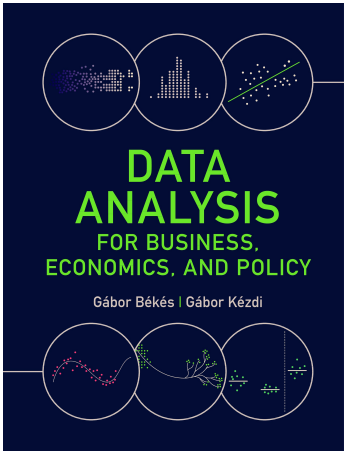
II: Generalizing from data and testing hypotheses

Alexandra Avdeenko

Data Analysis

2021

Literature



- Slides 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 lecture is based on Chapters 5 and 6
- Lecture slides for the book shared by G. Békés

Overview

Motivation

- *How likely is it that we shall experience losses on our investment portfolio? To answer this, you have collected and analyzed past financial information.*
- To predict the frequency of a loss of certain magnitude for the coming calendar year, you will need to make an inference and think hard about what can be different in the future.

Inference

Generalization

- Sometimes we analyze a dataset with the goal of learning about patterns in that dataset alone.
- In such cases there is no need to generalize our findings to other datasets.
- Example: We search for a good deal among offers of hotels, all we care about are the observations in our dataset.

- Often we analyze a dataset in order to learn about patterns that may be true in other situations.
- We are interested in finding the relationship between
  - Our dataset
  - The situation we care about

Generalization

- Generalize the results from a single dataset to other situations.
- The act of generalization is called *inference*: we infer something from our data about a more general phenomenon because we want to use that knowledge in some other situation.

- Aspect 1: statistical inference
- Aspect 2: external validity

Statistical inference

- Uses statistical methods to make inference.
- Well-developed and powerful toolbox that helps generalizing to situations similar to our data.
- Similar to ours = general pattern represented by our dataset.

- The general pattern is an abstract thing that may or may not exist.
- If we can assume that the general pattern exists, the tools of statistical inference can be very helpful.

General patterns 1: Population and representative sample

- The cleanest example of representative data is a representative sample of a well-defined *population*.
- A sample is representative of a population if the distribution of all variables is very similar in the sample and the population.
- Random sampling is the best way to achieve a representative sample.

General patterns 2: No population but general pattern

The concept of representation is less straightforward in other setups.

- Using data with observations from the past to uncover a pattern that may be true for the future.
- Generalizing patterns observed among some products to other, similar products.

There isn't necessarily a "population" from which a random sample was drawn on purpose. Instead, we should think of our data as one that represents a general pattern.

- There is a general pattern, each year is a random realization.
- There is a general pattern, each product is a random version, all represented by the same general pattern.

External validity

- Assessing whether our data represents the same general pattern that would be relevant for the situation we truly care about.
- Externally valid case: the situation we care about and the data we have represent the same general pattern
- With external validity, our data can tell what to expect.
- No external validity: whatever we learn from our data, may turn out to be not relevant at all.

The process of inference

The process of inference

1. Consider a statistic we may care about, such as the mean.
2. Compute its *estimated value* from a dataset
3. Infer the value in the population / in the general pattern, that our data represents.

It is good practice to divide the inference problem into two.

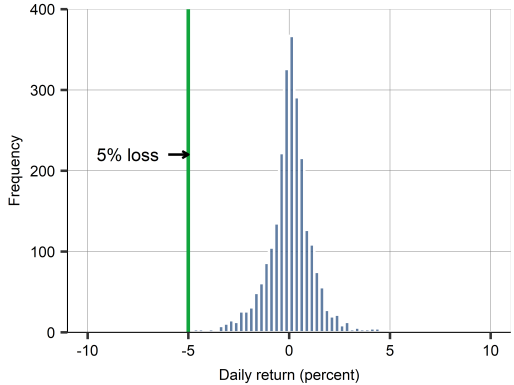
1. Use statistical inference to learn about the population, or general pattern, that our data represents.
2. Assess external validity: define the population, or general pattern we are interested in and assess how it compares to the population, or general pattern, that our data represents.

Stock market returns: Inference

- Task: Assess the likelihood of experiencing a loss of certain magnitude on an investment portfolio from one day to the next day
- Predict the frequency of a loss of certain magnitude for the coming calendar year

- The investment portfolio is the S&P 500, a US stock market index
- Data: day-to-day returns on the S&P 500, defined as percentage changes in the closing price of the index between two consecutive days
- 11 years: 25 August 2006 to 26 August 2016. It includes 2,519 days.

Histogram of daily returns



Note: *S&P 500 market index. Day to day (gaps ignored) changes, in percentage. From August 25 2006 to August 26 2016.*

Stock market returns: Inference

- To define "loss", we take a day-to-day loss exceeding 5 percent.
- "loss" is a binary variable, taking 1 when the day-to-day loss exceeds 5 percent and zero otherwise.
- The statistic in the data is the proportion of days with such losses.
- It is 0.5 percent in this dataset
  - the S&P500 portfolio lost more than 5 percent of its value on 0.5 percent of the days between August 25 2006 and August 26 2016.

- Inference problem: How can we generalize this finding? What can we infer from this 0.5 percent chance for the next calendar year?

Repeated samples

Repeated samples

- Repeated samples - the conceptual background to statistical inference
- Our data - one example of many datasets that could have been observed.
- Each datasets can be viewed as samples drawn from the population (general pattern)

- Easier concept: When our data is sample from a well-defined population - many other samples could have turned out instead of what we have.
- Harder concept: no clear definition of population. We think of a general pattern we care about.

Repeated samples

- The goal of statistical inference is learning the value of a statistic in the population, or general pattern, represented by our data.
- The statistic has a distribution: its value may differ from sample to sample.
- The distribution of the statistic of interest is called its sampling distribution

Repeated samples

- Standard deviation in this distribution: spread across repeated samples
- The standard error (SE) of the statistic = the standard deviation of the sampling distribution
- Any particular estimate is likely to be an erroneous estimate of the true value. The magnitude of that typical error is one SE.

Repeated samples properties

The sampling distribution of a statistic is the distribution of this statistic across repeated samples.
The sampling distribution has three important properties

1. Unbiasedness: The average of the values in repeated samples is equal to its true value (=the value in the entire population / general pattern).
2. Asymptotic normality: The sampling distribution is approximately normal. With large sample size, it is very very close.
3. Root-n convergence: The standard error (the standard deviation of the sampling distribution) is smaller the larger the samples, with a proportionality factor of the square root of the sample size.

Stock market returns: A simulation

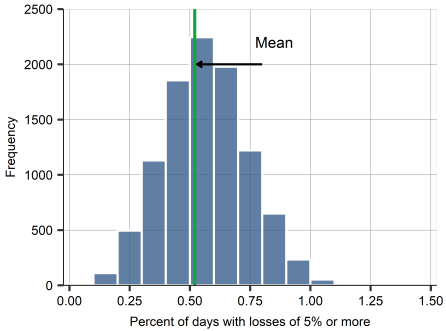
- We can not rerun history many many times...
- Simulation exercise - to better understand how repeated samples work
- Suppose the 11-year dataset is *the* population - the fraction of days with 5%+ losses is 0.5% in the entire 11 years' data. That's the true value.
- Assume we have only three years (900 days) of daily returns in our dataset.
- Task: estimate the true value of the fraction in the 11-year period from the data we have using a simulation exercise.
  1. many data table with three years' worth of observations may be created from the 11 years' worth of data,
  2. compute the fraction of days with 5%+ losses in data tables
  3. learn about the true value

Stock market returns: A simulation

- Do simple random sampling: days are considered one after the other and are selected or not selected in an independent random fashion.
  - This sampling destroys the time series nature
  - This is OK because daily returns are (almost) independent across days in the original dataset
- We do this 10,000 times....

Stock market returns: A simulation

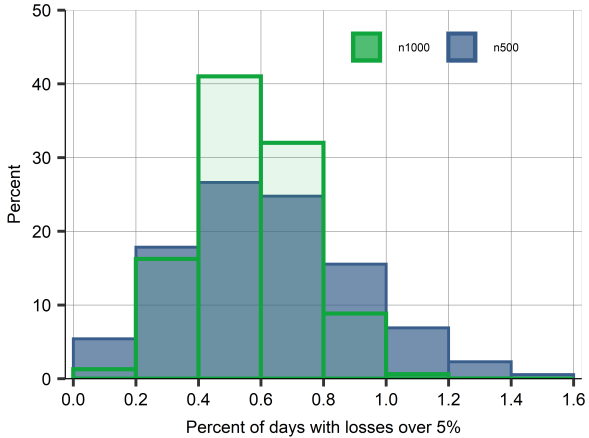
- percent of days with losses of 5% of more.
- histogram created from the 10,000 random samples, each w/ 900 obs, drawn from entire dataset
- distribution has some spread: smallest realization is 0.1 %, while the largest is smaller than 1.25 %



Histogram of the proportion of days with losses of 5 percent or more, across repeated samples of size n=900. 10,000 random samples. Source: `sandp-stocks` data. S&P 500 market index.

Stock market returns: Sampling distributions

- Proportion of days with losses of 5 percent or more
- Repeated samples in two simulation exercises, with n=500 and n=1,000. (10,000 random samples)
- Kernel density (goes to minus / can cut it at 0)
- Role of sample size: smaller sample: skewed; higher standard deviation



The Confidence Interval

The standard error and the confidence interval

- Confidence interval (CI) - measure of statistical inference.
  - Recall: Statistical inference - we analyze a dataset to infer the true value of a statistic: its value in the population, or general pattern, represented by our data.
- The CI defines a range where we can expect the true value in the population, or the general pattern.
- CI gives a range for the true value with a probability
- Probability tells how likely it is that the true value is in that range
- Probability - data analysts need to picks it, such as 95%

The standard error and the confidence interval

- The "95 percent CI" gives the range of values where we think that true value falls with a 95 percent likelihood.
- Viewed from the perspective of a single sample, the chance (probability) that the truth is within the CI measured around the value estimated from that single sample is 95 percent.

- Also: we think that with 5 percent likelihood, the true value will fall outside the confidence interval.

The standard error and the confidence interval

- Confidence interval - symmetric range around the estimated value of the statistic in our dataset.
  - Get estimated value.
  - Define probability
  - Calculate CI with the use of SE
- 95 percent CI is the $\pm 1.96SE$ (but we use $\pm 2SE$) interval around the estimate from the data.
  - 90% CI is the $\pm 1.6SE$ interval, the 99 % CI is the $\pm 2.6SE$

Calculating the standard error

An important consequence of evidence from the repeated sample exercise:

- In reality, we don't get to observe the sampling distribution. Instead, we observe a single dataset
- That dataset is one of the many potential samples that could have been drawn from the population, or general pattern
- Good news: We can get a very good idea of how the sampling distribution would look like - good estimate of the standard error - even from a single sample.
- Getting SE – Option 1: Use a formula
- Getting SE – Option 2: Simulate by a new method, called 'bootstrapping'

Calculating the standard error

Consider the statistic of the sample mean.

- Assume the values of x are independent across observations in the dataset.
- $\bar{x}$ is the estimate of the true mean value of x in the general pattern/population. (LLN)
- Sampling distribution is approximately normal, with the true value as its mean. (CLT)

$$\bar{x} \overset{a}{\sim} \mathcal{N}\left(E\left[x\right], \frac{1}{n} Var\left[\bar{x}\right]\right)$$

The standard error formula for the estimated $\bar{x}$ is

$$SE(\bar{x}) = \frac{1}{\sqrt{n}} Std[x] \tag{1}$$

where $Std[x]$ is the standard deviation of the variable x in the data and n is the number of observations in the data.

The standard error formula

- The standard error is larger...
  - the larger the standard deviation of the variable.
  - the smaller the sample and
- For intuition, consider $SE(\bar{x})$ vs $Std[x]$.
- Think back to the repeated samples simulation exercise:
  - $SE(\bar{x})$ = the standard error of $\bar{x}$ is the standard deviation of the various $\bar{x}$ estimates across repeated samples.
  - The larger the standard deviation of x itself, the more variation we can expect in $\bar{x}$ across repeated samples.

Stock market returns: The standard error formula

Let's consider our example of 11-years' of data on daily returns on the S&P 500 portfolio.

- The calculated statistics, $P(loss > 5\%) = 0.5\%$
- The $SE\left[P(loss > 5\%)\right]$ is calculated by,
  - The size of the sample is n = 2,519 so that $1/\sqrt{n} = 0.02$.
  - The standard deviation of the fraction of $SD\left[P(loss > 5\%)\right] = 0.07$.
  - So the $SE = 0.07 * 0.02 = 0.0014$ (0.14 percent).
- Can calculate the 95 percent CI:
  - $CI = [0.5 - 2 * SE, 0.5 + 2 * SE] = [0.22, 0.78]$
- This means that in the general pattern represented by the 11-year history of returns in our data, we can be 95 percent confident that daily losses of more than 5 percent occur with a 0.2 to 0.8 percent chance.

Take a quick stop to summarize the idea of CI

- We are interested in generalizing from our data. Statistical inference.
- Consider a statistic such as the sample mean $\bar{x}$
- Take a 95% confidence interval - where we can expect to see the true value
- CI=statistic $+/-2SE$.
- We have a formula for the SE calculated from our the data only using the standard deviation and sample size.
- Using the CI, we can now do statistical inference, generalize for the population / general pattern we care about.

The bootstrap SE

The bootstrap

- Bootstrap is a method to create synthetic samples that are similar but different
- An method that is very useful in general.
- It is essential for many advanced statistics application such as machine learning

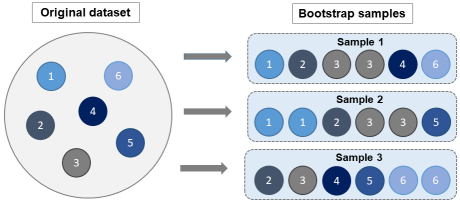
- More in Chapter 05

The bootstrap

- The bootstrap method takes the original dataset and draws many repeated samples of the size of that dataset.
- The trick is that the samples are drawn *with replacement*.
- The observations are drawn randomly one by one from the original dataset; once an observation is drawn it is "replaced" to the pool so that it can be drawn again, with the same probability as any other observation.
- The drawing stops when it reaches the size of the original dataset.
- The result is a sample of the same size as the original dataset, yielding a single *bootstrap sample*.

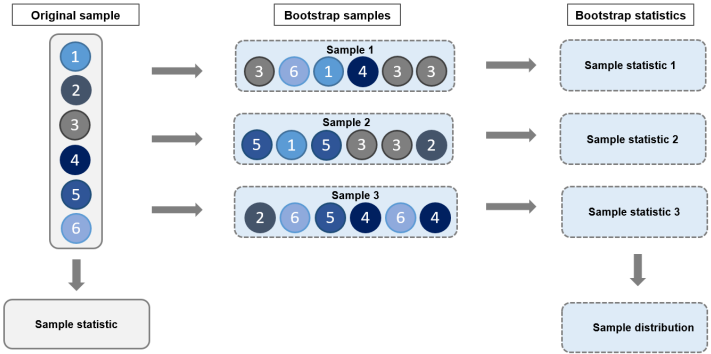
The bootstrap

- A bootstrap sample is always the same size the original
- it includes some of the original observations multiple times,
- it does not include some of other original observations.
- We typically create 500 - 10,000 samples
- Computationally intensive but feasible, relatively fast.



The bootstrap

- We have a dataset (the sample), can compute a statistic (e.g. mean)
- Create many bootstrap samples, and get a mean value for each sample
- Bootstrap estimate of SE = standard deviation of statistic based on bootstrap samples' estimates.



The bootstrap SE

- The bootstrap method creates many repeated samples that are different from each other, but each has the same size as the original dataset.
- Bootstrap gives a good approximation of the standard error, too.
- The bootstrap estimate (or the estimate from the bootstrap method) of the standard error is simply the standard deviation of the statistic across the bootstrap samples.

Stock market returns: The Bootstrap standard error

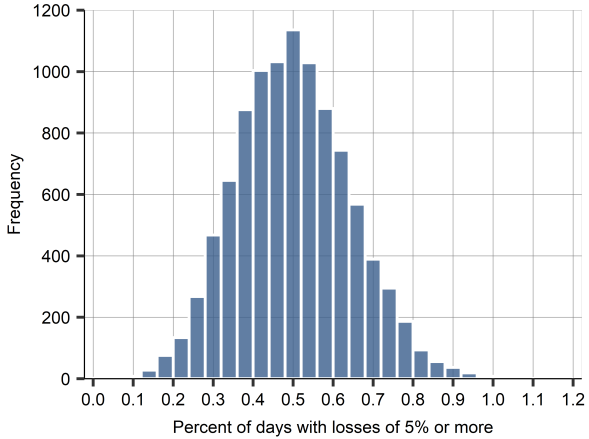
- We estimate the standard error by bootstrap.
- Let's consider our example of 11-years' of data on daily returns on the S&P 500 portfolio.
- Do the process ——————>
- End up with a new a dataset: one observations / bootstrap sample. Only variable is the estimated proportion in a sample
- The SE is simply the standard deviation of those estimated values in this new dataset.

The process

1. Take the original dataset and draw a bootstrap sample.
2. Calculate the proportions of days with 5%+ loss in that sample.
3. Save that value.
4. Then go back to the original dataset and take another bootstrap sample.
5. Calculate the proportion of days with 5%+ loss and save that value, too.
6. And so on, repeated many times.

Stock market returns: The Bootstrap standard error

- 10,000 bootstrap samples with 2,519 observations
- The proportion of days with 5+ percent loss.
- Varied 0.1 percent to 1.2 percent. Mean=Median= 0.5
- Standard deviation across the bootstrap samples = 0.14
- CI: the 95 percent CI is [0.22, 0.78].



Stock market returns: The Bootstrap standard error

- This means that in the general pattern represented by the 11-year history of returns in our data, we can be 95 percent confident that daily losses of more than 5 percent occur with a 0.22 to 0.78 percent chance.
- SE formula and bootstrap gave the same exact answer
- Under some conditions, this is what we expect
  - Large enough sample size
  - Observations independent
  - ... (other we overlook now)

External validity

External validity

- We discussed statistical inference: CI - uncertainty about the true value of the statistic in the population / general pattern that our data represents.
- What is the population, or general pattern, we care about?
- How close is our data to this?
- External validity is the concept that captures the similarity of our data to the population/general pattern we care about.
- High external validity: if our data is close to the population or the general pattern we care about.
- External validity is as important as statistical inference. However, it is not a statistical question.

External validity

- The most important challenges to external validity may be collected in three groups:

- Time: we have data on the past, but we care about the future
- Space: our data is on one country, but interested how a pattern would hold elsewhere in the world
- Sub-groups: our data is on 25-30 year old people. Would a pattern hold on younger / older people?

External validity in Big Data

- Big data: very large N
- Statistical inference not really important - CI becomes very narrow
- External validity remains as important

- 1.) Large sample DOES NOT mean representative sample
- 2.) Big data as result of actions - nature of things may change as people alter behavior, outside conditions change

Hypothesis

Generalization - Summary

- Generalization is a key task - finding beyond the actual dataset.
- This process is made up of discussing statistical inference and external validity.
- Statistical inference generalizes from our dataset to the population using a variety of statistical tools.
- External validity is the concept of discussing beyond the population for a general pattern we care about; an important but typically somewhat speculative process.

Testing Hypotheses: 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:

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

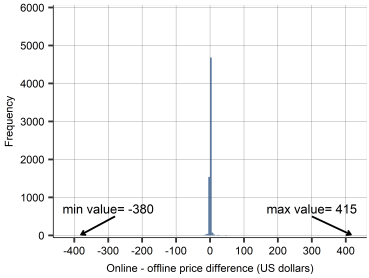
The statistic we are interested in with n observations is:

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

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 t-test

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

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

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

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

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

$$t = \frac{\bar{x}_A - \bar{x}_B}{SE(\bar{x}_A - \bar{x}_B)} \tag{4}$$

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.

Making a decision

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?

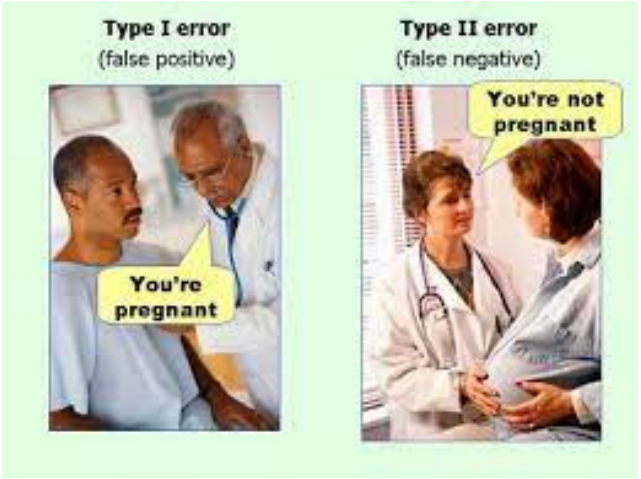
- Null rejected if the test statistic is larger than the critical value.
- Critical value governs the trade-off between being too strict or too lenient with our decision.

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

Being right or wrong



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.

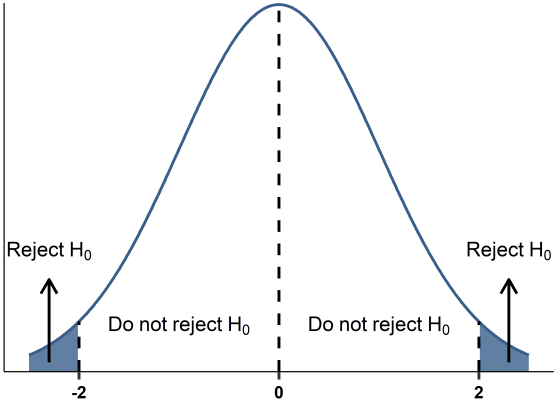
Rule of thumb when making a decision

- 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

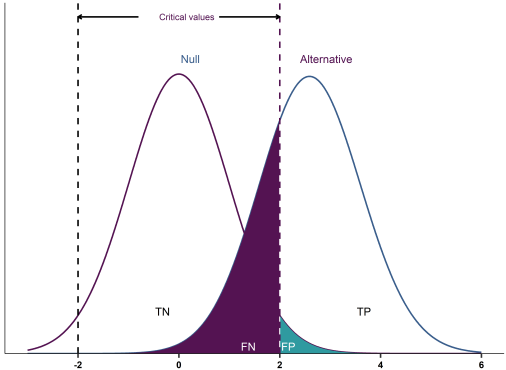


Critical values and generalization

- 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 (FP) affects the chance of false negatives (FN) 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.



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.

Multiple test

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