DESIGN OF EXPERIMENTS

1.0 STATISTICS REVIEW

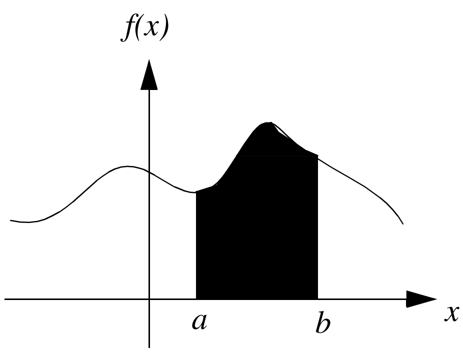
(updated Spring 2005)

Random Variable

- Discrete random variable: Number of "up spots" on a throw die; Exam score; etc.
- Continuous random variable: Time between car arrivals at a spot light (11.3, 51.2 etc.); Diameter of a machined part.

Continuous Probability Distributions

Let x be a continuous random variable characterized by f(x), which is called a **Probability Density Function**. It describes how the random variable arises in a frequency sense.

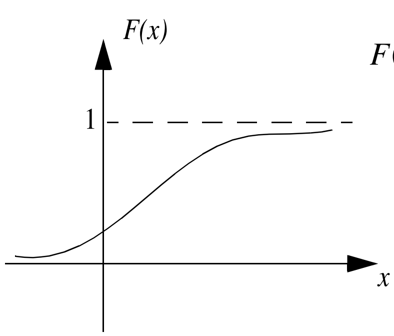


- $f(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- Probability of ($a \leq X \leq b$)

  $= \int_a^b f(x)dx$ = Shaded Area

Definition of **Probability Density Function**

A random variable can also be characterized by the **Cumulative Distribution Function (CDF).**



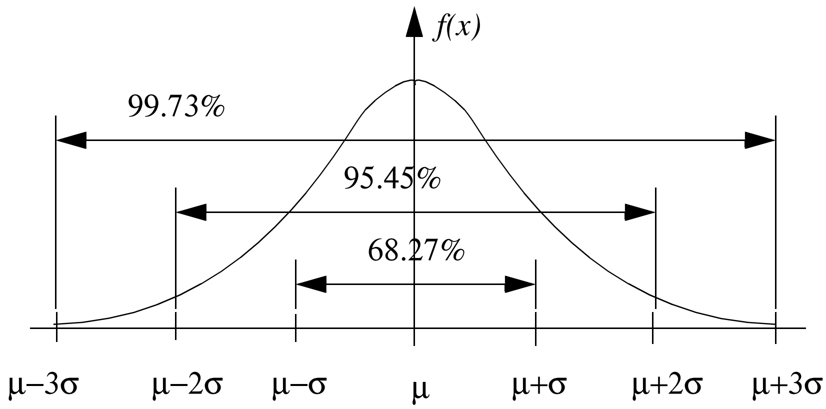
$F(x)$ = Probability of (random variable $\leq x$)

$$= \int_{-\infty}^{x} f(u)du$$

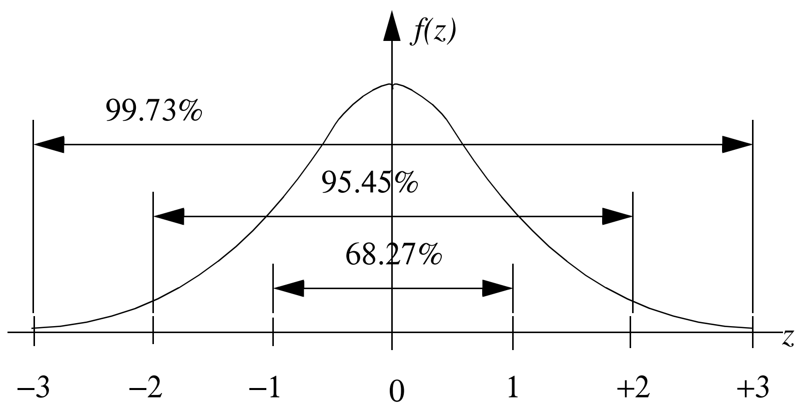
Definition of **Cumulative Distribution Function *F(x)***

Normal (Gaussian) Distribution

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



Normal Distribution Probability Function



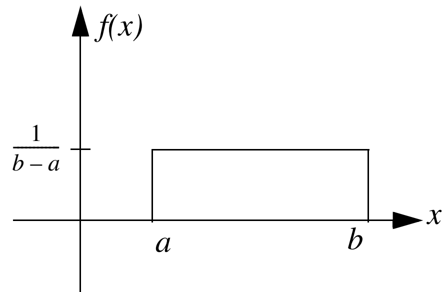
Standard Normal Distribution

Three things are needed to characterize a distribution:

- Mean
- Standard deviation
- Shape of the distribution function

For a normal distribution (shape), the *f(x)* can be uniquely defined if mean (μ) and standard deviation (σ) are known.

Uniform Distribution



Uniform Distribution Function

Theorems on Expectation

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \text{Expected value of X} = \text{Mean of X} = \mu_X$$

$$E\left[(x-\mu_x)^2\right] = \int_{-\infty}^{+\infty} (x-\mu_x)^2 f(x)dx = Var(X) = \sigma_X^2$$

where σ_X is the standard deviation of *x*.

Greek letters usually are used for the true parameters of the PDF of a random variable.

Properties of the expectation function:

- *E(cX)=cE(X),* where c is a constant
- E*(X+Y)=E(X)+E(Y)*
- *E(XY)=E(X)E(Y)* if *X* & *Y* are independent

Theorems on Variance

$$\sigma_X^2 = E[(x-\mu_X)^2] = E[x^2 - 2x\mu_X + \mu_X^2]$$
$$= (E[X^2] - 2\mu_X E[X] + \mu_X^2) = E[X^2] - \mu_X^2$$

- *Var(cX)*=c^2*Var(X)*
- *Var(X+Y)=Var(X)+Var(Y)* ($\sigma^2_{x+y}=\sigma^2_x+\sigma^2_y$) if *X* and *Y* are independent
- *Var(X-Y)=Var(X)+Var(Y)* ($\sigma^2_{x-y}=\sigma^2_x+\sigma^2_y$) if *X* and *Y* are independent

Example: For a uniform distribution if a=1 and b=7,

f(x)=1/6 for 1<x<7
0 others

then,

$$\mu_X = \int_1^7 x\left(\frac{1}{6}\right)dx = 4$$

and

$$\sigma_X^2 = \int_1^7 (x-4)^2\left(\frac{1}{6}\right)dx = 3$$

Probabilities

Areas under the PDF may be interpreted as probabilities.
For our uniform distribution function,

$$\Pr(1 \le X \le 7) = 1.0 = \int_1^7 \frac{1}{6}dx.$$

For the normal distribution function

$$Pr(a < X < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} dx$$

This integration can not be done analytically.

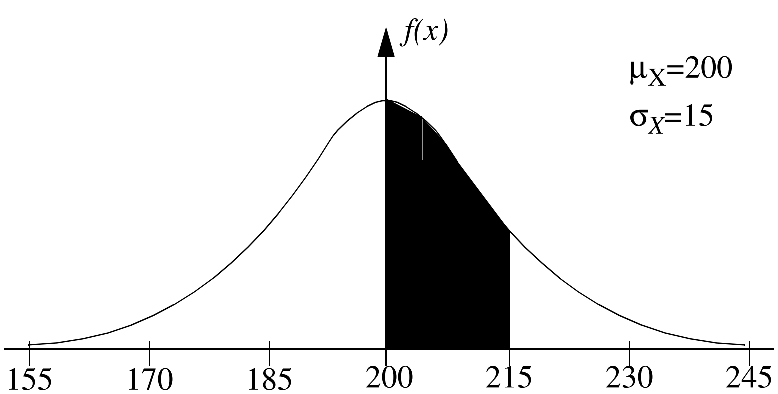
Any normal distribution function can be transformed into the "standardized/unit" normal dist$^{\underline{n}}$ by setting

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

and Z can be explained as the number of standard deviations from the mean. Areas under the unit normal distribution are tabulated in most statistics books.

Some Examples:

A random variable, X, describes the filled weight of a can of tomatoes



$Pr(X \geq 200) = 0.5$

$Pr(200 \leq X \leq 215) = Pr(0 \leq Z \leq 1) = 0.3413$

$Pr(200 \leq X \leq 220) = Pr(0 \leq Z \leq 1.33) = 0.4082$

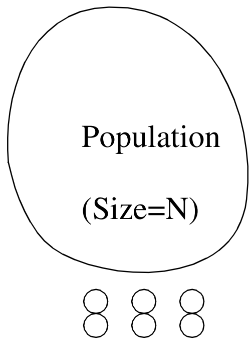
If we reach into the dist$^{\underline{n}}$ and pull out one can, how large (or small) can the weight be, before we believe something is wrong?

Let us select limits and if the weight of a can goes beyond the limits, we conclude the $mean \neq 200$.

In this case, $z = 1.96 = \dfrac{x_{CRIT} - \mu_X}{\sigma_X} = \dfrac{x_{CRIT} - 200}{15}$. Solving for the unknown gives $x_{CRIT} = 200 + (29.4)$.

If an x is beyond these limits, i.e., statistical significantly we conclude: "There is strong evidence to suggest that the true mean is not 200". For example, if a can is selected at random and, $x = 230$, this is evidence that the true mean is not 200.

Sampling



Sample size=n

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

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

Sample mean:

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

Sample variance:

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

and S is the sample standard deviation.

$\bar{X}$ and S are known as statistics.

Central Limit Theorem:

Sample means ($\bar{X}$'s) drawn from any type of dist$^{\underline{n}}$ tend to be normally distributed. The tendency is better at larger sample sizes.

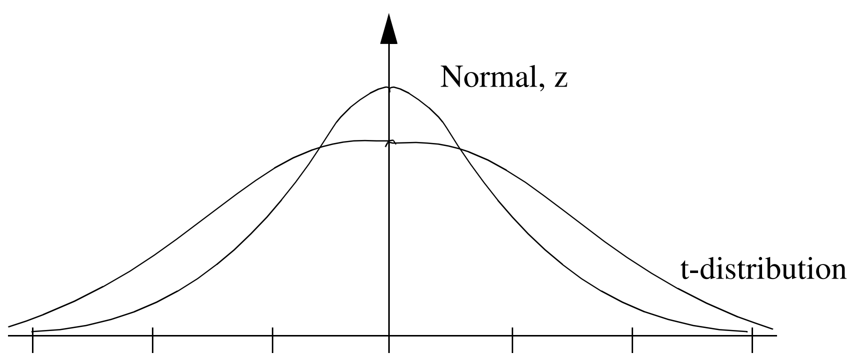
Population	Sample
Dist$^{\underline{n}}$(x's)	Means($\bar{x}$'s)

Mean=μ_x	Mean=$\mu_{\bar{X}} = \mu_X$
Var=σ_X^2	Var=$\sigma_{\bar{X}}^2 = \dfrac{\sigma_X^2}{n}$
Shape=Anything	Shape=Normal

Previously, we saw that for a normal dist$^{\underline{n}}$ with mean, μ, and standard deviation, σ, that the quantity z = $\dfrac{X-\mu}{\sigma}$ followed the unit normal dist$^{\underline{n}}$.

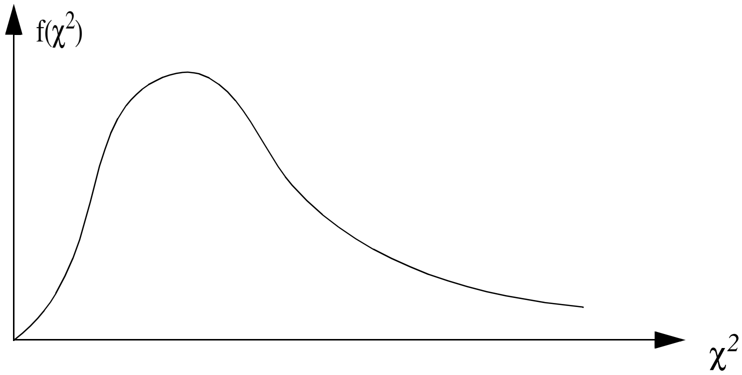
Consider a random variable, y which is normally distributed, unknown mean and variance. A sample of size n is collected to estimate the mean $\bar{Y}$ and standard deviation S_Y. The quantity will follow the t-distribution. $\dfrac{y-\mu_Y}{S_Y}$ follows the t - dist$^{\underline{n}}$ with v degrees of freedom, where v is the degrees of freedom that were used to calculate the variance.

Width of t - dist$^{\underline{n}}$ is greater than that of z because of additional uncertainty in using S in place of σ for the standard deviation.



Sampling Dist^n of S^2

If parent population is normal, the quantity $\dfrac{S^2(n-1)}{\sigma^2}$ follows the χ^2 distribution.



Probability Density Function of χ^2 Distribution

Sampling Distribution of Two Variances

If a sample of size $n_1(v_1=n_1-1)$ is drawn from Normal Dist$^{\underline{n}}$ with variance of σ_1^2 and a sample of size is $n_2(v_2=n_2-1)$ drawn from Normal Dist$^{\underline{n}}$ with variance of σ_2^2, estimates of Population variance S_1^2 and S_2^2 can be calculated. We know that $\frac{S_1^2}{\sigma_1^2}$ is $\frac{\chi_{v1}^2}{v_1}$ distributed and $\frac{S_2^2}{\sigma_2^2}$ is $\frac{\chi_{v2}^2}{v_2}$ distributed.

The ratio $\frac{\chi_{v1}^2 / v_1}{\chi_{v2}^2 / v_2}$ follows an F Distribution with ν_1, ν_2 degrees of freedom.

Therefore,

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{\nu_1,\nu_2}, \text{ or } \frac{S_1^2}{S_2^2} \sim \frac{\sigma_1^2}{\sigma_2^2} F_{\nu_1,\ \nu_2}.$$

If $\sigma_1^2 = \sigma_2^2$, then $\frac{S_1^2}{S_2^2} \sim F_{\nu_1,\ \nu_2}$.

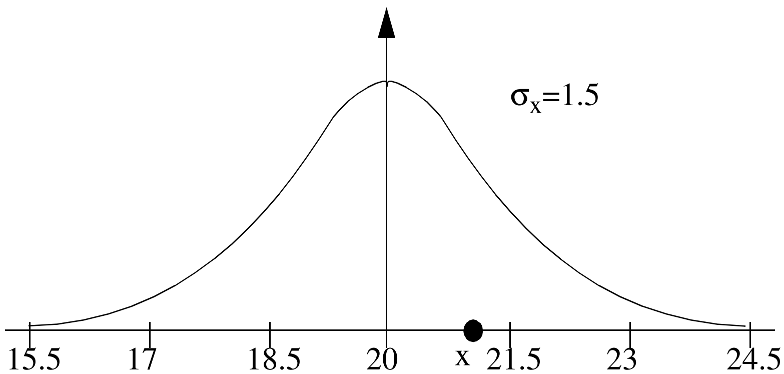
Decisions Concerning a Single Value

Statistical Test of Hypothesis

1. Define the statistic for the situation. State the Null (H_0) & Alternative (H_a) we hypotheses.
2. Select the risk/significance level.
3. Conduct the experiment and “calculate” the statistic.
4. Define dist$^{\underline{n}}$ for statistic. Select the appropriate test statistic: t, F, etc.
5. Make the statistical decision.

6. Draw the conclusions.
Example: X describes the filled weight of a sack of potatoes. The process is distributed normally with $\sigma_X = 1.5$. The manufacturer claims the average sack weight is 20 lbs. Is the claim true?

1. H_0: $\mu_X = 20$ and H_a: $\mu_X \neq 20$. Plan to collect a single x value.
2. Pick α =.05, tail area = 0.025.
3. Sack drawn at random, x=21.



Distribution of x under H_0

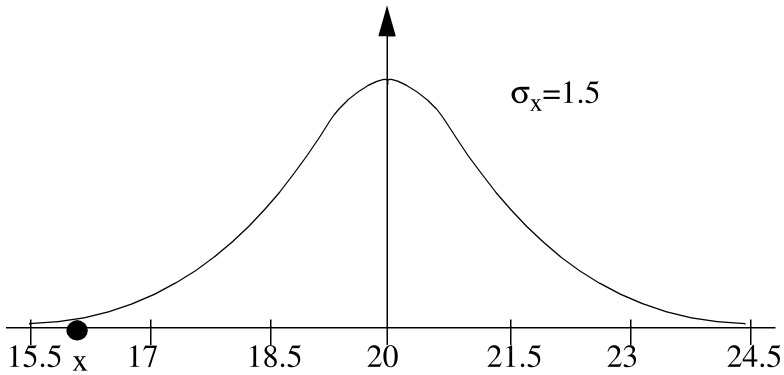
4. Test Statistic is

$$z = \frac{x - \mu_X}{\sigma_X} = \frac{21. - 20}{1.5} = 0.667$$

5. $Pr(Z \geq 0.667) = Pr(X \geq 21) = 0.2514$, so z is not statistically significant.
6. Can not reject H_0. The true mean may or may not be 20.

Another x is drawn, x=16

4. Test statistic is $z = \frac{x - \mu_X}{\sigma_X} = \frac{16 - 20}{1.5} = -2.67$

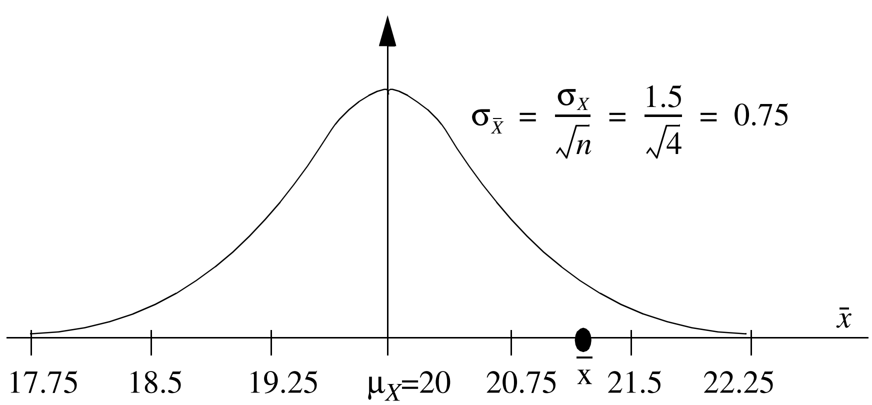


Distribution of *X* under H_0

5.$Pr(X \leq 16) = Pr(Z \leq -2.67) = .0038$. Therefore X and Z are statistically significant.
6.Reject H_0. The true mean is not 20.

Example: Once again focus on the potatoes. This time collect a sample of size n=4.
1. H_0: $\mu_X = 20$, H_a: $\mu_X \neq 20$. We will use sample mean, $\overline{X}$, to test hypothesis.
2. $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$
3. Sample collected: 20.5, 19.0, 22.0, & 22.5, and
4.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{(20.5 + 19.0 + 22.0 + 22.5)}{4} = 21$$



Test statistic $Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{21 - 20}{0.75} = 1.33$

5. $\Pr(\overline{X} \geq 21) = \Pr(Z \geq 1.33) = 0.0918$
6. Can't reject H_0. True mean may be 20.

Example: Examining larger sacks of potatoes which can be assumed to be normally distributed. We know $\mu_X = 40$, but σ_X is unknown. A sample of size n = 4 is collected. The sample is (41, 40, 42.5, 43.5).

1.H_0: $\mu_X = 40$, H_a: $\mu_X \neq 40$. Use $\overline{X}$ to test the hypothesis.

2. $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$

3.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{(41 + 40 + 42.5 + 43.5)}{4} = 41.75$$

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

$$= \frac{(41-41.75)^2 + (40-41.75)^2 + (42.5-41.75)^2 + (43.5-41.75)^2}{3} = 2.417$$

It uses $n-1$ instead n in the calculation of s_x^2, because it has only $n-1$ degrees of freedom. Note that s_x=1.5546
4.Test statistic:

$$t = \frac{\bar{X} - \mu_X}{S_{\bar{X}}}$$

in which $S_{\bar{X}} = \frac{S_X}{\sqrt{n}} = 0.777$. Therefore

$$t_{calc} = \frac{41.75-40}{0.777} = 2.252$$

5. From t - table Prob ($t_{3,\ .025} \geq 3.182$) = 0.025.
6. Can't reject manufacturer's claim.

Confidence Interval for μ_X

Example: A filling process is used to put cereal into boxes. The weight (oz.) of the boxes is normally distributed and has a standard deviation of 2. The manufacturer claims the process is centered at 22 oz. We will periodically test H_0 by drawing a box, and performing a statistical test of hypothesis.

1. 1.H_0: $\mu_X = 22$, H_a: $\mu_X \neq 22$

2. $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$

3. Draw a single x.

4. $z_{calc} = \frac{x - \mu_x}{\sigma_x}$

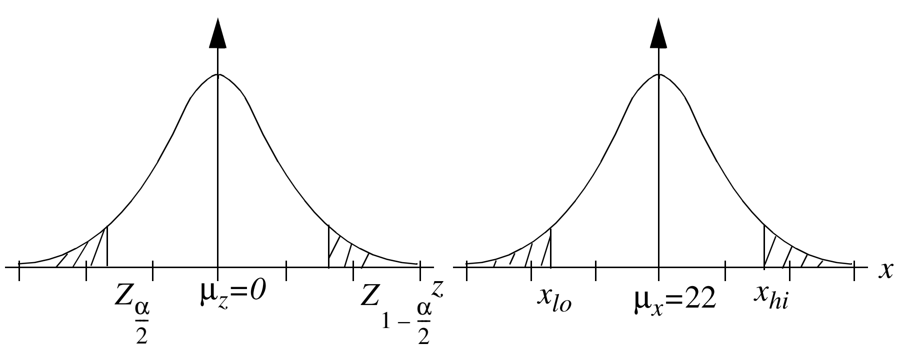
5. Evaluate probability of z_{calc}. If probability $\leq$ 0.025, z and x are statistically significant

6. Based on #5, make decision: Reject H_0 or Can't Reject H_0

Instead of calculating probability of z_{calc} and comparing it to $\frac{\alpha}{2}$, we can find the Z value associated with $\frac{\alpha}{2}$, i.e., $Z_{\frac{\alpha}{2}}, Z_{1-\frac{\alpha}{2}}$.

Prob $(Z \leq Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ and Prob $(Z \geq Z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}$, for instance $Z_{.025} = -1.96$ and $Z_{.975} = 1.96$.

If Z_{calc} is outside [-1.96, 1.96], then reject H_0
If Z_{calc} is within [-1.96, 1.96], can't reject H_0



$\frac{x_{lo} - \mu_X}{\sigma_X} = Z_{\frac{\alpha}{2}}$, $\frac{x_{hi} - \mu_X}{\sigma_X} = Z_{1-\frac{\alpha}{2}}$, $x_{lo} = \mu_X + Z_{\frac{\alpha}{2}}\sigma_X$ and $x_{hi} = \mu_X + Z_{1-\frac{\alpha}{2}}\sigma_X$.

Cutoff values are $\mu_X \pm Z_{1-\frac{\alpha}{2}}\sigma_X$

For our Example:

Cutoff values $\mu_X \pm Z_{1-\frac{\alpha}{2}}\sigma_X = 22 \pm 1.96 \times 2 = [18.08, 25.92]$

Draw an x = 19.3 $\rightarrow$ Can't reject H_0

Draw an x = 23.4 $\rightarrow$ Can't reject H_0

Draw an x = 26.8 $\rightarrow$ Reject H_0

Another process normally distributed with $\sigma_X = 3$. A single value drawn at random, $x = 25$. Can we guess or estimate where the dist$^{\underline{n}}$ of x's is truly centered (μ_X).

Let's assume the x we obtained was fairly typical, i.e., not a rare event. How low (or high) could μ_X be and still have this x within the cutoff values?

It is easy to see that $\frac{x-\mu_{lo}}{\sigma_X} = Z_{1-\frac{\alpha}{2}}$ and $\frac{x-\mu_{hi}}{\sigma_X} = Z_{\frac{\alpha}{2}}$,

so

$$\mu_{lo} = x - Z_{1-\frac{\alpha}{2}}\sigma_X$$

and

$$\mu_{hi} = x - Z_{\frac{\alpha}{2}}\sigma_X.$$

For an α level of 0.05, μ_{lo}=19.12 and μ_{hi}=30.88, thus we believe that 19.12 $\le \mu_X \le$ 30.88.

We have developed what is known as a confidence interval. In fact a 100(1- α)% confidence interval for the true mean. We are 95% confident that 19.12 $\le \mu_X \le$30.88.

In general,

$$x - Z_{1-\frac{\alpha}{2}}\sigma_X \le \mu_X \le x + Z_{1-\frac{\alpha}{2}}\sigma_X$$

Example: Sacks of potatoes. Develop a C.I. for average bag weight (μ_X) for n=8 and $\sigma_X = 1.5$.

Assume the weights are normally distributed. Sample is

(20,23,22,19,22,21,20,24).

$\bar{x}$=21.375 and $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1.5}{\sqrt{8}} = 0.53$. Let α=0.05, and $Z_{1-\frac{\alpha}{2}} = 1.96$, so

$$\mu_{lo} = \bar{x} - Z_{1-\frac{\alpha}{2}}\sigma_X = 20.34,$$

and

$$\mu_{hi} = \bar{x} + Z_{1-\frac{\alpha}{2}}\sigma_X = 22.41.$$

Therefore, 95% CI for μ_X is $20.34 \le \mu_X \le 22.41$. We are 95% confident that $20.34 \le \mu_X \le 22.4$. We can't reject any H_0 when μ_X has a value on this interval.

If we were to collect many $\bar{X}$ values, 95% of the C.I.'s developed from these $\bar{x}$' s would include the true μ_X.

For the potato example, let's say σ_X was unknown, and we estimated it from our sampled data:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{X})^2 = \frac{19.875}{7} = 2.84.$$

So s_x=1.685 and $v = 7$ are the degrees of freedom.

Previously, we calculated μ_{lo} and μ_{hi} when σ_X was known

Now, without σ_X available, $\frac{\bar{x}-\mu_{\bar{X}}}{s_{\bar{X}}} = \frac{\bar{x}-\mu_X}{s_X/\sqrt{n}} = t_v$. Therefore,

$$\frac{\bar{x}-\mu_{lo}}{s_{\bar{X}}} = t_{\nu, 1-\frac{2}{\alpha}}$$

$$\frac{\bar{x}-\mu_{hi}}{s_{\bar{X}}} = t_{\nu, \frac{2}{\alpha}} = -t_{\nu, 1-\frac{\alpha}{2}}$$

$$\mu_{lo}, \mu_{hi} = \bar{x} \pm t_{\nu, 1-\frac{2}{\alpha}} s_X = 21.375 \pm (2.365)(0.5957) = 21.375 \pm 1.409$$

$$19.97 \le \mu_X \le 22.78$$

or with $100\,(1-\alpha)\,\% = 95\%$ confidence, the true mean lies on the interval [19.97, 22.78]