Data Science in Medicine

Lecture 5: Hypothesis Testing – Part 2

Dr Areti Manataki

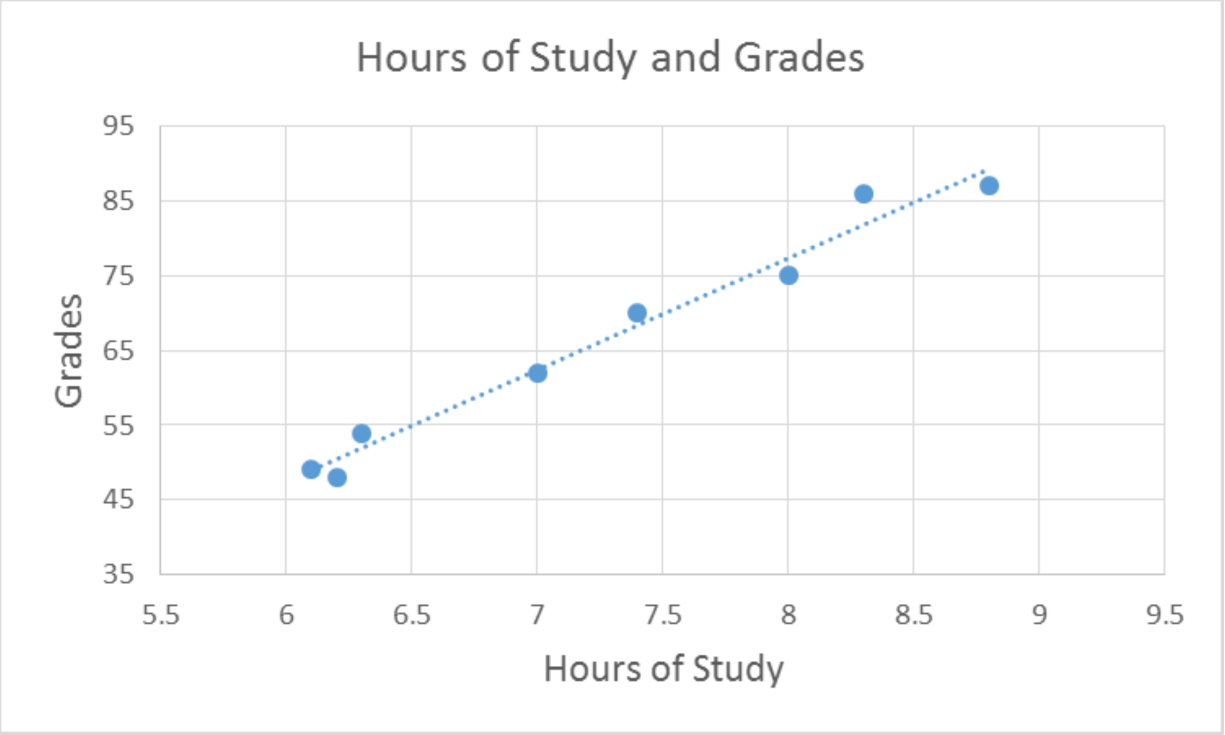
Usher Institute
The University of Edinburgh

In the previous lecture

- Correlation
  - e.g. temperature and ice cream sales in Edinburgh
  - Correlation does not imply causation!
- Arguing about correlation
  - Visualise your data
  - Calculate the correlation coefficient
  - Carry out hypothesis testing (using the correlation coefficient as a statistical test)

Example: correlation between hours of study and final grade

Weekly hours of study	Grades
8	75
7.4	70
8.3	86
6.2	48
6.3	54
7	62
8.8	87
6.1	49



Example: correlation between hours of study and final grade

- $\rho_{x,y} \simeq 0.988$
- Hypothesis testing:
  - H0: There is no correlation between weekly hours of study and final exam grades in Statistics.
  - H1: There is a correlation between weekly hours of study and final exam grades in Statistics.

ρ	$p = 0.10$	$p = 0.05$	$p = 0.01$	$p = 0.001$
$N = 7$	0.669	0.754	0.875	0.951
$N = 8$	0.621	0.707	0.834	0.925
$N = 9$	0.582	0.666	0.798	0.898
$N = 10$	0.549	0.632	0.765	0.872

In this lecture

- Correlation between two categorical variables
  - Chi-square test
- Comparing the means for two groups
  - t-test for independent samples

Correlation in categorical data

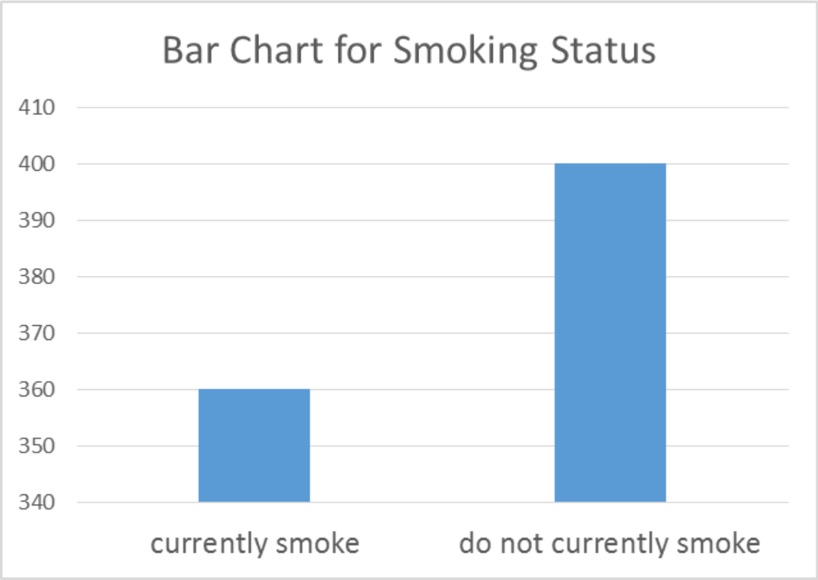
- Categorical variable examples:
  - sex: male, female
  - nationality: Vietnamese, Greek, Colombian, …
  - age: under 18 years, 18 to 29 years, 30 to 49 years, 50 to 64 years, 65 years or older
- Correlation between smoking status and lung cancer diagnosis?

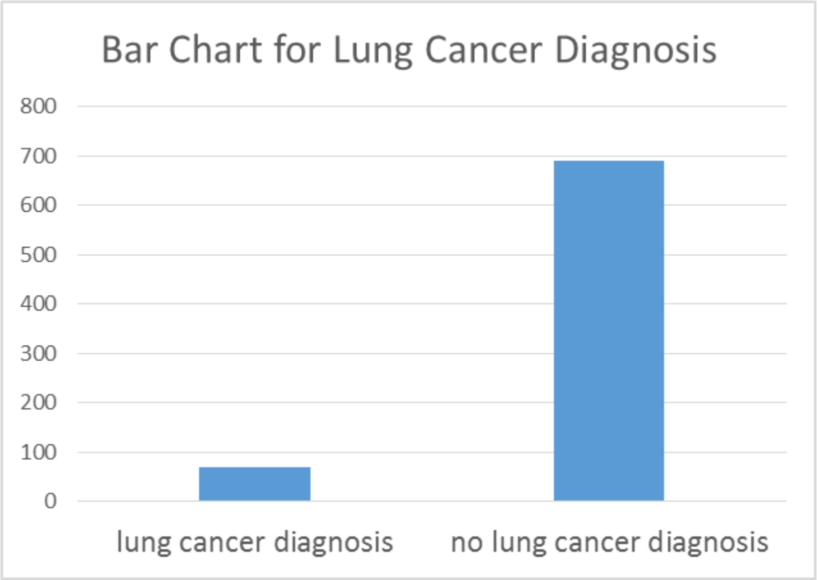
Data collected

Currently smoking?	Diagnosed with lung cancer?
Yes	No
No	No
No	No
Yes	Yes
No	Yes
Yes	Yes
No	No
…	…

Sample size: 760

Visualisations (not that informative)





These don’t tell us much about the relationship between the two variables…

Contingency table

Frequencies	Lung cancer diagnosis	No lung cancer diagnosis
Smoke	O11	O12
Not smoke	O21	O22

- O11: number of people that currently smoke and have been diagnosed with lung cancer
- O12: number of people that currently smoke and have not been diagnosed with lung cancer
- O21: number of people that do not currently smoke and have been diagnosed with lung cancer
- O22: number of people that do not currently smoke and have not been diagnosed with lung cancer

Contingency table for our example

Frequencies	Lung cancer diagnosis	No lung cancer diagnosis
Smoke	60	300
Not smoke	10	390

Contingency table (with marginals)

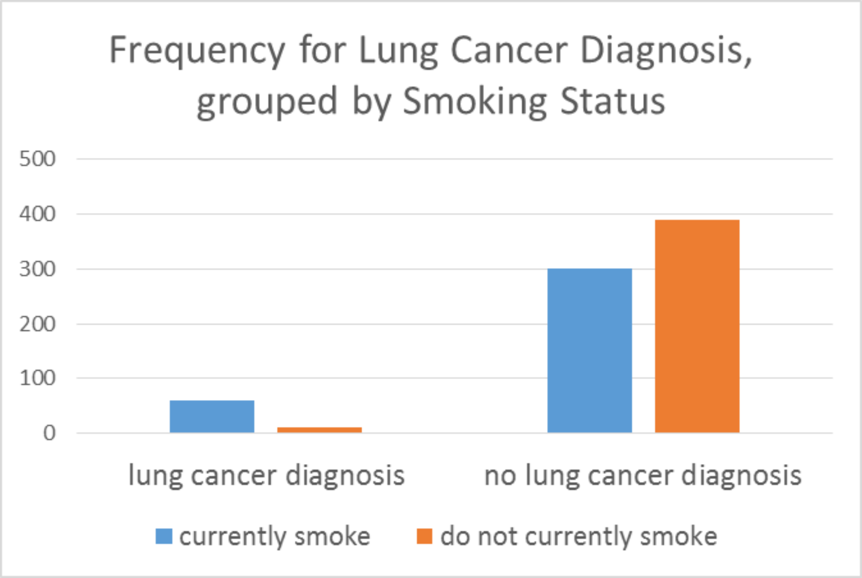
Frequencies	Lung cancer diagnosis	No lung cancer diagnosis	
Smoke	O11	O12	R1
Not smoke	O21	O22	R2
	C1	C2	N

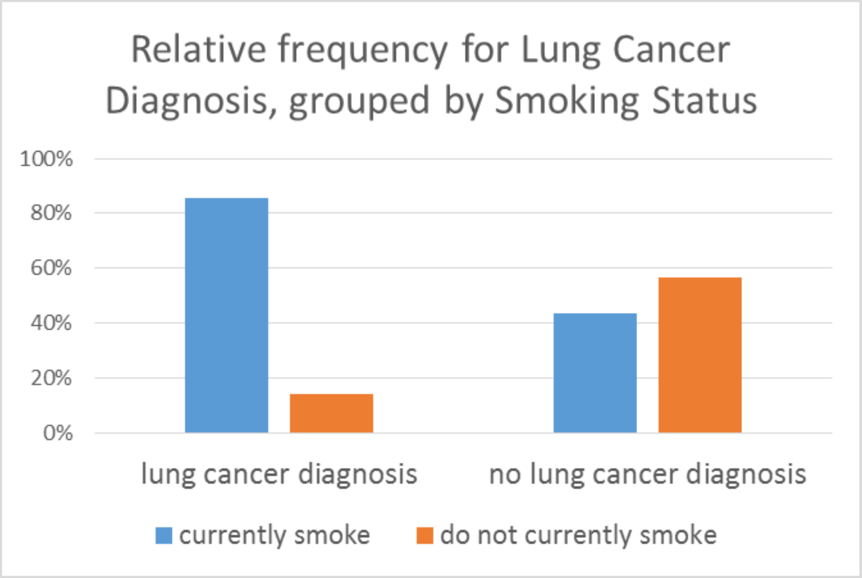
- R1 = O11 + O12 number of people that currently smoke
- R2 = O21 + O22 number of people that do not currently smoke
- C1 = O11 + O21 number of people that have been diagnosed with lung cancer
- C2 = O12 + O22 number of people that have not been diagnosed with lung cancer
- N = R1+ R2 = C1 + C2 sample size

Contingency table (with marginals) for our example

Frequencies	Lung cancer diagnosis	No lung cancer diagnosis	
Smoke	60	300	360
Not smoke	10	390	410
	70	690	760

Visualising our data





Main idea behind χ^2 test

- We have a table of observed frequencies Oij, and from these we calculate expected frequencies Eij, i.e. the numbers we would expect to see if the null hypothesis were true.
- The χ^2 value is calculated by comparing the actual frequencies to the expected frequencies.
- The larger the discrepancy between these two, the less probable it is that observations like this would occur were the null hypothesis true.
- More precisely, if the null hypothesis were true, then the χ^2 value would vary according to the χ^2 distribution.
- If the χ^2 is significantly large then we reject the null hypothesis.

Expected Frequencies

Expected frequencies	Lung cancer diagnosis	No lung cancer diagnosis	
Smoke	E11	E12	R1
Not smoke	E21	E22	R2
	C1	C2	N

- Expected frequencies: the values we would expect if the two variables were independent

$$E_{ij} = \frac{R_i \times C_j}{N}$$

Expected frequencies for our example

Expected frequencies	Lung cancer diagnosis	No lung cancer diagnosis	
Smoke	33.16	326.84	360
Not smoke	36.84	363.16	410
	70	690	760

- For example,

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{360 \times 70}{760} = 33.16$$

Combining observed and expected frequencies in a single table

Observed and expected frequencies	Lung cancer diagnosis	No lung cancer diagnosis	
Smoke	60 (33.16)	300 (326.84)	360
Not smoke	10 (36.84)	390 (363.16)	410
	70	690	760

Computing χ^2

The χ^2 statistic for a contingency table in general is defined as

$$\chi^2 = \sum_{i=1,j=1}^{i=R,j=C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Computing χ^2 for our example

Observed and expected frequencies	Lung cancer diagnosis	No lung cancer diagnosis	
Smoke	60 (33.16)	300 (326.84)	360
Not smoke	10 (36.84)	390 (363.16)	410
	70	690	760

$$\chi^2 = \sum_{i=1,j=1}^{i=R,j=C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(60-33.16)^2}{33.16} +$$

$$\frac{(300-326.84)^2}{326.84} + \frac{(10-36.84)^2}{36.84} + \frac{(390-336.16)^2}{363.16}$$

$$\approx 45.5$$

The χ^2 test

- The null hypothesis here is that there is no relationship between smoking status and lung cancer diagnosis.
- The χ^2 test indicates the probability p that data of the kind we actually see would turn up if the null hypothesis were true.
- If p is low, then we reject the null hypothesis and conclude that there is a correlation between smoking status and lung cancer diagnosis.

Critical Values for χ^2

- These are the critical values for different significance levels of the χ^2 distribution for a 2 x 2 table:

p	0.10	0.05	0.01	0.001
χ^2	2.71	3.84	6.64	10.83

- In our example $\chi^2 = 45.5$, meaning $p < 0.001$. This is evidence to suggest that there is a correlation, and we reject the null hypothesis at the 99% level. The result is statistically significant.

Interpreting the p-value in our example

- So it appears that in this data there is a correlation between smoking status and lung cancer diagnosis.
- Remember – this does not tell us whether there is any causal link between the two variables.
- But it gives a hypothesis that we could explore in further data.

Degrees of Freedom

- In tables of critical values for the χ^2 distribution, entries are usually classified by degrees of freedom.
- An r × c contingency table has $(r-1)\times(c-1)$ degrees of freedom.
- A 2 × 2 table has only one degree of freedom.

Low Frequencies

- The statistics underlying the χ^2 test become inaccurate when expected frequencies are small.
- Reasons include: inevitable differences up to 0.5 as observed values can only be whole numbers; and that χ^2 is only an approximation to the exact (but computationally more expensive) distribution.
- The test is usually considered unreliable for a 2×2 table if any cell has expected value below 5; or for a larger table, if more than 20% of cells have expected value below 5.
- For these cases there are more refined methods, such as Fisher's Exact Test.

t-tests for numerical data

One-sample t-test

- Purpose: compare the mean of a sample to a population with a known mean
- We calculate the one-sample t-test statistic by

$$t = \frac{m - \mu}{\frac{s}{\sqrt{N}}}$$

- We next consult the table of upper critical values for the t-distribution (e.g. as in this link) to see if we can reject the null hypothesis at the significance level of choice.

Assumptions in the one-sample t-test

- Normality: the population distribution is normal
- Independence: the observations in our sample are generated independently of one another

Independent samples t-test

- Main idea: compare the means of two samples that were independently drawn, with the purpose to determine whether the means of the corresponding populations are the same
- The t statistic is calculated as

$$t = \frac{m_1 - m_2}{\sqrt{{s_p}^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where

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

Independent samples t-test

- After calculating the t-statistic we consult the table of critical values for the t-distribution
- Assumptions of this test:
  - Normality: the population distribution is normal
  - Independence: the observations in our sample are generated independently of one another, both within and across samples
  - Homogeneity of variance: the population standard deviation is the same in both groups

Conclusions

- Chi-square test
  - State H0 and H1
  - Create contingency table
  - Calculate expected frequencies
  - Compute χ^2 statistic and consult table of critical values
- Tests for comparing two means
  - One-sample t-test
  - Independent samples t-test