CLT-based inference - hypothesis testing

Intro to Data Science

Shawn Santo

Announcements

- Homework 04 out today
- Hard deadline for Lab 06 is today
- Keep thinking about what data you want to work with for your project

Today's agenda

- Finish-up confidence intervals from last time
- CLT-based inference for hypothesis testing

Recall

The Central Limit Theorem

For a population with a well-defined mean μ and standard deviation σ, these three properties hold for the distribution of sample average $\bar{X}$, assuming certain conditions hold:

1. The mean of the sampling distribution is identical to the population mean μ,
2. The standard deviation of the distribution of the sample averages is $\sigma/\sqrt{n}$, or the **standard error** (SE) of the mean, and
3. For n large enough (in the limit, as $n \to \infty$), the shape of the sampling distribution of means is approximately *normal* (Gaussian).

Conditions

What are the conditions we need for the CLT to hold?

- **Independence:** The sampled observations must be independent. This is difficult to check, but the following are useful guidelines:
    - the sample must be random
    - if sampling without replacement, sample size must be less than 10% of the population size
- **Sample size / distribution:**
    - if data are numerical, usually n > 30 is considered a large enough sample, but if the underlying population distribution is extremely skewed, more might be needed
    - if we know for sure that the underlying data are normal, then the distribution of sample averages will also be exactly normal, regardless of the sample size
    - if data are categorical, at least 10 successes and 10 failures.

CLT results: $\bar{X}$, $\hat{p}$

Assuming the conditions for the CLT hold, $\bar{X}$ approximately has the following distribution:

$$\text{Normal}\left(\mu, \sigma/\sqrt{n}\right)$$

Equivalently, we can define the quantity Z, such that $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, where Z has the following distribution:

$$\text{Normal}\left(0, 1\right)$$

Assuming the conditions for the CLT hold, $\hat{p}$ approximately has the following distribution:

$$\text{Normal}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

We can standardize this in a similar way and define a quantity Z that is normally distributed with a mean of 0 and a standard deviation of 1.

The hypothesis testing framework

1. Start with two hypotheses about the population: the null hypothesis and the alternative hypothesis.

2. Choose a (representative) sample, collect data, and analyze the data.

3. Figure out how likely it is to see data like what we observed, **IF** the null hypothesis were in fact true.

4. If our data would have been extremely unlikely if the null claim were true, then we reject it and deem the alternative claim worthy of further study. Otherwise, we cannot reject the null claim.

The "errors"

Suppose we test a certain null hypothesis, which can be either true or false (we never know for sure!). We make one of two decisions given our data: either reject or fail to reject H_0.

We have the following four scenarios:

Decision	H_0 **is true**	H_0 **is false**
Fail to reject H_0	Correct decision	*Type II Error*
Reject H_0	*Type I Error*	Correct decision

It is important to weigh the consequences of making each type of error.

CLT-based testing

Testing comparison

What changes now that we plan to use a CLT-based approach in doing our testing?

We no longer have to simulate the null distribution. The Central Limit Theorem gives us an approximation for the distribution of our point estimate under the null hypothesis.

Rather than work directly with the sampling distribution of the point estimates, we'll use standardized versions that we'll call **test statistics**.

For tests of μ:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where μ_0 is the value of μ under the null hypothesis.

For tests of p:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

where p_0 is the value of p under the null hypothesis.

Test statistic and p-value

Recall step 3 of our testing framework: Figure out how likely it is to see data like what we observed, **IF** the null hypothesis were in fact true.

To do this:

1. Compute the test statistic's value - all information is obtained from the sample data or value of the parameter under the null hypothesis.

2. To quantify how likely it is to see this test statistic value given the null hypothesis is true, compute the probability of obtaining a test statistic as extreme or more extreme than what we observed. This probability is calculated from a known distribution.

Let's use the CLT to conduct hypothesis tests

- Create your personal private repository by clicking https://classroom.github.com/a/WhBQhIWu