Course 02402 Introduction to Statistics

Lecture 10: Inference for proportions

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

Overview

Different analysis/data situations in 02402

Mean of quantitative data

- Hypothesis test/CI for one mean (one sample)
- Hypothesis test/CI for two means (two samples)
- Hypothesis test/CI for several means (K samples)

Today: Proportions

- Hypothesis test/CI for one proportion
- Hypothesis test/CI for two proportions
- Hypothesis test for several proportions
- Hypothesis test for several “multi-categorical” proportions

Estimation of proportions

- Estimation of a proportion/probability, by observing how many times x an event has occurred in n (independent) trials:

$$\hat{p} = \frac{x}{n}$$

- Note that $\hat{p} \in [0;1]$.
- Example: A dice is thrown $n = 100$ times. In $x = 20$ cases the outcome was ⚅. Then, $\hat{p}$ is the estimated probability of throwing a ⚅.

Overview

Confidence interval for one proportion

Method 7.3

If we have a large sample, then a $(1-\alpha)100\%$ confidence interval for p is:

$$\frac{x}{n} - z_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}} < p < \frac{x}{n} + z_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}}$$

How?

Follows from approximating the binomial distribution by the normal distribution.

A rule of thumb

Suppose that $X \sim \text{binom}(n,p)$. The normal distribution is a good approximation of the binomial distribution if np and $n(1-p)$ (expected no. of successes and failures, respectively) are both greater than 15.

Confidence interval for one proportion

Mean and variance of binomial distribution, Chapter 2.21

$$\begin{aligned} E(X) &= np \\ Var(X) &= np(1-p) \end{aligned}$$

This means that

$$\begin{aligned} E(\hat{p}) = E\left(\frac{X}{n}\right) &= \frac{np}{n} = p \\ Var(\hat{p}) = Var\left(\frac{X}{n}\right) &= \frac{1}{n^2}Var(X) = \frac{p(1-p)}{n} \end{aligned}$$

Example 1

Left-handedness:

$$p = \text{Proportion of left-handed people in Denmark}$$

and/or:

Female engineering students:

$$p = \text{Proportion of female engineering students}$$

Example 1

Left-handedness:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{10/100(1-10/100)}{100}} = 0.03$$

$$0.10 \pm 1.96 \cdot 0.03 = 0.10 \pm 0.059 = [0.041, 0.159]$$

Better "small sample" method - the "plus 2-approach" (Remark 7.7)

Use the same formula on $\tilde{x} = 10+2 = 12$ and $\tilde{n} = 100+2+2 = 104$:

$$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} = \sqrt{\frac{12/104(1-12/104)}{104}} = 0.03$$

$$0.115 \pm 1.96 \cdot 0.031 = 0.115 \pm 0.061 = [0.054, 0.177]$$

The Margin of Error (ME)

The Margin of Error

with $(1-\alpha)100\%$ confidence becomes:

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

where p may be estimated using $\hat{p} = \frac{x}{n}$.

The Margin of Error:

- Corresponds to half the width of the $(1-\alpha)100\%$ confidence interval.
- Describes the "minimum desired precision" of the estimate $\hat{p}$.

Sample size determination

Design of experiments:

How large should the sample size n be in order to obtain the desired precision?

Method 7.13

If you want a Margin of Error, ME, with $(1-\alpha)100\%$ confidence, then you need the following sample size:

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

Sample size determination

Method 7.13

If you want a Margin of Error, ME, with $(1-\alpha)100\%$ confidence, and you do *not* have a reasonable guess of p, then you need the following sample size:

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

since the worst case approach is given by: $p = \frac{1}{2}$

Example 1 - continued

Left-handedness:

Suppose that we want $\text{ME} = 0.01$ (with $\alpha = 0.05$) – what should n be?

Assume $p \approx 0.10$:

$$n = 0.1 \cdot 0.9 \left(\frac{1.96}{0.01}\right)^2 = 3467.4 \approx 3468$$

Without any assumption on the size of p:

$$n = \frac{1}{4}\left(\frac{1.96}{0.01}\right)^2 = 9604$$

Overview

Steps of a hypothesis test - an overview (repetition)

1. Formulate the hypothesis and choose the level of significance α (i.e. the "risk-level").
2. Use the data to calculate the value of the test statistic.
3. Calculate the p-value using the test statistic and the relevant distribution. Compare the p-value to the significance level α and draw a conclusion.
4. (Alternatively, draw a conclusion based on the relevant critical value(s)).

Hypothesis test for one proportion

The null and alternative hypothesis for one proportion p:

$$H_0: \quad p = p_0$$
$$H_1: \quad p \neq p_0$$

We either accept H_0 or reject H_0.

Testing the hypothesis: The test statistic

Theorem 7.10 and Method 7.11

If the sample size is sufficiently large (if $np_0 > 15$ and $n(1-p_0) > 15$), we use the following test statistic:

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

Under the null hypothesis, the random variable Z (approximately) follows a standard normal distribution, $Z \sim N(0, 1^2)$.

Testing the hypothesis: p-value and conclusion (Method 7.11)

Find the p-value (evidence against the null hypothesis):

- $2P(Z > |z_{\text{obs}}|)$

Test using the critical value:

Reject null hypothesis if $z_{\text{obs}} < -z_{1-\alpha/2}$ or $z_{\text{obs}} > z_{1-\alpha/2}$.

Example 1 - continued

Is half of the Danish population left handed?

$$H_0: p = 0.5, \;\; H_1: p \neq 0.5$$

Test statistic:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{10 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.5(1-0.5)}} = -8$$

p-value:

$$2 \cdot P(Z > 8) = 1.2 \cdot 10^{-15}$$

There is very strong evidence against the null hypothesis - we reject it (with $\alpha = 0.05$).

Example 1 - continued

Testing the hypothesis in R

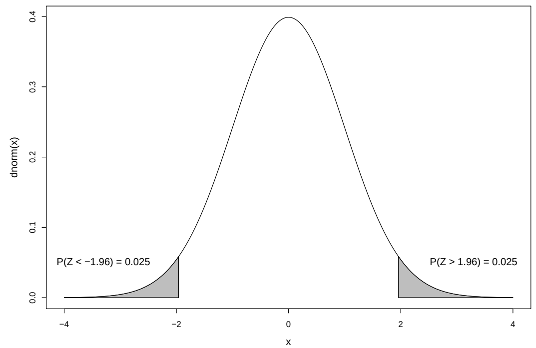
```
prop.test(10, 100, p = 0.5, correct = FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  10 out of 100, null probability 0.5
## X-squared = 64, df = 1, p-value = 1e-15
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.05523 0.17437
## sample estimates:
##   p
## 0.1
```


Example 1 - continued

Using the critical value instead:

$$z_{0.975} = 1.96$$

As $z_{\text{obs}} = -8$ is (much) less than -1.96 we reject the hypothesis.



Overview

Confidence interval for (the difference between) two proportions

Method 7.15

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

where

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

Rule of thumb:

Both $n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i = 1, 2$.

Hypothesis test for two proportions, Method 7.18

Two sample proportions hypothesis test

Comparing two proportions (shown here for a two-sided alternative)

$$H_0: \quad p_1 = p_2$$
$$H_1: \quad p_1 \neq p_2$$

The test statistic:

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \quad \textit{where} \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

And for large samples:

Use the standard normal distribution again.

Example 2

Is there a relation between ths use of birth control pills and the risk of a blood clot in the heart?

In a study (USA, 1975) the connection between birth control pills and the risk of a blood clot in the heart was investigated.

	Blood clot	No blood clot
B. C. pill	23	34
No B. C. pill	35	132

Carry out a test to check if there is any connection between the use of birth control pills and the risk of a blood clot in the heart. Use a significance level of $\alpha = 0.05$.

Example 2

In a study (USA, 1975) the connection between birth control pills and the risk of blood clot in the heart was investigated.

	Blood clot	No blood clot
B. C. pill	23	34
No B. C. pill	35	132

Estimates within each sample

$$\hat{p}_1 = \frac{23}{57} = 0.4035, \;\; \hat{p}_2 = \frac{35}{167} = 0.2096$$

Common estimate:

$$\hat{p} = \frac{23 + 35}{57 + 167} = \frac{58}{224} = 0.2589$$

Example 2 - continued

prop.test for equality of two proportions in R

```r
# Read data table into R
pill.study <- matrix(c(23, 34, 35, 132),
                     ncol = 2, byrow = TRUE)
colnames(pill.study) <- c("Blood Clot", "No Clot")
rownames(pill.study) <- c("Pill", "No pill")
pill.study

# Test whether probabilities are equal for the two groups
prop.test(pill.study, correct = FALSE)
```

Example 2 - continued

prop.test for equality of two proportions in R

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

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  pill.study
## X-squared = 8.3, df = 1, p-value = 0.004
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.05239 0.33546
## sample estimates:
## prop 1 prop 2
## 0.4035 0.2096
```


Overview

Hypothesis test for several proportions

The comparison of c proportions

In some cases, we might be interested in determining whether two or more binomial distributions have the same parameter p. That is, we are interested in testing the null hypothesis:

$$H_0: \quad p_1 = p_2 = \ldots = p_c = p$$

vs. the alternative that at least two proportions are different.

Hypothesis test for several proportions

Table of observed counts for c samples:

	Sample 1	Sample 2	...	Sample c	Total
Succes	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

Common (average) estimate:

Under the null hypothesis, the estimate of p is:

$$\hat{p} = \frac{x}{n}$$

Hypothesis test for several proportions

Common (average) estimate:

Under the null hypothesis the estimate of p is:

$$\hat{p} = \frac{x}{n}$$

"Use" this common estimate in each group:

If the null hypothesis is true, we expect that the j'th group/sample has e_{1j} successes and e_{2j} failures, where

$$e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot x}{n}$$

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

Hypothesis test for several proportions

Make a table with the *expected* counts for the c samples:

e_{ij}	Sample 1	Sample 2	...	Sample c	Total
Succes	e_{11}	e_{12}	...	e_{1c}	x
Failure	e_{21}	e_{22}	...	e_{2c}	$n-x$
Total	n_1	n_2	...	n_c	n

General way to find the expected counts in freqency tables:

$$e_{ij} = \frac{(i\text{'th row total}) \cdot (j\text{'th column total})}{\text{total}}$$

Computation of the test statistic - Method 7.20

The test statistic becomes

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

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

Find the p-value or use the critical value - Method 7.20

Sampling distribution for test statistic under H_0:

χ^2-distribution with $(c-1)$ degrees of freedom (approx.)

Critical value method:

If $\chi^2_{\text{obs}} > \chi^2_{\alpha}(c-1)$ the null hypothesis is rejected.

Rule of thumb for validity of the test:

All expected values $e_{ij} \geq 5$.

Example 2 - continued

The *observed* values o_{ij}

Observed	Blood clot	No Blood clot
B. C. pill	23	34
No B. C. pill	35	132

Example 2 - continued

Compute the *expected* values e_{ij}

Expected	Blood clot	No Blood clot	Total
B. C. pill			57
No B. C. pill			167
Total	58	166	224

Example 2 - continued

Use "the rule" for expected values four times, e.g.:

$$e_{22} = \frac{167 \cdot 166}{224} = 123.76$$

The *expected* values e_{ij}:

Expected	Blood clot	No Blood clot	Total
B. C. pill	14.76	42.24	57
No B. C. pill	43.24	123.76	167
Total	58	166	224

Example 2 - continued

The test statistic (remember to include all cells):

$$\chi^2_{\text{obs}} = \frac{(23-14.76)^2}{14.76} + \frac{(34-42.24)^2}{42.24} + \frac{(35-43.24)^2}{43.24} + \frac{(132-123.76)^2}{123.76}$$
$$= 8.33$$

Critical value:

```
qchisq(0.95, 1)
```

```
[1] 3.841
```

Conclusion:

We reject the null hypothesis - there *is* a significantly higher risk of blood clots in the birth control pill group.

Example 2 - continued

`chisq.test` for equality of two proportions in R

```
# Test whether probabilities are equal for the two groups
chisq.test(pill.study, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data:  pill.study
## X-squared = 8.3, df = 1, p-value = 0.004

# Expected values
chisq.test(pill.study, correct = FALSE)$expected

##          Blood Clot No Clot
## Pill          14.76    42.24
## No pill       43.24   123.76
```


Overview

Example 3: Analysis of contingency tables

A 3×3 table - 3 samples, 3-category outcomes

	4 weeks bef	2 weeks bef	1 week bef
Candidate I	79	91	93
Candidate II	84	66	60
Undecided	37	43	47
	$n_1 = 200$	$n_2 = 200$	$n_3 = 200$

Are the votes equally distributed?

$$H_0: \quad p_{i1} = p_{i2} = p_{i3}, \quad i = 1,2,3.$$

Analysis of contingency tables

A 3×3 table - 1 sample, two 3-category variables:

	bad	average	good
bad	23	60	29
average	28	79	60
good	9	49	63

Is there independence between the row and column variables?

$$H_0: \; p_{ij} = p_{i\cdot} p_{\cdot j}$$

Computation of the test statistic – no matter the type of table 7.22

In a contingency table with r rows and c columns, the test statistic is:

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

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

General way to find the expected counts in freqency tables:

$$e_{ij} = \frac{(i\text{'th row total}) \cdot (j\text{'th column total})}{\text{total}}$$

Find p-value or use critical value - Method 7.22

Sampling distribution for test-statistic under H_0:

χ^2-distribution with $(r-1)(c-1)$ degrees of freedom (approx.).

Critical value method:

If $\chi^2_{\text{obs}} > \chi^2_{\alpha}$ with $(r-1)(c-1)$ degrees of freedom, the null hypothesis is rejected.

Rule of thumb for validity of the test:

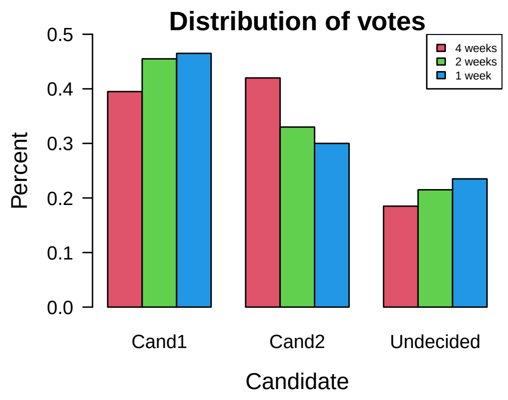
All expected values $e_{ij} \geq 5$.

Example 3 - continued

`chisq.test` for contingency tables

```
# Read data table into R
poll <-matrix(c(79, 91, 93, 84, 66, 60, 37, 43, 47),
              ncol = 3, byrow = TRUE)
colnames(poll) <- c("4 weeks", "2 weeks", "1 week")
rownames(poll) <- c("Cand1", "Cand2", "Undecided")

# Show column percentages
prop.table(poll, 2)

##           4 weeks 2 weeks 1 week
## Cand1       0.395   0.455  0.465
## Cand2       0.420   0.330  0.300
## Undecided   0.185   0.215  0.235
```


Example 3 - continued



Example 3 - continued

```
# Testing for same distribution in the three populations
chisq.test(poll, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data:  poll
## X-squared = 7, df = 4, p-value = 0.1

# Expected values
chisq.test(poll, correct = FALSE)$expected

##            4 weeks 2 weeks 1 week
## Cand1        87.67   87.67  87.67
## Cand2        70.00   70.00  70.00
## Undecided    42.33   42.33  42.33
```


Overview