EPSE 581C: Bayesian Methods

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

September 9, 2019

Syllabus and Course Policies

- Check website: ekroc.weebly.com → "Teaching"

8 pillars of statistics

- Design
- Sampling
- Measurement
- Estimation
- Inference
- Modelling
- Computation
- Communication

Note: some items may be repeated or concurrent

Probability & statistics

- **Probability** is a mathematical abstraction. It is usually used to describe the likelihood, or chance, of observing some event; e.g. sun or rain tomorrow, the height of the tallest person in the room.

Probability & statistics

- **Probability** is a mathematical abstraction. It is usually used to describe the likelihood, or chance, of observing some event; e.g. sun or rain tomorrow, the height of the tallest person in the room.
    - The most essential feature of a random phenomenon is its *distribution*; i.e. a function that tells us how likely is each possible outcome of the random phenomenon.

Probability & statistics

- **Probability** is a mathematical abstraction. It is usually used to describe the likelihood, or chance, of observing some event; e.g. sun or rain tomorrow, the height of the tallest person in the room.
    - The most essential feature of a random phenomenon is its *distribution*; i.e. a function that tells us how likely is each possible outcome of the random phenomenon.
- **Statistics** can be thought of as the discipine of *applied probability*. More specifically, it is the study of *uncertainty*.

Probability & statistics

- **Probability** is a mathematical abstraction. It is usually used to describe the likelihood, or chance, of observing some event; e.g. sun or rain tomorrow, the height of the tallest person in the room.
  - The most essential feature of a random phenomenon is its *distribution*; i.e. a function that tells us how likely is each possible outcome of the random phenomenon.
- **Statistics** can be thought of as the discipine of *applied probability*. More specifically, it is the study of *uncertainty*.
  - It is the goal of descriptive statistics to describe some random distribution via a *sample* of observations; e.g. sample mean as a "typical" value of the distribution.
  - It is the goal of inferential statistics to infer properties of some random distribution via a *sample* of observations and to *quantify* our confidence in these inferences; e.g. a 95% confidence interval for the mean.

Classical (frequentist) methodology

Most of the tools of classical statistics comprise the *frequentist* methodology: p-values, confidence intervals, hypothesis testing, Type I and Type II errors (power), ordinary least squares, maximum likelihood estimation, etc.

- These tools all rely on *conditional probabilities* of a very particular form:

$$\Pr(\text{data} \mid H), \quad \text{or} \quad \Pr(\text{statistic} \mid H).$$

- That is, we *suppose* a particular hypothesis H holds, then calculate the probability of observing our data (statistic) under this hypothesis.
- This is a *counterfactual* method of inference.

Classical (frequentist) methodology

- Relies on *counterfactuals* to make inferences:

$$\Pr(\text{data} \mid H), \quad \text{or} \quad \Pr(\text{statistic} \mid H).$$

- Therefore, treats hypotheses (parameters) as *fixed quantities*; i.e. all inference proceeds by *assuming* a particular hypothesis is *true*.
- At the same time, treats data as *random*; i.e. how likely is it that we observe a certain data set (or statistic) *given* that some hypothesis is true?
- Produces estimates that (theoretically) have good *frequency* properties: e.g. p-values, CIs.

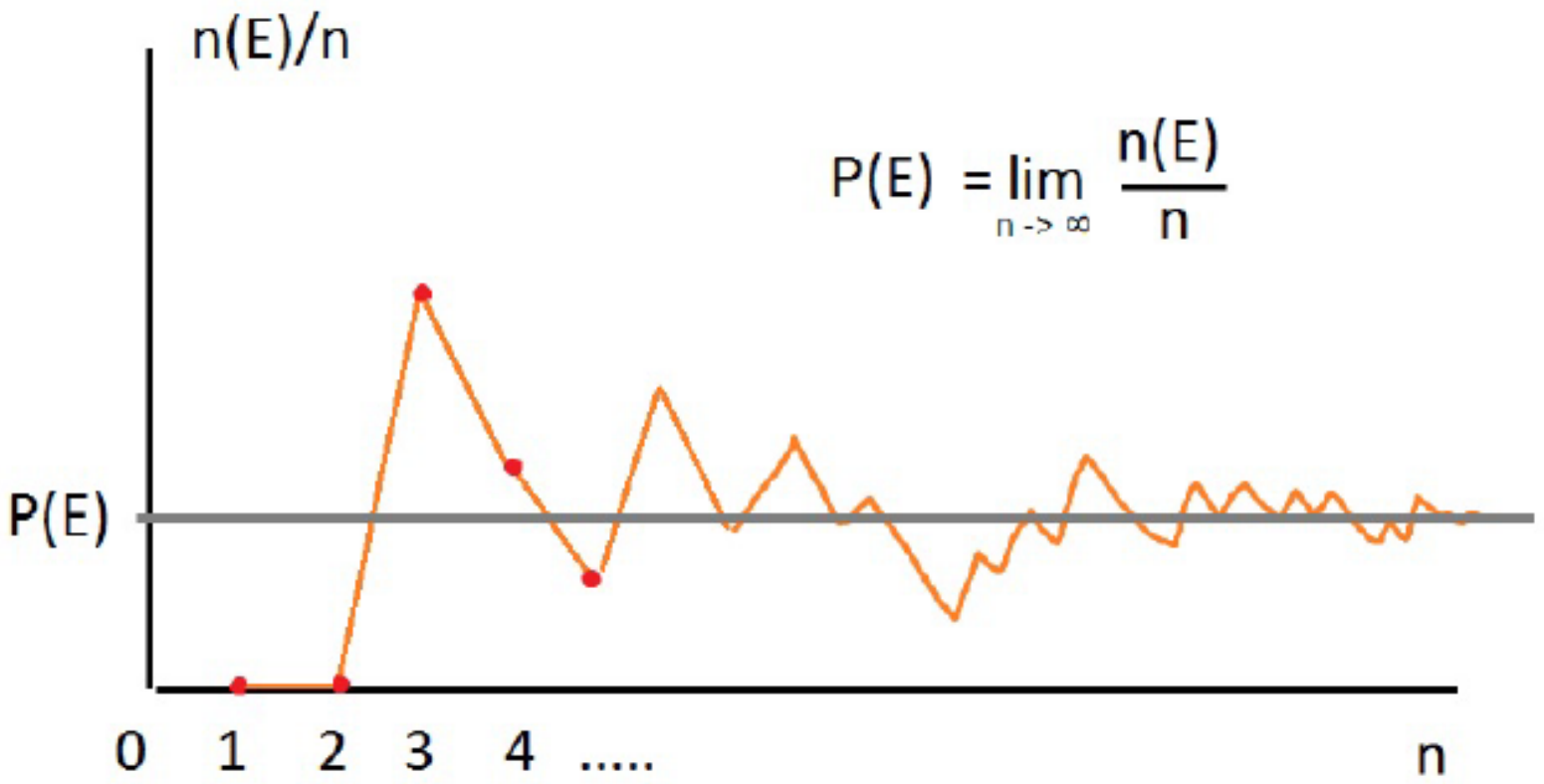
A formal definition of the probability of an event E

Consider repeating a random experiment (observation) many times; e.g. flipping a coin. If the sequence of trials are *independent* (i.e. the result of one trial of the experiment (observation) does not affect the result of any other trials), then the probability that an event E will occur is given by

$$\Pr(E) = \lim_{n \to \infty} \frac{n(E)}{n}$$

where n is the number of repetitions, and $n(E)$ is the number of times the event E occurs in n repetitions. $\Pr(E)$ is the limiting or long-run relative frequency.

A formal definition of the probability of an event E



Note that we must always have $0 \leqslant \Pr(E) \leqslant 1$.

Other definitions of probability?

- Does the previous definition always make sense?

Other definitions of probability?

- Does the previous definition always make sense?
  - Probability that it will rain tomorrow?
  - Certainty of agreement with the statement *"I support the implementation of a federal carbon tax"*?
  - Probability of a fair coin coming up heads in a finite universe (space-time)?
  - Et al.

Frequency properties

All frequentist-based quantities (classical statistics) inherently rely on the idea of *direct replication*:

- p-value gives the probability of observing a test statistic as or more extreme than the one we actually observe, given that the null hypothesis is true.
- If you repeat the same experiment many times, and each time calculate a sample mean and corresponding 95% CI, then about 95% of those CIs will contain the true population mean.
- Type I error (Type II error / power): the chance of making a false positive (false negative) upon repeated testing.

In general, Bayesian quantities do *not* have these frequentist properties or interpretations because the Bayesian paradigm does not rely on the notion of direct replication.

Bayesian methodology

Consider: we do science by collecting data and then using these data to inform our belief in a particular hypothesis.

- This scientific paradigm also relies on *conditional probabilities*, but of a different particular form:

$$\Pr(H \mid \text{data}).$$

  These quantities are called *posterior probabilities*.

- *Bayesian* methodology quantifies these posterior probabilities directly.
- We can relate counterfactual (frequentist) probabilities to these posterior probabilities by means of Bayes' Theorem.
- Bayesian methodology interprets the abstract notion of probability as a *degree of belief*, not a long-run frequency of occurence.

Bayesian methodology

- Relies on *posterior probabilities* to make inferences:

$$\Pr(H \mid \text{data})$$

- Therefore, treats hypotheses (parameters) as *random variables*.
- At the same time, treats data as *fixed*; i.e. how likely is it that a certain hypothesis is true *given* that we observe some particular data (statistic)?
- As we will see, Bayesian methods produce estimates that explicitly incorporate *prior* believability of hypotheses via Bayes' Theorem.

Bayes' Theorem

All of Bayesian methodology is predicated upon a simple mathematical relation: Bayes' Theorem (proven independently by Rev. Thomas Bayes circa 1760 and Pierre-Simon Laplace circa 1774).

Bayes' Theorem

Let the sets $F_1, F_2, \ldots, F_n$ be disjoint (no overlap). Further, suppose that they partition the entire universe of events: $S = \{F_1 \text{ or } F_2 \text{ or } \ldots \text{ or } F_n\}$. Then:

$$\Pr(F_i \mid E) = \frac{\Pr(E \mid F_i)\Pr(F_i)}{\sum_{j=1}^{n} \Pr(E \mid F_j)\Pr(F_j)}$$

- Jeffreys (1973): "Bayes' theorem is to the theory of probability what the Pythagorean theorem is to geometry."
- To understand this theorem, we need to understand *conditional probability*.

Unconditional vs. conditional probability

- Experiment: toss a fair coin twice: set of all outcomes is $S = \{HH,\ HT,\ TH,\ TT\}$
- $\Pr(\textit{tossing 2 heads}) = 1/4$ (all outcomes equally likely)

Unconditional vs. conditional probability

- Experiment: toss a fair coin twice: set of all outcomes is $S = \{HH,\ HT,\ TH,\ TT\}$
- $\Pr(\textit{tossing 2 heads}) = 1/4$ (all outcomes equally likely)
- Suppose you are given *additional information* that at least 1 head was tossed in this experiment. What is the probability of tossing 2 heads, *given this extra information*?

Unconditional vs. conditional probability

- Experiment: toss a fair coin twice: set of all outcomes is $S = \{HH,\ HT,\ TH,\ TT\}$
- $\Pr(\textit{tossing 2 heads}) = 1/4$ (all outcomes equally likely)
- Suppose you are given *additional information* that at least 1 head was tossed in this experiment. What is the probability of tossing 2 heads, *given this extra information*?
- With the given information, the only possibilites left from S are: $\{HH,\ HT,\ TH\}$.

Unconditional vs. conditional probability

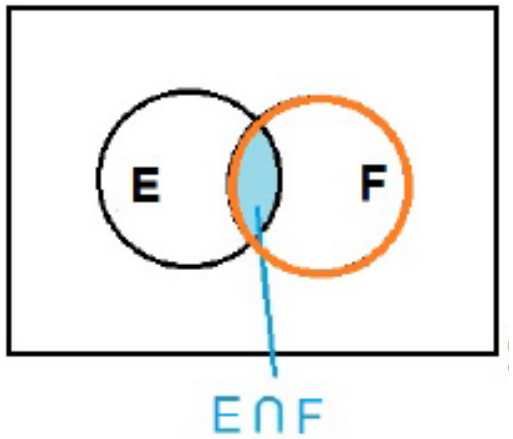
- Experiment: toss a fair coin twice: set of all outcomes is $S = \{HH,\ HT,\ TH,\ TT\}$
- $\Pr(\textit{tossing 2 heads}) = 1/4$ (all outcomes equally likely)
- Suppose you are given *additional information* that at least 1 head was tossed in this experiment. What is the probability of tossing 2 heads, *given this extra information*?
- With the given information, the only possibilites left from S are: $\{HH,\ HT,\ TH\}$.
- Thus, the probability that we toss 2 heads, *given that at least 1 head was tossed*, is...?

Unconditional vs. conditional probability

- Experiment: toss a fair coin twice: set of all outcomes is $S = \{HH,\ HT,\ TH,\ TT\}$
- $\Pr(\textit{tossing 2 heads}) = 1/4$ (all outcomes equally likely)
- Suppose you are given *additional information* that at least 1 head was tossed in this experiment. What is the probability of tossing 2 heads, *given this extra information*?
- With the given information, the only possibilites left from S are: $\{HH,\ HT,\ TH\}$.
- Thus, the probability that we toss 2 heads, *given that at least 1 head was tossed*, is...?
- $= \frac{1}{3}$.

Conditional Probability

- In general, given that an event F has occurred, the probability that another event E occurs is called the *conditional probability of E given F*.



- Notation and formula:

$$\Pr(E \mid F) = \frac{\Pr(E \cap F)}{\Pr(F)} = \frac{\Pr(E \text{ and } F)}{\Pr(F)}$$

Conditional Probability

Note, in general:

$$\Pr(E \mid F) \neq \Pr(F \mid E)$$

or:

$$\frac{\Pr(E \text{ and } F)}{\Pr(F)} \neq \frac{\Pr(E \text{ and } F)}{\Pr(E)}$$

If you assume different information, you should not expect to end up with the same probabilities!

Independence of Events

Definition

Two events E and F are said to be **independent** if and only if $\Pr(E \mid F) = \Pr(E)$ or $\Pr(F \mid E) = \Pr(F)$.

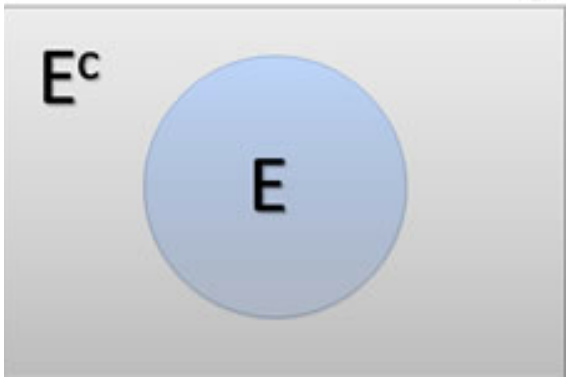
- By definition of conditional probability then, we have

$$\Pr(E \text{ and } F) = \Pr(E) \cdot \Pr(F) \quad \text{if and only if } E,\ F \text{ are independent.}$$

- In words: Conditional probabilities of events can only equal their unconditional probabilities when they are *independent* of the information being conditioned upon.

Complements of Events

- We use the notation E^c to denote the *complement* of the event E; i.e., E^c contains all events that are not in E.



- Since probabilities of all events must sum to 1, we have

$$\Pr(E^c) = 1 - \Pr(E).$$

- Also, we can always write the whole universe of events as

$$U = E \cup E^c = E \text{ or } E^c.$$

The Prosecutor's Fallacy

The Prosecutor's Fallacy is a common probability *misconception*, intimately related to p-value misconceptions: it is the fallacy of thinking that $\Pr(A \text{ and } B)$ is the same as $\Pr(A \mid B)$.

This is obviously false in general! Recall:

$$\Pr(A \mid B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)}$$

So can only be true if $\Pr(B) = 1$ or if $\Pr(A \text{ and } B) = 0$.

This fallacy is quite common and can have many distressing consequences...

The Case of Sally Clark

- In 1998, Sally Clark was accused of murdering her two infant sons. One died in 1996 at eleven weeks old. The second died a year later at eight weeks of age.

- Sir Roy Meadow, pediatrician and expert witness for the prosecution, testified that the chance of two children in the same family dying from Sudden Infant Death Syndrome (SIDS) was about $(1/8500)^2$, or 1 in 73 million.

- On the strength of this testimony alone, Clark was convicted in 1999. The Royal Statistical Society then pointed out the flaws in the argument. What are they?

The Case of Sally Clark

- Flaw #1: The events of two *siblings* dying from SIDS are *not* independent. There is a genetic component! In reality, the probability of two children from the same family dying of SIDS is much closer to $1/8500$ than to $(1/8500)^2$.

- Flaw #2: Meadow confused the conditional and unconditional probabilities (the Prosecutor's Fallacy).

  Let I : event that Clark is innocent of murder, E : event of two dead children (the evidence).

  We know that in general,

$$\Pr(I \mid E) \neq \Pr(E \text{ and } I).$$

The Case of Sally Clark

Now,

$$\begin{aligned}\Pr(I \mid E) &= \frac{\Pr(I \text{ and } E)}{\Pr(E)} \\ &= \frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I^c \text{ and } E)}\end{aligned}$$

What are the events I and E and I^c and E?

- I and E is the event of the two chidren dying by SIDS.
- I^c and E is the event of the two children dying by murder.

Double SIDS is rare, but double murder is much, much rarer! So,

$$\Pr(I^c \text{ and } E) \ll \Pr(I \text{ and } E).$$

The Case of Sally Clark

$$\Pr(I^c \text{ and } E) \ll \Pr(I \text{ and } E)$$

implies:

$$\Pr(I \text{ and } E) + \Pr(I^c \text{ and } E) \ll \Pr(I \text{ and } E) + \Pr(I \text{ and } E)$$

which implies:

$$\frac{1}{\Pr(I \text{ and } E) + \Pr(I^c \text{ and } E)} \gg \frac{1}{\Pr(I \text{ and } E) + \Pr(I \text{ and } E)}$$

which implies:

$$\frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I^c \text{ and } E)} \gg \frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I \text{ and } E)}$$

The Case of Sally Clark

$$\frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I^c \text{ and } E)} \gg \frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I \text{ and } E)}$$

But LHS equals $\Pr(I \mid E)$ and RHS equals $1/2$:

$$\Pr(I \mid E) = \frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I^c \text{ and } E)}$$
$$\gg \frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I \text{ and } E)} = \frac{1}{2}$$

So, $\Pr(I \mid E) \approx 1$!

Moral of the story 1: **circumstantial evidence of a rare event is very weak evidence.**
Moral of the story 2: **conditional information is radically different from unconditional information.**

The Case of Sally Clark

- Sally Clark's conviction was overturned in 2003, after she had already spent four years in jail.
- In prison, Clark developed psychological problems and an alcohol dependency.
- Sally Clark died of alcohol poisoning in 2007.

Bayes' Theorem

Recall:

Bayes' Theorem

Let the sets $F_1, F_2, \ldots, F_n$ be disjoint (no overlap). Further, suppose that they partition the entire universe of events: $S = \{F_1 \text{ or } F_2 \text{ or } \ldots \text{ or } F_n\}$. Then:

$$\Pr(F_i \mid E) = \frac{\Pr(E \mid F_i)\Pr(F_i)}{\sum_{j=1}^{n} \Pr(E \mid F_j)\Pr(F_j)}$$

- Now let's prove this theorem:

Bayes' Theorem

Start with the simplest case of $n = 2$, so S is partitioned by $F_1 = F$ and $F_2 = F^c$.

Using the defn. of conditional probability again (twice), we have:

$$\begin{aligned} \Pr(F \mid E) &= \frac{\Pr(F \cap E)}{\Pr(E)} \\ &= \frac{\Pr(E \mid F)\Pr(F)}{\Pr(E)} \end{aligned}$$

Bayes' Theorem

Start with the simplest case of $n = 2$, so S is partitioned by $F_1 = F$ and $F_2 = F^c$.

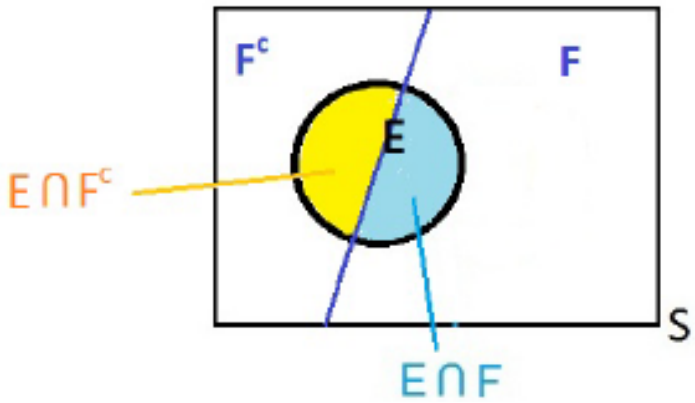
Using the defn. of conditional probability again (twice), we have:

$$\begin{aligned} \Pr(F \mid E) &= \frac{\Pr(F \cap E)}{\Pr(E)} \\ &= \frac{\Pr(E \mid F)\Pr(F)}{\Pr(E)} \end{aligned}$$

Now we need to show that $\Pr(E)$ can be decomposed into a sum of appropriately weighted conditional probabilities.

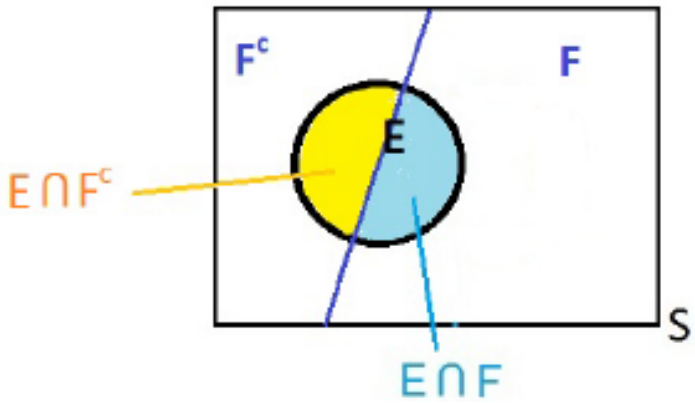
Bayes' Theorem

In this simplest case, the universe is partitioned by only two sets: F and F^c.



Bayes' Theorem

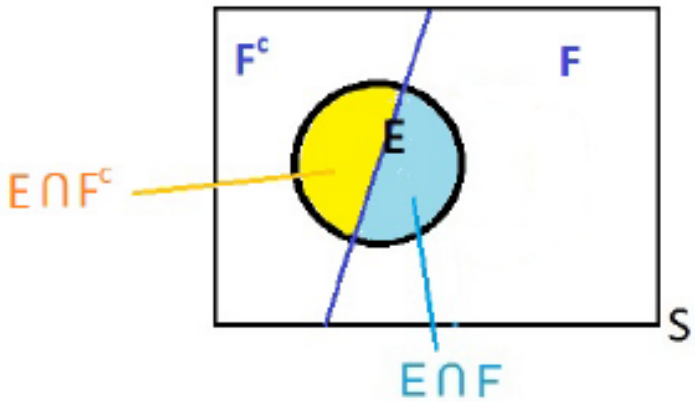
In this simplest case, the universe is partitioned by only two sets: F and F^c.



$$\Pr(E) = \Pr((E \cap F) \cup (E \cap F^c))$$

Bayes' Theorem

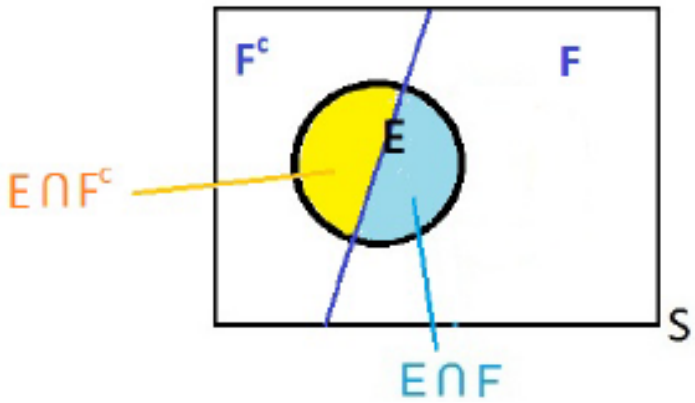
In this simplest case, the universe is partitioned by only two sets: F and F^c.



$$\begin{aligned}\Pr(E) &= \Pr((E \cap F) \cup (E \cap F^c)) \\ &= \Pr(E \cap F) + \Pr(E \cap F^c) \quad \text{(disjoint events)}\end{aligned}$$

Bayes' Theorem

In this simplest case, the universe is partitioned by only two sets: F and F^c.



$$\begin{aligned}
\Pr(E) &= \Pr((E \cap F) \cup (E \cap F^c)) \\
&= \Pr(E \cap F) + \Pr(E \cap F^c) \quad \text{(disjoint events)} \\
&= \Pr(E \mid F)\Pr(F) + \Pr(E \mid F^c)\Pr(F^c) \quad \text{(defn. of cond. prob.)}
\end{aligned}$$

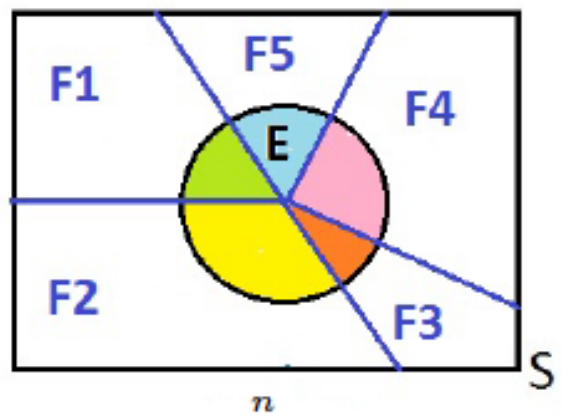
Bayes' Theorem

Now plugging in our expression for $\Pr(E)$ into the previous slide, we arrive at Bayes' Theorem in its simplest form:

$$\Pr(F \mid E) = \frac{\Pr(E \mid F)\Pr(F)}{\Pr(E \mid F)\Pr(F) + \Pr(E \mid F^c)\Pr(F^c)} \quad \checkmark$$

Bayes' Theorem

In general, $F_1, \ldots, F_n$ partition the universe (n=5 in the figure below).



- Can repeat above argument for any *finite* partition of the universe of events to find:

$$\Pr(E) = \sum_{i=1}^{n} \Pr(E \mid F_i)\Pr(F_i).$$

Bayes' Theorem

- But what about when there are infinitely many disjoint sets (hypotheses) that partition the universe of events?
- Think of the simple t-test: the mean μ of a random variable can be *any* real number: an infinite *continuum* of possibilities!
- Simple summation is not going to work....
- Next time: we will do a little calculus-based probability to sort this out.

Bayes' theorem to Bayesian methodology

- Bayes' Theorem gives us an explicit way to decompose the probability of a hypothesis given some data:

$$\Pr(H \mid \text{data}) = \frac{\Pr(\text{data} \mid H) \cdot \Pr(H)}{\Pr(\text{data})}$$

$$\Pr(H_1 \mid \text{data}) = \frac{\Pr(\text{data} \mid H_1) \cdot \Pr(H_1)}{\sum_{i=1}^{n} \Pr(\text{data} \mid H_i) \Pr(H_i)}$$

Bayes' theorem to Bayesian methodology

$$\Pr(H \mid \text{data}) = \frac{\Pr(\text{data} \mid H) \cdot \Pr(H)}{\Pr(\text{data})}$$

$$\Pr(H_1 \mid \text{data}) = \frac{\Pr(\text{data} \mid H_1) \cdot \Pr(H_1)}{\sum_{i=1}^{n} \Pr(\text{data} \mid H_i) \Pr(H_i)}$$

Critical terminology:

- Likelihood
- Prior probability
- Posterior probability
- Normalizing factor

Bayes' theorem to Bayesian methodology

$$\Pr(H \mid \text{data}) = \frac{\Pr(\text{data} \mid H) \cdot \Pr(H)}{\Pr(\text{data})}$$

$$\Pr(H_1 \mid \text{data}) = \frac{\Pr(\text{data} \mid H_1) \cdot \Pr(H_1)}{\sum_{i=1}^{n} \Pr(\text{data} \mid H_i) \Pr(H_i)}$$

- The posterior probability has a natural interpretation: it gives an explicit measure of certainty to a hypothesis given some evidence for or against that hypothesis.
- This is the quantity we are always most interested in in practice for scientific inquiry.

P-values vs. posterior probabilities

So why don't we build statistical tests to give us information about the posterior probability rather than about the p-value?

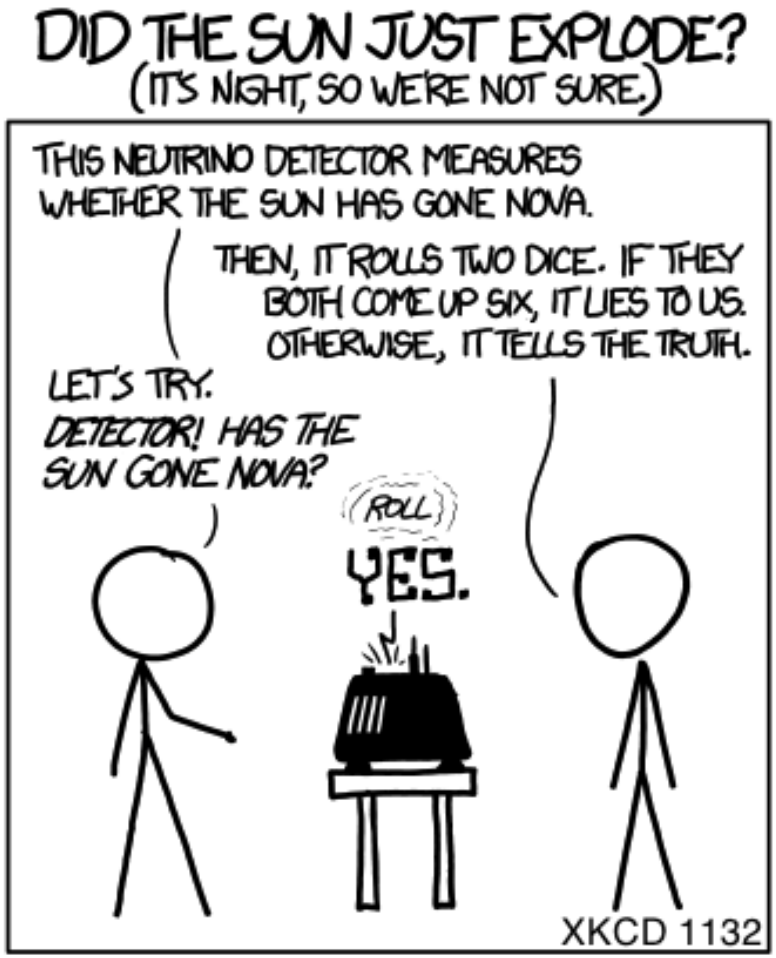
- Classical statistics is built entirely around the p-value out of necessity: the p-value is easy to calculate because we assume a hypothesis is true; i.e. we assume *extra information* than we have.
- The posterior probability relies on three quantities:
  - the *likelihood*, $\Pr(\text{data} \mid H)$
  - the *prior probability*, $\Pr(H)$
  - and the *normalizing factor*, $\Pr(\text{data})$

$$\Pr(H \mid \text{data}) = \frac{\Pr(\text{data} \mid H) \cdot \Pr(H)}{\Pr(\text{data})}$$

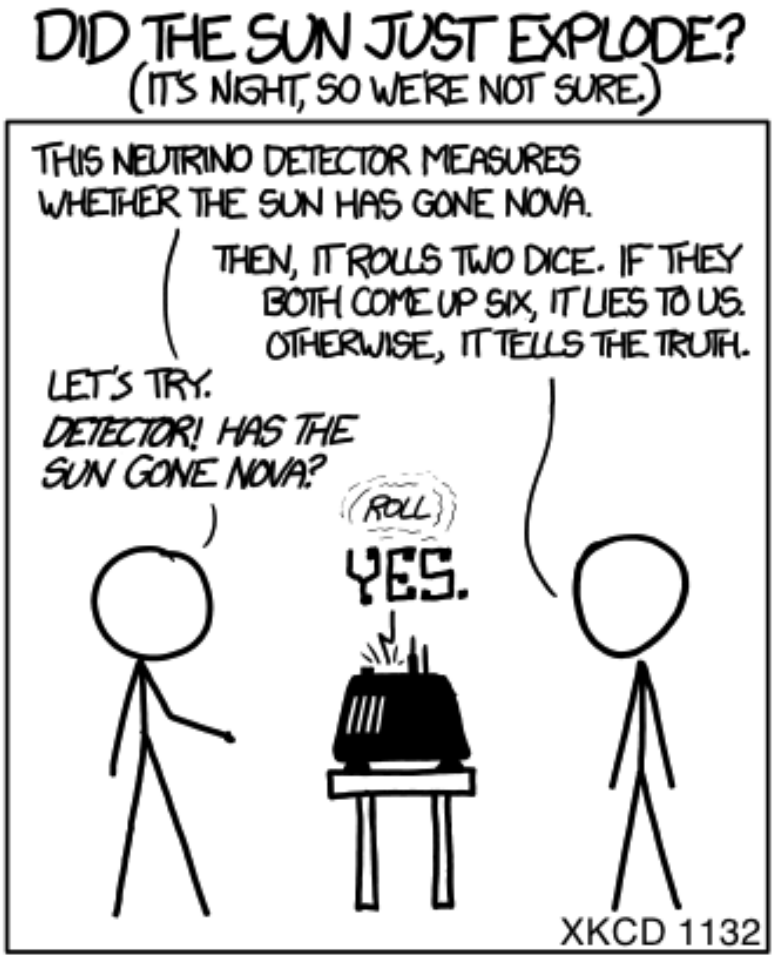
Likelihoods and priors

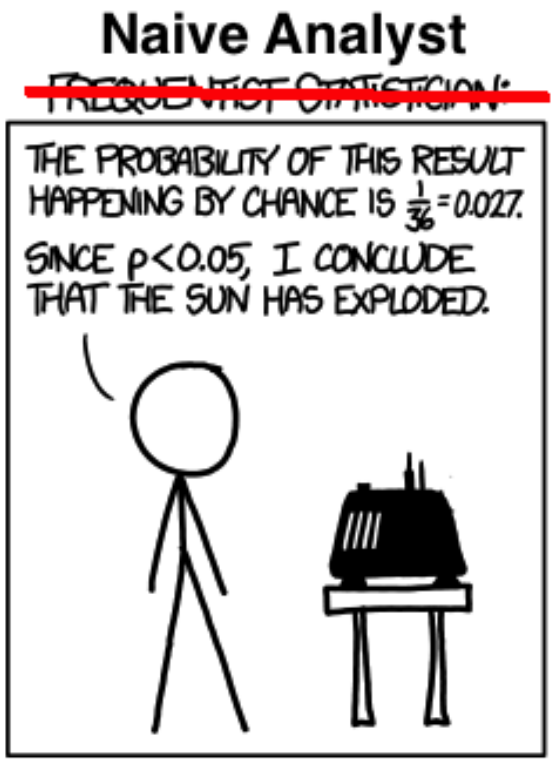
- The likelihood is easy to calculate (more next time); it is basically the same as the p-value.
- The prior is *not* determined by the data; the researcher must set its value using prior information.
- The role of the prior has always been the most controversial aspect of Bayesian methodology.
- However, the frequentist framework *also* requires consideration of prior probabilities when exact replications are not available or possible.
- There are *many* ways to "best" specify a prior; decades of research has been devoted to this topic. We will talk a *lot* more about this in future classes.

The role of prior probabilities



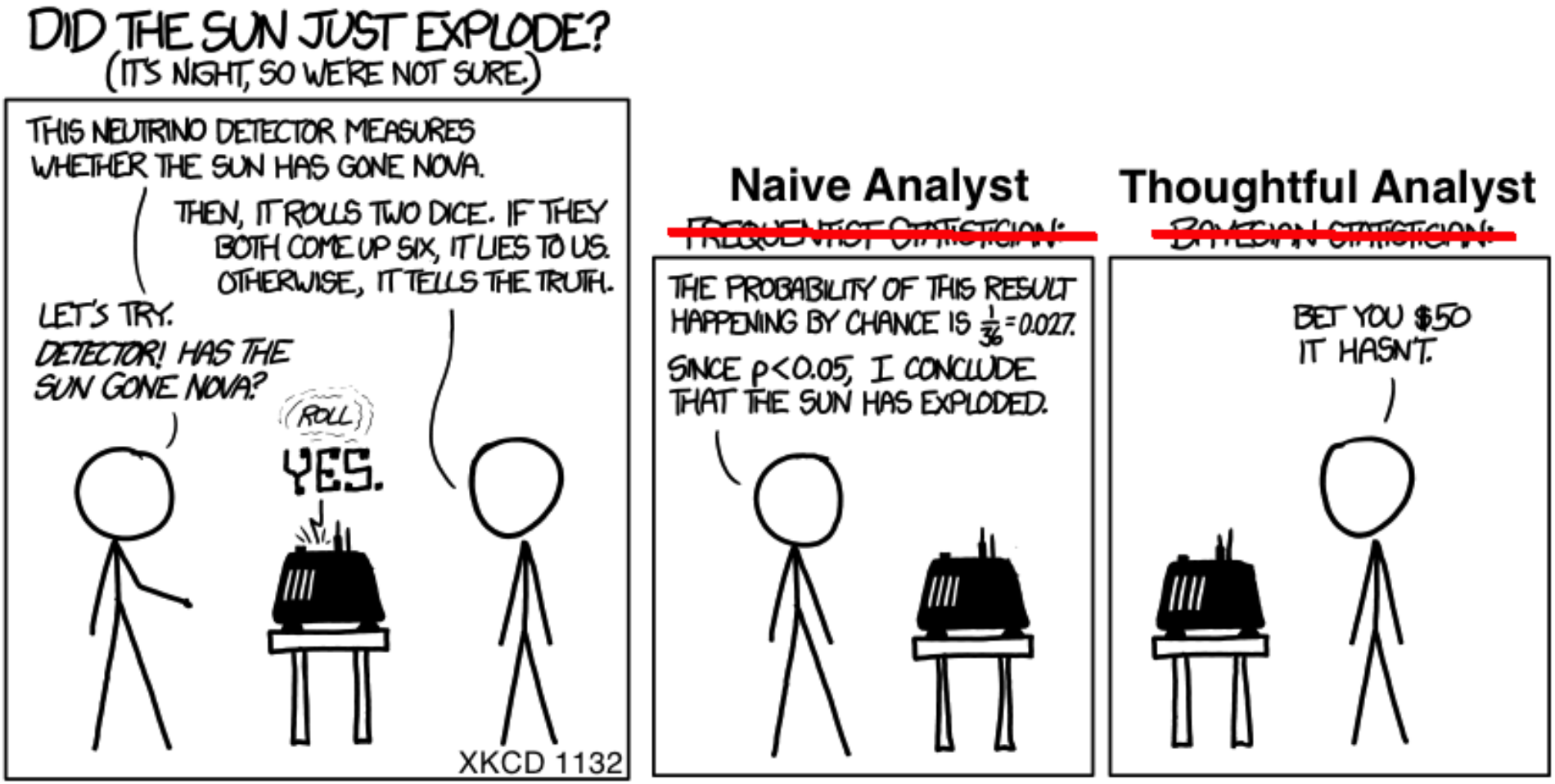
The role of prior probabilities





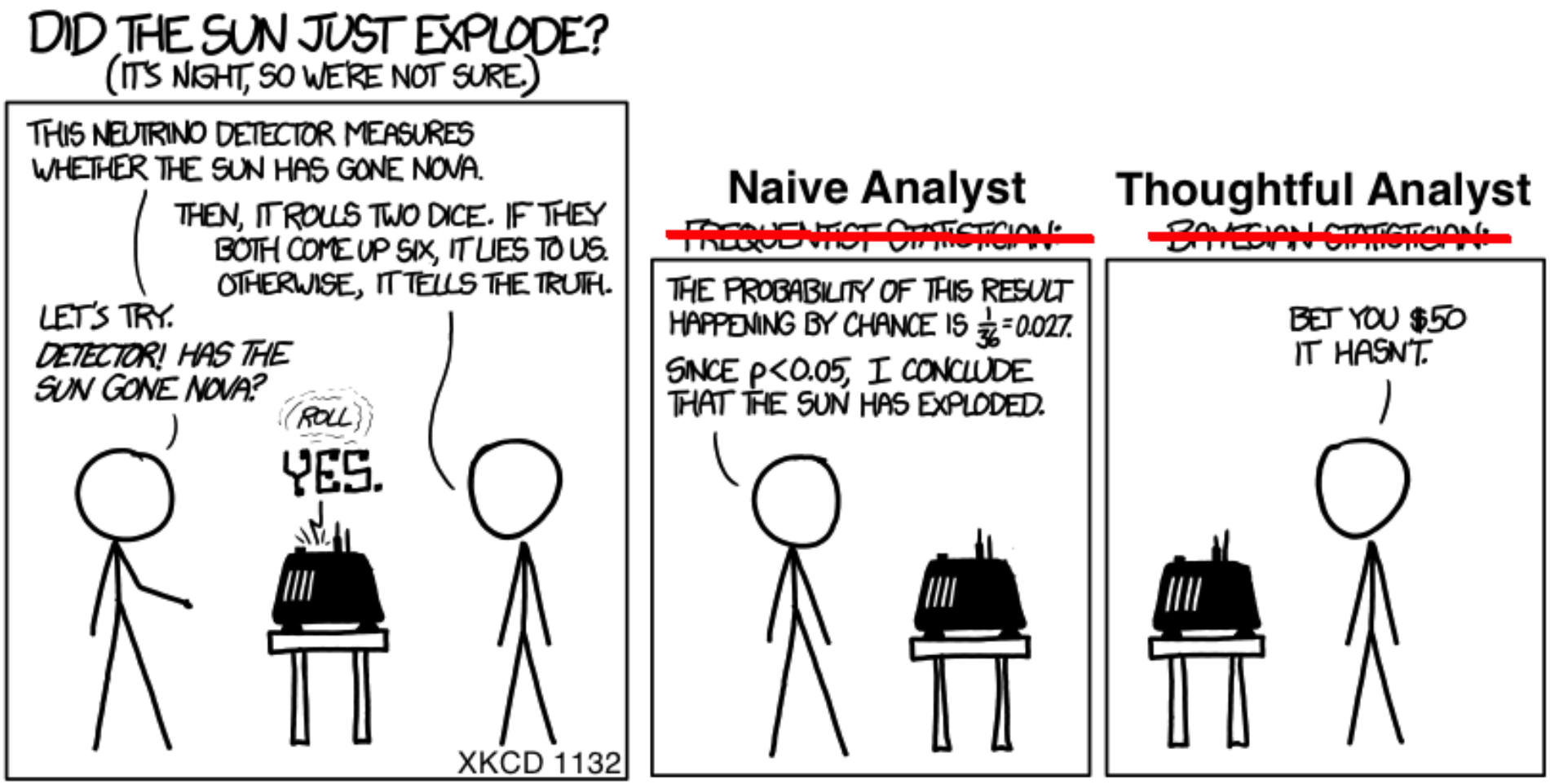
Replication?

The role of prior probabilities



What if we repeat this experiment 10 times *and observe the same outcome each time?*

The role of prior probabilities



What if we repeat this experiment 10 times *and observe the same outcome each time?* **Prior will still dominate.**

The role of prior probabilities

Why will the prior still dominate?

- Suppose: $\Pr(H_0) = \Pr(\text{sun has not exploded}) = \frac{10^{100}}{10^{100}+1}$
- Observing the same outcome (double-sixes) under the null 10 times in a row happens with probability about $\Pr(\text{data} \mid H_0) = 10^{-17}$.
- Now,

$$\Pr(\text{data}) = \Pr(\text{data} \mid H_0) \cdot \Pr(H_0) + \Pr(\text{data} \mid H_A) \cdot \Pr(H_A).$$

  The first term is very small, $10^{-17} \cdot \frac{10^{100}}{10^{100}+1} \approx 10^{-17}$, but the second term is *much, much smaller*: $\approx 10^{-100}$

- Same conclusion as with Sally Clark:

$$\frac{10^{-17}}{10^{-17} + 10^{-100}} \approx 1.$$

Posterior probabilities

- The posterior probability relies on three quantities:
  - the *likelihood*, $\Pr(\text{data} \mid H)$
  - the *prior probability*, $\Pr(H)$
  - and the *normalizing factor*, $\Pr(\text{data})$

$$\Pr(H \mid \text{data}) = \frac{\Pr(\text{data} \mid H) \cdot \Pr(H)}{\Pr(\text{data})}$$

The normalizing factor

- In general, the normalizing factor is the hardest term to deal with (compute) in practice.
- Ironically, it is the *least* important factor theoretically (because it doesn't depend on the quantity of interest: the *hypothesis* in question).
- There are some special cases where the normalizing factor can be computed analytically (i.e. by hand), but for most complex modelling scenarios, this would be hopeless.
- Hence, Bayesian methods could not really take off before the advent of *cheap and fast computing technology*. More on this later.
- Thus, classical statistics was *forced* to develop the p-value based methodology: *frequentist* statistics.

Bayesian vs. Frequentist approaches

- Bayesian methodology relies on *posterior distributions* rather than test statistics.
- Bayesian methodology has no (direct) concept of Type I or Type II errors (or power).
- Bayesian methodology relies heavily on *prior* (unconditional) probabilities of hypotheses; frequentist methodology only *indirectly* relies on priors.
- Bayesian methodology does *not* rely on p-values or counterfactuals. However, there is an evil little quantity called a *Bayes factor* that has attempted to play the role of the p-value. It suffers from many of the same problems as the p-value (more later).