Lecture 3.2: Sampling + Introduction to Bayesian Statistics

2014/02/12

R Homework Question 1

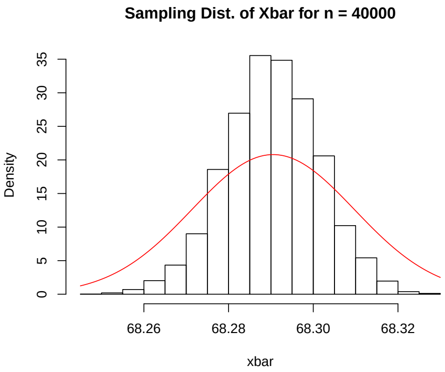
- The red curve is the theoretical sampling distribution of $\overline{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$

R Homework Question 1

- The red curve is the theoretical sampling distribution of $\overline{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$
- The histogram is `num.sim=10000` simulations of the experiment:
  - Sample $n = 40000$ values from the population
  - Compute $\overline{x}$

R Homework Question 1

- The red curve is the theoretical sampling distribution of $\overline{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$
- The histogram is `num.sim=10000` simulations of the experiment:
  - Sample $n = 40000$ values from the population
  - Compute $\overline{x}$



Independence of the Sample

The results that for $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$, regardless of n:

Independence of the Sample

The results that for $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$, regardless of n:

- $\overline{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$

Independence of the Sample

The results that for $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$, regardless of n:

- $\overline{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$
- $\frac{(n-1)}{\sigma^2} S^2 \sim \chi_{n-1}$

Independence of the Sample

The results that for $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$, regardless of n:

- $\overline{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$
- $\frac{(n-1)}{\sigma^2} S^2 \sim \chi_{n-1}$

assumes independence of the sample.

Independence of the Sample

The results that for $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$, regardless of n:

- $\overline{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$
- $\frac{(n-1)}{\sigma^2} S^2 \sim \chi_{n-1}$

assumes independence of the sample. Also implicit is that the population is infinite.

Independence of the Sample

The results that for $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$, regardless of n:

- $\overline{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$
- $\frac{(n-1)}{\sigma^2} S^2 \sim \chi_{n-1}$

assumes independence of the sample. Also implicit is that the population is infinite.

However, in many real-life settings with finite populations, once we've sampled someone / something, we don't typically put them back into the pool of potential samples. Think of a survey.

Independence of the Sample

The results that for $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$, regardless of n:

- $\overline{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$
- $\frac{(n-1)}{\sigma^2} S^2 \sim \chi_{n-1}$

assumes independence of the sample. Also implicit is that the population is infinite.

However, in many real-life settings with finite populations, once we've sampled someone / something, we don't typically put them back into the pool of potential samples. Think of a survey.

i.e. sampling with replacement vs sampling without replacement.

Independence of the Sample

Say we are interested in the probability of sampling Wayne and Mario. For independence to hold we need

$$P(\text{ Wayne \& Mario }) = P(\text{ Wayne }) \times P(\text{ Mario })$$

Independence of the Sample

Say we are interested in the probability of sampling Wayne and Mario. For independence to hold we need

$$P(\text{ Wayne \& Mario }) \;=\; P(\text{ Wayne }) \times P(\text{ Mario })$$

also, by conditional independence

$$P(\text{ Wayne \& Mario }) \;=\; P(\text{ Mario }) \times P(\text{ Wayne } \mid \text{ Mario })$$

Independence of the Sample

Say we are interested in the probability of sampling Wayne and Mario. For independence to hold we need

$$P(\text{ Wayne \& Mario }) \quad = \quad P(\text{ Wayne }) \times P(\text{ Mario })$$

also, by conditional independence

$$P(\text{ Wayne \& Mario }) \quad = \quad P(\text{ Mario }) \times P(\text{ Wayne } \mid \text{ Mario })$$

Compare $N = 4$ & $N = 10000$ and assume WLOG we pick Mario first.

Independence of the Sample

Say we are interested in the probability of sampling Wayne and Mario. For independence to hold we need

$$P(\text{ Wayne \& Mario }) \;=\; P(\text{ Wayne }) \times P(\text{ Mario })$$

also, by conditional independence

$$P(\text{ Wayne \& Mario }) \;=\; P(\text{ Mario }) \times P(\text{ Wayne } | \text{ Mario })$$

Compare $N = 4$ & $N = 10000$ and assume WLOG we pick Mario first.

$$P(\text{ Wayne \& Mario }) \;=\; \frac{1}{4} \times \frac{1}{3}$$

Independence of the Sample

Say we are interested in the probability of sampling Wayne and Mario. For independence to hold we need

$$P(\text{ Wayne \& Mario }) = P(\text{ Wayne }) \times P(\text{ Mario })$$

also, by conditional independence

$$P(\text{ Wayne \& Mario }) = P(\text{ Mario }) \times P(\text{ Wayne } \mid \text{ Mario })$$

Compare $N = 4$ & $N = 10000$ and assume WLOG we pick Mario first.

$$P(\text{ Wayne \& Mario }) = \frac{1}{4} \times \frac{1}{3} \neq \frac{1}{4} \times \frac{1}{4}$$

Independence of the Sample

Say we are interested in the probability of sampling Wayne and Mario. For independence to hold we need

$$P(\text{ Wayne \& Mario }) \quad = \quad P(\text{ Wayne }) \times P(\text{ Mario })$$

also, by conditional independence

$$P(\text{ Wayne \& Mario }) \quad = \quad P(\text{ Mario }) \times P(\text{ Wayne } | \text{ Mario })$$

Compare $N = 4$ & $N = 10000$ and assume WLOG we pick Mario first.

$$\begin{aligned} P(\text{ Wayne \& Mario }) &= \frac{1}{4} \times \frac{1}{3} \neq \frac{1}{4} \times \frac{1}{4} \\ P(\text{ Wayne \& Mario }) &= \frac{1}{10000} \times \frac{1}{9999} \end{aligned}$$

Independence of the Sample

Say we are interested in the probability of sampling Wayne and Mario. For independence to hold we need

$$P(\text{ Wayne \& Mario }) = P(\text{ Wayne }) \times P(\text{ Mario })$$

also, by conditional independence

$$P(\text{ Wayne \& Mario }) = P(\text{ Mario }) \times P(\text{ Wayne } \mid \text{ Mario })$$

Compare $N = 4$ & $N = 10000$ and assume WLOG we pick Mario first.

$$\begin{aligned} P(\text{ Wayne \& Mario }) &= \frac{1}{4} \times \frac{1}{3} \neq \frac{1}{4} \times \frac{1}{4} \\ P(\text{ Wayne \& Mario }) &= \frac{1}{10000} \times \frac{1}{9999} \approx \frac{1}{10000} \times \frac{1}{10000} \end{aligned}$$

Independence of the Sample

The finite population correction factor (FPC) to the SE of $\overline{X}$ accounts for this issue:

$$SE_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

where

Independence of the Sample

The finite population correction factor (FPC) to the SE of $\overline{X}$ accounts for this issue:

$$SE_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

where

- N is the study population size
- n is the sample size

Independence of the Sample

The finite population correction factor (FPC) to the SE of $\overline{X}$ accounts for this issue:

$$SE_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

where

- N is the study population size
- n is the sample size

Say $N = 10000$.

- For $n = 100$

$$\sqrt{\frac{10000-100}{10000-1}} = 0.99$$

- For $n = 9000$

$$\sqrt{\frac{10000-9000}{10000-1}} = 0.32$$

Independence of the Sample

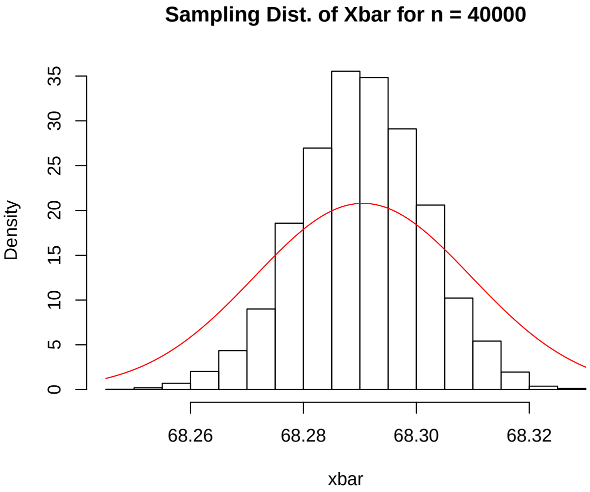
- So while sampling without replacement still yields an unbiased estimate of μ
- The SE is off, therefore confidence intervals and hypothesis tests will be off.

Independence of the Sample

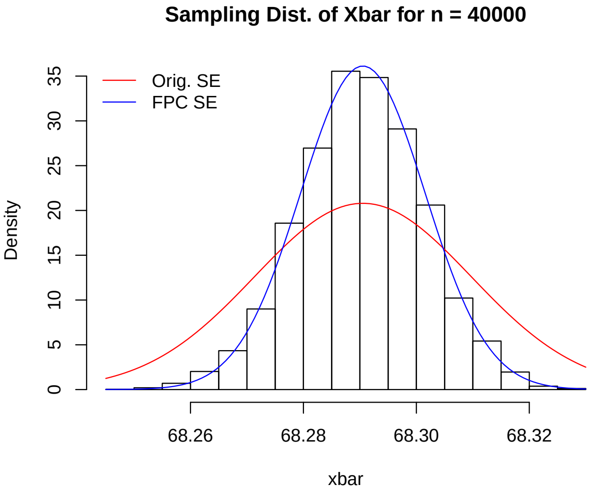
- So while sampling without replacement still yields an unbiased estimate of μ
- The SE is off, therefore confidence intervals and hypothesis tests will be off.

Capping n to be less than 10% of the population is a rule of thumb for keeping the correction factor “close” to 1.

R Homework Question 1



R Homework Question 1



Independence of the Sample

Using this fact, we can tie the conceptual notion and the mathematical notion of sampling:

Independence of the Sample

Using this fact, we can tie the conceptual notion and the mathematical notion of sampling:

- Conceptual: If we sample everybody in our study population, then we don't need to use statistics because we know exactly what the true μ is.

Independence of the Sample

Using this fact, we can tie the conceptual notion and the mathematical notion of sampling:

- Conceptual: If we sample everybody in our study population, then we don't need to use statistics because we know exactly what the true μ is.
- Mathematical: If $n = N$, then $FPC = \sqrt{\frac{N-n}{N-1}} = 0$, hence $SE_{\overline{X}} = 0$.

Independence of the Sample

Using this fact, we can tie the conceptual notion and the mathematical notion of sampling:

- Conceptual: If we sample everybody in our study population, then we don't need to use statistics because we know exactly what the true μ is.
- Mathematical: If $n = N$, then $FPC = \sqrt{\frac{N-n}{N-1}} = 0$, hence $SE_{\overline{X}} = 0$. i.e. there is no variability in our sampling procedure. If we repeat the procedure many times, we get the exact same value every time: the true μ.

Bayesian Statistics

Switching Gears: Thomas Bayes

Philosophical Schools of Thought (My Understanding)

Philosophy for dummies:

Philosophical Schools of Thought (My Understanding)

Philosophy for dummies:

- (Logical) Positivism: philosophy of science based on the view that information derived from logical and mathematical treatments and reports of sensory experience is the exclusive source of all authoritative knowledge,

Philosophical Schools of Thought (My Understanding)

Philosophy for dummies:

- (Logical) Positivism: philosophy of science based on the view that information derived from logical and mathematical treatments and reports of sensory experience is the exclusive source of all authoritative knowledge, and that there is valid knowledge (truth) only in scientific knowledge.

Philosophical Schools of Thought (My Understanding)

Philosophy for dummies:

- (Logical) Positivism: philosophy of science based on the view that information derived from logical and mathematical treatments and reports of sensory experience is the exclusive source of all authoritative knowledge, and that there is valid knowledge (truth) only in scientific knowledge. The only truth is what we can observe.

Philosophical Schools of Thought (My Understanding)

Philosophy for dummies:

- (Logical) Positivism: philosophy of science based on the view that information derived from logical and mathematical treatments and reports of sensory experience is the exclusive source of all authoritative knowledge, and that there is valid knowledge (truth) only in scientific knowledge. The only truth is what we can observe.
- Subjectivism: the philosophical tenet that "our own mental activity is the only unquestionable fact of our experience"

Probability (Backbone of Statistics)

What does probability even mean?

Probability (Backbone of Statistics)

What does probability even mean?

- Positivist View - Frequentist Probability: it defines an event's probability as the limit of its relative frequency in a large observable number of trials.

Probability (Backbone of Statistics)

What does probability even mean?

- Positivist View - Frequentist Probability: it defines an event's probability as the limit of its relative frequency in a large observable number of trials.
- Subjectivist View - Bayesian Probability: a subjective status by regarding it as a measure of the degree of belief of the individual assessing the uncertainty of a particular situation.

Statistics In General

(Parametric) statistics is inferring about an unknown parameter θ.

Statistics In General

(Parametric) statistics is inferring about an unknown parameter θ.

- Frequentist Statistics: the true θ is a single value that if we had an infinite sample size, we can compute it exactly. This has been the predominant view of statistics for a long time.

Statistics In General

(Parametric) statistics is inferring about an unknown parameter θ.

- Frequentist Statistics: the true θ is a single value that if we had an infinite sample size, we can compute it exactly. This has been the predominant view of statistics for a long time.
- Bayesian Statistics: the true θ is a distribution of values that reflects our belief in the plausibility of different values.

Specific Example

Concrete example: the probability of flipping heads

Specific Example

Concrete example: the probability of flipping heads

- Frequentist Statistics: the true probability p is a single value

Specific Example

Concrete example: the probability of flipping heads

- Frequentist Statistics: the true probability p is a single value
- Bayesian probability: the true probability p is a distribution of values

The Bayesian Procedure

To express our belief about θ from as a Bayesian, we have:

The Bayesian Procedure

To express our belief about θ from as a Bayesian, we have:

1. A prior distribution $\Pr(\theta)$. It reflects our prior belief about θ.

The Bayesian Procedure

To express our belief about θ from as a Bayesian, we have:

1. A prior distribution $\Pr(\theta)$. It reflects our prior belief about θ.
2. The likelihood function $\Pr(X|\theta)\,(= L(\theta|X))$. This is the mechanism that generates the data.

The Bayesian Procedure

To express our belief about θ from as a Bayesian, we have:

1. A prior distribution $\Pr(\theta)$. It reflects our prior belief about θ.
2. The likelihood function $\Pr(X|\theta)\,(= L(\theta|X))$. This is the mechanism that generates the data.
3. A posterior distribution $\Pr(\theta|X)$. We update our belief about θ after observing data X.

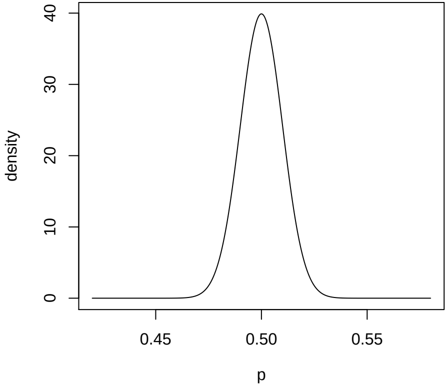
$$\Pr(\theta|X) = \frac{\Pr(X|\theta) \cdot \Pr(\theta)}{\Pr(X)}$$

The Issue: The Bayesian Procedure

Where do you come up with $\Pr(\theta)$? It's completely subjective! You decide!

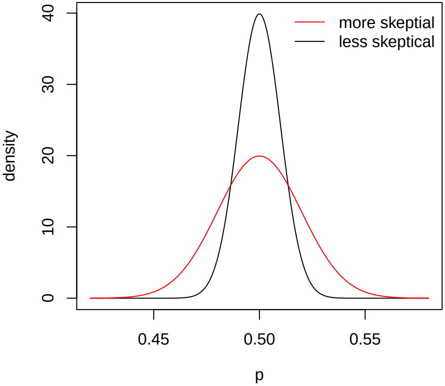
Prior Distribution

This distribution can reflect someone's prior belief of p.



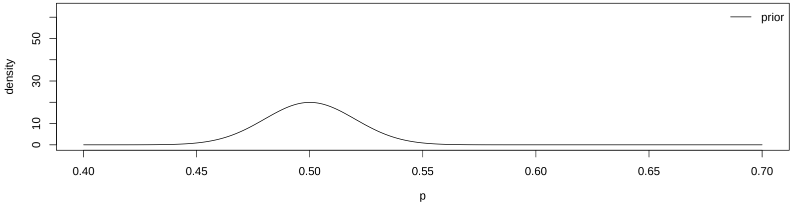
Prior Distribution

Say someone is more skeptical that $p = 0.5$, we can lower it.



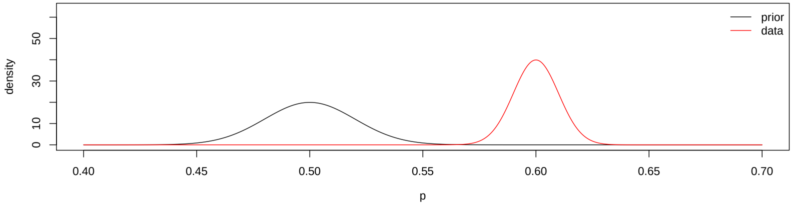
The Bayesian Procedure

Say we have the following prior belief centered at $p = 0.5$



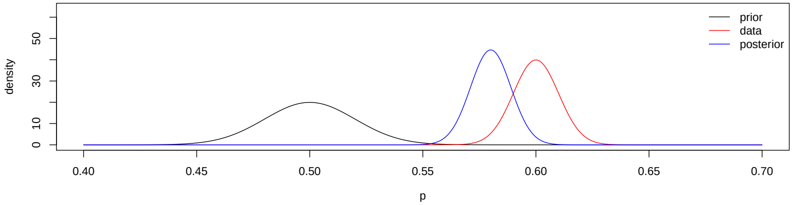
The Bayesian Procedure

Say we collect data, represented by the red line, suggesting $p = 0.6$



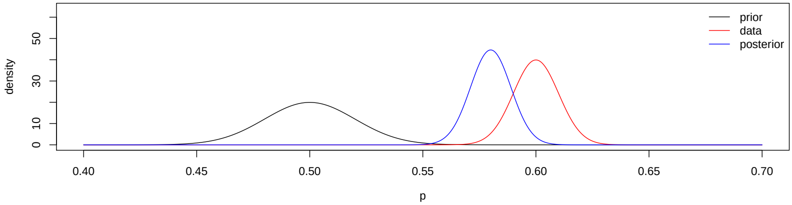
The Bayesian Procedure

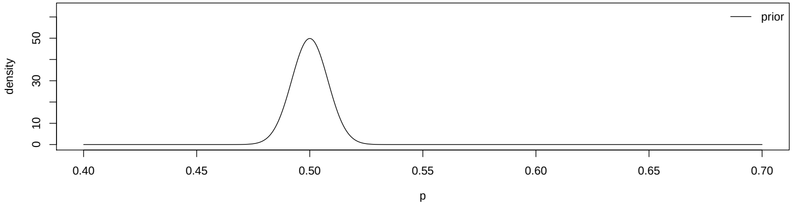
We then update our belief, as reflected in the posterior distribution!



The Bayesian Procedure

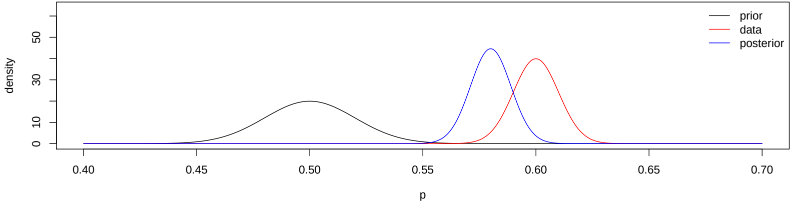
Now say we have a stronger prior belief that $p = 0.5$

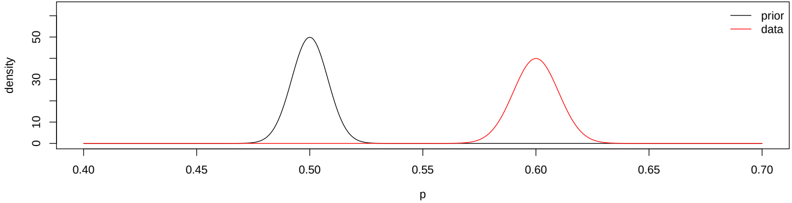




The Bayesian Procedure

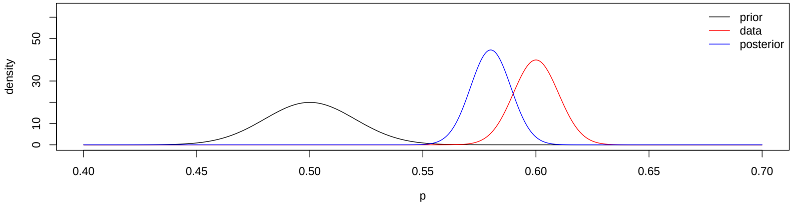
Say we observed the same data (as represented in red).

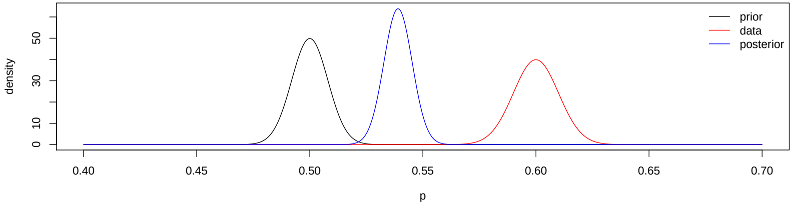




The Bayesian Procedure

The posterior in this case is pulled left due to the sharper prior.





Back to Debate

Frequentists believe statistics should be completely objective and therefore do not accept the premise of a subjective prior.

Back to Debate

Frequentists believe statistics should be completely objective and therefore do not accept the premise of a subjective prior.

The debate rages on...