Statistical data analysis lab

Statistical significance 2

Statistical significance

Some hypothesis H

-what is a significant result

-what question would actually like to answer

Statistical significance

Some hypothesis H

-what is a significant result

The probability of such data (observed statistics), or something even more extreme-lookgin, is low, if hypothesis H were true.

-what question would actually like to answer

Is hypothesis H true

Statistical significance

Some hypothesis H

-what question would actually like to answer

Is hypothesis H true

-so what we get:

Here is some evidence against hypothesis H

The p-value is **NOT** a measure:

- of the strength of the effect

- of the probability of the hypothesis we are testing

The p-value is **NOT** a measure:

- of the strength of the effect
- of the probability of the hypothesis we are testing

The p-value is **NOT** a measure:

of the probability of the hypothesis we are testing

A hypothesis **cannot have probability.**

The p-value is **NOT** a measure:

of the probability of the hypothesis we are testing

A hypothesis **cannot have probability.**
A hypothesis is not a random variable, it is not data. It is a statement about distribution parameter.

It is either true (probability 1), or it isn't (probability 0).

The p-value is NOT the probability of H

A hypothesis **cannot have probability.**
A hypothesis is not a random variable, it is not data. It is a statement about distribution parameter.

However, I will consider the Bayes Formula:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Event A and B

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B \mid A) \cdot P(A) + P(B \mid \neg A) \cdot P(\neg A)}$$

Conditional probability and Bayes formua

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)}$$

where:

$P(A)$ = The probability of A occurring

$P(B)$ = The probability of B occurring

$P(A|B)$ = The probability of A given B

$P(B|A)$ = The probability of B given A

$P\left(A \bigcap B\right))$ = The probability of both A and B occurring

Conditional probability and Bayes formua

Ex.
13*4=52 cards
11 number
3 face cards

Probability that card is a queen?
Probability that card is a queen, given that we know the card is a face card?

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)}$$

where:

$P(A)$ = The probability of A occurring

$P(B)$ = The probability of B occurring

$P(A|B)$ = The probability of A given B

$P(B|A)$ = The probability of B given A

$P\left(A \bigcap B\right))$ = The probability of both A and B occurring

Conditional probability and Bayes formua

Ex.
A is probability that a person in a room is a woman
B is a probability that a person in a room has long hair

P(A)	0.5
P(B|A)	0.7
P(B| not A)	0.1

Find probability that a person is a woman, if you know the person has long hair

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)}$$

where:

$P(A)$ = The probability of A occurring

$P(B)$ = The probability of B occurring

$P(A|B)$ = The probability of A given B

$P(B|A)$ = The probability of B given A

$P\left(A \bigcap B\right))$ = The probability of both A and B occurring

Conditional probability and Bayes formua

Ex.
A is probability that a person in a room is a woman
B is a probability that a person in a room has long hair

P(A)	0.5
P(B|A)	0.7
P(B| not A)	0.1

Find probability that a person is a woman, if you know the person has long hair

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)}$$

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B \mid A) \cdot P(A) + P(B \mid \neg A) \cdot P(\neg A)}$$

Conditional probability and Bayes formua

Test screening - positive or negative test

sensitivity 92% (percentage of patients with positive test; true positive)
specificity of 94% (percentage of healthy indiviuals with negative test; true negative)

incidence of diseas in population: 0.089%

calculate: probability of person having disease, if thest is positive

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)}$$

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B \mid A) \cdot P(A) + P(B \mid \neg A) \cdot P(\neg A)}$$

answer

1.35% (still 15 times higher)

answer

1.35% (still 15 times higher)

Other numbers:

Incidence 0.351% - answer 5.12%
Incidence 15 % - answer 46.37%

The case of the p-value

$$p = \Pr(T \leq t \mid H_0)$$

* However, keep in mind that in classical probability

A hypothesis **cannot have probability.**
It is either true (probability 1), or it isn't (probability 0).

The case of the p-value

$$p = \Pr(T \leq t \mid H_0)$$

$$\Pr(\, H_0 \mid T \leq t\,)$$

* However, keep in mind that in classical probability

A hypothesis **cannot have probability.**
It is either true (probability 1), or it isn't (probability 0).

The case of the p-value

$$p = \Pr(T \leq t \mid H_0)$$

$$\Pr(H_0 \mid T \leq t)$$

$$P(H|E) = \frac{(P(H) * P(E|H))}{((P(H) * P(E|H)) + ((1 - P(H)) * P(E|\neg H)))}$$

p-value

T ≤ t replaced by E "evidence"
H_0 replaced by H

* However, keep in mind that in classical probability

A hypothesis **cannot have probability.**
It is either true (probability 1), or it isn't (probability 0).

The case of the p-value

Similarly to the case of a test for a rare disease, if the hypothesis H is very unlikely, than even if we get small p-value P(E | H), the P(H | E) can be very very small, ex. still <1%

This also explains the arbitrary choice of p-value cutoff depending on the science field

Analogy

Let’s say we have a robbery.

The evidence we have is a blurry photo where we can see a silhouette of a person.

The argument is as follows: the person is too tall to be a woman. So the photo is evidence that the robbery was performed by a man.

Analogy

Let’s say we have a robbery.

The evidence we have is a blurry photo where we can see a silhouette of a person, but the person is green.

The argument is as follows: the person looks too green for a human. So the photo is evidence that the robbery was performed by an alien.

Analogy

Let’s say we have a robbery.

The evidence we have is a blurry photo where we can see a silhouette of a person, but the person is green.

The argument is as follows: the person looks too green for a human. So the photo is evidence that the robbery was performed by an alien.

How do you estimate probability ot theses hypotheses, given similar evidence:

-the robbery was performed by a man
-the robbery was performed by an alien

Analogy

What if the evidence is “stronger” (i. e. smaller p-value), for example the photo is higher resolution?

How do you estimate probability ot theses hypotheses, given similar evidence:

-the robbery was performed by a man
-the robbery was performed by an alien

The p-value

The p-value is **NOT** a measure of the probability of the hypothesis we are testing

A small p-value only indicates evidence against the null hypothesis

Smaller p-value = stronger evidence

The p-value is **NOT** a measure:

- of the strength of the effect

- of the probability of the hypothesis we are testing

How can we get smaller p-value?

P-values are not a measure of effect

P-values depend directly on samples size.

If one study has small sample size, and the second one has large sample sizes, then we will get different p-values even if the underlying effects are exactly the same.

Measuring the “effect” in a study can compare results of different studies.

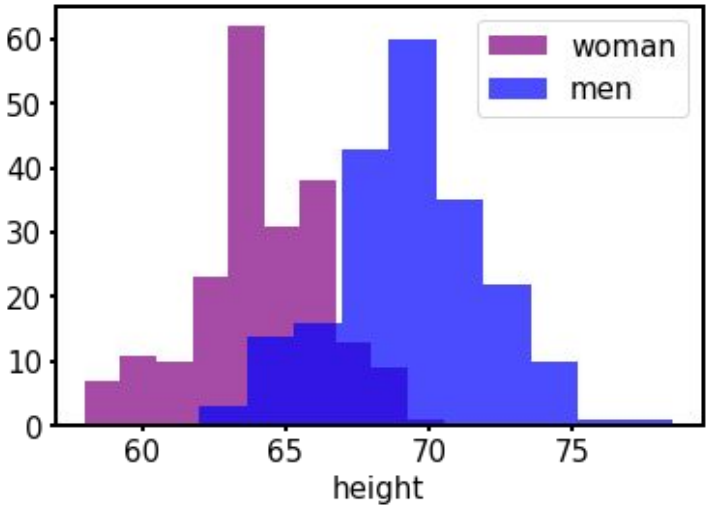
Effect size

A measure strength of the effect that doesn’t depend on the sample size.

Often, but not always, standardized in scale.

Example of height

P-value <0.0001 highly statistically significant result

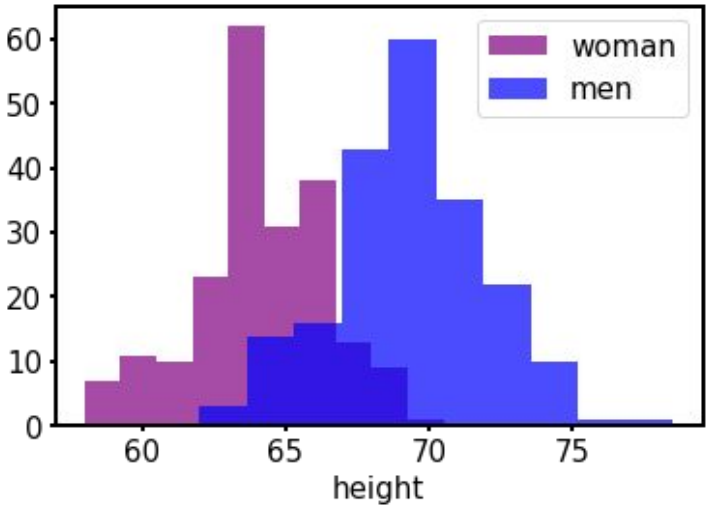


Example of height

Only 3% of men are shorter than the average woman

or for ex.

The average man is 5.3 inches (13 cm) higher than the average woman



Cohen's d

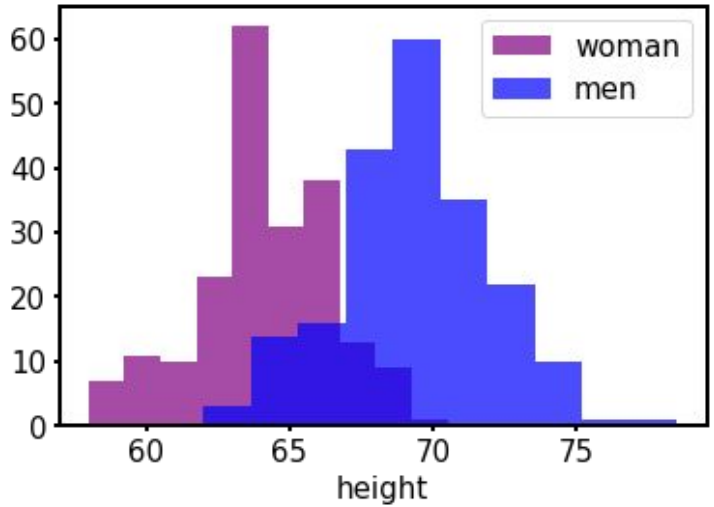
"Standardized" difference in means

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}.$$

where :

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$\bar{x}_1$ s_1^2 -sample mean and sample variance



In this case Cohen's d = 2.15
It will be similar regardless of sample size.

Cohen's d

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}.$$

where :

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$\bar{x}_1$ s_1^2 -sample mean and sample variance

t-statistics

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

Cohen’s d

Caution:
The validity of the cohen’s d as a measure of effect depends on the same assumption as the t-test (independence, normality, equal variance).

If you do Welch test, Mann-Whitney U test, it doesn’t really make sense to report Cohen’s

(note: t-test vs Welch test)

$$t = \frac{\Delta \overline{X}}{s_{\Delta \bar{X}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s^2_{\bar{X}_1} + s^2_{\bar{X}_2}}}$$

$$s_{\bar{X}_i} = \frac{s_i}{\sqrt{N_i}}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s^2_{X_1} + (n_2 - 1)s^2_{X_2}}{n_1 + n_2 - 2}}$$

$\bar{x}_1$ s_1^2

-sample mean and sample variance

Mann-Whitney U test

Effect sizes?

Often, people will just report differences in mean.

Alternative:

$$1 - \frac{2U}{n1\ n2}$$

where U is the Mann-Whitney U test statistics, n1 an n2 are the sample sizes

Correlation coefficient

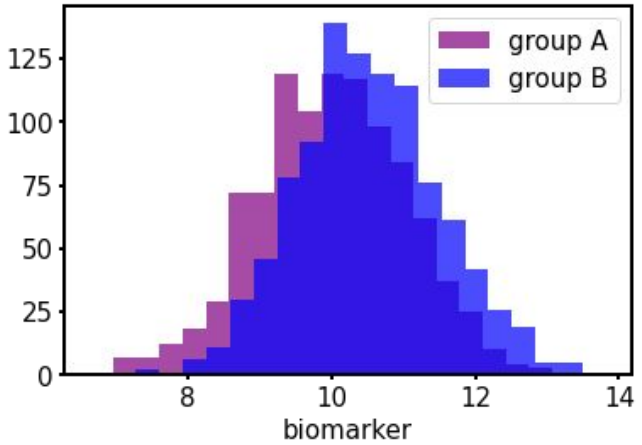
Both pearson and spearman can be reported as effect sizes.

Statistical significance vs practical significance

Ex. we are looking for a biomarker of a disease.
Two different studies report “significant results”:

p-value of t-test = 3e-62
N samples = 1000
Group B has significantly higher levels of the biomarker than group A.

But, 30% of samples from group A have higher levels of the biomarker than average person from group B, vs 70% lower.



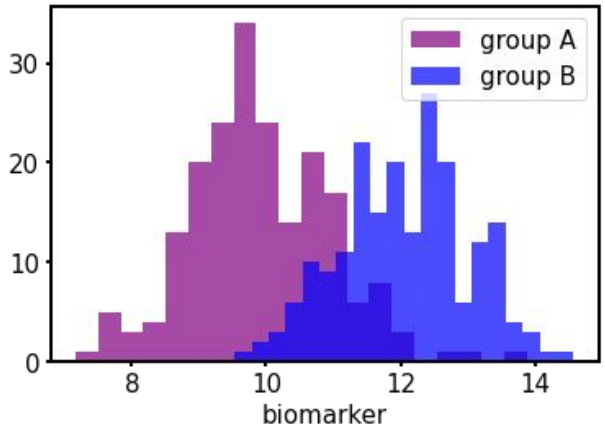
Small “significant” p-value doesn’t mean actual significance in real lift

Statistical significance vs practical significance

Ex. we are looking for a biomarker of a disease.
Two different studies report “significant results”:

p-value of t-test = 2e-66
N samples = 200
Group B has significantly higher levels of the biomarker than group A.

But, only 3% of samples from group A have higher levels of the biomarker than average person from group B, vs 97% lower.



Small “significant” p-value doesn’t mean actual significance in real lift

Statistical significance vs practical significance

So in this example,

just looking at the distributions and making some intuitive conclusions about how “different” the distributions of the two groups are,

is a better inference method than just relying on the p-values alone.

Statistical significance vs practical significance

So in this example,

just looking at the distributions and making some intuitive conclusions about how “different” the distributions of the two groups are,

is a better inference method than just relying on the p-values alone.

So what is the point of statistical testing, and in which cases are we interested in a small, or maybe “very small” p-value?

P-values are still usefull

- P-value below 5%, small effect, large sample size

P-values are still usefull

- P-value below 5%, small effect, large sample size

ex. two buttons red and blue in web interface, 10000 cases and 5% shift

P-values are still usefull

- P-value below 5%, small effect, large sample size

ex. two buttons red and blue in web interface, 10000 cases and 5% shift

- P-value below 5%, large effect, small sample size

P-values are still usefull

- P-value below 5%, small effect, large sample size

ex. two buttons red and blue in web interface, 10000 cases and 5% shift

- P-value below 5%, large effect, small sample size

ex. you do a “pilot” study before deciding to study a blood marker in disease (reasonable hypothesis)

P-values are still usefull

- P-value below 5%, small effect, large sample size

ex. two buttons red and blue in web interface, 10000 cases and 5% shift

- P-value below 5%, large effect, small sample size

ex. you do a “pilot” study before deciding to study a blood marker in disease (reasonable hypothesis)

A significant p-value can be used, for ex. to answer the following question: Is something there? Should i investigate more?

P-values are still usefull

- P-value below 5%, small effect, large sample size

ex. two buttons red and blue in web interface, 10000 cases and 5% shift

- P-value below 5%, large effect, small sample size

ex. you do a “pilot” study before deciding to study a blood marker in disease (reasonable hypothesis)

- Collecting more data to get a very small p-value, much smaller than 5%

P-values are still usefull

- P-value below 5%, small effect, large sample size

ex. two buttons red and blue in web interface, 10000 cases and 5% shift

- P-value below 5%, large effect, small sample size

ex. you do a “pilot” study before deciding to study a blood marker in disease (reasonable hypothesis)

- Collecting more data to get a very small p-value, much smaller than 5%

ex. the hypothesis is very “outlandish” and unlikely. Or maybe this is a pilot study before implementing something very expensive, so the chance of a 5% error is very high

Because the main value of our research is the estimation of effect sizes and of their uncertainty, our emphasis should shift to the clear and comprehensive presentation of point estimates and their associated interval estimates. A straightforward way of doing so is interpreting the classical confidence intervals as compatibility intervals (Amrhein et al., 2019a, b; Gelman & Greenland, 2019; McElreath, 2020; Rafi & Greenland, 2020).

https://pmc.ncbi.nlm.nih.gov/articles/PMC9322409/