```
import numpy as np
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.stats.proportion as smprop
```

Chapter 7

Inference for Proportions

Contents

7.1 Categorical data

Until now we have mainly focused on continuous outcomes such as the height of students. In many applications the outcome that we wish to study is categorical (7.1). For example, one could want to study the *proportion* of defective components in a sample, hence the outcome has two categories: "defect" and "non-defect". Another example could be a study of the caffeine consumption among different groups of university students, where the consumption could be measured via a questionnaire in levels: none, 1-3 cups per day, more than 3 cups per day. Hence the categorical variable describing the outcome has three categories.

In both examples the key is to describe the *proportion* of outcomes in each category.

Remark 7.1

A variable is categorical if each outcome belongs to a category, which is one of a set of categories.

7.2 Estimation of single proportions

We want to be able to find estimates of the population category proportions (i.e. the "true" proportions). We sometimes refer to such a proportion as the probability of belonging to the category. This is simply because the probability that a randomly sampled observation from the population belongs to the category, is the proportion of the category in the population.

Example 7.2

In a survey in the US in 2000, 1154 people answered the question whether they would be willing to pay more for petrol to help the environment. Of the 1154 participants 518 answered that they would be willing to do so.

Our best estimate of the proportion of people willing to pay more (p) is the observed proportion of positive answers

$$\hat{p} = \frac{\text{"Number of positive answers"}}{\text{"Total number of participants"}} = \frac{518}{1154} = 0.4489.$$

This means that our best estimate of the proportion of people willing to pay more for petrol to help the environment is 44.89%.

In the above example we can think of $n = 1154$ trials, where we each time have a binary outcome (yes or no), occurring with the unknown probability p. The random variable X counts the number of times we get a yes to the question, hence X follows a binomial distribution $B(n, p)$ with the probability of observing x successes given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{7-1}$$

As mentioned in Example 7.2, our best estimate of the unknown p is the proportion

$$\hat{p} = \frac{x}{n}, \quad \hat{p} \in [0, 1]. \tag{7-2}$$

From Chapter 2 we know that if $X \sim B(n, p)$, then

$$\mathrm{E}(X) = np, \tag{7-3}$$
$$\mathrm{V}(X) = np(1-p). \tag{7-4}$$

This means that

$$\mathrm{E}(\hat{p}) = \mathrm{E}\left(\frac{X}{n}\right) = \frac{np}{n} = p, \tag{7-5}$$
$$\mathrm{V}(\hat{p}) = \mathrm{V}\left(\frac{X}{n}\right) = \frac{1}{n^2}\mathrm{V}(X) = \frac{p(1-p)}{n}. \tag{7-6}$$

From Equation (7-5) we see that $\hat{p}$ is an unbiased estimator of the unknown p and from Equation (7-6) that the standard error (the (sampling) standard deviation) of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. It is important to quantify the uncertainty of the calculated estimate using confidence intervals. For large samples, the Central Limit Theorem gives us that the sample proportion $\hat{p}$ is well approximated by a normal distribution, and thus a $(1-\alpha)100\%$ confidence interval for the population proportion p is

$$\hat{p} \pm z_{1-\alpha/2}\, \sigma_{\hat{p}}. \tag{7-7}$$

However, $\sigma_{\hat{p}}$ depends on the unknown p, which we do not know. In practice we will have to estimate the standard error by substituting the unknown p by the estimate $\hat{p}$.

|||| Method 7.3 Proportion estimate and confidence interval

The best estimate of the probability p of belonging to a category (the population proportion) is the sample proportion

$$\hat{p} = \frac{x}{n}, \tag{7-8}$$

where x is the number of observations in the category and n is the total number of observations.

A large sample $(1-\alpha)100\%$ confidence interval for p is given as

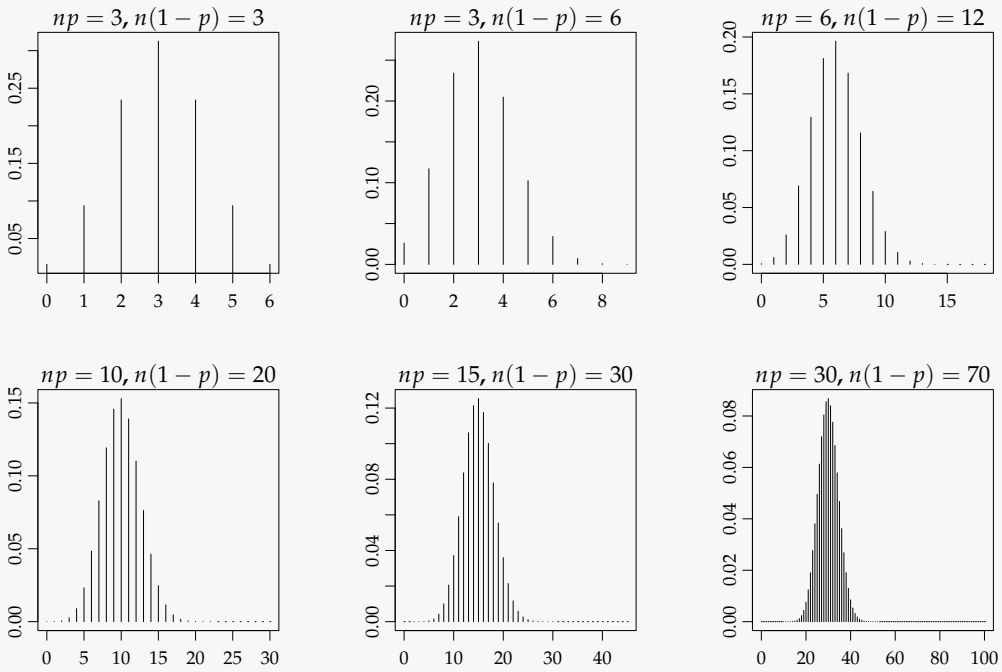
$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{7-9}$$

|||| Remark 7.4

As a rule of thumb the normal distribution is a good approximation of the binomial distribution if np and $n(1-p)$ are both greater than 15.

|||| Example 7.5

In the figure below we have some examples of binomial distributions. When we reach a size where $np \geq 15$ and $n(1-p) \geq 15$ it seems reasonable that the bell-shaped normal distribution will be a good approximation.



Example 7.6

If we return to the survey in Example 7.2, we can now calculate the 95% confidence interval for the probability (i.e. the proportion willing to pay more for petrol to help the environment).

We found the estimate of p by the observed proportion to $\hat{p} = \frac{518}{1154} = 0.45$. The standard error of the proportion estimate is

$$\hat{\sigma}_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{0.45 \cdot 0.55/1154} = 0.0146.$$

Since we have $n\hat{p} = 1154 \cdot 0.45 = 519.3$ and $n(1-\hat{p}) = 1154 \cdot 0.55 = 634.7$, both greater than 15, we can use the expression from Method 7.3 to get the 95% confidence interval

$$\hat{p} \pm 1.96 \cdot \hat{\sigma}_{\hat{p}} = 0.45 \pm 1.96 \cdot 0.0146 = [0.42, 0.48].$$

From this we can now conclude that our best estimate of the proportion willing to pay more for petrol to protect the environment is 0.45, and that the true proportion with 95% certainty is between 0.42 and 0.48. We see that 0.5 is not included in the confidence interval, hence we can conclude that the proportion willing to pay more for petrol is less than 0.5 (using the usual $\alpha = 0.05$ significance level). We will cover hypothesis testing for proportions more formally below.

|||| Remark 7.7 What about small samples then?

There exist several ways of expressing a valid confidence interval for p in small sample cases, that is, when either $np \leq 15$ or $n(1-p) \leq 15$. We mention three of these here - only for the last one we give the explicit formula:

Continuity correction

The so-called *continuity correction* is a general approach to making the best approximation of discrete probabilities (in this case the binomial probabilities) using a continuous distribution, (in this case the normal distribution). We do not give any details here.

Exact intervals

Probably the most well known of such small sample ways of obtaining a valid confidence interval for a proportion is the so-called *exact* method based on actual binomial probabilities rather than a normal approximation. It is not possible to give a simple formula for these confidence limits, and we will not explain the details here, but simply note that they can be obtained by the Python function `stats.binomtest`. These will be valid no matter the size of n and p.

"Plus 2"-approach

Finally, a simple approach to a good small sample confidence interval for a proportion, will be to us the simple formula given above in Method 7.3, but applied to $\tilde{x} = x + 2$ and $\tilde{n} = n + 4$.

|||| Remark 7.8 Confidence intervals for single proportions in Python

In Python we can either use the function `smprop.proportions_ztest` or `stats.binomtest` to find the confidence interval of a single proportion (and some hypothesis testing information to be described below).

The `stats.binomtest` function uses the exact approach. The `smprop.proportions_ztest` does not use continuity correction, but assumes normality.
Therefore: none of these intervals calculated by Python coincides exactly with the formula given in Method 7.3, neither applied to x and n nor applied to $\tilde{x} = x + 2$ and $\tilde{n} = n + 4$. And vice versa: the exact computational details of the different intervals calculated by Python are not given in the text here.

7.2.1 Testing hypotheses

Hypothesis testing for a single proportion (or probability) p is presented in this section.

The first step is to formulate the null hypothesis and the alternative as well as choosing the level of significance α. The null hypothesis for a proportion has the form

$$H_0 : p = p_0 \tag{7-10}$$

where p_0 is a chosen value between 0 and 1. In Example 7.2, we could be interested in testing whether half of the population, from which the sample was taken, would be willing to pay more for petrol, hence $p_0 = 0.5$.

The alternative hypothesis is the two-sided alternative

$$H_1 : p \neq p_0. \tag{7-11}$$

|||| Remark 7.9

As for the t-tests presented in Chapter 3, we can also have one-sided tests for proportions, i.e. the "less than" alternative

$$H_0 : p \geq p_0 \tag{7-12}$$
$$H_1 : p < p_0, \tag{7-13}$$

and the "greater than" alternative

$$H_0 : p \leq p_0 \tag{7-14}$$
$$H_1 : p > p_0, \tag{7-15}$$

however these are not included further in the material, see the discussion in Section 3.1.7 (from page 145 in the book), which applies similarly here.

The next step is to calculate a test statistic as a measure of how well our data fits the null hypothesis. The test statistic measures how far our estimate $\hat{p}$ is from the value p_0 relative to the uncertainty – under the scenario that H_0 is true.

So, under H_0 the true proportion is p_0 and the standard error is $\sqrt{p_0(1-p_0)/n}$, thus to measure the distance between $\hat{p}$ and p_0 in standard deviations we calculate the test statistic

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}. \tag{7-16}$$

When H_0 is true, the test statistic seen as a random variable is

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}, \tag{7-17}$$

and follows approximately a standard normal distribution $Z \sim N(0,1)$, when n is large enough:

|||| Theorem 7.10

In the large sample case the random variable Z follows approximately a standard normal distribution

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0,1), \tag{7-18}$$

when the null hypothesis is true. As a rule of thumb, the result will be valid when both $np_0 > 15$ and $n(1-p_0) > 15$.

We can use this to make the obvious explicit method for the hypothesis test:

Method 7.11 One sample proportion hypothesis test

1. Compute the test statistic using Equation (7-16)

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

2. Compute evidence against the *null hypothesis*

$$H_0 : p = p_0, \tag{7-19}$$

vs. the *the alternative hypothesis*

$$H_1 : p \neq p_0, \tag{7-20}$$

by the

$$p\text{-value} = 2 \cdot P(Z > |z_{\text{obs}}|). \tag{7-21}$$

where the standard normal distribution $Z \sim N(0, 1^2)$ is used

3. If the p-value $< \alpha$ we reject H_0, otherwise we accept H_0,

or

The rejection/acceptance conclusion can equivalently be based on the critical value(s) $\pm z_{1-\alpha/2}$:
if $|z_{\text{obs}}| > z_{1-\alpha/2}$ we reject H_0, otherwise we accept H_0

Example 7.12

To conclude Example 7.2 we want to test the null hypothesis

$$H_0 : p = 0.5,$$

against the alternative

$$H_1 : p \neq 0.5.$$

We have chosen $\alpha = 0.05$, hence the critical value is the 0.975 quantile in the standard normal distribution $z_{1-\alpha/2} = 1.96$. Thus we get the observed value of the test statistic by

$$z_{\text{obs}} = \frac{518 - 577}{\sqrt{1154 \cdot 0.5 \cdot (1 - 0.5)}} = -3.47.$$

Since $z = -3.47 < -1.96$ then we reject H_0. The p-value is calculated as the probability of observing z_{obs} or more extreme under the null hypothesis

$$2 \cdot P(Z \geq 3.47) = 0.0005.$$

We can get this directly using Python:

```
# Testing the probability = 0.5 with a two-sided alternative
# We have observed 518 out of 1154
# Do it without continuity corrections
z_obs,p_value = smprop.proportions_ztest(518, 1154, value=0.5,
prop_var=0.5)
print(z_obs)

-3.473594375515837


print(p_value)

0.0005135367279608199
```

Note that the results are exactly the same as when calculated by hand even though the test statistic used is actually $Z^2 \sim \chi^2$ with one degree of freedom, since this is the same as saying $Z \sim N(0,1)$. This is explained in detail later in the chapter.

7.2.2 Sample size determination

Before conducting a study, it is important to consider the sample size needed to achieve a wanted precision. In the case with a single probability to estimate, we see that the error we make when using the estimator $\hat{p} = \frac{x}{n}$ is given by $\left|\frac{x}{n} - p\right|$. Using the normal approximation (see Theorem 7.3) we can conclude that the error will be bounded by

$$\left|\frac{x}{n} - p\right| < z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \tag{7-22}$$

with probability $1 - \alpha$. Thus the *Margin of Error* (ME) of the estimate becomes

$$ME = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}. \tag{7-23}$$

Similar to the method given for quantitative data in Method 3.63, we can use

Equation (7-23) to determine the needed sample size in a single proportions setup. Solving for n we get:

Method 7.13 Sample size formula for the CI of a proportion

Given some "guess" (scenario) of the size of the unknown p, and given some requirement to the ME-value (required expected precision) the necessary sample size is then

$$n = p(1-p)\left(\frac{z_{1-\alpha/2}}{ME}\right)^2. \tag{7-24}$$

If p is unknown, a worst case scenario with $p = 1/2$ is applied and necessary sample size is

$$n = \frac{1}{4}\left(\frac{z_{1-\alpha/2}}{ME}\right)^2. \tag{7-25}$$

The expression in Equation (7-25) for n when no information about p is available is due to the fact that $p(1-p)$ is largest for $p = 1/2$, so the required sample size will be largest when $p = 1/2$.

Method 7.13 can be used to calculate the sample size for a given choice of ME.

7.3 Comparing proportions in two populations

For categorical variables we sometimes want to compare the proportions in two populations (groups). Let p_1 denote the proportion in group 1 and p_2 the proportion in group 2. We will compare the groups by looking at the difference in proportions $p_1 - p_2$, which is estimated by $\hat{p}_1 - \hat{p}_2$.

Example 7.14

In a study in the US (1975) the relation between intake of contraceptive pills (birth control pills) and the risk of blood clot in the heart was investigated. The following data were collected from a participating hospital:

	Contraceptive pill	No pill
Blood clot	23	35
No blood clot	34	132
Total	57	167

We have a binary outcome blood clot (yes or no) and two groups (pill or no pill). As in Section 7.2 we find that the best estimates of the unknown probabilities are the observed proportions

$$\hat{p}_1 = \frac{\text{"Number of blood clots in the pill group"}}{\text{"Number of women in the pill group"}} = \frac{23}{57} = 0.4035, \tag{7-26}$$

$$\hat{p}_2 = \frac{\text{"Number of blood clots in the no pill group"}}{\text{"Number of women in the no pill group"}} = \frac{35}{167} = 0.2096. \tag{7-27}$$

The difference in probabilities is estimated to be

$$\hat{p}_1 - \hat{p}_2 = 0.4035 - 0.2096 = 0.1939. \tag{7-28}$$

Thus the observed probability of getting a blood clot, was 0.1939 higher in the contraceptive pill group than in the no pill group.

We have the estimate $\hat{p}_1 - \hat{p}_2$ of the difference in probabilities $p_1 - p_2$ and the uncertainty of this estimate can be calculated by:

|||| Method 7.15

An estimate of the standard error of the estimator $\hat{p}_1 - \hat{p}_2$ is

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}. \tag{7-29}$$

The $(1-\alpha)100\%$ confidence interval for the difference $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}_1 - \hat{p}_2}. \tag{7-30}$$

This confidence interval requires independent random samples for the two groups and large enough sample sizes n_1 and n_2. A rule of thumb is that $n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i = 1, 2$, must be satisfied.

Remark 7.16

The standard error in Method 7.15 can be calculated by

$$V(\hat{p}_1 - \hat{p}_2) = V(\hat{p}_1) + V(\hat{p}_2) = \hat{\sigma}^2_{\hat{p}_1} + \hat{\sigma}^2_{\hat{p}_2}, \tag{7-31}$$

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{V(\hat{p}_1 - \hat{p}_2)} = \sqrt{\hat{\sigma}^2_{\hat{p}_1} + \hat{\sigma}^2_{\hat{p}_2}}. \tag{7-32}$$

Notice, that the standard errors are added (before the square root) such that the standard error of the difference is larger than the standard error for the observed proportions alone. Therefore in practice the estimate of the difference $\hat{p}_1 - \hat{p}_2$ will often be further from the true difference $p_1 - p_2$ than $\hat{p}_1$ will be from p_1 or $\hat{p}_2$ will be from p_2.

Example 7.17

Returning to Example 7.14 where we found the estimated difference in probability to be

$$\hat{p}_1 - \hat{p}_2 = 0.4035 - 0.2096 = 0.1939. \tag{7-33}$$

The estimated standard error of the estimate is

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{0.4035(1 - 0.4035)}{57} + \frac{0.2096(1 - 0.2096)}{167}} = 0.0722. \tag{7-34}$$

A 99% confidence interval for this difference is then

$$(\hat{p}_1 - \hat{p}_2) \pm z_{0.995} \cdot \hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = 0.1939 \pm 2.5758 \cdot 0.0722 = [0.0079, 0.3799]. \tag{7-35}$$

Hence our best estimate of the difference is 0.19 and with very high confidence the true difference is between 0.008 and 0.38.

We find that 0 is not included in the confidence interval, so 0 is not a plausible value for the difference $p_1 - p_2$. The values in the confidence interval are all positive and therefore we can conclude that $(p_1 - p_2) > 0$, that is $p_1 > p_2$, i.e. the probability of blood clot is larger in the contraceptive pill group than in the no pill group.

We can also compare the two proportions p_1 and p_2 using a hypothesis test. As in Method 7.11, there are four steps when we want to carry out the test. The first step is to formulate the hypothesis and the alternative.

The null hypothesis is $H_0 : p_1 = p_2$ and we will denote the common proportion p, and choose a two-sided alternative $H_1 : p_1 \neq p_2$.

In the second step we calculate a test statistic measuring how far $\hat{p}_1 - \hat{p}_2$ falls from 0, which is the value of $p_1 - p_2$ under H_0.

Under H_0, we only have one proportion p (since $p_1 = p_2 = p$). The best estimator for this common proportion is the overall observed proportion

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}. \tag{7-36}$$

When the two sample sizes n_1 and n_2 are similar, this pooled estimate of the overall proportion will be approximately half way between $\hat{p}_1$ and $\hat{p}_2$, but otherwise the pooled estimate will be closest to the estimate from the largest sample size.

|||| Method 7.18 Two sample proportions hypothesis test

The two-sample hypothesis test for comparing two proportions is given by the following procedure:

1. Compute, with $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$, the test statistic

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{7-37}$$

2. Compute evidence against the *null hypothesis*

$$H_0 : p_1 = p_2, \tag{7-38}$$

vs. the *the alternative hypothesis*

$$H_1 : p_1 \neq p_2, \tag{7-39}$$

by the

$$p\text{-value} = 2 \cdot P(Z > |z_{\text{obs}}|). \tag{7-40}$$

where the standard normal distribution $Z \sim N(0, 1^2)$ is used

3. If the p-value $< \alpha$ we reject H_0, otherwise we accept H_0,

or

The rejection/acceptance conclusion can equivalently be based on the critical value(s) $\pm z_{1-\alpha/2}$:
if $|z_{\text{obs}}| > z_{1-\alpha/2}$ we reject H_0, otherwise we accept H_0

Example 7.19

In Example 7.17 we tested whether the probability of blood clot is the same for the group taking the pills as for the group without pills using the CI. The null hypothesis and alternative are

$$H_0 : p_1 = p_2,$$
$$H_1 : p_1 \neq p_2.$$

This time we will test on a 1% significance level ($\alpha = 0.01$).

The pooled estimate of the probability of blood clot under H_0 is

$$\hat{p} = \frac{23 + 35}{57 + 167} = 0.259,$$

which is closest to the estimate from the largest group $\hat{p}_2 = 0.210$.

According to Method 7.15 the test statistic is

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0.194}{\sqrt{0.259(1-0.259)(\frac{1}{57} + \frac{1}{167})}} = 2.89.$$

The p-value is calculated by looking up z_{obs} in a standard normal distribution (i.e. $N(0,1)$)

$$2P(Z \geq 2.89) = 0.0039 < 0.01.$$

As the p-value is less than 0.01 we can reject the null hypothesis of equal probabilities in the two groups.

Instead of doing all the calculations in steps, we can use the function `smprop.proportions_ztest()` to test the hypothesis.

```
# Testing that the probabilities for the two groups are equal
z_obs, p_val = smprop.proportions_ztest([23, 35], [57, 167], value=0,
prop_var=0)
print(z_obs)

2.8859712586466184


print(p_val)

0.003902077897925702
```

7.4 Comparing several proportions

In the previous Section 7.3, we were interested in comparing proportions from two groups. In some cases we might be interested in proportions from two or more groups, or in other words if several binomial distributions share the same parameter p. The data can be setup in a $2 \times c$ table, where "Success" is the response we are studying (e.g. a blood clot occurs) and "Failure" is when the response does not occur (e.g. no blood clot).

	Group 1	Group 2	...	Group c	*Total*
Success	x_1	x_2	...	x_c	x
Failure	$n_1 - x_1$	$n_2 - x_2$	...	$n_c - x_c$	$n - x$
Total	n_1	n_2	...	n_c	n

We are then interested in testing the null hypothesis

$$H_0 : p_1 = p_2 = \ldots = p_c = p \tag{7-41}$$

against the alternative hypothesis: that the probabilities are not equal (or more precisely: that that at least one of the probabilities is different from the others).

Under H_0 the best estimator for the common p is the overall observed proportion

$$\hat{p} = \frac{x}{n}. \tag{7-42}$$

To test the null hypothesis, we need to measure how likely it is to obtain the observed data (or more extreme) under the null hypothesis. So, under the scenario that the null hypothesis is true, we can calculate the expected number of successes in the jth group as

$$e_{1j} = n_j \cdot \hat{p} = n_j \cdot \frac{x}{n}, \tag{7-43}$$

and the expected number of failures is

$$e_{2j} = n_j \cdot (1 - \hat{p}) = n_j \cdot \frac{(n - x)}{n}. \tag{7-44}$$

Notice, that the expected number for a cell is calculated by multiplying the row and column totals for the row and column, where the cell belongs and then dividing by the grand total n.

|||| Method 7.20 The multi-sample proportions χ^2-test

The hypothesis

$$H_0 : p_1 = p_2 = \ldots = p_c = p, \tag{7-45}$$

can be tested using the test statistic

$$\chi^2_{\text{obs}} = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}, \tag{7-46}$$

where o_{ij} is the observed number in cell (i, j) and e_{ij} is the expected number in cell (i, j).

The test statistic χ^2_{obs} should be compared with the χ^2-distribution with $c - 1$ degrees of freedom.

The χ^2-distribution is approximately the sampling distribution of the statistics under the null hypothesis. The rule of thumb is that it is valid when all the computed expected values are at least 5: $e_{ij} \geq 5$.

The test statistic in Method 7.20 measures the distance between the observed number in a cell and what we would expect if the null hypothesis is true. If the hypothesis is true then χ^2 has a relatively small value, as most of the cell counts will be close to the expected values. If H_0 is false, some of the observed values will be far from the expected resulting in a larger χ^2.

|||| Example 7.21

Returning to Example 7.19 we can consider a 2×2 table as a case of a $2 \times c$ table. We can organize our table with "Success" and "Failure" in the rows and groups as the columns.

	Contraceptive pill	No pill	*Total*
Blood clot	23	35	58
No blood clot	34	132	166
Total	57	167	224

Here $x = 23 + 35 = 58$ and $n = 224$

For each cell we can now calculate the expected number if H_0 is true. For the pill and blood clot cell we get

$$e_{1,1} = \frac{58 \cdot 57}{224} = 14.76, \tag{7-47}$$

but we only observed 23 cases.

For the no pill and blood clot cell we get

$$e_{1,2} = \frac{58 \cdot 167}{224} = 43.24, \tag{7-48}$$

which is more than the observed 35 cases.

In the following table we have both the observed and expected values.

	Birth control pill	No birth control pill	*Total*
Blood clot	$o_{11} = 23$ $e_{11} = 14.76$	$o_{12} = 35$ $e_{12} = 43.24$	$x = 58$
No blood clot	$o_{21} = 34$ $e_{21} = 42.24$	$o_{22} = 132$ $o_{22} = 123.8$	$(n - x) = 166$
Total	$n_1 = 57$	$n_2 = 167$	$n = 224$

The observed χ^2 test statistic can be calculated

$$\chi^2_{\text{obs}} = \frac{(23 - 14.76)^2}{14.76} + \frac{(35 - 43.24)^2}{43.24} + \frac{(34 - 42.24)^2}{42.24} + \frac{(132 - 123.8)^2}{123.8} = 8.33. \tag{7-49}$$

We then find the p-value, by calculating how likely it is to get 8.33 or more extreme if the null hypothesis is true, using the χ^2 distribution with $c - 1 = 2 - 1 = 1$ degrees of freedom

$$p\text{-value} = P(\chi^2 \geq 8.33) = 0.0039, \tag{7-50}$$

which is exactly the same as the result in Example 7.14. Do the same with the `stats.chi2_contingency()` function in Python:

```
# Reading the data into Python
pill_study = np.array([[23, 35], [34, 132]])
# Using Pandas
pill_study = pd.DataFrame(pill_study, index=['Blood Clot', 'No Clot'],
columns=['Pill', 'No pill'])
print(pill_study)
```

```
            Pill  No pill
Blood Clot    23       35
No Clot       34      132
```

```
# Chi^2 test for testing that the distribution for the two groups are
equal
chi2, p_val, dof, expected = stats.chi2_contingency(pill_study,
correction=False)
# Test Statistic
print(chi2)
```

```
8.328830105734347
```

```
# P value
print(p_val)
```

```
0.0039020778979257016
```

```
# Degrees of freedom
print(dof)
```

```
1
```

```
# Expected frequencies under the null hypothesis
# Output will not be pandas DataFrame, but we can use pandas to display
it nicely
print(pd.DataFrame(expected, index=['Blood Clot', 'No Clot'],
columns=['Pill', 'No pill']))
```

```
                Pill     No pill
Blood Clot  14.758929   43.241071
No Clot     42.241071  123.758929
```

In Section 7.3 we presented a z-test for the hypothesis $H_0 : p_1 = p_2$, where

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

and in this section we have just seen a χ^2 test that can also be used for 2×2 tables. Using some algebra it turns out that the two tests are equivalent

$$\chi^2_{\text{obs}} = z^2_{\text{obs}}, \tag{7-51}$$

and they give exactly the same p-value for testing $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$.

7.5 Analysis of Contingency Tables

Until now we have been looking at $2 \times c$ tables, but we can also have a more general setup with $r \times c$ tables that arise when two categorical variables are cross-tabulated. Such tables usually arise from two kinds of studies. First, we could have samples from several groups (as in Section 7.4), but allowing for more than two outcome categories. An example of this could be an opinion poll, where three samples were taken at different time points by asking randomly selected people whether they supported either: Candidate 1, Candidate 2 or were undecided. Here we want to compare the distribution of votes for the three groups (i.e. over time).

The other setup giving rise to an $r \times c$ table is when we have samples with two paired categorical variables with same categories (i.e. both variables are measured on each observational unit). This might happen if we had a sample of students and categorized them equivalently according to their results in English and mathematics (e.g. good, medium, poor). These tables are also called contingency tables.

The main difference between the two setups is: in the first setup the column totals are the size of each sample (i.e. fixed to the sample sizes), whereas in the second setup the column totals are not fixed (i.e. they count outcomes and the grand total is fixed to the sample size). However, it turns out that both setups are analysed in the same way.

7.5.1 Comparing several groups

In the situation comparing several groups, the hypothesis is that the distribution is the same in each group

$$H_0 : p_{i1} = p_{i2} = \ldots = p_{ic} = p_i, \text{ for all rows } i = 1, 2, \ldots, r. \tag{7-52}$$

So the hypothesis is that the probability of obtaining an outcome in a row category does not depend on the given column.

As in Section 7.4 we need to calculate the expected number in each cell under H_0

$$e_{ij} = \text{"}j\text{th column total"} \cdot \frac{\text{"}i\text{th row total"}}{\text{"grand total"}} = n_j \cdot \frac{x_i}{n}. \tag{7-53}$$

|||| Method 7.22 The $r \times c$ frequency table χ^2-test

For an $r \times c$ table the hypothesis

$$H_0 : p_{i1} = p_{i2} = \ldots = p_{ic} = p_i, \text{ for all rows } i = 1, 2, \ldots, r, \tag{7-54}$$

is tested using the test statistic

$$\chi^2_{\text{obs}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}. \tag{7-55}$$

where o_{ij} is the observed number in cell (i, j) and e_{ij} is the expected number in cell (i, j). This test statistic should be compared with the χ^2-distribution with $(r-1)(c-1)$ degrees of freedom and the hypothesis is rejected at significance level α if

$$\chi^2_{\text{obs}} > \chi^2_{1-\alpha}\big((r-1)(c-1)\big). \tag{7-56}$$

From Method 7.22, we see that we use the same test statistic as for $2 \times c$ tables measuring the distance between the observed and expected cell counts. The degrees of freedom $(r-1)(c-1)$ occurs because only $(r-1)(c-1)$ of the expected values e_{ij} need to be calculated – the rest can be found by subtraction from the relevant row or column totals.

Example 7.23

An opinion poll has been made at three time points (4 weeks, 2 weeks and 1 week before the election) each time 200 participants was asked who they would vote for: Candidate 1, Candidate 2 or were undecided. The following data was obtained:

	4 weeks before	2 weeks before	1 week before	*Row total*
Candidate 1	79	91	93	263
Candidate 2	84	66	60	210
Undecided	37	43	47	127
Column total	200	200	200	600

Note, that in this poll example the sample sizes are equal (i.e. $n_1 = n_2 = n_3 = 200$), however that is not a requirement.

We want to test the hypothesis that the votes are equally distributed in each of the three polls

$$H_0 : p_{i1} = p_{i2} = p_{i3}, \text{ for all rows } i = 1, 2, 3. \tag{7-57}$$

The expected number of votes under H_0 is calculated for the "Candidate 2" - "2 weeks before" cell of the table

$$e_{22} = \text{"2'nd column total"} \cdot \frac{\text{"2'nd row total"}}{\text{"grand total"}} = \frac{210 \cdot 200}{600} = 70. \tag{7-58}$$

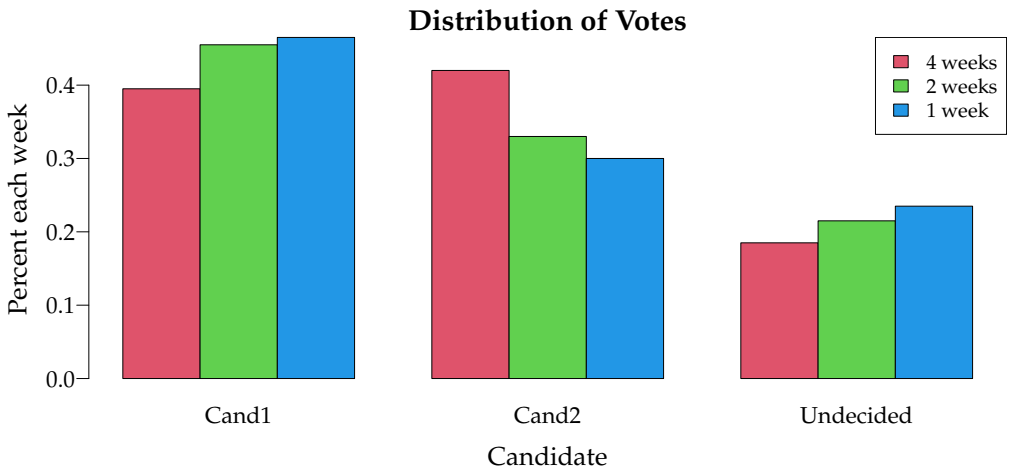
Continuing in the same way we can calculate all the expected cell counts:

	4 weeks before	2 weeks before	1 week before
Candidate 1	$o_{11} = 79$ $e_{11} = 87.67$	$o_{12} = 91$ $e_{12} = 87.67$	$o_{13} = 93$ $e_{13} = 87.67$
Candidate 2	$o_{21} = 84$ $e_{21} = 70.00$	$o_{22} = 66$ $e_{22} = 70.00$	$o_{23} = 60$ $e_{23} = 70.00$
Undecided	$o_{31} = 37$ $e_{31} = 42.33$	$o_{32} = 43$ $e_{32} = 42.33$	$o_{33} = 47$ $e_{33} = 42.33$

Looking at this table, it seems that 4 weeks before, Candidate 1 has less votes than expected while Candidate 2 has more, but we need to test whether these differences are statistically significant.

We can test the hypothesis in Equation (7-52) using a χ^2 test with $(3-1)(3-1) = 4$ degrees of freedom.

However, first we will calculate the observed column percentages and plot them:



From the bar plot it could seem that the support for Candidate 2 decreases as the election approaches, but we need to test whether this is significant. In the following Python code the hypothesis, stating that the distribution at each time point is the same, is tested:

```
# Reading the data into Python
poll = np.array([[79, 91, 93], [84, 66, 60], [37, 43, 47]])
poll = pd.DataFrame(poll, index=['Cand1', 'Cand2', 'Undecided'],
                    columns=['4 weeks', '2 weeks', '1 week'])

# testing same distribution in the three populations
chi2, p_val, dof, expected = stats.chi2_contingency(poll,
correction=False)
# Test statistic
print(chi2)

6.961978041718169


# p-value
print(p_val)

0.1379112060673381


# Degrees of Freedom
print(dof)

4


# Expected frequencies under the null hypothesis
print(pd.DataFrame(expected, index=['Cand1', 'Cand2', 'Undecided'],
                    columns=['4 weeks', '2 weeks', '1 week']))

             4 weeks    2 weeks     1 week
Cand1      87.666667  87.666667  87.666667
Cand2      70.000000  70.000000  70.000000
Undecided  42.333333  42.333333  42.333333
```

From the χ^2 test we get an observed test statistic of 6.96, and we must now calculate how likely it is to obtain this value or more extreme from a χ^2-distribution with 4 degrees of freedom. It leads to a p-value of 0.14, so we accept the null hypothesis and find that there is no evidence showing a change in distribution among the three polls.

7.5.2 Independence between the two categorical variables

When the only fixed value is the grand *total*, then the hypothesis we are interested in concerns independence between the two categorical variables

$$\begin{aligned} &H_0 : \text{"The two variables are independent"}, \\ &H_1 : \text{"The two variables are not independent (they are associated)"}. \end{aligned} \tag{7-59}$$

Using the cell proportions p_{ij} the null hypothesis can be written as:

|||| Theorem 7.24

To test if two categorical variables are independent the null hypothesis

$$H_0 : p_{ij} = p_{i.}p_{.j} \text{ for all } i, j, \tag{7-60}$$

where $p_{i.} = \sum_{j=1}^{c} p_{ij}$ is the proportion of row i and $p_{.j} = \sum_{i=1}^{r} p_{ij}$ is the proportion of column j, is tested.

The p-value for the observed result under this null hypothesis is calculated using the χ^2 test statistic from Method 7.22.

|||| Example 7.25

A group of 400 students have had an English test and a mathematics test. The results of each test a categorized as either bad, average or good.

English	Mathematics			
	Bad	Average	Good	*Row total*
Bad	23	60	29	112
Average	28	79	60	167
Good	9	49	63	121
Column total	60	188	152	400

We want to test the hypothesis of independence between results in English and mathematics. First we read the data into Python and calculate proportions and totals:

```
# Reading the data into Python
results = np.array([[23, 60, 29], [28, 79, 60], [9, 49, 63]])
results_df = pd.DataFrame(results, index=['EngBad', 'EngAve',
'EngGood'],
                          columns=['MathBad', 'MathAve',
'MathGood'])
```

```
# Percentages
prop = results_df/results_df.sum().sum()
print(prop)

          MathBad  MathAve  MathGood
EngBad     0.0575   0.1500    0.0725
EngAve     0.0700   0.1975    0.1500
EngGood    0.0225   0.1225    0.1575

# Row totals
print(results_df.sum(axis=1))

EngBad     112
EngAve     167
EngGood    121
dtype: int64

# Column totals
print(results_df.sum(axis=0))

MathBad      60
MathAve     188
MathGood    152
dtype: int64
```

We want to calculate the expected cell count if H_0 is true. Consider the events "good English result" and "good mathematics result" corresponding to cell $(3,3)$. Under the hypothesis of independence, we have

$$p_{33} = P(\text{"Good English and Good Maths"}) = P(\text{"Good English"}) \cdot P(\text{"Good Maths"}) \tag{7-61}$$

From the calculated row and column totals, we would estimate

$$\hat{p}_{33} = \left(\frac{121}{400}\right) \cdot \left(\frac{152}{400}\right), \tag{7-62}$$

and out of 400 students we would expect

$$e_{33} = 400 \cdot \hat{p}_{33} = 400 \cdot \left(\frac{121}{400}\right) \cdot \left(\frac{152}{400}\right) = 121 \cdot \frac{152}{400} = 45.98. \qquad (7\text{-}63)$$

The method of calculating the expected cell counts is exactly as before. For the "Good English and Good Mathematics" cell the expected value is less than the observed 63. Continuing in this way, we can calculate all the expected cell counts:

English	Mathematics		
	Bad	Average	Good
Bad	$o_{11} = 23$ $e_{11} = 16.80$	$o_{12} = 60$ $e_{12} = 52.64$	$o_{13} = 29$ $e_{13} = 42.56$
Average	$o_{21} = 28$ $e_{21} = 25.05$	$o_{22} = 79$ $e_{22} = 78.49$	$o_{23} = 60$ $e_{23} = 63.46$
Good	$o_{31} = 9$ $e_{31} = 18.15$	$o_{32} = 49$ $e_{32} = 56.87$	$o_{33} = 63$ $e_{33} = 45.98$

We can see that we have more students than expected in the Good - Good cell and less than expected in the two Bad - Good cells. We can now test the hypothesis of independence between English and mathematics results:

```
# Testing independence between english and maths results
chi2, p, dof, expected = stats.chi2_contingency(results,
correction=False)
# Test statistic
print(chi2)

20.178903582087926


# p-value
print(p)

0.00046038041384262443


# Degrees of Freedom
print(dof)

4


# Expected frequencies under the null hypothesis
print(pd.DataFrame(expected, index=['EngBad', 'EngAve', 'EngGood'],
                   columns=['MathBad', 'MathAve', 'MathGood']))

          MathBad  MathAve  MathGood
EngBad      16.80    52.64     42.56
EngAve      25.05    78.49     63.46
EngGood     18.15    56.87     45.98
```

The χ^2-test gives a test statistic of 20.18, which under H_0 follows a χ^2-distribution with 4 degrees of freedom leading to a p-value of 0.0005. This means that the hypothesis of independence between English and mathematics results is rejected.

Even though the hypothesis were formulated differently in the first setup when *comparing several groups*, compared to the second setup with the hypothesis on *independence of two categorical variables*, it turns out that the first hypothesis (7-52) is also about independence. Two events are independent if

$$P(A \text{ and } B) = P(A) \cdot P(B), \tag{7-64}$$

which expresses: the probability of both event A and event B occurring is equal to the probability of event A occurring times the probability of event B occuring.

Another way of defining independence of two variables is through conditioning. Two events are independent if

$$P(A|B) = P(A), \tag{7-65}$$

which states: the probability of event A does not change if we have information about B. In the first Example 7.23 the probability of voting for Candidate 1 is the same irrespective of week and therefore the distribution in one week is independent of the results from the other weeks.

Glossaries

Alternative hypothesis [Alternativ hypotese] The alternative hypothesis (H_1) is ofter the negation of the null hypothesis 6, 15

Binomial distribution [Binomial fordeling] If an experiment has two possible outcomes (e.g. failure or success, no or yes, 0 or 1) and is repeated more than one time, then the number of successes is binomial distributed 2, 3, 15

χ^2-distribution [χ^2-fordeling (udtales: chi-i-anden fordeling)] 20

Continuity correction The so-called Continuity correction is a general approach to make the best approximation of discrete probabilities 5

Critical value *Kritisk værdi* As an alternative to the p-value one can use the so-called critical values, that is the values of the test-statistic which matches exactly the significance level 8, 13

Degrees of freedom [Frihedsgrader] The number of "observations" in the data that are free to vary when estimating statistical parameters often defined as $n - 1$ 16, 17, 20, 21, 23, 27

Independence [Uafhængighed] 24–28

Null hypothesis [Nulhypotese (H_0)] 6–8, 12–17, 23, 24

One-sided (test) [Énsidet test] Is also called directional (test) 6

P-value [p-værdi (for faktisk udfald af en teststørrelse)] 9, 14, 17, 19, 23, 24, 27

Standard normal distribution [Standardiseret normalfordeling ($N(0,1)$)] 8

Two-sided (test) [Tosidet test (test med tosidet alternativ)] Is also called non-directional (test) 6, 12

Acronyms

ANOVA Analysis of Variance *Glossary:* Analysis of Variance

cdf cumulated distribution function *Glossary:* cumulated distribution function

CI confidence interval 2, 4, 5, 11, 12, *Glossary:* confidence interval

CLT Central Limit Theorem *Glossary:* Central Limit Theorem

IQR Inter Quartile Range *Glossary:* Inter Quartile Range

LSD Least Significant Difference *Glossary:* Least Significant Difference

pdf probability density function *Glossary:* probability density function