# Unit 7 Day 1 – Theoretical vs. Empirical

Name _____

Empirical Probability: (directly observed)
estimate that the event will happen based on how often it occurs from data collection

Theoretical Probability:

$\frac{\text{\# ways event can occur}}{\text{total \# outcomes}}$ * what should happen

**Examples:** You roll a die 100 times. The outcomes are recorded in the table below.

| Outcome | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Frequency | 18 | 22 | 15 | 16 | 14 | 14 |

1. What is the theoretical probability of rolling a 5?

$\frac{1}{6}$

2. What is the empirical probability of rolling a 6?

$\frac{14}{100} = \frac{7}{50}$

A survey was conducted to determine students' favorite breeds of dogs. Each student chose only one breed.

| Dog | Collie | Spaniel | Lab | Boxer | PitBull | Other |
|---|---|---|---|---|---|---|
| # | 10 | 15 | 35 | 8 | 5 | 12 |

3. What is the probability that a student's favorite dog breed is Lab?

$\frac{35}{85} = \frac{7}{17}$

4. Is this an example of empirical or theoretical probability?

empirical

5. Russell and Ryan roll 2 dice 50 times and record their results.

| Sum of the rolls of two dice |
|---|
| 3, 5, 5, 4, 6, 7, 7, 5, 9, 10, 12, 9, 6, 5, 7, 8, 7, 4, 11, 6, 8, 8, 10, 6, 7, 4, 4, 5, 7, 9, 9, 7, 8, 11, 6, 5, 4, 7, 7, 4, 3, 6, 7, 7, 7, 8, 6, 7, 8, 9 |

a. What is their empirical probability of rolling a 7?

$\frac{13}{50} = 26\%$

b. What is the theoretical probability of rolling a 7?

$\frac{1}{6} \approx 17\%$

c. How do a) and b) compare?

empirical is higher

6. Geologists say that the probability of a major earthquake occurring in the San Francisco Bay area in the next 30 years is about 90%. Is this empirical or theoretical probability?

based on previously observed data

Vocabulary:

**Statistics:** collection + analysis of data

**Population:** entire set of individuals/objects in which we're interested

**Sample:** a subset of a population

qL **Qualitative Data:** Cant be measured, descriptions, qualities

qN **Quantitative Data:** data can be measured, numbers, quantity

Determine if qualitative or quantitative:

qL 1. Gender qN 2. Temperature qL 3. Zip code qL 4. How different foods taste

qN 5. Number of days during the past week a 21 year old had a least one alcoholic beverage

D **Discrete variable:** countable # of values

C **Continuous variable:** no spaces btw values, not countable

Determine if discrete or continuous:

D 6. Number of heads after 5 flips of a coin D 7. # of cars at a drive through between 12-1

C 8. Distance a 2007 Toyota Prius can travel in city driving conditions with a full tank of gas

C 9. Weight of a newborn C 10. Time to complete a task

D 11. Number of accidents on Hwy 64

**Types of Gathering Data**

I. Survey: collections of information about items in a population or sample

II. Observational Study: Sample being studied is measured as is (no influence by researchers)

Investigators observe subjects and measure variables of interest **without assigning treatments** to the subjects.

III. Experimental Study: researchers apply treatments to sample, then observes the effect

IV. Simulations: The use of a mathematical model to recreate a situation, often repeatedly, so that the likelihood of various outcomes can be more accurately estimated.

Identify each as one of the four from above.

Simulation a. On average, suppose a baseball player hits a home run once in every 10 times at bat, and suppose he gets exactly two "at bats" in every game. Estimate the likelihood that the player will hit 2 home runs in a single game.

experimental b. Forty volunteers suffering from insomnia were divided into two groups. The first group was assigned to a special no-desserts diet while the other continued desserts as usual. Half of the people in these groups were randomly assigned to an exercise program, while the others did not exercise. Those who ate no desserts and engaged in exercise showed the most improvement.

observational c. In 2001, a report in the *Journal of the American Cancer Institute* indicated that women who work nights have a 60% greater risk of developing breast cancer. Researchers based these findings on the work histories of 763 women with breast cancer and 741 women without the disease.

experimental d. Scientists at a major pharmaceutical firm investigated the effectiveness of an herbal compound to treat the common cold. They exposed each subject to a cold virus, and then gave him or her either the herbal compound or a sugar solution known to have no effect. Several days later, they assessed the patient's condition, using a cold severity scale of 0 to 5.

**Types of Sampling Design (methods used to choose the sample from the population)**

I. Simple Random Sample: a sample of ppl chosen so every person has equal chance of being selected

II. Stratified Random Sample: pop. divided into groups w/ similar traits then SRS

III. Cluster Sample: based on location, pick a spot + sample w/in

IV. Convenience Sample: easy method to gather data, not random

Identify the sampling design with the above choices.

SRS 1. Suppose we were to take 100 APEX students – put each students' name in a hat. Then randomly select 100 names from the hat.

Stratified 2. Suppose were to take ALL APEX students and divide them by grade level. Put their names in a hat and randomly select 25 names from each grade.

cluster 3. Suppose we were to take all classrooms during 2nd period and randomly select students in 10 of those classrooms.

convenience 4. Surveys left on tables at restaurants

convenience 5. Stand at the main entrance of Apex High School and stop friendly-looking students to survey.

Stratified 6. The Educational Testing Service (ETS) needed a sample of colleges. ETS first divided all colleges into groups of similar types (small public, small private, etc.) Then they randomly selected 3 colleges from each group.

cluster 7. A county commissioner wants to survey people in her district to determine their opinions on a particular law up for adoption. She decides to randomly select blocks in her district and then survey all who live on those blocks.

SRS 8. The names of 70 contestants are written on 70 cards. The cards are placed in a bag, and three names are picked from the bag.

convenience 9. To avoid working late, the quality control manager inspects the last 10 items produced that day.

**Bias:** anything that causes the data to be wrong (outcomes favored)

AFM Unit 7 Day 1 HW

Introduction to Statistics Worksheet

**I. Classify the variable as qualitative or quantitative.**

1. Number of siblings
2. Grams of carbohydrates in a doughnut
3. Number on a football player's jersey
4. Assessed value of a house
5. Number of un-popped kernels in a bag of ACT microwave popcorn
6. Phone number
7. Student ID number

**II. Determine whether the quantitative variable is discrete or continuous.**

8. Runs scored in a season by Albert Pujols
9. Volume of water lost each day through a leaky faucet.
10. Length (in minutes) of a country song
11. Number of sequoia trees in a randomly selected acre of Yosemite National Park
12. Temperature on a randomly selected day in Memphis, Tennessee
13. Internet connection speed in kilobytes per second
14. Points scored in an NCAA basketball game
15. Air pressure in pounds per square inch in an automobile tire

***III. Determine whether the study depicts an observation study, experimental study, simulation or survey.***

16. Researchers wanted to know if there is a link between proximity to high-tension wires and the rate of leukemia children. To conduct the study, researchers compared the incidence rate of leukemia for children who lived

# Unit 7 Day 2 – Frequency Distribution

**Frequency:** how often something occurs

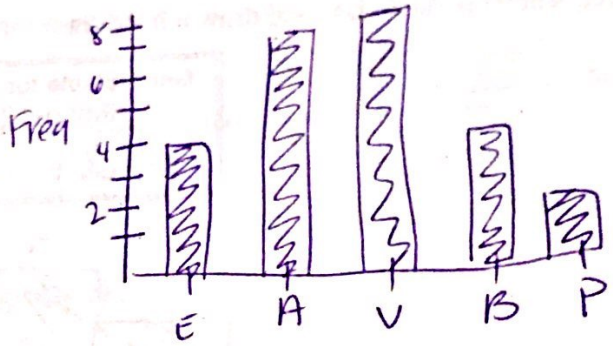
**Frequency Distribution:** a set of data & their frequencies

**Relative Frequency Distribution:** a set of data & their percentages

Ex. 1: A television network has asked 25 viewers to evaluate a new police drama. The possible evaluations are (E)xcellent, (A)bove average, a(V)erage, (B)elow average, (P)oor. After the show, the 25 evaluations were as follows: A,V,V,B,P,E,A,E,V,V,A,E,P,B,V,V,A,A,A,E,B,V,A,B,V.

a. Construct a frequency table and a relative frequency table for this list of evaluations.
b. Draw a bar graph of the frequency distribution of TV viewers' responses from #1.

| Eval | Freq | Rel Freq |
|---|---|---|
| E | 4 | 16% |
| A | 7 | 28% |
| V | 8 | 32% |
| B | 4 | 16% |
| P | 2 | 4% |
| Total | 25 | 100% |




Ex. 2: The bar graph shows the number of Atlantic hurricanes over a period of years. Use it to answer the following questions.

a. What was the smallest number of hurricanes in a year during this period? Largest? 4 19

b. What number of hurricanes per year occurred most frequently? 11

c. How many years were the hurricanes counted? 58 years

d. In what percentage of the years were there more than ten hurricanes? 25/58 = 43.1%

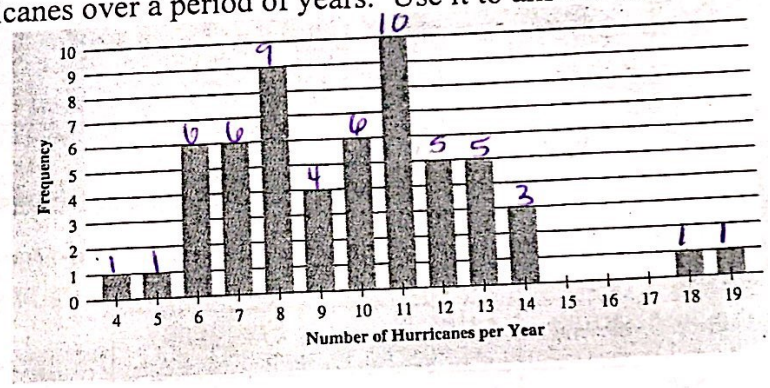


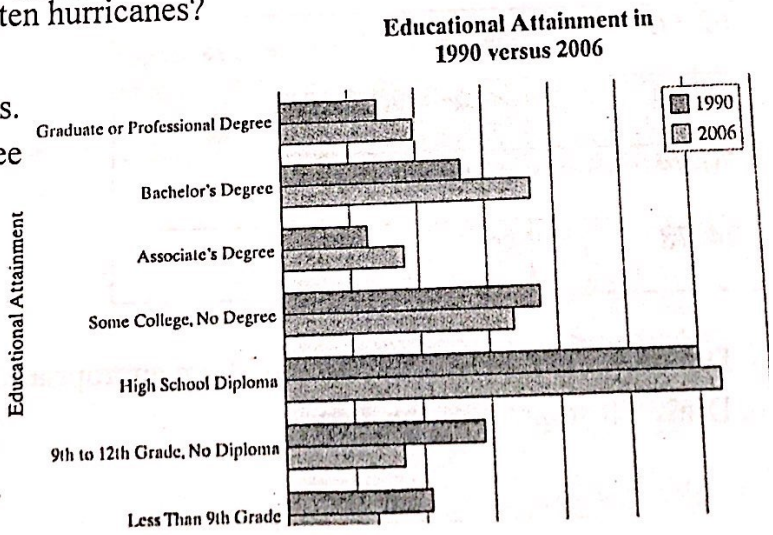

FIGURE 14.3 Number of hurricanes per year.

3. Use the side by side bar graph to answer the questions.

a. What percent of Americans had their associate's degree in 1990? 6-7% 2006? 8-9%

b. What percent of Americans at least had a high school diploma in 1990? 75% 2006? 85%

c. Are a greater proportion of Americans dropping out of college before earning a degree? no



**Histogram:** graph w/ bars w/ no spaces, intervals

**Class:** each interval

**Class Interval:** range of each class (should be equal)

**Class Limits:** upper + lower values in each interval

**Class Marks:** midpoints of class limits

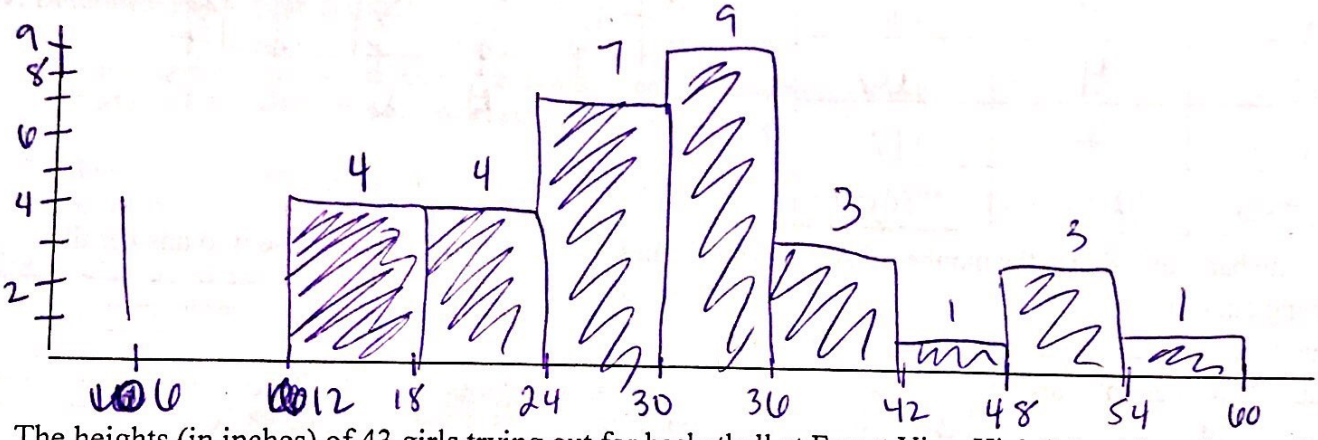
5. The winning scores for the first 33 Super Bowls are: 35, 33, 16, 23, 16, 24, 14, 24, 16, 21, 32, 27, 35, 31, 27, 26, 27, 38, 46, 39, 42, 20, 55, 20, 37, 52, 30, 49, 27, 35, 31, 34, and 23.

a. Determine an appropriate class interval for this data. $\frac{55-14}{7} = 5.8$ so 6

b. Find the frequency for each class and draw a histogram for the data.

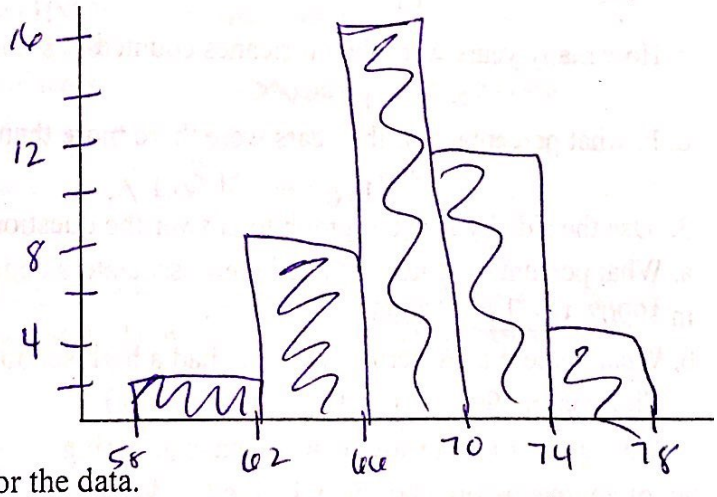
**General rule for determining intervals:**
(Range)/(# of desired bars)=(max-min)/( # of desired bars)

**Typically the number of desired bars will be 7.




6. The heights (in inches) of 43 girls trying out for basketball at Forest View High School have been tallied in the chart below.

| Class Limits | Tally | Frequency |
|---|---|---|
| 58-62 | \|\| | 2 |
| 62-66 | 卌 \|\|\| | 8 |
| 66-70 | 卌 卌 卌 \| | 16 |
| 70-74 | 卌 卌 \|\| | 12 |
| 74-78 | 卌 | 5 |




a. Determine if the class intervals of 4 are appropriate for the data.

b. Draw a histogram of the data.

**Box and Whisker Plots:** displays mean, quartiles, extreme values

**Median:** middle # of data (L to G)

**Minimum:** lowest # **Maximum:** highest #

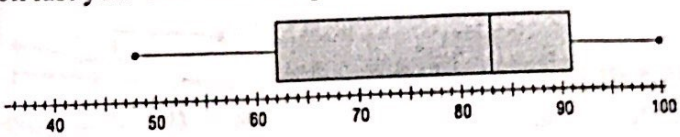
**Quartile 1:** med of lower ½ **Quartile 3:** median of upper ½

**Interquartile Range:** Q3 − Q1 (½ of the data)

**Outliers:** extreme values 1.5IQR beyond Q1 & Q3

More than: Q3 + 1.5(IQR) less than: Q1 − 1.5(IQR)

7. Use the following box plot of student test scores on last year's advanced algebra mid-year exam.




a. What is the median score?

83

b. What is the interquartile range?

91 − 62 = 29

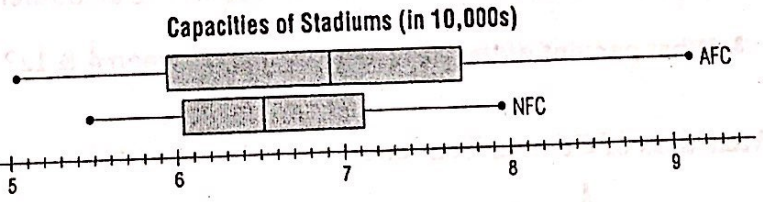
c. What **percent** of the students scored between 62 and 91?

50%

d. What is the interval of scores of students who ranked **below** the lower quartile?

48 to 62

8. The National Football League is separated into two parts-the American Football Conference (AFC) and the National Football Conference (NFC). Here are separate box plots of the capacities of the football stadiums used by the AFC and NFC.




a. What is the median capacity in each conference?

AFC → 69,000 NFC → 65,000

b. What is the size of the largest stadium in each conference?

AFC → 91,000 NFC → 80,000

c. About what percent of the stadiums in the AFC hold fewer than 60,000 people?

25%

9. Below is a stem plot of the amount of money spent by 25 shoppers at a grocery store. The stem is in $10 units.

| | |
|---|---|
| 0 | 3 8 |
| 1 | 0 1 7 8 9 |
| 2 | 0 0 3 6 8 |
| 3 | 1 3 4 7 |
| 4 | 2 5 5 |
| 5 | 0 |
| 6 | 5 |
| 7 | 2 6 |
| 8 | |
| 9 | 7 |
| 10 | |
| 11 | 3 |

a. Find the median. $31

b. Find the lower quartile. $18.5

c. Find the upper quartile. $47.3

d. Determine if there are any outliers. yes, $97, $113

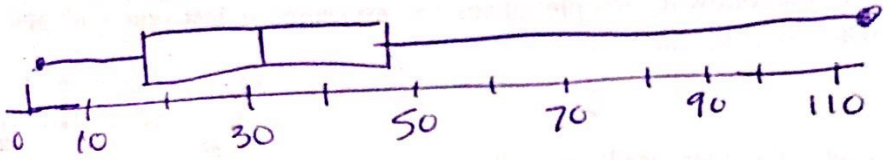
e. Construct a box and whiskers plot.

IQR = 29

1.5(IQR) = 43.5

47.3 + 43.5 = $90.8

anything above is outlier




**Steps in the Calculator:**

1. Enter the data in L1
2. Stat > Calc > 1 – 1-Var Stats (enter twice) **Scroll down to see what you need!**

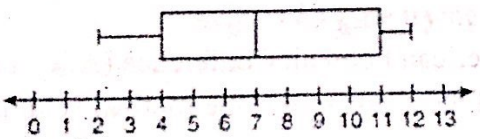
n = the number of values — minX = the smallest value

Q1 = first quartile — Med = median

Q3 = third quartile — maxX = the largest value

## Measures of Central Tendency

For the box plot shown:




IQR = 11 − 4 = 7

1.5(7) = 10.5

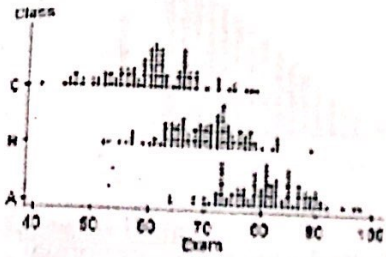
1. If -1 were a data value in this set, would it be an outlier? no
2. If 22 were a data value in this set, would it be an outlier? yes, 11 + 10.5 = 21.5
3. What percent of the data have values between 4 & 12? 75%

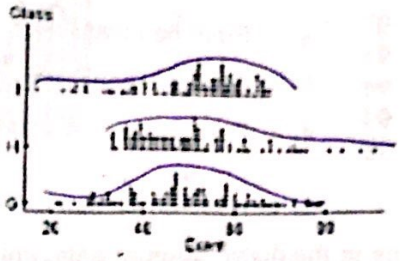
## Measures of Central Tendency

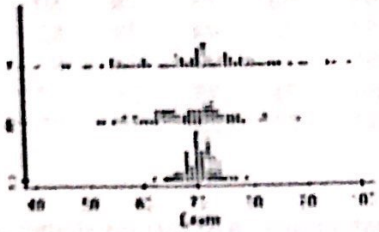
**Mean:** Average **Median:** Middle #

**Mode:** # most often **Range:** max − min

**What strikes you as the most distinctive difference among the distributions of exam scores in the different classes?**







1. data shifted 2. shape 3. spread

**Deviation:** how far from mean a data value is

**Variance:** the sum of the squares of deviations from mean divided by n or n-1 * the average of squared diff from mean

**Standard Deviation:** square root of variance
*measures how spread out #s are

Population Standard Deviation versus Sample **Standard Deviation**

Population: Divide by 'n' is used when the sample is the population $\sigma x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

Sample: Divide by 'n-1' is used for a sample because it gives a better estimate of the population mean

$$Sx = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

"n" = # data values

4. Find the mean and standard deviation for the following data: {3,5,6,7,9,11,22}. Use the table on the right.

mean = 9

$\sigma x$ Std dev = 5.83

| Value | Mean | Deviation from mean | Square of deviation from mean |
|---|---|---|---|
| 3 | 9 | -6 | 36 |
| 5 | 9 | -4 | 16 |
| 6 | 9 | -3 | 9 |
| 7 | 9 | -2 | 4 |
| 9 | 9 | 0 | 0 |
| 11 | 9 | 2 | 4 |
| 22 | 9 | 13 | 169 |
| | | | 238 |

**USING CALC: Stat > Calc** **1: 1-Var Stats**

Symbol for mean: $\overline{X}$

Sx = sample standard deviation calculated using (n-1)

σx = population standard deviation calculated using "n"

5. Use the following scores of a mathematics class on the midyear exam.

| | | | | | | |
|---|---|---|---|---|---|---|
| 43 | 68 | 73 | 78 | 80 | 88 | 92 |
| 52 | 70 | 74 | 78 | 82 | 89 | 93 |
| 65 | 70 | 75 | 78 | 85 | 90 | 94 |
| 66 | 71 | 75 | 78 | 87 | 90 | 94 |
| 67 | 72 | 76 | 79 | 87 | 90 | 98 |

a. Find the mean. 78.5

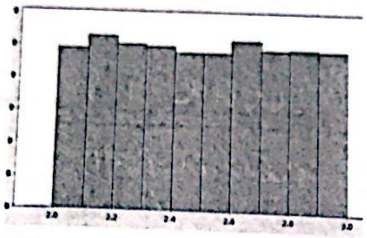
b. Find the range. 55

c. Find the standard deviation.

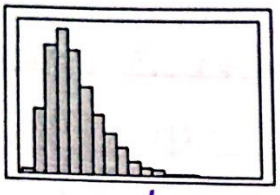
≈12

Histograms have various shapes according to the distribution of data: uniform, skewed, symmetric, or normal.

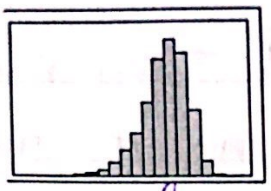
**Uniform**



**Skewed**



Skewed Right

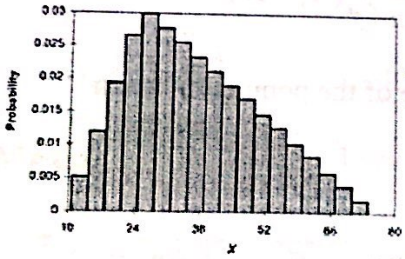


Skewed Left

**Symmetric**

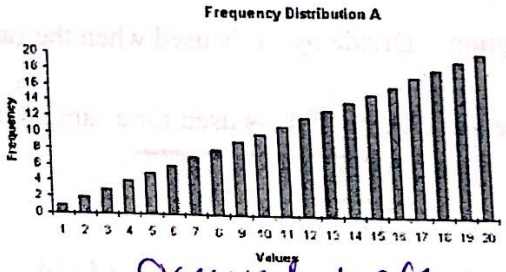
**Identify the distribution as right skewed, left skewed, symmetric, or normal.**

7.



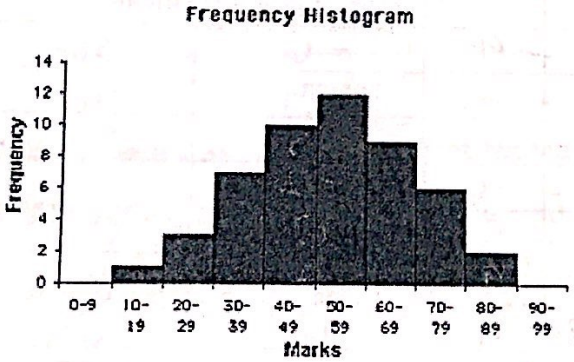

Skewed Right

8.



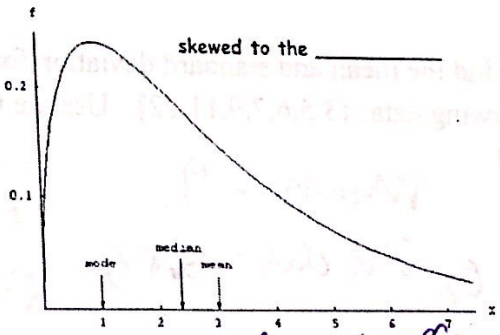

Skewed Left

9.



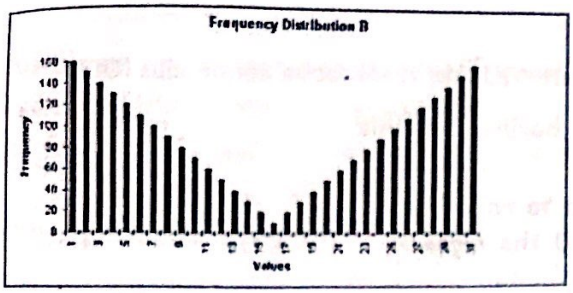

Symmetric

10.




Skewed ~~Left~~ Right

11.




Bimodal

**General Rule:**

*If the data is skewed right, the mean is greater than the median

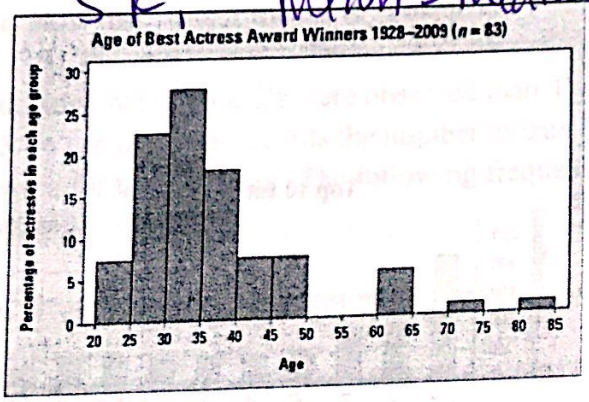
*If the data is skewed left, the mean is less than the median

*If the data is close to symmetric, the mean and median are close to each other.

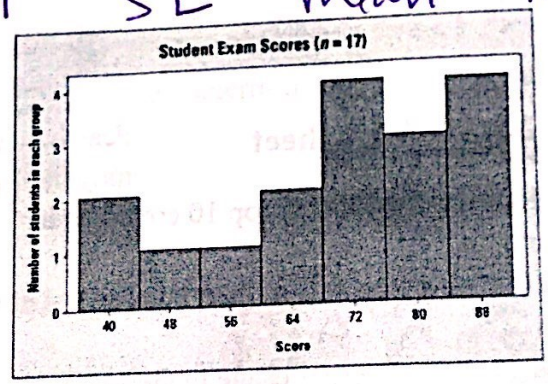
*The mean gets pulled towards to tail!

**Tell if the data is symmetric or skewed and which is greater the mean or the median in each graph.**

SR, mean > median

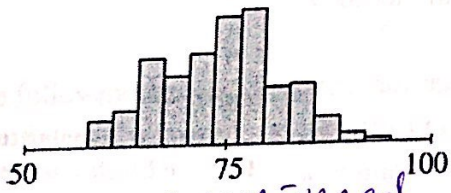



SL mean < median




Which histogram appears to have the largest **value** for the center?

Statistic A



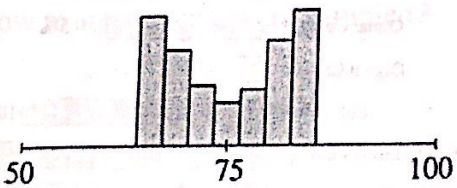

mean = med

Statistic B

B

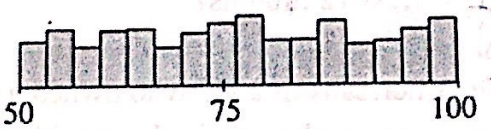

mean = med

Statistic C




mean

Statistic D



# Unit 7 Day 4 – Normal Distribution

Name_____

**Normal Distribution:** The distribution of data along a bell-shaped curve, symmetric curve that reaches its maximum height at the mean.

**Properties of Normal Curve:**
68% of the distribution is within 1 standard deviation of the mean.
95% of the distribution is within 2 standard deviations of the mean.
99.7% of the distribution is within 3 standard deviations of the mean.

**Standard Normal Distribution:** the normal distribution having a mean of 0 and a standard deviation of 1. The total area under the curve (and above the x-axis) is 1.

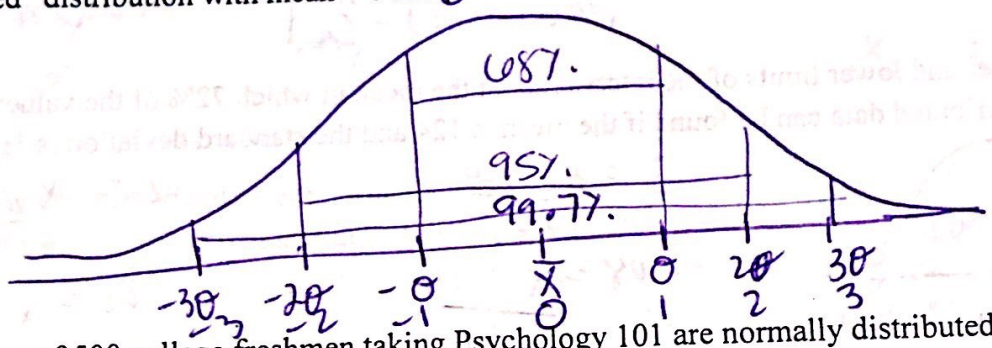
**z-score:** the # of standard deviations a data value is away from the mean

**Formula:**

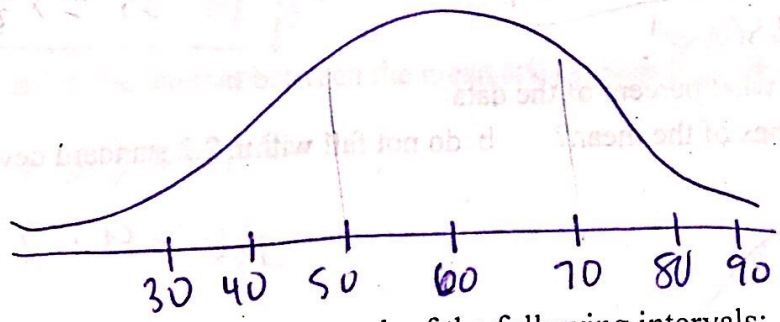
$$z = \frac{x - \bar{x}}{\sigma}$$

*The higher the standard deviation is, the more the data varies.

Draw a "normalized" distribution with mean = 0 and $\sigma$ = 1.




1. Suppose the scores of 500 college freshmen taking Psychology 101 are normally distributed. The mean score is 60 out of 100, and the standard deviation is 10. Sketch a normal curve that represents the frequency scores.




Estimate how many grades will fall between each of the following intervals:

a. 50 – 70 68%

500(.68) = 340

b. 40 – 80 95%

500(.95) = 475

c. 30 – 90 99.7%

500(.99.7) = 498 or 499

Find z score!

*d. 55 – 65

$z_{55} = \frac{55-60}{10} = -.5$

$z_{65} = \frac{65-60}{10} = .5$

$Area_{65} - Area_{55}$ = Total Area

$.69146 - .30854 = .38292(500) = 191.46$ students

*e. 45 – 75

$z_{45} = \frac{45-60}{10} = -1.5$ $z_{75} = \frac{75-60}{10} = 1.5$

$A_{75} - A_{45}$ = Total Area

$.93319 - .06681$

$= .86638(500) = 433.19$

f. P(grade < 72)

$z_{72} = \frac{72-60}{10} = 1.2$

Area $= .88493(500)$

$= 442.47$

g. P(grade > 85)


$z_{85} = \frac{85-60}{10} = 2.5$

$z = 2.5$, area $= .99379$

$1 - .99379 = .00621$

$.00621(500) = 3.1$

h. P(73 < grade < 83)

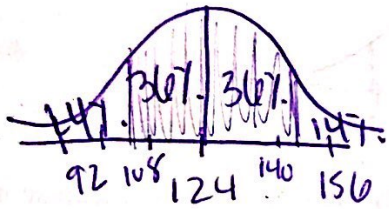
$z_{73} = \frac{73-60}{10} = 1.3$

$z_{83} = \frac{83-60}{10} = 2.3$

$.98928 - .90320 = .08608$

$.08608(500) = 43.04$

2. Find the upper ($X_2$) and lower ($X_1$) limits of the interval about the mean in which 72% of the values of a set of normally distributed data can be found if the mean is 124 and the standard deviation is 16.




$z = \frac{X_1 - 124}{16}$

$-1.08 = \frac{X_1 - 124}{16}$

$\boxed{106.72 = X_1}$

$z = \frac{X_2 - 124}{16}$

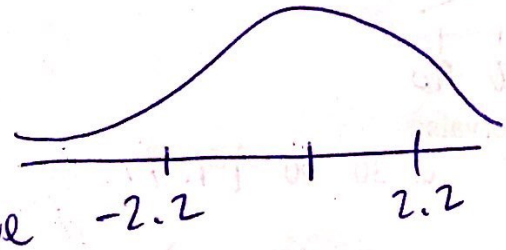
$1.08 = \frac{X_2 - 124}{16}$

$\boxed{141.28 = X_2}$

*use chart to find zscore!

3. In a normal distribution, about what percent of the data:

a. fall within 2.2 standard deviations of the mean?




Area under curve

$.98610 - .01390 = .9722$

97.22%

b. do not fall within 2.2 standard deviations of the mean?

$100 - 97.22$

$= 2.78\%$

use table for area!

4. A day is selected at random at a post office whose daily letter-handling rate is normally distributed. The mean number of letters per day is 10,000 and the standard deviation is 350. What is the probability that the post office handles between 9000 and 11,000 letters per day?

$.99788 - .00212 = \boxed{.99576}$

or

99.576%.

$z_{9000} = \frac{9000-10000}{350} = -2.857$

$z_{11000} = \frac{11000-10000}{350} = 2.857$

5. In a certain large school district, the set of all standardized mathematics scores is normally distributed with mean $\bar{x} = 540$ and standard deviation of 64. What is the probability that a student chosen at random scores between 580 and 620 on that test?

$.89435 - .73565$

$= \boxed{.1587}$

15.87%.

$z_{580} = \frac{580-540}{64} = .625$

$z_{620} = \frac{620-540}{64} = 1.25$

6. Suppose 300 values in a set of data were normally distributed.

a. How many values are within one standard deviation of the mean?

$300(.68) = 204$

b. How many values are two standard deviations of the mean?

$300(.95) = 285$

c. How many values fall in the interval between the mean and one standard deviation above the mean?

$300(.34) = 102$

# Unit 7 Day 5 – Normal Distribution

Name_____

1. A company producing stereos find they have a mean life of 7 years and a standard deviation of 1.2 years. What guarantee should the company make so they will only have to exchange fewer than 10% of all stereos sold?

$-1.28 = \frac{X - 7}{1.2}$ $\boxed{X = 5.464}$

4 area under the curve -1.28
z-score = -1.28

2. A teacher marks some exams and finds the mean is 54% and the standard deviation is 8%. The teacher then adjusts the marks by raising the mean to 60% and the standard deviation to 9%. The z-scores are kept constant. If a student scored a 76%, what would their new mark be?

$z = \frac{76 - 54}{8}$

$z = 2.75$

$2.75 = \frac{X_2 - 60}{9}$ $\boxed{X_2 = 84.75}$

* **When looking for a percent (probability):**
2nd – VARS – DIST – 2 – normalcdf

Normalcdf (X1, X2, μ, σ)

X1 lower bound X2 upper bound

μ mean σ standard dev.

* **When looking for a data value and you know the percent:**
2nd – VARS – invNorm (area to the left, μ, σ)

decimal

3. A set of 1000 values has a normal distribution. The mean of the data is 120, and the standard deviation is 20.

a. How many values are in one standard deviation from the mean?

680

b. What percent of data is in the range 110 to 130?

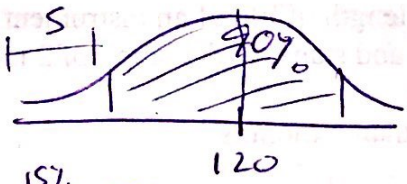
.38292

c. What percent of the data is in the range 90 to 110?

.24173

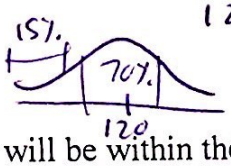
d. Find the range about the mean that includes 90% of the data?

87.1 to 152.897




e. Find the range about the mean that includes 70% of the data?

99.27 to 140.73




f. Find the probability that a value selected at random from the data will be within the limits 100 and 150?

.77454

g. Find the probability that a value selected at random from the data will be greater than 140?

.15866

h. Find the point below which 90% of the data lie?

145.63

normcdf

4. X is a normally distributed variable with mean $\mu = 30$ and standard deviation $\sigma = 4$. Find

a) $P(x < 40)$ b) $P(x > 21)$ c) $P(30 < x < 35)$

normalcdf .99379 .98778 .34435

5. A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is travelling at more than 100 km/hr?

normcdf .15866

6. For a certain type of computers, the length of time between charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. John owns one of these computers and wants to know the probability that the length of time will be between 50 and 70 hours.

normcdf .40879

7. Entry to a certain University is determined by a national test. The scores on this test are normally distributed with a mean of 500 and a standard deviation of 100. Tom wants to be admitted to this university and he knows that he must score better than at least 70% of the students who took the test. Tom takes the test and scores 585. Will he be admitted to this university?

invnorm .70 = 552.44 yes he will

8. The length of similar components produced by a company are approximated by a normal distribution model with a mean of 5 cm and a standard deviation of 0.02 cm. If a component is chosen at random:

a) What is the probability that the length of this component is between 4.98 and 5.02 cm?

normcdf .6827

b) What is the probability that the length of this component is between 4.96 and 5.04 cm?

normcdf .9545

9. The length of life of an instrument produced by a machine has a normal distribution with a mean of 12 months and standard deviation of 2 months. Find the probability that an instrument produced by this machine will last:

a) less than 7 months.

normcdf .00621

b) between 7 and 12 months.

normcdf .49379

10. The time taken to assemble a car in a certain plant is a random variable having a normal distribution with a mean of 20 hours and a standard deviation of 2 hours. What is the probability that a car can be assembled at this plant in a period of time: norm cdf

a) less than 19.5 hours?

.40129

b) between 20 and 22 hours?

.34134

A large group of students took a test in Physics and the final grades have a mean of 70 and a standard deviation of 10. If we can approximate the distribution of these grades by a normal distribution, what percent of the students:

a) scored higher than 80?

normcdf

.15866

b) should pass the test (grades≥60)?

.84134

c) should fail the test (grades<60)?

.15866

12. The annual salaries of employees in a large company are approximately normally distributed with a mean of $50,000 and a standard deviation of $20,000.

a) What percent of people earn less than $40,000?

.30233

b) What percent of people earn between $45,000 and $65,000?

.37208

c) What percent of people earn more than $70,000?

.16486

13. In a city, it is estimated that the maximum temperature in June is normally distributed with a mean of 23° and a standard deviation of 5°. Calculate the number of days in this month in which it is expected to reach a maximum of between 21° and 27°.

normcdf .44350(30 days) = 13.3 days so ~13 days

14. The mean weight of 500 college students is 70 kg and the standard deviation is 3 kg. Assuming that the weight is normally distributed, determine how many students weigh:

a) Between 60 kg and 75 kg.

normcdf

.95178(500)

≈475 students

b) More than 80 kg.

normcdf

.000429(500)

~0 students

c) Less than 64 kg.

.02275(500)

≈11 students