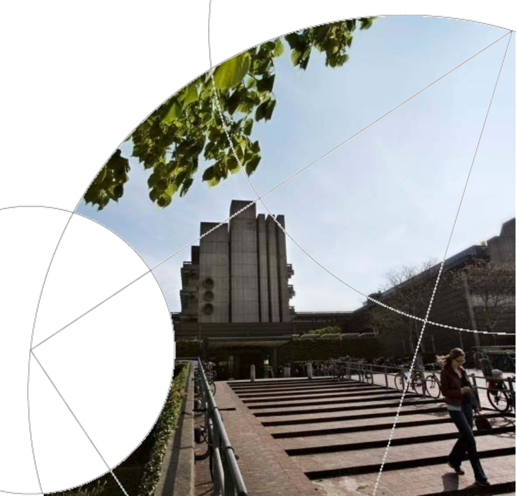
Test in 2by2 tables, RCTs, and power

Theis Lange
Department of Biostatistics, University of Copenhagen
&
Center for Statistical Science, Peking University.

Mail: thlan@sund.ku.dk



TESTS IN COUNT TABLES

A recent study

- Is the hospital mortality different in weekends compared to weekdays?
- The following data on 1500 randomly selected patients were collected from Danish hospitals.

	Survived	Dead in hospital
Admited on weekday	1338	128
Admited on weekend / holiday	27	7

Independence hypothesis – count tables

Our null hypothesis is:

H_0: the two categorical variables are independent of each other.

Example: H_0 : mortality is independent of time of admission.

Alternative formulations of the same hypothesis:

- Mortality does not depend admission time.
- Mortality rate is the same for weekend and weekday patients.

$P_{weekend}$ = 7/(7+27) =0.21

P_{week} = 128/(128+1338) =0.087

Add marginal tables

	Survived	Dead in hospital	Total
Admited on weekday	1338	128	1466
Admited on weekend/holiday	27	7	34
Total	1365	135	1500

Expected values/counts

- How would the table look if mortality and admission time was indeed independent?

- In row nr i and column nr j would the expected values be

$$E_{ij} = n \cdot \frac{n_i}{n} \cdot \frac{d_j}{n} = \frac{n_i \cdot d_j}{n}.$$

Expected value = total count × row-% × column-%

- **Example:** The expected number of dead in weekend group would be:

  1500*(34/1500)*(135/1500) = 3.06

  i.e. somewhat fewer than the observed number of 7 persons.

Expected counts in our example

- Expected (*E*) count are added in parentheses.

	Survived	Dead in hospital	Total
Admitted on weekday	1338 (1334)	128 (132)	1466
Admitted on weekend	27 (31)	7 (3)	34
Total	1365	135	1500

The chi-sq test

To test the hypotheses about independence:

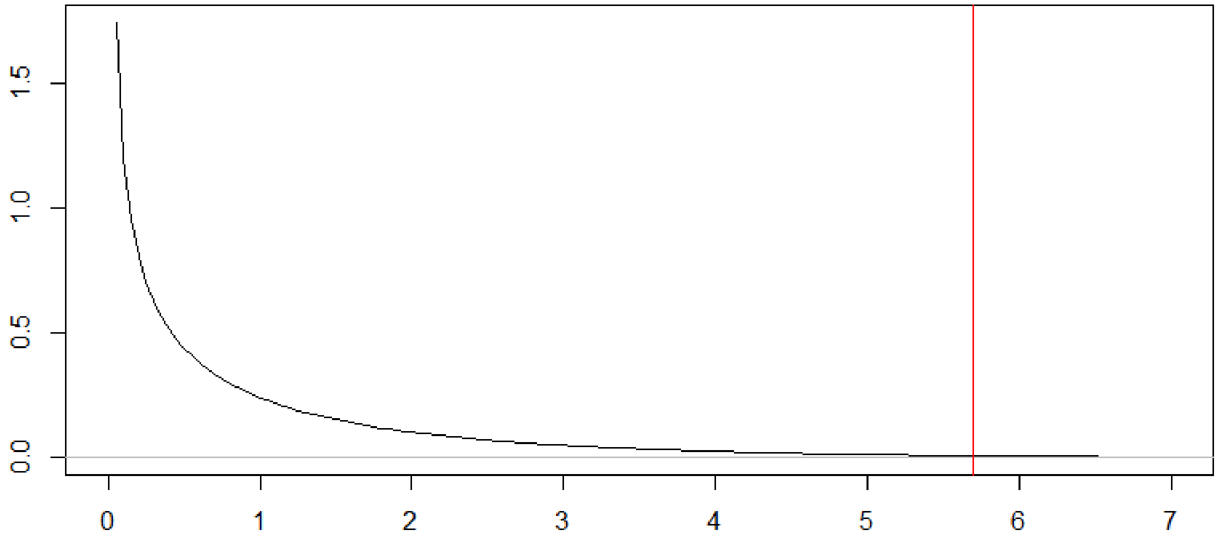
$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

If the null hypothesis is true the quantity approximately follows a χ^2-distribution with $\text{df} = (r-1) \cdot (c-1)$ degrees of freedom

You reject when the χ^2 is large.

Example: χ^2 =(1338-1334)2/1334 +... = 5.7

To get p-value

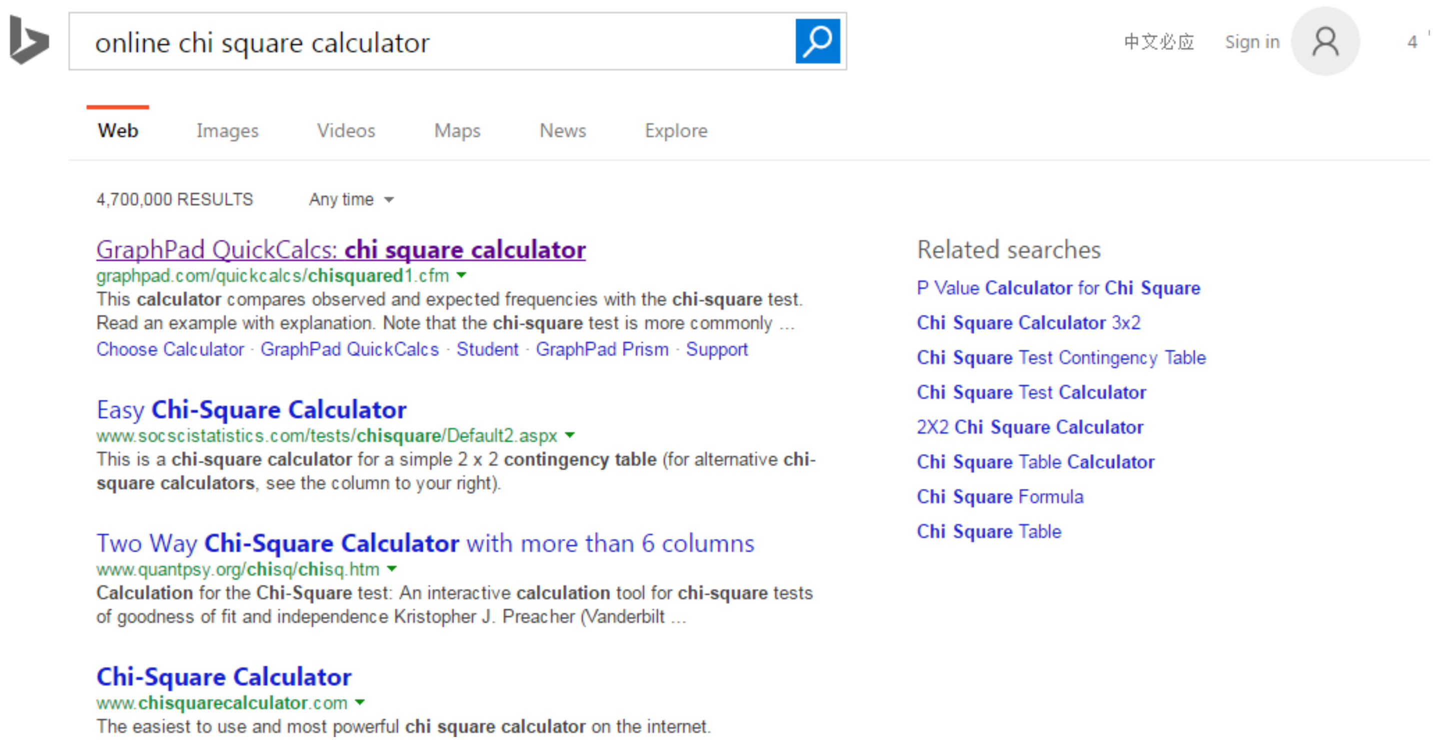


Right tail is: 0.01693

Conclusion???

Chi-sq test in computers

- `chisq.test` function in R.
- Cross tab in SPSS.
- Or online:



The chi-sq test in R

Console ~/

```
> tab
     [,1] [,2]
[1,] 1338  128
[2,]   27    7
> chisq.test(tab, correct = F)

	Pearson's Chi-squared test

data:  tab
X-squared = 5.7041, df = 1, p-value = 0.01693

Warning message:
In chisq.test(tab, correct = F) :
  Chi-squared approximation may be incorrect
> 
```

Tests in sparse tables

- Fisher's exact test calculate exact probability for all possible outcomes (ie. All possible tables) with certain margins:

- Alternatively use traditional chi-sq, but bootstrap p-values.

```
> fisher.test(tab)

        Fisher's Exact Test for Count Data

data:  tab
p-value = 0.02783
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.9757602 6.5347754
sample estimates:
odds ratio
   2.70743

>
>
> chisq.test(tab, simulate.p.value = TRUE, B = 10^7)

        Pearson's Chi-squared test with simulated p-value (based on 1e+07 replicates)

data:  tab
X-squared = 5.7041, df = NA, p-value = 0.02786
```

Conclusion?

- What does our analysis say about mortality in weekends?
- Is this surprising?
- Could there be an alternative explanation?

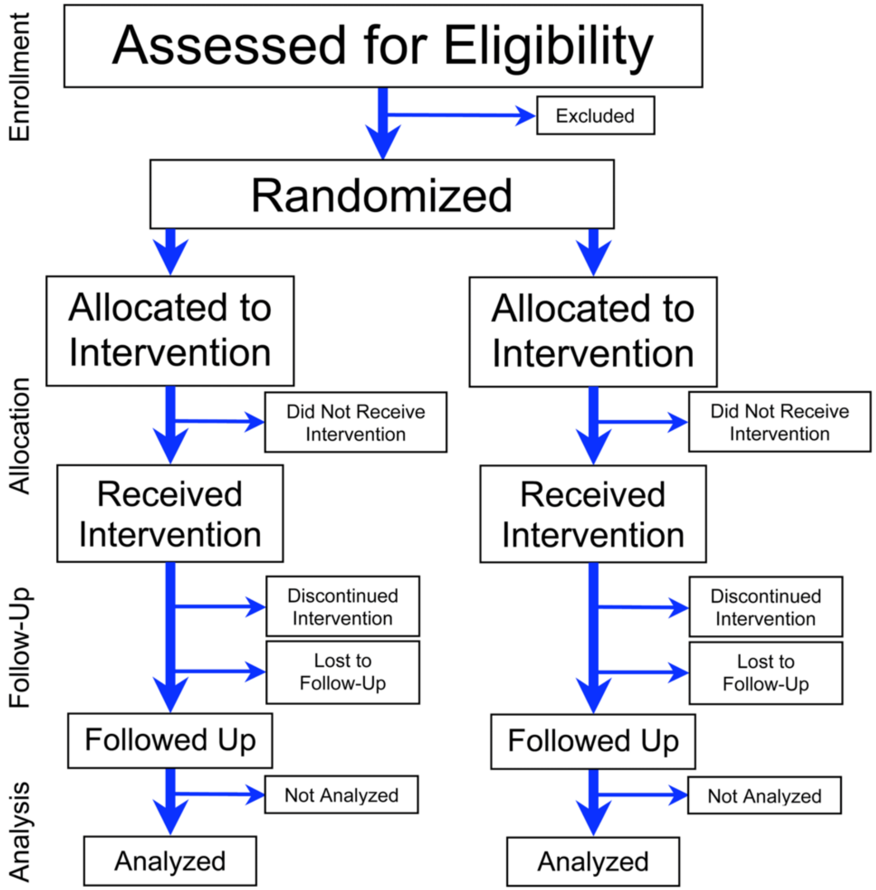
Dead by weekend – the real data.

	In-hospital mortality				
	Patient time at risk		Deaths		Mortality rate ratio
	Days (1,000s)	Percent	Number	Percent	
BED OCCUPANCY RATE (%)					
0–79	3,800	20	22,182	20	1.00
80–89	3,320	18	18,736	17	1.02
90–99	4,507	24	25,670	23	1.00
100–109	4,543	24	27,129	24	1.02
110 or more	2,618	14	17,455	16	1.09**
ADMITTED DURING NORMAL WORKING HOURS[a]					
Yes	17,148	91	48,100	43	1.00
No	1,640	9	63,072	57	10.63**
ADMITTED DURING WEEKDAYS					
Yes	17,080	91	75,963	68	1.00
No[b]	1,709	9	35,209	32	2.23**
SEX					

RANDOMIZED TRIALS

Randomized experiments

- We can ensure comparable groups by allocating to groups ourselves.
- This must be done “by the flip of a coin”.
- This is a so-called RCT.
- First done in 1948 by Sir Bradford Hill to examine “Streptomycin treatment of pulmonary tuberculosis”



Design of a RCT

Before including any patient the following must be written down.

Non-statistical:

- Define the two interventions.
- Define exclusion and inclusion criterions.

Statistical:

- Decide on the tool(s) used to analyze data.
- Determine the number of patients to randomize.

Power of a test

- The power of a statistical test is defined as

  the probability of rejecting a false null hypothesis

- Recall that *size* of a test is probability of falsely rejecting a true null hypothesis.

- Power clearly depend on how false the null hypothesis is.

- In designing an RCT one will decide the minimal difference between treatment groups that the study should be able to detect.
  - Power is typically set to 90%.
  - For binary outcomes you also need to guess risk in baseline group.
  - For continuous outcomes you need to guess standard deviation.

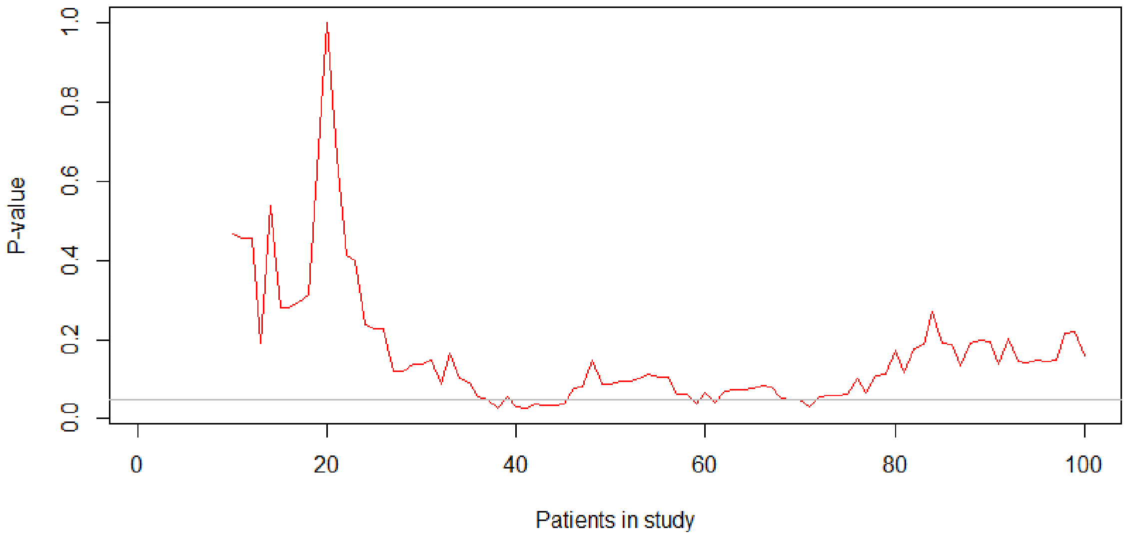
Formula for power in R

```
Console ~/
> power.prop.test(p1 = 0.02, p2 = 0.04, power=0.9)

     Two-sample comparison of proportions power calculation

              n = 1526.752
             p1 = 0.02
             p2 = 0.04
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group

> 
```

- Here we assume equal proportions in each group.
- More complex designs can only be assessed by simulations.
  1. Simulate data according to assumptions.
  2. Conduct pre-determined test on simulated data
  3. Record how often the test is rejected (ie. repeat 1-2 many times).
  4. Repeat until you find the sample size that gives you 90% power to rejec.

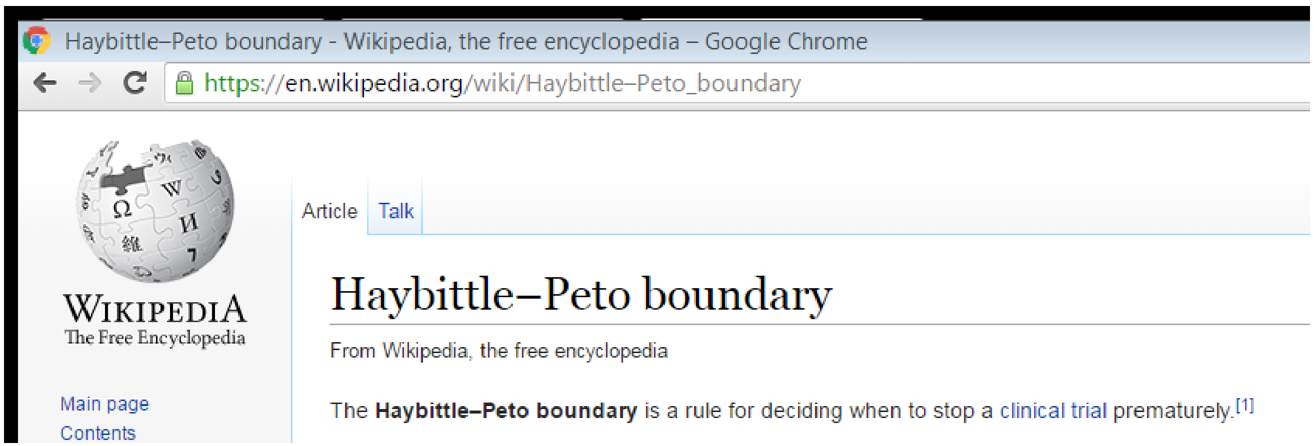
Early stopping of trial

- If treatment turns out to be either very beneficial or very harmful you might want to stop the trial early.
- Can you just test the null-hypothesis every time you get a outcome for a new patient?
- Let us try:
  - Simulate binary outcome
  - P(Y=1)=0.5 for both groups (ie. no treatment effect)
  - Original sample size is N=100.
  - Test after each patient is included.
  - P-value trajectories are depicted below.



Pragmatic solution

- Only do a few interim analyses.
- In interim analyses use much lower value for significance.



List of *p*-values used at each interim analysis, assuming the overall *p*-value for the trial is 0.05

Number of planned analyses	Interim analysis	*p*-value threshold
2	1	0.001
	2 (final)	0.05
3	1	0.001
	2	0.001
	3 (final)	0.05
4	1	0.001

A case: Comment by Shawn and Lange in Lancet 2015

- Bubble continuous positive airway pressure (CPAP) might be a very effective treatment for severe pneumonia in children.
- Mohammond Chisti and colleagues report a randomised trial carried out in Bangladesh, in which 225 children aged under 5 years with severe pneumonia were enrolled.
- The original power analysis indicated 650 children should be included.
- The trial was stopped after only 146 children had been included.

- When the trial was stopped this was the data:

	Failure	Success
CPAP	5	74
Low Flow	16	51

What was p-value at time of stopping?

	Failure	Success
CPAP	5	74
Low Flow	16	51

Done in the paper:

- Using simple chi-sq: 0.0026
  But are the approximation still good here?

Better:

- Using Fisher's: 0.00381

Ethics of early stopping

- Should the trialed had continued?
  (more children would die)

- We must balance
  A. The sacrifices already made by the included patients.
  B. Coming patients to be included in the trial.
  C. Future patients which will never be in this trial, but could benefit from the result of the trial.

- As I wrote in the Lancet I think the authors got it wrong here.