Bayesian Data Analysis Session 1

Edwin Thoen

10/2/2017

Overview

Session 1: Edwin

What is Bayesian statistics? Theory and simple examples.

Session 2: Rick

Introduction MCMC and building hierarchical models with Stan.

Introduction

Intuition of Bayesian Statistics

A Statistician:

Describes the world in probality distributions, these distributions have parameters θ.

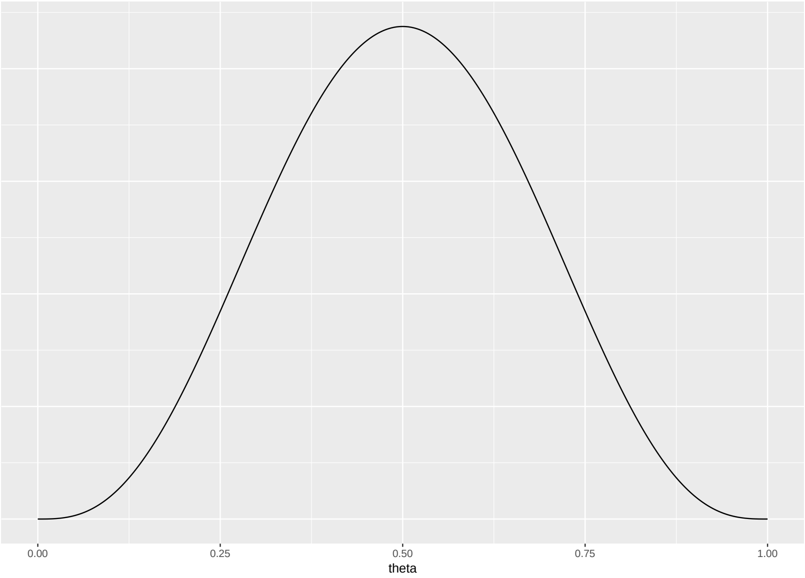
Collects data to learn about the distributions: $\hat{\theta}$.

How do we deal with uncertainty due to estimation?

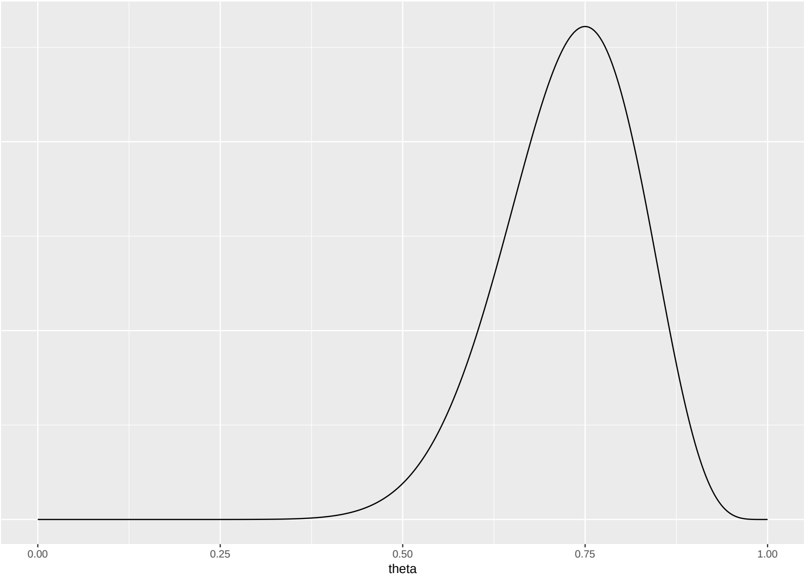
A Bayesian:

- Sets a probabilty distribution on all θ.
- Updates his beliefs with data.

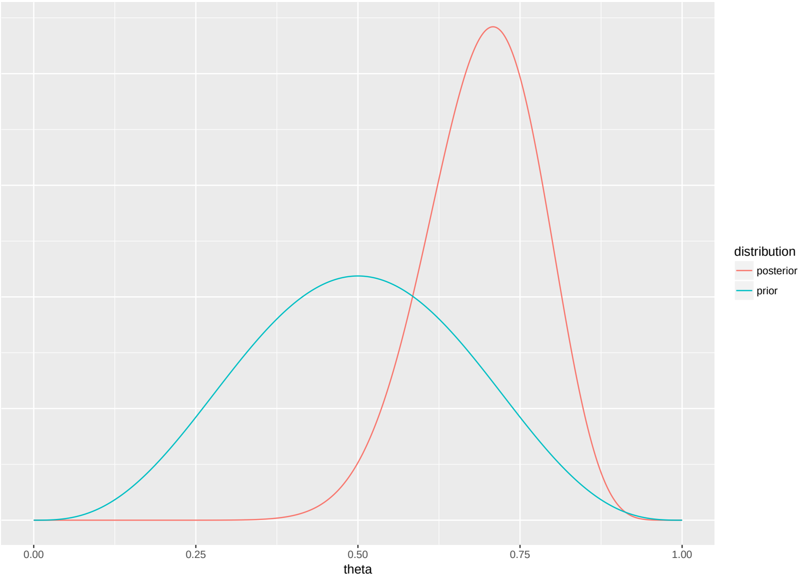
Set a prior: $P(\theta)$



Get the likelihood function: $P(D|\theta)$



Update prior to posterior with likelihood: $P(\theta|D)$



Likelihood not in this plot, on different scale (why?).

Bayesian data analysis

The essence of BDA is **credibility (re)allocation**.

We have an a priori idea about θ:

- expert opinion
- previous research
- educated guess

Data provides evidence of the parameter value.

The posterior is a compromise between prior and likelihood. It reflects the current knowledge.

Bayesian vs frequentist

- Frequentist only consider the likelihood.
- Frequentits have an objective view of probability. For Bayesians it is a subjective best guess.
- Frequentists: data random, parameters fixed. Bayesians: data fixed, parameters random.

Why do we want BDA in the first place?

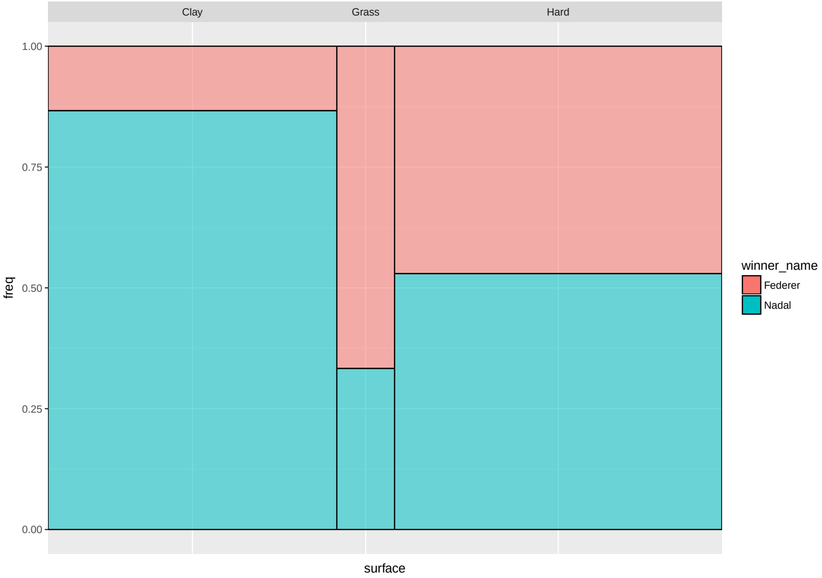
- Elegant and intuitive paradigm.
- Incorparation of previous knowledge and allowing for updating.
- Describe complex relationships without huge amounts of data.

Probability

Considered known

- sample space Ω.
- probability functions.
- discrete and continuous random variables.
- expected value and variance of random variables.
- from probability distribution to likelihood.

Probability of multiple events



Joints, marginals and conditionals

(We assume the probabilities here as given, not as estimated.)

The joint is the probability two events coincide. $P(A = a \cap B = b)$ or for brevity $P(A \cap B)$

winner_name	Clay	Grass	Hard
Federer	0.057	0.057	0.229
Nadal	0.371	0.029	0.257

Joints, marginals and conditionals

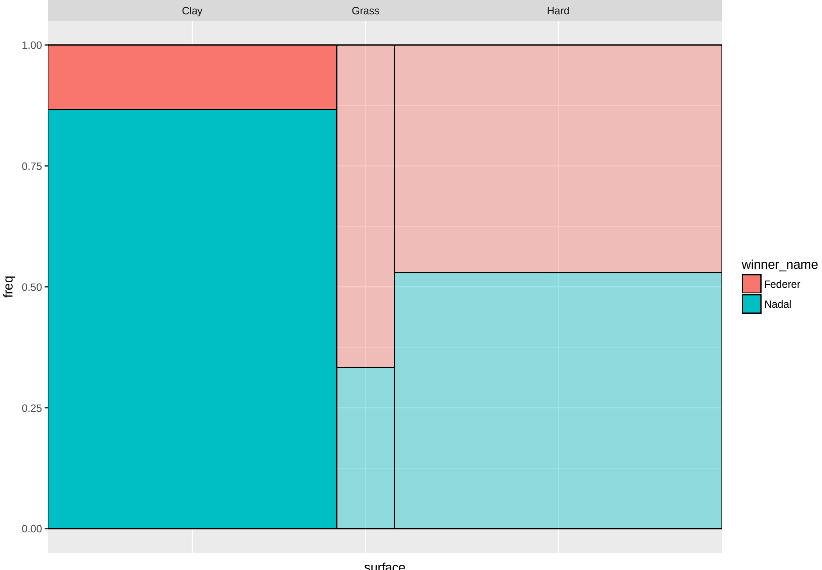
The marginals are the univariate distributions of A and B. Sum over the other (or integrate them out when continuous).

Clay	Grass	Hard
0.428	0.086	0.486

Federer	Nadal
0.343	0.657

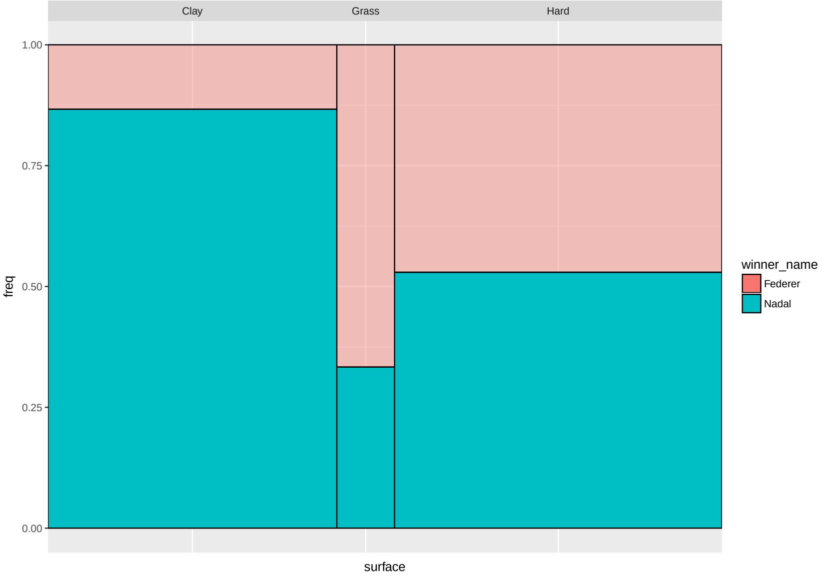
Joints, marginals and conditionals

Conditionals, like $P(A|B)$, redefine Ω, it is now a subset of the joint.
$P(A|B) = P(A \cap B)/P(B)$



Joints, marginals and conditionals

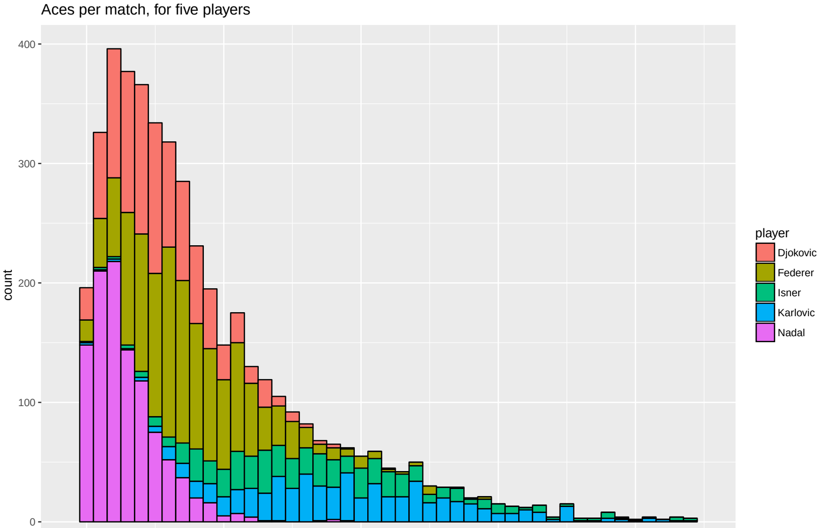
For $P(B|A)$ this looks



Joints, marginals and conditionals

With a continuous and discrete variable, a way to graph the data is

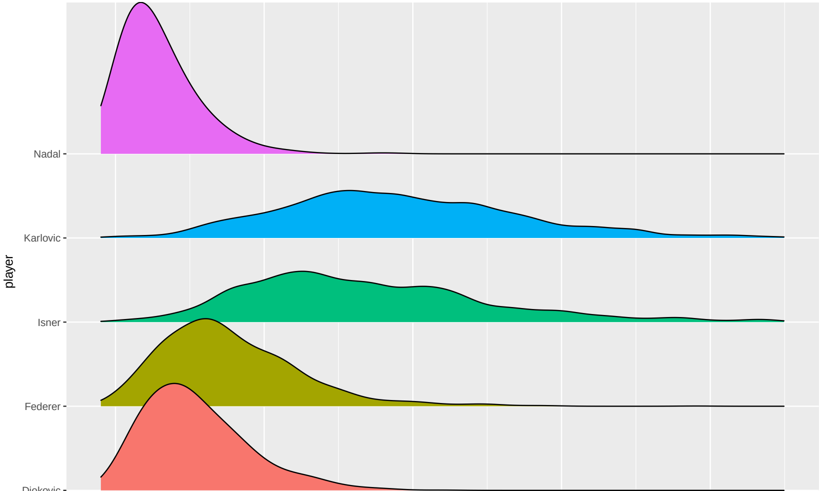
```
## Warning: Removed 11 rows containing non-finite values (s
```



Conditionals

$P(ace|player)$

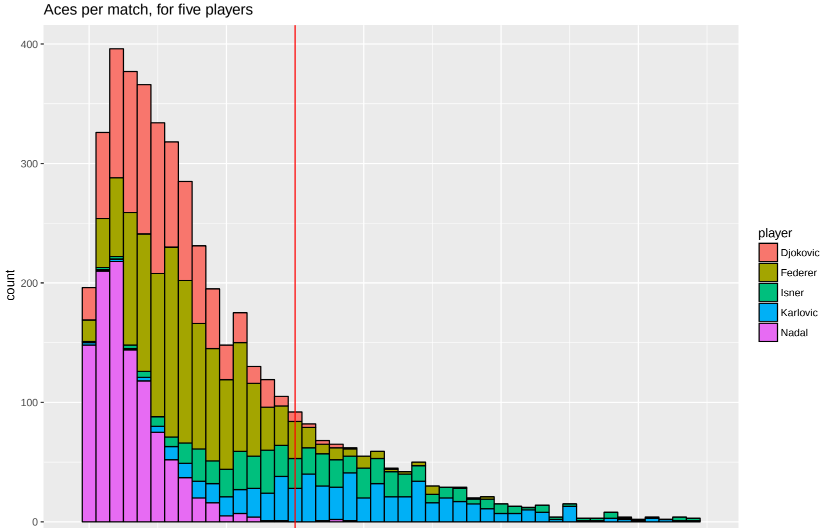
```
## Warning: Removed 11 rows containing non-finite values
## (stat_density_ridges).
```



Conditionals

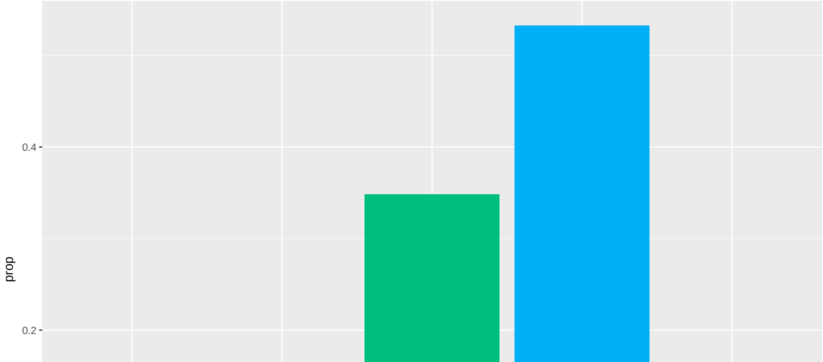
$P(player|ace > 15)$

```
## Warning: Removed 11 rows containing non-finite values (s
```

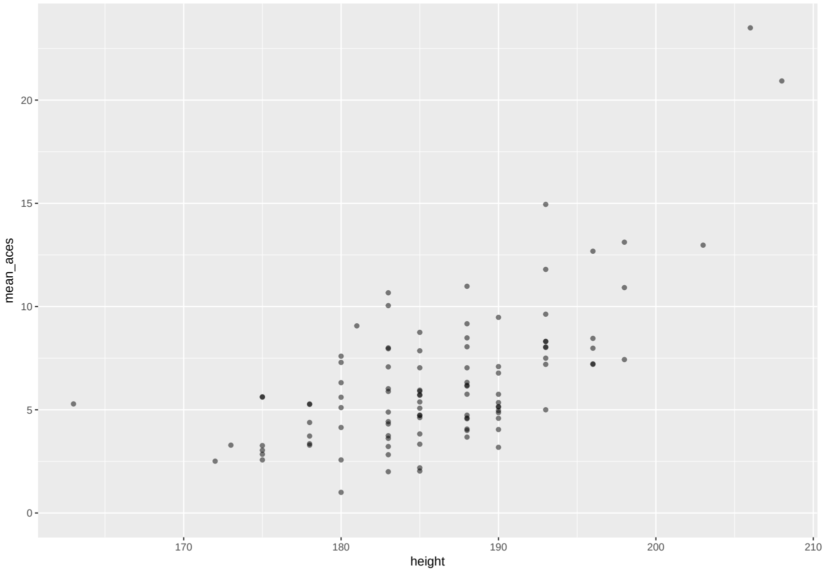


Conditionals

```
aces_set %>%
  filter(ace > 15) %>%
  count(player) %>%
  mutate(prop = n / sum(n)) %>%
  ggplot(aes(player, prop)) +
  geom_bar(aes(fill = player), stat = "identity") +
  guides(fill = FALSE)
```



Two continuous variables



Two continuous variables

The marginals are the univariate densities. With discrete variables we could sum over B to get marginal A. With continuous variables we need to integrate the other out.

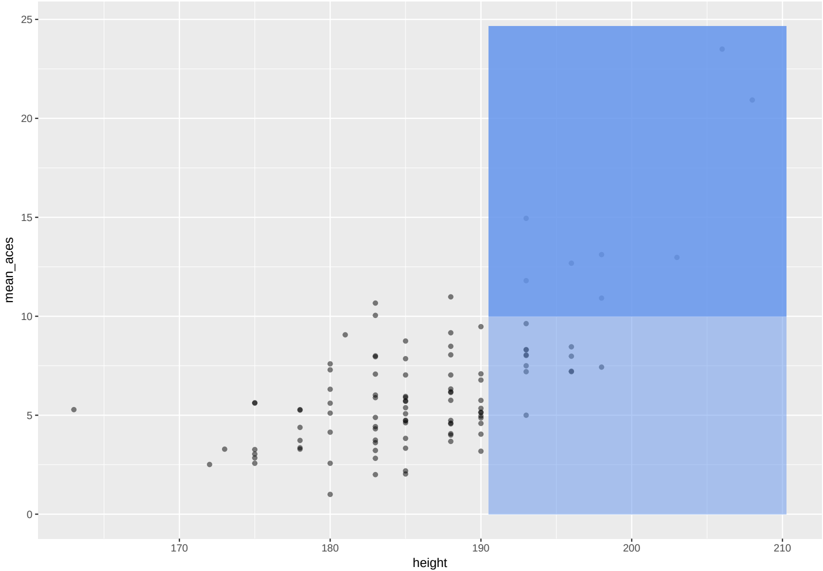
$P(A) = \int P(A, B) db$

Also here we normalize by the marginal to get the conditional.

$P(A|B) = P(A, B)/P(B)$

Two continuous variables

$P(a > 10 | h > 190)$



Bayes Rule

We all learned it in Introduction to Stats.

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

follows from combining

$P(A, B) = P(B|A)P(A)$

and

$P(A|B) = P(A, B)/P(B)$

Bayes Rule

Intuition:

- take the known conditional
- convert it to the joint by multiplying by the marginal
- obtain the desired conditional by dividing by the other marginal

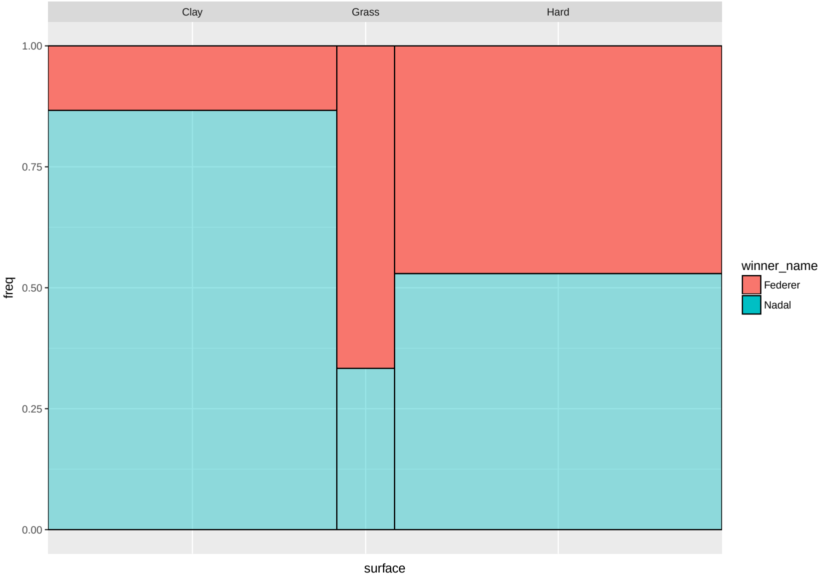
Bayes Rule

Example: Probability Federer wins the next match on Clay.

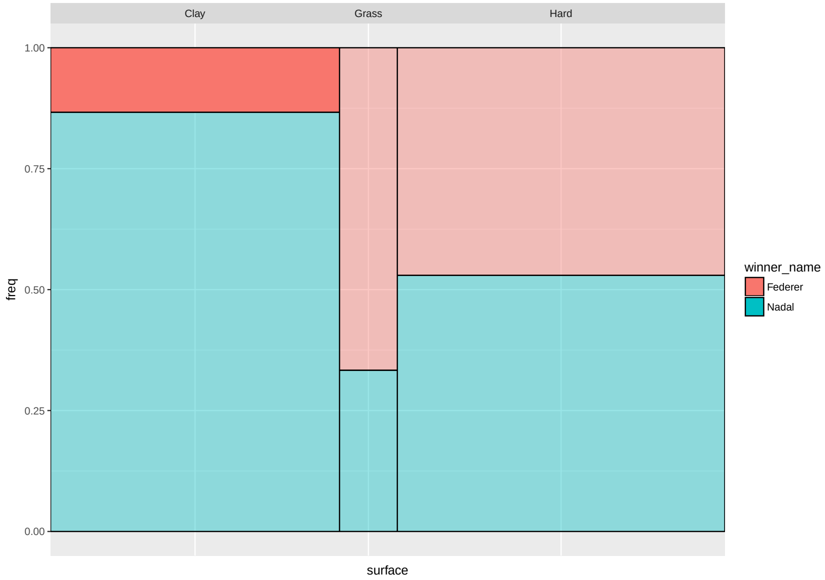
We only know: proportion Federer wins(.343), proportion of Federer wins were on Clay (.167), and proportion of matches played on clay (.429).

Bayes Rule

Going from the conditional to the joint:

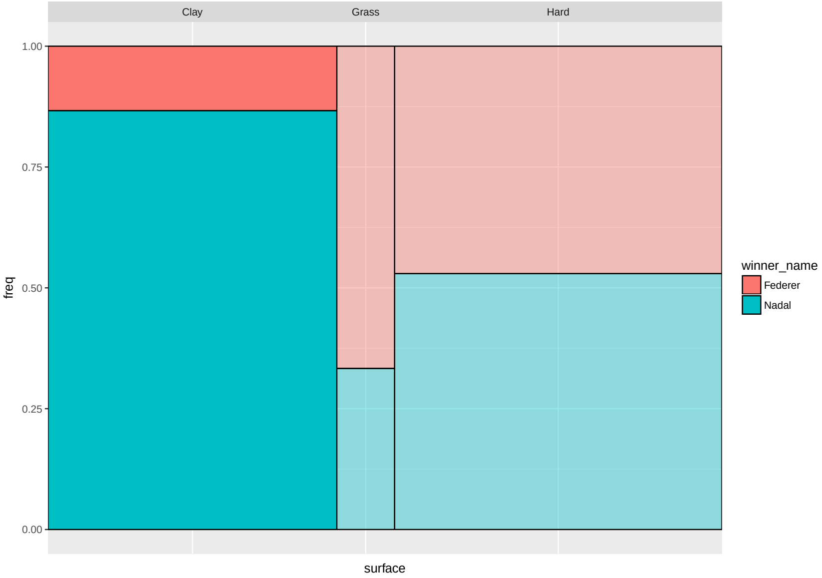


Bayes Rule



Clay
Grass
Hard
1.00
0.75
0.50
0.25
0.00
freq
surface
winner_name
Federer
Nadal

Bayes Rule



Clay
Grass
Hard
1.00
0.75
0.50
0.25
0.00
freq
surface
winner_name
Federer
Nadal

Bayes Rule

We can rewrite the marginal in the denominator for a discrete case as:

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

And for a continuous case this is:

$$P(A = a|B) = \frac{P(B|A = a)P(A = a)}{\int P(B|A)p(A)da}$$

Bayesian Analysis

Remember we put a prior distribution on a parameter.

We then use the data to obtain the likelihood.

We multiply the two into to obtain the posterior.

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

Proprotional because AUC is not equal to 1.

Bayesian Analysis

Now following Bayes Rule, to normalize to a probability distribution we have divide by the likelihood.

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

Which we can rewrite as

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$$

Note that in $P(\theta|D)$ and in the numerator we refer to a specific value of θ. In the demoninator it is a function of θ (remember summing over all the values of B).

Conjugate prior

We want to test if one player is better than the other. We can do this by estimating the bernoulli probability of player A beating player B.

Bernoulli density: $p(X = 1) = \theta^x (1 - \theta)^{(1-x)}$

makes the Bernoulli likelihood: $p(X = 1) = \theta^z (1 - \theta)^{(n-z)}$

where $z = \Sigma x = 1$ and n is number of observations.

Conjugate prior

What prior to set on this probability?

θ must be $[0,1]$, so distribution must be limited.

The beta distribution: $p(X=x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$

We are placing this distribution on θ, so $x=\theta$ in the above.

Conjugate prior

The denominator is just a normalizing constant, does not depend on θ

Thus:

$$P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

and

$$P(\theta|D) \propto \theta^{z}(1-\theta)^{n-z}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{z+\alpha-1}(1-\theta)^{n-z+\beta-1}$$

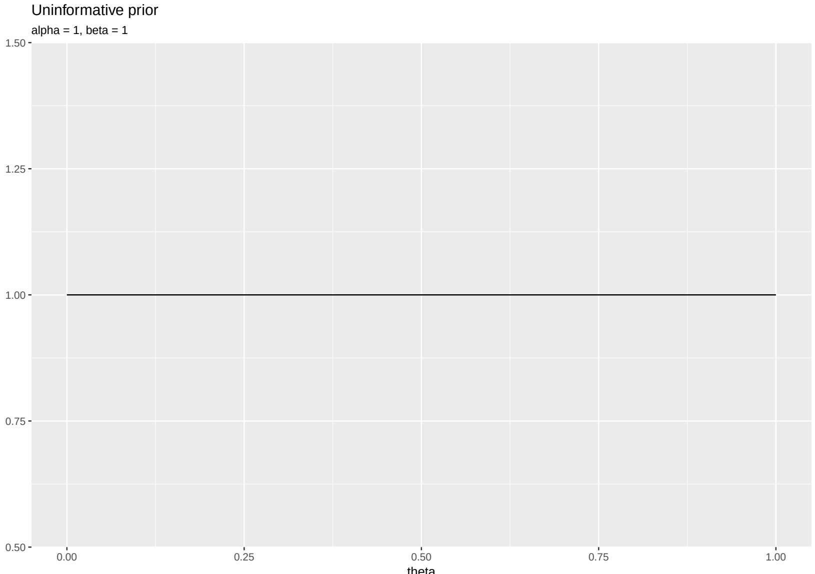
Note that this is again the denominator of a Beta, we use the same parameters to normalize.

The posterior is thus $Beta(z+\alpha-1, n-z+\beta-1)$.

In conjugacy the prior has the same functional form as the likelihood, and can be updated without solving the integral.

Conjugate prior

$\theta = P(Federer)$, no a priori idea. Set an uninformative prior, $Beta(1, 1)$

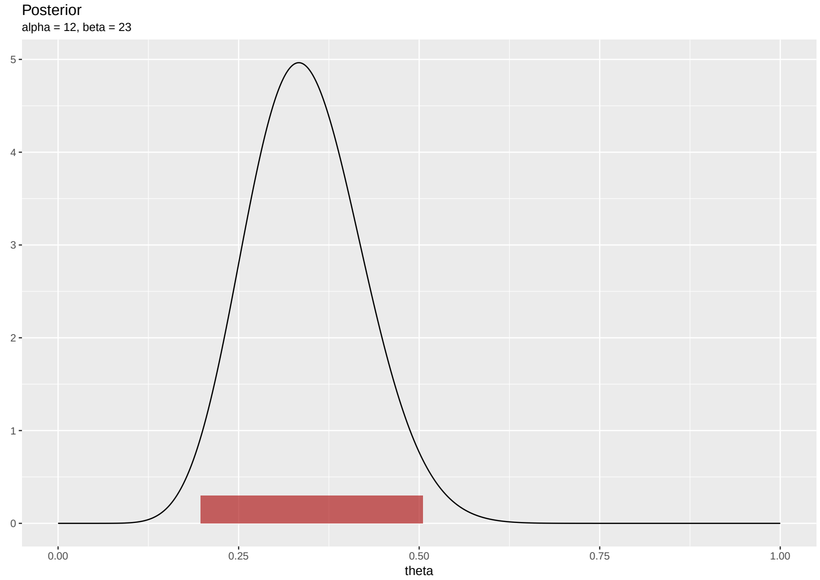


Conjugate prior

Federer won 12 out of 35 matches. So posterior is $Beta(12, 23)$.

HDI

95% highest posterior density interval.



The challenge of Bayesian estimation

Conjugacy makes it very easy to obtain posterior. However, only works for very simple situation.

Usually models involve many θs.

Integral in denominator cannot be solved.

Made Bayesian statistics a solely theoretical exercise for decades.

Normalizing by sampling

We can no longer normalize by the marginal, because of too complex integrals.

Draws from a function proportional to a distribution are the same as draws from the actual distribution.

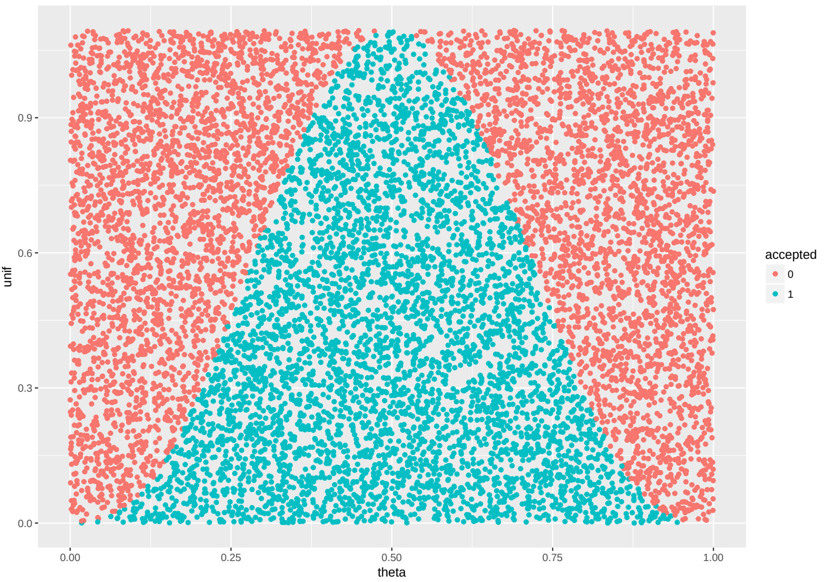
1. Draw many samples from the proportional distribution.
2. Calculate summary statistics, these descrice the posterior.

Acceptance-Rejection sampling

For a function $f(x)$

1. Determine the x-range.
2. Draw p samples from the x-range.
3. Draw p samples from $U(0, max(f(x)))$.
4. Compare 2. and 3. on index. If $f(2.) >= 3.$ accept, else reject.

Acceptance-Rejection sampling



Acceptance-Rejection sampling

```
stats <- acc_rej_data %>%
  filter(accepted == '1') %>%
  summarise(mn = mean(theta),
            sd = sd(theta))

est_beta <- function(mu, sd) {
  var   <- sd^2
  alpha <- ((1 - mu) / var - 1 / mu) * mu ^ 2
  beta  <- alpha * (1 / mu - 1)
  c(alpha = alpha, beta = beta)
}

est_beta(stats$mn, stats$sd) %>% round(3)
```

```
## alpha  beta
## 3.791 3.824
```