Bayesian Statistics

Vera Demberg

Saarland University

25.1.2022

In our reasonings concerning matter of fact, there are all imaginable degrees of assurance, from the highest certainty to the lowest species of moral evidence. A wise man, therefore, proportions his belief to the evidence. – David Hume

Where are we?

1. Bayes Theorem
2. Frequentist statistics vs. Bayesian statistics
3. Bayesian hypothesis tests
4. Why is Bayesian stats interesting?

Table of Contents

1. Bayes Theorem
2. Frequentist statistics vs. Bayesian statistics
3. Bayesian hypothesis tests
4. Why is Bayesian stats interesting?

Bayesian analysis

Take the prior, combine it with the data, get out the posterior.

$$P(hypothesis|data) = \frac{P(data|hypothesis) * P(hypothesis)}{P(data)}$$

An Example (which we can calculate by hand)

Data: Dan Navarro is carrying an umbrella.

Question: Is it raining?

Prior and Likelihood

Prior: what we believe before: P(h)

Dan Navarro lives in Australia, and at the moment it's their summer, so they have a lot of good weather over there.

Hypothesis	Degree of Belief
Rainy day	0.15
Dry day	0.85

Likelihood: $P(d|h)$ conditional probability of Dan carrying an umbrella when it rains / doesn't rain. Actually, Dan rarely carries an umbrella, she sometimes forgets it even if rain is forecasted: P(umbrella|rainy)=0.3.

	Data	
Hypothesis	Umbrella	No umbrella
Rainy day	0.30	0.70
Dry day	0.05	0.95

Belief updating

Now, we find out that Dan actually *is* carrying an umbrella. What is the probability that it's raining?

We use Bayes' rule:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

Hypothesis	Degree of Belief
Rainy day	0.15
Dry day	0.85

	Data	
Hypothesis	Umbrella	No umbrella
Rainy day	0.30	0.70
Dry day	0.05	0.95

So $P(rain|umbrella) = \frac{P(umbrella|rain)*P(rain)}{P(umbrella)}$

$$P(umbrella) = P(umbrella|rain) * P(rain) + P(umbrella|no\_rain) * P(no\_rain)$$
$$= 0.3 * 0.15 + 0.05 * 0.85 = 0.0875$$

$P(rain|umbrella) = \frac{0.3*0.15}{0.0875} = 0.51$

Taking into account a prior: advantage of Bayesian modelling

The point about the prior is central:

- common example: what's the likelihood of having breast cancer when you get a positive result at mammography? $\mathrm{P(c \mid m)} = \frac{P(m|c)*P(c)}{P(m)}$
- probability of positive mammography given cancer is 90%: $\mathrm{P(m \mid c) = 0.9}$
- let's assume that prior probability of having breast cancer is 0.4%: $\mathrm{P(c) = 0.004}$
- P(m)? by law of total probability: $P(m) = P(m, c) + P(m, \neg c)$
- $\mathrm{P(m,c) = 0.9{*}0.004 = 0.0036}$
- $\mathrm{P(m,\neg c) = P(m \mid \neg c) * P(\neg c) = 0.12 * 0.996 = 0.1195}$
- $\mathrm{P(c|m)} = \frac{0.9*0.004}{0.1195} = 0.03.$

Table of Contents

1. Bayes Theorem
2. Frequentist statistics vs. Bayesian statistics
3. Bayesian hypothesis tests
4. Why is Bayesian stats interesting?

Frequentist Statistics

What we have done so far is also known as “frequentist statistics”.
The logics for frequentist statistics work by thinking what proportion of many random samples drawn you would expect to show a result at least as extreme as the observation you made.

Bayesian statistics

The way of thinking is very different in Bayesian statistics.

In Bayesian statistics it's all about prior beliefs, and belief revision in the face of new evidence.

Frequentist vs. Bayesian Statistics

Frequentist vs. Bayesian

- Frequentist: parameters are fixed, data is random
- Bayesian: parameters are random, data is fixed

What's a p-value?

- Probability of test statistic, given null hypothesis

So what do Bayesians get?

- No p values
- Probabilities for certain parameter values, given the observed data

How does this connect to the Bayes rule?

Bayes Rule

A and B are *observable events*.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Note that $P(\cdot)$ is the probability of an event.

Bayesian Statistics

Instead of looking at events, we will be looking at *probability distributions*.

$$f(\theta|data) = \frac{f(data|\theta)f(\theta)}{f(data)}$$

$f(\cdot)$ is a probability density.

How does this connect to the Bayes rule?

Bayesian Statistics

Instead of looking at events, we will be looking at *probability distributions*.

$$f(\theta|data) = \frac{f(data|\theta)f(\theta)}{f(data)}$$

$f(\cdot)$ is a probability density.

What is $f(data)$?

How does this connect to the Bayes rule?

Bayesian Statistics

Instead of looking at events, we will be looking at *probability distributions*.

$$f(\theta|data) = \frac{f(data|\theta)f(\theta)}{f(data)}$$

$f(\cdot)$ is a probability density.

What is $f(data)$? Just a normalizing constant.

$$f(data) = \sum_{i=1}^{n} f(data|\theta_i)P(\theta = \theta_i)$$

How does this connect to the Bayes rule?

Bayesian Statistics

Instead of looking at events, we will be looking at *probability distributions*.

$$f(\theta|data) = \frac{f(data|\theta)f(\theta)}{f(data)}$$

$f(\cdot)$ is a probability density.

What is $f(data)$? Just a normalizing constant.

$$f(data) = \sum_{i=1}^{n} f(data|\theta_i)P(\theta = \theta_i)$$

Without this normalizing constant, we get: $f(\theta|data) \propto f(data|\theta)f(\theta)$

Posterior $\propto$ Likelihood $\times$ Prior

Example

Example: Fake news detection model

You have developed a new fake news detector, and found that on a set of 100 news items, it detected 24 correctly. A previously existing fake news detector typically detects 20% of fakes. Is your new one better?

Example

Example: Fake news detection model

You have developed a new fake news detector, and found that on a set of 100 news items, it detected 24 correctly. A previously existing fake news detector typically detects 20% of fakes. Is your new one better?

How would you go about this question following Frequentist logic?

Example

Example: Fake news detection model

You have developed a new fake news detector, and found that on a set of 100 news items, it detected 24 correctly. A previously existing fake news detector typically detects 20% of fakes. Is your new one better?

How would you go about this question following Frequentist logic?

Calculate the binomial distribution for n=100, with success probability 0.2. (This comes from the logic of repeating an experiment with 100 biased coins that come up heads with p=0.2 many times and observing how often we obtain events with 24 heads or more.)

Example

Example: Fake news detection model

You have developed a new fake news detector, and found that on a set of 100 news items, it detected 24 correctly. A previously existing fake news detector typically detects 20% of fakes. Is your new one better?

How would you go about this question following Frequentist logic?

Calculate the binomial distribution for n=100, with success probability 0.2. (This comes from the logic of repeating an experiment with 100 biased coins that come up heads with p=0.2 many times and observing how often we obtain events with 24 heads or more.)
Then sum up the probabilities for all events such that $X > 23$.
Multiply by 2 if you want to do a 2-tailed test, then check whether the result is smaller than α.

Example

Example: Fake news detection model

You have developed a new fake news detector, and found that on a set of 100 news items, it detected 24 correctly. A previously existing fake news detector typically detects 20% of fakes. Is your new one better?

How would you go about this question following Frequentist logic?

Calculate the binomial distribution for n=100, with success probability 0.2. (This comes from the logic of repeating an experiment with 100 biased coins that come up heads with p=0.2 many times and observing how often we obtain events with 24 heads or more.)
Then sum up the probabilities for all events such that $X > 23$.
Multiply by 2 if you want to do a 2-tailed test, then check whether the result is smaller than α.
pbinom(23,100,0.2, lower.tail=FALSE)$*$2 $<$ 0.05?

Example

Example: Fake news detection model

You have developed a new fake news detector, and found that on a set of 100 news items, it detected 24 correctly. A previously existing fake news detector typically detects 20% of fakes. Is your new one better?

How would you go about this question following Frequentist logic?

Calculate the binomial distribution for n=100, with success probability 0.2. (This comes from the logic of repeating an experiment with 100 biased coins that come up heads with p=0.2 many times and observing how often we obtain events with 24 heads or more.)
Then sum up the probabilities for all events such that $X > 23$.
Multiply by 2 if you want to do a 2-tailed test, then check whether the result is smaller than α.
pbinom(23,100,0.2, lower.tail=FALSE)$*2 < 0.05$?
nope, pbinom(23,100,0.2, lower.tail=FALSE) $= 0.19$, so accept H_0.

In Bayesian Statistics....

.

Example: Fake news detection model

You have developed a new fake news detector, and found that on a set of 100 news items, it detected 24 correctly. A previously existing fake news detector typically detects 20% of fakes. Is your new one better?

In Bayesian Stats, the logics work differently

- We construct a *generative model*, and calculate with what probability its parametrisation will give rise to the data we observed.

In Bayesian Statistics....

.

Example: Fake news detection model

You have developed a new fake news detector, and found that on a set of 100 news items, it detected 24 correctly. A previously existing fake news detector typically detects 20% of fakes. Is your new one better?

In Bayesian Stats, the logics work differently

- We construct a *generative model*, and calculate with what probability its parametrisation will give rise to the data we observed.
- What is the "generative model"?
  The process that gives rise to the observations, i.e. a binomial process that success / failure of fake news detection: `rbinom(1, 100, x)` (generate one time a sample of 100 instances, which are true detections with probability x). What is probability x? This is what we mean when we talk about "estimating parameter values".

In Bayesian Statistics....

- One way for estimating the parameter x would be to run the generative model many times with all possible parameter settings for x, and observe what parameter settings of x lead to the observed result with what probability.
- Additionally, we can have some prior ideas about sensible parameter settings for x which we can put into the model as a prior (we will get back to this later).
- Then we can compare it to a generative model for the previously existing fake news detector to find out whether the new detector is reliably better than the old one.

Bayesian Statistics Formula

$$f(\theta|data) \propto f(data|\theta)f(\theta)$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- θ is the unknown parameter which describes the true proportion of fake news that our detector can classify correctly.

Bayesian Statistics Formula

$$f(\theta|data) \propto f(data|\theta)f(\theta)$$

Posterior $\propto$ Likelihood $\times$ Prior

- θ is the unknown parameter which describes the true proportion of fake news that our detector can classify correctly.
- *data* is our observation of 24 correct detections out of 100.

Bayesian Statistics Formula

$$f(\theta|data) \propto f(data|\theta)f(\theta)$$

Posterior $\propto$ Likelihood $\times$ Prior

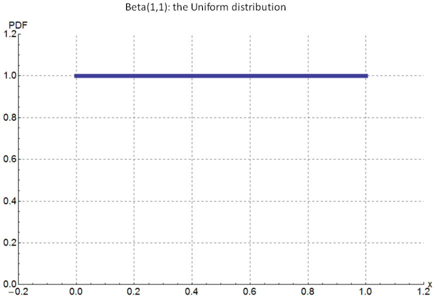
- θ is the unknown parameter which describes the true proportion of fake news that our detector can classify correctly.
- *data* is our observation of 24 correct detections out of 100.
- $f(data|\theta)$ (the likelihood function) is a binomial distribution

Bayesian Statistics Formula

$$f(\theta|data) \propto f(data|\theta)f(\theta)$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- θ is the unknown parameter which describes the true proportion of fake news that our detector can classify correctly.
- *data* is our observation of 24 correct detections out of 100.
- $f(data|\theta)$ (the likelihood function) is a binomial distribution
- $f(\theta)$ is the prior.
  Here we can choose a flat prior (uniform distribution, with range of 0 to 1), which we can describe mathematically as Beta(1,1).



So let's calculate the posterior!

Bayesian Statistics Formula

$$f(\theta|data) \propto f(data|\theta)f(\theta)$$

Posterior $\propto$ Likelihood $\times$ Prior

- $f(data|\theta)$ binomial distribution: $P(data|\theta) = \theta^{24}(1-\theta)^{76}$
- $f(\theta)$ is the prior: Beta(1,1).

$$Beta(a,b) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1} \quad \text{if } 0 < x < 1; 0 \text{ otherwise.}$$

$\frac{1}{B(a,b)}$ is a normalizing constant that we won't worry about for now.
So our posterior can be calculated as:

$$f(\theta|data) \propto \theta^{24}(1-\theta)^{76} * \theta^{1-1}(1-\theta)^{1-1} = \theta^{24}(1-\theta)^{76}$$

So let's calculate the posterior!

Bayesian Statistics Formula

$$f(\theta|data) \propto f(data|\theta)f(\theta)$$

Posterior $\propto$ Likelihood $\times$ Prior

- $f(data|\theta)$ binomial distribution: $P(data|\theta) = \theta^{24}(1-\theta)^{76}$
- $f(\theta)$ is the prior: Beta(1,1).

$$Beta(a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \quad \text{if } 0 < x < 1; 0 \text{ otherwise.}$$

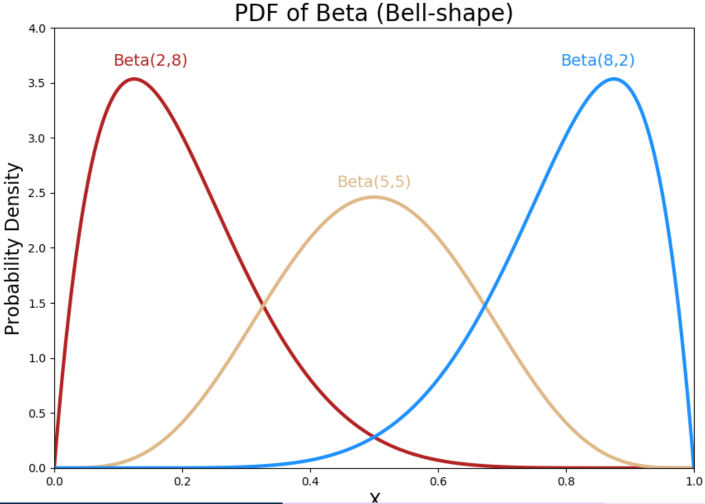
$\frac{1}{B(a,b)}$ is a normalizing constant that we won't worry about for now.
So our posterior can be calculated as:

$$f(\theta|data) \propto \theta^{24}(1-\theta)^{76} * \theta^{1-1}(1-\theta)^{1-1} = \theta^{24}(1-\theta)^{76}$$

In this case, the posterior is purely determined by the data.

Beta distribution

But we could choose our prior differently, in order to express a belief about what value the parameter is most likely to have.



Beta distribution

- Prior: Beta(2,8)

$$f(\theta|data) \propto \theta^{24}(1-\theta)^{76} * \theta^{2-1}(1-\theta)^{8-1} = \theta^{25}(1-\theta)^{83}$$

- Prior: Beta(5,5)

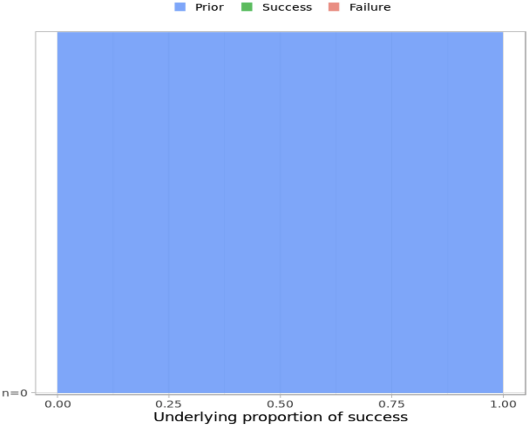
$$f(\theta|data) \propto \theta^{24}(1-\theta)^{76} * \theta^{5-1}(1-\theta)^{5-1} = \theta^{28}(1-\theta)^{80}$$

- Prior: Beta(8,2)

$$f(\theta|data) \propto \theta^{24}(1-\theta)^{76} * \theta^{8-1}(1-\theta)^{2-1} = \theta^{31}(1-\theta)^{77}$$

Let's visualise this.

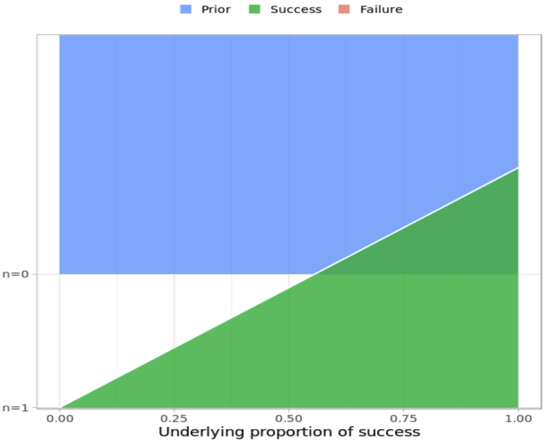
100 fake news items, 24 of them were identified correctly;
let's simulate the process of evaluating those 24 news items:



(no observations yet)

Let's visualise this.

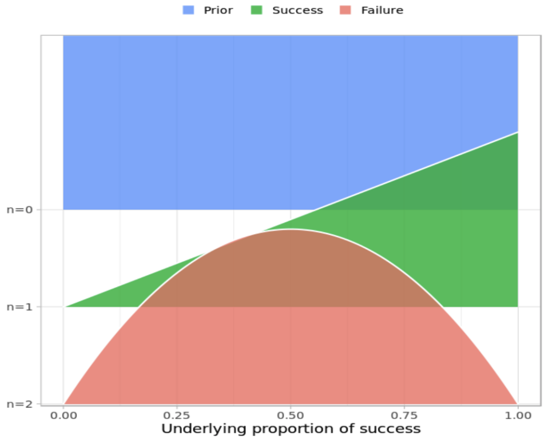
100 fake news items, 24 of them were identified correctly;
let's simulate the process of evaluating those 24 news items:



(first observation: correct)

Let's visualise this.

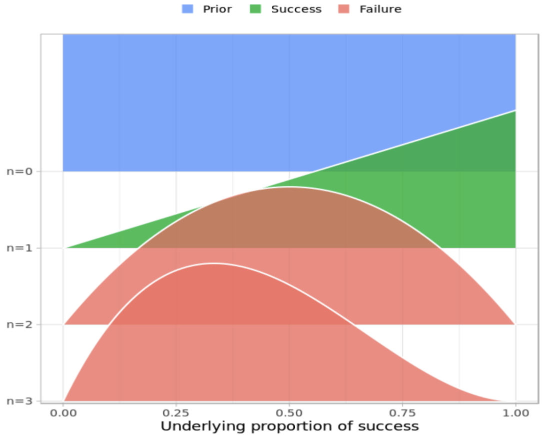
100 fake news items, 24 of them were identified correctly;
let's simulate the process of evaluating those 24 news items:



(second observation: false)

Let's visualise this.

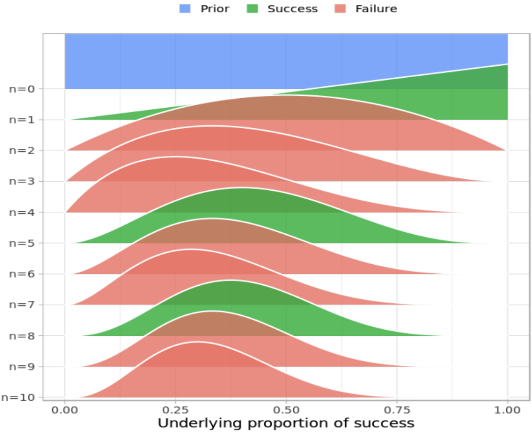
100 fake news items, 24 of them were identified correctly;
let's simulate the process of evaluating those 24 news items:



(third observation: false)

Let's visualise this.

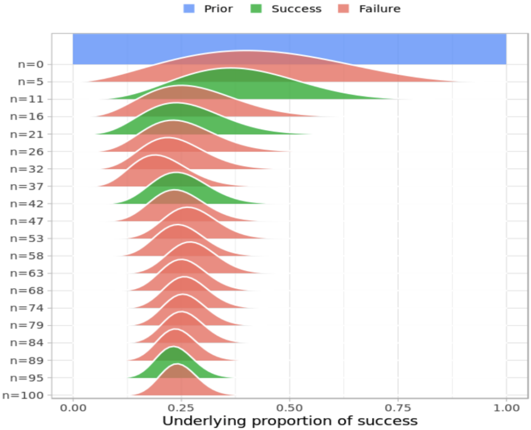
100 fake news items, 24 of them were identified correctly;
let's simulate the process of evaluating those 24 news items:



(10 obs, 3 correct, 7 false)

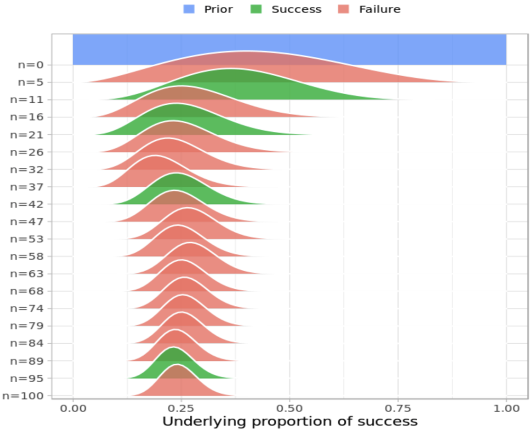
Let's visualise this.

100 fake news items, 24 of them were identified correctly;
let's simulate the process of evaluating those 24 news items:



(all 100 observations are in)

How compatible is this with a success prob of 0.2?



We can "read off" the probability of "underlying probability of success" from the graph (or from Beta(24,76)).

In this case, we accept the null hypothesis.

Calculating the posterior

- In this case, we saw that the posterior could easily be calculated analytically from the likelihood and the prior: Binomial likelihood $\times$ Beta prior $=$ Beta posterior.
- But depending on the distribution of the likelihood function and the prior, this might be a lot more difficult.
- Another example of an easy case: Poisson likelihood $\times$ Gamma prior $=$ Gamma posterior.
- But in most realistic cases, the posterior cannot be specified by a particular density. Then we need to sample from the posterior distribution.
- Sampling Techniques (how these work is not covered in this course):
    - Gibbs sampling
    - Metropolis-Hasting
    - Hamiltonian Monte Carlo

Bayesian analysis

What have we learned so far?

- Via Bayes rule, we can combine prior information with observational data
- In Bayesian statistics we do this in order to estimate the value of the parameter underlying a distribution from which we have observed data.
- Prior and posterior are combined, think of it as a weighted mean.
- How to estimate the shape of the posterior: calculate it in simple cases, estimate it through sampling otherwise.

Bayesian analysis

What have we learned so far?

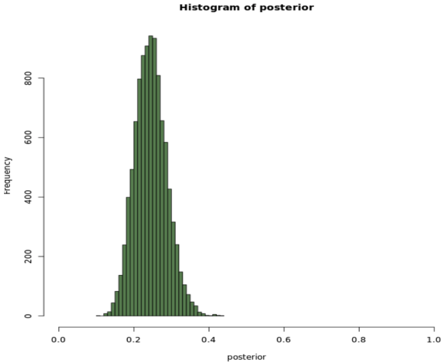
- Via Bayes rule, we can combine prior information with observational data
- In Bayesian statistics we do this in order to estimate the value of the parameter underlying a distribution from which we have observed data.
- Prior and ~~posterior~~ likelihood are combined, think of it as a weighted mean.
- How to estimate the shape of the posterior: calculate it in simple cases, estimate it through sampling otherwise.

Discussion points:

- How to choose a prior?
- What is the effect of choosing a prior? (as a function of the amount of data that we have observed)

Bayesian analysis

How do we get the probability of the parameter from the posterior?
One option is **sampling** from the posterior.
Then, we can calculate an expected value (the mean) and the 95% credible interval from that sample.



mean $= 0.245$; $\mathrm{CI}_{95} = [0.167, 0.333]$

Bayesian linear regression in R

Let's now take a look at how to run Bayesian linear regressions in R...

Credible interval vs. confidence interval

Evaluating Bayesian parameters

Confidence interval: Probability that a range contains the true value

- There is a 90% probability that range contains the true value `confint()`

Credible interval: Probability that the true value is within a range

- There is a 90% probability that the true value falls within this range `posterior_interval()`

Probablility of point lying between two values (Bayesian) vs. probability of the two points capturing the true value (Frequentist)

→ only Bayesian inference allows to make inferences about the parameters.

Table of Contents

1. Bayes Theorem
2. Frequentist statistics vs. Bayesian statistics
3. Bayesian hypothesis tests
4. Why is Bayesian stats interesting?

Bayesian hypothesis tests

Probability of null hypothesis given the data:

$$P(h_0|d) = \frac{P(d|h_0)P(h_0)}{P(d)}$$

Bayesian hypothesis tests

Probability of null hypothesis given the data:

$$P(h_0|d) = \frac{P(d|h_0)P(h_0)}{P(d)}$$

Probability of alternative hypothesis given the data:

$$P(h_1|d) = \frac{P(d|h_1)P(h_1)}{P(d)}$$

Posterior odds

In Bayesian statistics, you will often see reference to the "Bayes factor". So what is that?

It's the data-based update on the prior to obtain the posterior odds ratio.

Posterior odds if $P(h_1|d) = 0.75$ and $P(h_0|d) = 0.25$:

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{0.75}{0.25} = 3$$

Posterior odds

In Bayesian statistics, you will often see reference to the "Bayes factor". So what is that?
It's the data-based update on the prior to obtain the posterior odds ratio.
Posterior odds if $P(h_1|d) = 0.75$ and $P(h_0|d) = 0.25$:

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{0.75}{0.25} = 3$$

Bayes factor:

$$\underset{\text{Posterior odds}}{\underset{\uparrow}{\frac{P(h_1|d)}{P(h_0|d)}}} = \underset{\text{Bayes factor}}{\underset{\uparrow}{\frac{P(d|h_1)}{P(d|h_0)}}} \times \underset{\text{Prior odds}}{\underset{\uparrow}{\frac{P(h_1)}{P(h_0)}}}$$

Bayes factor

Bayes factor:

$$\underset{\text{Posterior odds}}{\underset{\uparrow}{\frac{P(h_1|d)}{P(h_0|d)}}} = \underset{\textbf{Bayes factor}}{\underset{\uparrow}{\frac{P(d|h_1)}{P(d|h_0)}}} \times \underset{\text{Prior odds}}{\underset{\uparrow}{\frac{P(h_1)}{P(h_0)}}}$$

The Bayes factor thus shows how much the data changes your beliefs.

Hypothesis testing with Bayesian stats

So if we set the prior probability as equal (= we don't know whether null hypothesis or alternative hypothesis is true), then the posterior odds are the same as the Bayes Factor. This is why the Bayes Factor (BF) is usually reported.

$$\underbrace{\frac{P(h_1|d)}{P(h_0|d)}}_{\uparrow\ \text{Posterior odds}} = \underbrace{\frac{P(d|h_1)}{P(d|h_0)}}_{\uparrow\ \textbf{Bayes factor}} \times \underbrace{\frac{P(h_1)}{P(h_0)}}_{\uparrow\ \text{Prior odds}}$$

Interpretation of the Bayes Factor

Bayes factor	Interpretation
1 - 3	Negligible evidence
3 - 20	Positive evidence
20 - 150	Strong evidence
>150	Very strong evidence

Easy interpretation

If there is more support for the null hypothesis compared to the alternative hypothesis, the Bayes factor will be < 1. In that case, you could change it around and report the odds of the null hypothesis:

$$\mathrm{BF} = \frac{P(d|h_1)}{P(d|h_0)} = \frac{0.1}{0.2} = 0.5$$

$$\mathrm{BF}' = \frac{P(d|h_0)}{P(d|h_1)} = \frac{0.2}{0.1} = 2$$

Table of Contents

1. Bayes Theorem
2. Frequentist statistics vs. Bayesian statistics
3. Bayesian hypothesis tests
4. Why is Bayesian stats interesting?

More intuitive

You find that in your study, you get a p value of 0.04 – please formulate for yourself what this means.

More intuitive

You find that in your study, you get a p value of 0.04 – please formulate for yourself what this means.

It's tempting to interpret the p value as the "probability of the null hypothesis being true".

More intuitive

You find that in your study, you get a p value of 0.04 – please formulate for yourself what this means.

It's tempting to interpret the p value as the "probability of the null hypothesis being true".

But as you hopefully have learned during the past few months, this is wrong, it's not the correct interpretation.

More intuitive

You find that in your study, you get a p value of 0.04 – please formulate for yourself what this means.

It's tempting to interpret the p value as the "probability of the null hypothesis being true".

But as you hopefully have learned during the past few months, this is wrong, it's not the correct interpretation.

Advantage 1: With Bayesian statistics however, we CAN talk about the probability of a hypothesis being true.

Avoiding common methodological failures in data collection and analysis.

Consider the following example situation:

You've come up with a really exciting research hypothesis and you design a study to test it. You're very diligent, so you run a power analysis to work out what your sample size should be, and you run the study. You run your hypothesis test and out pops a p-value of 0.072. Really bloody annoying, right?

Avoiding common methodological failures in data collection and analysis.

Consider the following example situation:

You've come up with a really exciting research hypothesis and you design a study to test it. You're very diligent, so you run a power analysis to work out what your sample size should be, and you run the study. You run your hypothesis test and out pops a p-value of 0.072. Really bloody annoying, right?

What do you do?

1. You conclude that there is no effect, and try to publish it as a null result
2. You guess that there might be an effect, and try to publish it as a "borderline significant" result
3. You give up and try a new study
4. You collect some more data to see if the p value goes up or (preferably!) drops below the "magic" criterion of $p < .05$

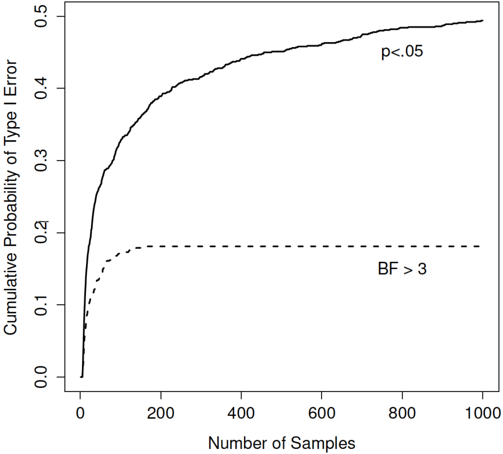
all bad choices :-(

Unfortunately, the theory of null hypothesis testing forbids one to follow option 4. The reason is that the theory assumes that the experiment is finished and all the data is in. It considers only two possible decisions (reject or retain null hypothesis), "continue experiment" is not an option.

Advantage 2: with Bayesian statistics, this problem is much less bad.

Peeking at data

What happens if you take decisions on whether to stop your experiment or continue collecting data after looking at the data?



In R

You can use the `BayesFactor` package (see chapter 17, Navarro), or the `rstan` package.

To summarize

- theoretical underpinnings of Bayesian stats
- Bayesian hypothesis testing / Bayes Factor
- Bayesian statistics vs. Frequentist statistics
  $\rightarrow$ problem with peaking at data
- the DataCamp chapters will cover some basics about the inference process of estimating the parameters given the data.