Y-DATA
SCHOOL OF DATA SCIENCE

Probability and Statistics for Data Science

Lecture 5 – Hypothesis testing

Today

- Terminology of hypothesis testing
- Type I and type II errors
- Simple hypotheses
- Choice of critical value
- p-value
- Neyman-Pearson lemma (most powerful test)
- Composite hypotheses
- Power function
- Hypothesis testing and confidence intervals

Introduction

- In lecture 4 we talked about point estimation and confidence intervals.
- **Point estimation** – estimation of one or more (but a finite amount) unknown parameters.
- **Confidence intervals** – construction of an interval in which the true parameter lays with high probability.

In both cases, we were interested in the value of the true parameter (exactly or its range).

Today, we will learn how to answer ”yes”/”no” questions about the unknown parameter.

Hypothesis testing – motivation

- In many cases, the research questions yields one of two possible answers (“yes”/”no”).
- We are required to choose the correct answer from the data.
- For example:
  - Is the proportion of newborn girls is 0.5 or not?
  - Is the expected weight of a cheese pack is indeed 200g or less?
  - Is the new drug effective?
  - Is the signal appearing on a radar comes from an aircraft?
- Let’s see how can we formalize the framework to deal with such questions.

Terminology

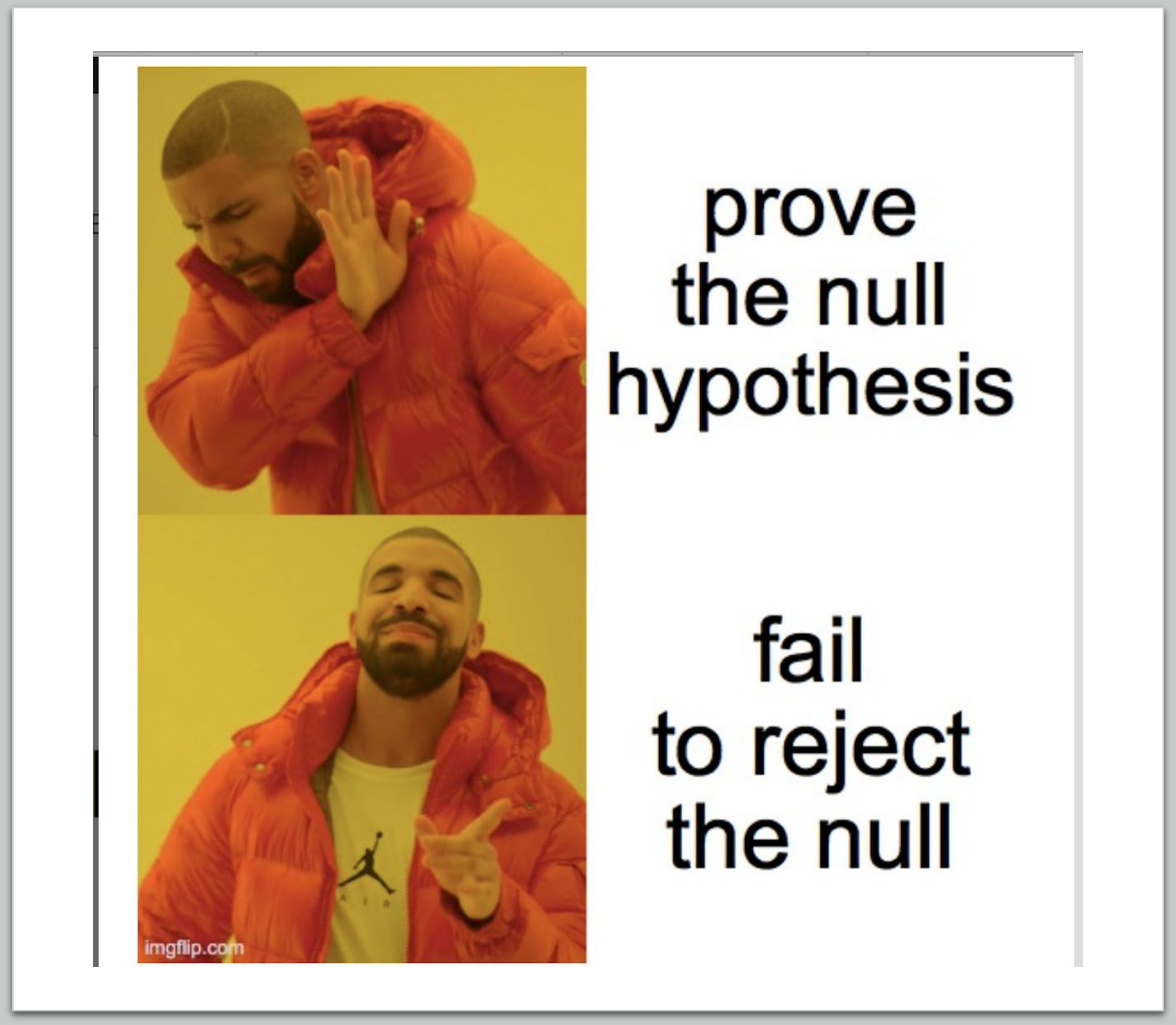
- We have $X_1, \dots, X_n \sim F_\theta$ (i.i.d.) where the parameter $\theta \in \Theta$ is unknown.
- We would like to decide whether θ belongs to a given subset Θ_0.
- The subset Θ_0 is called the *null* space and its complement $\Theta_1 = \Theta \setminus \Theta_0$ is called the *alternative*.
- In statistical terminology, we want to test

$$H_0: \theta \in \Theta_0 \;\; (null\ hypothesis)$$

$$H_1: \theta \in \Theta_1 \; (alternative)$$

- We want to find a *test statistic* $T(X)$ to decide whether we accept or reject the null hypothesis.

Remark: It is customary to consider the null as some usual theory or absence of effect and to test it against the alternative, viewed as new theory.



Example

- We are required to decide whether a coin is real or fake, where the real coin is known to be fair, and the fake is known to have 2/3 probability of "heads". The coin is tossed independently n times and the decision is to be based on the obtained sample $X_1, \ldots, X_n$.
- In our case, the model is $Ber(\theta)$ where $\Theta = \left\{\frac{1}{2}, \frac{2}{3}\right\}$.

$$H_0: \theta = \frac{1}{2} \qquad vs. \qquad H_1: \theta = \frac{2}{3}$$

- If such precise information about the fake coin is unavailable, it is reasonable to consider the problem with $\Theta = (0,1)$ and

$$H_0: \theta = \frac{1}{2} \qquad vs. \qquad H_1: \theta \neq \frac{1}{2}$$

Example (continued)

- Now that we have the model and the hypothesis to test, how can we actually test the hypothesis?
- Informally, we would like that for large n, the empirical mean is close to the true value of the parameter θ.
- It makes sense to reject the null if the empirical mean $\bar{X}_n$ exceeds a certain value c. That is, reject H_0 if $\{\bar{X}_n > c\}$ with $c > \frac{1}{2}$.
- For the scenario where $H_1: \theta \neq \frac{1}{2}$, we can propose the test: reject H_0 if $\{|\bar{X}_n - 0.5| > c\}$ (i.e. if the proportion of heads in the sample is too different from 0.5).

What’s next?

- We have the terminology and we know how to describe a problem in terms of hypothesis testing.

BUT

- We still don’t know how to calculate the value of c (AKA the *critical value*)
- Is our test good? How can we measure that?

Types of errors

- Since the inference on hypotheses is based on random data, there is always a room for a wrong decision.
- Type I error: when we erroneously reject the null hypothesis (AKA false positive, α, significance level)
- Type II error: the null hypothesis is not rejected, when it is actually false (AKA false negative, β)

Types of errors (continued)

- So, the key question in hypothesis testing is a proper choice of a test statistic that would minimize the probability of an erroneous decision.
- Unfortunately, it is generally impossible to minimize the probability of errors of both types. Decreasing one of them comes at the expense of increasing the other.
- Consider the extreme cases:
  - If we decide to always reject the null regardless of the data, there is a zero probability of a type II error, but the probability of type I error is 1.

  out of the possible errors

  - If one always accepts the null, the probability of type I error is zero but the probability of type II error is one.
- A good test will make errors of both types with small probabilities.

Simple Hypotheses

Simple hypotheses

- Let $\mathrm{X} = (X_1, \dots, X_n) \sim F_\theta(x)$ and we wish to test two simple hypotheses:
$$H_0: \theta = \theta_0 \ \ vs.\ H_1: \theta = \theta_1 \text{ (assume w.l.o.g. that } \theta_1 > \theta_0\text{)}$$
- Choose a statistic $T(X)$ and reject the null if $T(X) \geq c$ for some critical value c.
- This induces a partition of the sample space Ω into two disjoint regions:
  - rejection region: $\Omega_1 = \{x: T(x) \geq c\}$
  - acceptance region: $\Omega_0 = \{x: T(x) < c\}$
- In this case,
$$\alpha = P_{\theta_0}(T(X) \geq c) \ \ \beta = P_{\theta_1}(T(X) < c)$$
- The *power* of the test is $\pi = 1 - \beta$.

=Pr(Reject H_0 | H_1 true)

Example

A car company introduces a new car model and advertises that it consumes less fuel than other cars of the same class. In particular, the company claims that the new car on average will run 15 kilometers per liter on the highway, while its existing competitors only 12. “Auto-World Journal” wants to check the credibility of the claim by testing a random sample of 5 new cars produced by the company.

Assume that $X_1, \ldots, X_5 \sim N(\mu, 4)$.

In this case, we need to test

$$H_0: \mu = 12 \;\; vs. \; H_1: \mu = 15$$

Example (continued)

- We have

$$H_0 : \mu = 12 \quad H_1 : \mu = 15$$

- If we decide to reject the null if $\bar{X}_n \geq 14$, then the corresponding error probabilities are

$$\alpha = P_{H_0}(\bar{X} \geq 14) = 1 - \Phi\left(\frac{14 - 12}{\frac{2}{\sqrt{5}}}\right) = 0.012$$

X_i~N(12,4)

prob. for false positive.
i.e. the prob. that the new car consumes fule as old cars but we flasly get that it consumes less

and

$$\beta = P_{H_1}(\bar{X} < 14) = \Phi\left(\frac{14 - 15}{\frac{2}{\sqrt{5}}}\right) = 0.132$$

X_i~N(15,4)

prob. for false negative

Choice of critical value

- Up until now, we assumed that c is some value for us to choose, from which we can compute the desired probabilities of error.
- In practice, we choose c differently.
- To choose a critical value c, we usually fix some α that we are willing to accept and then solve the equation
$$\alpha = P_{\theta_0}(T(X) \geq c)$$
- Note that for each value of α we will get a different critical value c.

Example

- Consider the car example with $X_1, \dots, X_5 \sim N(\mu, 4)$ and
$$H_0\colon \mu = 12 \quad vs.\ H_1\colon \mu = 15$$
- Instead of fixing $c = 14$, we will find the critical value corresponding to $\alpha = 0.05$.
- To this end, we solve the equation
$$\alpha = P_{\mu_0}(\bar{X}_n \geq c) = 1 - \Phi\left(\frac{c - 12}{\frac{2}{\sqrt{5}}}\right)$$

Therefore,

$$c = 12 + \overset{\substack{\boxed{\text{inverse of phi}} \\ \downarrow}}{z_{0.95}} \frac{2}{\sqrt{5}}$$

The p-value

- Let $t_{obs} = T(x)$ be the actual value of the statistic $T(X)$ for a sample x. The p-value is defined as

~Def (for the simple case): $p - value = P_{\theta_0}(T(X) \geq t_{obs})$

- The p-value measures the "unlikeliness" of t_{obs} under the null.
- The smaller the p-value, the more "unlike"/"extreme" t_{obs} is under the null hypothesis and the stronger evidence against it.
- Recall that $P_{\theta_0}(T(X) \geq c) = \alpha$ while $P_{\theta_0}(T(X) \geq t_{obs}) = p - value$. Therefore, $t_{obs} \geq c$ if and only if $p - value \leq \alpha$.
- That is, we can obtain a decision rule in terms of p-value and reject the null if $p - value \leq \alpha$.
- The p-value can be viewed as the minimal significance level for which we reject the null for a given value of the statistic.

Ok… so why p-value?

- If $t_{obs} \geq c$ if and only if $p - value \leq \alpha$, why bother ourselves with computing the p-value?
- The p-value can be viewed as the minimal significance level for which we reject the null for a given value of the statistic.
- A p-value has a clear and absolute probabilistic meaning.
- As opposed to the critical value, the p-value does not change as a function of α.

Example

- Consider the car example with $X_1, \dots, X_5 \sim N(\mu, 4)$ and
$$H_0 : \mu = 12 \quad H_1 : \mu = 15$$
and suppose that the empirical mean is 13.8.
- Calculate the p-value
$$p - value = P_{\mu_0}(\bar{X}_n \geq 13.8) = 1 - \Phi\left(\frac{13.8 - 12}{\frac{2}{\sqrt{5}}}\right) = 0.022$$
- Meaning: We would accept the company’s claim at a significance level of 0.05, but not at a significance level of 0.01 (or any significance level lower than 0.022).

Maximal power test

- How do we find $T(X)$ properly?
- We saw the tradeoff between the types of error, so our strategy will be to find a test statistic that achieves the minimal probability of a type II error (equivalently, maximal power) among all the tests at a given level α.

Neyman-Pearson lemma

- **Neyman-Pearson lemma:** Assume that the observed data $X \sim F_\theta$ and consider the two simple hypotheses:
$$H_0: \theta = \theta_0 \;\; vs. H_1: \theta = \theta_1$$
and define the *likelihood ratio*
$$\lambda(x) = \frac{L(\theta_1; x)}{L(\theta_0; x)}$$
then the likelihood ratio test with rejection region $\{\lambda(x) \geq c\}$ has the maximal power among all tests at significance level not larger than α.
- <u>Remark:</u> the LR $\lambda(x)$ shows how much θ_1 is more likely than θ_0. The larger $\lambda(x)$, the stronger the evidence against the null.

Composite Hypotheses

One-sided hypothesis

- We will consider the model $X_1, \dots, X_n \sim N(\theta, \sigma^2)$ with known σ^2 to explain composite hypotheses testing for simplicity.
- A hypothesis testing with composite alternative has the form
$$H_0: \theta = \theta_0 \; vs. H_1: \theta > \theta_0$$
- Notice that for the simple hypothesis we didn't use the value of θ_1, but just the fact that $\theta_1 > \theta_0$. Therefore, the test statistic and the p-value would be the same. We just need to redefine the power.
- **Def:** The *power function* is given by $\pi(\theta) = P_\theta(reject\ H_0)$
- In a way, $\pi(\theta)$ is the analogue of significance level for $\theta \in \Theta_0$ and of power for $\theta \in \Theta_1$.

Significance and p-value in composite hypotheses

- **Def:** The significance level of a given test with a power function $\pi(\theta)$ is $\alpha = \sup_{\theta\in\Theta_0} \pi(\theta)$.
- **Def:** For the observed value t_{obs} of a given test statistic $T(X)$,
$$p - value = \sup_{\theta\in\Theta_0} P_\theta(T(X) \geq t_{obs})$$
- It turns out that for the case of
$$H_0 : \theta \leq \theta_0 \; vs. H_1 : \theta > \theta_0$$
(or vice versa) this can be treated as the case of simple null. Therefore, it is not complicated to compute the significance and the p-value.

Two-sided hypothesis

- Consider the model $X_1, \dots, X_n \sim N(\theta, \sigma^2)$ with known σ^2.
- A hypothesis testing with composite (two-sided) alternative has the form

$$H_0: \theta = \theta_0 \; vs. H_1: \theta \neq \theta_0$$

- We reject H_0 is $\bar{X}_n$ is too far from θ_0, i.e. $|\bar{X}_n - \theta_0| \geq c$.
- Find the critical value: solve for c,

$$P_{\theta_0}(|\bar{X}_n - \theta_0| \geq c) = \alpha$$

- From calculations we get that we reject H_0 if $|\bar{X}_n - \theta_0| \geq z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

How to choose the test statistic?

- It turns out that a UMP (uniformly most powerful) test for testing composite hypotheses usually does not exist.
- Therefore, a “reasonable” statistic to use is a generalization of the likelihood ratio test.
- Generalized likelihood ratio test (GLRT): We compute the statistic

$$\lambda^*(x) = \frac{\sup_{\theta \in \Theta} L(\theta; x)}{\sup_{\theta \in \Theta_0} L(\theta; x)}$$

All space

And we reject the null if $\lambda^*(x) \geq c$, where c is such that $\sup_{\theta \in \Theta_0} P_\theta(\lambda^*(x) \geq c) = \alpha$.

How to calculate the GLRT?

1. Find the MLE $\hat{\theta}$ for θ to calculate the numerator
2. Find the MLE $\hat{\theta}_0$ for θ under the restrictions $\theta \in \Theta_0$ to calculate the denominator
3. Form the generalized likelihood ratio $\lambda(X)$ and find an equivalent simpler test statistic $T(X)$ (if possible).
4. Find the corresponding critical value for $T(X)$ solving

$$\sup_{\theta \in \Theta_0} P_\theta(T(X) \geq c) = \alpha$$

- There is a theorem (Wilks' theorem) that provides the asymptotic distribution of $2 \log \lambda_n(X)$ under the null.

which is chi^2

Different models

There are a lot more hypotheses to test that have well-known statistical test:

- One and two-sample t-test for normal means
- Chi squared tests for normal variance
- F-test for normal variances
- Pearson’s chi-squared test for goodness of fit
- Etc.

They are all examples of GLRTs!

Hypothesis testing and confidence intervals

- Note that the acceptance region in the two sided test is

$|\bar{X}_n - \theta_0| \geq z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

- Recall that a $1-\alpha$ CI for θ is $\left[\overline{\mathrm{X}}_n \pm \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right]$.
- After some algebra, we can say that we can use the $1-\alpha$ confidence interval for θ (expectation of i.i.d. normal RV's with known variance) to test the two-sided hypotesis testing with significance level α by rejecting the null if $\mu_0 \notin \left[\overline{\mathrm{X}}_n \pm \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right]$.

References

- Abramovich, Felix, and Ya'acov Ritov. “Statistical theory: a concise introduction. CRC Press, 2013.
- Micha Mandel’s lecture notes in course 52221 at the Hebrew University.
- Pavel Chigansky’s lecture notes in course 52325 at the Hebrew University (available on his website).