Preface

Major branches of Statistics:

- Descriptive Statistics
- Inferential Statistics
What is Inferential Statistics?

The objective is to make inferences about a population parameters based on information contained in a sample.
What is Inferential Statistics?

The objective is to make inferences about a population parameters based on information contained in a sample.

Mean, Median, Standard Deviation, Proportion
Inferences about Population Parameters

What is Inferential Statistics?

Statistical inference-making procedures differ from ordinary procedures in that we not only make an inference but also provide a measure of how good that inference is.
What is Inferential Statistics?

Methods for making inferences about parameters fall into two categories:

- **Estimating** the population parameter
- **Hypothesis Testing** about a population parameter
Point Estimation

The first step in statistical inference is ‘Point Estimation’, in which we compute a single value (statistic) from the sample data to estimate a population parameter.
Inferences about Population Parameters

Estimation

Interval Estimation

\[ \hat{\theta} \pm A \times SE(\hat{\theta}), \text{ where } 'A' \text{ is based on sampling distribution of } \hat{\theta} \]
Estimation of ‘μ’: 

- **Point Estimation**: Sample Mean $\bar{y}$
Inferences about Population Parameters

Estimation

**Estimation of ‘μ’:**

- **Point Estimation:** Sample Mean $\bar{y}$
- **Interval Estimation:**
Inferences about Population Parameters

Estimation

Estimation of ‘μ’: 

- **Point Estimation:** Sample Mean \(\overline{y}\)

- **Interval Estimation:**
  
  For large ‘n’, \(\overline{y}\) is approximately normally distributed with mean ‘μ’ and standard error \(\frac{\sigma}{\sqrt{n}}\)
Inferences about Population Parameters

Estimation

Estimation of ‘μ’:

- **Point Estimation:** Sample Mean $\bar{y}$
- **Interval Estimation:** $(\bar{y} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{y} + 1.96 \frac{\sigma}{\sqrt{n}})$

with level of confidence 95% when $\sigma$ is known
Estimation of ‘μ’:

- **Point Estimation:** Sample Mean \( \bar{y} \)
- **Interval Estimation:** \((\bar{y} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{y} + 1.96 \frac{\sigma}{\sqrt{n}})\)

with level of confidence **95%** when \( \sigma \) is known

In 95% of the times in repeated sampling, the interval contains the mean ‘μ’
Inferences about Population Parameters

Estimation

Estimation of ‘μ’:

- **Point Estimation**: Sample Mean $\bar{y}$

- **Interval Estimation**: $(\bar{y} - 2.09 \frac{s}{\sqrt{n}}, \bar{y} + 2.09 \frac{s}{\sqrt{n}})$

  with level of confidence 95% when σ is unknown
Estimation of ‘μ’:

- **Point Estimation:** Sample Mean \( \bar{y} \)
- **Interval Estimation:** \((\bar{y} - 2.09 \frac{s}{\sqrt{n}}, \bar{y} + 2.09 \frac{s}{\sqrt{n}})\)

*Is good approximation if population distribution is not too non-normal and sample size is large enough*
Inferences about Population Parameters

Estimation

$100(1 - \alpha)\%$ confidence Interval for ‘$\mu$’

(‘$\sigma$’ known) when sampling from a normal population or ‘$n$’ large

\[
\left( \bar{y} - z\alpha \frac{s}{\sqrt{n}}, \bar{y} + z\alpha \frac{s}{\sqrt{n}} \right)
\]
100(1 − α)% confidence Interval for ‘μ’
(‘σ’ unknown) when sampling from a normal population or ‘n’ large

\[
\left( \bar{y} - t_{\alpha} \frac{s}{\sqrt{n}}, \bar{y} + t_{\alpha} \frac{s}{\sqrt{n}} \right)
\]
Goodness of inference for interval estimation:

1. Confidence coefficient
2. Width of the confidence interval
Goodness of inference for interval estimation:

1. Confidence coefficient: Higher
2. Width of the confidence interval
Goodness of inference for interval estimation:

1. Confidence coefficient
   - Higher

2. Width of the confidence interval
   - Smaller
Sample Size Required for a $100(1 - \alpha)\%$ Confidence Interval for $\mu$ of the Form $\bar{y} \pm E$

$$n = \frac{(z\alpha)^2 \sigma^2}{E^2}$$
Inferences about Population Parameters

Estimation

Sample Size Required for a $100(1 - \alpha)\%$ Confidence Interval for $\mu$ of the Form $\bar{y} \pm E$

\[ n = \frac{\left(\frac{Z\alpha}{2}\right)^2 \sigma^2}{E^2} \]

Estimate using information from prior survey
Sample Size Required for a $100(1 - \alpha)\%$ Confidence Interval for $\mu$ of the Form $\bar{y} \pm E$

\[ n = \frac{(\frac{Z\alpha}{2})^2 \sigma^2}{E^2} \]

Estimate using
\[ s = \frac{\text{range}}{4} \]
Inferences about Population Parameters

Estimation

Sample Size for Estimation of \( \mu \):

**Consideration 1:**
Desired Confidence Level (z-value)

\[
(\bar{y} - z \alpha \frac{\sigma}{\sqrt{n}}, \bar{y} + z \alpha \frac{\sigma}{\sqrt{n}})
\]

**Consideration 2:**
the tolerable error that establishes desired width of the interval

**Consideration 3:**
Standard Deviation (\(\sigma\))

appropriate sample size for estimating \(\mu\) using a confidence interval
Inferences about Population Parameters
Estimation

Sample Size for Estimation of ‘μ’:

Consideration 1:
High Confidence Level (z-value)

Consideration 2:
Narrow width of the interval

Consideration 3:
Large Standard Deviation (σ)

Large Sample Size

(\bar{y} - z\alpha \frac{\sigma}{\sqrt{n}}, \bar{y} + z\alpha \frac{\sigma}{\sqrt{n}})
Inferences about Population Parameters

Estimation

Sample Size for Estimation of 'μ':

Consideration 1:
Confidence Level is traditionally set at 95%.

Consideration 2:
Width of interval depends heavily on context of problem.

Consideration 3:
Based on a guess about population standard deviation or Estimate from an initial sample.

Large Sample Size

\[
(\bar{y} - z\alpha \frac{\sigma}{\sqrt{n}}, \bar{y} + z\alpha \frac{\sigma}{\sqrt{n}})
\]
Sample Size for Estimation of ‘μ’:

- **desired accuracy of the sample statistic** as an estimate of the population parameter
- **required time and cost to achieve this degree of accuracy**
Statistical Tests

Using sampled data from the population, we are simply attempting to determine the value of the parameter.

In hypothesis testing, there is a idea about the value of the population parameter.
Inferences about Population Parameters

Hypothesis Testing

**Statistical Tests**

A statistical test is based on the concept of proof and composed of **five** parts:

- **$H_a$:** Research Hypothesis (Alternative Hypothesis)
- **$H_0$:** Null Hypothesis
- Test Statistic
- Rejection Region
- Check assumptions and draw conclusions
Guidelines for Determining $H_0$ and $H_a$ in Statistical Tests

- $H_0$: the statement that parameter equals a specific value
- $H_a$: the statement that researcher is attempting to support or detect using the data
- The null hypothesis is presumed correct unless there is strong evidence in the data that supports the research hypothesis.
Test Statistic

The quantity computed from the sample data, that helps to decide whether or not the data support the research hypothesis.
Rejection Region

The rejection region (based on the sampling distribution) contains the values of test statistic that support the research hypothesis and contradict the null hypothesis.
Inferences about Population Parameters
Hypothesis Testing

Statistical Tests

- One-Tailed Test
- Two-Tailed Test
Inferences about Population Parameters
Hypothesis Testing

Statistical Tests

- **One-Tailed Test**

The rejection region is located in only one tail of the sampling distribution of test statistic

\[ H_a: \theta < \theta_0 \text{ Or } H_a: \theta > \theta_0 \]
Inferences about Population Parameters
Hypothesis Testing

Statistical Tests

- **Two-Tailed Test**

  The rejection region is located in **both tails** of the sampling distribution of test statistic

  $$H_a: \theta \neq \theta_0$$
Inferences about Population Parameters
Hypothesis Testing

Errors in Statistical Tests

- Type I Error
- Type II Error
Errors in Statistical Tests

- Type I Error
  - Rejecting the null hypothesis when it is true.
  - The probability of a Type I error is denoted by ‘\( \alpha \)’.

- Type II Error
Inferences about Population Parameters
Hypothesis Testing

Errors in Statistical Tests

- Type I Error
  - Accepting the null hypothesis when it is false. The probability of a Type II error is denoted by ‘β’.

- Type II Error
### Errors in Statistical Tests

<table>
<thead>
<tr>
<th>Decision</th>
<th>Null Hypothesis</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>Reject</td>
<td>Type I ( \alpha )</td>
<td>Correct ( 1-\alpha )</td>
</tr>
<tr>
<td>Accept</td>
<td>Correct ( 1-\beta )</td>
<td>Type II ( \beta )</td>
</tr>
</tbody>
</table>
### Errors in Statistical Tests

<table>
<thead>
<tr>
<th>Decision</th>
<th>Truth</th>
<th>Null Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td>Reject Type I $\alpha$</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td>Accept Correct $(1-\beta)$</td>
</tr>
</tbody>
</table>

The probabilities associated with Type I and Type II errors are inversely related. For a fixed sample size ‘$n$’, when ‘$\alpha$’ decreases ‘$\beta$’ will increase and vice versa.
### Inferences about Population Parameters

#### Hypothesis Testing

#### Errors in Statistical Tests

<table>
<thead>
<tr>
<th>Decision</th>
<th>Null Hypothesis</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>Reject</td>
<td>Type I ( \alpha )</td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1-( \alpha ))</td>
</tr>
<tr>
<td>Accept</td>
<td>Correct ( (1-\beta) )</td>
<td>Type II ( \beta )</td>
</tr>
</tbody>
</table>

Usually ‘\( \alpha \)’ is specified to locate the **Rejection Region**
# Errors in Statistical Tests

## Inferences about Population Parameters

### Hypothesis Testing

<table>
<thead>
<tr>
<th>Decision</th>
<th>Null Hypothesis</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td>False</td>
<td></td>
</tr>
<tr>
<td>Reject</td>
<td>Type I α</td>
<td>Correct (1-α)</td>
<td></td>
</tr>
<tr>
<td>Accept</td>
<td>Correct (1-β)</td>
<td>Type II β</td>
<td></td>
</tr>
</tbody>
</table>

**Power of Test**
Effectiveness of a statistical test is measured by:

Maginitudes of
Type I Error and Type II Error
Effectiveness of a statistical test is measured by:

For a fixed ‘α’, as the sample size increases, ‘β’ decreases
Inferences about Population Parameters
Hypothesis Testing

**Statistical Tests (Drawing Conclusion)**

**Traditional Approach:**
- Using Statistic Test, two types of errors, their probability $\alpha$, $\beta$
Inferences about Population Parameters
Hypothesis Testing

Statistical Tests

Traditional Approach:
- Using Statistic Test, two types of errors, their probability $\alpha$, $\beta$

The problem with this approach is that if other researchers want to apply the results of your study using a different value for $\alpha$ then they must compute a new rejection region.
Inferences about Population Parameters
Hypothesis Testing

Statistical Tests – Alternative Approach

Using Level of Significance/P-Value

Smallest size of ‘α’ at which H₀ can be rejected, based on the observed value of the test statistic.
Statistical Tests – Alternative Approach

Using Level of Significance/P-Value

Smallest size of ‘\( \alpha \)’ at which \( H_0 \) can be rejected, based on the observed value of the test statistic.

The probability of observing a sample outcome more contradictory to \( H_0 \) than the observed sample result.
Statistical Tests – Alternative Approach

Using Level of Significance/P-Value

The weight of evidence for rejecting the null hypothesis.
Inferences about Population Parameters
Hypothesis Testing

Statistical Tests – Alternative Approach

Using Level of Significance/P-Value
The weight of evidence for rejecting the null hypothesis

The smaller the value of this probability,
the heavier the weight of the sample evidence against $H_0$. 
Inferences about Population Parameters
Hypothesis Testing

Statistical Tests – Alternative Approach

Decision Rule for Hypothesis Testing Using P-Value

- **P - Value ≤ α**
  - Reject H₀

- **P - Value > α**
  - Fail to Reject H₀
Inferences about Population Parameters

Hypothesis Testing

Statistical Test for population mean ‘μ’

(‘σ’ is known, when sampling from a normal population or ‘n’ large)

Test Statistic:

\[ z = \frac{\bar{y} - \mu_0}{\sigma / \sqrt{n}} \sim N(0, 1) \]

\[ H_0 : \mu \leq \mu_0 \quad H_a : \mu > \mu_0 \]

• Reject \( H_0 \) if \( z \geq z_\alpha \)

\[ H_0 : \mu \geq \mu_0 \quad H_a : \mu < \mu_0 \]

• Reject \( H_0 \) if \( z \leq -z_\alpha \)

\[ H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0 \]

• Reject \( H_0 \) if \( |z| \geq z_{\alpha/2} \)
Statistical Test for population mean ‘μ’

(‘σ’ is known, when sampling from a normal population or ‘n’ large)

Power of the test:

One-Tailed Test

\[ 1 - \beta(\mu_a) = 1 - Pr(z \leq z_\alpha - \frac{|\mu_0 - \mu_a|}{\sigma / \sqrt{n}}) \]

Two-Tailed Test

\[ 1 - \beta(\mu_a) \approx 1 - Pr(z \leq \frac{z_\alpha}{2} - \frac{|\mu_0 - \mu_a|}{\sigma / \sqrt{n}}) \]
Inferences about Population Parameters

Hypothesis Testing

Statistical Test for population mean ‘μ’

(‘σ’ is known, when sampling from a normal population or ‘n’ large)

\[
\begin{align*}
\{ & H_0: \mu \leq \mu_0 \\
& H_a: \mu > \mu_0 \}
\end{align*}
\]

• P-Value: \( Pr(z \geq \text{computed } z) \)

\[
\begin{align*}
\{ & H_0: \mu \geq \mu_0 \\
& H_a: \mu < \mu_0 \}
\end{align*}
\]

• P-Value: \( Pr(z \leq \text{computed } z) \)

\[
\begin{align*}
\{ & H_0: \mu = \mu_0 \\
& H_a: \mu \neq \mu_0 \}
\end{align*}
\]

• P-Value: \( 2Pr(z \geq |\text{computed } z|) \)
Statistical Test for population mean ‘μ’

(‘σ’ is unknown, when sampling from a normal population or ‘n’ large)

Test Statistic:

\[ T = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \sim t(n - 1) \]

- **Hypothesis Testing**
  - **H_0**: \( \mu \leq \mu_0 \)
  - **H_a**: \( \mu > \mu_0 \)
    - Reject \( H_0 \) if \( t \geq t_\alpha \)
  - **H_0**: \( \mu \geq \mu_0 \)
  - **H_a**: \( \mu < \mu_0 \)
    - Reject \( H_0 \) if \( t \leq -t_\alpha \)
  - **H_0**: \( \mu = \mu_0 \)
  - **H_a**: \( \mu \neq \mu_0 \)
    - Reject \( H_0 \) if \(|t| \geq t_{\alpha/2}\)
Statistical Test for population mean ‘μ’

(‘σ’ is unknown, when sampling from a normal population or ‘n’ large)

\[
\begin{align*}
\{ H_0 : \mu &\leq \mu_0 \\
H_a : \mu &> \mu_0
\end{align*}
\quad \text{P-Value: } Pr(t \geq \text{computed } t)
\]

\[
\begin{align*}
\{ H_0 : \mu &\geq \mu_0 \\
H_a : \mu &< \mu_0
\end{align*}
\quad \text{P-Value: } Pr(t \leq \text{computed } t)
\]

\[
\begin{align*}
\{ H_0 : \mu &= \mu_0 \\
H_a : \mu &\neq \mu_0
\end{align*}
\quad \text{P-Value: } 2Pr(t \geq |\text{computed } t|)
\]
Non-normal distribution of population

‘2’ ISSUES to consider:
- Skewed Distribution
- Heavy-tailed Distribution
Inferences about Population Parameters
Hypothesis Testing

Non-normal distribution of population

‘2’ ISSUES to consider:
- Skewed Distribution
- Heavy-tailed Distribution

Tests of Hypothesis tend to have smaller ‘\( \alpha \)’ than specified level, so test has lower power
Inferences about Population Parameters

Hypothesis Testing

Non-normal distribution of population

‘2’ ISSUES to consider:
- Skewed Distribution
- Heavy-tailed Distribution

Tests of Hypothesis tend to have smaller ‘α’ than specified level, so test has lower power

Robust Methods
Inferences about Median

When the population distribution is “highly skewed” or “very heavily tailed” or “sample size is small”, median is more appropriate than the mean as a representation of the center of the population.
Statistical Test for population median
(Sign Test)

Test Statistic: \( W_i = y_i - M_0 \), \( B = \text{No. of Positive } W_i\text{s} \)
\( B \sim \text{Binom}(n, \pi) \)

\[
\begin{align*}
\{H_0: M \leq M_0, H_a: M > M_0\} & \quad \text{Reject } H_0 \text{ if } B \geq n - C_{\alpha(1),n} \\
\{H_0: M \geq M_0, H_a: M < M_0\} & \quad \text{Reject } H_0 \text{ if } B \leq C_{\alpha(1),n} \\
\{H_0: M = M_0, H_a: M \neq M_0\} & \quad \text{Reject } H_0 \text{ if } B \leq C_{\alpha(2),n} \text{ or } B \geq n - C_{\alpha(2),n}
\end{align*}
\]
Inferences about Population Parameters

Hypothesis Testing

Statistical Test for population median

(Approximation)

Test Statistic:  \( B_{st} = \frac{B - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \),  \( B_{st} \sim N(0, 1) \)

\( H_0: M \leq M_0 \)
\( H_a: M > M_0 \)

- Reject \( H_0 \) if \( B_{st} \geq z_\alpha \) with P-value \( Pr(z \geq B_{st}) \)

\( H_0: M \geq M_0 \)
\( H_a: M < M_0 \)

- Reject \( H_0 \) if \( B_{st} \leq z_\alpha \) with P-value \( Pr(z \leq B_{st}) \)

\( H_0: M = M_0 \)
\( H_a: M \neq M_0 \)

- Reject \( H_0 \) if \( |B_{st}| \geq \frac{z_\alpha}{2} \) with P-value \( 2Pr(z \geq |B_{st}|) \)
Inferences about Population Parameters
Hypothesis Testing

**Statistical Test for Mean (one-population)**

<table>
<thead>
<tr>
<th>Mean of One-Population</th>
<th>Normal Population, or ‘n’ large</th>
<th>‘σ’ known</th>
<th>Test using z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of One-Population</td>
<td>Normal Population, or ‘n’ large</td>
<td>‘σ’ unknown</td>
<td>T-Test</td>
</tr>
<tr>
<td>Mean of One-Population</td>
<td>Non-normal Population</td>
<td></td>
<td>Sign-Test</td>
</tr>
</tbody>
</table>
Inferences about Population Parameters

Hypothesis Testing

Statistical Test for population Variance

(Normal Population)

Test Statistic:

\[ \chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n - 1), \quad s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \]

- **Reject H$_0$ if $\chi^2 > \chi^2_{U,\alpha}$**
- **Reject H$_0$ if $\chi^2 < \chi^2_{L,\alpha}$**
- **Reject H$_0$ if $\chi^2 > \chi^2_{U,\alpha/2}$ or $\chi^2 > \chi^2_{L,\alpha/2}$**
100(1 – \(\alpha\))% confidence Interval for \(\sigma^2\) (or \(\sigma\))

\[
\frac{(n - 1)s^2}{\chi^2_U} < \sigma^2 < \frac{(n - 1)s^2}{\chi^2_L}
\]

\[
\sqrt{\frac{(n - 1)s^2}{\chi^2_U}} < \sigma < \sqrt{\frac{(n - 1)s^2}{\chi^2_L}}
\]
The inferences we have made so far have concerned a parameter of a single population.

Quite often we are faced with an inference involving a comparison of parameters of different populations.
Theorem

If two independent random variables $y_1$ and $y_2$ are normally distributed with means and variances $(\mu_1, \sigma_1^2)$ and $(\mu_2, \sigma_2^2)$ respectively, then

$$(y_1 - y_2) \sim N \left( (\mu_1 - \mu_2), \sqrt{\sigma_1^2 + \sigma_2^2} \right)$$
Inferences about Population Parameters
Hypothesis Testing

Sampling Distribution for $\bar{y}_1 - \bar{y}_2$

Two independent large samples

$$(\bar{y}_1 - \bar{y}_2) \sim N\left((\mu_1 - \mu_2), \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$
Inferences about Population Parameters
Hypothesis Testing

Statistical Test for \( \mu_1 - \mu_2 \)

(Independent samples, \( y_1 \) and \( y_2 \) approximately normal, \( \sigma_1^2 = \sigma_2^2 \))

Test Statistic:

\[
\begin{align*}
t &= \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)
\end{align*}
\]

where:

\[
s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}
\]
Inferences about Population Parameters
Hypothesis Testing

Statistical Test for \( \mu_1 - \mu_2 \)

(Independent samples, \( y_1 \) and \( y_2 \) approximately normal, \( \sigma_1^2 = \sigma_2^2 \))

**Test Statistic:**

\[
t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)
\]

where:

\[
s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}
\]

\( D_0 \) is a specified value, often ‘0’
**Statistical Test for 'μ₁ − μ₂'**

(Independent samples, y₁ and y₂ approximately normal, \( \sigma^2_1 = \sigma^2_2 \))

**Test Statistic:**

\[
\begin{align*}
t & = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\
& \sim t(n_1 + n_2 - 2)
\end{align*}
\]

where:

\[
s_p = \sqrt{\frac{(n_1-1)s^2_1 + (n_2-1)s^2_2}{n_1+n_2-2}}
\]

's_p' is a weighted average, that combine (pools) two independent estimates of 'σ'
Inferences about Population Parameters

Hypothesis Testing

**Statistical Test for 'μ₁ − μ₂'**

(Independent samples, \(y_1\) and \(y_2\) approximately normal, \(\sigma_1^2 = \sigma_2^2\))

**Test Statistic:**

\[
t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)
\]

- **Reject \(H_0\) if \(t \geq t_\alpha\)**
- **Reject \(H_0\) if \(t \leq -t_\alpha\)**
- **Reject \(H_0\) if \(|t| \geq t_\alpha^2\)**
100(1 - \alpha)\% confidence Interval for ‘\mu_1 - \mu_2’

(Independent samples, \(y_1\) and \(y_2\) approximately normal, \(\sigma_1^2 = \sigma_2^2\))

\[
(\overline{y}_1 - \overline{y}_2) \pm t_\alpha \frac{s_p}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]
Inferences about Population Parameters

Hypothesis Testing

- **Cluster-Effect**
  - Serial or Spatial Correlation, Paired Sample

- **Dependency in observations**

- **Skewness**
  - Heavy-tailed

- **Non-Normality**

- **Un-equal variances**

- **More Advanced methods such as longitudinal or spatial analysis, Paired Tests**

- **Non-Parametric Tests**

- **Approximate T-test**

Training Workshop on Statistical Data Analysis

8-21 July 2011

Afsaneh Yazdani
Inferences about Population Parameters
Hypothesis Testing

Statistical Test for \( \mu_1 - \mu_2 \)

(Independent samples, \( y_1 \) and \( y_2 \) approximately normal, \( \sigma_1^2 \neq \sigma_2^2 \))

**Test Statistic:**

\[
t' = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(df)
\]

where:

\[
df = \frac{(n_1-1)(n_2-1)}{(1-c)^2(n_1-1)+c^2(n_2-1)} \quad \text{and} \quad c = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}
\]
Inferences about Population Parameters
Hypothesis Testing

**Statistical Test for \( \mu_1 - \mu_2 \)**

(Independent samples, \( y_1 \) and \( y_2 \) approximately normal, \( \sigma_1^2 \neq \sigma_2^2 \))

**Test Statistic:** \( t' \sim t(df) \)

- Reject \( H_0 \) if \( t' \geq t_\alpha \)
- Reject \( H_0 \) if \( t' \leq -t_\alpha \)
- Reject \( H_0 \) if \( |t'| \geq \frac{t_\alpha}{2} \)
100(1 − α)% confidence Interval for ‘μ₁ − μ₂’

(Independent samples, y₁ and y₂ approximately normal, σ₁² ≠ σ₂²)

\[
(\bar{y}_1 - \bar{y}_2) \pm t'_{\alpha} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]
Statistical Test for ‘μ₁ − μ₂’

Independent Samples, Wilcoxon Rank Sum Test

1- Sort the data and replace the data value with its rank
2- Make the Test Statistic:

- when \( n_1, n_2 \leq 10 \) then \( T = \text{sum of the ranks in sample 1} \)
- when \( n_1, n_2 > 10 \) then \( z = \frac{T - \mu_T}{\sigma_T} \)

\[
\mu_T = \frac{n_1(n_1+n_2+1)}{2}, \quad \sigma_T = \sqrt{\frac{n_1n_2}{12}(n_1 + n_2 + 1)}
\]
Statistical Test for ‘$\mu_1 - \mu_2$’

Independent Samples, Wilcoxon Rank Sum Test

1- Sort the data and replace the data value with its rank

2- Make the Test Statistic:

- when $n_1, n_2 \leq 10$ then $T=$ sum of the ranks in sample 1

- when $n_1, n_2 > 10$ then $z = \frac{T - \mu_T}{\sigma_T}$

$\mu_T = \frac{n_1(n_1+n_2+1)}{2}$, and $\sigma_T = \frac{\sqrt{n_1n_2}}{12} (n_1 + n_2 + 1)$
Statistical Test for \( \mu_1 - \mu_2 \)

**Independent Samples, Wilcoxon Rank Sum Test**

1- Sort the data and replace the data value with its rank

2- Make the Test Statistic:
   - when \( n_1, n_2 \leq 10 \) then \( T = \text{sum of the ranks} \)
   - when \( n_1, n_2 > 10 \) then 
     \[
     z = \frac{T - \mu_T}{\sigma_T}
     \]

\[
\mu_T = \frac{n_1(n_1+n_2+1)}{2}, \quad \text{and} \quad \sigma_T = \sqrt{\frac{n_1n_2}{12}(n_1 + n_2 + 1)}
\]

Provided there are no tied ranks
Inferences about Population Parameters
Hypothesis Testing

Statistical Test for \( \mu_1 - \mu_2 \)
Independent Samples, Wilcoxon Rank Sum Test

Test Statistic:
\[
z = \frac{T - \mu_T}{\sigma_T}
\]

\( H_0 \): Two populations are identical

\( H_a \): Population 1 is shifted to the right of population 2
- Reject \( H_0 \) if \( z \geq z_{\alpha} \)

\( H_a \): Population 1 is shifted to the left of population 2
- Reject \( H_0 \) if \( z \leq -z_{\alpha} \)

\( H_a \): Population 1 and 2 are shifted from each other
- Reject \( H_0 \) if \( |z| \geq \frac{z_{\alpha}}{2} \)
Inferences about Population Parameters
Hypothesis Testing

**Statistical Test for \( \mu_d \)**
(Paired samples, \( y_1 - y_2 \) approximately normal)

**Test Statistic:**
\[
t = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}} \\
\sim t(n - 1)
\]

- \( H_0: \mu_d \leq D_0 \) \( H_a: \mu_d > D_0 \) \( \Rightarrow \) Reject \( H_0 \) if \( t \geq t_\alpha \)
- \( H_0: \mu_d \geq D_0 \) \( H_a: \mu_d < D_0 \) \( \Rightarrow \) Reject \( H_0 \) if \( t \leq -t_\alpha \)
- \( H_0: \mu_d = D_0 \) \( H_a: \mu_d \neq D_0 \) \( \Rightarrow \) Reject \( H_0 \) if \( |t| \geq \frac{t_\alpha}{2} \)
100(1 − \( \alpha \))\% confidence Interval for 
\( \mu_d \)

(Paired samples, \( y_1 - y_2 \) approximately normal)

\[
\bar{d} \pm t_{\alpha} \frac{s_d}{\sqrt{\frac{n}{2}}}
\]
Statistical Test for ‘μ₁ – μ₂’
Paired Samples, Wilcoxon Signed-Rank Test

1- Calculate differences of the pairs, subtract them from ‘D₀’ keep non-zero differences (n), sort the absolute values in increasing order and rank them.

2- Make the Test Statistic:
   - when n ≤ 50 then ‘T−’, ‘T+’, or ‘min(T−, T+)’ depending on Hₐ
   - when n > 50 then \( Z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \)
Inferences about Population Parameters

Hypothesis Testing

Statistical Test for $\mu_1 - \mu_2$

Paired Samples, Wilcoxon Signed-Rank Test

- Reject $H_0$ if $z < -z_{\alpha}$
- Reject $H_0$ if $z < -z_{\alpha}$
- Reject $H_0$ if $|z| < -z_{\alpha/2}$
### Inferences about Population Parameters

#### Hypothesis Testing

#### Statistical Test for Mean (Two-population)

<table>
<thead>
<tr>
<th>Means of Two-Populations</th>
<th>Independent Samples</th>
<th>Normal Population, or ‘n’ large</th>
<th>$\sigma_1 = \sigma_2$</th>
<th>T-Test (Equal Variances)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Means of Two-Populations</th>
<th>Independent Samples</th>
<th>Normal Population or ‘n’ large</th>
<th>$\sigma_1 \neq \sigma_2$</th>
<th>T-Test (Unequal Variances)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Means of Two-Populations</th>
<th>Independent Samples</th>
<th>Non-normal Population</th>
<th>Wicoxon Rank Sum Test</th>
</tr>
</thead>
</table>
Inferences about Population Parameters

Hypothesis Testing

**Statistical Test for Mean (Two-population)**

- Means of Two-Populations
  - Paired Samples
  - Normal Population, or ‘n’ large
  - Paired T-Test

- Means of Two-Populations
  - Paired Samples
  - Non-normal Population
  - Wicoxon Signed Rank Test
Inferences about Population Parameters
Hypothesis Testing

Statistical Test for \( \frac{\sigma_1^2}{\sigma_2^2} \)

(Normal Population)

Test Statistic:

\[
F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1)
\]

\[
\begin{align*}
H_0: \sigma_1^2 &\leq \sigma_2^2 \\
H_a: \sigma_1^2 &> \sigma_2^2
\end{align*}
\]

- Reject \( H_0 \) if \( F \geq F_{\alpha, df_1, df_2} \)

\[
\begin{align*}
H_0: \sigma_1^2 &= \sigma_2^2 \\
H_a: \sigma_1^2 &\neq \sigma_2^2
\end{align*}
\]

- Reject \( H_0 \) if \( F \leq F_{1-\alpha/2, df_1, df_2} \) or if \( F \geq F_{\alpha/2, df_1, df_2} \)

Training Workshop on Statistical Data Analysis
8-21 July 2011
Afsaneh Yazdani
Inferences about Population Parameters

100(1 – α)% confidence Interval for $\frac{\sigma_1^2}{\sigma_2^2}$

$$\frac{s_1^2}{s_1^2 F_{\alpha, df_1, df_2}} \frac{1}{s_2^2 F_{\frac{\alpha}{2}, df_2, df_1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_1^2 F_{\frac{\alpha}{2}, df_2, df_1}}$$