Statistical Reasoning

Week 7

Sciences Po - Louis de Charsonville

Spring 2018

Outline

Research Paper

Chi-Square

Correlation

Research Paper

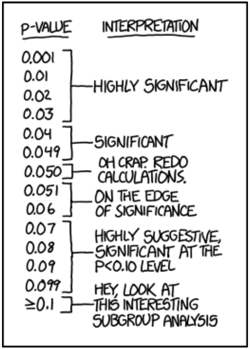
Research Paper

Timeline

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

Last Week Reminder 1/3

- Statistical tests provide **proof by contradiction**.
- We posit that there is *no association*, H_0
- We run test to try to reject H_0.
- A **p value** measure the probability to observe the result, plus more extreme results under H_0.



P-VALUE	INTERPRETATION
0.001 0.01 0.02 0.03	HIGHLY SIGNIFICANT
0.04 0.049	SIGNIFICANT
0.050	OH CRAP. REDO CALCULATIONS.
0.051 0.06	ON THE EDGE OF SIGNIFICANCE
0.07 0.08 0.09 0.099	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Last Week Reminder 2/3

Steps in hypothesis testing

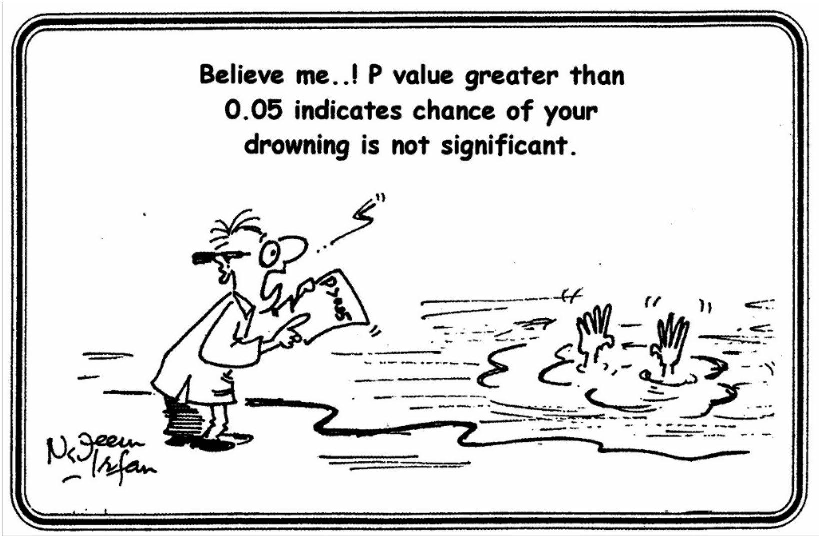
1. Specify the null hypothesis H_0. A typical null hypothesis is $\mu_1 = \mu_2$.
2. Specify the α level = significance level. Typical values are 0.05 or 0.01.
3. Compute the p-value
4. Compare the p-value with α.
5. If $p < \alpha$, reject H_0

⚠ **Failure to reject the null hypothesis does not constitute support for the null hypothesis**

Last Week Reminder 3/3

Keep in mind

- Association is not causation.
- Statistical significance is not substantive significance.
- Statistical significance does not provide an order of magnitude to the substantive significance of the association. Statistical strength of an association relies on data and sample size and does not indicate that association is theoretically more important.
- Do not accept the null hypothesis when you do not reject it.


Believe me..! P value greater than 0.05 indicates chance of your drowning is not significant.
P>0.05
Naveen Irfan

P-Values misconceptions

Non-exhaustive list of false statements on p values

1. If $p = 0.05$, the null hypothesis has only a 5% chance of being true.
2. A nonsignificant difference (eg, $p \geq 0.05$) means there is no difference between groups.
3. A statistically significant finding is substantively important.
4. $p = 0.05$ means that we have observed data that would occur only 5% of the time under the null hypothesis.
5. $p = 0.05$ means that if you reject the null hypothesis, the probability of a type I error is only 5%.
6. With a $p = 0.05$ threshold for significance, the chance of a type I error will be 5%.

Find a complete list here.

t test - Wrapping up

- *t* test : statistical similarity between the means of two independent samples on a single variable
- Independent variable is always a **dummy**.
- Dependent variable is quantitative `ttest`
- Dependent variable is a dummy `prtest`

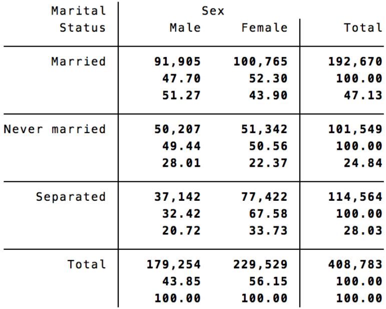
Chi-Square

Cross-Tabulations

- Break down dependent variable according to categories of the independent variable.
- Conventionally, the **dependent variable** is placed in columns and the independent variables in rows.

Stata command

```
tab y x, row col
```



```
      Marital |          Sex
       Status |      Male     Female |     Total
--------------+----------------------+----------
      Married |    91,905    100,765 |   192,670
              |     47.70      52.30 |    100.00
              |     51.27      43.90 |     47.13
--------------+----------------------+----------
Never married |    50,207     51,342 |   101,549
              |     49.44      50.56 |    100.00
              |     28.01      22.37 |     24.84
--------------+----------------------+----------
    Separated |    37,142     77,422 |   114,564
              |     32.42      67.58 |    100.00
              |     20.72      33.73 |     28.03
--------------+----------------------+----------
        Total |   179,254    229,529 |   408,783
              |     43.85      56.15 |    100.00
              |    100.00     100.00 |    100.00
```

Chi-Square Test or χ^2 test

- A significance test measuring whether two variables are independent.
- For comparing **qualitative variables**.
- H_0 : the two variables are independent.
- If there is no difference (= no distance) between the observed and theoretical situation.
- We usually want to reject H_0 and conclude that the variables are dependent.

Chi-Square test in practice 1/5

Question : **Do men and women differ in their selection of college majors ?**

- Population : liberal arts college in the US.
- Sample : random selection of 100 men and 100 women.

	Psychology	**English**	**Biology**
Men	35	50	15
Women	30	25	45

- Statistically significant difference between genders in college major choices ?
- Is the distribution different from what we would expect if everything were left to chance ?

Chi-Square test in practice

- H_0 : there is no difference between genders in college major choices
- Under H_0, what is the expected distribution among college major ?

	Psychology	**English**	**Biology**	*Row Total*
Men	35	50	15	100
Women	30	25	45	100
Column Total	65	75	60	200

$$\begin{aligned}\text{Expected men in Psychology} &= \frac{\text{Nb of men}}{\text{sample size}} * \frac{\text{Nb of Psychology majors}}{\text{sample size}} * \text{sample size}\\ &= \frac{100}{200} * \frac{65}{200} * 200\\ &= 32.5\end{aligned}$$

Expected Frequencies

	Psychology	English	Biology	Row Totals
Men	32.5	37.5	30	100
Women	32.5	37.5	30	100
Column totals	65	75	60	200

Formula for calculating χ^2

$$\chi^2 = \sum \left(\frac{(O-E)^2}{E} \right)$$

Where O is the observed value in each cell and E the expected value in each cell.

Computing χ^2

	Psychology	English	Biology
Men	$\frac{(35-32.5)^2}{32.5}$	$\frac{(50-37.5)^2}{37.5}$	$\frac{(15-30)^2}{30}$
Women	$\frac{(30-32.5)^2}{32.5}$	$\frac{(25-37.5)^2}{37.5}$	$\frac{(45-30)^2}{30}$

$$\chi^2 = 23.7$$

Chi-Square test in practice 5/5

Interpretation of χ^2

- We use Chi-Square table of distributions to determine the p value.
- Calculation depends on **degrees of freedom** : $df =$ (nb of rows -1) $*$ (nb of cols -1)
- Compare this χ^2 to critical χ^2 value.
- For an alpha-level of 0.05, critical value is 5.99.
- Our observed χ^2 value is 23.7. **We reject** H_0

χ^2 test in Stata

χ^2 test in Stata

- `tab y x, exp chi2`
- use `tabchi` to inspect residuals
- use `tabodds` for odds ratios

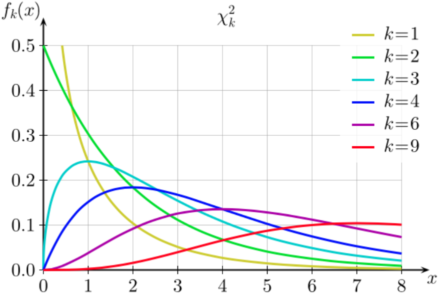
- Stata gives p-value for $\alpha = 0.05$.
- The p-value is the probability of observing this Chi-Square value (or more) if the null hypothesis were true.
- The smaller the p-value, the lower the probability of obtaining this Chi-Square value with this sample under H_0
- In practice : p-value is very small → we can reject the null hypothesis
- Small p-value means $p < \alpha$

Assumptions behind χ^2 test

- Random sampling
- Large sample size ($\approx$ 5 per cell)
- No particular shape of the distribution

χ^2 distribution

- The χ^2 distribution is the distribution of the sum of squared standard normal deviates.
- The degrees of freedom of the distribution is equal to the number of the standard normal deviates being summed.
- $\mu_{\chi^2} = d.f.$, the mean of the distribution equals the number of degrees of freedom.
- χ^2 distributions are positively skewed. Skew increase with degrees of freedom.



Writing up a Chi-Square test

Be clear and **concise** in your writing :

"A Chi-Square analysis was performed to determine whether college major choice among U.S. college students is dependent on gender. The results of the test show a significant chi-square value (p-value lower than 0,001). We can thus reject the null hypothesis and confirm that there is a relationship between gender and college majors. Women tend to choose biology as a major more frequently than men, while men are more concentrated in English major".

Cramer's V

Definition

- Cramer's V is a correlation coefficient
- measure the **strength** of the relationship between two **qualitative** variables.

Properties

- Varies between 0 and 1
- If it equals 0, there is no association.
- From 0.1, we consider that the association is moderately strong.
- Cramer's V is **not** a measure of statistical significance.

Stata

add `V` to measure the association with Cramer's V :

```
tab y x, exp V
```

Correlation

What it is and when to use it

When to use it

- How two variables relate to each other.
- Both variables are **quantitative**.

Characteristics

- Correlation coefficient can be either positive or negative
- Strength or magnitude of the correlation.

Positive vs negative correlation

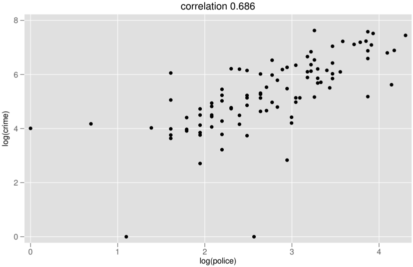


Figure – Positive correlation

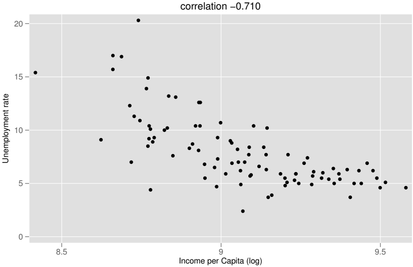


Figure – Negative correlation

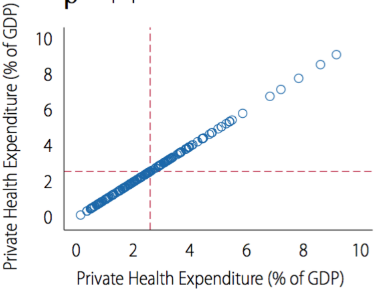
Correlation - What it tells us ?

Positive vs Negative correlation

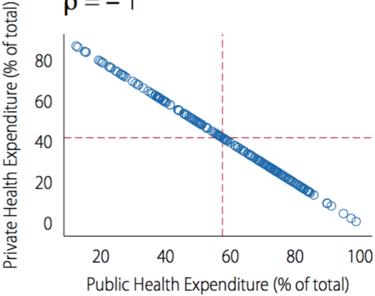
- A **positive correlation** indicates that the values on the two variables being analyzed move in the same direction.
- A **negative correlation** indicates that the values on the two variables being analyzed move in opposite directions

Strength of relationship

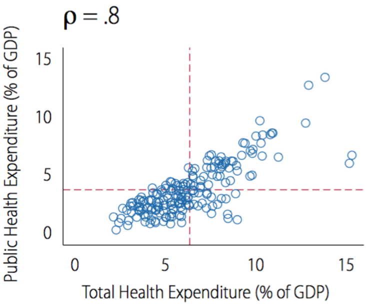
- Correlation coefficients range in strength from -1 to 1.
- A perfect negative correlation of -1 indicates that for every member of the sample, a higher score on one variable is related to a lower score on the other variable
- A perfect positive correlation of 1 reveals that for every member of the sample or population, a higher score on one variable is related to a higher score on the other variable.

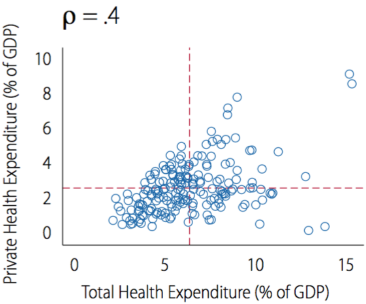

Private Health Expenditure (% of GDP)
10
8
6
4
2
0
0
2
4
6
8
10
Private Health Expenditure (% of GDP)

$\rho = +1$

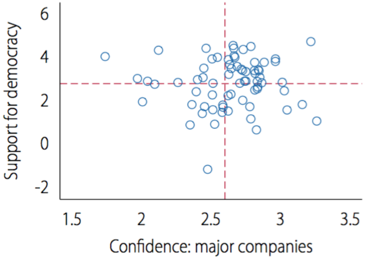

Private Health Expenditure (% of total)
80
60
40
20
0
20
40
60
80
100
Public Health Expenditure (% of total)

$\rho = -1$

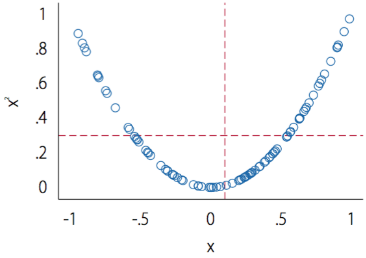

ρ = .8
Public Health Expenditure (% of GDP)
15
10
5
0
0
5
10
15
Total Health Expenditure (% of GDP)


ρ = .4
Private Health Expenditure (% of GDP)
10
8
6
4
2
0
0
5
10
15
Total Health Expenditure (% of GDP)

$\rho = .02$


Support for democracy
6
4
2
0
-2
1.5
2
2.5
3
3.5
Confidence: major companies

$\rho = .16$


x^2
1
.8
.6
.4
.2
0
-1
-.5
0
.5
1
x