STAT 010 : Introduction to Statistics

— MONIKA BALODA

Chapter 1 Introduction to Data

1.1. Case Study : using stents to prevent strokes

1.2. Data basics
- Categorical Data : Ordinal / Nominal
- Numerical Data : Discrete / continuous.

1.3. Sampling principles and strategies.
- Population and Samples.
- Sampling methods.

1.4. Experiments.

Chapter 2 Summarizing Data

Part 1 : Numerical Data

- Plots - Scatterplots, Dot plots, Histograms, Maps
- Measure of center - Mean, Median, Mode
- Measure of Spread - Variance, Standard dev., IQR.
- Boxplots, outliers, Robust Statistics

Part 2: Categorical Data

- One variable plot – Bar plots, Pie charts.
- Contingency table
- Two variables plot – Side-by-side Bar plots, Stacked Bar Plots, Mosaic plots, side-by-side Box plots, Stacked Histogram.

Examples :–

2.6 Sleeping in college. A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night. Identify which value represents the sample mean and which value represents the claimed population mean.

2.6 Population mean = 5.5

Sample mean = 6.25

2.15 Distributions and appropriate statistics, Part I. For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Number of pets per household.

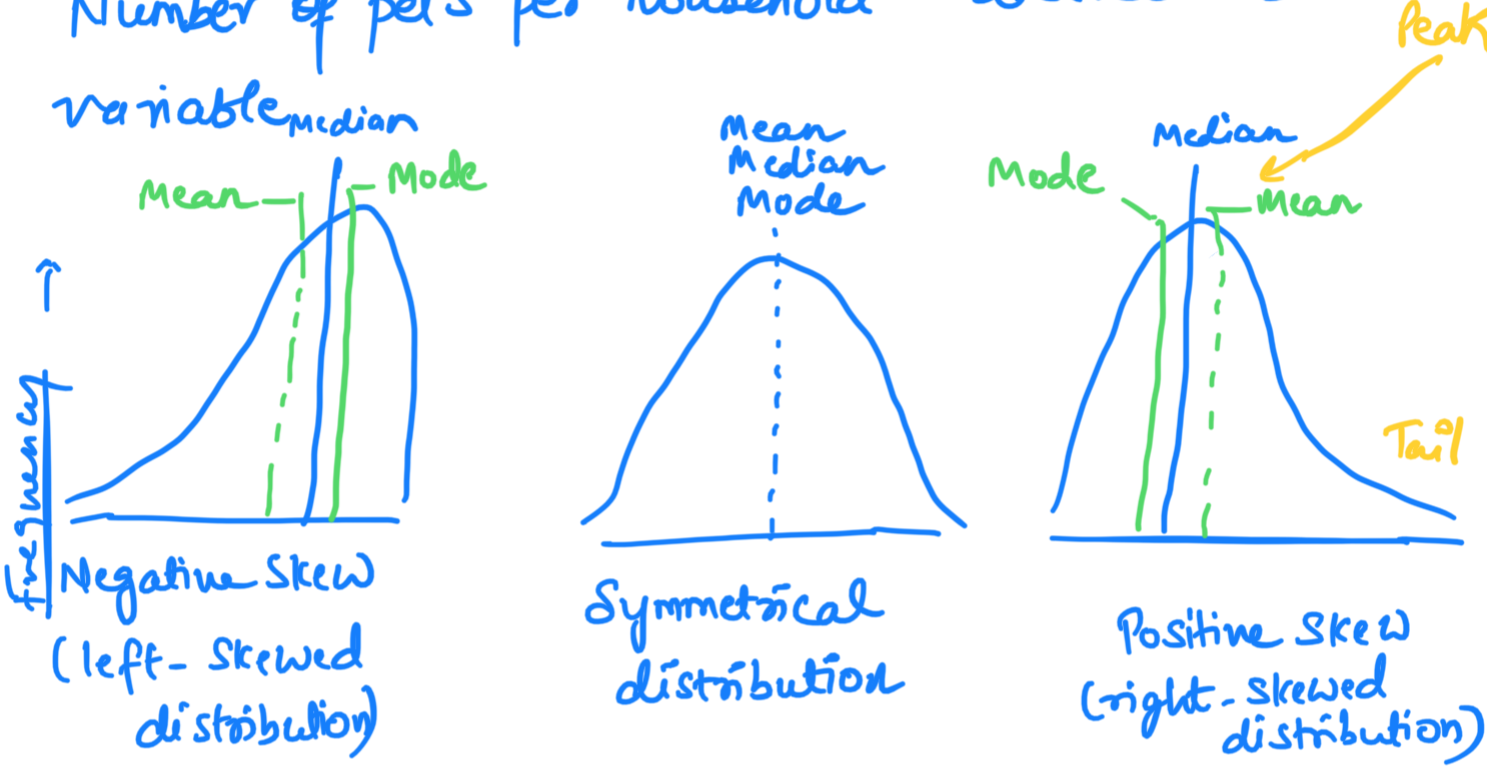
(b) Distance to work, i.e. number of miles between work and home.

(c) Heights of adult males.

2.15

a.) The distribution of number of pets per household is likely right skewed as there is a natural boundary at 0 and only a few people have many pets. Therefore the center would be best described by the median, and variability would be best described by the IQR.

Number of pets per household - discrete random variable

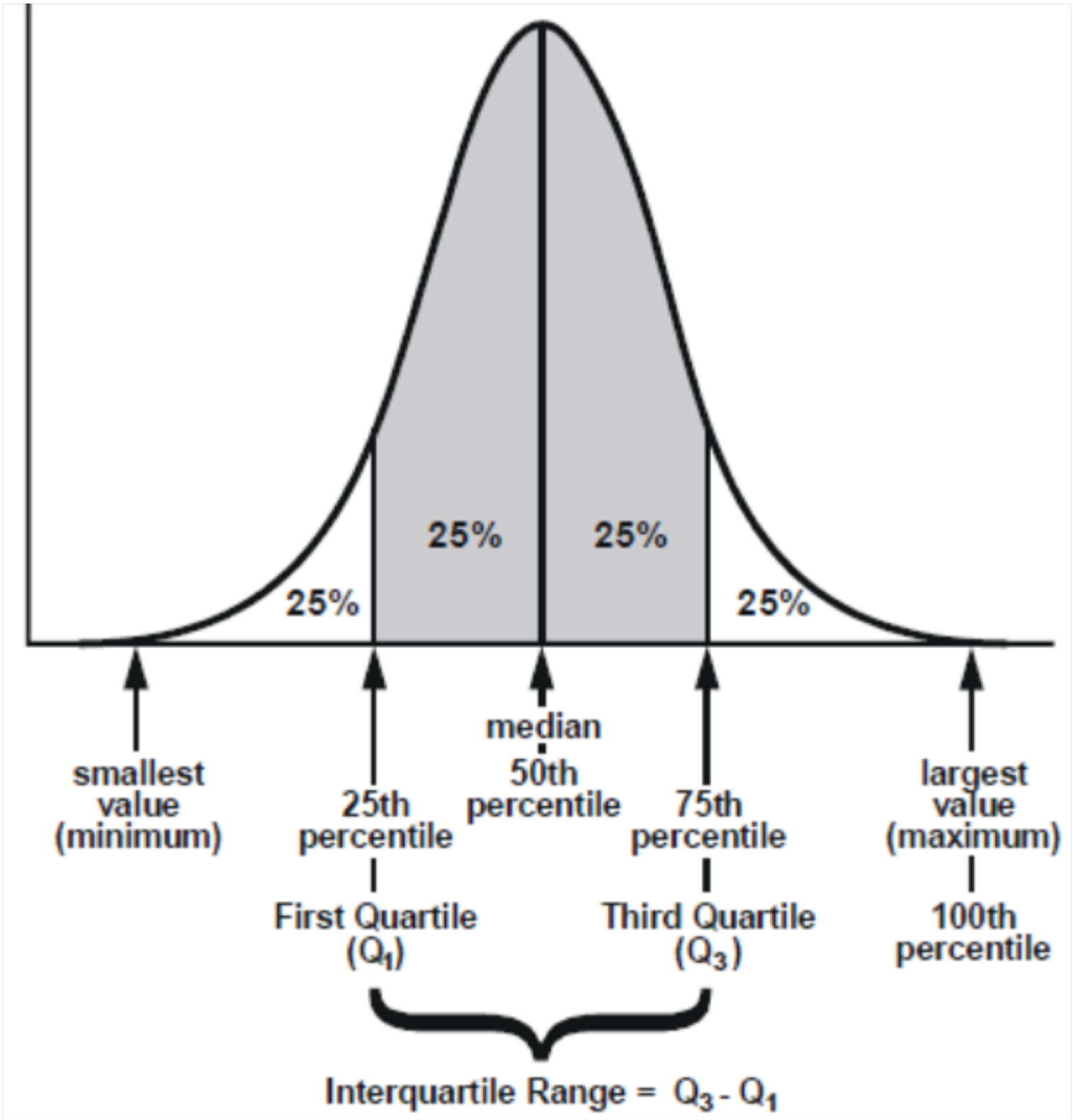


mean < mode

mean > mode

b.) The distribution of number of distance to work is likely right skewed as there is a natural boundary at 0 and only a few people live a very long distance from work. Therefore the center would be best described by the median, and variability would be best described by IQR (not influenced by extreme value)

IQR → Range of values that resides in the middle of of the scores (difference between the 75th (Q_3) and 25th (Q_1) percentile of data)


25%
25%
25%
25%
smallest value (minimum)
25th percentile
median 50th percentile
75th percentile
largest value (maximum)
First Quartile (Q_1)
Third Quartile (Q_3)
100th percentile
Interquartile Range = $Q_3 - Q_1$

c.) The distribution of heights of males is likely Symmetric. Therefore the center would be the best described by the mean, and variability would be best described by standard deviation. (influenced by extreme values)

2.30 A new statistic. The statistic $\frac{\bar{x}}{median}$ can be used as a measure of skewness. Suppose we have a distribution where all observations are greater than 0, $x_i > 0$. What is the expected shape of the distribution under the following conditions? Explain your reasoning.

(a) $\frac{\bar{x}}{median} = 1$

(b) $\frac{\bar{x}}{median} < 1$

(c) $\frac{\bar{x}}{median} > 1$

2.30) a.) If midrange = 1, then $\bar{x}$ = median. This is most likely to be the case for Symmetric distributions.

b.) If midrange < 1, then $\bar{x}$ < median. This is most likely to be the case for left skewed distributions, since mean is affected (and pulled down) by the lower values more so than the median.

c.) If midrange > 1, then $\bar{x}$ > median. This is most likely to be the case for right skewed distribution, since the mean is affected (and pulled up) by the higher values more so than the median.

Q→ A survey of licensed drivers asked whether they had received a speeding

ticket in the last year and whether their car is red. The results are shown in the contigency table.

	Speeding ticket	No speeding ticket	Total
Red car	15	135	150
Not Red car	45	470	515
Total	60	605	665

Find the probability that a randomly selected survey participant.

(a) has a red car.

$$P(\text{red car}) = \frac{150}{665} \approx 0.226 \text{ or } 22.6\%.$$

(b) has had a speeding ticket in last year.

$$P(\text{speeding ticket}) = \frac{60}{665} \approx 0.09 \text{ or } 9\%$$

(c) has had a speeding ticket in the last year given they have a red car.

$P(\text{Speeding ticket given red Gr}) = \frac{15}{150} = 0.10$ or 10%.

Q→ compare distributions based on means and standard deviation.

(a.) 1.) $-10, 0, 0, 0, 15, 25, 30, 30$

2.) $-20, 0, 0, 0, 15, 25, 30, 30$

Distribut^n 1) has higher mean since $-10 > -20$, and Dist 2 has a higher standard deviation since -20 is farther away from the rest of data than -10

(b) 1.) $4, 6, 6, 6, 9, 12, 12, 12, 14$

2.) $4, 6, 6, 6, 9, 12, 12, 12, 21$

Distribution 2) has higher mean since $21 > 14$, and a higher Standard deviation (spread ↑)

Since 21 is further from the rest of the data than 14.