Section 2.6 Statistics (reference 2.6)

Definitions:

Population: The set of all possible observations

Sample: Any subset of a population

Homogeneous Sample: The sample comes from 1 population only

Random Sample: Equal probability of selecting any member of the population Independence (of events A and B): P(A and B) = P(A)*P(B)

Sample and Population Mean (Average value): $\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i$

Mode (Most commonly occurring value in a sample)

Median (middle value, 50th percentile. Half of the sample values are greater and half are smaller)

Deviation (from the mean value): $d_i = x_i - \bar{x}$

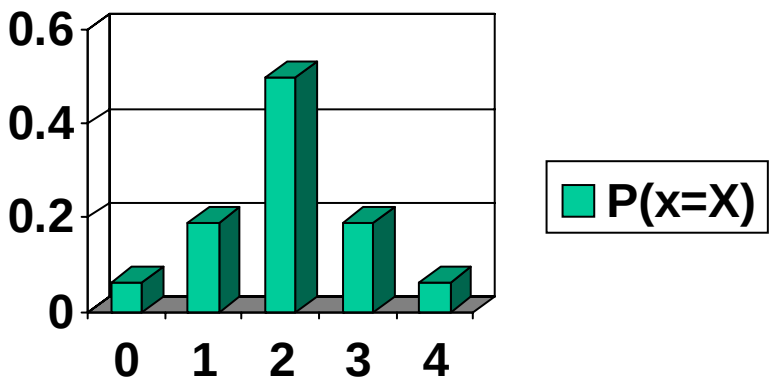
Population Variance (from the mean value): $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} d_i^2$

Population Standard Deviation (from the mean value): $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}$

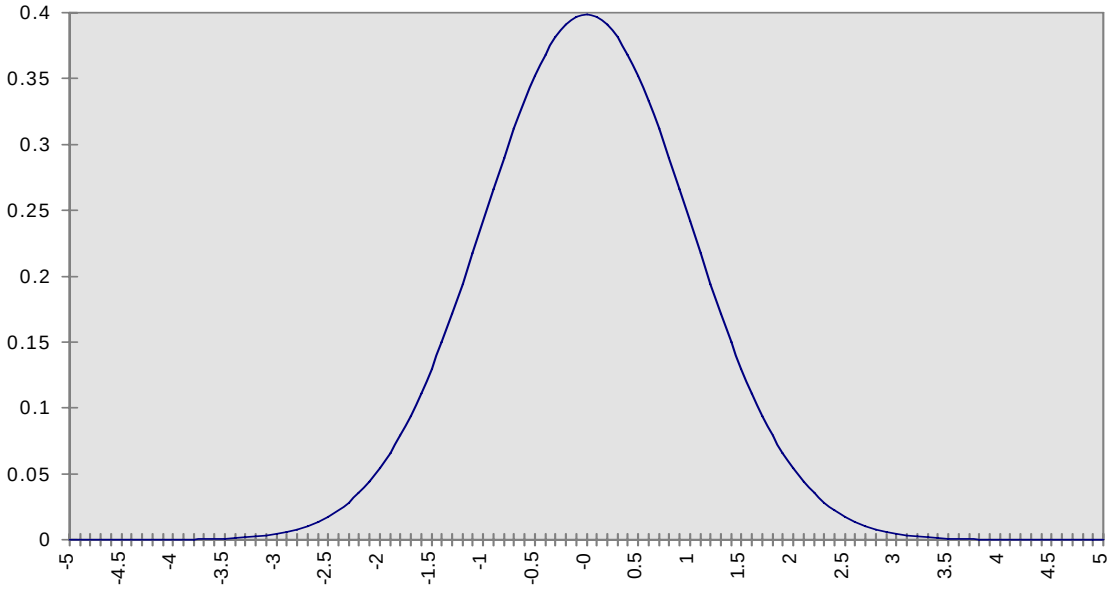
Sample Standard Deviation (from the mean value): $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} d_i^2}$

Discrete Probability Distributions:

Binomial: N independent events, each having probability *p* of success, and1-*p* of failure. For example, tossing a fair coin N times where *p* = the probability of getting a head on any toss. If the random variable x indicates the number of heads in N=2 tosses, then P(x=0) = 1/4, P(x=1) = 1/2, P(x=2) = 1/4. If N=4, then the probabilities are illus trated in the following graph:



As N approaches infinity ...



So**,** the binomial distribution is the discrete case of the Normal distribution.

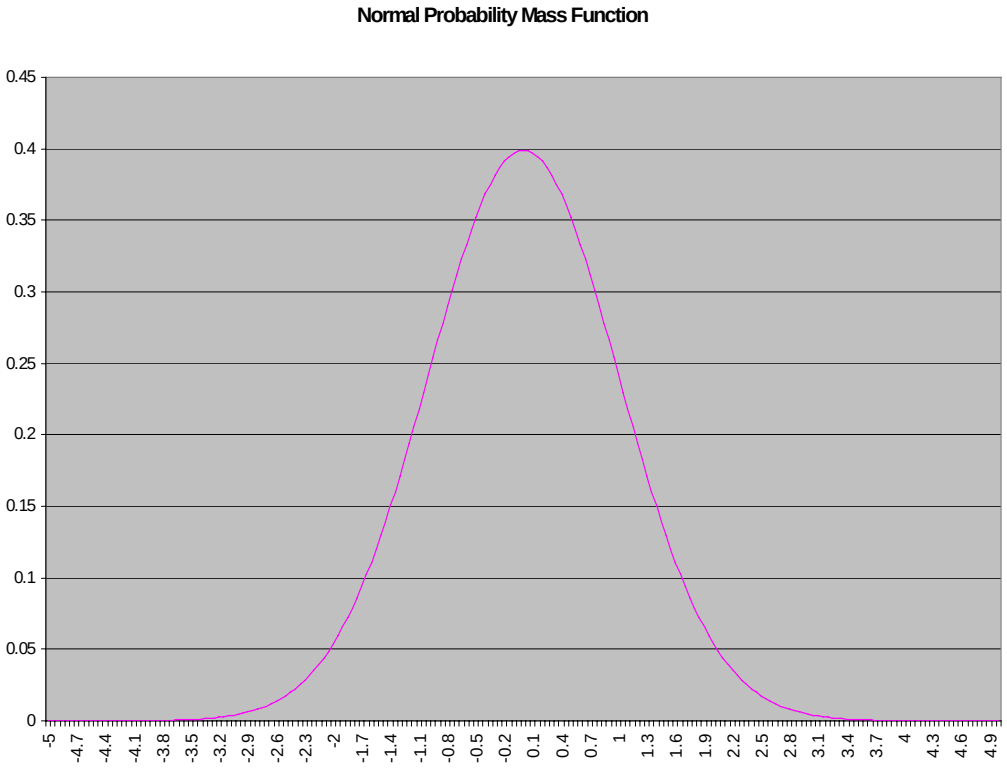
Continuous Distributions: As the number of samples increases and the width of the Discrete sample intervals shrink to zero, discrete distributions become continuous.

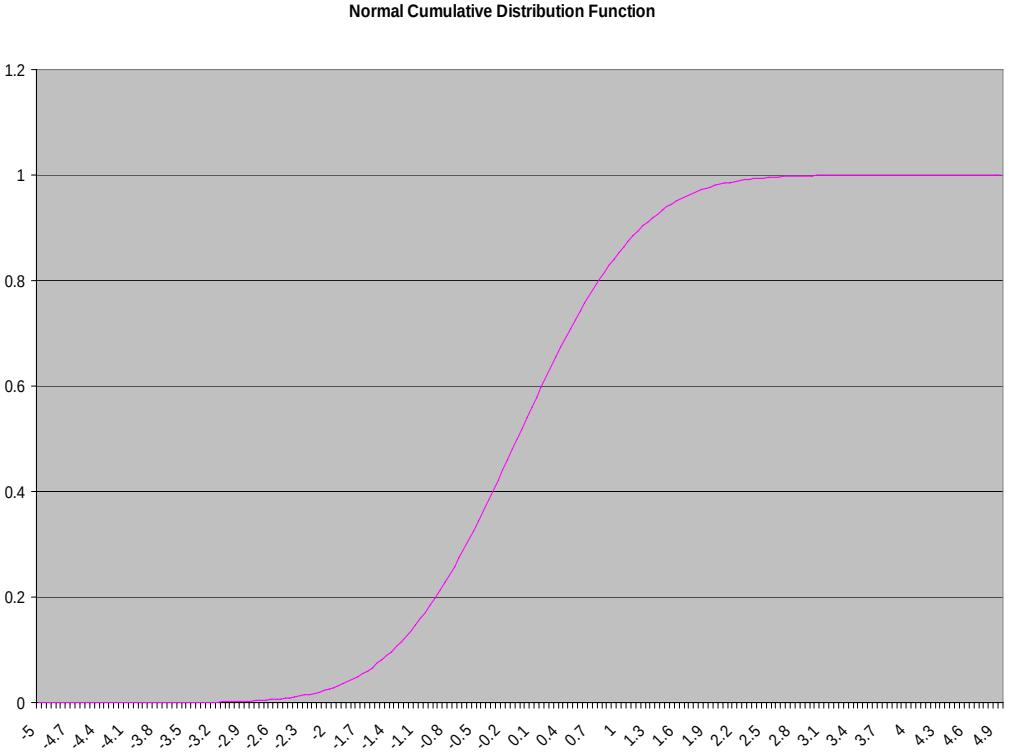
P(x=X) = 0
Must talk about intervals, e.g. P(a < x < b)

The Normal Distribution:

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

Normal Distribution:





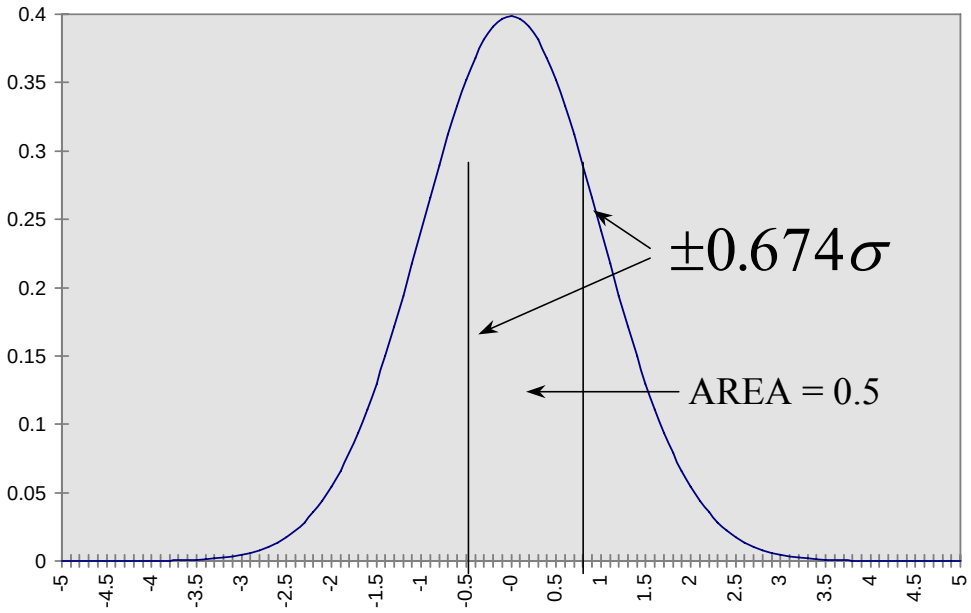
The Standard Normal Distribution:

$$\mu = 0, \sigma = 1$$

$$z = \frac{x - \mu}{\sigma}, dz = \frac{1}{\sigma} dx$$

$$P(a < z < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

Error Probable: An error budget that would contain half of the population data points. Assumes that events are independent and identically distributed (iid). **Also assumes N is large (greater than 30), or population is normally distributed.**



Circular Error Probable – the 2 Dimensional Case (X error and Y error):

If $\sigma_x < \sigma_y$ and $\sigma_x / \sigma_y \leq 0.28$ then CEP = $0.562\sigma_x + 0.615\sigma_y$

If $\sigma_x > \sigma_y$ and $\sigma_y / \sigma_x \leq 0.28$ then CEP = $0.615\sigma_x + 0.562\sigma_y$

Otherwise CEP = $0.5887(\sigma_x + \sigma_y)$

Confidence Intervals: In practice, we take a sample from population. The sample mean and variance will differ from the population mean and variance. Confidence Intervals express how certain we are that the population statistics lie in a region around the sample statistics.

Central Limit Theorem: Given a population Normally distributed, (μ, σ^2)

then the distribution of successive sample means from samples of n observations

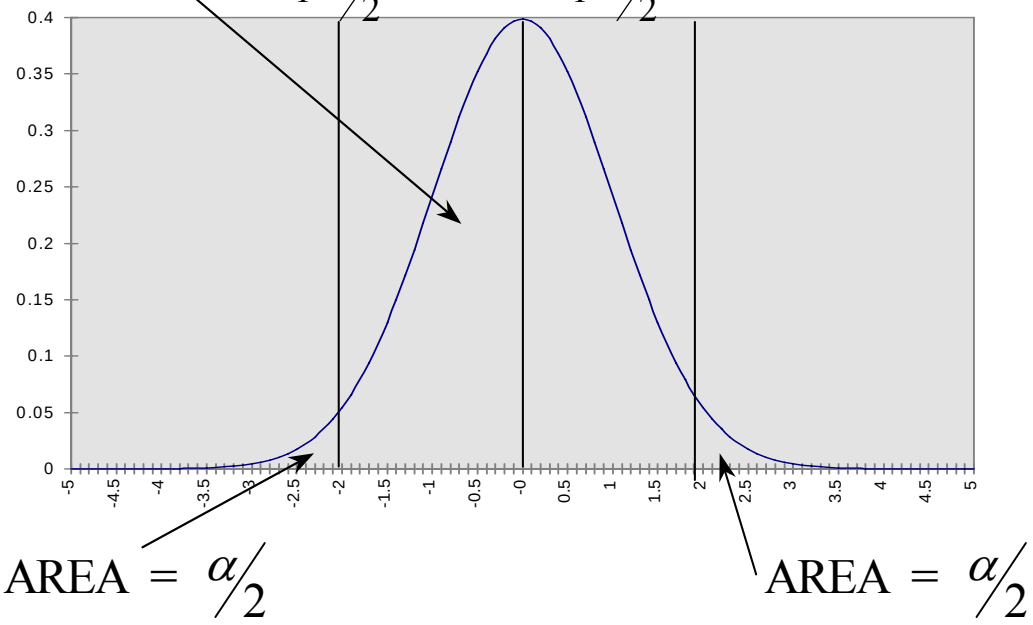
Approaches a Normal distribution with parameters $(\mu, \sigma^2/\sqrt{n})$

We want 1- α level of confidence that a region around our sample mean

value contains the actual population mean.

$$\text{AREA} = 1-\alpha$$

$$P(-z_{1-\alpha/2} < x < z_{1-\alpha/2}) = 1-\alpha$$



$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}}$$

$$P(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}) = 1-\alpha$$

If n < 30, we must use Student's T Distribution instead of the Standard Normal

$$P(\bar{x} - t_{n,1-\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n,1-\alpha/2}\frac{s}{\sqrt{n}}) = 1-\alpha$$

Determining Sample Size: For the population mean to fall into an interval defined by

$$(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}) < \mu < (\bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}})$$

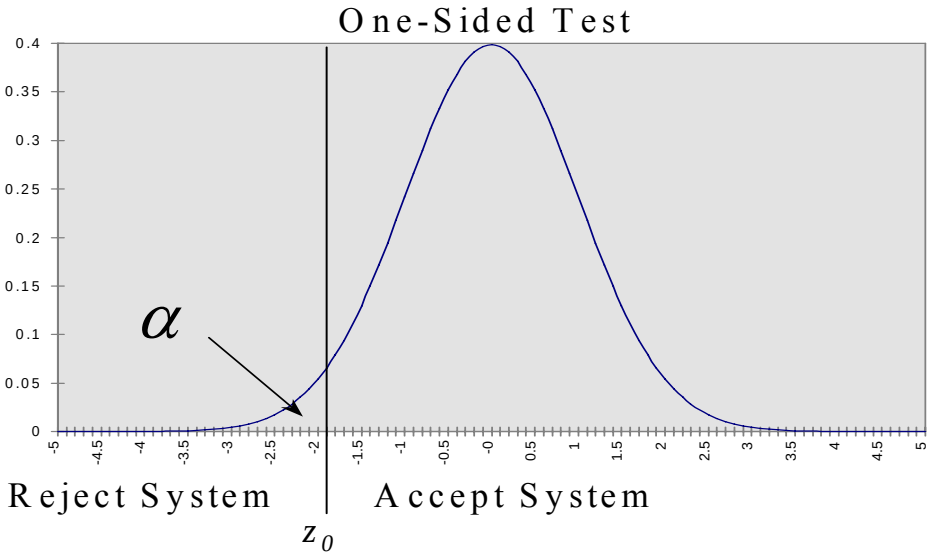
$$\left|\mu - \bar{x}\right| \le z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

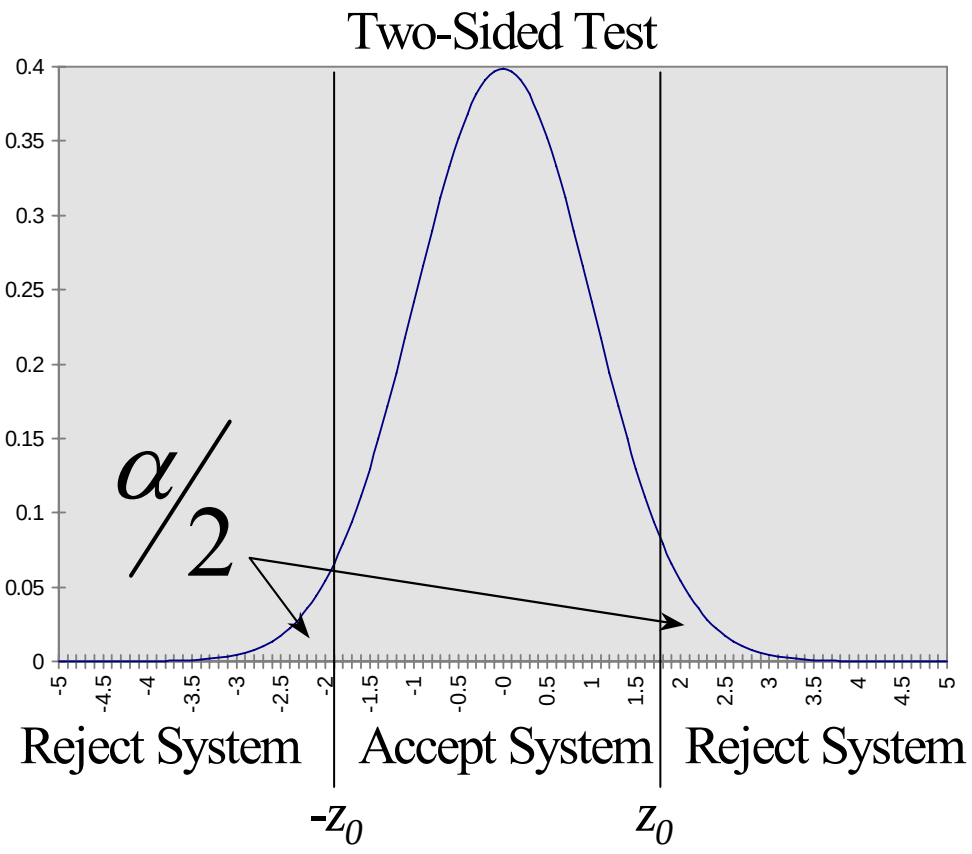
Where $\left|\mu - \bar{x}\right|$ is the accuracy desired (or the error that can be tolerated).

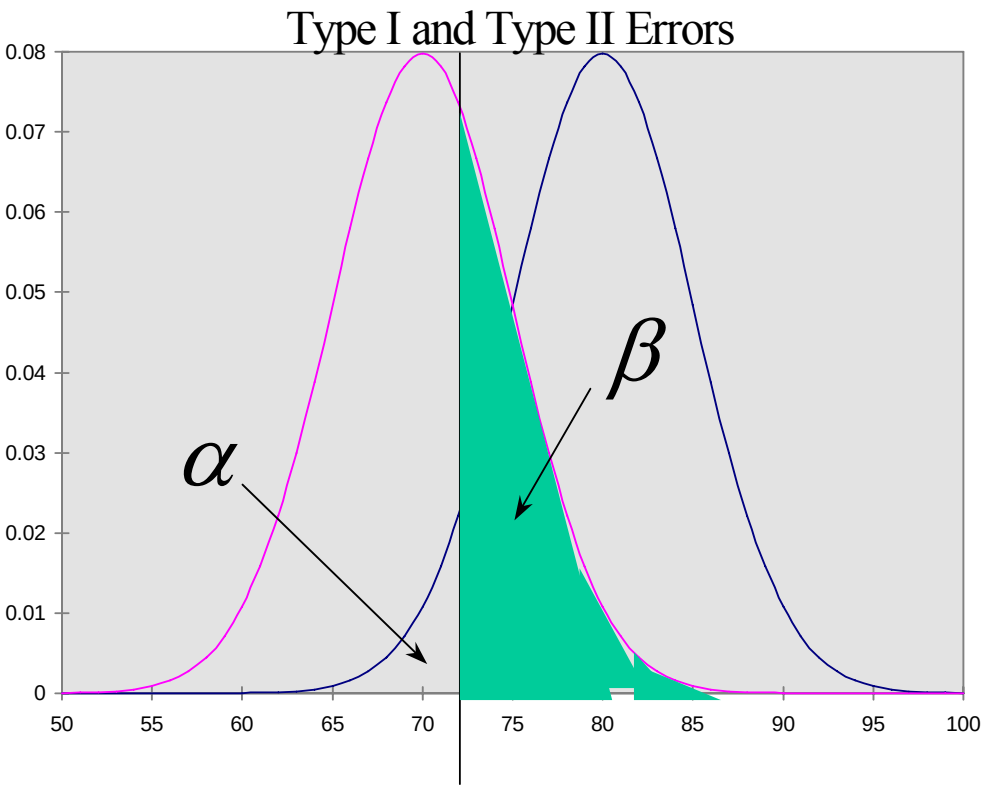
Since the sample size decision must be made prior to the test, an estimate must be made for the population standard deviation. Using the estimate we can solve for N

$$N \ge \left|\frac{z_{1-\alpha/2}\sigma}{error}\right|^2$$

Hypothesis Testing: Begins with an assumption (hypothesis), usually about the underlying population distribution of some measured quantity or computed error. Select values for the hypothesis and alternate hypothesis(es) that partition the sample space. Collect N samples of the population test statistic or parameter. There are two types of errors: Type 1 errors reject the hypothesis when it is true; Type II accept the hypothesis when in is false.




Two-Sided Test
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
α/2
-5 -4.5 -4 -3.5 -3 -2.5 -2 -1.5 -1 -0.5 -0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
Reject System
Accept System
Reject System
-z0
z0


Type I and Type II Errors
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
α
β
50 55 60 65 70 75 80 85 90 95 100

Large Samples, Unknown Variance use $s = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}}$ for σ

$$z' = \frac{\bar{x} - \mu'}{\sigma / \sqrt{n}}$$

$$z' = z + \frac{(\mu - \mu')}{\sigma / \sqrt{n}}$$

Small Samples, Unknown Variance use: $t = \frac{\bar{x} - \mu}{s / \sqrt{n-1}}$

$$t' = \frac{\bar{x} - \mu'}{s / \sqrt{n-1}}$$

$$t' = t + \frac{\mu - \mu'}{s / \sqrt{n-1}}$$

Adjusting α and β

Adjust the size of the Error we wish to Detect Change the sample size *n*

$H_0 : T_j = 0, \forall j$

NormalEquations:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{k} m + \sum_{i=1}^{n}\sum_{j=1}^{k} t_j = nkm + n\sum_{j=1}^{n} t_j, \text{ but } \sum_{j=1}^{k} t_j = 0$$

$$\text{so } \sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij} = nkm$$

$$\sum_{i=1}^{n} X_{ij} = \sum_{i=1}^{n} m + \sum_{i=1}^{n} t_j = nm + nt_j$$

m is the least squares estimate of
t_j is the least squares estimate of T_j

$$SS_r(m, t_j) = m\sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij} + \sum_{j=1}^{n} t_j \sum_{i=1}^{n} X_{ij}$$

Assuming H_0 is True, the model is :

$$X_{ij} = \mu + \varepsilon_{ij}$$

$$SS_r(m') = m'\sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}$$

Between Treatments : $SS_r(m, t_j) - SS_r(m')$

$$SS_e = \sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}^2 - SS_r(m, t_j)$$

Test Statistic is : $F_{k-1,(n-1)k} = \dfrac{SS_t / (k-1)}{SS_e / ((n-1)k)}$