BarCharts, Inc.®

WORLD'S #1 ACADEMIC OUTLINE

Quick Study® ACADEMIC

STATISTICS

parameters.variables.intervals.proportions

THE BASIC PRINCIPLES OF STATISTICS FOR INTRODUCTORY COURSES

DEFINITIONS

❑ **STATISTICS** - A set of tools for collecting, organizing, presenting, and analyzing numerical facts or observations.

1. **Descriptive Statistics** - procedures used to organize and present data in a convenient, useable, and communicable form.
2. **Inferential Statistics** - procedures employed to arrive at broader generalizations or inferences from sample data to populations.

❑ **STATISTIC** - A number describing a sample characteristic. Results from the manipulation of sample data according to certain specified procedures.

❑ **DATA** - Characteristics or numbers that are collected by observation.

❑ **POPULATION** - A complete set of actual or potential observations.

❑ **PARAMETER** - A number describing a population characteristic; typically, inferred from sample statistic.

❑ **SAMPLE** - A subset of the population selected according to some scheme.

❑ **RANDOM SAMPLE** - A subset selected in such a way that each member of the population has an equal opportunity to be selected. **Ex.** ***lottery numbers in a fair lottery***

❑ **VARIABLE** - A phenomenon that may take on different values.

❑ **MEAN** - The point in a distribution of measurements about which the summed deviations are equal to zero. ***Average value of a sample or population.***

POPULATION MEAN

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

SAMPLE MEAN

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Note: The mean is very sensitive to extreme measurements that are not balanced on both sides.

❑ **WEIGHTED MEAN** - Sum of a set of observations multiplied by their respective weights, divided by the sum of the weights:

$$\text{WEIGHTED MEAN} \quad \frac{\sum_{i=1}^{G} w_i x_i}{\sum_{i=1}^{G} w_i}$$

where w_i = weight; x_i = observation; G = number of observation groups. Calculated from a population, sample, or groupings in a frequency distribution.

Ex. ***In the FrequencyDistribution below, the mean is 80.3; calculated by using frequencies for the w_i's. When grouped, use class midpoints for x_i's.***

❑ **MEDIAN** - Observation or potential observation in a set that divides the set so that the same number of observations lie on each side of it. For an odd number of values, it is the middle value; for an even number it is the average of the middle two.

Ex. ***In the Frequency Distribution table below, the median is 79.5.***

❑ **MODE** - Observation that occurs with the greatest frequency. **Ex.** ***In the Frequency Distribution table below, the mode is 88.***

MEASURES OF DISPERSION

❑ **SUM OF SQUARES *(SS)*** - Deviations from the mean, squared and summed:

$$\text{Population SS} = \Sigma(x_i - \mu_x)^2 \text{ or } \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{N}$$

$$\text{Sample SS} = \Sigma(x_i - \bar{x})^2 \text{ or } \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$$

❑ **VARIANCE** - The average of square differences between observations and their mean.

POPULATION VARIANCE

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

SAMPLE VARIANCE

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

VARIANCES FOR GROUPED DATA

POPULATION

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{G} f_i(m_i - \mu)^2$$

SAMPLE

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{G} f_i(m_i - \bar{x})^2$$

❑ **STANDARD DEVIATION** - Square root of the variance:

Ex. ***Pop. S.D.*** $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

FREQUENCY DISTRIBUTION

Shows the number of times each observation occurs when the values of a variable are arranged in order according to their magnitudes.

FREQUENCY DISTRIBUTION

Frequency Distribution of student scores on an exam

x	*f*	*x*	*f*	*x*	*f*	*x*	*f*
100	1	83	11	74	111	65	0
99	1	84	11111	75	1111	66	1
98	0	85	1	76	11	67	11
97	0	86	0	77	111	68	1
96	11	87	1	78	1	69	111
95	0	88	1111111	79	11	70	1111
94	0	89	111	80	1	71	0
93	1	90	11	81	11	72	11
92	0	91	1	82	1	73	111

x = observation *f* = frequency

❑ **GROUPED FREQUENCY DISTRIBUTION** - A frequency distribution in which the values of the variable have been grouped into classes.

GROUPED FREQUENCY DISTRIBUTION

CLASS	f	CLASS	f
98-100	2	80-82	4
95-97	2	77-79	6
92-94	1	74-76	9
89-91	6	71-73	5
86-88	8	68-70	8
83-85	8	65-67	3

GROUPING OF DATA

CUMULATIVE FREQUENCY/PERCENTAGE DISTRIBUTIONS

❑ **CUMULATIVE FREQUENCY DISTRIBUTION** - A distribution which shows the total frequency through the upper real limit of each class.

❑ **CUMULATIVE PERCENTAGE DISTRIBUTION** - A distribution which shows the total percentage through the upper real limit of each class.

CUMULATIVE FREQUENCY / PERCENTAGE DISTRIBUTION

CLASS	f	Cum f	%
65-67	3	3	4.84
68-70	8	11	17.74
71-73	5	16	25.81
74-76	9	25	40.32
77-79	6	31	50.00
80-82	4	35	56.45
83-85	8	43	69.35
86-88	8	51	82.26
89-91	6	57	91.94
92-94	1	58	93.55
95-97	2	60	96.77
98-100	2	62	100.00

GRAPHING TECHNIQUES

❑ **BAR GRAPH** - A form of graph that uses bars to indicate the frequency of occurrence of observations.

- **Histogram** - a form of bar graph used with interval or ratio-scaled variables.

- Interval Scale- a quantitative scale that permits the use of arithmetic operations. The zero point in the scale is arbitrary.

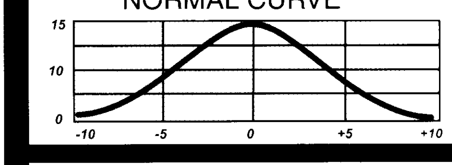
- Ratio Scale- same as interval scale except that there is a true zero point.

❑ **FREQUENCY CURVE** - A form of graph representing a frequency distribution in the form of a continuous line that traces a histogram.

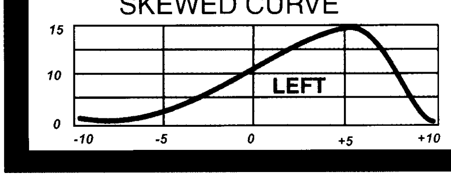
- **Cumulative Frequency Curve** - a continuous line that traces a histogram where bars in all the lower classes are stacked up in the adjacent higher class. It cannot have a negative slope.
- **Normal curve** - bell-shaped curve.
- **Skewed curve** - departs from symmetry and tails-off at one end.

FREQUENCY CURVES

NORMAL CURVE



SKEWED CURVE



PROBABILITY

The long term relative frequency with which an outcome or event occurs.

Probability of occurrence of Event A $\rho(A) = \dfrac{\textit{Number of outcomes favoring Event A}}{\textit{Total number of outcomes}}$

❑ **SAMPLE SPACE** - All possible outcomes of an experiment.

❑ **TYPE OF EVENTS**

- **Exhaustive** - two or more events are said to be exhaustive if all possible outcomes are considered. Symbolically, ρ (A or B or...) = 1.
- **Non-Exhaustive** -two or more events are said to be non-exhaustive if they do not exhaust all possible outcomes.
- **Mutually Exclusive** - Events that cannot occur simultaneously: ρ (A and B) = 0; and ρ (A or B) = ρ (A) + ρ (B). **Ex. *males, females***
- **Non-Mutually Exclusive** - Events that can occur simultaneously: ρ (A or B) = ρ (A) + ρ (B) - ρ (A and B). **Ex. *males, brown eyes.***
- **Independent** - Events whose probability is unaffected by occurrence or nonoccurrence of each other: $\rho(A|B) = \rho(A)$; $\rho(B|A) = \rho(B)$; and ρ(A and B) = $\rho(A)\,\rho(B)$. **Ex. *gender and eye color***
- **Dependent** - Events whose probability changes depending upon the occurrence or non-occurrence of each other: $\rho(A|B)$ differs from $\rho(A)$; $\rho(B|A)$ differs from $\rho(B)$; and ρ(A and B) = $\rho(A)\,\rho(B|A) = \rho(B)\,\rho(A|B)$ **Ex. *race and eye color.***

❑ **JOINT PROBABILITIES** - Probability that 2 or more events occur simultaneously.

❑ **MARGINAL PROBABILITIES** or Unconditional Probabilities = summation of probabilities.

❑ **CONDITIONAL PROBABILITIES** - Probability of ***A*** given the existence of ***S***, written, ρ ***(A\S)***.

❑ **EXAMPLE**- Given the numbers 1 to 9 as observations in a sample space:

•Events mutually exclusive and exhaustive-
Example: ρ *(all odd numbers)*; ρ *(all even numbers)*
•Events mutually exclusive but not exhaustive-
Example: ρ *(an even number)*; ρ *(the numbers 7 and 5)*
•Events neither mutually exclusive or exhaustive-
Example: ρ *(an even number or a 2)*

FREQUENCY TABLE

	EVENT C	EVENT D	TOTALS
EVENT E	52	36	87
EVENT F	62	71	133
TOTALS	114	106	220

Ex. Joint Probability Between C and E
ρ (C & E) = 52/220 = 0.24

JOINT, MARGINAL & CONDITIONAL PROBABILITY TABLE

	EVENT C	EVENT D	MARGINAL PROBABILITY	CONDITIONAL PROBABILITY
EVENT E	0.24	0.16	0.40	(C/E)=0.60 (D/E)=0.40
EVENT F	0.28	0.32	0.60	(C/F)=0.47 (D/F)=0.53
MARGINAL PROBABILITY	0.52	0.48	1.00	
CONDITIONAL PROBABILITY	(E/C)=0.46 (F/C)=0.54	(E/D)=0.33 (F/D)=0.67		

❑ **SAMPLING DISTRIBUTION** - A theoretical ***probability distribution*** of a statistic that would result from drawing all possible samples of a given size from some population.

THE STANDARD ERROR OF THE MEAN

A theoretical standard deviation of sample mean of a given sample size, drawn from some specified population.

❑When based on a very large, known population, the standard error is: $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$

❑When estimated from a sample drawn from very large population, the standard error is:

$$\sigma_{\overline{x}} \approx \frac{s}{\sqrt{n}}$$

❑The dispersion of sample means decreases as sample size is increased.

RANDOM VARIABLES

A mapping or function that assigns one and only one numerical value to each outcome in an experiment.

❑ **DISCRETE RANDOM VARIABLES** - Involves rules or probability models for assigning or generating only distinct values (not fractional measurements).

❑ **BINOMIAL DISTRIBUTION** - A model for the sum of a series of **n** independent trials where trial results in a 0 (failure) or 1 (success). **Ex. *Coin toss*** $p(s) = \binom{n}{s}\pi^s(1-\pi)^{n-s}$

where **p(s)** is the probability of **s** success in **n** trials with a constant π probability per trials, and where $\binom{n}{s} = \dfrac{n!}{s!(n-s)!}$

Binomial mean: $\mu = n\pi$

Binomial variance: $\sigma^2 = n\pi\,(1-\pi)$

As n increases, the Binomial approaches the Normal distribution.

❑ **HYPERGEOMETRIC DISTRIBUTION** - A model for the sum of a series of **n** trials where each trial results in a 0 or 1 and is drawn from a small population with **N** elements split between $\mathbf{N_1}$ successes and $\mathbf{N_2}$ failures. Then the probability of splitting the **n** trials between $\mathbf{x_1}$ successes and $\mathbf{x_2}$ failures is:

$$p(x_1 \text{ and } x_2) = \frac{\dfrac{N_1!}{x_1!(N_1-x_1)!}\;\dfrac{N_2!}{x_2!(N_2-x_2)!}}{\dfrac{N!}{n!(N-n)!}}$$

Hypergeometric mean: $\mu_1 = E(x_1) = \dfrac{nN_1}{N}$

and variance: $\sigma^2 = \dfrac{N-n}{N-1}\left[\dfrac{nN_1}{N}\right]\left[\dfrac{N_2}{N}\right]$

❑ **POISSON DISTRIBUTION** - A model for the number of occurrences of an event x = 0,1,2,..., when the probability of occurrence is small, but the number of opportunities for the occurrence is large, for **x = 0,1,2,3,...** and **λ > 0** , otherwise ***P*(*x*) = 0.**

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Poisson mean and variance: λ.

For continuous variables, frequencies are expressed in terms of areas under a curve.

❑ **CONTINUOUS RANDOM VARIABLES** - Variable that may take on any value along an uninterrupted interval of a numberline.

❑ **NORMAL DISTRIBUTION** - bell curve; a distribution whose values cluster symmetrically around the mean (also median and mode).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$

where $f\,(x)$ = frequency at a given value
σ = standard deviation of the distribution
π = approximately 3.1416
e = approximately 2.7183
μ = the mean of the distribution
x = any score in the distribution

❑ **STANDARD NORMAL DISTRIBUTION** - A normal random variable Z, that has a mean of 0, and standard deviation of 1.

❑ **Z-VALUES** - The number of standard deviations a specific observation lies from the mean:

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

TESTING STATISTICAL HYPOTHESES

❑ **LEVEL OF SIGNIFICANCE** - A probability value considered rare in the sampling distribution. specified under the null hypothesis where one is willing to acknowledge the operation of chance factors. Common significance levels are 1%, 5%, 10%. Alpha (α) level = the lowest level for which the null hypothesis can be rejected. The significance level determines the critical region.

❑ **NULL HYPOTHESIS (H_0)** - A statement that specifies hypothesized value(s) for one or more of the population parameter. **[Ex. H_0** = a coin is unbiased. That is p = 0.5.]

❑ **ALTERNATIVE HYPOTHESIS (H_1)** - A statement that specifies that the population parameter is some value other than the one specified under the null hypothesis. **[Ex. H_1**= a coin is biased. That is $p \neq 0.5$.]

1. NONDIRECTIONAL HYPOTHESIS - an alternative hypothesis (H_1) that states only that the population parameter is different from the one specified under H_0. **Ex.** $H_1\!: \mu \neq \mu_0$ Two-Tailed Probability Value is employed when the alternative hypothesis is non-directional.

2. DIRECTIONAL HYPOTHESIS - an alternative hypothesis that states the direction in which the population parameter differs from the one specified under H_0. **Ex.** $H_1\!: \mu > \mu_0$ *or* $H_1\!: \mu < \mu_0$ One-Tailed Probability Value is employed when the alternative hypothesis is directional.

❑ **NOTION OF INDIRECT PROOF** - Strict interpretation of hypothesis testing reveals that the null hypothesis can never be proved. **[Ex.** If we toss a coin 200 times and tails comes up 100 times. it is no guarantee that heads will come up exactly half the time in the long run; small discrepancies might exist. A bias can exist even at a small magnitude. We can make the assertion however that NO BASIS EXISTS FOR REJECTING THE HYPOTHESIS THAT THE COIN IS UNBIASED. (*The null hypothesis is not rejected*). When employing the 0.05 level of significanc reject the null hypothesis when a given resu occurs by chance 5% of the time or less.]

❑ **TWO TYPES OF ERRORS**

- Type 1 Error (Type α Error) = the rejection of H_0 when it is actually true. The probability of a type 1 error is given by α.
- Type II Error (Type β Error) = The acceptance of H_0 when it is actually false. The probability of a type II error is given by β.

Statistical Hypotheses		True Status of Ho	
		Ho True	Ho False
Decision:	Accept Ho	Correct (1-α)	Type II error (β)
	Reject Ho	Type I error (α)	Correct (1-β)

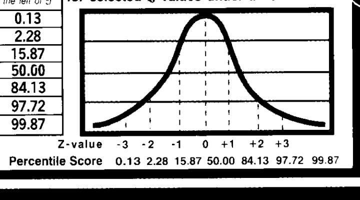
CENTRAL LIMIT THEOREM

(for sample mean $\overline{x}$)

•If $x_1, x_2, x_3, \ldots x_n$, is a simple random sample of n elements from a large (infinite) population, with mean mu(μ) and standard deviation σ, then the distribution of $\overline{x}$ takes on the bell shaped distribution of a normal random variable as n increases and the distribution of the ratio: $\dfrac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ approaches the standard normal distribution as n goes to infinity. In practice, a normal approximation is acceptable for samples of 30 or larger.

Percentage Cumulative Distribution
for selected z values under a normal curve

z	% Cum (total area to the left of z)
-3	0.13
-2	2.28
-1	15.87
0	50.00
+1	84.13
+2	97.72
+3	99.87



QuickStudy

INFERENCE FOR PARAMETERS

UNBIASEDNESS - Property of a reliable estimator being estimated.

- **Unbiased Estimate of a Parameter** - an estimate that equals on the average the value of the parameter.

Ex. ***the sample mean is an unbiased estimator of the population mean.***

- **Biased Estimate of a Parameter** - an estimate that does not equal on the average the value of the parameter.

Ex. ***the sample variance calculated with n is a biased estimator of the population variance, however, when calculated with n-1 it is unbiased.***

❑ **STANDARD ERROR** - The standard deviation of the estimator is called the standard error.

Ex. The standard error for x̄'s is. $\sigma_{\overline{x}} = \sigma/\sqrt{n}$

This has to be distinguished from the STANDARD DEVIATION OF THE SAMPLE:

$$s = \sqrt{\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- The standard error measures the variability in the x̄'s around their expected value E(x̄) while the standard deviation of the sample reflects the variability in the sample around the sample's mean (x̄).

❑ **USED WHEN THE STANDARD DEVIATION IS UNKNOWN** -Use of **Student's *t*.** When σ is not known, its value is estimated from sample data.

- ***t*-ratio-** the ratio employed in the testing of hypotheses or determining the significance of a difference between means (two-sample case) involving a sample with a t-distribution. The formula is:

$\dfrac{\overline{x} - \mu}{s_{\overline{x}}}$ where μ = population mean under H_0

and $s_{\overline{x}} = s/\sqrt{n}$

- **Distribution**-symmetrical distribution with a mean of zero and standard deviation that approaches one as degrees of freedom increases (i.e., approaches the Z distribution).

Assumption and condition required in assuming *t*-distribution: Samples are drawn from a normally distributed population and σ (population standard deviation) is unknown.

- **Homogeneity of Variance-** If 2 samples are being compared, the assumption in using t-ratio is that the variances of the populations from where the samples are drawn are equal.

- Estimated $\sigma_{\overline{X}_1 - \overline{X}_2}$ (that is $s_{\overline{X}_1 - \overline{X}_2}$) is based on the unbiased estimate of the population variance.

- **Degrees of Freedom (*df*)**- the number of values that are free to vary after placing certain restrictions on the data.

Example. ***The sample (43,74,42,65) has n = 4. The sum is 224 and mean = 56. Using these 4 numbers and determining deviations from the mean, we'll have 4 deviations namely (–13,18,–14,9) which sum up to zero. Deviations from the mean is one restriction we have imposed and the natural consequence is that the sum of these deviations should equal zero. For this to happen, we can choose any number but our freedom to choose is limited to only 3 numbers because one is restricted by the requirement that the sum of the deviations should equal zero. We use the equality:***

$$(x_1-\overline{x})+(x_2-\overline{x})+(x_3-\overline{x})+(x_4-\overline{x})=0$$

So given a mean of 56, if the first 3 observations are 43, 74, and 42, the last observation has to be 65. This single restriction in this case helps us determine df. The formula is n less number of restrictions. In this case, it is n–1= 4-1=3df.

- **_*t*-Ratio is a robust test-** This means that statistical inferences are likely valid despite fairly large departures from normality in the population distribution. If normality of population distribution is in doubt, it is wise to increase the sample size.

USING THE Z - STATISTIC

❑ **USED WHEN THE STANDARD DEVIATION IS KNOWN:** When σ is known it is possible to describe the form of the distribution of the sample mean as a Z statistic. The sample must be drawn from a normal distribution or have a sample size (n) of at least 30.

$z = \dfrac{\overline{x} - \mu}{\sigma_{\overline{x}}}$ where u = population mean (either known or hypothesized under H_o) and $\sigma_{\overline{x}} = \sigma/\sqrt{n}$.

- **Critical Region** - the portion of the area under the curve which includes those values of a statistic that lead to the rejection of the null hypothesis.

- The most often used significance levels are 0.01, 0.05, and 0.1. For a one-tailed test using *z*-statistic, these correspond to z-values of 2.33, 1.65, and 1.28 respectively. For a two-tailed test, the critical region of 0.01 is split into two equal outer areas marked by z-values of |2.58|.

Example 1. ***Given a population with μ=250 and σ = 50, what is the probability of drawing a sample of n=100 values whose mean (x̄) is at least 255? In this case, Z=1.00. Looking at Table A, the given area for Z=1.00 is 0.3413. To its right is 0.1587(=0.5-0.3413) or 15.85%.***

Conclusion: there are approximately 16 chances in 100 of obtaining a sample mean = 255 from this population when n = 100.

Example 2. ***Assume we do not know the population mean. However, we suspect that it may have been selected from a population with μ = 250 and σ = 50, but we are not sure. The hypothesis to be tested is whether the sample mean was selected from this population. Assume we obtained from a sample (n) of 100, a sample mean of 263. Is it reasonable to assume that this sample was drawn from the suspected population?***

1. H_0:μ = 250 (that the actual mean of the population from which the sample is drawn is equal to 250) H_1: μ not equal to 250 (the alternative hypothesis is that it is greater than or less than 250, thus a two-tailed test).

2. *z*-statistic will be used because the population σ is known.

3. Assume the significance level (α) to be 0.01. Looking at Table A, we find that the area beyond a *z* of 2.58 is approximately 0.005.

To reject H_0 at the 0.01 level of significance, the absolute value of the obtained *z* must be equal to or greater than $|z_{0.01}|$ or 2.58. Here the value of z corresponding to sample mean = 263 is 2.60.

❑ **CONCLUSION-** Since this obtained *z* falls within the critical region, we may reject H_o at the 0.01 level of significance.

❑ **CONFIDENCE INTERVAL-** Interval within which we may consider a hypothesis tenable. Common confidence intervals are 90%, 95%, and 99%. Confidence Limits: limits defining the confidence interval.

(1- α)100% confidence interval for μ :

$$\overline{x}_i - z_{\alpha/2}\left(\sigma/\sqrt{n}\right) \le \mu \le \overline{x} + z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$$

where $Z_{\alpha/2}$ is the value of the standard normal variable z that puts α/2 percent in each tail of the distribution. The confidence interval is the complement of the critical regions.

A t-statistic may be used in place of the z-statistic when σ is unknown and s must be used as an estimate. (But note the caution in that section.)

Critical region for rejection of H_0 when α = 0.01, two-tailed test

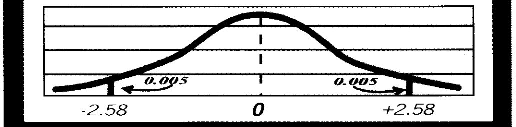


Table A Normal Curve Areas

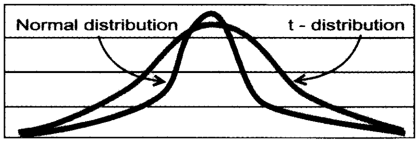
area from mean to z
0
z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Example. Given x̄=108, s=15, and n=26 estimate a 95% confidence interval for the population mean. Since the population variance is unknown, the t-distribution is used. The resulting interval, using a t-valve of 2.060 from Table B (row 25 of the middle column), is approximately 102 to 114. Consequently, any hypothesized μ between 102 to 114 is tenable on the basis of this sample. Any hypothesized μ below 102 or above 114 would be rejected at 0.05 significance.

❑ **COMPARISON BETWEEN *t* AND *z* DISTRIBUTIONS**

Although both distributions are symmetrical about a mean of zero, the *t*-distribution is more spread out than the normal distribution (*z*-distribution).

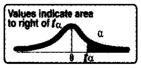


Thus a much larger value of *t* is required to mark off the bounds of the critical region of rejection.

As *df* increases, differences between *z*- and *t*- distributions are reduced. Table A (*z*) may be used instead of Table B (*t*) when n>30. To use either table when n<30, the sample must be drawn from a normal population.

Table B Critical Values of *t*

A*=Level of significance for one-tailed test
B*=Level of significance for two-tailed test



df A*	0.1	0.05	0.025	0.01	0.005
df B*	0.2	0.1	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
inf.	1.282	1.645	1.960	2.326	2.576

QuickStudy

TESTING INDEPENDENT SAMPLES

❑ **SAMPLING DISTRIBUTION OF THE DIFFERENCE BETWEEN MEANS-** If a number of pairs of samples were taken from the same population or from two different populations, then:

- The distribution of differences between pairs of sample means tends to be normal (z-distribution).
- The mean of these differences between means $\mu_{\bar{x}_1-\bar{x}_2}$ is equal to the difference between the population means, that is $\mu_1-\mu_2$.

❑ **Z-DISTRIBUTION: σ_1 and σ_2 are known**

- The standard error of the difference between means

$$\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{(\sigma_1^2)/n_1+(\sigma_2^2)/n_2}$$

- Where $(\mu_1 - \mu_2)$ represents the hypothesized difference in means, the following statistic can be used for hypothesis tests:

$$Z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}}$$

- When n_1 and n_2 are >30, substitue s_1 and s_2 for σ_1 and σ_2, respectively.

$$S_{\bar{x}_1-\bar{x}_2}=\sqrt{\left(\frac{SS_1+SS_2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

(*To obtain sum of squares (SS) see Measures of Central Tendency on page 1*)

❑ **POOLED *t*-TEST**

- **Distribution is normal**
- ***n* < 30**
- **σ_1 and σ_2 are *not* known but assumed equal**

- The hypothesis test may be 2 tailed (= vs. ≠) or 1 tailed: $\mu_1 \le \mu_2$ and the alternative is $\mu_1 > \mu_2$ (or $\mu_1 \ge \mu_2$ and the alternative is $\mu_1 < \mu_2$.)
- degrees of freedom(*df*): $(n_1-1)+(n_2-1)=n_1+n_2-2$.
- Use the given formula below for estimating $\sigma_{\bar{x}_1-\bar{x}_2}$ to determine $s_{\bar{x}_1-\bar{x}_2}$.
- Determine the critical region for rejection by assigning an acceptable level of significance and looking at the *t*-table with $df= n_1 + n_2-2$.

- **Use the following formula for the estimated standard error:**

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{\left[\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\right]\left[\frac{n_1+n_2}{n_1n_2}\right]}$$

F - TEST

❑ **HETEROGENEITY OF VARIANCES** may be determined by using the F-test:

$$F=\frac{S_1^2(\textit{larger variance})}{S_2^2(\textit{smaller variance})}$$

❑ **NULL HYPOTHESIS-** Variances are equal and their ratio is one.

❑**ALTERNATIVE HYPOTHESIS-**Variances differ and their ratio is not one.

❑ Look at "**Table C**" below to determine if the variances are significantly different from each other. Use degrees of freedom from the 2 samples:(n_1-1, n_2-1).

Table C Critical Values of *F*

Top row=.05, Bottom row=.01 points for distribution of *F*

Degrees of freedom for numerator

Degrees of freedom for denominator	1	2	3	4	5	6	7	8	9	10
1	161	200	216	225	230	234	237	239	241	242
	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	1937	19.38	19.39
	98.49	99.01	99.17	99.25	99.30	99.33	99.34	99.36	99.38	99.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78
	34.12	30.81	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23
4	7.71	6.94	6.59	6.39	6.26	6.16	6.08	6.04	6.00	5.96
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63
	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34
	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97
	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.08	4.95	4.85

CORRELATED SAMPLES

❑ **STANDARD ERROR OF THE DIFFERENCE** between Means for Correlated Groups. The general formula is:

$$S_{\bar{x}_1-\bar{x}_2}=\sqrt{s_{\bar{x}_1}^2+s_{\bar{x}_2}^2-2rs_{\bar{x}_1}s_{\bar{x}_2}}$$

where r is Pearson correlation

- By matching samples on a variable correlated with the criterion variable, the magnitude of the standard error of the difference can be reduced.
- The higher the correlation, the greater the reduction in the standard error of the difference.

ANALYSIS OF VARIANCE (*ANOVA*)

❑ **PURPOSE-** Indicates possibility of overall mean effect of the experimental treatments before investigating a specific hypothesis.

❑ **ANOVA-** Consists of obtaining independent estimates from population subgroups. It allows for the partition of the sum of squares into known components of variation.

❑ **TYPES OF VARIANCES**

- **Between-Group Variance (BGV)-** reflects the magnitude of the difference(s) among the group means.
- **Within-Group Variance (WGV)-** reflects the dispersion within each treatment group. It is also referred to as the error term.

❑ **CALCULATING VARIANCES**

- Following the F-ratio, when the BGV is large relative to the WGV, the F-ratio will also be large.

$$BGV=\frac{n\Sigma(\bar{x}_i-\bar{x}_{tot})^2}{k-1}$$

where x_i = mean of i^th treatment group and x_{tot} = mean of all n values across all k treatment groups.

$$WGV=\frac{SS_1+SS_2+...+SS_k}{n-k}$$

where the SS's are the sums of squares (see *Measures of Central Tendency* on page 1) of each subgroup's values around the subgroup mean.

❑ **USING F-RATIO-** $F = BGV/WGV$

- Degrees of freedom are k-1 for the numerator and n-k for the denominator.
- If ***BGV > WGV***, the experimental treatments are responsible for the large differences among group means. Null hypothesis: the group means are estimates of a common population mean.

PROPORTIONS

In random samples of size *n*, the sample proportion *p* fluctuates around the proportion mean = π with a proportion variance of $\frac{\pi(1-\pi)}{n}$ proportion standard error of $\sqrt{\pi(1-\pi)/n}$

As the sampling distribution of *p* increases, it concentrates more around its target mean. It also gets closer to the normal distribution. In which case: $z=\frac{p-\pi}{\sqrt{\pi(1-\pi)/n}}$

ISBN 157222649-8

50495

9 781572 226494

quickstudy.com

6 54614 20649 7

CORRELATION

Definition - Correlation refers to the relationship between two variables. The Correlation Coefficient is a measure that expresses the extent to which two variables are related.

❑ **"PEARSON r"METHOD** (Product-Moment Correlation Coefficient) - Corelation coefficient employed with interval- or ratio-scaled variables.

Ex.: Given observations to two variables *X* and *Y*, we can compute their corresponding *z* values: $Z_x=(x-\bar{x})/s_x$ and $Z_y=(y-\bar{y})/s_y$.

- The formulas for the Pearson correlation (r):

$$r=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{SS_x\cdot SS_y}}$$

- Use the above formula for large samples.
- Use this formula (also known as the ***Mean-Deviation Method*** of computing the Pearson r) for small samples.

$$r=\frac{\Sigma(z_x z_y)}{n}$$

❑ **RAW SCORE METHOD** is quicker and can be used in place of the first formula above when the sample values are available.

$$r=\frac{\Sigma xy-\frac{(\Sigma x)(\Sigma y)}{N}}{\sqrt{\left[\Sigma x^2-\frac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2-\frac{(\Sigma y^2)}{n}\right]}}$$

- Most widely-used non-parametric test.
- The χ^2 mean = its degrees of freedom.
- The χ^2 variance = twice its degrees of freedom.
- Can be used to test one or two independent samples.
- The square of a standard normal variable is a chi-square variable.
- Like the t-distibution, it has different distributions depending on the degrees of freedom.

❑ **DEGREES OF FREEDOM (*d.f.*) COMPUTATION**

- If ***chi-square*** tests for the goodness-of-fit to a hypothesized distribution,

$$d.f. = g - 1 - m, \text{ where}$$

g = number of groups, or classes, in the frequency distribution.

m = number of population parameters that must be estimated from sample statistics to test the hypothesis.

- If ***chi-square*** tests for homogeneity or contingency:

$$d.f. = (rows-1)\ (columns-1)$$

❑ **GOODNESS-OF-FIT TEST-** To apply the chi-square distribution in this manner, the critical chi-square value is expressed as:

$$\sum\frac{(f_o-f_e)^2}{f_e} \textbf{ where}$$

f_0 = observed frequency of the variable

f_e = expected frequency (based on hypothesized population distribution).

❑ **TESTS OF CONTINGENCY-** Application of Chi-square tests to two separate populations to test statistical independence of attributes.

❑ **TESTS OF HOMOGENEITY-** Application of Chi-square tests to two samples to test if they came from populations with like distributions.

❑ **RUNS TEST-** Tests whether a sequence (to comprise a sample) is random. The following equations are applied:

$$(\bar{R})=\frac{2n_1n_2}{n_1+n_2}+1 \textit{ and } S_R\sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

Where

$\bar{R}$ = mean number of runs

n_1 = number of outcomes of one type

n_2 = number of outcomes of the other type

S_R = standard deviation of the distribution of the number of runs.

U.S.$4.95 / CAN.$7.50 **February 2003**

NOTICE TO STUDENT: This *QUICK STUDY*™ chart covers the basics of Introductory Statistics. Due to its condensed format, however, use it as a Statistics *guide* and **not as a replacement for assigned course work.**

Customer Hotline # 1.800.230.9522

Manufactured by Laminating Services, Inc., Largo, FL, and sold under license from Fast M[illegible] Holdings, LLC., Palo Alto, CA USA under one or more of the following Patents: U.S. Patent No. 5,063,637 or 5,273,432