Section 12: Difference of means

STOR 155.02, Spring ’21

updated 2021-04-20

What you will learn

- is in the title

Resources

- Textbook ch 6.2

Eyes on the schedule

Three lectures remain

- 20, 27, May 4

Final project due May 11, 3p.m. sharp

- **no excuses for lateness**
- **except** by approved university exam excuses or exceptional circumstances
- assigned one week prior

Question:

Do women who have taken hormonal birth control have a higher risk for breast cancer than those who haven't?

Contemporary Hormonal Contraception and the Risk of Breast Cancer

a large-scale study in Denmark from a few years ago

How to answer this question?

Mathematical formulation

'Treatment group'

hormonal birth control (BC) users

all biologically female

random sample size n_1

from a 'true' but unknown r.v. X

$$X_i = 1\{\text{ith person got cancer}\}$$

true cancer probability for BC users

$$m_1 = E(X)$$

'Control' group

females who never have used BC

random sample of size n_0

from unknown r.v. Y

$$Y_i = 1\{\text{ith person got cancer}\}$$

true cancer prob for non-BC group

$$m_0 = E(Y)$$

Do a hypothesis test!

$$H_0 : m_1 = m_0, \qquad H_1 : m_1 > m_0, \quad \text{or} \quad m_1 \neq m_0$$

would allow us to say something like

> > with xx% confidence we reject the null hypothesis that birth control users have the same chance of contracting cancer as non-users.

Problem

previously, we needed to know the value m_0

here m_0 **also is not known!**

Recap: calculation for 'two-sided test' when null-hypothesis value is known

point estimates from data

$$\hat{s}_n/\sqrt{n} = \text{ point estimate for s.d. of } \bar{X}_n$$

$$\hat{m}_n = \text{ point estimate for mean of X}$$

p-value

$$p = 2P\left(Z < -\frac{|\hat{m}_n - m_0|}{\hat{s}_n/\sqrt{n}}\right) \approx P\left(\frac{|\bar{X}_n - m_0|}{\hat{s}_n/\sqrt{n}} > \frac{|\hat{m}_n - m_0|}{\hat{s}_n/\sqrt{n}}\right)$$

decision rule for level α test

$$\text{reject } H_0 \text{ if} \quad p \leq \alpha$$

example: $\alpha = 0.03$

$$\text{reject } H_0 \text{ with 97\% confidence if} \quad p \leq 0.03$$

Difference of means tests

Want to test

$$H_0 : m_1 = m_0, \qquad H_1 : m_1 \neq m_0$$

Reformulate

$$H_0 : m_1 - m_0 = 0, \qquad H_1 : m_1 - m_0 \neq 0$$

now it's the same as we saw previously, with

$$m = m_1 - m_0$$

Ingredients

$$\hat{m}_{n_1,n_0} = \text{ point estimate for the mean of the difference } \bar{X}_{n_1} - \bar{Y}_{n_0}$$

$$\hat{s}_{n_1,n_0} = \text{ point estimate for the s.d. of the difference } \bar{X}_{n_1} - \bar{Y}_{n_0}$$

Review Q:

If V_1, V_0 are independent random variables with s.d. s_1, s_0, what is

$$sd(V_1 + V_2) = \ ?$$

Apply this to the difference of means test problem

$$V_1 = \bar{X}_{n_1}, \qquad V_2 = -\bar{Y}_{n_0}$$

$$sd\left(\bar{X}_{n_1} - \bar{Y}_{n_0}\right) = \ ?$$

Calculating p-values for difference of means

point estimates from data

$$\hat{m}_{n_1,n_0} = \hat{m}_{n_1} - \hat{m}_{n_0} = \text{ difference of sample means, treatment vs. control}$$

$$\hat{s}_{n_1,n_0} = \sqrt{\frac{\hat{s}_{n_1}^2}{n_1} + \frac{\hat{s}_{n_0}^2}{n_0}}$$

p-value for two-sided test

Same formula, different point estimates.

$$p = 2P\left(Z < -\frac{|\hat{m}_{n_1} - \hat{m}_{n_0}|}{\hat{s}_{n_1,n_0}}\right) \approx P\left(\frac{|\bar{X}_{n_1} - \bar{Y}_{n_0}|}{\hat{s}_{n_1,n_0}} > \frac{|\hat{m}_{n_1} - \hat{m}_{n_0}|}{\hat{s}_{n_1,n_0}}\right)$$

could also do a confidence interval

$$\text{high, low} = (\hat{m}_{n_1} - \hat{m}_{n_0}) \pm z^* \hat{s}_{n_1,n_0}$$

When does it make sense to use this test?

recap: assumptions for one-sample test

CLT/LLN are valid

meaning random sample

from a common population

that is b̀ig enough'

two-sample assumptions

one-sample assumptions apply to both treatment and control groups

treatment and control are independent from each other

Example: Breast cancer study

From breast cancer study in Denmark, table 2.

- **Treatment**: Any hormonal contraception in current/recent use
- **Control**: Never used hormonal contraception

	n	cancer_per_100k	mhat	shat	shat_mean
control	78.15	4298.35	68	27.76	3.25
treatment	73.08	4969.74	55	27.76	3.14

Units in the data

I've simplified a few things to make it fit in our class's discussion, but **our conclusions (p-values) will match theirs.**

`n` is in 100,000 'person years', one year of life for a person in the group. Think of this as one observation. `shat` values are $\hat{s}_{n_0}$ and $\hat{s}_{n_1}$ used in the calculation of $\hat{s}_{n_1,n_0}$. `shat_mean` is `shat` / `sqrt(n)`.

One-sided test

Analogous to one-sided version of one-sample test.

$$H_1 : m_1 - m_0 > 0$$

$$p = P\left(Z < -\frac{\hat{m}_{n_1} - \hat{m}_{n_0}}{\hat{s}_{n_1,n_0}}\right) = P\left(Z < -\frac{68 - 55}{\sqrt{3.25^2 + 3.14^2}}\right)$$

p
0.0020093

Same p-value as the paper

See note about about how I simplified this for class.

Q: How would you interpret this result?

Use language from class

Aside: Is this bad news for me (you, all of us)?

Absolute risk, relative risk and 'significance'

relative

- are birth-control users more at risk for breast cancer **compared to non-BC users**?
- is this difference **statistically** significant, at some α level?

That's why we do hypothesis tests.

A statistical question

absolute

- is birth-control use a **substantial contributor to my overall risk** of injury/death?
- compared to **other sources, e.g. car accidents**?

even the people in the study who used birth control 10+ years had risk of *contracting* breast cancer that was ~15 smaller than the risk of dying in a car accident in the US

- Cancer rate for that group: ~ 1 per 100k 'life years'
- motor vehicle fatalities in the US: 12-15 per 100k people/year

PollEv.com/brendanbrown849

poll closes at _________


five more minutes