

SDS 321: Introduction to Probability and Statistics

Lecture 26: Frequentist statistics: Significance testing

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin
www.cs.cmu.edu/~psarkar/teaching

Roadmap

- ▶ Hypothesis testing setup
- ▶ Type 1 and type 2 error
- ▶ One vs two sided tests
- ▶ Testing if two means are equal.

Binary hypothesis testing

- ▶ You are given an observation vector $X = (X_1, \dots, X_n)$ of random variables
- ▶ We want to choose between two hypothesis H_0 and H_1 about the parameters of the underlying distribution
- ▶ We will call H_0 the **null** hypothesis and H_1 the **alternative** hypothesis.
- ▶ $P(X \in A; H_0)$ will denote the probability of $X \in A$ **under the null hypothesis**, i.e. when the null hypothesis is true.
- ▶ Now, based on the data, we will decide whether H_0 is true or false.

A decision rule

		reality	
		$H_0 = \text{true}$	$H_0 = \text{false}$
conclusion	H_0 is not rejected	OK	type II error
	H_0 is rejected	type I error	OK

- ▶ We first calculate an appropriate summary (or test statistic) of the data, e.g. the mean/variance/median etc.
- ▶ Based on the data distribution, we come up with a rejection region R for this statistic.
- ▶ If the statistic falls in the rejection region R , we reject H_0 . Otherwise we accept it.

Type I and type II errors

- ▶ Reject H_0 even though H_0 is true. We will denote this by **Type I** error or a **false rejection**. The probability of this for a given rejection region is denoted by $\alpha(R)$

$$\alpha(R) = P(X \in R; H_0)$$

- ▶ Accept H_0 even though H_0 is false. This is denoted by a **Type II** error or a **false acceptance**. The probability of this for a given rejection region is denoted by $\beta(R)$.

$$\beta(R) = P(X \notin R; H_1)$$

- ▶ Ideally we want a test such that both these errors are small.

Type I and type II errors

You have X_1, \dots, X_n i.i.d $\sim N(\mu, 1)$. $H_0 : \mu = 0$ and $H_1 : \mu = 1$. Let $n = 16$. Set up the rejection region $R = \{\bar{X} > \xi\}$ at 10% significance level. Calculate the type 2 error.

▶ We want $P(\bar{X} \geq \xi; H_0) = 0.1$.

▶ Remember: $\bar{X} \sim \begin{cases} N(0, 1/n) & \text{Under } H_0 \\ N(1, 1/n) & \text{Under } H_1 \end{cases}$.

▶ $P(\bar{X} > \xi; H_0) = P\left(\frac{\bar{X}}{1/\sqrt{n}} > \frac{\xi}{1/\sqrt{n}}\right) = P(Z > \frac{\xi}{1/\sqrt{n}}) = 0.1$

$$\rightarrow \sqrt{n}\xi = 1.28 \rightarrow \xi = 1.28/\sqrt{n} = 1.28/4 = .32$$

▶ Question to ask: Why is the ξ going down with larger n ?

Significance testing: The general approach

Lets do an example first. We toss a coin independently a thousand times. Each toss is a Bernoulli(θ) random variable.

$$H_0 : \theta = 1/2 \quad H_1 : \theta \neq 1/2$$

- ▶ However, we can note that if one looks at the average \bar{X} , then under the null, it will be very close to $1/2$.
- ▶ Why?

Significance testing: The general approach

Lets do an example first. We toss a coin independently a thousand times. Each toss is a Bernoulli(θ) random variable.

$$H_0 : \theta = 1/2 \quad H_1 : \theta \neq 1/2$$

- ▶ However, we can note that if one looks at the average \bar{X} , then under the null, it will be very close to $1/2$.
- ▶ Why? Weak law of large numbers!

Significance testing: The general approach

Lets do an example first. We toss a coin independently a thousand times. Each toss is a Bernoulli(θ) random variable.

$$H_0 : \theta = 1/2 \quad H_1 : \theta \neq 1/2$$

- ▶ However, we can note that if one looks at the average \bar{X} , then under the null, it will be very close to $1/2$.
- ▶ Why? Weak law of large numbers!
- ▶ If $|\bar{X} - 1/2| > \xi$, then it makes sense that the null is not true. In this case, $|\bar{X} - 1/2|$ is the sample statistic you have chosen.
- ▶ So the rejection rule is Reject H_0 if $|\bar{X} - 1/2| > \xi$

Significance testing: The general approach

Lets do an example first. We toss a coin independently a thousand times. Each toss is a Bernoulli(θ) random variable.

$$H_0 : \theta = 1/2 \quad H_1 : \theta \neq 1/2$$

- ▶ However, we can note that if one looks at the average \bar{X} , then under the null, it will be very close to $1/2$.
- ▶ Why? Weak law of large numbers!
- ▶ If $|\bar{X} - 1/2| > \xi$, then it makes sense that the null is not true. In this case, $|\bar{X} - 1/2|$ is the sample statistic you have chosen.
- ▶ So the rejection rule is Reject H_0 if $|\bar{X} - 1/2| > \xi$
- ▶ We pick the critical value ξ such that, $P(\text{reject } H_0; H_0) = \alpha$

Significance testing: The general approach

Lets do an example first. We toss a coin independently a thousand times. Each toss is a Bernoulli(θ) random variable.

$$H_0 : \theta = 1/2 \quad H_1 : \theta \neq 1/2$$

- ▶ However, we can note that if one looks at the average \bar{X} , then under the null, it will be very close to $1/2$.
- ▶ Why? Weak law of large numbers!
- ▶ If $|\bar{X} - 1/2| > \xi$, then it makes sense that the null is not true. In this case, $|\bar{X} - 1/2|$ is the sample statistic you have chosen.
- ▶ So the rejection rule is Reject H_0 if $|\bar{X} - 1/2| > \xi$
- ▶ We pick the critical value ξ such that, $P(\text{reject } H_0; H_0) = \alpha$
- ▶ Here α is also called the **level of significance**

Significance testing: The general approach

Lets do an example first. We toss a coin independently a thousand times. Each toss is a Bernoulli(θ) random variable.

$$H_0 : \theta = 1/2 \quad H_1 : \theta \neq 1/2$$

Significance testing: The general approach

Lets do an example first. We toss a coin independently a thousand times. Each toss is a Bernoulli(θ) random variable.

$$H_0 : \theta = 1/2 \quad H_1 : \theta \neq 1/2$$

- ▶ We pick the critical value ξ such that, $P(\text{reject } H_0; H_0) = \alpha$

Significance testing: The general approach

Lets do an example first. We toss a coin independently a thousand times. Each toss is a Bernoulli(θ) random variable.

$$H_0 : \theta = 1/2 \quad H_1 : \theta \neq 1/2$$

- ▶ We pick the critical value ξ such that, $P(\text{reject } H_0; H_0) = \alpha$
- ▶ Here α is also called the **level of significance**

Significance testing: The general approach

Lets do an example first. We toss a coin independently a thousand times. Each toss is a Bernoulli(θ) random variable.

$$H_0 : \theta = 1/2 \quad H_1 : \theta \neq 1/2$$

- ▶ We pick the critical value ξ such that, $P(\text{reject } H_0; H_0) = \alpha$
- ▶ Here α is also called the **level of significance**
- ▶ Say $\alpha = 0.05$. We want $P(|\bar{X} - 1/2| > \xi; H_0) \approx 0.05$. We use the fact that under the null, the distribution of $(\bar{X} - 1/2)/\sqrt{1/4000}$ is approximately standard normal.

Significance testing: The general approach

Lets do an example first. We toss a coin independently a thousand times. Each toss is a Bernoulli(θ) random variable.

$$H_0 : \theta = 1/2 \quad H_1 : \theta \neq 1/2$$

- ▶ We pick the critical value ξ such that, $P(\text{reject } H_0; H_0) = \alpha$
- ▶ Here α is also called the **level of significance**
- ▶ Say $\alpha = 0.05$. We want $P(|\bar{X} - 1/2| > \xi; H_0) \approx 0.05$. We use the fact that under the null, the distribution of $(\bar{X} - 1/2)/\sqrt{1/4000}$ is approximately standard normal.
- ▶ So $P(|\bar{X} - 1/2| > \xi; H_0) \approx P(|Z| > \xi/0.016) = 0.05$. Since $\Phi(1.96) = 0.05$, $\xi = 0.031$.

Significance testing: The general approach

Lets do an example first. We toss a coin independently a thousand times. Each toss is a Bernoulli(θ) random variable.

$$H_0 : \theta = 1/2 \quad H_1 : \theta \neq 1/2$$

- ▶ We pick the critical value ξ such that, $P(\text{reject } H_0; H_0) = \alpha$
- ▶ Here α is also called the **level of significance**
- ▶ Say $\alpha = 0.05$. We want $P(|\bar{X} - 1/2| > \xi; H_0) \approx 0.05$. We use the fact that under the null, the distribution of $(\bar{X} - 1/2)/\sqrt{1/4000}$ is approximately standard normal.
- ▶ So $P(|\bar{X} - 1/2| > \xi; H_0) \approx P(|Z| > \xi/0.016) = 0.05$. Since $\Phi(1.96) = 0.05$, $\xi = 0.031$.
- ▶ You got some data such that $\bar{X} = 0.8$. Do you accept or reject?

Significance testing methodology

You are given $X = (X_1, \dots, X_n)$ and you want to test a given null hypothesis H_0 .

- ▶ Choose a statistic $S = h(X)$, which is a scalar random variable that summarizes the data, e.g. the mean.
- ▶ Determine the shape of the rejection region by specifying the set R_ξ , such that if $h(X) \in R_\xi$ then H_0 is rejected. In the last example this was $\{X : |\bar{X} - 1/2| \geq \xi\}$
- ▶ Specify a significance level α . The false acceptance probability equals (exactly or approximately) α .
 - ▶ A common choice is 0.01.
 - ▶ If you make α smaller, you reject more often and also raise the probability of false rejection.
- ▶ Find ξ such that $P(X \in R_\xi; H_0) \approx \alpha$.

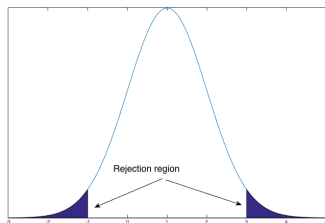
Once you have the data, calculate S , and reject H_0 if $S \in R_\xi$

Things to remember

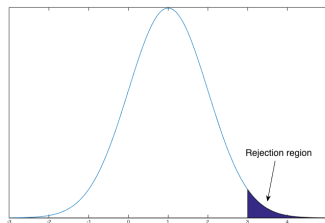
- ▶ Note that we do not say that we accept the alternative hypothesis anymore. Because the alternative is basically the complement of the null. We only say that we reject the null.
- ▶ There is no best statistic.
- ▶ Often the choice is natural, like the mean.
- ▶ Sometimes the statistic is chosen such that one can easily approximate its distribution under the null.

One sided vs two sided tests

- ▶ Sometimes the natural choice of the alternative is $H_1 : \theta \neq \theta_0$, where $H_0 : \theta = \theta_0$.
- ▶ This is called a two sided test, since the decision rule looks like **Reject H_0 if $|S - c| > \xi$** . c was $1/2$ in the last example.
- ▶ Sometimes it may be more suitable to have an alternative of the form $H_1 : \theta > \theta_0$.
- ▶ In this situation the rejection region looks like **Reject H_0 if $S - c > \xi$** .
- ▶ This is called an one sided test. All that changes is the calculation of the critical value.



Two sided test



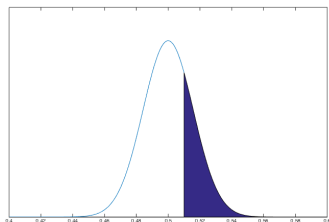
One sided test

Statistical significance

- ▶ In significance testing, really you are asking the question, is it possible that what I think is signal is just a random artifact?
- ▶ If in the last example $\bar{X} = 0.51$, then could it be possible, that even though the null is true, the sample mean is 0.51?
- ▶ You only reject the null, if the difference between \bar{X} you see from data and $1/2$ is **statistically significant**.
- ▶ When we say that H_0 is rejected at 0.01 significance level, what does it mean?
- ▶ Does not mean that probability that H_0 is true is less than α .
- ▶ Means that if you were given many datasets drawn under H_0 , then 1 percent of the time you will reject the null hypothesis (even though it is true).
- ▶ i.e. The observed data is highly unlikely under H_0 , since such data would only arise happen 1% of the time

p-value

- ▶ In the last example, say your data gives $\bar{x} = 0.51$ and for simplicity say $H_1 : \theta > 1/2$. If you wanted to ask how surprising this is, what would you look at?
- ▶ One option: you look at how often data generated from the null gives rise to a statistic value bigger than 0.51? If this is large, then we do not want to reject the null.



- ▶ Often a statistician skips the steps of calculating ξ for a fixed significance level, and instead reports a p-value for the calculated statistic S .

Example: testing for equal means

You are given (X_1, \dots, X_n) iid from $Bernoulli(\theta_X)$ and (Y_1, \dots, Y_m) iid from $Bernoulli(\theta_Y)$. You want to test $H_0 : \theta_X = \theta_Y$ vs $H_1 : \theta_X \neq \theta_Y$. You may be interested in finding whether a flu-shot is equally effective for two different age groups.

- ▶ $\hat{\theta}_X = \sum_{i=1}^n X_i/n$ is approximately normal with mean θ_X and variance $\theta_X(1 - \theta_X)/n$
- ▶ $\hat{\theta}_Y = \sum_{i=1}^m X_i/m$ is approximately normal with mean θ_Y and variance $\theta_Y(1 - \theta_Y)/m$
- ▶ What is a reasonable statistic to measure if the two means are different?
- ▶ $S = (\theta_X - \theta_Y)!$
- ▶ Decision rule: reject H_0 if $|\theta_X - \theta_Y| > \xi$. This is a **two sided test!**

Example: testing for equal means

You are given (X_1, \dots, X_n) iid from $Bernoulli(\theta_X)$ and (Y_1, \dots, Y_m) iid from $Bernoulli(\theta_Y)$. You want to test $H_0 : \theta_X = \theta_Y$ vs $H_1 : \theta_X \neq \theta_Y$. You may be interested in finding whether a flu-shot is equally effective for two different age groups.

- ▶ Decision rule: reject H_0 if $|\hat{\Theta}_X - \hat{\Theta}_Y| > \xi$. This is a **two sided test!**
- ▶ We need the distribution of S .
- ▶ If we knew the common mean $\theta = \theta_X = \theta_Y$ under the null, we would be done.
- ▶ Then, S is approximately normal with mean zero and variance $\text{var}(\hat{\Theta}_X) + \text{var}(\hat{\Theta}_Y) = \theta(1 - \theta)/n + \theta(1 - \theta)/m$
- ▶ But we do not know θ
- ▶ But we can approximate it by $\hat{\Theta} = (\sum_i X_i + \sum_j Y_j)/(m + n)$.

- ▶ Now $\frac{\hat{\Theta}_X - \hat{\Theta}_Y}{\sqrt{\hat{\Theta}(1 - \hat{\Theta})(1/m + 1/n)}}$ is approximately a standard normal and we can calculate ξ to get the rejection region at level α significance level.