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 0.5), we can reject H_0 and accept H_a, as the evidence against H_0 is strong.
- If $p-\text{value} > \alpha$ (usually 0.5), we do not have enough evidence against H_0 and cannot reject it.