Statistical Reasoning

Week 6

Sciences Po - Louis de Charsonville

Spring 2018

Outine

Research Paper

Research Paper

Timeline

1^{st} draft	Done
2^{nd} draft	**10 April**
Final draft	**24 April**

Statistical Hypotheses

Last Week Reminder

- Sample $\neq$ Population
- A short video on Central Limit Theorem

Univariate statistics $\Rightarrow$ Bivariate statistics

Bivariate statistics

- Describing the **association** between two variables
- Measuring the **strength** of the relationship.

Comparing means across groups

- Dependent Variable is quantitative → you can calculate a **mean**
- The value of a mean is not always interesting in itself ...
- But **comparing the means of two groups** might highlight relationships between variables

- **Descriptive statistics** : describe and compare the difference in means
- **Inferential statistics** : determine whether or not the difference in means is due to random chance or is it statistically significant.

Statistical significance
Likelihood that a statistic derived from a *sample* represents some genuine phenomenon in the *population*.

Example

- Sample : 1000 women
- $\bar{X} = 9$, $\sigma_X = 2$
- Population mean : 8.

1. *What is the probability to observe a mean of 9 ?*
2. *Is this difference statistically significant ?*

Sampling error

$$s_{\bar{X}} = \frac{2}{\sqrt{1000}} = 0.06$$

t Value

$$t = \frac{9-8}{0.06} = 16.67$$

- Using the *t distribution*, the probability associated with a t value of 16.67 is less than 0.001 (see Table of critical values for the *Student* distribution in Annex).
- This probability is known as **p-value**.

Is this difference statistcally significant ?

- We need to delve into *hypothesis testing.*

No Booze? You May Loose, 2006

"A number of theorists assume that drinking has harmful economic effects, but data show that drinking and earnings are positively correlated. ***We hypothesize that drinking leads to higher earnings by increasing social capital****. If drinkers have larger social networks, their earnings should increase. Examining the General Social Survey, we find that self-reported drinkers earn 10-14 percent more than abstainers, which replicates results from other data sets"*

H_1 : "An increase in social drinkings leads to an increase in earnings."

Hypotheses

Substantive hypotheses

H_1 : $\pm$ social drinking $\rightarrow$ $\pm$ social capital $\rightarrow$ $\pm$ earnings
H_2 : $\pm$ earnings $\rightarrow$ $\pm$ disposable income $\rightarrow$ $\pm$ social drinking

Rejecting the null hypothesis H_0

H_0 : no relationship between social drinkings and earnings
H_1 : any relationship between social drinking and earnings

Proof by contradiction

- Reject of retain H_0 with a certain level of confidence (usually 95% or 99%)

Null hypothesis

Null hypothesis H_0

- The null hypothesis always suggests that there will be an absence of effect *in the population*. Denoted H_0.
- Usually, researchers want to reject H_0 with a certain level of confidence.

Alternate hypothesis H_1

- Alternative to the null hypothesis. Usually,it is the hypothesis that there is some effect present in the population. Denoted H_1.

Type I and II Errors

Errors in hypothesis testing 1/2

Type I Error - Rejecting H_0 when it is actually true

"Last year executed man proven innocent by DNA evidence."

- H_0 : innocent...
- H_1 : ... until proven guilty

H_0 wrongly rejected

Type II Error - Retaining H_0 when its actually false

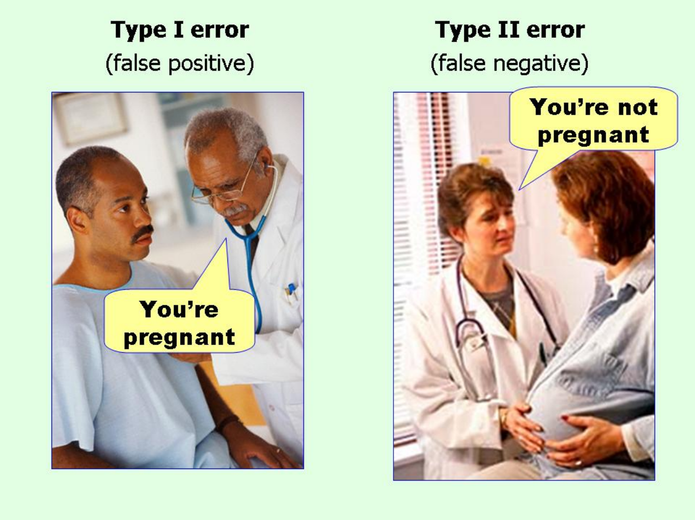
"I didn't see your email Professor, it was in the spam folder."

- H_0 : email is spam
- H_1 : email is non-spam

H_0 wrongly accepted.

Errors in hypothesis testing 2/2

H_0 : **Patient is *not* pregnant**



t-Tests

Different type of t-tests

Different t-tests

- **Independent samples t-test** : compare the means of two *independent* samples on a given variable.
  - Example : compare the average earnings between 100 randomly chosen social drinker and 100 randomly chosen nondrinker.
- **Dependent samples t-test** : compare two means on a single dependent variable of two matched or paired samples.
  - Example : compare average grade for first draft and final draft

What do t-test measure ?

- Do two independent samples differ from each other **significantly** in their average scores on some variable ?
- **Significantly** = the difference in **samples**' average is large enough to suggest a difference in **population**.
- Inferential statistics : *Sample* ⇒ *Population*.
- What difference can I expect to observe given **random sampling error** ?
- What is the average expected difference between the means of two samples of this size selected randomly ?
- What is the **standard error of the difference** between the means ?
- Is our observed difference between our two sample means is large relative to the standard error of the difference between the means.

t-test tells us how likely it is to observe a difference between two values in a sample if no real difference actually exists in the population.

Example - Earnings and drinkings

1. Assume H_0 is true : No difference between in earnings of drinkers and nondrinkers in the *population.*
2. Under H_0, how likely (e.g. what is the probability) is it to observe such a difference in group means ? (**p-value,** p).
3. Two cases :
   - $p < 0.05$ (e.g. it is very unlikely) : reject $H_0 \rightarrow$ there is a difference in earnings between drinkers and non-drinkers.
   - $p > 0.05$: We can't reject H_0.

Mathematics of the t-tests

We compute the t statistics :

$$t = \frac{\text{observed difference between sample means}}{\text{standard error of the difference between the means}} \quad (1)$$

$$= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}}} \quad (2)$$

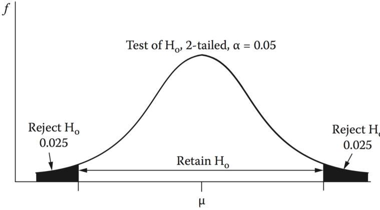
With :

- $\bar{X}_1$ is the mean for the first sample
- $\bar{X}_2$ is the mean for the second sample
- $s_{\bar{x}_1}$ is the standard error of the mean for the first sample
- $s_{\bar{x}_2}$ is the standard error of the mean for the second sample

One-tailed vs Two-tailed tests

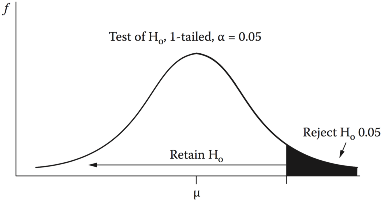
Two-tailed t-tests

- $H_0 : \bar{X} = A$
- $H_1 : \bar{X} \neq A$



One-tailed t-tests

- $H_0 : \bar{X} = A$
- $H_1 : \bar{X} < A$



In Stata

Comparing differences

- Comparing means $H_0 : \Delta = \bar{X} - \bar{Y} = 0$ `ttest`
- Comparing proportions $H_0 : \Delta = Pr X - Pr Y = 0$ `prtest`

`ttest y, by(x)+`

- `y` is continuous, `x` is a dummy
- use `prtest` if `y` is also a dummy (proportions test).
- use `tab, gen()` to create dummies from categories variables

Bivariate statistics - Which procedure ?

The choice of the appropriate procedure depends on the **type of variables** you are working with

- Two quantitative variables ?
  - Correlation coefficient
- One quantitative / one qualitative variable ?
  - t-Tests
- Two qualitative variables ?
  - Cross tables ; Chi-square ; Cramer's V

Practice

Practice

Opposition to Torture in Israel

What matters

- Age ? Gender ? Income ? Education ?
- Religious faith ?
- Media exposure ?

Before the session

- Run `setup.do` (do it again !)
- Check that all packages are installed for this session : `fre`, `renvars`, `spineplot`, `tab_chi`
- `doedit code/week6.do`

Annex

	α Level for Two-Tailed Test					
	.20	.10	.05	.02	.01	.001
	α Level for One-Tailed Test					
df	.10	.05	.025	.01	.005	.0005
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.474	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Note: To be significant the *t* value obtained from the data must be equal to or greater than the value shown in the table.