Homework 3

Bayesian Statistics

Jaiden Atterbury

02-20-2023

Variance of Sample Mean from Finite Population vs. Infinite Population

1. Consider the sample mean $\bar{X}$ obtained from an iid sample of size n from a Normal (μ, σ^2) (infinite) population.

(a) Using known results, write down $E[\bar{X}]$ and $Var[\bar{X}]$.

The $E[\bar{X}]$ and $Var[\bar{X}]$ follow closely from results for the expected value and variance of linear combinations of independent and identically distributed random variables, we will derive $E[\bar{X}]$ and $Var[\bar{X}]$ below.

If $X_i \sim Norm(\mu, \sigma^2)$, consider the sample mean $\bar{X}$ obtained from an iid sample of size n, then $E[\bar{X}]$ and $Var[\bar{X}]$ are:

$$\begin{aligned}
E[\bar{X}] &= E[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)] \\
&= E[\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n] \\
&= \frac{1}{n}E[X_1] + \frac{1}{n}E[X_2] + \cdots + \frac{1}{n}E[X_n] \quad \text{(Linearity of expectation)} \\
&= \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu \quad \text{(Since } E[X_i] = \mu\text{)} \\
&= \mu
\end{aligned}$$

$$\begin{aligned}
Var[\bar{X}] &= Var[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)] \\
&= Var[\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n] \\
&= \frac{1}{n^2}Var[X_1] + \frac{1}{n^2}Var[X_2] + \cdots + \frac{1}{n^2}Var[X_n] \quad \text{(Independence of the } X_i\text{'s)} \\
&= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 \quad \text{(Since } Var[X_i] = \sigma^2\text{)} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Using R or another language, set a random seed and then draw 100K samples of size $n = 1024$, from a Normal population with $\mu = 0$ and $\sigma^2 = 25$. Then compute the sample mean $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_i$ for each sample.

To draw 100K samples of size $n = 1024$, from a Normal population with $\mu = 0$ and $\sigma^2 = 25$, we will use rnorm and the replicate function.

```
set.seed(123)

draws = 100000
n = 1024

norm_samples <- tibble(
  sample_means = replicate(draws, expr = mean(rnorm(n, mean = 0, sd = 5)))
)
```

(b) Report the random seed that you used. Then compute the sample mean and sample variance of the 100K sample means. Use this to confirm your answer from (a).

The seed used above when making the 100K samples and calculating the mean for each sample was 123. We will use R to calculate the sample variance and mean of the 100K sample means, then use these results to confirm our answer from (a).

```
mean_samples <- mean(norm_samples$sample_means)
var_samples <- var(norm_samples$sample_means)
```

As calculated above, the sample mean of the 100K sample means was 8.0773869×10^{-4}, and the sample variance of the 100K sample means was 0.0244666. In this example, each sample $X_i \sim Norm(\mu = 0, \sigma^2 = 25)$ with $n = 1024$. Thus $E[\bar{X}] = 0$ and $Var[\bar{X}] = \frac{25}{1024} = 0.024414$. As can be seen from the following results, our sample mean and sample variance of our 100K sample means confirm our answer from part (a).

Read in the data in this file finite-population-size-1200.csv. View this as a fixed finite population of size $N = 1200$.

```
finite_pop <- read_csv("finite-population-size-1200.csv")
```

(c) Using the data from the file, find the finite population mean μ and variance σ^2:

$$\mu \equiv \frac{1}{N}\sum_{i=1}^{N} y_i \qquad \sigma^2 \equiv \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$$

where the population is $y_1, \dots, y_N$.

Hint: σ^2 can be computed in R via a simple adjustment to the sample variance function var; see also comments regarding "the denominator" in help(var) documentation.

To find the finite mean and variance according to the above formulas, we will have to make a simple adjustment to the sample variance function in R. In particular, this adjustment will be of the form sample variance × $\frac{N-1}{N}$. The code that performs this is show below.

```
N = nrow(finite_pop)

finite_mean <- mean(finite_pop$y)
finite_var <- var(finite_pop$y) * ((N - 1)/N)
```

As calculated above, the finite population mean is $4.4489875 \times 10^{-17}$, and the finite population variance is 25.

Set a random seed and then draw 100K samples each of size $n = 1024$ (drawing each sample without replacement) from the finite population of size 1200; compute the sample mean $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ for each sample. *Hint: use the sample command with replace=FALSE.*

```
set.seed(123)

finite_samples <- tibble(
  sample_means = replicate(draws, expr = mean(sample(x = finite_pop$y,
                                                     size = n,
                                                     replace = FALSE)))
)
```

(d) Report the random seed that you used. Then compute the sample mean and sample variance of the 100K sample means that you obtained from the finite population.

The seed used above when making the 100K samples and calculating the mean for each sample was 123. We will use R to calculate the sample variance and mean of the 100K sample means of the finite population.

```
mean_finite_resamples <- mean(finite_samples$sample_means)
var_finite_resamples <- var(finite_samples$sample_means)
```

As calculated above, the sample mean of the 100K sample means was 1.8524408×10^{-4}, and the sample variance of the 100K sample means was 0.0035593.

(e) Using your results from (d) verify the claims in this **Theorem 1 (Sample Mean from sample of size n taken from a Finite Popn. of size N)** *The mean and variance of the sample mean $\bar{X}$ of a sample of size n from a population of size N with population mean μ and variance σ^2 are:*

$$E[\bar{X}] = \mu \qquad Var[\bar{X}] = \frac{(N-n)}{n(N-1)}\sigma^2$$

As shown in part c, the population mean is $4.4489875 \times 10^{-17}$, and the population variance is 25. Thus we can see that $E[\bar{X}] = 4.4489875 \times 10^{-17}$ (which is esentially zero) and $Var[\bar{X}] = \frac{(N-n)}{n(N-1)}\sigma^2 = \frac{(1200-1024)}{1024(1200-1)} \cdot 25 = 0.003584$. Since we found that the mean of the 100K resample means was 1.8524408×10^{-4}, and the variance of the 100K resample means was 0.0035593, we have successfully verified the claim that the mean and variance of the sample mean $\bar{X}$ of a sample of size n from a population of size N with population mean μ and variance σ^2 are: $E[\bar{X}] = \mu$ and $Var[\bar{X}] = \frac{(N-n)}{n(N-1)}\sigma^2$.

(f) Compute the ratio of the theoretical variances obtained in (e) over that obtained in (a); also compare this to the ratio of the variance of the 100K sample reported in (d) over that reported in (b).

```
theor_ratio <- ((N-n)*finite_var/(n*(N-1))) / ((25)/(n))
sample_ratio <- var_finite_resamples / var_samples
```

As computed above, the ratio of the theoretical variances obtained in (e) over that obtained in (a) is $\frac{\frac{(N-n)}{n(N-1)} \cdot \sigma^2}{\frac{\sigma^2}{n}} = \frac{\frac{(1200-1024)}{1024(1200-1)} \cdot 25}{\frac{25}{1024}} = 0.146789$. While the ratio of the variance of the 100K samples reported in (d) over that reported in (b) is 0.1454749.

Bayesian Statistics

2. Suppose a medical test has the following characteristics:

$$\Pr(\text{Test Positive} \mid \text{Patient Diseased}) = 0.99$$

$$\Pr(\text{Test Negative} \mid \text{Patient Not Diseased}) = 0.98$$

(a) Find Pr(Test Negative | Patient Diseased) and Pr(Test Positive | Patient Not Diseased). Suppose that 1 in $5,000$ people have this disease so

$$\Pr(\text{Patient Diseased}) = 0.0002$$

In order to find Pr(Test Negative | Patient Diseased) and Pr(Test Positive | Patient Not Diseased), we will create a table to better visualize what's going on in this scenario, we will make a table of probabilities.

	Test Positive	Test Negative
Patient Diseased	0.99	?
Patient Not Diseased	?	0.98

Since $P(\text{Test Positive} \mid \text{Patient Diseased}) + P(\text{Test Negative} \mid \text{Patient Diseased}) = 0.99 + ? = 1$, we can see that $P(\text{Test Negative} \mid \text{Patient Diseased}) = 0.01$. Similarly, since $P(\text{Test Positive} \mid \text{Patient Not Diseased}) + P(\text{Test Negative} \mid \text{Patient Not Diseased}) = ? + 0.98 = 1$, we can see that $P(\text{Test Positive} \mid \text{Patient Not Diseased}) = 0.02$.

Thus our table is now:

	Test Positive	Test Negative
Patient Diseased	0.99	0.01
Patient Not Diseased	0.02	0.98

(b) Compute Pr(Test Positive). *Hint: Find Pr(* Test Positive, Patient Diseased*) and Pr(* Test Positive, Patient Not Diseased

Just like in the previous part, in order to find Pr(Test Positive), we will create a table to better visualize what's going on in this scenario, we will make a table of proportions.

	Test Positive	Test Negative	
Patient Diseased	0.000198	0.000002	0.0002
Patient Not Diseased	0.019996	0.979804	0.9998
	0.020194	0.979806	1

Thus, Pr(Test Positive) = Pr(Test Positive, Patient Diseased) + Pr(Test Positive, Patient Not Diseased) = $0.000198 + 0.019996 = 0.020194$.

(c) Use Bayes' rule to find Pr(Patient Diseased | Test Positive).

Bayes' rule states that $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$, thus in our scenario, $P(\text{Patient Diseased} \mid \text{Test Positive}) = \frac{P(\text{Test Positive} \mid \text{Patient Diseased}) \cdot P(\text{Patient Diseased})}{P(\text{Test Positive})}$. However, we already know $P(\text{Test Positive} \mid \text{Patient Diseased})$, $P(\text{Test Positive})$, and $P(\text{Patient Diseased})$, thus all we need to do is to put it all together. Therefore, $P(\text{Patient Diseased} \mid \text{Test Positive}) = \frac{0.99 \cdot 0.0002}{0.201949} = 0.009805$.

(d) Give an intuitive explanation for the discrepancy between Pr(Patient Diseased | Test Positive) and Pr(Test Positive | Patient Diseased).

An intuitive explanation for the discrepancy between Pr(Patient Diseased | Test Positive) and Pr(Test Positive | Patient Diseased) revolves around two important aspects: how small the probability is to actually observe someone who has the disease, and the appearance of false positives (Type I errors). To further expand on the important aspects above, because the probability of having the disease is so low at 0.0002, even with a low significance level, the sheer amount of individuals who don't have the disease leads to an overwhelming amount of false positives which completely overshadows the true positives. For example, consider our situation with $50,000$ individuals, we would see 9.9 individuals test positive out of the 10 who have the disease, and we would also see 999.8 individuals test positive out of the $49,990$ who don't have the disease. As you can see, there are 100 times more false positives than there are true positives, hence why we see such a small likelihood of knowing someone has the disease given that they tested positive.

3. A researcher is interested in the proportion p of voters who are in favor of a ballot measure. She gathers data over two days. She contacts a random sample of n_1 voters on day 1 of whom y_1 are in favor; on day 2 she contacts another n_2 of whom y_2 are in favor. Before seeing the data the researcher's prior distribution for p is $Beta(\alpha, \beta)$, where $\alpha, \beta > 0$.

(a) Write down the researcher's posterior distribution for p after seeing the data from day 1;

In this simple example, the proportion p of voters who are in favor of a ballot measure are represented by a Binomial likelihood since you can either be in favor of a ballot measure or you can be opposed (technically you could have no opinion but we will disregard this option for simplicity). In this particular example, before seeing the data a researcher's prior distribution for p is $Beta(\alpha, \beta)$, where $\alpha, \beta > 0$. Suppose she contacts a random sample of n_1 voters on day 1 of whom y_1 are in favor. Since the prior is Beta and the likelihood is Binomial, it follows that the posterior will also have a Beta distribution, this is due to the fact that the Beta family is said to be conjugate to the Binomial likelihood. With that said, since the posterior in this Binomial likelihood/Beta prior case takes the form of $p|x \sim Beta(\alpha + x, \beta + (n - x))$, it follows that the researcher's posterior distribution for p after seeing the data from day 1 is $p|y_1 \sim Beta(\alpha + y_1, \beta + (n_1 - y_1))$.

(b) Write down the researcher's posterior distribution after seeing the data from both day 1 and day 2.

Suppose after day 1, on day 2 she contacts another n_2 of whom y_2 are in favor. Following from the above logic, and taking the old posterior as our prior, it follows that the the researcher's posterior distribution after seeing the data from both day 1 and day 2 is $p|y_1, y_2 \sim Beta(\alpha + y_1 + y_2, \beta + (n_1 - y_1) + (n_2 - y_2))$.

(c) Give simple expressions for means of the two posterior distributions that you found in (a) and (b); you may appeal to properties of the Beta distribution here.

In order to give simple expressions for the means of the two posterior distributions that we found in (a) and (b), we will use the simple property of the Beta distribution that sates: if $X \sim Beta(\alpha, \beta)$, then $E[X] = \frac{\alpha}{\alpha+\beta}$.

Thus we can see for the posterior after day 1, $E[p|y_1] = \frac{\alpha+y_1}{\alpha+y_1+\beta+(n_1-y_1)} = \frac{\alpha+y_1}{\alpha+\beta+n_1}$. Likewise, for the posterior after day 2, $E[p|y_1, y_2] = \frac{\alpha+y_1+y_2}{\alpha+y_1+\beta+(n_1-y_1)+(n_2-y_2)} = \frac{\alpha+y_1+y_2}{\alpha+\beta+n_1+n_2}$.

Suppose that $\alpha = \beta = 1$, $n_1 = 50$, $y_1 = 29$, $n_2 = 60$, $y_2 = 38$.

With these given parameters and data points, it follows that $p|y_1, y_2 \sim Beta(1 + 29 + 38, 1 + (50 - 29) + (60 - 38)) = Beta(68, 44)$.

(d) Find a symmetric 95% credible interval for p based on the data from both days. *Hint: Use the qbeta command in R.*

Day 1:

In order to find the 95% credible interval for p based on the data from day 1, we will use qbeta to solve for the 0.025th quantile and the 0.975th quantile of $p|y_1 \sim Beta(30, 22)$.

```r
lower <- qbeta(p = 0.025, shape1 = 30, shape2 = 22)
upper <- qbeta(p = 0.975, shape1 = 30, shape2 = 22)
```

Thus, as calculated above, our 95% credible interval for p based on the data from day 1 is [0.4416928, 0.7065451].

This interval tells us that there is a 95% probability that the true estimate would lie within the interval [0.4416928, 0.7065451], given the evidence provided by our observed data from day 1.

Day 1 and Day 2:

Similarly, in order to find the 95% credible interval for p based on the data from both days (day 1 and day 2), we will use qbeta to solve for the 0.025th quantile and the 0.975th quantile of $p|y_1, y_2 \sim Beta(68, 44)$.

```r
lower2 <- qbeta(p = 0.025, shape1 = 68, shape2 = 44)
upper2 <- qbeta(p = 0.975, shape1 = 68, shape2 = 44)
```

Thus, as calculated above, our 95% credible interval for p based on the data from both days is [0.5154685, 0.6951917].

This interval tells us that there is a 95% probability that the true estimate would lie within the interval [0.5154685, 0.6951917], given the evidence provided by our observed data from both days.