Basic concepts and statistical packages

stats544-1-slr Applied simple linear regression

Gunnar Stefansson

July 18, 2019

Introduction

A review of concepts: tralla

- Confidence intervals
- Hypotheses test
- p-values

Also the R statistical package

Content: Simple linear regression (SLR) and matrix representation of SLR. This is a preparation for multiple linear regression.

Population and sample

Inferential statistics: To generalize from a sample to a larger group of people or items

We take a **sample** from the population and use that to generalize about an underlying **population**.

If the sampling scheme is appropriate then the generalization works.

R

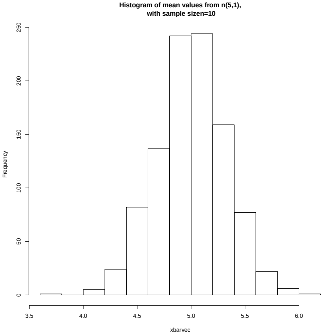
- Students are expected to obtain and install R.
- R is freely available and can be downloaded from the Internet (http://www.r-project.org).
- Although R is free, it is very extensive.
- It is designed for easy extensibility and the emphasis is on easy graphical display and model searches.
- The built-in help system is well-designed and is recommended for all users.
- R-studio can be used to edit R-scripts. It is freely available and can be downloaded from the Internet (http://rstudio.org)
- A short course in R can be found on the tutor-web: STATS240.

Estimators and estimates

Estimator

An **estimator** is a statistic which estimates parameters of a population/probability distributions.

- Estimators for parameters of normal distribution, Poisson distribution and binomial distribution.
- μ, σ, λ and π.
- Estimators are denoted e.g. by $\hat{\mu}$, $\hat{\sigma}$, $\hat{\lambda}$ and $\hat{\pi}$.
- The outcome of an estimator is an estimate, typically $\bar{x}$, $s, \ldots$.



Confidence intervals and confidence level

Usually the probability is zero of the estimate becoming exactly the true value of the parameter.

Confidence intervals

An interval which contains the true value with a confidence level 1 - α.

Confidence level

The proportion of cases where the confidence interval contains the true parameter, in repeated experiments.

Confidence limits

are the endpoints of the confidence interval, called the **lower and upper confidence limit** (or bounds)

The ideology behind hypothesis tests

The ideology behind hypothesis tests

A hypothesis is found which describes what we want to demonstrate and another that describes a neutral (null) case.
A statistic is found which has a known probability distribution in the neutral case. This statistic is our test statistic.
It is defined what values of the test statistic are "improbable" according to the probability distribution in the neutral case.
If the retrieved estimate classifies as "improbable" the hypothesis for the neutral stage is rejected and the hypothesis we want to demonstrate is claimed.
If the estimate is not "improbable" no claims are made.

Hypotheses

Null hypothesis A **null hypothesis** is a hypothesis that can be rejected with observed data. It can never we be claimed. It is usually denoted with H_0.

Alternative hypothesis

An **alternative hypothesis** is the hypothesis we wish confirm with the experiment. It can only be claimed but not rejected. It is either denoted with H_1 or H_a.

Test statistics

Test statistic

A **test statistic** is a statistic that can be used to reject a null hypothesis if the measurements allow.

Null hypothesis rejected

A null hypothesis is **rejected** if the test statistic receives a improbable value compared to the probability distribution it should have if the null hypothesis would be true.

Rejection areas and α-levels

α-level

The α **level** of a hypothesis test is the highest acceptable probability that we receive an improbable value when the null hypothesis is true.

Rejection areas of hypothesis tests

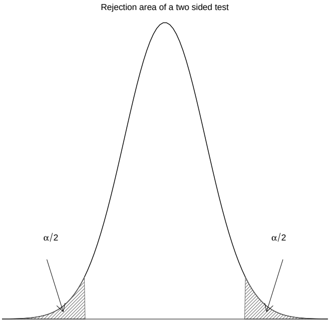
Rejection areas of hypothesis tests are the intervals that contain **all** of the improbable values and **only** those values.
If the test statistics falls within the rejection interval of the hypothesis test, we reject the null hypothesis.
If it does not fall within the rejection interval of the hypothesis test, we make no claims

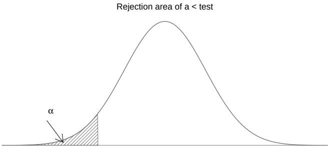
Directions of tests

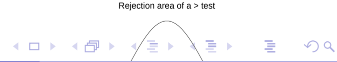
Most tests are "two-sided"

Two-sided tests

In **two-sided tests**, the alternative hypothesis states that the population parameter of interest is **not equal** to a given value (or two parameters are different).


Rejection area of a two sided test


Rejection area of a < test


Rejection area of a > test

One-sided tests

There are two types of **one-sided tests**:
Those who claim that one parameter of the probability distribution is **larger** then another parameter or a certain value, if the measurements allow.
Those who claim that one parameter of the probability

p-values

p-values

A **p-value** is the probability of receiving as improbable value or an value even more improbable as the one received with the measurements if the null hypothesis is true. The H_0 shall be rejected if the p-value is less then α. If the p-value is greater then α the null hypothesis cannot be rejected.

Conducting hypothesis tests

Conducting hypothesis tests

1 Decide which hypothesis test is appropriate for our measurements.

2 Decide the α-level.

3 Propose a null hypothesis and decide the direction of the test (one- or two-sided).

4 Calculate the test statistic for the hypothesis test.

5a See whether the test statistic falls within the rejection interval.

5b Look at the p-value of the test statistic.

6 Draw conclusions.