vliruos

SHARING MINDS, CHANGING LIVES

This course was developed as a part of the VLIR-UOS Cross-Cutting projects:

- Statistics: 2011-2016, 2017.
- Statistics: 2017.
- Statistics for development : 2018-2022.

The >eR-Biostat initiative

Making R based education materials in statistics accessible for all

Introduction to Statistical inference and estimation using R: Inference for categorical data

Developed by

Ziv Shkedy (Hasselt Univesrsity) and Tadesse Awoke (Gondar University)

LAST UPDATE: 2022

Visit us on Facebook ER-BioStat

GitHub https://github.com/eR-Biostat

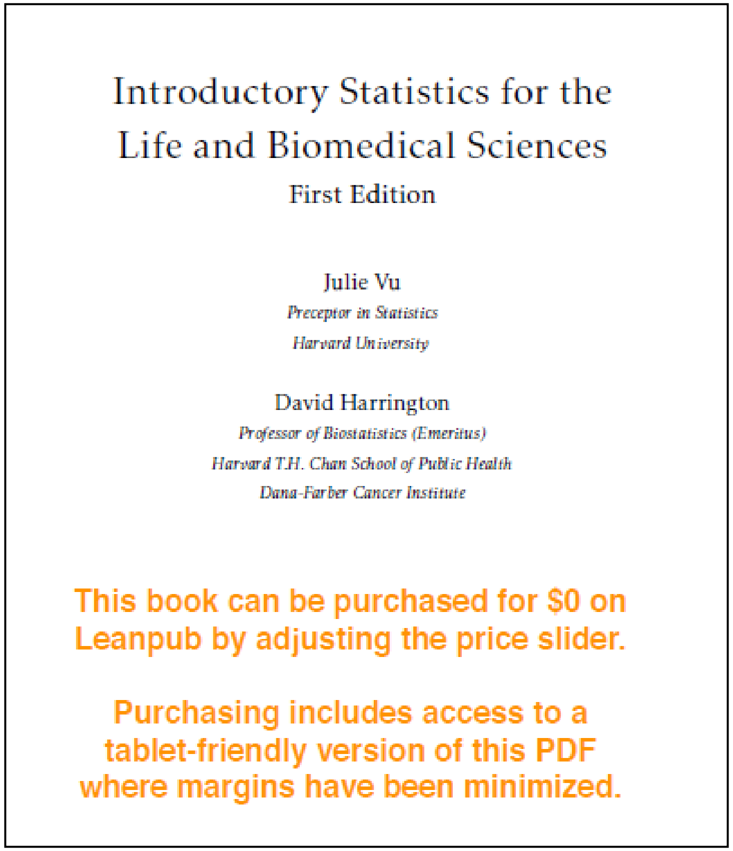
Email: erbiostat@gmail.com

twitter @erbiostat

Development team

- Tadele Worku Mengesha (Gondar University).
- Abdisa Gurmessa (Jmma University).
- Ziv Shkedy (Hasselt Univesrsity).
- Tadesse Awoke (Gondar University).
- Adetayo Kasim (Durham University).

Recommended reading



Introductory Statistics for the Life and Biomedical Sciences

First Edition

Julie Vu

Preceptor in Statistics

Harvard University

David Harrington

Professor of Biostatistics (Emeritus)

Harvard T.H. Chan School of Public Health

Dana-Farber Cancer Institute

This book can be purchased for $0 on Leanpub by adjusting the price slider.

Purchasing includes access to a tablet-friendly version of this PDF where margins have been minimized.

Chapter 8: Inference for categorical data

- We cover mainly Chapter 8 (Section 8.1-8.3).
- The examples that are used for illustration are not the same as the examples in the book.

Software

- R functions:
  - prop.test().
  - chisq.test().

YouTube tutorials

- YouTube tutorials are available for:
  - Inference on a Proportion in R... using prop.test() (host: Ed Boone): https://www.youtube.com/watch?v=-msRQ0YZtAY.
  - Confidence intervals on proportions in R (host: Ed Boone): https://www.youtube.com/watch?v=l-n8PAnEbN0&t=57s.

Datasets

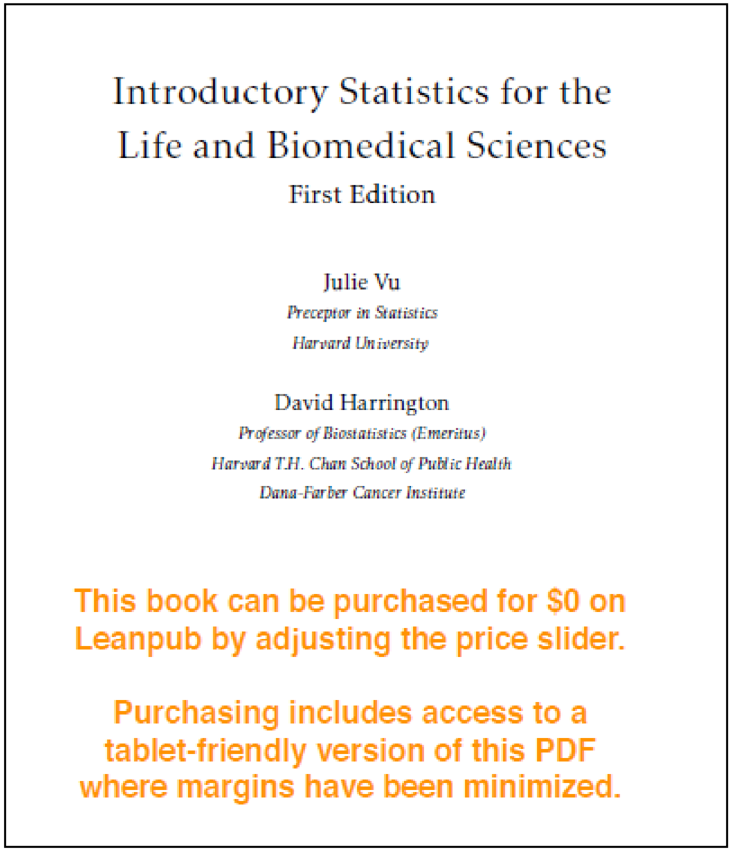
- Data are given as a part of R programs for the course.

- External datasets (which are not given as a part of the R code) and used for illustration are available online.

Topics

1. Inference for a single proportion.
2. Confidence intervals for a single proportion.
3. Inference for two independent samples.
4. Chi-squared test for independence.

Inference for a single proportion



Introductory Statistics for the Life and Biomedical Sciences

First Edition

Julie Vu

Preceptor in Statistics

Harvard University

David Harrington

Professor of Biostatistics (Emeritus)

Harvard T.H. Chan School of Public Health

Dana-Farber Cancer Institute

This book can be purchased for $0 on Leanpub by adjusting the price slider.

Purchasing includes access to a tablet-friendly version of this PDF where margins have been minimized.

Section 8.1

Population

- Population: random variable with two categories.
- Example: in the Belgian population

$$X_i = \begin{cases} Woman \\ Man \end{cases} \qquad X_i = \begin{cases} 1 & W \\ 0 & M \end{cases}$$

What is the number of women in Belgium? $X = \sum_{i=1}^{N} X_i$

What is the proportion of women in Belgium? $\pi = \frac{1}{N} \sum_{i=1}^{N} X_i$

The population: Bernoulli distribution

Thus, the unknown parameter of the population is :

$$\pi = \frac{1}{N} \sum_{i=1}^{N} X_i = P(X = 1)$$

We say that X is a Bernoulli distribution with parameter π follows:

$$X = \begin{cases} 1 & \text{with probability } \pi \\ 0 & \text{with probability } 1 - \pi \end{cases}$$

Example

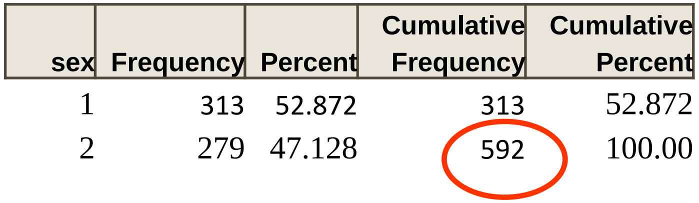
HairEyeColor data were used to show the proportion of the gender. Suppose X is the sex.

$$X_i = \begin{cases} 1 & \text{Female} \\ 0 & \text{Male} \end{cases}$$

$$\pi = P(X = 1) = P(Female)$$

→ We want to estimate π

Sample from the population

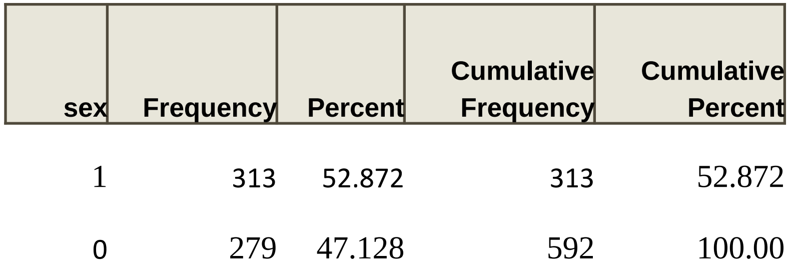


sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	313	52.872	313	52.872
2	279	47.128	592	100.00

- where:
  - 1: female
  - 2: male

Thus

$$X = \begin{cases} 1 & \text{When sex is female} \\ 0 & \text{When sex is male} \end{cases}$$



sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	313	52.872	313	52.872
0	279	47.128	592	100.00

Bernoulli distribution

Population mean and population variance :

$$X = \begin{cases} 1 & \pi \qquad\qquad \text{When sex is female} \\ 0 & 1-\pi \qquad \text{When sex is male} \end{cases}$$

$$\left.\begin{array}{l} \boxed{\mu = E(X) = 0 \times (1-\pi) + 1 \times \pi = \pi} \\ \boxed{\sigma^2 = Var(X) = 0^2 \times (1-\pi) + 1^2 \times \pi - \pi^2 = \pi(1-\pi)} \end{array}\right\} \text{Unknown parameter}$$

The sample mean

- A sample is, as always, a row of random measurements X1, ..., Xn are independent and have the same distribution as X.
- The sample mean is:

$$\overline{P} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\textit{total number of 1's in the sample}}{n}$$

•The sample proportion (notation$\overline{P}$).

•Point Estimator of π

Frequency Table of “gender”

$$X = \begin{cases} 1 & \text{When sex is female} \\ 0 & \text{When sex is male} \end{cases}$$

spine	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	313	52.872	313	52.872
0	279	47.128	592	100.00

The sample proportion is a point estimator for the unknown proportion of the population in our example:

52.872% of the sample were females.

Point estimate of population proportion using R

```
> malefemale=apply(HairEyeColor,3,sum)
> n=sum(malefemale)
> pbar=malefemale[2]/n
> Pbar
```

```
 Female
0.5287162
```

Point Estimator for population mean

As it is generally true that

$$E(\overline{X}) = \mu \qquad\qquad Var(\overline{X}) = \frac{\sigma^2}{n}$$

$$\boxed{E(\overline{X}) = E(\overline{P}) = \pi}$$

i.e. $\overline{P}$ an unbiased estimator for π

Variance of $\overline{P}$

As it is generally true that

$$E(\overline{X}) = \mu \qquad\qquad Var(\overline{X}) = \frac{\sigma^2}{n}$$

$$\boxed{\sigma^2 = Var(X) = 0^2 \times (1-\pi) + 1^2 \times \pi - \pi^2 = \pi(1-\pi)}$$

$$\boxed{Var(\overline{X}) = Var(\overline{P}) = \frac{\pi(1-\pi)}{n}}$$

Central limit theorem

From former properties of the sample mean we get in particular for $\overline{P}$ that for n large

$$\boxed{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\overline{P}-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)}$$

Since π in most cases is not known, we replace, if necessary π (1 - π) by $\overline{P}(1-\overline{P})$

What does the sample size(n) large?

Our previous condition for such an approach was: n ≥ 30.

Although in most cases turns out well, there is a rule of thumb that is more specific for proportions.

This says that the approach is good if both true:

$$\boxed{n\pi \geq 5 \quad or \quad n(1-\pi) \geq 5}$$

In the normal approximation, we replace the unknown π (1-π) by

$$\boxed{\bar{P}(1-\bar{P})}$$

This gives

$$\boxed{\frac{\bar{P}-\pi}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n}}} \sim N(0,1)}$$

Interval Estimation for population proportions

Confidence interval

- The (1 - α) x 100% confidence interval for the population proportion

$$\boxed{\left[\bar{p}-z\times\sqrt{\frac{\vec{p}(1-\vec{p})}{n}},\bar{p}+z\times\sqrt{\frac{\vec{p}(1-\vec{p})}{n}}\right]}$$

$P(-z \le Z \le z)=1-\alpha$ with Z~N(0,1)

Example

Determine a 95% C.I. for the proportion of female in the population

Step 1: *choose a confidence level $1-\alpha = 0.95$*

Step 2: *decision on the basis of the data in which case you are* :

no normal distribution and unknown σ^2 but large sample size

(592 x 0.5287 >> 5 and 592 x (1-0.52872) >> 5)

→ case 2, thus normal distribution

Example: 95% C.I for the population proportion

$$X = \begin{cases} 1 & \text{When sex is female} \\ 0 & \text{When the sex is male} \end{cases}$$

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	313	52.872	313	52.872
0	279	47.128	592	100.00

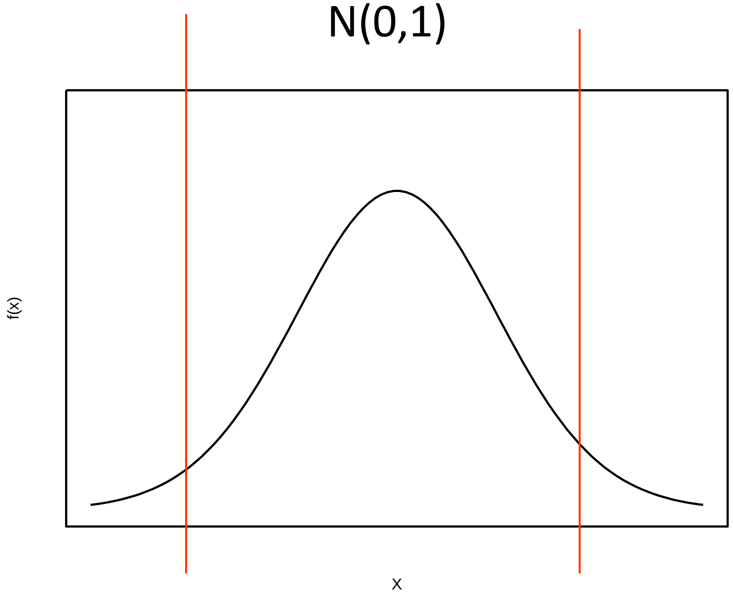
Step 4: *calculate the point estimators for μ and σ ²*

$$\bar{p} = 0.2139$$

$$\frac{\bar{p}(1-\bar{p})}{n} = \frac{0.2139(1-0.2139)}{173}$$

Step 3: *Find the critical points in the correct table: -1.96 and 1.96.*

N(0,1)



-1.96 **1.96**

Step 5: *calculate the confidence interval :*

$$\boxed{\left[\bar{p}-z\times\sqrt{\frac{\vec{p}(1-\vec{p})}{n}},\ \bar{p}+z\times\sqrt{\frac{\vec{p}(1-\vec{p})}{n}}\right]}$$

$$\left[0.5287-1.96\underbrace{\sqrt{\frac{0.5287(1-0.5287)}{592}}}_{},0.5287-1.96\underbrace{\sqrt{\frac{0.5287(1-0.5287)}{592}}}_{}\right]$$

$$\boxed{L=0.4885} \qquad \boxed{R=0.5689}$$

Example

Decision:

The 95% CI for the proportion of females in the population, is [0.4885,0.5689]

interpretation:

We are 95% confident that between 48.85% and 56.89% of the population are females.

Confidence interval for population proportion using R

```
> malefemale=apply(HairEyeColor,3,sum)
> n=sum(malefemale)
> pbar=malefemale[2]/n
> SE=sqrt(pbar*(1-pbar)/n)
> E=qnorm(0.975)*SE
> pbar+c(-E,E)
```

```
 Female    Female
0.4885057 0.5689267
```

OR

```
> library(TeachingDemos)
> malefemale=apply(HairEyeColor,3,sum)
> prop.test(malefemale[2], sum(malefemale), correct=F)
```

Hypothesis testing for a proportion

example 1

facebook accounts of students in Belgium

- According to "World web stat" there are 10,431,477

resident in Belgium (2011), 77.8% are Internet users and 4,444,500 have FACEBOOK account (42.60% in December 2011) (http://www.internetworldstats.com/stats4.htm#europe)

- A researcher wants the proportion of students in Belgium with FACEBOOK account estimating and testing the hypotheses that more than 40% of the students have FACEBOOK account.

A hypothesis about a population proportion

- The general principles on keys remain here valid.
- Typical test problems :

$$\begin{aligned}
&(a) \quad H_0 : \pi = \pi_{H_0} \quad versus \quad H_1 : \pi < \pi_{H_0} \\
&(b) \quad H_0 : \pi = \pi_{H_0} \quad versus \quad H_1 : \pi > \pi_{H_0} \\
&(c) \quad H_0 : \pi = \pi_{H_0} \quad versus \quad H_1 : \pi \neq \pi_{H_0}
\end{aligned}$$

π_{H_0} which is the value of the population proportion if H_0 is true.

The distribution of the sample proportion

For large samples, we use the property that, if H_0 is true,

$$\boxed{\frac{\overline{P}-\pi_{H_0}}{\sqrt{\frac{\pi_{H_0}(1-\pi_{H_0})}{n}}} \sim N(0,1)}$$

example 1

facebook accounts of students in Belgium

- According to "World web stat" there are 10,431,477 resident in Belgium (2011), 77.8% are Internet users and 4,444,500 have FACEBOOK account (42.60% in December 2011) (http://www.internetworldstats.com/stats4.htm#europe
- A researcher wants the proportion of students in Belgium with FACEBOOK account estimating and testing the hypotheses that more than 40% of the students have FACEBOOK account.

$$H_0 : \pi = 0.40$$

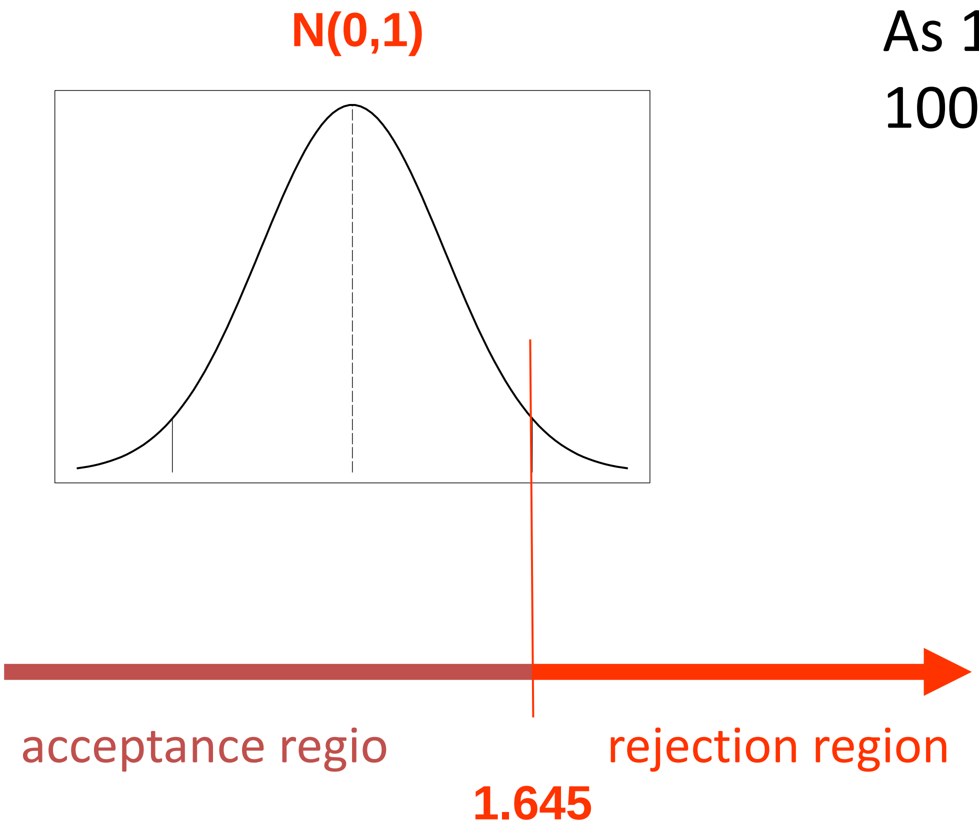
$$H_1 : \pi > 0.40$$

The sample

- The researcher draws a sample of 100 students (Internet users) in Belgium.
- In one sample, 47 students have a Face book account.
- On the basis of this sample, we can reject the null hypothesis that the proportion of students in Belgium with face book account is 40%?

Example: the rejection region

For a significance level α = 0.05 is the rejection region given by



As 100 x 0.4> 5 x 0.6 and 100> 5, we use that

$$\boxed{\frac{\overline{P}-0.4}{\sqrt{\frac{0.4(1-0.4)}{100}}} \sim N(0,1)}$$

example

The observed value of the statistic is

$$\boxed{\frac{\bar{p}-\pi_{H_0}}{\sqrt{\frac{\pi_{H_0}(1-\pi_{H_0})}{n}}}=\frac{0.47-0.4}{\sqrt{\frac{0.4\times 0.6}{100}}}=1.428869}$$

For a significance level α = 0.05: 1.4288 <1.645.

Conclusion?

R code

```
> pbar = 0.47
> prop = 0.4
> n = 100
> z=(pbar-prop)/sqrt((prop*(1-prop))/n)
> z
[1] 1.428869
> alpha=0.05
> crit.point=qnorm(1-alpha)#p=0.05 one tailed (upper)
> crit.point
[1] 1.644854
```

Test statistic

Critical point

The checklist

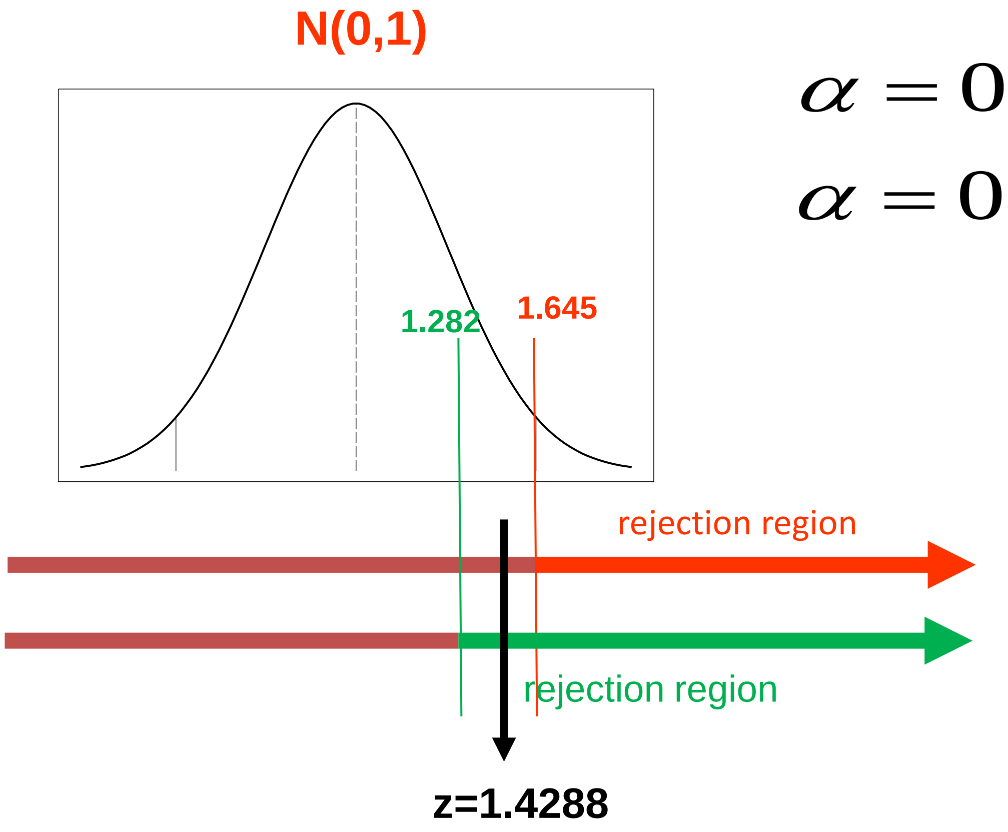
Stap	Information	example
1	The hypotheses (the testing problem)	$H_0 : \pi = 0.4$ $H_1 : \pi > 0.4$ One-sided test
2	The level of significance	$\alpha = 0.05$
3	The test statistic	$\boxed{\dfrac{\bar{P} - \pi_{H_0}}{\sqrt{\dfrac{\pi_{H_0}(1-\pi_{H_0})}{100}}} \sim N(0,1)}$
4	The distribution of the test statistic under H_0	
5	The critical point (or points)	1.645 N(0,1)

Example: the rejection region

α=0.05 , z=1.645

α=0.1 , z=1.282

$$\boxed{z_p = 1.428869}$$



$$\alpha = 0.05 : 1.4288 < 1.645$$

$$\alpha = 0.10 : 1.4288 > 1.282$$

EXAMPLE: p-value

The answer expressed with a p-value is as follows:

$$\boxed{p - value = P(Z > 1.4288) = 0.07652094}$$

$$\alpha = 0.05 : p - value > \alpha$$

$$\alpha = 0.10 : p - value < \alpha$$

R code

```
> pbar = 0.47
> prop = 0.4
> n = 100
> z=(pbar-prop)/sqrt((prop*(1-prop))/n)
> z
[1] 1.428869
> alpha=0.1
> crit.point1=qnorm(1-alpha)#p=0.1 one tailed (upper)
> crit.point1
[1] 1.281552
> pval = 1-pnorm(z, lower.tail=TRUE)  # upper tail
> pval
[1] 0.07652094
```

Example 2: two-tailed test

- In a sample of 1000 women from the aged 50 to 54 whose mother had breast cancer, 40 were found with breast cancer.
- Suppose that the overall prevalence rate for breast cancer in women of that age (regardless of their family history) 2%.

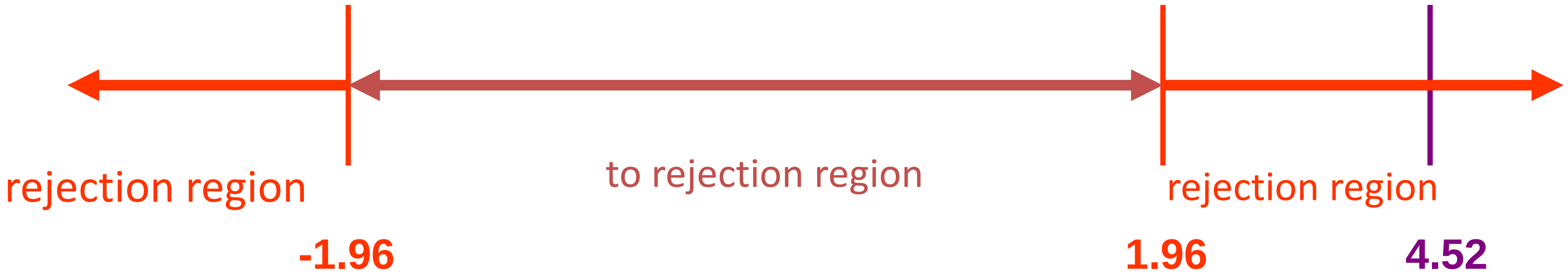
The checklist

Step	information	example
1	The hypotheses (the qualifying problem)	two-tailed test
2	The level of significance	
3	The test statistic	
4	The distribution of the test statistic under H_0	
5	The critical point (or points)	

Example: the rejection region

Since 4.52 in the rejection region is (4.52> 1.96), we reject the null hypothesis at 5% significance level.

$$\boxed{\frac{\overline{P}-\pi_{H_0}}{\sqrt{\frac{\pi_{H_0}(1-\pi_{H_0})}{1000}}}=4.52}$$



EXAMPLE: p-value

The answer expressed with a p-value is as follows:

$$\boxed{p-value = 2\times P(Z > 4.52) = 2\times[1-\Phi(4.52)] \approx 0.000}$$

The result is (very) significant.

R code

```
> pbar1 = 0.04      # sample proportion
> prop1 = 0.02      # hypothesized value
> n = 1000          # sample size
> z1=(pbar1-prop1)/sqrt((prop1*(1-prop1))/n)
> z1                # test statistic
[1] 4.51754
> pval = 2*pnorm(z1, lower.tail=FALSE)
> pval
[1] 6.256236e-06
```

Example

- the number of hours of sleep for each of 24 students in class.
- if the student got at least 9 hours of sleep(yes).

```
>sleep=c(7.75,8.5,8,6,8,6.33,8.17,7.75,7,6.5,8.75,8,7.5,3,6.25,8.5,9,6.5,9,9.5,9,8,8,9.5)
> nine.hrs=ifelse(sleep>=9,"yes","no")
> table(nine.hrs)
nine.hrs
 no yes
 19   5
> y=5;n=24
> test=prop.test(y,n,p=0.5,alternative="two.sided",correct=FALSE)
> test
```

```
  > test

 1-sample proportions test without continuity correction

data:  y out of n, null probability 0.5
X-squared = 8.1667, df = 1, p-value = 0.004267
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.09244825 0.40470453
sample estimates:
        p
0.2083333
```

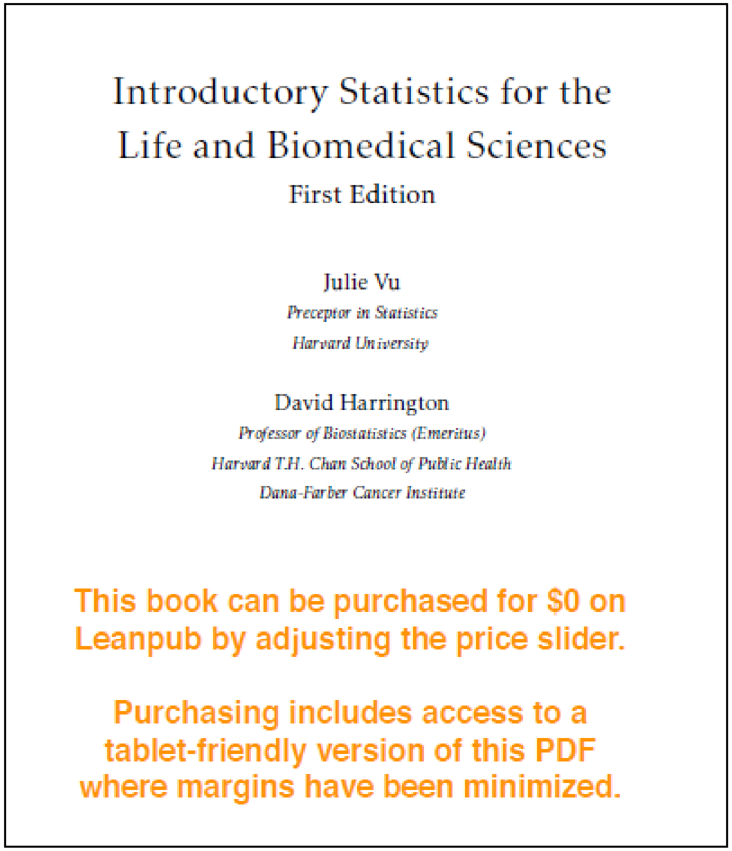
8.2: Inference for the difference of two proportions

Hypothesis tests and Confidence intervals for Multiple populations

Objectives

- To distinguish between a problem associated with measurements and a two-sample problem using example.
- To perform a test of hypothesis about the difference of two population means and two population proportions.
- To calculate a confidence interval for the difference of two population means and the difference of two population proportions.
- The tests and confidence intervals can perform and interpret using R.

Inference for a difference of two proportion



Introductory Statistics for the
Life and Biomedical Sciences

First Edition

Julie Vu
Preceptor in Statistics
Harvard University

David Harrington
Professor of Biostatistics (Emeritus)
Harvard T.H. Chan School of Public Health
Dana-Farber Cancer Institute

This book can be purchased for $0 on Leanpub by adjusting the price slider.

Purchasing includes access to a tablet-friendly version of this PDF where margins have been minimized.

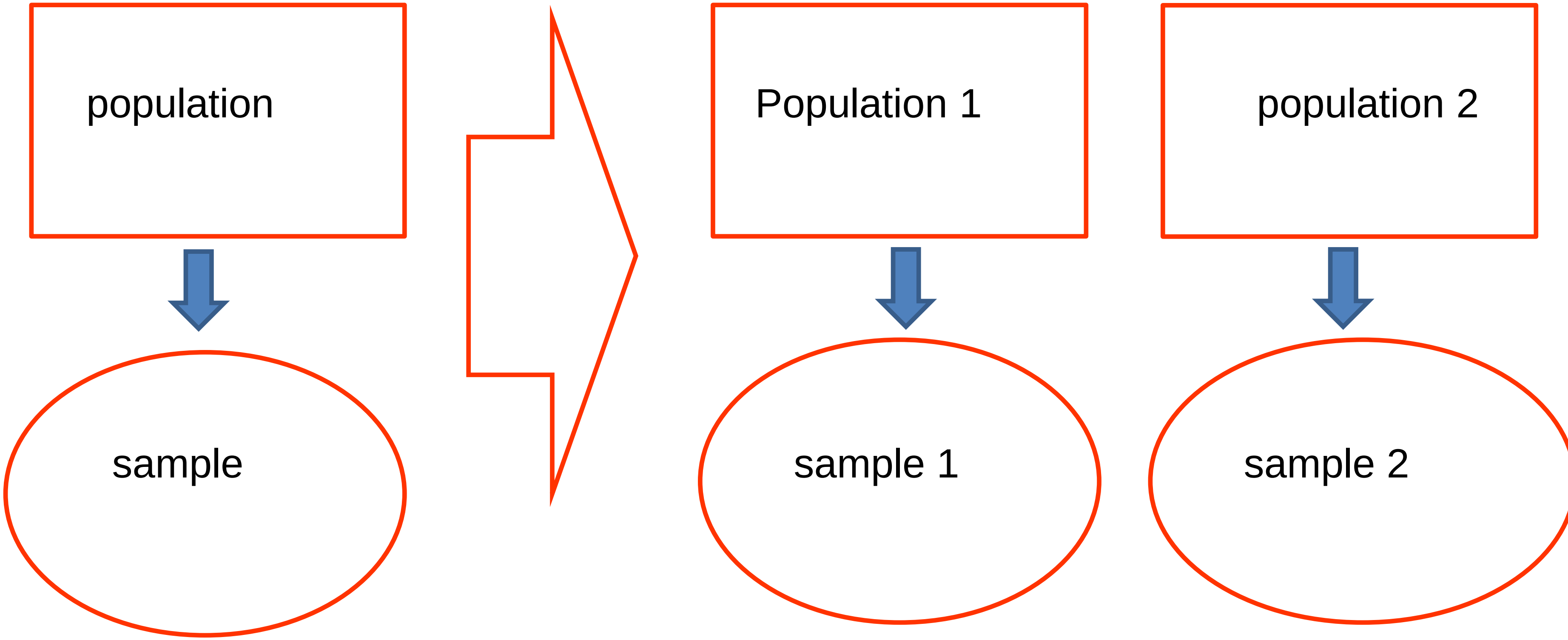
Section 8.2

Section 8.1 → Section 8.2

Section 8.1

Section 8.2

Twee populaties



population

sample

Population 1

population 2

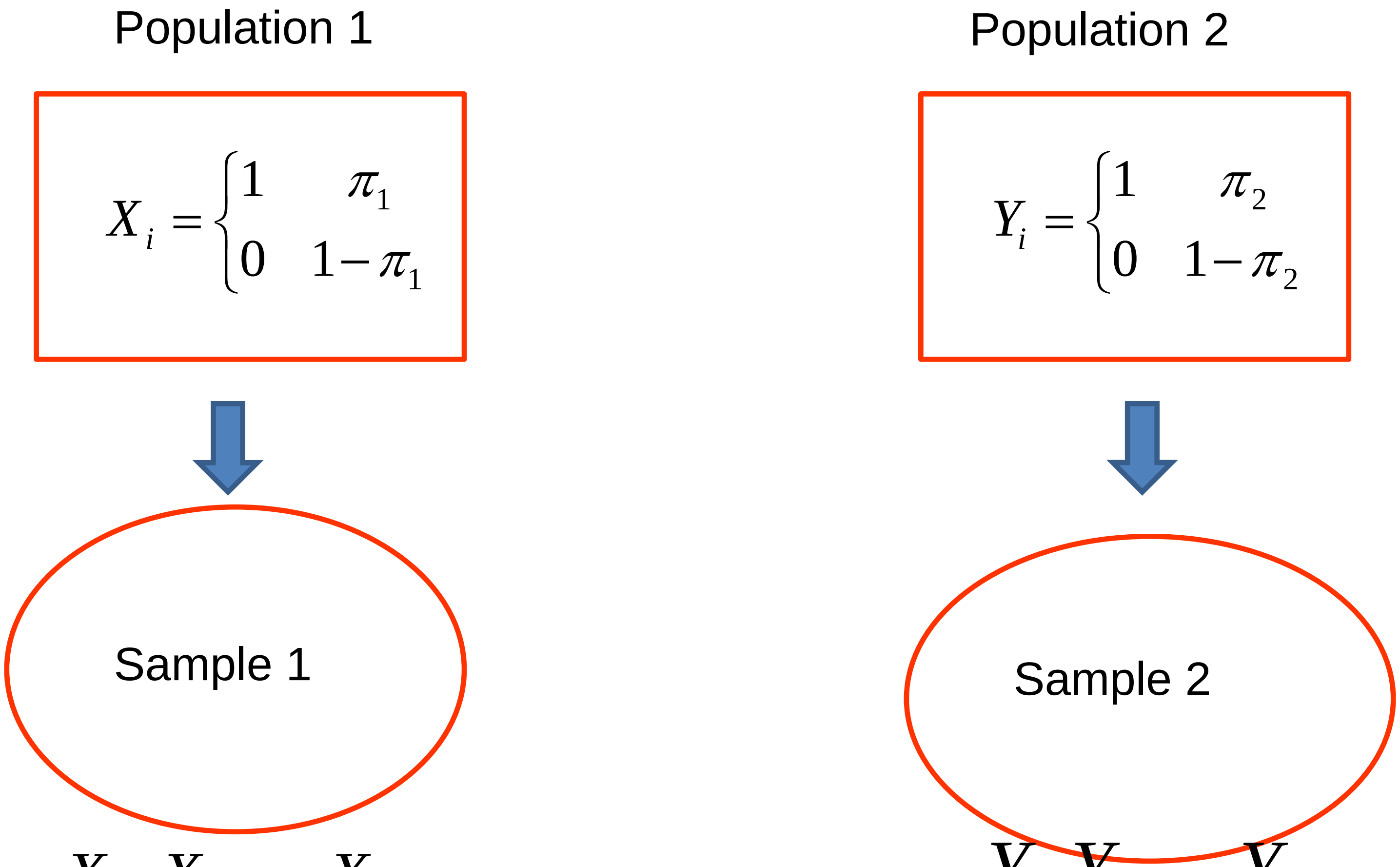
sample 1

sample 2

Two independent samples

Comparing two population proportions

Two populations and two proportions



Population 1

$$X_i = \begin{cases} 1 & \pi_1 \\ 0 & 1-\pi_1 \end{cases}$$

Sample 1

$$X_1, X_2, \ldots, X_{n_1}$$

Population 2

$$Y_i = \begin{cases} 1 & \pi_2 \\ 0 & 1-\pi_2 \end{cases}$$

Sample 2

$$Y_1, Y_2, \ldots, Y_{n_2}$$

The testing problem

The null hypothesis that we want to test

$$H_0 : \pi_2 - \pi_1 = 0$$

versus an alternative hypothesis

$$\left.\begin{array}{ll} (a) & H_1 : \pi_2 - \pi_1 < 0 \\ (b) & H_1 : \pi_2 - \pi_1 > 0 \end{array}\right\} \quad \text{One sided}$$

$$(c) \quad H_1 : \pi_2 - \pi_1 \neq 0 \qquad \text{Two sided}$$

8.2.1: Sampling distribution of the difference of two proportions

The sample proportions

- Suppos S_1 the number of successes in the first sample and S_2 the number of successes in the second sample.
- The proportions of the sample are then given by

$$\overline{P}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{the\,number\,of\;times\,1\,in\,sample_1}{n_1}$$

$$\boxed{\overline{P}_1 = \frac{S_1}{n_1}} \qquad \boxed{\overline{P}_2 = \frac{S_2}{n_2}}$$

a sample from the first population

$$E(\overline{P}_1) = \pi_1$$

$$Var(\overline{P}_1) = \frac{\pi_1(1-\pi_1)}{n_1}$$

a sample from the second population

$$E(\overline{P}_2) = \pi_2$$

$$Var(\overline{P}_2) = \frac{\pi_2(1-\pi_2)}{n_2}$$

The average and the variance of the difference

$$E(\overline{P}_2 - \overline{P}_1) = \pi_2 - \pi_1$$

$$Var(\overline{P}_1 - \overline{P}_2) = Var(\overline{P}_2) + Var(\overline{P}_1) = \frac{\pi_2(1-\pi_2)}{n_2} + \frac{\pi_1(1-\pi_1)}{n_1}$$

The proportion under H_0

The null hypothesis says that $\pi_1=\pi_2=\pi$ (for some unknown value π).

So H_0 is true:

$$E(\overline{P}_2-\overline{P}_1)=0$$

$$Var(\overline{P}_2-\overline{P}_1)=Var(\overline{P}_2)+Var(\overline{P}_1)=\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

The pooled sample proportion

Problem is that π is not known. An estimator for p is given by the pooled sample proportion

$$\boxed{\overline{P} = \frac{n_1\overline{P}_1 + n_2\overline{P}_2}{n_1 + n_2} = \frac{S_1 + S_2}{n_1 + n_2}}$$

The test statistic

$$\boxed{\frac{\overline{P}_2-\overline{P}_1-(p_2-p_1)_{H_0}}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}=\frac{\overline{P}_2-\overline{P}_1}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\approx N(0,1)}$$

unknown parameter

$$\min(n_1,n_2)\overline{p}>5$$

$$\min(n_1,n_2)(1-\overline{p})>5$$

8.2.3: Comparing two population proportions

The test statistic

If n_1 and n_2 are sufficiently large, then

$$\boxed{\frac{\overline{P}_2 - \overline{P}_1}{\sqrt{\overline{P}(1-\overline{P})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1)}$$

Example 5:
Two-sided testing problem

- A study on the effect of oral contraceptives (OC) on the occurrence of myocardial infarction (MI) in women (40-44 years) produced the following results (for an observation period of 3 years):

1. From 5000 OC users 13 women were with MI
2. 10000 non - OC users 7 women were with MI

- What is the statistical significance of these results?

solution

- π_1 is the population proportion of MI in OC users
- π_2 is the population proportion of MI in non-OC users.

$$H_0 : \pi_2 - \pi_1 = 0$$ Null hypothesis

$$H_1 : \pi_2 - \pi_1 \neq 0$$ Alternative hypothesis

Sample proportion

$$n_1 = 5000 \qquad \bar{p}_1 = \frac{13}{5000} = 0.0026$$

$$n_2 = 10000 \quad \bar{p}_2 = \frac{7}{10000} = 0.0007$$

$$\bar{p} = \frac{13+7}{15000} = 0.00133$$

the pooled sample proportion

The test statistic

$$\boxed{\frac{\overline{P}_1 - \overline{P}_2}{\sqrt{\overline{P}(1-\overline{P})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1)}$$

$$\min(n_1, n_2)\,\overline{p} = 5000 \times 0.00133 > 5$$

$$\min(n_1, n_2)(1-\overline{p}) = 5000 \times 0.99867 > 5$$

The test statistic

the value of test of statistic is:

$$\boxed{\frac{0.0007-0.0026}{\sqrt{0.00133\times 0.99867\left(\frac{1}{5000}+\frac{1}{10000}\right)}}=-3.01}$$

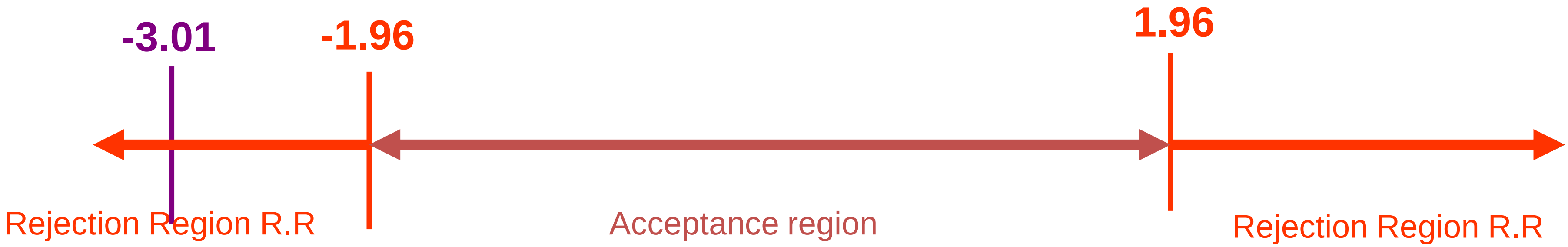
$$\boxed{\frac{\overline{p}_1-\overline{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\sim N(0,1)}$$

The rejection region (two-tailed test)

$$\frac{0.0007-0.0026}{\sqrt{0.00133\times 0.99867\left(\frac{1}{5000}+\frac{1}{10000}\right)}}=-3.01$$

The test statistic

-3.01< -1.96 ⇨ we reject H_0 at significance level 0.05.



2-sided test

The p-value of the two-sided test is given by :

$$p = 2 \times P(Z < -3.01) = 2 \times \left[1 - \Phi(3.01)\right] = 2 \times 0.0013 = 0.0026 < \alpha$$

So there is a very significant difference between the occurrence of MI in OC users and non-OC users (α = 0.05).

Comparing two population proportions using R

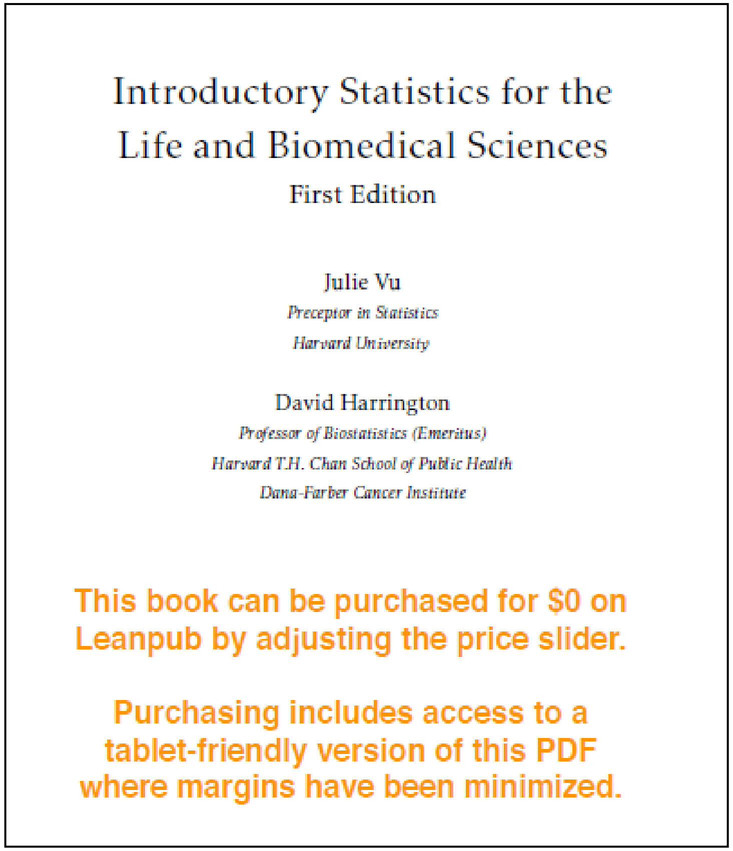
```
> library(MASS)
> prop.test(c(7,13),c(10000,5000), correct = F)
```

```
2-sample test for equality of proportions without
        continuity correction

data:  c(7, 13) out of c(10000, 5000)
X-squared = 9.037, df = 1, p-value = 0.002646
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.0034036884 -0.0003963116
sample estimates:
prop 1 prop 2
0.0007 0.0026
```

The checklist

Step	Information	example
1	Test of Hypothesis	$H_0 : \pi_2 - \pi_1 = 0$ $H_1 : \pi_2 - \pi_1 \neq 0$
2	Determine case	$\min(n_1, n_2)\bar{p} = 5000 \times 0.00133 > 5$ $\min(n_1, n_2)(1 - \bar{p}) = 5000 \times 0.99867 > 5$
3	The test statistic The distribution of the test statistic under the null hypothesis	$\boxed{\dfrac{\bar{P}_1 - \bar{P}_2}{\sqrt{\bar{P}(1-\bar{P})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim N(0,1)}$
4	The level of significance	$\alpha = 0.05$
5	The critical point (or points) & R.R	-1.96 & 1.96 N(0,1)
6	Calculate the test statistic	-3.01
7	Conclusion	Reject Ho

Inference for two or more groups



Introductory Statistics for the Life and Biomedical Sciences

First Edition

Julie Vu

Preceptor in Statistics

Harvard University

David Harrington

Professor of Biostatistics (Emeritus)

Harvard T.H. Chan School of Public Health

Dana-Farber Cancer Institute

This book can be purchased for $0 on Leanpub by adjusting the price slider.

Purchasing includes access to a tablet-friendly version of this PDF where margins have been minimized.

- Chi-squared test for independence.
- Analysis of I x J contingency tables.

Section 8.3

Analysis of IxJ contingency tables

- The main goal of analysing a contingency table is to test independence between rows and columns.

- In our case study, the null hypothesis is that there is no association between anaemia prevalence and socio-economic status. Therefore, the distribution of outcome categories should be independent of the explanatory variable

Analysis of IxJ contingency tables

- **2 x 2 contingency table**

Explanatory	Outcome		Total
	Yes	No	
A	n_{11}	n_{12}	n_{1+}
B	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

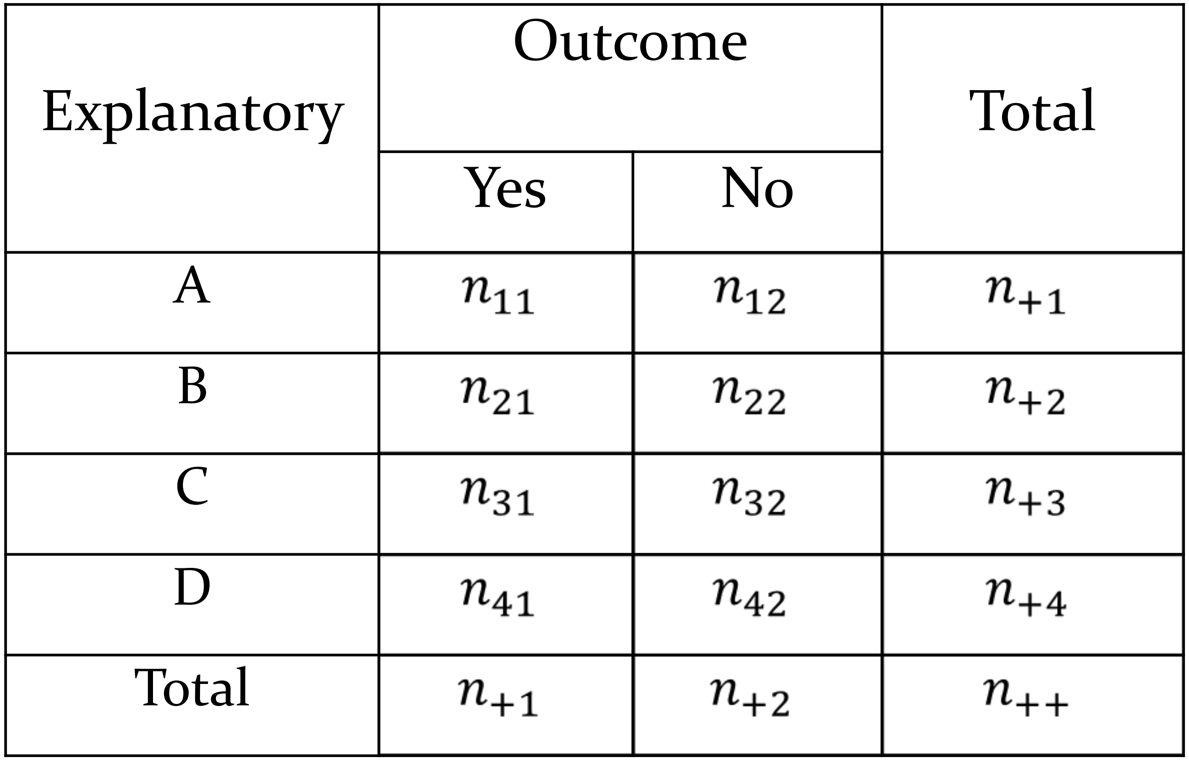
Analysis of IxJ contingency tables

- **2 x 2 contingency table**

Explanatory	Outcome		Total
	Yes	No	
A	n_{ij}		n_{i+}
B			
Total	n_{+j}		n_{++}

Analysis of IxJ contingency tables

- **4 x 2 contingency table**



Explanatory	Outcome		Total
	Yes	No	
A	n_{11}	n_{12}	n_{+1}
B	n_{21}	n_{22}	n_{+2}
C	n_{31}	n_{32}	n_{+3}
D	n_{41}	n_{42}	n_{+4}
Total	n_{+1}	n_{+2}	n_{++}

Analysis of IxJ contingency tables

- **4 x 3 contingency table**

Explanatory	Outcome			Total
	Large	Medium	Small	
A	n_{11}	n_{12}	n_{13}	n_{1+}
B	n_{21}	n_{22}	n_{23}	n_{2+}
C	n_{31}	n_{32}	n_{33}	n_{3+}
D	n_{41}	n_{42}	n_{43}	n_{4+}
Total	n_{+1}	n_{+2}	n_{+3}	n_{++}

Analysis of IxJ contingency tables

- Independence test in a generalised two-way contingency tables of **nominal** outcomes can be tested using;

$$H_0 \colon \pi_{ij} = \pi_{i+}\pi_{+j}$$

$$\pi_{ij} = \frac{n_{ij}}{n_{++}} \quad ; \quad \pi_{i+} = \frac{n_{i+}}{n_{++}} \quad ; \quad \pi_{+j} = \frac{n_{j+}}{n_{++}}$$

- If the independent assumptions holds, then the distribution of the cell counts is independent of the rows and the columns.

Probability under independence

$$H_0\colon \pi_{ij} = \pi_{i+}\pi_{+j}$$

- For two independent events:

$$P(A \cap B) = P(A) \times P(B)$$

- In a I X J table :

$$P(X = i \cap Y = j) = P(X = i) \times P(Y = j)$$

$$\pi_{ij} = \pi_{i+} \times \pi_{+j}$$

Analysis of IxJ contingency tables

- Under the null model we can calculate the expected cell frequencies ($\hat{\mu}_{i\,i}$) as:

$$n_{++} \times \hat{\pi}_{ij} = n_{++} \times \left(\hat{\pi}_{j+} \hat{\pi}_{+j} \right) = n_{++} \times \frac{n_{i+}}{n_{++}} \times \frac{n_{+j}}{n_{++}} \quad \Rightarrow \quad \hat{\mu}_{ij} = \frac{n_{i+} n_{+j}}{n_{++}}$$

- We can use $Chi - square\ test$ to compare the expected frequencies under the null model with the observed frequencies:

Analysis of IxJ contingency tables

- **Pearson Chi-square statistics**

$$X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}; \qquad X^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

- ➢ O_{ij} = observed cell counts for row i and column j
- ➢ E_{ij} = Expected cell counts for row i and column j
- ➢ X^2 ~Chi-Square distribution with $(I-1)(J-1)$ degree of freedom (df)

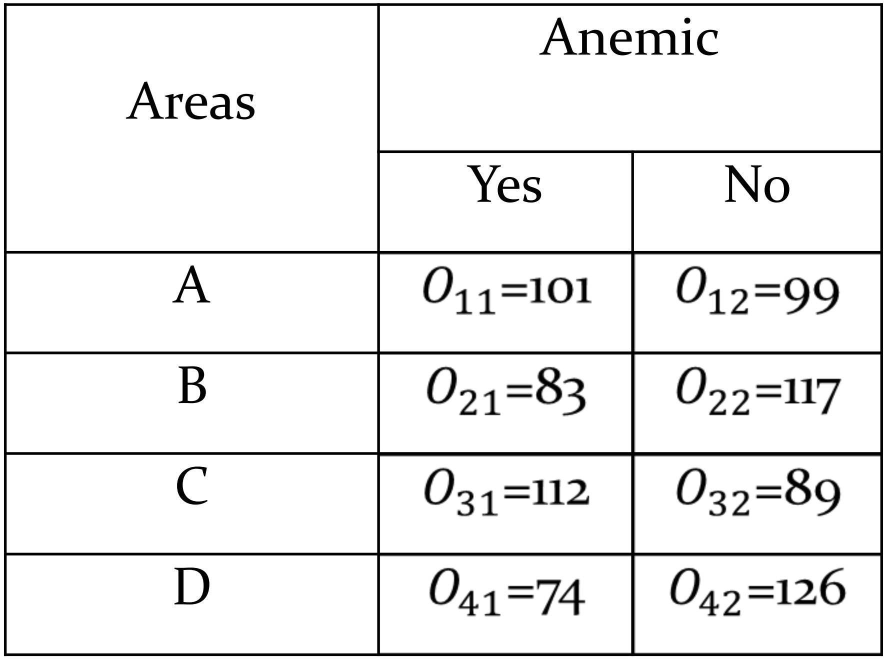
Analysis of IxJ contingency tables

- Investigate whether there is association between child location and child anaemia.

Areas	Anemic		Total
	Yes	No	
A	101	99	200
B	83	117	200
C	112	89	201
D	74	126	200
Total	370	431	801

Analysis of IxJ contingency tables

- Matrix of the observed cell counts (O_{ij})



Areas	Anemic	
	Yes	No
A	O_{11}=101	O_{12}=99
B	O_{21}=83	O_{22}=117
C	O_{31}=112	O_{32}=89
D	O_{41}=74	O_{42}=126

Analysis of IxJ contingency tables

- Matrix of the expected values (E_{ij}).

Areas	Anemic	
	Yes	No
A	$E_{11} = \frac{200*370}{801}$=92.4	$E_{12} = \frac{200*431}{801}$=107.6
B	$E_{21} = \frac{200*370}{801}$=92.4	$E_{22} = \frac{200*431}{801}$=107.6
C	$E_{31} = \frac{201*370}{801}$=92.8	$E_{32} = \frac{201*431}{801}$=108.2
D	$E_{41} = \frac{200*370}{801}$=92.4	$E_{42} = \frac{200*431}{801}$=107.6

Chi-square test in R

```
areaAnemic <- table(nonMissingAnemic$Areas, nonMissingAnemic$Child_Anemic,
        exclude=FALSE)
nplus. <- rowSums(areaAnemic)
n.plus <- colSums(areaAnemic)
npluplus <- sum(areaAnemic)
Oij <- areaAnemic
Eij <- (nplus.%*%t(n.plus))/npluplus
tmp  <- ((Oij-Eij)^2)/Eij
X2 <- sum(tmp)
df <- (nrow(areaAnemic)-1)*(ncol(areaAnemic)-1)
pvalue <- pchisq(X2, df,lower.tail = FALSE))
```

Chi-square test in R

- **Results**

- $X^2 = 17.4$
- Pvalue = 0.0006

- **Interpretation**

There is a significant association between child location and child anaemia.

Chi-square test in R

- Definition of the variables

```
> Anemic<- as.factor(c(rep("Yes",101),rep("No",99),rep("Yes",83),rep("No",117)
                    ,rep("Yes",112),rep("No",89),rep("Yes",74),rep("No",126)))
> Areas<-as.factor(c(rep("A",101),rep("A",99),rep("B",83),rep("B" ,117),rep("C",112),rep("C"
      ,89),rep("D",74),rep("D" ,126)))
```

Chi-square test in R

- Chi-square for independence

```
> areaAnemic<-table(Anemic,Areas)
> areaAnemic
      Areas
Anemic   A   B   C   D
   No   99 117  89 126
   Yes 101  83 112  74
```

```
> chiArea <- chisq.test(areaAnemic,correct = FALSE)
> chiArea

        Pearson's Chi-squared test

data:  areaAnemic
X-squared = 17.4074, df = 3, p-value = 0.0005827
```

Example: A 2 X 2 table

- Suppose we are interested in investigating whether younger children were more prone to anaemia than the older children.
- Data were collected about 779 subjects.
- Variables of interest: Anaemia (Yes/No), age.
- We need to create a contingency or a cross tabulation table with the outcome variable (Anaemia) on the columns and the explanatory variable (Age category of children) on the rows.

Example: a 2 X 2 table

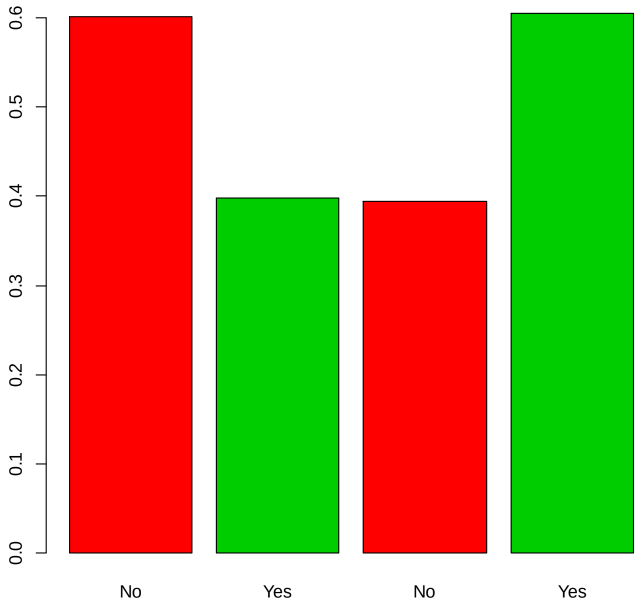
Age categories	Anemic		Total
	Yes	No	
6-23 months	259	169	428
24-59 months	310	469	779

$$H_0 : P_M = P_F$$

$$H_1 : P_M \neq P_F$$

$$P_M = P_M\left(\text{Child Anemic}\right)$$

$$P_F = P_F\left(\text{Child Anemic}\right)$$



Risk Difference: estimation

- $Risk\ Difference\ (RD) = \widehat{p_1} - \widehat{p_2}$

$$\widehat{p_1} = \frac{n_{11}}{n_{1+}} = \frac{259}{428} = 0.605$$

$$\widehat{p_2} = \frac{n_{21}}{n_{2+}} = \frac{310}{779} = 0.397$$

Age categories	Anemic		Total
	Yes	No	
6-23 months	259	169	428
24-59 months	310	469	779

- $Risk\ Difference\ (RD) = \widehat{p_1} - \widehat{p_2} = 0.605 - 0.379$

$$RD = 0.208$$

Risk Difference

- Test for independence

$$Z = \frac{p_{1-}\, p_2}{\sqrt{\frac{p_1(1-p_1)}{n_{1+}} + \frac{p_2(1-p_2)}{n_{2+}}}} = -7.041399$$

- Two sided test, $\alpha = 0.05$, p-value= <0.001

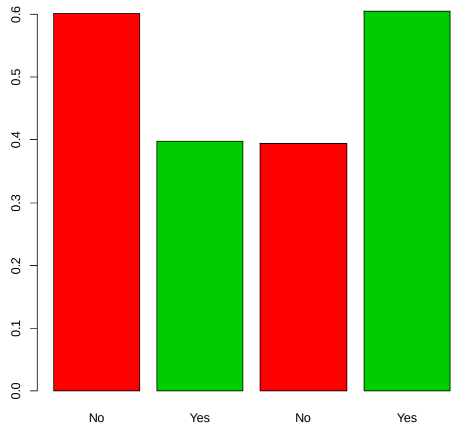
- Note that Z is approximated with standard Normal distribution $N(0,1)$

Risk Difference in R

```
> RDanemic <- prop.test(x=ageAnemic[,2],  n= rowSums(ageAnemic) ,
correct = FALSE)
>
> RDanemic

        2-sample test for equality of proportions without
continuity
        correction

data:  ageAnemic[, 2] out of rowSums(ageAnemic)
X-squared = 47.5894, df = 1, p-value = 5.255e-12
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2648663 -0.1495219
sample estimates:
   prop 1    prop 2 
0.3979461 0.6051402 
```

Example: a 2 X 2 table

Age categories	Anemic		Total
	Yes	No	
6-23 months	259	169	428
24-59 months	310	469	779

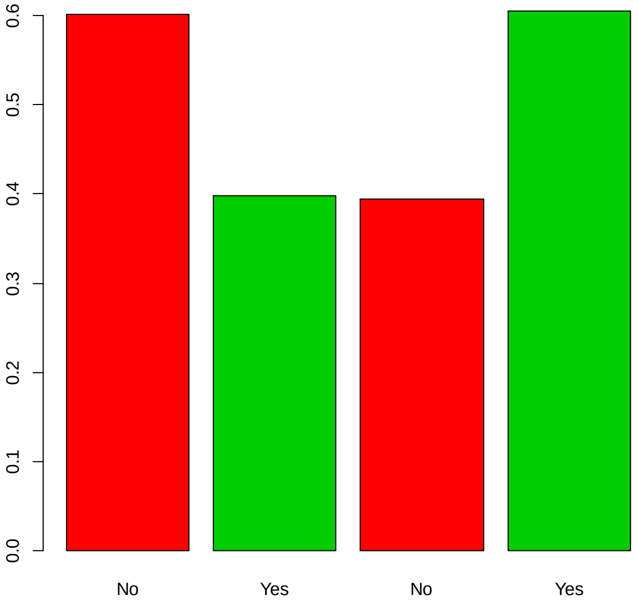


```
> Oij <- ageAnemic
> Oij

                No Yes
  24-59 months 469 310
  6-23 months  169 259
>
> nplus. <- rowSums(ageAnemic)
> nplus.
24-59 months  6-23 months
         779          428
> n.plus <- colSums(ageAnemic)
> n.plus
 No Yes
638 569
> npluplus <- sum(ageAnemic)
> npluplus
[1] 1207
```

```
> Eij <- (nplus.%*%t(n.plus))/npluplus
> Eij
          No      Yes
[1,] 411.7664 367.2336
[2,] 226.2336 201.7664
>
> tmp  <- ((Oij-Eij)^2)/Eij
> X2 <- sum(tmp)
> X2
[1] 47.58941
```

Example: a 2 X 2 table

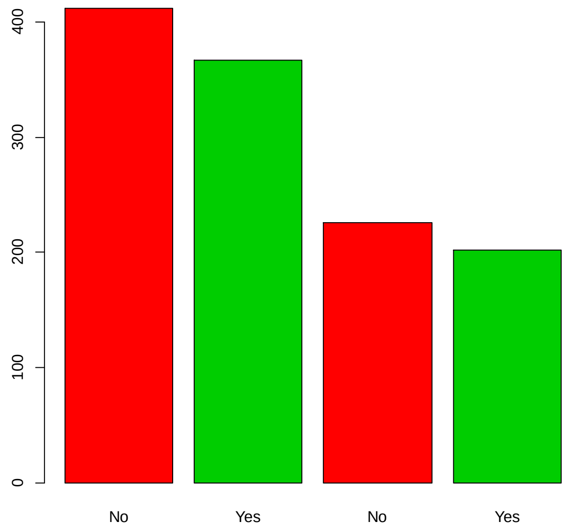
Age categories	Anemic		Total
	Yes	No	
6-23 months	259	169	428
24-59 months	310	469	779



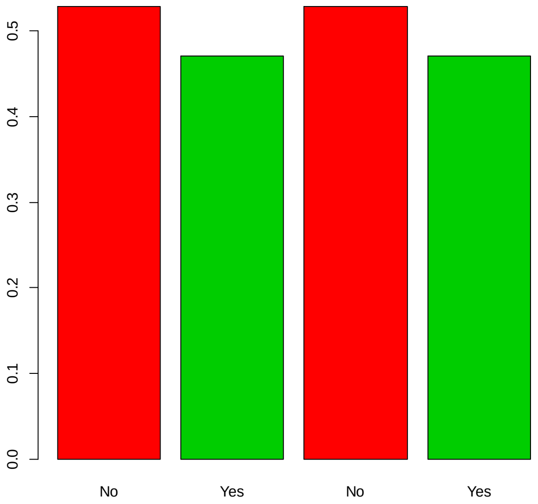
```
 chi.sq <- chisq.test(ageAnemic,correct = FALSE)
> chi.sq

        Pearson's Chi-squared test

data:  ageAnemic
X-squared = 47.5894, df = 1, p-value = 5.255e-12
```

Example: a 2 X 2 table – OR for rhe expected table

counts



proportions

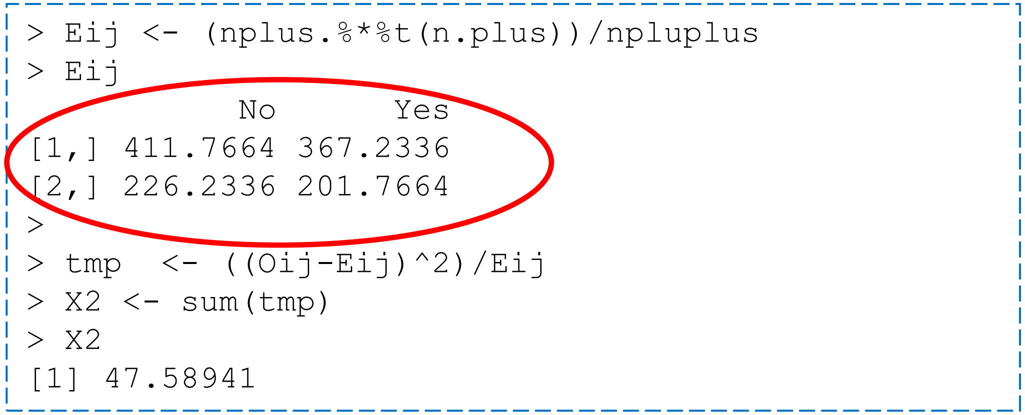


```
> Eij <- (nplus.%*%t(n.plus))/npluplus
> Eij
           No      Yes
[1,] 411.7664 367.2336
[2,] 226.2336 201.7664
>
> tmp  <- ((Oij-Eij)^2)/Eij
> X2 <- sum(tmp)
> X2
[1] 47.58941
```

Expected value

$$\hat{\mu}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$$

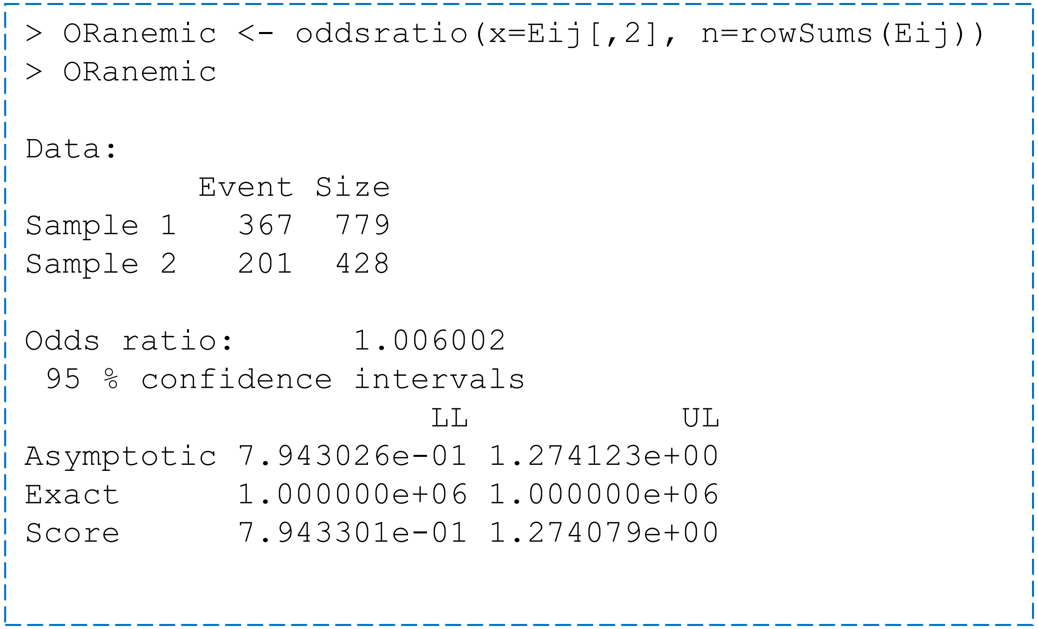
Example: a 2 X 2 table – OR for the expected table



```
> Eij <- (nplus.%*%t(n.plus))/npluplus
> Eij
              No       Yes
[1,] 411.7664 367.2336
[2,] 226.2336 201.7664
>
> tmp  <- ((Oij-Eij)^2)/Eij
> X2 <- sum(tmp)
> X2
[1] 47.58941
```

Expected value

$$\hat{\mu}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$$



```
> ORanemic <- oddsratio(x=Eij[,2], n=rowSums(Eij))
> ORanemic

Data:
         Event Size
Sample 1   367  779
Sample 2   201  428

Odds ratio:       1.006002
 95 % confidence intervals
                     LL           UL
Asymptotic 7.943026e-01 1.274123e+00
Exact      1.000000e+06 1.000000e+06
Score      7.943301e-01 1.274079e+00
```

OR for the expected table !!

Why OR=1 ?