



INM373 - RMPI. Lecture 6: Quantitative Methods (part 2) Statistical Hypothesis Testing.

Andrey A. Povyakalo
andrey@city.ac.uk

v1.0

Outline¹

- Inferential statistics
- How we do this (CAD in mammography)
- Confidence intervals
- Sally Clark case: Prosecutor's fallacy
- Alarm systems: Base rate fallacy
- Testing hypotheses
- Example: Testing hypotheses about the mean
- MMR story: Meaning of Statistical significance. Power of the test.
- Statistical 'Paradoxes'
 - Simpson's paradox
 - Regression towards the mean
 - Statistical independence and covariance
- Conclusions
- Revision exercises

¹<http://moodle.city.ac.uk/mod/resource/view.php?id=1079292>

Inferential Statistics

- Rowntree (2000):
 - going beyond what has been observed
 - use observations as a basis for making estimates or predictions
 - * inferences about a situation that has not been observed
 - generalising from a sample to the population

“comparing what we have observed with what we might expect”

Most Important Concepts to take away

- Statistical test
- Likelihood
- Nil hypothesis. Alternative
- Statistical significance. P-value
- Power of the test
- Confidence interval

Steps in inferential statistical analysis

1. Formulation of hypotheses
2. Specification of statistical test
3. Computation of the test statistic
4. Identification of a level of significance (confidence)
5. Construction of a decision rule
6. Decision

<https://www.graphpad.com/support/faqid/1790/>

| | Type of Data | | | |
|---|--|--|--|---|
| Goal | Measurement (from Gaussian Population) | Rank, Score, or Measurement (from Non-Gaussian Population) | Binomial (Two Possible Outcomes) | Survival Time |
| Describe one group | Mean, SD | Median, interquartile range | Proportion | Kaplan Meier survival curve |
| Compare one group to a hypothetical value | One-sample t test | Wilcoxon test | Chi-square or Binomial test ** | |
| Compare two unpaired groups | Unpaired t test | Mann-Whitney test | Fisher's test (chi-square for large samples) | Log-rank test or Mantel-Haenszel* |
| Compare two paired groups | Paired t test | Wilcoxon test | McNemar's test | Conditional proportional hazards regression* |
| Compare three or more unmatched groups | One-way ANOVA | Kruskal-Wallis test | Chi-square test | Cox proportional hazard regression** |
| Compare three or more matched groups | Repeated-measures ANOVA | Friedman test | Cochrane Q** | Conditional proportional hazards regression** |
| Quantify association between two variables | Pearson correlation | Spearman correlation | Contingency coefficients** | |
| Predict value from another measured variable | Simple linear regression or Nonlinear regression | Nonparametric regression** | Simple logistic regression* | Cox proportional hazard regression* |
| Predict value from several measured or binomial variables | Multiple linear regression* or Multiple nonlinear regression** | | Multiple logistic regression* | Cox proportional hazard regression* |

How we do this . . .

Example: CAD²in mammography (1/7)

Povyakalo, A. A., Alberdi, E., Strigini, L. and Ayton, P. (2013). How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. *Medical Decision Making*, 33(1), pp. 98-107. doi: 10.1177/0272989X12465490

²Computer **A**ided **D**etection

Example: CAD in mammography (Sensitivity and Specificity) (2/7)

| | Cancer | Normal |
|------------|-------------------------------|-------------------------------|
| Recall | TP $Sens = TP / (TP + FN)$ | FP |
| Not recall | FN | TN $Spec = TN / (TN + FP)$ |

- $Eff.sens = Sens.prompted - Sens.unprompted$;
- $Eff.spec = Spec.prompted - Spec.unprompted$;
- Hypotheses to test:
 - Sensitivity: $H_0: Eff.sens = 0$ (simple) vs $H_1: Eff.sens \neq 0$; (complex)
 - Sensitivity: $H_0: Eff.spec = 0$ (simple) vs $H_1: Eff.spec \neq 0$; (complex)
- One can test hypotheses using confidence intervals.

Example: CAD in mammography (3/7)

Background: Computer aids can affect decisions in complex ways, potentially even making them worse; common assessment methods may miss these effects. We developed a method for estimating the quality of decisions and how computer aids affect it, and applied it to computer-aided detection (CAD) of cancer, re-analysing data from a published study where 50 professionals (“readers”) interpreted 180 mammograms, both with and without computer support.

Method: We used stepwise regression to estimate how CAD affected the probability of a reader making a correct screening decision on a patient with cancer (sensitivity), thereby taking into account the effects of the difficulty of the cancer (proportion of readers who missed it) and the reader’s discriminating ability (Youden’s Determinant.) Using regression estimates we obtained thresholds for classifying a posteriori the cases (by difficulty) and the readers (by discriminating ability).

Example: CAD in mammography (4/7)

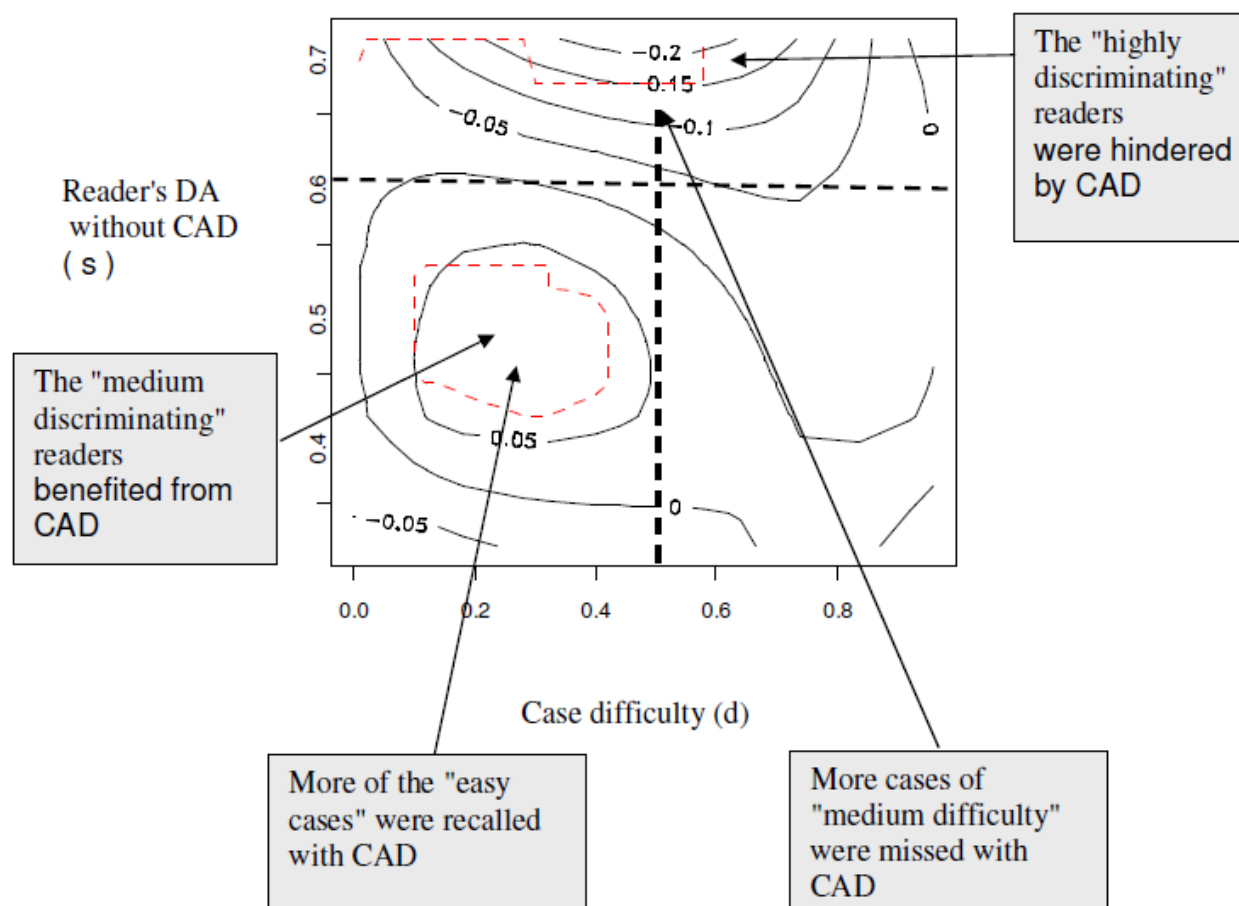
Results: Use of CAD was associated with a 0.016 increase in sensitivity (95% CI: 0.003, 0.028) for the 44 least discriminating radiologists for 45 relatively easy, mostly CAD-detected, cancers. However, for the 6 most discriminating radiologists, with CAD sensitivity decreased by 0.145 (95% CI: 0.034, 0.257) for the 15 relatively difficult cancers.

- H_0 rejected, because $Eff.sens = 0$ is outside the confidence intervals.

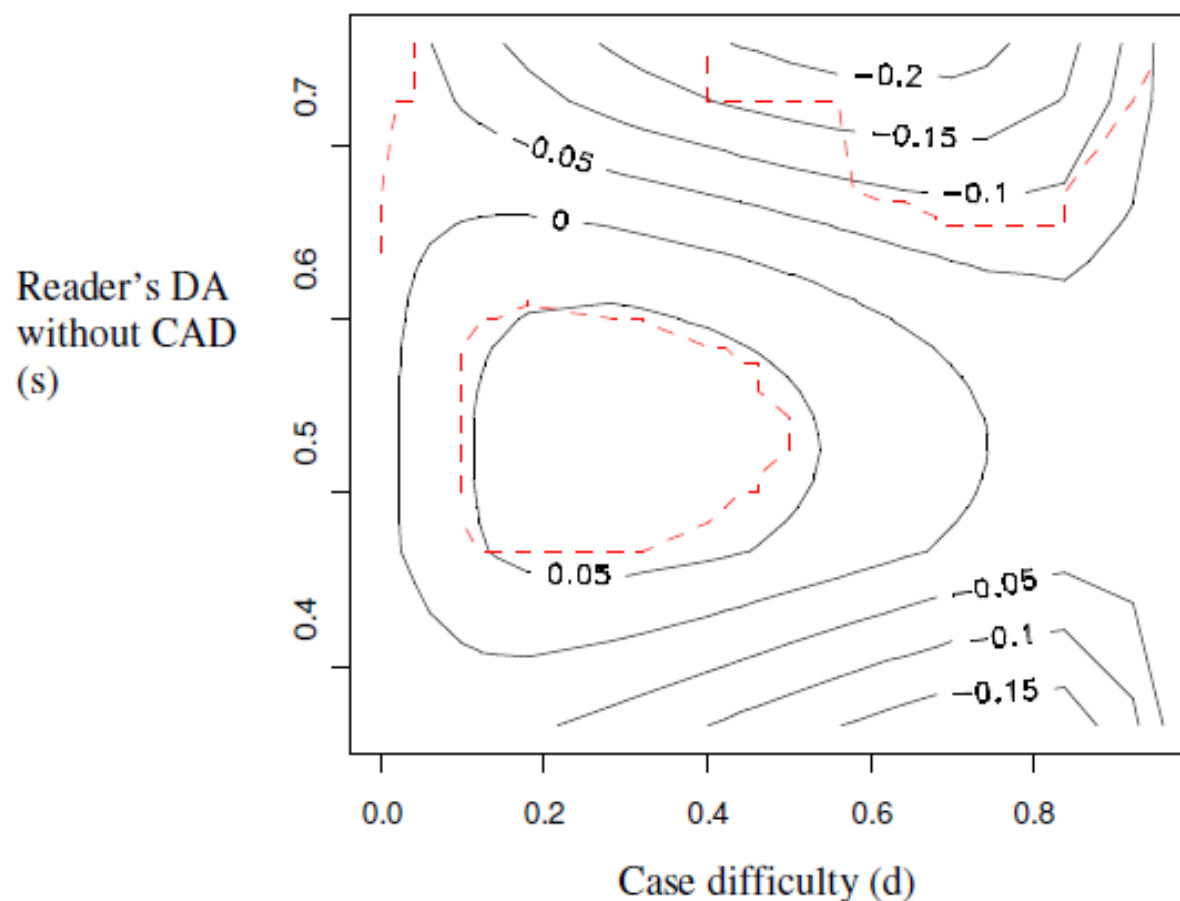
Example: CAD in mammography (Confidence intervals) (5/7)

- **Confidence interval** is an interval random estimate of the parameter of interest (guess), based on data (observation)
- For a given experiment outcome, a confidence interval is either right (covers the true unknown value of the parameter of interest) or wrong (does not cover)
- Confidence (e.g. 0.95) is a reliability measure for the confidence procedure
- Confidence interval includes values of the parameter of interest (e.g. sensitivity) which, are in some sense, consistent with the observations

Example: CAD in mammography (Impact) (6/7)



Example: CAD in mammography (Paired outcomes) (7/7)



Case study: Sally Clarke case ³

- Clark's first son died suddenly within a few weeks of his birth in 1996. After her second son died in a similar manner, she was arrested in 1998 and tried for the murder of both sons. Her prosecution was controversial due to statistical evidence presented by pediatrician Professor Sir Roy Meadow, who testified that the chance of two children from an affluent family suffering sudden infant death syndrome was 1 in 73 million, which was arrived at by squaring 1 in 8500 for likelihood of a cot death in similar circumstance.

³http://en.wikipedia.org/wiki/Sally_Clark

Sally Clarke case: Sir Roy Meadow ⁴

- Sir Samuel Roy Meadow (born 1933) is a British paediatrician and professor, who rose to initial fame for his 1977 academic paper on the now controversial Munchausen Syndrome by Proxy (MSbP). He was knighted for this work. He endorsed the dictum that “*one sudden infant death is a tragedy, two is suspicious and three is murder, until proved otherwise*” in his book ABC of Child Abuse ⁵ and this became known as Meadow’s Law and at one time was widely adopted by social workers and child protection agencies (such as the NSPCC) in Britain

⁴http://en.wikipedia.org/wiki/Roy_Meadow

⁵Meadow, Roy (May 1997). ABC of Child Abuse. BMJ books. p. 100. ISBN 0-7279-1106-6.

Prosecutor's fallacy⁶ (1/2)

- E is the observed evidence
- I stands for "accused is innocent"
- $P(E|I)$ is the probability that the "damning evidence" would be observed even when the accused is innocent (a "false positive")
- $P(I|E)$ is the probability that the accused is innocent, despite the evidence E .
- With forensic evidence, $P(E|I)$ is tiny. The prosecutor wrongly concludes that $P(I|E)$ is comparatively tiny.

⁶https://en.wikipedia.org/wiki/Prosecutor%27s_fallacy

Prosecutor's fallacy (2/2)

- In fact, $P(E|I)$ and $P(I|E)$ are quite different;
- using Bayes' theorem:

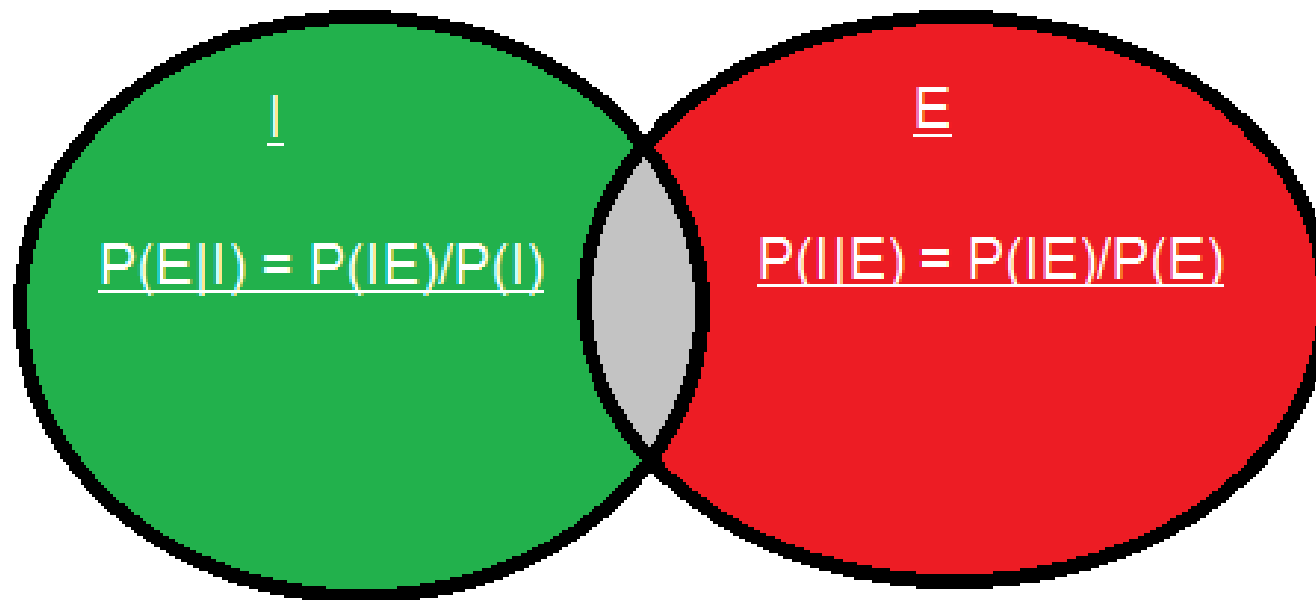
$$P(I|E) = P(E|I) \cdot P(I) / P(E)$$

where:

$P(I)$ is the probability of innocence independent of the test result (i.e. from all other evidence) and $P(E)$ is the prior probability that the evidence would be observed (regardless of innocence).

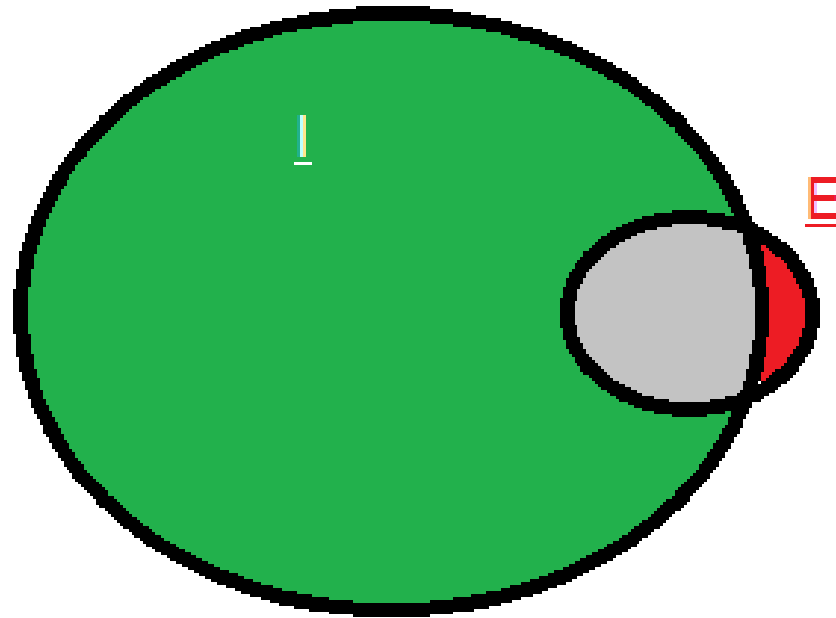
⁶https://en.wikipedia.org/wiki/Prosecutor%27s_fallacy

In diagrams: What RM, apparently, thought . . .

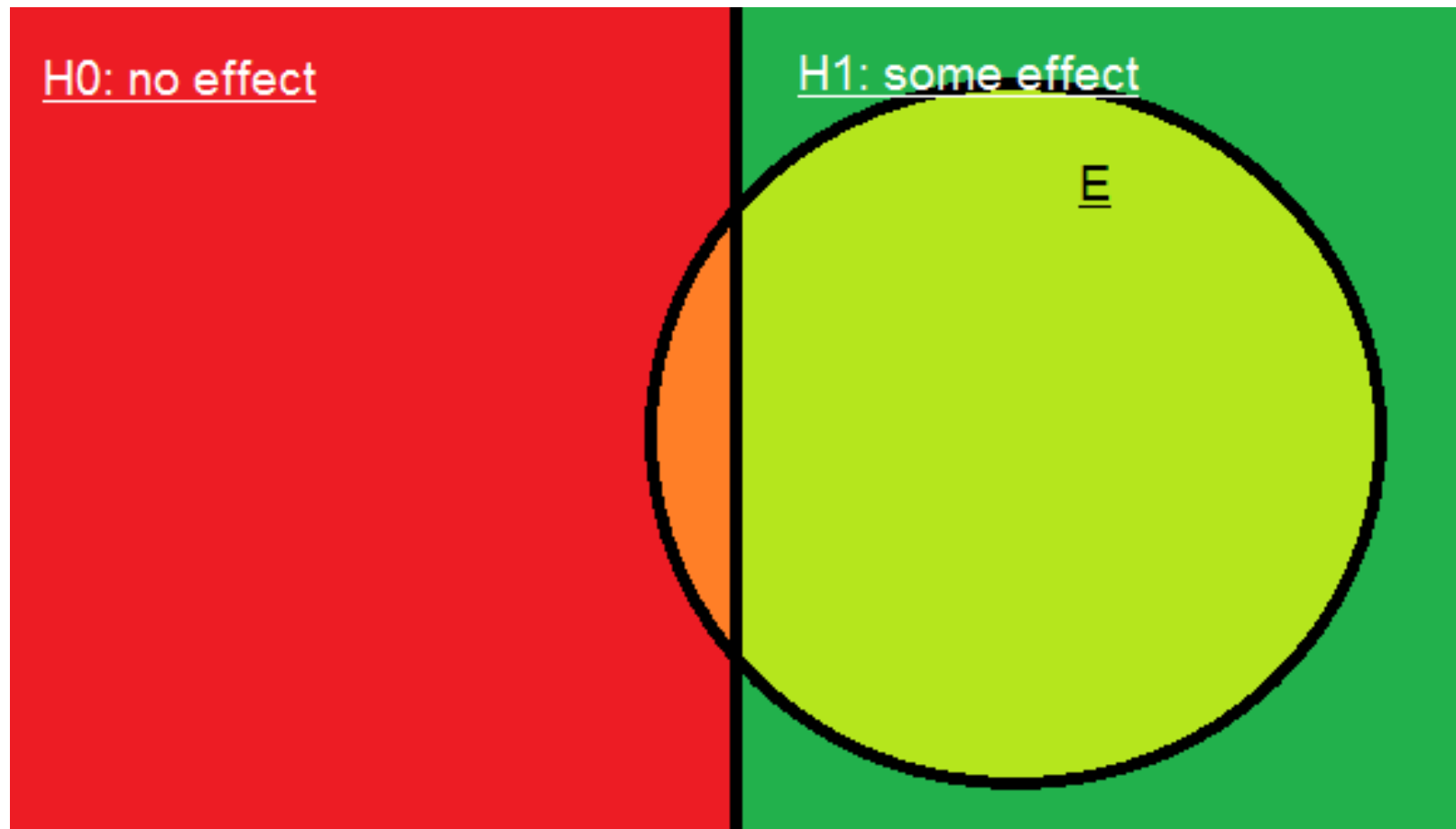


$P(E|I)$ small implies $P(I | E)$ small

In diagrams: What RM, apparently, missed . . .



In Diagrams: How scientists think . . .



Reject H_0 , because $P(E|H_0)$ small implies $P(H_0|E)$ small

What do you think?

Respond at [PollEv.com/andreypovak436](https://poll-ev.com/andreypovak436)

Text a **CODE** to **22333**

Scientists reduce chances for committing the Prosecutor's Fallacy by:

Ensuring that $P(H_0)$ is
as large as possible **49502**

Ensuring that $P(H_1)$ is
as large as possible **49556**

Ensuring that $P(E)$ is
as large as possible **10815**

None of the above **17495**

Example: Alarm system⁷

- In a city of 1 million inhabitants let there be 100 terrorists and 999,900 non-terrorists ...
- In an attempt to catch the terrorists, the city installs an alarm system with a surveillance camera and automatic facial recognition software:
 - If the camera scans a terrorist, a bell will ring 99% of the time, and it will fail to ring 1% of the time.
 - If the camera scans a non-terrorist, a bell will not ring 99% of the time, but it will ring 1% of the time.
- Suppose now that an inhabitant triggers the alarm. What is the chance that the person is a terrorist?

⁷http://en.wikipedia.org/wiki/Base_rate_fallacy

What do you think?

Respond at PolleEv.com/andreypovyak436

Text a **CODE** to 22333

Suppose now that an inhabitant triggers the alarm. What is the chance that the person is a terrorist?

The chances are small **341328**

The chances are 50/50 **341329**

The chances are large **341330**

Likelihood as a measure of goodness of fit

- D is observed data (e.g. the system passed 1000 pen tests with no failures)
- $H0$ and $H1$ are competing hypotheses (theories)
- $P(D | H0) = 0.9$ // likelihood of observations for $H0$
- $P(D | H1) = 0.2$ // likelihood of observations for $H1$
- Which hypothesis does fit the data better?
- Likelihood shows how (relatively) well the theory “predicts” (fit) the past.
- Likelihood ratio is a criterion for comparing theories (model selection)

What do you think?

Respond at PolleEv.com/andreypovyak436

Text a **CODE** to 22333

Which hypothesis (H0 or H1) does fit the data (D) better?

H0 with $P(D | H0) = 0.9$ **65334**

H1 with $P(D | H1) = 0.2$ **155493**

Testing H_0 vs H_1 : types of errors

| | H_0 is true | H_1 is true |
|--------------------|---------------|---------------|
| H_0 rejected | Type 1 error | Correct |
| H_0 not rejected | Correct | Type 2 error |

- $\alpha = P(\text{Type 1 error})$ - significance level;
- $1 - \beta = 1 - P(\text{Type 2 error})$ - power of the test

Run R-code online: <https://rdr.io/snippets/>

rdr.io Find an R package R language docs Run R in your browser R Notebooks packages, doctext, code...

Home / Snippets

Snippets

Run any R code you like. There are over *twelve thousand* R packages preloaded.

[Privacy information](#)
[Embed this on your website](#)
[List of installed packages](#)

```
# Number of failures
m <- 3
# Number of tests
n <- 10
# Number of simulated repeated experiments
N <- 1000
# Sequence of numbers of failures in the simulated repeated experiments
SRE <- rbinom(1000, 10, prob=m/n)
# Histogram
hist(SRE)
# binomial test
binom.test(m, n, p=0.3)
```

Run (Ctrl-Enter)

Any scripts or data that you put into this service are public.

```
Exact binomial test

data: m and n
number of successes = 3, number of trials = 10, p-value = 1
alternative hypothesis: true probability of success is not equal to 0.3
95 percent confidence interval:
 0.06673951 0.65245285
sample estimates:
probability of success
               0.3
```

Histogram of SRE

Improve this page

Case study: Pen tests with no failures (1/2)

- Example: Adversaries (Pen testers) committed 1000 attempted attacks and none has succeeded
- Does this mean that $P(\text{protection failure}) = 0$?
- We could have observed no failures in the finite sequence of tests with comparatively high probability even if $P(\text{protection failure}) > 0$
- For instance, if $P(\text{protection failure}) = q = 10^{-4}$, then the probability (likelihood) of observing no failures in a sequence of $n = 1000$ independent tests is $(1 - q)^n$, i.e.

```
> (1 - 1e-4)^1e3
```

```
[1] 0.9048329
```

- The smaller q , the higher the likelihood
- We might be interested in the range of hypothetical values of q with the likelihood of observing n successful tests better than $1 - \text{conf} = 0.05$, i.e.
- $(1 - q)^n > 1 - \text{conf}$ or $q < q_u(n, \text{conf}) = 1 - (1 - \text{conf})^{1/n}$

```
> 1 - (1-0.95)^(1/1000)
```

```
[1] 0.00299125
```

Case study: Pen tests with no failures (2/2)

- Thus, if $q > q_u(n, conf)$ the likelihood of observing $n = 1000$ failure free tests is less than 0.05 and we can use the interval $[0, q_u(n, conf)]$ as $conf \times 100\%$ confidence interval:

```
> binom.test(0,1000, p=0, alternative = "less")
```

Exact binomial test

```
data: 0 and 1000
```

```
number of successes = 0, number of trials = 1000, p-value = 1
```

```
alternative hypothesis: true probability of success is less than 0
```

```
95 percent confidence interval:
```

```
0.000000000 0.00299125
```

```
sample estimates:
```

```
probability of success
```

```
0
```

```
>
```

Testing Hypotheses (1/5)

- Simple hypothesis $H0 : q = q_0$
- Complex alternative: $H1 : q < q_0$
- Observation: n tests, K observed failures
- If k given n is large enough, then we must not reject $H0$ against $H1$
- If k given n is small enough, then we must reject $H0$ in favour of $H1$
- Thus we need to find k_{crit} such that $P(K \leq k_{crit} \mid n, q_0)$ was small,
- Normally, k_{crit} is chosen such that

$$P(K \leq k_{crit} \mid n, q_0) = pbinom(k, n, q_0) = \sum_{i=0}^k \binom{n}{i} q_0^i (1 - q_0)^{n-i} < \alpha = 0.05$$

- Markedly $\alpha = P(H0 \text{ is rejected} \mid H0 \text{ is correct}) = P(\text{1st type error})$ is called *significance level*
- $k_{crit}(\alpha, n, q_0) = qbinom(\alpha, n, prob = q_0)$

Testing Hypotheses (2/5)

- ```
> qbinom(0.05, 5000, prob=1e-3)
[1] 2
>
```
- i.e. if in the sequence of 5000 tests we observe 2 failures or less, then we reject  $H_0$  in favour of  $H_1$
- Observation:  $K < k_{crit}$  iff  $pbinom(K, n, q_0) < \alpha = pbinom(k_{crit}, n, q_0)$
- **Transformed rejection rule:**
  - $H_0$  is rejected in favour of  $H_1$  at the significance level  $\alpha$  if  $pbinom(K, n, q_0) < \alpha$
- The probability  $pbinom(K, n, q_0)$ , of observing  $K$  failures or less in a repeated experiment is called **p-value**
- Example:  $K = 1$ 

```
> pbinom(1, 5000, prob=1e-3)
[1] 0.04036031
>
```
- Thus, after observing  $K = 1$ ,  $H_0$  is rejected in favour of  $H_1$  at the significance level 0.05



## Testing Hypotheses (3/5)

- Easy way:

```
> binom.test(1,5000,alternative = "less",p=1e-3)
```

Exact binomial test

data: 1 and 5000

number of successes = 1, number of trials = 5000, p-value = 0.04036

alternative hypothesis: true probability of success is less than 0.001

95 percent confidence interval:

0.0000000000 0.0009484178

sample estimates:

probability of success

2e-04

>

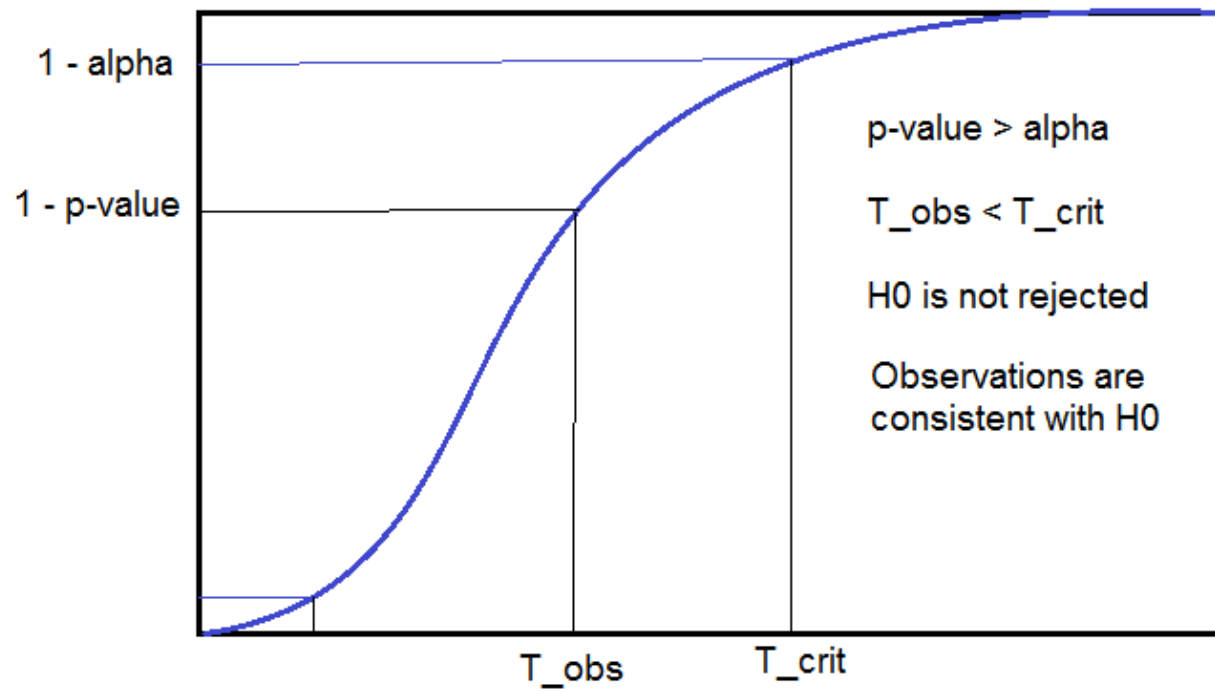
- Observation:  $p = 10^{-3} > 0.0009484178$  i.e. it is outside 95% conf.interval

## Testing Hypotheses (4/5)

General idea:

- Calculate the test statistics:  $T = T(data, H_0)$  with known distribution  $F_T(x)$
- $T$  is a random measure of coherence between  $H_0$  and observations (goodness of fit)
- Define the *critical area*:  $[T_{lower}(H_1, \alpha), T_{upper}(H_1, \alpha)]$ :
  - $F(T_{lower}(H_1, \alpha)) = \alpha/2$
  - $1 - F(T_{upper}(H_1, \alpha)) = \alpha/2$
- if  $T(data, H_0) < T_{lower}(H_1, \alpha)$  or  $T(data, H_0) > T_{upper}(H_1, \alpha)$ , then  $H_0$  is rejected in favour of  $H_1$  at the significance level  $\alpha$ .
- For two-sided alternatives *p-value* calculations are a bit more complicated

## Testing Hypotheses: a diagram (5/5)



Cumulative Distribution Function of the test statistics

## Example: Testing hypotheses about mean

- Simulated losses from security related incidents. Observations don't support the 'true' claim

```
> data<-rnorm(12, 900, 300); data
[1] 1214.1273 1268.2659 659.5749 724.2138 703.6727 874.5503 1444.9907
[8] 716.1635 961.2057 914.1676 999.0021 786.2387
> t.test(data,mu=1000, alternative="less")
```

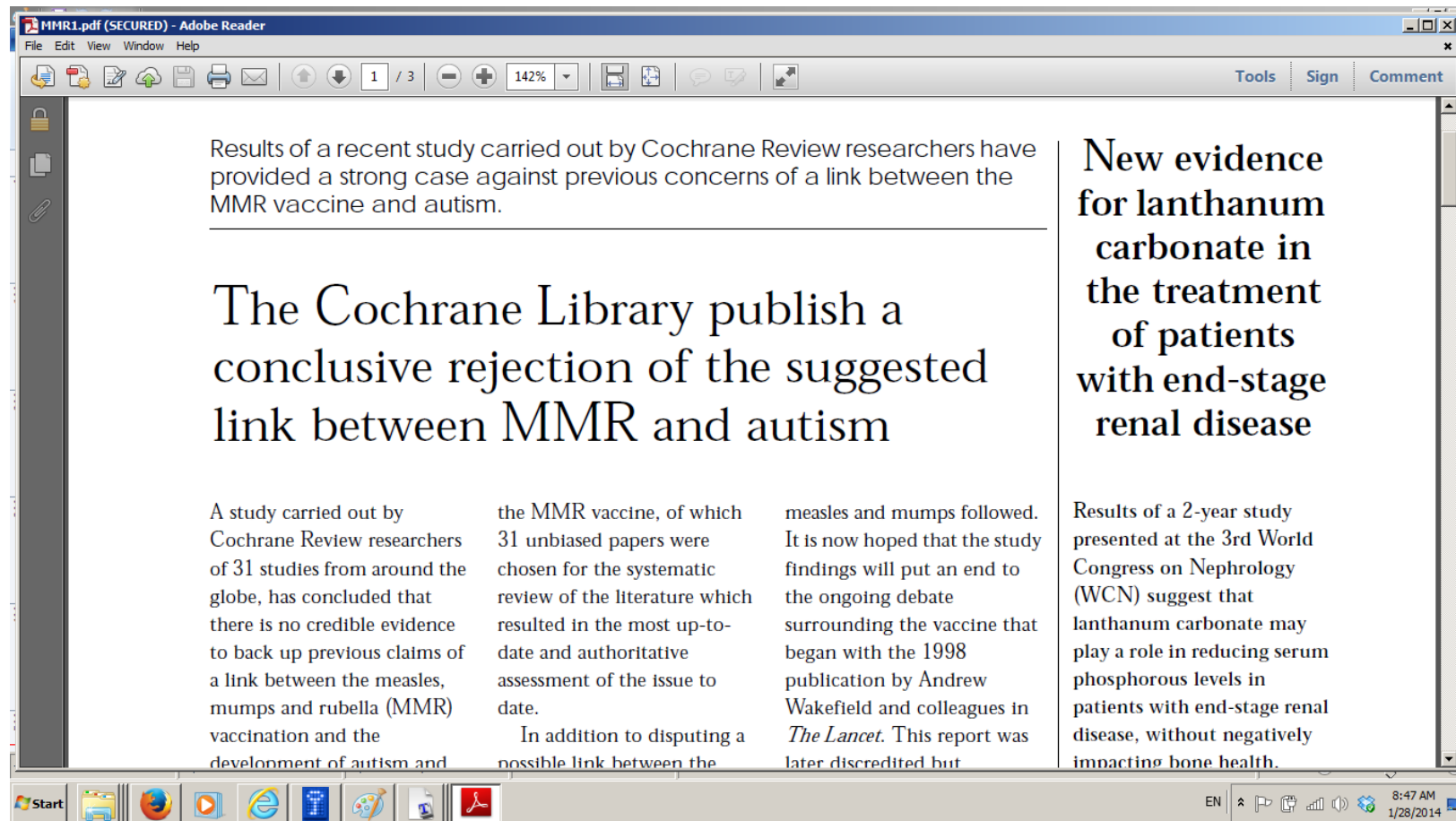
One Sample t-test

```
data: data
t = -0.8388, df = 11, p-value = 0.2097
alternative hypothesis: true mean is less than 1000
95 percent confidence interval:
-Inf 1069.775
sample estimates:
mean of x
938.8478
```

## Power of the test

- Type 2 error:  $H_0$  is accepted when it is not true
- Power of the test:  $\beta = 1 - P(\text{type 2 error})$
- In our example the test was not powerful enough to discriminate between the hypothetical value 1000 and the real value 900
- Experiments have limited “resolution”. Small effects are easy to miss.
- Power of the test for complex alternative (e.g.  $effect > 0$ ) is practically impossible to estimate.
- Therefore, a simple alternative for an “expected” effect is used (e.g.  $effect = \epsilon$ )

# MMR story<sup>8</sup>(1/7)



<sup>8</sup><http://www.futuremedicine.com/doi/pdf/10.2217/14750708.2.6.819>

# MMR story<sup>9</sup> (2/7)

## Vaccines for measles, mumps and rubella in children (Review)

Demicheli V, Jefferson T, Rivetti A, Price D



This is a reprint of a Cochrane review, prepared and maintained by The Cochrane Collaboration and published in *The Cochrane Library* 2008, Issue 4

<http://www.thecochranelibrary.com>

---

<sup>9</sup><http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004407.pub3/abstract>

# MMR story<sup>10</sup> (3/7)

## Authors' conclusions

The design and reporting of safety outcomes in MMR vaccine studies, both pre- and post-marketing, are largely inadequate. The evidence of adverse events following immunisation with MMR cannot be separated from its role in preventing the target diseases.

## PLAIN LANGUAGE SUMMARY

### Using the combined vaccine for protection of children against measles, mumps and rubella

Measles, mumps and rubella are three very dangerous infectious diseases which cause a heavy disease, disability and death burden in the developing world. Researchers from the Cochrane Vaccines Field reviewed 139 studies conducted to assess the effects of the live attenuated combined vaccine to prevent measles, mumps and rubella (MMR) in children. MMR protects children against infections of the upper airways but very rarely may cause a benign form of bleeding under the skin and milder forms of measles, mumps and rubella. No credible evidence of an involvement of MMR with either autism or Crohn's disease was found. No field studies of the vaccine's effectiveness were found but the impact of mass immunisation on the elimination of the diseases has been demonstrated worldwide.

---

<sup>10</sup><http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004407.pub3/abstract>



# MMR story<sup>11</sup> (4/7)

## Vaccines for measles, mumps and rubella in children (Review)

Demicheli V, Rivetti A, Debalini MG, Di Pietrantonj C



This is a reprint of a Cochrane review, prepared and maintained by The Cochrane Collaboration and published in *The Cochrane Library* 2012, Issue 2

<http://www.thecochranelibrary.com>

---

<sup>11</sup><http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004407.pub3/abstract>

# MMR story<sup>12</sup> (5/7)

## **Authors' conclusions**

The design and reporting of safety outcomes in MMR vaccine studies, both pre- and post-marketing, are largely inadequate. The evidence of adverse events following immunisation with the MMR vaccine cannot be separated from its role in preventing the target diseases.

---

<sup>12</sup><http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004407.pub3/abstract>

## MMR story<sup>13</sup> (6/7)

We could assess no significant association between MMR immunisation and the following conditions: autism, asthma, leukaemia, hay fever, type 1 diabetes, gait disturbance, Crohn's disease, demyelinating diseases, or bacterial or viral infections. The methodological quality of many of the included studies made it difficult to generalise their results.

---

**Vaccines for measles, mumps and rubella in children (Review)**

Copyright © 2012 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

2

---

<sup>13</sup><http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004407.pub3/abstract>

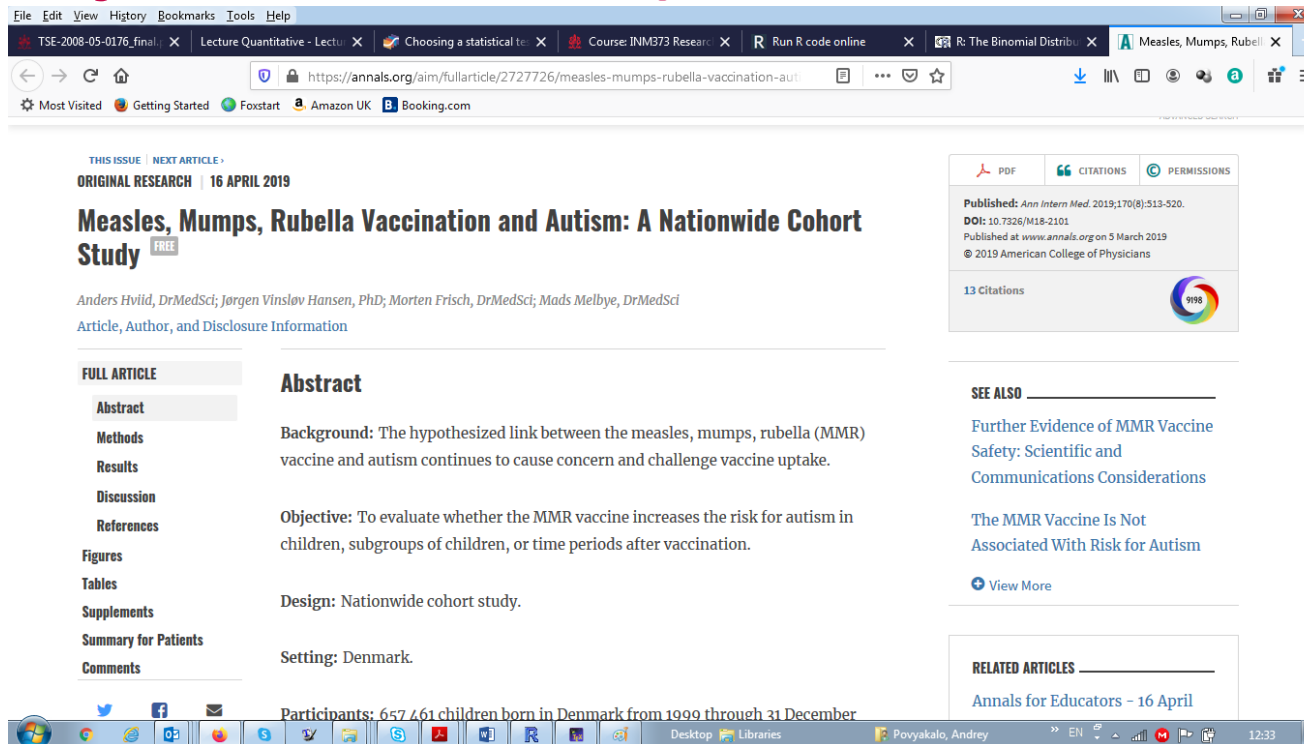
## MMR story<sup>14</sup> (7/7)

The range of differing study designs used by authors is partly a reflection on the lack of control children not exposed to MMR, due to the population nature of vaccination programmes. As MMR vaccine is universally recommended, recent studies are constrained by the lack of a non-exposed control group. This is a methodological difficulty which is likely to be encountered in all comparative studies of established childhood vaccines. We were unable to include a majority of the retrieved studies because a comparable, clearly-defined control group or risk period was not available. The exclusion may be a limitation of our review or may reflect a more fundamental methodological dilemma: how to carry out meaningful studies in the absence of a representative population not exposed to a vaccine that is universally used in public health programmes. Whichever view is chosen, we believe that meaningful inferences from individual studies lacking a non-exposed control group are difficult to make.

---

<sup>14</sup><http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD004407.pub3/abstract>

# MMR story: latest developments<sup>15</sup>



The screenshot shows a web browser window displaying the article 'Measles, Mumps, Rubella Vaccination and Autism: A Nationwide Cohort Study' on the Annals of Internal Medicine website. The article is dated 16 April 2019 and is marked as 'ORIGINAL RESEARCH'. The authors listed are Anders Hviid, DrMedSci; Jørgen Vinsløv Hansen, PhD; Morten Frisch, DrMedSci; and Mads Melbye, DrMedSci. The article is available for free. The abstract states: 'Background: The hypothesized link between the measles, mumps, rubella (MMR) vaccine and autism continues to cause concern and challenge vaccine uptake. Objective: To evaluate whether the MMR vaccine increases the risk for autism in children, subgroups of children, or time periods after vaccination. Design: Nationwide cohort study. Setting: Denmark. Participants: 657 461 children born in Denmark from 1999 through 31 December'. The right sidebar shows 13 citations and a 'SEE ALSO' section with links to 'Further Evidence of MMR Vaccine Safety: Scientific and Communications Considerations' and 'The MMR Vaccine Is Not Associated With Risk for Autism'. The bottom of the page shows a Windows taskbar with various application icons and the system clock at 12:33.

Hviid A, Hansen JV, Frisch M, et al. Measles, Mumps, Rubella Vaccination and Autism: A Nationwide Cohort Study. *Ann Intern Med.* 2019;170:513-520. [Epub ahead of print 5 March 2019]. doi: 10.7326/M18-2101

<sup>15</sup><https://annals.org/aim/fullarticle/2727726/measles-mumps-rubella-vaccination-autism-nationwide-cohort-study>

## MMR story: latest developments

- Hviid A, Hansen JV, Frisch M, et al. Measles, Mumps, Rubella Vaccination and Autism: A Nationwide Cohort Study. *Ann Intern Med.* 2019;170:513?520. [Epub ahead of print 5 March 2019]. doi: 10.7326/M18-2101
- "Results: During 5 025 754 person-years of follow-up, 6517 children were diagnosed with autism (incidence rate, 129.7 per 100 000 person-years). Comparing MMR-vaccinated with MMR-unvaccinated children yielded a fully adjusted autism hazard ratio of 0.93 (95% CI, 0.85 to 1.02). Similarly, no increased risk for autism after MMR vaccination was consistently observed in subgroups of children defined according to sibling history of autism, autism risk factors (based on a disease risk score) or other childhood vaccinations, or during specified time periods after vaccination."

# MMR story: latest developments (My observations)

- The base rate of autism in the population is around 1%
- The study design considers increase in number of autism cases less than 10% statistically insignificant (approx 651 cases).
- The control population is 20 times smaller than the vaccinated population.
- The control population contains 50% more high autism risk cases than the vaccinated population.
- The control population contains 22% less lower autism risk cases than the vaccinated population
- In high autism risk subpopulation 22% increase in number of autism cases is considered insignificant.
- The observed hazard ratio for this subgroup is outside the confidence area for the whole population.
- The good authors didn't bother to discuss the power of their study with numbers, or claim that that power is sufficient.

## Case study: Coordinated network activities (Conspiracies)

- During the period between 15/11/06 and 04/12/06 two hosts from China: 60.191.231.93 and 218.75.24.230 were most active
- Labeling the host 60.191.231.93 with “1” and host 218.75.24.230 with “2”,
- The sequence of attacks from the both hosts can be represented with the following sequence of 70 labels: 1 2 2 2 2 2 1 1 2 2 2 1 2 2 2 1 2 1 2 2 1 2 1 2 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 1 2 2 2 1 2 2 1 2 2 1 2 2 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1
- Automated analysis found different sub-sequences of labels (patterns) repeated at least twice.
- Among them, the sub-sequence: 2 2 1 2 2 2 was repeated 5 times: 1 2 2 2 2 2 1 1 2 **2 2 1 2 2** 2 1 2 1 2 2 1 2 1 **2 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 1 2 2 2 1 2 1 2 2 1 2 2 2 1 1 1 1 1 1 1** 2 1 1 1 1 1 1 1 1
- Accidental?
- Exact test difficult to apply
- Who was the leading host?



## Approximate test

- Statistical simulation showed that the estimated probability that in the random permutation of the original “long” sequence of seventy “1”-s and “2”-s the pattern 2 2 1 2 2 2 occurs at least 5 times, is smaller than 0.05:

```
> hndata.ar.patterns.prob(dt,pt,nocr=5)
```

```
Data array: 1 2 2 2 2 2 1 1 2 2 2 1 2 2 2 1 2 1 2 2 1 2 1 2 2
1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 1 2 2 2 1 2 1 2 2 1 2 2 2 1 1 1
1 1 1 1 2 1 1 1 1 1 1 1 1
```

```
Pattern array: 2 2 1 2 2 2
```

```
Minimum number of occurrences: 5
```

```
Estimated probability: 0.0071
```

```
95 % Confidence Interval: (0.005549202 0.008947365)
```

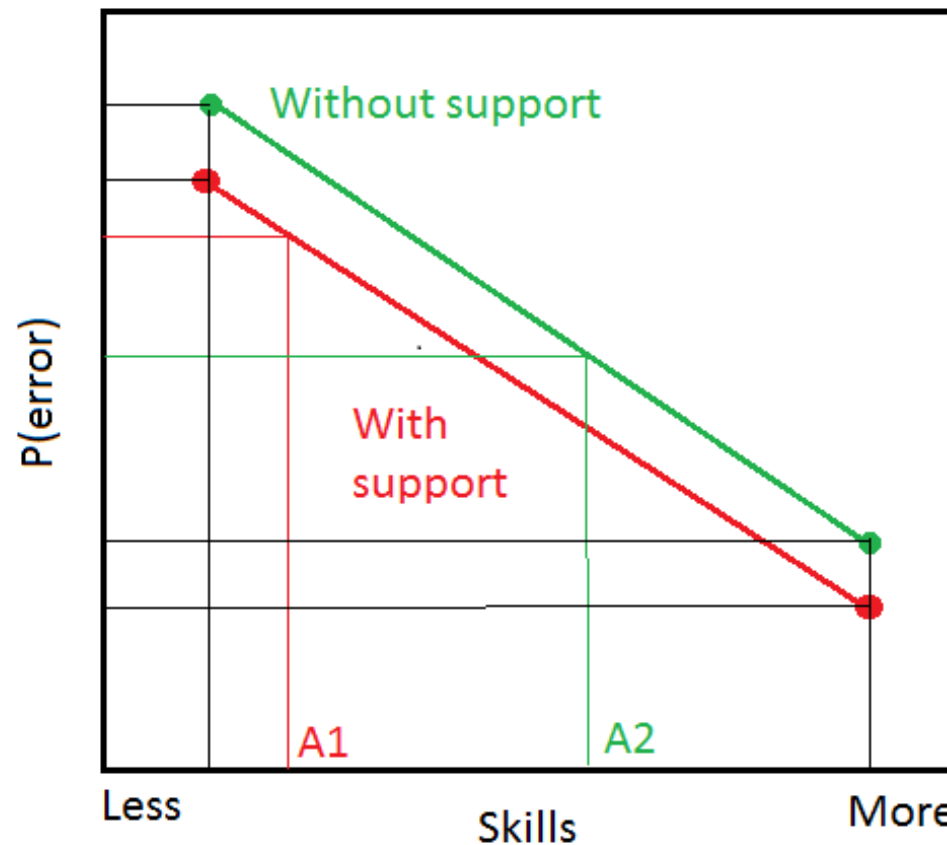
```
>
```

- This is unlikely to happen merely by chance.
- Presence of such a repeated sub-sequence can be considered as an indication of co-ordination between activities of two hosts (“conspiracy”).



# Statistical 'Paradoxes'

# Simpson's paradox<sup>16</sup>



<sup>16</sup>Simpson, Edward H. (1951). "The Interpretation of Interaction in Contingency Tables". Journal of the Royal Statistical Society, Series B 13: 238–241.

## Example from the book<sup>17</sup>

important to make sure that the groups really are comparable.

It's easy to make up an example showing that averaging across very different values or groups can give absurd results. Here's how that might work. Suppose there are two pilots, Moe and Jill. Moe argues that he's the better pilot of the two, since he managed to land 83% of his last 120 flights on time compared with Jill's 78%. But let's look at the data a little more closely. Here are the results for each of their last 120 flights, broken down by the time of day they flew:

|       |      | Time of Day          |                      |                       |
|-------|------|----------------------|----------------------|-----------------------|
|       |      | Day                  | Night                | Overall               |
| Pilot | Moe  | 90 out of 100<br>90% | 10 out of 20<br>50%  | 100 out of 120<br>83% |
|       | Jill | 19 out of 20<br>95%  | 75 out of 100<br>75% | 94 out of 120<br>78%  |

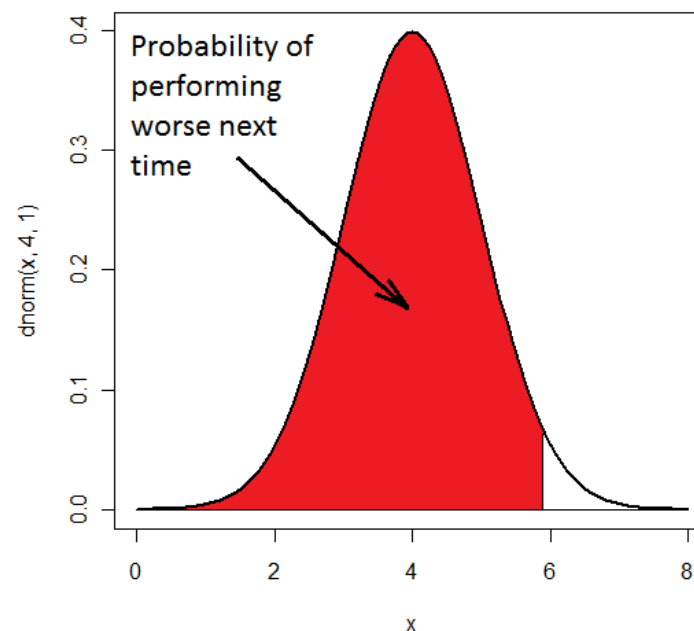
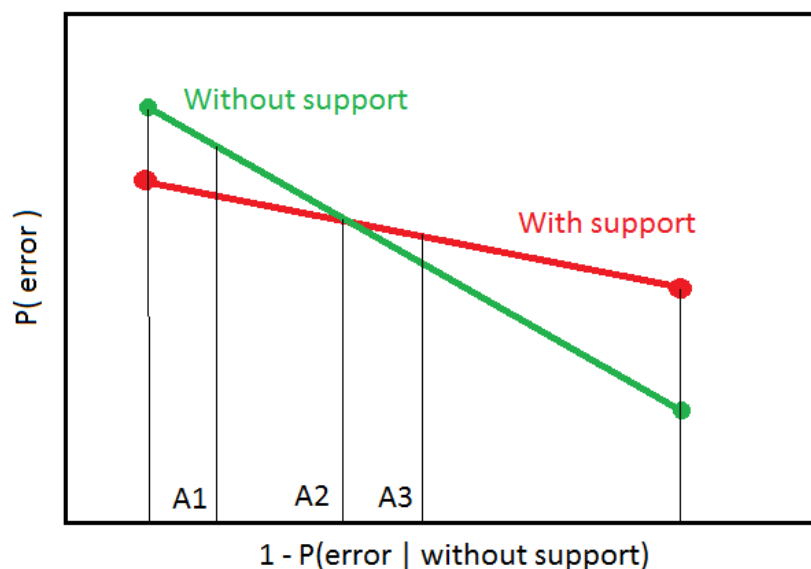
On-time flights by *Time of Day* and *Pilot*. Look at the percentages within each *Time of Day* category. Who has a better on-time record during the day? At night? Who is better overall? **Table 3.10**

s paradox arose  
on rates for men  
ifornia at  
ported in an ar-  
a applicants

Look at the day and nighttime flights separately. For day flights, Jill had a 95% on-time rate, and Moe only a 90% rate. At night, Jill was on time 75% of the time, and Moe only 50%. So Moe is better "overall," but Jill is better both during the day and at night. How can this be?

<sup>17</sup><http://www.laurel.kyschools.us/userfiles/545/Classes/60146/Ch03.pdf>

# Regression towards the mean<sup>18</sup>

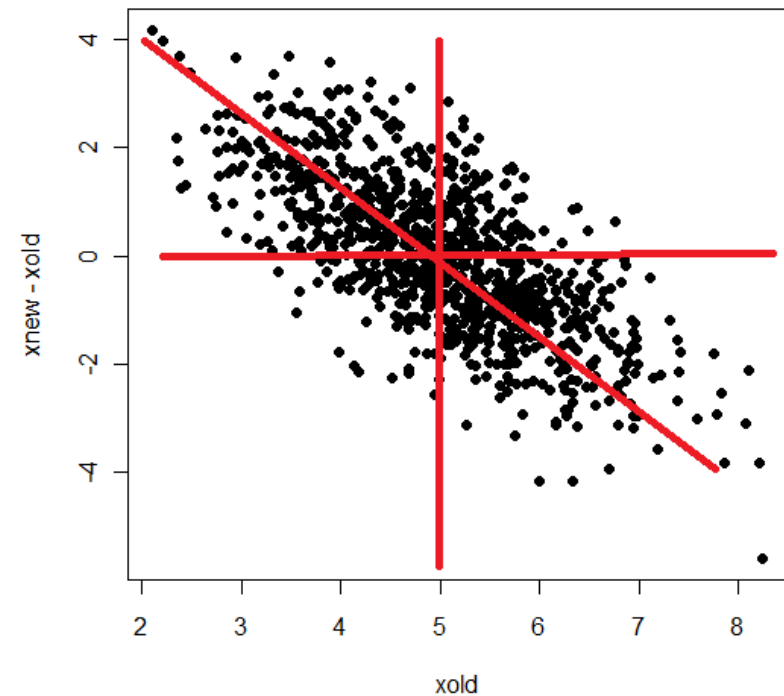


- Wrong choice of the 'independent' variable. Correlated 'noise'

<sup>18</sup>Francis Galton (1886). "Regression towards mediocrity in hereditary stature". The Journal of the Anthropological Institute of Great Britain and Ireland (The Journal of the Anthropological Institute of Great Britain and Ireland, Vol. 15) 15: 246–263.

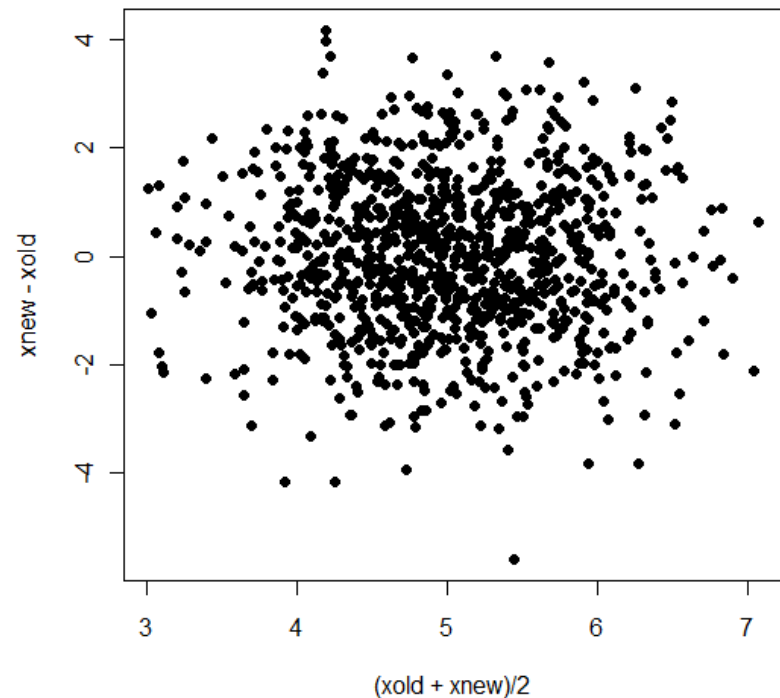
# Regression towards the mean (Simulation)

```
> xold <- rnorm(1000,0,1)+ 5
> xnew <- rnorm(1000,0,1)+5
> plot(xold, xnew - xold,pch=19)
>
```



## Regression towards the mean (Solution?)

```
> xold <- rnorm(1000,0,1)+ 5
> xnew <- rnorm(1000,0,1)+5
> plot(xold, xnew - xold,pch=19)
> plot((xold+xnew)/2,xnew-xold,pch=19)
>
```



- Does the value  $(xold + xnew)/2$  measures what we want to be measured?

# Statistical Independence and Covariance

- $X, Y$  are random variables
- $X$  and  $Y$  are called statistically independent if

$$F_{XY}(x, y) =$$

$$P(X \leq x \text{ and } Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y)$$

- If  $X$  and  $Y$  are statistically independent then

$$E(XY) = E(X)E(Y)$$

- Covariance:

$$\text{cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

- Interesting property: If  $Y = AX + B$ , then

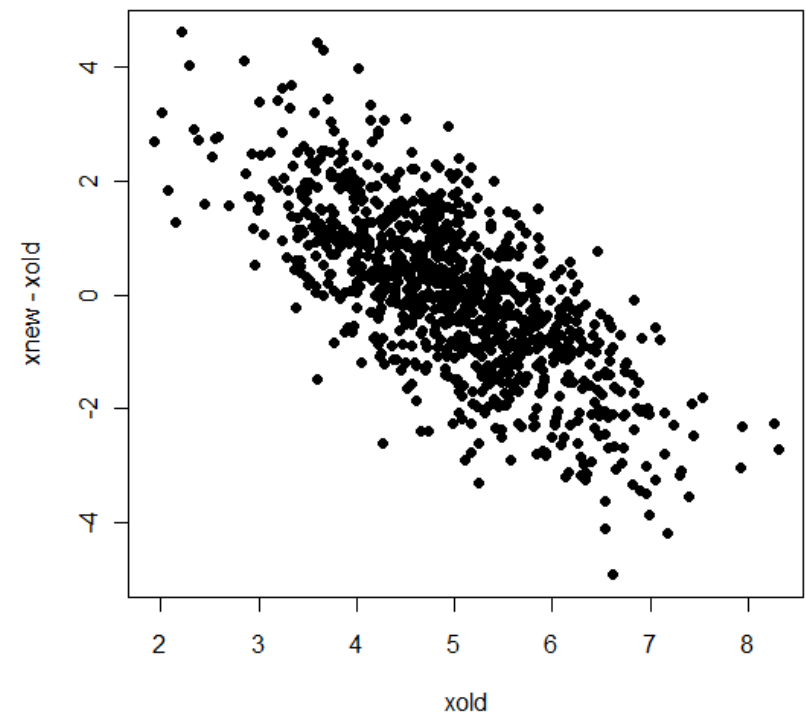
$$\text{cov}(X, Y) = E(XY) - E(X)E(Y) =$$

$$E(AX^2 + BX) - AE(X)^2 - BE(X) = A(E(X^2) - E(X)^2) = A \cdot \text{Var}(X)$$



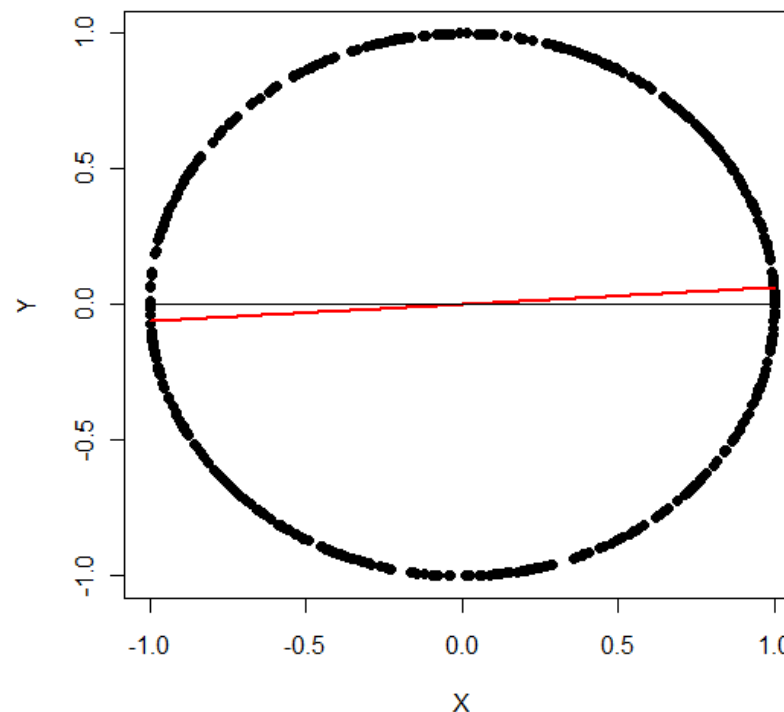
# Regression towards the mean: negative correlation

```
> xold <- rnorm(1000,0,1)+ 5
> xnew <- rnorm(1000,0,1)+5
> plot(xold, xnew - xold,pch=19)
> cov(xnew - xold, xold)
[1] -1.021144
```



## Example: $\text{cov}(X, Y) = 0$ does not mean statistical independence of $X$ and $Y$

```
> phi <- runif(500, 0, 2*pi)
> X <- cos(phi)
> Y <- sin(phi)
> plot(X, Y, pch=19)
> cov(X, Y)
[1] 0.03073386
>
> curve(cov(X, Y) * x / var(X),
+ from=-1, to=1, col="red",
+ lwd=2, add=TRUE)
> line(c(-1, 1), c(0, 0))
```



# Conclusions

- Statistics are good for making inferences
    - about “things in general”
    - and “in the long run”
  - but NOT so good for inferences
    - about particular things
    - on a particular occasion
  - Statistics deal with “probabilities”
    - conclusions should be tentative rather than absolute
    - expectations based on numerical generalisations are bound to be confounded much of the time
- ( “There are lies, damn lies? and statistics!” )

## Revision exercises

- Contrast and compare the Prosecutor's fallacy and the Base Rate Fallacy
- Calculate FP rate for the alarm system from slide 23
- What is a weak spot of my "conspiracy" test?
- Use R to reproduce the calculations from the lecture. Play around with the parameters. Observe the effects.

## Reading suggestions

- “Statistics without tears: a primer for non-mathematicians” - Rowntree, Derek, c2004
  - Particularly relevant chapters are: 4 (on the Normal distribution), 6 & 7 (on comparing samples and significance testing) and 8 (on correlation).
- “A gentle guide to research methods” - Rugg, Gordon, Petre, Marian, 2007
  - E.g. ”Inferential statistics: What are the Odds Against Your Findings Being Due to Random Chance?”