# Basic Statistics: Inference about proportions and rates

# Topics

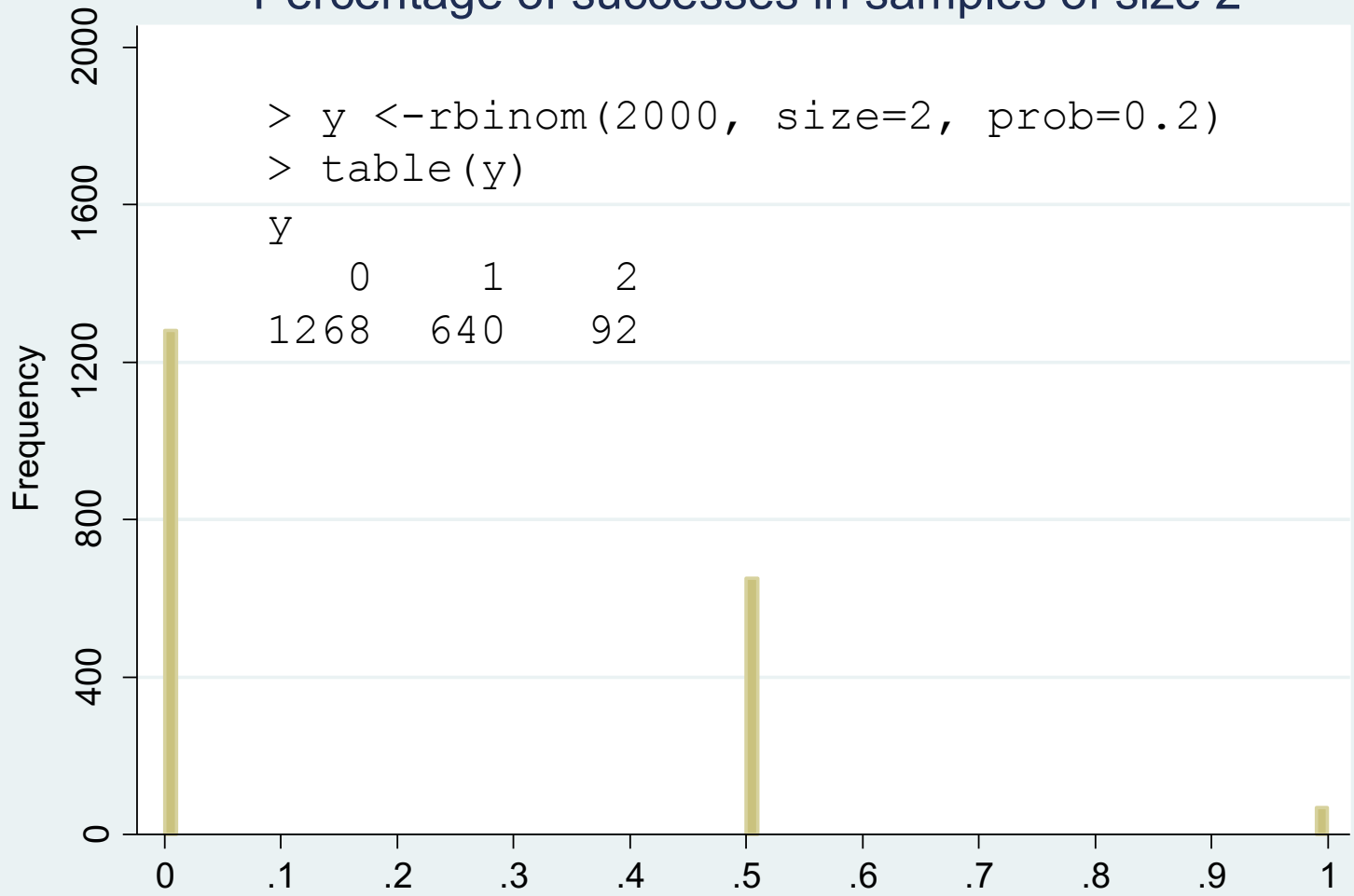
- Dichotomous data: binomial distribution, proportions, risk ratios, odds ratios
- Contingency tables, chi-squared test
- Events rates, rate ratios
- Time to event data: Lifetables, Kaplan Meier survival curves, log-rank test

# A game with a 20% success probability

- Let's play a chance game with a 20% probability of success
- We repeat the game with different length of sequence
  - The number of "single games" (sample size) will vary
  - We look how many times we win in a given number of single games and calculate the fraction of wins
  - We repeat that 2000 times and look at histograms of the fraction of wins


Percentage of successes in samples of size 1
Frequency
0
400
800
1200
1600
2000
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1

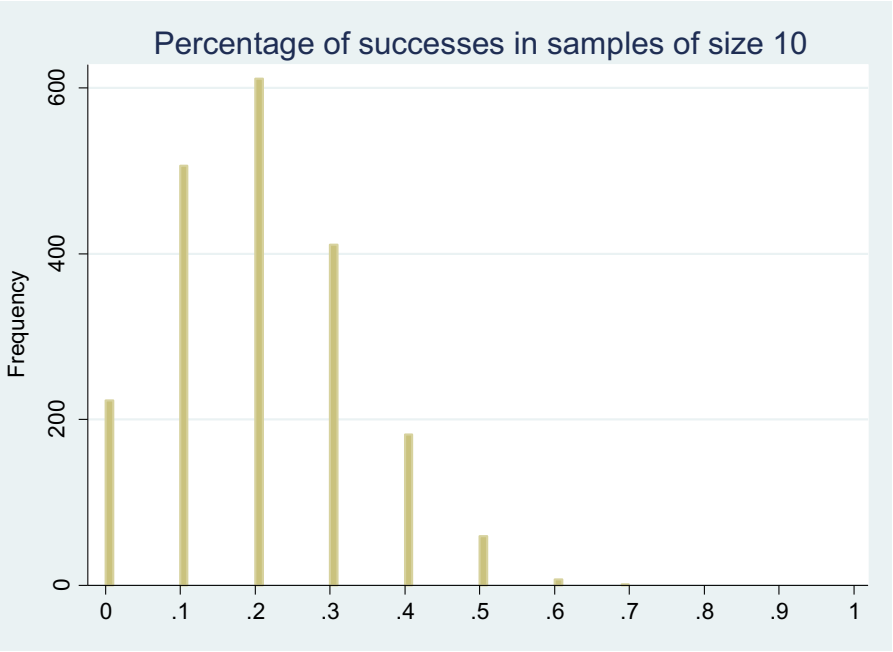
## Percentage of successes in samples of size 2

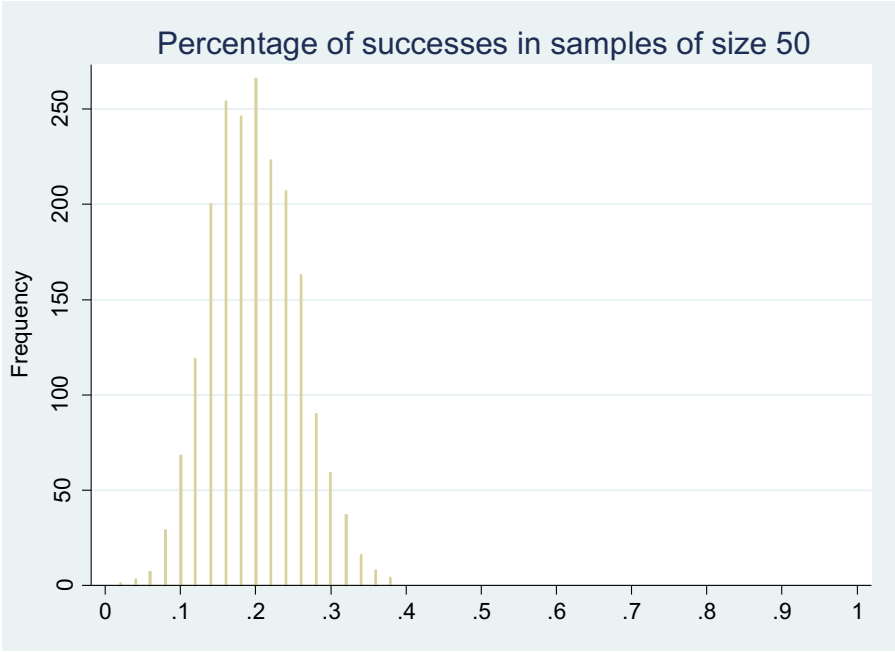


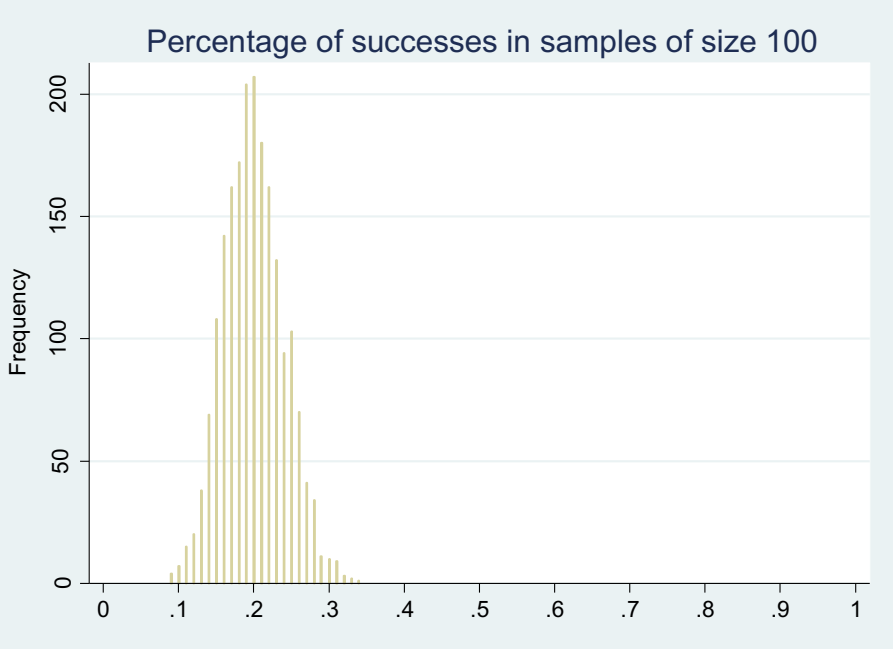

```
> y <-rbinom(2000, size=2, prob=0.2)
> table(y)
y
   0    1    2
1268  640   92
```

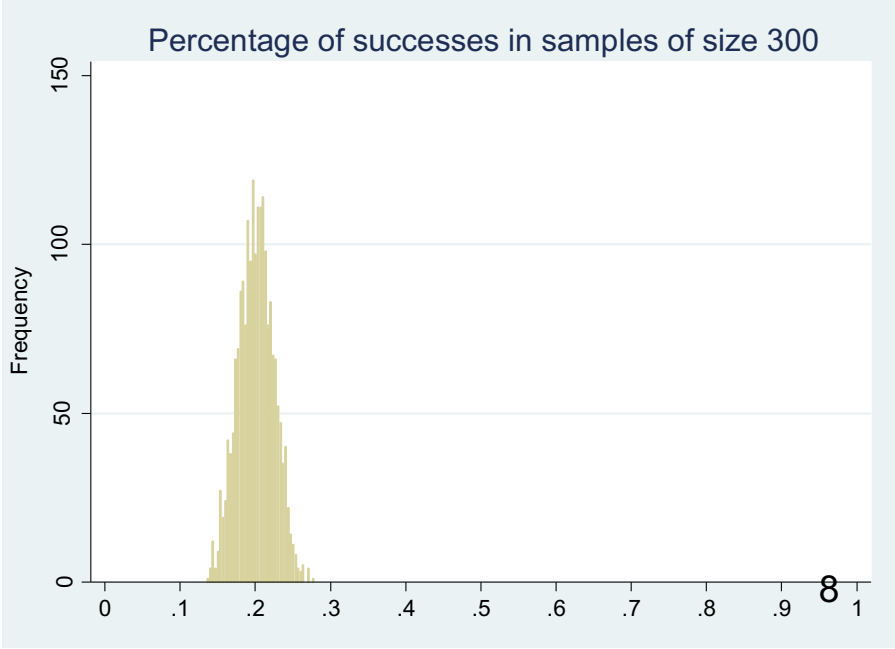

Percentage of successes in samples of size 3
Frequency
0
200
400
600
800
1000
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1

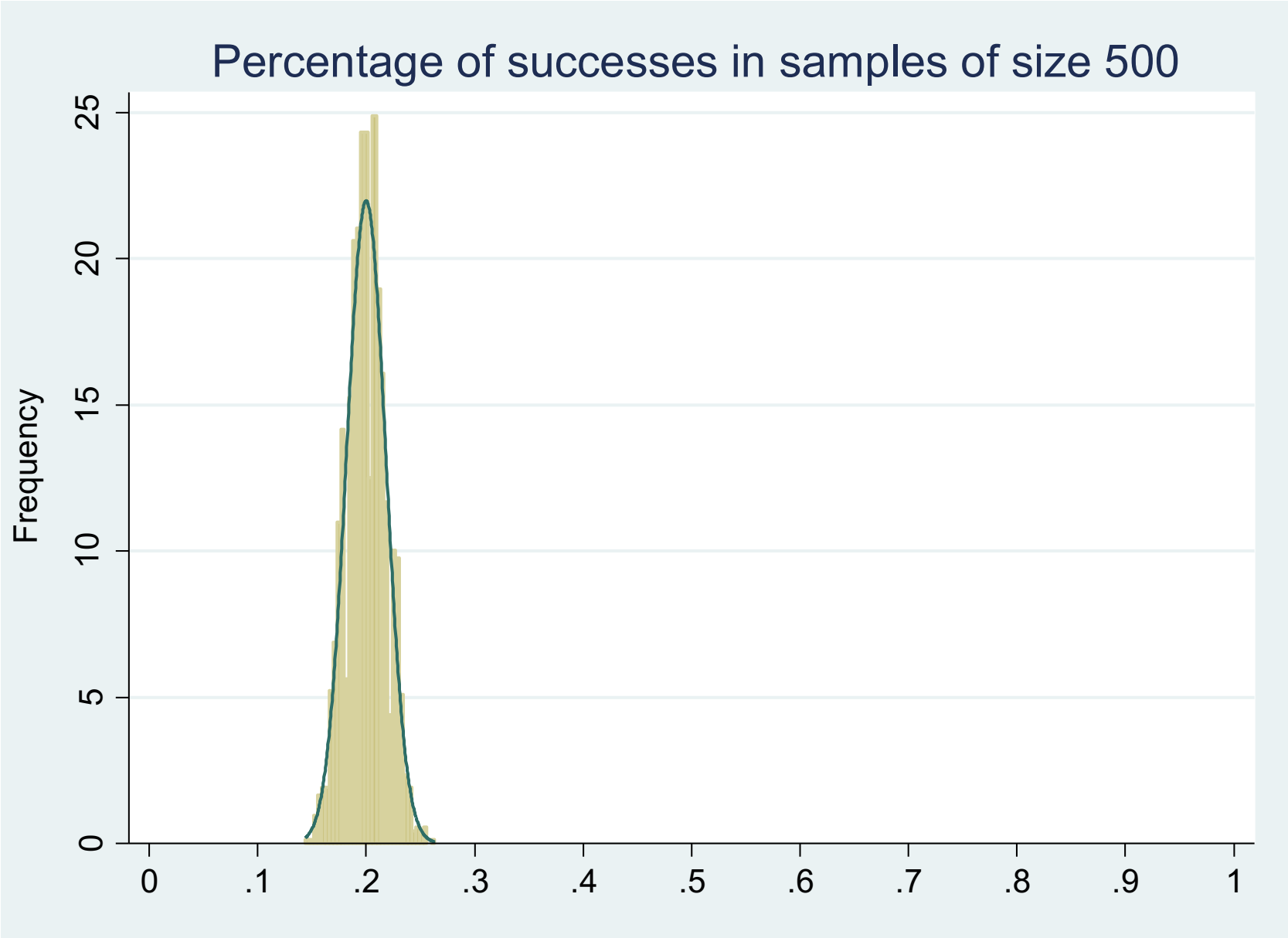

Percentage of successes in samples of size 5
Frequency
0
200
400
600
800
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1


Percentage of successes in samples of size 10
Frequency
0
200
400
600
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1


Percentage of successes in samples of size 50
Frequency
0
50
100
150
200
250
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1

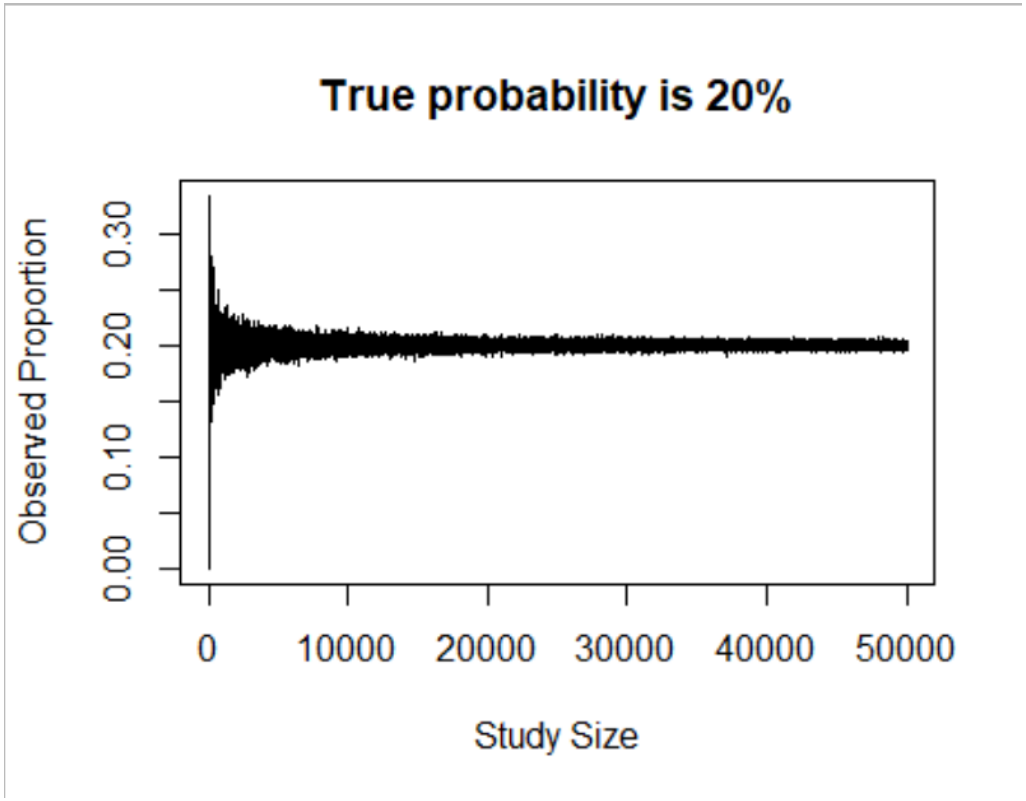

Percentage of successes in samples of size 100
Frequency
0
50
100
150
200
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1


Percentage of successes in samples of size 300
Frequency
0
50
100
150
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1


Percentage of successes in samples of size 500
Frequency
0
5
10
15
20
25
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1

# The larger N

The closer the observed percentage of «events» is to the true probability of event



# Die Binomial Distribution

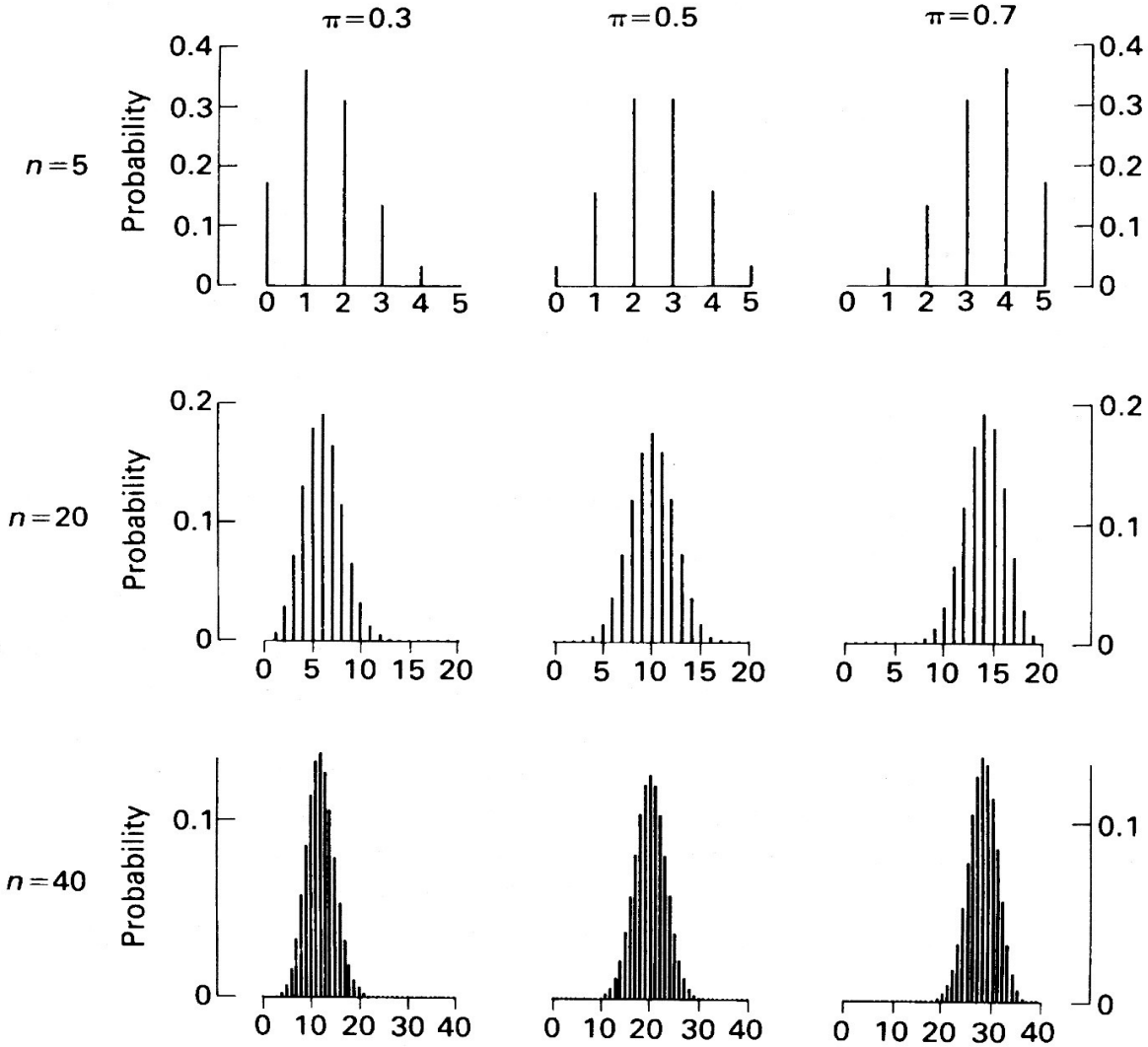



**Fig. 15.2** Binomial distribution for various values of $\pi$ and $n$. The horizontal scale in each diagram shows values of $d$.

# Binomial Distribution with success probability $\pi$

X= # of successes, N= # number of draws

$$\text{Prob}(X = k) = \frac{N!}{k!\,(N-k)!} \cdot \pi^k \cdot \pi^{N-k}$$

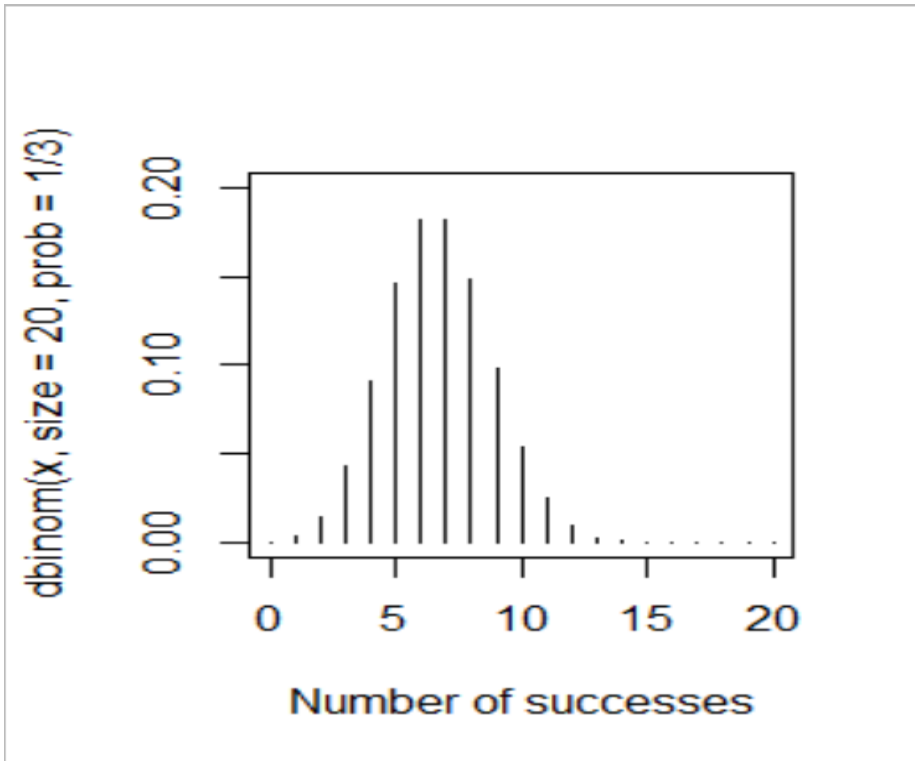
Mean: $E(X) = N \cdot \pi$

Variance: $\text{Var}(X) = N \cdot \pi \cdot (1 - \pi)$

```
In R implemented in binom «family» dbinom, rbinom,
pbinom, qbinom (as for norm «family»
```

# Binomial Distribution for true success probability $\pi$=0.3

```
x<- 0:20
plot(x, dbinom(x, size=20, prob=1/3), type="h",
ylim=c(0,0.2),xlab = ("Number of successes"))
```



# The proportion is an an unbiased estimator of the success probability $\pi$

Data: $k$ events among $N$ persons

Proportion: $p = \frac{k}{N}$

Expectation of the proportion under repeated sampling with size $N$

$$E(P) = E\left(\frac{X}{N}\right) = \frac{1}{N}E(X) = \frac{1}{N}N\pi = \pi$$

# Standard error of a proportion

$$\mathrm{SE(p)} = \sqrt{\frac{\pi \cdot (1 - \pi)}{\mathrm{N}}}$$

# The proportion is a mean

Let's code the the event of interest as

0: no event

1: event

Example data (N=10): 0, 1, 0, 0, 0, 1, 1, 0, 1, 0

$$p = \frac{\#\text{ events}}{N} = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{4}{10} = 0.4$$

Formula for the sample mean

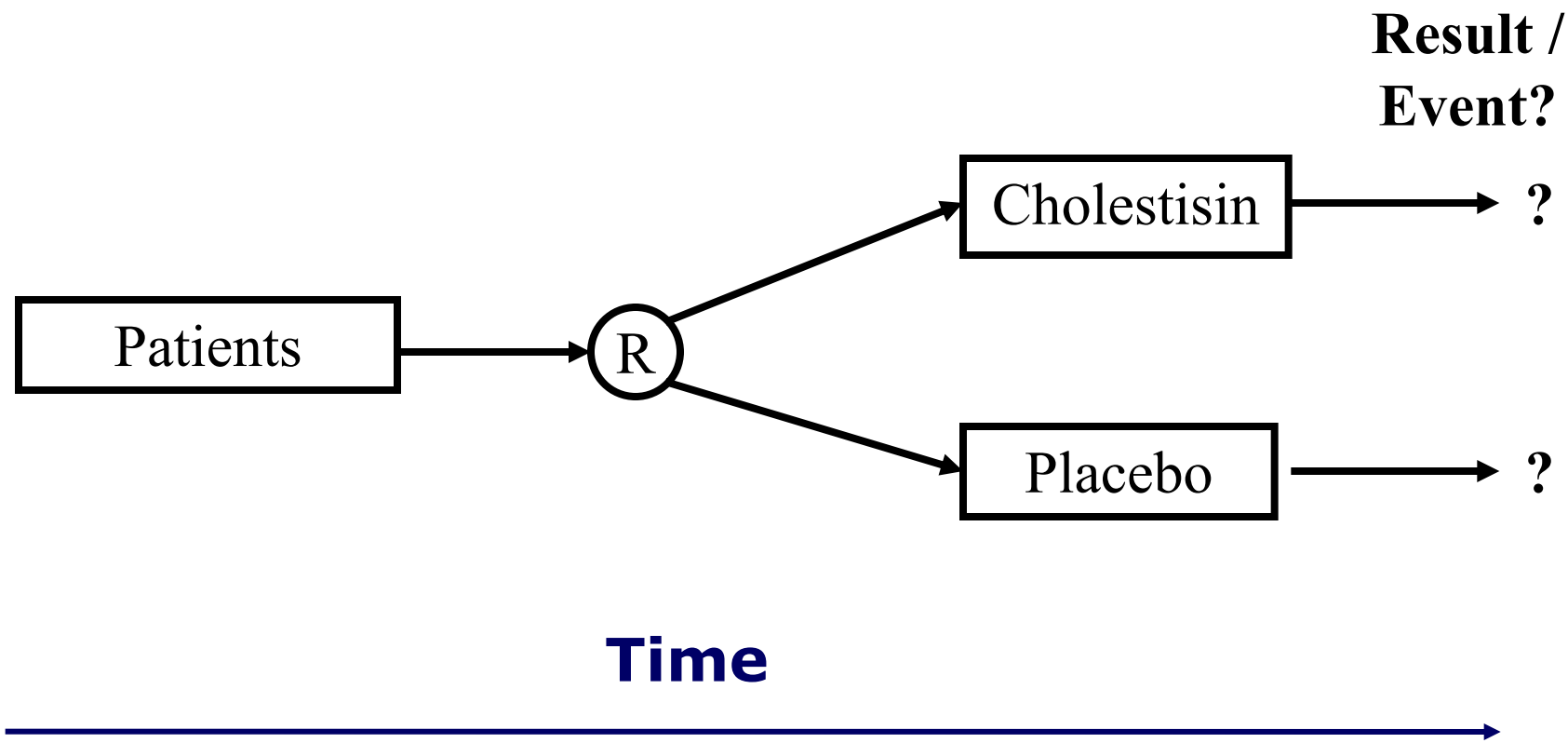
# The proportion is a mean

This means:

According to the central limit theorem the sampling distribution of the proportion is approximately normal for large samples.

$\Rightarrow$ We can apply standard methods for inference about means using the normal approximation.

# Comparing two risks from a RCT on lowering LDL cholesterol



# Randomized Study (i.e. treamtent is randomly allocated)

- Does Cholestesin reduce mortality ?
- Group1 get's Cholestesin
- Group 2 get's Placebo
- Deaths over 2 years : 20/150 versus 31/150

# Results

| | died | did not die | |
|---|---|---|---|
| Cholestisin | 20 (13.3%) | 130 | 150 |
| Placebo | 31 (20.7%) | 119 | 150 |

# 95% confidence interval for a proportion

observed proportion

**Lower end of the 95% CI :** $p - 1.96\sqrt{\frac{p \cdot (1-p)}{N}}$

**Upper end of the 95% CI :** $p + 1.96\sqrt{\frac{p \cdot (1-p)}{N}}$

# Calculations for placebo

## P= 20.67%

SE(p for placebo) $= \sqrt{\frac{0.2067 \cdot (1 - 0.2067)}{150}} = \sqrt{\frac{0.2067 \cdot 0.7933}{150}} =$

$$= \sqrt{\frac{0.163875}{150}} = \sqrt{0.001093} = 0.03306$$

# 95% CI for Placebo

**lower end of 95% CI :** $0.2067 - 1.96 \cdot 0.03306 = 0.142$

**upper end of 95% CI :** $0.2067 + 1.96 \cdot 0.03306 = 0.271$

# Calculations for Cholestisin

## P= 13.33%

$$\text{SE(p for Cholestisin)} = \sqrt{\frac{0.1333 \cdot (1 - 0.1333)}{150}} = \sqrt{\frac{0.1333 \cdot 0.8667}{150}} =$$

$$= \sqrt{\frac{0.115531}{150}} = \sqrt{0.000770} = 0.027756$$

# 95% CI for Cholestisin

**lower end for 95% CI :** $0.1333 - 1.96 \cdot 0.027756 = 0.079$

**upper end of 95% CI :** $0.1333 + 1.96 \cdot 0.027756 = 0.188$

# CI's for proportions in R

The normal approximation (also called Wald method) is poor if $N*p$ or $N*(1 - p) < 10$.

There are more precise methods:

`prop.test` uses the "Wilson score method"

`binom.test` uses the exact CI's based on the binomial distribution

For a comparison of methods see: Newcombe. Stat Med 1998;17:857-72

# Calculations for Cholestisin

```
> prop.test(x = 20, n = 150)
  95 percent confidence interval:
  0.08529927 0.20077120
> binom.test(x = 20, n = 150)
  95 percent confidence interval:
  0.08338381 0.19838697
```

Our result: `0.079 0.188`

Caution: the normal approximation will sometimes result in CI upper bounds >1 or lower bounds <0.

# Difference of two proportions

$$SE(p_{Placebo} - p_{Cholestisin}) = \sqrt{SE(p_{Placebo})^2 + SE(p_{Cholestisin})^2}$$

$$= \sqrt{\frac{p_{Placebo} \cdot (1 - p_{Placebo})}{N_{Placebo}} + \frac{p_{Cholestisin} \cdot (1 - p_{Cholestisin})}{N_{Cholestisin}}}$$

$$= \sqrt{0.03306^2 + 0.027756^2} = 0.043167$$

# CI for the difference in risk between Placebo and Cholestisin

$$\text{lower end } 95\% \text{ CI} = (0.2067 - 0.1333) - 1.96 \cdot 0.043167$$
$$= 0.0733 - 1.96 \cdot 0.043167 = -0.011$$

$$\text{upper end } 95\% \text{ CI} = (0.2067 - 0.1333) + 1.96 \cdot 0.043167$$
$$= 0.0733 + 1.96 \cdot 0.043167 = 0.158$$

95% CI for the difference in mortality risk ranges from

-1.1% to 15.8%

# Interpretation

- Patients treated with placebo had a 7.3% higher risk of death compared to patients treated with Cholestisin.
- With 95% confidence the difference in risk of death is between -1.1% und 15.8%.

⇒ We expect a p-value >0.05

# Getting the P-value

$$Z - value = \frac{0.0733}{0.043167} = 1.7$$

```
z <- pdiff/pdiff.se
[1] 1.698821
p <- 2*pnorm(-abs(z))
[1] 0.08935299
```

**→ P-value = 0.09**

# Interpretation of the p-value as conditional probability

Assuming no treatment effect, there is a 9% probability to observe a difference of 7.3% or greater.

# Two sample test for proportions in R

```
> prop.test(cbind(c(31,20), c(119,130)))

      2-sample test for equality of proportions with
continuity correction

data:  cbind(c(31, 20), c(119, 130))
X-squared = 2.3624, df = 1, p-value = 0.1243
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.0179395  0.1646062
sample estimates:
   prop 1    prop 2
0.2066667 0.1333333
```

# Risk ratios

| | dead | not dead | |
|---|---|---|---|
| Cholestisin | 20 (13.3%) | 130 | 150 |
| Placebo | 31 (20.7%) | 119 | 150 |

Mortality risk for patients with cholestisin = 20 / 150 = 13.3 pro 100
Mortality risk for patients with placebo = 31 / 150 = 20.7 pro 100

Risk Ratio (RR) for death = 0.133 / 0.207 = 0.64

# How now to get the 95% CI for RR?

Would be easy if we could use

$$\mathrm{SE}\,(\mathrm{quant}_{G1} - \mathrm{quant}_{G2}) = \sqrt{\mathrm{SE}^2(\mathrm{quant}_{G1}) + \mathrm{SE}^2(\mathrm{quant}_{G2})}$$

# Logarithms

- The division becomes a substraction...
- $\ln(RR) = \ln(risk_1/risk_0) = \ln(risk_1) - \ln(risk_0)$

- If we now would have a st.error(ln(risk)), we could use our rule how to combine SE's for differences of quantities of interest.

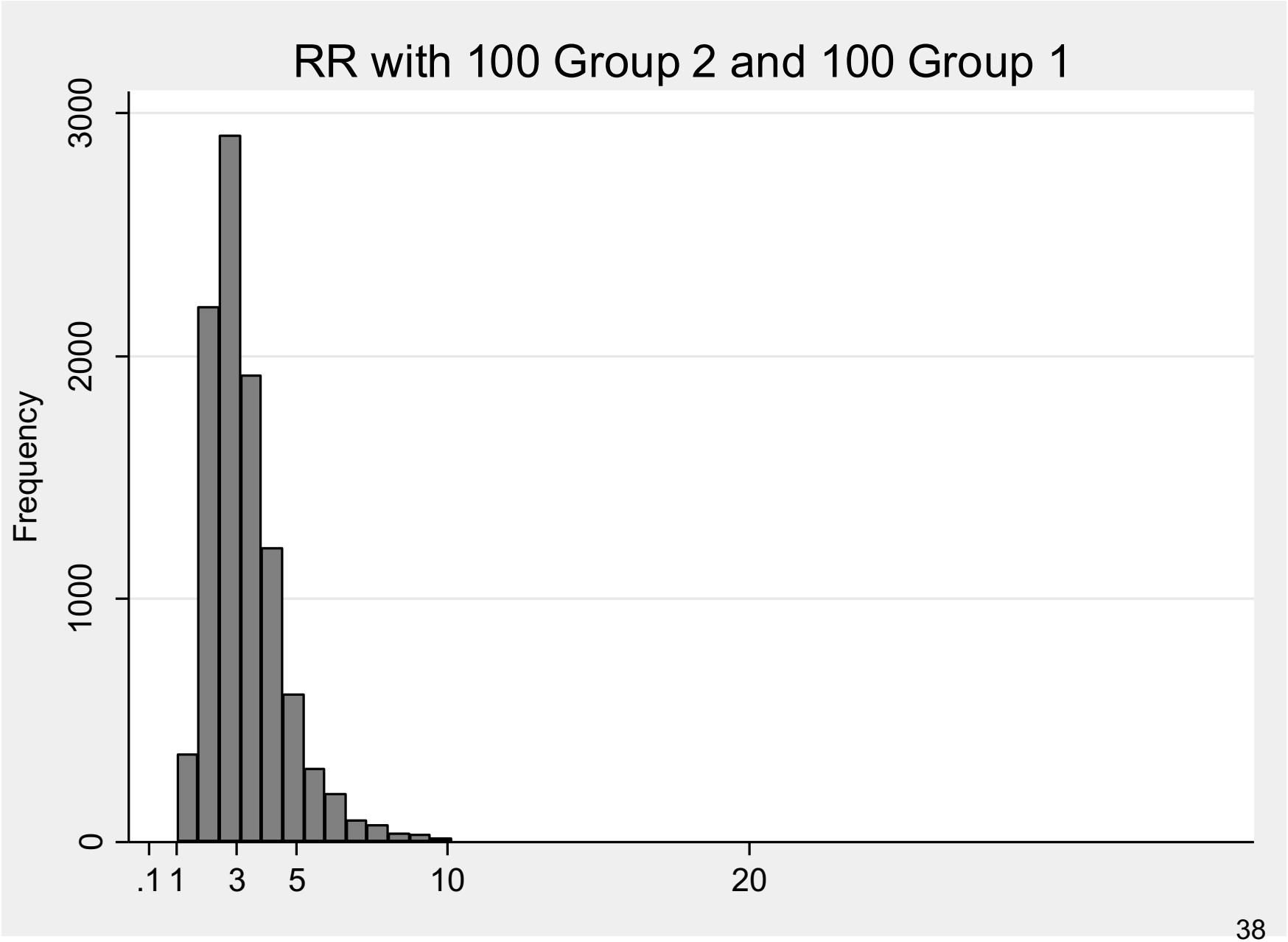
$$SE\,(quant_{G1} - quant_{G2}) = \sqrt{SE^2(quant_{G1}) + SE^2(quant_{G2})}$$

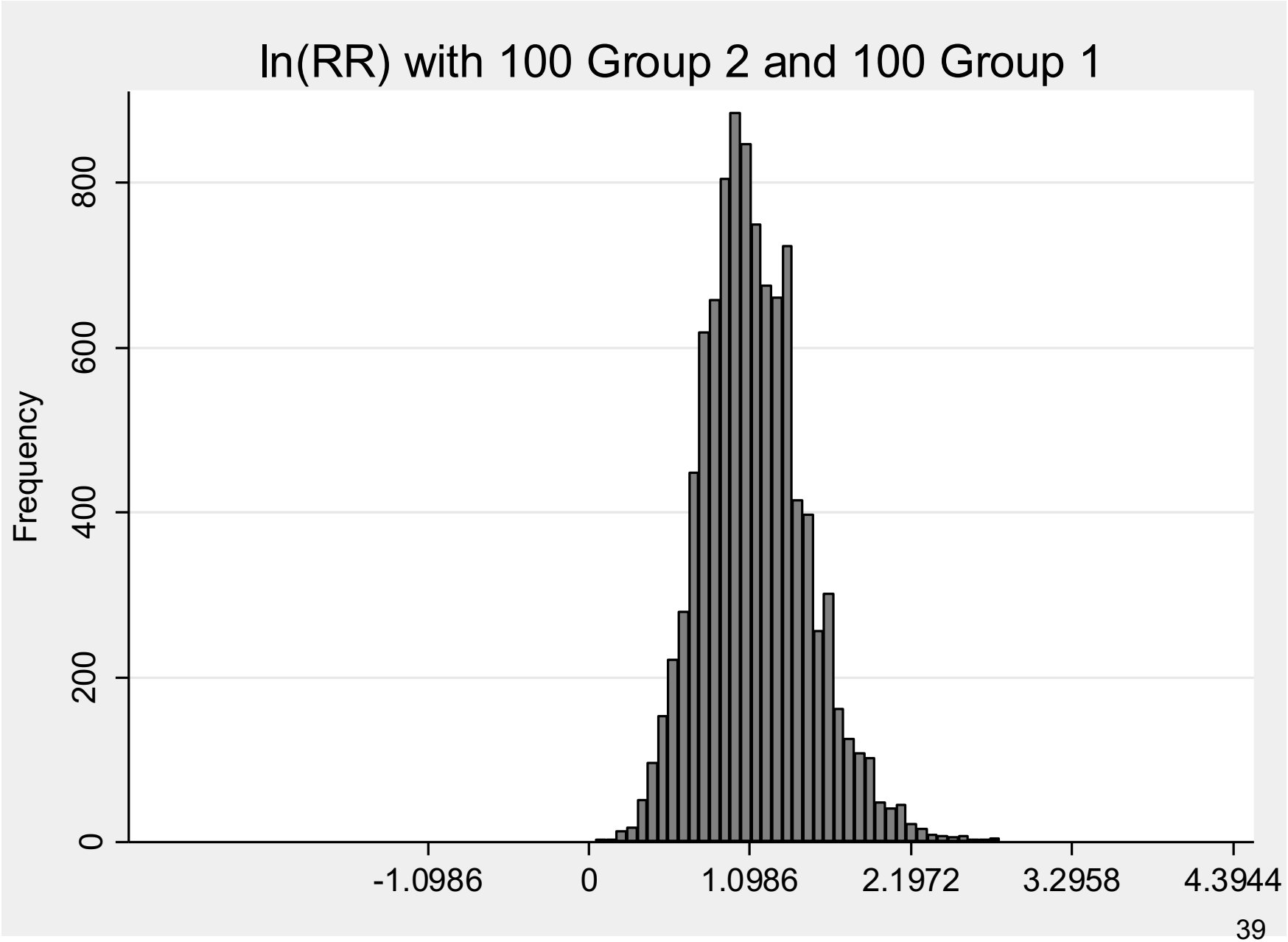
# Informal reason for using ln(RR) Simulation of studies

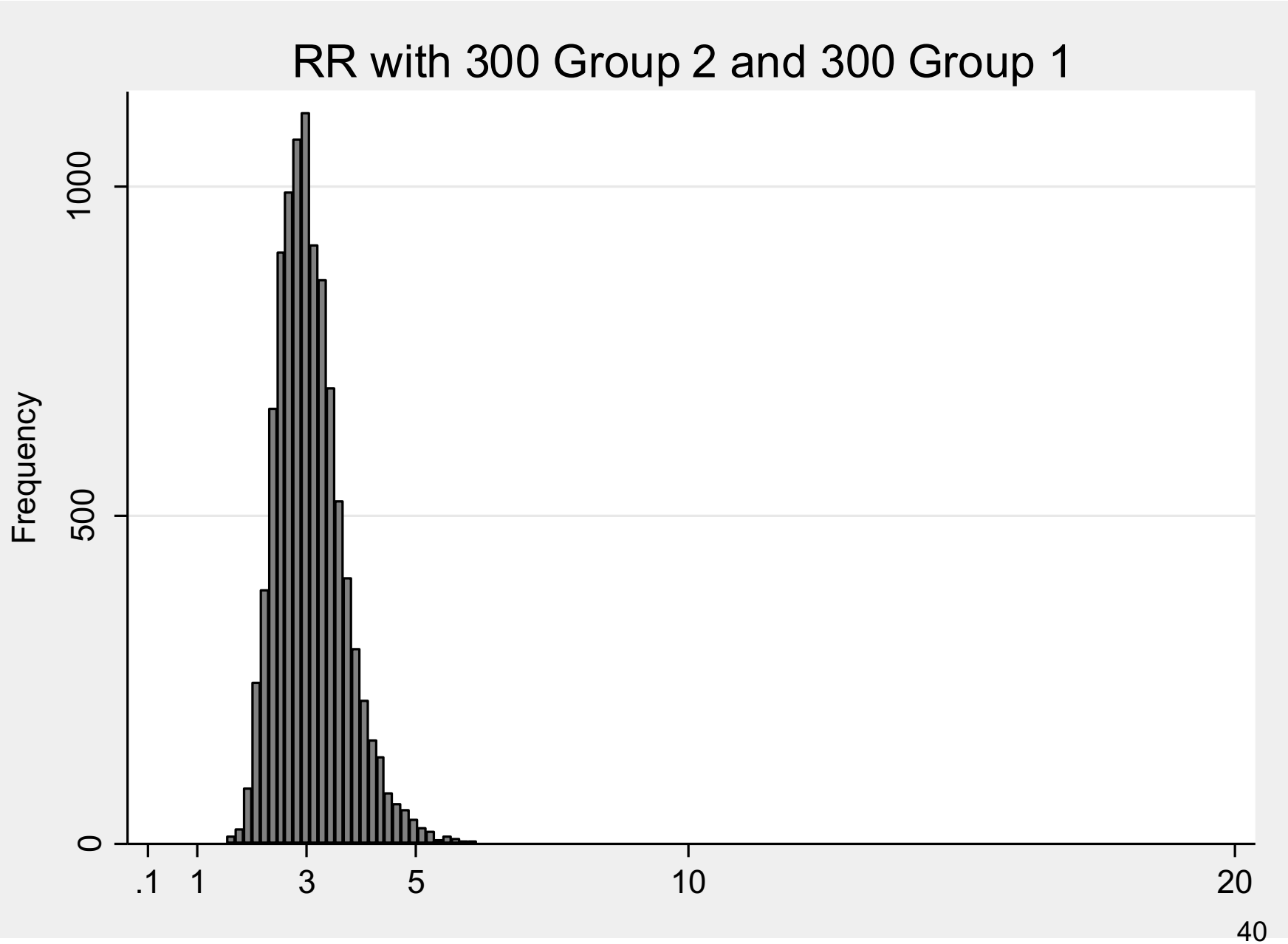
- 10’000 simulated studies
- 2 treatments (groups) are compared
- True risk in group 1: 10%

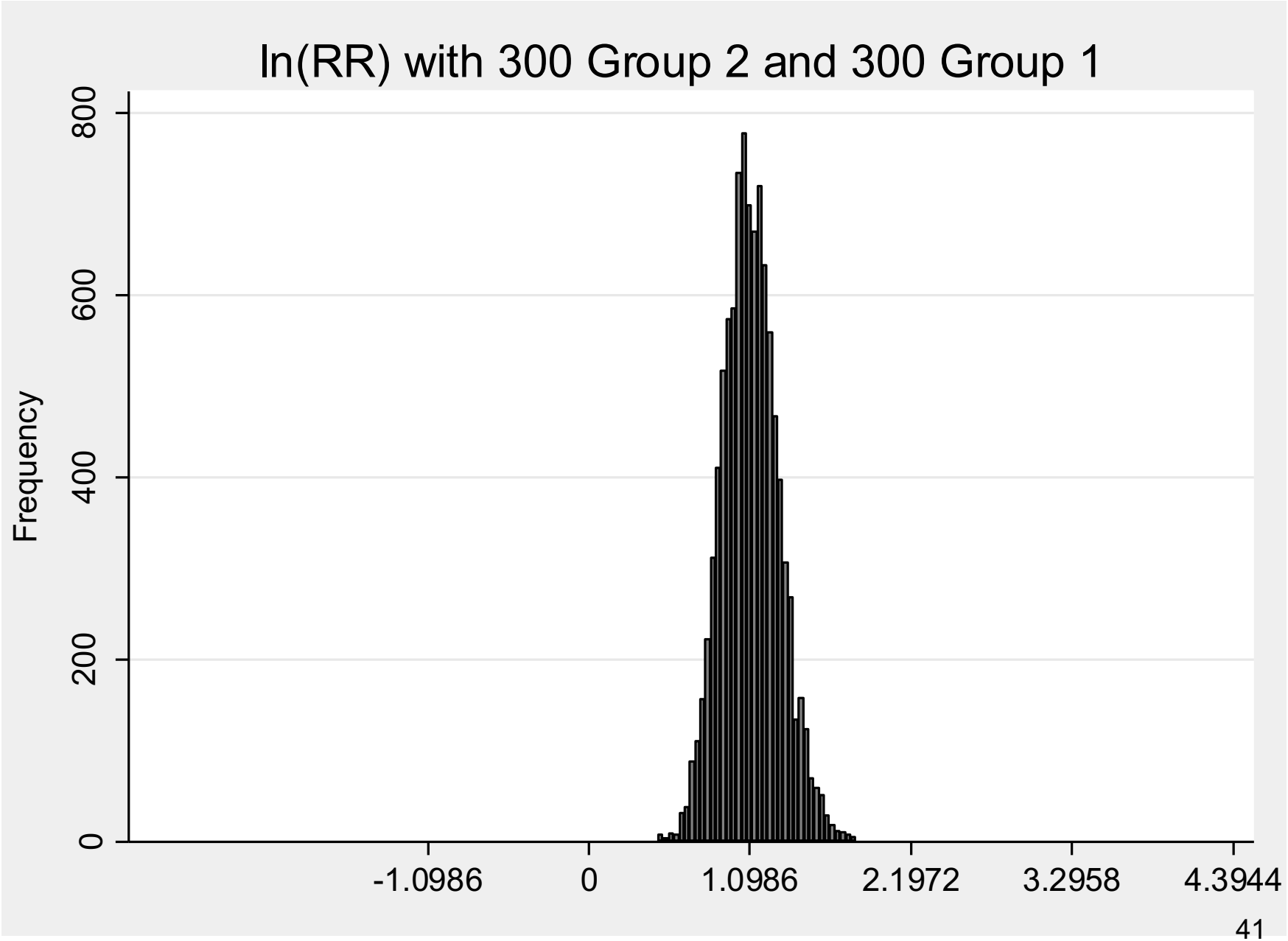
  -> RR = 3

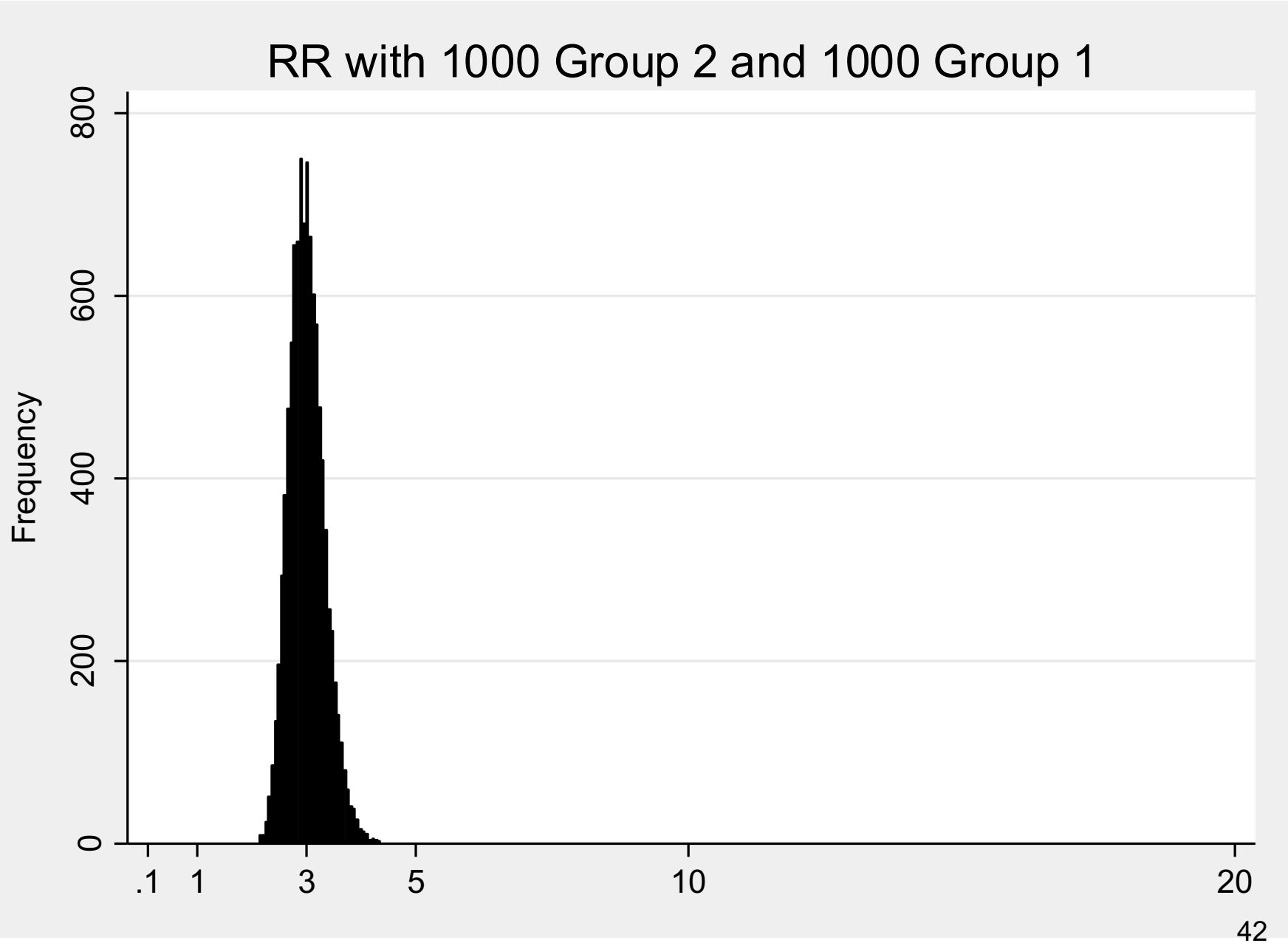
- True risk in group 2: 30%
- Number of persons per group : 100, 300 or 1000

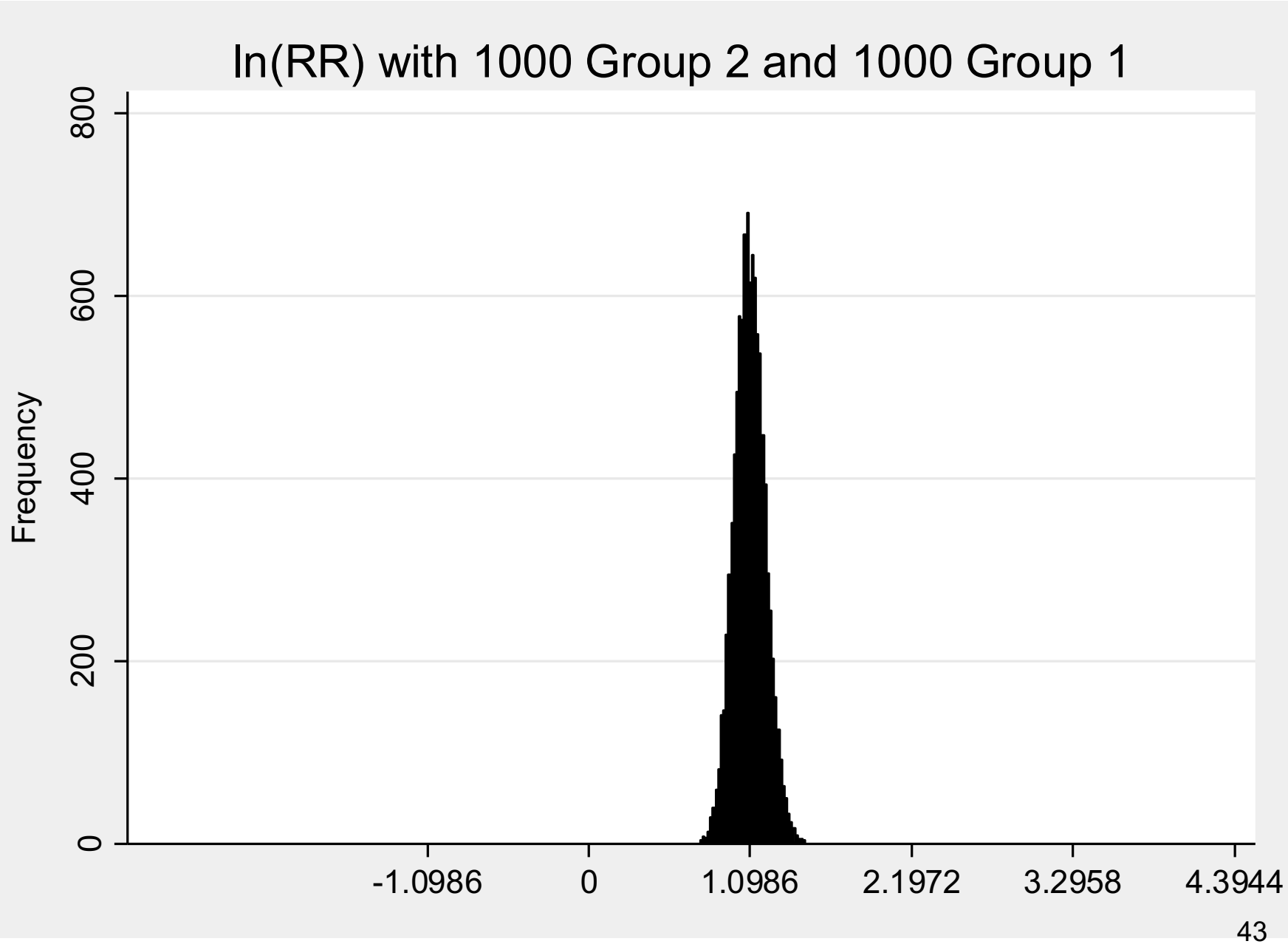

RR with 100 Group 2 and 100 Group 1
Frequency
0
1000
2000
3000
.1
1
3
5
10
20


ln(RR) with 100 Group 2 and 100 Group 1
Frequency
0
200
400
600
800
-1.0986
0
1.0986
2.1972
3.2958
4.3944


RR with 300 Group 2 and 300 Group 1
Frequency
0
500
1000
.1
1
3
5
10
20


ln(RR) with 300 Group 2 and 300 Group 1
Frequency
0
200
400
600
800
-1.0986
0
1.0986
2.1972
3.2958
4.3944


RR with 1000 Group 2 and 1000 Group 1
Frequency
0
200
400
600
800
.1
1
3
5
10
20


ln(RR) with 1000 Group 2 and 1000 Group 1
Frequency
0
200
400
600
800
-1.0986
0
1.0986
2.1972
3.2958
4.3944

# The delta method

- Allows to obtain approximately a st.error for ln(**p**) from the SE(p)

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

- Using the notation p=d/n, we obtain

$$st.error(\ln(p)) = \sqrt{1/d - 1/n}$$

# The 95 % confidence interval for a relative risk

RR = $risk_1$ / risk0 = **(** $d_1$ / $n_1$ ) /($d_0$ / $n_0$ )

On ln scale: ln(risk1/risk0) = ln(risk1) – ln(risk0)

and: s.e.(ln(risk1/risk0)) = s.e.(ln(risk1) – ln(risk0))

Using: $$\mathrm{SE}\,(\mathrm{quant}_{G1} - \mathrm{quant}_{G2}) = \sqrt{\mathrm{SE}^2(\mathrm{quant}_{G1}) + \mathrm{SE}^2(\mathrm{quant}_{G2})}$$

$$st.error(\ln(p)) = \sqrt{1/d - 1/n}$$

**produces a 95% CI for the ln(RR)**

$$\ln(RR) \pm 1.96 * \sqrt{1/d_1 - 1/n_1 + 1/d_0 - 1/n_0}$$

# Cholestisin Study

| | dead | not dead | |
|---|---|---|---|
| Cholestisin | 20 (13.3%) | 130 | 150 |
| Placebo | 31 (20.7%) | 119 | 150 |

Mortality risk for patients with cholestisin = 20 / 150 = 13.3 pro 100
Mortality risk for patients with placebo = 31 / 150 = 20.7 pro 100

Risk Ratio (RR) for death = 0.133 / 0.207 = 0.64

# 95% CI for RR

$$\text{SE for ln(RR)} = \sqrt{\left(\frac{1}{31}-\frac{1}{150}\right)+\left(\frac{1}{20}-\frac{1}{150}\right)}$$

$$= \sqrt{(0.03226-0.006667)+(0.05-0.006667)}$$

$$= \sqrt{0.025591+0.043333} = \sqrt{0.068925} = 0.26254$$

$$\text{lower end 95\% CI of ln(RR)} = \text{ln(RR)- } 1.96\cdot \text{ SE(of ln(RR))}$$

$$= \text{ -0.438255- } 1.96\cdot 0.26254$$

$$= \text{-.9528333}$$

$$\text{lower end 95\% CI of RR} = \exp(\text{lower end 95\% CI of ln(RR)})$$

$$= \exp(\text{-.9528333}) = \text{ 0.3856}$$

# 95% CI of RR

$$\text{upper end 95\% CI of } \ln(\text{RR}) = \ln(\text{RR}) + 1.96 \cdot \text{SE(of } \ln(\text{RR}))$$
$$= -0.438255 + 1.96 \cdot 0.26254$$
$$= 0.0763235$$

$$\text{upper end 95\% CI of RR} = \exp(\text{upper end 95\% CI of } \ln(\text{RR}))$$
$$= \exp(0.07632347) = 1.0793$$

# Getting a p-value

$$\text{Z-value} = \frac{\ln(RR)}{\text{SE(of ln(RR))}}$$

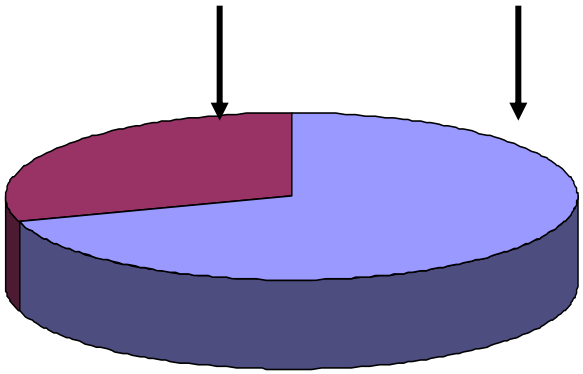
$$= \frac{-0.438255}{0.26254} = -1.67$$

Almost identical Z-value as for the risk difference

# Odds

$$a/b = 0.3/0.7 = 0.43$$

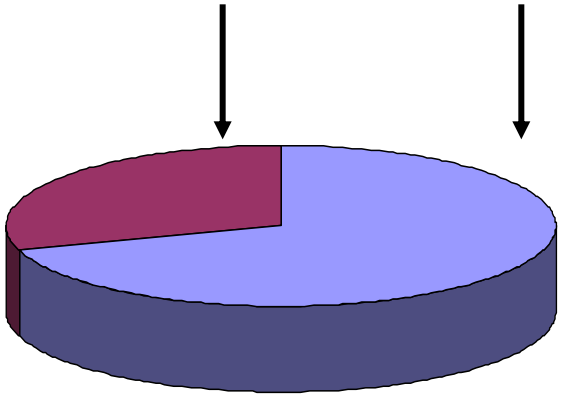
a=0.3 b=0.7



# Risk

$$a/(a+b) = 0.3/1 = 0.3$$

a=0.3 b=0.7



# Odds Ratio for the cholestisin study

| | dead | not dead | |
|---|---|---|---|
| Cholestisin | 20 (13.3%) | 130 | 150 |
| Placebo | 31 (20.7%) | 119 | 150 |

Odds of death in cholestisin group = 20 / 130 = 0.15384615

Odds of death in placebo group = 31 / 119 = 0.2605042

Odds Ratio of death = 0.15384615 / 0.2605042 = 0.59

# st.error for the (ln(Odds))

- Again using the delta method and the notation p=d/n → Odds = d/h where h = n-d

$$st.error(\ln(Odds)) = \sqrt{1/d + 1/h}$$

# The 95 % CI for the Odds Ratio

**Odds Ratio =** $(d_1 / h_1)$ / $(d_0 / h_0)$

**The 95% confidence interval for the ln(OR) is then**

$$\ln(OR) \pm 1.96 * \sqrt{1/d_1 + 1/h_1 + 1/d_0 + 1/h_0}$$

# Example

| Categories of BMI | Myocardial infarction or death 1 | 0 | Total |
|---|---|---|---|
| Overweight | 200 | 716 | 916 |
| Normal | 95 | 477 | 572 |
| Total | 295 | 1193 | 1,488 |

**OR = Odds$_1$ /Odds$_0$ =** ( $d_1$ / $h_1$ ) /($d_0$ / $h_0$ ) = .2793 / .1992 = 1.403

ln(OR) = 0.3383

st.error(ln(OR))=√[(1/200)+(1/716)+(1/95)+(1/477)]=

= ..... = 0.1379

# Then

95%-CI ln(OR) =

lower bound of CI ln(OR) = 0.3383 -1.96* 0.1379 = 0.0680

upper bound of CI ln(OR) = 0.3383 +1.96* 0.1379 = 0.6086

95%-lower bound of CI OR = exp(0.0680) = 1.07

95%-upper bound of CI OR = exp(0.6086) = 1.84

# A short list of useful s.e. formulas "approximations"

| | |
|---|---|
| $p = d / n$ | $s.e.(\ln p) = \sqrt{[1/d - 1/n]}$ |
| $RR = p_1 / p_0$ | $s.e.(\ln RR) = \sqrt{\{[1/d_1 - 1/n_1]+[1/d_0 - 1/n_0]\}}$ |
| $odds = d / h$ | $s.e.(\ln odds) = \sqrt{[1/d + 1/h]}$ |
| $OR = odds_1 / odds_0$ | $s.e.(\ln OR) = \sqrt{\{[1/d_1 + 1/h_1]+[1/d_0 + 1/h_0]\}}$ |

# Contincency tables, Pearsons chi-Squared test of independence

# A general approach to testing independence between categorical variables

| | died | did not die | |
|---|---|---|---|
| Cholestisin | 20 (13.3%) | 130 | 150 |
| Placebo | 31 (20.7%) | 119 | 150 |

# Calculate the expected cell frequencies assuming the treatment and outcome are independent

| | Placebo | Cholestisin | Total |
|---|---|---|---|
| Cases | 31 | 20 | 51 (17%) |
| | E=51*50%=25.5 | E=25.5 | |
| Noncases | 119 | 130 | 249 (83%) |
| | E=124.5 | E=124.5 | |
| Total | 150 | 150 | 300 |
| | (50%) | (50%) | |

Remember P(A and B) = P(A)*P(B) if A and B are independent

# The Chi-Square statistic

Describes and uses the differences between observed and expected cell counts

The larger “in total” the differences the stronger we have evidence against the assumption of “no difference"

$$\chi^2 = \sum_{\text{all cells}} \frac{(O - E)^2}{E}$$

O –observed frequency per cell

E –expected frequency per cell when the hypothesis of no difference in distribution is assumed

If indeed there is no difference then the Chi-Quadrat statistic (in a 2x2 table) follows a Chi-Square distribution with 1 degree of freedom

# Example

$$\sum_{\text{i over all cells}} \frac{(O_i - E_i)^2}{E_i} =$$

$$\frac{(31-25.5)^2}{25.5} + \frac{(20-25.5)^2}{25.5} + \frac{(119-124.5)^2}{124.5} + \frac{(130-124.5)^2}{124.5}$$

$$= 2.8585$$

# Chi-Square Statistics for larger tables (r x c tables)

| Current smoker | Categories of BMI: Underweig | Normal | Overweigh | Obese | Total |
|---|---|---|---|---|---|
| No | 13 | 212 | 441 | 130 | 796 |
| Yes | 62 | 360 | 475 | 93 | 990 |
| Total | 75 | 572 | 916 | 223 | 1,786 |

**Observed**

$$\text{Chi-sq} = \sum_{\text{i over all cells}} \frac{(O_i - E_i)^2}{E_i}$$

**With (r-1)*(c-1) degrees of freedom**

$$E_i = \frac{\text{total in column * total in row}}{\text{grand total}}$$

# Chi-Square Statistics for larger tables (r x c tables)

| Current smoker | Categories of BMI: Underweig | Normal | Overweigh | Obese | Total |
|---|---|---|---|---|---|
| No | 33.4 | 254.9 | 408.3 | 99.4 | 796 |
| Yes | 41.6 | 317.1 | 507.7 | 123.6 | 990 |
| Total | 75 | 572 | 916 | 223 | 1,786 |

Expected

$$\text{Chi-sq} = \sum_{\text{i over all cells}} \frac{(O_i - E_i)^2}{E_i}$$

**With (r-1)*(c-1) degrees of freedom**

$$E_i = \frac{\text{total in column * total in row}}{\text{grand total}}$$

# Chi-Square Statistics for larger tables (r x c tables)

| Current smoker | Categories of BMI: Underweig | Normal | Overweigh | Obese | Total |
|---|---|---|---|---|---|
| No | 13 | 212 | 441 | 130 | 796 |
| Yes | 62 | 360 | 475 | 93 | 990 |
| Total | 75 | 572 | 916 | 223 | 1,786 |

**Chi-sq =** $\sum_{\text{i over all cells}} \frac{(O_i - E_i)^2}{E_i}$ **= 57.311 , df = (r-1)*(c-1) = 3**

**p-value = 2.205e-12**

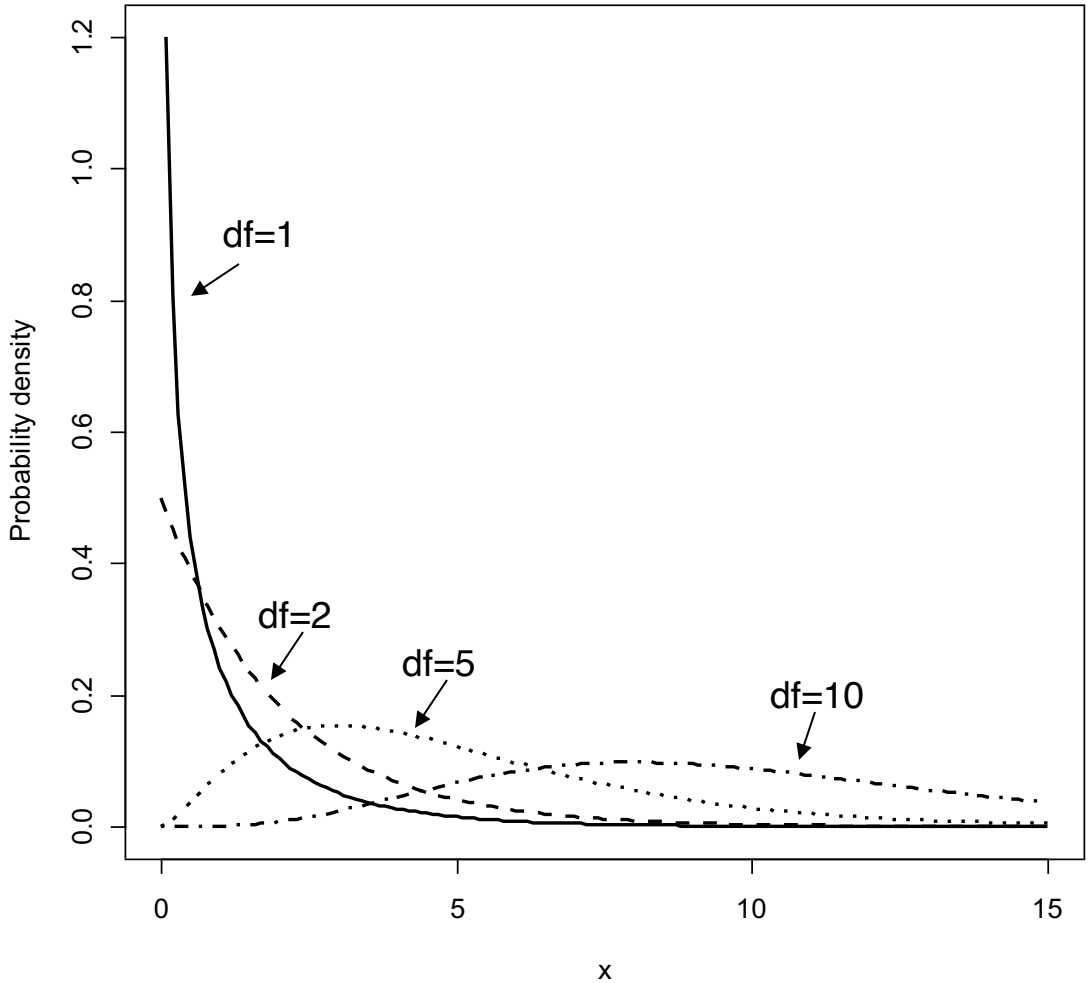
# Why (r-1)*(c-1) degrees of freedom?

| Current smoker | Categories of BMI: Underweig | Normal | Overweigh | Obese | Total |
|---|---|---|---|---|---|
| No | 13 | 212 | 441 | 130 | 796 |
| Yes | 62 | 360 | 475 | 93 | 990 |
| Total | 75 | 572 | 916 | 223 | 1,786 |

We use the column and row sums to estimate the expected counts und $H_0$ of no association.

Given column and row sums, only (r-1)(c-1)=3 observed counts are 'free' to vary.

# Chi-Square Distributions




df stands for the „degrees of freedom“

# Rates and rate ratios

# Longitudinal studies

**Studies where subjects are <u>followed over time</u>**

- **cohort studies** in which a group of initially disease-free individuals is followed over time, and the incidence of disease is recorded.
- **survival studies** in which individuals are followed from the time of an event such as the diagnosis of disease, and disease recurrence or death is recorded.
- **intervention studies** in which subjects are randomised to two or more treatment regimens, and the occurrence of pre-specified outcomes is recorded.

We assume that subjects experience only one disease endpoint (it is always possible to examine time until the **first** occurrence).

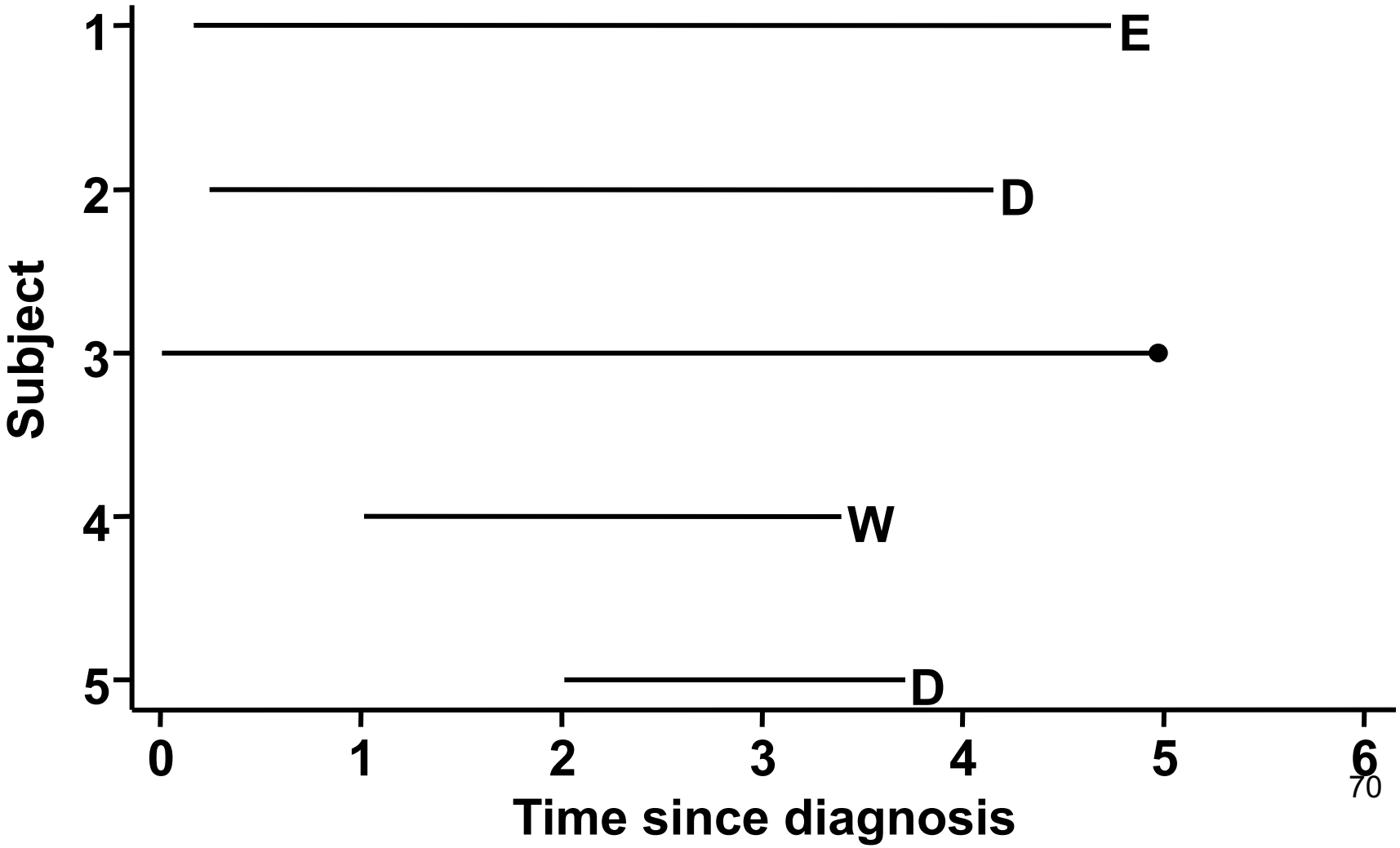
# Variable follow up times

In most longitudinal studies, individuals are followed for different lengths of time:

- logistic reasons: individuals recruited over time, but followed to the same end date
- new individuals may be enrolled during the study, because they have moved into the study area
- survival studies: delay between diagnosis and recruitment
- loss to follow up: e.g. emigration or withdrawal
- death from causes other than the one that is the focus of interest
- if the population of interest is defined by their age: e.g. women of child bearing age (ie.15-44 years)

# Example – 5 year study of prostate cancer

- subjects were recruited to the study at varying times after diagnosis, and exited at different points in time



In follow up studies we observe at least **two** pieces information for each individual:

- whether they experience the **disease event D**, and
- the length of time for which they were followed (the **observation time**).

## An individual's observation time:

- **starts** when the subject joins the study
- **stops** at the earlier of:
  - the time they develop the disease
  - the time they are lost to follow-up
  - the time the follow-up period ends

**i.e. the time during which, if the subject experienced an event, the event would be recorded in the study**.

# Rates

The **rate** of disease measures the occurrence of new events per unit time

To estimate a rate ($\lambda$), we:

1) Calculate the total number of events observed among all individuals, $d$

2) Calculate the sum of the individual observation times, $T$

3) Estimate the rate as:

$$\text{rate}, \lambda = \frac{\text{number of events}}{\text{total person years of observation}} = \frac{d}{T}$$

When $T$ is measured in years it is called **person-years-at-risk** or **pyar**

# Estimation of rates - example

57 lower respiratory infections were recorded during a 2-year study of 500 children. The total child-years of follow-up was $T$=873.

The rate of lower respiratory infection was estimated to be:

$$\lambda = 57/873 = 0.0653 \text{ per year}$$

This can also be expressed per 1000 child–years at risk, as:

$$\lambda = 57/873 \times 1000 = 65.3 \text{ per 1000 child-years}$$

# Estimation of rates - example

In a data set from the Caerphilly study 796 of the participants were non smokers:

- 12'182.46 person–years at risk were observed and
- 118 events of myocardial infarction or death occurred

The rate is therefore $\lambda = 118/\ 12'182.46 = 0.0096861$

This can also be expressed per 1000 person-years at risk, as:

$\lambda = 118/\ 12.18246 = 9.6861$ per 1000 person-years

# Poisson Distribution

- X be the number of independent events being observed in a fixed time span T:
- possible values are 0, 1, 2, …

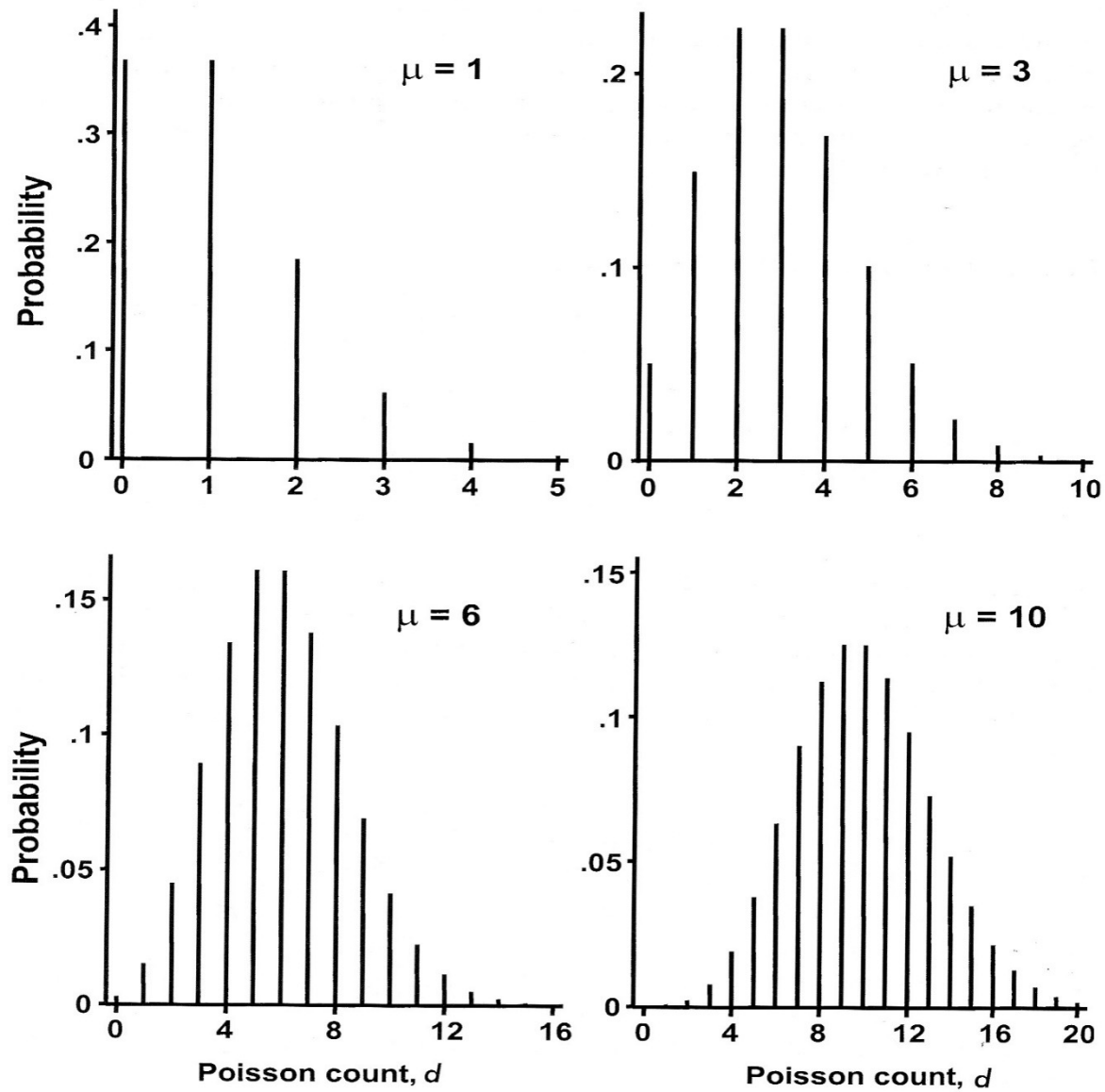



**Fig. 22.3** Poisson distribution for various values of $\mu$. The horizontal scale in each diagram shows values of the number of events, $d$.

# Poisson Distribution

The Poisson distribution is described by one parameter : $\mu$

Poisson($\mu$): mean = $\mu$; variance = $\mu$ → SD= $\sqrt{\mu}$

Probability (to observe N events ) =

$$e^{-\mu} \cdot \frac{\mu^{N}}{N!}$$

Important: We assume a fixed follow-up period T:

$\mu$= rate of events *T

# st.err. of incidence rate

$$\text{incidence rate } \lambda = \frac{d}{\text{Total Time}}$$

$$\text{st.error (incidence rate)} = \frac{\sqrt{d}}{\text{Total Time}}$$

# Confidence interval for a rate working on the log scale

1. Derive a confidence interval for the log rate
2. Antilog this to give a confidence interval for the rate

The standard error of the log rate is estimated by:

$$\text{s.e. of } \ln(\text{rate}) = \sqrt{\frac{1}{d}} = \frac{1}{\sqrt{d}}$$

This depends only on $d$, (not on $T$)

# 95% CI of rate : example

- We had the event rate for non-smokers $\lambda = 9.6861$ per 1000 pyar
- log rate = log(9.6861) = 2.27
- se(log rate) = $1/\sqrt{d} = 1/\sqrt{118} = 1/10.86 = 0.092$
- 95% CI (log rate) = $2.27-(1.96\times 0.092)$ to $2.27+(1.96 \times 0.092)$
  = 2.09 to 2.45
- 95% CI (rate) = exp(2.09) to exp(2.45)
  = 8.085 to 11.59 events per 1000 person–years

# Use of cipossion from survival package for getting 95% CI of rate

```
require(survival)# or library(survival)

(rate = 118/12182.46)
[1] 0.009686057
cipoisson(k=118, time = 12182.46)
      lower       upper
0.008017405 0.011599589
1000*cipoisson(k=118, time = 12182.46)
    lower     upper
 8.017405 11.599589
```

# Comparing rates

$$\text{Rate ratio} = \frac{\text{rate in exposed}}{\text{rate in unexposed}} = \frac{\lambda_1}{\lambda_0} = \frac{d_1 / T_1}{d_0 / T_0}$$

We use the standard error of the <u>log</u> rate ratio to derive confidence intervals and tests of the null hypothesis:

$$\text{s.e. of } \log(\text{rate ratio}) = \sqrt{1/d_0 + 1/d_1}$$

To test the null hypothesis that the rates in the two groups are equal:

$$z = \frac{\log(\text{rate ratio})}{\text{s.e. of } \log(\text{rate ratio})}$$

# Example 95% CI for rate ratio

For the rate among the 990 smokers we have:

230 events in 13‘978.48 years of observation,

thus $\lambda = 16.45$ per 1000 pyar

$$\text{Rate ratio} = \frac{16.45 \text{ per } 1000 \text{ py}}{9.6861 \text{ per } 1000 \text{ py}} = 1.7 \text{ and } \ln(\text{RR}) = 0.53$$

$$\text{se of } \ln(\text{RR}) = \sqrt{\frac{1}{230} + \frac{1}{118}} = 0.113$$

95% CI (log rate ratio) = 0.53-(1.96× 0.113) to 0.53+(1.96× 0.113 )
= 0.3085 to 0.751

95% CI (rate ratio) = exp(0.3085) to exp(0.751)
= 1.36 to 2.12

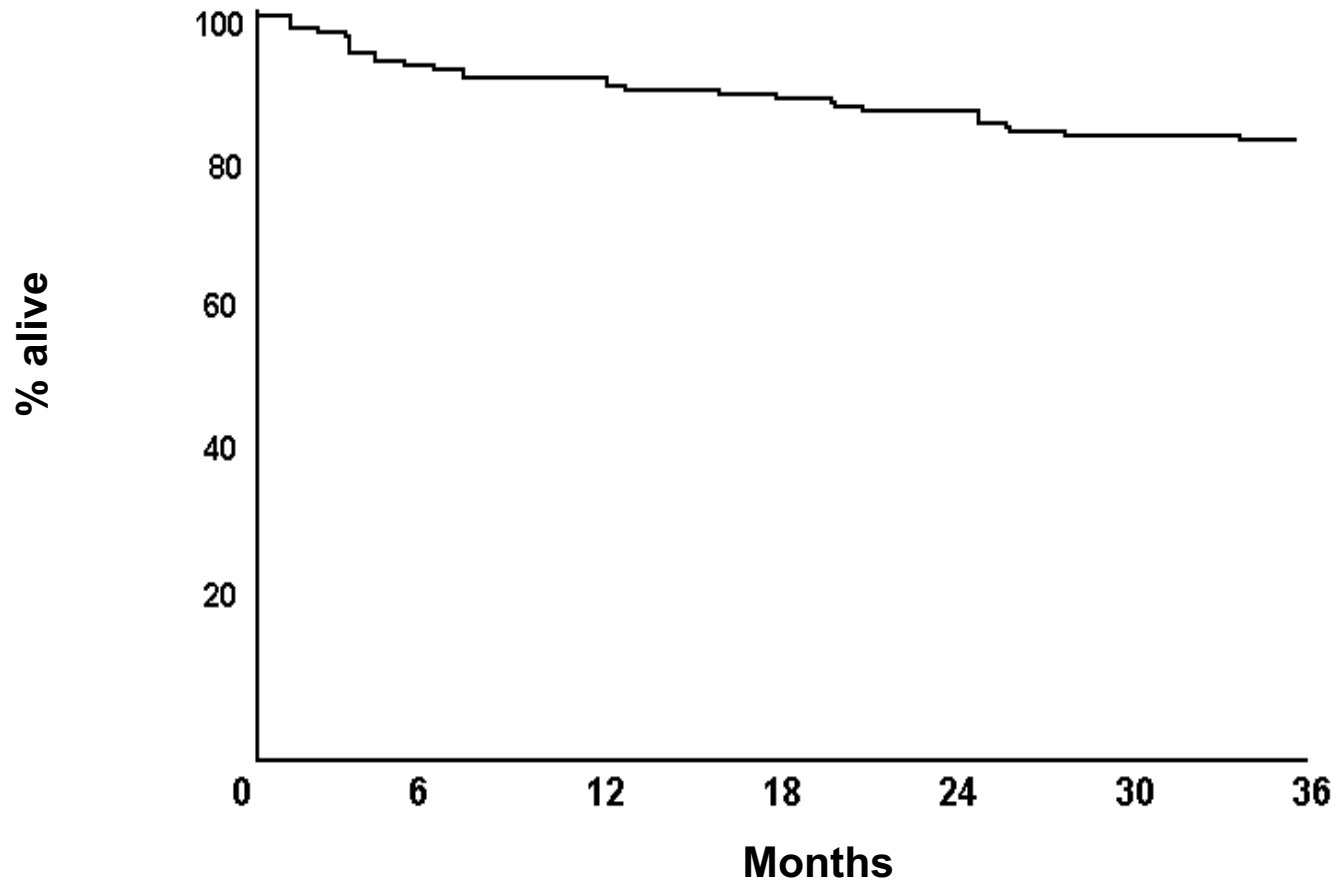
# Survival analysis avoiding the assumption of a constant rate

- Analysis of *time* to an event
- Rate *not* assumed to be constant over time
- Concentrates on survival curve

# Outline

- Life-table calculation on grouped information
- Kaplan-Meier survival calculation
- Log-rank test for comparing two survival curves

# Describing prognosis using life-table calculations and the Survival curve



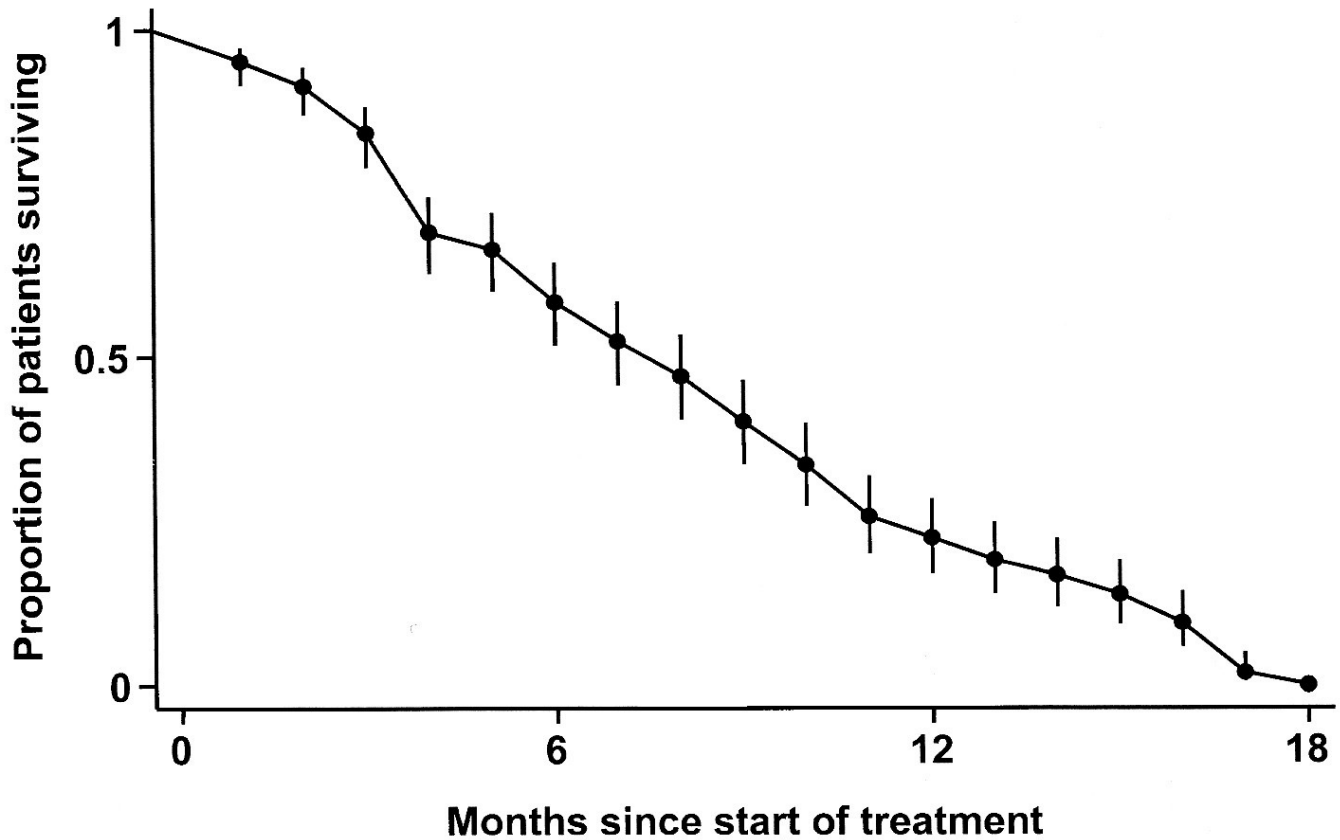



**Fig. 26.1** Survival curve for patients with small-cell carcinoma of the bronchus treated with radiotherapy, drawn from life table calculations presented in Table 26.1.

# A group of patients followed up after diagnosis

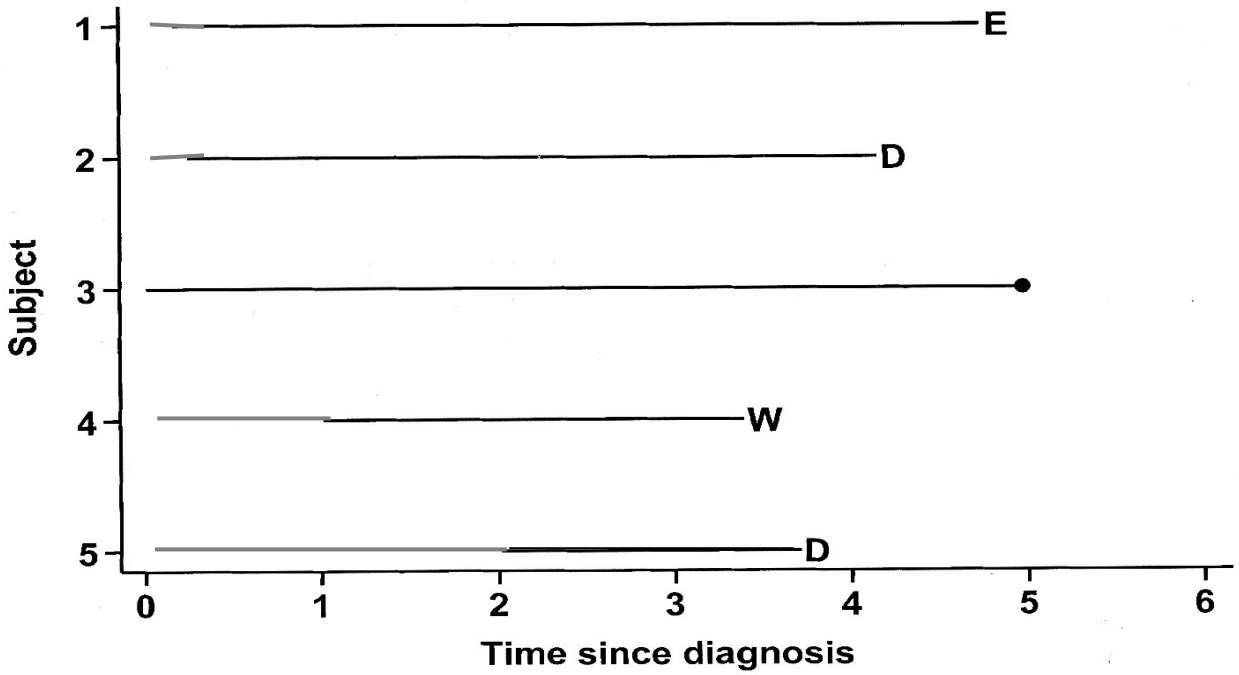



**Fig. 22.1** Follow-up histories for 5 subjects in a study of mortality after a diagnosis of prostate cancer (D = died, E = emigrated, W = withdrew, • = reached the end of follow-up without experiencing the disease event).

## Life tables

at start of the intervals

| Inter-val | Number surviving | Number of deaths | Number censored |
|---|---|---|---|
| 1 | **240** | 12 | 0 |
| 2 | | 9 | 0 |
| 3 | | 17 | 1 |
| ⋮ | | ⋮ | ⋮ |
| 16 | | 7 | 1 |
| 17 | | 12 | 0 |
| 18 | | 3 | 0 |

# Life tables

| Inter-val | Number surviving | Number of deaths | Number censored |
|---|---|---|---|
| 1 | 240 | **12** | **0** |
| 2 | **228** | 9 | 0 |
| 3 | | 17 | 1 |
| ⋮ | | ⋮ | ⋮ |
| 16 | | 7 | 1 |
| 17 | | 12 | 0 |
| 18 | | 3 | 0 |

Number surviving at 2. interval= 240 - 12 = 228

# Life tables

| Inter-val | Number surviving | Number of deaths | Number censored |
|---|---|---|---|
| 1 | 240 | 12 | 0 |
| 2 | 228 | 9 | 0 |
| 3 | 219 | 17 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 16 | 23 | **7** | **1** |
| 17 | **15** | 12 | 0 |
| 18 | 3 | 3 | 0 |

Number surviving at begin of 17th interval = 23 - 7 - 1 = 15

# Life tables

| Inter-val | Number surviving | Number of deaths | Number censored | P(death) |
|---|---|---|---|---|
| 1 | **240** | **12** | 0 | **0.0500** |
| 2 | 228 | 9 | 0 | |
| 3 | 219 | 17 | 1 | |
| ⋮ | ⋮ | ⋮ | ⋮ | |
| 16 | 23 | 7 | 1 | |
| 17 | 15 | 12 | 0 | |
| 18 | 3 | 3 | 0 | |

Probability (P) to die in 1st interval = 12/240 = 0.05

# Life tables

| Inter-val | Number surviving | Number of deaths | Number censored | Number at risk | P(death) |
|---|---|---|---|---|---|
| 1 | 240 | 12 | 0 | 240.0 | 0.0500 |
| 2 | 228 | 9 | 0 | 228.0 | 0.0395 |
| 3 | 219 | 17 | 1 | 218.5 | 0.0778 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 16 | 23 | 7 | 1 | 22.5 | 0.3111 |
| 17 | 15 | 12 | 0 | 15.0 | 0.8000 |
| 18 | 3 | 3 | 0 | 3.0 | 1.0000 |

= All persons at begin of interval – 0.5*(number of persons censored)

# Life tables

| Inter-val | Number surviving | Number of deaths | Number censored | Number at risk | P(death) | P(survival) |
|---|---|---|---|---|---|---|
| 1 | 240 | 12 | 0 | 240.0 | **0.0500** | **0.9500** |
| 2 | 228 | 9 | 0 | 228.0 | 0.0395 | |
| 3 | 219 | 17 | 1 | 218.5 | 0.0778 | |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |
| 16 | 23 | 7 | 1 | 22.5 | 0.3111 | |
| 17 | 15 | 12 | 0 | 15.0 | 0.8000 | |
| 18 | 3 | 3 | 0 | 3.0 | 1.0000 | |

Probability to survive the 1. interval = 1 – P(death) in 1. interval = 1 - 0.0500 = 0.9500

# Life tables

| Inter-val | Number surviving | Number of deaths | Number censored | Number at risk | P(death) | P(survival) |
|---|---|---|---|---|---|---|
| 1 | 240 | 12 | 0 | 240.0 | 0.0500 | 0.9500 |
| 2 | 228 | 9 | 0 | 228.0 | 0.0395 | 0.9605 |
| 3 | 219 | 17 | 1 | 218.5 | 0.0778 | 0.9222 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 16 | 23 | 7 | 1 | 22.5 | 0.3111 | 0.6889 |
| 17 | 15 | 12 | 0 | 15.0 | 0.8000 | 0.2000 |
| 18 | 3 | 3 | 0 | 3.0 | 1.0000 | 0.0000 |

# Life tables

| Inter-val | Number surviving | Number of deaths | Number censored | Number at risk | P(death) | P(survival) | Cumulative survival |
|---|---|---|---|---|---|---|---|
| 1 | 240 | 12 | 0 | 240.0 | 0.0500 | **0.9500** | **0.9500** |
| 2 | 228 | 9 | 0 | 228.0 | 0.0395 | 0.9605 | **0.9125** |
| 3 | 219 | 17 | 1 | 218.5 | 0.0778 | 0.9222 | |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |
| 16 | 23 | 7 | 1 | 22.5 | 0.3111 | 0.6889 | |
| 17 | 15 | 12 | 0 | 15.0 | 0.8000 | 0.2000 | |
| 18 | 3 | 3 | 0 | 3.0 | 1.0000 | 0.0000 | |

= Cumulative survival probability to survive up to the previous interval * probability to survive the current interval

# Life tables

| Inter-val | Number surviving | Number of deaths | Number censored | Number at risk | P(death) | P(survival) | Cumulative survival |
|---|---|---|---|---|---|---|---|
| 1 | 240 | 12 | 0 | 240.0 | 0.0500 | 0.9500 | 0.9500 |
| 2 | 228 | 9 | 0 | 228.0 | 0.0395 | 0.9605 | 0.9125 |
| 3 | 219 | 17 | 1 | 218.5 | 0.0778 | 0.9222 | 0.8415 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 16 | 23 | 7 | 1 | 22.5 | 0.3111 | 0.6889 | 0.0943 |
| 17 | 15 | 12 | 0 | 15.0 | 0.8000 | 0.2000 | 0.0189 |
| 18 | 3 | 3 | 0 | 3.0 | 1.0000 | 0.0000 | 0.0000 |

# Life-Table at one glance

**Table 26.1** Life table showing the survival pattern of 240 patients with small-cell carcinoma of bronchus treated with radiotherapy.

| (1) Interval (months) since start of treatment $i$ | (2) Number alive at beginning of interval $a_i$ | (3) Deaths during interval $d_i$ | (4) Number censored (lost to follow-up) during interval $c_i$ | (5) Number of persons at risk $n_i = a_i - c_i/2$ | (6) Risk of dying during interval $r_i = d_i/n_i$ | (7) Chance of surviving interval $s_i = 1 - r_i$ | (8) Cumulative chance of survival from start of treatment $S(i) = S(i-1) \times s_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 240 | 12 | 0 | 240.0 | 0.0500 | 0.9500 | 0.9500 |
| 2 | 228 | 9 | 0 | 228.0 | 0.0395 | 0.9605 | 0.9125 |
| 3 | 219 | 17 | 1 | 218.5 | 0.0778 | 0.9222 | 0.8415 |
| 4 | 201 | 36 | 4 | 199.0 | 0.1809 | 0.8191 | 0.6893 |
| 5 | 161 | 6 | 2 | 160.0 | 0.0375 | 0.9625 | 0.6634 |
| 6 | 153 | 18 | 7 | 149.5 | 0.1204 | 0.8796 | 0.5835 |
| 7 | 128 | 13 | 5 | 125.5 | 0.1036 | 0.8964 | 0.5231 |
| 8 | 110 | 11 | 3 | 108.5 | 0.1014 | 0.8986 | 0.4700 |
| 9 | 96 | 14 | 3 | 94.5 | 0.1481 | 0.8519 | 0.4004 |
| 10 | 79 | 13 | 0 | 79.0 | 0.1646 | 0.8354 | 0.3345 |
| 11 | 66 | 15 | 4 | 64.0 | 0.2344 | 0.7656 | 0.2561 |
| 12 | 47 | 6 | 1 | 46.5 | 0.1290 | 0.8710 | 0.2231 |
| 13 | 40 | 6 | 0 | 40.0 | 0.1500 | 0.8500 | 0.1896 |
| 14 | 34 | 4 | 2 | 33.0 | 0.1212 | 0.8788 | 0.1666 |
| 15 | 28 | 5 | 0 | 28.0 | 0.1786 | 0.8214 | 0.1369 |
| 16 | 23 | 7 | 1 | 22.5 | 0.3111 | 0.6889 | 0.0943 |
| 17 | 15 | 12 | 0 | 15.0 | 0.8000 | 0.2000 | 0.0189 |
| 18 | 3 | 3 | 0 | 3.0 | 1.0000 | 0.0000 | 0.0000 |

# The Kaplan-Meier Graph : The life-table calculation with infinitesimal small intervals

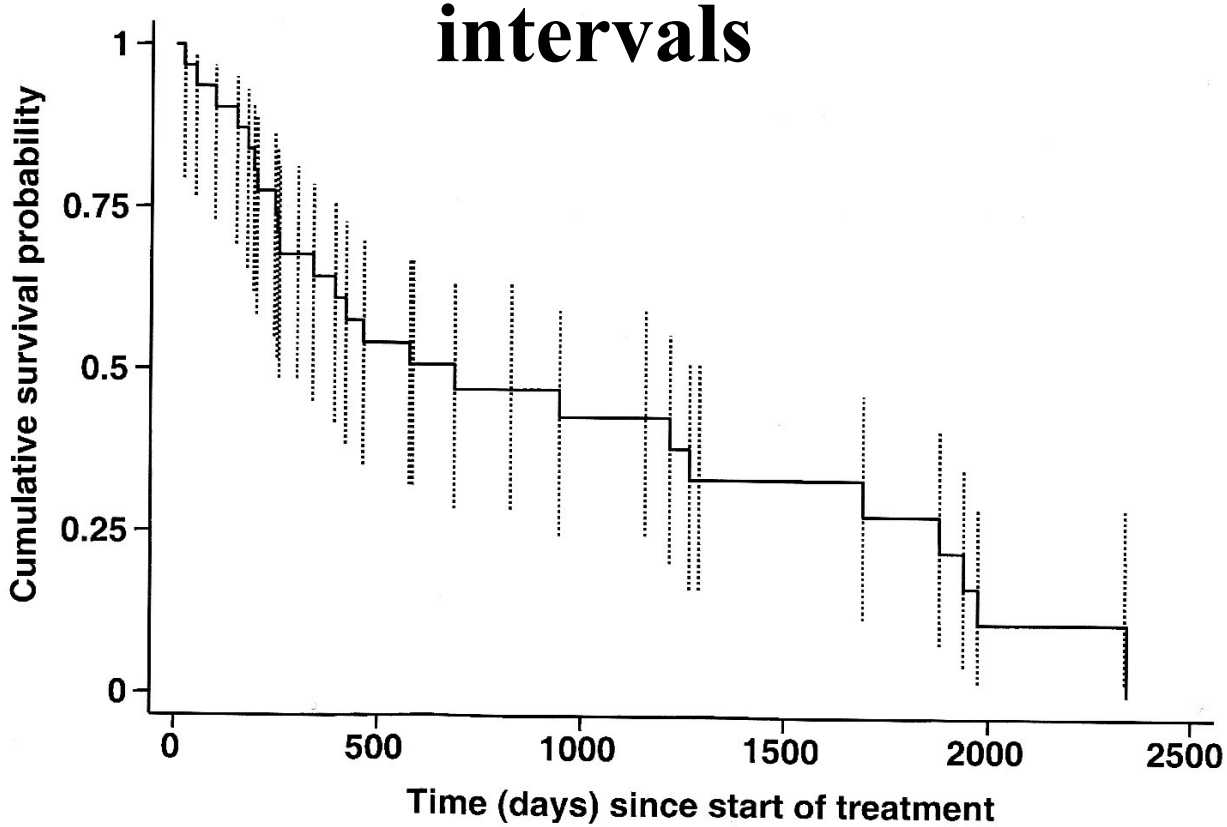



**Fig. 26.2** The Kaplan–Meier estimate of the survivor function, *S(t)*, together with upper and lower confidence limits, for 31 patients with primary biliary cirrhosis and central cholestasis.

# Kaplan-Meier estimates of the survival curve

- Standard way to estimate and display the survival curve $S(t)$
- Assume that we know the exact follow up time for each individual
- Based on a *conditional probability* argument

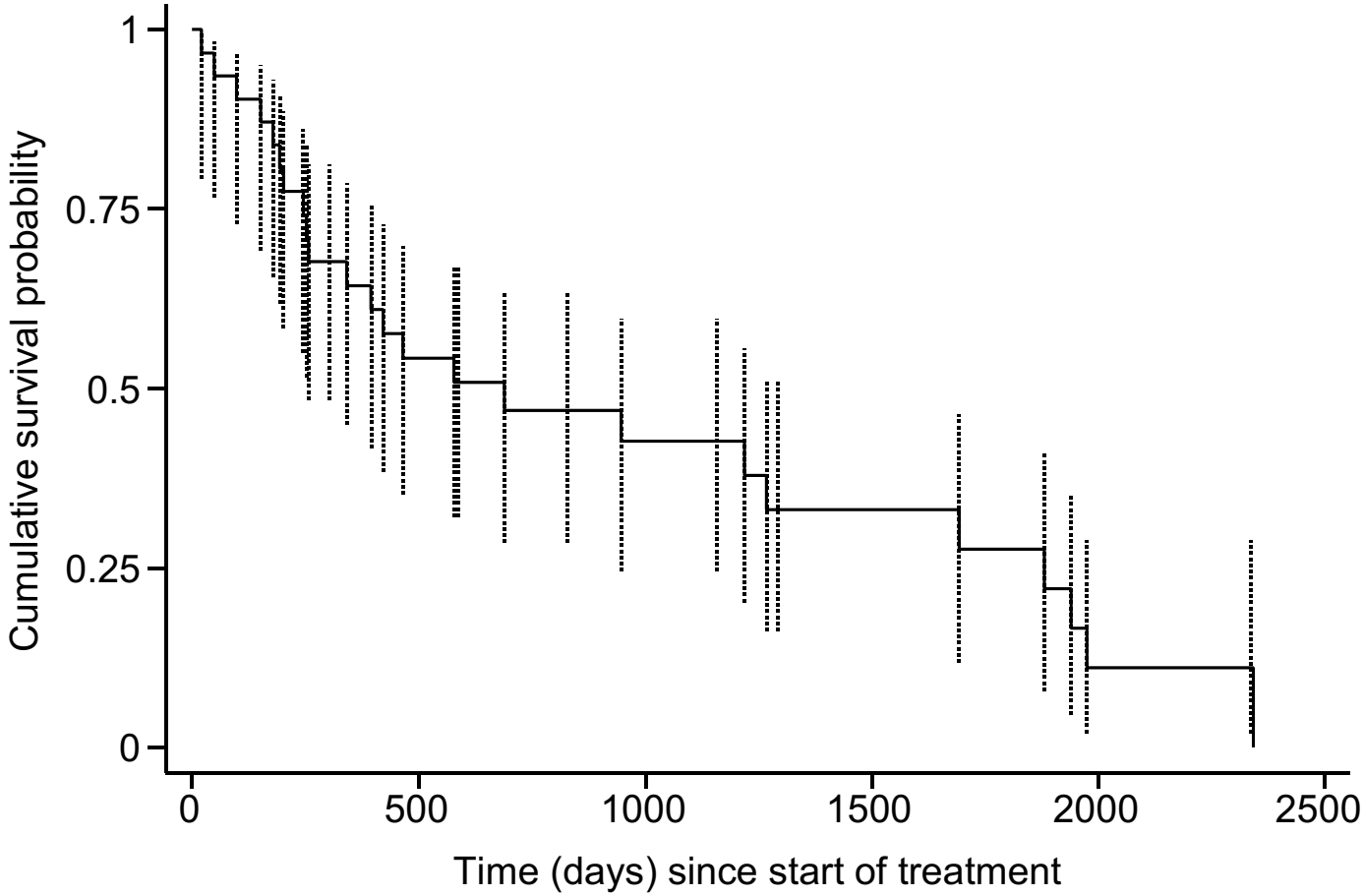
# Kaplan-Meier

| Time | Number at risk | Number of deaths | Number censored | Pr(death) | Pr(survival) |
|---|---|---|---|---|---|
| 19 | 31 | 1 | 0 | 0.0323 | 0.9677 |
| 48 | 30 | 1 | 0 | 0.0333 | 0.9667 |
| 96 | 29 | 1 | 0 | 0.0345 | 0.9655 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1975 | 3 | 1 | 0 | 0.3333 | 0.6667 |
| 2338 | 2 | 0 | 1 | 0.0000 | 1.0000 |
| 2343 | 1 | 1 | 0 | 1.0000 | 0.0000 |

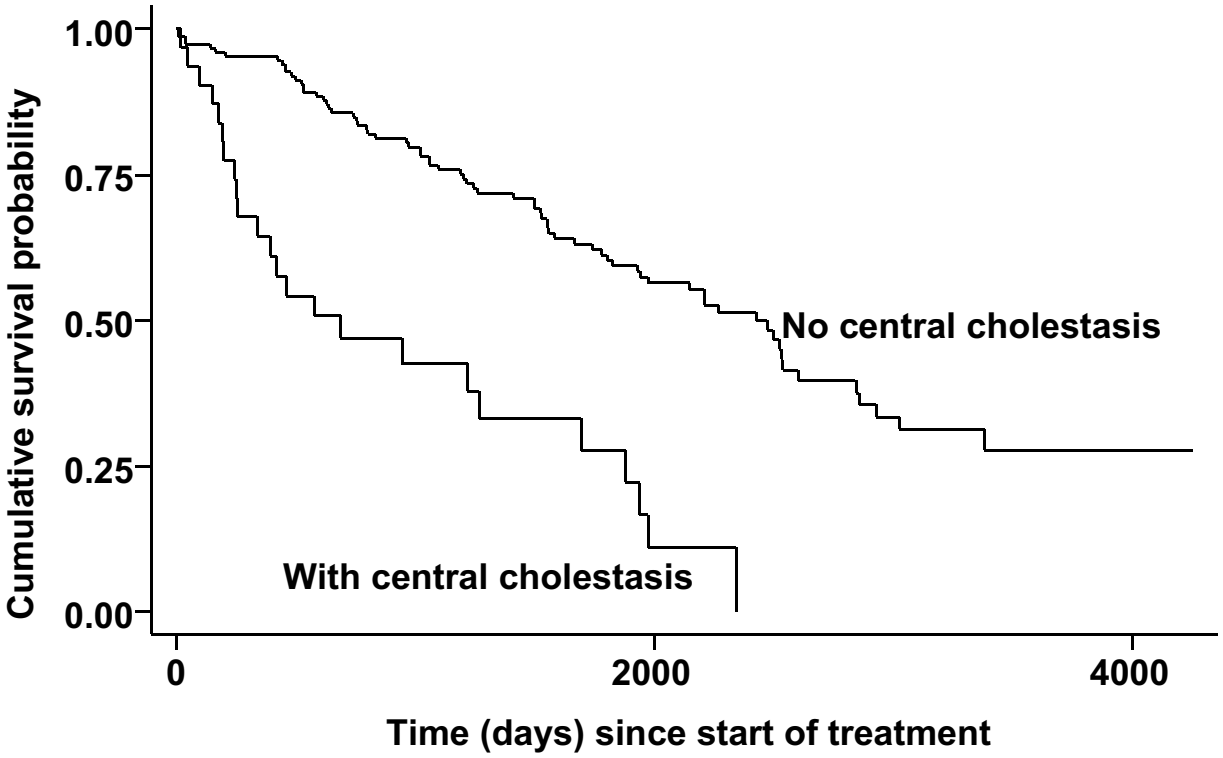
# Kaplan-Meier

| Time | Number at risk | Number of deaths | Number censored | Pr(death) | Pr(survival) | Survivor function **S(t)** |
|---|---|---|---|---|---|---|
| 19 | 31 | 1 | 0 | 0.0323 | 0.9677 | 0.9677 |
| 48 | 30 | 1 | 0 | 0.0333 | 0.9667 | 0.9355 |
| 96 | 29 | 1 | 0 | 0.0345 | 0.9655 | 0.9032 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1975 | 3 | 1 | 0 | 0.3333 | 0.6667 | **0.1105** |
| 2338 | 2 | 0 | 1 | 0.0000 | 1.0000 | **0.1105** |
| 2343 | 1 | 1 | 0 | 1.0000 | 0.0000 | 0.0000 |

# Kaplan-Meier



# Mantel-Cox method (log rank test)



# Mantel-Cox method (log rank test)

Extension of Mantel-Haenszel procedure:

- Construct 2×2 table for each time at which an event occurs
- Derive contributions from table
- Combine across all times (strata)

# Log rank test

| Day | $n_0$ | $d_0$ | $n_1$ | $d_1$ |
|---|---|---|---|---|
| 9 | **152** | **2** | **31** | **0** |
| 19 | 150 | 0 | 31 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

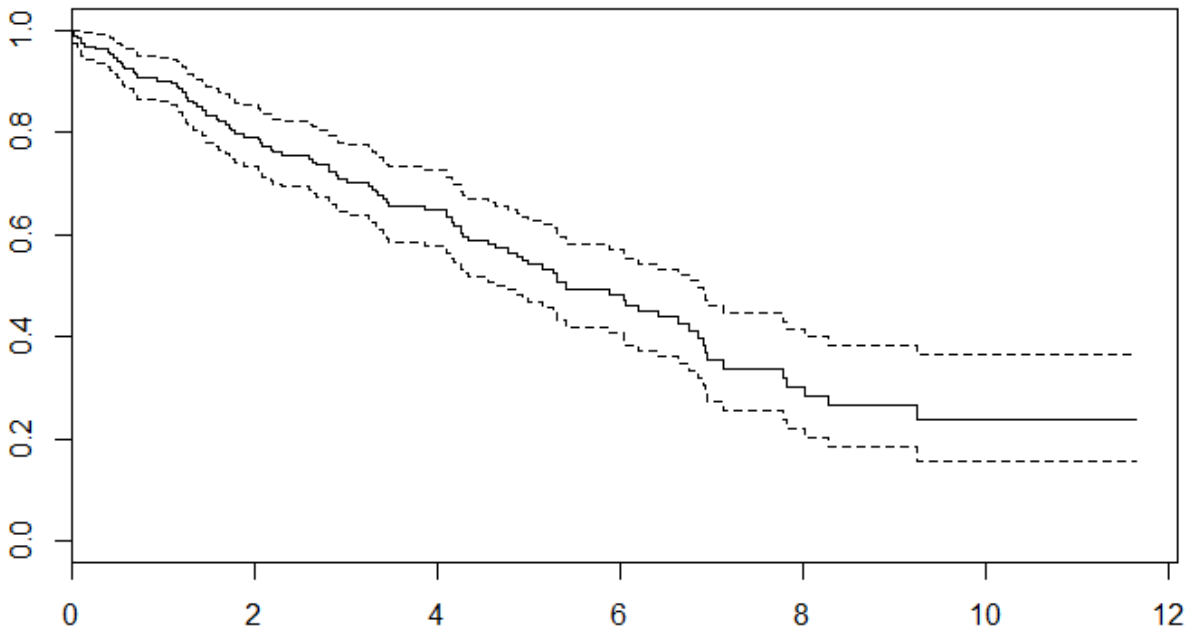
Day 9

| | Unexposed | Exposed | Total |
|---|---|---|---|
| Events | **2** | **0** | 2 |
| At risk | **152** | **31** | 183 |

Details of calculations are to hard for us

# Data on mortality from primary biliary cirrhosis (PBC)

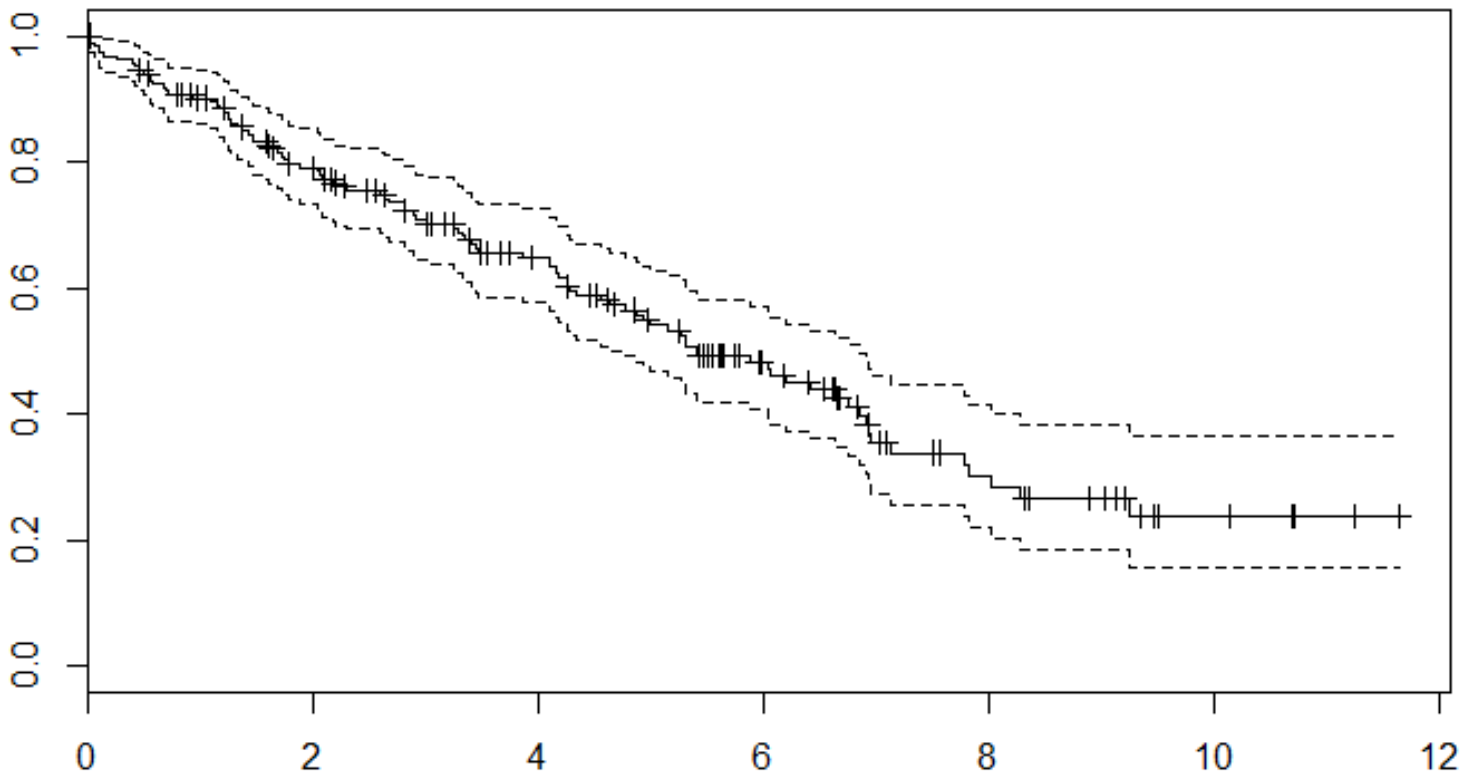
```
pbcdata <- read.table("pbc1bas.csv", sep="‟,", header=TRUE)


pbc.surv <- Surv(pbcdata$time, pbcdata$d==1)
surv.all <- survfit(pbc.surv ~ 1 )
plot(surv.all)
```



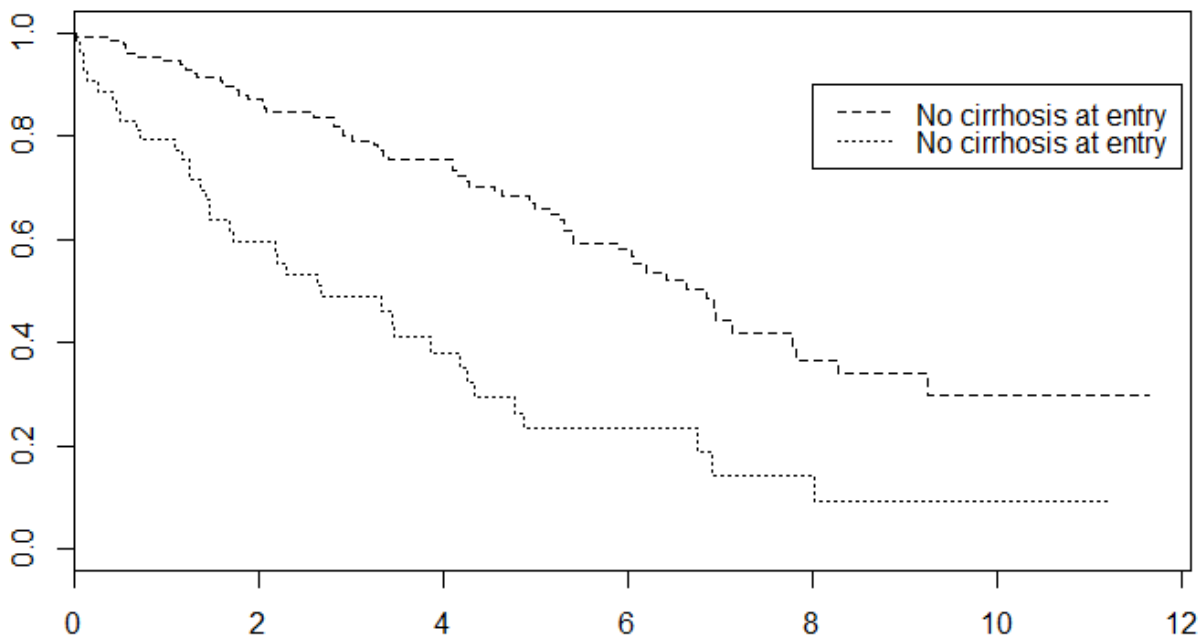

This is not 48%

```
summary(pbc.surv)
     time              status
Min.   : 0.0219   Min.   :0.0000
1st Qu.: 1.6728   1st Qu.:0.0000
Median : 3.5962   Median :1.0000
Mean   : 4.0600   Mean   :0.5217
3rd Qu.: 5.9945   3rd Qu.:1.0000
Max.   :11.6441   Max.   :1.0000
```

# Quite a bit of censored observations



```
surv.cirrhosis <- survfit(pbc.surv ~ pbcdata$cir0 )
plot(surv.cirrhosis, lty = 2:3)
legend(8, .9, c("No cirrhosis at entry", "No cirrhosis at
entry"),lty = 2:3)
```



. Log-Rank Test

```
survdiff(pbc.surv ~ pbcdata$cir0)
Call:
survdiff(formula = pbc.surv ~ pbcdata$cir0)

                  N Observed Expected (O-E)^2/E (O-E)^2/V
pbcdata$cir0=0 131       58     77.5       4.9        26
pbcdata$cir0=1  53       38     18.5      20.5        26

 Chisq= 26  on 1 degrees of freedom, p= 3.48e-07
```