

Key Characteristics of a Data Set

- Every data set is accompanied by important background information. In a statistical study, **always ask the following questions:**
- **Who?** What **cases** do the data describe? **How many** cases does a data set have (sample size)?
- **What?** How many **variables** does the data set have? Are these variables continuous or categorical?
- **Why?** **What purpose** do the data have? Do the data contain the information needed to answer the questions of interest?

Table to Summarize Study Cohort

Table I

	Positive (N=22)	Negative (N=158)	p value
Age			0.614 ¹
Mean (SD)	65.77 (7.09)	65.00 (7.31)	
Median (Q1, Q3)	66.50 (61.25, 71.00)	66.00 (60.25, 70.00)	
Min - Max	50.00 - 76.00	40.00 - 80.00	
Race			0.075 ²
White	16 (72.7%)	136 (89.5%)	
AA	5 (22.7%)	14 (9.2%)	
other	1 (4.5%)	2 (1.3%)	
FamilyHist			0.472 ³
No	13 (59.1%)	107 (67.7%)	
Yes	9 (40.9%)	51 (32.3%)	
GradeGroup			0.830 ²
low	6 (27.3%)	54 (34.2%)	
medium	4 (18.2%)	30 (19.0%)	
high	12 (54.5%)	74 (46.8%)	

Data Types

- Qualitative (**Categorical**): a variable places each case into one of several groups, or categories.
 - Not ordered (e.g., marriage status, race, gender, smoking status)
 - Ordered (e.g., tumor stage, age group)
- Quantitative (**Continuous or Discrete**): a variable takes numerical values for which arithmetic operations such as adding and averaging make sense. **Discrete** data can take on only integer values whereas **continuous** data can take on any value.
 - Examples: age, blood pressure; BMI; tumor size
 - length of hospital stay

Describe a Categorical Variable

- **Numeric summaries:**

- Frequency/Count
- Percentage

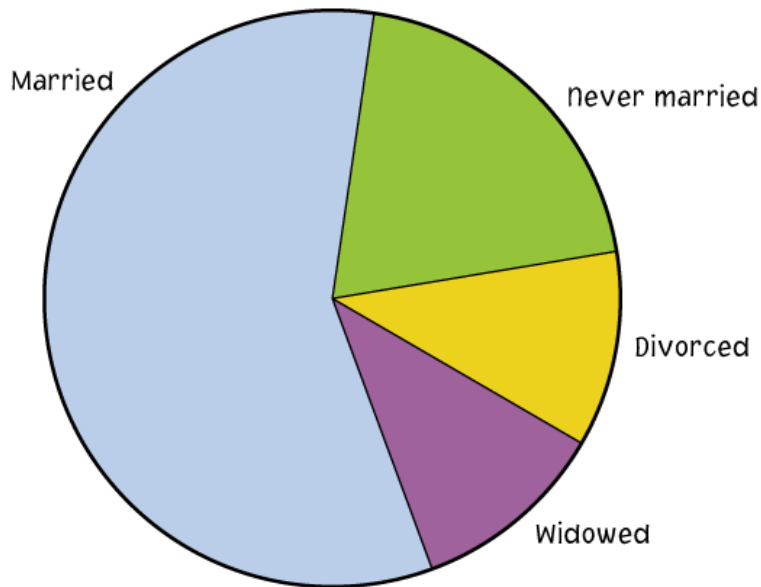
- **Graphs:**

- Bar graphs
- Pie charts

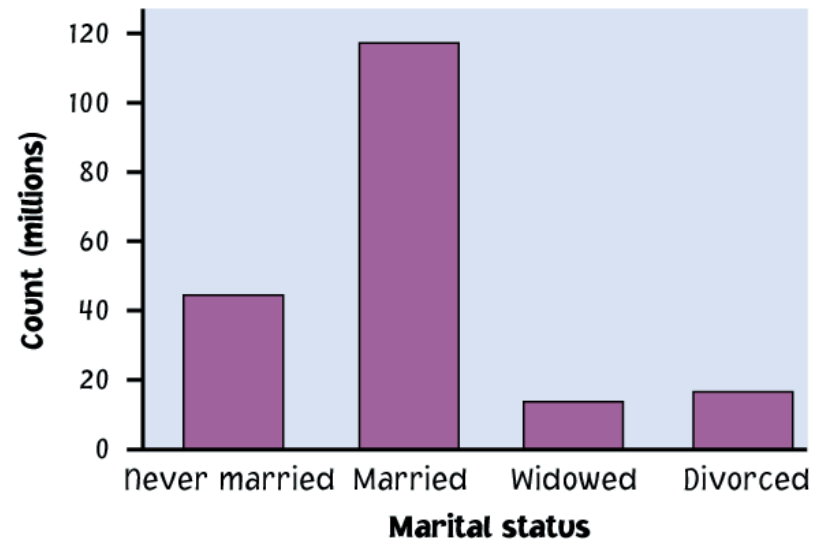


Graph for Categorical Data

Pie chart



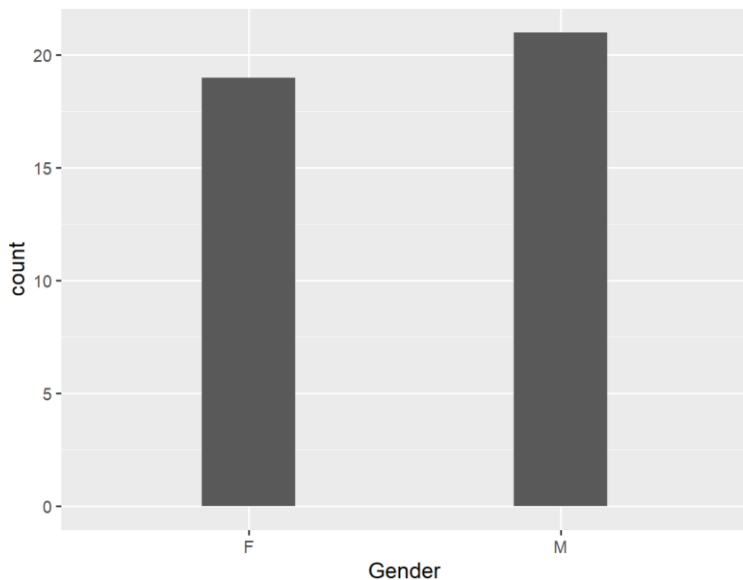
Bar graph



Describe Gender Variable Using R

R code:

```
ggplot(data = data, aes(x = Gender, y =  
after_stat(count))) + geom_bar(width=0.3)
```



R code:

```
table(Gender)
```



F	M
19	21

R code:

```
prop.table (table(Gender) )
```




F	M
0.475	0.525

Describe a Continuous Variable

- **Describing the data with numeric summaries:**
 - Measures of center:
 - Mean
 - Median
 - Measures of spread:
 - Range
 - Quartiles
 - Standard deviation (Variance= Standard deviation²)

Graph for Continuous variables

- Boxplots 
- Histograms
- Density plots

Measures of Central

- **Mean (average)**

- Sum all the observations
- divide by the total number of observations

Q: What's the mean survival times of guinea pigs in control group?

Data: 21 23 24 25 31 33 33 54

$$(21+23+24+25+31+33+33+54)/8=244/8=30.5$$

- **Median**

- The “number” such that 50% of the observations are less than the “number” and 50% are greater than the “number”

Q: Median survival time of guinea pigs?

$$(25+31)/2=28$$

Mean vs Median

- Median is robust to outliers and less influenced by extreme observations.
- For example, suppose the observation 54 in guinea pig example were 131 then median remains unchanged, but mean changes dramatically

Data: 21 23 24 25 31 33 33 ~~54~~ **131**

Spread

Measures the extent of variability in the data

- **Range** = largest – smallest value
- **Quartiles:**
 - Second quartile (**Q2**) is the median of the data
 - First quartile (**Q1**) is the median of the lower half of the data
 - Third quartile (**Q3**) is the median of the upper half of the data
- **Interquartile range (IQR):** **Q3** – **Q1**

The $1.5 \times \text{IQR}$ Rule for Outliers

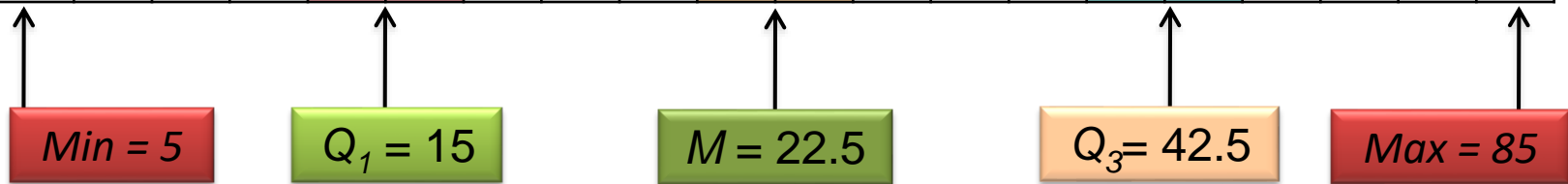
Call an observation an outlier if it falls more than $1.5 \times \text{IQR}$ above Q3 or below Q1.

Example

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45
----	----	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

Order the data

5	10	10	15	15	15	15	20	20	20	25	30	30	40	40	45	60	60	65	85
---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----



*This is an outlier
by the
 $1.5 \times IQR$ rule*

$$IQR = Q_3 - Q_1 = 42.5 - 15 = \mathbf{27.5}$$

$$Q_1 - 1.5 \times IQR = 15 - 41.25 = \mathbf{-26.25} \text{ (lower fence)}$$

$$Q_3 + 1.5 \times IQR = 42.5 + 41.25 = \mathbf{83.75} \text{ (upper fence)}$$

Any data smaller than -26.25 or larger than 83.75 is considered an outlier.

Example: Glucose Data

Random blood glucose levels (mmol/liter) from a groups of 40 first year medical students

4.7 3.6 3.8 2.2 4.7 4.1 3.6 4.0 4.4 5.1
4.2 4.1 4.4 5.0 3.7 3.6 2.9 3.7 4.7 3.4
3.9 4.8 3.3 3.3 3.6 4.6 3.4 4.5 3.3 4.0
3.4 4.0 3.8 4.1 3.8 4.4 4.9 4.9 4.3 6.0

Calculate Summary Statistics

R code:

```
summary(data$Glucose)
```



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.200	3.600	4.000	4.055	4.525	6.000

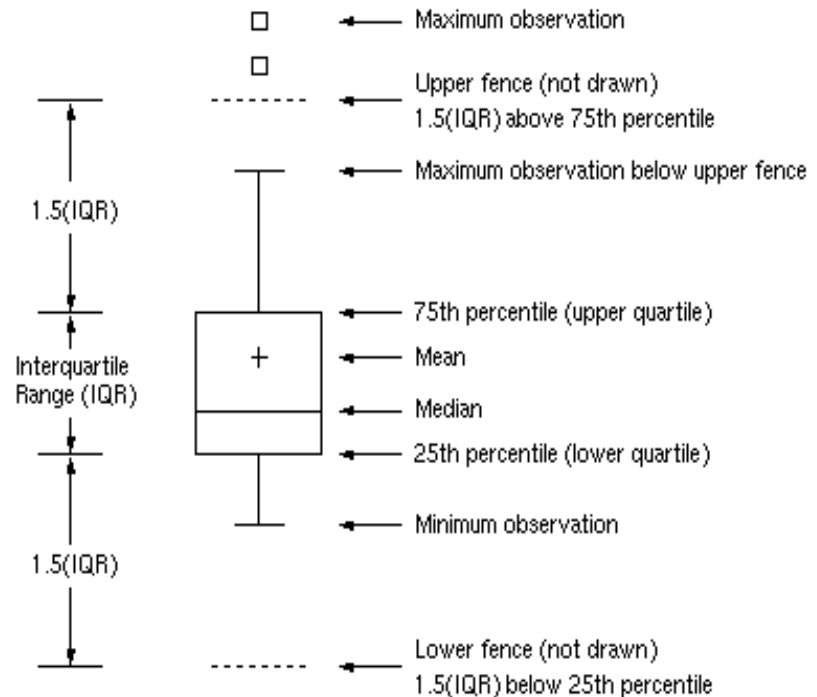
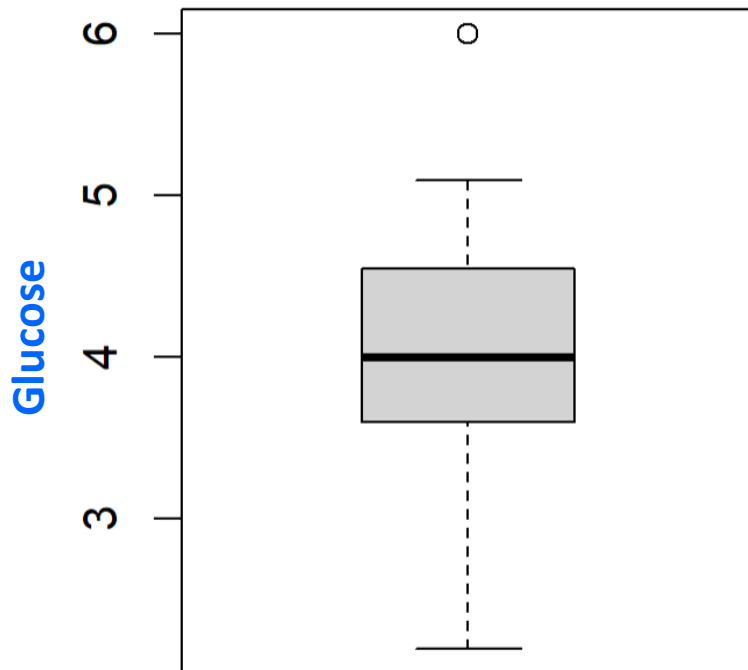
Graphically depicting these number summaries using **Boxplot**

Boxplot

R code:

Boxplot (Glucose, **outline** = TRUE)

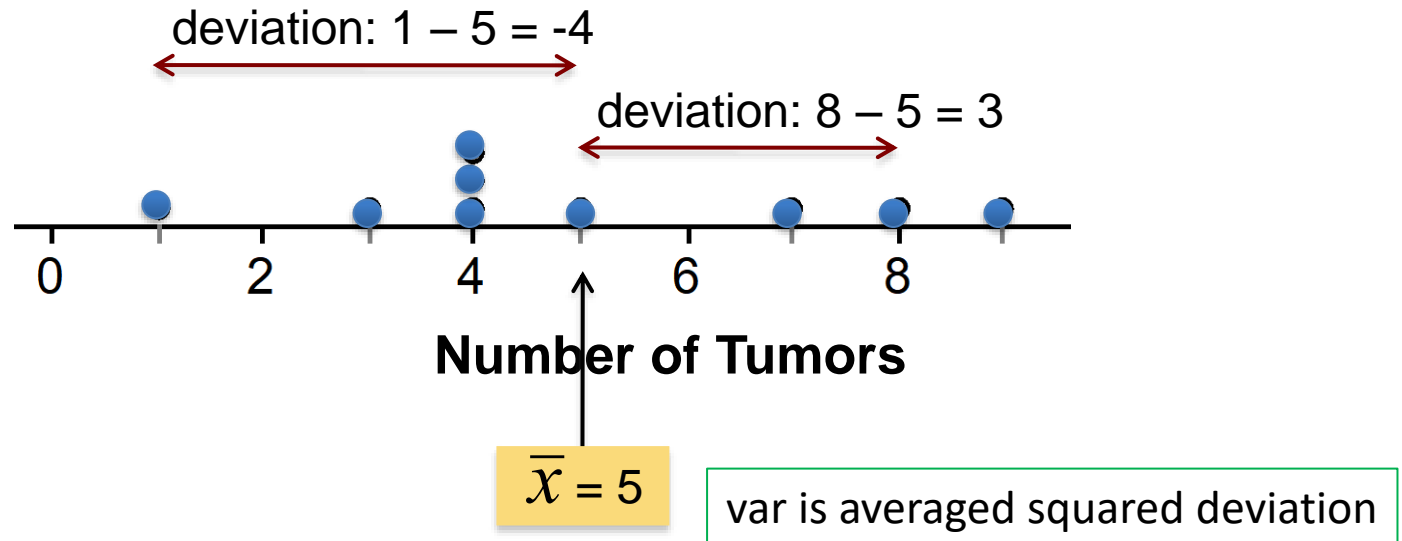
*if **outline** is not true, the outliers are not drawn*



Measuring Spread: the Standard Deviation

The most common measure of spread looks at how far each observation is from the mean. This measure is called the **standard deviation**.

Example: Consider the following data on the number of tumors in nice mice.



Questions

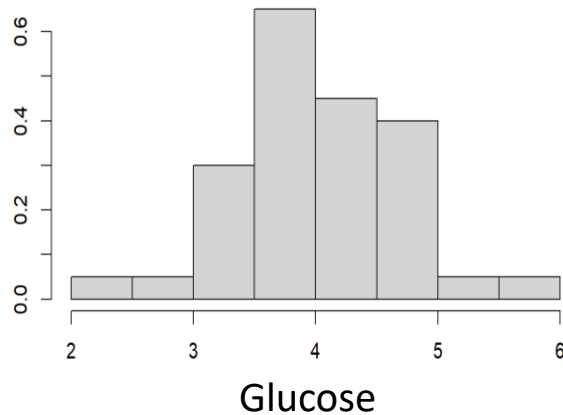
- When all observations have the same value, what is the standard deviation?
- We have learned **Mean**, **Median**, **Q1**, **Q3**, **range**, **IQR**, **variance** and **standard deviation**. Which of those measures are robust to the outlier?

Histogram and Density Curve

Histogram and Density for Glucose

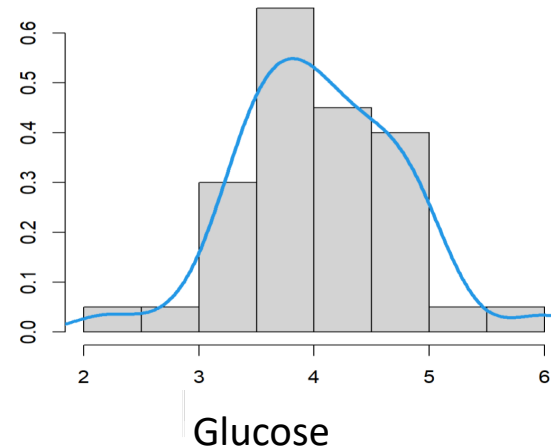
R code:

```
hist(Glucose, freq=FALSE)
```



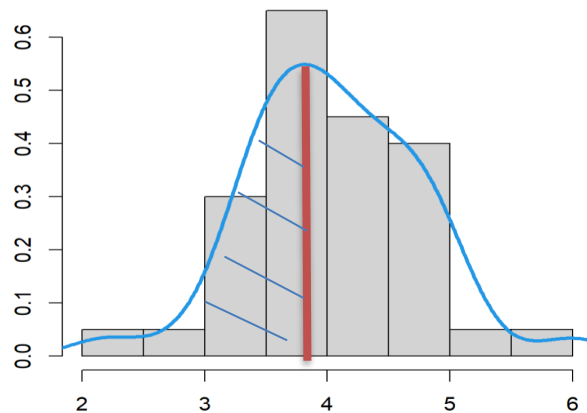
R code:

```
hist(Glucose, freq=FALSE)  
lines(density(Glucose), col=4)
```



- **Density curve** is a smooth approximation of the irregular bars of a histogram
 - It describes the overall pattern of the data: **center**, **shape** and **spread**
 - It tells us the proportion of subjects in a certain range

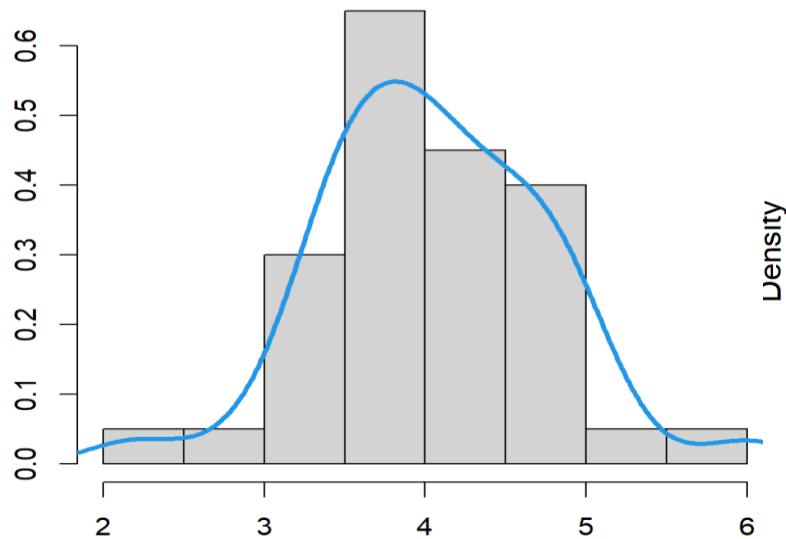
Density Plot for Glucose



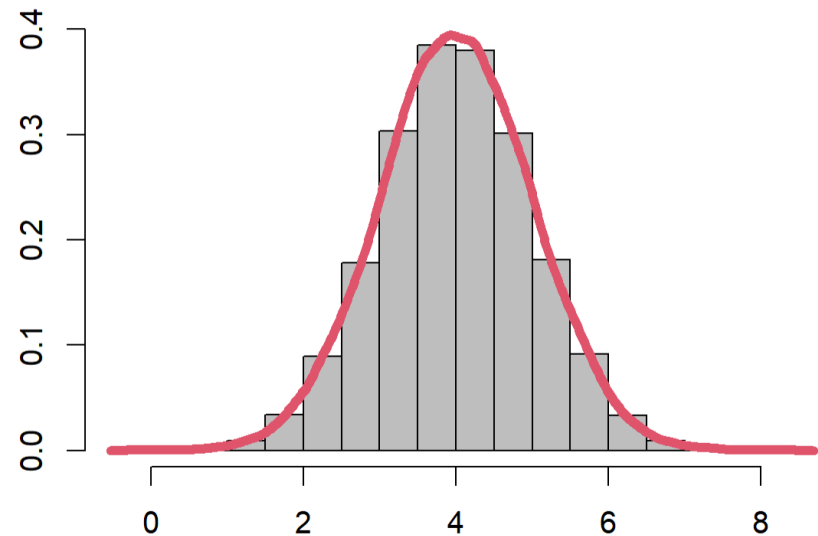
The area to the left of the red bar represents the proportion of subjects in the observed data that are less than or equal to 3.9 (low)

A Density Curve (sample vs population)

Glucose data from 40 medical students (**sample**)



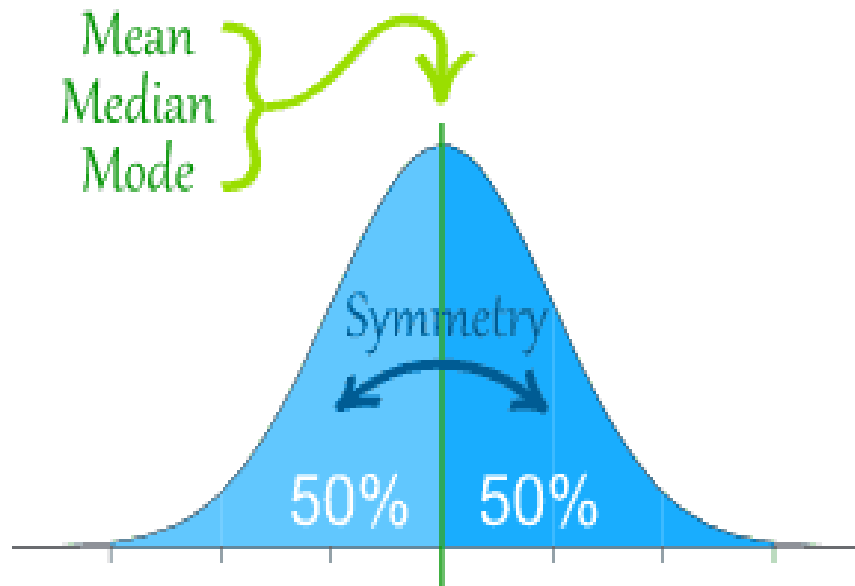
Glucose data from 1000 medical students (**population**)



Normal Distributions

One particularly important class of density curves is the class of Normal curves, which describe Normal distributions.

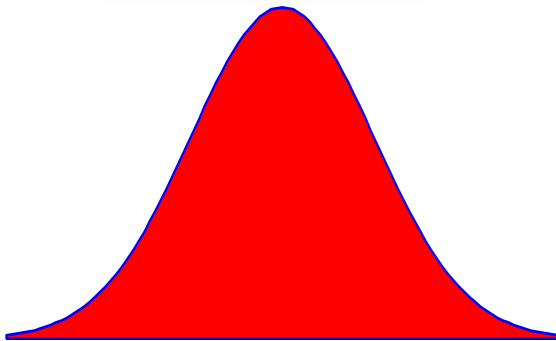
- All Normal curves are symmetric, single-peaked, and bell-shaped.



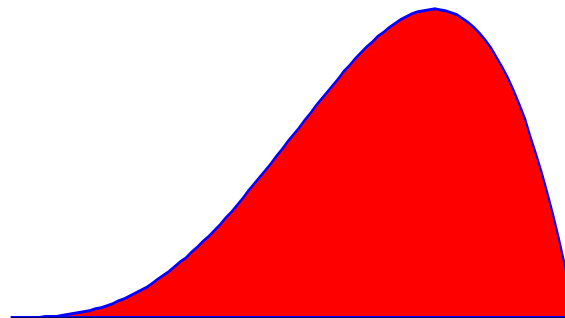
Shape

- A distribution is **symmetric** if the right and left sides of the graph are approximately mirror images of each other.
- A distribution is **skewed to the right** if the right side of the graph is much longer than the left side.
- It is **skewed to the left** if the left side of the graph is much longer than the right side.

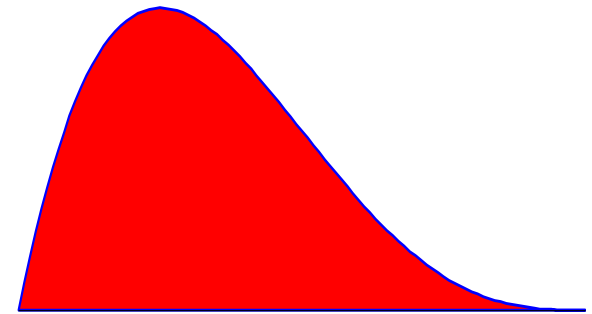
Symmetric



Left-skewed

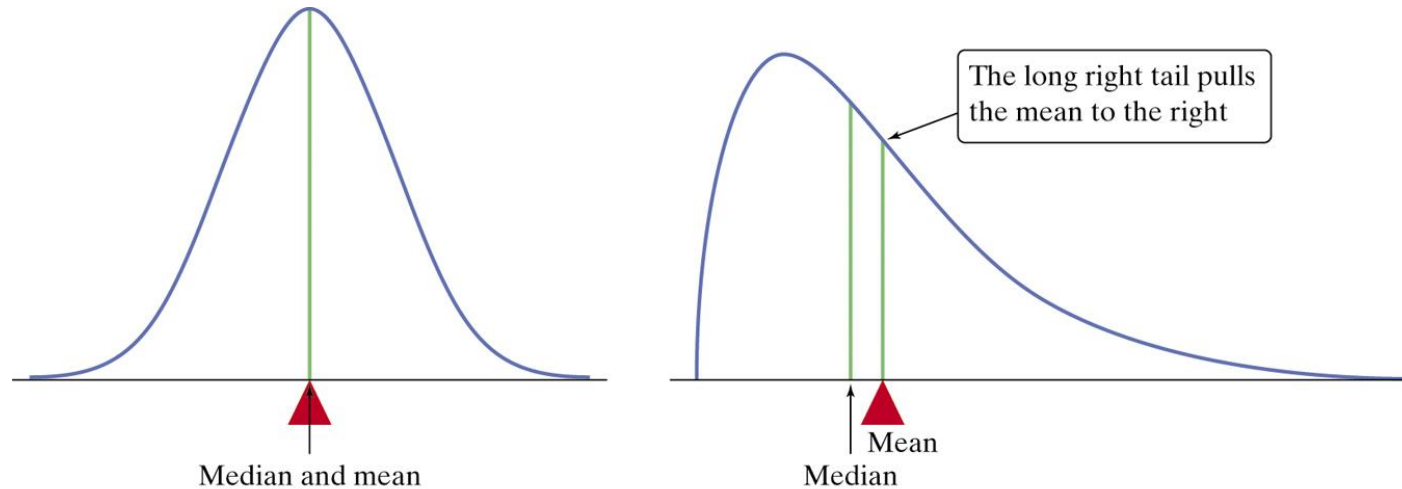


Right-skewed



Mean and Median of a Density Curves

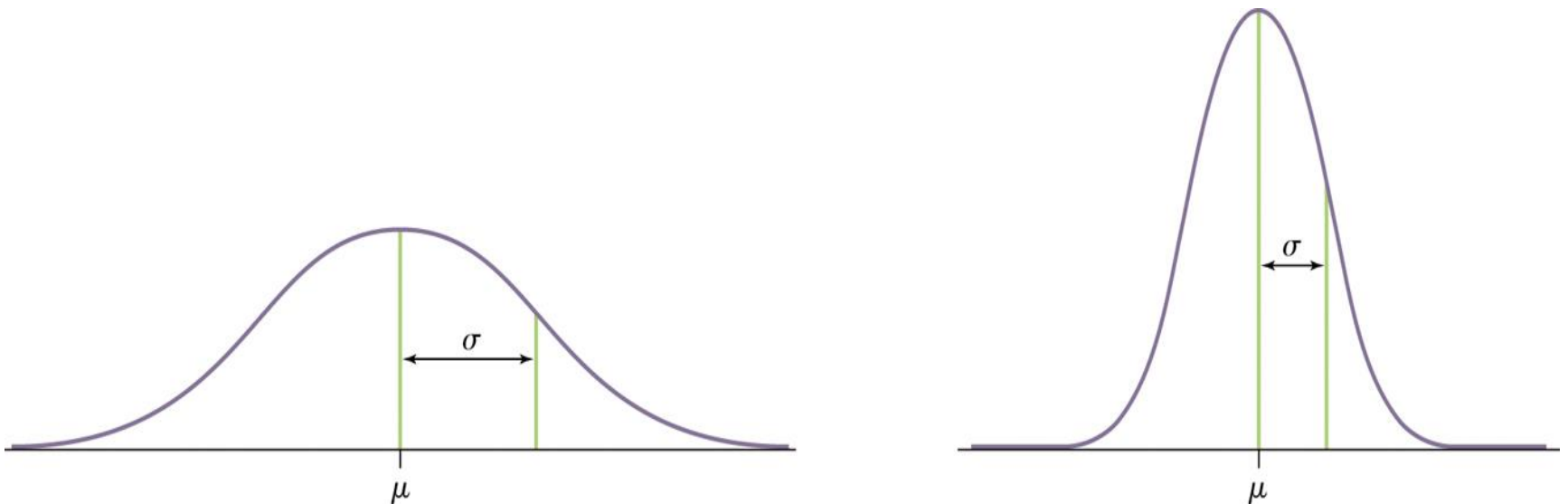
- The **median** of a density curve is the point that divides the area under the curve in half.
- The median and the mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.



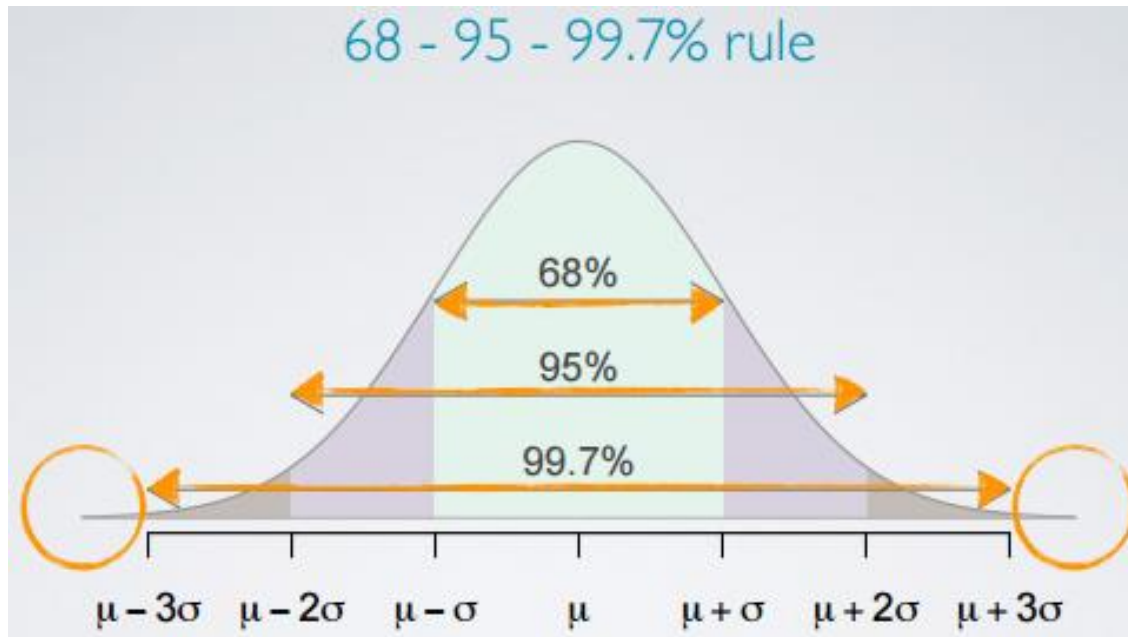
Use Mean for a symmetric distribution and median for a skewed distribution

Normal Distributions

- Any particular Normal distribution **is completely specified by two numbers**: its mean μ and standard deviation σ
- We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$



Normal Distributions



In the Normal distribution with mean μ and standard deviation σ :

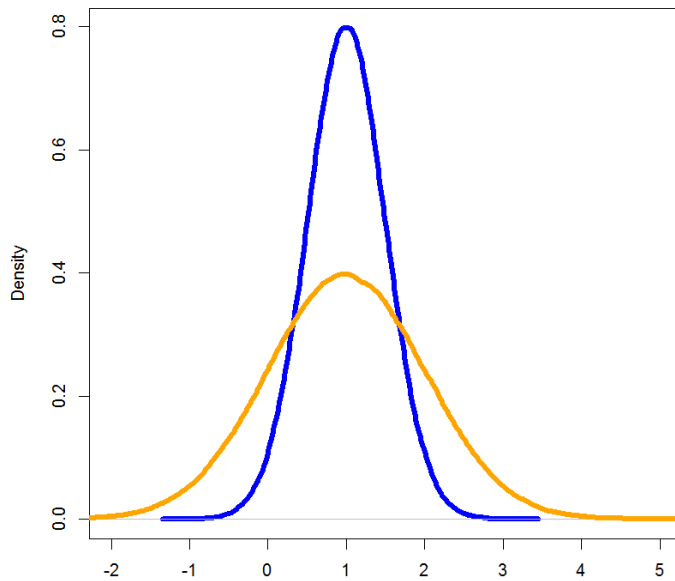
- Approximately **68%** of the observations fall within σ of μ .
- Approximately **95%** of the observations fall within 2σ of μ .
- Approximately **99.7%** of the observations fall within 3σ of μ .

Exercise: Sketch Normal Distributions

a) $N(1, 0.5)$ & $N(1, 1)$ in one figure b) $N(1, 0.5)$ & $N(10, 1)$ in one figure

Exercise: Sketch Normal Distributions

$N(1,0.5)$ & $N(1,1)$



$N(1,0.5)$ & $N(10,1)$

