DESCRIPTIVE STATISTIC

Sorana D. Bolboacă

OBJECTIVES

Measures of Centrality	Measures of Spread
✓ Mean	✓ Range (amplitude)
✓ Mediana	✓ Variance
✓ Mode	✓ Standard deviation
✓ Central value	✓ Coefficient of variance
	✓ Standard error
Measures of Symmetry	**Measures of Localization**
✓ Skewness	✓Quartile
✓ Kurtosis	✓Percentiles

CENTRALITY MEASURES

- Mean (arithmetic average)
- Median: midpoint of the distribution (50th percentile)
- Mode: most frequent observation

Population → parameter

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

Sample → statistics

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Odd (3, 5, 7 …)

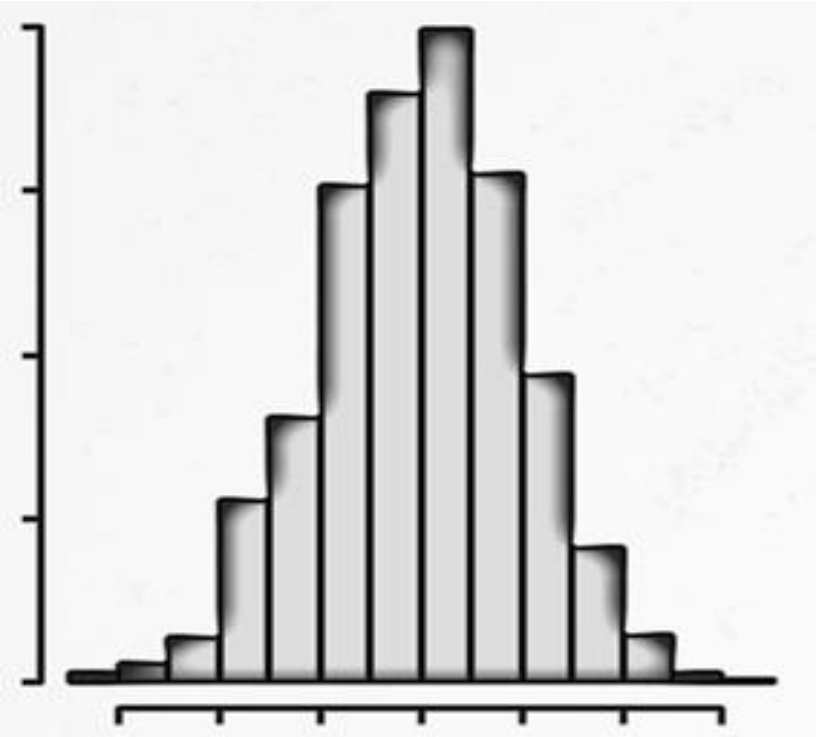
$$Me = X_{\frac{n+1}{2}}$$

$$Me = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

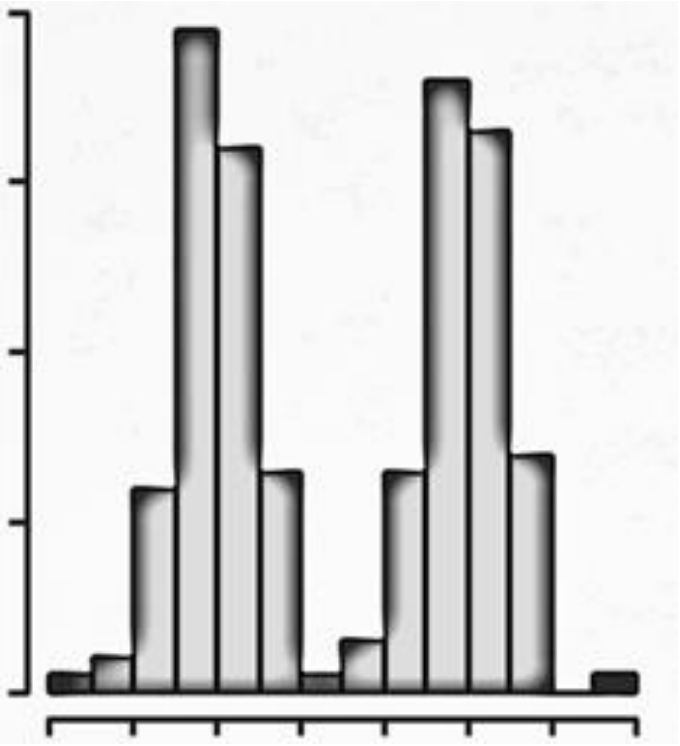
Even (2, 4, …)

CENTRALITY MEASURES

Mode / Modal value



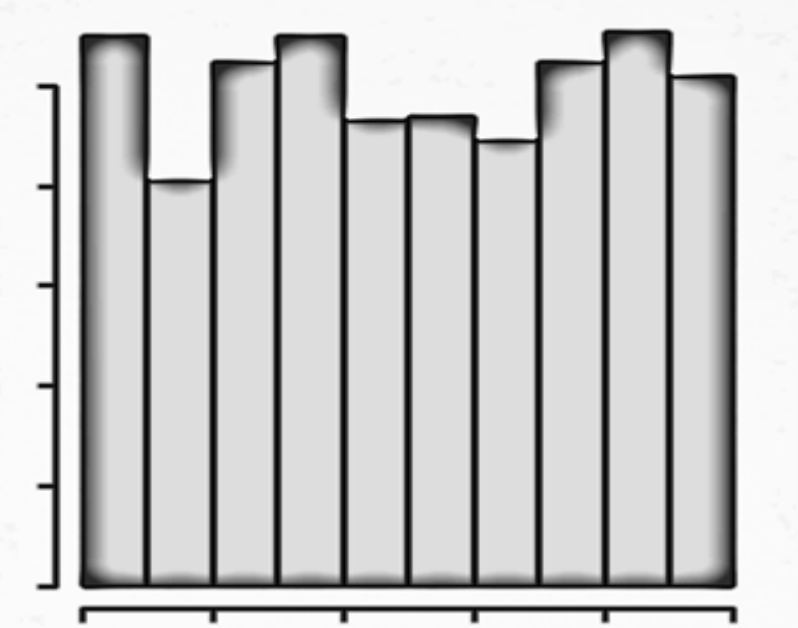
Unimodal



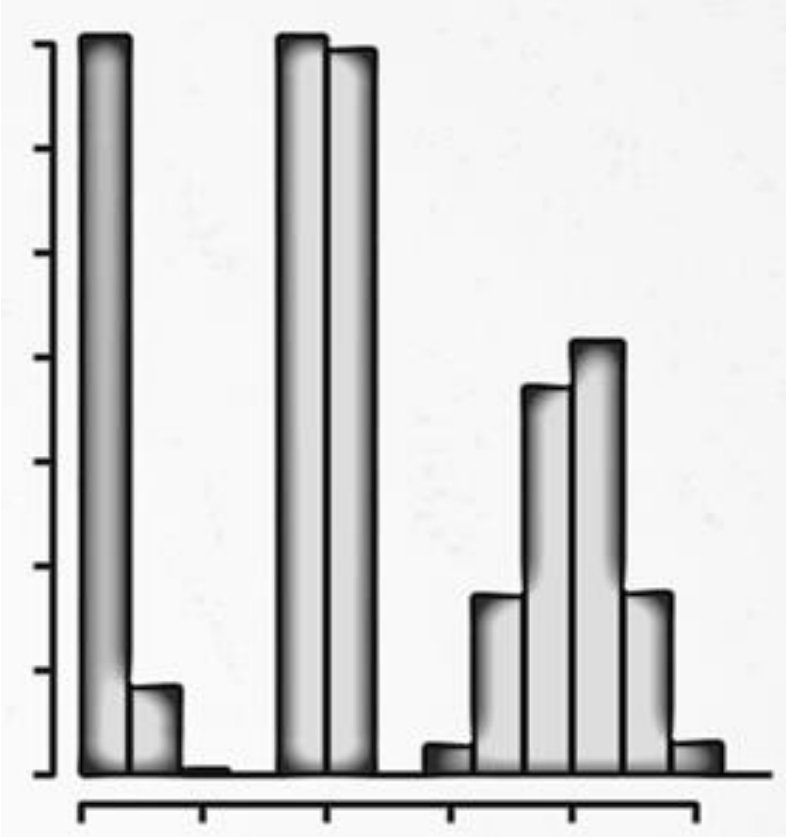
Bimodal

CENTRALITY MEASURES

Mode / Modal value



Uniform



Multimodal

EXAMPLE

11 student's practical exam scores:

4, 9, 5, 8, 6, 7, 9, 10, 8, 6, 5

- Mean = (4+9+5+8+6+7+9+10+8+6+5)/11 = 7
- Multimodal: 5, 6, 8, 9
- Median: 4, 5, 5, 6, 6, 7, 8, 8, 9, 9, 10
  - n (sample size) = 11
  - Me = $X_{(n+1)/2} = X_6 = 7$

EXAMPLE

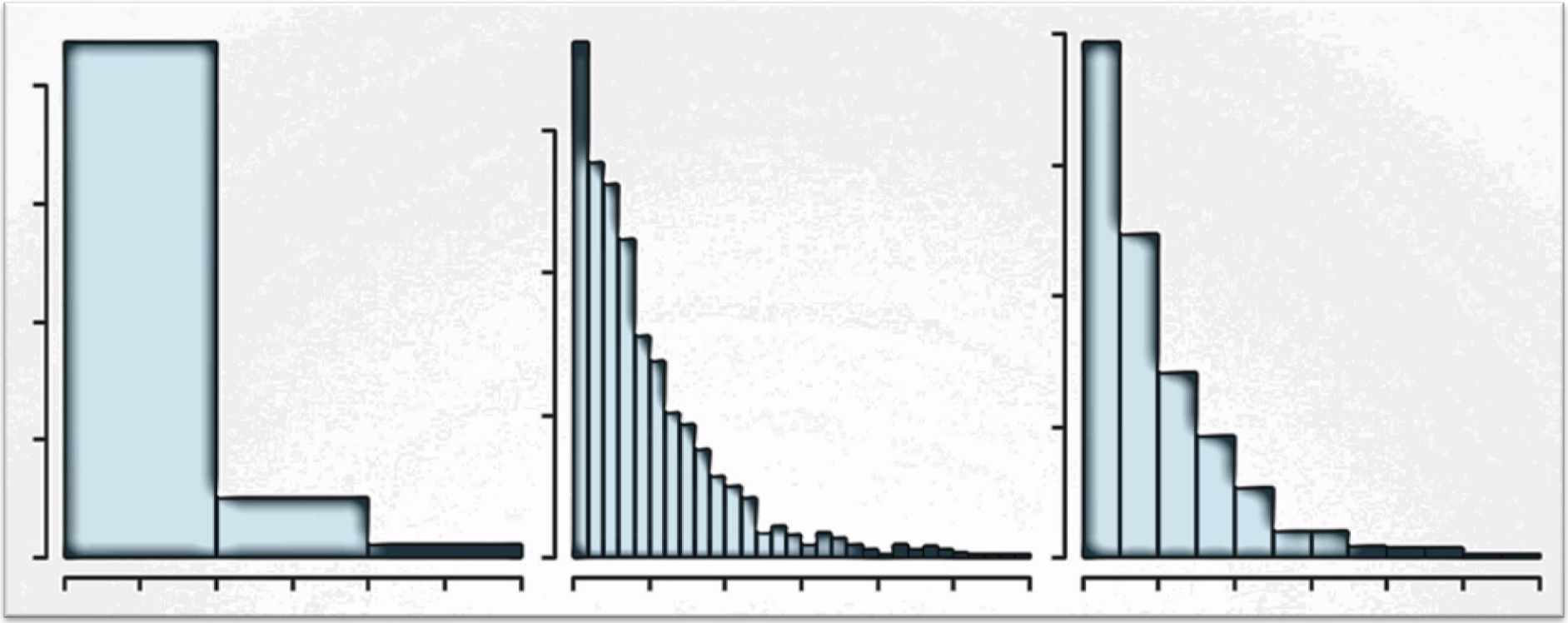
12 student's practical exam scores:

4, 9, 5, 8, 6, 4, 9, 10, 8, 6, 5, 4

- Mean = (4+9+5+8+6+4+9+10+8+6+5+4)/12 = 6.5
- Unimodal: 4
- Median: 4, 4, 4, 5, 5, 6, 6, 8, 8, 9, 9, 10,
  - n (sample size) = 12
  - Me = $(X_{n/2} + X_{n/2+1})/2 = (X_6 + X_7)/2 = (6+6)/2 = 6$

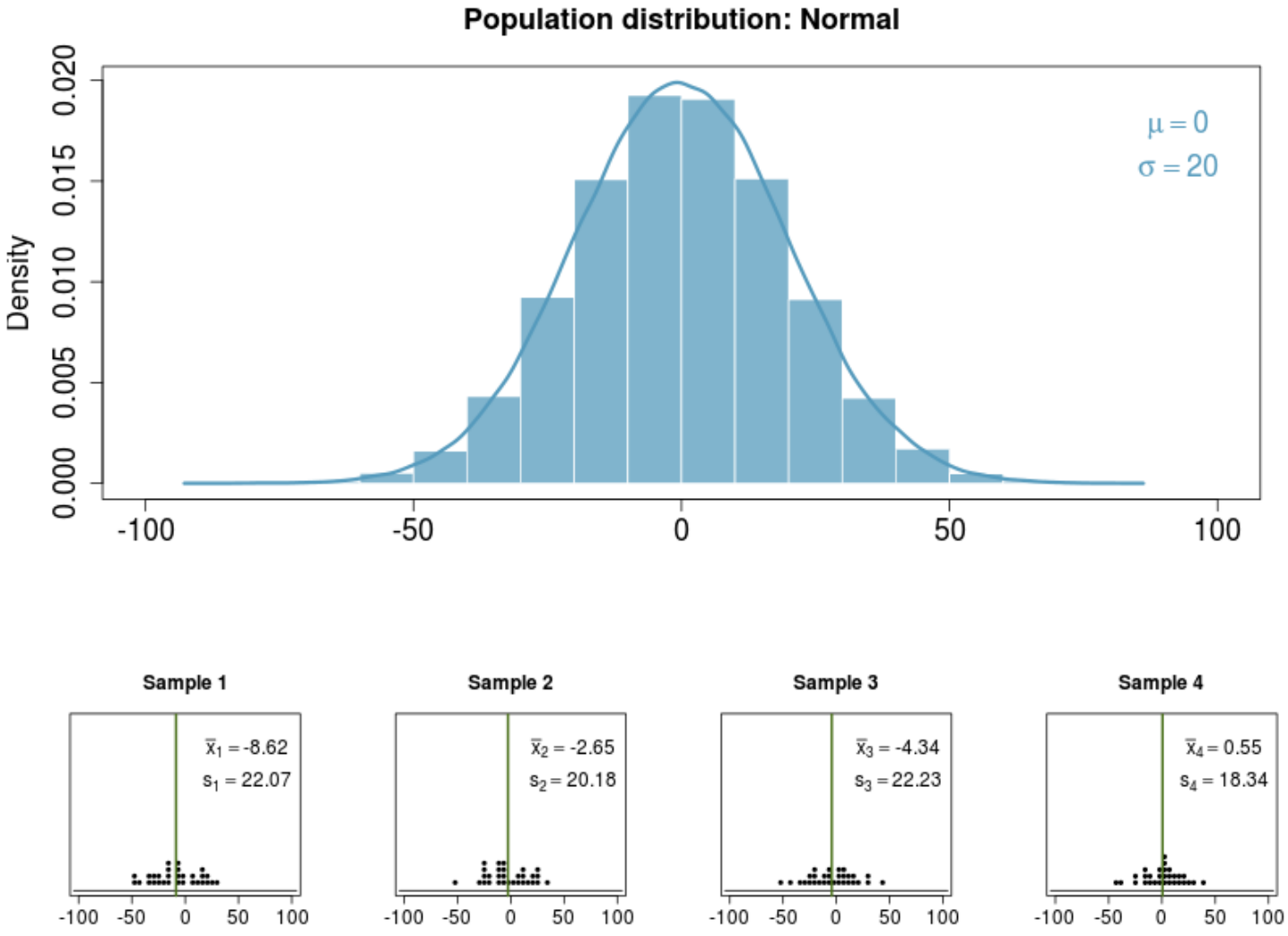
CENTRALITY MEASURES

Mode: different width of the bin alter the distribution of data and what the histogram tell us

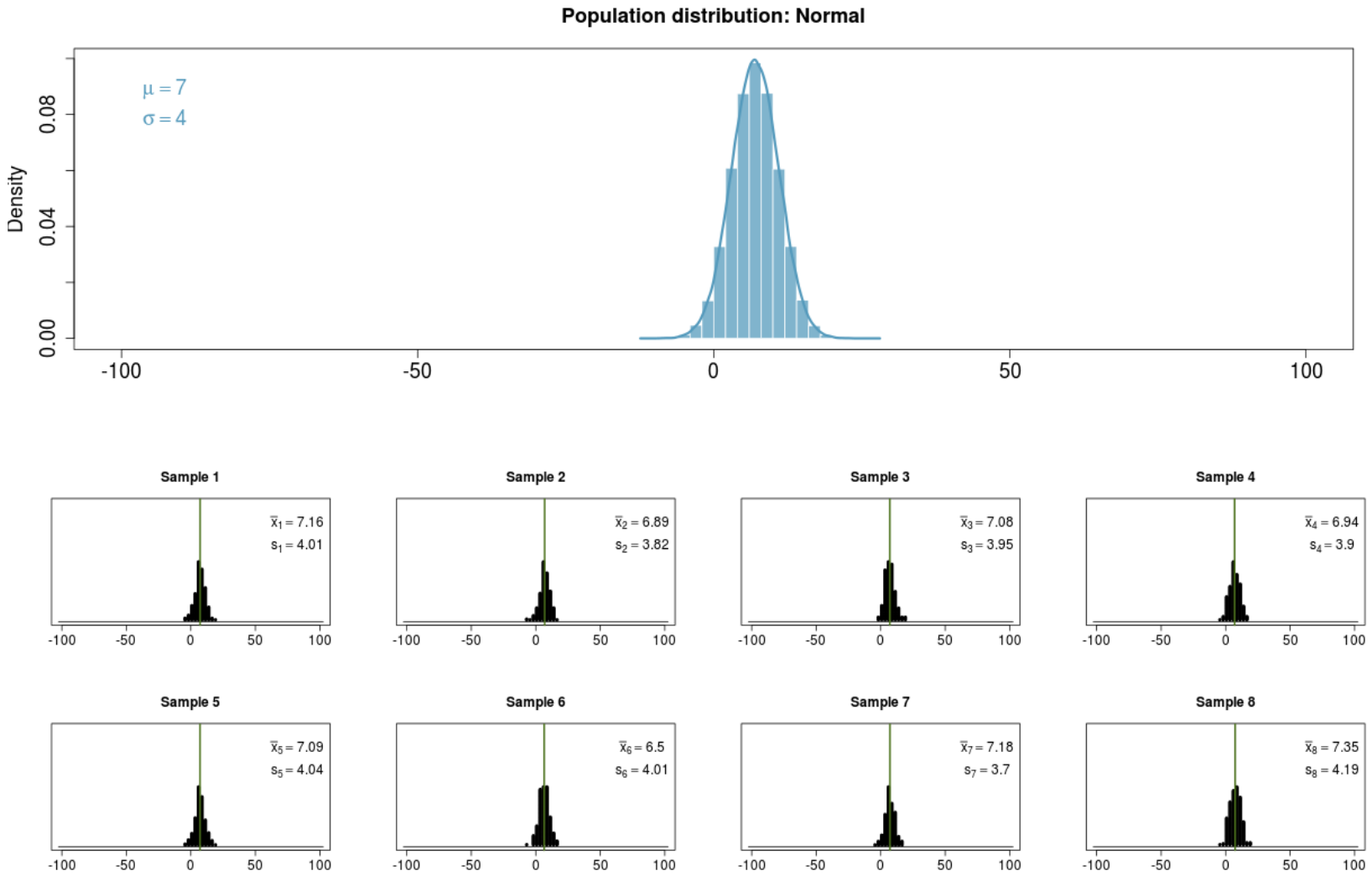


CENTRALITY MEASURES

Arithmetic average: http://spark.rstudio.com/minebocek/CLT_mean/



CENTRALITY MEASURES



CENTRALITY MEASURES

- Weighted mean

$$m_X = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Arithmetic mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Arithmetic mean is a special case of the weighted mean ($W_i = 1$).

Dispersion Measures

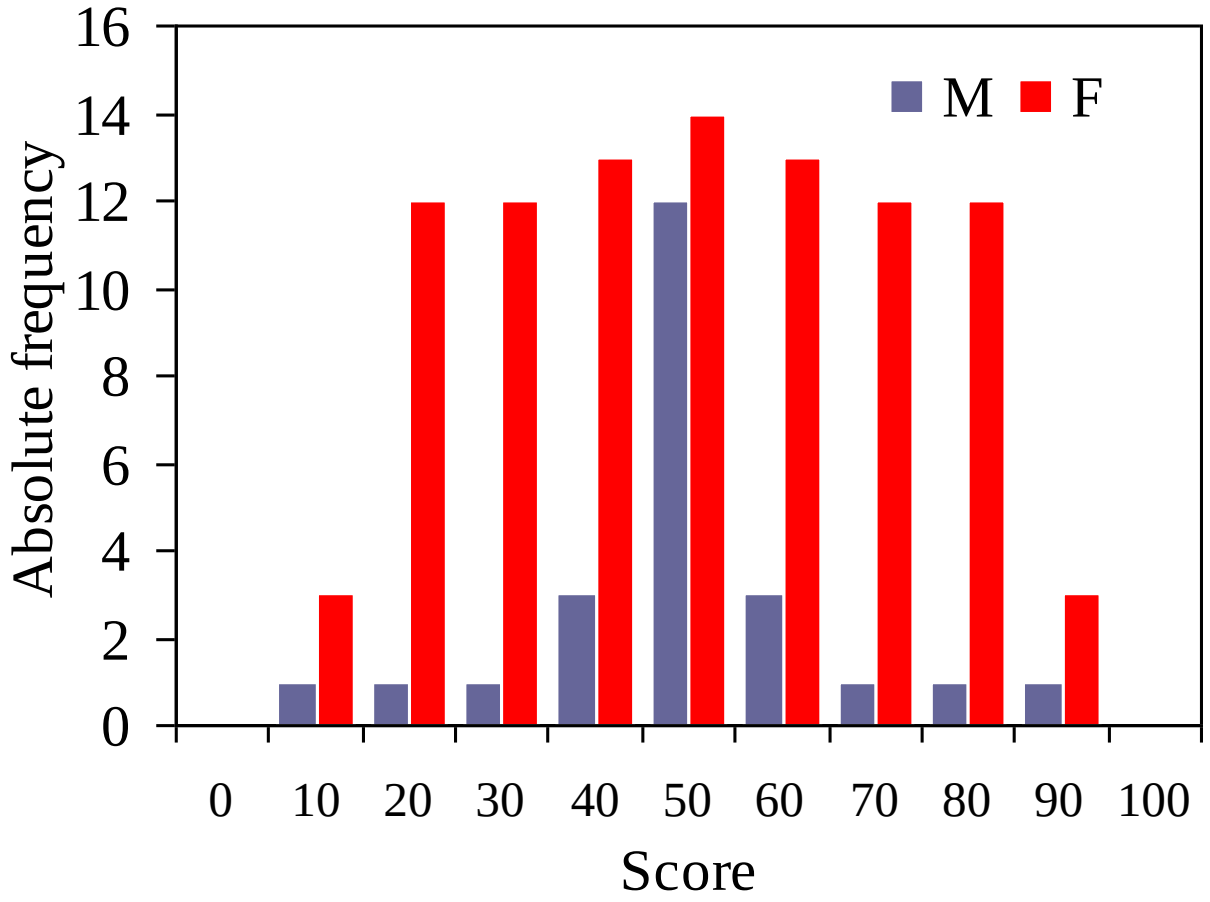
- Spread related to the central value
- The data are more spread as their values are more different by each other

DISPERSION MEASURES

$R_M = 90\text{-}10 = 80$

$R_F = 90\text{-}10 = 80$

$$R = X_{max} - X_{min}$$



DISPERSION MEASURES

- **Population** variance:

$$\sigma^2 = \frac{SS}{n} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}$$

- **Sample** variance (the sample variance tend to sub estimate the population variance):

$$s^2 = \frac{SS}{n-1} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

DISPERSION MEASURES

- Standard deviation (sd) = square root of variance
  - Describe variability
- Is useful when considering how close the data are to the mean
  - Population (σ)
  - Sample (s)

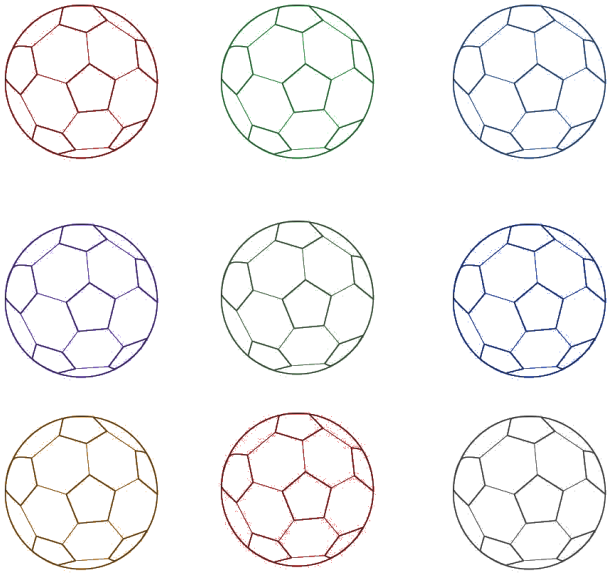
$$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}}$$

DISPERSION MEASURES

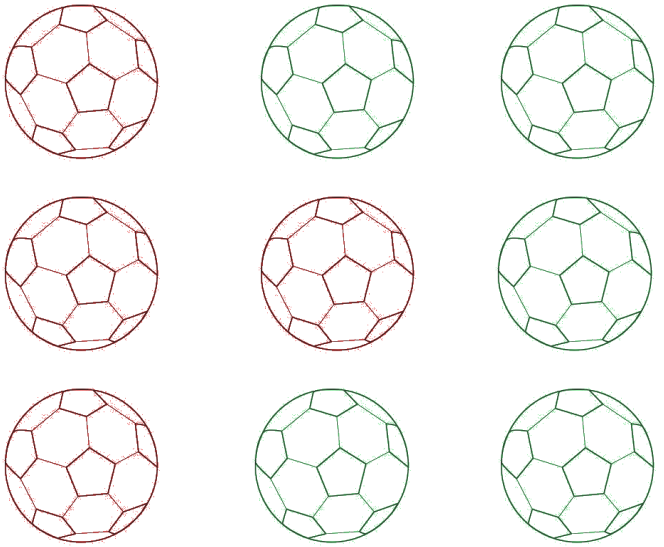
Variability vs. Diversity

Which of the following sets of ball has a more diverse composition of colors?

Set 1



Set 2

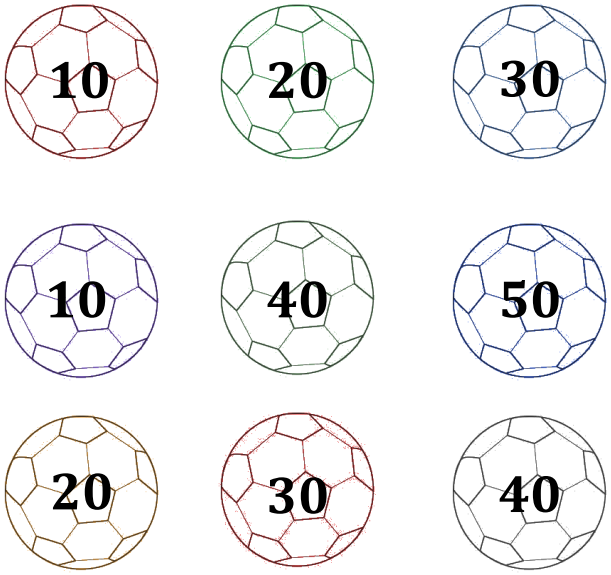


Dispersion Measures

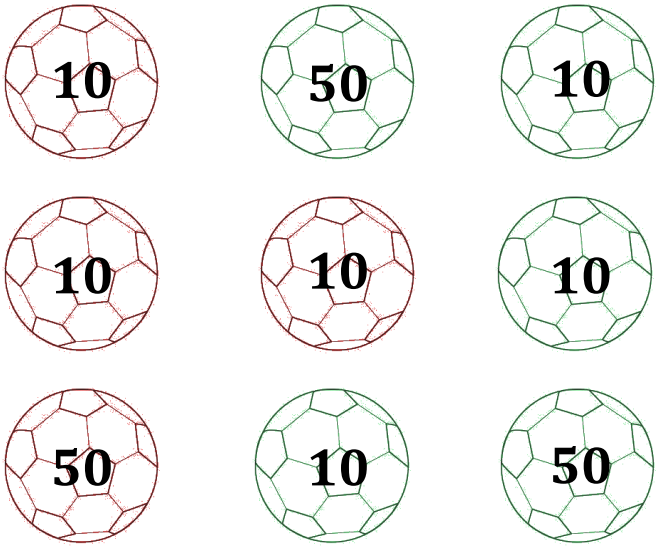
Variability vs. Diversity

Which of the following sets of ball has more variable hours of use?

Set 1



Set 2

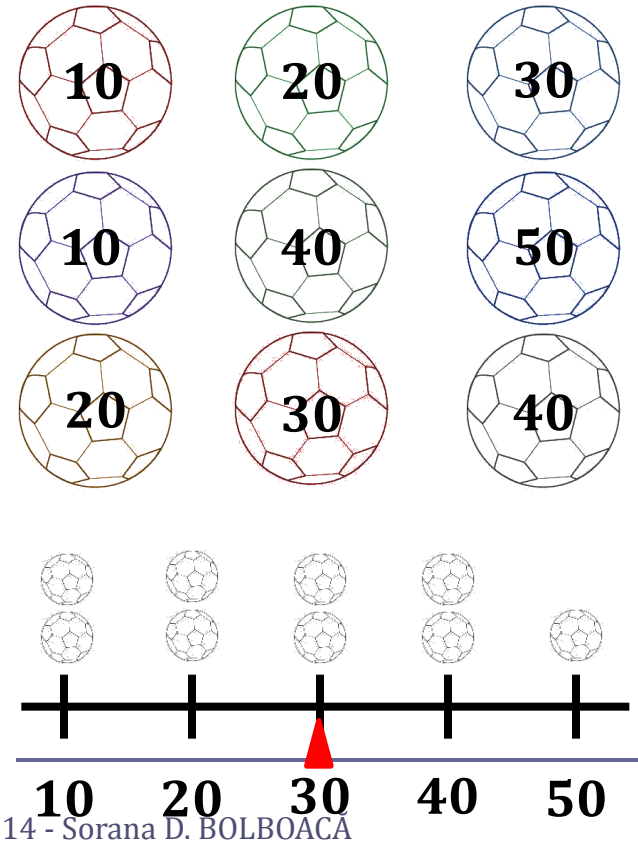


DISPERSION MEASURES

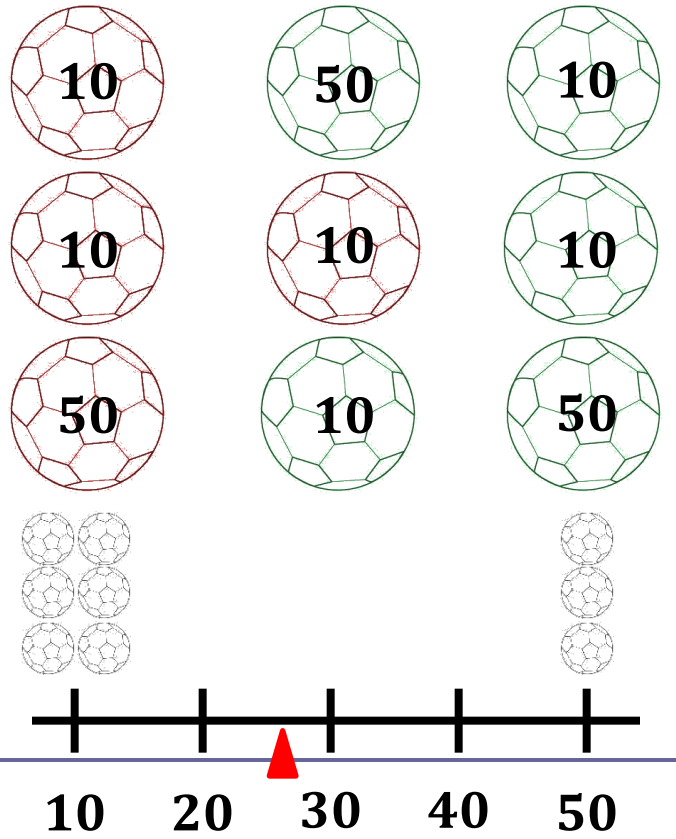
Variability vs. Diversity

Which of the following sets of ball has more variable hours of use?

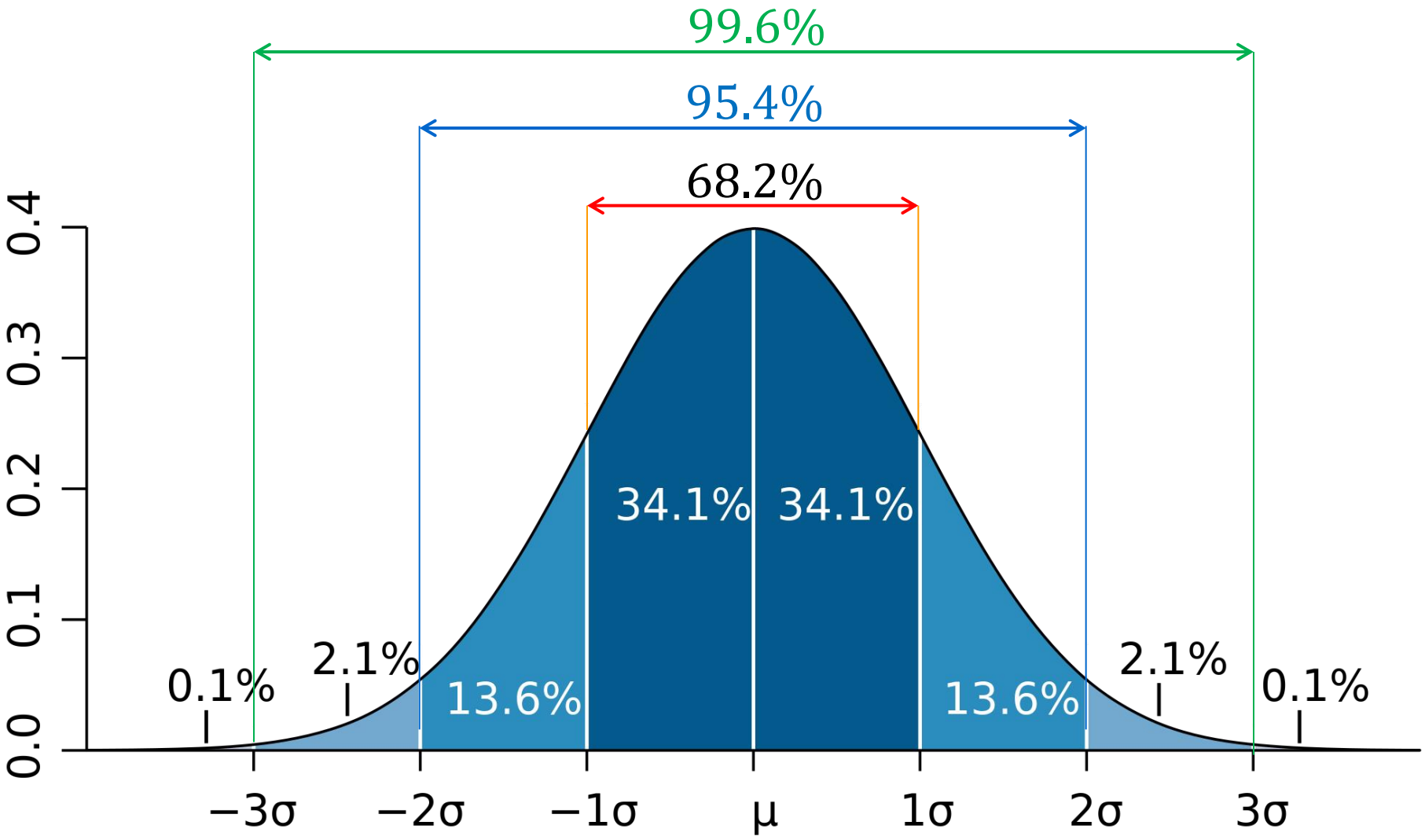
Set 1⟶ s = 15.81



Set 2 ⟶ s = 21.91

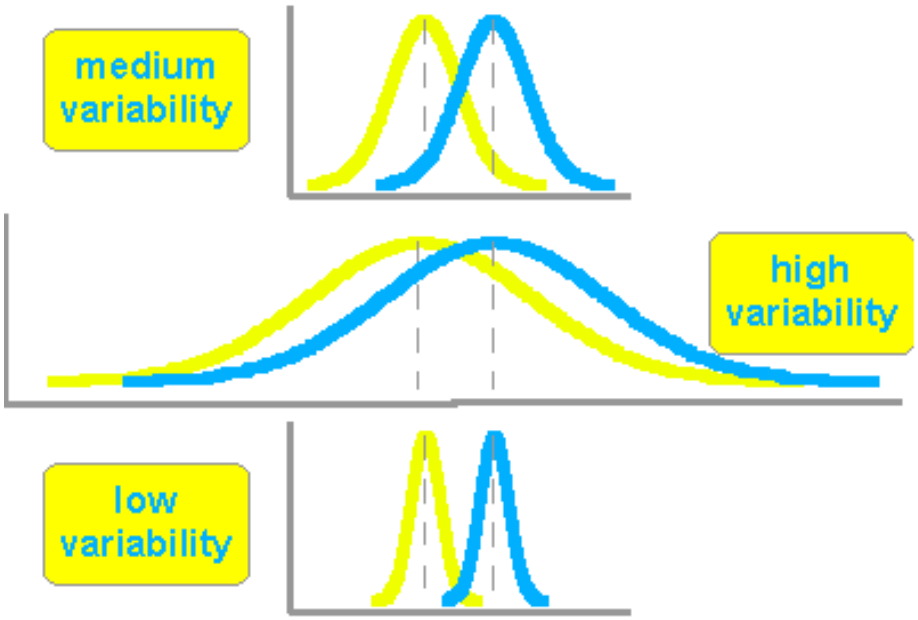


Dispersion Measures



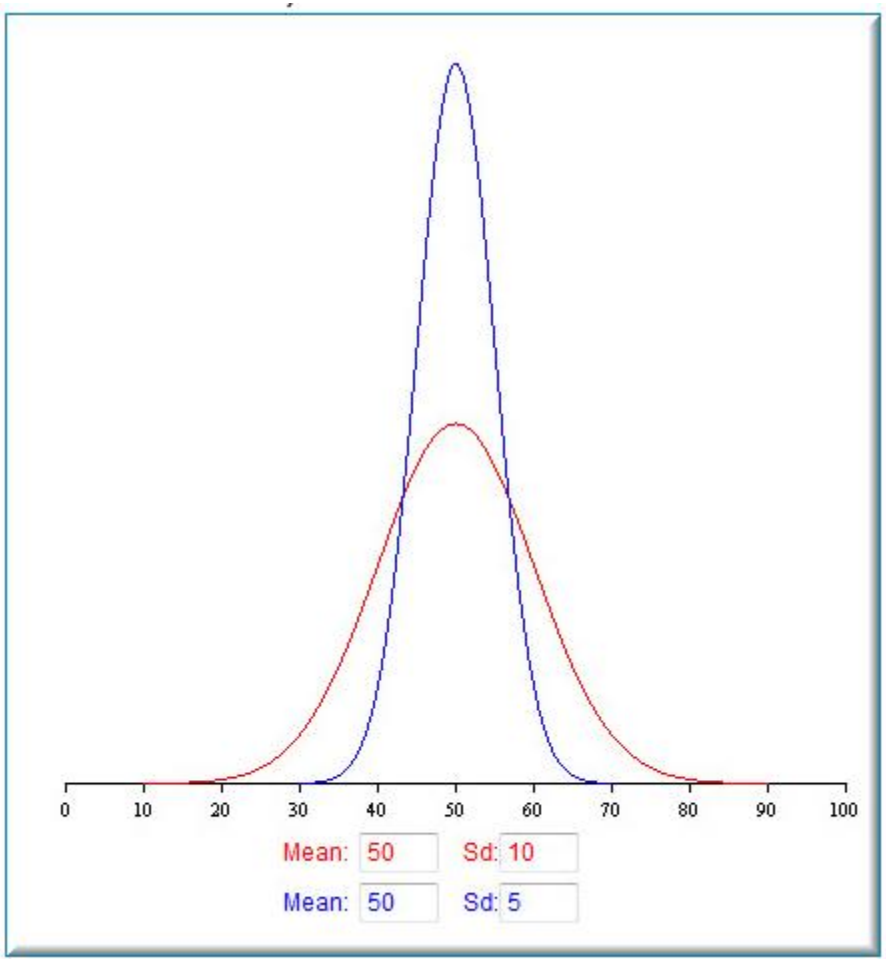
DISPERSION MEASURES

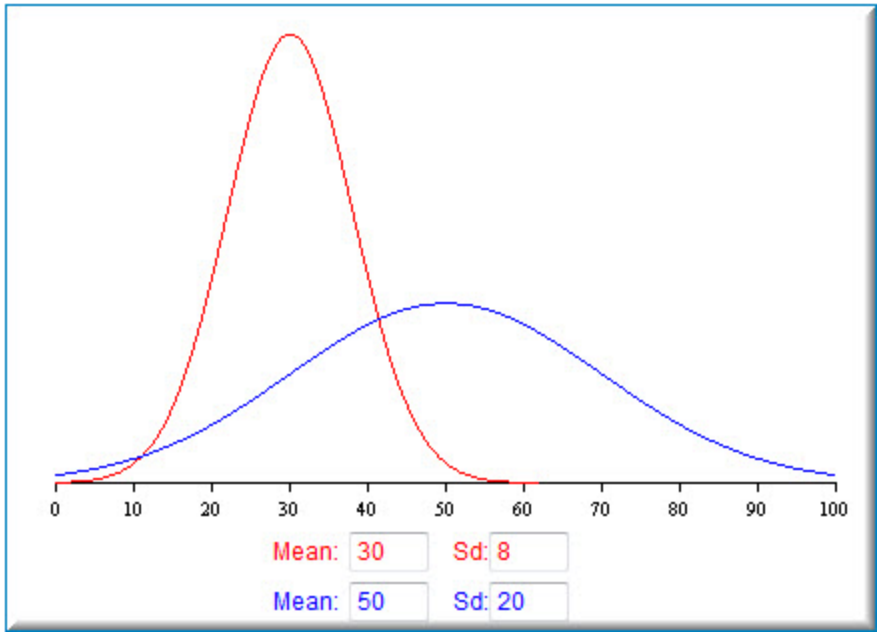
- ↓ s ⟶ data is clustered closely around the mean value
- ↑ s ⟶ a wider spread around the mean



DISPERSION MEASURES

http://onlinestatbook.com/2/summarizing_distributions/spread_sim.html





DISPERSION MEASURES

COEFICIENT OF VARIATION

- A normalized measure of dispersion of a probability distribution
- Ratio of the standard deviation to the mean
- Computed on data measured on ratio scale which can have just positive values
- Is a dimensionless number

$CV < 0.10$	Homogenous
$0.10 \leq CV < 0.20$	Relative homogenous
$0.20 \leq CV < 0.30$	Relative heterogeneous
> 0.20	Heterogeneous

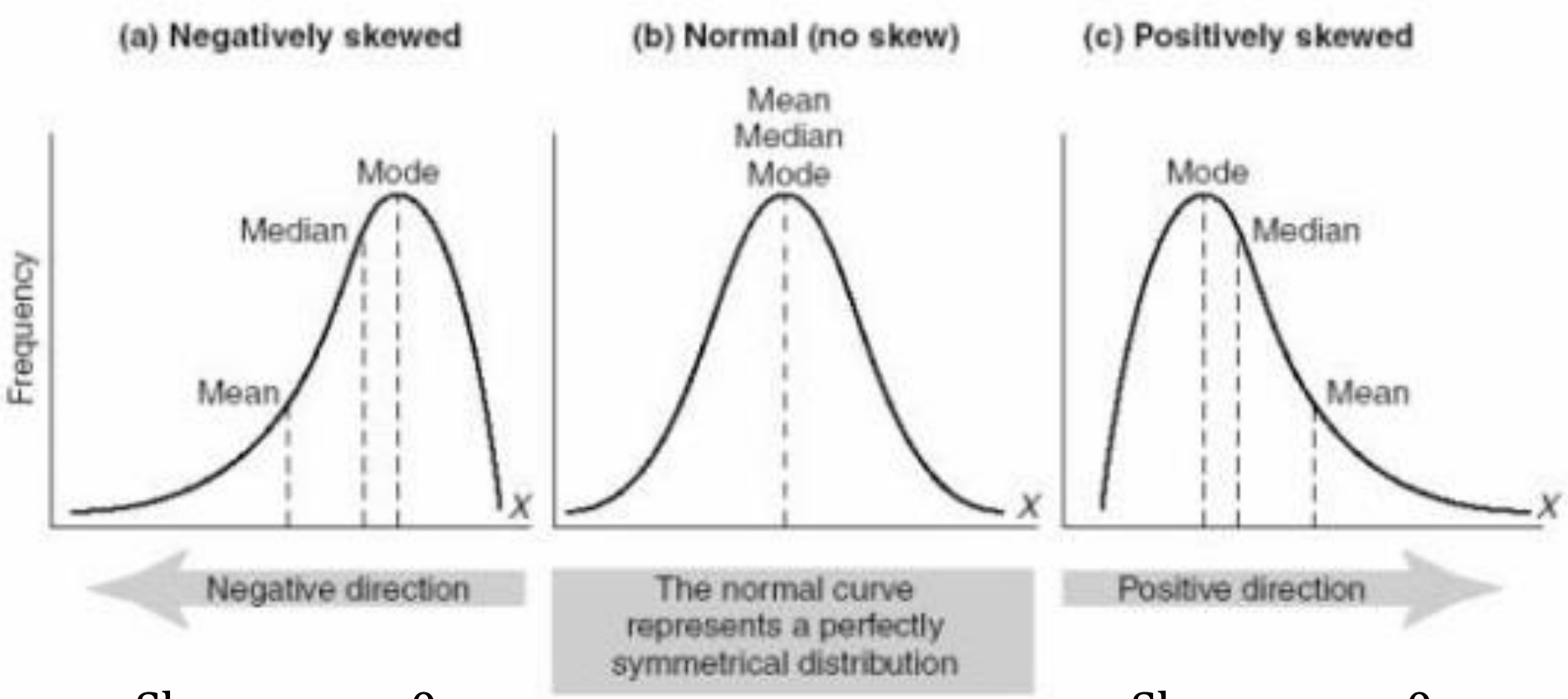
DISPERSION MEASURES

STANDARD ERROR (SEM)

- indicates the accuracy of the sample mean: $SEM = s/\sqrt{n}$
- n increase ⟶ SEM decrease

SHAPE MEASURE

http://chubbyrevision.weebly.com/representation-of-data.html

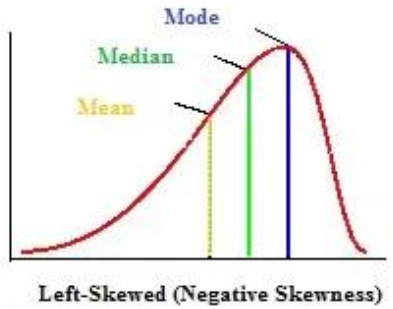


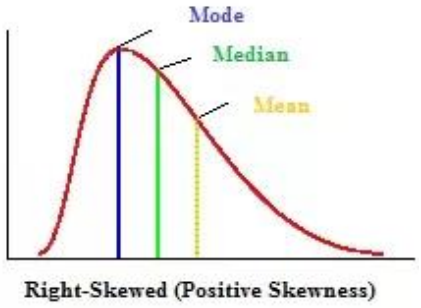
Skewness < 0

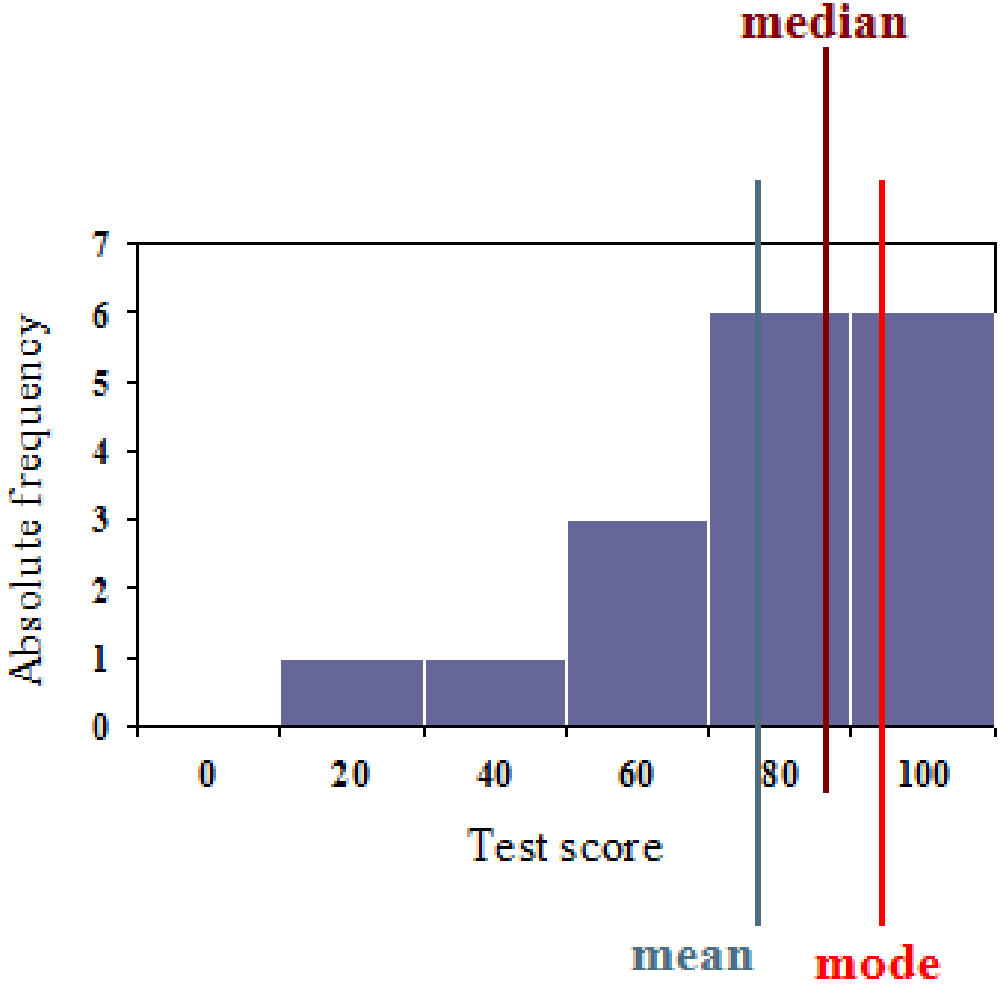
Skewness > 0

Skewness = 0

SHAPE MEASURE



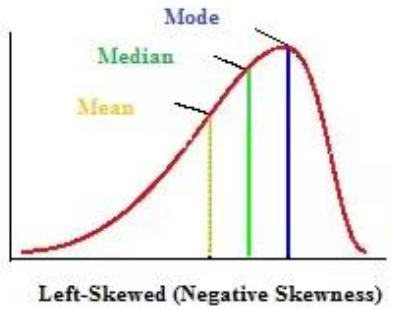


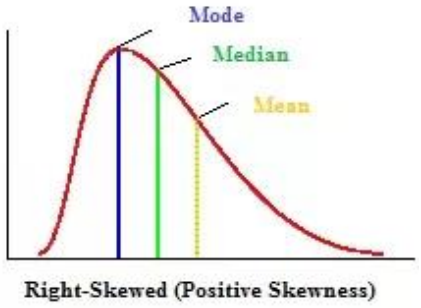


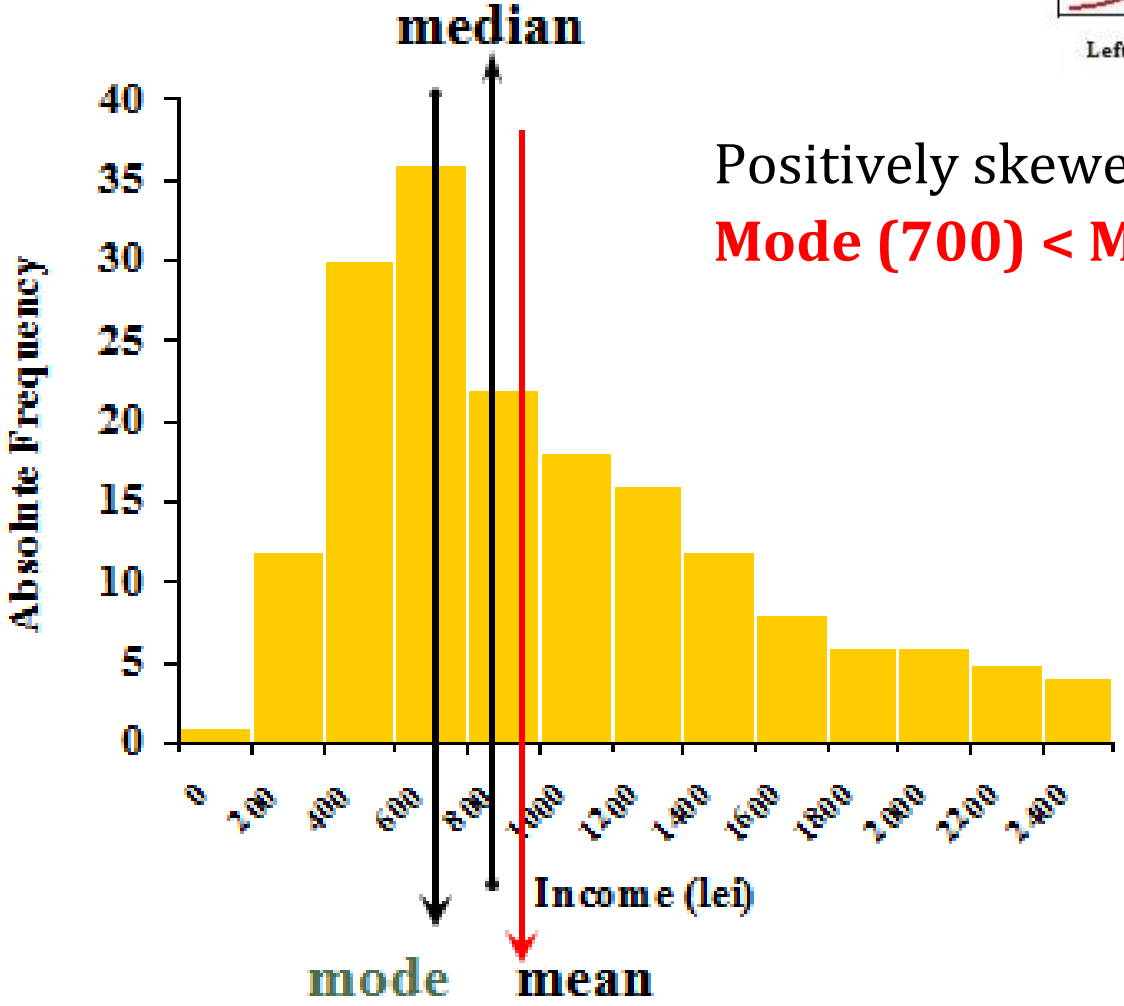
Negatively skewed

Mode > Median > Mean

SHAPE MEASURES







Positively skewed:

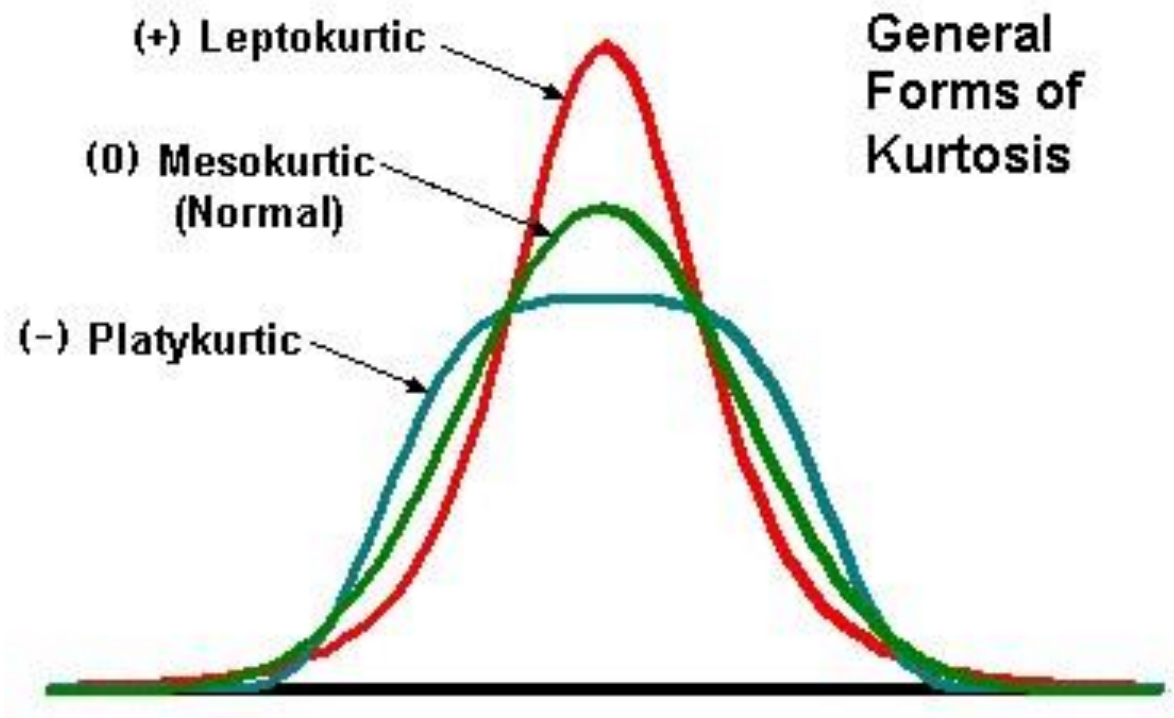
Mode (700) < Median (887) < Mean (936)

SHAPE MEASURES: SKEWNESS

- Interpretation [Bulmer MG. Principles of Statistics. Dover, 1979.] – applied to population
  - If skewness is less than −1 or greater than +1, the distribution is **highly skewed**.
  - If skewness is between −1 and −½ or between +½ and +1, the distribution is **moderately skewed**.
  - If skewness is between −½ and +½, the distribution is **approximately symmetric**.
- Can you conclude anything about the population skewness looking to the skewness of the sample? → Inferential statistics

SHAPE MEASURES

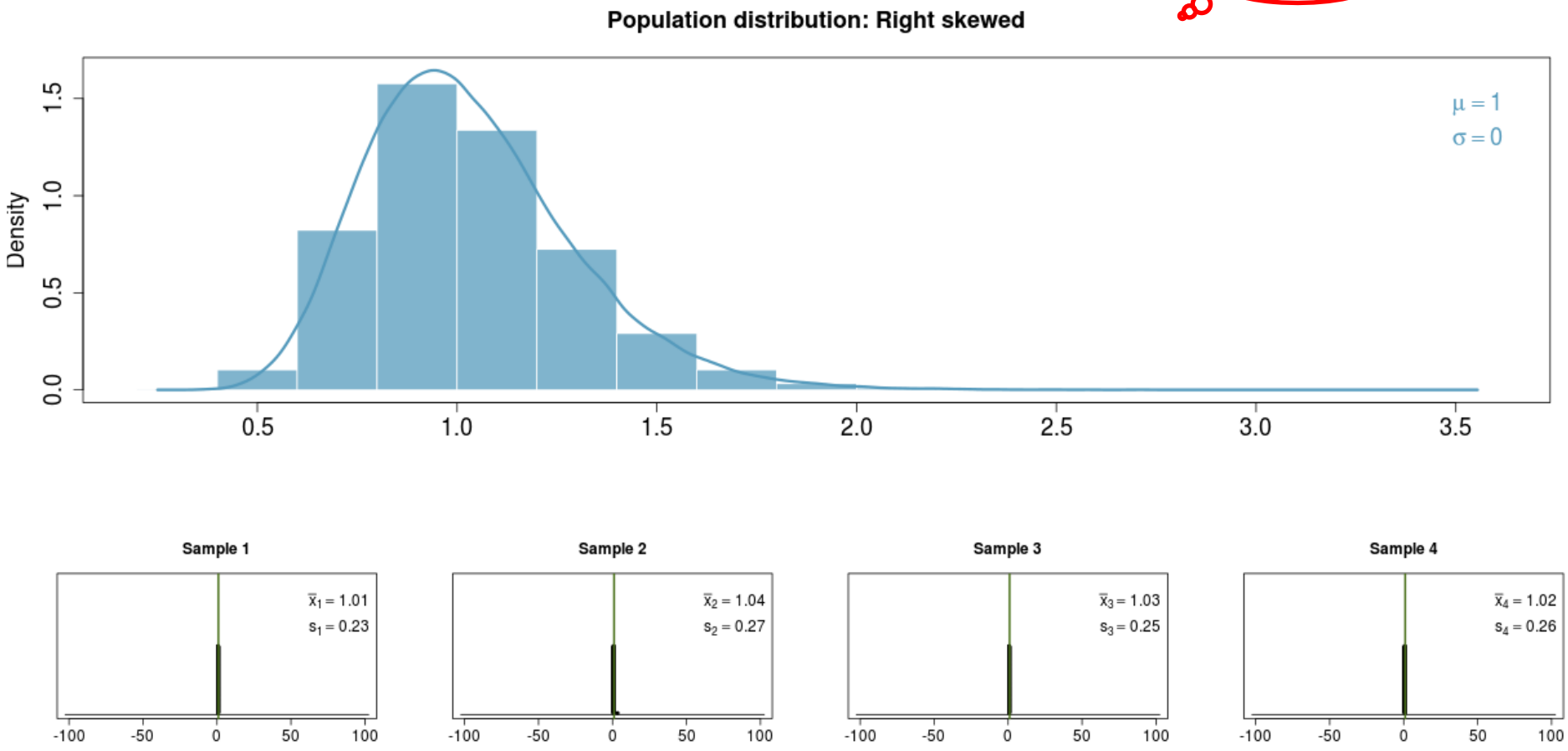
http://mvpprograms.com/help/mvpstats/distributions/SkewnessKurtosis



SHAPE MEASURES: KURTOSIS

- The reference standard is a normal distribution, which has a kurtosis of 3.
- Excess kurtosis (kurtosis in Excel) = kurtosis – 3
  - A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≅3 (excess ≅0) is called **mesokurtic**.
  - A distribution with kurtosis <3 (excess kurtosis <0) is called **platykurtic**. Compared to a normal distribution, its central peak is lower and broader, and its tails are shorter and thinner.
  - A distribution with kurtosis >3 (excess kurtosis >0) is called **leptokurtic**. Compared to a normal distribution, its central peak is higher and sharper, and its tails are longer and fatter.

SKEWNESS VS CENTRALITY MEASURES

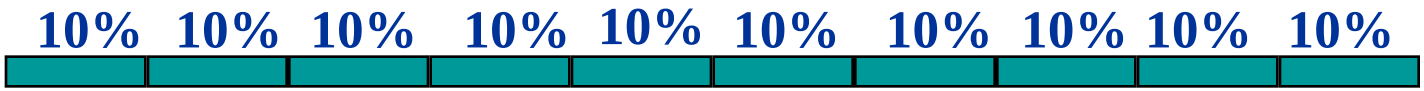


LOCALIZATION MEASURES

- **Quatiles**:
  - Split the series in 4 equal parts:

(minimum) **(median)** **(maximum)**

- **Decile**:
  - Split the series in 10 equal parts:



LOCALIZATION MEASURES

The symmetry of a distribution could be analyzed using quartiles

Let Q_1, Q_2 and Q_3 be 1st (1/3), 2nd (1/2) and 3rd (3/4) quartiles:

- $Q_2-Q_1 \approx Q_3-Q_2$ ($\approx$ almost equal) $\rightarrow$ the distribution is almost symmetrical
- $Q_2-Q_1 \neq Q_3-Q_2 \rightarrow$ the distribution is asymmetrical (through left or right)

MEASURES OF LOCALIZATION: QUARTILES

2.80	2.97	3.05	3.25	3.40	3.45	3.80	4.10	4.30	4.40
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}

- $Q_1 = 3.03$
- $Q_2 = 3.43$
- $Q_3 = 4.15$

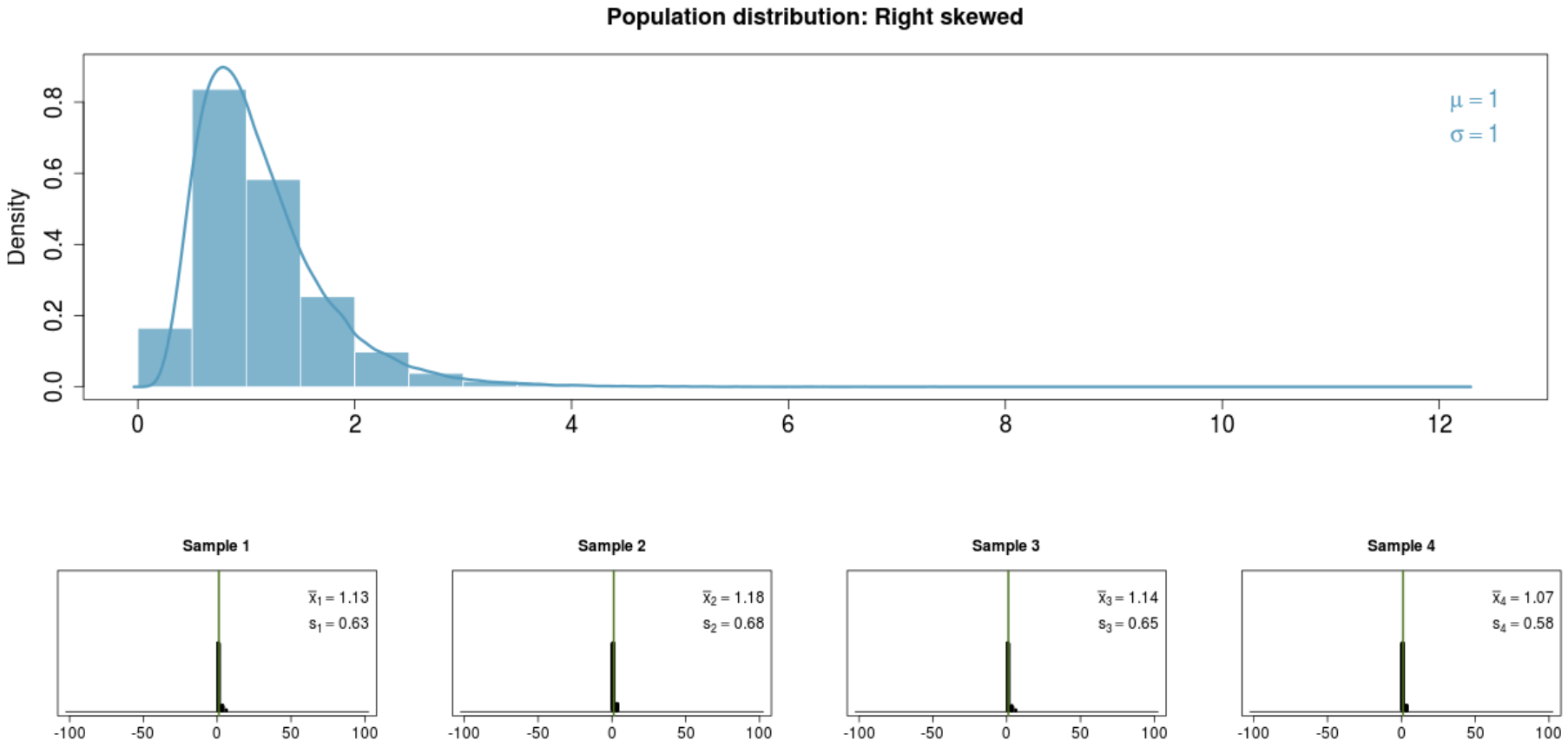
$Q_2 - Q_1 = 3.43 - 3.03 = 0.40$

$Q_3 - Q_2 = 4.15 - 3.43 = 0.72$

How do you interpret this result???

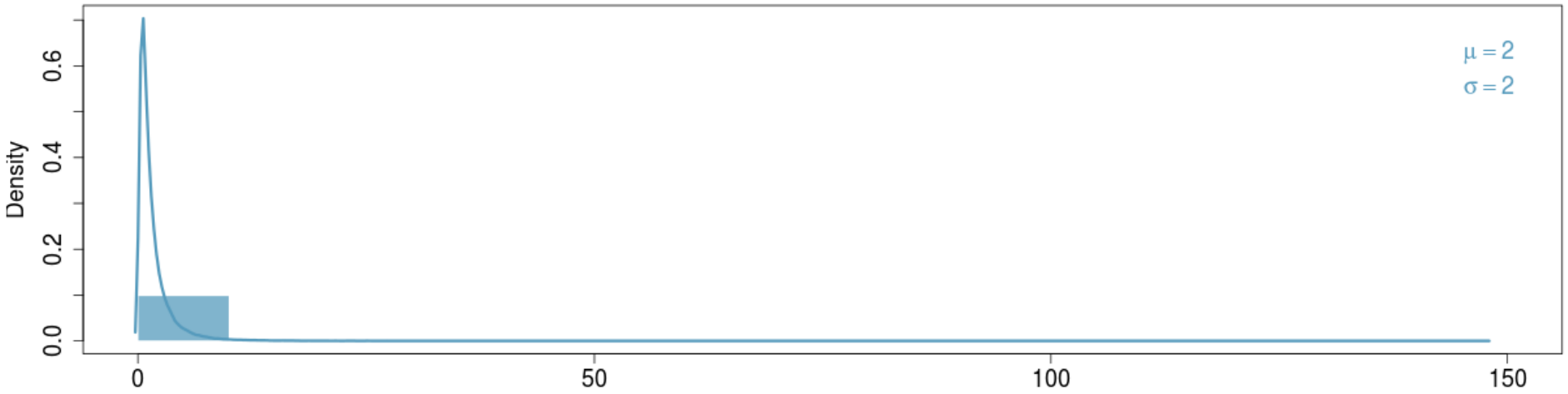
SKEWNESS VS CENTRALITY MEASURES

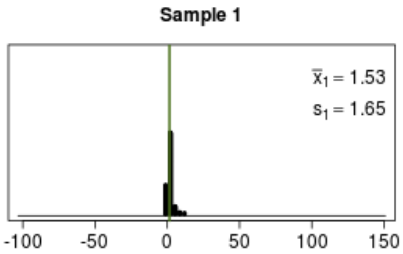
Medium skew

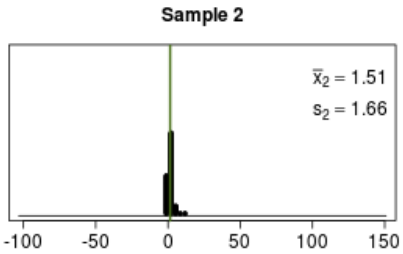


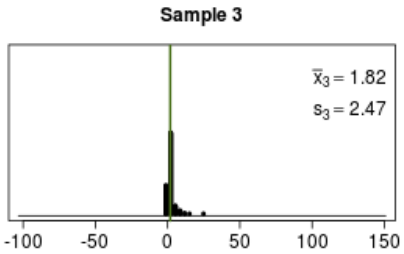
SKEWNESS VS CENTRALITY MEASURES

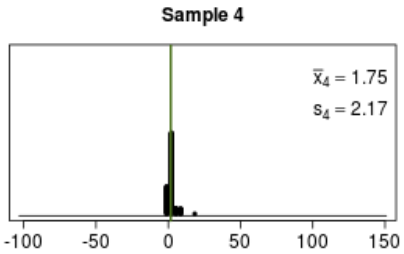
High skew



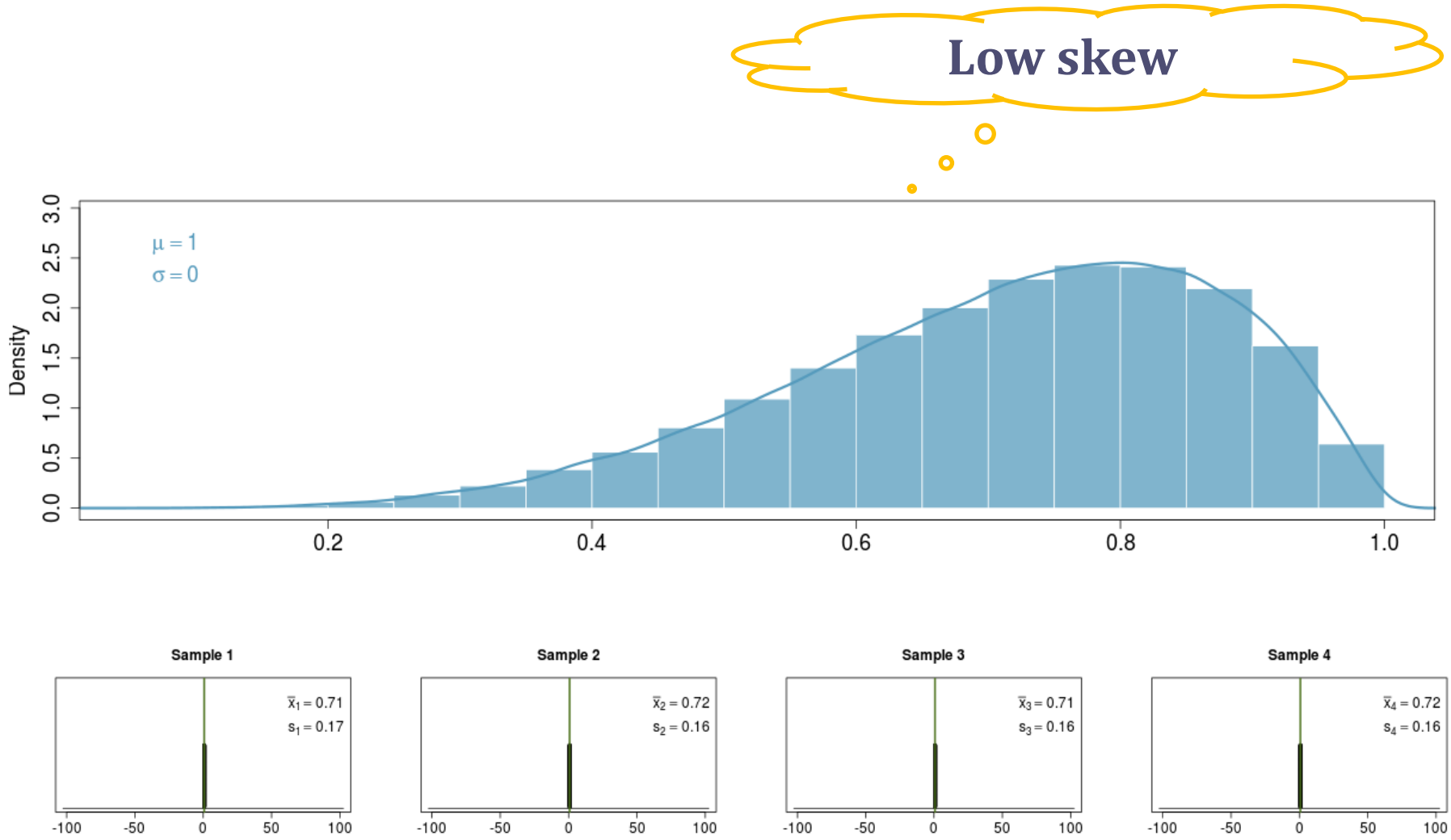






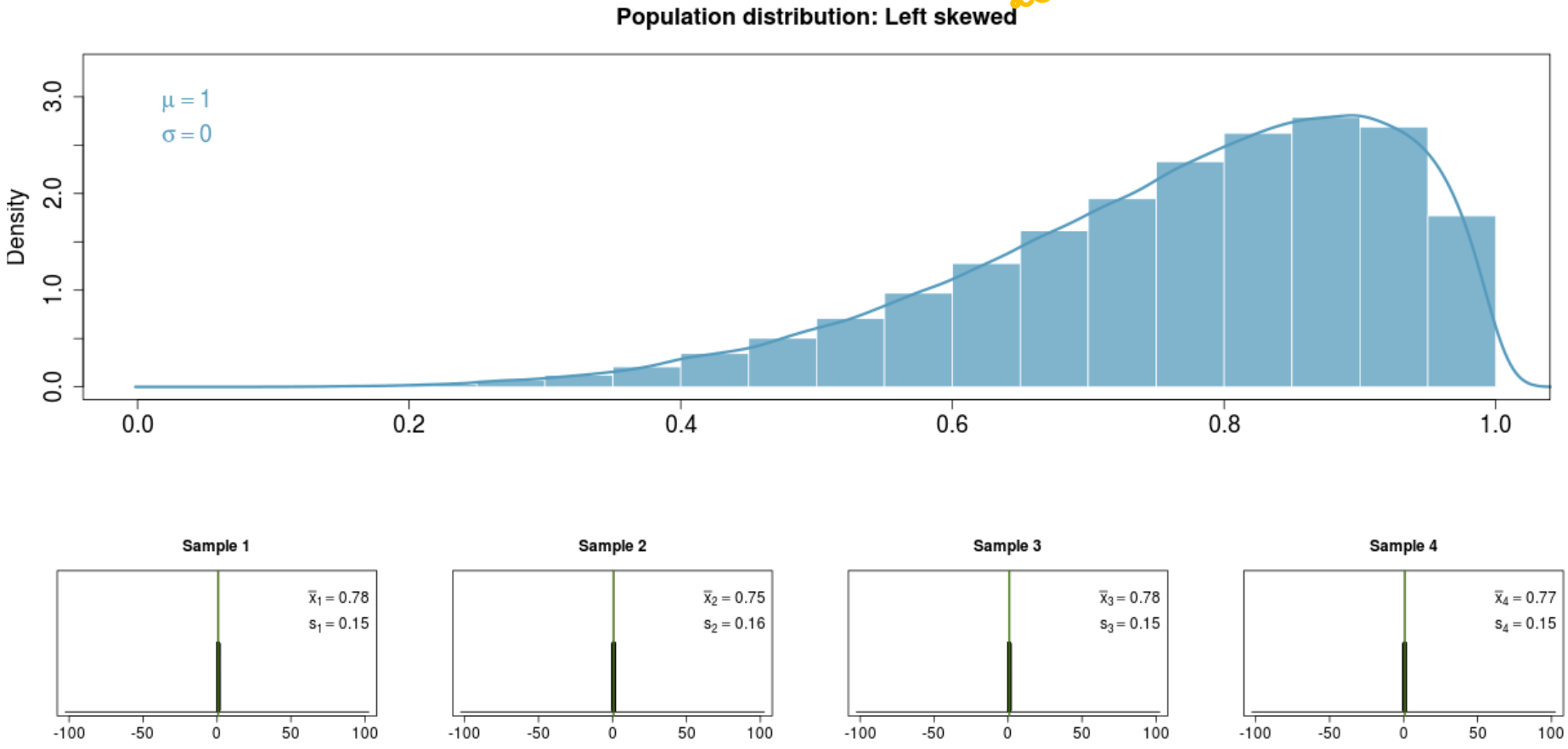


SKEWNESS VS CENTRALITY MEASURES



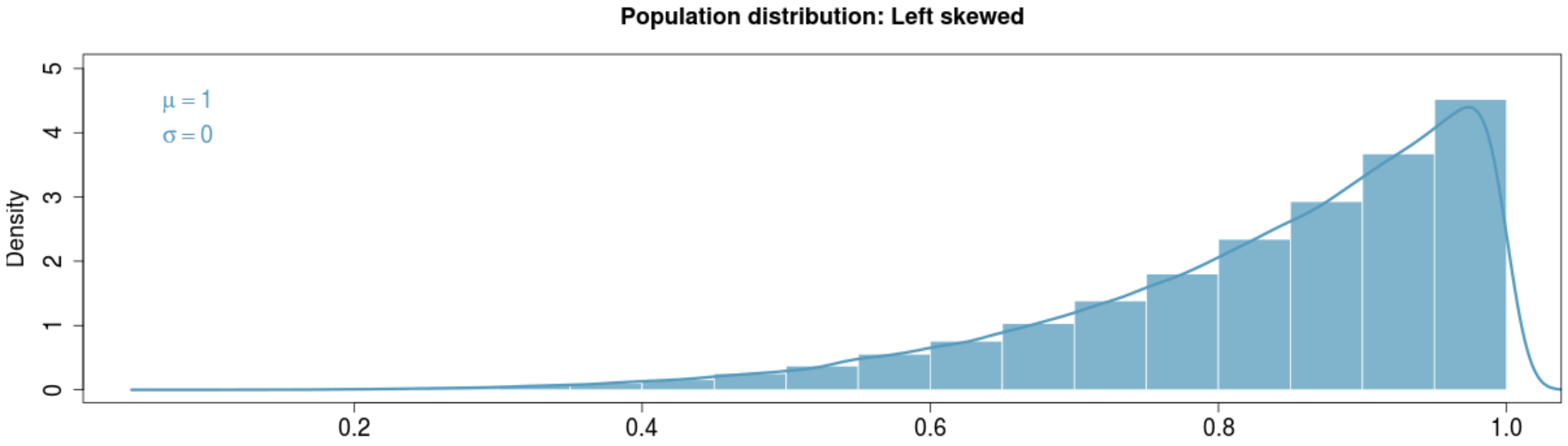
SKEWNESS VS CENTRALITY MEASURES

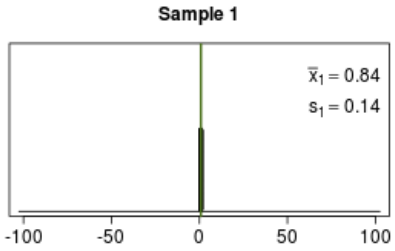
Medium skew

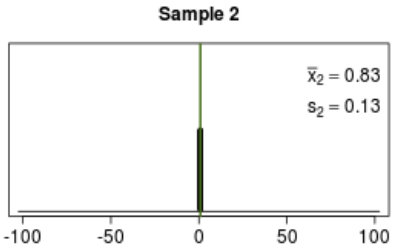


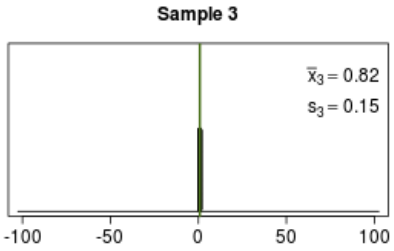
SKEWNESS VS CENTRALITY MEASURES

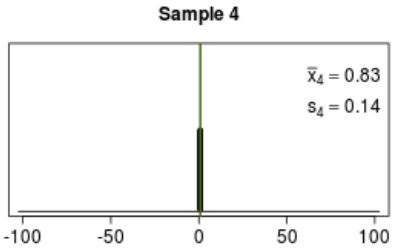
High skew











REMEMBER!

http://www.sagepub.com/upm-data/43350_4.pdf

Table 4.1 Measures of Central Tendency and Dispersion by Level of Measurement

Level of Measurement	Measures of Central Tendency	Measures of Dispersion
nominal	**mode**	**percent distribution**
ordinal	**median** mode	**minimum and maximum** **range** **percentiles** percent distribution
interval/ratio	**mean** median mode	**variance** **standard deviation** minimum and maximum range percentiles percent distribution