DA503 Applied Statistics

Lecture 02

Descriptive Statistics

Fall 2021 – H . Sait Ölmez

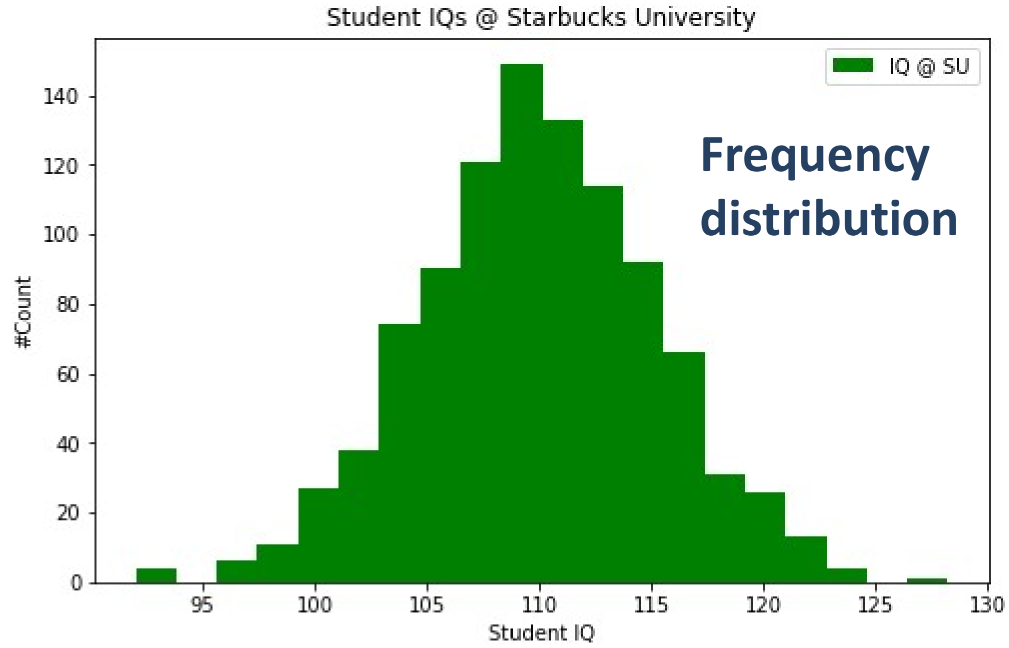
Agenda

- **Descriptive Statistics**
  - **Univariate Analysis**
    - Tabular representation of data and frequency distributions (histograms)
    - Relative and cumulative frequency distributions
    - Common shapes of frequency distributions
    - Measures of central tendency
      - Mean, mode and median
    - Measure of spread (quantifying variability)
      - Variance and standard deviation, range
    - Quartiles and percentiles
  - **Bivariate (multivariate) Analysis**
    - Relations between any combination of categorical and continuous variables: continuous-continuous, categorical-categorical and continuous-categorical

Tabular presentation of data and frequency

IQ	Freq.
90-92	1
92-94	3
94-96	0
. . .	. . .
106-108	125
108-110	168
110-112	146
. . .	. . .
128-130	1
Total	1000

Histogram is a graphical representation of the **frequency distribution** which shows the number of observations in each class



Frequency distribution

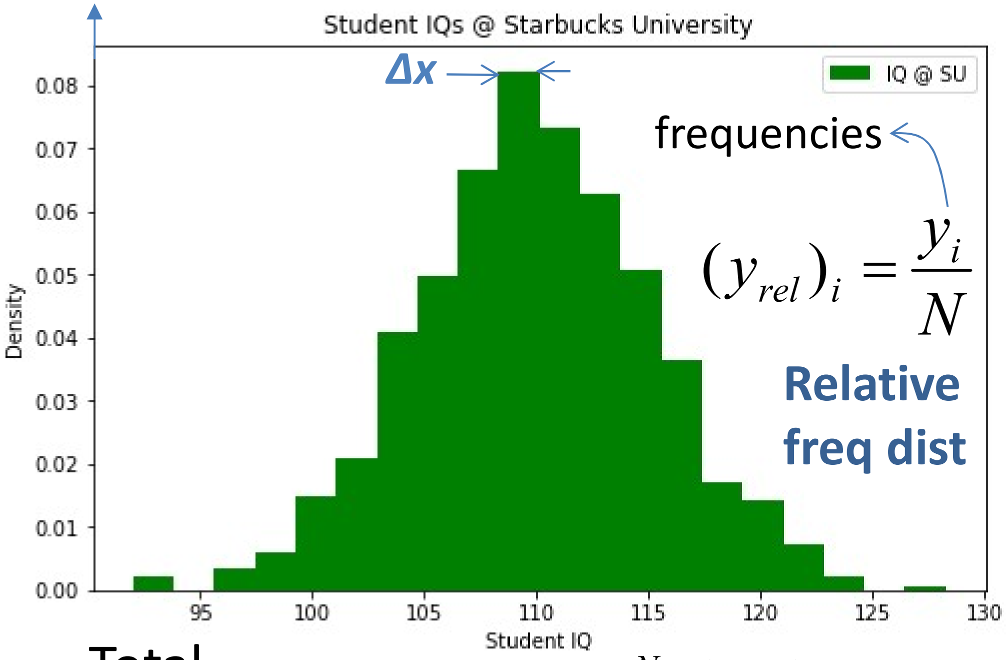
Python code

```
df = pd.read_csv('IQSU.txt',header=None)
IQ = df.iloc[:, 0]
plt.hist(IQ, bins=20, color='g', label=‘IQ @ SU')
plt.title(‘Student IQs @ Starbucks University')
plt.xlabel(‘Student IQ') ; plt.ylabel('#Count')
plt.legend(loc='upper right') ; plt.show()
```

Relative frequency

IQ	Freq	Relative freq
90-92	1	0.001
92-94	3	0.003
94-96	0	0.000
. . .	. . .	. . .
106-108	125	0.125
108-110	168	0.168
110-112	146	0.146
. . .	. . .	. . .
128-130	1	0.001
Total	**1000**	**1.000**

(y_{rel} = normalized count)



Δx

frequencies

$$(y_{rel})_i = \frac{y_i}{N}$$

Relative freq dist

Total normalized count = $$\sum_{i=1}^{N} \frac{(y_{rel})_i}{N} = 1$$

```
plt.hist(IQ, bins=20, normed=True, color='g', label='IQ @ SU')
```

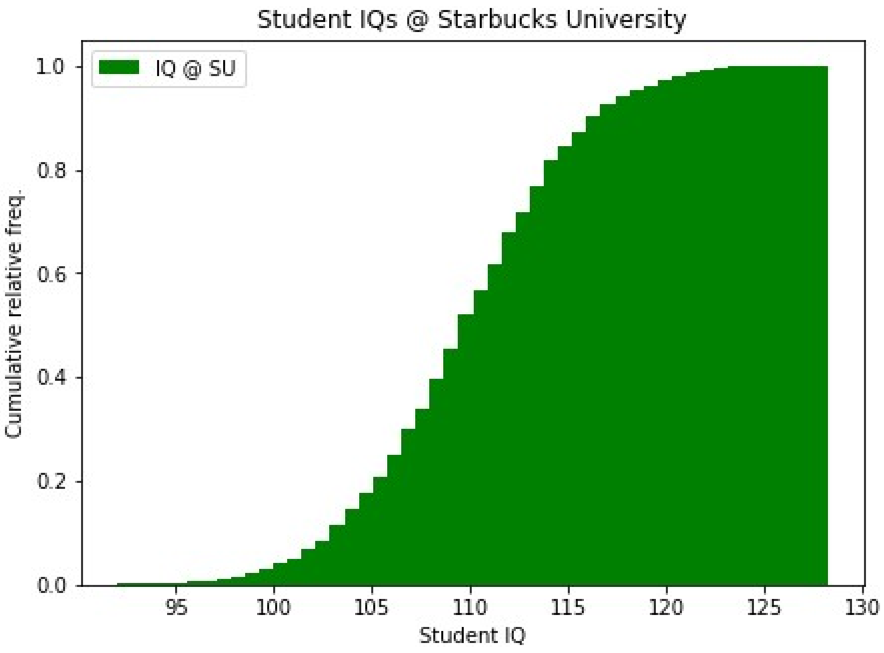
Python code

Density histogram (discrete): y axis is density value (normalized count divided by bin width) such that the bar areas sum to 1.

Cumulative relative frequency

- Accumulation of the previous relative frequencies

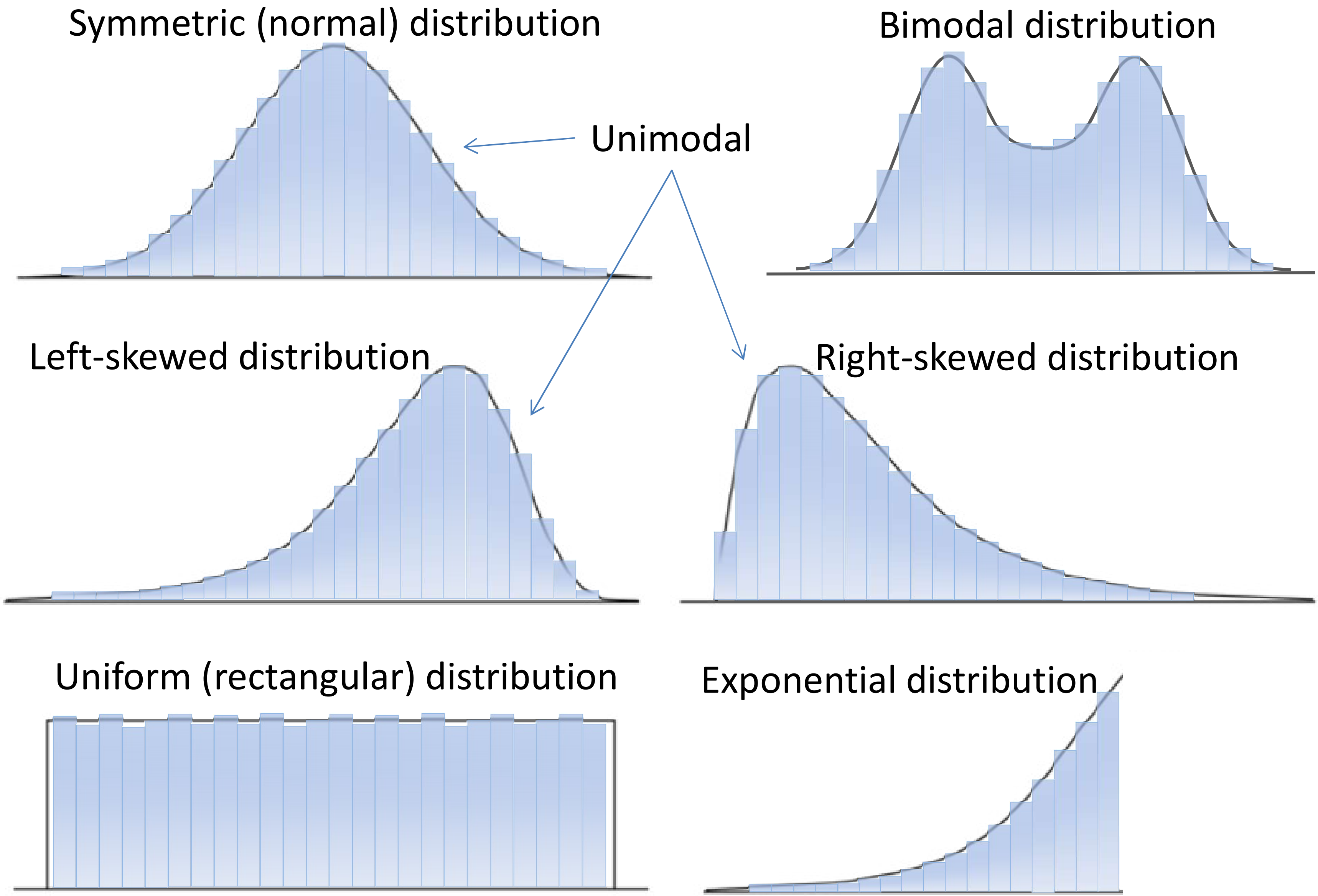
IQ	Freq	Relative freq	Cumulative rel. freq
90-92	1	0.001	0.001
92-94	3	0.003	0.004
94-96	0	0.000	0.004
. . .	. . .	. . .	. . .
106-108	125	0.125	0.346
108-110	168	0.168	0.514
110-112	146	0.146	0.660
. . .	. . .	. . .	. . .
128-130	1	0.001	1.00000
Total	**1000**	**1.000**	



```
plt.hist(height, bins=50, normed=True, cumulative=1,
color='g', label=‘IQ @ SU')
```

Python code

Common shapes of frequency distributions



Symmetric (normal) distribution

Bimodal distribution

Unimodal

Left-skewed distribution

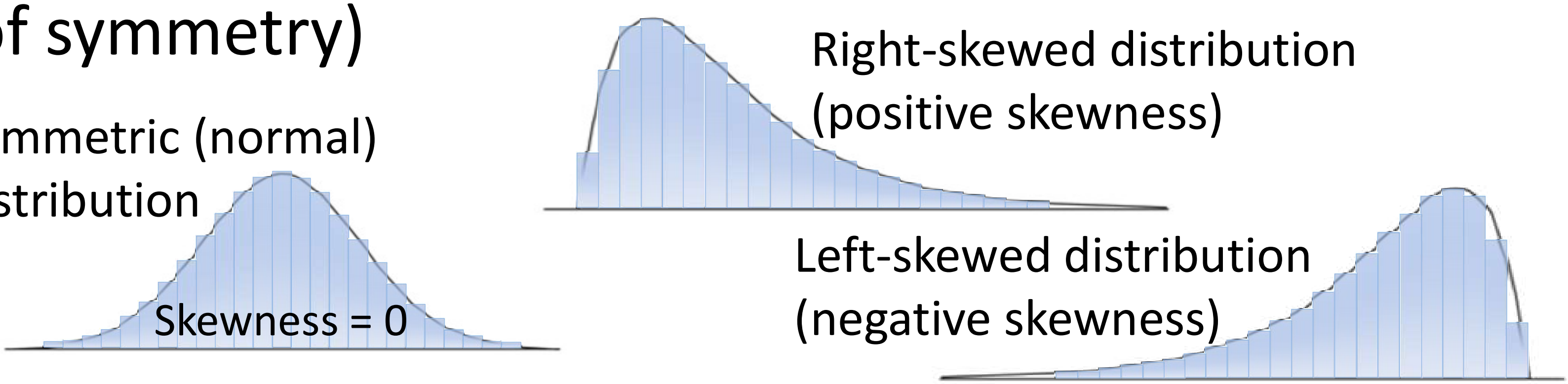
Right-skewed distribution

Uniform (rectangular) distribution

Exponential distribution

Skewness

- **Skewness**: A measure of dataset’s symmetry (or lack of symmetry)



Symmetric (normal) distribution

Skewness = 0

Right-skewed distribution (positive skewness)

Left-skewed distribution (negative skewness)

- A symmetrical distribution has a 0 skewness
- A general rule of thumb for skewness:
  - If < −1 or > +1, **highly skewed**.
  - If between (−1 , −½) or (+½ , +1), **moderately skewed**.
  - If between −½ and +½, **approximately symmetric**.

Population

$$\frac{1}{n}\frac{\sum_{i=1}^{n}(X_i-\bar{X})^3}{\sigma^3}$$

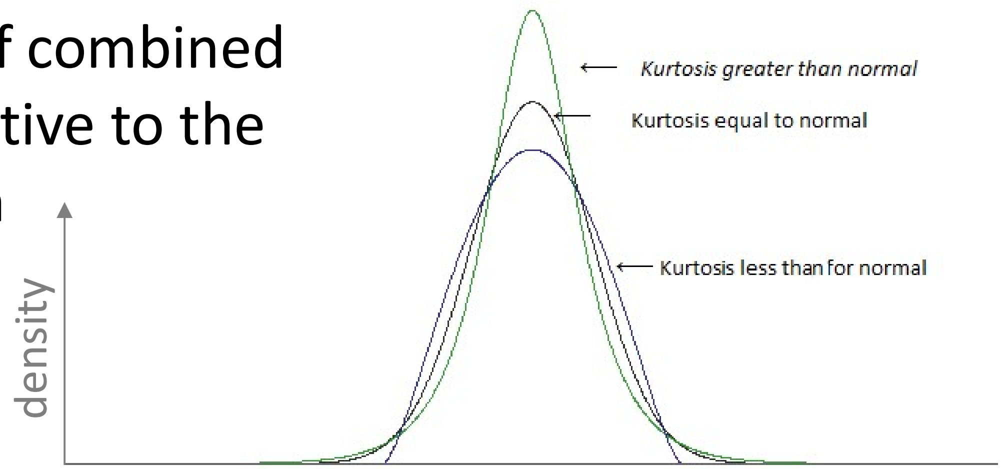
Sample

$$\frac{n}{(n-1)(n-2)}\frac{\sum_{i=1}^{n}(X_i-\bar{X})^3}{s^3}$$

```
import scipy.stats as stats
stats.skew(IQ)
-0.045182690
```

Kurtosis

- **Kurtosis**: A measure of combined weight of the tails relative to the rest of the distribution



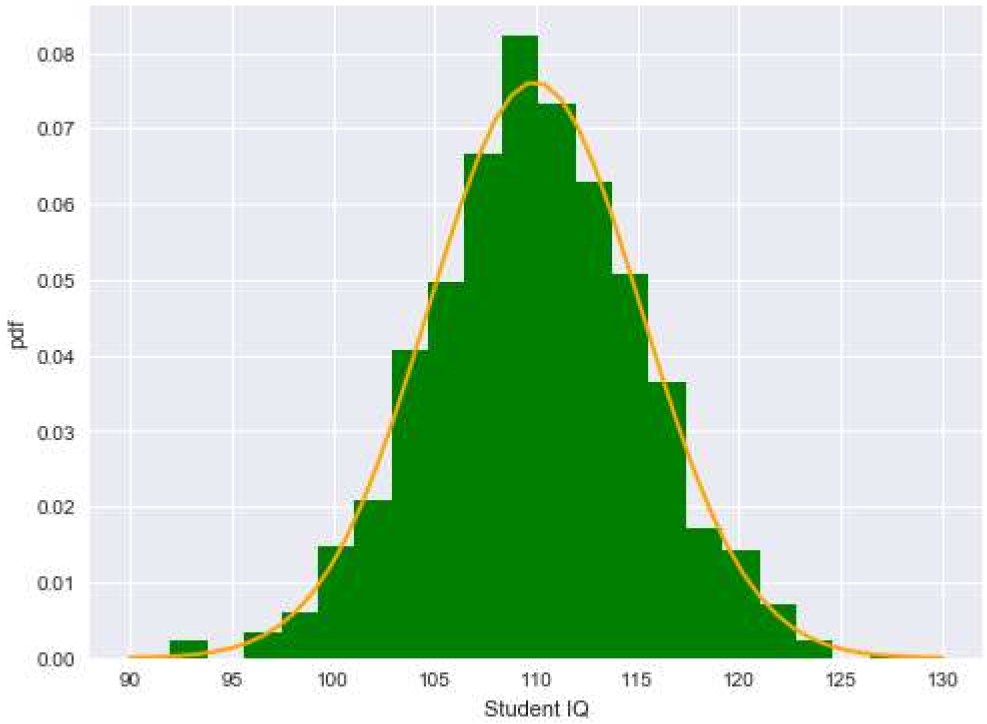
- A standard normal distribution has 3 kurtosis. A rule of thumb:
  - Kurtosis = 3 : **mesokurtic** (normal dist)
  - Kurtosis < 3 : **platykurtic** (compared to a normal dist., tails are shorter and thinner, and often central peak is lower and broader)
  - Kurtosis > 3 : **leptokurtic** (compared to a normal dist., tails are longer and fatter, and often its central peak is higher and sharper.

Population

$$\frac{1}{n}\frac{\sum_{i=1}^{n}(X_i-\bar{X})^4}{\sigma^4}$$

```
import scipy.stats as stats
stats.kurtosis(IQ)+3
3.124896214
```

Back to our “IQ @ SU” problem



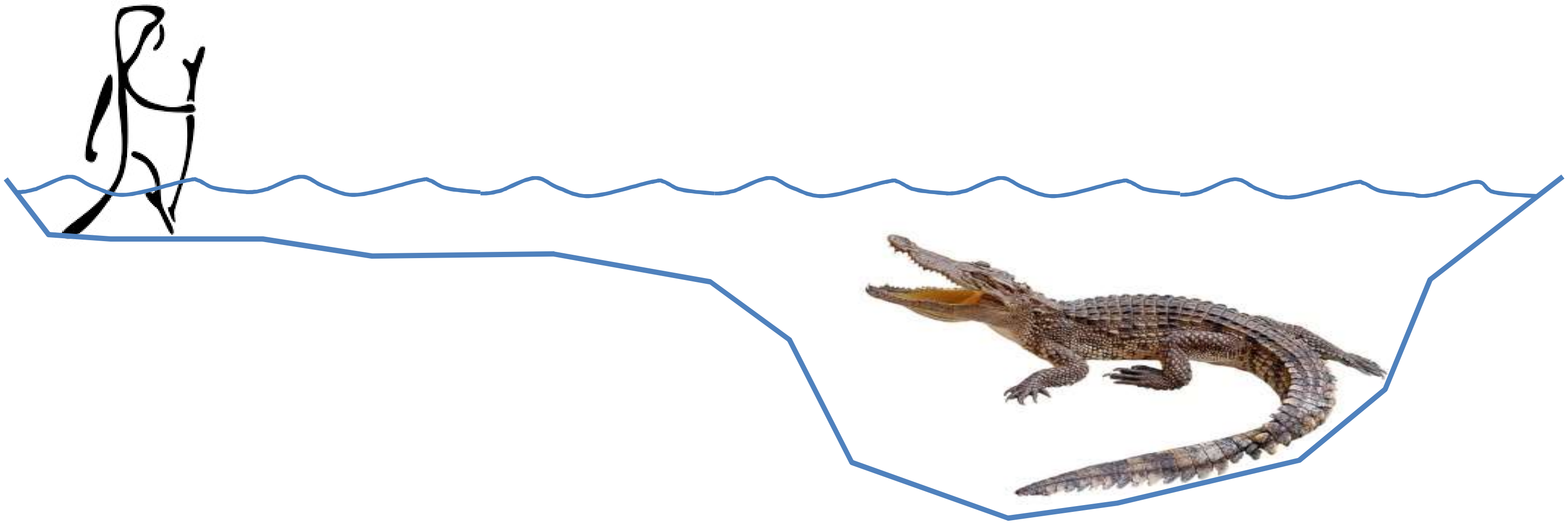
Python code

```
import scipy.stats as stats
meanH=IQ.mean() ; sdH=IQ.std()
plt.hist(IQ, bins=20 ,normed=True, facecolor='green')
rv = stats.norm(meanH,sdH)
x = np.linspace(90,130)
plt.plot(x, rv.pdf(x), color=‘orange’, lw=2)
plt.xlabel('Student IQ') ; plt.ylabel('pdf')
plt.show()
```

Measure of central tendency

"Never try to walk across a river just because it has an average depth of 120 cm"

M. Friedman

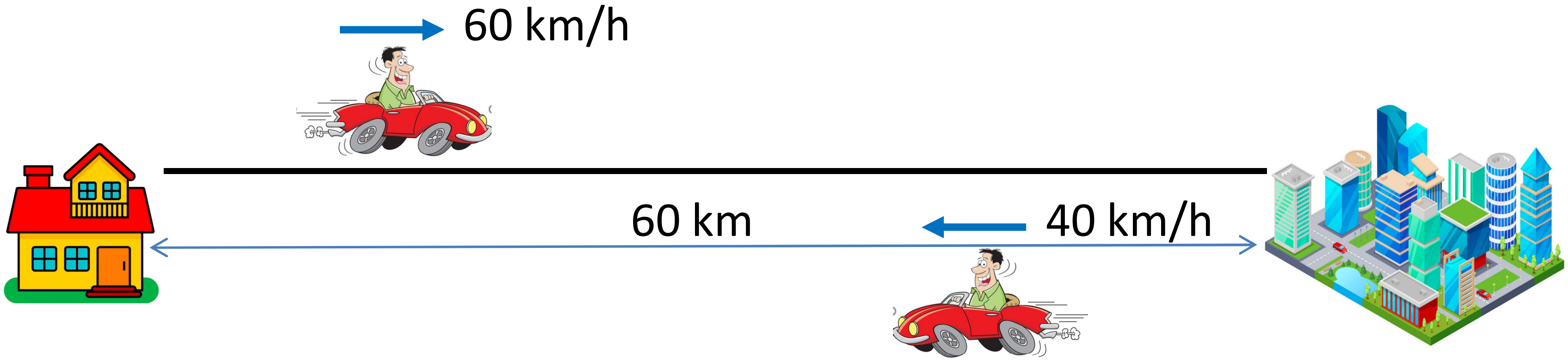


Measure of central tendency

- Given the data set: X = { 1, 3, 5, 5, 6, 7, 9, 11, 24 }

Mean: 7.9

- **Mean**: The average of the data set → $\overline{X} = (1/N)\sum_{i=1}^{N} X_i$
- How can you go wrong with the mean?



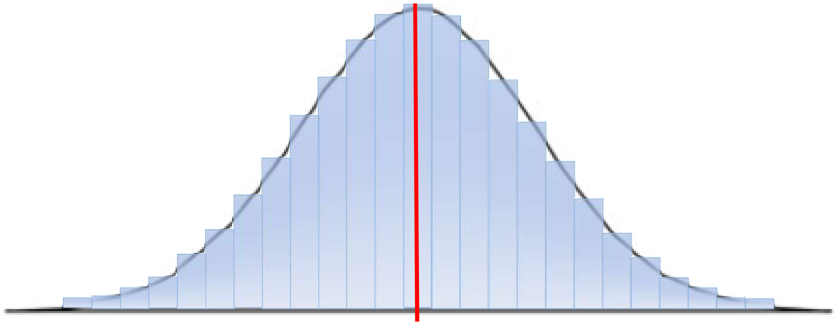
- What's the average speed? (60+40) / 2 = 50 km/h?
- Total driving time = (60/60) + (60/40) = 1 + 1.5 = 2.5 hrs
- Total distance traveled= 2 x 60 = 120 km
- Avg speed = 120 / 2.5 **= 48 km/h**

Measure of central tendency

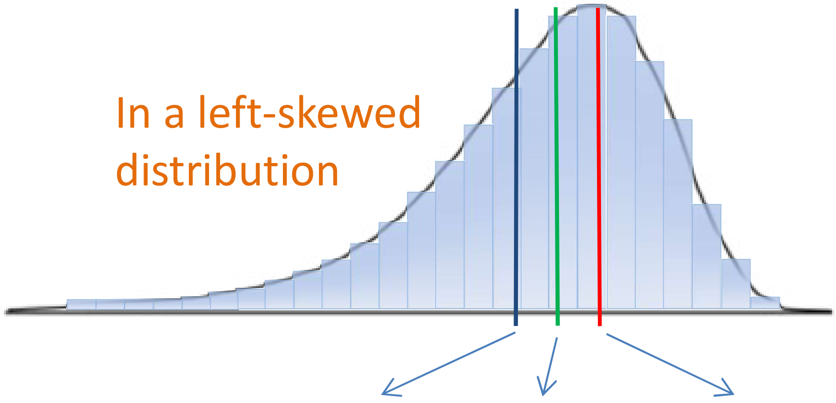
- Given the data set: X = { 1, 3, 5, 5, 6, 7, 9, 11, 24 }

Mode Median Mean: 7.9

- **Median**: Measure of the center of the data set (50th percentile)
- Mode: Point with the **highest** frequency
- Comparison of mean, median and mode:



Mean, Median, Mode



In a left-skewed distribution

Mean < Median < Mode

Mean vs Median vs Mode

- For skewed distributions, the mean is dragged in the direction of the skew. In such cases, the median is a much better representative of the central location of the data.
- Median is resistant to outliers as well. Here is an example:

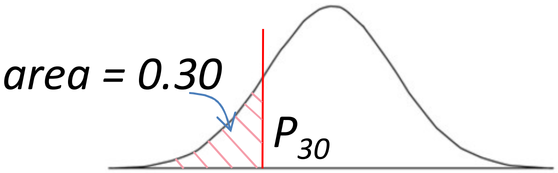
Data Set 1	Data Set 2
{4,4,5,5,5,6,6,6,7,7}	{4,4,5,5,5,6,6,6,7,7,300}
Mean = 5.5	Mean = **32.3**
Median = 5.5	Median = 6
σ = 1.08	σ = **88.8**

Type of variable	**Best measure of central tendency**
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

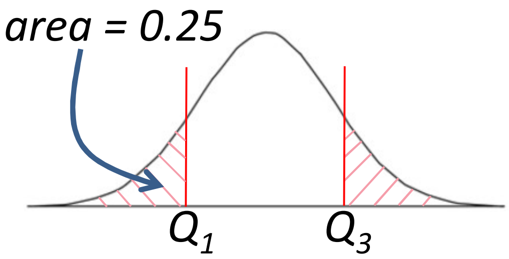
Quantiles

- **Quantile:** A portion of total number of observations. Quantiles are usually named according to the number of portions into which the range is divided.
- **Percentile**: divides a data set into 100 equal groups

The 30th percentile of a continuous distribution:

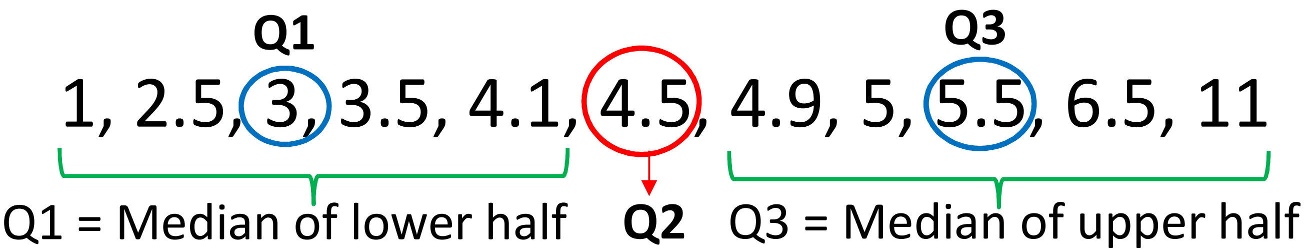


- **Decile:** divides a data set into 10 equal groups
- **Quintile:** divides a data set into 5 equal groups
- **Quartile**: divides a data set into 4 equal groups
  - Lower quartile (Q1) : 25th percentile
  - Middle quartile (Q2) : 50th percentile
  - Upper quartile (Q3) : 75th percentile
  - Interquartile range : IQR = $Q_3 - Q_1$



Boxplots

- A sample of N=11 observations:
- Dataset = { 1, 2.5, 3, 3.5, 4.1, 4.5, 4.9, 5, 5.5, 6.5, 11 }

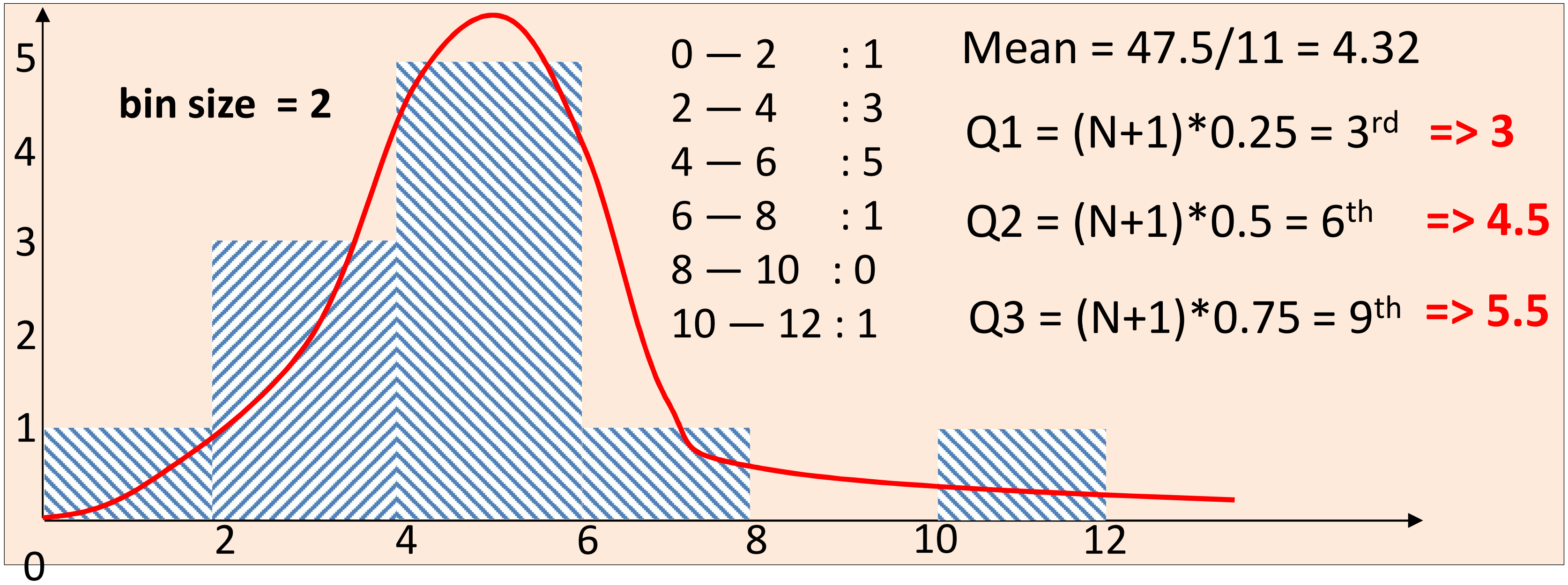


Q1 Q3

1, 2.5, 3, 3.5, 4.1, 4.5, 4.9, 5, 5.5, 6.5, 11

Q1 = Median of lower half **Q2** Q3 = Median of upper half

Histogram:



0 — 2	: 1
2 — 4	: 3
4 — 6	: 5
6 — 8	: 1
8 — 10	: 0
10 — 12	: 1

Mean = 47.5/11 = 4.32

Q1 = (N+1)*0.25 = 3rd => 3

Q2 = (N+1)*0.5 = 6th => 4.5

Q3 = (N+1)*0.75 = 9th => 5.5

Boxplots – cont'd

- **Boxplots (cont'd)**
- Dataset = { 1, 2.5, 3, 3.5, 4.1, 4.5, 4.9, 5, 5.5, 6.5, 11 }

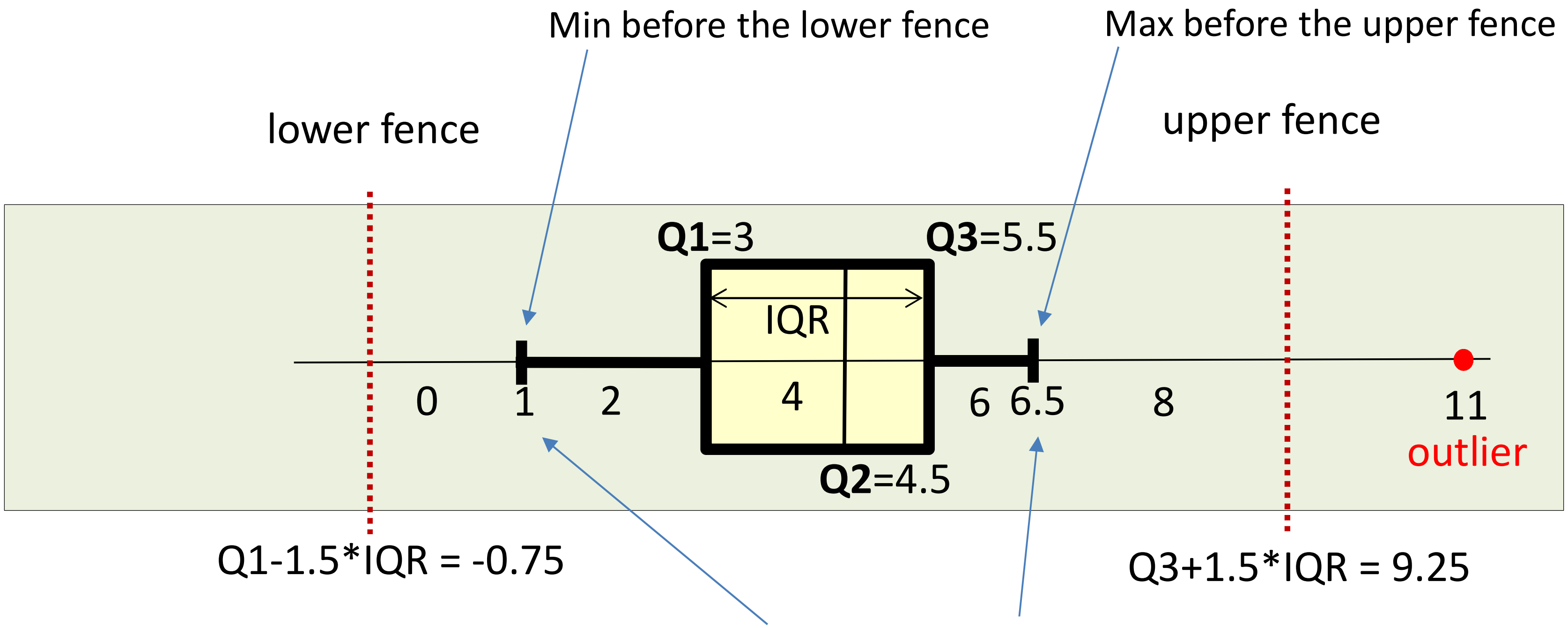
Q1 = 3

Q2 = 4.5

Q3 = 5.5

IQR = Q3-Q1 = 2.5

Inter**Q**uartile **R**ange



Min before the lower fence

Max before the upper fence

lower fence

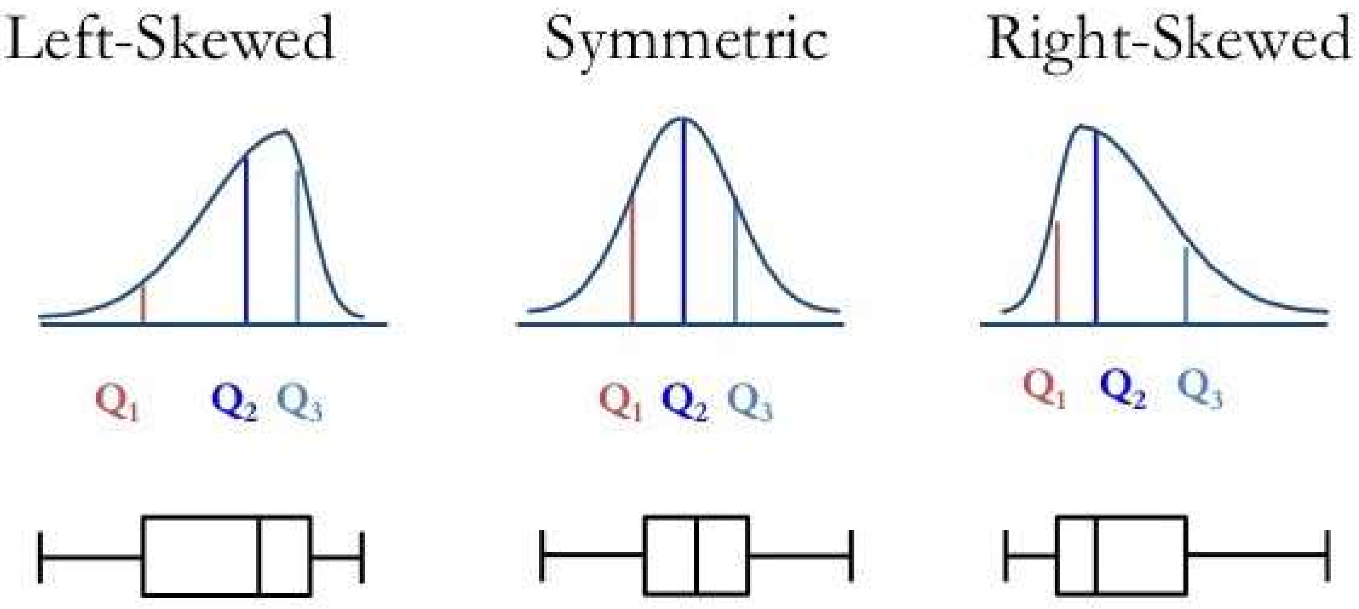
upper fence

Q1-1.5*IQR = -0.75

Q3+1.5*IQR = 9.25

The "**whiskers**" extend to the **smallest** and **largest** observations within the upper and lower fences

Interpretation of boxplots



Source: https://www.simplypsychology.org/box-plots-distribution.jpg

For **normal** distribution, the mean will be nearly the same as the median and the boxplot will look symmetric with equally long whiskers on the sides.

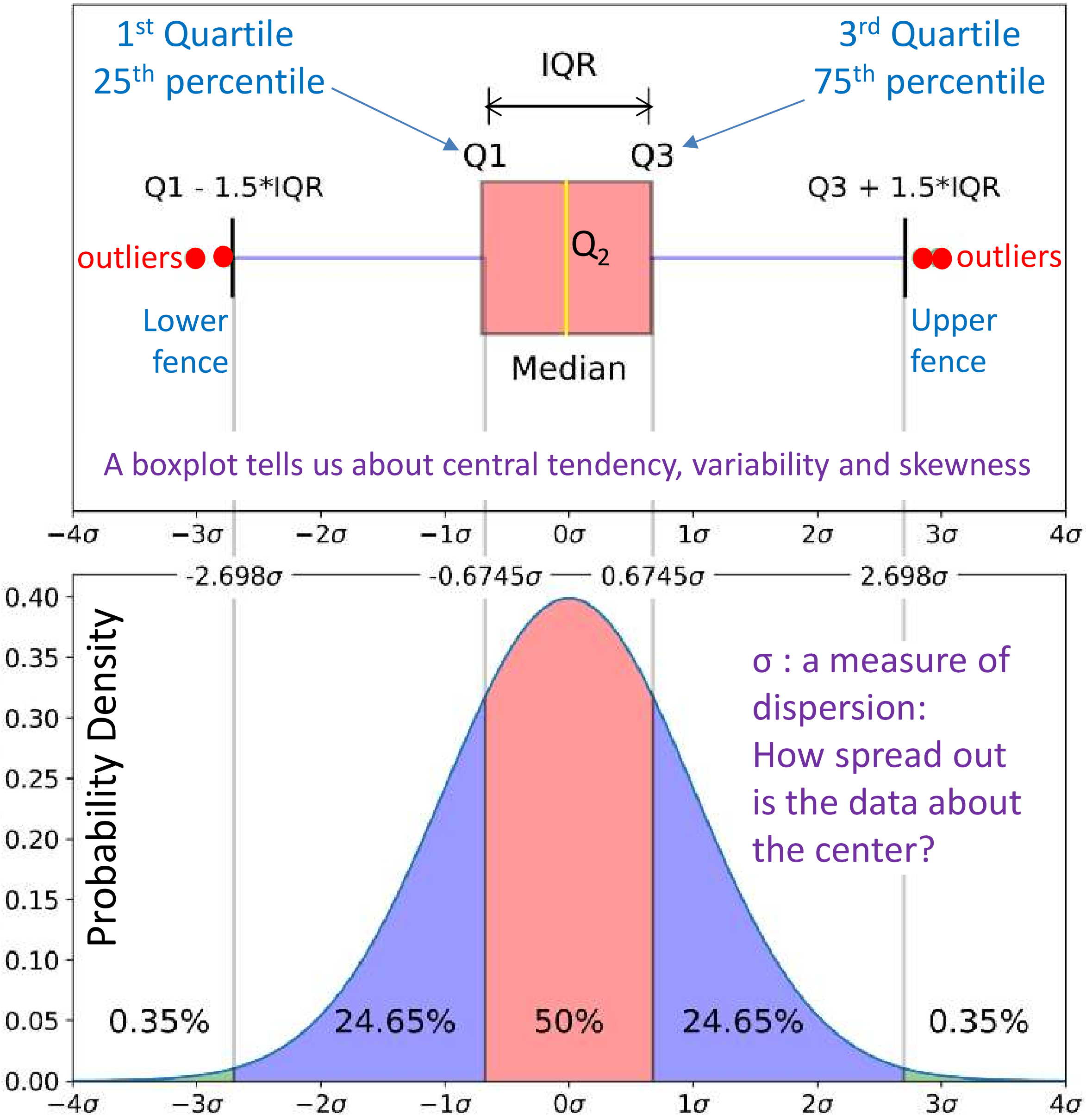
If most of the data points are small, the distribution becomes **right-skewed** (Median > Mean) and this will make the boxplot shifted to the left with a long right whisker.

If most of the data points are large, the distribution becomes **left-skewed** (Mean > Median), and this will make the boxplot shifted to the right with a long left whisker.

Interpretation of boxplots

What a box plot tells us:

- **Short?** Much of your data points are similar (many values in a small range)
- **Tall?** Much of data points are quite different (values spread over a wide range)
- **Median closer to Q1?** Most data has lower values
- **Closer to Q3?** Most data has higher values (media (median not being in the middle sign of skewness)
- **Very long whiskers**? Data has a high **standard deviation** and **variance** (values are spread out and highly varying)
- **Outliers?** Any value beyond [Q_1 - 1.5 x IQR] or [Q_3 + 1.5 x IQR] are considered outliers



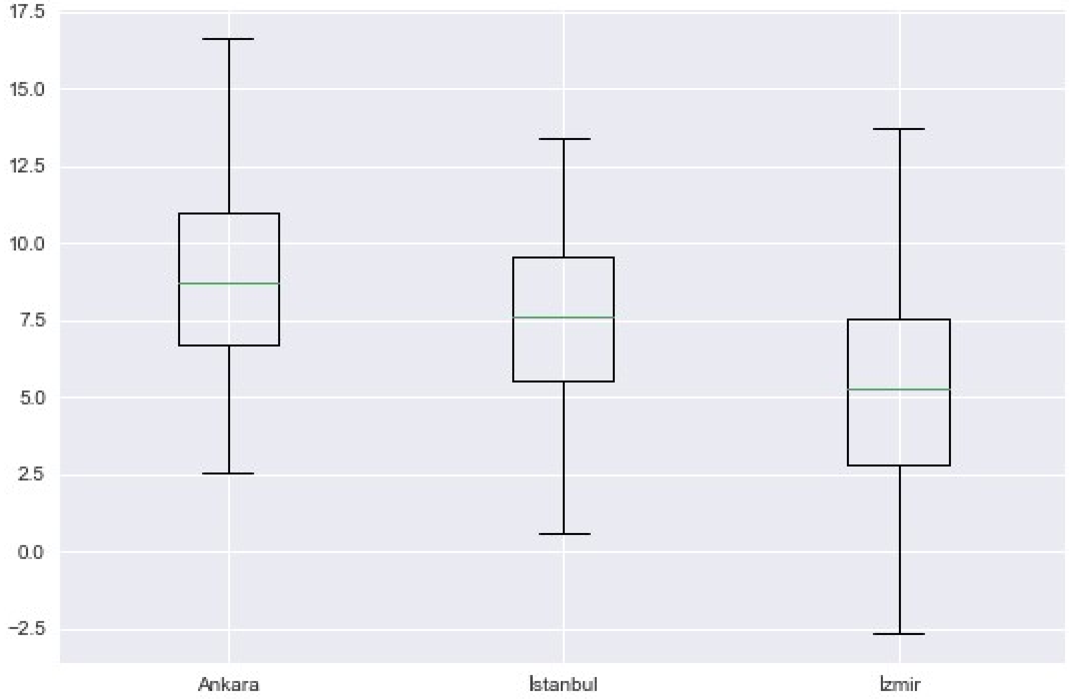
A boxplot tells us about central tendency, variability and skewness

σ : a measure of dispersion: How spread out is the data about the center?

Ref: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Multiple box plots

A simple way to visualize the positive association between the **city** and the CO_2-level **measurement**



Python code

```
ankara = np.random.normal(9,3,120)
istanbul = np.random.normal(7,3,120)
izmir = np.random.normal(5,3,120)
location = [ankara,istanbul,izmir]
fig = plt.figure(1, figsize=(9, 6))
ax = fig.add_subplot(111)
bp = ax.boxplot(location)
ax.set_xticklabels(['Ankara', 'İstanbul', 'İzmir'])
plt.show()
```

Measure of spread (dispersion)

- There are 3 basic ways of measuring dispersion: **Range, Interquartile range** and **variance** (or standard deviation)
- Measure of location tells us nothing about how much variability (spread of a distribution) exists in the data
- Distance measure of dispersion:

  Range = Max – Min

  – Simple to calculate, but can be greatly influenced by any outliers (large or small) distorting the measurement of variability in data

  Interquartile Range (IQR) = Q3 – Q1

  – Difference between the 75th and 25th percentiles

  – Less sensitive to outlier(s) than range as it doesn't use the smallest and largest values
- Range only looks at extremes whereas the **variance** (next) looks at the whole data distribution.

Measure of spread (dispersion)

- **Variance** = A measure of deviation from the mean (distribution of data around the mean)

Population variance: $$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

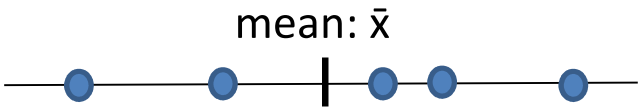
Standard deviation: σ

- More spread out (more variability) higher variance
- σ is more appropriate for visuals & has the same units as X_i
- The standard deviation is heavily influenced by outliers just like the mean
- **CV**: Coefficient of Variation (aka relative std dev):
  - Particularly useful in comparing the standard deviations of 2 different data sets:

$$CV = \frac{\text{Std dev}}{\text{Mean}}$$

Why squared distance in variance?

- Variance is a measure of spread from a reference point



- Why not use (x_i- x̄)? This may sum up to zero!
- Use |x_i- x̄|? Possible, but squaring the difference has some nice mathematical properties:
  – Squaring always gives a positive value, so the sum is never 0
  – Squaring emphasizes larger differences
  – A square gives a nice continuous differentiable function
  – An important property (won't work with mean absolute deviation): Var($x_1 + x_2 +…+ x_n$) = Var(x_1) + var(x_2) +…+ Var(x_n)
  – Variance is defined as the 2nd moment of the deviation (the RV here is (x−μ)) and thus the square as moments are simply the expectations of higher powers of the RV's (more later)

Data Visualization in Bivariate Analysis

- Commonly used visualization methods with respect to data types of predictors and response variables

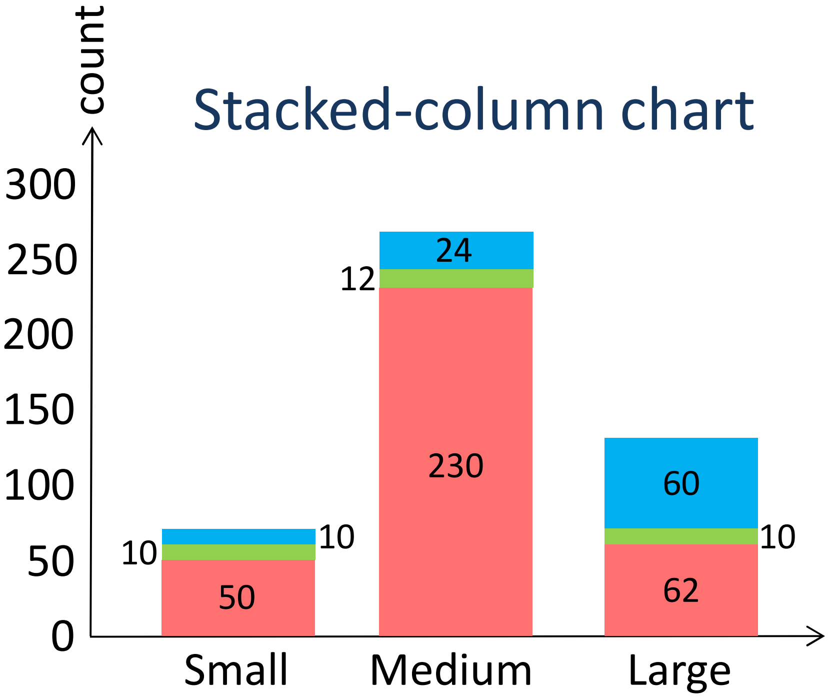
Predictor	Response	Visualization
Categorical	Categorical	Mosaic plots (stacked charts)
Categorical	Numerical	Box plots, Density plots
Numerical	Categorical	Box plots, Density plots
Numerical	Numerical	Scatter plots

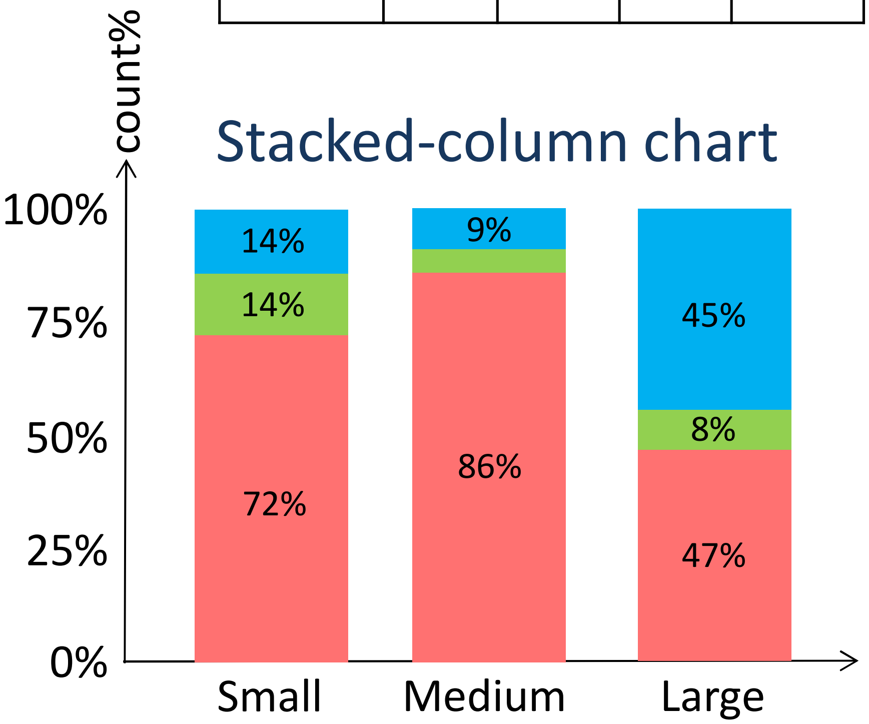
Categorical vs Categorical

- Suppose a store sold a total of 468 shirts in S, M and L sizes with three different colors (**Red**, **Green**, **Blue**) in a month. "Size vs Color"

Two-way table:

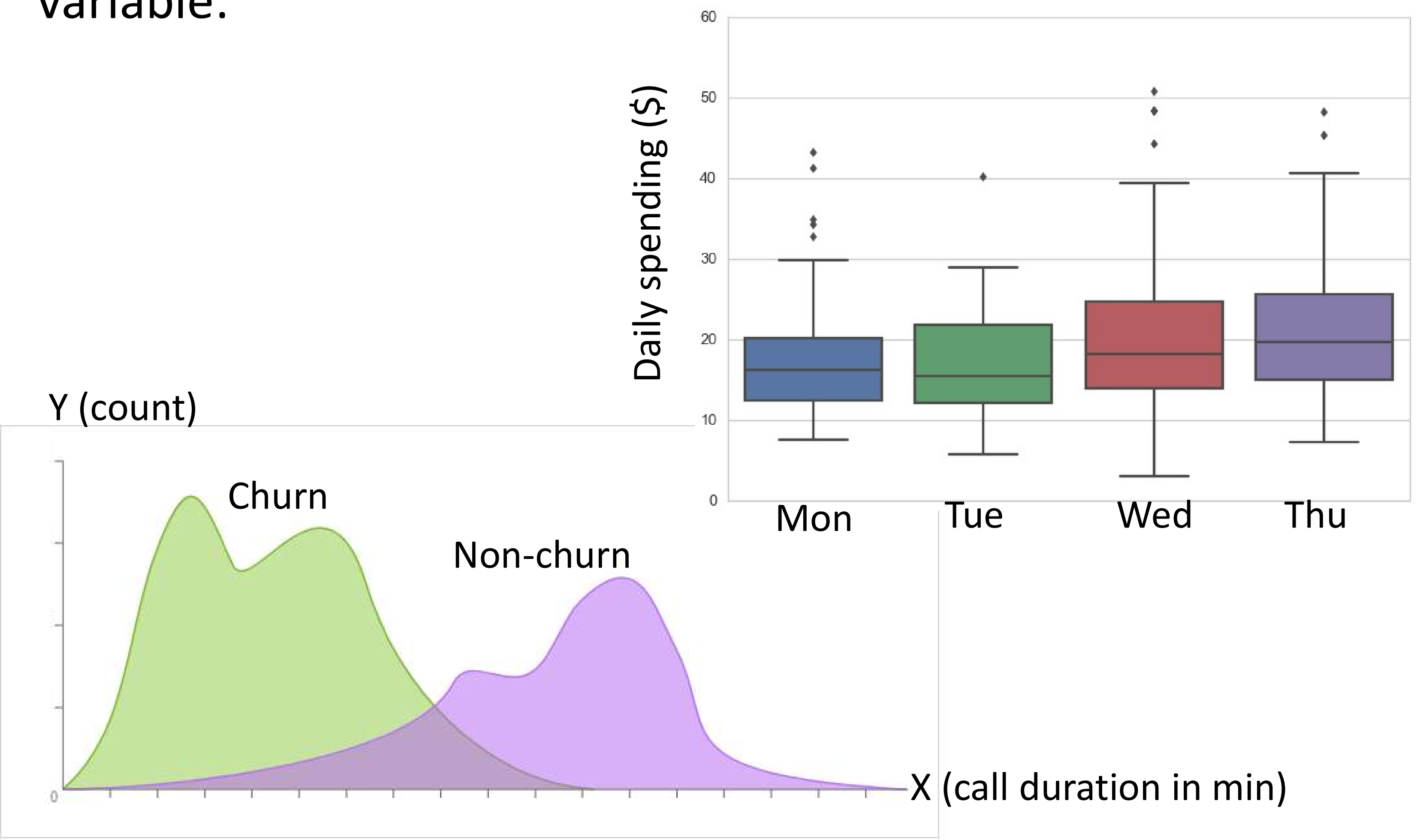
	S	**M**	**L**	total
Red	50	230	62	342
Green	10	12	10	32
Blue	10	24	60	94
total	70	266	132	468





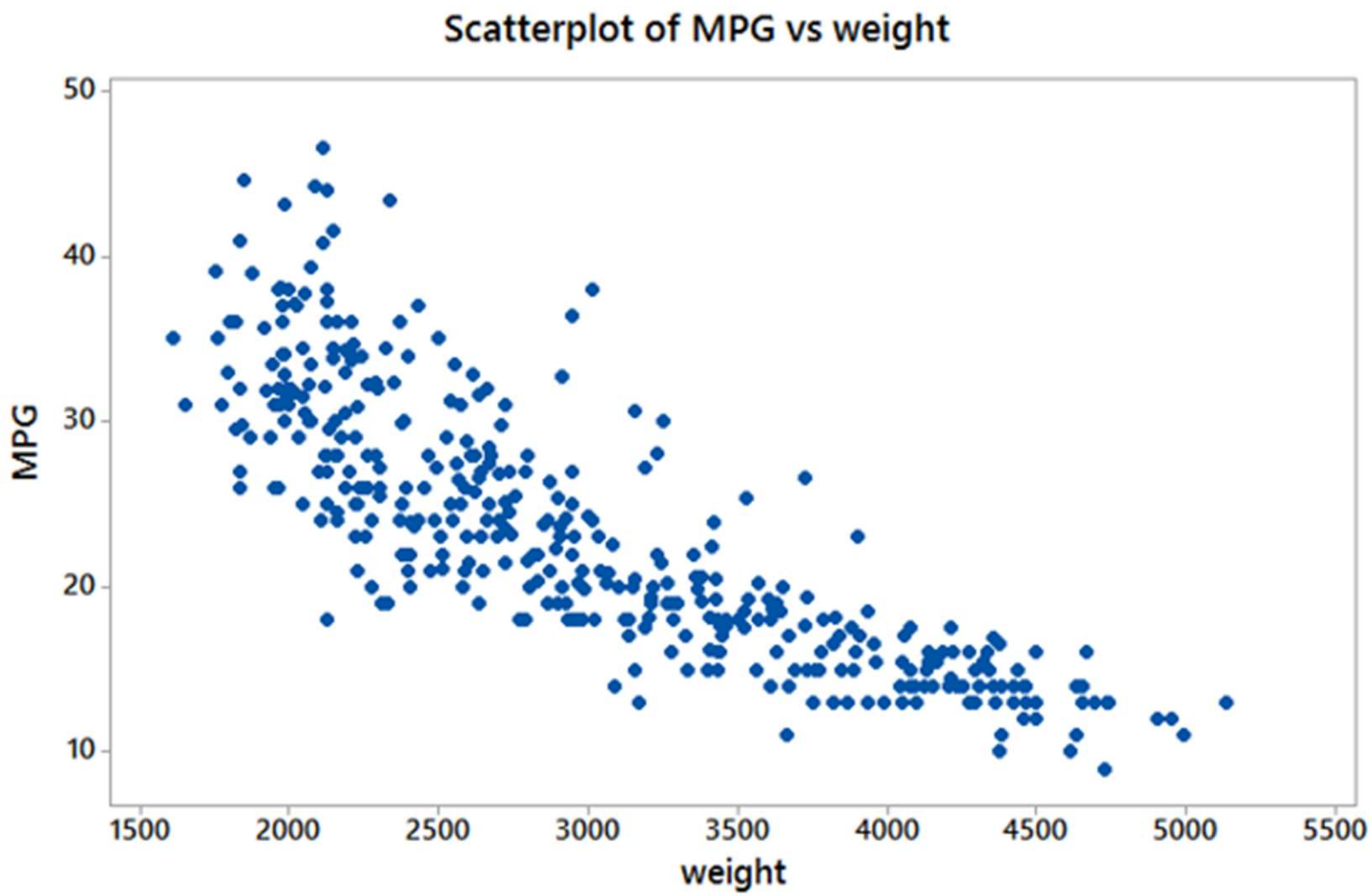
Categorical—Numerical

- In exploration of relations between categorical and continuous variables, we can examine multiple boxplots for each level of categorical variables, or have density plots for every categorical variable:



Numerical—Numerical

- Scatter plots for numerical-numerical variables:



Summary

- Quantitative features used in Descriptive Statistics:
  - **Mean**
  - Median
  - Mode

  Related to accuracy: Where is the data concentrated and what are the typical values?

  - Range (max and min)
  - Quantiles (e.g. **IQR**)
  - **Variance** (σ^2) and **Std deviation** (σ)

  Related to precision: precision: Has implications on estimation or inference errors

- The Descriptive Statistics Report:
  - Form of the distributions
  - Central tendency and spread of the distribution
  - Graphical representations