Lab 9

Statistics 10 S2022

Statistical Inference
Hypothesis Test of Population Mean

Dave Zes
April 15, 2022

1 Introduction

1.1 Hypothesis Testing

Formal statistical hypothesis testing facilitates the following scientific scenario.

We need to make a binary decision concerning the value, or range of values, of a parameter within a population that can be only partially observed.

The binary decision is guided by two hypotheses concerning the population parameter, the null hypothesis, which in statistics circles should always be a strict equality, and an alternative hypothesis, which will be a space of values that are in opposition to the equality expressed in the null.

The null and alternative hypotheses are stated prior to analysis and preferably prior to collecting data. This step also includes stating a significance level, α. The significance level can be thought to be related to our perceived cost of making the wrong decision when we're done. If the cost of wrongly concluding the alternative is true (called a "Type I Error") is high, we may want to set our significance level very small, say, $\alpha = 0.001$.

We call the "partially observed" portion of our population, of course, our "sample".

By the way, a "parameter" can be a vector — that is, we can test many scalar parameter values all at the same time.

In an instructional setting, we may sample from a known population just to illustrate technique and results. However, in practice — in real application — we use statistical hypothesis tests when the entirety of the population is not practically accessible. After all, if we have access to the full population, there would be no need to undertake a formal hypothesis test since we can directly calculate the parameter values of interest.

1.2 Hypothesis Testing of Population Mean

As we've seen, a proportion is an average. If we have a 2-level categorical variable, say, Win or Loss, we can recode this into a numerical variable with 1 for Win and 0 for Loss. In math circles, this is called an "indicator" variable — 1 indicating Win. In stats, we can call this a "dummy" variable. In machine learning, this may be referred to as a "one-hot encoded" variable. The proportion of heads in our original variable equals the average of our indicator variable.

There are two nice consequences of a proportion.

1. The variance of this indicator is a direct function of its mean — the proportion: $\mathsf{Var}[p_W] = p_W\,(1 - p_W)$
2. Given certain conditions, we know the sampling distribution of a proportion will be closely approximated by a normal distribution

Alas, with numerical attributes more broadly, the variance will not be a function of the mean — for the sake of inference, the sample variance will need to be estimated separately. Also, the sampling distribution of the sample mean may be difficult to characterize. This difficulty can be pronounced when the parent distribution (the distribution of our numerical attribute of interest within the larger population) is heavily skewed.

1.2.1 Hypothesis Testing

Same basic deal as with proportions. Our null and alternative hypotheses are statements about the unobserved population mean. Our null hypothesis should be a strict equality, and our alternative will be in opposition to the null.

For example,

$$H_0 : \mu_x = 10$$
$$H_a : \mu_x > 10$$

or,

$$H_0 : \mu_x = 0$$
$$H_a : \mu_x \neq 0$$

or,

$$H_0 : \mu_x = -2.124$$
$$H_a : \mu_x < -2.124$$

1.2.2 Sampling Distribution

A sampling distribution is something with which we're already familiar. It is the distribution of an estimator or a test statistic estimator. A sampling distribution may be either a PDF (probability density function) or a PMF (probability mass function).

A major part of statistical inference is characterizing the sampling distribution of some estimator of interest, i.e., defining its shape. Why? Because we need to be able to determine a critical region, or calculate a p-value, or construct a confidence interval. If we can't characterize our sampling distribution, we can do none of these — at least not with accuracy.

All this takes on special meaning in the case of characterizing the sampling distribution of a sample mean. This is so because the shape of our sampling distribution will be dependent upon the shape of the distribution of the attribute of interest within our population. Said simply, if our parent distribution is skewed, especially if heavy skewed, the normal model, or the t-distribution model, may do a poor job of approximating the true sampling distribution.

1.3 Etc.

Note that in RStudio we can directly execute code from the editor by selecting the code and, on Mac, pressing Command+Return; or, on Windows, Shift+Alt+T.

Please also consult our lab textbook by Peter Dalgaard, [1].

Avoid plagiarism! It is very important that if you borrow code from examples from other sources, like the internet (this is very common, even for skilled programmers) or from anyone else, you should give attribution.

2 For Your Lab 9 Submission

In the folder "yourLab", examine the .Rmd (Rmarkdown) file. When this file is knit to PDF, the corresponding .pdf file is created.

Follow instructions provided in Section "Your Work".

References

[1] Dalgaard Peter. *Introductory Statistics with R.* Springer, 2008.