PhD course in Basic Biostatistics – Day 4

Erik Parner, Department of Biostatistics, Aarhus University©

One sample from a binomial
- Model, estimate, exact and approximate inference

Two independent binomial samples
- Model, estimates, measures of association
- Exact and approximate inference
- Sample size and power
- The **Chi-squared test** for 2x2 tables
- **Fishers exact test** for 2x2 tables

One sample of **paired binary data**
- Estimation, McNemars test

The **Chi-squared test** for R x C tables

Test for trend in an ordered R x C table (**Spearman rank**)

Overview

Data to analyse	Type of analysis	Unpaired/Paired	Type	Day
Continuous	One sample mean	Irrelevant	Parametric	Day 1
			Nonparametric	Day 3
	Two sample mean	Non-paired	Parametric	Day 2
			Nonparametric	Day 2
		Paired	Parametric	Day 3
			Nonparametric	Day 3
	Regression	Non-paired	Parametric	Day 5
	Several means	Non-paired	Parametric	Day 6
			Nonparametric	Day 6
Binary	**One sample mean**	**Irrelevant**	**Parametric**	**Day 4**
	Two sample mean	**Non-paired**	**Parametric**	**Day 4**
		Paired	**Parametric**	**Day 4**
	Regression	Non-paired	Parametric	Day 7
Time to event	One sample: Cumulative risk	Irrelevant	Nonparametric	Day 8
	Regression: Rate/hazard ratio	Non-paired	Semi-parametric	Day 8

One sample from a binomial

Ex. 15.3: Smoking among 15-16 year olds in Birmingham

Question: What is the prevalence of smoking among 15-16 year olds in Birmingham and how does it compare to the target 13%?

Design/Data: Self-reported smoking habits (current smoker: Yes/No) among 1000 randomly chosen 15-16 year olds living in Birmingham.

Note, the data for each teenager is **binary** – it can only take two values Yes or No.
One will often code a Yes as 1 and a No as 0.

The total number of Yes's will be a whole number in the range 0 to n=1000.

Result: **123** out of the 1000 teenagers said they were current smokers.

One sample from a binomial

We will make the following **four assumptions**:

1. The sample size n **does not depend** on the observations (e.g. the number of Yes's)
2. The observations are **independent**.
3. There is exactly **the same two possible** outcomes for each teenager: Yes (current smoker) No (not current smoker)
4. The **probability** of being a smoker is **the same** for all the teenagers. Let us denote this unknown probability, π.

The last three assumptions correspond to:

"n independent tosses with the same coin".

If the four assumptions are true, then the number of Yes's, x, follows a **binomial distribution**.

$$x \sim b(n, \pi)$$

Comments to the assumptions behind the binomial model

1. The **sample size** does not need to be determined before we collect the data.
   But we are not allowed to base our decision on how much data to collect, on the number of positive answers.

2. **Independency** is checked, as usual, by going through the design.

3. It does not make sense to analyze the data, if the teenagers did not have exactly the same choice of answers.

4. If the unknown probability, π, of being a current smoker differ in **subgroups**, then it may not be appropriate to analyze the pooled data and report just one number.

Note, the **four assumptions lead** to a binomial distribution.
One **does not** need any additional 'graphical check' like the QQ-plot for the normal model.

Properties of the binomial distribution

If x follows a binomial distribution with sample size n and probability π,

then $$\Pr(x=k;n,\pi)=\frac{n!}{k!(n-k)!}\pi^{k}(1-\pi)^{(n-k)} \qquad k=0,1,\ldots,n$$

The **expected** number of x: $n\cdot\pi$

and the **standard deviation** $\sqrt{n\cdot\pi\cdot(1-\pi)}$

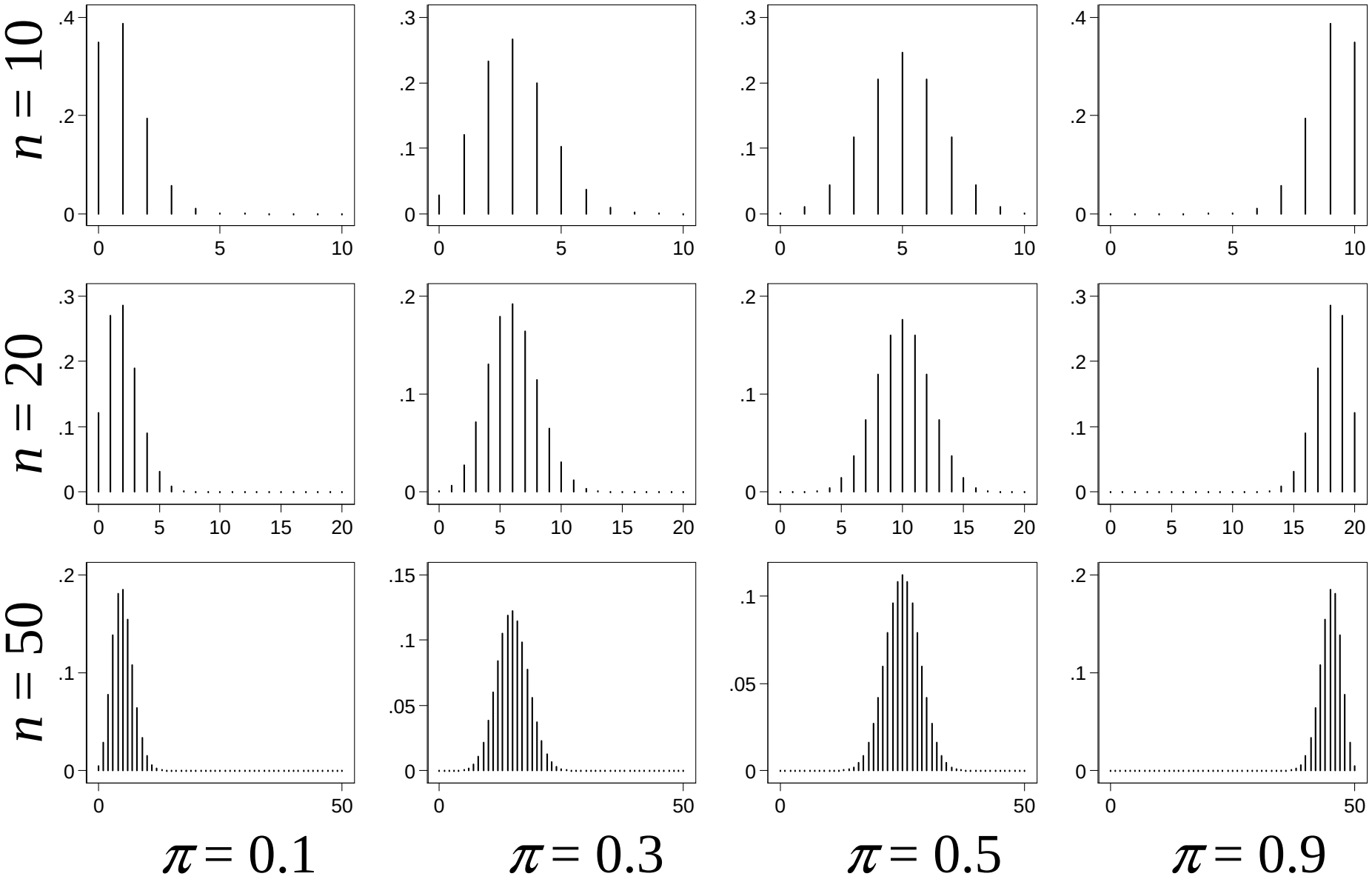
Note, if we know π (and the sample size), then we also know the standard deviation!

Estimation

The unknown probability of Yes is estimated by: $\hat{\pi}=\frac{x}{n}$

- the **observed relative frequency** of Yes.

Some different binomial distributions



Approximate inference in the binomial distribution

There are **many approximate formulas** for the standard error (and test) for the estimate of π in the binomial distribution.

The most simple is: $$se(\hat{\pi}) = \sqrt{\hat{\pi} \cdot (1-\hat{\pi})/n}$$

Based on that one can construct an **approx. 95% CI**:

$$\hat{\pi} \pm 1.96 \cdot se(\hat{\pi})$$

The **hypothesis** that π has a specific value: $\pi = \pi_0$ is tested as usually:

$$z_{obs} = \frac{\hat{\pi} - \pi_0}{se(\hat{\pi})}$$

and a approx. p-value as $$2 \cdot \Pr\left(\text{standard normal} \geq \left|z_{obs}\right|\right)$$

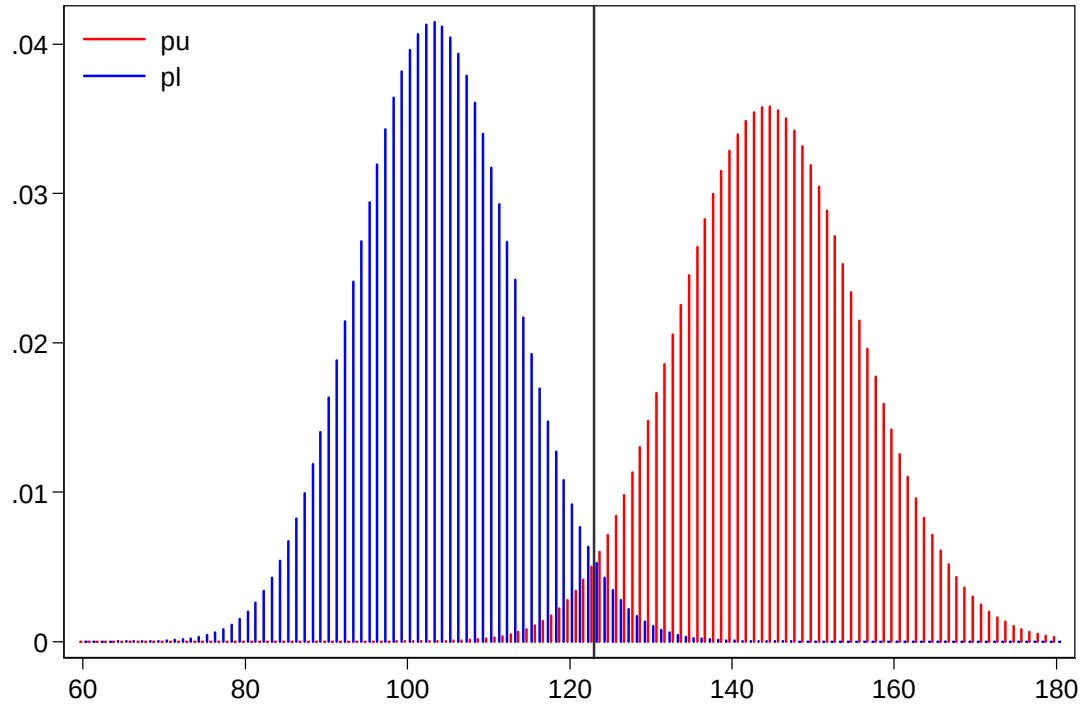
In Stata this is done by `prtest`.
The approximations work ok if the expected number is larger **than 10**.

Exact inference in the binomial distribution - CI

The limits of the **exact** 95%-confidence intervals for π is not based on a standard error, but on solving the equations:

$$\Pr\left(x \geq x_{obs}; \pi = \pi_{Lower}\right) = 0.025$$

$$\Pr\left(x \leq x_{obs}; \pi = \pi_{Upper}\right) = 0.025$$



In Stata 14: "`ci prop` *`variable`*", -13: "`ci` *`variable`*`, bin`

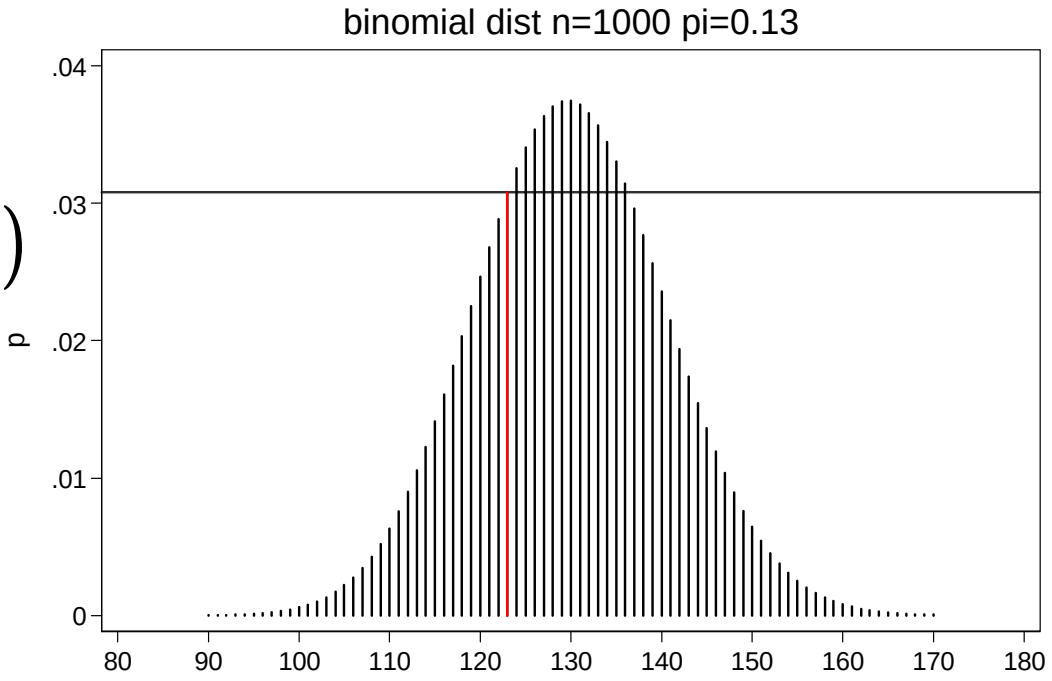
Exact inference in the binomial distribution - test

The hypothesis : $\pi = \pi_0$

The p-value can be defined in different ways –
in Stata (`bitest`) it is done as follows:

The p-value is the probability of observing an event, which is just as or less probable than, what you have seen, given the hypothesis is true, i.e.

$$\text{p-val} = \sum_{\Pr(x;n,\pi_0)\le\Pr(x_{obs};n,\pi_0)} \Pr(x;n,\pi_0)$$



Smoking among 15-16 year olds in Birmingham

Here $n = 1000$ and $x_{obs} = 123$ giving :

$$\hat{\pi} = \frac{123}{1000} = 0.123 = 12.3\%$$

Exact 95% CI: (0.1033; 0.1450)
Approx 95% CI: (0.1026; 0.1434)

The hypothesis: $\pi = 13\% = 0.13$ has the:

Exact p-value: p=0.541
Approx p-value: p=0.510

Stata: One sample from a binomial

Exact analysis (more commands in: Day4.do).

```
. use smokers, clear
. * The exact confidence interval
. ci prop smoker
. * In Stata 13 and prior: ci smoker,bin
                                                    -- Binomial Exact --
Variable |   Obs      Mean    Std. Err.    [95% Conf. Interval]
---------+--------------------------------------------------------
  smoker |  1000      .123    .0103861     .1032769    .1449722

. * Testing the hypothesis pi=0.13  -exact p-value
. bitest smoker=0.13
Variable |    N  Observed k  Expected k  Assumed p   Observed p
---------+--------------------------------------------------------
  smoker | 1000         123         130     0.13000      0.12300

  Pr(k >= 123)                 = 0.757843  (one-sided test)
  Pr(k <= 123)                 = 0.272961  (one-sided test)
  Pr(k <= 123 or k >= 137) = 0.541104  (two-sided test)
```

Smoking among 15-16 year olds in Birmingham - formulations

Methods:
Data was analyzed using exact methods for binomial data. Estimates are given with 95% confidence intervals.

Results:
The prevalence of smoking was 12.3(10.3;14.5)%. This was not statistically different (p=54%) from the target of 13.0%.

Conclusion:
Between 10 and 15 percent of the 15-16 year olds in Birmingham are smoking. The present study is not large enough to determine whether or not the smoking habits in Birmingham satisfies the goal that less than thirteen percent should smoke.

Two independent binomial samples
Example 16.1: Influenza vaccination

Question: What is the effect of vaccination against influenza?

Design/Data: A placebo controlled randomized trial of influenza vaccine on 460 adults. Follow-up period three months after inclusion.

Data:

	Influenza			
	Yes	No	Total	% Yes
Vaccine	20	220	240	8.33%
Placebo	80	140	220	36.36%
Total	100	360	460	21.74%

First impression - the vaccine reduces the risk!

Two independent binomial samples

Statistical model:
Two independent samples from two binomials:

$$x_V \sim b(n_V, \pi_V) \qquad n_V = 240$$

$$x_P \sim b(n_P, \pi_P) \qquad n_P = 220$$

That is, **within** the two groups the design should fulfill the four assumptions on page **4**.

Furthermore, **the two samples** should be independent.

Under this model the two probabilities are, of course, estimated by:

$$\hat{\pi}_V = \frac{x_V}{n_V} \quad \text{and} \quad \hat{\pi}_P = \frac{x_P}{n_P}$$

and the two estimates are **independent**.

Two independent binomial samples

Statistical model:
Two independent samples from two binomials .

This trial will only make sense if the persons in the study are **exposed to influenza virus**!

Effect of the vaccine will depend on the size of this exposure.

Data might not be independent as the exposure to the **virus might cluster**.

Two independent binomial samples

Focus is on comparing the two probabilities π_V and π_P. This can be done by considering one of three measures of association:

Risk difference: $RD = \pi_V - \pi_P$

Risk ratio: $RR = \dfrac{\pi_V}{\pi_P}$

Odds ratio: $OR = \dfrac{\pi_V \cdot (1-\pi_P)}{\pi_P \cdot (1-\pi_V)}$

Note, the hypothesis of no difference between the groups:

$\pi_V = \pi_P$ is equivalent to, $RD = 0$, $RR = 1$ and $OR = 1$.

RR's are often used when studying etiology, RDs when making public health statements and ORs in case-control studies or when the outcome is rare.

Two independent binomial samples

Example

$\pi_1 = 0.10, \pi_2 = 0.15$, RR=1.50

Group 2 has a 50% increase in risk compared to group 1.

$\pi_1 = 0.10, \pi_2 = 0.60$, RD=0.50

Group 2 has a 50% increase in risk compared to group 1.

The two statement sounds similar!

Therefore, we could emphasize that the latter is an absolute difference, for example by

Group 2 has a 50% **point** increase in risk compared to group 1.

The Risk Difference

Risk difference: $RD = \pi_V - \pi_P$

The estimate: $\widehat{RD} = \hat{\pi}_V - \hat{\pi}_P$

The approximative standard error:

$$se\left(\widehat{RD}\right) = \sqrt{se\left(\hat{\pi}_V\right)^2 + se\left(\hat{\pi}_P\right)^2}$$

$$= \sqrt{\hat{\pi}_V \cdot \left(1-\hat{\pi}_V\right)/n_V + \hat{\pi}_P \cdot \left(1-\hat{\pi}_P\right)/n_P}$$

Approx 95%CI(RD): $\widehat{RD} \pm 1.96 \cdot se\left(\widehat{RD}\right)$

It is **not possible** to make exact inference for RD !

The Risk Ratio

Risk ratio: $RR = \pi_V / \pi_P$

The estimate: $\widehat{RR} = \hat{\pi}_V / \hat{\pi}_P$

Inference is made on the **log-scale**.

The approx. stand. error: $se\left(\ln\left(\widehat{RR}\right)\right) = \sqrt{\frac{1}{x_V} - \frac{1}{n_V} + \frac{1}{x_P} - \frac{1}{n_P}}$

Approx 95%CI $\ln(RR)$:

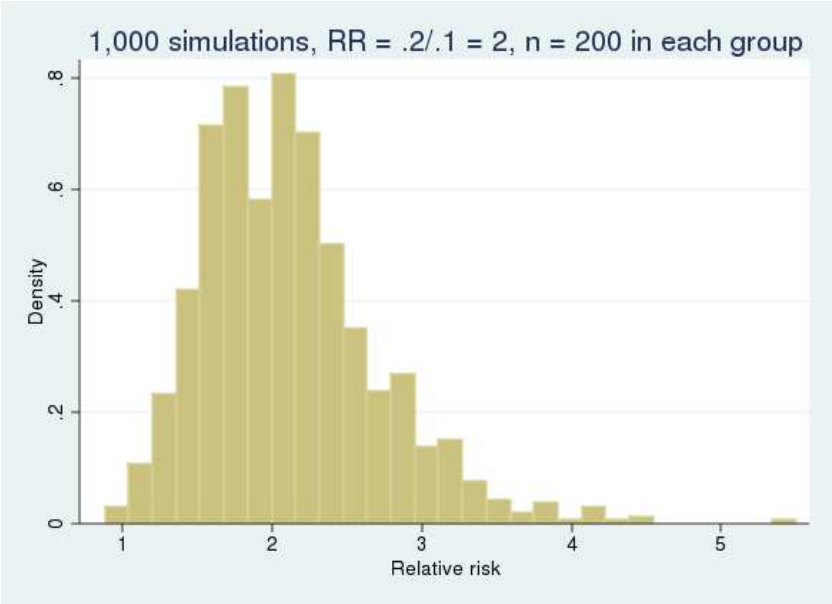
$$\ln\left(\widehat{RR}\right) \pm 1.96 \cdot se\left(\ln\left(\widehat{RR}\right)\right) = \left(\ln\left(\widehat{RR}\right)_{lower}; \ln\left(\widehat{RR}\right)_{upper}\right)$$

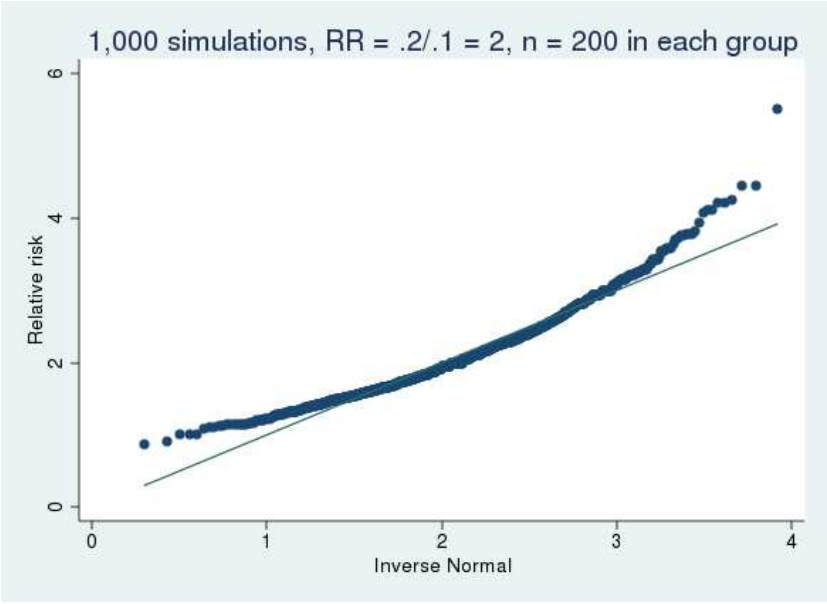
exp

Approx 95%CI RR: $= \left(\exp\left(\ln\left(\widehat{RR}\right)_{lower}\right); \exp\left(\ln\left(\widehat{RR}\right)_{upper}\right)\right)$

It is **not possible** to make exact inference for RR !

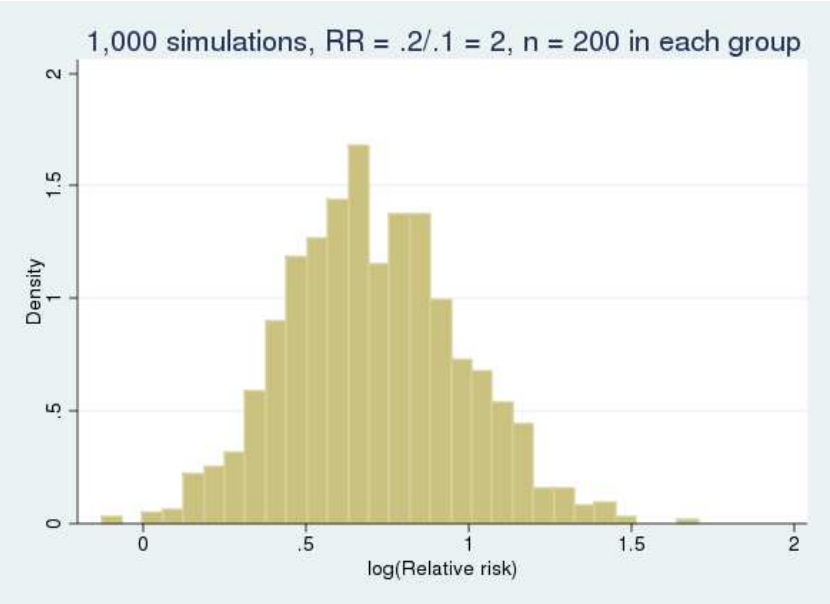
Why analyze Risk Ratio on a log-scale?

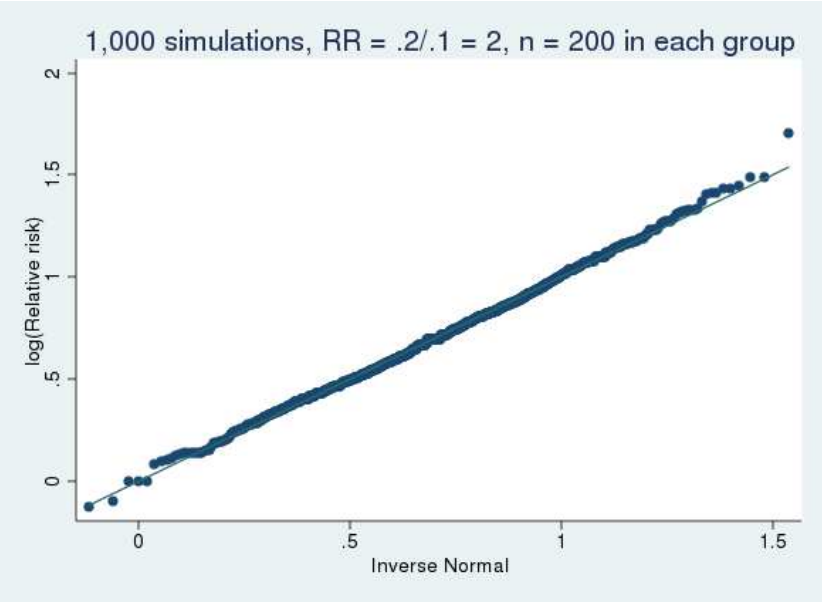




Normality assumption of RR violated on original scale

Why analyze Risk Ratio on a log-scale?





Normality assumption of RR very good on log-scale

The Odds Ratio

Odds ratio: $OR = \dfrac{\pi_V \cdot (1-\pi_P)}{\pi_P \cdot (1-\pi_V)}$ and $\widehat{OR} = \dfrac{\hat{\pi}_V \cdot (1-\hat{\pi}_P)}{\hat{\pi}_P \cdot (1-\hat{\pi}_V)}$

Inference is made on the **log-scale**.

The approx. stand. error:

$$se\left(\ln\left(\widehat{OR}\right)\right) = \sqrt{\frac{1}{x_V} + \frac{1}{n_V - x_V} + \frac{1}{x_P} + \frac{1}{n_P - x_P}}$$

Approx 95%CI $\ln(OR)$:

$$\ln\left(\widehat{OR}\right) \pm 1.96 \cdot se\left(\ln\left(\widehat{OR}\right)\right) = \left(\ln\left(\widehat{OR}\right)_{lower}; \ln\left(\widehat{OR}\right)_{upper}\right)$$

exp

Approx 95%CI OR:

$$= \left(\exp\left(\ln\left(\widehat{OR}\right)_{lower}\right); \exp\left(\ln\left(\widehat{OR}\right)_{upper}\right)\right)$$

It is **possible** to make exact inference for OR ! see later

Changing the event

In the example we considered the risk/probability of getting influenza.

We might instead have considered the risk/probability of **not** getting influenza.

If we do that then three measures of association will change:

$$RD_{\text{not flu}} = -RD_{\text{flu}}$$

$$RR_{\text{not flu}} \neq -RR_{\text{flu}}$$ Not a simple relation

$$OR_{\text{not flu}} = \frac{1}{OR_{\text{flu}}}$$

Comparing the unexposed to the exposed

In the example we compared the risk of getting influenza among vaccinated to that of the placebo-group

We could have compared the placebo-group to the vaccinated.

If we did that then the three measures of association would change:

$$RD_{\text{placebo vs vaccine}} = -RD_{\text{vaccine vs placebo}}$$

$$RR_{\text{placebo vs vaccine}} = \frac{1}{RR_{\text{vaccine vs placebo}}}$$

$$OR_{\text{placebo vs vaccine}} = \frac{1}{OR_{\text{vaccine vs placebo}}}$$

Influenza vaccination - estimates

	estimate	95% CI		
Vaccine influenza	0.0833	0.0516	0.1258	Exact
Placebo influenza	0.3636	0.3000	0.4310	Exact
Risk difference	-0.2803	-0.3529	-0.2078	Approx.
Risk ratio	0.2292	0.1455	0.3610	Approx.
Odds ratio	0.1591	0.0933	0.2713	Approx.

The risk difference is an **additive/absolute** measure.

The risk ratio is a **multiplicative/relative** measure.

2x2 table test of no association

Often one would like to test the **hypothesis of no difference** in the risk in two groups, i.e.:

$$\pi_V = \pi_P \text{ , } RD = 0, RR = 1 \text{ and } OR = 1.$$

This could be done by using one of the **three** estimates and the standard errors as we have seen before.

If one uses this method, then one should remember that the analysis based on the two relative measures RR and OR should be done on the log scale, see next slide.

The **three** tests will give **almost identical p-values**.
If this is not the case, then you have too few data to use any of them.

2x2 table test of no association based on estimates

$$z_{RD} = \frac{\widehat{RD} - 0}{se\left(\widehat{RD}\right)} = \frac{-0.2803 - 0}{\sqrt{\frac{0.0833(1-0.0833)}{240} + \frac{0.3636(1-0.3636)}{220}}}$$

$$= \frac{-0.2803}{0.0370} = -7.57$$

$$z_{RR} = \frac{\ln\left(\widehat{RR}\right) - \ln(1)}{se\left(\ln\left(\widehat{RR}\right)\right)} = \frac{\ln(0.2292) - 0}{\sqrt{\frac{1}{20} - \frac{1}{240} + \frac{1}{80} - \frac{1}{220}}} = \frac{-1.4733}{0.2319} = -6.35$$

$$z_{OR} = \frac{\ln\left(\widehat{OR}\right) - \ln(1)}{se\left(\ln\left(\widehat{OR}\right)\right)} = \frac{\ln(0.1591) - 0}{\sqrt{\frac{1}{20} + \frac{1}{220} + \frac{1}{80} + \frac{1}{140}}} = \frac{-1.8383}{0.2724} = -6.75$$

P<0.0001

2x2 table test of no association
the chi-squared test

Often one would test the hypothesis of no association by the **chi-squared test**.

This test will compare the **observed** cell counts with the **expected** under the hypothesis

$$X^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Large values are critical. The p-value is found by the χ^2 distribution with 1 degree of freedom: $\Pr(\chi^2(1) \geq X^2)$

Observed	Yes	No	Total
Vacine	20	220	240
Placebo	80	140	220
Total	100	360	460

Ecpected	Yes	No	Total
Vacine	52.17	187.83	240
Placebo	47.83	172.17	220
Total	100	360	460

$X^2 = 53.01$ p<0.0001 the hypothesis is rejected.

Stata: Two independent binomial samples

```
. use vaccine, clear
. cs influenza vaccine,or woolf
                | vaccine                |
                |  Exposed   Unexposed   |     Total
----------------+------------------------+-----------
          Cases |       20           80  |       100
      Noncases  |      220          140  |       360
----------------+------------------------+-----------
          Total |      240          220  |       460
                |                        |
           Risk |  .0833333    .3636364  |  .2173913

                |   Point estimate   |  [95% Conf.Interval]
                |--------------------+-----------------------
Riskdifference  |      -.280303      | -.3528516 -.2077545
     Risk ratio |      .2291667      |  .1454585  .3610472
Prev.frac. ex.  |      .7708333      |  .6389528  .8545415
Prev.frac. pop  |      .4021739      |
     Odds ratio |      .1590909      |  .0932823  .2713261 (Woolf)
                +-------------------------------------------
                           chi2(1) =   53.01  Pr>chi2 = 0.0000
```

Stata: Two independent binomial samples

```
. * Chi- squared test using table
. tab2 vaccine influenza,chi2
-> tabulation of vaccine by influenza
           |       influenza
   vaccine |        No        yes |     Total
-----------+----------------------+----------
        No |       140         80 |       220
       yes |       220         20 |       240
-----------+----------------------+----------
     Total |       360        100 |       460
          Pearson chi2(1) =  53.0084   Pr = 0.000
* Fisher's exact test (later)
* tab2 vaccine influenza,exact
```

The influenza vaccine – RD formulations

Methods:
The effect of the vaccine is measured as absolute reduction in risk compared to the placebo group. A Chi-squared test is used to asses the hypothesis of no difference in risk. Estimates are given with 95% confidence intervals.

Results:
In the vaccine group 8.5(5.2;12.6)% acquired influenza compared to 36.4(30.0;43.1)% in the placebo group. This reduction of 28(21;35)% was statistically significant ($p<0.0001$).

Conclusion:
The vaccine decreases the risk of acquired influenza with between 21 and 35 percent points during the influenza season in 199 …..

The influenza vaccine – RR formulations No 1

Methods:
The effect of the vaccine is measured as relative risk of acquiring influenza in the vaccine group compared to the placebo group. A Chi-squared test is used to asses the hypothesis of no difference in risk. Estimates are given with 95% confidence intervals.

Results:
In the vaccine group 8.5(5.2;12.6)% acquired influenza compared to 36.4(30.0;43.1)% in the placebo group. This relative risk of 0.23(0.14;0.36) was statistically significant ($p<0.0001$).

Conclusion:
The vaccine reduced the risk of acquired influenza with between 64 and 86 percent during the influenza season in 199… .

The influenza vaccine – RR formulations No 2

Methods:
The effect of the vaccine is measured as relative risk of acquiring influenza in the placebo group compared to the vaccine group. A Chi-squared test is used to asses the hypothesis of no difference in risk. Estimates are given with 95% confidence intervals.

Results:
In the placebo group 36.4(30.0;43.1)% acquired influenza compared to 8.5(5.2;12.6)% in the vaccine group. This relative risk of 4.4(2.8;6.9) was statistically significant ($p<0.0001$).

Conclusion:
This randomized trial shows that the risk of acquired influenza was between 3 and 7 times higher among the non-vaccinated during the influenza season in 199… ..

Sample size for the two sample binary data – testing no difference

The basis for the power considerations are these five quantities:

$\pi_1 =$ The probability in group one

$\pi_2 =$ The probability in group two

$\alpha =$ The significance level (typically 5%)

$\beta =$ The risk of type 2 error = 1-the power

$n =$ The sample size in each group

The formulas are complicated - use a computer!

Note you can also base it on π_1 and RR, or π_1 and OR using:

$$\pi_2 = RR \cdot \pi_1 \qquad \pi_2 = \frac{OR}{OR + (1-\pi_1)/\pi_1}$$

Sample size for the two sample binary data – testing no difference

Consider the planning of a randomized trial comparing a new treatment with an old standard.

With the old treatment the one-year mortality is 5%.
You suspect that the new treatment will reduce this with 30% that is RR=0.7.

This corresponds to a one-year mortality of 0.05*0.7=0.035.

How many should you include in each arm, if you want a power of 85%?

$$\pi_1 = 0.05, \pi_2 = 0.035, Power = 85\%, \alpha = 5\%$$

Using Stata you get that n= 6494 (per group = 3247)

Stata: Sample size for the two sample binary data

```
. * In Stata 13 and later.
. power twoproportions 0.05 0.035, power(0.85)

Performing iteration ...

Estimated sample sizes for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1
Study parameters:
        alpha =    0.0500
        power =    0.8500
        delta =   -0.0150  (difference)
           p1 =    0.0500
           p2 =    0.0350
Estimated sample sizes:
            N =      6494
  N per group =      3247
. * In Stata 12, and prior.
. * sampsi 0.05 0.035, power(0.85)
```

Exact inference for a two by two table

If you have **few observations** then the approximate methods will not give valid confidence intervals and p-value.

A rule-of-thumb: Few obs. = the smallest expected cell counts is ≤ 5.

It is only possible to find exact confidence intervals for the **Odds Ratio**. The calculation is complicated and we will skip them here.

Furthermore, this is only implemented in a few programs (in Stata in the "cc" command).

The exact test for the hypothesis of no association is called **Fisher's exact test**.

Fisher's exact test for a two by two table

	Bleeding complications		
Treatment	Yes	No	Total
A	1	12	13
B	3	9	12
Total	4	21	25

The idea behind the test is that under the hypothesis the 4 patients will be randomly divided in treatment A and B.

	Bleeding complications		
Treat,	Yes	No	Total
A	0	13	13
B	4	8	12
Total	4	21	25

Prob= 0.039

	Bleeding complications		
Treat,	**Yes**	**No**	**Total**
A	**1**	**12**	**13**
B	**3**	**9**	**12**
Total	**4**	**21**	**25**

Prob= 0.226

	Bleeding complications		
Treat,	Yes	No	Total
A	2	11	13
B	2	10	12
Total	4	21	25

Prob= 0.407

	Bleeding complications		
Treat,	Yes	No	Total
A	3	10	13
B	1	11	12
Total	4	21	25

Prob= 0.271

	Bleeding complications		
Treat,	Yes	No	Total
A	4	9	13
B	0	12	12
Total	4	21	25

Prob= 0.057

$$P-val = 0.039 + 0.226 + 0.057 = 0.322$$

Treatment A vs B – formulations

Methods:
Chi-squared tests are used to test the hypothesis of no association, except when the data are sparse, in which case Fisher's exact test is applied. Estimates are given with 95% confidence intervals.

Results:
One in 13 patients in group A and 3 in 12 in group B experienced bleeding. The difference was not statistically significant (p=32%).

Conclusion:
This study was too small ! .....

Example: Severe cold – paired binary data

Question: Describe the difference in risk of severe cold among 12 and 14 year old boys.

Design: The medical journals for 1319 boys were checked for symptoms of severe cold at the age 12 and 14.

Data: Two observations for each boy. Two different representations of the data:

Severe cold at age 12	Severe cold at age 14	Count
Yes	Yes	212
Yes	No	144
No	Yes	256
No	No	707

Severe cold	Age 14		
Age 12	Yes	No	Total
Yes	212	144	356
No	256	707	963
Total	468	851	1319

Paired binary data – some considerations

The data is the **cross classification** of 1319 observations.

There are **four different possibilities** for each child.

Let us introduce some notation:

Probabilities	Age 14		
Age 12	Yes	No	Sum
Yes	π_{YesYes}	π_{YesNo}	π_{Yes*}
No	π_{NoYes}	π_{NoNo}	π_{No*}
Sum	π_{*Yes}	π_{*No}	1

$$\Pr(\text{cold at 14}) = \pi_{*Yes} = \pi_{YesYes} + \pi_{NoYes}$$

$$\Pr(\text{cold at 12}) = \pi_{Yes*} = \pi_{YesYes} + \pi_{YesNo}$$

$$\Pr(\text{cold at 14}) - \Pr(\text{cold at 12}) = \pi_{NoYes} - \pi_{YesNo}$$

Paired binary data – estimation

A common measure of difference is the risk difference:

$$RD = \Pr(\text{cold at 14}) - \Pr(\text{cold at 12}) \quad = \pi_{\text{NoYes}} - \pi_{\text{YesNo}}$$

That is of course estimated as:

$$\widehat{RD} = \hat{\pi}_{\text{NoYes}} - \hat{\pi}_{\text{YesNo}} = x_{\text{NoYes}} / n - x_{\text{YesNo}} / n$$

There exist several approximate formulas for the standard error. Here is one of them:

$$\text{se}\left(\widehat{RD}\right) = \frac{1}{n}\sqrt{n\cdot\left(\hat{\pi}_{\text{NoYes}} + \hat{\pi}_{\text{YesNo}}\right) - n\cdot\widehat{RD}^2}$$

$$\widehat{RD} = 256/1319 - 144/1319 = 0.1941 - 0.1092 = 0.0849$$

$$\text{se}\left(\widehat{RD}\right) = \frac{1}{1319}\sqrt{1319\cdot(0.1941+0.1092) - 1319\cdot 0.0849^2} = 0.0150$$

$$95\%CI : 0.0849 \pm 1.96 \cdot 0.0150 = (0.0555; 0.1143)$$

Paired binary data – The hypothesis of no difference

The hypothesis of the same risk of severe cold is equivalent to:

$$\Pr(\text{cold at 12}) = \Pr(\text{cold at 14}) \Leftrightarrow$$

$$\pi_{\text{NoYes}} = \pi_{\text{YesNo}} \Leftrightarrow \frac{\pi_{\text{YesNo}}}{\pi_{\text{NoYes}} + \pi_{\text{YesNo}}} = \frac{1}{2}$$

That is the discordant pairs should be divided **fifty-fifty** in the **YesNo** and the **NoYes** cells.
The test of this is called the **McNemar's** test.
There exists both an exact version based on the binomial distribution as well as an approximate one.

Exact test: 144 out of 400=256+144 : pval=0.0001

Stata: Paired binary data

We first compute the prevalences to help interpret the paired binary analysis output on the next overhead:

```
. use cold, clear
. ci prop pold14
                                                    Binomial Exact
 Variable |   Obs      Mean    Std. Err.   [95% Conf. Interval]
----------+-----------------------------------------------------
   cold14 |  1319  .3548143    .0131741     .3289622   .3813151
. ci prop cold12
                                                    Binomial Exact
 Variable |   Obs      Mean    Std. Err.   [95% Conf. Interval]
----------+-----------------------------------------------------
   cold12 |  1319  .2699014    .0122228     .2461006    .294732
* In Stata 13 and prior: "ci cold14, bin" and "ci cold12, bin".
```

Stata: Paired binary data

```
. mcc cold14 cold12
                    | Controls                |
Cases               |   Exposed   Unexposed   |      Total
--------------------+-------------------------+-----------
          Exposed   |       212         256   |        468
        Unexposed   |       144         707   |        851
--------------------+-------------------------+-----------
            Total   |       356         963   |       1319

McNemar's chi2(1) =     31.36    Prob > chi2 = 0.0000
Exact McNemar significance probability       = 0.0000
Proportion with factor
        Cases         .3548143
        Controls      .2699014     [95% Conf. Interval]
                    ---------     --------------------
        difference    .0849128      .0547911   .1150345
        ratio         1.314607      1.194231   1.447116
        rel. diff.    .1163032      .0780381   .1545684
        odds ratio    1.777778      1.443859   2.195911   (exact)
```

Severe cold - formulations

Methods:
The difference in incidence of severe cold at age 14 compared to at age 12 was described by a risk difference. The hypothesis of no difference in risk was tested by McNemar's test. Estimates are given with 95% confidence intervals.

Results:
The incidence of severe cold was 35.5(31.9;38.1)% at age 14 and 27.0(26.6;29.5)% at age 12, corresponding to a difference in incidence of 8.5(5.5;11.5)%. The difference was highly statistically significant ($p<0.0001$).

Conclusion:
The incidence of severe cold is between 5.5 and 11.5 percent points higher at age 14……

Sample size for paired binary data – testing no difference

There are three ways of specifying the assumptions

1. $\pi_{YesNo}/(\pi_{YesNo}+\pi_{NoYes})$ (one sample binary problem)
2. π_{YesNo} and π_{NoYes}
3. π_{*No}, π_{Yes*} and $Corr(Y_1,Y_2)$

where $Corr(Y_1,Y_2)$ is the Pearson correlation (see Day 5) between the paired binary data for one individual.

Suppose we assume that

P(cold 12 years)=π_{Yes*} =0.30

P(cold 14 years)=π_{*No} =0.40

$Corr(Y_1,Y_2)$=0.30

Stata: Sample size for paired binary data

```
. power pairedproportions 0.30 0.40 , corr(0.30)
Performing iteration ...
Estimated sample size for a two-sample paired-proportions test
Large-sample McNemar's test
Ho: p+1 = p1+  versus  Ha: p+1 != p1+
Study parameters:
        alpha =    0.0500
        power =    0.8000
        delta =    0.1000  (difference)
          p1+ =    0.3000
          p+1 =    0.4000
         corr =    0.3000
Estimated sample size:
            N =       253
```

Test of no association in a RxC table

Example 17.3: 150 households cross tabulated into village and water source.
Hypothesis: No association between village and water source.

$$X^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Large values are critical.
The p-value is found in a χ^2 distribution with df=(R-1)x(C-1).

Observed	Water source			
Village	River	Pond	Spring	Total
A	20	18	12	50
B	32	20	8	60
C	18	12	10	40
Total	70	50	30	150

Excepted	Water source			
Village	River	Pond	Spring	Total
A	23.33	16.67	10.00	50
B	28.00	20.00	12.00	60
C	18.67	13.33	8.00	40
Total	70	50	30	150

$$X^2 = 3.54,\ df = (3-1)\cdot(3-1) = 4,\ p = 0.47$$

The hypothesis of no association cannot be rejected!

Test of no association in a RxC table

Comments:

The test is **valid** no matter whether data is collected:

with only the total number known in advance
- 150 households cross tabulated

with the **row sums fixed**
- the number of households in each village is fixed

with the **column sums fixed**
- the number of households at each water source is fixed

The **expected number** in each cell should be **above five** – otherwise one should use a test like Fisher's exact test.

It is only a test!
If the hypothesis is rejected then look at the discrepancies between the observed and the expected cell counts to understand why!

Test of no association in a RxC table
Ordered categories

Example 17.4: 583 women cross tabulated into age at menarche and triceps skinfold group.

Hypothesis: Age at menarche and size of triceps skinfold.

Note, the triceps skinfold groups are ordered and if one expects that deviations from the hypothesis will follow this ordering, then one should apply some kind of **test for trend**.

There exists several of these.
One is based on **Spearman's rank correlation**, see next week.

Age at menarche	Triceps skinfold group Small	Intermediate	Large	Total
<12	15	29	36	80
12+	156	197	150	503
Total	171	226	186	583
Percentage	9%	13%	19%	14%

Spearman's rank corr. $= -0.12$

$p=0.0035$

The hypothesis is rejected. Skinfold decrease with age at menarche.

Stata: Test of no association - ordered categories

```
. use triceps,clear
. spearman age triceps
 Number of obs =     583
Spearman's rho =      -0.1209
Test of Ho: age and triceps are independent
    Prob > |t| =       0.0035
```

Test of no association in a RxC table
Ordered categories

Comments to Spearman's rank correlation test:

The test is **valid** no matter whether data is collected:

- with only the total number known in advance
- with the **row sums fixed**
- with the **column sums fixed**

The test will work even on data with sparse cells.

To make sense both the columns and rows should be ordered or binary.

There are several other "tests for trend in RxC tables" - these will typically give comparable p-values.

If the hypothesis is rejected then look at the discrepancies between the observed and the expected cell counts to understand why!