SOC6707 Intermediate Data Analysis

Monica Alexander

Week 8: Introduction to Bayesian probability and inference

Overview

What is probability?

- Probability is all about measuring the chance of something (an event happening or observing a particular thing)
- There is uncertainty associated with the event or observation, and probability helps us to quantify this
- We use the word ‘probability’ in everyday language, as well as ‘chance’ or ‘likelihood’
- But there are formal definitions of probability

Different interpretations of probability

- Just like there are different types or interpretations of reasoning (inductive, deductive), there are different types of probability
- Up until now (perhaps without realizing it) all our interpretations of probability (and how they relate to statistical inference) have been **objective** or **classical** or **frequentist**

Frequentist probability

Frequentist probability

- The probability of an event is defined as the long run relative frequency of that event in (infinitely many) trials
- If we had infinite time and were able to repeat the process that we're interested in over and over, we would calculate the probability as the number of times the event happened divided by the total number of trials
- E.g. the probability of heads on a coin, the probability of tossing 3 on a die, the probability of a toddler drinking a cup of water without spilling any
- Frequentist probability is devoid of opinion, the interpretation is literally just counting events

Frequentist probability

We are interested in some parameter, call it θ. E.g. θ is the probability of a head, the probability of a 3, the probability of no water spilled

- In frequentist probability, we treat θ as fixed
- I.e. there is only one true value of θ, and we want to estimate it
- We collect data (tossing a coin or die, giving the toddler water)
- Then to estimate θ we are fundamentally interested in

$$P(\text{Data}|\theta)$$

i.e. given a certain value of our parameter of interest, how likely is it that we observed that data?

Bayesian probability

Problems (?) with the objective view

- A lot of things that we are interested in cannot be repeated over and over
- For example, thinking before last year's US election, we may have been interested in asking "What is the probability that Biden will win?"
- In a frequentist set-up, this is hard to interpret; we can't repeat the 2020 election over and over
- Biden either wins or he doesn't

But there's another way

- Frequentist probability is just one interpretation of what probability is
- An alternative: **subjective** or **Bayesian** probability
- Here, probability is interpreted as a state of knowledge about the world, where this could be a reasonable expectation, or a personal belief
- Before seeing any data, we can have an opinion about the probability of an event happening
- We can then update that belief based on any data we do see
- Probability is a measure of strength of belief

Bayesian probability

- We are still interested in parameter θ as before
- But now θ is not treated as fixed, but random, with its own set of likely values based on our knowledge / the data that we see
- In the Bayesian framework we can say something about θ even before seeing any data
- We can then update our beliefs about θ once we've seen data

To estimate θ we are interested in

$$P(\theta|\text{Data})$$

i.e. conditional on seeing a set of observations, what is the probability of θ equaling certain values?

- Notice this is the opposite to previous, and as such was initially called the 'inverse probability problem'

A brief history

Thomas Bayes

"*Given:* the number of times in which an unknown event has happened and failed: *Required:* the chance that the probability of it happening in a single trial lies somewhere between any two degrees of probability that can be named."

Thomas Bayes

- Presbyterian minister (1701-1761)
- Studied logic and theology, interested in probability
- "An Essay towards solving a Problem in the Doctrine of Chances" published in 1763 by his friend Richard Price

Pierre-Simon Laplace

- Did a bunch of stuff (1749-1827)
- Largely responsible for development of Bayesian interpretation of probability
- set out a mathematical system of inductive reasoning based on probability

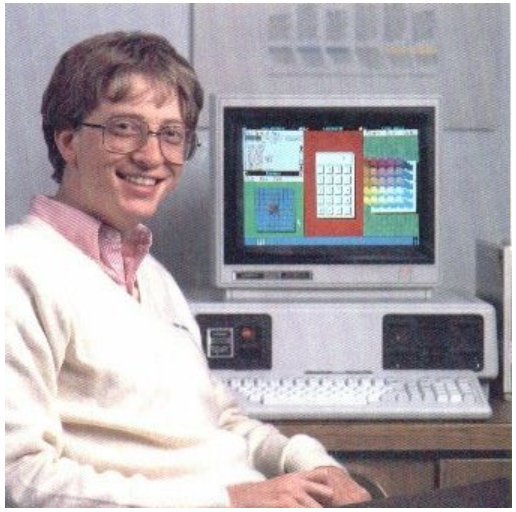
Another reason to dislike Fisher?

- Until early 1920s, the inverse probability method, which is based on what is now called Bayes's Theorem, was pretty much the predominant point of view of statistics.
- R. A. Fisher and Jerzy Neyman, criticized Bayesian inference for the use of subjective elements in an objective discipline. In Fisher's words "The theory of inverse probability is founded upon an error, and must be wholly rejected"
- Frequentist methods became the norm; hypothesis testing developed by Fisher, Pearson, Neyman in early 20th century

Bayesian bubblings

- Beyond the agenda of a few influential statisticians, a big problem with using Bayesian probability for statistical inferences was the lack of computing power
- Bayesian inference often ends up in complicated expressions that cannot be solved on paper
- Need computational algorithms to solve a lot of problems
- Work by physicists in the 1950s and 1960s laid the foundations of algorithms used today (Ulam, Metropolis (Arianna Rosenbluth), Hastings)

The golden age



- Bayesian methods reappeared in the 1980s/1990s (important paper by Geman and Geman in 1980)
- Rise and rise of Markov Chain Monte Carlo (MCMC) algorithms, which allow for complex Bayesian models to be estimated
- Coupled with rise of more complex data structures and data problems

Probability review and Bayes rule

Probability review and Bayes rule

- **Probability function**: a rule that assigns a value $P(A)$ to each event such that
  - $P(A)$ is greater than or equal to zero ($P(A) \geq 0$)
  - $P(A)$ is less that or equal to one ($P(A) \leq 1$)
  - the sum of all $P(A)$ is equal to one for a finite sample space. ($\sum_i^N P(A) = 1$)

Multiplicative / Intersection rule

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

- $P(B|A)$ is conditional probability i.e. the probability of B given that A is true

Bayes rule

Bayes rule for events

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

Bayes rule in practice

Example: breast cancer screening (using mammograms) in Germany. Imagine we know

- The probability an asymptomatic woman has breast cancer is 0.8%.
- If she has breast cancer, the probability is 90% that she has a positive mammogram
- If she does not have breast cancer, the probability is 7% that she still has a positive mammogram.

Suppose a woman has a positive mammogram: What is the probability she actually has breast cancer?

Breast cancer

Use Bayes rule for events:

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

Let

- C be the cancer outcome (=1 if cancer, 0 otherwise)
- M be the mammogram outcome (=1 if mammogram is positive, 0 otherwise)

"Suppose a woman has a positive mammogram: What is the probability she actually has breast cancer?"

A somewhat famous example because physicians had no idea what the answer should be.

We want to know $P(C = 1|M = 1)$.

Breast cancer

We want to know $P(C = 1|M = 1)$.

- $P(C = 1) = 0.008$.
- $P(M = 1|C = 1) = 0.9$.
- $P(M = 1|C = 0) = 0.07$.
- so $P(M = 1) =?$

Slide for working

Bayesian updating

Use Bayes rule, get $P(C = 1|M = 1) = 9.4\%$.

What did we do? Updated **prior** probability $P(C = 1)$ based on observing **data** (mammograms) to get the **posterior** probability $P(C = 1|M = 1)$.

Bayesian inference about parameters

Happiness example

Hoff (Chapter 3):

- Each female aged 65+ in 1998 General Social Survey was asked about being happy.
- Data: Out of $n = 129$ women, $y = 118$ women (91%) reported being happy.
- What is $\theta =$ the proportion of 65+ women who are happy?
- Goal: inference about $\theta =$ happiness parameter.

Happiness example

What's our usual approach? (frequentist)

1. Relate data to parameter of interest through a likelihood function, e.g. assume $Y|\theta \sim Bin(n, \theta)$ where y is the number of women who report to be happy out of the sample of n women.
2. Maximum likelihood estimate: Find a point estimate θ that maximizes the likelihood function ($\hat{\theta} = 0.91$)
3. Construct a confidence interval for θ (CI: [0.87, 0.96])
4. Interpretation of frequentist CI: If repeated samples were taken and the 95% confidence interval was computed for each sample, 95% of the intervals would contain the population mean.

Happiness example

The Bayesian approach:

- Also assume a likelihood, as before $Y|\theta \sim Bin(n, \theta)$

But now we proceed differently. In Bayesian inference, unknown parameters (like θ) are considered **random variables**. This means information/knowledge about these random variables can be summarized using probability distributions.

- Have existing knowledge/info about θ, summarized by the prior probability distribution
- Observe some data that gives more info about θ
- Update our previous knowledge to obtain the posterior distribution using Bayes' rule

Happiness example

The Bayesian approach:

1. Also assume a likelihood $p(y|\theta)$, as before $Y|\theta \sim Bin(n, \theta)$
2. Set a prior distribution for θ, $p(\theta)$
3. Use Bayes rule to update the prior into the posterior distribution
$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$
4. Use the posterior to provide summaries of interest, e.g. point estimates and uncertainty intervals, called credible intervals.

Bayes rule for parameters

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- $p(\theta|y)$ is called the posterior distribution of θ: the probability distribution of parameter θ given observed data y
- $p(y|\theta)$ is the likelihood distribution: the probability distribution of data y given θ.
- $p(\theta)$ is the prior distribution of θ
- $p(y)$ is the marginal distribution of the data y

Happiness example

1. Likelihood is $Y|\theta \sim Bin(n, \theta)$ so

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

2. Now we need to pick a prior $p(\theta)$

- Suppose any outcome between 0 and 1 for θ is equally likely, what prior can be used to describe these beliefs?
- $\theta \sim U(0, 1)$ so $p(\theta) = 1$

3. Now we calculate the posterior distribution

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta')d\theta'}$$

Up to a constant

Turns out to work out the posterior we only to consider the terms that include θ

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

In the Happiness case, it turns out that the posterior distribution of θ is a Beta distribution.

Say it with me

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

The posterior is proportional to the likelihood times the prior

Inference about θ based on posterior distribution

Bayesian point estimates are often given by:

- The posterior mean $E(\theta|y)$
- The posterior median θ^* $P(\theta < \theta^*|y) = 0.5$.

Uncertainty is quantified with credible intervals (CIs), e.g. for 95% CIs:

- An interval is called a 95% Bayesian CI if the posterior probability that θ is contained in the interval is 0.95.

Happiness findings

```r
y <- 118 # number of successes in data set
n <- 129
# set a and b
# Use uniform prior, thus Beta(1,1)
a <- b <- 1

# mean
(a+y)/(a+b+n)
```

```
## [1] 0.9083969
```

```r
# Bayes quantile-based CI:
round(qbeta(c(0.025,0.975), y+a, n-y+b),2)
```

```
## [1] 0.85 0.95
```

```r
# extra: freq CI:
p <- y/n; round(p,2)
```

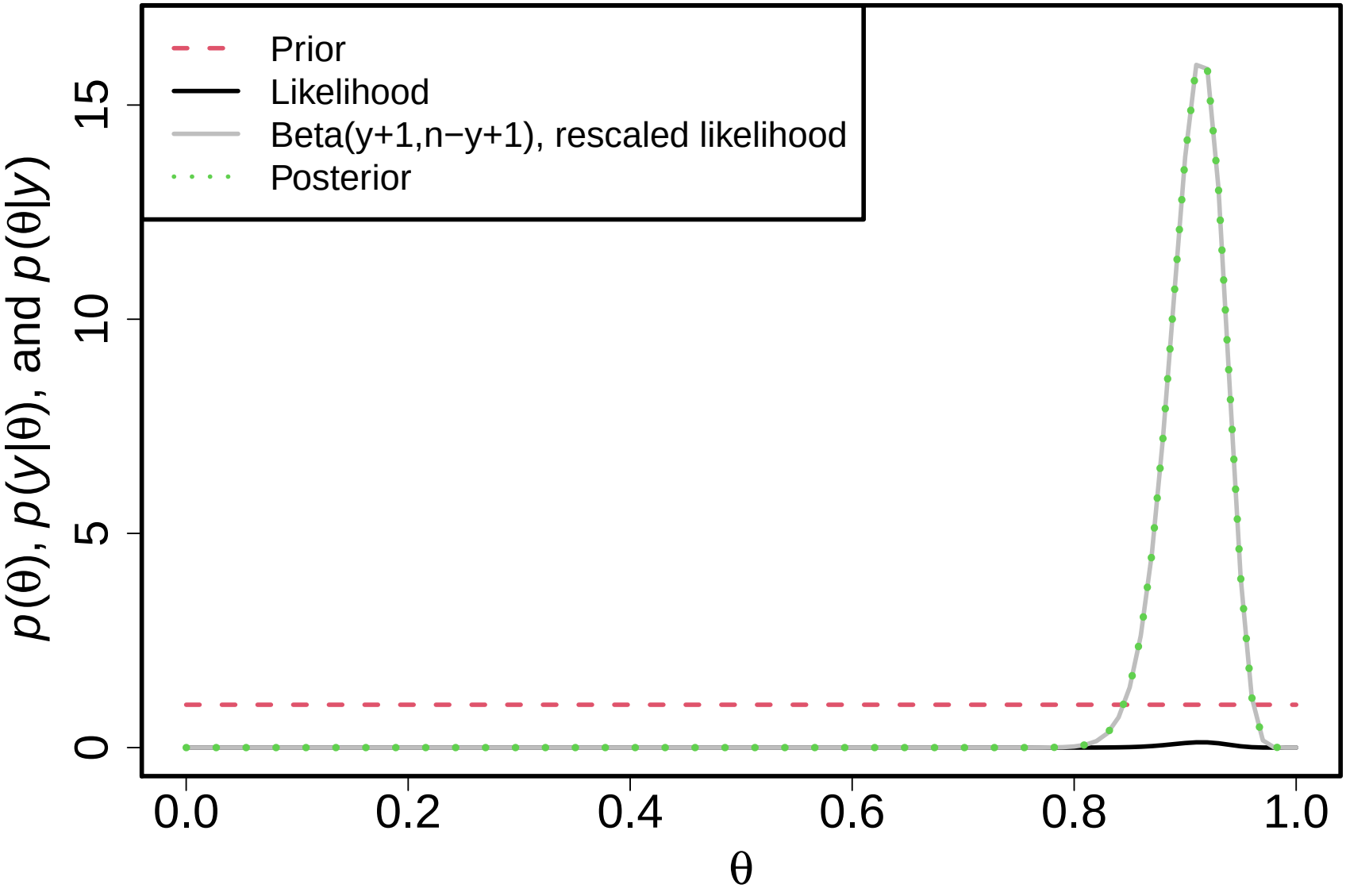
```
## [1] 0.91
```

```r
round(p + qnorm(0.975)*sqrt(p*(1-p)/n),2)
```

```
## [1] 0.96
```

```r
round(p - qnorm(0.975)*sqrt(p*(1-p)/n),2)
```

```
## [1] 0.87
```


Prior
Likelihood
Beta(y+1,n−y+1), rescaled likelihood
Posterior
p(θ), p(y|θ), and p(θ|y)
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
θ

Take-aways

- In the happiness case, we combined our prior knowledge (not much) with observed data to come up with a posterior estimate of the probability of being happy
- The key difference is that we treat parameters we are interested in as random, not fixed, which allows us to more intuitively make probability statements about the likely values

In the regression context

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

In the frequentist set-up:

- estimate $\hat{\beta}_0$, $\hat{\beta}_1$ using OLS
- Make assumptions about normality
- This allows us to write down sampling distribution for, say $\hat{beta}_1$
- This allows us to perform hypothesis testing about likely value of true β_1

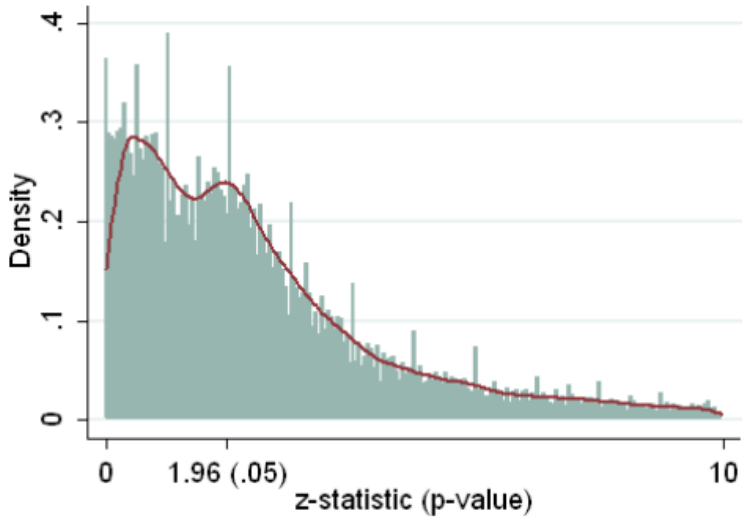
In the regression context

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

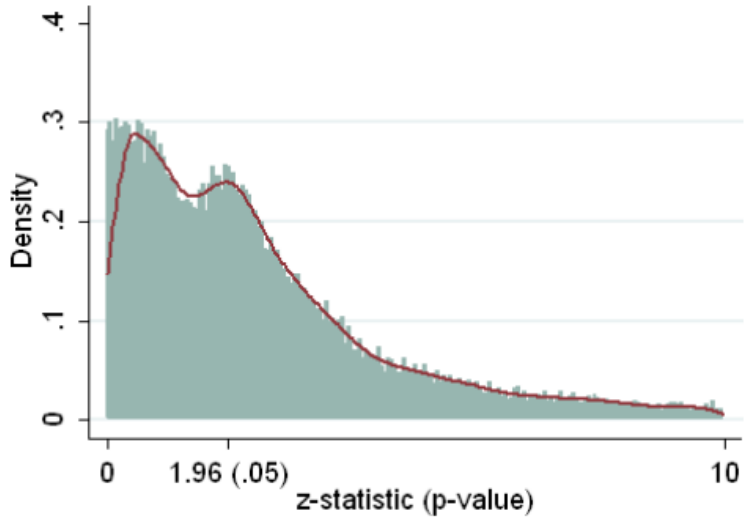
In Bayesian set-up:

- β_0 and β_1 are random variables
- We have some prior knowledge about their values, which we can encode in a prior probability distribution (if we don't know anything, then use non-informative priors)
- After seeing data, use Bayes rule to estimate posterior probability distribution of $\hat{\beta}_0$, $\hat{\beta}_1$
- Use this distribution to get values of expected value of β_1, variance of β_1, etc

Why go Bayes?



(a) Raw distribution of z-statistics.



(b) Unrounded distribution of z-statistics.

- Moving away from p-values, p-hacking (source of graph)
- Embracing uncertainty, model checking
- In my work: intuitive framework to deal with data sparsity, missing values, multiple data sources