DESIGN OF EXPERIMENTS

1.0 STATISTICS REVIEW
(updated Spring 2001)

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.

\[
\int_{-\infty}^{\infty} f(x) dx = 1
\]

\[
F(x) = \int_{-\infty}^{x} f(u) du
\]

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

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

Three things are needed to characterize a distribution:
- Mean
- Standard deviation (Variance)
- Shape of the distribution function

For a normal distribution (shape), the \( f(x) \) can be uniquely defined if mean \( (\mu) \) and standard deviation \( (\sigma) \) are known.

Uniform Distribution

\[ f(x) = \frac{1}{b - a} \quad a < x < b \]

Uniform Distribution Function

Theorems on Expectation

\[ E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \text{Expected value of } X = \text{Mean of } X = \mu_X \]
\[ E \left[ (x - \mu_x)^2 \right] = \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x) \, dx = \text{Var}(X) = \sigma_x^2 \]

where \( \sigma_x \) 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**

\[
\sigma_x^2 = E[(x - \mu_x)^2] = E[x^2 - 2x\mu_x + \mu_x^2] = E[X^2] - 2\mu_x E[X] + \mu_x^2 = E[X^2] - \mu_x^2
\]

- \( \text{Var}(cX) = c^2 \text{Var}(X) \)
- \( \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) \) (\( \sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 \)) if \( X \) and \( Y \) are independent
- \( \text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) \) (\( \sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 \)) if \( X \) and \( Y \) are independent

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

\[
f(x) = \frac{1}{6} \quad \text{for } 1 < x < 7
\]

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,

\[ \Pr(1 \leq X \leq 7) = 1.0 = \int_1^7 \frac{1}{6} \, dx. \]

For the normal distribution function
This integration can not be done analytically.

Any normal distribution function can be transformed into the “standardized/unit” normal dist\(^2\) 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

\[
Pr(a < X < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx
\]

If we reach into the dist\(^3\) 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, \(z = 1.96 = \frac{x_{CRIT} - \mu_X}{\sigma_X} = \frac{x_{CRIT} - 200}{15}\). Solving for the unknown gives \(x_{CRIT} = 200 + (29.4)\).
If an x is beyond these limits, i.e., statistically 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**

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

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

Sample mean:

\[
\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

Sample variance:

\[
S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})^2
\]

and S is the sample standard deviation. \( \overline{X} \) and S are known as statistics.

**Central Limit Theorem:**

Sample means (\( \overline{X} \)'s) drawn from any type of dist\( \mu \) tend to be normally distributed. The tendency is better at larger sample sizes.
Previously, we saw that for a normal dist\(\alpha\) with mean, \(\mu\), and standard deviation, \(\sigma\), that the quantity \(z = \frac{X - \mu}{\sigma}\) followed the unit normal dist\(\alpha\).

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.

\[
\frac{y - \mu_y}{S_y}
\]

follows the t-dist\(\alpha\) with \(v\) degrees of freedom, where \(v\) is the degrees of freedom that were used to calculate the variance.

Width of t-dist\(\alpha\) is greater than that of \(z\) because of additional uncertainty in using \(S\) in place of \(\sigma\) for the standard deviation.

**Sampling Dist\(\alpha\) of \(S^2\)**

If parent population is normal, the quantity \(\frac{S^2(n-1)}{\sigma^2}\) follows the \(\chi^2\) distribution.
Sampling Distribution of Two Variances

If a sample of size \( n_1 = n_1 - 1 \) is drawn from Normal Dist with variance of \( \sigma_1^2 \) and a sample of size is \( n_2 = n_2 - 1 \) drawn from Normal Dist with variance of \( \sigma_2^2 \), 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 distributed and \( \frac{S_2^2}{\sigma_2^2} \) is distributed.

The ratio \( \frac{\chi^2_{v_1}}{\chi^2_{v_2}} \) follows an F Distribution with \( v_1, v_2 \) degrees of freedom. Therefore,

\[
\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{v_1, v_2}, \quad \text{or} \quad \frac{S_1^2}{S_2^2} \sim \frac{\sigma_1^2}{\sigma_2^2} F_{v_1, v_2}.
\]

If \( \sigma_1^2 = \sigma_2^2 \), then \( \frac{S_1^2}{S_2^2} \sim F_{v_1, v_2} \).

Decisions Concerning a Single Value

Statistical Test of Hypothesis
1. Define the statistic for the situation. State the Null (\( H_0 \)) & Alternative (\( H_a \)) hypotheses.
2. Select the risk/significance level.
3. Conduct the experiment and “calculate” the statistic.
4. Define dist 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 $\alpha = 0.05$, tail area = 0.025.

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

$$z = \frac{x - \mu_X}{\sigma_X} = \frac{16 - 20}{1.5} = -2.67$$
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, \(\bar{X}\), 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. 

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

\[
\sigma_x = \frac{\sigma_x}{\sqrt{n}} = \frac{1.5}{\sqrt{4}} = 0.75
\]

Test statistic \(Z = \frac{\bar{X} - \mu_X}{\sigma_X} = \frac{21-20}{0.75} = 1.33\)
5. \( \text{Pr}(\bar{X} \geq 21) = \text{Pr}(Z \geq 1.33) = 0.0918 \)

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 \( \sigma_X \) 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 \( \bar{X} \) 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:

\[
t = \frac{\bar{X} - \mu_X}{S_x/\sqrt{n}}
\]

in which \( S_x = \frac{S_x}{\sqrt{n}} = 0.777 \). Therefore

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

5. From \( t \)-table \( \text{Prob}(t_3, .025 \geq 3.182) = 0.025 \).

6. Can’t reject manufacturer’s claim.

**Confidence Interval for \( \mu_X \)**

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

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., $Z_{\frac{\alpha}{2}}, Z_{1-\frac{\alpha}{2}}$.

$\text{Prob} (Z \leq Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ and $\text{Prob} (Z \geq Z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}$, for instance $Z_{0.025} = -1.96$ and $Z_{0.975} = 1.96$.

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 $H_0$
Cutoff values are $\mu_X \pm Z_{\frac{1-\alpha}{2}} \sigma_X$

For our Example:

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

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 distribution of $x$’s is truly centered ($\mu_X$).

Let’s assume the $x$ we obtained was fairly typical, i.e., not a rare event. How low (or high) could $\mu_X$ 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 $\alpha$ level of 0.05, $\mu_{lo} = 19.12$ and $\mu_{hi} = 30.88$, thus we believe that $19.12 \leq \mu_X \leq 30.88$.

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

In general,

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

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

Assume the weights are normally distributed. Sample is
(20,23,22,19,22,21,20,24).

\[ \bar{x} = 21.375 \text{ and } \sigma_X = \frac{\sigma_X}{\sqrt{n}} = \frac{1.5}{\sqrt{8}} = 0.53. \] Let \( \alpha = 0.05, \) and \( Z_{1-\alpha} = 1.96, \) so

\[ \mu_{lo} = \bar{x} - Z_{1-\alpha} \sigma_X = 20.34, \]

and

\[ \mu_{hi} = \bar{x} + Z_{1-\alpha} \sigma_X = 22.41. \]

Therefore, 95% CI for \( \mu_X \) is \( 20.34 < \mu_X < 22.41. \) We are 95% confident that \( 20.34 < \mu_X < 22.4. \) We can’t reject any \( H_0 \) when \( \mu_X \) 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 \( \mu_X \).

For the potato example, let’s say \( \sigma_X \) was unknown, and we estimated it from our sampled data:

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

So \( s_X = 1.685 \) and \( \nu = 7 \) are the degrees of freedom.

Previously, we calculated \( \mu_{lo} \) and \( \mu_{hi} \) when \( \sigma_X \) was known

Now, without \( \sigma_X \) available, \( \frac{\bar{x} - \mu_X}{s_X} = \frac{\bar{x} - \mu_X}{\sigma_X/\sqrt{n}} = t_\nu. \) Therefore,

\[
\frac{\bar{x} - \mu_{lo}}{s_X} = t_{\nu, 1-\alpha/2}
\]

\[
\frac{\bar{x} - \mu_{hi}}{s_X} = t_{\nu, \alpha/2} = -t_{\nu, 1-\alpha/2}
\]

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

\[ 19.97 \leq \mu_X \leq 22.78 \]

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