

Statistical Notations

Notations:

Measure	Sample Notation	Population Notation
Mean	$\bar{Y}$	$\mu_y$
Proportion	$p$	$\pi$
Standard Deviation	$S_y$	$\sigma$
Variance	$S^2_y$	$\sigma^2_y$

•  $H_1$ . = research hypothesis

Symbol	Symbol Name	Meaning / definition	Example
$P(A)$	probability function	probability of event A	$P(A) = 0.5$
$P(A \cap B)$	probability of events intersection	probability that of events A and B	$P(A \cap B) = 0.5$
$P(A \cup B)$	probability of events union	probability that of events A or B	$P(A \cup B) = 0.5$
$P(A   B)$	conditional probability function	probability of event A given event B occurred	$P(A   B) = 0.3$
$f(x)$	probability density function (pdf)	$P(a \leq x \leq b) = \int f(x) dx$	
$F(x)$	cumulative distribution function (cdf)	$F(x) = P(X \leq x)$	

Symbol	Symbol Name	Meaning / definition	Example
$E(X)$	expectation value	expected value of random variable X	$E(X) = 10$
$E(X   Y)$	conditional expectation	expected value of random variable X given Y	$E(X   Y=2) = 5$
$var(X)$	variance	variance of random variable X	$var(X) = 4$
$\sigma^2$	variance	variance of population values	$\sigma^2 = 4$
$std(X)$	standard deviation	standard deviation of random variable X	$std(X) = 2$
$\sigma_X$	standard deviation	standard deviation value of random variable X	$\sigma_X = 2$
$\tilde{x}$	median	middle value of random variable x	$\tilde{x} = 5$
$cov(X,Y)$	covariance	covariance of random variables X and Y	$cov(X,Y) = 4$
$corr(X,Y)$	correlation	correlation of random variables X and Y	$corr(X,Y) = 0.6$
$\rho_{X,Y}$	correlation	correlation of random variables X and Y	$\rho_{X,Y} = 0.6$
$\sum$	summation	summation - sum of all values in range of series	$\sum_{i=1}^4 x_i = x_1 + x_2 + x_3 + x_4$

Symbol	Symbol Name	Meaning / definition	Example
$\Sigma\Sigma$	double summation	double summation	$\sum_{j=1}^2 \sum_{i=1}^8 x_{i,j} = \sum_{i=1}^8 x_{i,1} + \sum_{i=1}^8 x_{i,2}$
$Mo$	mode	value that occurs most frequently in population	
$MR$	mid-range	$MR = (x_{max} + x_{min}) / 2$	
$Md$	sample median	half the population is below this value	
$Q_1$	lower / first quartile	25% of population are below this value	
$Q_2$	median / second quartile	50% of population are below this value = median of samples	
$Q_3$	upper / third quartile	75% of population are below this value	
$\bar{x}$	sample mean	average / arithmetic mean	$\bar{x} = (2+5+9) / 3 = 5.333$
$s^2$	sample variance	population samples variance estimator	$s^2 = 4$
$s$	sample standard deviation	population samples standard deviation estimator	$s = 2$
$z_x$	standard score	$z_x = (x - \bar{x}) / s_x$	
$X \sim$	distribution of $x$	distribution of random variable $X$	$X \sim N(0,3)$

Symbol	Symbol Name	Meaning / definition	Example
$N(\mu, \sigma^2)$	normal distribution	gaussian distribution	$X \sim N(0,3)$
$U(a,b)$	uniform distribution	equal probability in range a,b	$X \sim U(0,3)$
$exp(\lambda)$	exponential distribution	$f(x) = \lambda e^{-\lambda x}, x \geq 0$	
$gamma(c, \lambda)$	gamma distribution	$f(x) = \lambda^c x^{c-1} e^{-\lambda x} / \Gamma(c), x \geq 0$	
$\chi^2(k)$	chi-square distribution	$f(x) = x^{k/2-1} e^{-x/2} / (2^{k/2} \Gamma(k/2))$	
$F(k_1, k_2)$	F distribution		
$Bin(n,p)$	binomial distribution	$f(k) = {}_n C_k p^k (1-p)^{n-k}$	
$Poisson(\lambda)$	Poisson distribution	$f(k) = \lambda^k e^{-\lambda} / k!$	
$Geom(p)$	geometric distribution	$f(k) = p(1-p)^k$	
$HG(N,K,n)$	hyper-geometric distribution		
$Bern(p)$	Bernoulli distribution		

## Chapter 4 - Measures of Variability

### Intro:

- measures of variability are alternatives to measures of central tendency (mean, median, mode)
  - Using measures of variability allows us to see differences between respondents.
- Measures of variability describe diversity or variability in the distribution
  - Shows variation & diversity within the answers we received.
- Measures of variability:
  - 1- **Index of qualitative variation (IQV)** - Measure of variability for nominal variables. Index can range from 0.00 (if all the cases in the distribution are in the same category, meaning no variation) to 1.00 (if all the cases in the distribution are distributed evenly across the categories, meaning max variation).
  - 2- **Range** - Highest score - lowest score. Can be misleading indicator of variation as it's based on 2 values.
  - 3- **Interquartile range (IQR)** - Measure of variation for ordinal variables. IQR uses the middle scores of the distribution, those at the 50% mark. IQR is based on middle scores, so we don't face the problem of misrepresentation of the distribution as we do w/the range.
  - 4- **Standard Deviation** - Square root of the variance. Measures variability in interval-ratio variables.
  - 5- **Variance** - Average of the squared deviations from the mean of the distribution. Measures how spread out a distribution is. Measures variability in interval-ratio variables.
- For nominal vars:
  - **only** use Index of qualitative variation (IQV)
- For ordinal vars:
  - 3/5 can be used: Index of qualitative variation (IQV), Interquartile range (IQR), and range can be calculated but Interquartile range (IQR) provides more info about the var
- For interval-ratio vars:
  - You can use all 5 (Index of qualitative variation (IQV), Interquartile range (IQR), variance, standard deviation), and range Though, the standard deviation & variance provide the most info.

## Index of Qualitative Variation (IQV)

### **IQV Definition:**

- The index of qualitative variation is a measure of variability for nominal variables based on the ratio of the total actual number of differences in the distribution to the max number of possible differences within the same distribution.
- The index of IQV can vary from 0.00 to 1.00
  - If all the cases in the distribution are in 1 category, there is no variation (or diversity) & the IQV is 0.00.
  - If all the cases in the distribution are distributed evenly across the categories, there is a max variation (or diversity) & the IQV is 1.00.
  - Basically, the closer to zero the IQV is, the less diverse/variation there is.

### **Calculating IQV:**

$$IQV = \frac{K(100^2 - \sum Pct^2)}{100^2(K-1)}$$

$K$  = the number of categories

$\sum Pct^2$  = the sum of all squared percentages in the distribution

- Square each percent first, then add them all up.
- **NOTE:** If the scores are represented as frequencies, you use this equation:

$$■ IQV = \frac{K(n^2 - \sum f^2)}{n^2 (K-1)}$$

→

→  $N$  = total number of cases

→  $K$  = number of categories

- To express the final IQV score as a percentage rather than a proportion, multiply it by 100

## ***The Range***

### **Definition:**

- The range is a measure of variation for interval-ratio variables, it's the difference between the highest (max) and the lowest (min) scores in the distribution
- *Range = Highest Score - Lowest Score*

# ***The Interquartile Range (IQR)***

## **IQR Definition:**

- The IQR is the width of the middle 50% of the distribution. It's the difference between the lower and upper quartiles (Q1 & Q3). IQR can be calculated for interval-ratio & ordinal data.
- $IQR = Q3 - Q1$

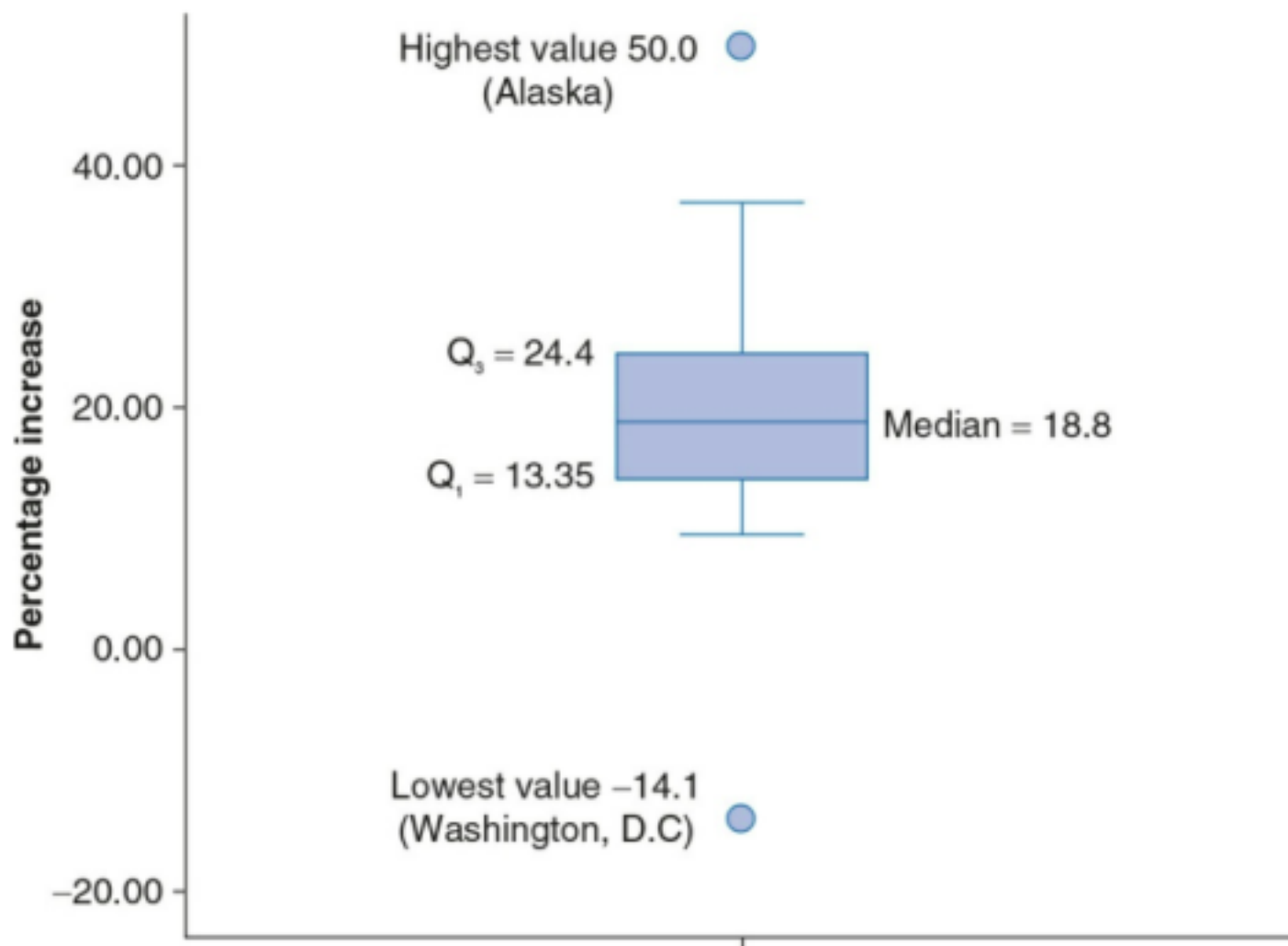
## **Calculating IQR:**

1. Order the scores in the distribution from highest to lowest or vice versa.
2. Identify the 1st quartile, Q1 or the 25th percentile by multiplying the total number of cases (N) by 0.25.
  - $(N)(0.25)$
  - If you get a decimal, you take the average of the 2 numbers the percentile falls into in terms of the category
3. Identify the 3rd quartile, Q3 or the 75th percentile by multiplying the total number of cases (N) by 0.75.
  - If you get a decimal, take the average of the 2 numbers the percentile falls into in terms of the category.
4. Once you find the quartile values, line the values up least to greatest and find the value that the quartile equals
  - **EX:** If the 25th quartile is at 25.5, find the values at 25 and take the average of them. That is your Q1 value you will subtract from Q3 later. Do the same for the 3rd quartile.
6.  $IQR = Q3 - Q1$  to finally find the IQR

## **The Box Plot:**

- A box plot can visually present the range, the IQR, the median, the lowest (min) score, and the highest (max) score.





## Variance & Standard Deviation

### Variance Definition:

- The variance is a measure of variation for interval-ratio variables. It's the average of the squared deviations from the mean

### Standard Deviation Definition:

- The standard deviation is a measure of variation for interval-ratio & ordinal variables. It's the square root of the variance.

### Calculating The Variance:

$$s_Y^2 = \frac{\sum (Y - \bar{Y})^2}{N - 1}$$

- 1) Calculate the mean.
- 2) Subtract the mean from each score to find the deviation. *Deviation = Score - Mean.*
- 3) Square each deviation.
- 4) Add all the squared deviations.
- 5) Divide the sum by N-1 (total cases - 1).
- 6) The answer is the variance.

### Calculating Standard Deviation:

- Simply **square** the variance
- Easier to interpret since the variance is expressed in squared percentages.
- Standard deviation tells you how much deviation from the mean to expect for a distribution.

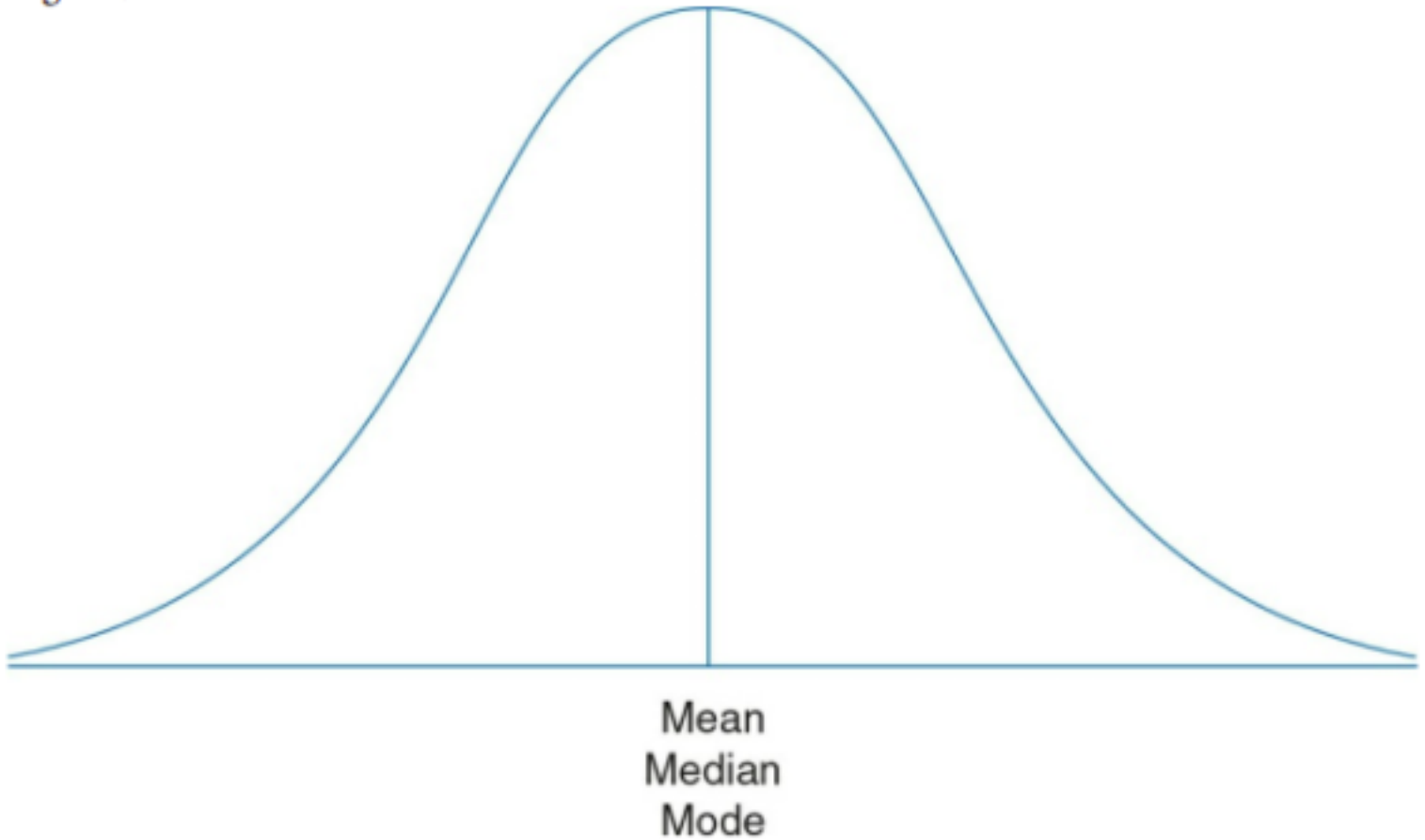
## ***Chapter 5 - The Normal Distribution***

## Normal Distribution Definition

### Definition of Normal Distribution:

- A bell-shaped & symmetrical theoretical distribution w/the mean, median & mode all at the peak & with the frequencies gradually decreasing at both ends of the curve.

Figure 5.1 The Normal Curve



- Notice that **most of the observations are clustered around the middle** with the **frequencies gradually decreasing at both ends of the distribution.**

# Properties

## Normal Distribution Properties:

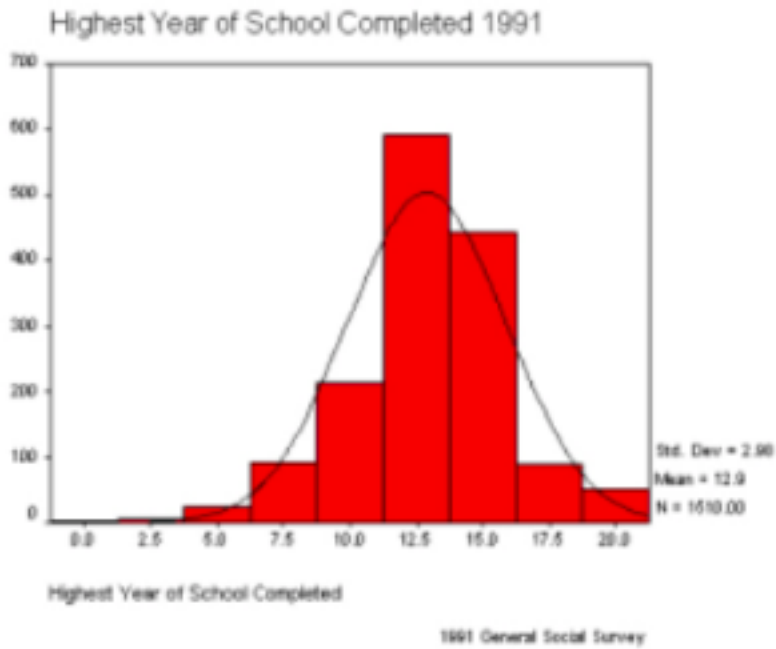
- Perfectly symmetric (see page before this for a normal curve)
  - Since it's perfectly symmetric, precisely half the observations fall on each side of the middle of the distribution.
- The mid point is the max frequency
  - The peak is also where the **three** measures are:
    - the **mode**
    - the **median**
    - the **mean**
  - The frequencies gradually decrease at both ends of the curve.
- The normal curve is pure theory, real-life distributions never match this model perfectly.
  - so if we say a distribution is normal, we mean the distribution closely resembles this theoretical curve.
- Like an empirical distribution (which is based on real data), a theoretical distribution can be organized into frequency distributions, displayed using graphs, and described by its central tendency and variation using measures such as the mean & standard deviation
  - But unlike an empirical distribution, a theoretical distribution is based on theory rather than real data.

## Practice

Final Grades in Social Statistics of 44 Students			
Midpoint Score	Frequency Bar Chart	Freq	%
50	*****	5	11
60	*****	10	22
70	*****	15	33
80	*****	10	22
90	*****	5	11

- The above is an example of a 'normal' distribution
  - Look at the bar chart & calculate the **mean, median, & mode**:
    - Calculate the **mean** by multiplying each score by the frequency & adding up all the values
    - Calculate the **median** by looking for the middle value in the frequency bar chart
    - Calculate the **mode** by looking for the score that has the most frequencies.
- Mean - 70
- Median - 70
- Mode - 70
  - Since all three measures of central tendency are equal & this is reflected in the bar graph, we can assume that this data is "**normally distributed**"
  - Also, since the median = mean, we know that the distribution is **symmetrical**

## Practice

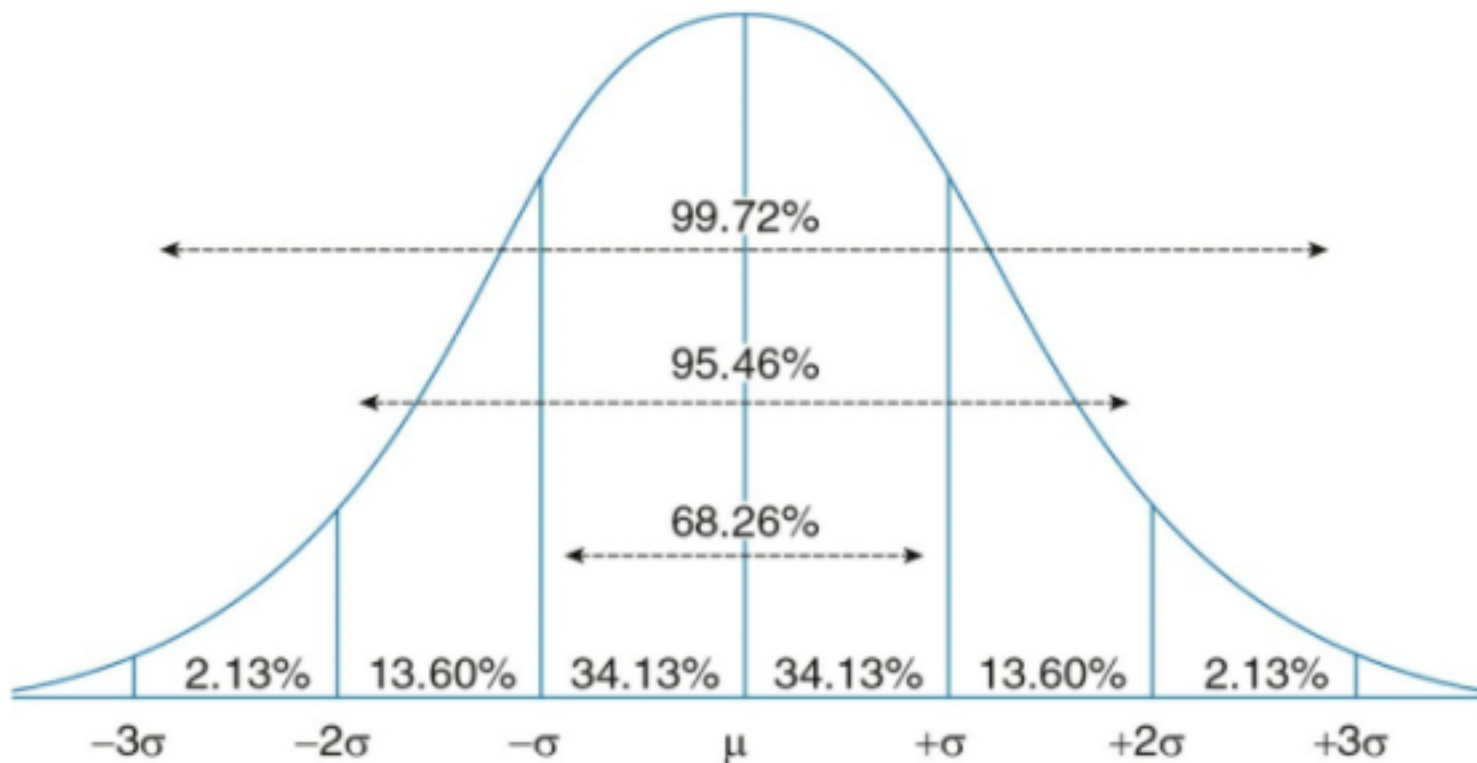


- Some shapes of normal curves in graphs are tall & thin or short & wide
- **All normal distributions are wider in the middle & symmetrical**

## Areas Under Normal Curve

### Areas Under Normal Curve:

- The area under the normal curve can be thought of as a proportion or percentage of the number of observations in the sample



- $\mu$  = Mean
- $-\sigma$  = 1 standard deviation below mean
- $+\sigma$  = 1 standard deviation above mean
- So between the mean and above or below 1 standard deviation, 68.26% of all observations in the distribution occur
  - Between the mean and above or below 2 standard deviations, 95.46% of all observations in the distribution occur.
  - Between the mean and above or below 3 standard deviations, 99.72% of all observations in the distribution occur.



## ***Interpreting Standard Deviation***

### **Interpreting Standard Deviation:**

- The relationship between the distance from the mean & the areas under the curve represents a property of the normal curve that has highly practical applications.
  - As long as the distribution is normal & we know the mean and the standard deviation, we can **determine the proportion or percentage of cases that fall between any score & the mean.**
  - For empirical distributions, when we know the mean & the standard deviation, we can determine the percentage or proportion of scores that are within any distance from that distribution's mean. (measured in standard deviation units)
  - Note that this relationship between the distance from the mean & the areas under the curve only applies to normal or approximately normal distributions.

## Transforming Raw Score into Z Score

### Transformation:

- A raw score can be turned into a *Z score* (aka a *standard score*) in order to see how many standard deviations it is above or below the mean
  - **Standard (Z) score** - The number of standard deviations that a given raw score is above or below the mean.
- Steps to transform a raw score into a Z score:
  1. Raw score - Mean
  2. Divide the difference by the standard deviation
  3. You have your Z score

$$Z = \frac{Y - \bar{Y}}{s}$$

- Formula:
- The Z score tells us how far a given raw score is from the mean in standard deviation units.
  - A positive Z score says the score > mean
  - A negative Z score says the score < mean
  - The larger the Z score, the larger the difference between the score & the mean.

## ***Transforming Z Score into Raw Score***

### **Transformation:**

- If you have a Z score but need to know the raw score, you can go backwards
- Formula:

$$Y = \bar{Y} + Z(S_y)$$

■

$$\bar{Y}$$

■ = mean

■ Z = Z score

$$(S_y)$$

■ = standard deviation

## ***Standard Normal Distribution***

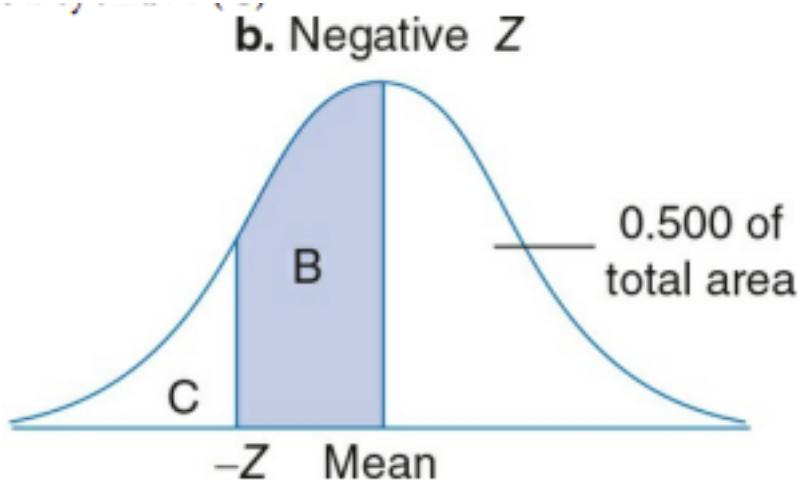
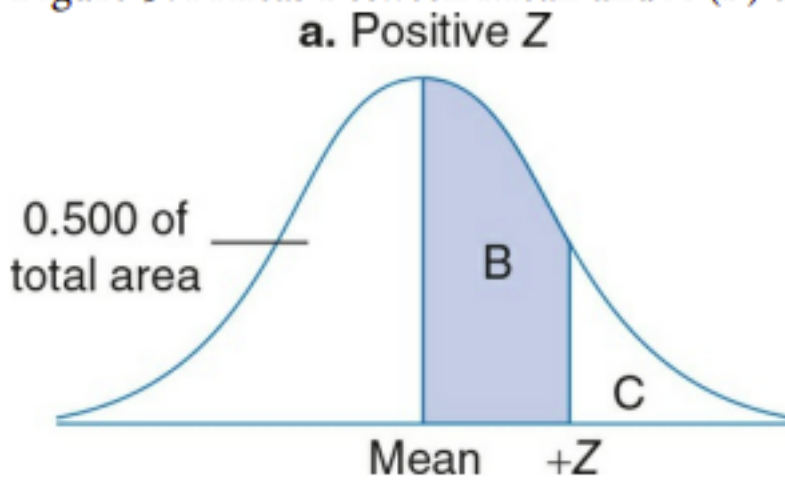
### **Standard Normal Distribution:**

- When a normal distribution is represented in standard scores (Z scores), it's called the standard normal distribution
  - Z scores are the #s that tell us the distance between an actual score & the mean in terms of standard deviation units.
- The standard normal distribution has a mean of 0.0 & a standard deviation of 1.0

## Standard Normal Table

### Standard Normal Table:

- This is a table showing the area (as a proportion, which can be translated into a percentage) under the standard normal curve corresponding to any Z score or its fraction.
- Z scores can be used to determine the proportion of cases that are included between the mean & any Z score in a normal distribution.



- Column B shows the area included between the mean and the Z score listed in Column A.
  - When Z is positive, the area is located on the right side of the mean whereas for a negative Z score, the same area is located on the left side of the mean.
- Column C shows the proportion of the area that is beyond the Z score listed in Column A.
  - Areas corresponding to positive Z scores are on the right side of the curve whereas areas corresponding to negative Z scores are identical except they're on the left side of the curve.

## ***Finding Area Between Mean & Positive or Negative Z Score***

- The standard normal table can be used to find the area between the mean & specific Z scores
- The standard normal table is located in Appendix B of the book for reference
- To find the area between a mean of 475 and a raw score of 675, follow these steps:
  - 1> Convert 675 to a Z score
  - 2> You get 1.83
  - 3> Search for 1.83 in Appendix B
  - 4> Multiply 0.4664 by 100 to get 46.64%
  - 5> 46.64% of the total area lies between 475 and 675
  - 6> Say you wanted to find the actual number of students who scored between 475 & 675, multiply the proportion 0.4664 by the total number of students. So say 1,108,165 students, you'd multiply 0.4664 & 1,108,165 which would equal 516,848 students scoring between 475 & 675.
- Now say you wanted to find the area for a score lower than the mean, indicating the z score would be negative:
  - The score being 305, the mean being 475
  - You first find the Z score which is **-1.56**
  - Since the proportions that correspond to positive Z scores are identical to those corresponding to negatives, ignore the negative sign & look it up like normal & proceed as if it wasn't a negative Z score.

## ***Finding Area Above a Positive Z Score or Below a Negative Z Score***

- The normal distribution table in appendix B could be used to find the area beyond a Z score too

## ***Transforming Proportions & Percents into Z Scores***



## Turning percentile into raw score

### Finding Z Score Which Bounds an Area Above It:

- Find the Z score for the percentile you're searching for (if it's the 20th percentile, you find the Z score corresponding to .20 in Appendix B column C or whatever is closest to .20)
- After you have the Z score, use this formula to find the raw score:

$$Y = \bar{Y} + Z(s)$$

- Where Y is the raw score you're searching for

$\bar{Y}$

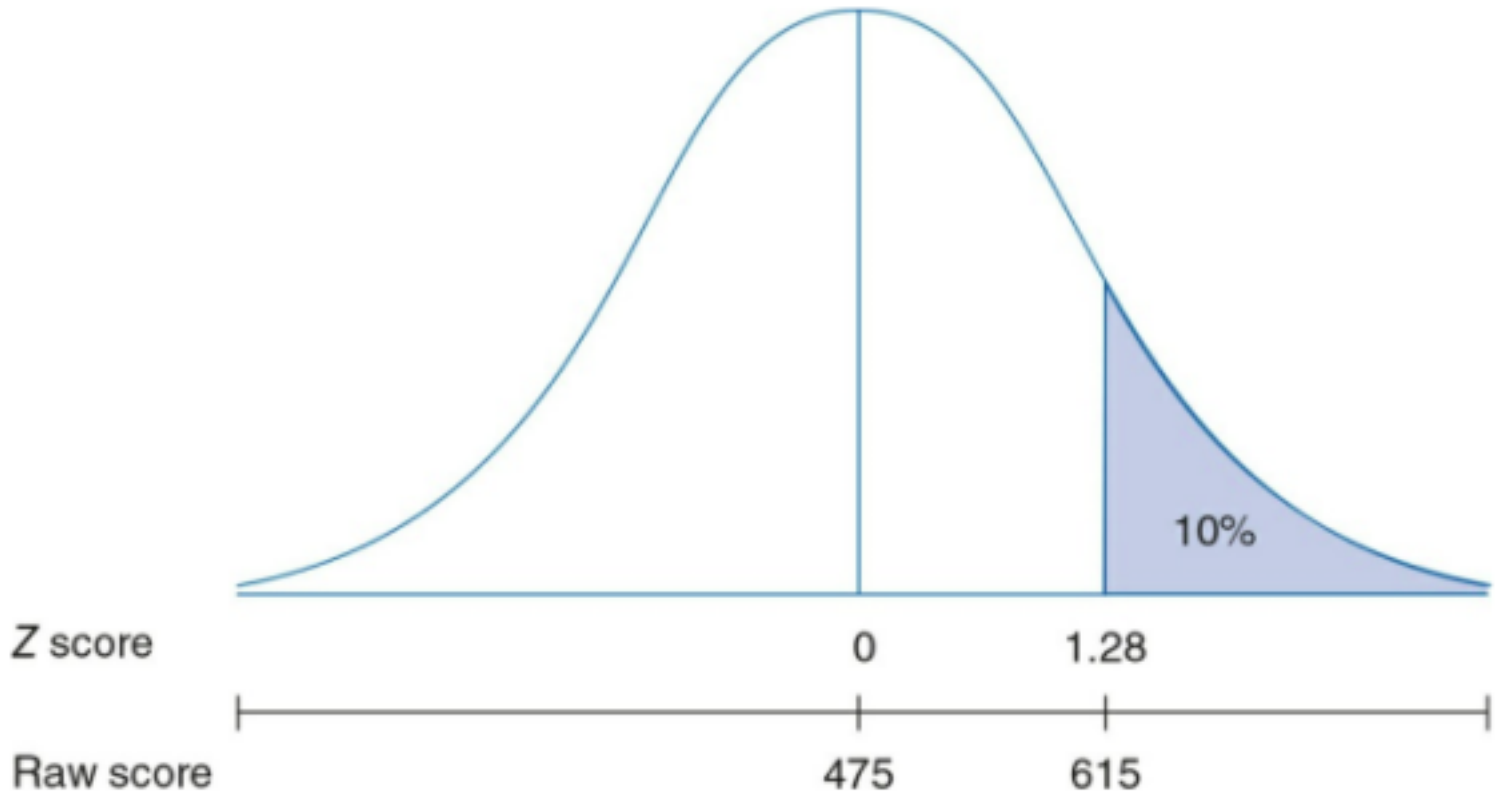
- = mean

- Z = Z score

- S = standard deviation

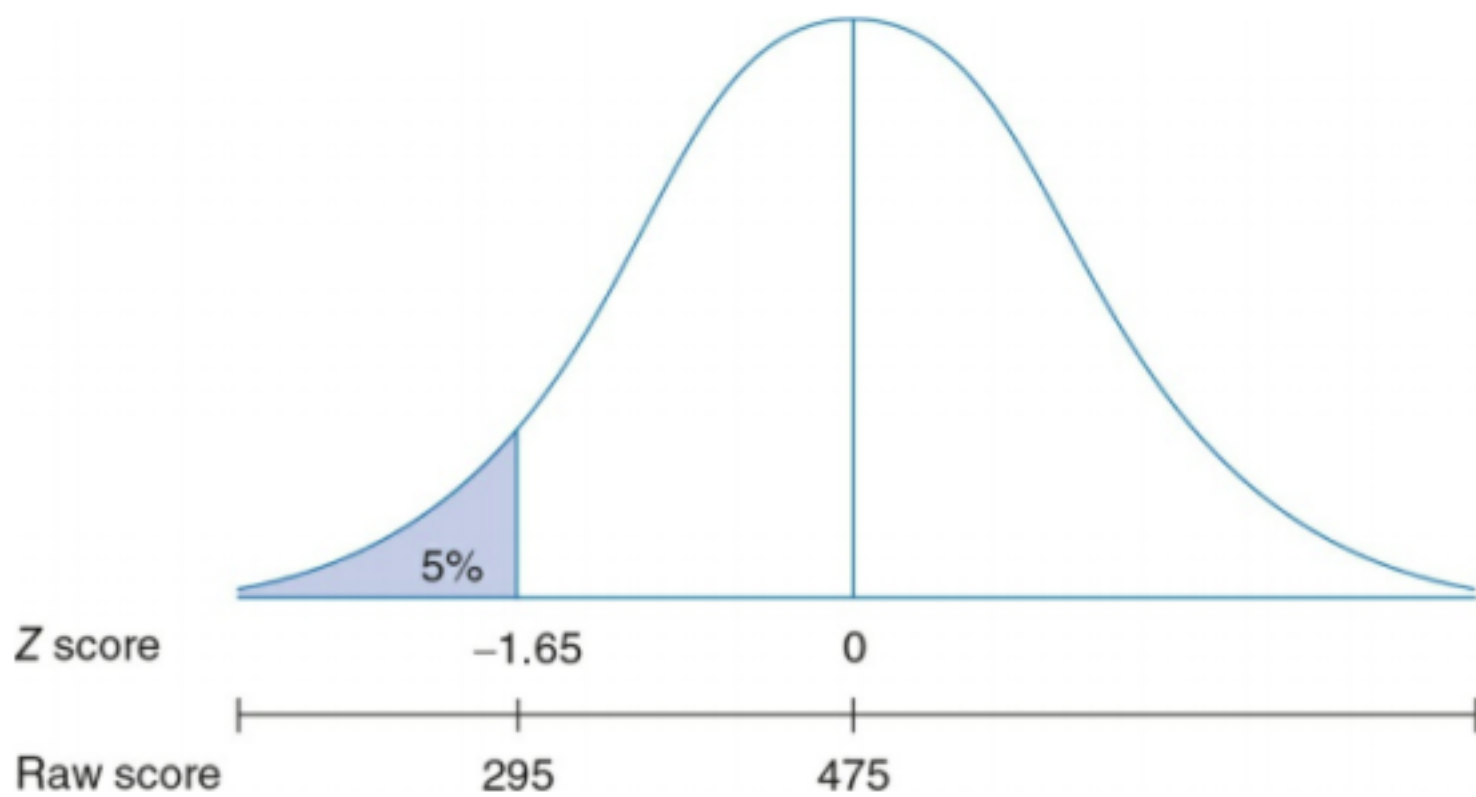
- By finding the percentile, you're able to identify the cut off point

■ **EX:** Say you were attempting to find the cut off point for the top 10% of an SAT exam. You'd find that the Z score is 1.28 & multiply it by a mean of let's say 475 and a standard deviation of 109 & plug these values into the equation to find that the cut off point for the top 10% of the SAT exam takers is 615.



### Finding a Z Score Which Bounds an Area Below It:

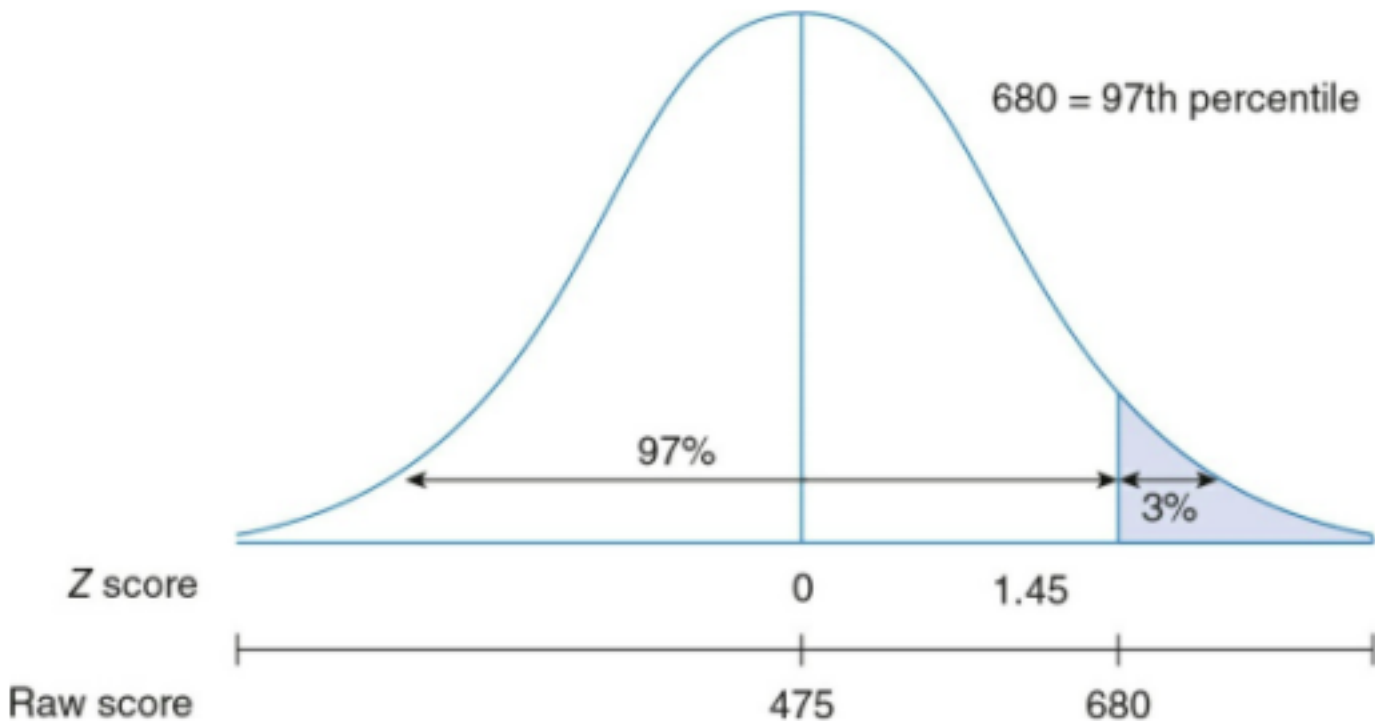
- Let's say you wanted to find the score that corresponds to the bottom 5% of test takers:
  - 1- Find the Z score that corresponds to 0.05 or the closest to it in Column C. Then locate the Z in Column A that corresponds to this proportion. In this case, it'd be 1.65 & **since the area we're looking for is on the left side of the curve (below the mean), the Z score is negative.** So the Z score associated with 0.05 is -1.65.
  - 2- Use the same equation as above to transform a Z score into a raw score.
  - 3- If you use the values 475 for the mean, 109 for the standard deviation, and -1.65 then you'd find the raw score to be 295.
- The cutoff for the lowest 5% of SAT scores is 295.



## Working w/Percentiles in Normal Distribution

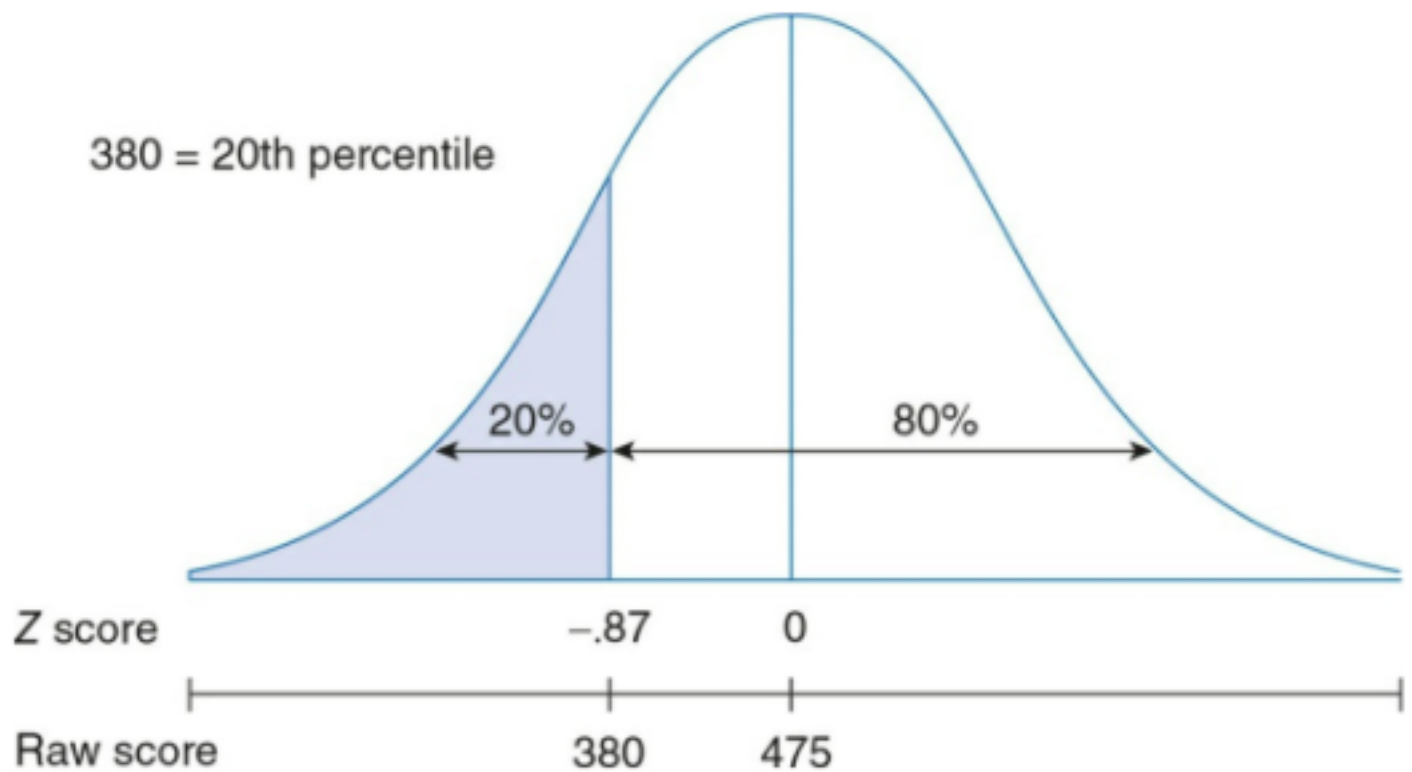
### Finding Percentile Rank of a Score Higher Than the Mean:

- Say you took the SAT & scored a 680 but how did you do relative to other students who took the exam?
- To find the percentile rank of a score higher than the mean, follow:
  - 1- Convert the raw score to a Z score using the formula from earlier (or the script)
  - 2- Find the area beyond Z in Appendix B Column C
  - 3- Subtract the area from 1.00 & multiply by 100 to get the percentile rank
- In this case, you'd find that the Z score is 1.88 and the area is 0.0301 and by using the formula to find the percentile rank, you get 97% which tells you that 97% of all test takers scored lower than 680 & 3% scored higher than 680.
- Note, the mean & standard deviation used in this example are 475 & 109 respectively.



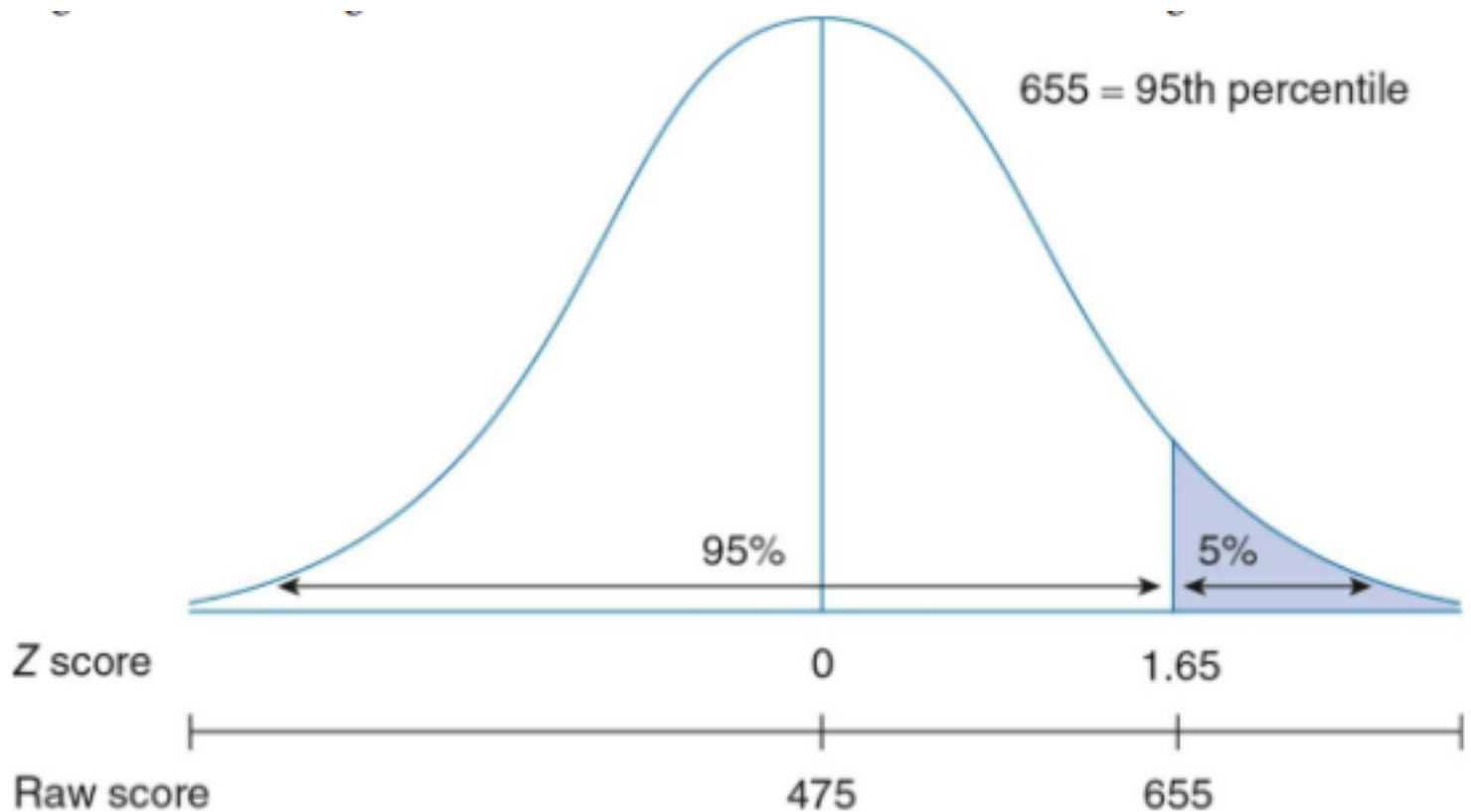
### Finding Percentile Rank of a Score Lower Than the Mean:

- If your SAT score is 380 (which is below the mean of 475), what's your percentile rank?
- 1. Convert the raw score of 380 to a Z score
  - The Z score would be -0.87
- 2. Find the area beyond Z in Appendix B Column C, the area beyond a Z score of -0.87 is 0.1992.
- 3. Multiply the area by 100 to obtain the percentile rank
- In this case, it's the 20th percentile rank which means that 20% of all test takers scored lower than you but 80% scored the same or higher.



## Finding the Raw Score Associated w/a Percentile Higher Than 50:

- Let's assume a uni only admits students who score at or above the 95th percentile in an SAT exam. What is the cut off point required for acceptance?
- To find the score associated w/a percentile higher than 50, follow:
  - 1- Divide the percentile by 100 to find the area below the percentile rank (0.95 in this case)
  - 2- Subtract the area below the percentile rank from 1.00 to find the area above the percentile rate (0.05 in this case, meaning 0.05 get accepted)
  - 3- Find the Z score associated w/the area above the percentile rank. Refer to Appendix B in Column C. Then locate Column A which corresponds with the proportion (in this case, 1.65).
  - 4- Convert the Z score to a raw score:
    - $Y = \text{Mean} + Z \text{ score}(\text{standard deviation})$
  5. In this case, it'd be 654.85
    - The final SAT associated with the 95th percentile is 654.85 which means the lowest score you can get to be admitted is 654.85



### **Finding Raw Score Associated w/a Percentile Lower Than 50:**

1. Divide the percentile by 100 to find the area below the percentile rank
2. Find the Z score associated w/this area. Refer to Appendix B in Column C then locate the Z in column A that corresponds to this proportion.
3. Convert the Z score to a raw score using the same formula as in the previous sect.
  - Recall that since it'd be lower than the mean, the Z score would be negative

## ***Chapter 6 - Sampling & Sampling Distributions***

# Aims of Sampling

## Aims of Sampling:

- Researchers in the social sciences rarely have enough time or money to collect info about the entire group that interests them
  - **Population** - A group that includes all the cases (individuals, objects, or groups) in which the researcher is interested.
  - Despite this limitation, we can learn a lot about a population if we carefully select a subset of it, this subset is called a sample.
  - **Sampling** - Process of selecting a subset of observations from the population.
  - Through sampling, we attempt to generalize the characteristics of the larger group (population) based on what we learn from the smaller group (the sample).
    - This is the basis for *inferential statistics*, making predictions or inferences about a population from observations based on a sample.
- Characteristics used to describe the population we're interested in are called *parameters*
  - **Parameters** - A measure (**EX**: mean, standard deviation) used to describe the **population** distribution.
  - We use the term **statistic** when referring to a corresponding characteristic calculated for the **sample**. **EX**: A sample mean, a sample proportion, & a sample standard deviation are all statistics.
- When referring to sample statistics & population parameters, we will use different symbols:

Measure Notation	Sample Notation	Population
Mean	$\bar{Y}$	$\mu$
Proportion	$p$	$\pi$
Standard deviation	$s$	$\sigma$
Variance	$s^2$	$\sigma^2$

## ***Basic Probability Principles***

### **Basic Probability Principles:**

- Techniques used by social scientists to select samples from populations follow a general approach called probability sampling
- **Probability** - A quantitative measure that a particular event will occur. Represented by a lower case  $p$  & expressed as a ratio of the number of times an event will occur relative to the set of all possible & equally likely outcomes
  - $p = \text{Number of times an event will occur} / \text{Total number of events}$
  - Probabilities range in value from 0 (the event will not occur) to 1 (the event will certainly occur).
  - Probabilities can be expressed as percents or proportions.
- Sometimes we use info from past events to help us predict the likelihood of future events, this method is called the **relative frequency method**.



# ***Probability Sampling***

## **Probability Sampling:**

- **Probability sampling** - A method of sampling that enables the researchers to specify for each case in the population the probability of it being included in the sample.

- The purpose of probability sampling is to select a sample that is as representative as possible of the population, the sample is selected in a way that allows the use of the principles of probability to evaluate the generalizations made from the sample to the population.

- Only the general approach of probability sampling allows researchers to use the principles of statistical inferencing to generalize from the sample to the population.

- **Nonprobability samples** are often used since they're more convenient & cheaper to collect

- But their limitation is that they don't allow the use of the method of inferential statistics to make generalizations about the population from the sample.

## ***The Simple Random Sample***

### **The Simple Random Sample:**

- Most basic probability sampling design
- Designed in such a way to ensure that:
  - Every member of the population has an equal chance of being chosen
  - Every combination of members has an equal chance of being chosen
- **EX:** Let's say I wanted to study this class w/a sample of 10. I would put everyone's name on a piece of paper & draw 10 names out of a hat.
  - This is a simple random sample because it was pure chance that determined who was chosen. Every student had the same chance & every combination of students was possible
  - Every combination meaning that student A & B had the same chance of being chosen as student B & C since there'd be a 1/10 chance for each student.

## ***Systematic Random Sample***

### **Systematic Random Sample:**

- Easier to implement than a simple random sample.
- The systematic random sample, although not a true probability sample, provides results very similar to those obtained with a simple random sample.
  - It uses a ratio,  $K$ , obtained by dividing the population size by the desired population size.

$$K = \frac{\text{Population size}}{\text{Sample size}}$$

- In systematic random sampling, every  $K$ th member in the total population is chosen to be included in the sample after the 1st member of the sample is selected at random from among the first  $K$  members in the population.

■ **EX:** Say we had a population of 15,000 students and our sample was limited to 500. Our  $K$  would be 30. Using the systematic random sampling method, follow these steps:

1. First choose any one student at random from among the first  $K$ th students on the list of students (in this case, it'd be from among the first 30 students)
2. Then we'd select every 30th student until we reached our sample size (which is 500). So suppose our 1st selection at random would be the 8th student on the list. The 2nd student in our sample would be the 38th on the list ( $8 + 30 = 38$ ) and this continues on.
3. The 3rd student in our sample would be the 68th on the list ( $38 + 30 = 68$ ).

## ***Stratified Random Sample***

### **Stratified Random Sample:**

- We obtain a stratified random sample by doing the following:
  - 1- Dividing the population into subgroups based on 1 or more variables central to our analysis.
    - The choice of subgroups is based on what variables are known & what variables are of interest to us.
  - 2- Drawing a simple random sample from each of the subgroups.
- When you want to compare subgroups w/each other & when the size of the subgroups in the population is relatively small, you'd use **disproportionate stratified sampling**.
  - **Proportionate stratified sampling** on the other hand can result in the sample having too few from a small subgroup to yield reliable info about them.
- **Proportionate Stratified Sample** - The size of the sample selected from each subgroup is proportional to the size of that subgroup in the entire population.
- **Disproportionate Stratified Sample** - The size of the sample selected from each subgroup is disproportional to the size of the subgroup in the population.

## ***Sampling Distribution***

### **Sampling Distribution:**

- Helps estimate the likelihood of our sample statistics & therefore enables us to make generalizations about the population based on the sample.

## ***Example***

### **The Population:**

- Let's consider our population the 20 individuals in the table below.
- Our variable,  $Y$ , is the income of these 20 individuals & the parameter we're trying to estimate is the mean income.

**Table 6.3 The Population: Personal Income (in Dollars) for 20 Individuals (Hypothetical Data)**

Individual	Income (Y)
Case 1	11,350 ( $Y_1$ )
Case 2	7,859 ( $Y_2$ )
Case 3	41,654 ( $Y_3$ )
Case 4	13,445 ( $Y_4$ )
Case 5	17,458 ( $Y_5$ )
Case 6	8,451 ( $Y_6$ )
Case 7	15,436 ( $Y_7$ )
Case 8	18,342 ( $Y_8$ )
Case 9	19,354 ( $Y_9$ )
Case 10	22,545 ( $Y_{10}$ )
Case 11	25,345 ( $Y_{11}$ )
Case 12	68,100 ( $Y_{12}$ )
Case 13	9,368 ( $Y_{13}$ )
Case 14	47,567 ( $Y_{14}$ )
Case 15	18,923 ( $Y_{15}$ )
Case 16	16,456 ( $Y_{16}$ )
Case 17	27,654 ( $Y_{17}$ )
Case 18	16,452 ( $Y_{18}$ )
Case 19	23,890 ( $Y_{19}$ )
Case 20	25,671 ( $Y_{20}$ )
Mean ( $\mu$ ) = 22,766	Standard deviation ( $\sigma$ ) = 14,687

- The  $\mu$  symbol is used to represent the population mean, using the formula below we can calculate the population mean:

$$\mu = \frac{\sum Y}{Y} = \frac{Y_1 + Y_2 + Y_3 + Y_4 + Y_5 + \cdots + Y_{20}}{20}$$

- Add up all the Ys and divide by the total number of Ys



## Calculating Z score for Sampling Distribution

### Formula:

#### Learning Check 6.6



Suppose a population distribution has a mean  $\mu = 150$  and a standard deviation  $s = 30$ , and you draw a simple random sample of  $N = 100$  cases. What is the probability that the mean is between 147 and 153? What is the probability that the sample mean exceeds 153? Would you be surprised to find a mean score of 159? Why? (Hint: To answer these questions, you need to apply what you learned in [Chapter 5](#) about Z scores and areas under the normal curve [Appendix B].) To translate a raw score into a Z score we used this formula:

$$Z = \frac{Y - \bar{Y}}{S}$$

However, because here we are dealing with a sampling distribution, replace  $Y$  with the sample mean  $\bar{Y}$ ,  $\bar{Y}$  with the sampling distribution's mean  $\mu_{\bar{Y}}$ , and  $\sigma$  with the standard error of the mean.

$$Z = \frac{\bar{Y} - \mu_{\bar{Y}}}{\sigma / \sqrt{N}}$$



# Estimation

## Estimation:

- inferential statistics
- **Estimation** is a process where we select a random sample from a population & use a sample statistic to estimate a population parameter.
  - Parameter - measure used to describe an entire population
  - Statistic - any measure used to describe the sample pulled from population
- 2 types of **estimation**:
  - **Point estimation** - attempting to estimate the exact value of a population parameter.
  - **Interval estimation** - Creating a range of values which we believe that the population parameter is going to fall into.
- **Confidence interval** - range of values defined by the confidence level within which the population parameter is estimated to fall
  - for 90% confidence level, z value is 1.65
  - for 95% confidence level, z value is 1.96
  - for 99% confidence level, z value is 2.58
  - These were the **top 3 confidence levels** used in social sciences
  - You need confidence level to construct confidence interval
- **Confidence level** - Likelihood, expressed as a percent or probability that a specified interval will contain the population parameter. Expressed as a percentage or a probability. Basically, evaluates accuracy of the estimate of population parameters falling into the confidence interval.
  - To have more confidence, expand confidence interval
  - To be more precise, decrease confident level which results in a less expanded confidence interval
- To construct a **confidence interval** is:

$$CI = \bar{Y} \pm Z(\sigma_{\bar{y}})$$

- Confidence Interval = Sample Mean & add & subtract the z value (z value is based on our confidence level) times the standard error.
  - Calculate the product of the z value & the standard error first.
  - Line over the y + the o indicates standard error. Without the line over the y, it indicates standard deviation.
  - **Standard error** is calculated by taking *population standard deviation* & divide it by the square root of the sample size, here is the formula:

$$\sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{N}}$$

- → You may have to estimate the standard error instead of finding the actual standard error because the standard error isn't generally known. To calculate the **estimate standard error**, you divide the *sample standard deviation* by the square root of the sample size. Here is the formula:

$$s_{\bar{Y}} = \frac{s_Y}{\sqrt{N}}$$

- When calculating the confidence interval, the reason you add & subtract the values is to find the **upperbound** & the **lowerbound**

## ***Reducing Risk***

### **Reducing Risk:**

- To reduce the risk of being incorrect in terms of our interval not containing the true population mean, we can increase the level of confidence being used
  - By increasing the level of confidence, this widens our confidence interval
- When we use the highest level of confidence (99% confident), there is only a 1% risk that we're wrong that the specified interval doesn't contain the true population mean

### **Sample Size & Confidence Intervals:**

- By increasing the sample size, researchers increase the precision of their estimate, and as a result decreases the confidence interval size

## Confidence Intervals for Proportions & Percentages

### For Proportions:

- For Proportions, the formula is:

$$CI = p \pm Z (S_p)$$

- 
- CI = Confidence interval
- P = observed sample proportion
- Z = Z corresponding to the confidence level
- Sp = Estimated standard error of proportions

$$S_p =$$

→ Formula used for calculating **estimated standard error of proportions**:

### Interpreting Results:

- We are x% confident that the true population proportion is somewhere between x and x
- Express the results in percentages as it's easier to understand percents

### For Percentages:

- For percents, convert it into a proportion then use the proportions method of calculating the confidence interval.

## ***Reference (Exam 2, Ch 4-7)***

### **REMEMBER:**

- export to pdf to mobile

## Ch 4

### CH 4 Terms:

- **Variability**

- **Index of qualitative variation (IQV)** - Measure of variability for nominal vars, based on the ratio of the total number of differences in the distribution to the max number of possible differences within the same distribution. Total number of difference : max possible number of differences.

- Index for IQV varies from 0.00 to 1.00, the less variation/diversity, the closer to 0. The more variation/diversity, the closer to 1.

- When all cases in a distribution are in 1 category, there's no variation & the IQV is 0. But if they are evenly distributed among different categories it's 1.0 as there is more diversity.

- **Range** - Measure of variation for interval-ratio vars, difference between max & minimum scores in the distribution.

- **Interquartile Range (IQR)** - Calculated for interval-ratio & ordinal data. Defined as the difference between the lower & upper quartiles (Q1 - Q3). 1st quartile (Q1) = 25th percentile, the point at which 25% of the cases fall below it & 75% above it. The 3rd quartile (Q3) is the 75th percentile, the point at which 75% of the cases fall below it & 25% above it. So the IQR defines variation for the middle 50% of cases.

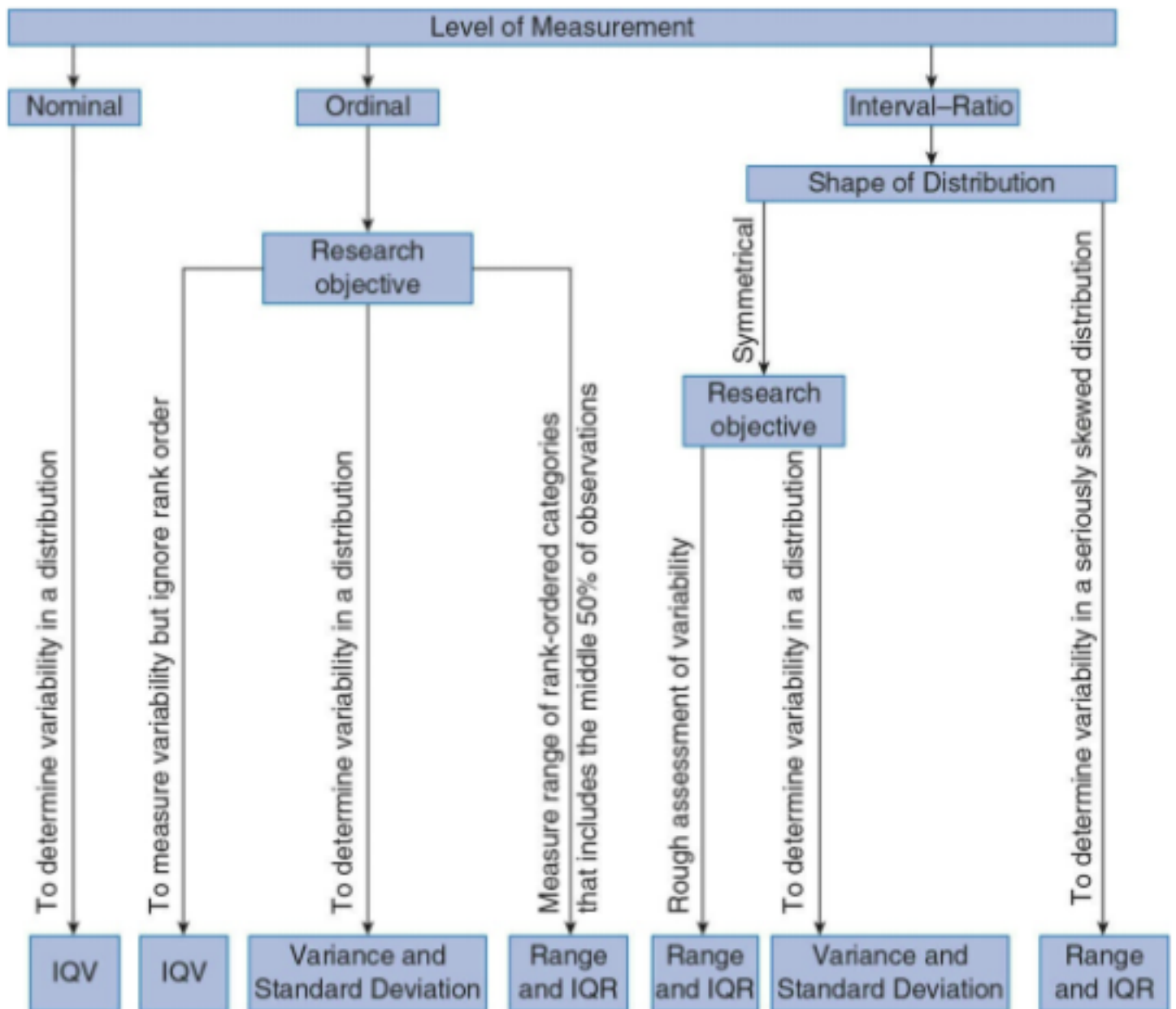
- IQR avoids instability associated w/range.

- **Box Plot** - Visually presents the range, the IQR, the median, the minimum score, & the max score (center, variation, & shape of distribution of interval-ratio vars)

- **Variance & Standard deviation** - Variance is the average of the squared deviations from the mean. Standard deviation is the square root of the variance. Both measure variability in interval-ratio & ordinal vars.

- Choosing which measure of variation out of the ones covered to use:





## Calculating IQV

### Formula:

$$IQV = \frac{K(100^2 - \sum Pct^2)}{100^2(K-1)}$$

- 

- K = Number of categories

- $\sum Pct^2$

= The sum of all squared percentages in the distribution

### Express as Percentage:

- Multiply IQV by 100, there's your percent
- As a percent, it reflects percentage of differences relative to the max possible differences in each distribution.

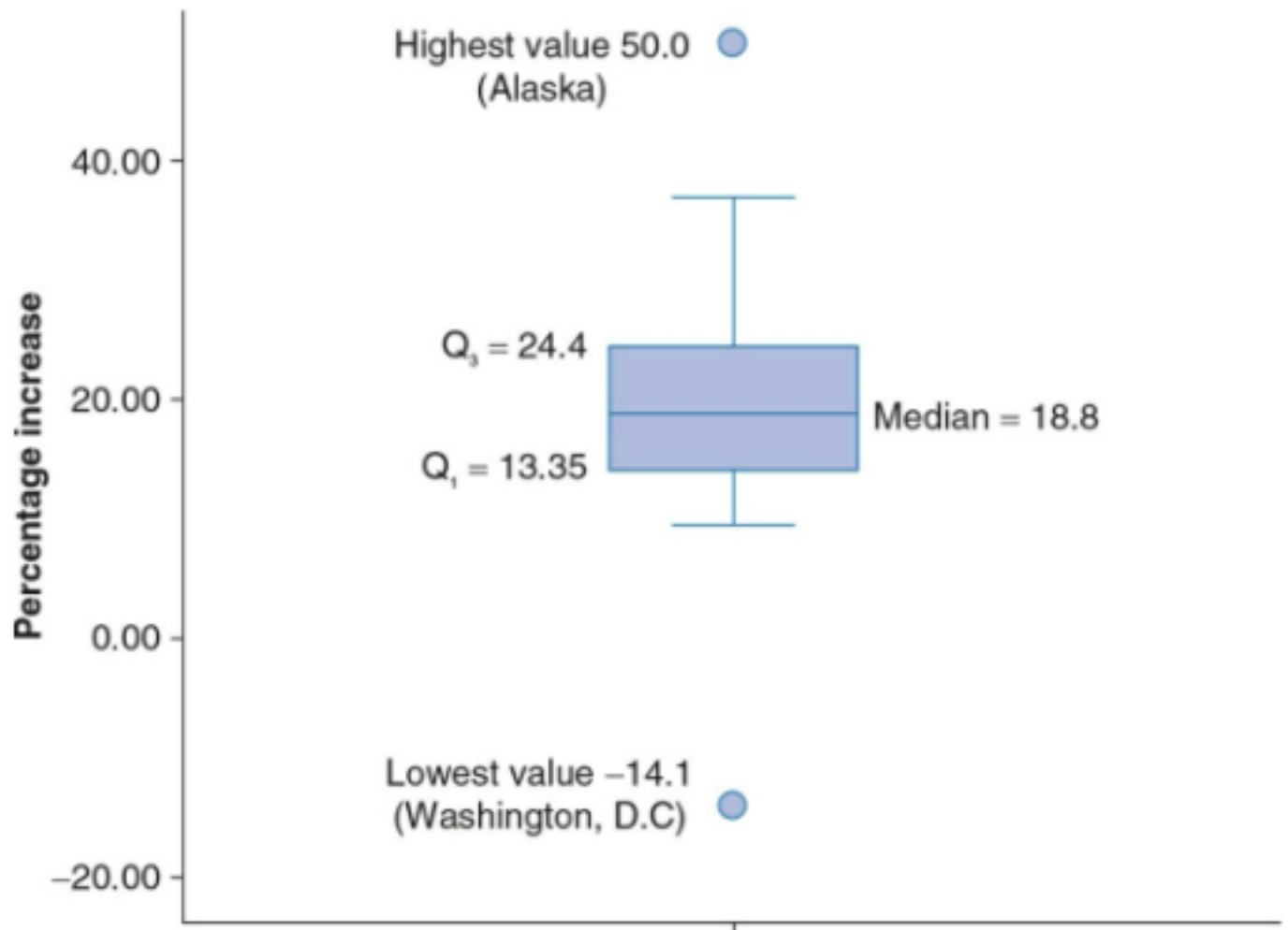
## ***Calculating IQR***

### **Steps:**

1. order scores in distribution from highest to lowest score or vice versa.
2. Identify 1st quartile (Q1)
  - $(N)(0.25)$  where  $N$  = total number of cases
  - If the result is a decimal, take the 2 numbers it falls between & take the average of their scores for interval-ratio data.
3. Identify 3rd quartile (Q3)
  - $(N)(0.75)$  where  $N$  = total number of cases
4.  $IQR = Q3 - Q1$ 
  - IQR tells us half the values in a distribution are clustered between the 3rd quartile & the 1st quartile

## Box Plot

### Example:



## Calculating Variance & Standard Deviation

### Calculating Deviation from Mean:

- Required for variance

$$\sum (Y - \bar{Y})$$

•

- (Score - Mean) for each score
- Then sum up all the results
- Since we may have negative values, we square the deviations from the mean then add them all together using this formula:

$$\rightarrow \sum (Y - \bar{Y})^2$$

### Calculating Variance:

- Variance is symbolized as  $S^2$
- Formula:

$$s^2 = \frac{\sum (Y - \bar{Y})^2}{N - 1}$$

■

$s^2$  = the variance

$(Y - \bar{Y})$  = the deviation from the mean

$\sum (Y - \bar{Y})^2$  = the sum of the squared deviations from the mean

$N$  = the number of scores

→

- 1- Calculate the mean (done above already)
  - 2- Subtract mean from each score to find the deviations (done above already)
  - 3- Square each deviation
  - 4- Add up all squared deviations
  - 5- Divide the sum by  $N - 1$
  - 6- There's your variance
- Remember, we often interpret the square root of the variance which gives us the **standard deviation**
    - So in order to calculate the standard deviation, use this formula;

$$s = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N - 1}}$$

→

→ Follow the same steps for variance, except as a last step you square it.

### Considerations for Choosing a Measure of Variability

---

- For **nominal variables**, you can only use IQV (Index of qualitative variation).
- For **ordinal variables**, you can calculate the IQV or the IQR (inter-quartile range.) Though, the IQR provides more information about the variable.
- For **interval-ratio variables**, you can use IQV, IQR, or variance/standard deviation. The standard deviation (also variance) provides the most information, since it uses all of the values in the distribution in its calculation.

## Ch 5

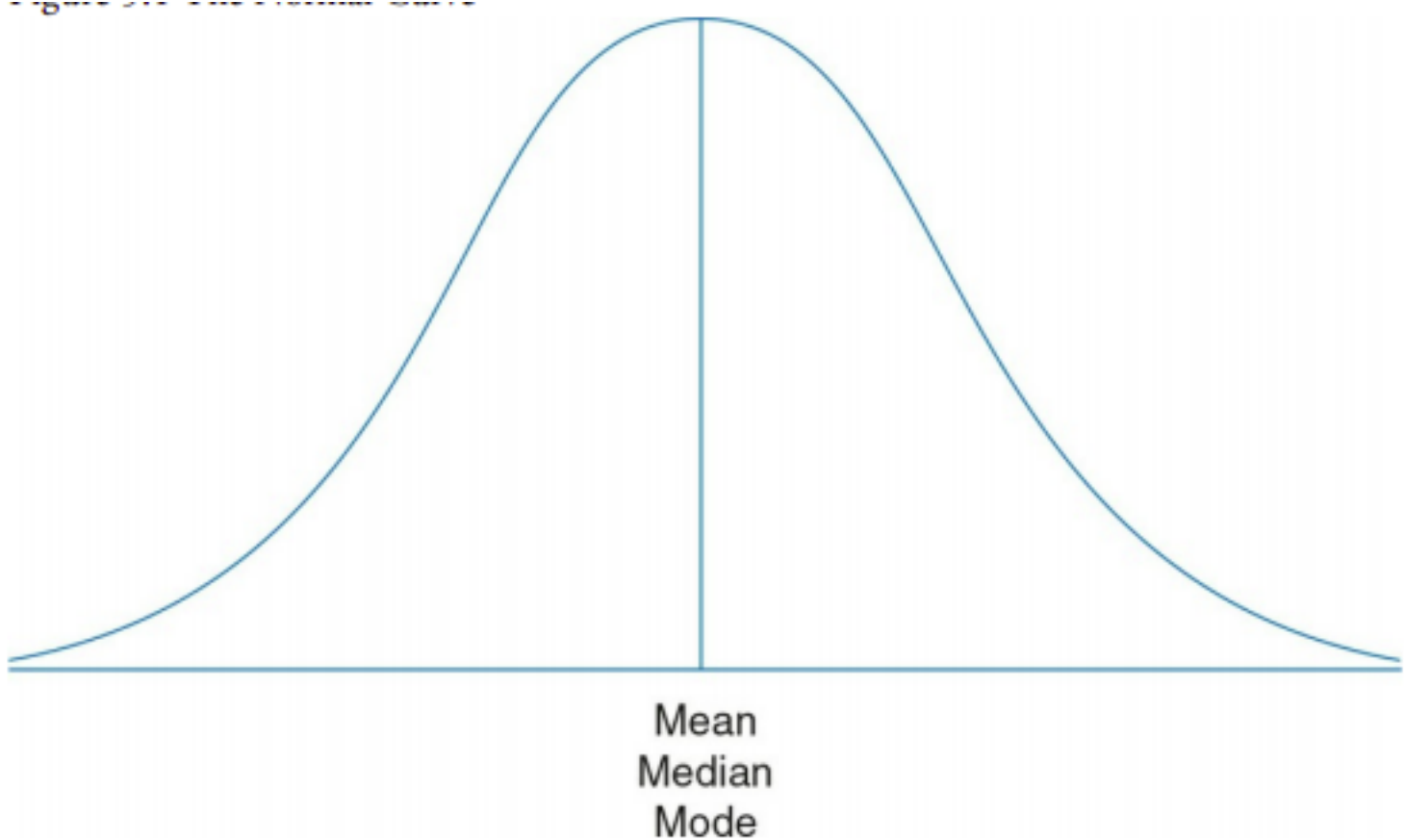
### CH 5 Terms:

- **Normal distribution** - A bell-shaped & symmetrical theoretical distribution w/the mean, median, & mode all at the peak & the frequencies gradually decreasing at both ends of the curve.
- **Empirical Distribution** - Used to describe distributions that are approximately bell-shaped & have symmetrical distributions closely resembling the idealized one pictured here.
- **Normal curve** -
- **Standard deviation**
- **Z Score**
- **Standard Normal Table** - A table showing the area (as a proportion that can be translated into a percentage) under the standard normal curve corresponding to any Z score or its fraction.

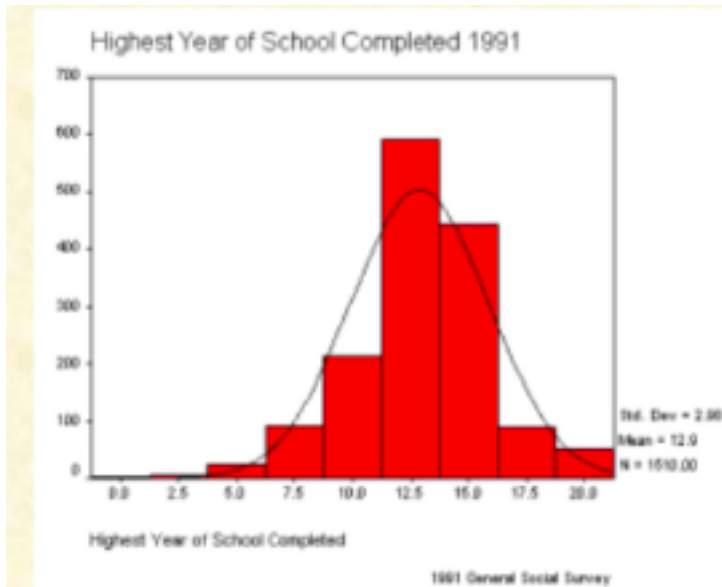
# Normal Distribution

## Properties:

- Perfectly symmetric
- Not extremely skewed
- Theoretical idea, real-life distributions never match this model perfectly.
- Midpoint of the normal curve is the point having the maximum frequency
  - Also where mode, median, & mean are located
- Most of the observations are clustered around the middle w/the frequencies gradually decreasing at both ends:



- 
- Mean, median, mode all equal to each other.
- Also a normal distribution:

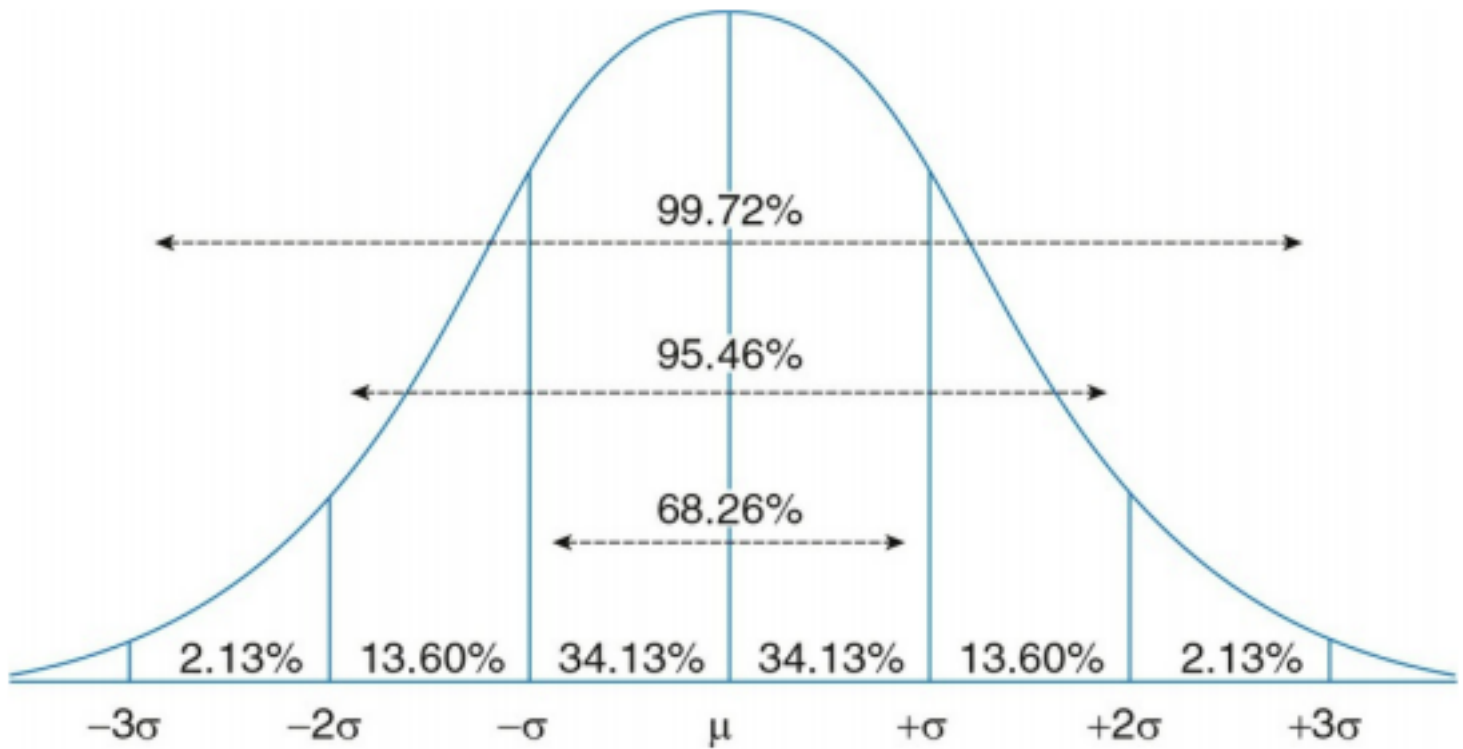




## Empirical Distribution

### Empirical Distribution:

- In all normal or nearly normal curves, we find a constant proportion of the area under the curve lying between the mean & any given distance from the mean when measured in standard deviation units:
  - The areas under the normal curve may be conceptualized as a proportion or percentage of the number of observations in the sample. So the entire area under the curve is equal to 1.00 or 100% of the observations
  - Since the curve is perfectly symmetrical, exactly 0.5 or 50% of the observations lie above or to the right of the center which is the mean of the distribution & the remaining 50% lie below or to the left of the mean:



## Interpreting Standard Deviation

### Interpreting Standard Deviation:

- As long as a distribution is normal & we know the mean & standard deviation, we can determine the proportion or percentage of cases that fall between any score & the mean
  - This provides an important interpretation for the standard deviation of empirical distributions that are approximately normal, **when we know the mean & standard deviation, we can determine the percentage/proportion of scores that are within any distance (in standard deviation units) from that distribution's mean.**
  - But remember, not every empirical distribution is normal and the fixed relationship between the distance from the mean & the areas under the curve apply only to normal or approximately normal distributions.

### Standard (Z) Score:

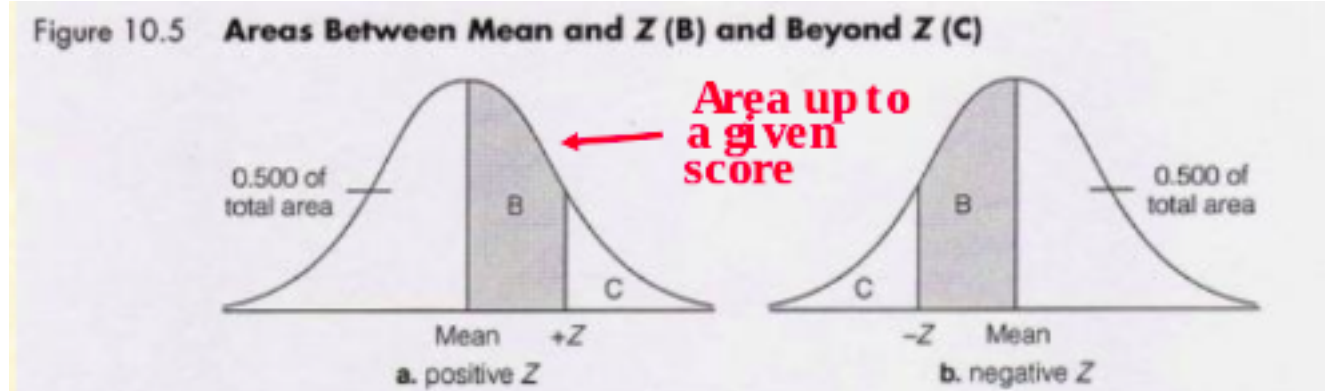
- A Z score is the number of standard deviations that a raw score is above or below the mean
- Positive Z Score - Above the mean
- Negative Z score - below the mean
- The larger the Z score, the further away from the mean you are
- **Formula:**

$$Z = \frac{Y - \bar{Y}}{S_y}$$

- After completing equation, look up the Z score in Appendix B, Column A, and the category you're looking in depends on what you're looking for (whether you're looking in area beyond Z or between mean & z)
- As for negative, you treat it as if it were positive when looking up its corresponding Z score

### Standard Normal Table:

- Table showing the area (as a proportion) under the standard normal curve corresponding to any Z score:



- C is the area beyond a given score

### Transforming Z Score Into Raw Score:

- Formula:

$$Y = \bar{Y} + Z(S_y)$$

## Finding Area Between 2 Z-Scores on the Same Side of the Mean:

- Find the percentage of students scoring between 74 & 84, when the mean is 70 & the standard deviation is 11.68
  - $Z = 74 - 70 / 11.68$  which is 0.34
  - $Z = 84 - 70 / 11.68$  which is 1.19
  - Then look up these values (0.34 & 1.19) in appendix B
  - 0.34 -> **0.1331** & 1.19 -> **0.3830**
  - Then, **subtract** the smaller area from the larger area ( $0.3830 - 0.1331 = .2499$ ) which is the proportion of students scoring **between** 74 & 84. The proportion can be converted to a percentage by multiplying by 100 as well.

## Finding Area Between 2 Z-Scores on Opposite Sides of the Mean:

- Find the percentage of students scoring between 62 & 72 when the mean is 70 & the standard deviation is 11.68
  - $62 - 70 / 11.68$  which is -0.68
  - $72 - 70 / 11.68$  which is 0.17
  - 0.17 -> **0.0675** & -0.68 -> **.2517**
  - Add** the two areas together ( $0.0675 + .2517 = .3192$ )
  - Now we know that .3192 students scored **between** 62 & 72, or 31.92%.

## Finding Area Above a Positive Z-Score or Below a Negative Z-Score:

- Find the percentage of students who did (a) very well, scoring above 85, & (b) those students who did poorly & scored below 50.
- With the mean being 70 & the standard deviation being 11.68
  - Convert 85 & 50 to a z score, then look up the value in **Column C** of Appendix B.
  - $85 - 70 / 11.68$  which is 1.28
  - $50 - 70 / 11.68$  which is -1.71
  - Look up these values (1.28 & -1.71) in Column C of appendix B
  - 1.28 -> **.1003** & -1.71 -> **0.0436**
  - Convert these Z-Scores to percents & you have your answer! (10.03% & 4.36%)

## Finding a Z-Score Bounding an Area Above it:

- Find the raw score that bounds the top 10% of the distribution
  - Reference the formula for transforming a Z-score into a raw score
  - Convert 10% to a decimal & search for that value in **Column C**, if you cannot find the exact value, choose the one that comes closest to it. Then take the Z-value in **Column A** that corresponds to the value in Column C.
  - The value coming closest to .1000 in Column C is .1003 & the corresponding Column A value is 1.28
  - Set up the equation which would be:  $Y = 70 + 1.28 (11.68)$  which equals **84.95**
    - $70 = \text{mean}$
    - $1.28 = \text{Z-score}$
    - $11.68 = \text{standard deviation}$
  - 84.95 is the raw score that bounds the upper 10% of the distribution. The Z-score associated w/ 84.95 in this distribution is 1.28.
- In this case, this area above it would belong in the C column of the Standard Normal Table

## Finding the Percentile Rank of a Score Higher than the Mean:

- Suppose your raw score was 85, you want to calculate the percentile (to see where in the class you rank)
  - 1- Convert the raw score (85) to a Z-Score using 70 as a mean & 11.68 as a standard deviation.
  - 2-  $Z = 85 - 70 / 11.68$  which is 1.28
  - 3- Find the area beyond Z in appendix B in Column C -> .1003
  - 4- Subtract the area from 1.00 for the percentile, since .1003 is **only** the area **not** below the score:  
 $1.00 - .1003 = .8997$  (this is the proportion of scores below 85)
  - 5- .8997 represents the proportion of scores less than 85 corresponding to a percentile rank of 90% (round .8997 up)

## Finding the Percentile Rank of a Score Lower than the Mean:

- Now suppose your raw score was 65, with a mean of 70 & a standard deviation of 11.68.
  - 1- Convert the raw score to a Z-Score:  $Z = (65 - 70) / 11.68 = -.42$
  - 2- Find the area beyond Z in appendix B, Column C which is .3372
  - 3- Multiply the value by 100 (.3372 x 100) to obtain the percentile rank which in this case is 33.72%.  
A score of 65 would be in the 34th percentile.

## Finding Raw Score of a Percentile Higher than 50:

- Say you need to score in the 95th percentile to be accepted in a school, what's the cut off for the 95th percentile?
  - 1- Turn the percentile into a decimal ( $95 / 100 = 0.9500$ )
  - 2- Subtract the area from 1.00 to find the area above & beyond the percentile rank ( $1.00 - 0.9500 = .0500$ )
  - 3- Find the Z-Score corresponding to .0500 in **Column C** (Area beyond Z) in appendix B which would be 1.65 (**NOTE:** find the value closest to .0500, there are 2 but I assume we may get the same answer with either we choose)
  - 4- Convert the Z score to a raw score:  $Y = 70 + 1.65(11.68) = 89.27$ , if we chose 1.64 which is the same distance from .0500 in appendix B, we would get **89.16**.
  - 5- So 89.27 is the 95th percentile.

## Ch 6

### CH 6 Terms:

- Sampling
- **NOTE:** Sampling that relies on random selection is more likely to result in a representative sample than sampling based on purposeful methods of selection
- Probability samples are affected by sampling error
- Sampling error occurs in probability samples because sample statistics always fluctuate in random ways around the value of population parameters
- Probability Principles
- Probability Sampling
  - Simple Random Sample
  - Systematic Random Sample
  - Stratified Random Sample
- Sampling Distribution
- **Central Limit Theorem** - If all possible random samples of size N are drawn from a population with a

mean ( $\mu$ ) & a standard deviation ( $\sigma$ ), then as N increases, the sampling distribution of sample means becomes approximately normal.

- Variability of the sampling distribution of the mean is affected by the sample size & the variability of the variable (aka the diversity)

# ***Central Limit Theorem***

## **Size of the Sample:**

- A general rule is that when  $N = 50$  or more, the sampling distribution of the mean will be **approximately normal** regardless of the shape of the distribution.
  - But we can assume that the sampling distribution will be normal even with samples as small as 30, if we know that the population distribution approximates normality.
- Central limit theorem describes the properties of the sampling distribution

## ***Z Score for Sampling Distribution***

**Formula:**

$$Z = \frac{\bar{Y} - \mu_{\bar{Y}}}{\sigma / \sqrt{N}}$$

## Ch 7

### CH 7 Terms:

- **Estimation** - A process where we select a random sample from a population & use a sample statistic to estimate a population parameter.
- **Point & Interval Estimation** - Point estimates are sample statistics used to estimate the exact value of a population parameter. And in Interval estimates (confidence intervals), we use a range of values within which the population parameter may fall.
  - Point estimates are more precise
  - Interval estimation is more of a spray & pray.
- **Confidence Interval** (links to formula for mean) - A range of values defined by the confidence level within which the population parameter is estimated to fall.
- **Confidence Level** - The likelihood, expressed as a percent or a probability, that a specified interval will contain the population parameter.
- **Standard Error of the mean** - measures amount of dispersion there is in the sampling distribution, or how much variability there is in the value of the mean from the sample to sample.
  - Measures the variability in the sampling distribution
  - The variability of the sampling distribution of the mean is affected by the sample size & the variability of the variable.
- **Confidence Interval for Proportions & Percents** - Procedure for estimating proportions & percents are identical. Sampling distribution of proportions underlies estimation of population proportions from the sample proportions which is the same theory used for the sampling distribution of the mean.



## ***Reason for Estimates***

### **Reason:**

- Too expensive & time consuming to find info on a population (& even impossible in some cases)
- Can learn a lot about a population by randomly selecting a sample from that population & obtaining an estimate of the population parameter.

## ***Confidence Level***

### **Confidence Level:**

- When we use confidence intervals to estimate population parameters we can also evaluate the accuracy of this estimate by assessing the likelihood that any given interval will contain the mean.
- Confidence level is expressed as a percent or probability.

## Confidence Intervals for Means

### Confidence Intervals:

- If we wanted to be 95% confident, all random sample means would fall within +/- 1.96 standard error(s) (which is our Z score)
- If we wanted to be 99% confident, all random sample means would fall within +/- 2.58 standard error(s) (which is our Z score)

Confidence Interval	Z-score
80%	1.28
85%	1.44
90%	1.65
95%	1.96
99%	2.58
99.5%	2.80
99.9%	3.29

- More here: <https://www.thecalculator.co/math/Confidence-Interval-Calculator-210.html>

- Confidence interval Formula:

$$CI = \bar{Y} \pm Z(\sigma_y)$$

- 1. You first need the standard error, here is the formula
  2. Decide on the level of confidence to use, this will lead you to the Z-score.
  3. Multiply the Z-score & the standard error
  3. Add & subtract the mean to get the **upperbound & lowerbound**

- **NOTE:** Remember, we can never be sure whether the population mean is actually contained within the confidence interval.

### Interpreting Results:

- We can be x% confident (confidence level here) that the actual mean value of the sample is not less than x and not greater than x

## ***Standard Error for Means***

### **Standard Error of the Mean:**

- The standard error is the standard deviation of a sampling distribution
- Formula:

$$\sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{N}}$$

- 
- But, since the standard error isn't generally known, we work instead with the **estimated standard error**, formula:

$$s_{\bar{Y}} = \frac{s_Y}{\sqrt{N}}$$

- 
- Simply divide the standard deviation by the square root of the total number of samples

## ***Reducing Risk***

### **Reducing Risk:**

- To reduce the risk of being incorrect in terms of our interval not containing the true population mean, we can increase the level of confidence being used
  - By increasing the level of confidence, this **widens our confidence interval** which **makes the estimate less precise**
    - **EX:** 95% confidence interval extends 1.96 standard errors above & below the sample mean but a 99% confidence interval extends 2.58 standard errors above & below the sample mean.
    - Estimate becomes less precise as the width of the interval increases.
- When we use the highest level of confidence (99% confident), there is only a 1% risk that we're wrong that the specified interval doesn't contain the true population mean

### **Sample Size & Confidence Intervals:**

- By increasing the sample size, researchers increase the precision of their estimate, and as a result decreases the confidence interval size
  - Larger samples result in smaller standard errors & therefore sampling distributions are more clustered around the population mean. A more tightly clustered sampling distribution means our confidence intervals will be narrower and more precise.

## ***Confidence Interval for Proportions/Percents***

### **Procedure for Estimating Proportions & Percents:**

- The formula for constructing confidence intervals for a proportion for any level of confidence is:

$$CI = p \pm Z (S_p)$$

- - Ci = Confidence interval
  - P = Observed sample proportion
  - Z = Z-score corresponding to the Ci
  - Sp = Estimated standard error of proportions
- 1> Calculate standard error of proportion
- 2> Decide on the level of confidence & find corresponding Z-score
- 3> Calculate confidence interval
- 4> Interpret results
- NOTE: You still have to do both addition & subtraction to find 2 values

### **Interpreting Results:**

- We are x% confident that the true population proportion is somewhere between x and x
- Express the results in percentages as it's easier to understand percents

### **Percents:**

- For percents, convert it into a proportion then use the proportions method of calculating the confidence interval.

## ***Standard Error of Proportion***

**Formula:**

$$S_p = \sqrt{\frac{(p)(1-p)}{n}}$$

- p = proportion
- $S_p$  = standard error of proportion
- n = total number of cases

## ***Chapter 8 - Hypothesis Testing***

### **Intro:**

- Since most estimates are based on single samples & different samples may result in different estimates, **sampling results can't be used directly to make statements about a population**

### **Homework:**

- Question 3 & 14 require 2-sample testing



# Terms

## Definitions:

- **Statistical hypothesis testing** - A procedure that allows us to evaluate hypotheses about population parameter based on sample statistics.

- **Research hypothesis** ( $H_1$ ) - A statement reflecting the substantive hypothesis. It's always expressed in terms of population parameter. The form varies from test to test. The research hypothesis usually specifies that the **population parameter** is one of the following:

- 1- Not equal to some specified value:  $\mu$  doesn't equal some specified value (we use  $\mu$  since it represents the population parameter). This is used when we have some theoretical basis to believe that there is a difference between groups, but we cannot anticipate the direction of that difference, here is where we conduct a two-tailed test & say the research hypothesis is not equal to some specified value.

- 2- Greater than some specified value:  $\mu >$  some specified value, this is considered a right-tailed test since the outcome is in the right tail of the sampling distribution.

- 3- Less than some specified value:  $\mu <$  some specified value, this is considered a left-tailed test since the outcome will be in the left tail of the sampling distribution.

- **One-tailed test** - A type of hypothesis test that involves a directional research hypothesis. It specifies that the values of 1 group are either larger or smaller than some specified population value

- **Right-tailed test** - a one-tailed test in which the sample outcome is hypothesized to be at the right tail of the sampling distribution

- **Left-tailed test** - A one-tailed test in which the sample outcome is hypothesized to be at the left tail of the sampling distribution

- **Two-tailed test** - A type of hypothesis test that involves a nondirectional research hypothesis. We're equally interested in whether the values are less than or greater than one another. The sample outcome may be located at both the lower & the higher ends of the sampling distribution.

- **Null Hypothesis** ( $H_0$ ) - A statement of "no difference" that contradicts the research hypothesis & is always expressed in terms of population parameters. Literally saying the research hypothesis is equal to some specified value, think of it as the opposite of the research hypothesis.

- In hypothesis testing, we hope to reject the null hypothesis to indirectly support the research hypothesis, this will strengthen our belief in the research hypothesis.

- **Z statistic** (obtained) - The test statistic computed by converting a sample statistic (EX: the mean) to a Z-score. The formula for obtaining Z varies from test to test. A negative Z indicates evaluation at the left tail of the distribution, while the opposite for a positive Z.

- **p value** - The probability associated w/the obtained value of Z. Measures how unusual or rare our obtained statistic is compared with what is stated in our null hypothesis. The larger the p value, we can assume that the null hypothesis is true.

- **Alpha** ( $\alpha$ ) - The level of probability at which the null hypothesis is rejected. It's customary to set alpha at the .05, .01, or .001 level. Null hypothesis is rejected when the p value ( $p$ ) is less than or equal to alpha ( $\alpha$ ).

- **Type 1 error** - The probability associated w/rejecting a null hypothesis when it's true.

- **Type 2 error** - The probability associated w/failing to reject a null hypothesis when it's false.

- **t statistic** (obtained) - The test statistic computed to test the null hypothesis about a population mean when the population standard deviation is unknown & is estimated using the sample standard deviation. Represents the number of standard deviation units (or standard error units) that our sample mean is

from the hypothesized value of  $\mu$  (the population mean)

- **t distribution** - A family of curves, each determined by its degrees of freedom (df). Used when the population standard deviation is unknown & the standard error is estimated from the sample standard deviation.

- **Degrees of freedom (df)** - The number of scores that are free to vary in calculating a statistic.

## Z Statistic (obtained)

### Probability Values & Alpha:

#### Formula:

- Before calculating, you first need the standard error, the formula for this is:

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{N}}$$

- Z Statistic (obtained) formula:

$$Z = \frac{\bar{Y} - \mu_{\bar{Y}}}{\sigma / \sqrt{N}}$$

■  $\bar{Y}$  = sample mean

■  $\mu_{\bar{Y}}$  = population mean

- Replace denominator with standard error (standard deviation)
- After calculation, look into C Column which tells us how improbable or probable (**p value**) it is to pull the sample mean from the population mean
- Converting the sample mean to a Z-score equivalent is called computing the **test statistic**, the Z value obtained is called the Z statistic
- This Z value gives us the number of standard deviations (standard errors) that our sample is from

the hypothesized value ( $\mu$  or  $\mu_{\bar{Y}}$ ), assuming the null hypothesis is true.

- Negative Z statistic would mean the difference would have to be evaluated at the left tail of the distribution
- Positive Z statistics would mean the difference would have to be evaluated at the right tail of the distribution
- **P value** is the probability of pulling that sample from the population, measure of how unusual or rare our obtained statistic is compared w/what's stated in our null hypothesis.
- The smaller the p value, the more evidence we have that the null hypothesis should be rejected in favor of the research hypothesis
- The larger the p value, we can assume null hypothesis is true & fail to reject it
- P value is compared to the **Alpha (a)**
- **Alpha (a)** is the level of probability at which the null hypothesis is rejected, generally set to .05, .01, or .001. If you could reject one alpha level, you can state that your research is statistically significant at the .05 level but not at the .01 or .001 alpha level.
- 0.05 -> 5 in 100 chance (most likely to make Type 1 error with, easiest to reject null)
- 0.001 -> .1 in 100 chance (most likely to make Type 2 error with, hardest to reject null)
- 0.01 -> 1 in 100 chance
- P value equal to or below the Alpha (a) level, we reject the null hypothesis.
- If the P value is above the Alpha (a) level, we fail to reject the null hypothesis.

#### Z For 2-tailed Test:

- After getting your **p value** that corresponds to the Z score in appendix B, you would multiply it by 2 to obtain the 2-tailed probability

## T Statistic

### T Statistic:

- Used instead of Z statistic since standard deviation won't be known in most cases
- Aka t-test
- Uses sample standard deviation
- Represents how many standard deviation units (or standard error units) our sample mean is from the hypothesized value of the population mean, assuming our null hypothesis is true.
- Formula for t statistic (obtained):

$$t = \frac{\bar{Y} - \mu_y}{\frac{S_y}{\sqrt{n}}}$$

$\bar{Y}$

- = sample mean

$\mu_y$

- = population mean

- Denominator is meant to calculate the estimated standard error


### T Distributions & Degrees of Freedom:

- T distributions is bell shaped
- T statistic can have positive & negative values: positive T statistics correspond to the right tail of the distribution while negative corresponds to the left tail.
- When the Degrees of Freedom (**df**) is small, the t distribution is much flatter than the normal curve.
  - As the df increases, the shape of the t distribution gets closer to the normal distribution, until the 2 are almost identical when the df is greater than 130
- Which distribution used is determined by degrees of freedom
- Degrees of freedom is calculated by taking sample size & subtracting the amount of restrictions you have
  - Amount of restrictions is defined as the number of samples
- Appendix C summarizes the T distribution
- The T table shows:
  - degrees of freedom
  - probabilities or alpha
  - significance levels
- **NOTE:** Since it's estimated from sample data, the denominator of the t statistic is subject to sampling error.
- After knowing your degrees of freedom (df), you go appendix C & grab the number that corresponds to the degrees of freedom & level of significance (aka Alpha), this is called the **t-critical**
  - With the **T-critical** & the T statistic (obtained), you can determine if you're going to reject or fail to reject your null hypothesis
    - If the value of the T-statistic is greater than the t-critical value, reject the null hypothesis
    - If the value of the T-statistic is less than the t-critical value, you fail to reject the null hypothesis

## For 2 Samples:

First steps:


- 1) Find mean
  - 2) Find N (number of samples)
  - 3) Find variance (to get variance from standard deviation, square the standard deviation)
- If one of the variances are more than double of the other, you use this equation to find the estimated standard error:


$$S_{\bar{Y}_1} - S_{\bar{Y}_2} = \sqrt{\frac{(N_1 - 1) S_{y_1}^2 + (N_2 - 1) S_{y_2}^2}{(N_1 + N_2) - 2}} \sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

- To calculate degrees of freedom:
  - Take N from 1st sample & add it to 2nd sample
  - Subtract the total by 2 (the number of samples there are)

$$df = (N_1 + N_2) - 2$$

- T-test formula:


$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{\bar{y}_1} - S_{\bar{y}_2}}$$

Difference between the Means

The estimated standard error

## **Lecture Notes**

### **Assumptions (Requirements) of Statistical Hypothesis Testing**

- Sample is a random sample
- dependent variable is measured at interval-ratio level
- population is normally distributed or that the sample size is larger than 50
- We can only at best, estimate the likelihood that the research hypothesis is true or false

### **Null Hypothesis:**

- Rejecting null hypothesis indirectly supports research hypothesis
- We can either **reject** or **fail to reject it**
  - We never accept it
- Rejecting null hypothesis for 2 tail test is more difficult than it is for 1 tail tests

## ***5 Steps to Hypothesis Testing***

### **5 Steps to Hypothesis Testing:**

- Statistical hypothesis testing can be organized into 5 basic steps:
  1. Making assumptions
  2. Stating the research & null hypotheses & selecting alpha
  3. Selecting the sampling distribution & specifying the test statistic
  4. Computing the test statistic
  5. Making a decision & interpreting the results