STAT 206: Chapter 3 (Numerical Descriptive Measures)

- **Ideas in Chapter 3**
  - Central Tendency
  - Variation
  - Shape

- **3.1 Central Tendency**
  - 3 different ways to consider the "center" of the distribution…
  - "Balancing point" (mean/average)
  - Value divides the upper half from the lower half of the data (median)
  - Value(s) occurs most often (mode)
  - Let's call the variable X
  - $\sum x_i$ means

- **MEAN – arithmetic average**
  - "Balancing point" (MEAN/average)
  - Let's use our variable X with
    x (or x_i) = value for one record (x_1, x_2, …, x_n) and n = number of values
  - $\bar{X}$ ("x-bar") is the mean of the variable X
  - **Sample mea**n is
  - $\overline{X} =$

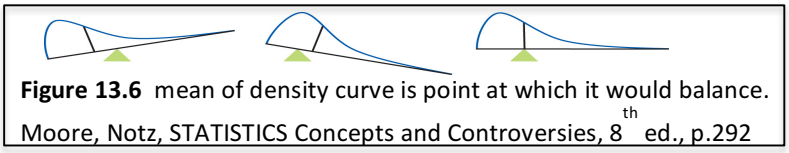


Figure 13.6 mean of density curve is point at which it would balance.
Moore, Notz, STATISTICS Concepts and Controversies, 8th ed., p.292

- **Example**: what is the (typical) MEAN time it takes you to get ready in the morning? Measure time between when you get up until you leave your home (rounded to the nearest minute) for 10 days.

Day	1	2	3	4	5	6	7	8	9	10
Time	39	29	43	52	39	44	40	31	44	35

$\overline{X} =$

  - Notice:
    -
    -
    -
    -

- **Example**: what is the (typical) MEAN time it takes you to get ready in the morning? Measure time between when you get up until you leave your home (rounded to the nearest minute) for 10 days.

Day	1	2	3	4	5	6	7	8	9	10
Time	39	29	**103**	52	39	44	40	31	44	35

$\overline{X} =$

  - Notice:
    -
    -

- **Median – middle value**
  - MEDIAN is

    - NOT affected by extreme values
    - MUST RANK IN ORDER
    - $\mathbf{MEDIAN} = \frac{n+1}{2}$ ranked value
      - If n is *ODD,*

      - If n is *EVEN*,

- **Example**: what is the (typical) MEAN time it takes you to get ready in the morning? Measure time between when you get up until you leave your home (rounded to the nearest minute) for 10 days.

Day	1	2	3	4	5	6	7	8	9	10
Time	39	29	43	52	39	44	40	31	44	35

Values	29	31	35	39	39	40	43	44	44	52
Ranks	1	2	3	4	5	6	7	8	9	10

  - Median =

    - Consider the extreme value example…

Ranks	1	2	3	4	5	6	7	8	9	10
Values	29	31	35	39	39	40	44	44	52	103

    - **Example 2**: Calories for 7 breakfast cereals. Compute the median.

Values	80	100	100	110	240	190	200
Ranks	1	2	3	4	5	6	7

- **MODE** is

  - Extreme values do NOT affect the MODE
  - There may be one mode, two modes (bi-modal), three modes (tri-modal), etc.
  - OR there may be NO mode if all values are unique

- **EXAMPLE**: times to get ready in the morning (again)

Values	29	31	35	39	39	40	43	44	44	52

- **GEOMETRIC MEAN**: When you want to measure the rate of change of a variable over time, use the GEOMETRIC MEAN (instead of the *arithmetic* mean)
  - **Geometric Mean** is

    - $\overline{X}_G =$

  - The **GEOMETRIC MEAN** rate of return measures the average percentage return of an investment per time period
  - $\bar{R}_G = [(1+R_1)\times(1+R_2)\times \ldots \times(1+R_n)]^{\frac{1}{n}} - 1$, where R_i = rate of return in period ***i***

  - **EXAMPLE**: The percentage change in the Russell 2000 Index of the stock prices of 2,000 small companies was -5.5% in 2011 and 14.6% in 2012. Compute the GEOMETRIC rate of return.
    - $\overline{R}_G =$

    - The geometric mean rate of return in the Russell 2000 Index

- **3.2 Variation and Shape**
  - **RANGE** =

  - **EXAMPLE**: times to get ready in the morning (and again)

Values	29	31	35	39	39	40	43	44	44	52

  - **VARIANCE, STANDARD DEVIATION**
    - **VARIANCE** is

$$s^2 = \frac{\sum(x - \bar{X})^2}{(n-1)}$$

    - **STANDARD DEVIATION** is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(x - \bar{X})^2}{(n-1)}}$$

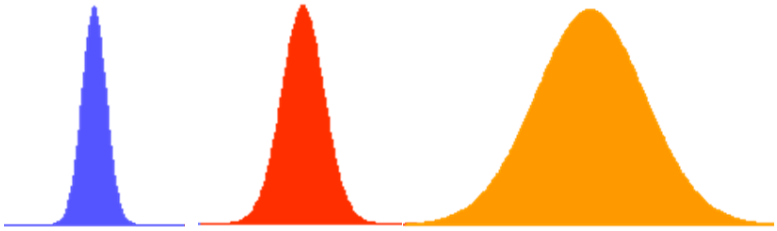
    - Can be thought of as

    - REMEMBER! Neither the variance nor the standard deviation can ever be NEGATIVE

- **Example**: Scores for CLASS A: 30, 65, 70, 76, 93, 99
  Scores for CLASS B: 68, 72, 73, 73, 74, 77
  What is the difference? Find the standard deviation for each class.

Class A		
x	**x-xbar**	**$(x\text{-}xbar)^2$**
30		
65		
70		
76		
93		
99		

- Standard deviation, σ or s, **controls the spread**. That is, the larger the value of s, the more spread out or "variable" the data are.



- **Coefficient of Variation** is

  - **CV =**

  - **EXAMPLE**: times to get ready in the morning (and again)

Values	29	31	35	39	39	40	43	44	44	52

    - $\bar{X}$ = 39.6 and s = 6.77

    - $CV =$

  - **Coefficient of Variation** measures the scatter in the data **relative to the mean.** That is, CV is expressed as a percentage since it is a relative measure
    - Useful when comparing two or more sets of data that are measured in different units (see EXAMPLE 3.7, p. 112)

- **Z score** is

  - **Z =**

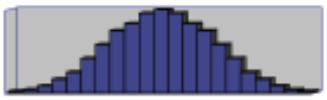
  - **Z** is a unit of measure of the number of standard deviations
    - If positive, ABOVE the mean
    - If negative, BELOW the mean
  - **Z** helps identify outliers
    - In general, Z < -3.00 or Z > 3.00 indicates an outlier value

  - **EXAMPLE**: times to get ready in the morning (and again). $\bar{X}$ = 39.6 and s = 6.77.
    What is the Z-score for 39 minutes to get ready?

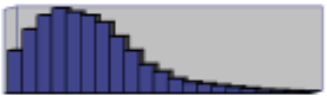
- **Shape of a variable – pattern of distribution**
  - **SKEWNESS**: extent to which data values are not symmetrical around the mean

Shape of a variable – pattern of distribution

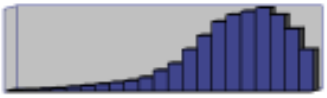
- SKEWNESS: extent to which data values are not symmetrical around the mean
- **Symmetric**



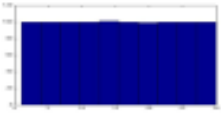
- **Skewed to the right**



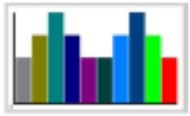
- **Skewed to the left**



- **Uniform**



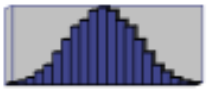
- **Bimodal**



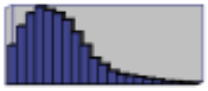
- **mean vs. median**

mean vs. median

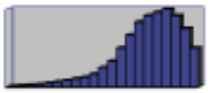
- The mean is strongly influenced by a few extreme observations
- The median is not strongly influenced by a few extreme observations
- If the distribution is
  - symmetric → mean median



  - skewed right → mean median



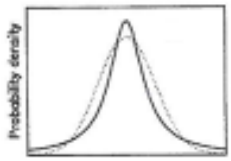
  - skewed left → mean median



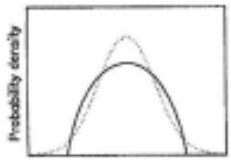
- KURTOSIS:

Shape of a variable – pattern of distribution

- KURTOSIS – extent to which values that are very different from the mean affect the shape of the distribution
  - For symmetric unimodal distributions
    ➔positive kurtosis indicates heavy tails and "peakedness"



➔negative kurtosis indicates light tails and flatness relative



relative to the normal distribution

- Normal distribution has kurtosis = 3

Illustrations Source: Lawrence DeCarlo, *On the Meaning and Use of Kurtosis*, ***Psychological Methods***, 1997, Vol. s, No.3, 292-307 21

- **3.3 Exploring Numerical Data**
  - Let's consider Measures of Position:
    PERCENTILES, QUARTILES, and 5-NUMBER SUMMARY
    - **PERCENTILE**:
    - **QUARTILES**:
    - **5-NUMBER SUMMARY** includes:

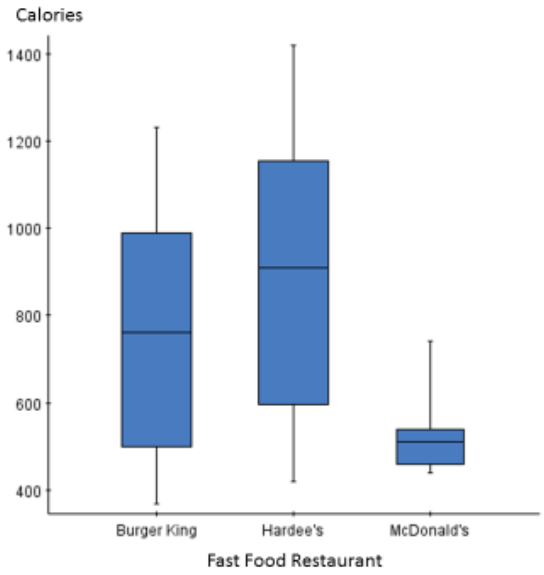
Example: Use the table (at the right) to find the 5-number summaries for calories in hamburgers for each of the three restaurants.

(source: www.aplorie counter.com/fast-food.php).

	Restaurant	Type	Calories
1	Burger King	Whopper Jr.	370
2	Burger King	Whopper Jr. (cheese)	410
3	Burger King	Double Hamburger	410
4	Burger King	Double Cheeseburger	500
5	Burger King	Double Stacker	610
6	Burger King	Whopper	670
7	Burger King	Whopper (cheese)	760
8	Burger King	Triple Stacker	800
9	Burger King	Double Whopper	900
10	Burger King	Double Whopper (cheese)	990
11	Burger King	Quad Stacker	1000
12	Burger King	Triple Whopper	1130
13	Burger King	Triple Whopper (cheese)	1230
1	Hardee's	Low Carb Thickburger	420
2	Hardee's	Double Hamburger	420
3	Hardee's	Double Cheeseburger	510
4	Hardee's	Cheeseburger	680
5	Hardee's	Mushroom N' Swiss Thickburger	720
6	Hardee's	Thickburger	910
7	Hardee's	Bacon Cheese Thickburger	910
8	Hardee's	Grilled Sourdough Thickburger	1030
9	Hardee's	Six Dollar Burger	1060
10	Hardee's	Double Thickburger	1250
11	Hardee's	Double Bacon Cheese Thickburger	1300
12	Hardee's	Monster Thickburger	1420
1	McDonald's	Double Cheeseburger	440
2	McDonald's	Big N' Tasty	460
3	McDonald's	Quarter Pounder (cheese)	510
4	McDonald's	Big N' Tasty (cheese)	510
5	McDonald's	Big Mac	540
6	McDonald's	Double Quarter Pounder (cheese)	740

- Boxplots

Use 5-Number summaries to construct boxplots. To compare different groups (e.g., restaurants), side-by-side boxplots can be constructed. (We will use Burger King's to identify the boxplot features with the 5-number summary.)

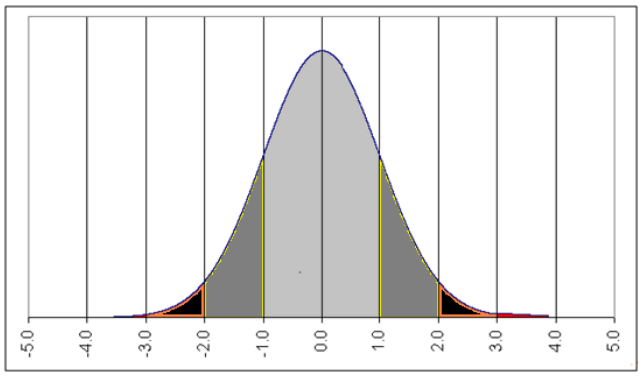


- Which restaurant has higher calories overall?

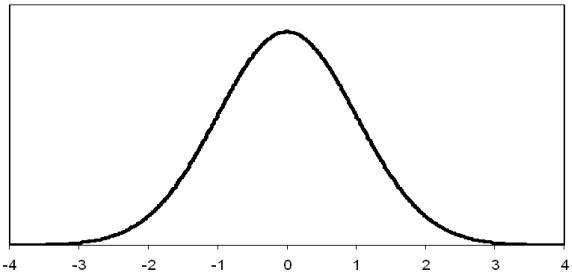
Which restaurant has the least variability in calories?

- **3.4 Numerical Descriptive Measures for a Population**
  - 3.1 and 3.2 discuss *statistics* for a *SAMPLE*
  - When data are collected for an **entire population**, analyze ***population PARAMETERS***
  - **POPULATION mean (μ)** is
  - $\mu =$
  - **POPULATION variance (σ^2)**
  - $\sigma^2 =$
  - **POPULATION standard deviation (σ)** is
  - $\sigma =$

- **Empirical Rule for normal distributions**
  - Remembering that in many/most(?) data sets, a large portion of the values tend to cluster somewhere near the mean
  - For normal (bell-shaped, symmetric) distributions, we are able to use the Empirical Rule
  - Within 1 std dev of the mean (gray area) ~ 68%
  - Within 2 std dev of the mean (gray + yellow) ~ 95%
  - Within 3 std dev of the mean (gray + yellow + orange) ~ 99.7%



- **Example**: The Health and Nutrition Examination Study of 1976-1980 (HANES) studied the heights of adults (aged 18-24) and found that the heights follow a normal distribution with the following:
  Women Mean (μ): 65.0 inches standard deviation (σ): 2.5 inches
  Men Mean (μ): 70.0 inches standard deviation (σ): 2.8 inches
  - Find the proportion of men with heights between 67.2 inches and 72.8 inches.



- **Chebyshev Rule**
  - Can't use the Empirical Rule for heavily skewed data sets
  - Chebyshev rule states that for any data set, regardless of shape, the percentage of values found within *k* standard deviations of the mean must be *at least*:
  - % (within *k* std dev) $= \left(1 - \frac{1}{k^2}\right) \times 100\%$
  - **% of Data Values Around the Mean**

Interval	**Chebyshev (any distribution)**	**Empirical (normal)**
$(\mu - \sigma, \mu + \sigma)$	*at least 0%*	*~ 68%*
$(\mu - 2\sigma, \mu + 2\sigma)$	*at least 75%*	*~ 95%*
$(\mu - 3\sigma, \mu + 3\sigma)$	*at least 88.89%*	*~ 99.7%*

  - **Example**: A population of 2-liter bottles of cola is known to have a mean fill-weight of 2.06 liter and a standard deviation of 0.02 liter. However, the shape of the population is unknown, and you cannot assume that it is bell-shaped. Describe the distribution of fill-weights.

- **3.6 Descriptive Statistics: Pitfalls and Ethical Issues**
  - Massive amounts of data available online
  - Should you report the mean or the median?
  - Should you report 5-number summary or the variance and standard deviation?
  - Unethical is pertinent findings are deliberately NOT reported if they are detrimental to a particular position