Biostatistics I: Hypothesis testing

Categorical data: Proportion tests

Eleni-Rosalina Andrinopoulou

Department of Biostatistics, Erasmus Medical Center

e.andrinopoulou@erasmusmc.nl

@erandrinopoulou

Erasmus MC
University Medical Center Rotterdam

In this Section

- z-test for proportions
- Bionomial test
- Examples

z-test for proportions: Theory

Assumptions

- The observations are independent of one another
- The sample size is large enough to use the normal approximation $\mathcal{N}(np, np(1-p))$
  - $np > 10$ and $n(1-p) > 10$, where n is the number of observations and p the proportion

One sample z-test for proportions: Theory

Scenario

Is the probability of being diagnosed with asthma now different than it was 50 years ago?

Hypothesis

$H_0 : \pi = \pi_0$
$H_1 : \pi \neq \pi_0$

One sample z-test for proportions: Theory

Hypothesis

If **one-tailed**
Is the probability of being diagnosed with asthma now higher than it was 50 years ago?
$H_0 : \pi = \pi_0$
$H_1 : \pi > \pi_0$

or

Is the probability of being diagnosed with asthma now lower than it was 50 years ago?
$H_0 : \pi = \pi_0$
$H_1 : \pi < \pi_0$

One sample z-test for proportions: Theory

Test statistic

For large sample sizes, the distribution of the test statistic is approximately normal

$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

- Sample proportion: p
- Population proportion: π_0
- Number of subjects: n

If continuity correction is applied: $z = \frac{p - \pi_0 + c}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$,

where

- $c = -\frac{1}{2n}$ if $p > \pi_0$
- $c = \frac{1}{2n}$ if $p < \pi_0$
- $c = 0$ if $|p - \pi_0| < \frac{1}{2n}$

One sample z-test for proportions: Theory

Sampling distribution

- z-distribution
- Critical values and p-value

Type I error

- Normally α = 0.05

Draw conclusions

- Compare test statistic (z) with the critical values$_{\alpha/2}$ or the p-value with α

If **one-tailed**: Compare test statistic with the critical value$_{\alpha}$

One sample z-test for proportions: Application

Scenario

Is the probability of being diagnosed with asthma now different than it was 50 years ago?

Hypothesis

$H_0 : \pi = \pi_0$
$H_1 : \pi \neq \pi_0$

One sample z-test for proportions: Application

Hypothesis

$H_0 : \pi = \pi_0$

$H_1 : \pi \neq \pi_0$

Collect and visualize data

x	Freq
No	47
Yes	53

50 years ago we had $\pi_0 = 0.6$

Test statistic

(with no continuity correction):

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.53 - 0.6}{\sqrt{\frac{0.6(1-0.6)}{100}}} = -1.43$$

Type I error

$\alpha = 0.05$

One sample z-test for proportions: Application

Critical values

Using R we get the critical values from the z-distribution:
critical value$_{\alpha/2}$ = critical value$_{0.05/2}$

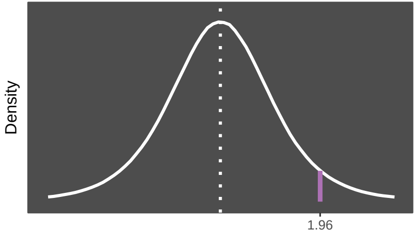
```
qnorm(p = 0.05/2, lower.tail = FALSE)
```

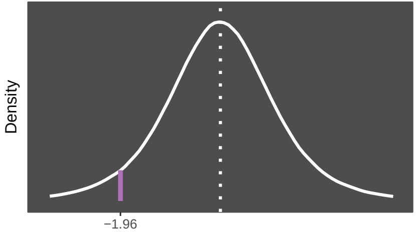
```
[1] 1.959964
```

-critical value$_{\alpha/2}$ = -critical value$_{0.05/2}$

```
qnorm(p = 0.05/2, lower.tail = TRUE)
```

```
[1] -1.959964
```





One sample z-test for proportions: Application

Critical values

If **one-tailed**

critical value$_\alpha$:

```
qnorm(p = 0.05, lower.tail = FALSE)
```

or

-critical value$_\alpha$:

```
qnorm(p = 0.05, lower.tail = TRUE)
```

One sample z-test for proportions: Application

Draw conclusions

We reject the H_0 if:

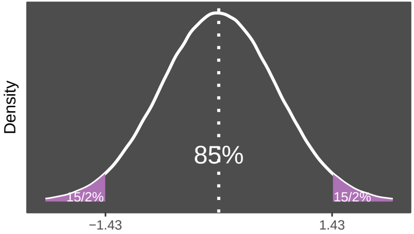
- $z >$ critical value$_{\alpha/2}$ or $z <$ - critical value$_{\alpha/2}$

We have -1.43 > -1.96 $\Rightarrow$ we do not reject the H_0

Using R we obtain the p-value from the z-distribution:

```
2 * pnorm(q = -1.43, lower.tail = TRUE)
```

```
[1] 0.152717
```



Two sample z-test for proportions: Theory

Scenario

Is the probability of being diagnosed with asthma in the Netherlands different than in Belgium?

Hypothesis

$H_0 : \pi_1 = \pi_2$
$H_1 : \pi_1 \neq \pi_2$

Two sample z-test for proportions: Theory

Test statistic

For large sample sizes, the distribution of the test statistic is approximately normal.

Pooled version:

$$z = \frac{(p_1 - p_2) - 0}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Unpooled version:

$$z = \frac{(p_1 - p_2) - 0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

- Sample proportion of group 1: p_1
- Sample proportion of group 2: p_2
- Number of subjects in group 1: n_1
- Number of subjects in group 2: n_2
- Total proportion: $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

Two sample z-test for proportions: Theory

Test statistic

If continuity correction is applied:

Pooled version:

$$z = \frac{(p_1 - p_2) + \frac{F}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Unpooled version:

$$z = \frac{(p_1 - p_2) + \frac{F}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

where

- $F = -1$ if $p_1 > p_2$
- $F = 1$ if $p_1 < p_2$

Two sample z-test for proportions: Theory

Sampling distribution

- z-distribution
- Critical values and p-value

Type I error

- Normally $\alpha = 0.05$

Draw conclusions

- Compare test statistic (z) with the critical values or the p-value with α

Two sample z-test for proportions: Application

Scenario

Is the probability of being diagnosed with asthma in the Netherlands different than in Belgium?

Hypothesis

$H_0 : \pi_1 = \pi_2$
$H_1 : \pi_1 \neq \pi_2$

Two sample z-test for proportions: Application

Collect and visualize data

Table 1: the Netherlands

x1	Freq
No	47
Yes	53

Table 2: Belgium

x2	Freq
No	62
Yes	38

Test statistic

(with no continuity correction and pooled version):

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{100\ 0.53 + 100\ 0.38}{100 + 100} = 0.46$$

$$z = \frac{(p_1 - p_2) - 0}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.53 - 0.38}{\sqrt{0.46(1-0.46)\left(\frac{1}{100} + \frac{1}{100}\right)}} = 2.13$$

Type I error

$\alpha = 0.05$

Two sample z-test for proportions: Application

Critical values

Using R we get the critical values from the z-distribution:
critical value$_{\alpha/2}$ = critical value$_{0.05/2}$

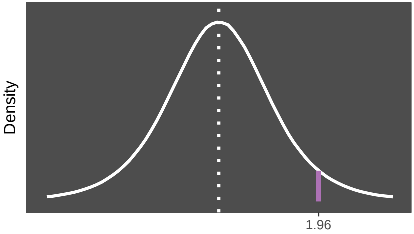
```r
qnorm(p = 0.05/2, lower.tail = FALSE)
```

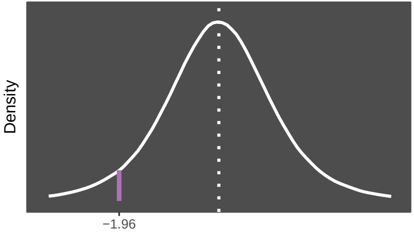
```
[1] 1.959964
```

-critical value$_{\alpha/2}$ = -critical value$_{0.05/2}$

```r
qnorm(p = 0.05/2, lower.tail = TRUE)
```

```
[1] -1.959964
```





Two sample z-test for proportions: Application

Draw conclusions

We reject the H_0 if:

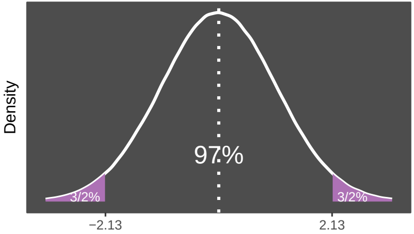
- z > critical value$_{\alpha/2}$ or z < - critical value$_{\alpha/2}$

We have 2.13 > 1.96 $\Rightarrow$ we reject the H_0

Using R we obtain the p-value from the z-distribution:

```
2 * pnorm(q = 2.13, lower.tail = FALSE)
```

```
[1] 0.03317161
```



Bionomial test: Theory

Assumptions

- Independent observations

Notes..

- The binomial test is an exact test

Bionomial test: Theory

Scenario

Is the probability of being diagnosed with asthma now different than it was 50 years ago?

Hypothesis

$H_0 : \pi = \pi_0$
$H_1 : \pi \neq \pi_0$

Bionomial test: Theory

If n is the sample size and k the successes: $Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ and ! indicates a factorial

- For any possible outcome of the binomial we obtain the corresponding probability
- We find the p-value by considering the probability of seeing an outcome as, or more, extreme
  - For a one-tailed test, $H_1 : \pi < \pi_0$
    $p - value = \Pr(X = 0) + \cdots + Pr(X = k) = \sum_{i=0}^{k} Pr(X = i) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$
  - Calculating a p-value for a two-tailed test is more complicated, since a binomial distribution is not symmetric if $\pi_0 \neq 0.5 \Rightarrow$ we cannot double the p-value from the one-tailed test

Type I error

- Normally $\alpha = 0.05$

Bionomial test: Application

Scenario
Is the probability of being diagnosed with asthma now lower than it was 50 years ago?

Hypothesis

$H_0 : \pi = \pi_0$

$H_1 : \pi < \pi_0$

Collect and visualize data

- $n = 10$
- $k = 3$
- $p = 0.3$
- $\pi_0 = 0.4$ the probability of being diagnosed with asthma 50 years ago

P-value

$Pr(X <= 3)$

Using R we get the p-value:

```
pbinom(q = 3, size = 10, prob = 0.4)
```

```
[1] 0.3822806
```

Draw conclusions

We do not reject the H_0

Bionomial test: Application

Scenario
Is the probability of being diagnosed with asthma now higher than it was 50 years ago?

Hypothesis

$H_0 : \pi = \pi_0$
$H_1 : \pi > \pi_0$

Collect and visualize data

- $n = 10$
- $k = 6$
- $p = 0.6$
- $\pi_0 = 0.4$ the probability of being diagnosed with asthma 50 years ago

P-value

$Pr(X >= 6) = 1 - Pr(X < 6) = 1 - Pr(X <= 5)$

Using R we get the p-value:

```
pbinom(q = 5, size = 10, prob = 0.4,
       lower.tail = FALSE)
```

```
[1] 0.1662386
```

Draw conclusions

We do not reject the H_0