Inference for categorical data

E. Pastucha

October 2024

Proportion

To describe categorical variables we usually employ proportion.

Population proportion: p

Sample Proportion; $\hat{p}$

Proportion

Proportion is usually regarded in category of 'Success' or 'Failure'.

In simple inference we simplify the data into two categories:

- category of interest - Success $\hat{p}$
- remaining categories - Failure $1 - \hat{p}$

Proportion

We want to check if majority of Danes supports a 4 day working week.

We have survey data of about 1000 Danes. They submitted they preferences in a survey.

What would be ‘Success’ group? What would be ‘Failure’ group?

Proportion - CLT

Proportion is also subjected to Central Limit Theorem.

Conditions:

- independent cases,
- number of cases for success and for failure should be more than 10: $n_{\hat{p}} > 10$, $n_{1-\hat{p}} > 10$

Proportion - SE

If CLT conditions are fulfield:

$$SE = \sqrt{\frac{p \cdot (1-p)}{n}}$$

Proportion - confidence interval

If we're using SE formula for confidence interval we use $\hat{p}$ as input.

Confidence interval:

$$\hat{p} \pm Z_{score} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

Where Z_{score} responds to probability we want confidence interval to represent.

Proportion - SE

When in doubt what p to use within SE always go for worse case scenario: $p = 0.5$

$$f(p) = p(1-p)$$

$$\downarrow$$

$$f(p) = p - p^2$$

Proportion - SE

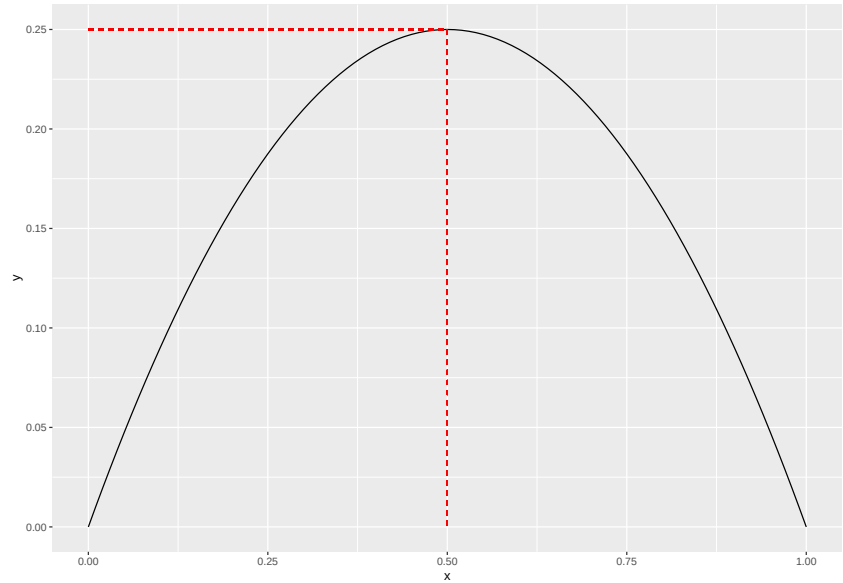
Derivative - to find maximum.

$$f(p) = p - p^2$$

$$\downarrow$$

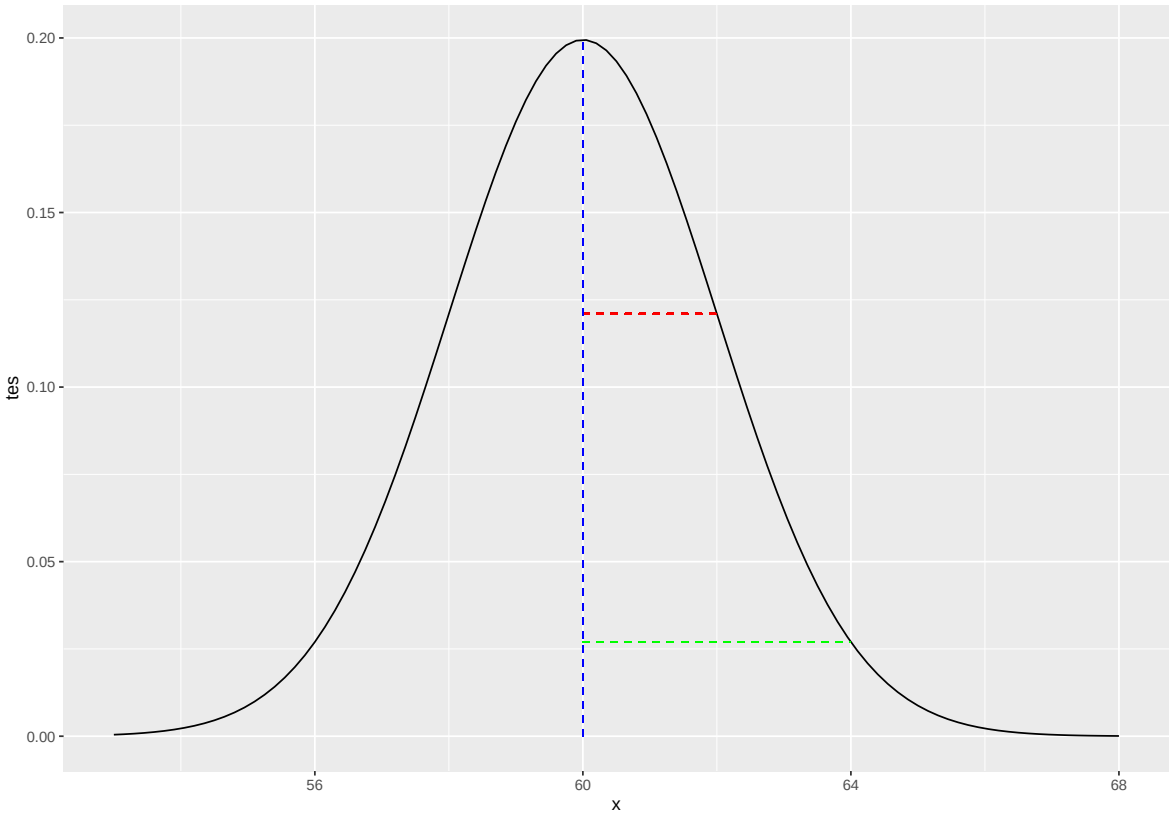
$$f'(p) = 1 - 2p$$

Zero point for $f(p) = p - p^2$ (maximum point) is $p = 0.5$



Proportion sampling distribution plot

$\hat{p} = 60\%$ with $SE = 2\%$



Proportion sampling distribution plot

$\hat{p} = 60\%$ with $SE = 2\%$ 95% confidence interval

Proportion - hypothesis testing example

Check whether over 50% of NBA players are over 200 cm tall.

Proportion - hypothesis testing example

1) Set up hypothesis:

H_0: 50% of NBA players is over 200 cm tall: $p = 0.5$

H_A: over 50% of NBA players is over 200 cm tall $p > 0.5$

2) Assume threshold values:

- α -significance level - 0.05

Proportion - hypothesis testing example

3) Check CLT conditions

- independence
- $n_{\hat{p}} > 10 \longrightarrow 302 > 10$
- $n_{1-\hat{p}} > 10 \longrightarrow 203 > 10$

Number of cases in the sample

```
## [1] 505
```

Number of players taller than 200 cm in the sample:

```
## [1] 302
```

Remaining players:

```
## [1] 203
```

Proportion - hypothesis testing example

4) Calculate the results

Proportion values - point estimates:

```
(p_success <- 302/505)
```

```
## [1] 0.5980198
```

```
(p_failure <- 203/505)
```

```
## [1] 0.4019802
```

Proportion - hypothesis testing example

4) Calculate the results

$$SE = \sqrt{\frac{p \cdot (1 - p)}{n}}$$

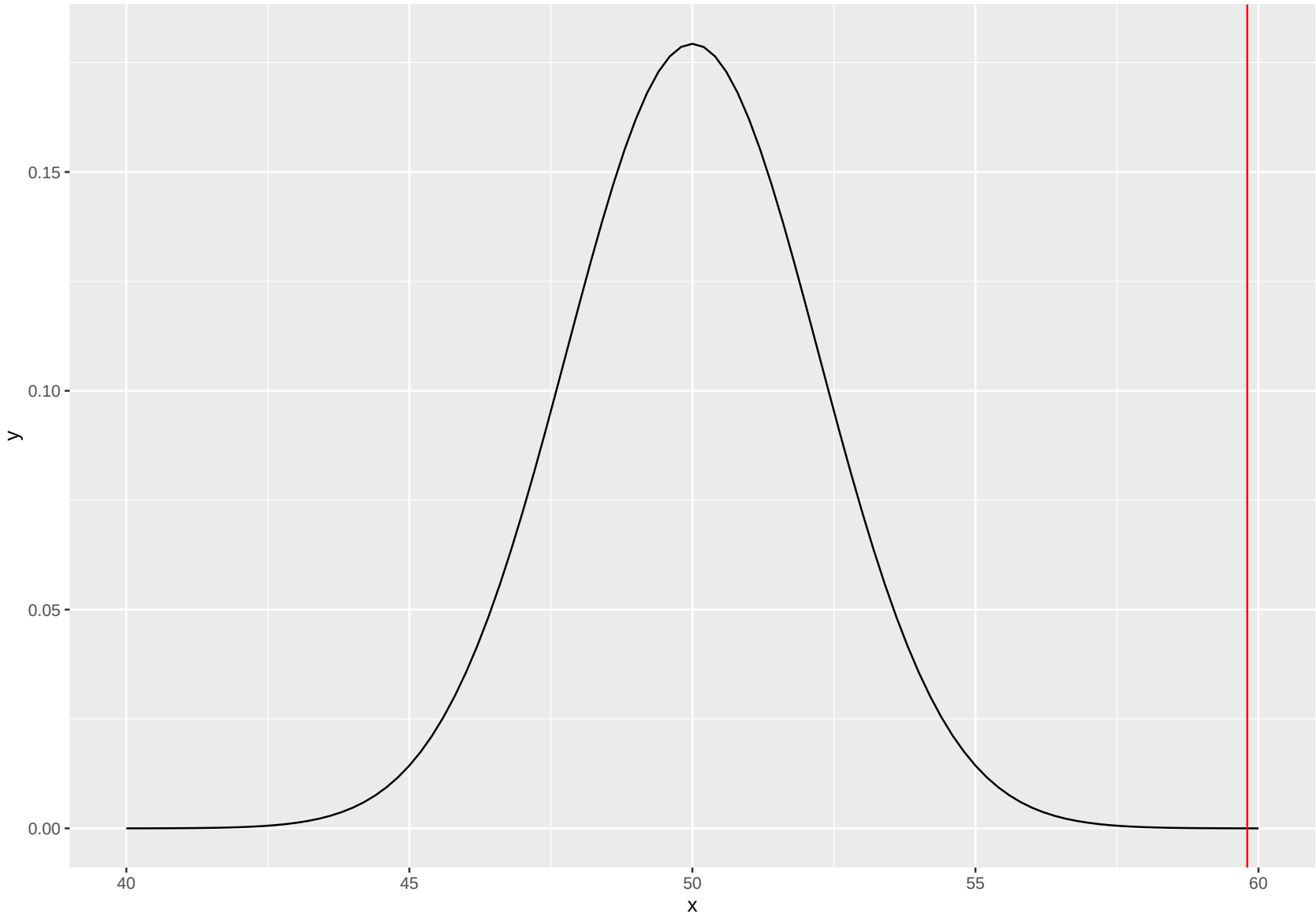
Here we will use null hypothesis p for calculation of SE.

```
(SE <- sqrt((0.5*0.5)/nrow(nba)))
```

```
## [1] 0.02224971
```

Proportion - hypothesis testing example

4) Calculate the results



Proportion - hypothesis testing example

4) Calculate the results

P-value:

```
(p_value <- (1-pnorm(p_success, mean = 0.5, sd = SE)))
```

```
## [1] 5.278415e-06
```

5) Form conclusions

We reject null hypothesis in favour of alternative. Due to the position of point estimate we can conclude that more than 50% of NBA players are over 200 cm tall.

Difference of two proportions

Difference of two proportions

CLT criteria:

- independent cases,
- number of cases for success and for failure should be more than 10: $n_{\hat{p}} > 10$, $n_{1-\hat{p}} > 10$

In this case we use pooled proportion p_{pool}

$$p_{pool} = \frac{n_{1success} + n_{2success}}{n_1 + n_2}$$

Difference of two proportions

SE formula:

$$SE = \sqrt{\frac{p_1 \cdot (1 - p_1)}{n_1} + \frac{p_2 \cdot (1 - p_2)}{n_2}}$$

H_0: poportion of success in population 1 differs from proportion of success within population 2: $p_1 = p_2 \longrightarrow p_1 - p_2 = 0$

H_A: poportion of success in population 1 differs from proportion of success within population 2: $p_1 \neq p_2 \longrightarrow p_1 - p_2 \neq 0$

Difference of two proportions - example

A clothes producer is looking for a new supplier of zippers. Two factories are frontrunners. The producer wants to decide based on one days production results.

First factory produced 23 935 zippers, out of which 132 were faulty. Second factory produced 22 312 zippers, out of which 111 were faulty.

We want to check if it's safe to assume that they have same proportion of faulty zippers, and due to that we can choose based squerly on price.

Difference of two proportions - example

1) Set up the hypothesis

H_0: poportion of faulty zippers is equal in both factories: $p_1 - p_2 = 0$

H_A: poportion of faulty zippers is different in between both factories: $p_1 - p_2 \neq 0$

2) Assume threshold values:

- α -significance level - 0.05

Difference of two proportions - example

3) Check conditions for CLT:

```
(p_pooled <- (132 + 111)/(23935 + 22312))
```

```
## [1] 0.005254395
```

```
p_pooled * 23935
```

```
## [1] 125.7639
```

```
(1- p_pooled) * 23935
```

```
## [1] 23809.24
```

```
p_pooled * 22312
```

```
## [1] 117.2361
```

```
(1- p_pooled) * 22312
```

Difference of two proportions - example

4) Calculate the results:

Calculate proportions:

```
(p_1 <- 132/23935)
```

```
## [1] 0.005514936
```

```
(p_2 <- 111/22312)
```

```
## [1] 0.004974901
```

Difference of two proportions - example

4) Calculate the results:

Point estimate:

```
(point_estimate <- p_1 - p_2)
```

```
## [1] 0.0005400349
```

And SE:

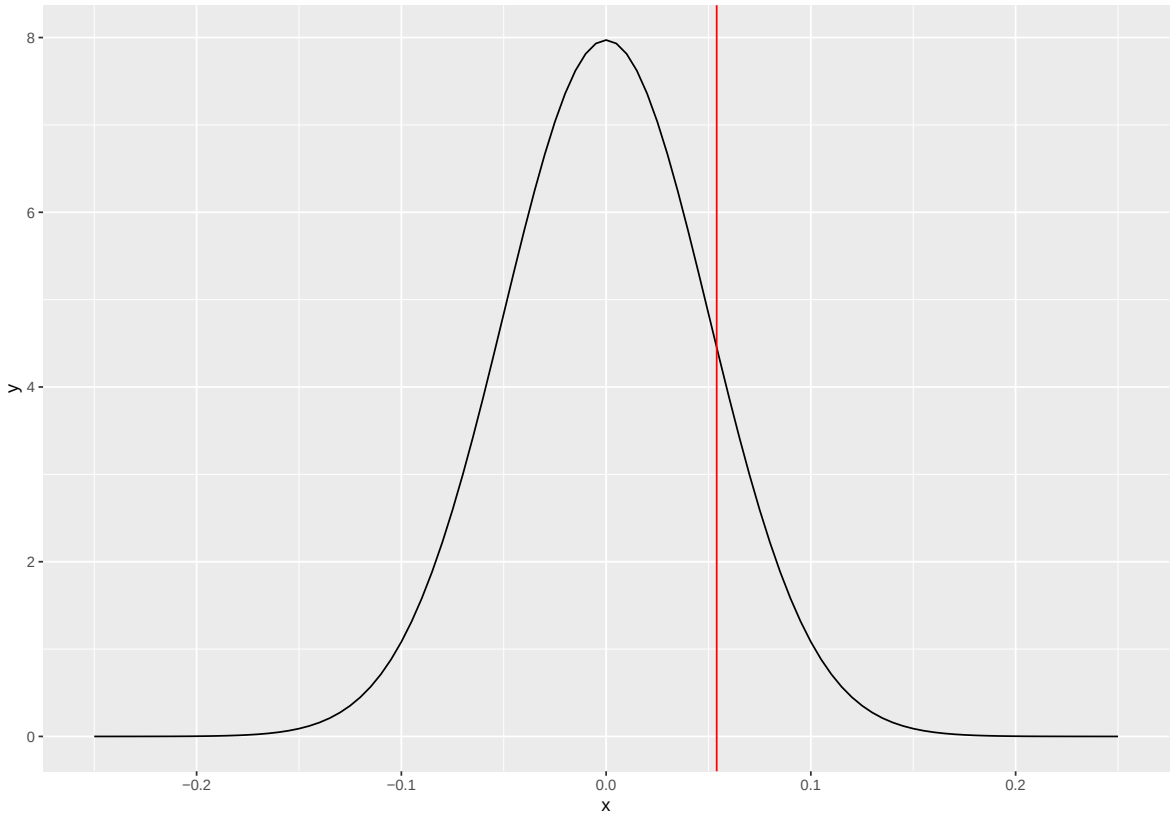
```
(SE <- sqrt((p_1*(1-p_1))/23935 + (p_2*(1-p_2))/232312))
```

```
## [1] 0.0005004503
```

Difference of two proportions - example

4) Calculate the results:

Plot null hypothesis distribution (view in percentages) with point estimate:



Difference of two proportions - example

4) Calculate the results:

Calculate p-value:

```
(p_value <- 2*(1-pnorm(point_estimate,
                       mean = 0, sd = SE)))
```

```
## [1] 0.2805441
```

5) Form conclusions:

We accept null hypothesis and reject alternative. Clothes producer should make a choice based on price alone, as there seem to be an equal proportion of faulty zippers in between factories.