MG Analytics

Inferential Statistics

By MG ANALYTICS

MG Analytics

Types:

- Descriptive Statistics:
- It consists of the collection, organization, summarization, and presentation of data. (It describes the situation as it is).

- Inferential Statistics:
- It consists of making inferences from samples to populations, hypothesis testing, determining relationships among variables, and making predictions. (It is based on probability theory.)

Why ?

Descriptive **statistics** describes data

inferential statistics allows you to make predictions (“inferences”) from that data.

With **inferential statistics**, you take data from samples and generalize about a population.

Why?

- Inferential statistics provide a way of going from a "sample" to a "population" inferring the "parameters" of a population from data on the "statistics" of a sample.
- It is usually necessary for a researcher to work with samples rather than a whole population. but one difficulty is that a sample is generally not identical to the population from which it comes.
- Another difficulty is that no two samples are the same. How can we know which best describes the population? We need rules that relate samples to population.

Estimation Techniques

- There are two different approaches for estimating: Point Estimation and Interval Estimation.
- • **For Point Estimation** you give one value for a characteristic, which is hopefully close to the true unknown value.
- We can not expect to find the precise value describing the population when only using data of a sample.
- • **For Interval Estimation** you give an interval of likely values, where the width of the interval will depend on the confidence you require to have in this interval.

Caution!!

- Statistics NEVER prove anything, instead, they indicate a relationship within a given probability of error.
- An association does not necessarily indicate a sure cause effect relationship.
- Statistics can always be wrong, however, there are things that researchers can do to improve the likelihood that the statistical analysis is correctly identifying a relationship between variables.

Variable Types

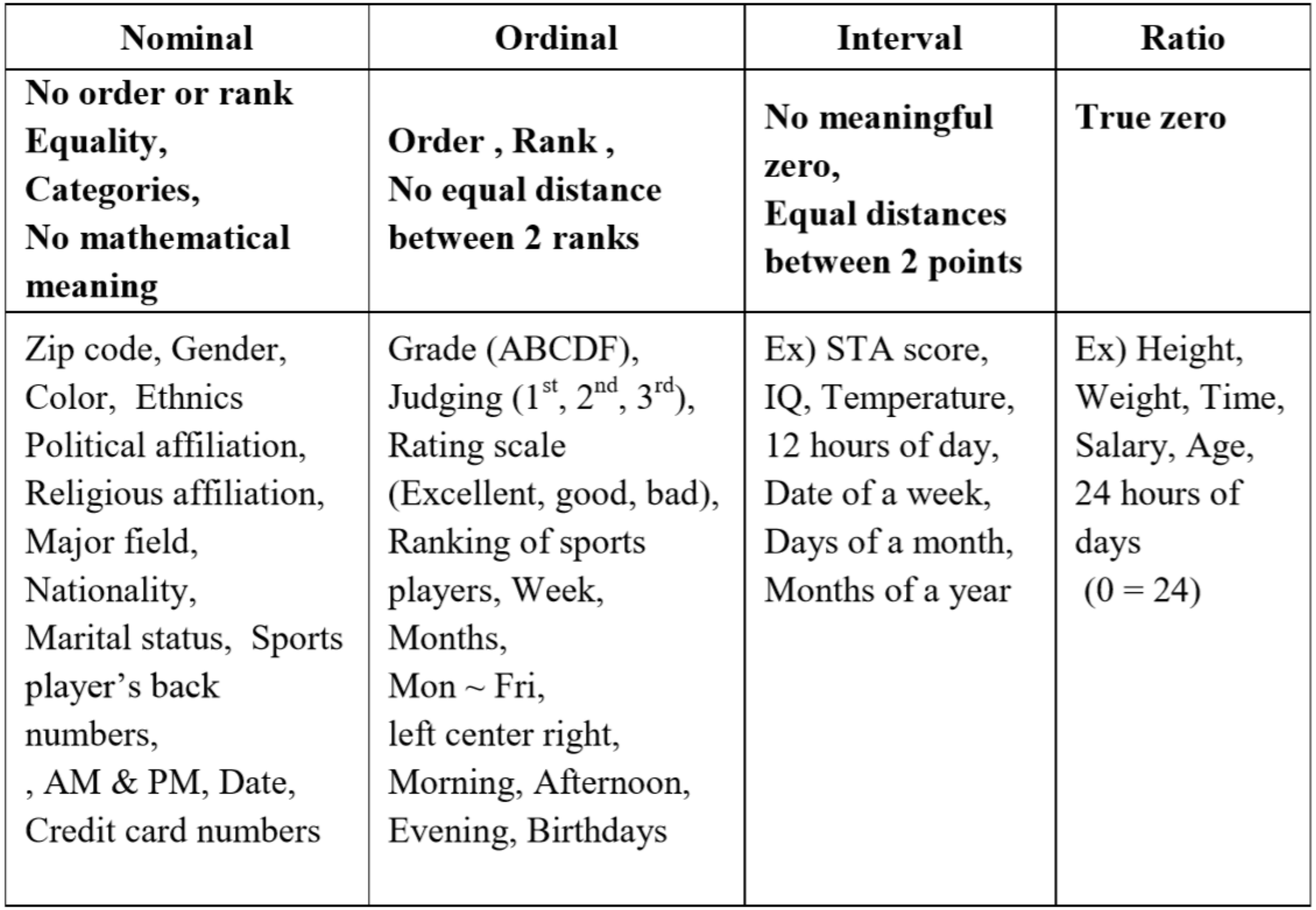
- **Qualitative Variables** – No mathematical meaning or Non-numerical variables that can be placed into distinct categories, according to some characteristic or attribute. Ex) gender, religious preferences, geographic locations, grades of a student, car"s tags, numbers on the uniforms of baseball players, etc.

- **Quantitative Variables -** numerical in nature and can be ordered or ranked. Ex) age, heights, weights, body temperatures, etc.
  - **Discrete Variables :** assume values that can be counted such as whole numbers.

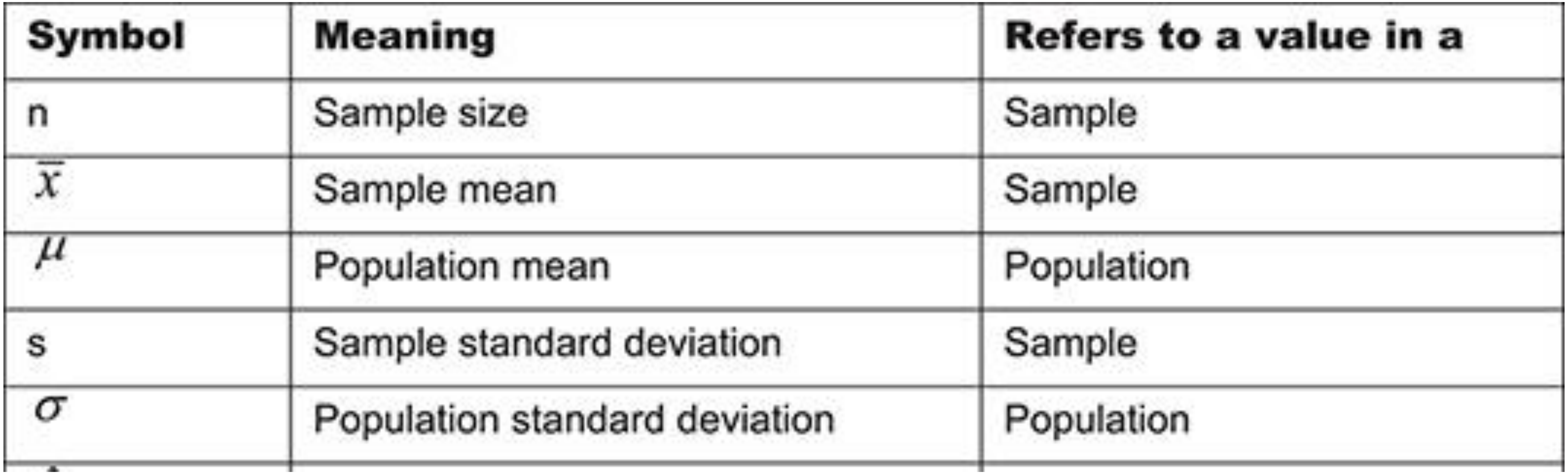
    Ex) the number of children in a family, the number of students in a class-room, the number of calls received by a switchboard operator each day for one month, batting order numbers of baseball, etc.
  - **Continuous Variables:** can assume all values between any two specific values by measuring.

    Ex) Temperature, height, weight, length, time, speed, etc.

Measurement Scales Of Data



Nominal	Ordinal	Interval	Ratio
No order or rank Equality, Categories, No mathematical meaning	**Order , Rank , No equal distance between 2 ranks**	**No meaningful zero, Equal distances between 2 points**	**True zero**
Zip code, Gender, Color, Ethnics Political affiliation, Religious affiliation, Major field, Nationality, Marital status, Sports player's back numbers, , AM & PM, Date, Credit card numbers	Grade (ABCDF), Judging (1st, 2nd, 3rd), Rating scale (Excellent, good, bad), Ranking of sports players, Week, Months, Mon ~ Fri, left center right, Morning, Afternoon, Evening, Birthdays	Ex) STA score, IQ, Temperature, 12 hours of day, Date of a week, Days of a month, Months of a year	Ex) Height, Weight, Time, Salary, Age, 24 hours of days (0 = 24)



Symbol	Meaning	Refers to a value in a
n	Sample size	Sample
$\bar{x}$	Sample mean	Sample
μ	Population mean	Population
s	Sample standard deviation	Sample
σ	Population standard deviation	Population

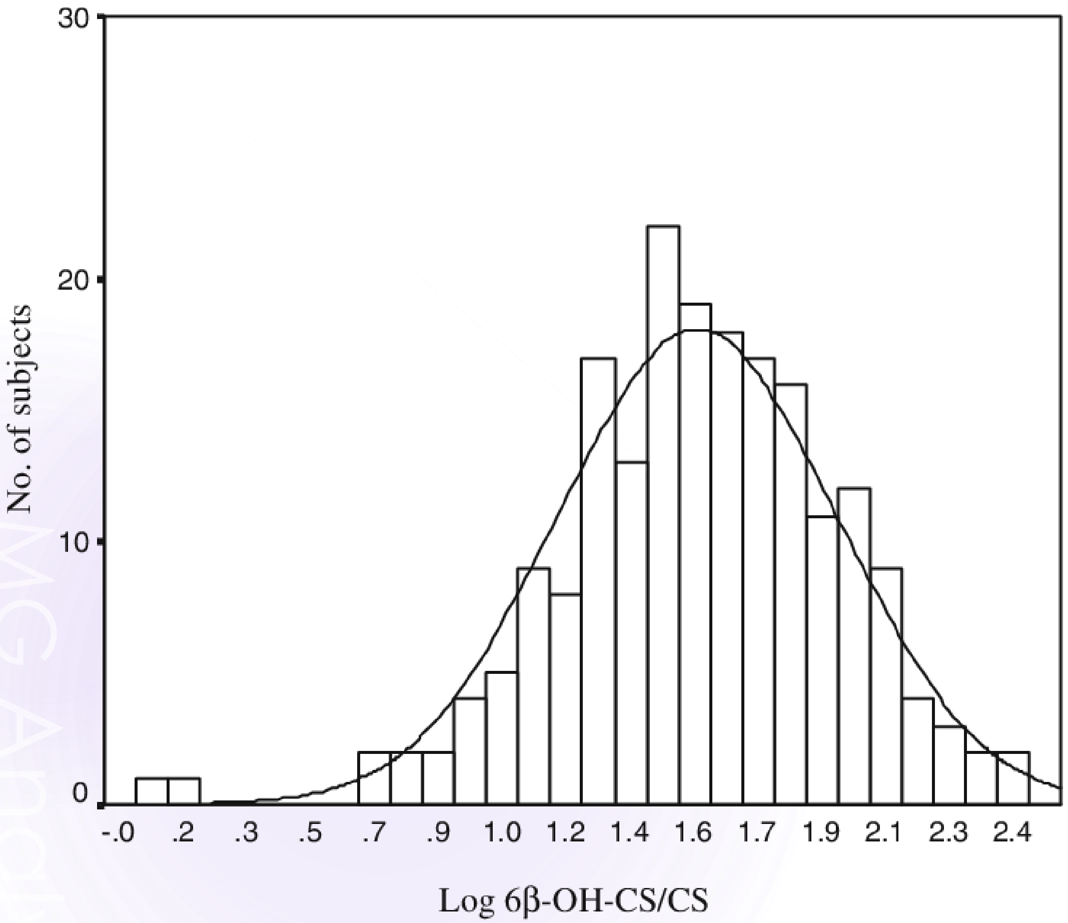
Frequency Distribution vs Probability Distribution.

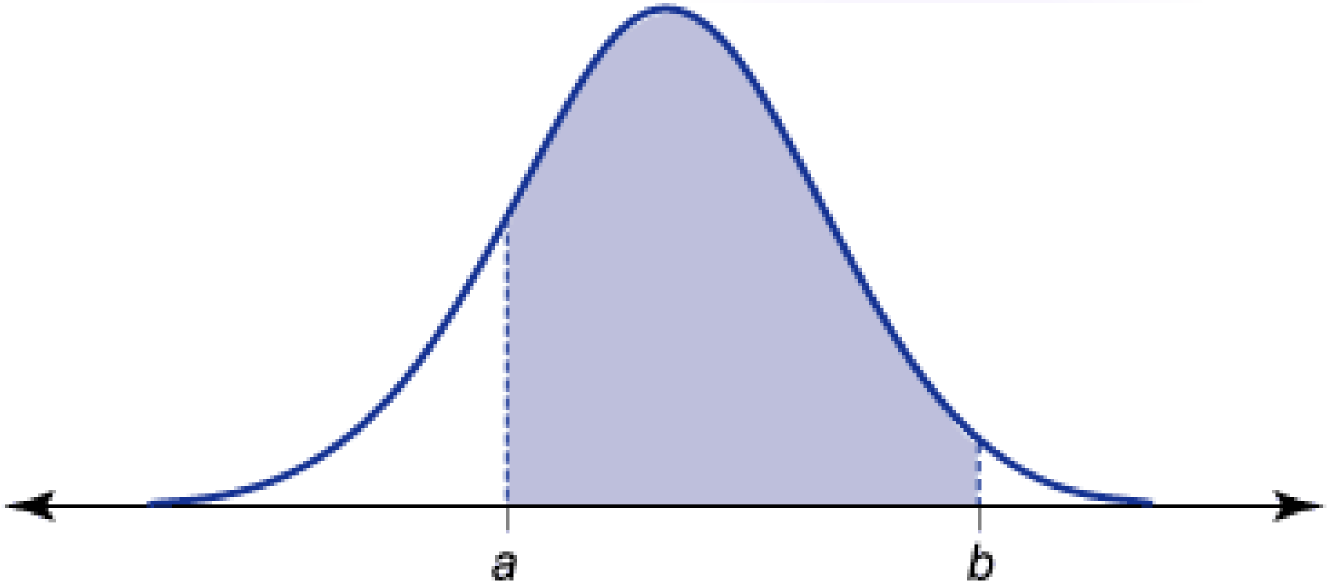
frequency distribution gives the exact **frequency** or the number of times a data point occurs

Probability distribution gives the **probability** of occurrence of the given data point.

When the number of test cases are large, the **frequency distribution** and the **probability distributions** are similar in shape.

Data	Frequency	Relative frequency	Probability
10	2	2/50	0.04
11	3	3/50	0.06
12	5	5/50	0.1
13	6	6/50	0.12
14	4	4/50	0.08
15	0	0/50	0
16	14	14/50	0.28
17	10	10/50	0.2
18	6	6/50	0.12
Total	**50**	**1**	**1**


30
20
10
0
No. of subjects
-.0
.2
.3
.5
.7
.9
1.0
1.2
1.4
1.6
1.7
1.9
2.1
2.3
2.4
Log 6β-OH-CS/CS


a
b

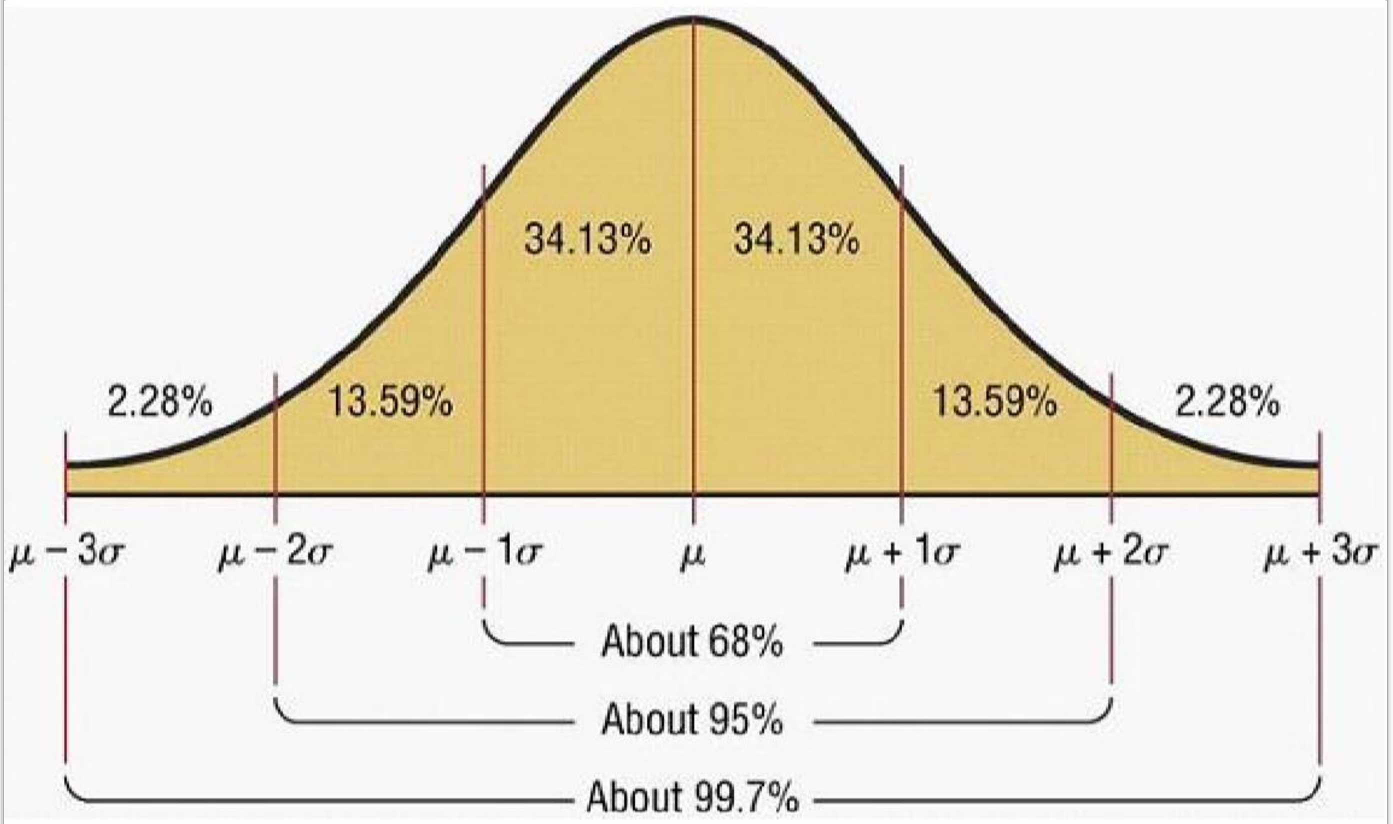
“The Distribution of Sample Means”

- Definition: the distribution of sample means is the collection of sample means for all the possible random samples of a particular size (n) that can be obtained from a population.
- It is not a distribution of scores, but a distribution of statistics.
- This distribution tends to be normal.
- It will be almost perfectly normal if either:
  - the population from which the sample is drawn is normal, or
  - the n of the sample is relatively large (30 or more). This distribution has a mean that is equal to the population mean;

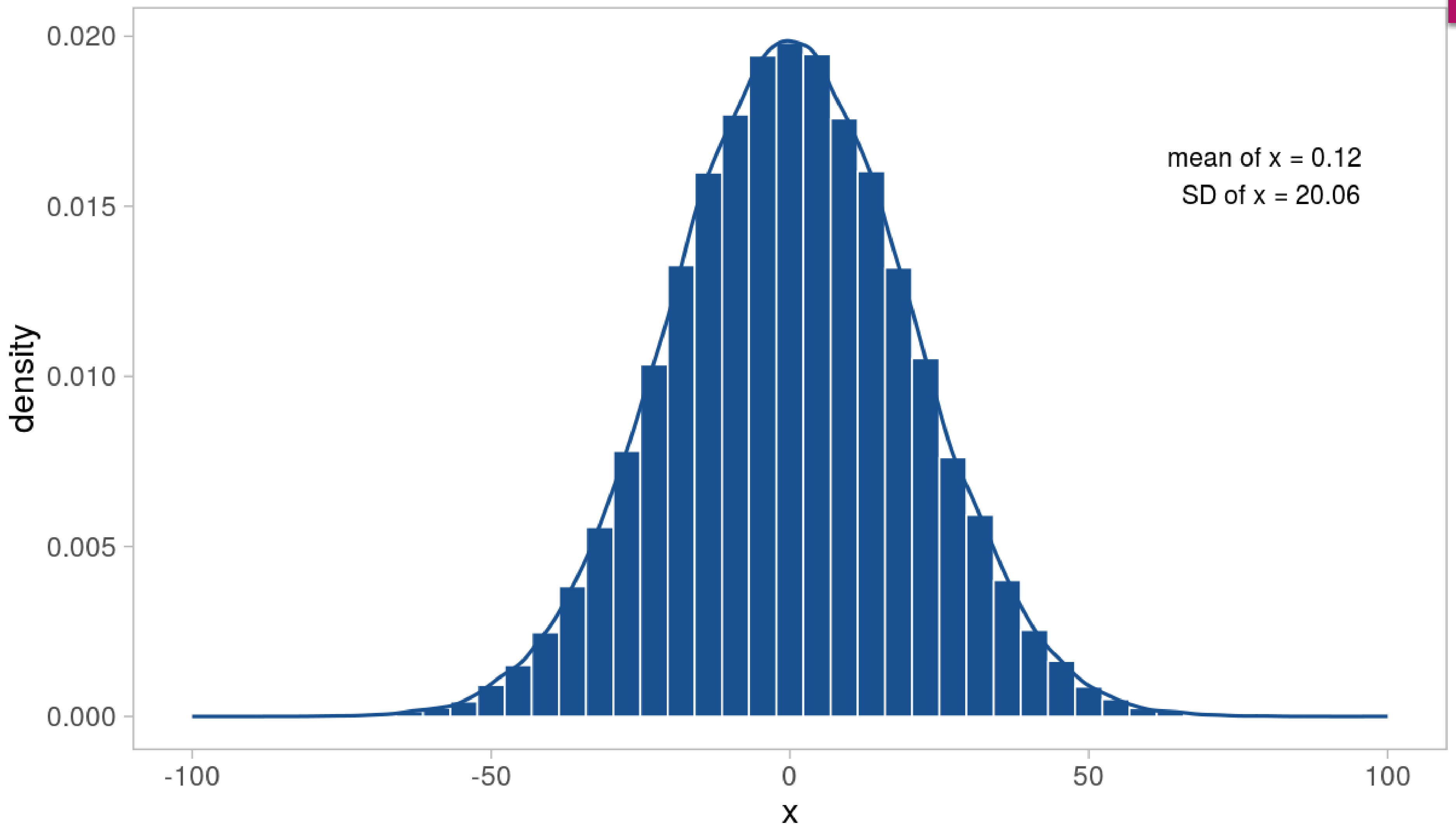
Properties of Normal Curve

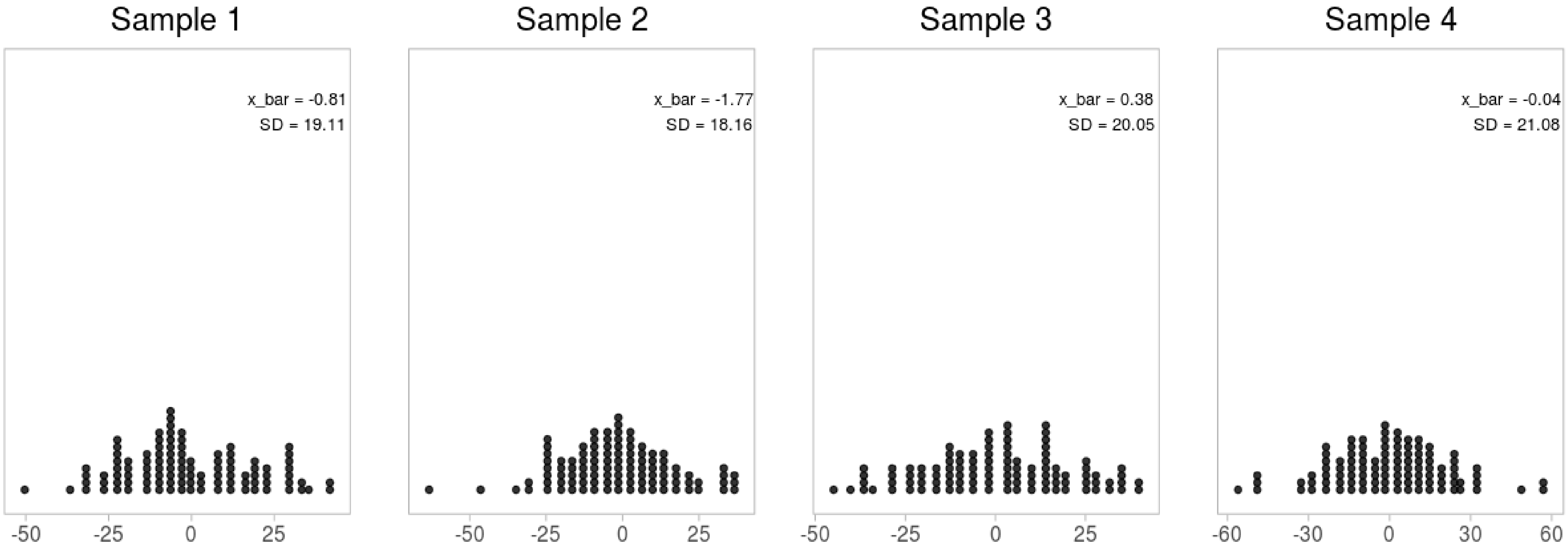
- 1. A normal distribution curve is bell-shaped
- The mean, median, and mode are equal and are located at the center of the distribution
- A normal distribution curve is unimodal (i.e., it has only one mode)
- The curve is symmetric about the mean, which is equivalent to saying that its shape is the same on both sides of a vertical line passing through the center
- The curve is continuous, that is, there are no gaps or holes. For each value of X, there is a corresponding value of Y
- The curve never touches the x axis. Theoretically, no matter how far in either direction the curve extends, it never meets the x axis – but it gets increasingly closer
- The total area under a normal distribution curve is equal to 1 or 100%
- Area under curve gives the probability of a point falling in the area

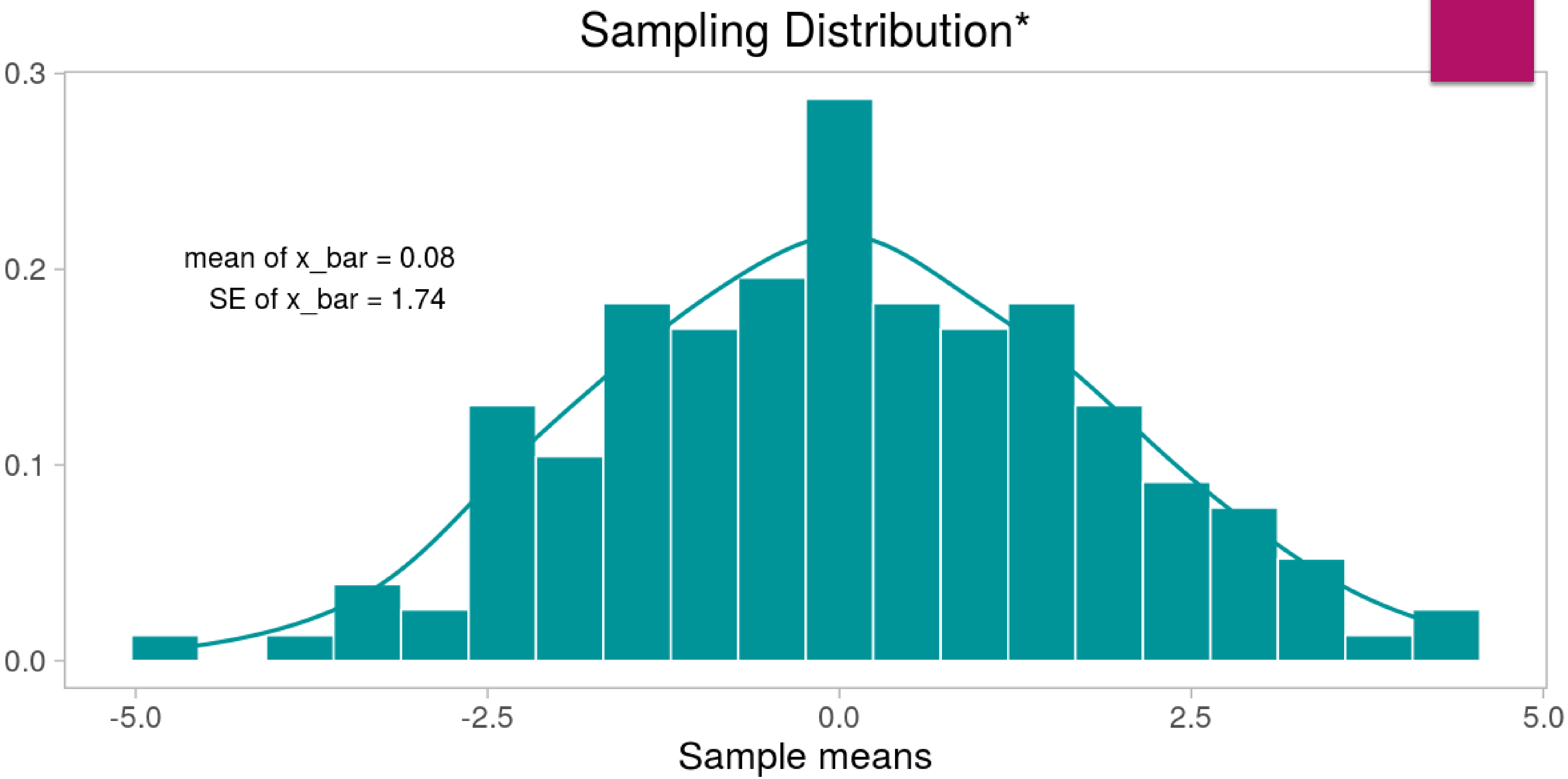
- The area under the part of a normal curve that lies within 1 standard deviation of the mean is approximately 0.68, or 68%
- within 2 standard deviations, about 0.95, or 95%;
- and within 3 standard deviations, about 0.997, or 99.7%. The Empirical rule applies.
-



MG Analytics


Population distribution: Normal


Sample 1
x_bar = -0.81
SD = 19.11
-50
-25
0
25
Sample 2
x_bar = -1.77
SD = 18.16
-50
-25
0
25
Sample 3
x_bar = 0.38
SD = 20.05
-50
-25
0
25
Sample 4
x_bar = -0.04
SD = 21.08
-60
-30
0
30
60



Sampling Distribution*

*Distribution of means of 160 random samples, each consisting of 103 observations from a normal population

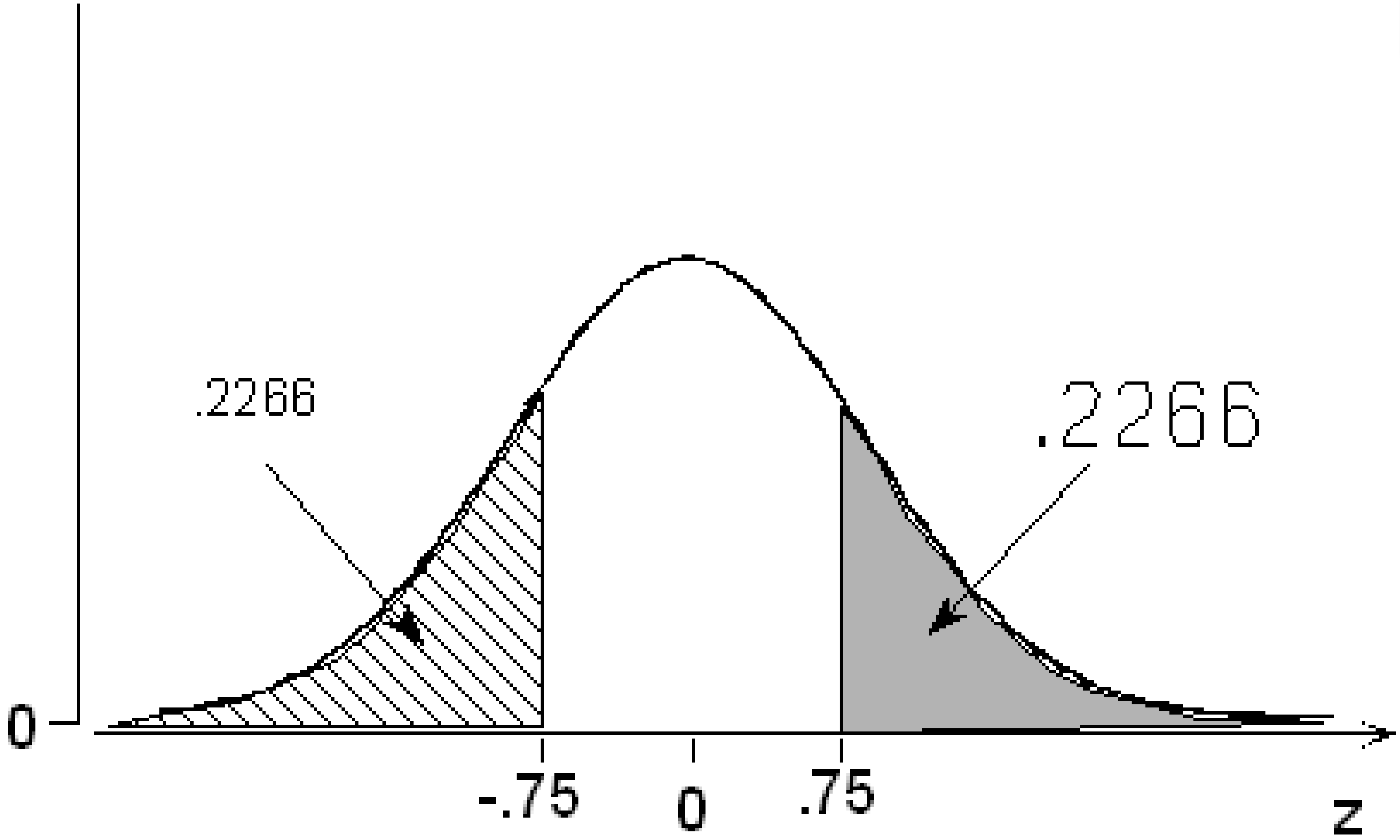
Central Limit Theorem

- The distribution of sample means (the sampling distribution) should be nearly normal.
- The mean of the sampling distribution should be approximately equal to the population mean (0.12)
- The standard error is the standard deviation of sample means
- SE = SD /SQRT(N) => (20.06/sqrt(103) = 1.98).

Z- score standardization

- It is used to scale the data so that all the variables have same range.
- Example : Age and Salary.
- For standard normal distribution, the mean is 0 and standard deviation in 1
- Z =(x-μ)/ σ

SNO	Age	Salary	Scaled Age	Scaled Salary
1	20	20500	-1.29352	-1.45646
2	24	25000	-0.92021	-0.92299
3	30	30000	-0.36041	-0.33024
4	34	34000	0.013072	0.143954
5	36	35000	0.199813	0.262503
6	41	40000	0.666667	0.855249
7	52	45000	1.693744	1.447995
Mean μ =	33.85714	32785.71429		
STDEV σ =	10.71492	8435.328204		


.2266
.2266
0
-.75
0
.75
z

Z SCORE

- **Standard Scores or z score (z) -** a comparison of a relative standard like both(the mean and standard deviations) can be made.
- Number of standard deviations a data value is above or below the mean for a specific distribution of values
- **z-score** is a very useful statistic because it allows us to calculate the probability of a **score** occurring within our normal distribution
- it enables us to compare two **scores** that are from different normal distributions.
- If a Z-score is 0, it indicates that the data point's score is identical to the mean score.
- A Z-score of 1.0 would indicate a value that is one standard deviation from the mean.
- Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

$$z = \frac{Value - Mean}{\text{Standard Deviation}} = \frac{X - \bar{X}}{s}\ (sample) = \frac{X - \mu}{\sigma}\ (population)$$

Finding probabilities (area) for a normally distributed variable by transforming it into a standard normal variable by using the formula..

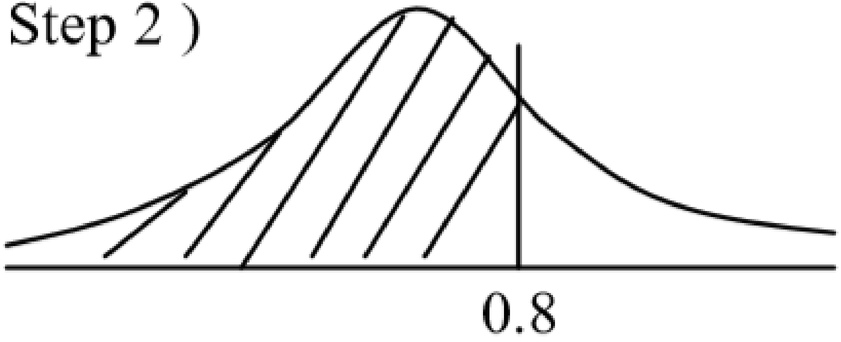
Ex 1] The average or the mean = 3.1 hours The standard deviation = 0.5

Find the percentage of less than 3.5 hours.

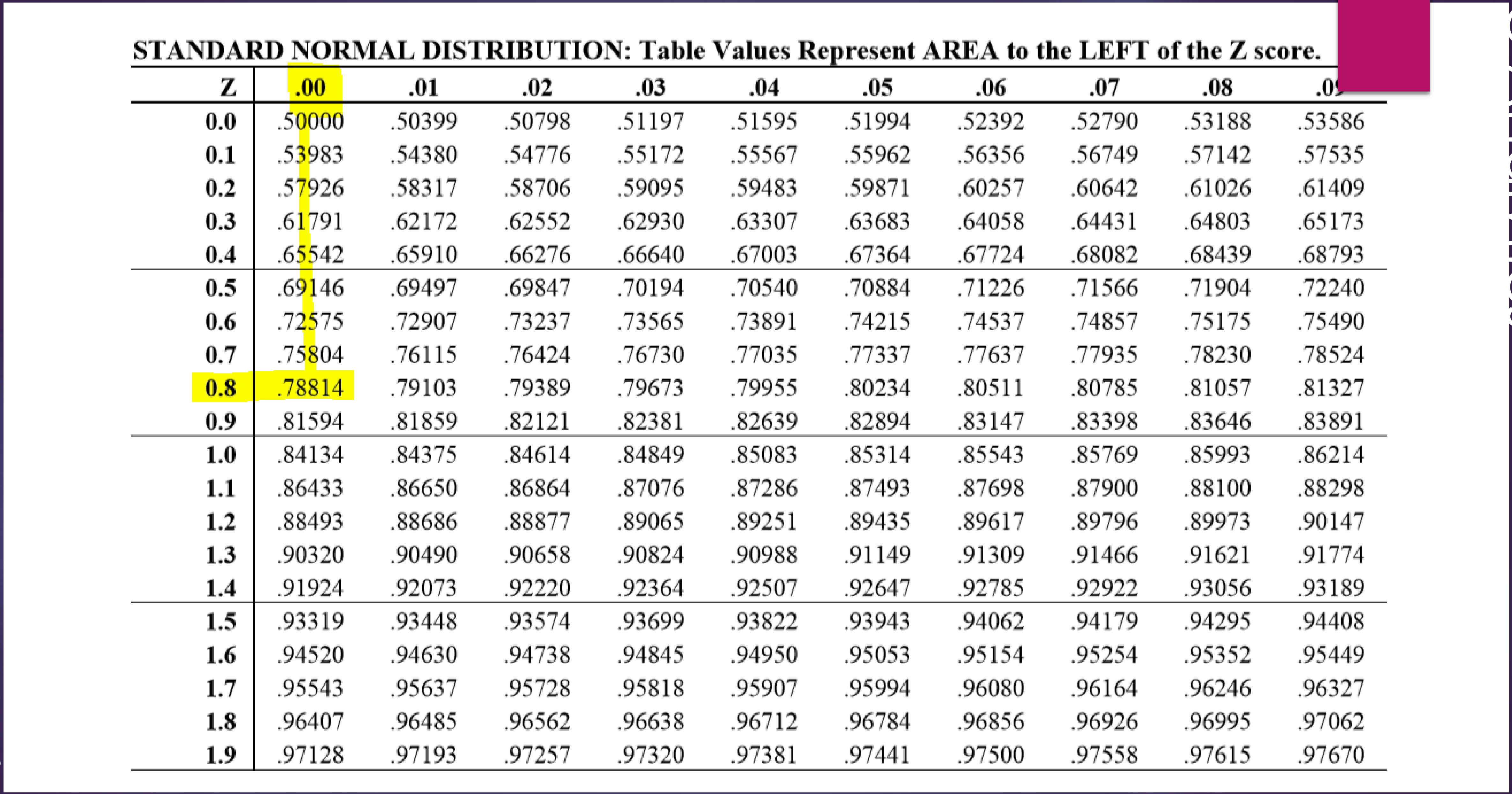
Step 1)

$$z = \frac{X - \mu}{\sigma} = z = \frac{3.5 - 3.1}{0.5} = 0.80$$

Step 2)

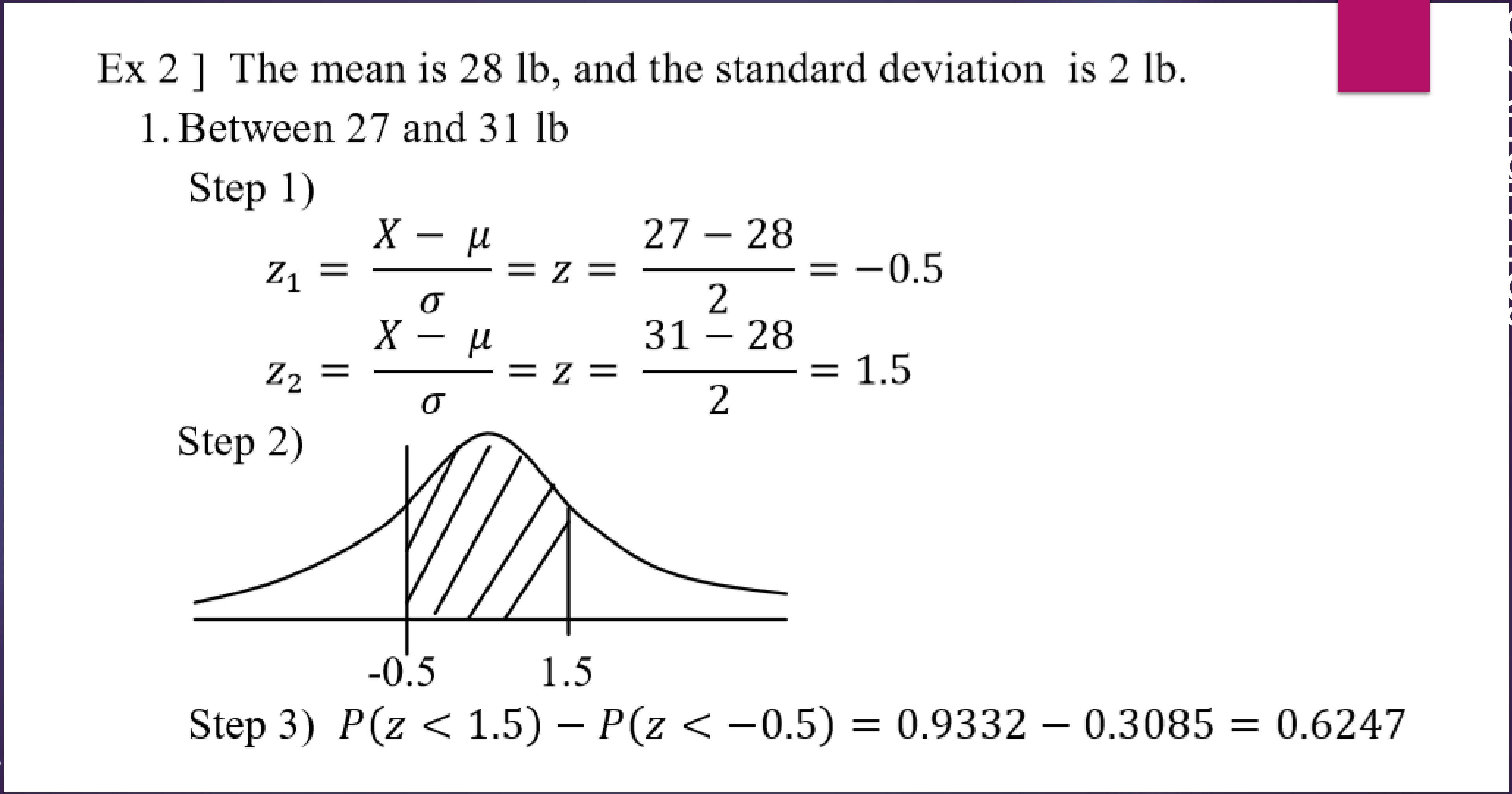


Step 3) 0.8 + .00 → 0.7881



STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670



Ex 2] The mean is 28 lb, and the standard deviation is 2 lb.

1. Between 27 and 31 lb

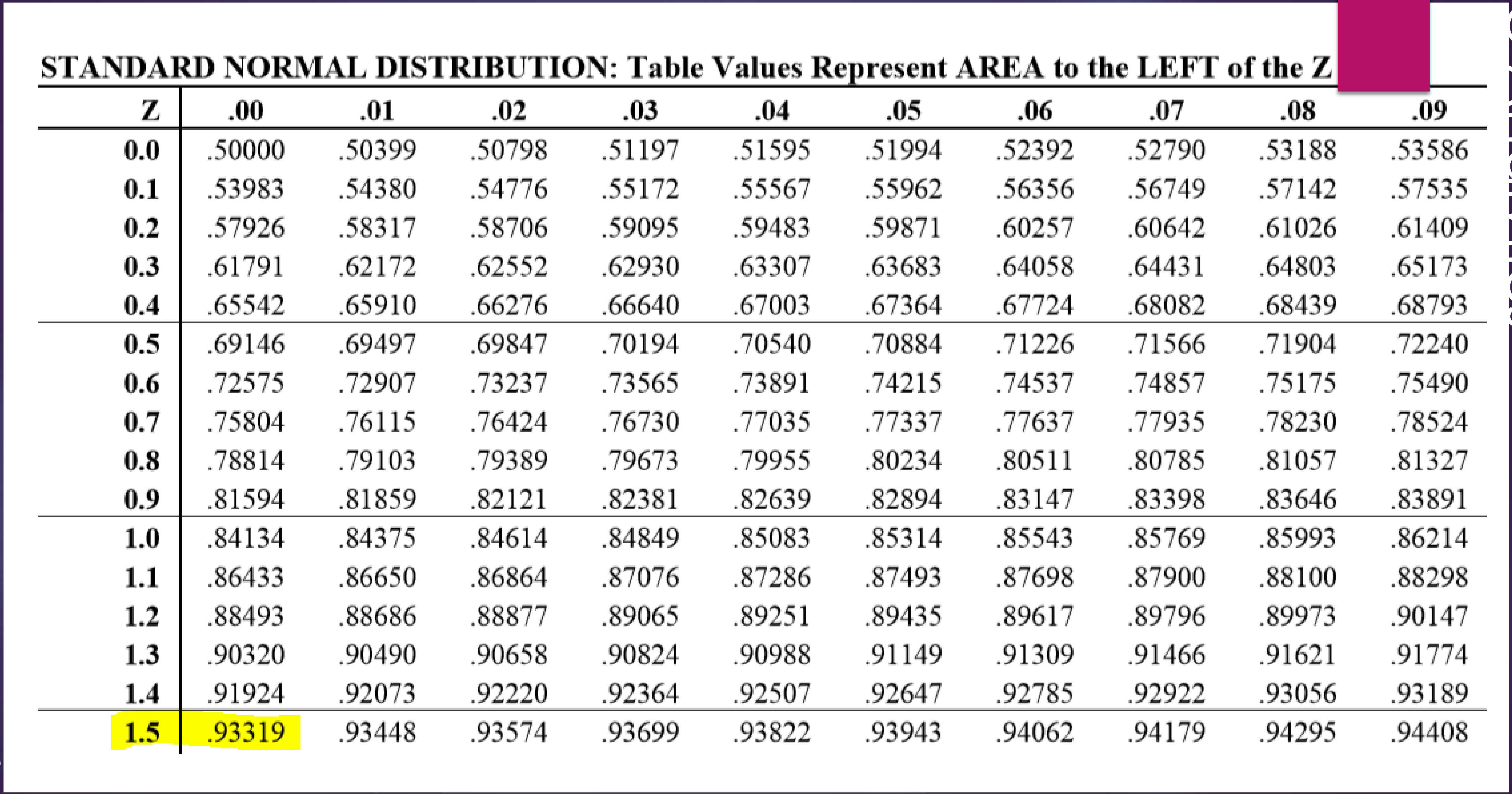
Step 1)

$$z_1 = \frac{X - \mu}{\sigma} = z = \frac{27 - 28}{2} = -0.5$$

$$z_2 = \frac{X - \mu}{\sigma} = z = \frac{31 - 28}{2} = 1.5$$

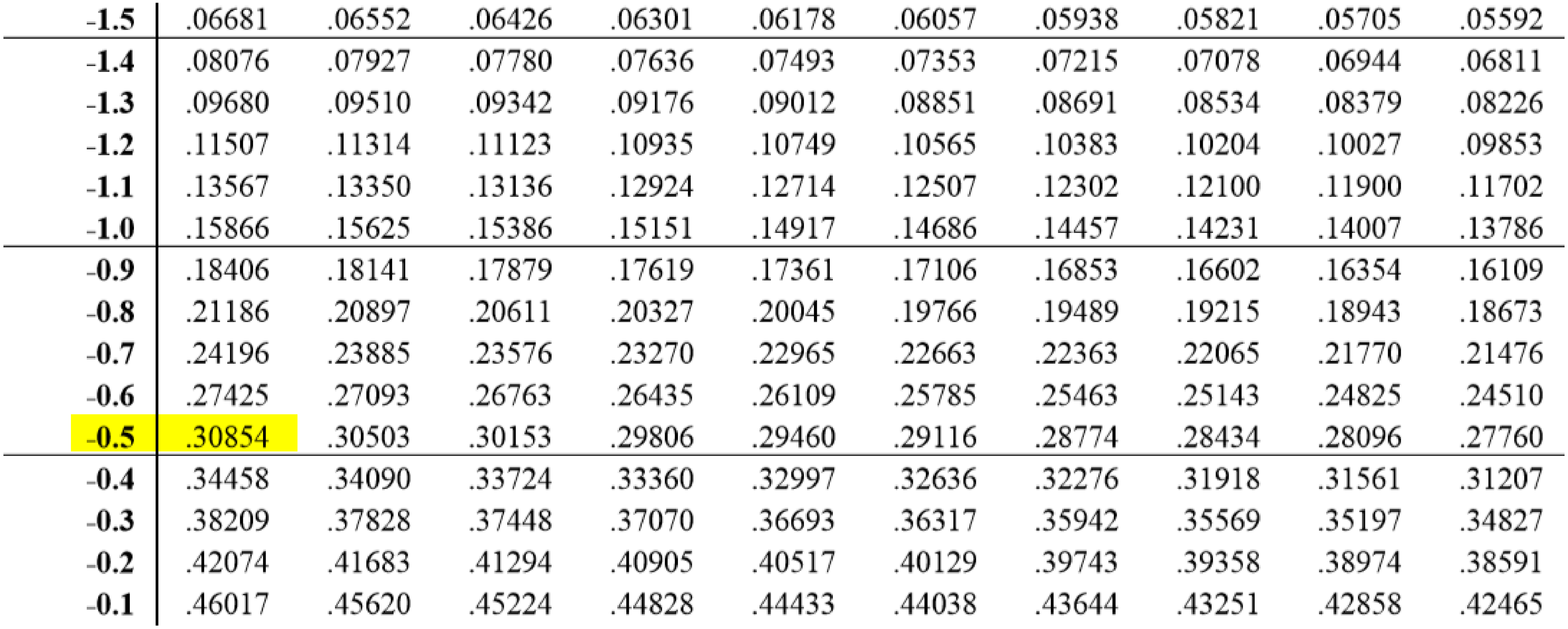
Step 2)

Step 3) $P(z < 1.5) - P(z < -0.5) = 0.9332 - 0.3085 = 0.6247$



STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408

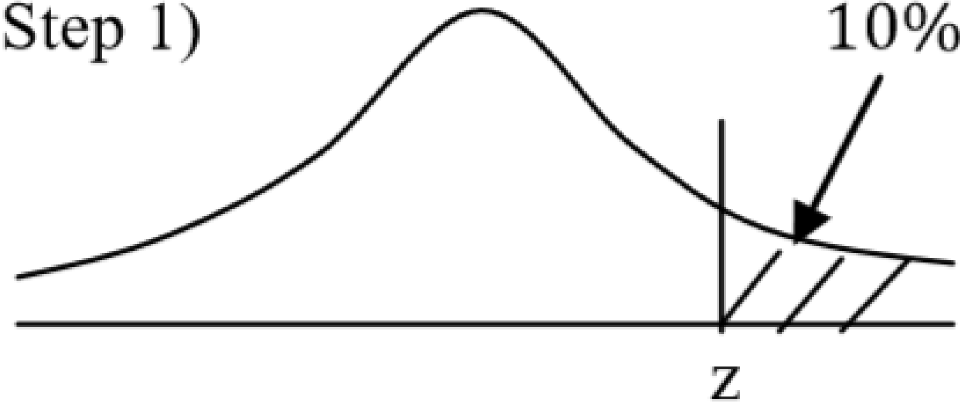


-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465

Finding specific data values for given percentage, using the standard normally distribution.

Ex 3) In the top 10%, the mean is 200 and the standard deviation is 20.
Find the lowest possible score to quality.

Step 1)

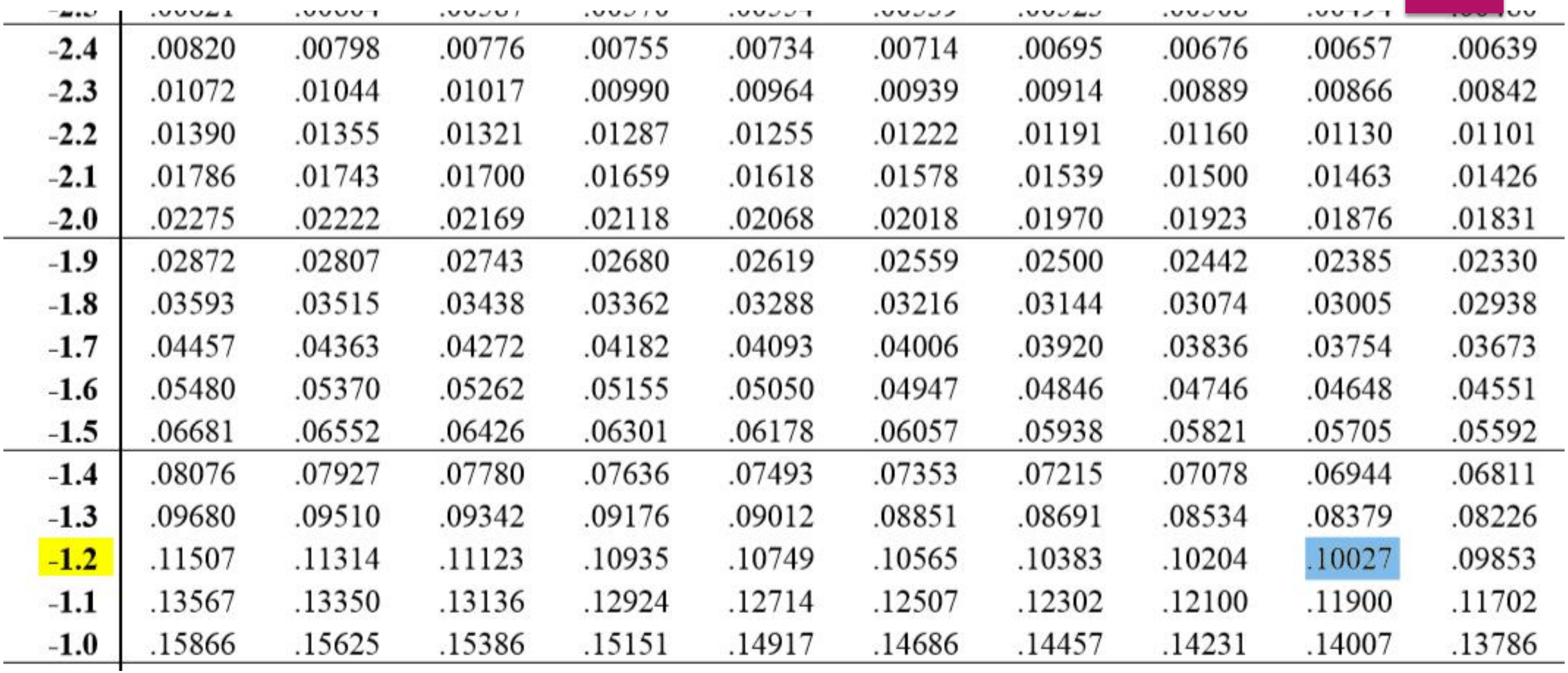


10% = 0.10

Step 2) Find the z in the table $0.10 \approx 0.003 \rightarrow -1.28 \rightarrow z = 1.28$

Step 3) $X_1 = z \cdot \sigma + \mu = 1.28 \cdot 20 + 200 = 226$

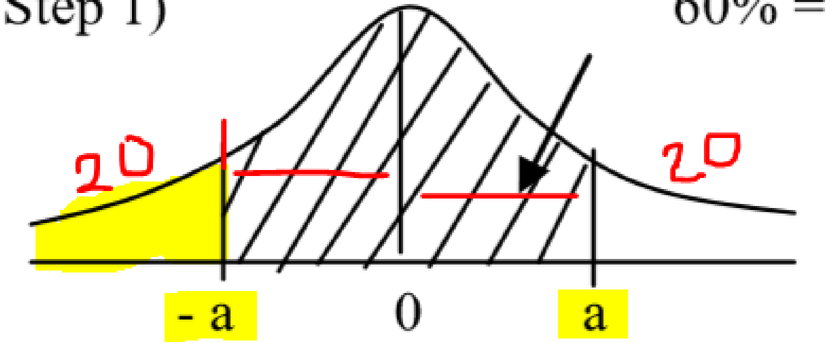
Step 4) $X > 226$



-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786

Ex4) To select in the middle 60% of the population, the mean is 120, and the standard deviation is 8. Find the upper and lower values

Step 1) 60% = 0.60



Step 2) $P\,(-a < z < a) = 60\% = 0.60$

$= 1 - [P(z < -a) \times 2]$ *Since it's symmetric*

$P\,(z < -a) = \left(1 - P\,(-a < z < a)\right) \div 2 =$

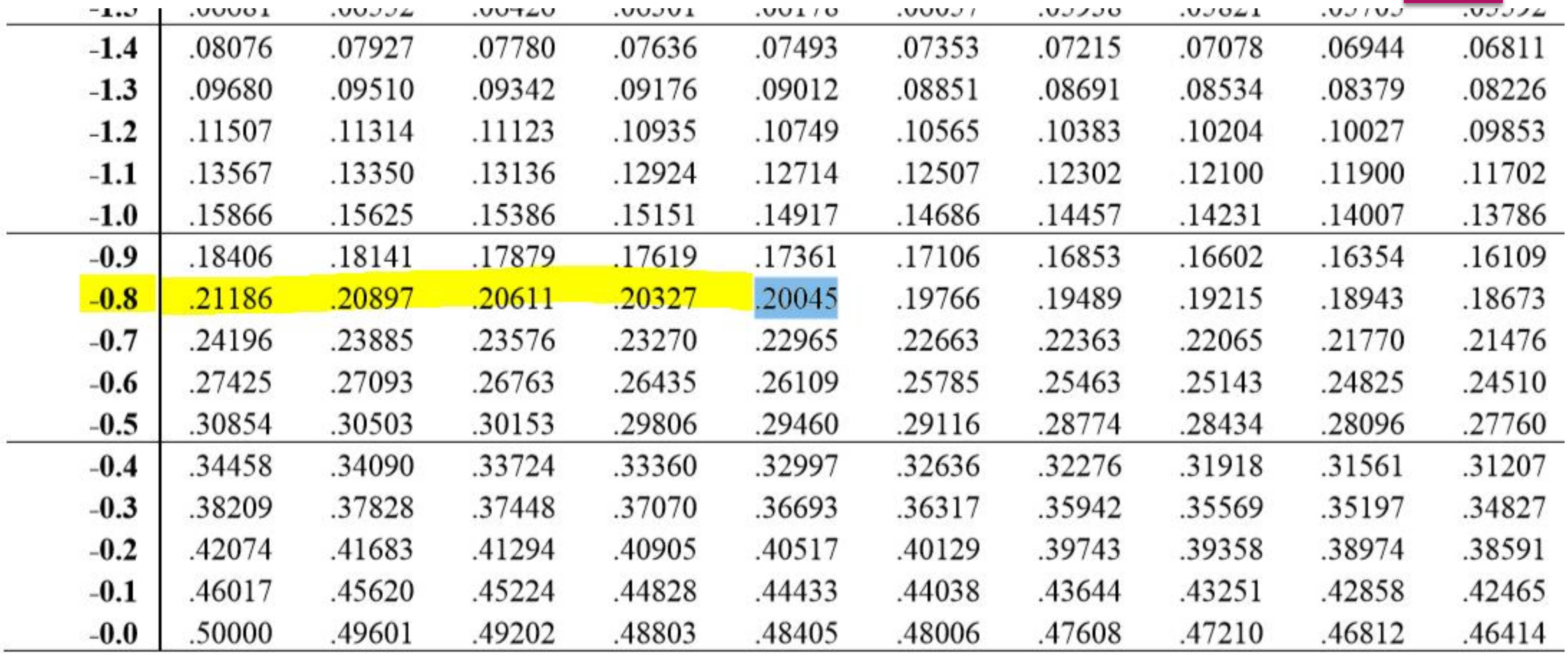
$= (1 - 0.60) \div 2 = 0.2 \rightarrow 0.205 \text{ or } 0.1977$

$0.205 \text{ is the closest z value } \rightarrow -0.84 \rightarrow \pm a = \pm 0.84$

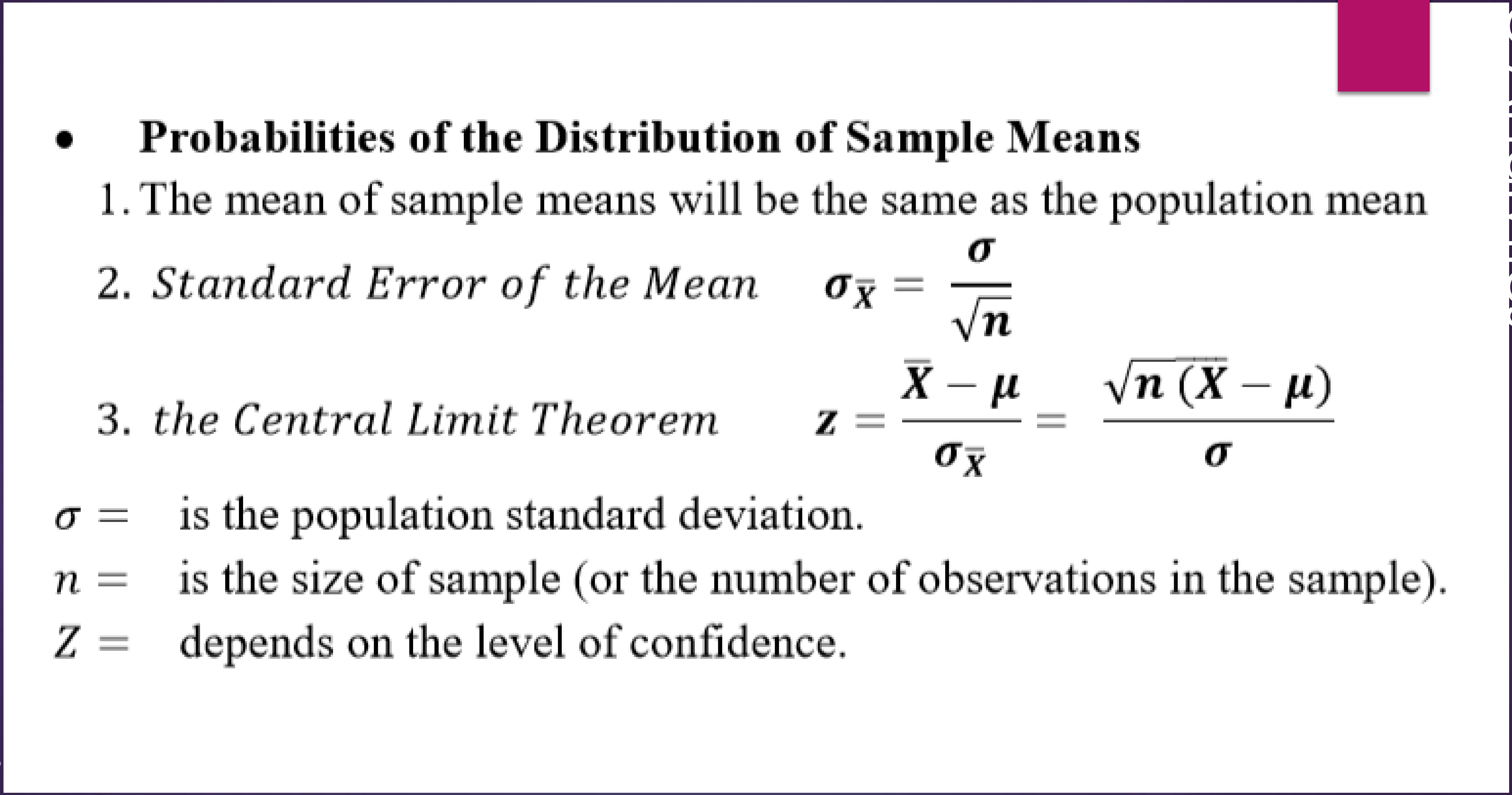
Step 3) $X_1 = z \cdot \sigma + \mu = 0.84 \cdot 8 + 120 = 12.72$

$X_2 = z \cdot \sigma + \mu = -0.84 \cdot 8 + 120 = 113.28$

Step 4) $113.28 < X < 126.72$



-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414



- **Probabilities of the Distribution of Sample Means**

1. The mean of sample means will be the same as the population mean

2. *Standard Error of the Mean* $$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

3. *the Central Limit Theorem* $$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\sqrt{n}\,(\bar{X} - \mu)}{\sigma}$$

$\sigma =$ is the population standard deviation.

$n =$ is the size of sample (or the number of observations in the sample).

$Z =$ depends on the level of confidence.

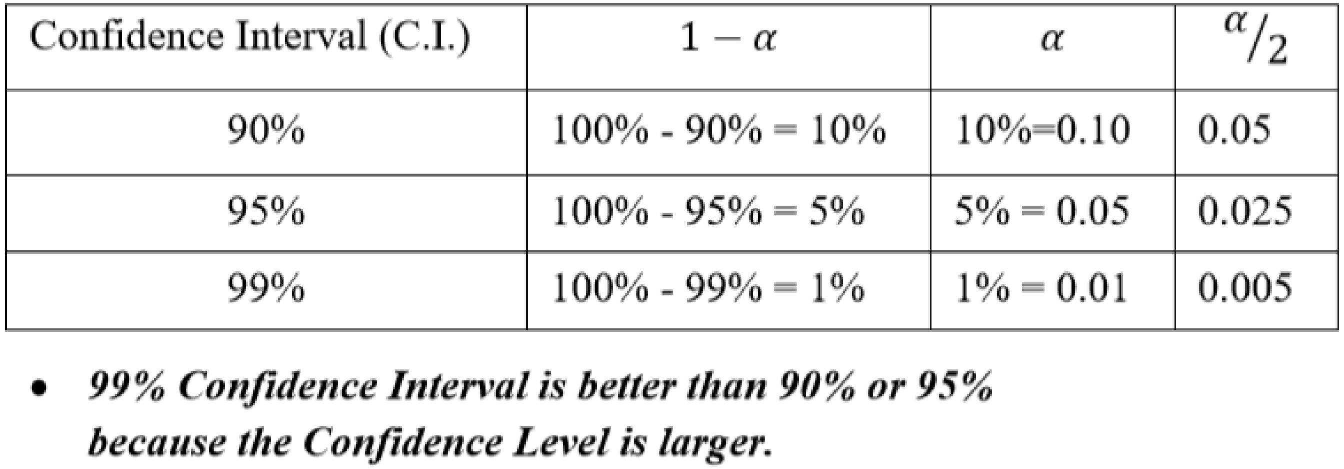
Confidence Intervals

A **confidence interval** refers to the probability that a population parameter will fall between two set values for a certain proportion of times.
Confidence intervals measure the degree of uncertainty or certainty in a sampling method.

- Symmetric about mean of data.
- CI can have different degrees of confidence 60%, 80%, 99% etc.
- Greater the sample size, smaller the Confidence Interval, i.e more accurate determination of population mean from the sample means.
- Different Sample means will have different CIs

Confidence Interval

- (1−alpha) gives the probability, that the true value falls into the calculated intervals.
- You also can give the confidence level in percent.
- Usual choices for the confidence level are 90%, 95%, or 99%.
- Most confidence intervals are of the form (point estimator) ± margin of error.



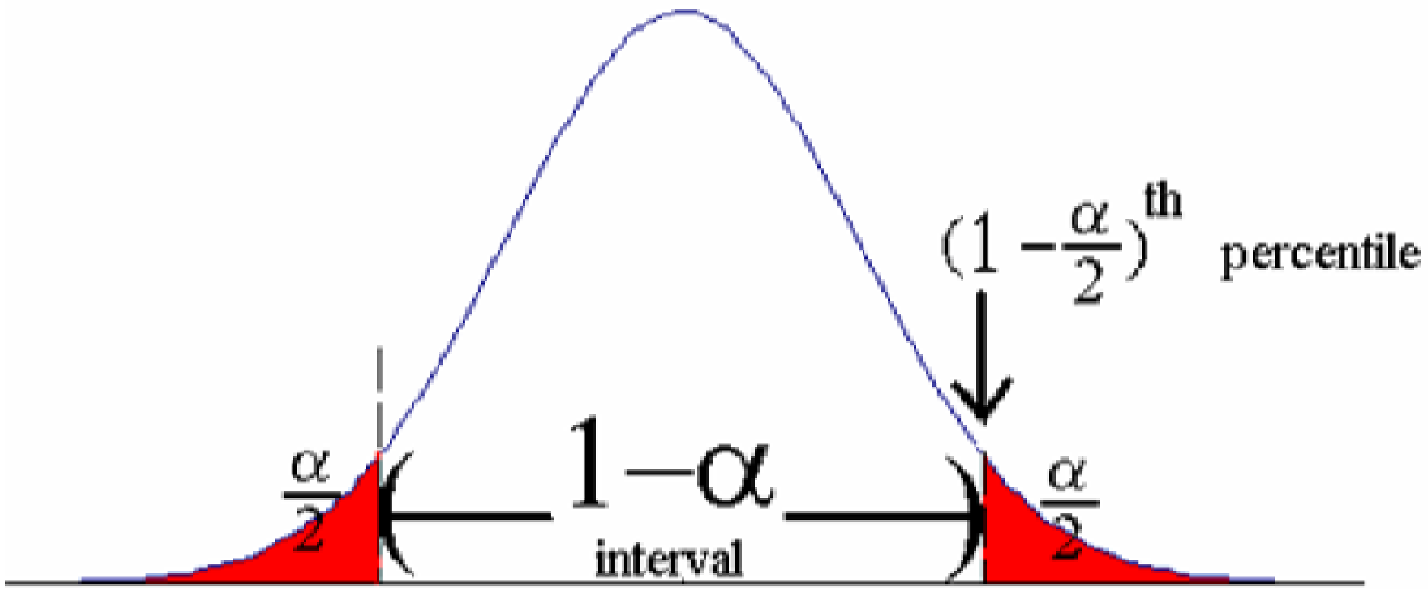
Confidence Interval (C.I.)	$1-\alpha$	α	$^{\alpha}/_{2}$
90%	100% - 90% = 10%	10%=0.10	0.05
95%	100% - 95% = 5%	5% = 0.05	0.025
99%	100% - 99% = 1%	1% = 0.01	0.005

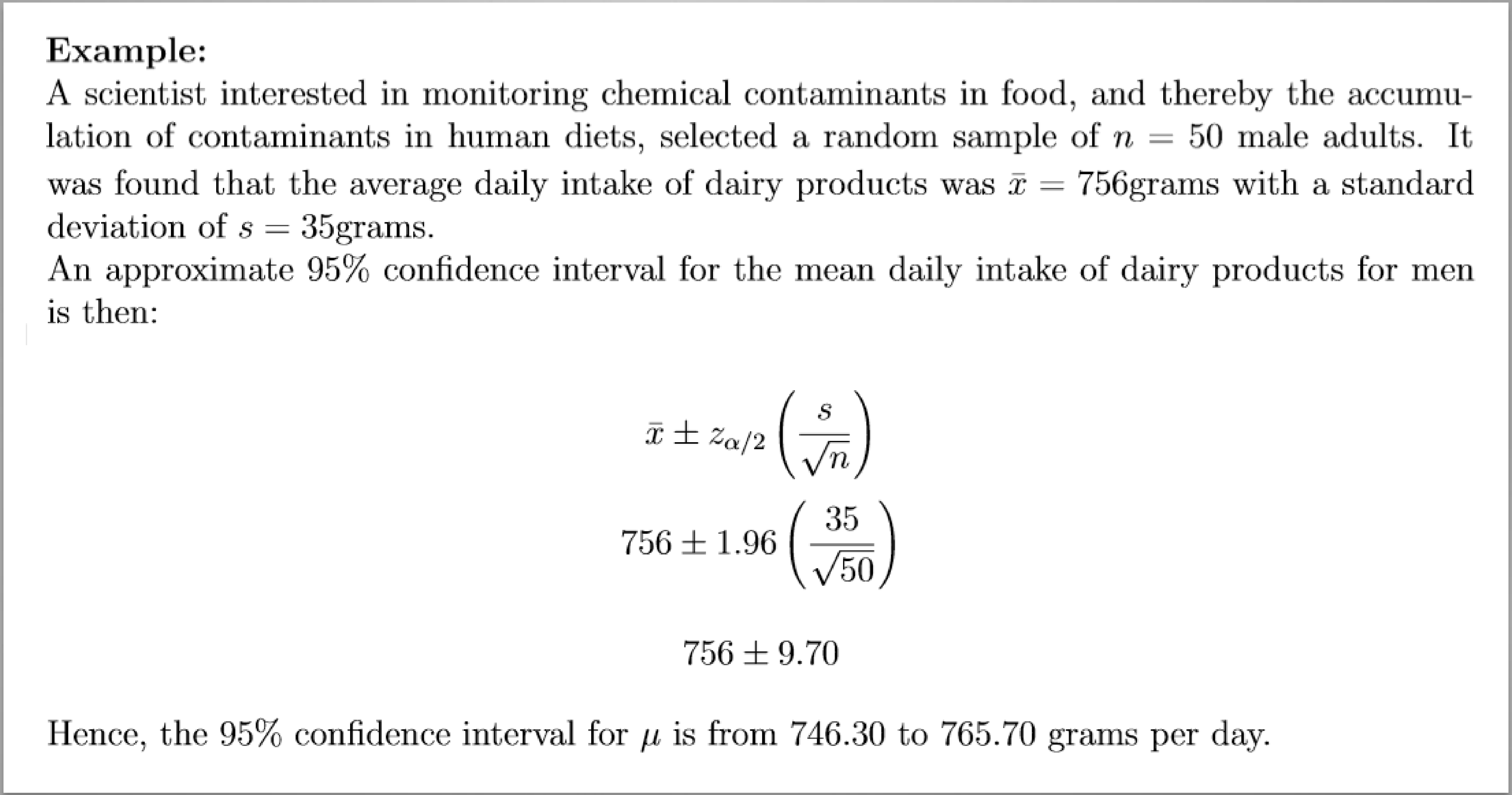
- ***99% Confidence Interval is better than 90% or 95% because the Confidence Level is larger.***

- **Range of Values**
  - Which may contain μ ; $\boldsymbol{a} < \mu < b$ (a, b)
  - It's called interval estimate of to find **confidence level** ; $\mathbf{1} - \boldsymbol{\alpha}$
  - Probability of success

$*$ α is e total area in both tailsof the standard normal distribution curve.

$*$ $\frac{\alpha}{2}$ is the area in each one of thetailes.





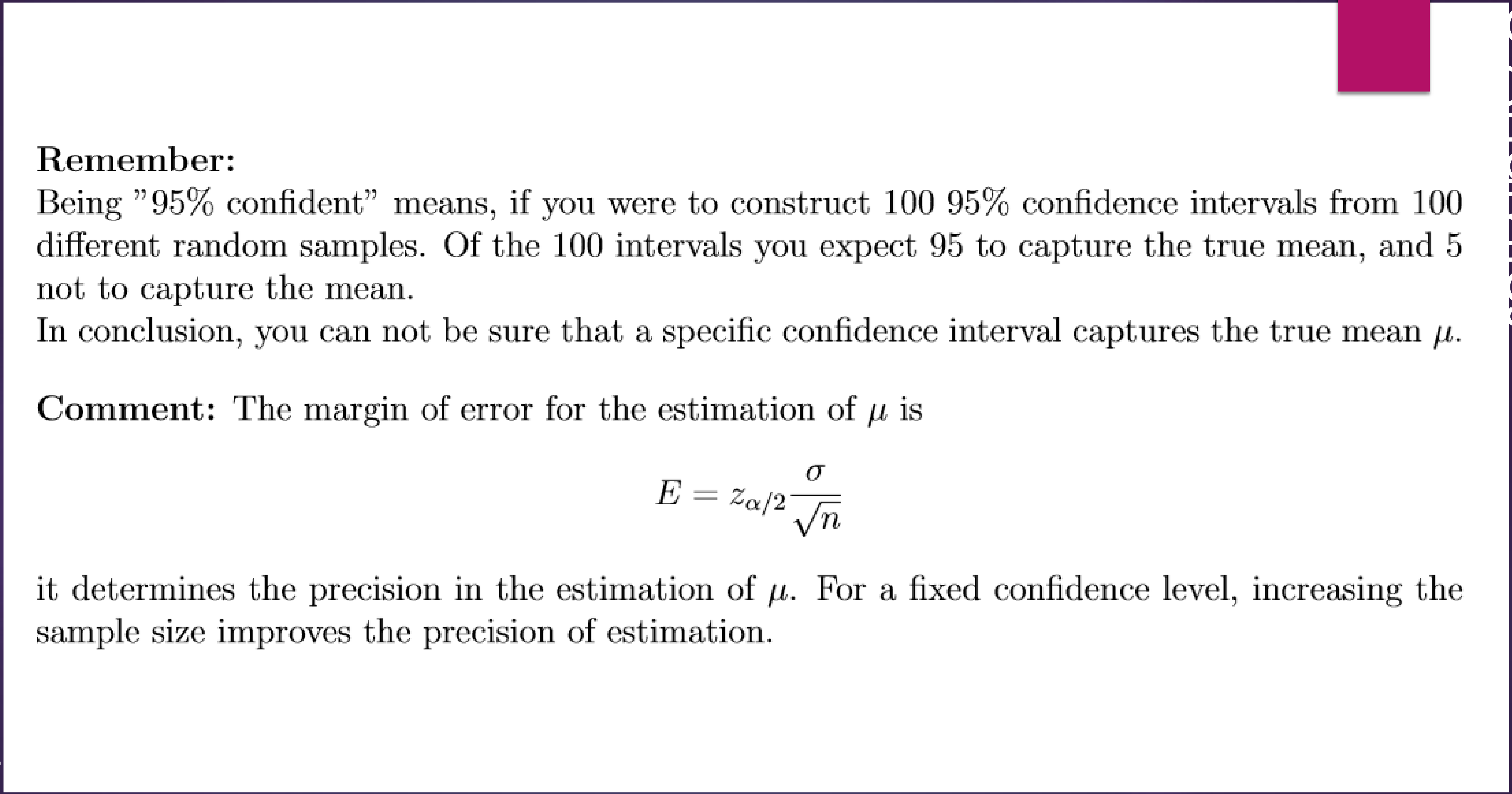
Example:
A scientist interested in monitoring chemical contaminants in food, and thereby the accumulation of contaminants in human diets, selected a random sample of $n = 50$ male adults. It was found that the average daily intake of dairy products was $\bar{x} = 756$grams with a standard deviation of $s = 35$grams.
An approximate 95% confidence interval for the mean daily intake of dairy products for men is then:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$756 \pm 1.96\left(\frac{35}{\sqrt{50}}\right)$$

$$756 \pm 9.70$$

Hence, the 95% confidence interval for μ is from 746.30 to 765.70 grams per day.



Remember:
Being "95% confident" means, if you were to construct 100 95% confidence intervals from 100 different random samples. Of the 100 intervals you expect 95 to capture the true mean, and 5 not to capture the mean.
In conclusion, you can not be sure that a specific confidence interval captures the true mean μ.

Comment: The margin of error for the estimation of μ is

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

it determines the precision in the estimation of μ. For a fixed confidence level, increasing the sample size improves the precision of estimation.

A random sample of 200 nurses is taken and each nurse asked his or her annual income in whole dollars. These 200 nurses have an average income of $ 35,000, SD of $ 5000.

The 90 % CI

will be : (34,415 to 35,585).

The interval width = 1170 $.

The 95 % CI

= 35,000±1.96 (5000/ √200) = 35,000 ±695. (34,305 to 35,695 $)

= this is the estimate of the average income of all nurses with a 95 % confidence.

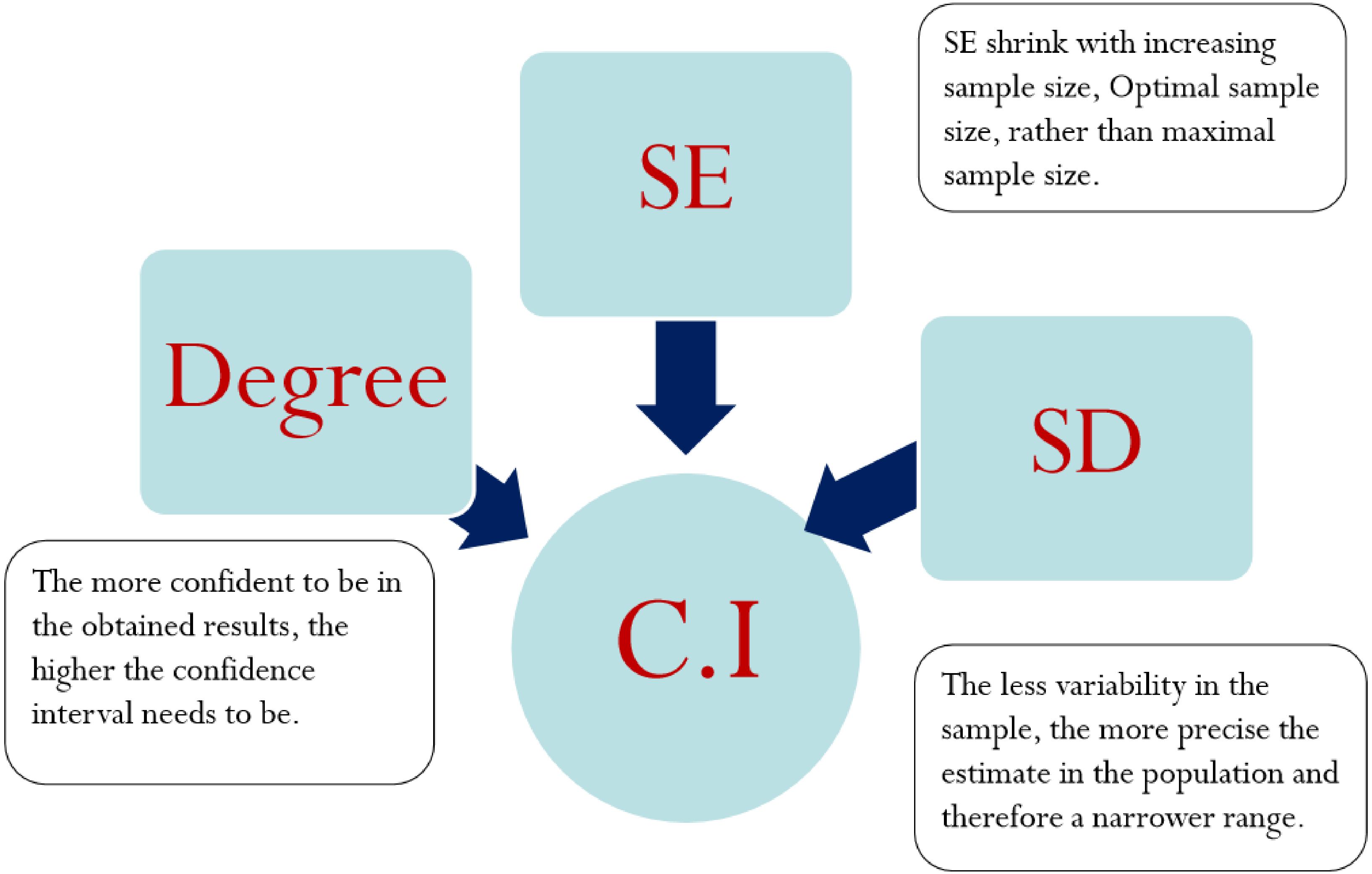
The interval width = 35,695-34,305 = $ 1390.

The 99 % CI

= 35,000 ±2.58 (5000 / √200) = 35,000 ±915 (34,085 to 35,915) =

the interval width is 1830.

Elements affecting the confidence intervals



SE shrink with increasing sample size, Optimal sample size, rather than maximal sample size.

The more confident to be in the obtained results, the higher the confidence interval needs to be.

The less variability in the sample, the more precise the estimate in the population and therefore a narrower range.

Hypothesis testing

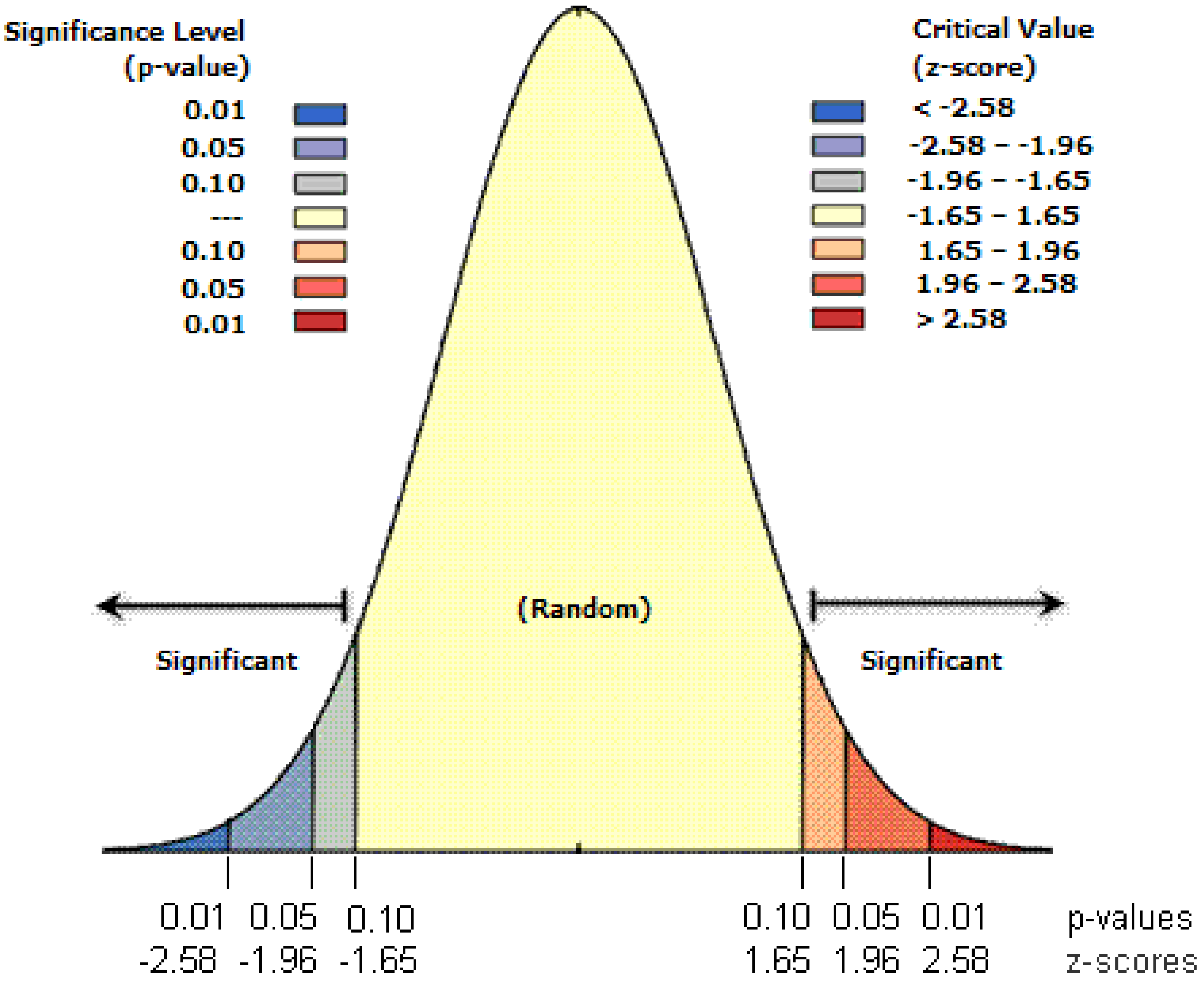
- Hypothesis testing is defined in two terms – Null Hypothesis and Alternate Hypothesis.
- **Null Hypothesis** being the sample statistic to be equal to the population statistic. For eg: The Null Hypothesis for the above example would be that the average marks after extra class are same as that before the classes.
- **Alternate Hypothesis** for this example would be that the marks after extra class are significantly different from that before the class.

Hypothesis testing

- **Hypothesis testing is an inferential procedure that uses sample data to evaluate the credibility of a hypothesis about a population.**
- A hypothesis is an **educated guess** about something in the world around you. It should be testable, either by experiment or observation. For example:
- A new medicine you think might work.
- A way of teaching you think might be better.
- A possible location of new species.
- A fairer way to administer standardized tests.
- It can really be *anything* as long as you can put it to the test.

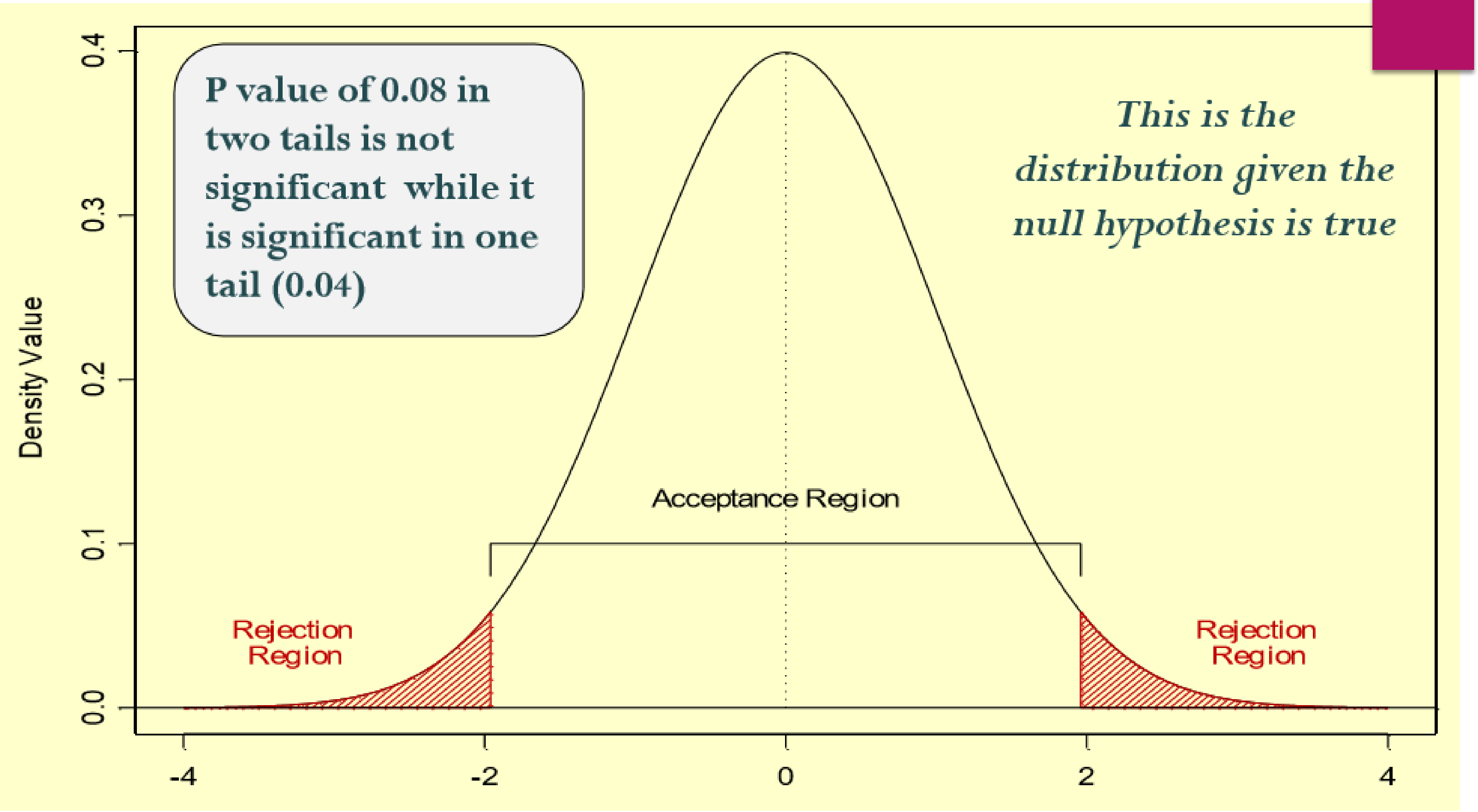
What is a Hypothesis Statement?

- If you are going to propose a hypothesis, it’s customary to write a statement. Your statement will look like this:
  “If I…(do this to an independent variable)….then (this will happen to the dependent variable).”
  For example:
- If I (decrease the amount of water given to herbs) then (the herbs will increase in size).
- If I (give patients counseling in addition to medication) then (their overall depression scale will decrease).
- If I (give exams at noon instead of 7) then (student test scores will improve).
- If I (look in this certain location) then (I am more likely to find new species).
- A good hypothesis statement should:
- Include an “if” and “then” statement (according to the University of California).
- Include both the independent and dependent variables.
- Be testable by experiment, survey or other scientifically sound technique.


Significance Level
(p-value)
0.01 🟦
0.05 🟦
0.10 ⬜
--- 🟨
0.10 🟧
0.05 🟥
0.01 🟥
Critical Value
(z-score)
< -2.58
-2.58 - -1.96
-1.96 - -1.65
-1.65 - 1.65
1.65 - 1.96
1.96 - 2.58
> 2.58
(Random)
Significant
Significant
0.01 0.05 0.10
-2.58 -1.96 -1.65
0.10 0.05 0.01
1.65 1.96 2.58
p-values
z-scores

General structure of hypothesis testing

- All hypothesis testing procedures follow the same general structure:
- Null Hyp. (H0): A specific statement about a population parameter (or parameters). We would like to prove this wrong if possible.
- Alt. Hyp. (H1): A general statement about a population parameter (or parameters) opposing H0.
- Data: Random sample(s) from the population(s)



P value of 0.08 in two tails is not significant while it is significant in one tail (0.04)

This is the distribution given the null hypothesis is true

Null Hypothesis

- A researcher thinks that if knee surgery patients go to physical therapy twice a week (instead of 3 times), their recovery period will be longer. Average recovery times for knee surgery patients is 8.2 weeks.
- **Null Hypothesis :** What happens if researcher is wrong…
  - if the researcher is wrong then the recovery time is less than or equal to 8.2 weeks
  - **$H_{0:}$ $\mu \leq 8.2$**
- **Alternate Hypothesis:** What happens if the treatment applied by researcher works.
  - Average recovery time is more than 8.2 weeks.
  - **H_1: $\mu > 8.2$**

Statement creation

- A study of the dental status of critically ill children in a Paediatric Intensive Care Unit examined 16 children with permanent teeth and found that the mean number of missing or filled teeth was 1.2 with a standard deviation of 1.9. Extensive analysis has established that the mean number of such teeth in the wider population of children is 1.4. Test whether the mean for critically ill children differs from this.
- Null Hyp. A specific statement about a population parameter (or parameters). We would like to prove this wrong if possible.

  H0 : μ = 1.4 missing/filled teeth
- Alt Hyp. A general statement about a population parameter (or parameters) opposing H0.

  H1 : μ 6= 1.4 missing/filled teeth

Probability theory:

- **Probability theory:** Allows us to calculate the exact probability that chance was the real reason for the relationship.
- Probability theory allows us to produce test statistics (using mathematical formulas)
- A test statistic is a number that is used to decide whether to accept or reject the null hypothesis.
- The most common statistical tests include:
- • Chi--square
- • T--test
- • ANOVA
- • Correlation
- • Linear Regression

Z - Test and T - Test

- P values is the value of probability obtained from Z value or t value from z/t tables.
- If the probability of obtaining a particular sample mean is less than the alpha value then it is called to be falling in **tail** or also known as **critical region.**
- If Z score or T score is greater than the critical level i.e. when p value is less than critical alpha value then it is said that the mean is **significantly different** than the population mean hence pointing that the population is entirely different.
- Which mean there was no impact from the treatment on the existing population thus the **null hypothesis** is rejected.
- We cannot accept the null hypothesis we can only reject it on the basis of evidence.

Example 1

- On a 25-point satisfaction scale, men and women differed by about 5 points (means were 18.75 and 23.5, respectively)
- They were not identical, but how likely is a 5 point difference to occur just by chance?
- An analysis was conducted, and the p-value for the gender comparison was p = .11
- Thus, there was about a 11% chance that this data (the 5 point difference) would occur by chance
- The p-value is greater than .05, so we would fail to reject the null (results are not significant)
- Thus, there is no evidence that males and females differ in their satisfaction

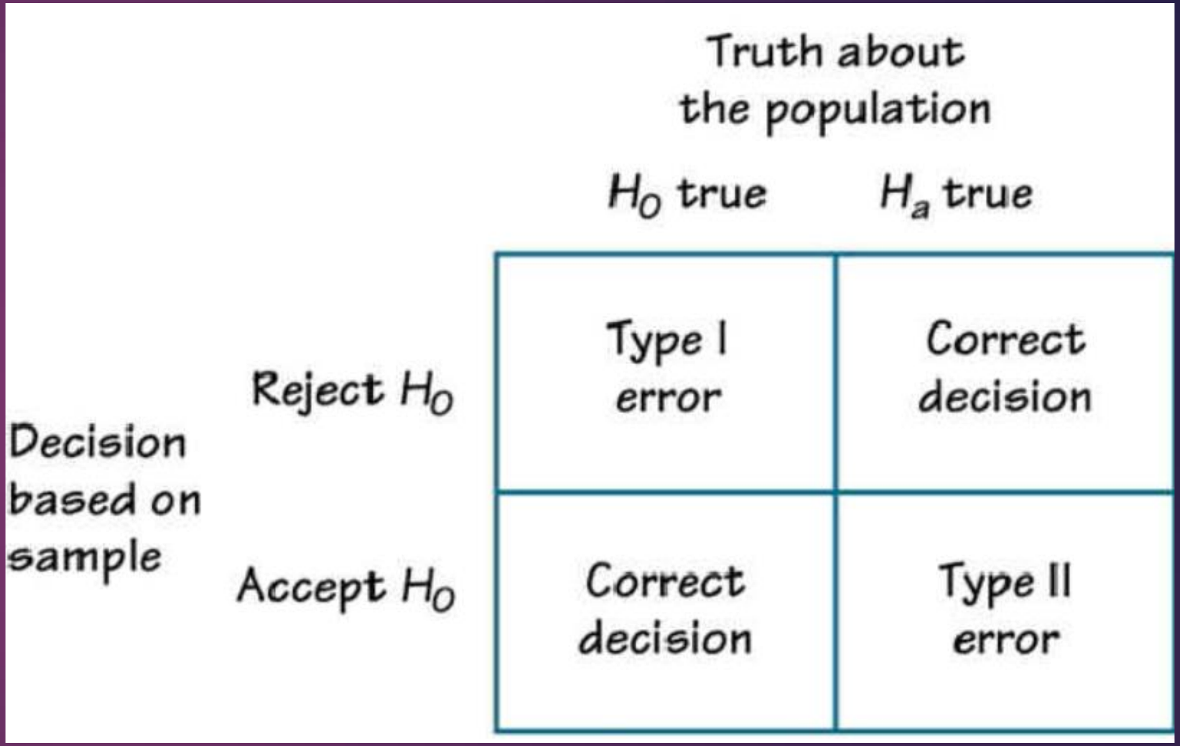
Example 2

- Suppose we were comparing how males and females differed with respect to how likely they would be to recommend an online course (measured on a 5 point scale)
- The null hypothesis states that there is no difference between men and women in their recommendation of an online course.
- (H0: X = Y)
- On a 5-point satisfaction scale, men and women differed by about 1 point (means were 4.3 and 3.1, respectively)
- They were not identical, but how likely is a 1 point difference to occur by chance?
- An analysis was conducted, and the p-value for the gender comparison was p = .03
- Thus, there was only a 3% probability that this data would occur by chance
- The p-value is less than .05, so we would reject the null (results are significant)
- Thus, there is evidence that males and females differ in their recommendations

There are two types of errors that are generally encountered while conducting Hypothesis Testing.

•**Type I error**: Look at the following scenario – A male human tested positive for being pregnant. Is it even possible? This surely looks like a case of False Positive. More formally, it is defined as the incorrect rejection of a True Null Hypothesis. The Null Hypothesis, in this case, would be – Male Human is not pregnant.

•**Type II error**: Look at another scenario where our Null Hypothesis is – A male human is pregnant and the test supports the Null Hypothesis. This looks like a case of False Negative. More formally it is defined as the acceptance of a false Null Hypothesis.



		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

The Meaning of Statistical Significance

- p-values tell how likely it was that our sample was drawn from a hypothetical population where "nothing was going on"
- Thus, the term "statistical significance" simply means that the obtained results are unlikely to represent a situation where there was no relationship between variables
- The difference is big enough to be unlikely to have happened simply due to chance
- The smaller the P-value, the greater your confidence in the statistical result.
- Alpha does not change whereas P values are dependent on the actual value of the statistic in question.

T - Test

- Using Z is not always possible as we would not have the population parameters always handy. Hence we use another version which is Student's T-Test.
- **Use T-test if:**
- Has a sample size below 30,
- Has an unknown population standard deviation
- **Assumptions** of the test :
- Data is independent.
- Data is collected randomly.
- The data is approximately normally distributed.
- Interpretation of t-ratio: If **the calculated t is less** than the tabulated values of t in table at 0.05 or 0.01 levels then the **null hypothesis is accepted**.
- If the **calculated t is greater** than the tabulated t at 0.05 or 0.01 levels then the **null hypothesis is rejected**

Types of T-Test

- A **One sample t-test tests** the mean of a **single** group against a known mean.
- An Independent **Samples t-test** compares the means for two independent groups for e.g. male and female.
- A **Paired sample t-test** compares means from the same group at **different** times (say, **one** year apart).

A study of the dental status of critically ill children in a Paediatric Intensive Care Unit examined 16 children with permanent teeth and found that the mean number of missing or filled teeth was 1.2 with a standard deviation of 1.9. Extensive analysis has established that the mean number of such teeth in the wider population of children is 1.4. Test whether the mean for critically ill children differs from this.

$$\mathrm{H}_0 : \mu = 1.4 \text{ missing/filled teeth} \qquad \mathrm{H}_1 : \mu \neq 1.4 \text{ missing/filled teeth}$$

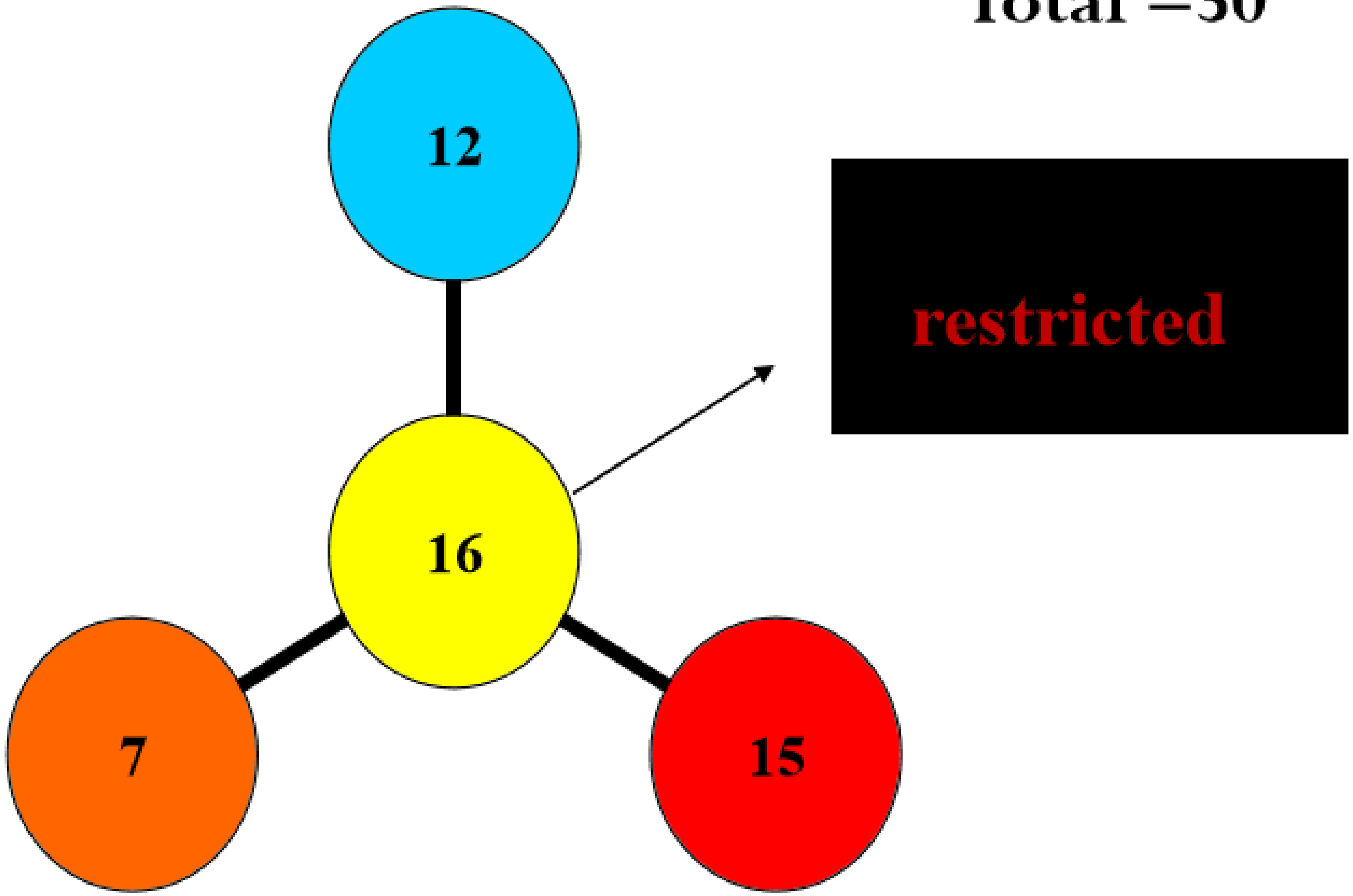
Data: $n =$ 16 critically ill children with permanent teeth.

Test Statistic: Suitable estimate of the population parameter (or combination of parameters) derived from these data.

$\overline{x} = 1.2$ teeth. In most Intro Stats units the only available testing procedure is the *one-sample t−test*. This uses a "standardized" version of $\overline{x}$:

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1.2 - 1.4}{\frac{1.9}{\sqrt{16}}} = -0.421$$

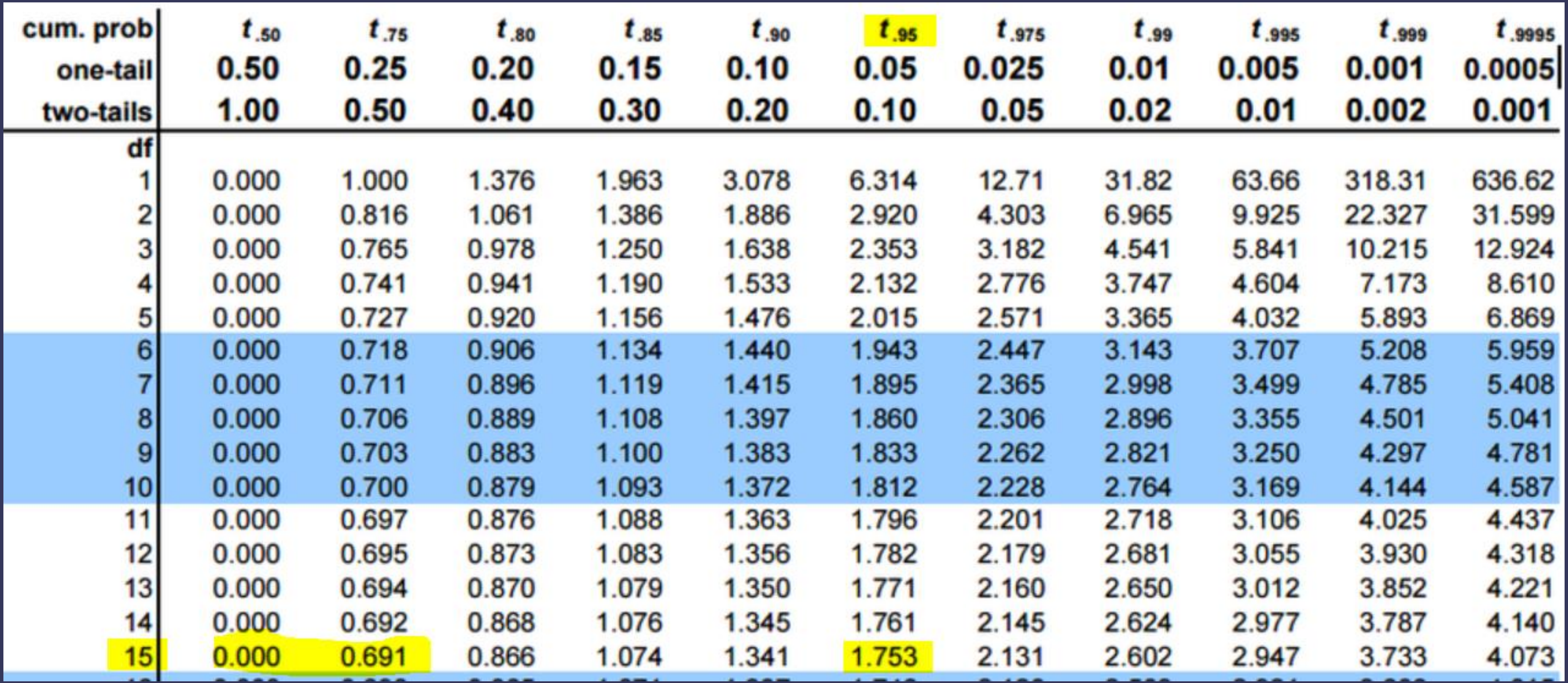
Total =50



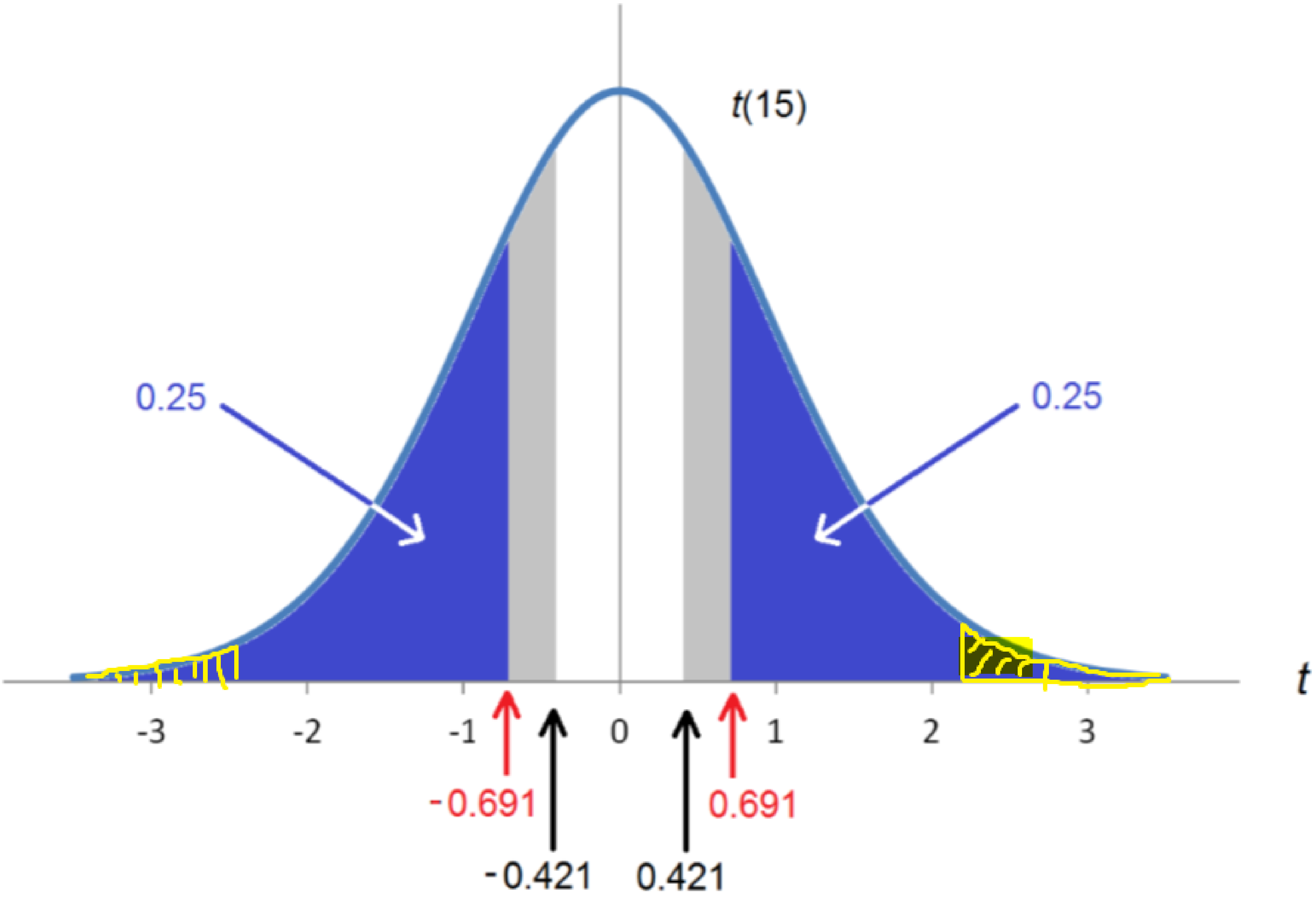
$df = \mathbf{n\text{-}1}$

Degrees of Freedom

- One sample t-test or paired t-test = N-1
- Independent t-test = N-2
- Chi-square test = (# rows -1) x (# columns – 1)
- ANOVA :
  - df between groups = (# levels or groups – 1)
  - df within groups = (# subjects -# of levels)
- Correlations = N-2



cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073


t(15)
0.25
0.25
-3
-2
-1
0
1
2
3
t
-0.691
0.691
-0.421
0.421

Calculator Links

http://www.z-table.com/

http://www.ttable.org/

https://www.graphpad.com/quickcalcs/contMenu/

Independent Sample Test

- Assumptions for the Independent Samples T Test
- Assumption of Independence: you need two independent, categorical groups that represent your independent variable. In the above example of test scores "males" or "females" would be your independent variable.
- Assumption of normality: the dependent variable should be approximately normally distributed. The dependent variable should also be measured on a continuous scale. In the above example on average test scores, the "test score" would be the dependent variable.
- For example, you might test two different groups of customer service associates on a business-related test or testing students from two universities on their English skills

Example:

Data set A: 1,2,2,3,3,4,4,5,5,6

Data set B: 1,2,4,5,5,5,6,6,7,9

Step 1: Sum the two groups:

A: 1 + 2 + 2 + 3 + 3 + 4 + 4 + 5 + 5 + 6 = 35

B: 1 + 2 + 4 + 5 + 5 + 5 + 6 + 6 + 7 + 9 = 50

Step 2: Square the sums from Step 1:

$35^2 = 1225$

$49^2 = 2500$

Step 3: Calculate the means for the two groups:

A: (1 + 2 + 2 + 3 + 3 + 4 + 4 + 5 + 5 + 6)/10 = 35/10 = 3.5

B: (1 + 2 + 4 + 5 + 5 + 5 + 6 + 6 + 7 + 9) = 50/10 = 5

Step 4: Square the individual scores and then add them up:

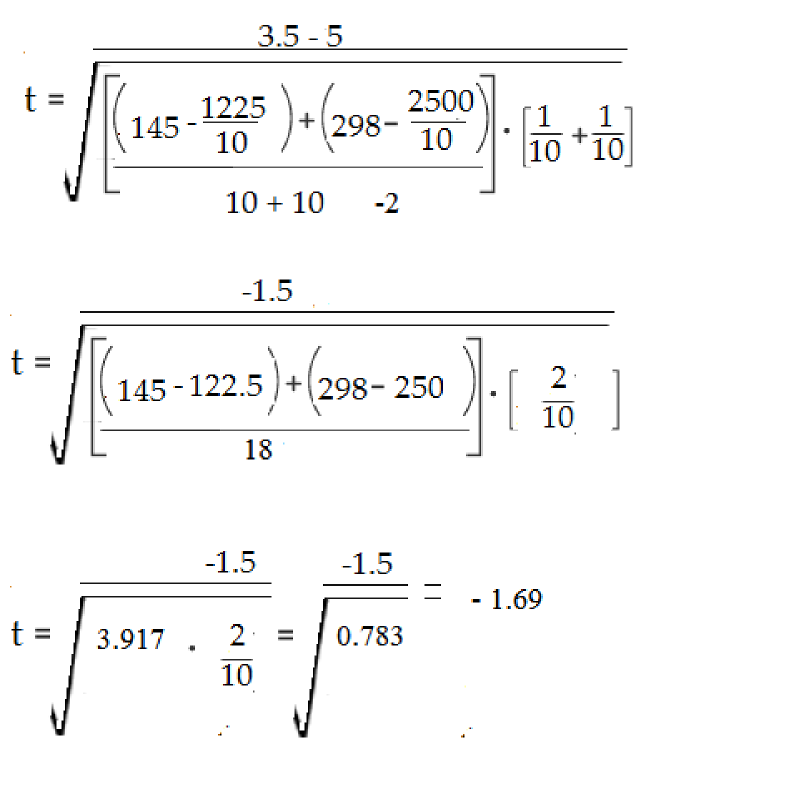
A: $1^1 + 2^2 + 2^2 + 3^3 + 3^3 + 4^4 + 4^4 + 5^5 + 5^5 + 6^6 = 145$

B: $1^2 + 2^2 + 4^4 + 5^5 + 5^5 + 5^5 + 6^6 + 6^6 + 7^7 + 9^9 = 298$

Step 5: Insert your numbers into the following formula and solve:

$$t = \frac{\mu_A - \mu_B}{\sqrt{\left[\frac{\left(\Sigma A^2 - \frac{(\Sigma A)^2}{n_A}\right) + \left(\Sigma B^2 - \frac{(\Sigma B)^2}{n_B}\right)}{n_A + n_B - 2}\right] \cdot \left[\frac{1}{n_A} + \frac{1}{n_B}\right]}}$$

$(\Sigma A)^2$: Sum of data set A, squared (Step 2).
$(\Sigma B)^2$: Sum of data set B, squared (Step 2).
μ_A: Mean of data set A (Step 3)
μ_B: Mean of data set B (Step 3)
ΣA^2: Sum of the squares of data set A (Step 4)
ΣB^2: Sum of the squares of data set B (Step 4)
n^A: Number of items in data set A
n^B: Number of items in data set B



$$t = \frac{3.5 - 5}{\sqrt{\left[\frac{\left(145 - \frac{1225}{10}\right) + \left(298 - \frac{2500}{10}\right)}{10 + 10 \quad -2}\right] \cdot \left[\frac{1}{10} + \frac{1}{10}\right]}}$$

$$t = \frac{-1.5}{\sqrt{\left[\frac{\left(145 - 122.5\right) + \left(298 - 250\right)}{18}\right] \cdot \left[\frac{2}{10}\right]}}$$

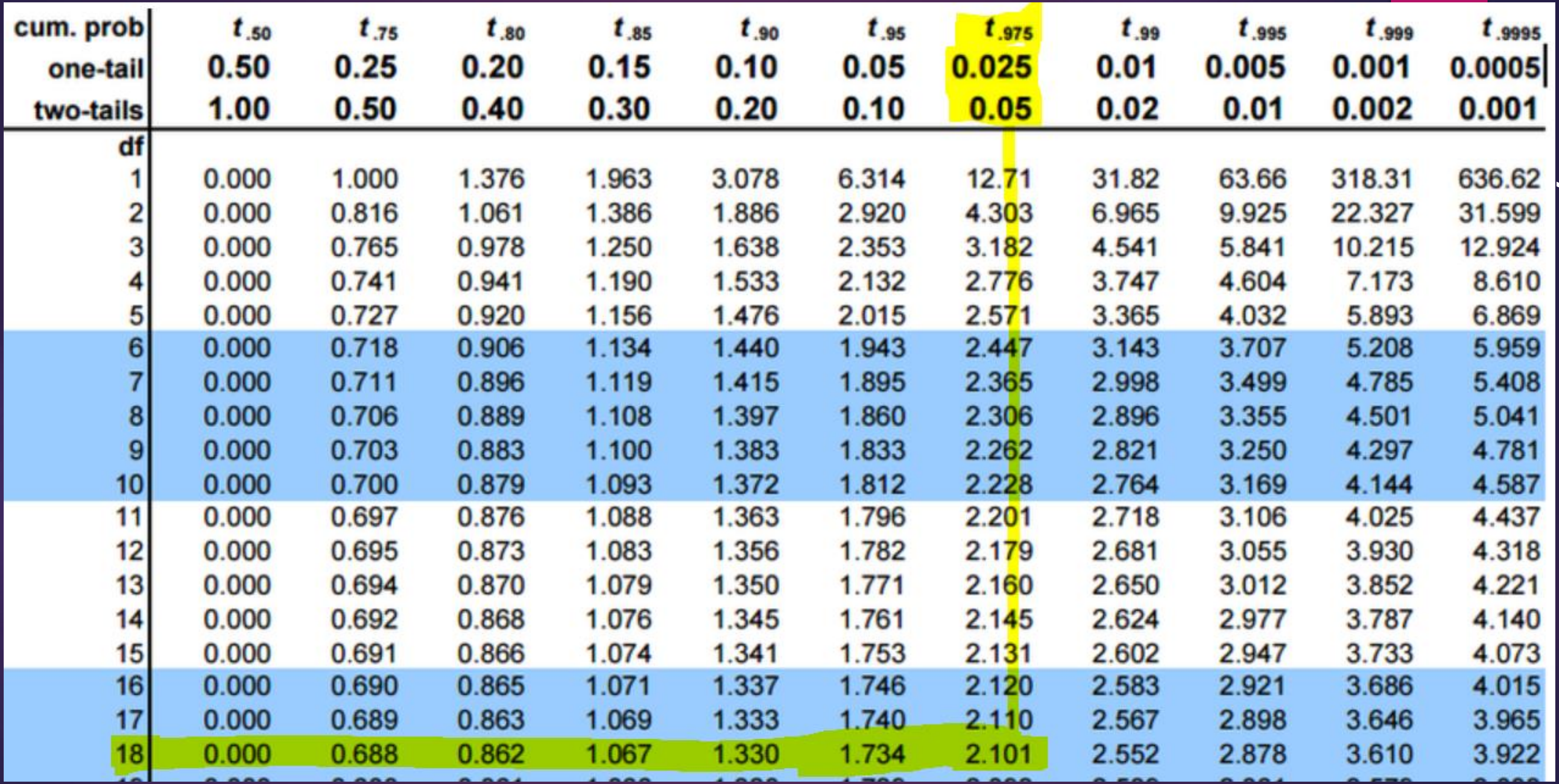
$$t = \frac{-1.5}{\sqrt{3.917 \cdot \frac{2}{10}}} = \frac{-1.5}{\sqrt{0.783}} = -1.69$$

Step 6: Find the Degrees of freedom (n_A-1 + n_B-1) = 18

Step 7: Look up your degrees of freedom (Step 6) in the t-table. If you don’t know what your alpha level is, use 5% (0.05).

18 degrees of freedom at an alpha level of 0.05 = 2.10.

Step 8: Compare your calculated value (Step 5) to your table value (Step 7). The calculated value of -1.79 is less than the cutoff of 2.10 from the table. Therefore p > .05. As the p-value is greater than the alpha level, we cannot conclude that there is a difference between means.



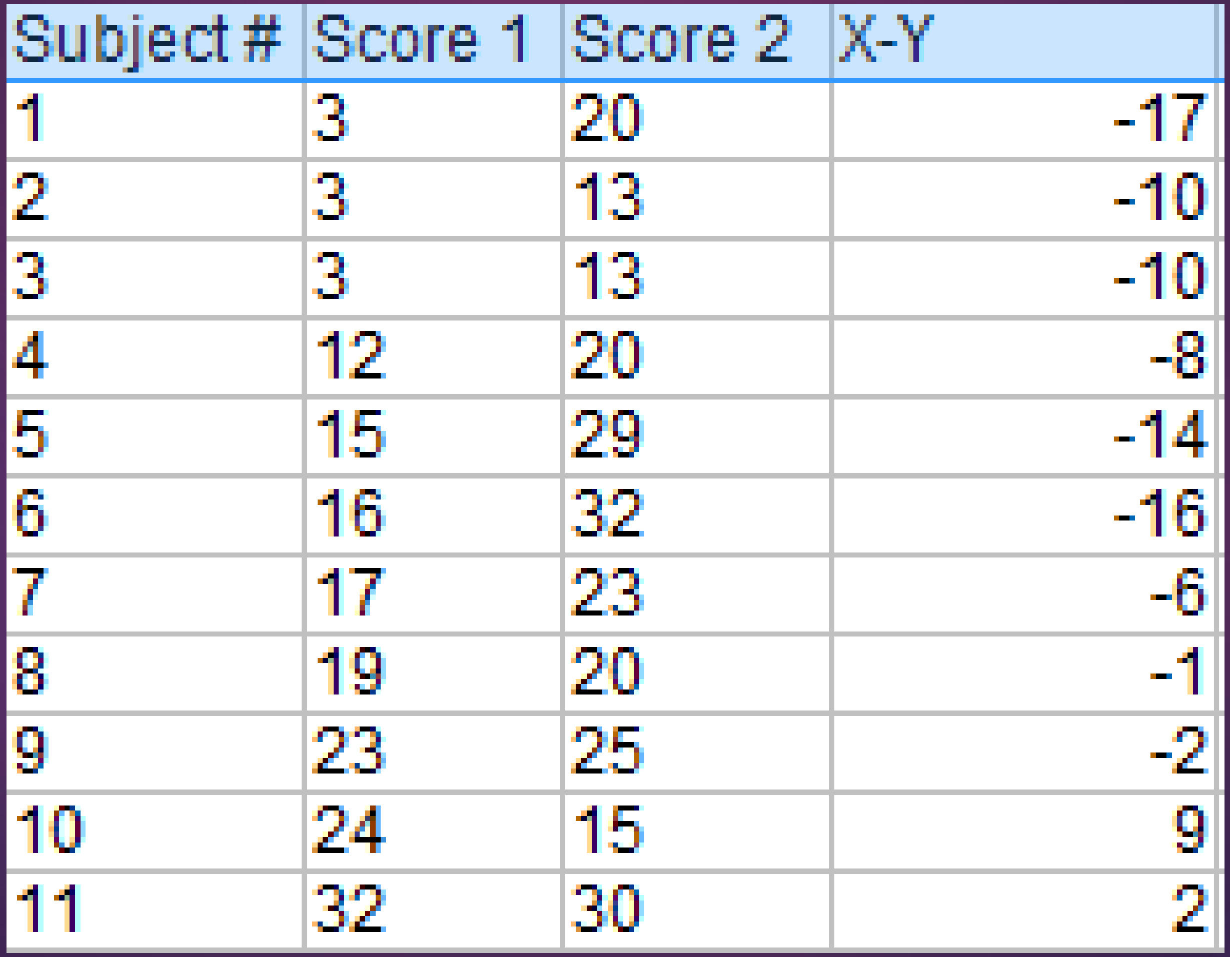
cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922

Paired T Test

- Choose the paired t-test if you have two measurements on the same item, person or thing. You should also choose this test if you have two items that are being measured with a unique condition.
- For example, you might be measuring car safety performance in Vehicle Research and Testing and subject the cars to a series of crash tests.
- The null hypothesis for the for the independent samples t-test is $\mu_1 = \mu_2$. In other words, it assumes the means are equal.
- With the paired t test, the null hypothesis is that the pairwise difference between the two tests is equal (H_0: $\mu_d = 0$).

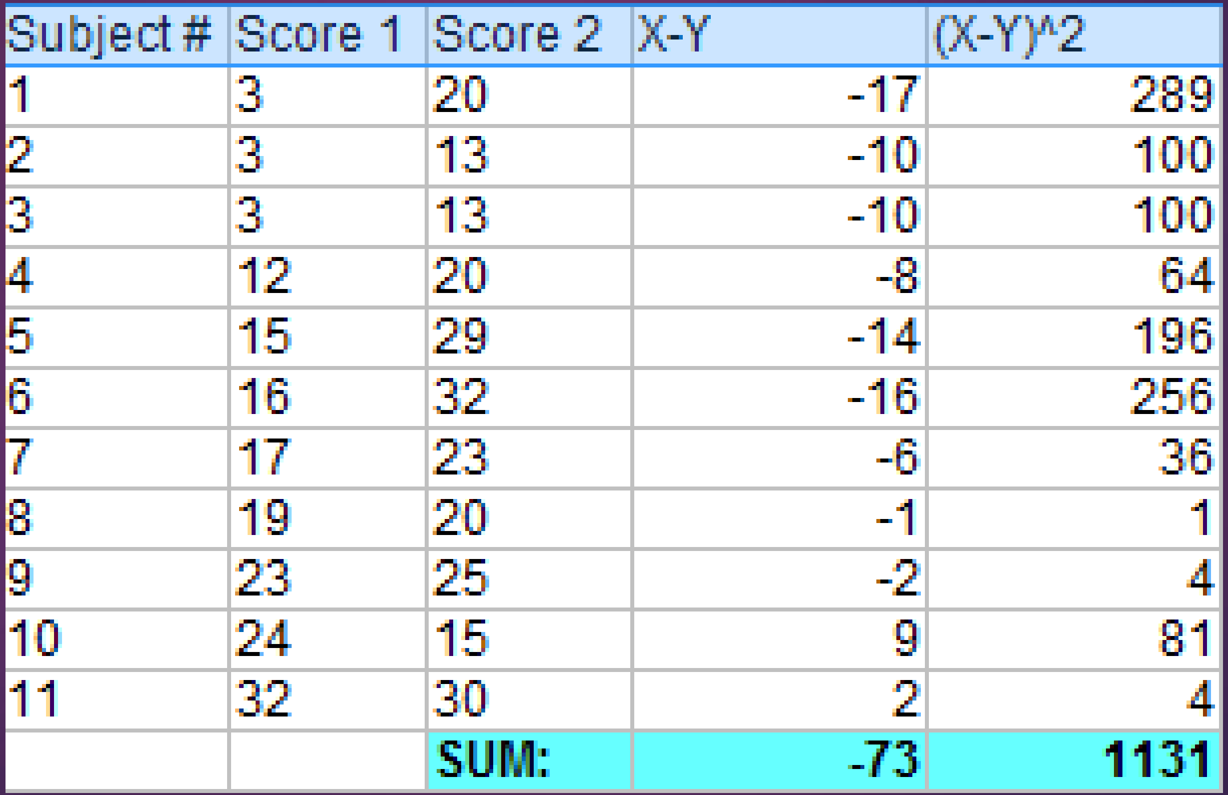
question: Calculate a paired t test by hand for the following data

Step 1: Subtract each Y score from each X score.



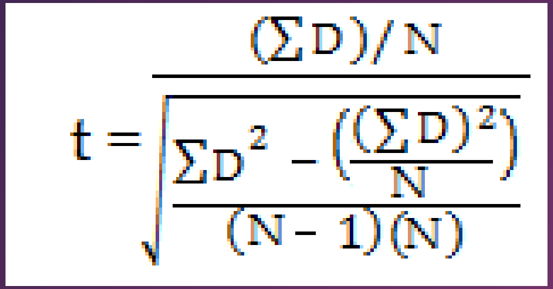
Subject #	Score 1	Score 2	X-Y
1	3	20	-17
2	3	13	-10
3	3	13	-10
4	12	20	-8
5	15	29	-14
6	16	32	-16
7	17	23	-6
8	19	20	-1
9	23	25	-2
10	24	15	9
11	32	30	2

Step 2: Add up all of the values from Step 1.
Step 3: Square the differences from Step 1.
Step 4: Add up all of the squared differences from Step 3.

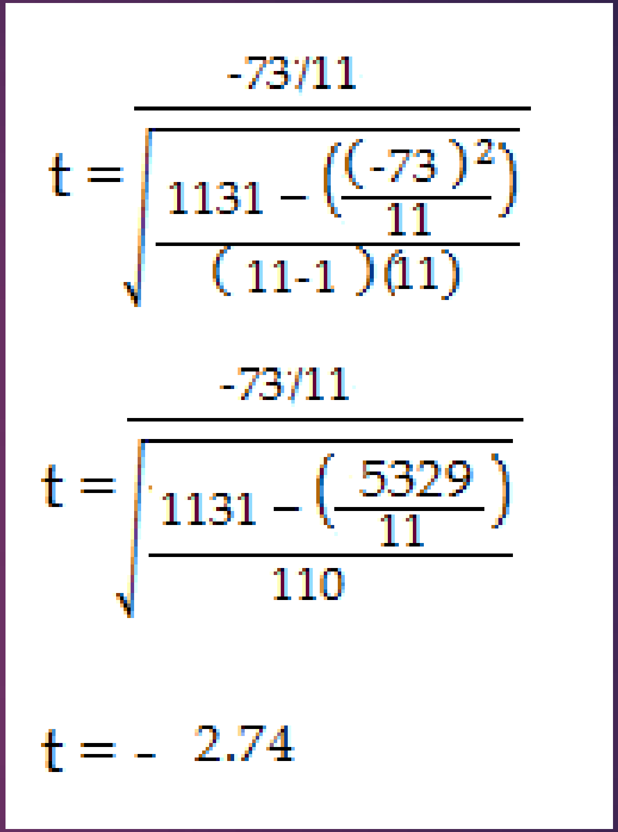


Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	-73	1131

Step 5: Use the following formula to calculate the t-score:



$$t = \frac{(\Sigma D)/N}{\sqrt{\dfrac{\Sigma D^2 - \left(\dfrac{(\Sigma D)^2}{N}\right)}{(N-1)(N)}}}$$



$$t = \frac{-73/11}{\sqrt{\dfrac{1131 - \left(\dfrac{(-73)^2}{11}\right)}{(11-1)(11)}}}$$

$$t = \frac{-73/11}{\sqrt{\dfrac{1131 - \left(\dfrac{5329}{11}\right)}{110}}}$$

$$t = -\ 2.74$$

ΣD: Sum of the differences (Sum of X-Y from Step 2)
ΣD^2: Sum of the squared differences (from Step 4)
$(\Sigma D)^2$: Sum of the differences (from Step 2), squared.

Step 6: Subtract 1 from the sample size to get the degrees of freedom. We have 11 items, so 11-1 = 10.
Step 7: Find the p-value in the t-table, using the degrees of freedom in Step 6. If you don't have a specified alpha level, use 0.05 (5%). For this sample problem, with df=10, the t-value is 2.228.

Step 8: Compare your t-table value from Step 7 (2.228) to your calculated t-value (-2.74). The calculated t-value is greater than the table value at an alpha level of .05. The p-value is less than the alpha level: p <.05. We can reject the null hypothesis that there is no difference between means.

Chi Square : Goodness of Fit

- **Non-parametric hypothesis tests** using the chi-square statistic:
  - the chi-square test for goodness of fit
  - the chi-square test for independence.

Non-parametric

- The term "non-parametric" refers to the fact that the chi-square tests do not require assumptions about population parameters nor do they test hypotheses about population parameters.
- Previous examples of hypothesis tests, such as the t tests, are **parametric tests** and they do include assumptions about parameters and hypotheses about parameters.
- The most obvious difference between the chi-square tests and the T-Test hypothesis test we have considered is the nature of the data.
- For chi-square, the data are frequencies rather than numerical scores.

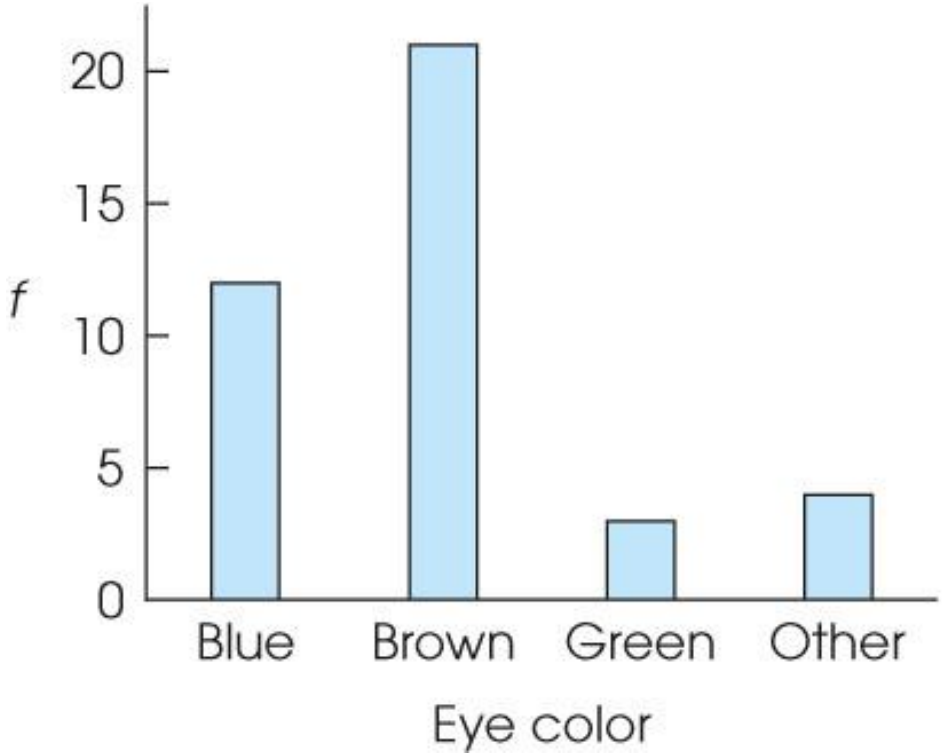
The Chi-Square Test for Goodness-of-Fit

- The chi-square test for goodness-of-fit uses frequency data from a sample to test hypotheses about the shape or proportions of a population.
- The chi-square test is used to test if a sample of data came from a population with a specific distribution.
- Every individual in the sample is classified into one category on the scale of measurement.
- The data, called observed frequencies, simply count how many individuals from the sample are in each category.

The Chi-Square Test for Goodness-of-Fit (cont.)

The null hypothesis specifies the proportion of the population that should be in each category.

The proportions from the null hypothesis are used to compute expected frequencies that describe how the sample would appear if it were in perfect agreement with the null hypothesis.



Eye color (X)	f
Blue	12
Brown	21
Green	3
Other	4

Blue	Brown	Green	Other
12	21	3	4

Example:

- Consider a standard package of milk chocolate M&Ms. There are six different colors: red, orange, yellow, green, blue and brown. Suppose that we are curious about the distribution of these colors and ask, do all six colors occur in equal proportion? This is the type of question that can be answered with a goodness of fit test.

Why Chi Square?

- We begin by noting the setting and why the goodness of fit test is appropriate. Our variable of color is categorical. There are six levels of this variable, corresponding to the six colors that are possible. We will assume that the M&Ms we count will be a simple random sample from the population of all M&Ms.

MG Analytics

Null and Alternate Hypothesis

- The null and alternative hypotheses for our goodness of fit test reflect the assumption that we are making about the population. Since we are testing whether the colors occur in equal proportions, our null hypothesis will be that all colors occur in the same proportion.
- More formally, if p_1 is the population proportion of red candies, p_2 is the population proportion of orange candies, and so on, then the null hypothesis is that $p_1 = p_2 = . . . = p_6 = 1/6$.
- The alternative hypothesis is that at least one of the population proportions is not equal to 1/6.

Actual and Expected Counts

- The actual counts are the number of candies for each of the six colors.
- The expected count refers to what we would expect if the null hypothesis were true.
- We will let n be the size of our sample. The expected number of red candies is p_1 n or n/6.
- In fact, for this example, the expected number of candies for each of the six colors is simply n times p_i, or n/6.

Chi-square Statistic for Goodness of Fit

$$X^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$$

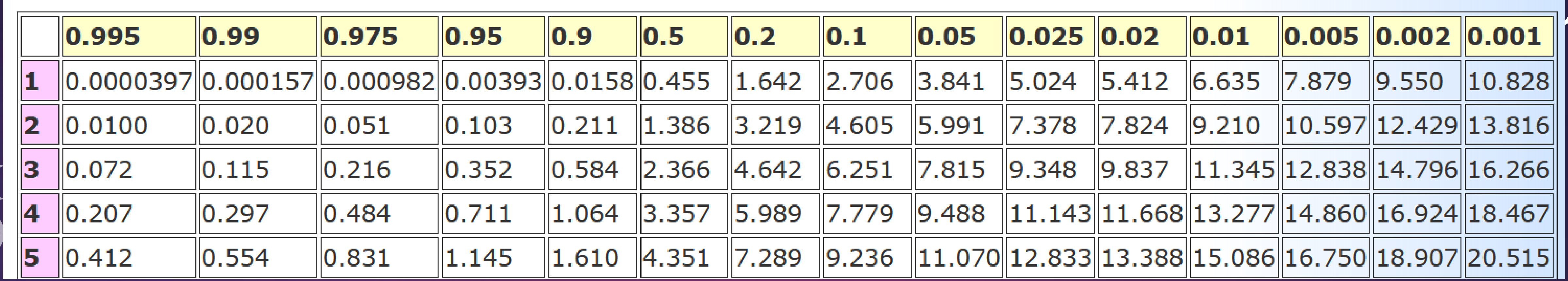
- If the null hypothesis were true, then the expected counts for each of these colors would be (1/6) x 600 = 100.
- We now use this in our calculation of the chi-square statistic.
- We calculate the contribution to our statistic from each of the colors. Each is of the form:

Color	Frequency	Chi Stat calc	Chi Value
Blue	212	$(212 – 100)^2/100$	125.44
Orange	147	$(147 – 100)^2/100$	22.09
Green	103	$(103 – 100)^2/100$	0.09
Red	50	$(50 – 100)^2/100$	25
Yellow	46	$(46 – 100)^2/100$	29.16
Brown	42	$(42 – 100)^2/100$	33.64
		Sum:	235.42

MG Analytics

Degrees of Freedom

The number of degrees of freedom for a goodness of fit test is simply one less than the number of levels of our variable. Since there were six colors, we have 6 – 1 = 5 degrees of freedom.



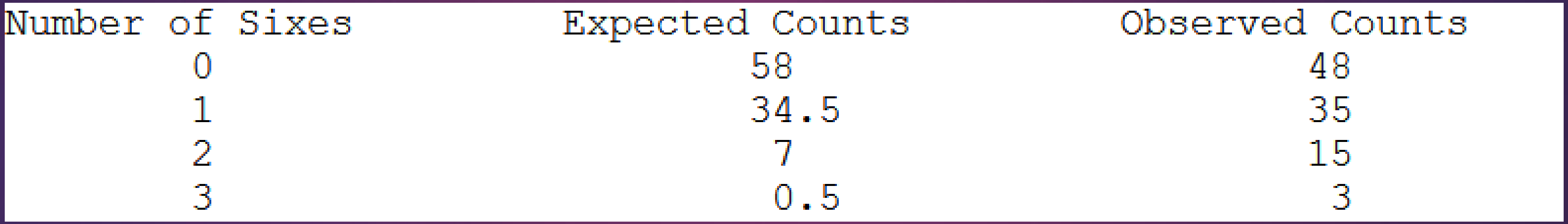
	0.995	0.99	0.975	0.95	0.9	0.5	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000397	0.000157	0.000982	0.00393	0.0158	0.455	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.020	0.051	0.103	0.211	1.386	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.072	0.115	0.216	0.352	0.584	2.366	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.297	0.484	0.711	1.064	3.357	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.554	0.831	1.145	1.610	4.351	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515

Since we have a very miniscule p-value, we reject the null hypothesis. We conclude that M&Ms are not evenly distributed among the six different colors.

Example 2:

A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 100 times, with the given observed counts.

The casino becomes suspicious of the gambler and wishes to determine whether the dice are fair. What do they conclude?



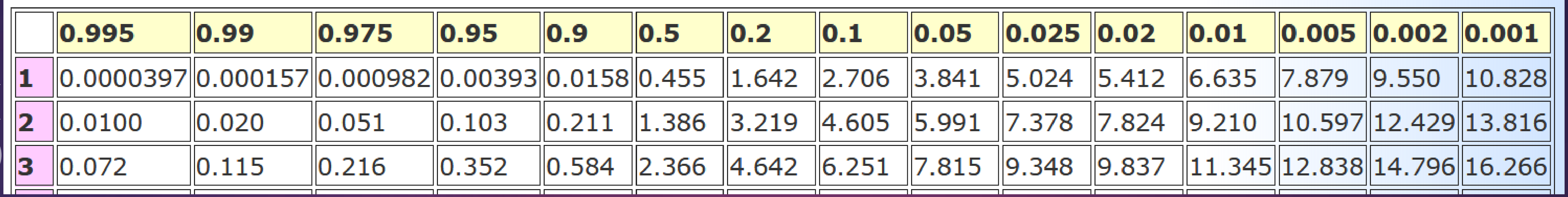
Number of Sixes	Expected Counts	Observed Counts
0	58	48
1	34.5	35
2	7	15
3	0.5	3

$(48\text{-}58)^2/58 + (35\text{-}34.5)^2/58 + (15\text{-}7)^2/7 + (3\text{-}0.5)^2/0.5$

$= 1.72 + 0.007 + 9.14 + 12.5 = 23.367.$

Conclusion:

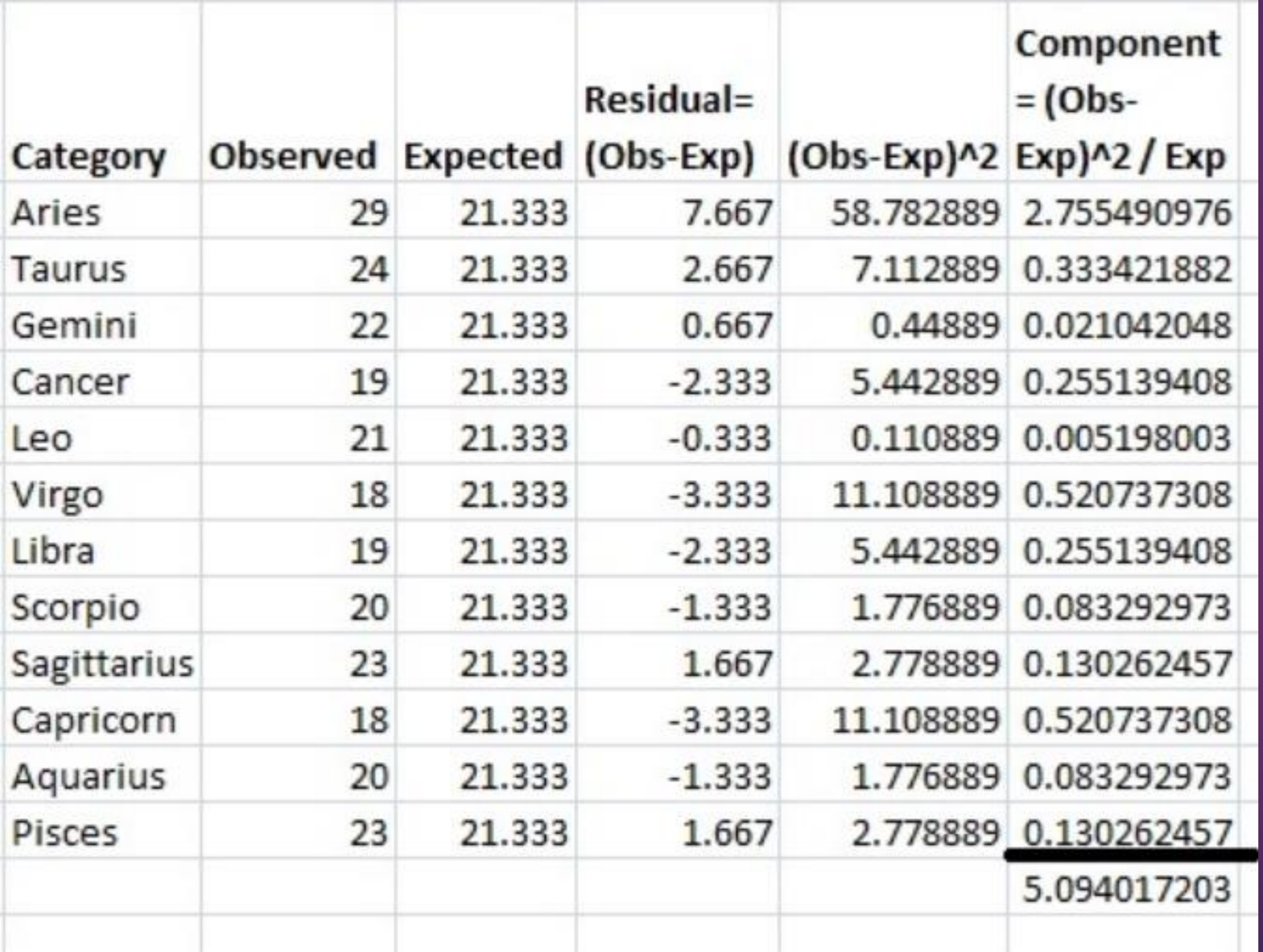
the chi-square test statistic was calculated to be 23.367. Since *k* = 4 in this case (the possibilities are 0, 1, 2, or 3 sixes), the test statistic is associated with the chi-square distribution with 3 degrees of freedom.



	0.995	0.99	0.975	0.95	0.9	0.5	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000397	0.000157	0.000982	0.00393	0.0158	0.455	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.020	0.051	0.103	0.211	1.386	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.072	0.115	0.216	0.352	0.584	2.366	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266

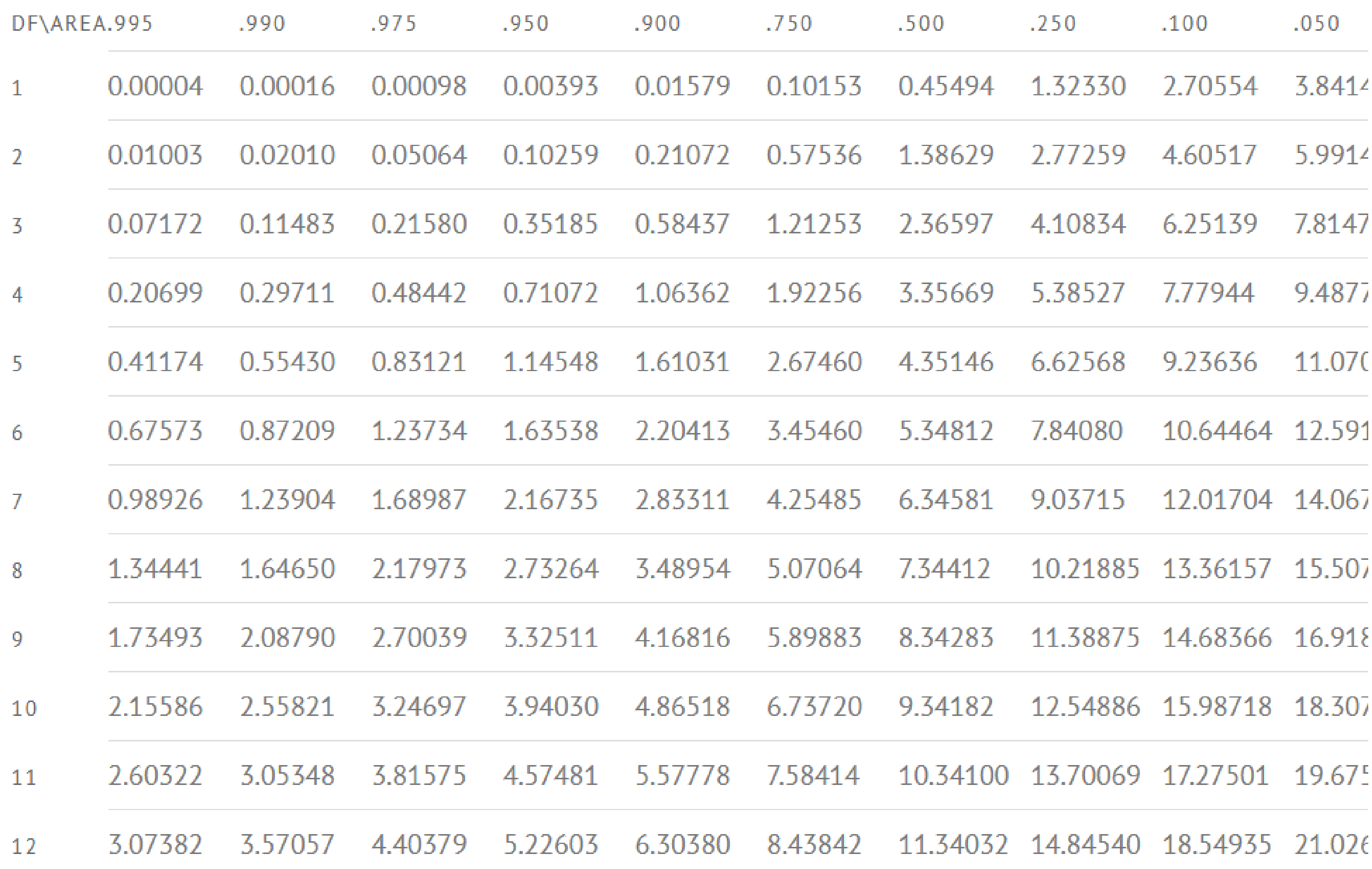
Since 23.367 is clearly greater than 7.815, we may reject the null hypothesis that the dice are fair at the 0.05 significance level.

256 visual artists were surveyed to find out their zodiac sign. Test the hypothesis that zodiac signs are evenly distributed across visual artists.



Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	2.755490976
Taurus	24	21.333	2.667	7.112889	0.333421882
Gemini	22	21.333	0.667	0.44889	0.021042048
Cancer	19	21.333	-2.333	5.442889	0.255139408
Leo	21	21.333	-0.333	0.110889	0.005198003
Virgo	18	21.333	-3.333	11.108889	0.520737308
Libra	19	21.333	-2.333	5.442889	0.255139408
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricorn	18	21.333	-3.333	11.108889	0.520737308
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457
					5.094017203

In general, small p-values (1% to 5%) would cause you to reject the null hypothesis. This very large p-value (92.65%) means that the null hypothesis should *not* be rejected.



DF\AREA	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050
1	0.00004	0.00016	0.00098	0.00393	0.01579	0.10153	0.45494	1.32330	2.70554	3.8414
2	0.01003	0.02010	0.05064	0.10259	0.21072	0.57536	1.38629	2.77259	4.60517	5.9914
3	0.07172	0.11483	0.21580	0.35185	0.58437	1.21253	2.36597	4.10834	6.25139	7.8147
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.92256	3.35669	5.38527	7.77944	9.4877
5	0.41174	0.55430	0.83121	1.14548	1.61031	2.67460	4.35146	6.62568	9.23636	11.070
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.45460	5.34812	7.84080	10.64464	12.591
7	0.98926	1.23904	1.68987	2.16735	2.83311	4.25485	6.34581	9.03715	12.01704	14.067
8	1.34441	1.64650	2.17973	2.73264	3.48954	5.07064	7.34412	10.21885	13.36157	15.507
9	1.73493	2.08790	2.70039	3.32511	4.16816	5.89883	8.34283	11.38875	14.68366	16.918
10	2.15586	2.55821	3.24697	3.94030	4.86518	6.73720	9.34182	12.54886	15.98718	18.307
11	2.60322	3.05348	3.81575	4.57481	5.57778	7.58414	10.34100	13.70069	17.27501	19.675
12	3.07382	3.57057	4.40379	5.22603	6.30380	8.43842	11.34032	14.84540	18.54935	21.026

MG Analytics

The Chi-Square Test for Independence

- The second chi-square test, the chi-square test for independence, can be used and interpreted in two different ways:
  - Testing hypotheses about the relationship between two variables in a population, or
  - Testing hypotheses about differences between proportions for two or more populations.

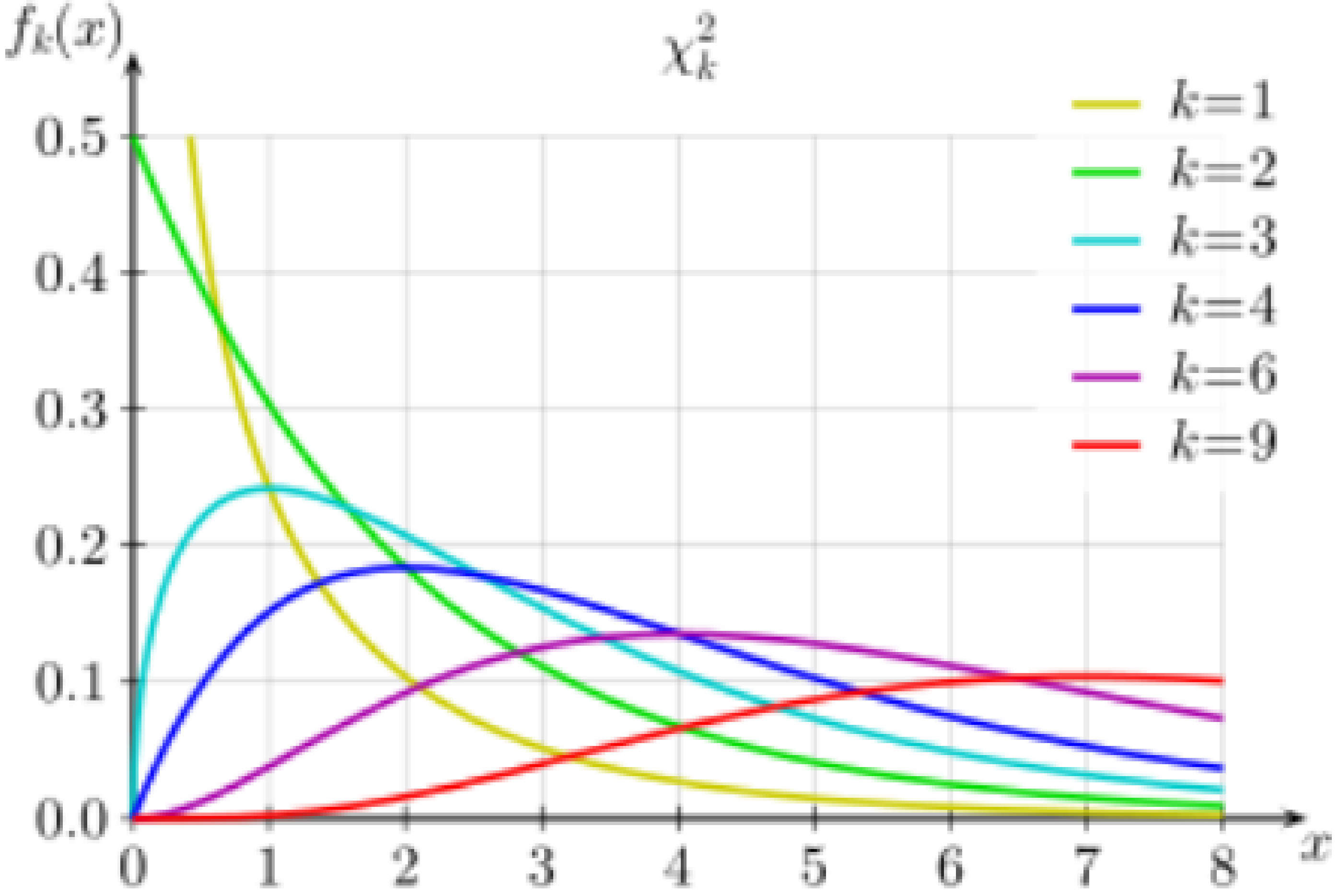
The Chi-Square Test for Independence (cont.)

- Although the two versions of the test for independence appear to be different, they are equivalent, and they are interchangeable.
- The first version of the test emphasizes the relationship between chi-square and a correlation, because both procedures examine the relationship between two variables.
- It determines whether there is an association between categorical variables, also known as: **Chi-Square Test** of Association.

A **very small chi square test statistic** means that your observed data fits your expected data extremely well. In other words, there is a relationship.
A **very large chi square test statistic** means that the data does not fit very well. In other words, there isn't a relationship.
The subscript "c" are the degrees of freedom. "O" is your observed value and E is your expected value.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The Chi-Square Distribution



Example

- A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency table.
- Is there a gender gap? Do the men's voting preferences differ significantly from the women's preferences? Use a 0.05 level of significance.

	Voting Preferences			Row total
	Rep	Dem	Ind	
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

Solution:

H_o: Gender and voting preferences are independent.
H_a: Gender and voting preferences are not independent.

- **Degrees of freedom**: DF = (r - 1) * (c - 1)

Where r is the number of levels for one categorical variable, and c is the number of levels for the other categorical variable.

- **Expected frequencies.** The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable.

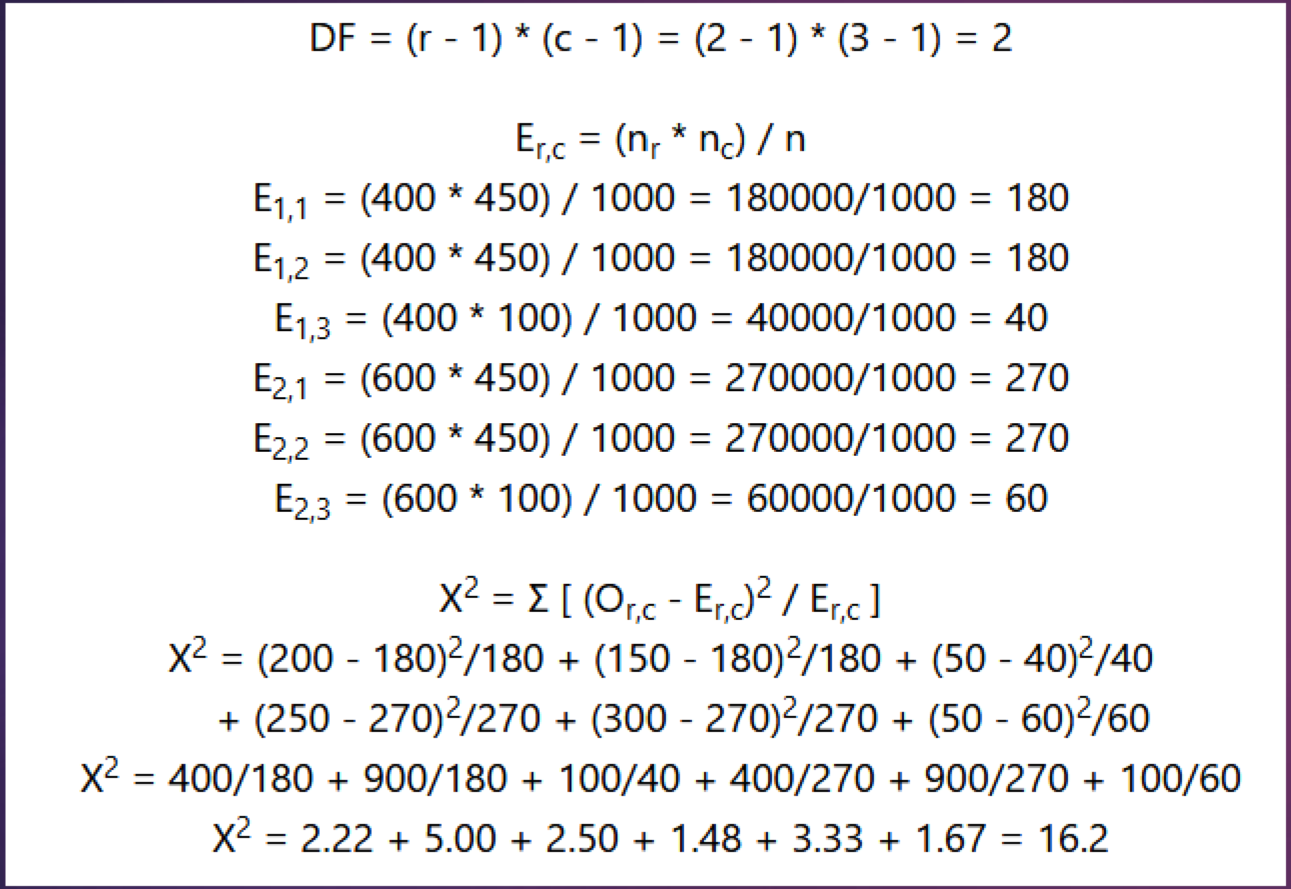
$$E_{r,c} = (n_r * n_c) / n$$

level *r* of Variable A, level c of Variable B,
n_r is the total number of sample observations at level r of Variable A,
n_c is the total number of sample observations at level c of Variable B,
n is the total sample size.

$$X^2 = \Sigma \left[(O_{r,c} - E_{r,c})^2 / E_{r,c} \right]$$

$O_{r,c}$ is the observed frequency count at level *r* of Variable A and level *c* of Variable B, and
$E_{r,c}$ is the expected frequency count at level *r* of Variable A and level *c* of Variable B.

Calculations



$$DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2$$

$$E_{r,c} = (n_r * n_c) / n$$
$$E_{1,1} = (400 * 450) / 1000 = 180000/1000 = 180$$
$$E_{1,2} = (400 * 450) / 1000 = 180000/1000 = 180$$
$$E_{1,3} = (400 * 100) / 1000 = 40000/1000 = 40$$
$$E_{2,1} = (600 * 450) / 1000 = 270000/1000 = 270$$
$$E_{2,2} = (600 * 450) / 1000 = 270000/1000 = 270$$
$$E_{2,3} = (600 * 100) / 1000 = 60000/1000 = 60$$

$$X^2 = \Sigma \left[(O_{r,c} - E_{r,c})^2 / E_{r,c} \right]$$
$$X^2 = (200 - 180)^2/180 + (150 - 180)^2/180 + (50 - 40)^2/40$$
$$+ (250 - 270)^2/270 + (300 - 270)^2/270 + (50 - 60)^2/60$$
$$X^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60$$
$$X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$$

	Voting Preferences			Row total
	Rep	Dem	Ind	
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

P-value (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.

MG Analytics

MG Analytics

ANOVA

- Analysis of variance, or ANOVA, is a technique from statistical interference that allows us to deal with several populations.

To conduct ANOVA with the M&M, we would test the null hypothesis H_0:$\mu_1 = \mu_2 = \mu_3 = \mu_4$. This states that there is no difference between the mean weights of the red, blue and green M&Ms. The alternative hypothesis is that there is some difference between the mean weights of the red, blue, green and orange M&Ms. This hypothesis is really a combination of several statements H_a:

The mean weight of the population of red candies is not equal to the mean weight of the population of blue candies, OR
The mean weight of the population of blue candies is not equal to the mean weight of the population of green candies, OR
The mean weight of the population of green candies is not equal to the mean weight of the population of orange candies, OR
The mean weight of the population of green candies is not equal to the mean weight of the population of red candies, OR
The mean weight of the population of blue candies is not equal to the mean weight of the population of orange candies, OR
The mean weight of the population of blue candies is not equal to the mean weight of the population of red candies.

p-value, we would utilize a probability distribution known as the F-distribution. Calculations involving the ANOVA F test can be done by hand, but are typically computed with statistical software.

Steps:

- Calculate the sample means for each of our samples as well as the mean for all of the sample data.
- Calculate the sum of squares of error. Here within each sample, we square the deviation of each data value from the sample mean. The sum of all of the squared deviations is the sum of squares of error, abbreviated SSE.
- Calculate the sum of squares of treatment. We square the deviation of each sample mean from the overall mean. The sum of all of these squared deviations is multiplied by one less than the number of samples we have. This number is the sum of squares of treatment, abbreviated SST.
- Calculate the degrees of freedom. The overall number of degrees of freedom is one less than the total number of data points in our sample, or n - 1. The number of degrees of freedom of treatment is one less than the number of samples used, or m - 1. The number of degrees of freedom of error is the total number of data points, minus the number of samples, or n - m.
- Calculate the mean square of error. This is denoted MSE = SSE/(n - m).
- Calculate the mean square of treatment. This is denoted MST = SST/m - `1.
- Calculate the F statistic. This is the ratio of the two mean squares that we calculated.
- F = MST/MSE.

Data and Sample Means

- Suppose we have four independent populations that satisfy the conditions for single factor ANOVA.
- We wish to test the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.
- For this example we will use a sample of size three from each of the populations being studied.

- The data from our samples is:
- Sample from population #1: 12, 9, 12. This has a sample mean of 11.
- Sample from population #2: 7, 10, 13. This has a sample mean of 10.
- Sample from population #3: 5, 8, 11. This has a sample mean of 8.
- Sample from population #4: 5, 8, 8. This has a sample mean of 7.

- The mean of all of the data is 9.

Sum of Squares of Error

- We now calculate the sum of the squared deviations from each sample mean. This is called the sum of squares of error.
- For the sample from population #1: $(12 – 11)^2 + (9– 11)^2 +(12 – 11)^2 = 6$
- For the sample from population #2: $(7 – 10)^2 + (10– 10)^2 +(13 – 10)^2 = 18$
- For the sample from population #3: $(5 – 8)^2 + (8 – 8)^2 +(11 – 8)^2 = 18$
- For the sample from population #4: $(5 – 7)^2 + (8 – 7)^2 +(8 – 7)^2 = 6$.
- We then add all of these sum of squared deviations and obtain 6 + 18 + 18 + 6 = 48.

Sum of Squares of Treatment

- Now we calculate the sum of squares of treatment. Here we look at the squared deviations of each sample mean from the overall mean, and multiply this number by one less than the number of populations:
- $3[(11 – 9)^2 + (10 – 9)^2 +(8 – 9)^2 + (7 – 9)^2] = 3[4 + 1 + 1 + 4] = 30.$

Population #	Samples	Mean	Mean-Sample Mean	SSE	SS of Treatment= (Mean- Sample Mean) * Treatment DF
1	12, 9, 12	11	11-9=4	$(12 – 11)^2 + (9– 11)^2 +(12 – 11)^2 =$ 6	
2	7, 10, 13	10	10-9=1	$(7 – 10)^2 + (10– 10)^2 +(13 – 10)^2 =$ 18	
3	5, 8, 11	8	8-9=1	$(5 – 8)^2 + (8 – 8)^2 +(11 – 8)^2 =$ 18	
4	5, 8, 8	7	7-9=4	$(5 – 7)^2 + (8 – 7)^2 + (8 – 7)^2 =$ 6	
		Mean: 9	**Sum: 10**	**SSE:** 6 + 18 + 18 + 6 = **48**	**SST:** **10*3 =30**

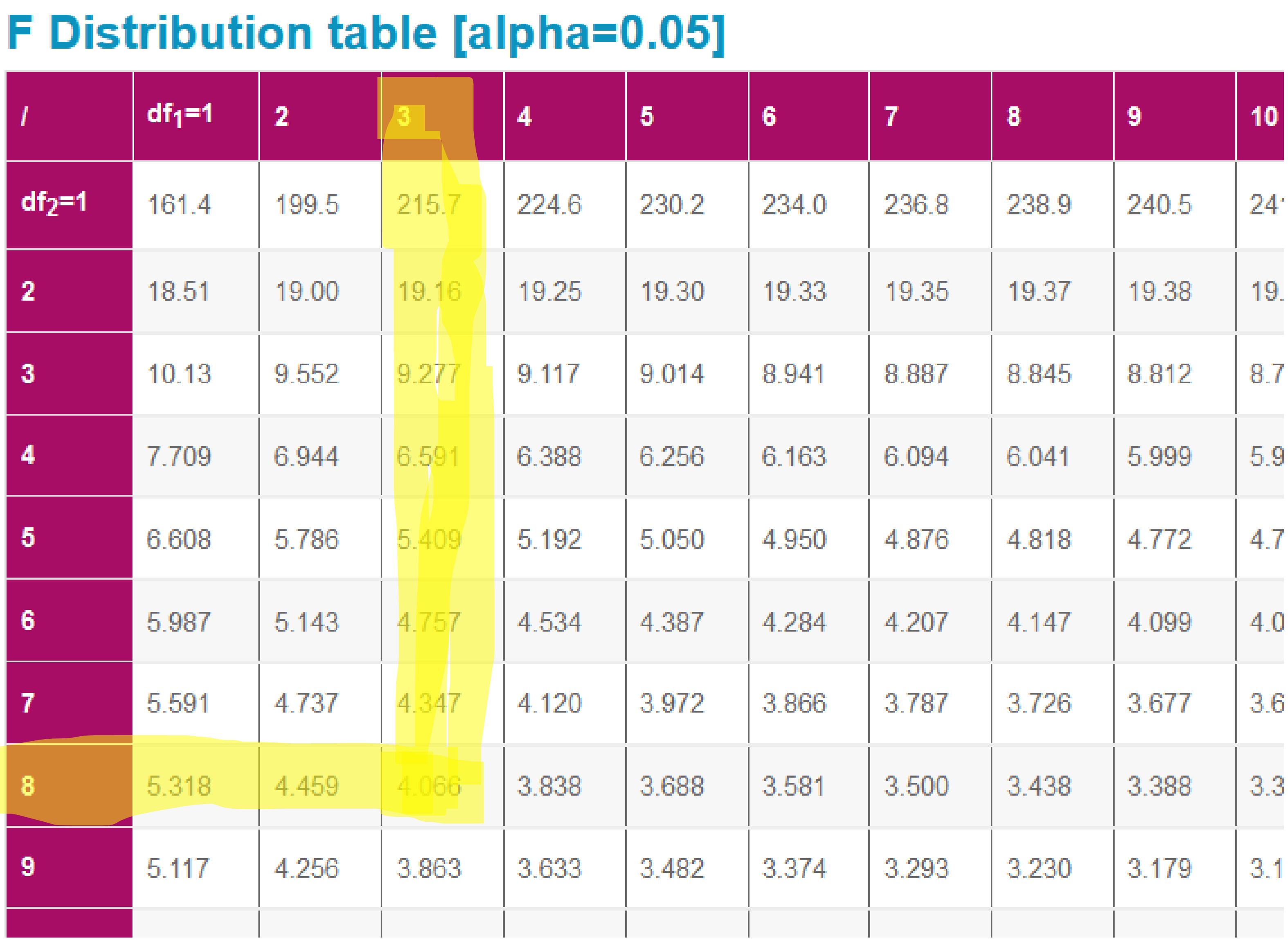
Degrees of Freedom

- Before proceeding to the next step, we need the degrees of freedom.
- There are 12 data values and four samples.
- Thus the number of degrees of freedom of treatment is 4 – 1 = 3.
- The number of degrees of freedom of error is 12 – 4 = 8.

Mean Squares
We now divide our sum of squares by the appropriate number of degrees of freedom in order to obtain the mean squares.
•The mean square for treatment is 30 / 3 = 10.
•The mean square for error is 48 / 8 = 6.

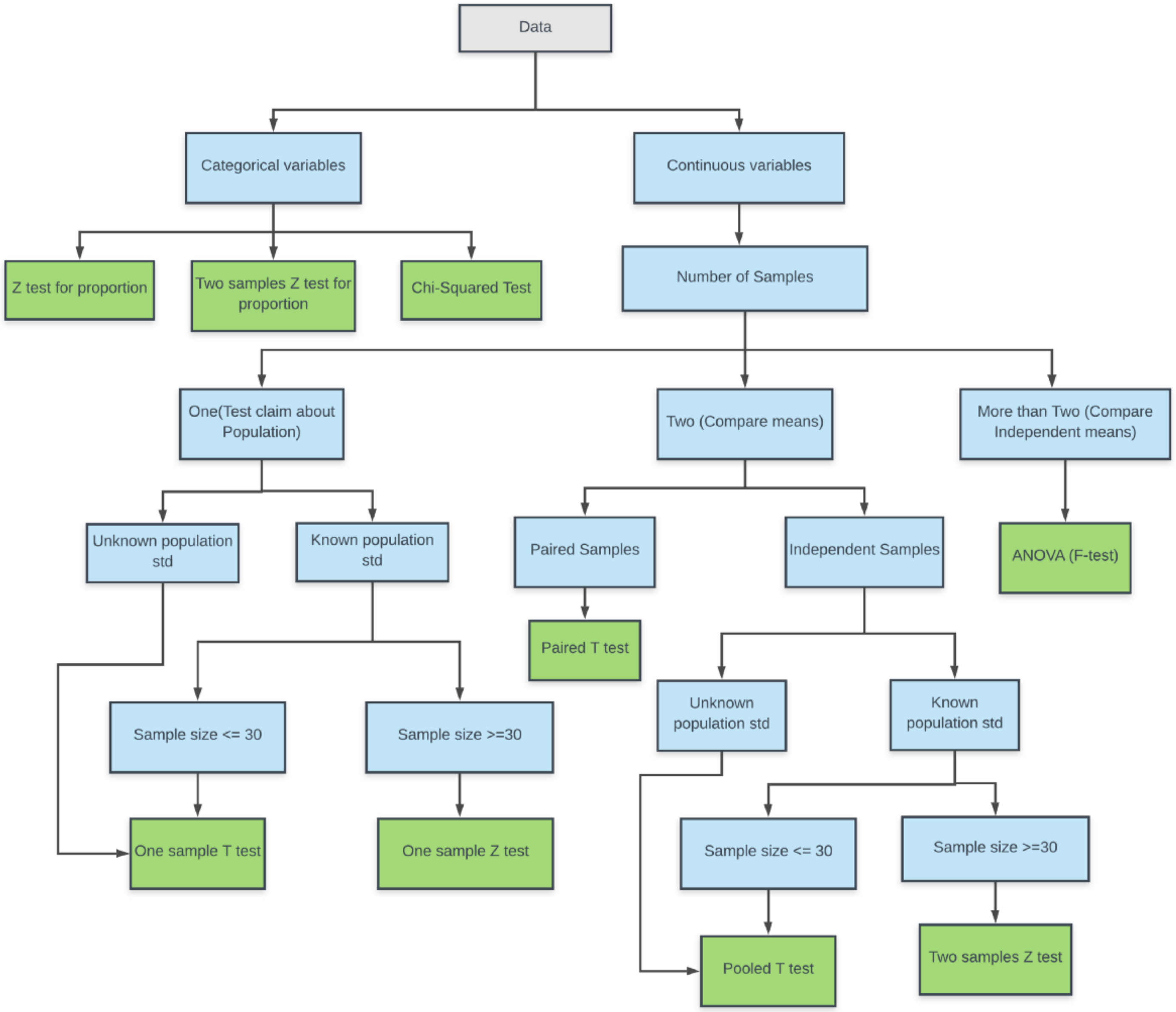
F statistic

- The final step of this is to divide the mean square for treatment by the mean square for error.
- This is the F-statistic from the data. Thus for our example
- $$F = \frac{\text{SST / DF of Treatment}}{\text{SSE / DF of Error}}$$
- $10/6 = 5/3 = 1.667.$



F Distribution table [alpha=0.05]

/	df_1=1	2	3	4	5	6	7	8	9	10
df_2=1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	24
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.
3	10.13	9.552	9.277	9.117	9.014	8.941	8.887	8.845	8.812	8.7
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.9
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.7
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.0
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.6
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.3
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.1



Data

Categorical variables

Continuous variables

Z test for proportion

Two samples Z test for proportion

Chi-Squared Test

Number of Samples

One(Test claim about Population)

Two (Compare means)

More than Two (Compare Independent means)

Unknown population std

Known population std

Paired Samples

Independent Samples

ANOVA (F-test)

Paired T test

Sample size <= 30

Sample size >=30

Unknown population std

Known population std

One sample T test

One sample Z test

Sample size <= 30

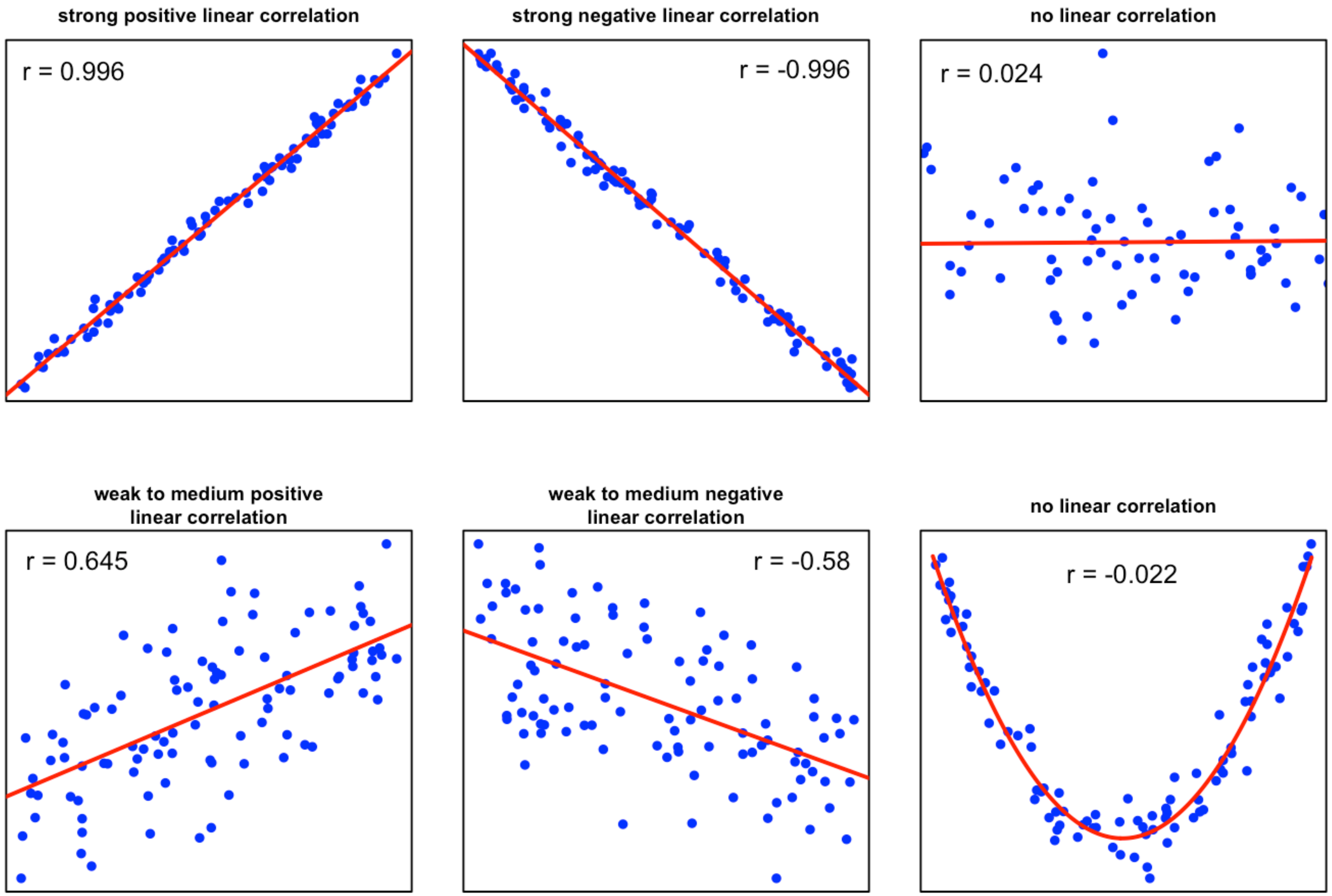
Sample size >=30

Pooled T test

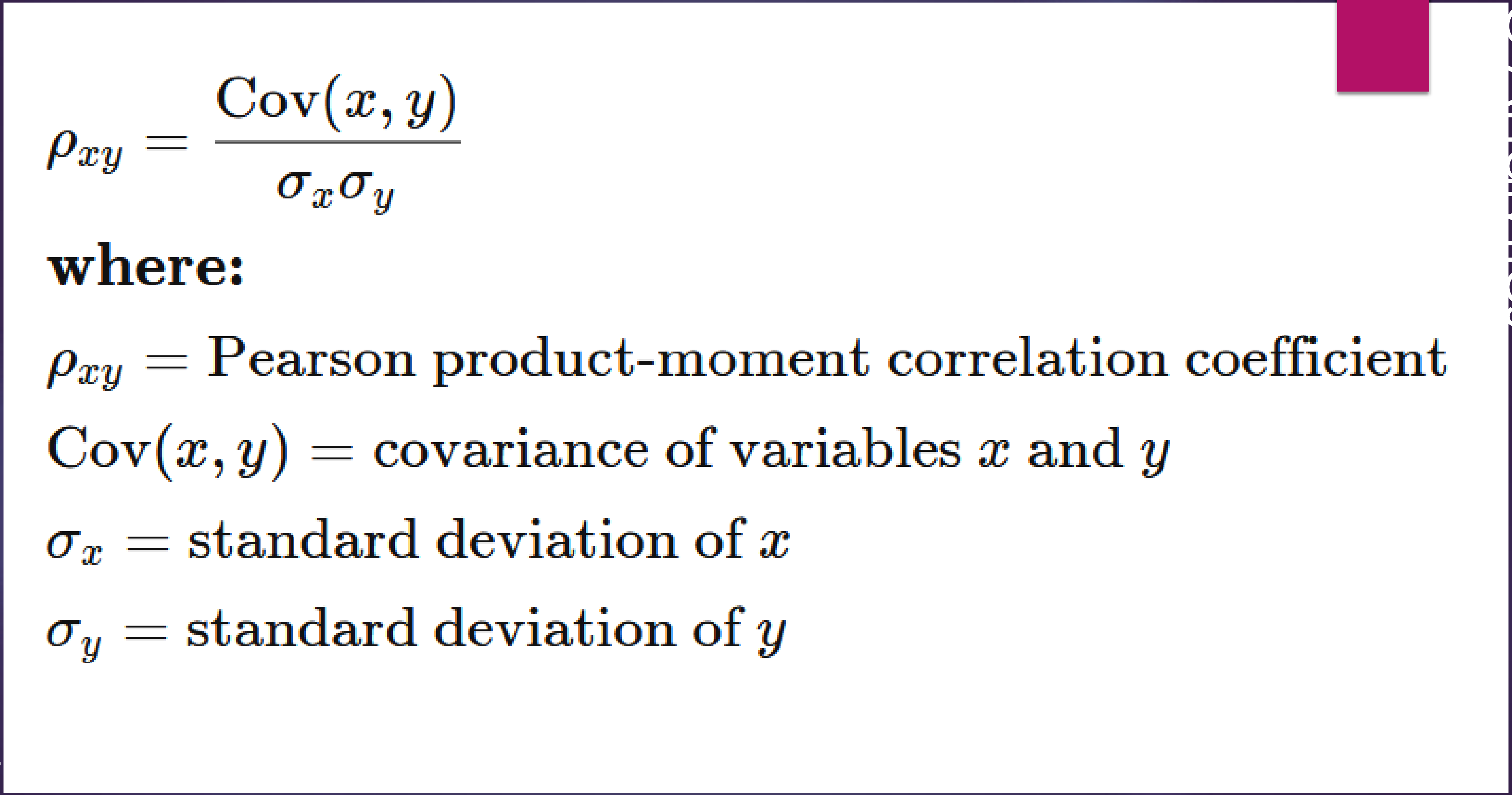
Two samples Z test

Correlation Coefficient

- The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables.
- The values range between -1.0 and 1.0.
- A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement.
- A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation.
- A correlation of 0.0 shows no linear relationship between the movement of the two variables.



MG Analytics



$$\rho_{xy} = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

where:

ρ_{xy} = Pearson product-moment correlation coefficient

$\mathrm{Cov}(x, y)$ = covariance of variables x and y

σ_x = standard deviation of x

σ_y = standard deviation of y