Probability and Statistics II

Sarah Bussmann

September 4, 2019

- Introduction
- PDFs
- Relationships between variables
- Estimation
- Hypothesis Testing
- Bootstrapping

Introduction

The real power of statistics comes from applying the concepts of probability to situations where you have data but not necessarily the whole population. The results, called **statistical inference**, give you probability statements about the population of interest based on that set of data.

There are two types of statistical inferences: Estimation and Statistical Tests

Estimation

Use information from the sample to estimate (or predict) the parameter of interest.

For instance, using the result of a poll about the president's current approval rating to estimate (or predict) his or her true current approval rating nationwide.

Statistical Tests

Use information from the sample to determine whether a certain statement about the parameter of interest is true. Statistical tests are also referred to as hypothesis tests.

For instance, suppose a news station claims that the Presidents current approval rating is more than 75%. We want to determine whether that statement is supported by the poll data.

Probability Distribution Functions

Recall that a random variable is a variable whose value is the outcome of a random event. For example, a random variable could be the outcome of the roll of a die or the flip of a coin.

A probability distribution is a list of all of the possible outcomes of a random variable along with their corresponding probability values

Probability Distribution Functions

Easy to describe as functions

The output of the function tells us the likelihood of certain outcomes in a random process

https://en.wikipedia.org/wiki/Probability$_{d}istribution$

Probability Distribution Functions

https://newonlinecourses.science.psu.edu/stat414/node/97/

parameter values - μ, λ, σ

Skewness is a property that describes the shape of a distribution. If the distribution is symmetric around its central tendency, it is unskewed. If the values extend farther to the right, it is right skewed and if the values extend left, it is left skewed.

Kernel density estimation(KDE) is an algorithm that takes a sample and finds an appropriately smooth PDF that fits the data

Relationships between variables - Scatterplot

The simplest way to check for a relationship between two variables is a scatter plot

Relationships between variables - Correlation

Correlation – statistic that quantifies the strength of the relationship between two variables

Challenge – variables we want to compare are not given in the same units, if they are in the same units they are from different distributions. To solve this problem we can either use Pearson correlation coef. or Spearman rank correlation coef.

Relationships between variables - Pearson correlation coefficient

Pearsons correlation works well if the relationship between variables is linear and if the variables are roughly normal. Not robust in the presence of outliers.

Relationships between variables - Spearman rank correlation coefficient

- Mitigates the effect of outliers and skewed distributions
- To compute Spearmans correlation, first compute the rank of each value, which is its index in the sorted sample.
- Then compute Pearsons correlation for the ranks
- Example: In sample [1, 2, 5, 7] the rank of the value 5 is 3, because it appears third in the sorted list.

Estimation - Sampling distributions

In theory, the idea of a sufficient statistic provides the basis of choosing a statistic (as a function of the sample data points) in such a way that no information is lost by replacing the full probabilistic description of the sample with the sampling distribution of the selected statistic

use the sample mean, x,to estimate the unknown population mean, μ
The mean of the sampling distribution is pretty close to the hypothetical value of μ, which means that the experiment yields the right answer, on average

Estimation - Sampling distributions

Two common ways to summarize the sampling distribution:

Standard error (SE) is a measure of how far we expect the estimate to be off, on average
A confidence interval (CI) is a range that includes a given fraction of the sampling distribution.

When you report an estimated quantity, it is useful to report standard error, or a confidence interval, or both, in order to quantify sampling error.

Estimation - Sampling bias

Telephone sampling - call and ask women their weight

Problems:

- the sample is limited to people whose telephone numbers are listed, so it eliminates people without phones (who might be poorer than average) and people with unlisted numbers (who might be richer)
- if you call home telephones during the day, you are less likely to sample people with jobs
- if you only sample the person who answers the phone, you are less likely to sample people who share a phone line

Estimation - Sampling bias ctd

If factors like income, employment, and household size are related to weightand it is plausible that the results of your survey would be affected one way or another.

This problem is called **sampling bias** because it is a property of the sampling process. This sampling process is also vulnerable to **self-selection**, which is a kind of sampling bias. Some people will refuse to answer the question, and if the tendency to refuse is related to weight, that would affect the results

Hypothesis Testing

Bootstrapping

- Bootstrapping resamples the original dataset with replacement many times to create simulated datasets. This process involves drawing random samples from the original dataset.
- The bootstrap method has an equal probability of randomly drawing each original data point for inclusion in the resampled datasets
- Can select a data point more than once i.e. sampling with replacement
- Creates resampled datasets that are the same size as the original dataset
- Bootstrapping procedures use the distribution of the sample statistics across the simulated samples as the sampling distribution

Bootstrapping

- Bootstrapping can be used when you believe you don't have the whole population
- Used to see how good your estimate is. Sometimes measuring variability is difficult and there is not an automatic output by statistical software.
- Example: Say we are estimating the variance of the original data set. Not sure how good our estimate is. In order to see if the estimate on the original data set is reasonable, we can bootstrap the original data set many times. With each new data set, we compute a variance statistic. Then after obtaining many variance statistics, we compute the overall standard error.