Inference

Madiba Hudson-Quansah

Contents

Chapter 1

Module 13: Inference

Definition 1.0.1: Statistical Inference

Inferring something about a population from a sample.

Definition 1.0.2: Point Estimation

Estimating an unknown parameter using a single number, calculated from the sample data.

Definition 1.0.3: Interval Estimation

Estimating an unknown parameter using an Interval of values that is likely to contain the true value of the parameter, and state how confident we are that the interval contains the true value.

Definition 1.0.4: Hypothesis Testing

Making decisions about the population parameter based on the sample data.

1.1 Inference for One Variable

Depending on the type of variable we are interested in the population parameter we infer about changes:

- Categorical : Population Proportion p
- Quantitative : Population Mean μ

Chapter 2

Module 14: Estimation

2.1 Point Estimation

Definition 2.1.1: Point Estimator

A statistic that provides an estimate of a population parameter.

The point estimator also changes based on the type of variable examined:

- Categorical: Sample Proportion / $\hat{p}$
- Quantitative: Sample Mean / $\bar{x}$

Note:-

The larger the sample size, the more accurate the point estimate.

2.2 Interval Estimation

Definition 2.2.1: Confidence Interval

An interval of values that is likely to contain the true value of the population parameter.

Interval estimation is based on the point estimate and the margin of error.

2.2.1 Confidence Intervals for the Population Mean

For a quantitative variable with a normally distributed sample mean distribution due to the Central Limit Theorem, construing a 95% confidence interval consists of the following steps:

- Identify mean $\overline{X}$, which for a sample mean distribution is approximately equal to μ
- Find the standard deviation S of the sample mean distribution, $\frac{\sigma}{\sqrt{n}}$
- Find $\overline{X} \pm 2S$, which are your upper and lower bounds of the confidence interval

Therefore generally the confidence interval is:

$$\overline{x} \pm 2 \times \frac{\sigma}{\sqrt{n}}$$

2.2.1.1 Other Levels of Confidence

Constructing a 99% confidence interval for μ can be done using:

$$\overline{x} \pm 2.576 \times \frac{\sigma}{\sqrt{n}}$$

And a 90% confidence interval for μ can be found using:

$$\overline{x} \pm 1.645 \times \frac{\sigma}{\sqrt{n}}$$

To calculate the confidence interval for any level of confidence, we use the z-score of the area of half the α of the confidence level, i.e.:

$$z^* = z_{\frac{\alpha}{2}}$$

Where alpha is

$$\alpha = 1 - C$$

Or

$$\alpha = 1 + C$$

2.2.2 General Structure of Confidence Intervals

A confidence interval has the following form:

$$\overline{x} \pm z^* \times \frac{\sigma}{\sqrt{n}}$$

Where z^* is general notation for the multiplier that depends on the level of confidence.
The confidence interval can then also be expressed in the form:

$$\overline{x} \pm m$$

Where $m = z^* \times \frac{\sigma}{\sqrt{n}}$ and $\overline{x}$ is the point estimator for the unknown population mean μ
m is called the margin of error, since it represents the maximum estimation error for a given level of confidence.

Note:-

A larger sample size makes for a smaller margin of error.

2.2.3 Sample Size Calculations

The sample size required to estimate the population mean μ with a margin of error m at a level of confidence C can be found using:

$$n = \left(\frac{z^*\sigma}{m}\right)^2$$

Which is rounded up to the nearest whole number.

2.2.4 When σ is unknown

When the population standard deviation σ is unknown, the sample standard deviation s is used instead, but as a result we need to use a different set of confidence multipliers t^*, associated with the t distribution. The interval is therefore:

$$\overline{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

These multipliers depend not only on the level of confidence, but also on the sample size n.
For large values of n, the t distribution approaches the standard normal distribution, and the t^* multipliers, therefore the z^* multipliers can be used, i.e. $t^* \approx z^*$ and the confidence interval becomes:

$$\overline{x} \pm z^* \times \frac{s}{\sqrt{n}}$$

2.3 Confidence Intervals for the Population Proportion

For a categorical variable, the population proportion p can be estimated using the sample proportion $\hat{p}$, and the margin of error m, i.e. the confidence interval is:

$$\hat{p} \pm m$$

Where m is:

$$m = z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Therefore:

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

2.3.1 Sample Size Calculations

The sample size required to estimate the population proportion p with a margin of error m at a level of confidence C can be found using:

$$n = p \times \left(1-p\right) \times \left(\frac{z^*}{m}\right)^2$$

Chapter 3

Module 15: Hypothesis Testing

3.1 Introduction

Definition 3.1.1: Hypothesis Testing

Assessing evidence provided by the data in favour of against some claim about the population.

The process of statistical hypothesis testing is as follows:

- We start with two claims about the behaviour of a population, where the claim usually contradict each other.
- Choose a sample and collect and summarize relevant data.
- Determine how likely it is to observe data, like the data we get had claim 1 been true
- Based on the results we make one of two conclusions:
  - If we find that if claim 1 were true it would extremely unlikely to observe the data we observed, then we have strong evidence against claim 1 and can reject it in favour of claim 2
  - If we find that if claim 1 were true it would not be extremely unlikely to observe the data we observed, then we do not have enough evidence against claim 1, and cannot reject it in favour of claim 2.

In the terminology of hypothesis testing Claim 1 is termed as the **null hypothesis**, denoted by H_0, and Claim 2 is termed as the **alternative hypothesis**, denoted by H_a.

Null Hypothesis No change from the status quo / No relationship

Alternative Hypothesis There is a change from the status quo / There is a relationship

Determining how likely it is to observe data like the data we would of gotten if claim 1 were true, is termed as finding its *p-value*

In making a decision about the null hypothesis, we use the p-value to determine the strength of the evidence against the null hypothesis. The smaller the p-value, the stronger the evidence against the null hypothesis, i.e.:

- If $p-\text{value} < \alpha$ (usually 0.05), we can reject H_0 and accept H_a, as the evidence against H_0 is strong.
- If $p-\text{value} > \alpha$ (usually 0.05), we do not have enough evidence against H_0 and cannot reject it.

3.2 Hypothesis Testing for Population Proportion

3.2.1 Step 1 - Stating the Hypothesis

In stating the null and alternative hypothesis for a population proportion, the null hypothesis always takes the form of equality, and the alternative hypothesis takes the form of inequality / difference that can either be one-sided or two-sided alternatives, where one-sided alternatives are used when we are interested in a specific direction of change, and two-sided alternatives are used when the direction of change is irrelevant.

- Null Hypothesis: $H_0 : p = p_0$
- Alternative Hypothesis:
  - $H_a : p \neq p_0$ (two-sided)
  - $H_a : p > p_0$ (one-sided)
  - $H_a : p < p_0$ (one-sided)

3.2.2 Step 2 - Collecting and Summarizing Data

3.3 Hypothesis Testing for Population Mean

In hypothesis testing for a population mean there are two cases:

- σ is known and we use the z-test for the population mean μ
- σ is unknown and we use the t-test for the population mean μ

3.3.1 z-test for the Population Mean

3.3.1.1 Step 1 - Stating the Hypothesis

The null and alternative hypothesis for the population mean μ takes the same form as the population mean, i.e.:

- Null Hypothesis: $H_0 : \mu = \mu_0$
- Alternative Hypothesis:
  - $H_a : \mu \neq \mu_0$ (two-sided)
  - $H_a : \mu < \mu_0$ (one-sided)
  - $H_a : \mu > \mu_0$ (one-sided)

Where μ_0 is the null value

3.3.1.2 Step 2 - Collecting and Summarizing Data

In this step we calculate the sample mean $\overline{x}$, and relevant sample statistic and summarize the data with a test statistic. This test statistic is the z-score of the sample mean, assuming H_0 is true, i.e:

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Where $\overline{x}$ is the sample mean, μ_0 is the null value, σ is the population standard deviation, And n is the sample size, where $n > 30$

The conditions needed to perform the z-test for the population mean are:

- The sample is random
- The variable varies normally in the population
- The variable does not vary normally in the population, but the sample size is large enough

3.3.1.3 Step 3 - Finding the p-value

The p-value is the probability of observing a sample mean as extreme as the one observed, assuming H_0 is true. The p-value is calculated using the z-score of the sample mean, i.e.:

- $H_a : \mu \neq \mu_0 \implies p\text{-value} = 2\mathbb{P}\left(Z \geqslant \mid z \mid\right)$
- $H_a : \mu < \mu_0 \implies p\text{-value } = \mathbb{P}\left(Z \leqslant z\right)$
- $H_a : \mu > \mu_0 \implies p\text{-value} = \mathbb{P}\left(Z \geqslant z\right)$

3.3.1.4 Step 4 - Drawing Conclusions

In drawing conclusions about the null hypothesis, we compare the p-value to the level of significance α, and state our conclusion as follows:

- p-value $> \alpha$ - We do not have enough evidence against H_0 and cannot reject it
- p-value $< \alpha$ - We have enough evidence against H_0 and can reject it in favour of H_a

3.3.2 t-test for Population mean

The first, second, and fourth steps in using the t-test for population mean are identical to that of the z-test, but the t-score being calculated as:

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Where the denominator $\frac{s}{\sqrt{n}}$ is the standard error of $\overline{X}$.
This value is then compared to the t-distribution with $n - 1$ degrees of freedom, and the p-value is calculated

3.3.2.1 Step 3 - Finding the p-value

The p-value is calculated using the t-score of the sample mean, i.e.:

- $H_a : \mu \neq \mu_0 \implies p\text{-value} = 2\mathbb{P}\left(t\left(n-1\right) \geqslant \mid t \mid\right)$
- $H_a : \mu < \mu_0 \implies p\text{-value} = \mathbb{P}\left(t\left(n-1\right) \leqslant t\right)$
- $H_a : \mu > \mu_0 \implies p\text{-value} = \mathbb{P}\left(t\left(n-1\right) \geqslant t\right)$

3.4 Type I and Type II Errors

3.4.1 Type I Error

Definition 3.4.1: Type I Error

Rejecting the null hypothesis when it is true

3.4.1.1 The Probability of Type I Error

The probability of making a Type I error is denoted by α, and is the level of significance of the test, i.e.:

$$\alpha = \mathbb{P}\left(\text{Type I Error}\right)$$

3.4.2 Type II Error

Definition 3.4.2: Type II Error

Failing to reject the null hypothesis when it is false

3.4.2.1 The Probability of a Type II Error

The probability of making a Type II error is inversely related to the probability of making a Type I error, and is denoted by β, i.e.:

$$\beta = \mathbb{P}\,(\text{Type II Error})$$

Chapter 4

Module 16 and 17: Inference for Relationships

Note:-

Large Enough - $n > 30$

4.1 Case $C \to Q$

In this case, we are interested in the relationship between a categorical variable and a quantitative variable. To make inferences about the relationship between the two variables, we compare the means of the quantitative variable across the categories of the categorical variable. The method used depends on the number of categories of the categorical variable., i.e:

- $k = 2$: We use the two-sample t-test
- $k > 2$ We use the ANOVA test

Where k is the number of categories of the categorical variable.

Furthermore when $k = 2$, there are two cases to consider:

- Independent groups /samples - Where the two categories are independent of each other.
- Paired groups / Dependent samples - Where the two categories are dependent on each other in some way or matched pairs.

4.1.1 Two sample t-test - Independent Groups

4.1.1.1 Step 1 - Stating the Hypothesis

In this case the hypothesis represents the difference in the population means $\left(\mu_1 \text{ and } \mu_2\right)$ of the two categories of the categorical variable.i.e.

- Null Hypothesis: $H_0 : \mu_1 = \mu_2$ / $H_0 : \mu_1 - \mu_2 = 0$
- Alternative Hypothesis:
  - $H_a : \mu_1 - \mu_2 \neq 0$ (two-sided)
  - $H_a : \mu_1 - \mu_2 > 0$ (one-sided)
  - $H_a : \mu_1 - \mu_2 < 0$ (one-sided)

4.1.1.2 Step 2 - Collecting and Summarizing Data

In this step we calculate the test statistic, which is in this case the t-score of the difference in the sample means, i.e.:

$$t = \frac{(\overline{y_1} - \overline{y_2}) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

$\overline{y_1}$ and $\overline{y_2}$ are the sample means of the samples from populations 1 and 2 respectively.

s_1 and s_2 are the sample standard deviations of the samples from populations 1 and 2 respectively.

n_1 and n_2 are the sample sizes of the two samples.

The conditions needed to perform the two-sample t-test for independent groups are:

- The two samples are independent
- Either:
  - Both populations are normal and both samples are random
  - Either population is not normal but the sample sizes are large enough

4.1.1.3 Step 3 - Finding the p-value

The p-value is the probability of observing a difference in sample means as extreme as the one observed, assuming H_0 is true.

4.1.1.4 Step 4 - Drawing Conclusions

In drawing conclusions about the null hypothesis, we compare the p-value to the level of significance α, and state our conclusion as follows:

- p-value $> \alpha$ - We do not have enough evidence against H_0 and cannot reject it.
- p-value $< \alpha$ - We have enough evidence against H_0 and can reject it in favour of H_a.

4.1.2 Two sample t-test - Matched Pairs

4.1.2.1 Step 1 - Stating the Hypothesis

For matched pairs, the hypothesis represents the difference in the population means (μ_d) of the two quantitative variables, i.e.:

- Null Hypothesis - $H_0 := \mu_d$
- Alternative Hypothesis:
  - $H_a : \mu_d \neq 0$ (two-sided)
  - $H_a : \mu_d > 0$ (one-sided)
  - $H_a : \mu_d < 0$ (one-sided)

4.1.2.2 Step 2 - Collecting and Summarizing Data

In this step we calculate the test statistic, which is in this case the t-score of the difference in the sample means, i.e.:

$$t = \frac{\overline{x_d} - 0}{\frac{s_d}{\sqrt{n}}}$$

Where $\overline{x_d}$ is the sample mean of the differences, and s_d is the sample standard deviation of the differences.

4.1.2.3 Step 3 - Finding the p-value

The p-value is the probability of observing a difference in sample means as extreme as the one observed, assuming H_0 is true, this is calculated using the t-score of the sample mean with the t-distribution with $n - 1$ degrees of freedom.

4.1.2.4 Step 4 - Drawing Conclusions

In drawing conclusions about the null hypothesis, we compare the p-value to the level of significance α, and state our conclusion as follows:

- p-value $> \alpha$ - We do not have enough evidence against H_0 and cannot reject it.
- p-value $< \alpha$ - We have enough evidence against H_0 and can reject it in favour of H_a.

4.1.2.5 Confidence Interval for μ_d

The point estimator used in this confidence interval is the sample mean of the differences $\overline{x_d}$

If the null value 0 falls outside the confidence interval, then we can reject H_0

If the null value 0 falls inside the confidence interval, then we cannot reject H_0

4.1.3 ANOVA (Analysis of Variance) Test

The ANOVA test is used to compare the means of three or more populations, and is used to determine if there is a difference in the means of the populations. The test statistic used in the ANOVA test is the F-score, which is the ratio of the variance between the sample means to the variance within the samples, i.e.:

4.1.3.1 Step 1 - Stating the Hypothesis

The null and alternative hypothesis for the ANOVA test are:

- Null Hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ / There is no relationship between the categorical variable and the quantitative variable
- Alternative Hypothesis: $H_a :$ not all the μ's are equal

4.1.3.2 Step 2 - Checking conditions and finding the test statistic

For the ANOVA test the test statistic is the F-score, which is the ratio of the variance between the sample means to the variance within the samples, i.e.:

$$F = \frac{\text{Variation among sample means}}{\text{Variation within samples}}$$

The larger the F-score, the stronger the evidence against the null hypothesis. If the variation within samples is large then the F-score will be small, and vice versa.

The conditions needed to perform the ANOVA test are:

- Random samples
- The sample size is large enough or the variable varies normally in the population

4.1.3.3 Step 3 - Find the p-value

The p-value is the probability of observing a difference in sample means as extreme as the one observed, assuming H_0 is true.

4.1.3.4 Step 4 - Drawing Conclusions

In drawing conclusions about the null hypothesis, we compare the p-value to the level of significance α, and state our conclusion as follows:

- p-value $> \alpha$ - We do not have enough evidence against H_0 and cannot reject it.
- p-value $< \alpha$ - We have enough evidence against H_0 and can reject it in favour of H_a.

4.2 Case $C \rightarrow C$

4.2.1 Step 1 - Stating the Hypothesis

In stating the null and alternative hypothesis for a relationship between two categorical variables, the null hypothesis and alternative hypothesis take the following forms:

- H_0 : There is no relationship between the two categorical variables / They are independent / $p_1 = p_2$
- H_a : There is a relationship between the two categorical variables / They are dependent / $p_1 \neq p_2$

4.2.2 Step 2 - Checking Conditions and Finding the Test Statistic

The test statistic of the chi-square test for independence is the χ^2-score, which is is a standardized measure of the difference between the observed and expected frequencies, i.e.:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed Count} - \text{Expected Count})^2}{\text{Expected Count}}$$

The conditions for the chi-square test for independence are:

- Random sample
- The expected count in each cell is at least 5

4.2.3 Step 3 - Finding the p-value

The p-value is the probability of observing a difference in sample means as extreme as the one observed, assuming H_0 were true.

4.2.4 Step 4 - Drawing Conclusions

In drawing conclusions about the null hypothesis, we compare the p-value to the level of significance α, and state our conclusion as follows:

- p-value $> \alpha$ - We do not have enough evidence against H_0 and cannot reject it.
- p-value $< \alpha$ - We have enough evidence against H_0 and can reject it in favour of H_a.

4.3 Case $Q \rightarrow Q$

4.3.1 Step 1 - Stating the Hypothesis

For a relationship between two quantitative variables, the null and alternative hypothesis take the form of, there is no linear relationship between the two variables, and there is a linear relationship between the two variables, respectively i.e.:

- Null Hypothesis : No linear relationship exists between X and Y
- Alternative Hypothesis: A linear relationship exists between X and Y

Where ρ is the population correlation coefficient

4.3.2 Step 2 - Collecting and Summarizing Data

In this step we calculate the sample correlation coefficient r, and relevant sample statistics and summarize the data with a test statistic. This test statistic is the t-score of the sample correlation coefficient, assuming H_0 is true.

4.3.3 Step 3 - Finding the p-value

In finding the p-value for the population correlation coefficient, we use the t-score of the sample correlation coefficient.

4.3.3.1 Conditions for the t-test for the Population Correlation Coefficient

The conditions needed to perform the t-test for the population correlation coefficient are:

- The observed data seems to have a linear relationship
- The observations are independent
- There are no extreme outliers in the data
- The sample size is fairly large

4.3.4 Step 4 - Drawing Conclusions

In drawing conclusions about the null hypothesis, we compare the p-value to the level of significance α, and state our conclusion as follows:

- p-value $> \alpha$ - We do not have enough evidence against H_0 and cannot reject it.
- p-value $< \alpha$ - We have enough evidence against H_0 and can reject it in favour of H_a.