Chapter 7

Inference for Proportions

Contents

```
# Import all needed python packages
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as stats
import statsmodels.formula.api as smf
import statsmodels.api as sm
import statsmodels.stats.proportion as smprop
```

7.1 Passing proportions

|||| Exercise 7.1 Passing proportions

To compare the level of 2 different courses at a university the following grades distributions (given as number of pupils who achieved the grades) were registered:

	Course 1	Course 2	*Row total*
Grade 12	20	14	34
Grade 10	14	14	28
Grade 7	16	27	43
Grade 4	20	22	42
Grade 2	12	27	39
Grade 0	16	17	33
Grade -3	10	22	32
Column total	108	143	251

The passing proportions for the two courses, p_1 and p_2 should be compared. As the grades -3 and 0 means not passed, we get the following table of the number of students:

	Course 1	Course 2	*Row total*
Passed	82	104	186
Not passed	26	39	65
Column total	108	143	251

a) Compute a 95% confidence interval for the difference between the two passing proportions.

|||| Solution

We use the formula for a (large sample) confidence band for the difference of two proportions (Method 7.15): with $x_1 = 82$, $n_1 = 108$, $x_2 = 104$, $n_2 = 143$ and $\alpha = 0.05$, so

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} = 0.032,$$

and $z_{0.025} = 1.96$. By the way

$$\sqrt{\frac{\frac{x_1}{n_1}\left(1-\frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1-\frac{x_2}{n_2}\right)}{n_2}} = \sqrt{\frac{x_1\left(n_1-x_1\right)}{n_1^3} + \frac{x_2\left(n_2-x_2\right)}{n_2^3}}.$$

Therefore the answer is:

$$0.032 \pm 1.96\sqrt{\frac{82\cdot 26}{108^3} + \frac{104\cdot 39}{143^3}} \Leftrightarrow [-0.0768; 0.141].$$

b) What is the critical values for the χ^2-test of the hypothesis $H_0 : p_1 = p_2$ with significance level $\alpha = 0.01$?

|||| Solution

This test has degrees of freedom $(2-1)(2-1) = 1$, with the critical value $\chi^2_{0.99}$, so the correct answer is:

```
print(stats.chi2.ppf(0.99, 1))
```

```
6.6348966010212145
```

c) If the passing proportion for a course given repeatedly is assumed to be 0.80 on average, and there are 250 students who are taking the exam each time, what is the expected value, μ and standard deviation, σ, for the number of students who do not pass the exam for a randomly selected course?

|||| Solution

If X is the number of students not passing a randomly selected course, this random variable follows the binomial distribution with $n = 250$ and $p = 0.20$, so we use the formulas from Theorem 2.21 for the mean and variance of the binomial distribution

$$\mu = np = 0.2 \cdot 250 = 50, \quad \sigma^2 = np(1-p) = 250 \cdot 0.2 \cdot 0.8 = 40 = 6.32^2.$$

Thus the answer is $\mu = 50$ and $\sigma = 6.32$.

7.2 Outdoor lightning

|||| Exercise 7.2 Outdoor lighting

A company that sells outdoor lighting, gets a lamp produced in 3 material variations: in copper, with painted surface and with stainless steel. The lamps are sold partly in Denmark and partly for export. For 250 lamps the distribution of sales between the three variants and Denmark/export are depicted. The data is shown in the following table:

	Country	
	Danmark	Export
Copper variant	7.2%	6.4%
Painted variant	28.0%	34.8%
Stainless steel variant	8.8%	14.8%

a) Is there a significant difference between the proportion exported and the proportion sold in Denmark (with $\alpha = 0.05$)?

|||| Solution

The situation asked about here is a "one sample proportion" case, where 110 (44% out of 250) are sold in Denmark (and hence 250-110=140 for export). Using Method 7.11 the standard statistic for the hypothesis test $H_0 : p = 0.5$, is

$$z_{\text{obs}} = \frac{140 - 250 \cdot 0.5}{\sqrt{250 \cdot 0.5 \cdot 0.5}},$$

the critical values are $\pm z_{0.975} = \pm 1.96$.

So the correct answer is: no, since $15/\sqrt{250/4} = 1.90$ is within ± 1.96.

b) The relevant critical value to use for testing whether there is a significant difference in how the sold variants are distributed in Denmark and for export is (with $\alpha = 0.05$)?

Solution

This is a so-called null hypothesis of homogeneity in a 3×2 frequency table ($r \times c$ table, see Method 7.22). The critical value for the χ^2-test is based on the χ^2 distribution with $(r-1)(c-1) = 2$ degrees of freedom. Hence the correct answer is: $\chi^2_{0.95}(2) = 5.991$.

7.3 Local elections

|||| Exercise 7.3 Local elections

At the local elections in Denmark in November 2013 the Social Democrats (A) had $p = 29.5\%$ of the votes at the country level. From an early so-called exit poll it was estimated that they would only get 22.7% of the votes. Suppose the exit poll was based on 740 people out of which then 168 people reported having voted for A.

a) At the time of the exit poll the p was of course not known. If the following hypothesis was tested based on the exit poll

$$H_0 : p = 0.295$$
$$H_1 : p \neq 0.295,$$

what test statistic and conclusion would then be obtained with $\alpha = 0.001$?

|||| Solution

The one-proportions test statistic found in Method 7.11 is

$$z_{\text{obs}} = \frac{168 - 740 \cdot 0.295}{\sqrt{740 \cdot 0.295 \cdot (1 - 0.295)}} = -4.05,$$

and the critical values are $\pm z_{0.9995} = \pm 3.291$ (in Python: `stats.norm.ppf(0.9995)`).

So the correct answer is:

Test statistic: -4.05. Conclusion: we reject the null hypothesis, since $-4.05 < -z_{0.9995} = -3.291$.

b) Calculate a 95%-confidence interval for p based on the exit poll.

Solution

We use Method 7.3

$$\hat{p} = \frac{x}{n} = \frac{168}{740} = 0.227, \tag{7-1}$$

and

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.227 \pm 1.96 \cdot \sqrt{\frac{0.227 \cdot 0.773}{740}}.$$

Which we calculate in Python by

```
CI = 0.227 + np.array([-1, 1]) * 1.96 * np.sqrt(0.227 * 0.773 / 740)
print(CI)
```

```
[0.197 0.257]
```

c) Based on a scenario that the proportion voting for particular party is around 30%, how large an exit poll should be taken to achieve a 99% confidence interval having a width of 0.01 in average for this proportion?

Solution

The proportion sample size formula, Method 7.13, using a guess of $p = 0.3$ is

$$0.3 \cdot 0.7 \cdot (z_{0.995}/ME)^2$$

where ME is the marginal error and since the confidence interval is plus/minus the marginal error, we should take $ME = 0.01/2$ and $z_{0.995} = 2.576$ (in Python: `stats.norm.ppf(0.995)`).

So, rounding up to nearest integer, the correct answer is:
$0.3 \cdot 0.7 \cdot (2.576/(0.01/2))^2 \approx 55741$ persons.

7.4 Sugar quality

|||| Exercise 7.4 Sugar quality

A wholesaler needs to find a supplier that delivers sugar in 1 kg bags. From two potential suppliers 50 bags of sugar are received from each. A bag is described as 'defective' if the weight of the filled bag is less than 990 grams. The received bags were all control weighed and 6 defective from supplier A and 12 defective from supplier B were found.

a) If the following hypothesis

$$H_0 : p_A = p_B,$$
$$H_1 : p_A \neq p_B.$$

is tested on a significance level of 5%, what is the p-value and conclusion?

|||| Solution

With

$$\hat{p}_A = \frac{6}{50} = 0.12, \quad \hat{p}_B = \frac{12}{50} = 0.24,$$

and the common

$$\hat{p} = \frac{6+12}{50+50} = 0.18.$$

Using the z-test for comparing two proportions Method 7.18 the hypothesis can be tested by

$$z_{obs} = \frac{\hat{p}_B - \hat{p}_A}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_B} + \frac{1}{n_B}\right)}} = \frac{0.24 - 0.12}{\sqrt{0.18 \cdot 0.82 \cdot (2/50)}} = 1.5617.$$

Finding the probability of observing this or more extreme that for a standard normal (in Python: `1 - stats.norm.cdf(1.5617)`) is

$$P(Z > 1.5617) = 0.0592,$$

thus the p-value becomes 0.118. This leads to the conclusion that the null hypothesis cannot be rejected, since the p-value is above the significance level p-value $= 0.118 > 0.05 = \alpha$.

Similarly one could have performed a 2-by-2 table χ^2-test that would give $\chi^2_{\text{obs}} = Z^2 = 2.439$ and the p-value found using the $\chi^2(1)$-distribution, or simply in Python run any of the two calls:

```
x1,x2 = 6,12
n1,n2 = 50,50
chi2_obs, p_val, (obs,exp) = smprop.proportions_chisquare([x1,x2],[n1,n2])
print(chi2_obs)


2.439024390243903


print(p_val)


0.11834981273562842


print(exp) # Expected values for table


[[ 9.000 41.000]
 [ 9.000 41.000]]


# Or z-test
z_obs, p_val = smprop.proportions_ztest([x1,x2],[n1,n2])
print(z_obs)


-1.5617376188860606


print(p_val)


0.11834981273562835
```

Notice that these functions assume as default the nulhypothesis $H0 : p1 = p2$. But be careful, `smprop.proportions_ztest()` assumes as default that the variance is from the sample proportion, whereas we normally want it based on the proportion under the nullhypothesis.

b) A supplier has delivered 200 bags, of which 36 were defective. A 99%

confidence interval for p the proportion of defective bags for this supplier is:

|||| Solution

The large sample proportion confidence interval formula in Method 7.3 is used

$$0.18 \pm z_{0.995} \cdot \sqrt{\frac{0.18 \cdot 0.82}{200}},$$

thus (in Python: `stats.norm.ppf(0.995)`)

$$0.18 \pm 2.576 \cdot \sqrt{\frac{0.18 \cdot 0.82}{200}}$$

So the answer becomes:

```
CI = 0.18 + np.array([-1, 1]) * 2.576 * np.sqrt(0.18 * 0.82 / 200)
print(CI)


[0.110 0.250]


# Or directly
smprop.proportion_confint(36, 200, alpha=0.01)


(0.11002462081879329, 0.2499753791812067)
```

c) Based on the scenario, that the proportion of defective bags for a new supplier is about 20%, a new study was planned with the aim of obtaining an average width, B, of a 95% confidence interval. The Analysis Department achieved the result that one should examine 1537 bags, but had forgotten to specify which value for the width B, they had used. What was the value used for B?

Solution

Method 7.13 holds the relevant sample size formular

$$n = 0.2 \cdot 0.8 \cdot \left[\frac{z_{0.975}}{ME}\right]^2$$

which when solved for ME becomes (and plugging in $n = 1537$)

$$ME = \sqrt{0.2 \cdot 0.8 \cdot \frac{z_{0.975}^2}{1537}} = 0.020$$

And since the width of the confidence interval is twice the margin of error, the answer is: $B = 0.040$.

7.5 Physical training

|||| Exercise 7.5 Physical training

A company wants to investigate whether the employees' physical training condition will affect their success in the job. 200 employees were tested and the following count data were found:

	Physical training condition		
	Below average	Average	Above average
Bad job succes	11	27	15
Average job succes	14	40	30
Good job succes	5	23	35

The hypothesis of independence between job success and physical training condition is to be tested by the use of the for this setup usual χ^2-test.

a) What is the expected number of individuals with above average training condition and good job success under H_0 (i.e. if H_0 is assumed to be true)?

|||| Solution

The expected number under the null hypothesis for each cell is found as

$$\text{"column total"} \cdot \frac{\text{"row total"}}{\text{"total"}},$$

for table cell (3,3), which is asked about, so the answer is

$$e_{33} = 80 \cdot \frac{63}{200} = 25.2.$$

b) For the calculation of the relevant χ^2-test statistic, identify the following two numbers:
   - A: the number of contributions to the test statistic
   - B: the contribution to the statistic from table cell (1,1)

|||| Solution

Since the contingency table is a 3-by-3 table the number of contributions to the test statistic is 9 (one for each cell), and the (1,1) contribution is

$$\frac{(o_{11} - e_{11})^2}{e_{11}},$$

where $o_{11} = 11$ and

$$e_{11} = \frac{30 \cdot 53}{200} = 7.95.$$

Hence

$$\frac{(o_{11} - e_{11})^2}{e_{11}} = \frac{(11 - 7.95)^2}{7.95} = 1.170.$$

So the answer becomes: A is 9 and B is 1.170.

c) The total χ^2-test statistic is 10.985, so the p-value and the conclusion will be (both must be valid):

|||| Solution

The p-value is found by the $\chi^2(4)$-distribution (degrees of freedom is $(r-1)\cdot(c-1) = (3-1)\cdot(3-1)$

$$P(\chi^2 > 10.985) = 0.027,$$

which can be found in Python by: `1-stats.chi2.cdf(10.985,4)`.

So the answer becomes:

The p-value $= 0.027 < 0.05 = \alpha$ and therefore H_0 is rejected on a 5% level, and thus on this significance level there is a significant dependence between job success and physical training condition.