Biostatistics and Informatics

Lecture 6:
Introduction to Hypothesis Testing
17th October 2018

Gergely AGÓCS

Sources: – Herényi L (2016): Statisztika és Informatika (in Hungarian): Chapters 15.0–15.5
– Reiczigel J, Harnos A, Solymosi N (2014): Biostatisztika nem statisztikusoknak (in Hungarian): Chapter 6
– WolframMathWorld: Probability and Statistics: http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html
– Stanford Online Lagunita: Satistics in Medicine

Goals of this Lecture

- Understanding the aim and process of **scientific decision making**
  - philosophical background: **the good scientific question**
  - **null hypothesis** (H_0) and **alternative hypothesis** (H_1)
  - **what are we supposed to prove?**
- Steps of hypothesis testing throughout **an example**
- Significance level and ***p*-value**
  - hypothesis testing vs. confidence intervals
  - what influences the p-value?
- Decision vs. Reality: Errors and their probability
- P-value interpretation
  - clinical relevance vs. statistical significance
  - multiple testing
  - H_0 not rejected ≠ H_0 proven
  - correlation ≠ causation
  - do not compare p-values
  - should p-values be used at all?

Philosophical Background

The Development of Science

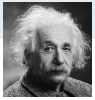
A scientific statement is one that can be independently reproduced.
But this does not tell us how science is created!

Francis **Bacon**
1561–1626

Inductivism:
(1) Take observations of nature.
(2) Create theory based on them to generalize a proposed pattern.
(3) Take more observations to...
(4) ...show the theory is (probably) true – or readjust it.
(5) Repeat steps (3) and (4).
This scientific method is based on **verification**.

How science is actually created:
(1) We have a problem.
(2) We *guess* a solution (theory) to explain it.
(3) Then we criticize the theory (by observations or by searching for inconsistencies)
This scientific method is based on **falsification**.

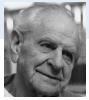
Philosophical Background



Albert **Einstein**
1879–1955

The Development of Science

"No amount of experimentation can ever prove me right;
a single experiment can prove me wrong."
after Albert Einstein: *Induction and Deduction*



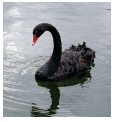
Karl **Popper**
1902–1994

"In so far as **a scientific statement** speaks about reality, it **must be falsifiable**;
and in so far as it is not falsifiable, it does not speak about reality."

"A theory which is not refutable by any conceptual event is non-scientific.
Irrefutability is not a virtue of a theory (as people often think) but a vice.
Every genuine **test of a theory is an attempt to falsify it**, or refute it."

"no matter how many instances of white swans we may have observed,
this does not justify the conclusion that all swans are white."

"**Induction is logically invalid**;
but refutation or falsification is
a logically valid way of arguing."
Karl Popper: *The Logic of Scientific Discovery*



Philosophical Background

Falsifiable (i.e. Scientific) Statements

"All swans are white."
(Karl Popper)
"The Earth is at the Center of the Universe."
(*Eppur si muove!*)
"Nothing can surpass the speed of light."
(Albert Einstein)
"The best teachers are usually those who are free, competent and willing to make original researches in the library and the laboratory."
(Daniel Coit Gilman; the Humboldtian model of higher education)

Non-Falsifiable (i.e. Non-Scientific) Statements
(may eventually be verifiable)

"There is a teapot orbiting the Sun somewhere between the Earth and the Mars ."
(Russell's teapot)
"There is a monster living in *Loch Ness*."

"A fire-breathing dragon lives in my garage."
(Carl Sagan)
"An extraterrestrial spacecraft crash-landed at a ranch near *Roswell*, *New Mexico*."

Philosophical Background

The Burden of Proof (*onus probandi*)

Onus probandi incumbit ei qui dicit, non ei qui negat: It is the obligation of someone coming up with a new idea to provide evidence to support it and than the scientific community will decide if that evidence is sufficient. If it is not, then the claim is dismissed and the opponents need not argue further in order to dismiss it

Quod gratis asseritur, gratis negatur: What can be asserted without evidence can be dismissed without evidence.

Way of Thinking in Hypothesis Testing

Indirect proof (*reductio ad absurdum*)

We have a box containing 100 marbles. Each of them are either red or white.
We want to figure out how many are red and how many are white.

Case #1:
Our hypothesis (*H*): all of them are white.
Experiment: We randomly take a marble out of the box.
Our observation: It is red.
Conclusion: The probability of our observation *given* our hypothesis is 0: Our hypothesis is for 100% sure wrong.

Impossible event

Case #2:
Our hypothesis (*H*): 99 are white and one is red.
Experiment: We randomly take a marble out of the box and put it back; we do this 5 times.
Our observation: All of them are red.
Conclusion: Our hypothesis is for *almost* 100% sure wrong: The probability of our observation *given* our hypothesis is $0.01^5 = 10^{-10}$: practically impossible.

Case #3:
Our hypothesis (*H*): 50 are white and 50 are red.
Experiment: We randomly take a marble out of the box and put it back; we do this 5 times.
Our observation: All of them are red.
Conclusion: Now we are not sure what to do: The probability of our observation *given* our hypothesis is $0.5^5 = 0.03125$: low but not that unlikely...

Case #4:
Our hypothesis (*H*): all of them are red.
Experiment: We randomly take a marble out of the box and put it back; we do this 5 times.
Our observation: All of them are red.
Conclusion: The probability of our observation *given* our hypothesis is $1^5 = 1$. Are we sure what to do now?

Sure event

Way of Thinking in Hypothesis Testing

Indirect proof (*reductio ad absurdum*)

We have a box containing 100 marbles. Each of them are either red or white.
We want to figure out how many are red and how many are white.

Case #1:
Our hypothesis (*H*): all of them are white.
Experiment: We randomly take a marble out of the box.
Our observation: It is red.
Conclusion: The probability of our observation *given* our hypothesis is 0: Our hypothesis is for 100% sure wrong.

Impossible event

✓

The hypothesis is false

Falsification works

Verification does not work

The hypothesis is true

Case #4:
Our hypothesis (*H*): all of them are red.
Experiment: We randomly take a marble out of the box and put it back; we do this 5 times.
Our observation: All of them are red.
Conclusion: The probability of our observation *given* our hypothesis is $1^5 = 1$. Are we sure what to do now?

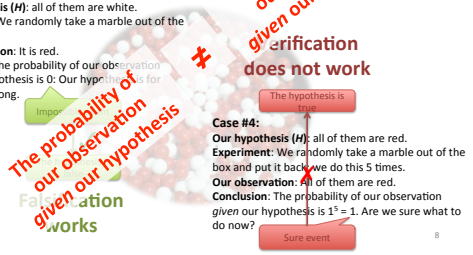
Sure event

Way of Thinking in Hypothesis Testing

Indirect proof (*reductio ad absurdum*)

We have a box containing 100 marbles. Each of them is either red or white.
We want to figure out how many are red and how many are white.

Case #1:
Our hypothesis (*H*): all of them are white.
Experiment: We randomly take a marble out of the box.
Our observation: It is red.
Conclusion: The probability of our observation *given* our hypothesis is 0: Our hypothesis is for 100% sure wrong.



Case #4:
Our hypothesis (*H*): all of them are red.
Experiment: We randomly take a marble out of the box and put it back. We do this 5 times.
Our observation: All of them are red.
Conclusion: The probability of our observation *given* our hypothesis is $1^5 = 1$. Are we sure what to do now?

Way of Thinking in Hypothesis Testing

Indirect proof (*reductio ad absurdum*)

Mathematical Logic:
We have a hypothesis (*H*).
If *H* is true, *E* event cannot occur.
E occurs.
So *H* is not true.
As we saw it before, **a hypothesis can only be rejected**.

Statistical Logic:
We have a hypothesis (*H*).
If *H* is true, *F* event is very unlikely to occur.
F occurs.
So we reject *H*. But we are not 100% sure if *H* is not true.
In this case **a hypothesis cannot even be rejected with 100% certainty**.

What Kind of Questions Can We Test?

The question...

Y/N **...must be a yes/no (a.k.a. dichotomous or polar) question.** **5W1H**

- Is the 5-years survival rate (i.e. probability) for myeloma 50%?
- Does the total blood cholesterol level of Cushing's syndrome patients differ from the general 200 mg/dL population mean? ✓

- What is the 5-years survival rate for myeloma?
- What is the expected value of total cholesterol level in Cushing's syndrome patients? ✗

...must refer to a set of observations, not to individual cases.
(And the question is aimed at a population, not a sample.)

- Is the 5-years survival rate for myeloma 50%? ✓

- Will this myeloma patient survive for 5 years? ✗

...must have at least one unambiguous answer.

- Is the 5-years survival rate for myeloma 50%? ✓

- Is the 5-years survival rate for myeloma less than 50%? ✗

What Kind of Answers Can We Test?

We have two answers for our question:

The null hypothesis (H_0)

- **Unambiguous**: can be realized in only one way. It contains some form of =.
  The 5-years survival rate for myeloma is 50%.
- Represents the current well-established, generally **accepted scientific knowledge**,
  The total blood cholesterol level of Cushing's syndrome patients is same as the population mean
  or something that is the **most trivial** with the least assumptions (Occam's razor).
  The probability of landing on heads in a coin tossing experiment is 50%.
- It is **not** necessarily the negative answer to the question.

The alternative hypothesis (H_1)

- Typically can be realized in more than one way.
  The 5-years survival rate for myeloma is not 50%.
  (can be a little more, a lot less etc.)
- Represent a **new statement** challenging the current scientific consensus,
  The total blood cholesterol level of Cushing's syndrome patients differs from the population mean
  or a set of all the **not-so-trivial** answers needing more or special assumptions.
  The probability of landing on heads in a coin tossing experiment is other than 50%.
- It is typically **complementary to H_0** (i.e., its negation).
  H_1 = not H_0

A Worked-out Example

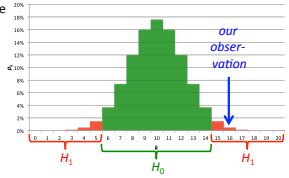
The current 5-years survival rate for myeloma patients is 50% (average between 2008–2012).

We have a new drug candidate that seems to be effective against myeloma in animal experiments. We want to test it on humans.

1.a Physicians question: Should we replace current treatment protocols with the new drug?
1.b Clinical question: Is the effect clinically relevant? Is the change in survival rate big enough?
1.c Statistician's question: Is there an effect?
2. H_0**: The drug has no effect:** Survival rate with the drug is same as with the conventional therapy.
3. H_1**: The drug has an effect:** Survival rate is different.
4. Test design: Select randomly 20 myeloma patients and treat them with the drug candidate. After 5 years, check the *number of patients still alive*. This number can be called here the **test statistic**.
5. Generate the H_0 distribution: It is a binomial distribution with $p = 0.5$ and $n = 20$. Same as a coin tossing experiment.
6.a Set up a significance level (α) which will also define your confidence level ($1-\alpha$): Be $\alpha = 5\%$. (For historical reasons 5% is often used in health sciences but it is our choice.)
6.b Define what change is clinically relevant: We expect at least a 20% higher survival rate.

A Worked-out Example

7. Determine your confidence interval using the H_0 distribution and α: In our binomial distribution the range of outcomes from 6 to 14 have a combined probability of just above 95%. This is the set of those outcomes, which represent too weak evidence against H_0.
8. Carry out the experiment:
(Note: *this is the 8th step!*)
Out of our 20 patients treated with the new drug 16 are still alive after 5 years.
...
10. Make a decision: Clinically speaking the effect is **relevant** (80% instead of 50% survival rate); Statistically speaking it is **significant** (our outcome is unlikely under the H_0 or, more precisely, it belongs to the set of 5% least probable outcomes).
11. Answer the question: Let's wait a little bit with this one...



A Worked-out Example

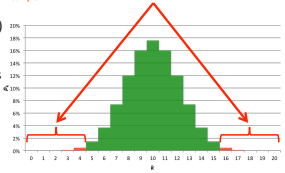
How significant is the outcome?

Obviously, a 16 out of 20 survival rate is more significant (= unlikely under H_0) than 14 out of 20 but less significant than 18 out of 20. How can we express this numerically?

9. Determine the sample *p*-value: The *p*-value is the Holy Grail of inferential statistics. It gives the **probability of your (or any less likely) observation** to occur ***given*** H_0 is true. In our case, it is the combined probability of all outcomes equally or less probable than getting 16 out of 20 survivals. The value is $p_{sample} = 0.0118$. This is shown in red in the graph.

11. Answer the question: "The group ($n = 20$) treated with the drug candidate had 80% survival rate that is significantly higher than that of the conventional treatment's 50% ($p = 1{,}18\%$)"

So we are sort of done. But again, can we be sure about the correctness of our decision?



Decision Errors

As you might have observed, the decision making procedure in **inferential statistics is** somewhat **similar to** the decision making process at a **court of justice**. Let's see:

- We have an accusation (H_1) confronting innocence (H_0)
- We have to assume innocence (assumption of ineffectiveness)
- Burden of Proof (*onus probandi*) lies on the plaintiff (who accuses)
- We collect evidence against H_0 (sampling).
- Based on the probability of the evidence given H_0 is true the defendant is either acquitted (H_0 not rejected) or convicted (H_0 rejected).
- Errors are of course possible in this decision making:

		Our Decision	
		H_0 not rejected	H_0 rejected
The Truth (never known)	H_0 true	**correct** $1-\alpha$	**α or type I error** $p(H_0 \text{ rejected} \mid H_0 \text{ true}) \le \alpha$
	H_0 false	**β or type II error** $p(H_0 \text{ not rejected} \mid H_0 \text{ false}) \le \beta$	**correct** $1-\beta$

Decision Errors

If H_0 is true...

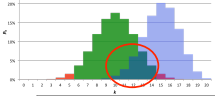
- **We can maximize the error:** We set up the margin of error ourselves. H_0 is only rejected if the sample *p*-value (p_{sample}) is less than the preset significance level (α a.k.a. p_{crit}). The significance level is our choice: we predefine the probability of decision error given H_0.
- A lower α will decrease the chance of type I error, it makes the test procedure more conservative. α is also called the **size** of the test.
- Besides the preset α level, nothing has effect on the probability of incorrectly rejecting the null (i.e. alpha error).

		Our Decision	
		H_0 not rejected	H_0 rejected
The Truth (never known)	H_0 true	**correct** $1 - \alpha$	**α** or **type I error** $p(H_0 \text{ rejected} \mid H_0 \text{ true}) \leq \alpha$
	H_0 false		

Decision Errors

If H_0 is false...

- **The probability of the error depends on the truth**, i.e. how much the truth differs from our expectations.
- Suppose in our example the *real* survival rate for the new drug was 75%, it is possible to calculate (using =BINOM.DIST() in Excel) the chances of getting an outcome between 6 and 14 (the H_0 region): =BINOM.DIST(14,20,75%,1)-BINOM.DIST(5,20,75%,1)
  This yields 38.28%.



		Our Decision	
		H_0 not rejected	H_0 rejected
The Truth (never known)	H_0 true		
	H_0 false	**β** or **type II error** $p(H_0 \text{ not rejected} \mid H_0 \text{ false}) \leq \beta$	**correct** $1 - \beta$

Decision Errors

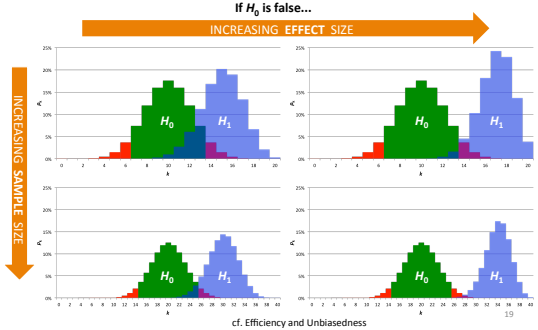
If H_0 is false...

- **Usually the beta error is not known**, but we can set up a "minimum" alternative hypothesis that reflects our expected minimum effect size (clinical relevance)
- If the **real effect is greater**, there is less overlap with the H_0 distribution, so **less chance of failing to reject the H_0 (beta error)**.
- If the **sample size is greater**, the error of the statistic will be lower, which again **decreases the chance of beta error**.
- the probability that a test finds a real effect is 1 – β and called the **power of the test**.

		Our Decision	
		H_0 not rejected	H_0 rejected
The Truth (never known)	H_0 true		
	H_0 false	**β** or **type II error** $p(H_0 \text{ not rejected} \mid H_0 \text{ false}) \leq \beta$	**correct** $1 - \beta$

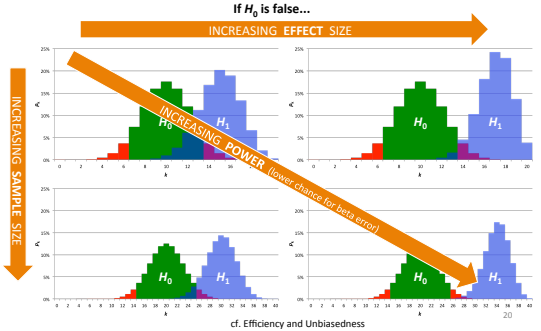
Decision Errors

If H_0 is false...



cf. Efficiency and Unbiasedness

Decision Errors



cf. Efficiency and Unbiasedness

p-value Pitfalls

- **clinical relevance vs. statistical significance:** something that's statistically significant is not necessarily clinically relevant and vice versa
- **multiple testing**: if we carry out multiple tests on a sample the chances of committing type one error increases; say alpha is 5%, that means that out of 20 tests where H_0 is true we will find approx. one significant outcome. This is a huge problem in science because people don't (cannot) publish insignificant results so usually no one really knows how many tests were carried out only how many were significant.
- **H_0 not rejected ≠ H_0 proven**: see the red-and-white-marbles example; we cannot verify only falsify.
- **correlation ≠ causation**: if two variables appear to influence each other somehow that does not automatically mean that there is a couse–effect relationship between them; see also: https://www.fastcompany.com/3030529/hilarious-graphs-prove-that-correlation-isnt-causation
- **Should p-values be used at all?** There is an ongoing debate whether p-values should be banished from scientific literature altogether. Main reason is frequent abuse, misunderstanding and misinterpretation. But other measures are not easier to grab either. The problem is that scientific reasoning itself is complex, not just the math associated to it.

Summary

- Only falsifiable statements can be considered scientific.
- The Burden of Proof lies on the one who makes the claim.
- The H_0 hypothesis should reflect the current scientific knowledge.
- The H_1 is our claim that contradicts current scientific consensus.
- The *p*-value gives the probability of our observation (or more extreme) given the H_0.
- A low *p*-value is an evidence against the H_0 but a high *p*-value is not an evidence in support of H_0
- If the *p*-value is high we won't "accept" H_0 but we rather fail to reject it. This is not just a play with words: this is to emphasize that the "validity" of the H_0 is supported by previous knowledge or other assumptions (remember the rules for formulating H_0) rather than the *p*-value calculated from our sample in the actual study.
- Decesion making comes with chances of error. But this error is measurable and controllable to some extent.

Follow-up Questions

– Why is a low *p*-value a proof against the H_0, but a high *p*-value is not a proof for the H_0?
– What are the steps of inductive thinking? Give an example as well.
– Why is induction logically invalid? Explain and give examples.
– Explain why scientific statements must be falsifiable?
– A scientific theory cannot be falsified. Is it a great scientific achievement or junk? Why?
– I think I have discovered a new drug. I start to advertise it heavily, but soon some sceptics put it in question if there is any effect at all. I want them to prove me wrong, they want that I prove I am right. How actually should prove what?
– Give example for falsifiable statements.
– Give example for non-falsifiable statements.
– What should be proven? That homeopathy works or that it has no effect?
– Iridology claims it can make diagnosis for non-ophtalmic conditions using only the iris of the eye. What should the null hypothesis be: that it this statement is true or that it is wrong?
– What is indirect proof? Give example.
– What is the similarity and the difference between mathematical and statistical logic?
– What are the properties of a good statistical question? Give some examples, too.
– What are the properties of a null hypothesis?
– What are the properties of an alternative hypothesis?
– Give example for statistical question, null, and alternative hypotheses.

Follow-up Questions

– Give the steps of null hypothesis testing.
– Why can we test the null but not the alternative hypothesis?
– What is the relationship between hypothesis testing and confidence interval calculation?
– Define *p*-value.
– What does the *p*-value depend on?
– What is a test statistic?
– The *p*-value calculated from a sample is less than the significance level. What is our decision?
– What is type one and type two error?
– What does the probability of type I and type II error depend on?
– Give some pitfalls of *p*-value.
– Why is the whole hypothesis testing procedure necessary instead of just presenting the results of our experiments?
– What is the meaning of Occam's razor?
– Why is 5% the usual significance level in medicin? Does something special happen at this level?
– What is the meaning of the size of a statistical test?
– What is the meaning of the power of a statistical test?