18/11/2011
Mohsen Mohammed Taqi Mohsen Al-Saleh

Statistics-6000

- **Variable:** are "characteristic that can take on different values" with respect to persons, time, and place *and types of variables are as follow:*

- **Independent (X)** you can choose and manipulate. Usually on x-axis

- **Dependent (Y)** is what you measure in the experiment and what is affected during the experiment. Usually on y-axis

- **Intermediate** is a variable in a causal pathway that causes variation in the dependent variable and is itself caused to vary by the independent variable

- **Confounder** is an extraneous variable in a statistical model that correlates (positively or negatively) with both the dependent variable and the independent variable. The methodologies of scientific studies therefore need to account for these variables - either through true experimental designs, in which case, one achieves control, or through statistical means. (Internal Validity)

- **Discrete Variable:** This is a whole number and countable variable. Ordinal, Ranking Type or Nominal Classificatory Categorical Type. (qualitative variable)

- **Continuous or measurable variable:** variables have *no gaps* between them. Have decimal points and units. (Quantitative variable)

- **Why statistic in general:** collection of data, summarization and analyzing of data set, evaluation, conduct a research and finally making conclusion (Testing hypothesis)

- **Specific goal of statistic:** define a normal range (μ and σ), correlation study (relationship), regression study (prediction), association (Qui-Square Test) & agreement testing (Cronblach Alpha & Kappa Cohen Correlation), testing hypothesis (z,t,f) and quality control (L G Chart)

- **Sample (n):** small random group of individuals or observations that is chosen for study from population. Sample is a part of population.

- **Random sample:** is the selection of the sample such that every member from the population has an equal chance of being included in the sample

- **Sampling unit:** A part from population, an individual, household, school, section, village
- **Sampling frame:** a complete list of sampling units in the population
- **Why we need sample study:**

    - Less time
    - Less personnel
    - Less resources
    - Less money
    - For in-depth study

- **Sample size:** the number of individuals or observations under study. ($n \geq 30$)
- **Sampling methods:**

    - **Simple Random Sampling:** Each unit in this method has an equal probability of being included in the sample. (Lottery sample) by using tables of random numbers. Is used when there is homogeneity in the study elements of the population. (N) is small
    - **Stratified Sampling:** The study elements of population are heterogeneous. (N) is lager. (Stratum). Precision (1/SE) of the estimate will be high (SE will be less)
    - **Systematic Sampling (convenience):** (N) is very large. (K)=N/n; is sampling interval. One number (X) is chosen randomly from (1 to K). X+0K, X+1K, X+2K X+3K… , X+ (n-1) K are included in the sample. Precision of the estimate will be less.
    - **Cluster Sampling:** (N) is large and it's not possible to get complete listing of the population unit. Precision of the estimate will be less.
    - **Multi Stage Sampling:** (N) is very large. Sampling is done in stages. Precision of the estimate will be less.
    - **Quota Sampling:** (Sampling of Convenience). (n) Is fixed and not probability sampling method. Not randomly selected. Results cannot be generalized but applicable to that area only. Not good sampling method.

- **Population (N):** Aggregate of subjects under consideration. Whole group is representative

- Parameters (μ and σ) ≠ Statistics ($\bar{x}$ and SD or s)

- **Statistical methods: <u>descriptive method</u> and <u>inference method</u>**

- **Descriptive method:** frequency tables, diagrams, graphs (bar chart, pie chart, pictogram, histogram, frequency polygon and curves-linerity), arithmetic or geometric or weighted mean, median, mode, range, quartile deviation(IQR), mean deviation, standard deviation(SD), coefficient of variation (CV%), correlation coefficient (r)-Pearson Product Moment Correlation, and regression analysis used for *predication*.

- **Inference analysis:** used to *generalize the results*, obtained from the random sample, for the population from which the representative sample was selected. Two main components of inference method are:

    - **Estimation of Parameters (population values)**
    - **Testing the Statistical Significance of the Hypothesis**

- **Measure of location:** mean, mode, and median. They are one single value to represent the distribution. When these values describe a population they called *parameters*. If the describe a sample then referred as *statistic(s)*.

- **Mean ($\bar{x}$ or μ) =** $\frac{\sum x}{n}$ **or** $\frac{\sum x}{N}$ **مجموع القيم على العدد**

- **Median:** is the middle most value of the *arrange data set* (continuous distribution). The value of it is not affected by the extreme values and therefore median is preferred to mean when there are extreme values. When sample not normally distributed

- **Mode:** the most frequent observation of data/distribution. Distribution *may have more than 1 mode*.

- **There are 2 types of data?** *Group data* and *Un-group data (very rich)*

- **Why we group the data?** Grouping the actual data collected will lose enrichment of the data set from its actual values but some time we need to hide the actual data from the public and other competitors or for simplification of data we handing large data set.

- **$\sum$f = n or N** ; total number of frequency = number of observations (sample size)

- **Number of classes or groups needed to make histogram: $2^k \geq$ n or N**

- **Class Interval Size =** $\frac{Maximum-Minimum}{k}$**;** *this is increment value that would be added*

- **For group data arithmetic mean;** $\bar{x} = \frac{\sum mf}{\sum f}$ **, where (m = mid-value of class interval)**

- **Mid-value = (Lower limit:L1 + Upper limit:L2) ÷ 2;** *these L = real limits only*
- **∑(x-$\bar{x}$) = Zero, always**
- **Variance for a group data; (SD^2 or σ^2) =** $\frac{\sum fm^2}{\sum f} - \bar{x}^2$
- **While computing arithmetic mean for a given grouped frequency distribution, it is assumed that all values falling in a particular group or class are located at the mid-point of the group.**
- **For group median=** $L_1 + \left(\frac{L2-L1}{f}\right) x \left(\frac{N}{2} - C\right)$**, f = median frequency, C=cumulative fre.**
- **Law of "next"**
- **If the given class limits are "score limits" then convert them to "real limits"**
- **Last group of cumulative frequency = N or n or ∑f**
- **For group mode =** $L_1 + \left(\frac{L2-L1}{2f-f1-f2}\right) x\ (f - f1)$ **;** ***class with maximum frequency***
- **Quartiles and Percentiles:** are the values in the continuous distribution showing the proportion/percentage of lying *below (or up to)* the given value
- $Q_i = L_1 + \left(\frac{L2-L1}{f}\right)$ **x** $\left(\frac{i\ x\ N}{4} - C\right)$**; i = 1,2,3** *(looks very likely to median formula)*
- **Interquartile range (IQR):** reflects the variability among the middle 50% of the observation of the data. Better than range (uses extreme values only)
- **Q_1 (25%) and Q_2 (50%) and Q_3 (75%)**
- **IQR = $Q_3 - Q_1$; better than 'range' = 75%-25%=50%**
- **$P_{50} = Q_2$ = Median;** of continuous data distribution
- **Real times limits used for group data for:** median, mode, quartiles, and percentiles

- **$P_i = L_1 + \left(\frac{L2-L1}{f}\right)$ x $\left(\frac{i\ x\ N}{100} - C\right)$; i = 1,2,3,….,99** *(looks very likely to median formula)*
- **Rule of "next" to locate the class interval from cumulative frequency distribution**
- **Measure of Variability** = Range, IQR, Variance, SD, and Coefficient of Variation
- **Measure of Variability** = *Scatter* or *dispersion* of data around the mean
- **Range = Largest observation – Smallest observation**
- **$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$ or $SD^2 = \frac{\sum(x-\bar{x})^2}{n-1}$; variance of ungroup data**
- **Group data σ^2 or $SD^2 = \left(\frac{\sum x^2}{n-1}\right) - \left(\frac{(\sum x)^2}{n(n-1)}\right)$; no need for $\bar{x}$**
- $\bar{x} = \frac{\sum fm}{\sum f}$
- **σ or SD** = $+\sqrt{\sigma^2 \text{or } SD^2}$; *unit of SD is similar to observation value*
- **CV = $\frac{SD}{\bar{x}}$ x 100 ; no unit its unitless quantity**
- **CV% is used to compare variation between same sample variables or different**
- **An event = outcome**
- **Probability of (A)** = is the proportion of times the outcomes would occur in a very long series of repetitions. (all events are equally likely)
- **$P(A) = \frac{m}{n}$ $(0 \leq m \leq n)$; when (n) is exhaustive, mutually exclusive**
- **Equally likely trials of (m) is possible**

- **Independent events:** two events are said to be independent if the presence or absence of one does not alter the chances of the other being present, or of the occurrence of one does not alter the chance of occurrence of the other. *(means that they can occur together)*

- **Mutually exclusive events:** if they cannot both occur together or be present at the same time. *No overlapping between the outcomes*. Coins flipping head or tail

- **Additive rule: mutually exclusive events –** the probability of occurrence of 2 or more mutually exclusive events is the *sum* of their probabilities of each outcome

- **P (A or B) = P (A) + P (B) e.g. throwing die for odd numbers- mutually exclusive ev.**

- **Multiplicative rule: Independent events** – probability of simultaneous occurrence of events A and B in a series of independent trails (i.e. chance of one outcome occurring is not affected by knowledge of whether or not the other occurred) is the product of their probabilities.

- **P (A and B) = P(A) x P(B)→Independent events**

- **General additive rule:** if the 2 events are *not* mutually exclusive, then the probability that either event A or B occurs is: **P(A or B or both) = P(A) + P(B) – P(A & B)**

- **Discrete Probability Distribution (DPD):** *sum of p(x)s = 1*, *probability of each outcome is between 0-1*, *outcomes are mutually exclusive*.

- $\mu=\sum(x_i\,p(x_i))$ **and** $\sigma^2=\sum((x_i-\mu)^2.p(x_i))$ **; for discrete probability distribution**

- **Conditional probability:**

- **Joint probability: P(A∩B)= P(A) x P(B) = multiplicative rule**

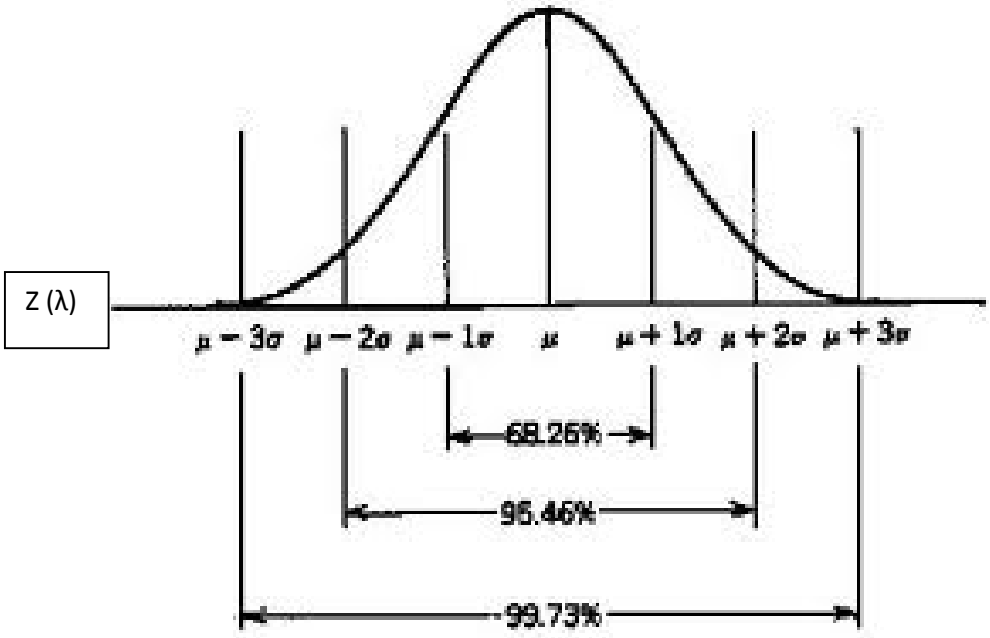
- **Binomial Distribution:** have two outcomes only one or zero. *Its discrete distribution*

- $p(x) = C_x^n\ p^x\ q^{n-x}$; C_x^n is called binomial coefficient. $(0 \le x \le n)$

- $C_0^n=1$ and $C_n^n = 1$ and $0! = 1$ and $(p+q)^n = 1$; p is the parameters and n is the degree of binomial distribution and *n and p is fixed*, *trails independent*, *2 outcomes possible*

- Its application when population is **dichotomized** or divided into 2 classes only
- (p) is the probability of success and (q) is the probability of failure. (p+q)=1
- The **mean** of the binomial distribution *(expected value)* = p(x) = **mean = n p**
- The **variance** of binomial distribution V(x) or σ^2 = **n p q; if n.p.q ≥ 10 we can use normal distribution to approximate binomial**
- **At least to 10 = P(10≤x≤n) =** *in the questions*
- **At most to 10 = P(0≤x≤10) =** *in the questions*
- **At least one will return: 1-p(x=0) in the binomial distribution =** *in the questions*
- **The Poisson distribution:** *discrete distribution*, *trails are independent*, *p is very small*, *n is very large*, *events are very rare.*
- **P(x) =** $\frac{x}{n}$
- **P(x) =** $\frac{e^{-\lambda}\,\lambda^{x}}{x!}$**;** x=0, 1, 2,…….∞. **λ (Aver.)= n.p;** is parameters **(Mean = Variance)**
- **e=2.7183**
- **Normal distribution:** for continuous distribution, *large number of observations*, curve is bell-shaped, symmetrical about the mean, mean=mode=median, *total area under the curve = 1sqr unit* and it approximate the histogram (frequency polygon).

- The mean of all possible sample mean is equal to the population mean, therefore sample mean is called unbiased estimation of population.



- **μ±1SD = 0.6826**
- **μ±2SD = 0.9544**
- **μ±3SD = 0.9973**

Empirical rule=Bell Curved-shaped

- The degree of flatness or peakness of the curve is determined by the value of **σ or SD**
- **Standard Normal Distribution(Z): μ=0, σ^2=1; σ = 1, Z or Z(λ)=** $\frac{X-\mu}{\sigma}$
- **λ** = area under the curve after transformation process. **Z(λ) is point on horizontal line**
- Estimation of discrete sample size = $\mathbf{n = \frac{Z^2 p\, q}{L^2}}$, Z = 1.96 (95% CI) or 2.58 (99% CI) or 3.29 (99.9%CI)
- **L:** is the *permissible error* on either side of the estimate ***(2L is the width of the interval)***
- If the permissible error on either side of the estimate is given in % L is calculate as **($\frac{\#}{100}$ x p)**; do pilot study to estimate p)
- The population proportion of the characteristic is expected to lie in the **interval (p_1-L, p_2+L)**

18/11/2011
Mohsen Mohammed Taqi Mohsen Al-Saleh

- Estimation of continuous sample size = $\mathbf{n = \frac{Z^2 SD^2}{d^2}}$, Z = 1.96 (95% CI) or 2.58 (99%CI) or 3.29 (99.9%CI)
- If the permissible error on either side of the estimate is given in % d is calculate **as** $\left(\frac{\#}{100} \mathbf{x}\ \bar{x}\right)$
- **When 95% of confidence interval:** $\bar{x} \pm 1.96\ (SE(\bar{x}))$ = $\frac{SD}{\sqrt{n}}$
- **When 95% of confidence interval:** $p \pm 1.96\ (SE(p))$ = $\sqrt{\frac{p.q}{n}}$ **(Prevalence rate)**
- **SD^2 = p q, Prevalence rate mean old and new cases together**
- **V(p) =** $\frac{p.q}{n}$ **then it follows that SE(p) =** $\sqrt{\frac{p.q}{n}}$ **for prevalence rate of the population**
- $\mathbf{SE\ (\bar{x}) = \frac{SD}{\sqrt{n}}}$
- **SD:** average amount of deviation of different sample values from the mean value
- **SE:** average amount of deviation of different means (of different samples) from the population mean
- **Average Mean Deviation =** $\frac{\Sigma|x - \bar{x}|}{n}$
- **Positive skew of the curve :** mean > median and the **right** side skewed **(positive)**
- **Geometric mean =** $\sqrt[n]{product\ of\ all\ \%\ values}$ **or =** $\sqrt[n]{\frac{value\ at\ end}{value\ at\ beging}} - 1$
- **Weighted mean =** $\frac{(n1 x \bar{x}1) + (n2\ x\ \bar{x}2)}{n1 + n2}$
- **An experienment:** the observation of some activity or the act of taking some measurement. *(having 3 children) by 3 pregnancies*
- **An outcome:** particular result of an experiment. All the (BBB, BBG…) = 8 outcomes
- **An event:** is the collection (subset) of one or more outcomes. E.g. Boy-Girl-Boy

A, B, C if we want 2 joints

- **Combinations** $(C_r^n) = \frac{n!}{(n-r)!\,r!}$ **- this is used in binomial probability: AB, BC, AC =3**

- **Permutations (P_r^n) =$\frac{n!}{(n-r)!}$; AB, AC, BA, BC, CA, CB = 6**

- **Simple Random Sample:** each unit or item has an equal chance of being selected

- **Sampling error = a sample statistic – population parameter**

- **We reject the null hypothesis, P<0.05 for testing of significance t-distribution**

- **We accept the null hypothesis, P>0.05 for testing of significance t-distribution**

- **P-value = α (5% or 1% or 0.1%) = rejection area= tailed area**

- **V ($\bar{X}_i$) = $\frac{N-n}{N-1}$ x $\frac{\sigma^2}{n}$ = SE($\bar{x}$)**

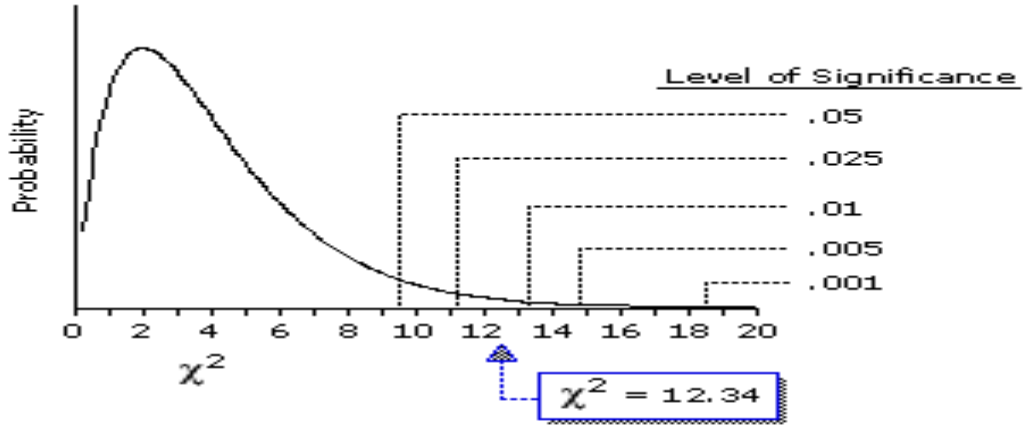
- **Central Limit Theory:** the mean of all possible samples mean is equal to the population mean. Therefore; "sample mean" is called unbiased estimation of "population mean".

- **V(X) = $\frac{N-n}{N-1}\left(\frac{\sigma^2}{n}\right)$ if the population is finite**

- **V(X) =$\left(\frac{\sigma^2}{n}\right)$ if the population is infinite (unlimited) = $(SE)^2$**

- **Chi-Square Test: x^2= $\sum\frac{(O-E)^2}{E}$; (No of column-1) (No of raw-1) =df**

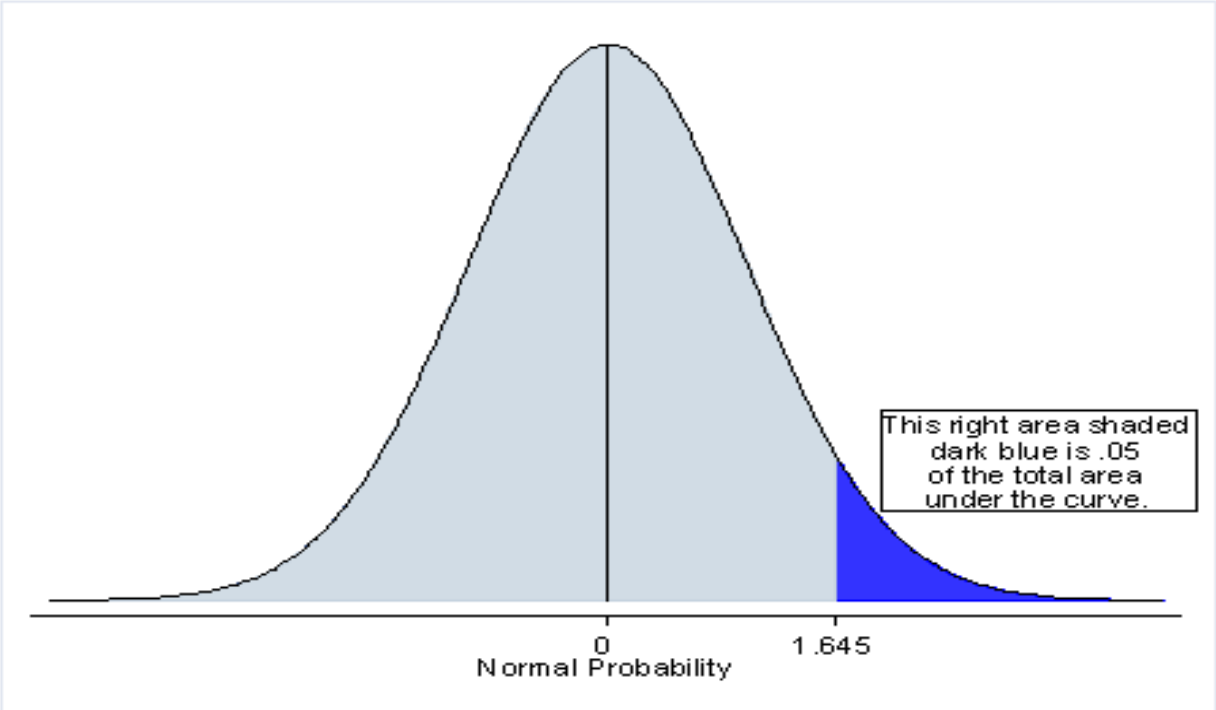
If calculated value is greater than tabular value then there is association



Level of Significance (non-directional test)

df	.05	.025	.010	.005	.001
4	9.49	11.14	13.28	14.86	18.47

critical values of chi-square for **df** = 4



Two-Tailed Versus One-Tailed Hyphothesis Tests

Figure A:
Two-Tailed Test

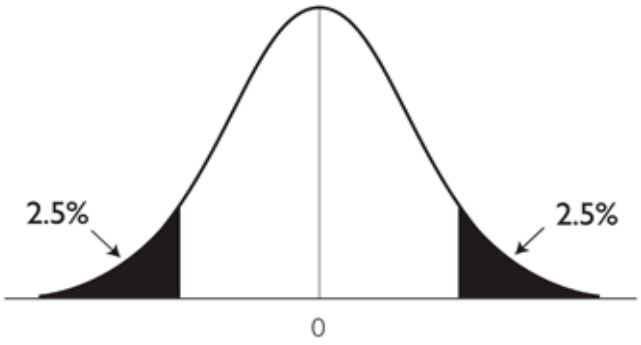
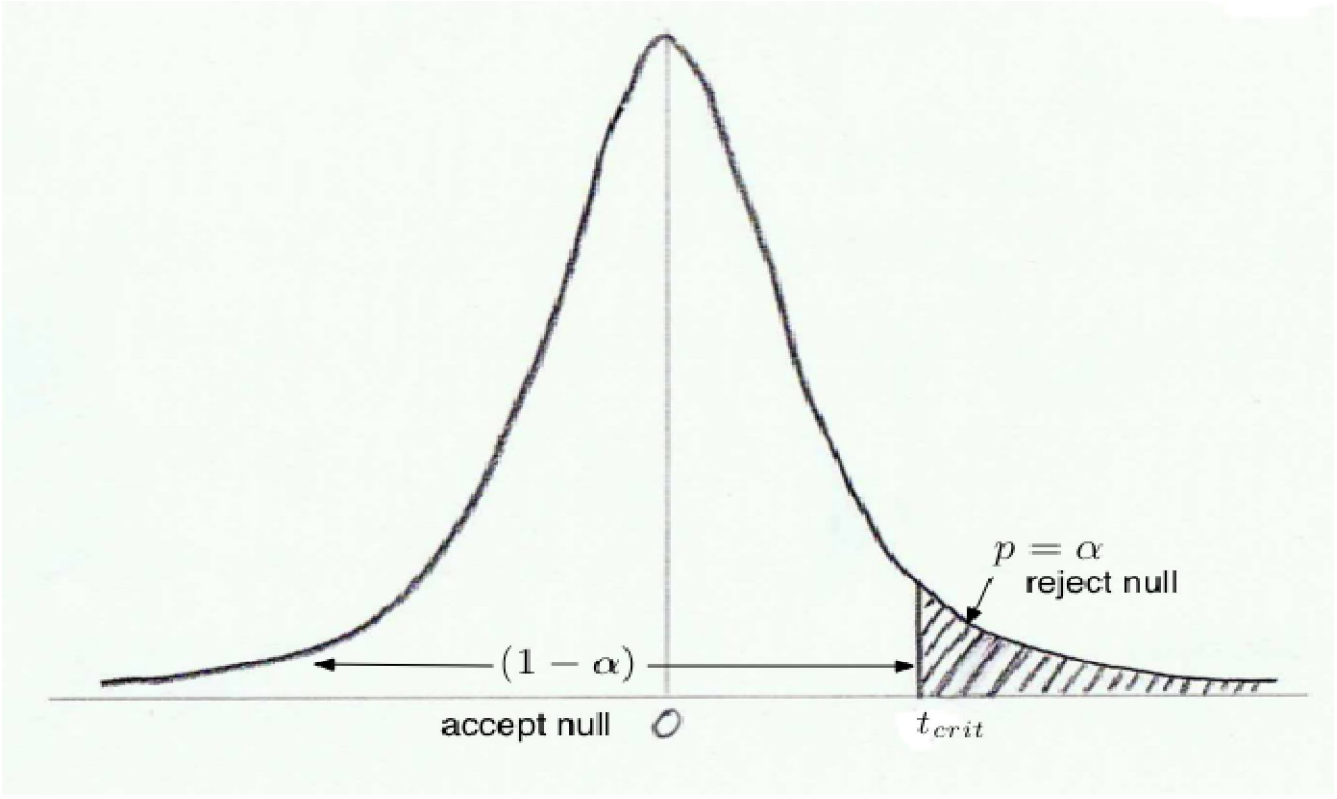
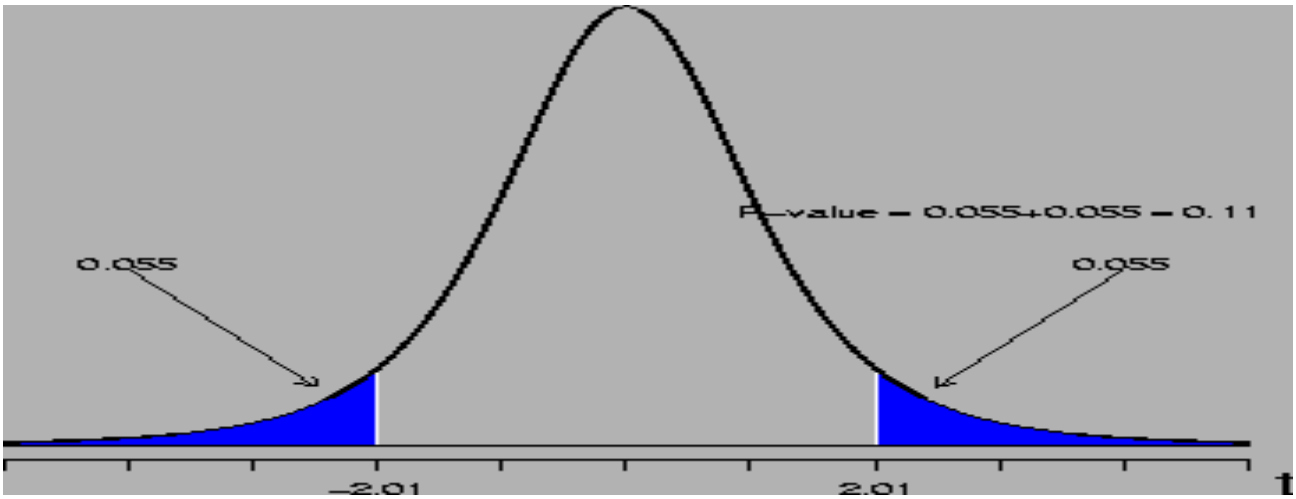


Figure B:
One-Tailed Test
(Left-Tailed Test)

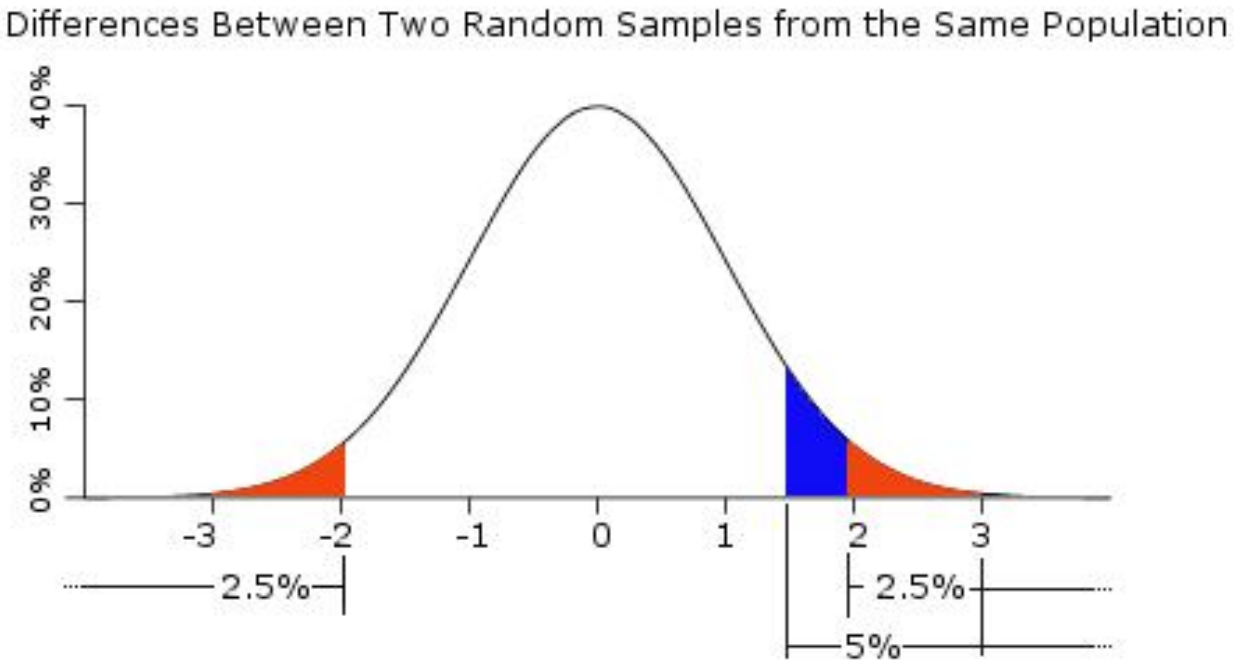
Source: The Heritage Foundation.





- P-value: Presuming H_0 is true, the likelihood of chance variation yielding a *t*-statistic more extreme than -2.01 on *either side* of 0 (since H_1 direction is both high and low) is .11.
- Conclusion: Since P-value > .05, we do not reject H_0.

- **Two-tailed t-test; H_0=0 and $H_1 \neq 0$**



- **One sample test:** Comparison of sample mean with population mean.

Degree of freedom = n-1 for t-test which is distribution of differences

If the calculated value of t > table value we reject the null hypothesis,

H_0: $\mu = \mu_0$ = # (no difference or they are same and equal)-type I error

$H_1 \neq 0$ or $H_1 > 0$ or $H_1 < 0$

$$Z = \frac{|\bar{x} - \mu 0|}{SE(\bar{x})}$$ **; here n<30 where assumption of SD = σ**

$$t = \frac{|\bar{x} - \mu 0|}{SE(\bar{x})}$$ **; here n<30 where SD ≠ σ, even (N) is normally distributed**

- **Unpaired two sample test:** Comparison of two independent sample means.

H_0:$\mu_1 = \mu_2$ = ($\mu_1 - \mu_2$ = Zero) they come from same population, samples are taken from the population

- $z = \frac{|\bar{x}1 - \bar{x}2|}{SE\ (\bar{x}1 - \bar{x}2)}$; n≥30

- $SE(\bar{x}1 - \bar{x}2) = \sqrt{\frac{SD1^2}{n1} + \frac{SD2^2}{n2}}$; n≥30

- $t = \frac{|\bar{x}1 - \bar{x}2|}{SE\ (\bar{x}1 - \bar{x}2)}$ **n<30 ; student t-distribution**

- $SE\ (\mu 1 - \mu 2) = s\sqrt{\frac{1}{n1} + \frac{1}{n2}}$ **; n<30**

- $S = \sqrt{\frac{(n1-1)SD1^2 + (n2-1)SD2^2}{n1+n2-2}}$ **; n<30**

Degree of freedom = (n1-1) + (n2-1) = n1+n2-2

- **Paired sample test:** Comparison of means of two correlated samples. Same subject in both groups. Mean difference for the values is Zero

H_0: μ_d = 0 (the mean of the difference in the population is zero

$$D = \frac{\sum di}{n} \quad \text{and} \quad SD_d = \sqrt{\frac{\sum (di - D)^2}{n-1}}$$

Degree of freedom = n-1

$$t = \frac{|D|}{SE(SDd)}$$

$$SE(SDd) = \frac{SDd}{\sqrt{n}}$$

- **"If (P-value) is low or equal the Null (H_0) must GO (Rejected)"**

- **Inference of proportions: $H_0 : P = P_0$**

$$Z = \frac{|p - P0|}{SE(p)} \text{ and SE (p)} = \sqrt{\frac{P0 \; x \; Q0}{n}} \text{ and } p = \frac{m}{n} \text{ m is prevalence}$$

Where $Q_0 = 1\text{-}P_0$ (remember this is population proportion)

- **(p) is calculate from (n)**

- **Two sample t-test is as follow:**

$$H_0: P_1 = P_2 \; (P_1 - P_2 = \text{Zero})$$

- $z = \frac{|p_A - p_B|}{SE\,(p_A - p_B)}$ **, for 2 sample test of proportion for any (n) sample #**

- $p = \frac{r_1 + r_2}{n_1 + n_2}$ **; weighted average for 2 sample test of proportion for any (n) sample**

- $SE\,(p_A - p_B) = \sqrt{pq\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$ **; for 2 sample test of <u>proportion</u> for any (n) sample #**

- **Correlation of (X,Y): DF= n-2**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Calculated t-value is greater than table t-value then X and Y significantly related to each other

- **Regression: a=is the y-intercept and b=slope**

$$Y = a + bX$$

Percentage of total variation in Y explained by X = 100 $(r)^2$

- **t=** $\frac{r*\sqrt{n-2}}{\sqrt{1-r^2}}$ **if t(calculated) > t(table) then variables (X,Y) related to each other**