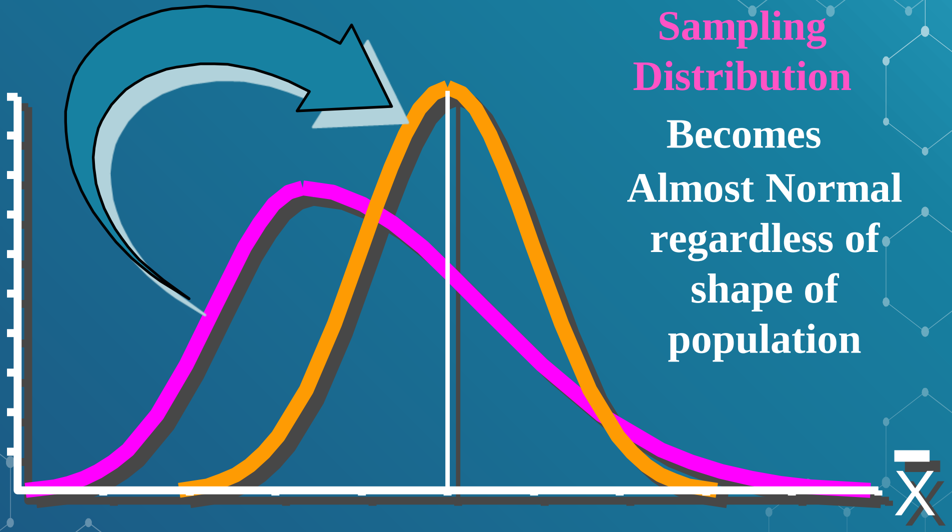
# Hypothesis Testing

# Central Limit Theorem

As **Sample Size** Gets Large Enough




**Sampling Distribution**

Becomes Almost Normal regardless of shape of population

# Central Limit Theorem

For almost all populations, the sample mean is normally or approximately normally distributed, and the mean of this distribution is equal to the mean of the population and the standard deviation of this distribution can be obtained by dividing the population standard deviation by the square root of the sample size

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$because, \; CLT \quad states \quad that$$

$$\mu_{\bar{X}} = \mu \quad and \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

# Central Limit Theorem

- If the original population is normal, a sample of only 1 case is normally distributed
- The further the original sample is from normal, the larger the sample required to approach normality
- Even for samples that are far from normal a modest number of cases will be approximately normal

# When the Population is Normal

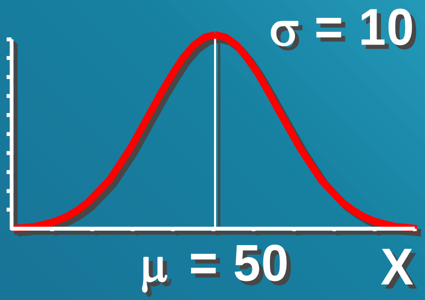
## Central Tendency

$$\mu_{\bar{x}} = \mu$$

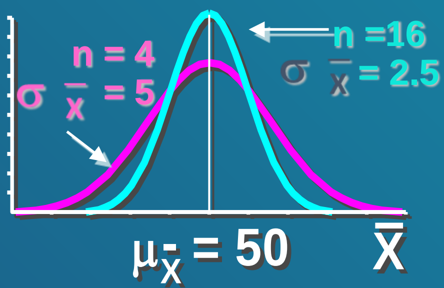
## Variation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Population Distribution




Sampling Distributions



# When The Population is Not Normal

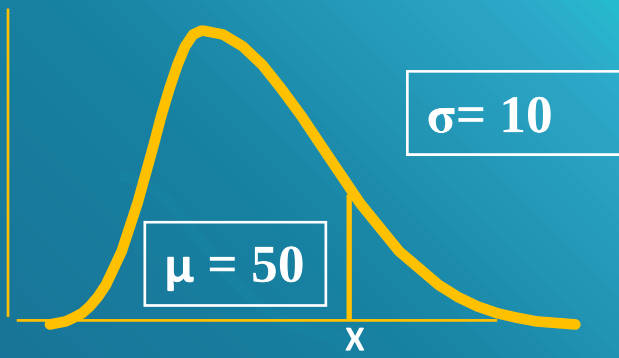
## Central Tendency

$$\mu_{\bar{x}} = \mu$$

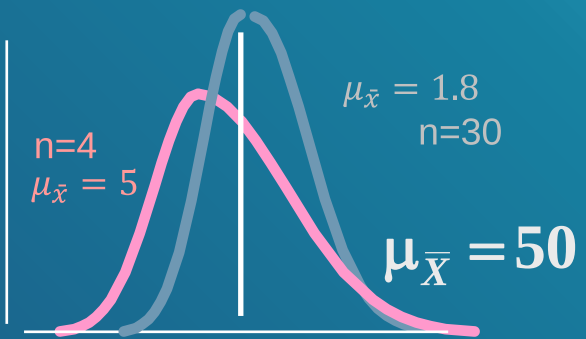
## Variation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

**Population Distribution**



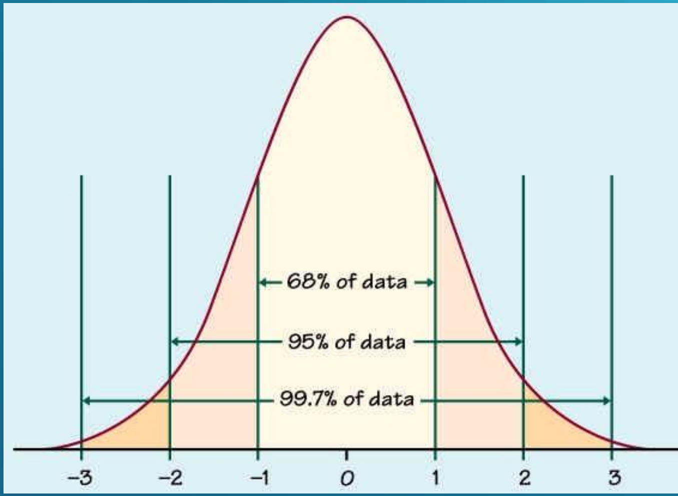

**Sampling Distributions**



# The Normal Distribution

- Along the X axis you see Z scores, i.e. standardized deviations from the mean

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

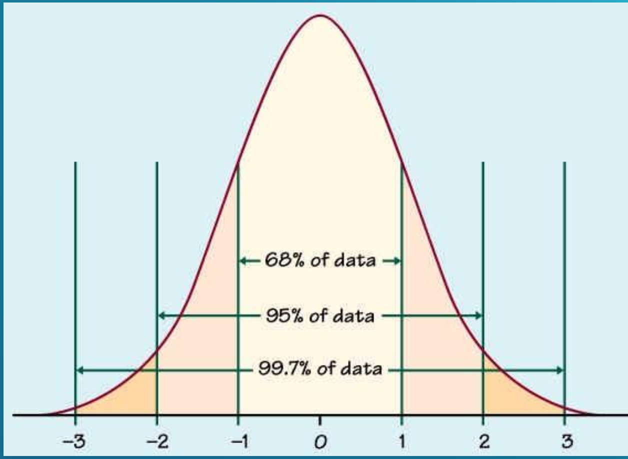



- Just think of Z scores as std. dev. denominated units.

- A Z score tells us how many std. deviations a case lies above or below the mean

# The Normal Distribution

- Note a property of the Normal distribution
- 68% of cases in a Normal distribution fall within 1 std. deviation of the mean
- 95% within 2 std. dev. (actually 1.96)
- 99.7% within 3 std. dev.



# Confidence Intervals

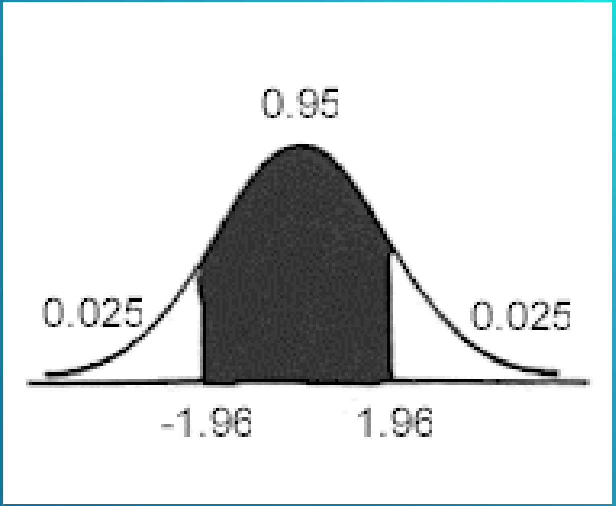
- We can use the Central Limit Theorem and the properties of the normal distribution to construct confidence intervals of the form:
  - The average salary is $40,000 plus or minus $1,000 with 95% confidence
  - Presidential support is 45% plus or minus 4% with 95% confidence.
- In other words, we can make our best estimate using a sample and indicate a range of likely values for what we wish to estimate

# Confidence Intervals

- Notice that our estimates of the population parameter are probabilistic.

- So we report our sample statistic with together with a measure of our (un)certainty

- Most often, this takes the form of a 95 percent **confidence interval** establishing a boundary around the sample mean (x bar) which will contain the true population mean (μ) 95 out of 100 times.




# Introduction to Hypothesis Testing

- Hypothesis testing is one of the most important concepts in Statistics
- Heavily used by **Statisticians**, **Machine Learning Engineers**, and **Data Scientists**
- Statistical tests are used to check whether the **null hypothesis** is rejected or not rejected.
- Statistical tests assume a null hypothesis **of no relationship or no difference between groups.**

- Definition:

**A hypothesis is defined as a formal statement, which gives the explanation about the relationship between the two or more variables of the specified population**

Example based on a sample data  we may wish to decide whether a serum is really effective in curing Corona

# Types of Hypothesis

- Simple
- Complex
- Null
- Alternative
- Empirical
- Statistical

# What is test of Hypothesis?

- If on the supposition that a particular hypothesis is true we find that results observed in a random sample differ markedly from those expected, we say that observed differences are significant and we reject the hypothesis
- Procedures that enable us to decide to accept or reject hypothesis are called **test of hypothesis, test of significance, decision rules**

# Type I and Type II Errors

- **Type I error**:- Rejecting a hypothesis when it happens to be true
- **Type II error**:- Accepting a hypothesis when it is to be rejected
- These errors have to be minimized but the decrease in one causes the increase in the other
- The best solution is to increase the sample size

# Type I, Type II Errors

| | Null hypothesis is TRUE | Null hypothesis is FALSE |
|---|---|---|
| Reject null hypothesis | Type I Error (False positive) | Correct outcome! (True positive) |
| Fail to reject null hypothesis | Correct outcome! (True negative) | Type II Error (False negative) |

| | | Reality | |
|---|---|---|---|
| | | Positive | Negative |
| Study Finding | Positive | **True Positive**<br>(Power)<br>($1-\beta$) | False Positive<br>**Type I Error**<br>($\alpha$) |
| | Negative | False Negative<br>**Type II Error**<br>($\beta$) | **True Negative** |

# Characteristics of Hypothesis

The important characteristics of the hypothesis are:

- The hypothesis should be short and precise
- It should be specific
- A hypothesis must be related to the existing body of knowledge
- It should be capable of verification

# Statistical Hypothesis

- It is a guess or assumption about the parameters of population distribution.
- It is established beforehand and may or may not be true
- Statistical Hypothesis can be either
  - Null Hypothesis
  - Alternative Hypothesis

# Null Hypothesis($H_0$)

- It is a statistical hypothesis which is to be actually tested for acceptance or rejection
- It is the hypothesis which is tested for possible rejection under the assumption that it is true
- It is expressed in the form of equality
- Example:- Independent variables have no effect on the dependent variables.

# Examples of Null Hypothesis

- Null hypothesis is always a simple hypothesis stated as an equality specifying an exact value of the parameter
- Examples
  - Population mean equals to a specified constant $\mu_0$
  - The difference between the sample means equals to a constant

# Alternate Hypothesis($H_1$)

- It is any other hypothesis other than null hypothesis
- It is expressed in the form of >.<.not =
- We can accept alternative hypothesis if there is sufficient evidence
- This was originated by Neyman
- Example: Independent events or variables have effect on dependent variables
- H1 : $\mu > \mu_0$

# Critical Region

- In any test of hypothesis, a test statistic S*, calculated from the sample data, is used to accept or reject the null hypothesis of the test
- The area under the probability curve of the sampling distribution of the test statistic S* which follows some known given distributions
- This area under probability curve is divided into two regions, **region of rejection** where null hypothesis is rejected and **region of acceptance**

- The **critical region** is the region of rejection of null hypothesis
- The area of the critical region equals to the level of significance $\alpha$
- Critical Region lies on the tail(s) of the distribution
- Depending upon the nature of alternate Hypothesis, critical region may lie on one side or both sides of the tail(s)

# Test of Significance

- This is the procedure to decide whether to accept or reject the null hypothesis
- This test is used to determine whether observed samples differ significantly from expected results
- **Acceptance** of hypothesis merely indicates that the data do not give sufficient evident to reject the hypothesis

- However **rejection** of hypothesis is a firm conclusion that the sample evidence rejects it
- **When null hypothesis is accepted the result is said to be non-significant** which means the observed differences are due to chance caused by the process of sampling
- **When null hypothesis is rejected which means the alternate hypothesis is accepted the result is said to be significant**
- Since the test is based on sample observation, the decision of acceptance or rejection of null hypothesis is subject to some error or risk

# Level of Significance

- Represented by $\alpha$
- This is the probability of committing the Type I error
- It measure the amount of risks associated in taking decisions
- This factor has to be chosen before sample information is collected
- It is either 0.01 or 0.05

# How to compute the level of significance?

- To measure the level of statistical significance of the result, the investigator first needs to calculate the **p-value**
- It defines the probability of identifying an effect which provides that the null hypothesis is true

**When the p-value is less than the level of significance (α), the null hypothesis is rejected.**

# Interpretation of p-value based on level of significance(10%)

- If p > 0.1, then there will be no assumption for the null hypothesis
- If p > 0.05 and p ≤ 0.1, it means that there will be a low assumption for the null hypothesis.
- If p > 0.01 and p ≤ 0.05, then there must be a strong assumption about the null hypothesis.
- If p ≤ 0.01, then a very strong assumption about the null hypothesis is indicated.

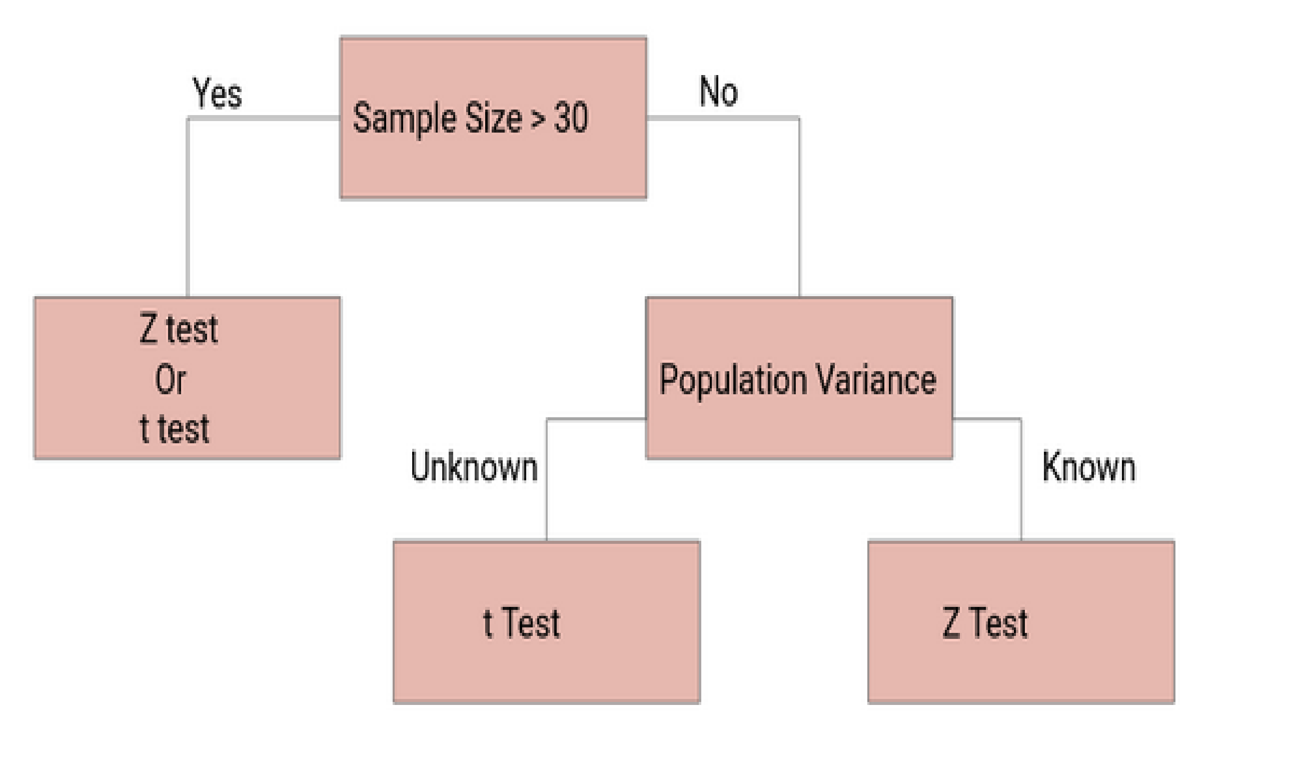
# Rejection rule of null hypothesis

- If $p < \alpha$, then one must reject the null hypothesis
- If $p > \alpha$, then one should not reject the null hypothesis

# Common Tests

| Z Test | T Test |
| --- | --- |
| z tests are a statistical way of testing a hypothesis when either:<br>1. We know the population variance or<br>2. We do not know the population variance but our sample size is large $n \geq 30$ | t-tests are a statistical way of testing a hypothesis when:<br>1. We do not know the population variance<br>2. Our sample size is small, $n < 30$ |

# Common Tests



# T-test

- The t-test is a basic test that is limited to two groups.
- For multiple groups, you would have to compare each pair of groups, for example with three groups there would be three tests (AB,AC, BC)
- The basic principle is to test the null hypothesis that the means of the two groups are equal.

# T-test

- The t-test assumes:
  - A normal distribution (parametric data)
  - Underlying variances are equal (if not, use Welch's test)
- It is used when there is random assignment and only two sets of measurement to compare.
- There are two main types of t-test:
  - Independent-measures t-test: when samples are not matched.
  - Matched-pair t-test: When samples appear in pairs (eg. before-and-after).
- A single-sample t-test compares a sample against a known figure, for example where measures of a manufactured item are compared against the required standard.

# T-test Applications

- To compare the mean of a sample with population mean.
- To compare the mean of one sample with the mean of another independent sample.
- To compare between the values (readings) of one sample but in 2 occasions.

# One Sample Test
## (Sample mean and population mean )

- Ho: Sample mean=Population mean.
- Degrees of freedom = n – 1

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

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

# Sample mean and population mean

**Example:** The following data represents hemoglobin values in gm/dl for 10 patients:
10.5 9 6.5 8 11 7 8.5 9.5 12 7.5
Is the mean value for patients significantly differ from the mean value of general population (12 gm/dl)?

$$t = \frac{8.95 - 12}{\frac{1.80201}{\sqrt{10}}} = -5.352$$

Df=9

Find tabulated value for 9df and % 0.05 level of significance= 2.262

Calculated value> tabulated value

Reject Ho.

There is a statistically significant difference between the mean of sample and population mean, and this difference is unlikely due to chance.

# T Table

Calculating p-value from t-value

| 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 |
| 19 | 0.000 | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 0.000 | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |
| 21 | 0.000 | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | 0.000 | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | 0.000 | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 | 3.768 |
| 24 | 0.000 | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |
| 25 | 0.000 | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 3.725 |
| 26 | 0.000 | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 | 3.707 |
| 27 | 0.000 | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 | 3.690 |
| 28 | 0.000 | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 | 3.674 |
| 29 | 0.000 | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 | 3.659 |
| 30 | 0.000 | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 | 3.646 |
| 40 | 0.000 | 0.681 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 | 3.551 |
| 60 | 0.000 | 0.679 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 | 3.460 |
| 80 | 0.000 | 0.678 | 0.846 | 1.043 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 3.195 | 3.416 |
| 100 | 0.000 | 0.677 | 0.845 | 1.042 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 | 3.390 |
| 1000 | 0.000 | 0.675 | 0.842 | 1.037 | 1.282 | 1.646 | 1.962 | 2.330 | 2.581 | 3.098 | 3.300 |
| **z** | 0.000 | 0.674 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 | 3.291 |
| | 0% | 50% | 60% | 70% | 80% | 90% | 95% | 98% | 99% | 99.8% | 99.9% |
| | Confidence Level | | | | | | | | | | |

# Example in R

***t.test(data$V1,mu=12)***

```
        One Sample t-test
data:  data$V1
t = -5.5678, df = 9, p-value = 0.0003484
alternative hypothesis: true mean is not equal to 12
95 percent confidence interval:
 7.640484 10.159516
sample estimates:
mean of x
     8.9
```

If p-value is less than 0.05
Reject Ho

# Confidence Interval

= population mean ± calculated t-value $(s/\sqrt{n})$

==8.9± 2.26$(s/\sqrt{n})$

=8.9 ± 2.26(1.8/ $\sqrt{10}$)

=8.9 ± 2.26(0.56)

=8.9 ± 1.28

10.1  7.62

# Two Sample Tests (Mean of Two samples)

- Ho:
  Mean of sample 1 = Mean of sample 2
- Degrees of freedom = n1+n2 – 2

# Two Sample Tests
# (Mean of Two samples)

The following data represents weight in Kg for 10 males and 12 females.

Males: 80 75 95 55 60 70 75 72 80 65

Females: 60 70 50 85 45 60 80 65 70 62 77 82

Is there a statistically significant difference between the mean weight of males and females. Let alpha = 0.01?

# Two Sample Tests (Mean of Two samples)

$$t = \frac{(x'_1 - x'_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Mean1=72.7
Variance1=128.46
Df = n1+n2-2=20
t = 1.074

Mean2=67.17
Variance2=157.787

The tabulated t, 2 sides, for alpha 0.01 is 2.845
Then accept Ho and conclude that there is no significant difference between the 2 means.
This difference may be due to chance. P>0.01

# Two Sample Test (Mean of Two samples)

```
>data1<-c(80,75, 95, 55, 60, 70, 75, 72, 80, 65)
>data3<-c(60, 70, 50, 85, 45, 60, 80, 65, 70, 62, 77, 82)
> t.test(data1,data3)

        Welch Two Sample t-test

data:  data1 and data3
t = 1.0853, df = 19.844, p-value = 0.2908
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.107295 16.173961
sample estimates:
mean of x mean of y
 72.70000  67.16667
```

# Diamonds -Two Sample Tests in R

```
t.test(diamonds2$x,diamonds2$y)

        Welch Two Sample t-test

data:  diamonds2$x and diamonds2$y
t = -0.28451, df = 107854, p-value = 0.776
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01524977  0.01138372
sample estimates:
mean of x mean of y
 5.731868  5.733801
```

# Pisa Score -Two Sample Tests in R

```
t.test(df$Maths.F,df$Maths.M)

        Welch Two Sample t-test

data:  df$Maths.F and df$Maths.M
t = -0.45289, df = 133.84, p-value = 0.6514
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.31807  14.00156
sample estimates:
mean of x mean of y
 458.1345  462.2928
```

# Paired Two Sample Test

## One sample in two occasions

Blood pressure of 8 patients, before & after treatment

| BP before | BP after | d | d2 |
|---|---|---|---|
| 180 | 140 | 40 | 1600 |
| 200 | 145 | 55 | 3025 |
| 230 | 150 | 80 | 6400 |
| 240 | 155 | 85 | 7225 |
| 170 | 120 | 50 | 2500 |
| 190 | 130 | 60 | 3600 |
| 200 | 140 | 60 | 3600 |
| 165 | 130 | 35 | 1225 |
| | | 465 | 29175 |

Mean d=465/8=58

# Paired Two Sample Test
## One sample in two occasions

- The df here = n – 1=7

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

=9.38

Tabulated t (df7), with level of significance 0.05, two tails, = 2.36

We reject Ho and conclude that there is significant difference between BP readings before and after treatment.

$P<0.05$.

# Paired Two Sample Test
## One sample in two occasions

```
t.test(data$V1,data$V2,paired = TRUE)

        Paired t-test
data:  data$V1 and data$V2
t = 9.3876, df = 7, p-value = 3.24e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 43.48397 72.76603
sample estimates:
mean of the differences
            58.125
```

**Example 14·2.** *A machinist is making engine parts with axle diameters of 0·700 inch. A random sample of 10 parts shows a mean diameter of 0·742 inch with a standard deviation of 0·040 inch. Compute the statistic you would use to test whether the work is meeting the specifications. Also state how you would proceed further.*

$\mu = 0{\cdot}700$ inches, $\bar{x} = 0{\cdot}742$ inches, $s = 0{\cdot}040$ inches and $n = 10$

*Null Hypothesis,* $H_0 : \mu = 0{\cdot}700$, *i.e.. the product is conforming to specifications.*

***Alternative Hypothesis,*** $H_1 : \mu \neq 0{\cdot}700$

- t=3.15

# Z test

Suppose we randomly sampled subjects from an honors program. We want to determine whether their mean IQ score differs from the general population. The general population's IQ scores are defined as having a mean of 100 and a standard deviation of 15.

Null (H0): $\mu = 100$

Alternative (HA): $\mu \neq 100$

IQ score sample mean ($\bar{x}$): 107

Sample size ($n$): 25

Hypothesized population mean ($\mu_0$): 100

Population standard deviation ($\sigma$): 15

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{107 - 100}{\frac{15}{\sqrt{25}}} = \frac{7}{3} = 2.333$$

- 2.333 is greater than the critical value of 1.960
- We can reject the null and conclude that the mean IQ score for the population of honors students does not equal 100. Based on the sample mean of 107, we know their mean IQ score is higher.

# Z test

| Significance Level | Type of Test | Critical Value(s) |
|---|---|---|
| 0.01 | Two-Tailed | ±2.576 |
| 0.01 | Left Tail | –2.326 |
| 0.01 | Right Tail | +2.326 |
| 0.05 | Two-Tailed | ±1.960 |
| 0.05 | Left Tail | +1.650 |
| 0.05 | Right Tail | –1.650 |

Suppose a teacher claims that his section's students will score higher than his colleague's section. The mean score is 22.1 for 60 students belonging to his section with a standard deviation of 4.8. For his colleague's section, the mean score is 18.8 for 40 students and the standard deviation is 8.1. Test his claim at $\alpha = 0.05$.

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

right-tailed two-sample z test.
H0: $\mu 1=\mu 2$, H1: $\mu 1>\mu 2$
$\mu 1-\mu 2=0$
Thus, $z = 2.32$
As $2.32 > 1.645$ thus, the null hypothesis can be rejected.

There is enough evidence to support the teacher's claim that the scores of students are better in his class.