inference for a proportion - Bayesian approach

Duke UNIVERSITY

Dr. Mine Çetinkaya-Rundel

framework

- consider the 20 total pregnancies
  - **question**: how likely is it that 4 pregnancies occur in the treatment group?
- if treatment and control are equally effective + sample sizes for the two groups are the same

P(pregnancy comes from treatment group) $= p = 0.5$

hypotheses, i.e. models

- delineate plausible models:
  - assume p could be

10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, or 90%

- consider 9 models, instead of 1 as in the frequentist paradigm
  - $p = 20\%$: given a pregnancy occurs, there is a 2:8 or 1:4 chance that it will occur in the treatment group

specifying the prior

- prior probabilities reflect state of belief prior to the current experiment
- incorporate information learned from all relevant research up to the current point in time, but not incorporate information from the current experiment
- suppose my prior probability for each of the 9 models is as presented below:

Model (p)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
Prior	0.06	0.06	0.06	0.06	0.52	0.06	0.06	0.06	0.06	1

  - benefit of treatment is symmetric — equally likely to be better or worse than the standard treatment
  - 52% chance that there is no difference between the treatments

likelihood

- calculate $P(data \mid model)$ for each model considered.
- this probability is called the **likelihood**:

$$P(data \mid model) = P(k = 4 \mid n = 20, p)$$

calculating the likelihood

```r
p <- seq(from = 0.1, to = 0.9, by = 0.1)
prior <- c(rep(0.06, 4), 0.52, rep(0.06, 4))
likelihood <- dbinom(4, size = 20, prob = p)
```

Model (p)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
Prior, $P(model)$	0.06	0.06	0.06	0.06	0.52	0.06	0.06	0.06	0.06	1
Likelihood, $P(data \mid model)$	0.0898	0.2182	0.1304	0.035	0.0046	0.0003	0	0	0	

posterior

use Bayes' rule to calculate the posterior probability, i.e. *P(model | data)*

$$P(model \mid data) = \frac{P(model \;\&\; data)}{P(data)}$$

$$= \frac{P(data \mid model) \times P(model)}{P(data)}$$

calculating the posterior

```r
numerator <- prior * likelihood
denominator <- sum(numerator)
posterior <- numerator / denominator
sum(posterior)
```

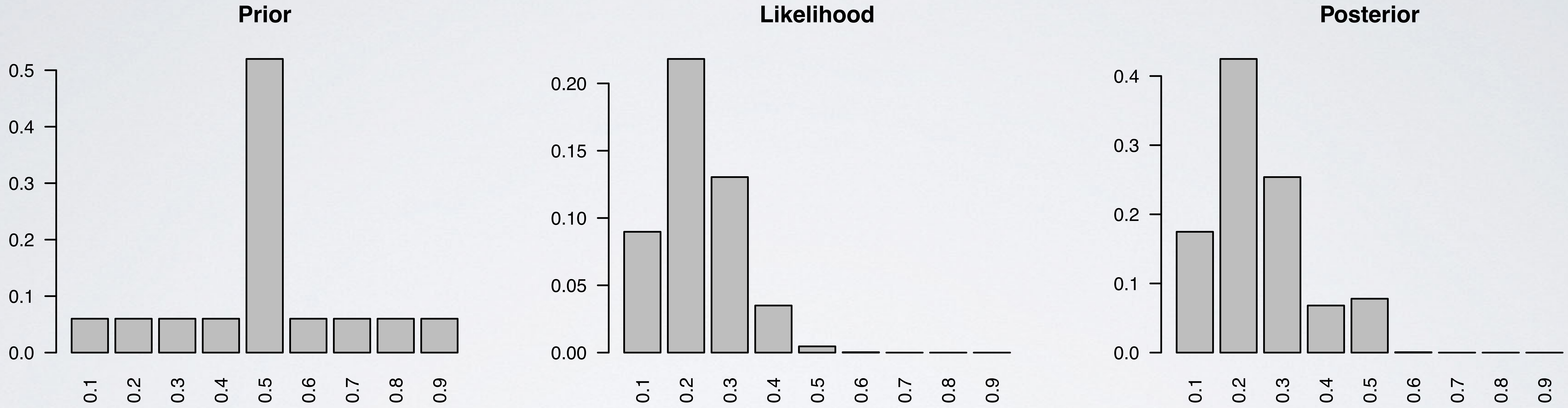
```
## [1] 1
```

Model (p)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
Prior, $P(model)$	0.06	0.06	0.06	0.06	0.52	0.06	0.06	0.06	0.06	1
Likelihood, $P(data \mid model)$	0.0898	0.2182	0.1304	0.035	0.0046	0.0003	0	0	0	
$P(data \mid model) \times P(model)$	0.0054	0.0131	0.0078	0.0021	0.0024	0	0	0	0	0.0308
Posterior, $P(model \mid data)$	0.1748	0.4248	0.2539	0.0681	0.0780	0.0005	0	0	0	1

decision making

- posterior probability that $p = 0.2$ is 42.48%
  - this model has the highest posterior probability
- calculation of the posterior incorporated prior information and likelihood of data observed
  - data "at least as extreme as observed" plays no part in the Bayesian paradigm
- note that probability that $p = 0.5$ dropped from 52% in the prior to about 7.8% in the posterior
  - this demonstrates how we update our beliefs based on observed data

prior, likelihood, and posterior, visualized



synthesis

- Bayesian paradigm allows us to make direct probability statements about our models
- we can also calculate the probability that RU-486 (the treatment) is more effective than the control
  - this is the sum of the posteriors of the models where $p < 0.5$

Model (p)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
Posterior, $P(model\|data)$	0.1748	0.4248	0.2539	0.0681	0.0780	0.0005	0	0	0	1

0.9216