"Your Honor, this was not a coincidence!"

On the (ab)use of statistics in the case against Lucia de Berk

Peter Grünwald
CWI / Leiden

The Case of Lucia de Berk

- In 2004, the court of appeals convicted Lucia de Berk, a nurse from Juliana Children's Hospital in the Hague, to life imprisonment for 7 murders and 3 murder attempts.
- In 2006, the supreme court upheld the conviction. The same day De Berk suffered a severe stroke.
- In 2008, the supreme court decided to reopen the case. Lucia was set free, after 6.5 years in prison
- In 2009, Lucia was cleared of all charges (except for never returning two library books) in what now counts as the biggest miscarriage of justice in Dutch history

---

2001 (september 16th) Juliana hospital notifies the police

2003 Court convicts L. of 4 murders, 3 attempts

2004 Court of appeals convicts L. of 7 murders,3 attempts

2006 Prof. **Ton Derksen** writes a book about the case

2001 (september 16th) Juliana hospital notifies the police

- –There were many more "incidents" during her shifts than during those of other nurses in her ward
- –Statistician calculates that the probability that something like this would happen by pure coincidence, is less than **1 in 342 million**
- –Trial gets (more than) substantial media attention

2003 Court convicts L. of 4 murders, 3 attempts

2004 Court of appeals convicts L. of 7 murders,3 attempts

2006 Prof. **Ton Derksen** writes a book about the case

2001 (september 16th) Juliana hospital notifies the police

- –There were many more "incidents" during her shifts than during those of other nurses in her ward
- –Statistician calculates that the probability that something like this would happen by pure coincidence, is less than **1 in 342 million**
- –Trial gets (more than) substantial media attention

2003 Court convicts L. of 4 murders, 3 attempts

- –Statistics is crucial part of the evidence
- –In appeal, experts for defense R. Meester (probabilist) and M. van Lambalgen (logician) claim calculation is flawed

2004 Court of appeals convicts L. of 7 murders,3 attempts

2006 Prof. **Ton Derksen** writes a book about the case

2001 (september 16th) Juliana hospital notifies the police

- –There were many more "incidents" during her shifts than during those of other nurses in her ward
- –Statistician calculates that the probability that something like this would happen by pure coincidence, is less than **1 in 342 million**
- –Trial gets (more than) substantial media attention

2003 Court convicts L. of 4 murders, 3 attempts

- –Statistics is crucial part of the evidence
- –In appeal, experts for defense R. Meester (probabilist) and M. van Lambalgen (logician) claim calculation is flawed

2004 Court of appeals convicts L. of 7 murders,3 attempts

- –Court says that this time, "statistics in the form of probability calculations has not been used"
- –In fact, **the court's report has flawed statistics written all over it**

2006 Prof. **Ton Derksen** writes a book about the case

2001 (september 16th) Juliana hospital notifies the police

- –There were many more “incidents” during her shifts than during those of other nurses in her ward
- –Statistician calculates that the probability that something like this would happen by pure coincidence, is less than **1 in 342 million**
- –Trial gets (more than) substantial media attention

2003 Court convicts L. of 4 murders, 3 attempts

- –Statistics is crucial part of the evidence
- –In appeal, experts for defense R. Meester (probabilist) and M. van Lambalgen (logician) claim calculation is flawed

2004 Court of appeals convicts L. of 7 murders,3 attempts

- –Court says that this time, “statistics in the form of probability calculations has not been used” – evidence is mostly *medical*
- –In fact, **the court’s report has flawed statistics written all over it**

2006 Prof. **Ton Derksen** writes a book about the case

- **– all deaths were most probably just natural!**
- **–Tide begins to turn…**

2006-2008

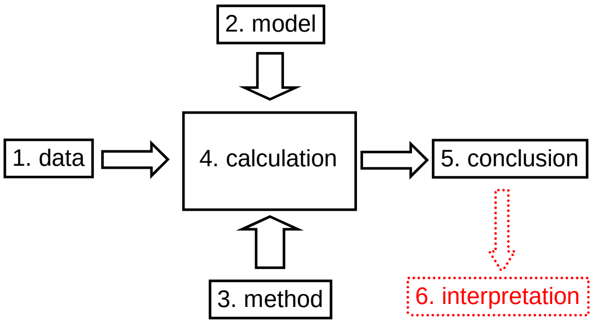
- Group of concerned citizens starts agressive media campaign to reopen the case. Main Heroes:
  -  **Metta de Noo**, medical doctor and sister-in-law of the chef de clinique of the hospital. The first person with (a) serious doubts and (b) the **guts** to say so
  -  **Ton Derksen**, Metta’s brother, retired professor of philosophy, University of Nijmegen. Writes very sharp book about the case
  -  **Richard Gill**, professor of statistics at Leiden University. Organizes a petition to reopen the case. Signed by, e.g., Gerard ‘t Hooft (nobel prize winner of physics) and almost all professors of probability theory and statistics in NL
  - Many smaller players (including myself)

A Case about Small Probabilities

Menu

1. The use of statistics
   - What the statistician did (evidence in first verdict)
   - What went wrong (everything)
   - The role of statistics in the final verdict
2. How should we deal with statistics in (court) cases like this anyway?

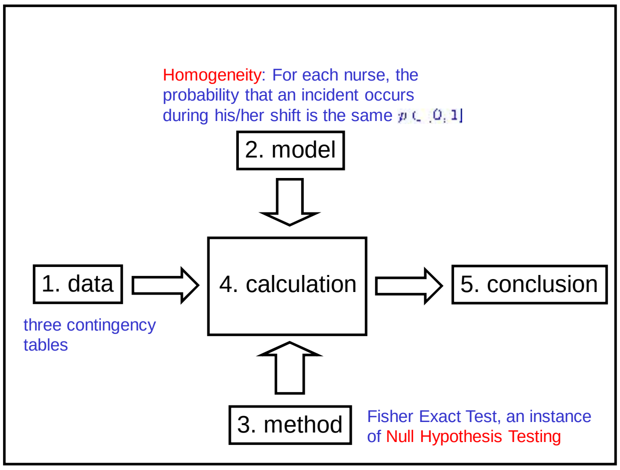
The Statistician’s Work



1. The Data

Juliana Hospital MCU-1, Oct 1 2000 – Sep 9 2001	no incident	incident	total
Nr of Shifts with L present	134	8	142
Nr of Shifts with L not present	887	**0**	887
Total Number of Shifts	1021	8	1029

Data from the Juliana Hospital Medium Care Unit-1, where suspicion first arose. “Incident” is sudden death or reanimation with no clear explanation



Homogeneity: For each nurse, the probability that an incident occurs during his/her shift is the same $p \in [0, 1]$

2. model

1. data

three contingency tables

4. calculation

5. conclusion

3. method

Fisher Exact Test, an instance of Null Hypothesis Testing

2. & 3. The Model and Method

- Statistician tested null hypothesis

  H0: “Lucia has same incident probability as other nurses”

  against the alternative

  H1: “Lucia has higher incident probability”

  using a standard test with a significance level of 1 in 10000.

2. & 3. The Model and Method

- Statistician tested null hypothesis

  H0: “Lucia has same incident probability as other nurses”

  against the alternative

  H1: “Lucia has higher incident probability”

  using a standard test with a significance level of 1 in 10000.
- i.e. he chooses some test statistic (a function of the data) T such that, as t increases, $\Pr_{H_0}(T \geq t)$ goes to 0.
- If the actually observed data t_{obs} is so extreme that

$$\text{p-value} := \Pr_{H_0}(T \geq t_{\text{obs}}) \leq \frac{1}{10000}$$

  then one “rejects” the null hypothesis.

The Method: Fisher’s Exact Test

- Statistician used Fisher’s Exact test
  - a “conditional” test with test statistic

    $T = \#\text{incidents with Lucia present}$
  - For convenience, define

    $\Pr^*(T \geq t) := \Pr_{H_0}(T \geq t$ #shifts with Lucia,
    #shifts total,
    #incidents)

The Method: Fisher’s Exact Test

- Statistician used Fisher’s Exact test
  - a “conditional” test with test statistic

    $T = \#\text{incidents with Lucia present}$
  - For convenience, define

    $\Pr^*(T \geq t) := \Pr_{H_0}(T \geq t$ #shifts with Lucia, ← 142
    #shifts total, ← 1029
    #incidents) ← 8

    ↑
    8

Under H0, Lucia has same probability q to witness an incident during her shift as any other nurse. Then q cancels out of equations and we can calculate this!

The Method: Fisher’s Exact Test

- Statistician used Fisher’s Exact test
  - a “conditional” test with test statistic

    $T = \#\text{incidents with Lucia present}$
  - For convenience, define

    $\Pr^*(T \geq t) := \Pr_{H_0}(T \geq t$ #shifts with Lucia, ← 142
    #shifts total, ← 1029
    #incidents) ← 8

    ↑
    8
  - We reject “Lucia is like the others” if $\Pr^*(T \geq 8) < 1/10000$
  - We find

    $\Pr^*(T \geq 8) = \Pr^*(T = 8) = .0000011057 \approx$ 1 in 9 million

Fisher’s Exact Test: Interpretation

- View nurse’s shifts as balls in an urn
  - $\Pr^*$ follows a hypergeometric distribution
- There are 1029 balls (shifts). 8 are black (incidents), the rest white
- We draw 142 balls without replacement from the urn (Lucia’s shifts)
- It turns out that all 8 black balls are among these 142
- $\Pr^*(T \geq 8) \approx$ 1 in 9 million is the probability that this happens.

$$\Pr^*(T = t) = \frac{\binom{\#\text{shifts L.}}{t}\binom{\#\text{shifts others}}{\text{total}\#\text{incidents}-t}}{\binom{\#\text{shifts total}}{\text{total}\#\text{incidents}}}$$

4. The Calculation

- Applying Fisher’s test to the Juliana data (first table) gives a *p*-value of 1 in 9 million

$$\text{p-val} := \Pr^*(T \geq 8) \approx \frac{1}{9 \cdot 10^6}$$

- Classical Problem: Same data that was used to suggest hypothesis was also used to test it

4. The Calculation

- Applying Fisher’s test to the Juliana data (first table) gives a *p*-value of 1 in 9 million

$$p := \Pr^*(T \geq 8) \approx \frac{1}{9 \cdot 10^6}$$

- Classical Problem: Same data that was used to suggest hypothesis was also used to test it
- Statistician recognizes this and performs a post-hoc correction, by considering the H0-probability that some nurse in Lucia’s ward experienced a pattern as extreme as Lucia’s

$$p' := \Pr^*(\text{there exists } j \in \{1, \ldots, 27\} \text{ s.t. } T_j \geq t_j) \approx 27 \Pr(T \geq 8) \approx \frac{3}{10^6}$$

4. The Calculation

- Another complication:There are three tables, giving rise to three p-values. How to combine these?
  - Using the fact that the data are independent, statistician combines them into one p-value by multiplying:

$$p_{new} :=$$

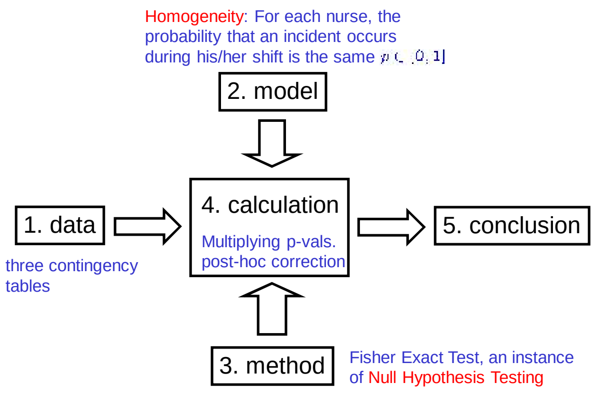
$$\Pr^*\left(\text{exists } j \in \{1, \ldots, 27\} \text{ s.t. } T_j^{(1)} \geq t_j^{(1)} \;\&\; T^{(2)} \geq t^{(2)} \;\&\; T^{(3)} \geq t^{(3)}\right)$$

$$= p'_1 \cdot p_2 \cdot p_3 = \text{1 in 342 million}$$

$$p'_1 := 27\Pr^*(T^{(1)} \geq 8)$$

$$p_2 := \Pr^*(T^{(2)} \geq 5)$$

$$p_3 := \Pr^*(T^{(3)} \geq 1)$$



Homogeneity: For each nurse, the probability that an incident occurs during his/her shift is the same $p \in [0, 1]$

2. model

1. data

three contingency tables

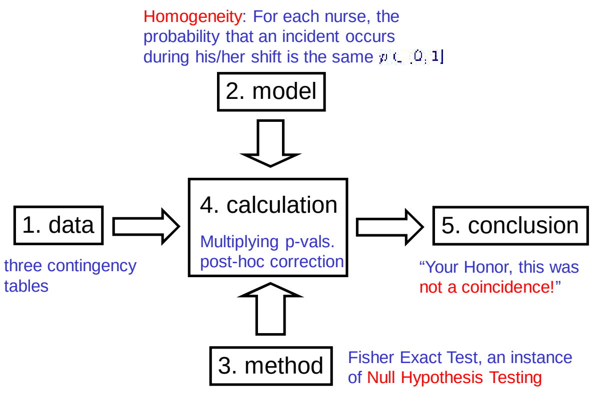
4. calculation

Multiplying p-vals.
post-hoc correction

5. conclusion

3. method

Fisher Exact Test, an instance of Null Hypothesis Testing



Homogeneity: For each nurse, the probability that an incident occurs during his/her shift is the same $p \in [0, 1]$

2. model

1. data

three contingency tables

4. calculation

Multiplying p-vals.
post-hoc correction

5. conclusion

“Your Honor, this was not a coincidence!”

3. method

Fisher Exact Test, an instance of Null Hypothesis Testing

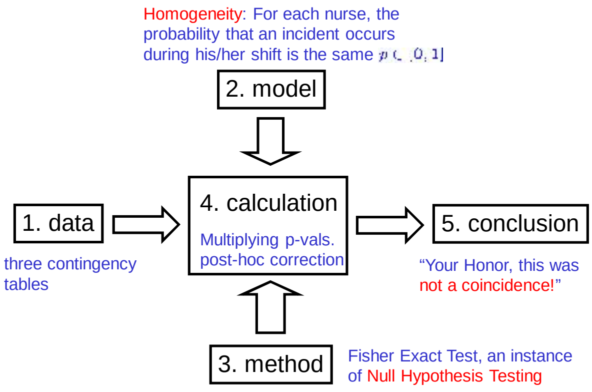
5. The Conclusion

- Statistician chose a significance level of 1 in 10000
- He observed a *p*-value of 1 in 342 million
- Therefore he rejects the null hypothesis
  "Lucia has same incidence probability as the others"
- Statistician explicitly mentions the *p*-value 1 in 342 million, and translates "rejection of null" into
  "your honor, this was not a coincidence!"
  (and the rest is up to you, your honor…)

The Conclusion - II

- Statistician does add a very explicit warning that this does *not* imply that Lucia is a murderer!
- He explicitly lists five alternative explanations:
  1. Lucia prefers to work together with another nurse. That nurse is really causing the incidents
  2. Lucia often does the night shift, during which more incidents happen
  3. Lucia is, quite simply, a bad nurse
  4. Lucia prefers to take on the most ill patients
  5. Somebody hates Lucia and tries to discredit her

What Went Wrong
(Everything!)



Homogeneity: For each nurse, the probability that an incident occurs during his/her shift is the same $p \in [0, 1]$

three contingency tables

Multiplying p-vals.
post-hoc correction

"Your Honor, this was not a coincidence!"

Fisher Exact Test, an instance of Null Hypothesis Testing

1. The Data

Juliana Hospital MCU-1, Oct 1 2000 – Sep 9 2001	no incident	incident	total
Nr of Shifts with L present	134	8	142
Nr of Shifts with L not present	887	0	887
Total Number of Shifts	1021	8	1029

Derksen: people die in hospitals. Why is there always a 0 in the middle? Maybe there was **bias** when gathering the data?

1. Relevant Additional Data

- The statistician and the defense and the court ignored following data that were available from the start:
  - From 1996-1998 (before Lucia worked there), there were seven deaths in her ward.
  - From 1999-2001 (when Lucia worked there), there were six deaths

2. Suspect-Driven Search

- In RKZ (other hospital from which tables were taken), explicit evidence that search was suspect-driven
  - More thorough search for incidents when she was present than for incidents when she wasn't

  "We were asked to make a list of incidents that happened during or shortly after Lucia's shifts"
- In JKZ, an attempt was made to be "objective", but
  - There was no record of reanimations. Doctors and nurses were asked whether they remembered such "incidents". Everybody knew why they were being asked…

3. Definition of "incident"

- "incident" was first defined as:

  a patient suddenly dies or needs reanimation
- Later the court changes this to

  a patient suddenly dies with no clear explanation, or reanimation is suspicious, i.e. without clear explanation
- This means that some sudden deaths and reanimations were not listed in the tables, because they were in no way suspicious
- All the people who have to report 'suspicious incidents' know that they are asked because Lucia may be a serial killer

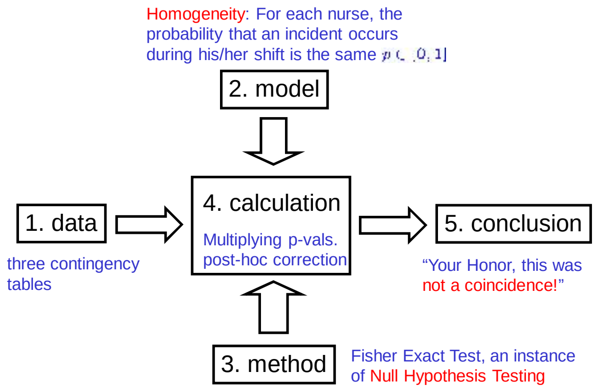
3. Definition of "incident"

- "incident" was first defined as:

  a patient suddenly dies or needs reanimation
- Later the court changes this to

  a patient suddenly dies with no clear explanation, or reanimation is suspicious, i.e. without clear explanation
- This means that some sudden deaths and reanimations were not listed in the tables, because they were in no way suspicious
- All the people who have to report 'suspicious incidents' know that they are asked because Lucia may be a serial killer

There is a considerable risk that "incident is suspicious" effectively becomes synonymous to "Lucia is present" (Van Zwet, grand old man of Dutch statistics)



Homogeneity: For each nurse, the probability that an incident occurs during his/her shift is the same $p \in [0, 1]$

2. model

1. data

three contingency tables

4. calculation

Multiplying p-vals. post-hoc correction

5. conclusion

"Your Honor, this was not a coincidence!"

3. method

Fisher Exact Test, an instance of Null Hypothesis Testing

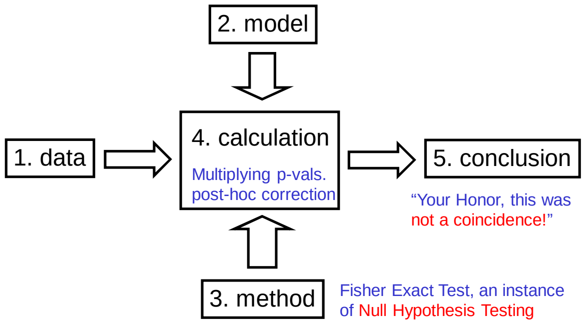
The Calculation

- Statistician combines three independent tests by *multiplying* the three p-values
- This is a beginner's mistake!
  - In this way, if you worked in 20 hospitals with a similar harmless incident pattern in each of them (say, a p-value of .5) you still end up with final p-value $(0.5)^{20} < 1/10^6$
  - if you change hospital a lot, you automatically become a suspect
  - Something close to .5 would be more reasonable!
- Various alternative, correct methods exist. A standard method such as Fisher's gives a p-value that is a factor 300 larger

The Method (and main issue)

- Even when combining *p*-values in a correct manner, final *p*-value (slightly) is smaller than 1 in 10000…so may we conclude "no coincidence" after all?
- **NO:**

  Neyman-Pearson style Null Hypothesis Testing cannot be used if the same data is used both for suggesting and testing a hypothesis. The results are essentially meaningless and there is *no* way a post-hoc correction can correct for this!
  - This will be explained in detail in final part of talk



Homogeneity: For each nurse, the probability that an incident occurs during his/her shift is the same $p \in [0,1]$

Multiplying p-vals.
post-hoc correction

"Your Honor, this was not a coincidence!"

Fisher Exact Test, an instance of Null Hypothesis Testing

The Model

- The assumption that there is no variation between ordinary nurses is wrong (but defensible –statistician implicitly warned for this).
- Following Lucy and Aitken (2002), A. de Vos and R. Gill propose the following model:
  - The nr of incidents witnessed by a nurse, is Poisson distributed with some parameter μ
  - For each nurse, μ is drawn independently from some distribution
  - Allows for **innocent heterogeneity** (e.g. clusters of shifts in time, caused by different vacation patterns, and so on)

The Model

- The assumption that there is no variation between ordinary nurses is wrong (but defensible).
- Following Lucy and Aitken (2002), A. de Vos and R. Gill propose the following model:
  - The nr of incidents witnessed by a nurse, is Poisson distributed with some parameter μ
  - For each nurse, μ is drawn independently from some distribution
  - Allows for **innocent heterogeneity** (e.g. clusters of shifts in time, caused by different vacation patterns, and so on)
- With this model, the *p*-value **increases dramatically**
  - combined with a correct combination method for the three tests and presumably correct data, it becomes as large as 1 in 9

---

5. The Conclusion

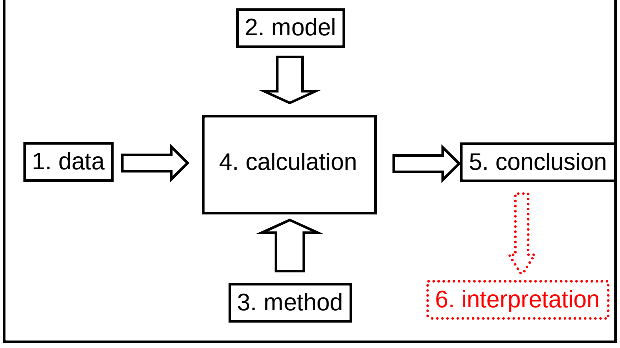
Here we are counterfactually assuming that calculations were correct in the first place

---

5. The Conclusion

- Statistician should have warned that the conclusion is extremely sensitive to the data being 100% correct
- The conclusion "this is not a coincidence" simply cannot be drawn – you cannot use a null hypothesis test here

---



The Interpretation - I

- First verdict:

  Court considers 4 murders and 3 attempts to be “lawfully and convincingly” proven, based on mix of medical, circumstantial and statistical evidence
- Statistics plays explicit role. However, their wording betrays that judges committ prosecutors fallacy:

Statistician writes: the probability that Lucia would witness so many incidents just by coincidence, is exceedingly small

Judge writes: the probability that it was all just a coincidence, is exceedingly small

The Prosecutor’s Fallacy

- What is the probability that a man is taller than 1m90, given that he is a professional basket ball player?
  - pretty large

The Prosecutor’s Fallacy

- What is the probability that a man is taller than 1m90, given that he is a professional basket ball player?
  - pretty large
- What is the probability that a man is a professional basket ball player, given that he is taller than 1m90?
  - pretty small

The Prosecutor’s Fallacy

- What is the probability that a man is taller than 1m90, given that he is a professional basket ball player?
  - pretty large
- What is the probability that a man is a professional basket ball player, given that he is taller than 1m90?
  - pretty small
- As a mathematical formula:

$$\Pr(A \mid B) \neq \Pr(B \mid A)$$

The probability of A given B is not the same as the probability of B given A

The Prosecutor’s Fallacy

- The probability that

  ‘Lucia has the same incidence probability as other nurses, given that she was present at so many incidents’ (perhaps not so small)

  is very different from the probability that

  ‘Lucia is present at so many incidents, given that she she has the same incidence probability as other nurses’ (very small)

$$\Pr(A \mid B) \neq \Pr(B \mid A)$$

Verdict Court of Appeal The Hague, 2004

On Page 1 (!) we read:

“....Er is geen statistisch bewijs in de vorm van toevalsberekeningen gebruikt. ...”

Verdict Court of Appeal The Hague, 2004

On page 1 we read:

> "....Er is geen statistisch bewijs in de vorm van toevalsberekeningen gebruikt. Wel is voor het bewijs van belang geacht dat de in het Juliana Kinderziekenhuis gepleegde delicten een betrekkelijk korte periode bestrijken en de meeste delicten voornamelijk op een gewone verpleegafdeling hebben plaatsgevonden en hiervoor geen verklaring is gevonden. ..."

The Interpretation - III

- One of the "murders" for which Lucia has been convicted concerned the death, in 1997, of a 73-year old woman suffering from terminal cancer
- In 2004, 6 medical experts testify regarding her death
  - 5 say it was natural
  - 1 (the one who in 1997 had given the 'natural death certificate') says: "at the time I thought it was a natural death, but, given all these other cases reported by the media, I now think it was unnatural"
- The court follows the single dissenting expert who has implicitly used statistics!

The Chain Proof Technique

- **Only one of the 7 murders was proven 'beyond reasonable doubt'**, based on (flawed) medical argument (digoxin poisoning)
- Court then uses the chain proof, an allowed 'proof technique' in Dutch law:
- K similar crimes have been committed
- It has been proven beyond doubt that X committed one of them
- The court is allowed to conclude that X has committed all of them based on much less evidence than would normally be needed
- In principle, this makes sense also from a probabilistic point of view. ***But only if the 'crimes' are similar!***

What Went Wrong - Summary

1. **Wrong data**
2. **Wrong model**
3. **Wrong method**
4. **Wrong calculation**
5. **Wrongly worded conclusion**
6. **Wrong interpretation of conclusion**

More on Statistics

- Small Probability Events Happen all the time!
- So how *should* we do statistics in cases like this?

A Case about Small Probabilities

- Main error made by statistician:
  apply standard null hypothesis test when you can't
- Main error made by media, courts, general public:
  "something with very small probability happened. **That cannot be a coincidence**! Something funny must be going on!"
- Both cases of the general fallacy underlying this (and many other) (potential) miscarriages of justice:

Small probability events happen!

- Report based on intuition that, when we observe something with “incredibly small probability”, this is a strong indication that something funny is going on
- Yet incredibly improbable things happen all the time

Small probability events happen!

- Report based on intuition that, when we observe something with “incredibly small probability”, this is a strong indication that something funny is going on
- Yet incredibly improbable things happen all the time
  - I met a good friend from high school in a coffee house in Marrakech
  - Somebody in my street won the lottery
- The reason is, quite simply, that very many things can happen. If all these things are equally likely, they must all have very small probability. So whatever actually happens, will have very small probability.

Small Probability Events Happen!

- Richard Feynman, one of the greatest physicists of the 20th century:

“You know, the most amazing thing happened to me tonight. I was coming here, on the way to the lecture, and I came in through the parking lot. And you won't believe what happened. I saw a car with the license plate ARW 357. Can you imagine? Of all the millions of license plates in the state, what was the chance that I would see that particular one tonight? Amazing!”

- The fact that

  “something with incredibly small probability happened”

  is totally insufficient to conclude

  “this is most probably not a coincidence”
- Underlies many (potential) miscarriages of justice: Sally Clark, Overzier, Sweeney….

How can we then do statistics?

- How can one ever draw valid statistical conclusions (e.g. “this new medication really works”) if small probability events happen all the time?
- There are two valid methods:
  1. Frequentist testing: not applicable in court
  2. Bayesian method: difficult to apply in court

Frequentist Analysis of Fallacy

Suppose Jim wins the lottery. I test two hypotheses:

Hypothese H0: there’s nothing special about Jim

vs.

Hypothese H1: not a coincidence

(e.g.. Jimmy is a fraud)

I state: the probability that Jim would win is *that* small (< 1 in 1 million), that it cannot be a coincidence!

Frequentist Analysis of Fallacy

Suppose Jim wins the lottery. I test two hypotheses:

Hypothese H0: there's nothing special about Jim

vs.

Hypothese H1: not a coincidence
(e.g.. Jimmy is a fraud)

I state: the probability that Jim would win is *so* small (< 1 in 1 million), that it cannot be a coincidence!

But somebody *has to* win the lottery! If Margaret had won, we would have decided that Margaret was a fraud! That can't be right!

Correct Frequentist Statistics

Jim has won the lottery. I know that this may be a coincidence, but I don't fully trust Jimmy, because I've seen him hanging around the lottery office with binoculars...and I still see him doing it. Therefore I test

Hypothese H0: Nothing special about Jimmy

vs.

Hypothese H1: Jimmy is a fraud

with the test "Jimmy wins the ***next*** lottery"

Correct Frequentist Statistics

Jim has won the lottery. I know that this may be a coincidence, but I don't fully trust Jimmy, because I've seen him hanging around the lottery office with binoculars...and I still him doing it. Therefore I test

Hypothese H0: Nothing special about Jimmy

vs.

Hypothese H1: Jimmy is a fraud

with the test "Jimmy wins the ***next*** lottery"

If Jimmy wins again, it does say something, because we ***predicted*** it!

Neyman-Pearson Testing

- The idea is to identify, *before seeing the data*, a definite event with probability smaller than 1/10000
  - If that event happens, you reject the null hypothesis
  - If you have already seen the data before you decided on your event, this only works if you do an additional experiment to gain additional data

Neyman-Pearson Guarantee

- NP Testing has been designed such that, if it is performed repeatedly (and **correctly**!), then the following guarantee holds:
- on average, at most 1 in 10000 times that we do a NP test, we say "null hypothesis rejected" even though null hypothesis is true

$$\Pr_{H_0}(\text{I say "reject"}) \leq \frac{1}{10000}$$

$$\Pr_{H_0}(\text{p-value} \leq \frac{1}{10000}) \leq \frac{1}{10000}$$

Multiplication issue!

Neyman-Pearson Guarantee

- NP Testing has been designed such that, if it is performed repeatedly (and **correctly**!), then the following guarantee holds:
- on average, at most 1 in 10000 times that we do a NP test, we say "null hypothesis rejected" even though null hypothesis is true

$$\Pr_{H_0}(\text{I say "reject"}) \leq \frac{1}{10000}$$

$$\Pr_{H_0}(\text{p-value} \leq \frac{1}{10000}) \leq \frac{1}{10000}$$

Neyman-Pearson Guarantee

- In Lucia's case, statistician effectively promises that, if his method is used repeatedly, then at most 1 in 10000 times one would say "not a coincidence", whereas in truth, it was just a coincidence
- But you can only hold such a promise if you specify your test *before* the data are generated!

Statistics in court..,

- ... Is nearly always *after-the-fact* statistics. It is unethical or even impossible to set up an experiment to see whether presumed effect is "replicable"!
- That's why it is difficult statistics and...
- "Standard" frequentist statistics simply *not applicable*
- Bayesian statistics is applicable, to some extent, but not without problems of its own
- Contrary to the judge's opinion in the Sally Clark case, it *is* ''rocket science"

The Prosecutor's Fallacy

- The probability of the evidence given a hypothesis is not the same as the probability of the hypothesis given the evidence

$$\Pr(A \mid B) \neq \Pr(B \mid A)$$

- The "Prosecutor's Fallacy" is to think that the two are equal
  - closely related to the 'small probability is suspicious' fallacy)
- The right way to connect both probabilities is Bayes' theorem

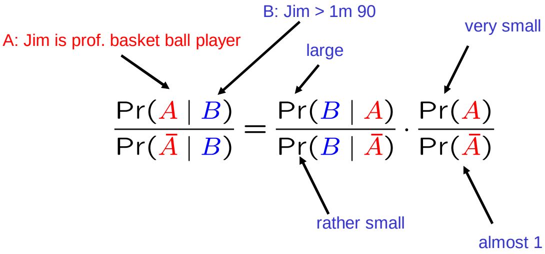
Bayes' Theorem

*Posterior odds = likelihood ratio * prior odds*

$$\frac{\Pr(A \mid B)}{\Pr(\bar{A} \mid B)} = \frac{\Pr(B \mid A)}{\Pr(B \mid \bar{A})} \cdot \frac{\Pr(A)}{\Pr(\bar{A})}$$

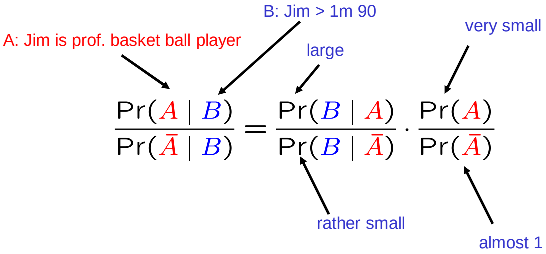
- A means "A happens"
- $\bar{A}$ means "A does not happen" $\Pr(\bar{A}) = 1 - \Pr(A)$

Bayes' Theorem, basket ball



$$\frac{\Pr(A \mid B)}{\Pr(\bar{A} \mid B)} = \frac{\Pr(B \mid A)}{\Pr(B \mid \bar{A})} \cdot \frac{\Pr(A)}{\Pr(\bar{A})}$$

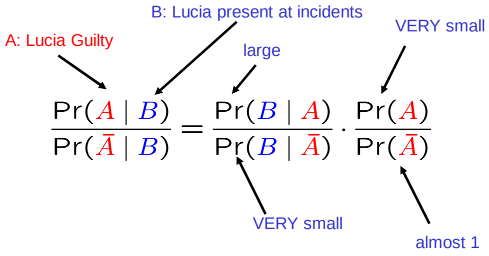
Bayes' Theorem, basket ball



$$\frac{\Pr(A \mid B)}{\Pr(\bar{A} \mid B)} = \frac{\Pr(B \mid A)}{\Pr(B \mid \bar{A})} \cdot \frac{\Pr(A)}{\Pr(\bar{A})}$$

Small = large * very small

Bayes' Theorem, Lucia



A: Lucia Guilty

B: Lucia present at incidents

large

VERY small

$$\frac{\Pr(A \mid B)}{\Pr(\bar{A} \mid B)} = \frac{\Pr(B \mid A)}{\Pr(B \mid \bar{A})} \cdot \frac{\Pr(A)}{\Pr(\bar{A})}$$

VERY small

almost 1

??? = very large * very small

Prior Probability that Lucia is Guilty?

- Frequentist Statisticians: ill-defined
  - there's no repeatable experiment
  - too narrow (no use of DNA evidence in court...)
- Hard-Core Bayesian: by thinking hard enough, you can come up with a number
  - problem: Lucia is not just an ordinary nurse (prostitute in youth, faked high school diploma) According to a an FBI agent witness "she fit the profile of serial killers". How should that influence the probabilities!?!?
- Cautious Bayesian:
  - consider *range* of values (say between 1 in 10000 and 1 in 1000000) and see what happens to the conclusion

Conclusion – Three Extreme Positions

1. Laplace (+- 1800): all legal "truth finding" should be based on Bayes 'theorem
   - Impossible. Judge/Jury often has to deal with uncertainty that can hardly be captured in terms of numbers
2. Worried Citizen: "statistic has no role in the courts: even if a probability is very small, it might still have been a coincidence"
   - Also wrong: numerical statistics sometimes useful and inevitable (good to know that 1 in 11 line-up identifications is wrong!)
3. The Legalist: when the judges uses the word "probability" without numbers, his reasoning is purely legal, and does not have to satisfy the basic rules of probability theory such as Bayes theorem
   - Also wrong: reasoning about uncertainty in a way contradicting mathematical theorems (e.g. Prosecutor's fallacy, chain proof in wrong circumstances) leads to wrong conclusions, even if 'probabilities' cannot be linked to frequencies

Conclusion

- Lucia isn't the Only One
- Sally Clark (UK), Henk Haalboom (NL), Kevin Sweeney (NL)...
- **Legal** profession has learned from Lucia though!
- **Medical** profession hasn't. To learn about their role, watch the **movie!**
- even though it doesn't acknowledge main heroes