statisticshowto.datasciencecentral.com/nominal-ordinal-interval-ratio/.

Data types – Practice Example

What is the type of these variables?

Features	Value set	Unit
Electric vehicle properties		
Vehicle type	Large sedan, Minivan, Small sedan, Large SUV, Small SUV, Small hatchback	
Range	120, 180, 240, 300, 360, 420, 480, 540	km
Recharge time	0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5	hours
Set up cost	1000, 1750, 2500, 3250	Dollars
Cost per km	3, 6, 9, 12	Cents
EV price	25000, 35000, 45000, 55000, 70000, 85000, 100000, 120000, 140000, 160000	Dollars
Governmental supports		
Charging station availability	5, 10, 15, 20	km
Bus lane access	Access to bus lane, No access to bus lane	
Rebates upfront costs	0, 3000, 6500, 10000	Dollars
Rebates parking fees	0, 100, 250, 400	Dollars
Energy bill discount	0, 25, 50, 75	Percent
Stamp duty discount	0, 5, 15, 25	Percent
Market penetration stage (in NSW)		
Percentage EV sold	1, 30, 60, 90	Percent

Features	Value set	Unit
Gender	Male, Female	
Annual gross household income	Continuous value	Dollars
Number of cars in household	0, 1, 2, more than 2	cars
Number of other driver licences in household	Continuous value	
Currently hold a driver licence	Yes, No	
Household type	Couple family with no children, Couple family with children, One parent family, Single person household, Group household, Other family	
Work status	Employed full time, Employed part time, Household duties, Retired, Student, Unemployed	

Data types (cont'd)

- **Time Series**: When a quantitative variable is recorded over time at equally spaced intervals (such as daily, weekly, monthly, quarterly, or yearly), the data set forms a **time series.**
- **Cross sectional data**: A cross-sectional study involves looking at data from a population at one specific point in time.
- **Panel data**: A panel data set (also longitudinal data) has both a cross-sectional and a time series dimension, where all cross-section units are observed during the whole time period.

Example: Which type of data that corresponds to each of the following statements?

- Data on daily sales volume, revenue, number of customers for the past month at each Highlands Coffee location in Ho Chi Minh City.
- Data on daily sales revenue and expenses over past 12 months at Crescent Mall Highlands Coffee location.
- Data on daily sales volume, revenue, number of customers for the past month at all Highlands Coffee locations in Ho Chi Minh City.
- Data on 2019 Christmas day sales revenue and expenses in all Highlands Coffee locations in Vietnam.

Measures of Centre

2 5 6 9 11
↑

2 5 $\boxed{6 \quad 9}$ 11 27
↑

- **Sample mean ($\bar{x}$):** $\bar{x} = \frac{\sum x_i}{n}$
  - What is the sample mean of [2, 9, 11, 5, 6, 27]?
  - What is the sample mean of [2, 9, 110, 5, 6, 27]?

- **Population mean (μ)**: usually unknown, estimated by $\bar{x}$

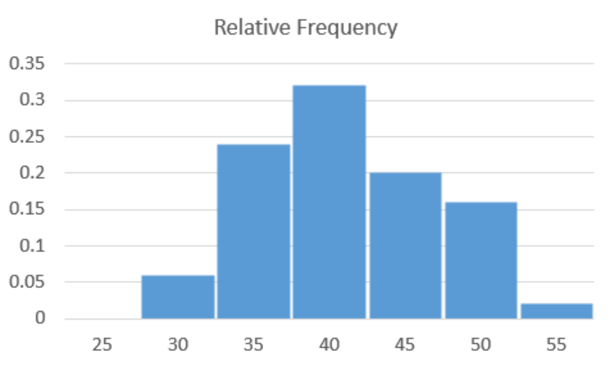
- **Median (m)**:
  - The value of x that falls in the middle position of an ordered sample
  - $\mathrm{m} = \boldsymbol{x_{0.5(n+1)}}$
  - What is the median of [2, 9, 110, 5, 6, 27]?

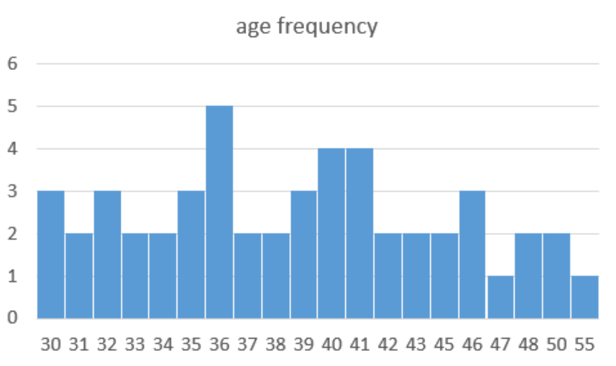
  -> Less sensitive to outliers

Measures of Centre

38	40	30	35	39	40	48	36	31	36
47	35	34	43	41	36	41	43	48	40
32	34	41	30	46	35	40	30	46	37
55	39	33	32	32	45	42	41	36	50
42	50	37	39	33	45	38	46	36	31

Bin	*Frequency*	*Relative Frequency*
25	0	0
30	3	0.06
35	12	0.24
40	16	0.32
45	10	0.2
50	8	0.16
55	1	0.02
Total	50	1





- **Mode:** "the category that occurs most frequently, or the most frequently occurring value of x"
- Relative frequency plot
  - Example: The ages (in months) at which 50 kids were first enrolled in a preschool
- Mode is generally used for large data sets, whereas mean and median can be used for any.

Measures of Variability

- **Range (R):** "the difference between the largest and smallest measurements"
- **Deviation:** difference between the sample mean and a measurement x_i, $x_i - \bar{x}$
- **Variance of a sample:** $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
- **Variance of a population:** $\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N}$
- **Standard deviation:** equals to square root of the variance

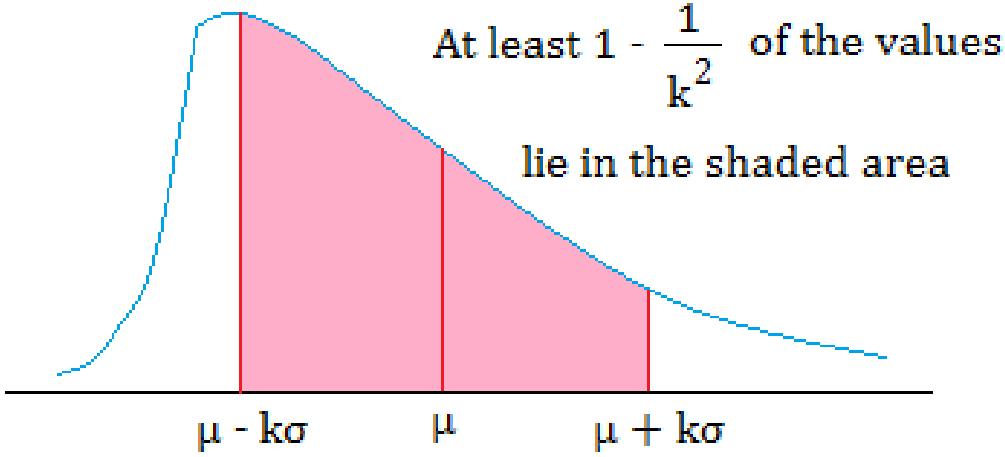
Measures of Centre and Measures of Variability

- **Practice Examples**
  - Calculate measures of centre and of variability of the 1985 Women’s Health Survey Data.
  - The Anscombe’s quartet dataset

Source: https://newonlinecourses.science.psu.edu/stat505/lesson/1/1.4

Tchebysheff's Theorem

- For **any** dataset
  - **At least** none of the measurements lie in the interval $\mu \pm \sigma$
  - **At least** 3/4 (75%) of the measurements lie in the interval $\mu \pm 2\sigma$
  - **At least** 8/9 (88.9%) of the measurements lie in the interval $\mu \pm 3\sigma$



Tchebysheff's Theorem

- Example: The ages (in months) at which 50 kids were first enrolled in a preschool

38	40	30	35	39	40	48	36	31	36
47	35	34	43	41	36	41	43	48	40
32	34	41	30	46	35	40	30	46	37
55	39	33	32	32	45	42	41	36	50
42	50	37	39	33	45	38	46	36	31

- Mean = 39.08 months, std = 5.99 months
  - Tchebysheff's theorem:

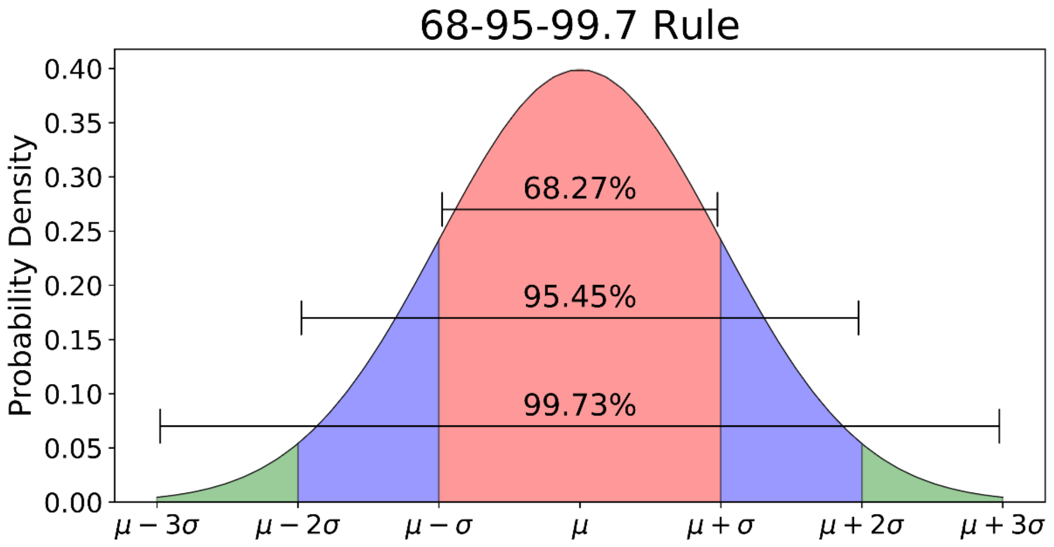
    At least ¾ of the kids (37.5 kids) are from 27.11 months to 51.05 months ($\mu \pm 2\sigma$)
  - Facts: 49 kids are from 33.09 months to 45.07 months.

  - Tchebysheff's theorem:

    At least 8/9 of the kids (44.4 kids) are from 21.12 months to 57.04 months ($\mu \pm 3\sigma$)
  - Facts: 50 kids are from 33.09 months to 45.07 months.

The Empirical Rule

- For an **approximately normal distribution** of measurements
  - 68% of the measurements lie in the interval $\mu \pm \sigma$
  - 95% of the measurements lie in the interval $\mu \pm 2\sigma$
  - 99.7% of the measurements lie in the interval $\mu \pm 3\sigma$

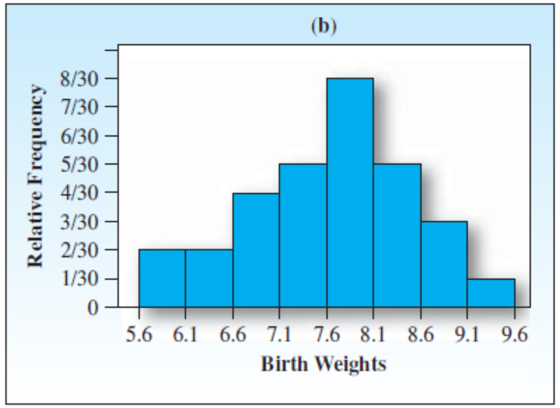


[Source: https://towardsdatascience.com/understanding-the-68-95-99-7-rule-for-a-normal-distribution-b7b7cbf760c2](https://towardsdatascience.com/understanding-the-68-95-99-7-rule-for-a-normal-distribution-b7b7cbf760c2)

The Empirical Rule

- Example: Birth weights (in pounds) of 30 full-term newborn babies

7.2	7.8	6.8	6.2	8.2
8.0	8.2	5.6	8.6	7.1
8.2	7.7	7.5	7.2	7.7
5.8	6.8	6.8	8.5	7.5
6.1	7.9	9.4	9.0	7.8
8.5	9.0	7.7	6.7	7.7



- Mean = 7.57 lbs, std = 0.95 lbs
  - *The Empirical Rule*:

    At least 68% of the babies (20.4 babies) are from 6.63 lbs to 8.52 lbs ($\mu \pm \sigma$)
  - Facts: 22 babies have weights between 6.63 lbs and 8.52 lbs.

  - *The Empirical Rule*:

    At least 95% of the babies (28.5 babies) are from 5.68 lbs to 9.47 lbs ($\mu \pm 2\sigma$)
  - Facts: 29 babies have weights between 5.68 lbs and 9.47 lbs.

Practice Examples

- Count the number of measurements in each variable within $\mu \pm 2\sigma$ in the 1985 Women's Health Survey Data
- Compare these counts with the Tchebysheff's Theorem and with the Empirical Rule.

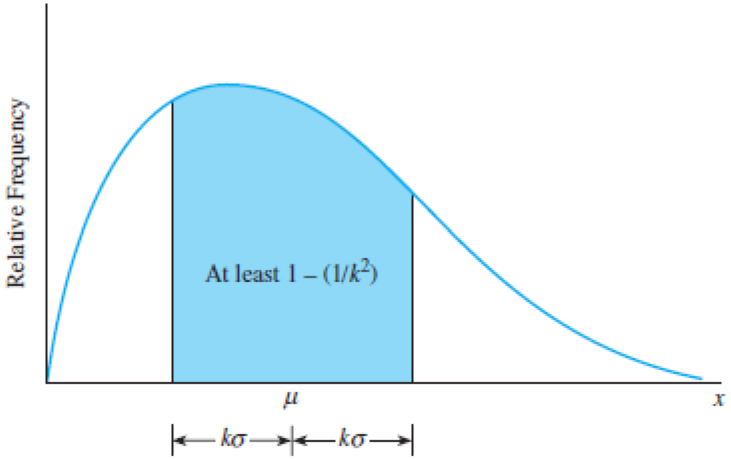
Source: https://newonlinecourses.science.psu.edu/stat505/lesson/1/1.4

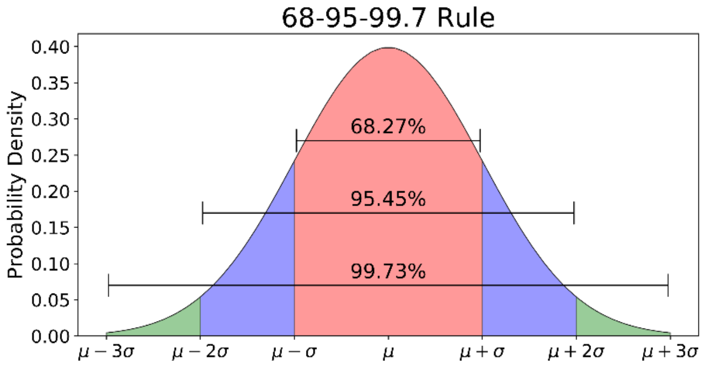
Measures of Relative Standing

- **Sample z-score**
  - “distance between an observation and the mean measured in units of standard deviation”

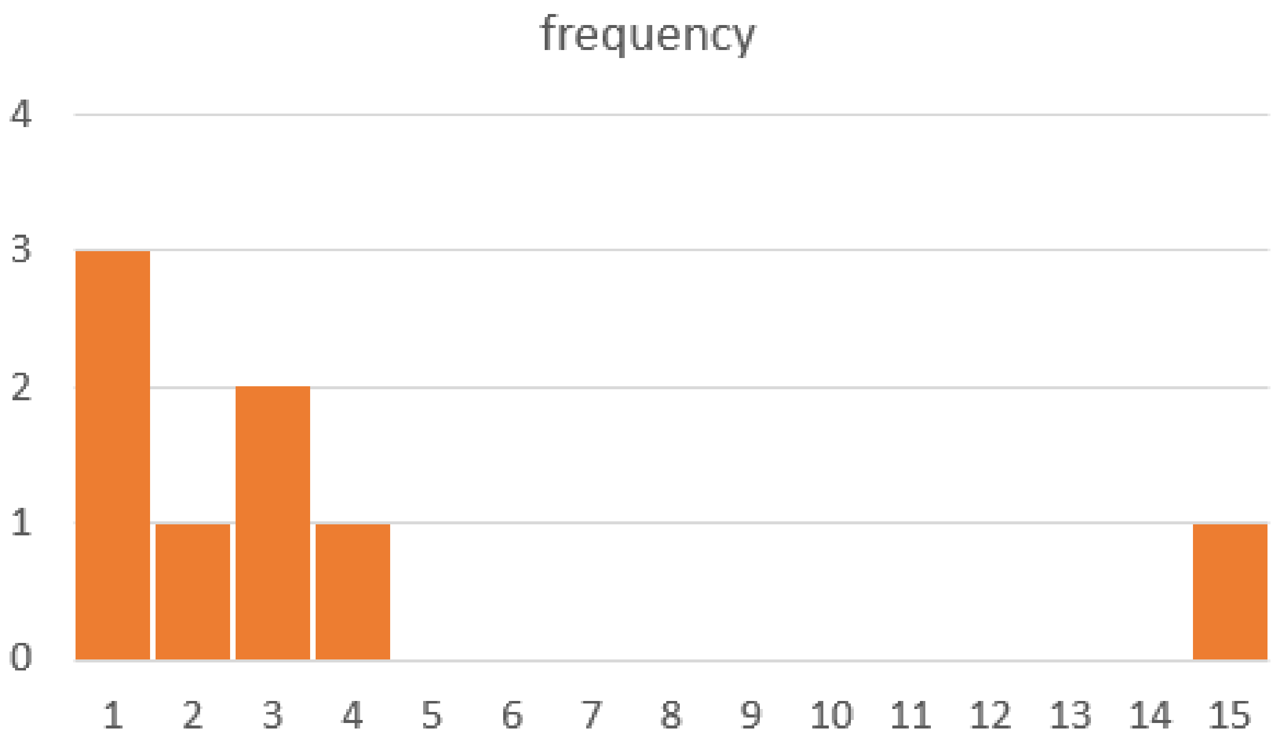
$$zscore = \frac{x - \bar{x}}{s}$$

  - A valuable tool in determining outliers. If z-score < -3 or z-score > 3 => outliers.





Measures of Relative Standing

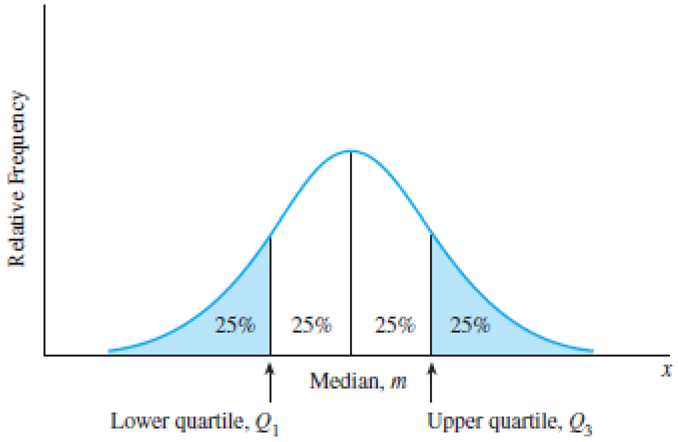


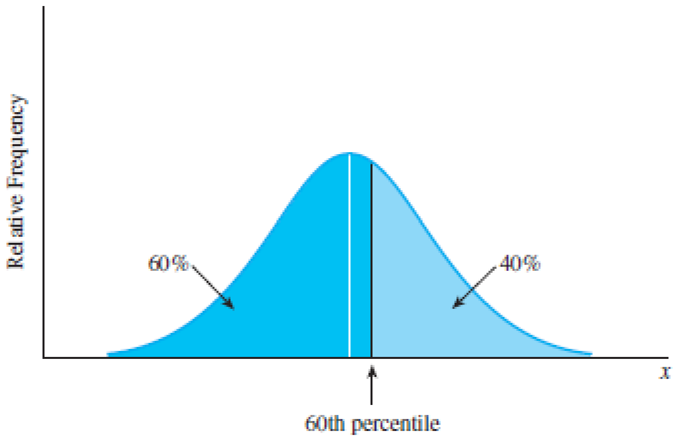
- Example: Calculate z-score of each observation for potential outliers in the list of measurements of [1, 1, 0, 15, 2, 3, 4, 0, 1, 3].
  - Mean = 3, std = 4.42
  - Z-score of x=15 is $\frac{15-3}{4.42} = 2.72$
  - 15 may be considered as an outlier

Measures of Relative Standing

- ***p*th percentile:** “the value of x that is greater than p% of the (ordered) measurements and is less than the remaining (100-p)%”
- Percentile of value x = (number of values less than x)/(number of values)*100

- **Lower quartile**, **upper quartile** and **interquartile range**





- **Q1 = .25(n+1)   Q3 = .75(n+1)**

Measures of Relative Standing

- Example: Consider the set of measurements [16, 25, 4, 18, 11, 13, 20, 8, 11, 9]
  - Sort the measurements [4, 8, 9, 11, 11, 13, 16, 18, 20, 25]
  - Value 18 is at 70^{th} percentile
  - Position of the 25^{th} percentile is 0.25*(10+1) = 2.75.

  Q1 value is therefore 8 + .75*(9-8) = 8.75

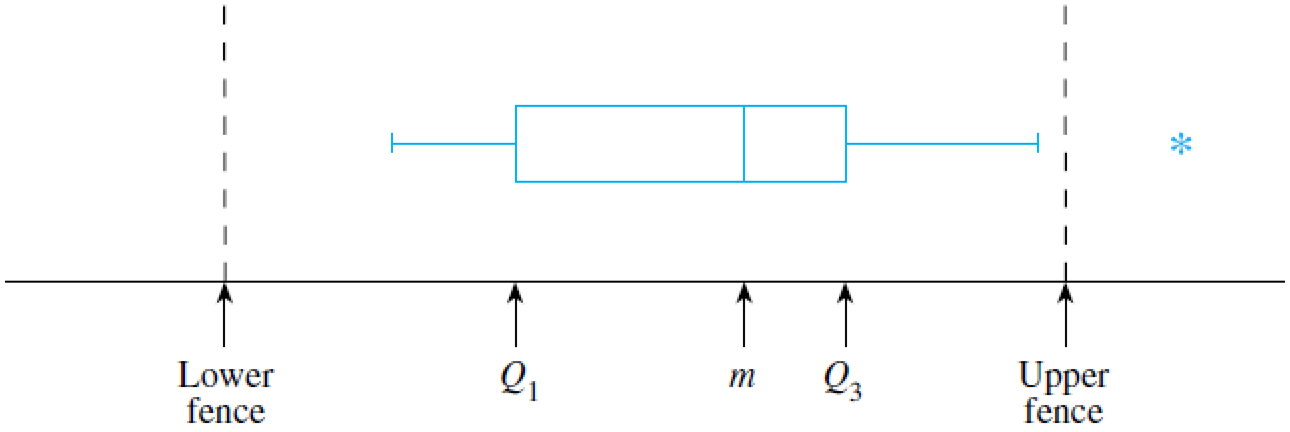
  - Position of the 75^{th} percentile is 0.75*(10+1) = 8.25.

  Q3 value is therefore 18 + .25(20-18) = 18.5

Note: Since these positions are not integers, the lower quartile is taken to be the value 3/4 of the distance between the second and third ordered measurements, and the upper quartile is taken to be the value 1/4 of the distance between the eighth and ninth ordered measurements.

The 5-number summary and Box Plots

- Five-number summary: Min, Q1, Median, Q3, Max
- A graphical tool "expressly designed" for isolating outliers from a sample.



- Lower fence = Q1 – 1.5(IQR)
- Upper fence = Q3 + 1.5(IQR)

The **interquartile range (IQR)** for a set of measurements is the difference between the upper and lower quartiles: IQR = Q_3 - Q_1.

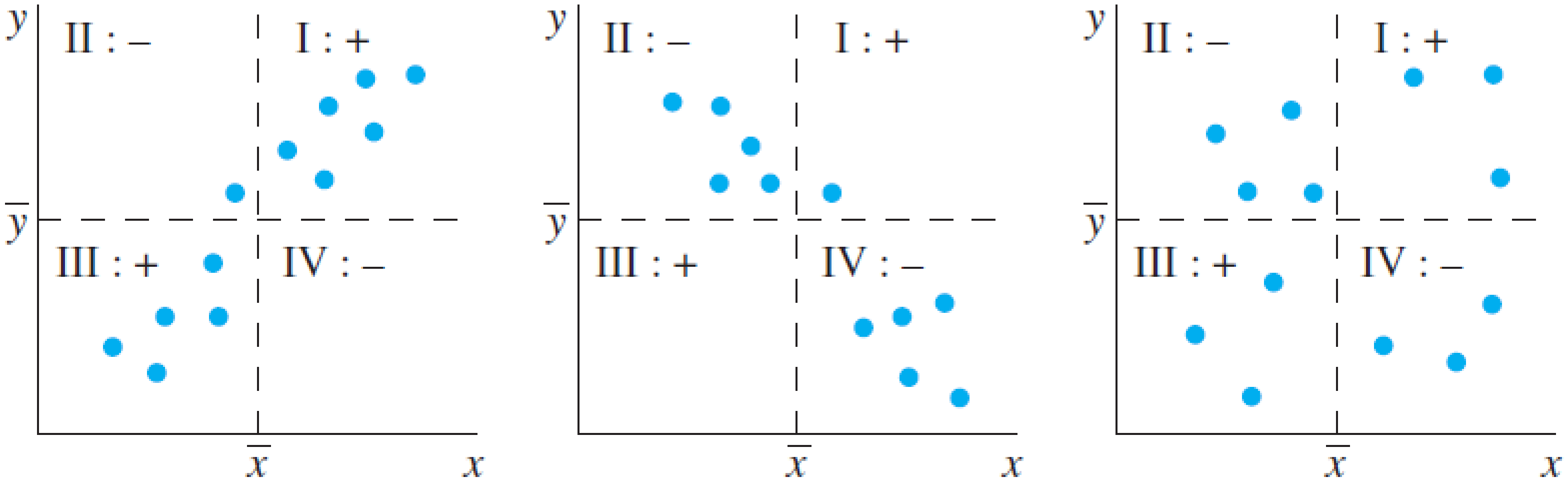
Practice Examples

- Produce a box plot of the 1985 Women's Health Survey Data in Excel.

Source: https://newonlinecourses.science.psu.edu/stat505/lesson/1/1.4

Describing Bivariate Data

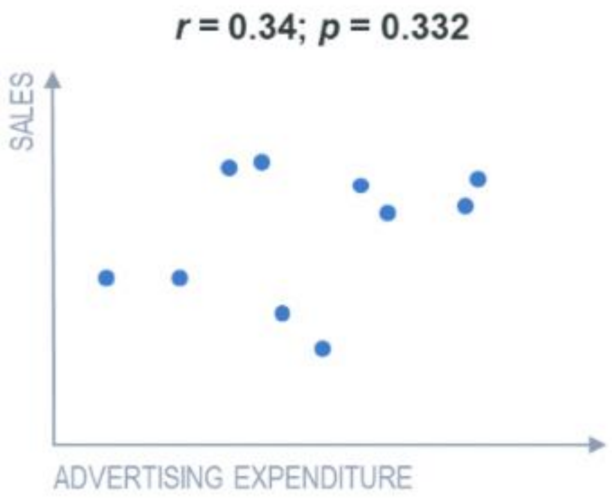
- Covariance between x and y in a bivariate sample, $s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
- Correlation coefficient, $r = \frac{s_{xy}}{s_x s_y}$

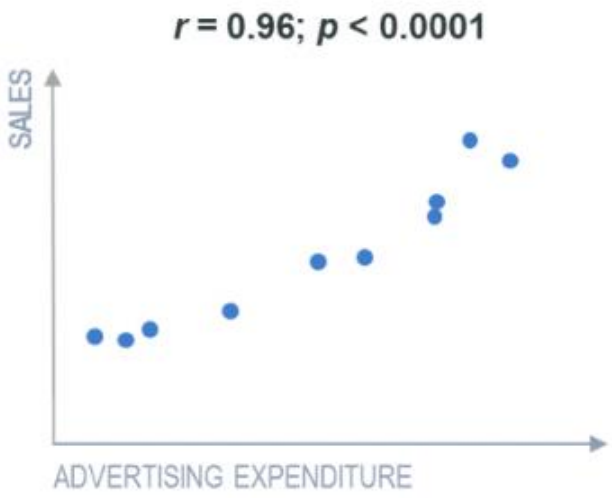


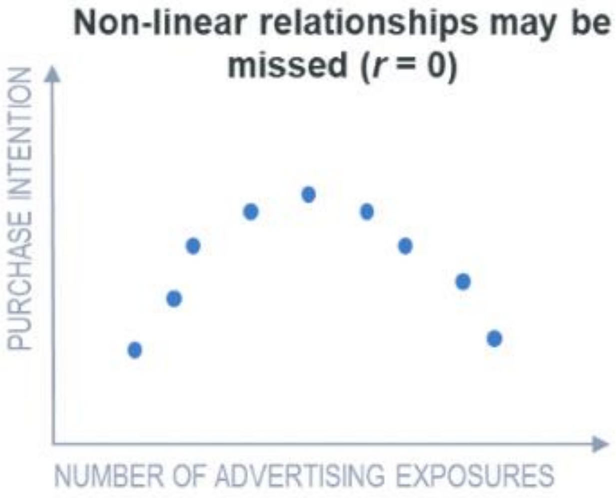
(a) Positive pattern (b) Negative pattern (c) No pattern

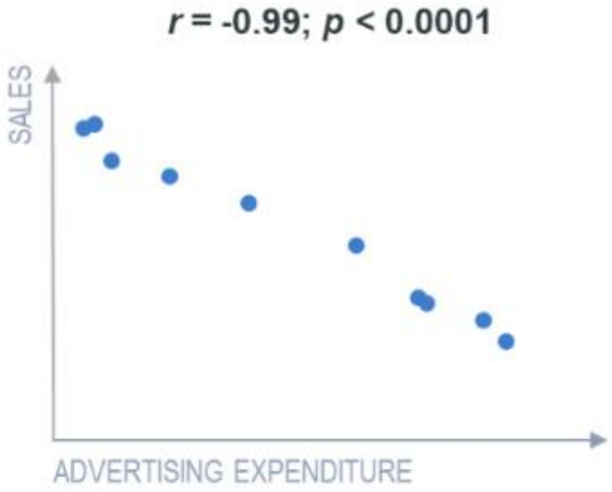
Describing Bivariate Data

- Correlation coefficient $-1 \leq r \leq 1$, indicating the strength of the correlation
- $r = 1$: perfect positive correlation
- $r = -1$: perfect negative correlation
- $r = 0$: no correlation between x and y (?)









Source: https://www.displayr.com/what-is-correlation/

Practice Examples

- Calculate covariance and correlation coefficients for each pair of variables in the USDA Women's Health Survey.

Source: https://newonlinecourses.science.psu.edu/stat505/lesson/1/1.4

Review

- Descriptive statistics and inferential statistics
- Sample vs Population
- Data types: nominal, ordinal, interval, ratio
- Measure of Centre: Mean, Median, Mode
- Measure of Variability: Range, Deviation, Variance, Standard Deviation
- Tchebysheff's Theorem, the Empirical Rule, and outlier detection
- Measures of relative standing: p^{th} percentile, quartiles, interquartile range
- Box plots
- Describing bivariate data: covariance and correlation coefficient