Introduction to

Statistical Data Analysis III

JULY 2011

Afsaneh Yazdani

Preface

Major branches of Statistics:

- Descriptive Statistics

- Inferential Statistics

Preface

What is Inferential Statistics?

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

Preface

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.

Inferences about Population Parameters

What is Inferential Statistics?

Methods for making inferences about parameters fall into two categories:

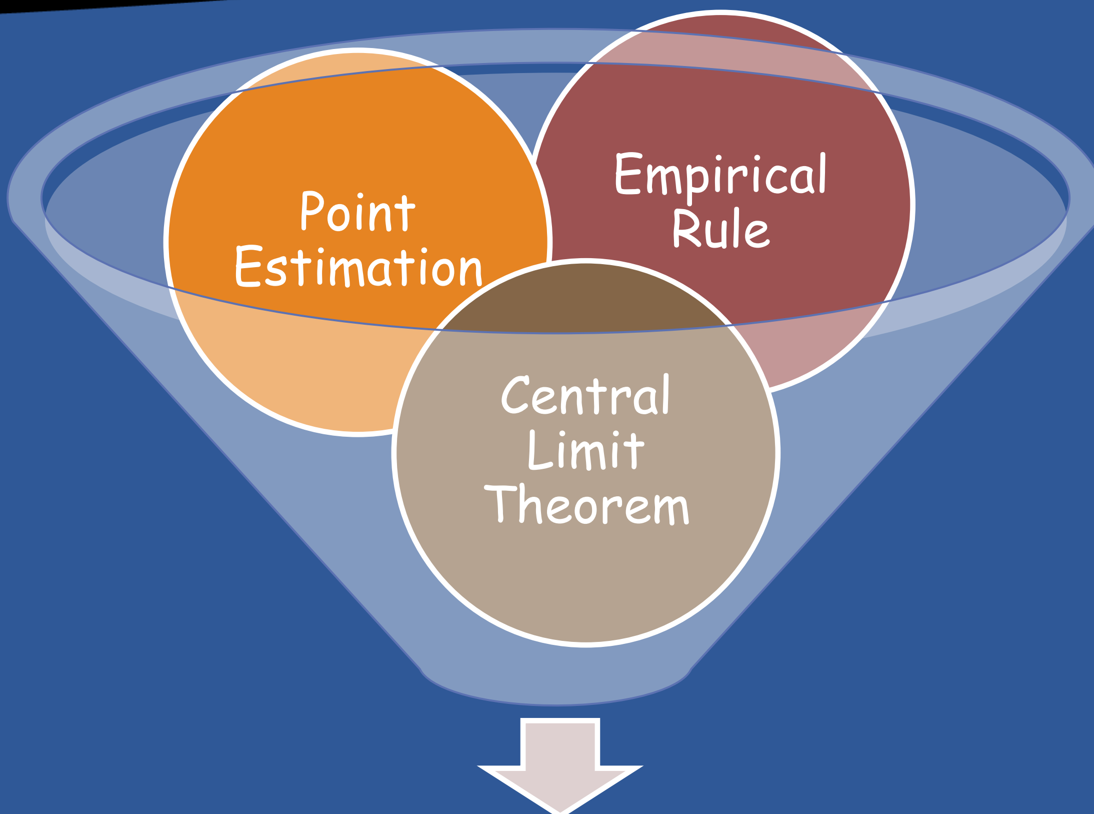
- Estimating the population parameter
- Hypothesis Testing about a population parameter

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

$\hat{\theta} \pm A * SE(\hat{\theta})$, **where 'A' 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:**

Estimation of 'μ':

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

For large 'n',
$\bar{y}$ is approximately normally distributed with mean 'μ' and standard error $\frac{\sigma}{\sqrt{n}}$

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 **σ is known**

Estimation of ‘μ’:

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

with level of confidence **95%** when **σ 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**

Inferences about Population Parameters

Estimation

Estimation of 'μ':

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

when **σ is unknown**

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 'μ'

('σ' known) when sampling from a normal population or 'n' large

$$\left(\bar{y} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\ ,\bar{y} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$$

Inferences about Population Parameters

Estimation

$100(1-\alpha)\%$ confidence Interval for 'μ'

('σ' unknown) when sampling from a normal population or 'n' large

$$\left(\bar{y} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\ ,\ \bar{y} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$$

Goodness of inference for interval estimation:

❶ Confidence coefficient

❷ Width of the confidence interval

Inferences about Population Parameters

Estimation

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**

Inferences about Population Parameters

Estimation

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

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 \sigma^2}{E^2}$$

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

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 \sigma^2}{E^2}$$

Estimate using information from prior survey

Inferences about Population Parameters

Estimation

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

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 \sigma^2}{E^2}$$

Estimate using

s= $\frac{\text{range}}{4}$

Sample Size for Estimation of ‘μ’:

Consideration 1:

Desired Confidence Level (z-value)

Consideration 2:

the tolerable error that establishes desired width of the interval

Consideration 3:

Standard Deviation (σ)

appropriate sample size for estimating ‘μ’ using a confidence interval

$$\left(\bar{y} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\ ,\bar{y} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

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

$$\left(\bar{y} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\ , \bar{y} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

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

$$\left(\bar{y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\ ,\bar{y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

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.

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.

Inferences about Population Parameters

Hypothesis Testing

Test Statistic

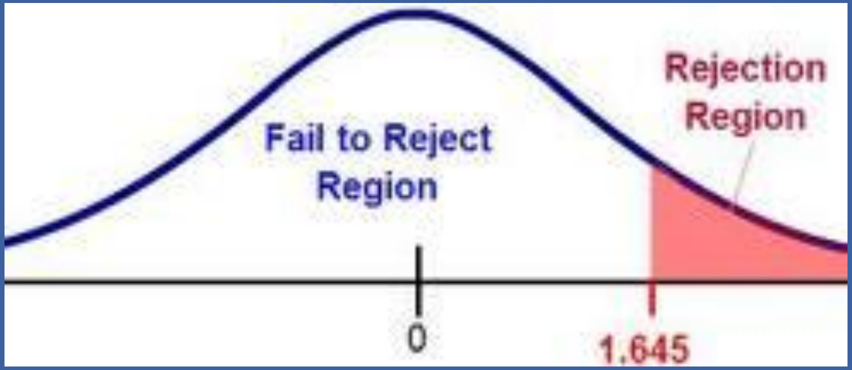
The quantity computed from the sample data, that helps to decide whether or not the data support the research hypothesis.

Inferences about Population Parameters

Hypothesis Testing

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

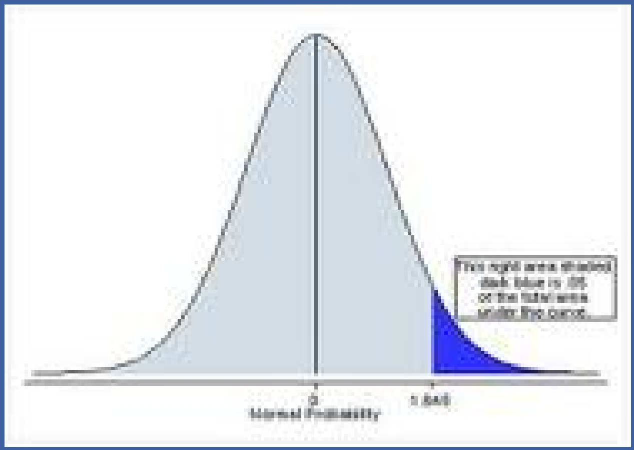
- One-Tailed Test
- Two-Tailed Test

Inferences about Population Parameters

Hypothesis Testing

Statistical Tests

- One-Tailed Test



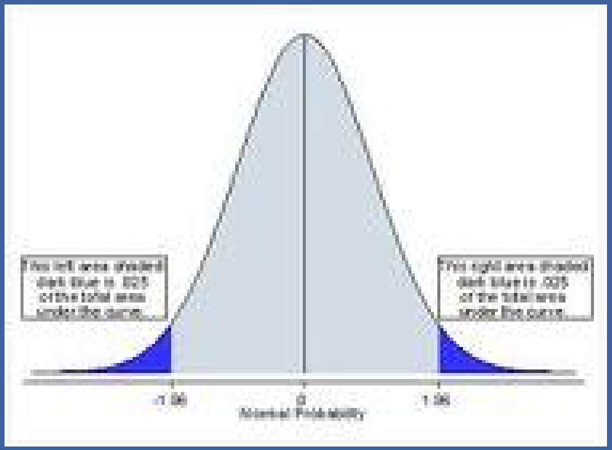
The rejection region is located in only one tail of the sampling distribution of test statistic

H_a: $\theta < \theta_0$ Or

H_a: $\theta > \theta_0$

Statistical Tests

- **Two-Tailed Test**



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

$$H_a\colon \theta \neq \theta_0$$

Errors in Statistical Tests

- Type I Error
- Type II Error

Errors in Statistical Tests

- Type I Error
- Type II Error

Rejecting the null hypothesis when it is true. The probability of a Type I error is denoted by ‘α’.

Errors in Statistical Tests

- Type I Error
- Type II Error

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

Errors in Statistical Tests

		Null Hypothesis	
		True	**False**
Decision	**Reject**	Type I α	Correct $(1-\alpha)$
	Accept	Correct $(1-\beta)$	Type II β

Errors in Statistical Tests

		Null Hypothesis	
		True	**False**
Decision	**Reject**	Type I α	Correct $(1-\alpha)$
	Accept	Correct $(1-\beta)$	Type II β

The probabilities associated with Type I and Type II errors are inversely related. For a fixed sample size '*n*', when 'α' decreases 'β' will increase and vice versa

Errors in Statistical Tests

		Null Hypothesis	
		True	**False**
Decision	**Reject**	Type I α	Correct $(1-\alpha)$
	Accept	Correct $(1-\beta)$	Type II β

Usually 'α' is specified to locate the Rejection Region

Errors in Statistical Tests

		Null Hypothesis	
		True	**False**
Decision	**Reject**	Type I α	Correct $(1-\alpha)$
	Accept	Correct $(1-\beta)$	Type II β

Power of Test

Effectiveness of a statistical test is measured by:

Magnitudes 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

Statistical Tests (Drawing Conclusion)

Traditional Approach:

- Using Statistic Test, two types of errors, their probability α, β

Statistical Tests

Traditional Approach:

- Using Statistic Test, two types of errors, their probability α, β

The problem with this approach is that if other researchers want to apply the results of your study using a different value for 'α' then they must compute a new rejection region.

Statistical Tests – Alternative Approach

Using Level of Significance/P-Value

Smallest size of 'α' at which H_0 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 ‘α’ 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.

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.

Statistical Tests – Alternative Approach

Decision Rule for Hypothesis Testing Using P-Value

$\mathbf{P-Value \leq \alpha}$	• Reject H_0
$\mathbf{P-Value > \alpha}$	• Fail to Reject H_0

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: $$z = \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}(0, 1)$$

$\begin{cases} H_0: \mu \le \mu_0 \\ H_a: \mu > \mu_0 \end{cases}$	• Reject H_0 if $z \ge z_\alpha$
$\begin{cases} H_0: \mu \ge \mu_0 \\ H_a: \mu < \mu_0 \end{cases}$	• Reject H_0 if $z \le -z_\alpha$
$\begin{cases} H_0: \mu = \mu_0 \\ H_a: \mu \ne \mu_0 \end{cases}$	• Reject H_0 if $\lvert z \rvert \ge z_{\frac{\alpha}{2}}$

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	$1-\beta(\mu_a) = 1 - Pr\left(z \le z_\alpha - \frac{\lvert\mu_0-\mu_a\rvert}{\frac{\sigma}{\sqrt{n}}}\right)$
Two-Tailed Test	$1-\beta(\mu_a) \approx 1 - Pr\left(z \le z_{\frac{\alpha}{2}} - \frac{\lvert\mu_0-\mu_a\rvert}{\frac{\sigma}{\sqrt{n}}}\right)$

Statistical Test for population mean ‘μ’

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

$\begin{cases} H_0: \mu \le \mu_0 \\ H_a: \mu > \mu_0 \end{cases}$	• P-Value: $\Pr(z \ge computed\ z)$
$\begin{cases} H_0: \mu \ge \mu_0 \\ H_a: \mu < \mu_0 \end{cases}$	• P-Value: $\Pr(z \le computed\ z)$
$\begin{cases} H_0: \mu = \mu_0 \\ H_a: \mu \ne \mu_0 \end{cases}$	• P-Value: $2\Pr(z \ge \lvert computed\ z \rvert)$

Statistical Test for population mean 'μ'

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

Test Statistic: $T = \dfrac{\bar{y}-\mu_0}{\frac{s}{\sqrt{n}}} \sim t(n-1)$

$\begin{cases} H_0: \mu \le \mu_0 \\ H_a: \mu > \mu_0 \end{cases}$	• Reject H_0 if $t \ge t_\alpha$
$\begin{cases} H_0: \mu \ge \mu_0 \\ H_a: \mu < \mu_0 \end{cases}$	• Reject H_0 if $t \le -t_\alpha$
$\begin{cases} H_0: \mu = \mu_0 \\ H_a: \mu \ne \mu_0 \end{cases}$	• Reject H_0 if $\lvert t\rvert \ge t_{\frac{\alpha}{2}}$

Statistical Test for population mean 'μ'

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

$\begin{cases} H_0: \mu \leq \mu_0 \\ H_a: \mu > \mu_0 \end{cases}$	• P-Value: $\Pr(t \geq computed\ t)$
$\begin{cases} H_0: \mu \geq \mu_0 \\ H_a: \mu < \mu_0 \end{cases}$	• P-Value: $\Pr(t \leq computed\ t)$
$\begin{cases} H_0: \mu = \mu_0 \\ H_a: \mu \neq \mu_0 \end{cases}$	• P-Value: $2\Pr(t \geq \lvert computed\ t \rvert)$

Non-normal distribution of population

'2' ISSUES to consider:

- Skewed Distribution
- Heavy-tailed Distribution

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

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: $W_i = y_i - M_0$, $B =$ No. of Positive $W_i s$

$B \sim Binom(n, \pi)$

$\begin{cases} H_0: M \le M_0 \\ H_a: M > M_0 \end{cases}$	• Reject H_0 if $B \ge n - C_{\alpha(1),n}$
$\begin{cases} H_0: M \ge M_0 \\ H_a: M < M_0 \end{cases}$	• Reject H_0 if $B \le C_{\alpha(1),n}$
$\begin{cases} H_0: M = M_0 \\ H_a: M \ne M_0 \end{cases}$	• Reject H_0 if $B \le C_{\alpha(2),n}$ or $B \ge n - C_{\alpha(2),n}$

Statistical Test for population median

(Approximation)

Test Statistic: $B_{st} = \dfrac{B - {}^{n}/_{2}}{\sqrt{{}^{n}/_{4}}}, \quad B_{st} \sim N(0, 1)$

$\begin{cases} \mathrm{H_0}: M \le M_0 \\ \mathrm{H_a}: M > M_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $B_{st} \ge z_{\alpha}$ with P-value $Pr(z \ge B_{st})$
$\begin{cases} \mathrm{H_0}: M \ge M_0 \\ \mathrm{H_a}: M < M_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $B_{st} \le z_{\alpha}$ with P-value $Pr(z \le B_{st})$
$\begin{cases} \mathrm{H_0}: M = M_0 \\ \mathrm{H_a}: M \ne M_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $\lvert B_{st}\rvert \ge z_{\frac{\alpha}{2}}$ with P-value $2Pr(z \ge \lvert B_{st}\rvert)$

Statistical Test for Mean (one-population)

Mean of One-Population → Normal Population, or ‘n’ large → $'\sigma'$ known → Test using z

Mean of One-Population → Normal Population, or ‘n’ large → $'\sigma'$ unknown → T-Test

Mean of One-Population → Non-normal Population → Sign-Test

Statistical Test for population Variance

(Normal Population)

Test Statistic:

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

$\begin{cases} H_0: \sigma^2 \le \sigma_0^2 \\ H_a: \sigma^2 > \sigma_0^2 \end{cases}$	• Reject H_0 if $\chi^2 > \chi^2_{U,\alpha}$
$\begin{cases} H_0: \sigma^2 \ge \sigma_0^2 \\ H_a: \sigma^2 < \sigma_0^2 \end{cases}$	• Reject H_0 if $\chi^2 < \chi^2_{L,\alpha}$
$\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_a: \sigma^2 \ne \sigma_0^2 \end{cases}$	• Reject H_0 if $\chi^2 > \chi^2_{U,\frac{\alpha}{2}}$ or $\chi^2 > \chi^2_{L,\frac{\alpha}{2}}$

Inferences about Population Parameters

$100(1-\alpha)\%$ confidence Interval for 'σ^2' (or σ)

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

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

Inferences about Population Parameters

Hypothesis Testing

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

Inferences about Population Parameters

Hypothesis Testing

Theorem

If two independent random variables y_1 and y_2 are normally distributed with means and variances (μ_1, σ_1^2) and (μ_2, σ_2^2) respectively, then

$$(y_1 - y_2) \sim N\left((\mu_1 - \mu_2), \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

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

Two independent large samples

$$(\bar{y}_1 - \bar{y}_2) \dot{\sim} N\left((\mu_1 - \mu_2), \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

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:

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

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'

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:

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

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

Statistical Test for '$\mu_1 - \mu_2$'

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

Test Statistic: $t = \dfrac{(\bar{y}_1 - \bar{y}_2) - D_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$

$\begin{cases} H_0: \mu_1 - \mu_2 \le D_0 \\ H_a: \mu_1 - \mu_2 > D_0 \end{cases}$	• Reject H_0 if $t \ge t_\alpha$
$\begin{cases} H_0: \mu_1 - \mu_2 \ge D_0 \\ H_a: \mu_1 - \mu_2 < D_0 \end{cases}$	• Reject H_0 if $t \le -t_\alpha$
$\begin{cases} H_0: \mu_1 - \mu_2 = D_0 \\ H_a: \mu_1 - \mu_2 \ne D_0 \end{cases}$	• Reject H_0 if $\lvert t \rvert \ge t_{\frac{\alpha}{2}}$

Inferences about Population Parameters

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

'$\mu_1 - \mu_2$'

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

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\frac{\alpha}{2}}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Inferences about Population Parameters

Hypothesis Testing

Cause	Problem	Solution
Cluster-Effect Serial or Spatial Correlation, Paired Sample	Dependency in observations	More Advanced methods such as longitudinal or spatial analysis, Paired Tests
Skewness Heavy-tailed	Non-Normality	Non-Parametric Tests
Un-equal variances		Approximate T-test

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:

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

where:

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(1-c)^2(n_1 - 1) + c^2(n_2 - 1)} \text{ and } c = \frac{s_1^2/n_1}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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: $t' \dot{\sim} \mathrm{t}(df)$

$\begin{cases} \mathrm{H_0}: \mu_1 - \mu_2 \leq D_0 \\ \mathrm{H_a}: \mu_1 - \mu_2 > D_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $t' \geq t_\alpha$
$\begin{cases} \mathrm{H_0}: \mu_1 - \mu_2 \geq D_0 \\ \mathrm{H_a}: \mu_1 - \mu_2 < D_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $t' \leq -t_\alpha$
$\begin{cases} \mathrm{H_0}: \mu_1 - \mu_2 = D_0 \\ \mathrm{H_a}: \mu_1 - \mu_2 \neq D_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $\lvert t' \rvert \geq t_{\frac{\alpha}{2}}$

Inferences about Population Parameters

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

'$\mu_1 - \mu_2$'

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

$$(\bar{y}_1 - \bar{y}_2) \pm t'_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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 $\sigma_T = \sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}$

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 $z = \frac{T - \mu_T}{\sigma_T}$ Normal Approximation

$\mu_T = \frac{n_1(n_1+n_2+1)}{2}$, and $\sigma_T = \sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}$

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

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

$$\mu_T = \frac{n_1(n_1+n_2+1)}{2}, \text{ and } \sigma_T = \sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}$$

Provided there are no tied ranks

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 $\lvert z \rvert \geq z_{\frac{\alpha}{2}}$

Statistical Test for 'μ_d'

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

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

$\begin{cases} H_0: \mu_d \le D_0 \\ H_a: \mu_d > D_0 \end{cases}$	• Reject H_0 if $t \ge t_\alpha$
$\begin{cases} H_0: \mu_d \ge D_0 \\ H_a: \mu_d < D_0 \end{cases}$	• Reject H_0 if $t \le -t_\alpha$
$\begin{cases} H_0: \mu_d = D_0 \\ H_a: \mu_d \ne D_0 \end{cases}$	• Reject H_0 if $\lvert t \rvert \ge t_{\frac{\alpha}{2}}$

Inferences about Population Parameters

$100(1-\alpha)\%$ confidence Interval for 'μ_{d}'

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

$$\bar{d} \pm t_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}$$

Statistical Test for '$\mu_1 - \mu_2$'

Paired Samples, Wilcoxon Signed-Rank Test

1- Calculate differences of the pairs, subtract them from 'D_0' keep non-zero differences (n), sort the absolute values in increasing order and rank them.

2- Make the Test Statistic:

- when $\mathbf{n} \leq 50$ then 'T_-' , 'T_+', or '$min(T_-, T_+)$' depending on H_a
- when $\mathbf{n} > 50$ then $\mathbf{z} = \dfrac{\mathrm{T} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$

Statistical Test for '$\mu_1 - \mu_2$'

Paired Samples, Wilcoxon Signed-Rank Test

$\begin{cases} \mathrm{H_0}: M = D_0 \\ \mathrm{H_a}: M > D_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $z < -z_\alpha$
$\begin{cases} \mathrm{H_0}: M = D_0 \\ \mathrm{H_a}: M < D_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $z < -z_\alpha$
$\begin{cases} \mathrm{H_0}: M = D_0 \\ \mathrm{H_a}: M \neq D_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $\lvert z \rvert < -z_{\frac{\alpha}{2}}$

Statistical Test for Mean (Two-population)

Means of Two-Populations	Independent Samples	Normal Population, or ‘n’ large	$\sigma_1 = \sigma_2$	T-Test (Equal Variances)
Means of Two-Populations	Independent Samples	Normal Population or ‘n’ large	$\sigma_1 \neq \sigma_2$	T-Test (Unequal Variances)
Means of Two-Populations	Independent Samples	Non-normal Population		Wicoxon Rank Sum Test

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

Statistical Test for ' $\frac{\sigma_1^2}{\sigma_2^2}$ '

(Normal Population)

Test Statistic: $F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1)$

$\begin{cases} H_0: \sigma_1^2 \le \sigma_2^2 \\ H_a: \sigma_1^2 > \sigma_2^2 \end{cases}$	• Reject H_0 if $F \ge F_{\alpha, df_1, df_2}$
$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_a: \sigma_1^2 \ne \sigma_2^2 \end{cases}$	• Reject H_0 if $F \le F_{1-\frac{\alpha}{2}, df_1, df_2}$ or if $F \ge F_{\frac{\alpha}{2}, df_1, df_2}$

Inferences about Population Parameters

$100(1-\alpha)\%$ confidence Interval for '$\frac{\sigma_1^2}{\sigma_2^2}$'

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