STAT2203 STAT7203

Probability models and data analysis

Lecture 8.2

Hypothesis testing

Hypothesis testing

We have seen how to estimate a parameter from sample data either by a single number (a point estimate) or an entire interval of plausible values (a confidence interval).

More often than not, however, the ultimate goal of an investigation is not to estimate a parameter but to decide which of the two opposing claims about the parameter is more plausible.

Methods for accomplishing this are often called **hypothesis testing**, which are examples of **statistical inference**.

Hypothesis testing

A *statistical hypothesis*, or just hypothesis, is a claim or assertion either about the value of a single parameter (population characteristic or characteristic of a probability distribution), about the values of several parameters, or about the form of an entire probability distribution.

Examples of hypothesis include:

- The average inside diameter of a certain type of PVC pipe is $0.8mm$.
- The proportion of defective circuit boards produced by a certain manufacturer is less than 10%.
- The claim that $\mu_1 - \mu_2 > 5$ where μ_1 and μ_2 denote the true average breaking strengths of two different types of cord.
- The vehicle braking distance under particular conditions has a normal distribution.

Hypothesis testing

In any hypothesis-testing problem, there are two contradictory hypotheses under consideration.

One of the hypothesis contains the “status quo” statement. It is the default assumption in the absence of any evidence to the contrary. This is often referred to as the *null hypothesis*, and is denoted by H_0.

The *alternative hypothesis*, denoted H_1, contains the statement that is contradictory to H_0, and we wish to find evidence/support for.

Innocent until proven guilty: The two hypotheses do not play equivalent roles: the null hypothesis is initially favored/deemed plausible, the burden of proof is placed on the investigators to demonstrate support for the alternative claim.

Hypothesis testing

The aim of hypothesis testing is to decide, on the basis of the observed data, which of two competing hypotheses is plausible.

We reject H_0 if the observed data is **inconsistent** with it.

Note: The *only* two possible conclusions from a hypothesis-testing analysis are *reject* H_0 (thus accepting H_1) or *fail to reject* H_0. In other words, we do not "prove" or "disprove" any of the claims.

The court of law analogy to this is that when an accused is acquitted, the innocence has not been established. In this case, the evidence was not strong enough to get conviction.

Hypothesis testing

A hypothesis testing procedure is specified in terms of a **test statistic** and a **rejection rule**:

- **Test statistic** is a function of the data only, so it is a random variable. The test statistic for testing H_0 about a parameter can be based on a point estimator of that parameter.
- The **rejection rule** prescribes when H_0 is to be rejected. It is based on the **p-value**, which is the probability that, under H_0, the test statistic takes a value as extreme as or more extreme than the one observed. H_0 is rejected when the observed data is deemed inconsistent with it, evidenced by a small p-value.

test statistic

For most problems we consider here, when testing H_0 about parameter(s), the test statistic takes the form

$$\text{Test Statistic} = \frac{\text{Estimator} - \text{Hypothesized Value}}{\text{SE of Estimator}},$$

which gives a measure of how far the estimator can be from our hypothesized value, relative to the variability of our estimator. We then use the (approximate) distribution of the statistic to decide how unlikely it is too see the observed value of the statistic.

This form of test statistic is almost identical to what we used for constructing confidence intervals. So, it should not be a surprise that the test statistic distributions, as well as the rules of thumb for approximating them, will be the same as those used for constructing CIs.

test statistic

Suppose we are hypothesis testing about a parameter θ of a population. Suppose $H_0 : \theta = \theta_0$. Our test statistic will take the form

$$T = \frac{\Theta - \theta_0}{\mathrm{SE}(\Theta)}.$$

The random variable Θ is an estimator of the unknown parameter θ and is constructed using the samples from the underlying population.

If $\theta = \theta_0$, i.e., H_0 is true, Θ is likely to assume values that are "similar" to θ_0, relative to the variability $\mathrm{SE}(\Theta)$. In other words, we are likely to observe values of T that are "tiny".

Suppose $H_1 : \theta > \theta_0$. Under H_1, Θ is likely to assume values that are "larger" than θ_0, relative to the variability $\mathrm{SE}(\Theta)$. In other words, we are likely to observe values of T that are "large".

test statistic

So, in contrast to H_1, large values of T are unlikely under H_0. In other words, under H_0, large values are considered "extreme".

After collecting data, we observe one realization of the test statistics, say t. Now, the hypothesis testing amounts to answering the following question:

Is t considered "extreme" under H_0?"

p-values

If so, then we have actually observed something that is unlikely to be observed under H_0! So, the observation is inconsistent with H_0, and we can reject H_0. Otherwise, if under H_0 it is quite likely to observe values as large as t, then we have no evidence to reject H_0.

p-value: The likelihood of t being "extreme" under H_0 is the p-value and when $H_1 : \theta > \theta_0$, it is given by

$$\text{p-value} = \mathbb{P}_{H_0}(T \geq t).$$

p-values

More generally, for hypothesis testing about a parameter θ of the population, we judge 'extreme' with respect to the alternative hypothesis. Suppose under H_0, we are assuming $\theta = \theta_0$. Then,

- $H_1 : \theta > \theta_0$

$$\text{p-value} = \mathbb{P}_{H_0}(T \geq t)$$

- $H_1 : \theta < \theta_0$

$$\text{p-value} = \mathbb{P}_{H_0}(T \leq t)$$

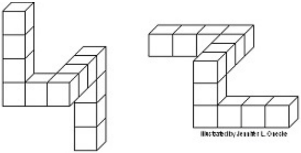
- $H_1 : \theta \neq \theta_0$

$$\text{p-value} = 2\min\{\mathbb{P}_{H_0}(T \leq t), \mathbb{P}_{H_0}(T \geq t)\}$$

Example

A recent study investigated whether games like Minecraft can be used to improve a persons 3D spatial reasoning skills. [1]

Spatial reasoning skill was assessed with the Mental Rotation Test (MRT) giving a score out of 40.



Illustrated by Jennifer L. Ouelle

Figure 1: Based on Shepard & Metzlar's 'Mental Rotation Task'

[1] Carbonell-Carrera et al. (2021) *Minecraft as a block building approach for developing spatial skills*, Entertainment Computing 38: 100427

Example (continued...)

34 participants were recruited to the study and randomly allocated to either Group A or Group B. (17 in each group)

All participants completed the MRT at the beginning of the study.

Two weeks later participants in Group A completed a number of tasks in Minecraft. The session lasted for approximately 4 hours. Participants in Group B did not.

All participants complete the MRT again at the end of the study and the change in their score recorded.

Example (continued...)

Question: Do MRT scores improve without the Minecraft training?

We first want to formulate a model for our data and then specify the hypothesis to be tested in terms of the model parameters.

We need to consider the change in scores for Group B. We represent the changes in the MRT score of Group B participants by $X_1, \ldots, X_{17}$ and assume they form an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

We would like to test:

$$H_0 : \mu = 0, \quad \text{against} \quad H_1 : \mu > 0.$$

Example (continued...)

Since we are testing about the population mean, μ, our estimator is naturally $\bar{X}$.

Since $\bar{X} \sim \mathcal{N}(0, \sigma^2/17)$, then

$$\frac{\bar{X} - 0}{\sigma^2/\sqrt{17}} \sim \mathcal{N}(0, 1).$$

But this test statistic depends on the unknown parameter σ^2, and we are unable to work with it, e.g., given a realization $\bar{x}$, we cannot calculate the corresponding values of this test statistic.

Recall: Student's t Distribution

Recall: Suppose $X_1, \ldots, X_n$ are independent random variables, each having a $\mathcal{N}(\mu, \sigma^2)$ distribution. If $\overline{X}$ and S are the sample mean and sample standard deviation from a random sample then the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}},$$

has a t_{n-1}-distribution.

So in our example, we can consider the following test statistic

$$T = \frac{\overline{X} - 0}{S/\sqrt{17}},$$

which under H_0, has a t_{16}-distribution.

Example (continued...)

The 17 participants in Group B had an average change in MRT score of 3.94 points with a sample standard deviation of 4.72. The realization of our test statistic is then

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{3.94 - 0}{4.72/\sqrt{17}} = 3.44$$

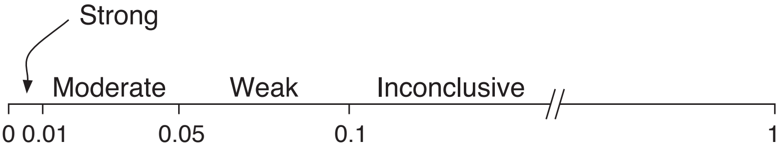
The p-value is

$$p\text{-value} = \mathbb{P}(T_{16} \geq 3.44) = 0.0017$$

What do we conclude?

Strength of Evidence

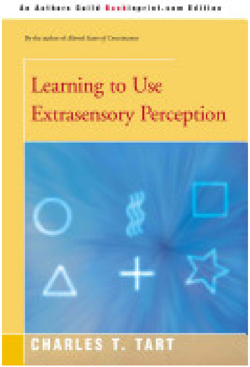
If we need to make a decision between the two hypotheses, then we will need to specify a *significance level*, denoted by α. If the p-value is smaller than α, we will reject H_0 in favour of H_1.



So, we have strong evidence ("beyond reasonable doubt" in legalistic terms) against H_0, suggesting the expected change in MRT score without Minecraft training is greater than zero.

Example

In the 70's Charles Tart at UC, Davis ran an experiment with the aim of demonstrating that specially selected subjects were clairvoyant. The experiment involved a machine which chooses one of four symbols with equal probability using a random number generator. The subject is not shown which. The subject then guesses which symbol was selected by pushing a button. The machine then shows which symbol was selected and records if the guess was correct.


An Authors Guild Backinprint.com Edition
By the author of Altered States of Consciousness
Learning to Use Extrasensory Perception
CHARLES T. TART

Example (continued...)

Tart selected 15 subjects and each subject made 500 guesses.

Out of the 7500 guesses made, 2006 correct guesses were recorded ($2006/7500 \approx 27\%$).

Is this evidence that the subjects are clairvoyant?

Example (continued...)

In the experiment, each guess can be considered a Bernoulli trial. There were $15 \times 500 = 7500$ trials and all trials are assumed to be independent with common probability of a correct guess p. Our model for the number of correct guesses X is therefore

$$X \sim \text{Binomial}(7500, p).$$

Null hypothesis: The null hypothesis is that the subjects are not clairvoyant. Since the four symbols are selected with equal probability by the machine, $H_0 : p = 1/4$.

Alternative hypothesis: If the subjects were clairvoyant, then they would be more likely to correctly guess the symbols. So the alternative hypothesis is $H_1 : p > 1/4$.

Example (continued...)

Under H_0, the probability of making at least 2006 correct guesses out of 7500 is

$$\mathbb{P}(X \geq 2006) = 0.0002739141$$

Since $\mathbb{P}(X \geq 2006) = 1 - \mathbb{P}(X < 2006) = 1 - \mathbb{P}(X \leq 2005)$, we can compute this in `R` as

```
> 1 - pbinom(2005,size=7500, prob=0.25)
[1] 0.0002739141
```

What do you conclude?

Example (continued...)

We used the exact formulas for binomial distribution to get the p-value.

Alternatively, under $H_0 : p = 1/4$, since $n \cdot \min\{p, 1-p\} \geq 5$, we can use CLT approximation

$$T = \frac{\hat{P} - p}{\sqrt{p\left(1-p\right)/n}} = \frac{\hat{P} - 1/4}{\sqrt{\frac{1}{4}\left(1-\frac{1}{4}\right)/7500}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1).$$

The observed values of test statistic in our experiment is

$$t = \frac{0.27 - 0.25}{\sqrt{0.25\left(1-0.25\right)/7500}} = 4.$$

Under H_0, the probability of observing a value of at least 4 for the test statistic is $\mathbb{P}(T \geq 4) \approx 3.17 \times 10^{-5}$.

Does our conclusion change?

Large sample hypothesis tests

Hypothesis test for population mean: For testing $H_0 : \mu = \mu_0$, if we drop the normality assumption of the underlying population, we can still approximate the test statistics

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$

with normal distribution (or t_{n-1}-distribution) according to the previous rules of thumb that we used for constructing confidence intervals; see page 17 of Lecture 7.3.

Hypothesis test for population proportion: For testing $H_0 : p = p_0$, as in the previous example, we can either use exact binomial formulas or alternatively use the normal approximation as long as $n \cdot \min\{p_0, 1 - p_0\} \geq 5$; see page 14 of Lecture 7.1.