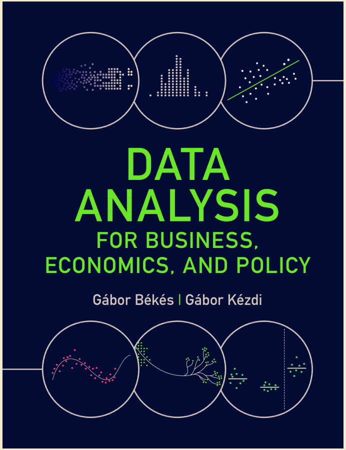
Békés-Kézdi: Data Analysis, Chapter 06: Hypotheses testing



Data Analysis for Business, Economics, and Policy

Gábor Békés (Central European University)
Gábor Kézdi (University of Michigan)

Cambridge University Press, 2021

gabors-data-analysis.com

Central European University

Any comments or suggestions:
gabors.da.contact@gmail.com

Motivation

- The internet allowed the emergence of specialized online retailers while brick-and-mortar shops also sell goods on the main street. How to measure price inflation in the age of these options?
- To help answer this, we can collect and compare online and offline prices of the same products and test if they are the same.

The logic of hypothesis testing

The logic of hypothesis testing

- A hypothesis is a statement about a general pattern, of which we are not sure if true or not.
- **Hypothesis testing** = analyze our data to make a decision 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 may not have enough evidence against a hypothesis
  - if the hypothesis is true
  - or it is not true only the evidence is weak
- Important asymmetry here: rejecting a hypothesis is a more conclusive decision than not rejecting it.

The logic of hypothesis testing: inference

- Testing a hypothesis: making inference with a focus on a specific statement.
- Can answer questions about the population, or general pattern, represented by our data.
- It is an inference: have to assess external validity

The logic of hypothesis testing: the setup

- Define the *the statistic we want to test*, s (e.g. mean).

- We are interested in the true value of s, s_{true}.
- This is statistical inference, so the true value means the value in the population, or general pattern represented by our data.

- 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/ range; the alternative specifies other possible values.
- Together, the null and the alternative cover all the possibilities we are interested in
- One example is null: s is zero, alternative: s is not zero.

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

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

The logic of hypothesis testing: H0 vs HA

- Our case study research question: Do the online and offline prices of the same products differ or are they the same?
- We have the price difference as our statistic and $H_0 : s_{true} = 0$
- Testing a hypothesis = see if there is enough evidence in our data to reject the null.

The logic of hypothesis testing: Null protected

- Testing a hypothesis = see if there is enough evidence in our data to reject the null.
- The null is protected: it has to be hard to reject it otherwise the conclusions of hypothesis testing would not be strong.

The logic of hypothesis testing: The criminal court example

- Logic of testing like a criminal court procedure.
  - Decide if the accused is guilty or innocent of a certain crime.
  - Assumption of innocence: accused judged guilty only if enough evidence against innocence
  - Even though the accused in court because of suspicion of guilt.
- To translate this procedure to the language of hypothesis testing,
  - H_0 is that the person is innocent
  - H_A is that the person is guilty.

The logic of hypothesis testing: H0 vs HA

- **Two-sided** alternative: The case when 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.

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

- **One-sided** alternative: interested if a statistic is positive or not.

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

Summary of the logic of hypothesis testing

- H_A is (often) what I wanna prove
- H_0 is what I wanna reject so that I can prove H_A

- H_0 is **not** rejected
  - not enough evidence or
  - true (ie H_A is false)
- I can never say H_0 is true.

Case Study - Comparing online and offline prices: Testing hypotheses

- Question: Do the online and offline prices of the same products differ?
- this 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 offline stores, and they were matched to the same products the same stores sold online.

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

Case Study - Comparing online and offline prices: Testing hypotheses

- Descriptive statistics of the difference
- Each product i has both an online and an offline price in the data, $p_{i,online}$ and $p_{i,offline}$, $pdiff$ is their difference:

$$pdiff_i = p_{i,online} - p_{i,offline} \tag{1}$$

The statistic with n observations (products) in the data, is:

$$s = \overline{pdiff} = \frac{1}{n}\sum_{i=1}^{n}(p_{i,online} - p_{i,offline}) \tag{2}$$

Case Study - Comparing online and offline prices: Testing hypotheses

- The average of the price differences is equal to the difference of the average prices
- s statistic also measures the difference between the average of online prices and the average of offline prices among products with both kinds of price

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

Case Study - Comparing online and offline prices: Testing hypotheses

Descriptive statistics of the difference

- The mean difference is USD -0.05: online prices are, on average, 5 cents lower in this dataset.
- Spread around this average: Std: USD 10
- Extreme values matter: Range: -380 — USD +415.
- Of the 6439 products, 64% have the same online and offline price, for 87%, the difference within ± 1 dollars.

Case Study - Comparing online and offline prices: the setup

External validity

- The products in the data may not represent all products sold at these stores.
    - Could be a bias. Example?
- Strictly: The general pattern of the statistic represented by this dataset is average online-offline price differences in large retail store chains for the kind of products that data collectors would select with a high likelihood.
- More broadly: price differences among *all* products in the U.S. sold both online and offline by the same retailers.
    - Need an assumption. What would it be?

Case Study - Comparing online and offline prices: the setup

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

$$H_0 : s_{true} = \bar{p}_{online\ true} - \bar{p}_{offline\ true} = 0 \quad (3)$$

$$H_A : s_{true} = \bar{p}_{online\ true} - \bar{p}_{offline\ true} \neq 0 \quad (4)$$

Testing

The logic of hypothesis testing

- The **t-test** is the testing procedure based on the **t-statistic**
- We compare the estimated value of the statistic $\hat{s}$ (our best guess of s) to zero.

- Evidence to reject the null = based on difference between $\hat{s}$ and zero.
- **Reject the null** if difference large = it is unlikely to be zero.
- **Not reject** the null if the difference is small = not enough evidence against it.

- Need to define "large"/"small" (*next*)

T-test

- The **test 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 estimated value of s ($\hat{s}$) and the standard error of estimate ($SE(\hat{s})$).
  - SE is the scaling (normalization)

- Consider $H_0 : s_{true} = 0, H_A : s_{true} \neq 0$. The t-statistic for this hypotheses is:

$$t = \frac{\hat{s}}{SE(\hat{s})} \tag{5}$$

- The test statistic summarizes all the information needed to make the decision.
- When hypotheses are about value of one coefficient the test statistic = t-statistic

T-test

When $\hat{s}$ is the average of a variable x, the t-statistic is simply

$$t = \frac{\bar{x}}{SE(\bar{x})} \tag{6}$$

When $\hat{s}$ is the average of a variable x minus a number, the t-statistic is

$$t = \frac{\bar{x} - number}{SE(\bar{x})} \tag{7}$$

When $\hat{s}$ is the difference between two averages, say, $\bar{x}_A$ and $\bar{x}_B$, the t-statistic is

$$t = \frac{\bar{x}_A - \bar{x}_B}{SE(\bar{x}_A - \bar{x}_B)} \tag{8}$$

T-test

- If $\hat{s} > 0 =$ the t-statistic is positive; if $\hat{s} < 0 =$ the t-statistic is negative.
- With a two-sided alternative ($H_A : s_{true} \neq 0$) it is the magnitude not the sign of the t-statistic that matters.

- If $\hat{s} = 0$ then $t = 0$.
  - In reality it's never *exactly* zero.
  - But expect $\hat{s}$ estimate to be *close* to zero.
- If the null is incorrect and thus s_{true} is *not* zero –> we expect the $\hat{s}$ estimate to be far from zero.

T-test

- We standardize distance with $SE(\bar{x})$
- May use $SE(\bar{x}) = \sqrt{\frac{1}{n}}Std[x]$.
- SE formula may be more complicated
- Sometimes no appropriate SE formula for a statistic interested in –> Need bootstrap estimation.

Generalization

Making a decision

- In hypothesis testing the decision is based on a clear rule specified in advance.
- A decision rule makes the decision straightforward + 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: decision rule

- 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 - between being too strict or too lenient.
- When we make the decision, we may be right or wrong, don't know: need to think

Making a decision

- We can be right in our decision in two ways:
  - we reject the null when it is not true,
  - or we do not reject the null when it is true.
- We can be wrong in our decision in two ways, too:
  - we reject the null even though it is true,
  - or we do not reject the null even though is not true.

	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 a decision

- 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.

Making a decision

- 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.

Making a decision

- A commonly applied critical value for a t-statistic is ± 2 (or 1.96):
  - 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 $+2$.
- 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 positive).
  - 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%
- If we make the critical values -2.6 and $+2.6$ the chance of the false positive is 1%.

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 being wrong when rejecting the null
- Data analysts avoid biases when testing hypotheses: use the same critical value regardless of the data and hypothesis they are testing.

Critical values and generalization

- Where does this 2SD – 5% come from?
- We can calculate the likelihood of a false positive because we know what the sampling distribution of the test statistic would be if the null were true.
- The sampling distribution of a statistic is its distribution across repeated samples
  - of the same size from the same population.
- The sampling distribution of an average is approximately normal, its mean is equal to the true mean, and its standard deviation is called the standard error.

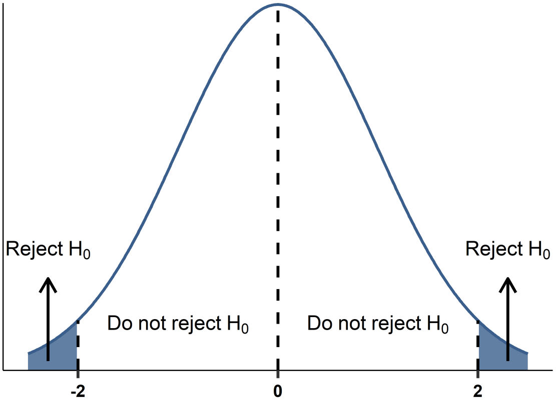
Critical values and generalization

- How would the sampling distribution look if the null hypothesis were true:
- Distribution of the t-statistic would be standard normal $N(0, 1)$

- The t-statistic has the average in its numerator, so that its distribution is also approximately normal,
- The t-statistic SD=1 because because the t-statistic is standardized – it has the SE of $\hat{s}$ in the denominator
  - Note: Small sample (<30), the normal approximation to the distribution of the t-statistic is not very good. Instead, the distribution is closer to "t-distribution")

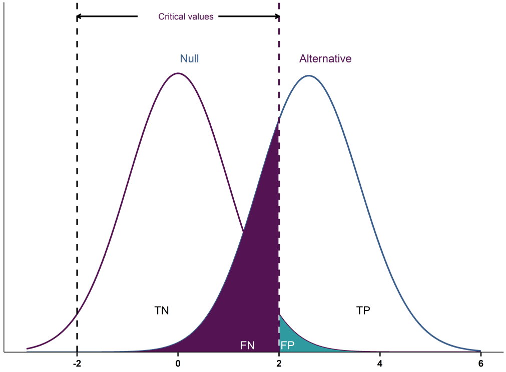
Sampling distribution of the test statistic when the null is true

- Distribution of the t-statistic close to $N(0, 1)$
- 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



False negative (FN)

- Fixing the chance of FP affects the chance of FN at the same time.
- A FN 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 FN call more likely when harder to make a decision
  - Sample is small
  - The difference between true value and null is small



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.

Making a decision

- We know the sampling distribution of the test statistic if the null is true–> can calculate the likelihood of a false positive
- Recall: sampling distribution of an average value is approximately normal,
  - mean= being equal to the true mean value,
  - the standard deviation being equal to its standard error.
- The distribution of the t-statistic is standard normal distribution N(0,1)
  - It has mean zero because $s_{true} = 0$ if the null is true.
  - It has standard deviation one because the standard deviation of the sampling distribution of $\hat{s}$ is SE($\hat{s}$), and the t-statistic is $\hat{s}/SE(\hat{s})$.

Recap

- In hypothesis testing we make decisions by a rule
  - A false positive is a decision to reject the null hypothesis when it is in fact true.
  - A false negative is a decision not to reject the null hypothesis when it is in fact not true.
- The level of significance is the maximum probability of a false positive that we tolerate.
- The power of the test is the probability of avoiding a false negative.

- In statistical testing we fix the level of significance of the test to be small (5%, 1%) and hope for high power.
- Tests with more observations have more power in general.

The p-value

- The p-value makes testing easier - captures info 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 is the probability that the test statistic will be as large as, or larger than, what we calculate from the data, if the null hypothesis is true.*
- The p-value tells us the largest probability of a false positive.
- The p-value depends on
  1. the test statistic,
  2. the critical value
  3. the sampling distribution of the test statistic

Recap: p vs power

- p-value = probability rejecting the null while it is true (probability of avoiding FP).
- Power = probability rejecting the null while it is false (probability of avoiding FN)

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 are not, we should not.
  - 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 certain! p 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.

What p-value to pick?

- p-value is about a trade-off. Large (10-15%) or small (1%) depends on scenarios
- Guilty beyond reasonable doubt?
- Proof of concept?

What p-value to pick?

- p-value is about a trade-off. Large (10-15%) or small (1%) depends on scenarios

- Guilty beyond reasonable doubt?
- Pick a conservative value, like 1% or lower

- Proof of concept?
- It's great if it works at 5%, but even 10-15% means it's much more likely to be true
  - May lead to doing more experimentation, increase sample size

One-sided t-test, calculating p-value

- One sided test: having an inequality in H_A
- $H_0 : s_{true} \geq 0$ against $H_A : s_{true} < 0$
- Equality always part of the null

- In order to reject H_0, we need to reject each and every value in favor of $s < 0$
- Hardest value to reject against is $s = 0$ against $s < 0$
  - this is why equality is part of the null
- Difference to two sided: we only care about being wrong on one side,
  - the probability of FP is smaller (=half)
  - t-test of two-sided hypotheses — the p-value as the sum of two probabilities – we only have half the probability of error
- Practically: run a two-sided test, calculate p-value and take its half.

Case Study - 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.
- The CI is $-0.054 \pm 2 * 0.124 = [-0.30, +0.19]$
  - 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 apart from zero.

Case Study - 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. A few cent difference would not matter economically ...

Case Study - 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 test

Multiple testing: motivation

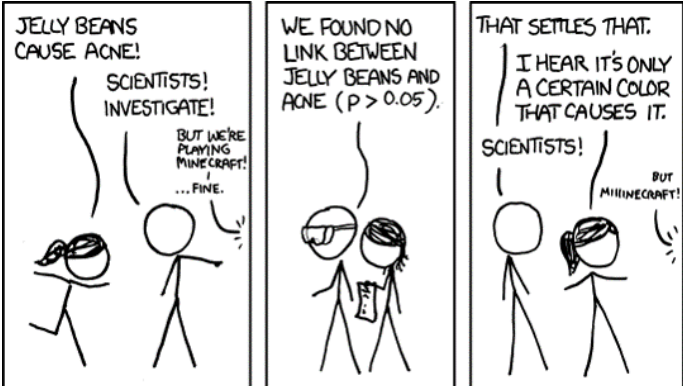
- Medical dataset: data on 400 patients
- A particular heart disease binary variable and 100 feature of life style (sport, eating, health background, socio-economic factors)
- Look for a pattern – is the heart disease equally likely for poor vs rich, take vitamins vs not, etc.
- You test one-by-one
-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

Multiple testing

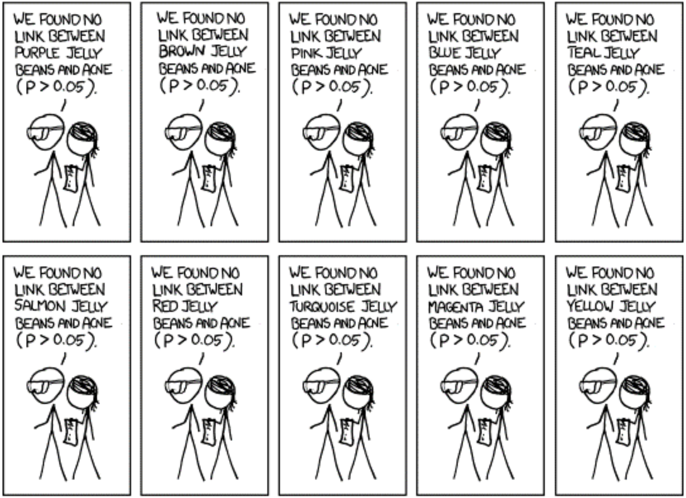
- The pre-set level of significance / p-value are defined for a single test
- In many cases, you will consider doing many many tests.
  - Different measures (mean, median, range, etc)
  - Different products, retailers, countries
  - Different measures of management quality

- For multiple tests, you cannot use the same approach as for a single one.

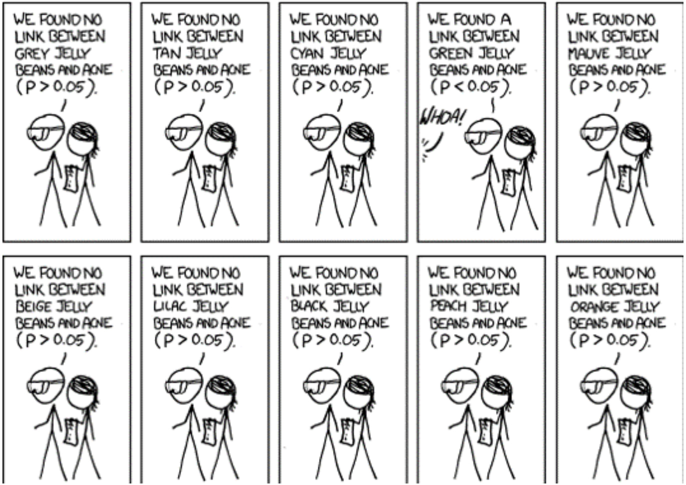
Multiple testing - a serious example



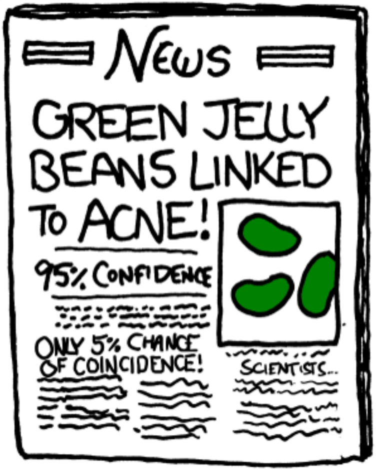
Multiple testing - a serious example



Multiple testing - a serious example



Multiple testing - a serious example



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

p-hacking

- Practice of doing many tests, and picking what works...

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 two strict

Testing when data is very big

- Very large datasets – statistical inference lose relevance.
- Millions of observations generalizing to the general pattern does not add much.
- That is true for testing hypotheses, too.
- 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)

AI: Asking for a formula

Can you show me the formula for a t-test for a difference in the means of a variable x in two samples. Also show in latex.

- What you need to know is that there should be one, different from what you have seen.
- Mostly correct, but be able to check...

AI answer

To get t-test for the difference in two sample means, let us consider two independent samples, $x1$ and $x2$:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{Std_{x1}^2}{n_1} + \frac{Std_{x2}^2}{n_2}}}$$

Where:

t is the t-statistic.

$\bar{x}_1$ and $\bar{x}_2$ are the sample means.

Std_{x1} and Std_{x2} are the standard deviations in $x1$ and $x2$.

n_1 and n_2 are the sample sizes of $x1$, $x2$.

Summary

Testing in statistics means making a decision about the value of a statistic in the general pattern represented by the data.

- Hypothesis starts with explicitly stating H0 and HA.
- A statistical test rejects H0 if there is enough evidence against it; otherwise it does not reject it.
- Testing multiple hypotheses at the same time is a tricky business; it pays to be very conservative with rejecting the null.

Extra

A special case in testing: the one sided-alternative

- Have only one of the inequalities in the alternative
- This leads to focusing on one side of the test statistic only
- Two most frequent examples are
  - $H_0 : s_{true} \leq 0$ against $H_A : s_{true} > 0$
  - $H_0 : s_{true} \geq 0$ against $H_A : s_{true} < 0$.

- Having zero is key. If we can reject zero, we can reject anything below (above)
  - Test $H_0 : s_{true} \leq 0$ vs $H_A : s_{true} > 0$ –> $H_0 : s_{true} = 0$ vs $H_A : s_{true} > 0$.
  - Test $H_0 : s_{true} \geq 0$ vs $H_A : s_{true} < 0$ –> $H_0 : s_{true} = 0$ vs $H_A : s_{true} < 0$.

One sided-alternative

- Focusing on deviations in one direction means that we care about one half of the sampling distribution of the test statistic.
- With $H_0 : s_{true} \leq 0$ against $H_A : s_{true} > 0$, we care about whether $\hat{s}$ is large positive enough in order to reject the null; if it is negative we don't reject it.
- The probability of a false positive is smaller in this case. We don't reject the null if the test statistic falls in the region that is specified in the null hypothesis.
- Thus, we make a false positive decision only half of the times.
- t-test of two-sided hypotheses — the p-value can be thought of as the sum of two probabilities
- So we only have half the probability of error

One sided-alternative

Therefore, the practical way to testing one-sided hypotheses is a two-step procedure.

1. If the test statistic is in the region of the null don't reject the null.

   This happens if $\hat{s}$ is in the region of the null (e.g., $\hat{s} < 0$ for $H_0 : s_{true} \leq 0$ against $H_A : s_{true} > 0$);

2. If the test statistic is in the region of the alternative proceed with testing the usual way with some modification.

   Ask the software to calculate the p-value of the null hypothesis of the equality (for example, $H_0 : s_{true} = 0$ if the true null is $H_0 : s_{true} \leq 0$) and divide the p-value by two.