Hypothesis testing

Vladimír Janiš

Department of Mathematics
Matej Bel University Banská Bystrica
Slovakia

December 13, 2011, Novi Sad
TEMPUS - Master programme in applied statistics

Short history

Short history

John Arbuthnot, 1710 - the first published statistical test

Short history

John Arbuthnot, 1710 - the first published statistical test

- observed, that the fraction of boys born is slightly larger than the fraction of girls

John Arbuthnot, 1710 - the first published statistical test

- observed, that the fraction of boys born is slightly larger than the fraction of girls
- calculated that assuming equal probabilities for boys and girls this emipirical fact would be **exceedingly unlikely** - probability $\frac{1}{4836000000000000000000000}$

John Arbuthnot, 1710 - the first published statistical test

- observed, that the fraction of boys born is slightly larger than the fraction of girls
- calculated that assuming equal probabilities for boys and girls this emipirical fact would be **exceedingly unlikely** - probability $\frac{1}{4836000000000000000000000}$
- argued that this was a proof of God's will - boys had higher risks of an early death

Short history

John Arbuthnot, 1710 - the first published statistical test

- observed, that the fraction of boys born is slightly larger than the fraction of girls
- calculated that assuming equal probabilities for boys and girls this emiprical fact would be **exceedingly unlikely** - probability $\frac{1}{4836000000000000000000000}$
- argued that this was a proof of God's will - boys had higher risks of an early death
- clearly a consideration possessing the basic characterizations of a hypothesis test

Tests in a modern sense

Tests in a modern sense

1900 - Karl Pearson

Tests in a modern sense

1900 - Karl Pearson

- chi-square test

Tests in a modern sense

1900 - Karl Pearson

- chi-square test
- compared observed frequency distribution to a theoretically assumed one

Tests in a modern sense

1900 - Karl Pearson

- chi-square test
- compared observed frequency distribution to a theoretically assumed one

1925 - R. A. Fisher

Tests in a modern sense

1900 - Karl Pearson

- chi-square test
- compared observed frequency distribution to a theoretically assumed one

1925 - R. A. Fisher

- data are regarded as the outcome of a random variable X

Tests in a modern sense

1900 - Karl Pearson

- chi-square test
- compared observed frequency distribution to a theoretically assumed one

1925 - R. A. Fisher

- data are regarded as the outcome of a random variable X
- X has a preassumed probability distribution

Tests in a modern sense

1900 - Karl Pearson

- chi-square test
- compared observed frequency distribution to a theoretically assumed one

1925 - R. A. Fisher

- data are regarded as the outcome of a random variable X
- X has a preassumed probability distribution
- null hypothesis - an assertion defining a subset of this family

Tests in a modern sense

1900 - Karl Pearson

- chi-square test
- compared observed frequency distribution to a theoretically assumed one

1925 - R. A. Fisher

- data are regarded as the outcome of a random variable X
- X has a preassumed probability distribution
- null hypothesis - an assertion defining a subset of this family
- test statistics $T = t(X)$ indicates the degree to which the data deviate from the null hypothesis

Tests in a modern sense

1900 - Karl Pearson

- chi-square test
- compared observed frequency distribution to a theoretically assumed one

1925 - R. A. Fisher

- data are regarded as the outcome of a random variable X
- X has a preassumed probability distribution
- null hypothesis - an assertion defining a subset of this family
- test statistics $T = t(X)$ indicates the degree to which the data deviate from the null hypothesis
- significance probability (p-value) 0.05

Tests in a modern sense

1900 - Karl Pearson

- chi-square test
- compared observed frequency distribution to a theoretically assumed one

1925 - R. A. Fisher

- data are regarded as the outcome of a random variable X
- X has a preassumed probability distribution
- null hypothesis - an assertion defining a subset of this family
- test statistics $T = t(X)$ indicates the degree to which the data deviate from the null hypothesis
- significance probability (p-value) 0.05
- Fisher was the first to recognise the arbitrary nature of this treshold

A competing approach to Fischer

A competing approach to Fischer

1928 - J. Neyman and Egon Pearson (the son of Karl Pearson)

A competing approach to Fischer

1928 - J. Neyman and Egon Pearson (the son of Karl Pearson)

- criticized the arbitrariness in Fisher's choice of of the test statistic

A competing approach to Fischer

1928 - J. Neyman and Egon Pearson (the son of Karl Pearson)

- criticized the arbitrariness in Fisher's choice of of the test statistic
- claimed that for rational choice of a test statistic not only null hypothesis, but also an alternative one is needed

A competing approach to Fischer

1928 - J. Neyman and Egon Pearson (the son of Karl Pearson)

- criticized the arbitrariness in Fisher's choice of of the test statistic
- claimed that for rational choice of a test statistic not only null hypothesis, but also an alternative one is needed
- formalized the testing problem as two decision problem

A competing approach to Fischer

1928 - J. Neyman and Egon Pearson (the son of Karl Pearson)

- criticized the arbitrariness in Fisher's choice of of the test statistic
- claimed that for rational choice of a test statistic not only null hypothesis, but also an alternative one is needed
- formalized the testing problem as two decision problem
- $H_0, H_1(H_a)$

A competing approach to Fischer

1928 - J. Neyman and Egon Pearson (the son of Karl Pearson)

- criticized the arbitrariness in Fisher's choice of of the test statistic
- claimed that for rational choice of a test statistic not only null hypothesis, but also an alternative one is needed
- formalized the testing problem as two decision problem
- $H_0, H_1(H_a)$
- decisions **reject** H_0 or **do not reject** H_0

A competing approach to Fischer

1928 - J. Neyman and Egon Pearson (the son of Karl Pearson)

- criticized the arbitrariness in Fisher's choice of of the test statistic
- claimed that for rational choice of a test statistic not only null hypothesis, but also an alternative one is needed
- formalized the testing problem as two decision problem
- $H_0, H_1(H_a)$
- decisions **reject** H_0 or **do not reject** H_0
- two types of resulting errors - rejecting a true H_0 (the first kind) and failing to reject a false H_0 (the second kind)

A competing approach to Fischer

1928 - J. Neyman and Egon Pearson (the son of Karl Pearson)

- criticized the arbitrariness in Fisher's choice of of the test statistic
- claimed that for rational choice of a test statistic not only null hypothesis, but also an alternative one is needed
- formalized the testing problem as two decision problem
- $H_0, H_1(H_a)$
- decisions **reject** H_0 or **do not reject** H_0
- two types of resulting errors - rejecting a true H_0 (the first kind) and failing to reject a false H_0 (the second kind)
- power of the test - probability of (correctly) rejecting H_0, if H_1 is true

Fisher formulation vs. Neyman-Pearson formulation

Fisher formulation vs. Neyman-Pearson formulation

Neyman-Pearson

- richer results at the cost of a more demanding model

Fisher formulation vs. Neyman-Pearson formulation

Neyman-Pearson

- richer results at the cost of a more demanding model
- we need to specify an alternative hypothesis

Neyman-Pearson

- richer results at the cost of a more demanding model
- we need to specify an alternative hypothesis
- testing problem as a two-decision situation

Fisher formulation vs. Neyman-Pearson formulation

Neyman-Pearson

- richer results at the cost of a more demanding model
- we need to specify an alternative hypothesis
- testing problem as a two-decision situation

Different philosophical positions summarised in

- Hacking, I.: Logic of Statistical Inference, Cambridge Univ. Press, Cambridge, 1965
- Gigerenzer, G., Swijtnik, Z., Porter, T., Daston, L., Beatty, J., Kruger, L.: The Empire of Chance, Cambridge Univ. Press, New York, 1989 - also dicsusses aspects in teaching and practise of statistics by a hybrid theory combining elements of both approaches

Problems in the interpretation and use of tests

Problems in the interpretation and use of tests

Hypothesis tests

- routinely applied

Problems in the interpretation and use of tests

Hypothesis tests

- routinely applied
- but there are continuing debates about their use from the philosophical point of view

Problems in the interpretation and use of tests

Hypothesis tests

- routinely applied
- but there are continuing debates about their use from the philosophical point of view
- Harlow, L.L., Mulaik, S.A., Steiger, J.H. (eds.): What if there were no Significance Tests?, Erlbaum, Mahwah, NJ
- Wilkinson, L.: Task force on statistical inference; Statistical methods in psychology journals, American Psychologist 54: 594-604, 1999.
- Nickerson, R.S.: Null hypothesis significance testing: A review of an old and continuing controversy, Psychological Methods 5:241-301, 2000.

Reasons for criticism

Reasons for criticism

- difficulty of reasoning with uncertain evidence

Reasons for criticism

- difficulty of reasoning with uncertain evidence
- natural human preference of strict rules and sharp decisions

Reasons for criticism

- difficulty of reasoning with uncertain evidence
- natural human preference of strict rules and sharp decisions

The empirical researcher would like to conclude whether a theory is true or false.

Reasons for criticism

- difficulty of reasoning with uncertain evidence
- natural human preference of strict rules and sharp decisions

The empirical researcher would like to conclude whether a theory is true or false. However, experimental and observational data are often so variable that the evidence produced by them is uncertain.

A serious and frequent misuse of hypothesis testing

$\alpha < p$ implies that the investigateg effect is absent, while $\alpha \geq p$ proves that the effect exists.

Reasons for criticism

- difficulty of reasoning with uncertain evidence
- natural human preference of strict rules and sharp decisions

The empirical researcher would like to conclude whether a theory is true or false. However, experimental and observational data are often so variable that the evidence produced by them is uncertain.

A serious and frequent misuse of hypothesis testing

$\alpha < p$ implies that the investigateg effect is absent, while $\alpha \geq p$ proves that the effect exists.

- better interpretation of hypothesis testing should be promoted

Reasons for criticism

- difficulty of reasoning with uncertain evidence
- natural human preference of strict rules and sharp decisions

The empirical researcher would like to conclude whether a theory is true or false. However, experimental and observational data are often so variable that the evidence produced by them is uncertain.

A serious and frequent misuse of hypothesis testing

$\alpha < p$ implies that the investigateg effect is absent, while $\alpha \geq p$ proves that the effect exists.

- better interpretation of hypothesis testing should be promoted
- tests should be used less mechanically and combined with other argumentations and other statistical procedures

Reasons for criticism

- difficulty of reasoning with uncertain evidence
- natural human preference of strict rules and sharp decisions

The empirical researcher would like to conclude whether a theory is true or false. However, experimental and observational data are often so variable that the evidence produced by them is uncertain.

A serious and frequent misuse of hypothesis testing

$\alpha < p$ implies that the investigateg effect is absent, while $\alpha \geq p$ proves that the effect exists.

- better interpretation of hypothesis testing should be promoted
- tests should be used less mechanically and combined with other argumentations and other statistical procedures
- user has to be aware that the random variability in the data cannot be filtered out of the results

Misinterpretations of hypothesis tests - 1

Nonrejection implies support for the null hypothesis.

Nonrejection implies support for the null hypothesis.

- nonrejection: there is not enough evidence against the null hypothesis

Misinterpretations of hypothesis tests - 1

Nonrejection implies support for the null hypothesis.

- nonrejection: there is not enough evidence against the null hypothesis
- the sample size may be small, error variability may be large, so that the data do not contain much information

Misinterpretations of hypothesis tests - 1

Nonrejection implies support for the null hypothesis.

- nonrejection: there is not enough evidence against the null hypothesis
- the sample size may be small, error variability may be large, so that the data do not contain much information
- therefore nonrejection often provides support for alternative hypothesis practically as strongly as for the null hypothesis itself

Misinterpretations of hypothesis tests - 1

Nonrejection implies support for the null hypothesis.

- nonrejection: there is not enough evidence against the null hypothesis
- the sample size may be small, error variability may be large, so that the data do not contain much information
- therefore nonrejection often provides support for alternative hypothesis practically as strongly as for the null hypothesis itself
- nonrejection may not be interpreted as a support for the null hypothesis

Misinterpretations of hypothesis tests - 2

A nonsignificant result supports null hypothesis more in cases of a high test power.

A nonsignificant result supports null hypothesis more in cases of a high test power.

- statistical power is the probability to reject the null hypothesis **if a given effect is present**

Misinterpretations of hypothesis tests - 2

A nonsignificant result supports null hypothesis more in cases of a high test power.

- statistical power is the probability to reject the null hypothesis **if a given effect is present**
- this cannot be inverted as a support of null hypothesis in the case of non-significance

Misinterpretations of hypothesis tests - 2

A nonsignificant result supports null hypothesis more in cases of a high test power.

- statistical power is the probability to reject the null hypothesis **if a given effect is present**
- this cannot be inverted as a support of null hypothesis in the case of non-significance
- power studies are important while planning an experiment

Misinterpretations of hypothesis tests - 2

A nonsignificant result supports null hypothesis more in cases of a high test power.

- statistical power is the probability to reject the null hypothesis **if a given effect is present**
- this cannot be inverted as a support of null hypothesis in the case of non-significance
- power studies are important while planning an experiment
- appropriate procedures **after** the experiment are confidence intervals

Misinterpretations of hypothesis tests - 2

A nonsignificant result supports null hypothesis more in cases of a high test power.

- statistical power is the probability to reject the null hypothesis **if a given effect is present**
- this cannot be inverted as a support of null hypothesis in the case of non-significance
- power studies are important while planning an experiment
- appropriate procedures **after** the experiment are confidence intervals

Test results tell us about the probabilities of null and alternative hypotheses.

Misinterpretations of hypothesis tests - 3

Misinterpretations of hypothesis tests - 3

Rejecting the null hypothesis the alternative theory is confirmed.

Misinterpretations of hypothesis tests - 3

Rejecting the null hypothesis the alternative theory is confirmed.

- the alternative hypothesis is not the same as a scientific theory

Rejecting the null hypothesis the alternative theory is confirmed.

- the alternative hypothesis is not the same as a scientific theory
- alternative hypotheses are deduced from the theory, but alway under asssumption that the study was well designed and usually also other assumptions (normality of distributions)

Misinterpretations of hypothesis tests - 3

Rejecting the null hypothesis the alternative theory is confirmed.

- the alternative hypothesis is not the same as a scientific theory
- alternative hypotheses are deduced from the theory, but alway under asssumption that the study was well designed and usually also other assumptions (normality of distributions)
- alternative hypotheses are consequences of the theory, not its sufficient conditions

Misinterpretations of hypothesis tests - 3

Rejecting the null hypothesis the alternative theory is confirmed.

- the alternative hypothesis is not the same as a scientific theory
- alternative hypotheses are deduced from the theory, but alway under asssumption that the study was well designed and usually also other assumptions (normality of distributions)
- alternative hypotheses are consequences of the theory, not its sufficient conditions
- n the other hand, it is possible that the theory is true, but the alternative hypothesis deduced from it is not true, e.g. because of the wrong experimental design

Misinterpretations of hypothesis tests - 3

Rejecting the null hypothesis the alternative theory is confirmed.

- the alternative hypothesis is not the same as a scientific theory
- alternative hypotheses are deduced from the theory, but alway under asssumption that the study was well designed and usually also other assumptions (normality of distributions)
- alternative hypotheses are consequences of the theory, not its sufficient conditions
- n the other hand, it is possible that the theory is true, but the alternative hypothesis deduced from it is not true, e.g. because of the wrong experimental design

... The more you reject the null hypothesis, the more likely it is that you'll get {a title, a permanent position, ...}

4 citations

4 citations

- *Overemphasis on tests of significance at the expense especially of interval estimation has long been condemned (Cox 1977)*

4 citations

- *Overemphasis on tests of significance at the expense especially of interval estimation has long been condemned (Cox 1977)*
- *The continued very extensive use of significance tests is alarming (Cox 1986)*

4 citations

- *Overemphasis on tests of significance at the expense especially of interval estimation has long been condemned (Cox 1977)*
- *The continued very extensive use of significance tests is alarming (Cox 1986)*
- *The author believes that tests provide a poor model of most real problems, usually so poor that their objectivity is tangential and often too poor to be useful (Pratt 1976)*

4 citations

- *Overemphasis on tests of significance at the expense especially of interval estimation has long been condemned (Cox 1977)*
- *The continued very extensive use of significance tests is alarming (Cox 1986)*
- *The author believes that tests provide a poor model of most real problems, usually so poor that their objectivity is tangential and often too poor to be useful (Pratt 1976)*
- *We do not perform an experiment to find out if two varieties of wheat or two drugs are equal. We know in advance, without spending a dollar on an experiment, that they are not equal (Deming 1975)*

Consequence for our project

Consequence for our project

- Critical reading of applied studies

Consequence for our project

- Critical reading of applied studies
- A deep knowledge of a particulart application is welcome (a statistically significant result need not necessarily be a noteworthy result)

Consequence for our project

- Critical reading of applied studies
- A deep knowledge of a particulart application is welcome (a statistically significant result need not necessarily be a noteworthy result)
- Where is the optimal ratio between quantity of taught statistical methods and the depth of understanding their nature?