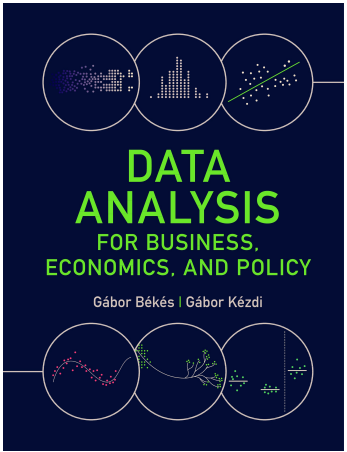
06 Testing hypotheses

Davide Del Prete

Data Analysis 1: Exploration

2024

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021
- gabors-data-analysis.com
  - Download all data and code: gabors-data-analysis.com/data-and-code/
- This slideshow is for Chapter 06

Motivation

- *The internet allowed the emergence of specialized online retailers while larger shops also sell goods on the main street as well. How to measure price differentiation by online and offline prices? To help answer this, we can collect and compare online and offline prices of the same products and test if the averages are the same.*

The logic of hypothesis testing

- A hypothesis is a statement about a general pattern, of which we are not sure if it is true or not.
- Hypothesis testing = analyze our data to make a decision based on the hypothesis.

- Reject the hypothesis if there is enough evidence against it.
- Don't reject it if there isn't enough evidence against it.
- We do not have enough evidence against a hypothesis
  - if the hypothesis is in fact true
  - or it is not true, but our evidence is weak
- Important asymmetry: rejecting a hypothesis is a more conclusive decision than not rejecting it!

The logic of hypothesis testing: the setup

- Define the *the statistic we want to test*. Let us call it s (e.g. mean).
- We are interested in the true value of s noted as s_{true}.
- The value the statistic in our data is its estimated value, denoted by a hat on top $\hat{s}$.

The logic of hypothesis testing: H0 vs HA

- Formally stating the question as two competing hypotheses of which only one can be true: a null hypothesis H_0 and an alternative hypothesis H_A.
- Formulated in terms of the unknown true value of the statistic.
- The null specifies some value or range; the alternative specifies *all other* possible values.
- Together, the null and the alternative cover all the possibilities we are interested in.
- One example:

$$H_0 : s_{true} = 0$$

$$H_A : s_{true} \neq 0$$

The logic of hypothesis testing: two vs. one-sided test

- Two-sided alternative:
  - We test if $H_A : s_{true} \neq 0$ - allows for s_{true} to be either greater than zero or less than zero. Not interested if the difference is positive or negative.
- One-sided alternative
  - interested if a statistic is positive or not.
- Different setup: the hypothesis we are testing is focusing to the alternative set.

One-sided test:

$$H_0 : s_{true} \leq 0$$

$$H_A : s_{true} > 0$$

Comparing online and offline prices: Testing hypotheses

- Question: Do the online and offline prices of the same products differ on average?
- Data includes 10 to 50 products in each retail store included in the survey (the largest retailers in the U.S. that sell their products both online and offline).
- The products were selected by the data collectors in stores, and they were matched to the same products the same stores sold online.

- Let define our statistic as the difference in average prices.

Comparing online and offline prices: Testing hypotheses

- Statistics we are interested: difference in prices
- Each product i has both an online ($p_{i,online}$) and an offline ($p_{i,offline}$) price in the data.
- The difference is:

$$pdiff_i = p_{i,online} - p_{i,offline}$$

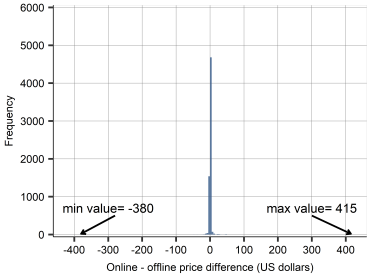
The statistic we are interested in with n observations is:

$$\hat{s} = \overline{pdiff} = \overline{p_{online}} - \overline{p_{offline}} = \frac{1}{n}\sum_{i=1}^{n}(p_{i,online} - p_{i,offline})$$

Comparing online and offline prices: Testing hypotheses

Descriptive statistics of the difference

- The mean difference is -0.05$: online prices are, on average, 5 cents lower in this dataset.
- Spread around this average: Standard deviation is 10$
- Extreme values matter: Range: -380$ — +415$.
- Out of the 6,439 products, 64% have the same online and offline price, for 87%, the difference within ± 1 dollars.



Comparing online and offline prices: Formalizing the question

Do average prices differ in the general pattern represented by the data?

$$H_0 : s_{true} = \bar{p}_{online\ true} - \bar{p}_{offline\ true} = 0$$

$$H_A : s_{true} = \bar{p}_{online\ true} - \bar{p}_{offline\ true} \neq 0$$

The logic of hypothesis testing

- The t-test is the testing procedure based on the *t-statistic*
  - t-statistic comes from a sampling distribution which is distributed as a standardized 'Student's-t' distribution.
- We compare the estimated value of the statistic $\hat{s}$ (our best guess of s) to our null-hypothesis.
- Evidence to reject the null: difference between $\hat{s}$ and our null-hypothesis is large.
- Not reject the null if the estimate is not very far, i.e., when there is not enough evidence against it.

T-test

- The *t-statistic* is a statistic that measures the distance of the estimated value from what the true value would be if H_0 was true.
- Uses sample estimates $\hat{s}$ and the standard error of the estimate, $SE(\hat{s})$. Let

$$H_0 : s_{true} = 0,$$
$$H_A : s_{true} \neq 0$$

- The t-statistic for this hypotheses is:

$$t = \frac{\hat{s}}{SE(\hat{s})}$$

- The test statistic summarizes all the information needed to make the decision.

Most important t-tests

When the H_0 is: the average is equal to zero, the t-statistic is simply

$$t = \frac{\bar{x}}{SE(\bar{x})} \tag{1}$$

When the H_0 is: the average is equal to a specific number, the t-statistic is

$$t = \frac{\bar{x} - number}{SE(\bar{x})} \tag{2}$$

When H_0 compares two averages: $\bar{x}_A - \bar{x}_B = 0$, the t-statistic is

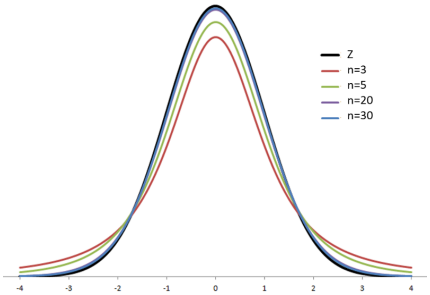
$$t = \frac{\bar{x}_A - \bar{x}_B}{SE(\bar{x}_A - \bar{x}_B)} \tag{3}$$

t-statistics under the null

"Under the null" or if the H_0 is true, the *t-statistics* comes from a sampling distribution which is distributed as 'Student's-t' distribution.

- Student's-t is similar to the standard normal distribution.
- It has mean zero and standard deviation of 1.
- It has a third parameter 'degree of freedom' which in our case relates to the number of observations.

What is the likelihood that our t-statistic (t) comes from this distribution?



Making a decision

- In hypothesis testing the decision is based on a clear rule *specified in advance*.
- A decision rule makes the decision straightforward and transparent.
- Helps avoid personal bias: put more weight on the evidence that supports our prejudices.
- Clear decision rules are designed to minimize the room for such temptations.

Making a decision

- The decision rule = comparing the test statistic to a pre-defined *critical value*.
- Is test statistic is large enough to reject the null?

- Null rejected if the test statistic is larger than the critical value.
- Critical value governs the trade-off between being too strict or too lenient with our decision.

Being right or wrong

When we make the decision, we may be right or wrong in two ways.

	H_0 is true	H_0 is false
Don't reject the null	True negative	False negative - Type II error
Reject the null	False positive - Type I error	True positive

Making an error

- We say that our decision is a *false positive* if we reject the null when it is true.
  - "positive" because we take the active decision to reject the protected null.
  - medical: person has the condition that they were tested against
  - False positive = type-I error;
- Our decision is a *false negative* if we do not reject the null even though we should.
  - "negative" because we do not take the active decision
  - medical: result is "negative" = not have the condition
  - False negative =type-II error.

Protecting against Type-I error

- False positives and false negatives: both wrong, but not equally.
- Testing procedure protects the null: reject it only if evidence is strong
- The background assumption - wrongly rejecting the null (a false positive) is a bigger mistake than wrongly accepting it (a false negative).
- Decision rule (critical value) is chosen in a way that makes false positives rare.

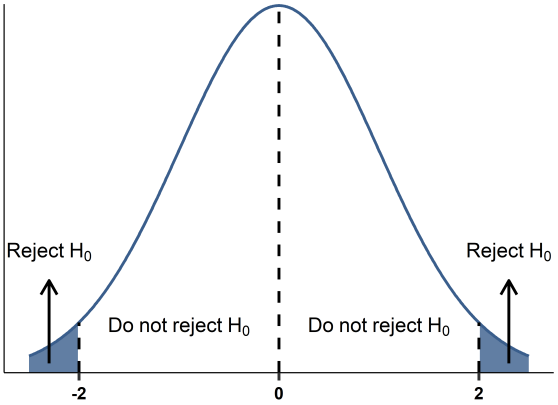
Rule of thumb when making a decision

- A commonly applied critical value for a t-statistic is ± 2:
  - reject the null if the t-statistic is smaller than -2 or larger than $+2$;
  - don't reject the null if the t-statistic is between -2 and $+2$.
- Prob(t-statistic < -2) or Prob(t-statistic > 2) are both appr 2.5%
- If the null is true: Probability t-statistic is below -2 or above $+2$ is 5%
- With ± 2 critical value - 5% is the probability of false positives - we have 5% as the probability that we would reject the null if it was true (False positve).

- If we make the critical values -2.6 and $+2.6$ the chance of the false positive is 1%.

Sampling distribution of the test statistic when the null is true

- Distribution of the t-statistic would be close to standard normal $N(0, 1)$, if we have medium sample size
- Prob t-statistic < -2 or > 2 is approximately 2.5%. Prob t-statistic is < -2 or $> +2$ is 5% if the null is true. (Two-sided alternative)
- 5% = probability of false positives if we apply the critical values of ± 2

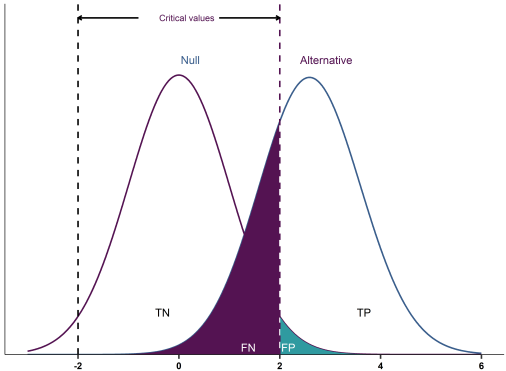


Critical values and generalization

- Can set other critical values that correspond to different probabilities of a false positive.
- That choice of 5% means that we tolerate a 5% chance for committing false positive error
- Data analysts avoid biases when testing hypotheses: use the same critical value regardless of the data and hypothesis they are testing.

False negative

- Fixing the chance of false positives affects the chance of false negatives at the same time.
- A false negative arises when the t-statistic is within the critical values and we don't reject the null even though the null is not true.
- Making a false negative call is more likely when it is harder to make a decision



Size and power of the test

Under the null:

- *Size* of the test: the probability of committing a false positive.
- *Level of significance*: The maximum probability of false positives we tolerate.

When we fix the level of significance at 5% and end up rejecting the null, we say that the statistic we tested is significant at 5%

Under the alternative:

- *Power* of the test: the probability of avoiding a false negative
- Being different from the null can be in many ways...
- High power is more likely when
  - The sample is large and the dispersion is small.
  - The further away the true value is from what's in a null.

We usually fix the level of significance at 5% and hope for a high power of the test.

The p-value

- The p-value makes testing easier - captures information for reject/accept calls.
  - Instead of calculating test statistics and specify critical values, we can make an informed decision based on the p-value only.

- p-value is the smallest significance level at which we can reject H_0 given the value of the test statistic in the sample.
- The p-value tells us the largest probability of a false positive.
- The p-value depends on
  1. the test statistic,
  2. the sampling distribution of the test statistic

The p-value

- If the p-value is 0.05 the maximum probability that we make a false positive decision is 5%.
- If we are willing to take that chance, we should reject the null; if we aren't, we shouldn't.
- If the p-value is, say, 0.001 there is at most a 0.1% chance of being wrong if we were to reject the null.
- We can never be absolute certain! p-value is never zero.
- For a reject/accept decision, one should pick a level of significance before the test.
- What we can accept depends on the setting: what is the cost of a false positive.

Comparing online and offline prices: Testing hypotheses

- Let's fix the level of significance at 5%.
- Doing so we tolerate a 5% chance for a false positive.
- Allow a 5% chance to be wrong if we reject the null hypothesis of zero average price difference.
- A 5% level of significance translates to ± 2 bound for the t-statistic.
- The value of the statistic in the dataset is -0.054. Its standard error is 0.124.
- Thus the t-statistic is 0.44. This is well within ± 2.
- Don't reject the null hypothesis of zero difference.
- We do *not* say we proved it's zero!!! We showed we cannot tell it different from zero.

Comparing online and offline prices: Testing hypotheses

- Conclude that the average price difference is not different from zero in the general pattern represented by the data.
- Large dataset, good power. What we see in t-statistic is not because of very small sample size.
- It is still possible that prices are indeed different, just the difference is very small.
- Economically speaking, a few cent difference would not matter.

Comparing online and offline prices: Testing hypotheses

- The p-value of the test is 0.66.
- That means that the smallest level of significance at which we can reject the null is 66%.
- The chance that we would make a mistake if we rejected the null is at most 66%.
- So we don't reject the null

Multiple testing

- Medical dataset: data on 400 patients
- A particular heart disease binary variable and 100 feature of life style (sport, eating, health background, socio-economic factors)
- Look for a pattern – is the heart disease equally likely for poor vs rich, take vitamins vs not, etc.
- You test one-by-one
- You find that for half a dozen factors, there is a difference
- Any special issue?

Multiple testing

- The pre-set level of significance / p-value are defined for a single test
- In many cases, you will consider doing many many tests.
  - Different measures (mean, median, range, etc)
  - Different products, retailers, countries
  - Different measures of management quality
- For multiple tests, you cannot use the same approach as for a single one.

Multiple testing

- Consider a situation in which we test 100 hypotheses.
- Assume that all of those 100 null hypotheses are true.
- Set significance - we accept 5% chance to be wrong when rejecting the null. That means that we tolerate if we are wrong 5 out of 100 times.
- We can expect the null to be rejected 5 times when we test our 100 null hypotheses, all of which are true.
- In practice that would appear in 5 out of the 100 tests
- We could pick those five null hypotheses and say there is enough evidence to reject.
- But that is wrong: we started out assuming that all 100 nulls are true.

- Simply by chance, we will see cases when we would reject the null, but we should not -> committing false positive error!

Multiple testing

- There are various ways to deal with probabilities of false positives when testing multiple hypotheses.
- Often complicated.

- Solution 1: If you have a few dozens of cases, just use a strict criteria (such as 0.1-0.5% instead than 1-5%) for rejecting null hypotheses.
- A very strict such adjustment is the Bonferroni correction that suggests dividing the single hypothesis value by the number of hypotheses.
  - For example, if you have 20 hypotheses and aim for a p=.05
  - reject the null only if you get a p=0.05/20=0.0025
  - It is typically too strict

p-hacking

P-hacking: Showing only results of hypothesis tests that suggest one decision when different choices in the data wrangling would lead to different conclusions.

- Handling of missing values
- Dealing with extreme values
- Conditioning on such x's
- Dropping observations

All these leads to 'set' the p-value such that you have your wanted result...

These methods are wrong and undermine the profession!

Testing when data is very big

- With very large dataset some aspects of statistical inference lose their relevance.
- When the data has millions of observations generalizing to the general pattern does not add much.
- That is true for testing hypotheses, too.
- If, for example, two averages calculated from millions of observations are different to a meaningful extent they are almost surely different in the general pattern represented by the dataset.
- So: if you have millions of observations, just look at meaningful difference - do not worry about hypotheses testing (unless you care about very very small differences)