Notes for 09/12

Nicholas Ray

2024-09-12

Let's say there exists a population with a feature we are interested in estimating the average of. After deciding how to map observations of the population onto $\mathbb{R}$ (i.e., defining a random variable), we went out to observe this population five times to obtain the sample $X = \{-1, -1, 0, 1, 2\}$. Note that we were careful to ensure that our observations were independent of each other, s.t. they satisfy the i.i.d requirement for inference (Aronow and Miller 2019, 92). In other words, observing one member of the population had no impact on the probability of observing another member (independent) and all observations were from the same population (identical).

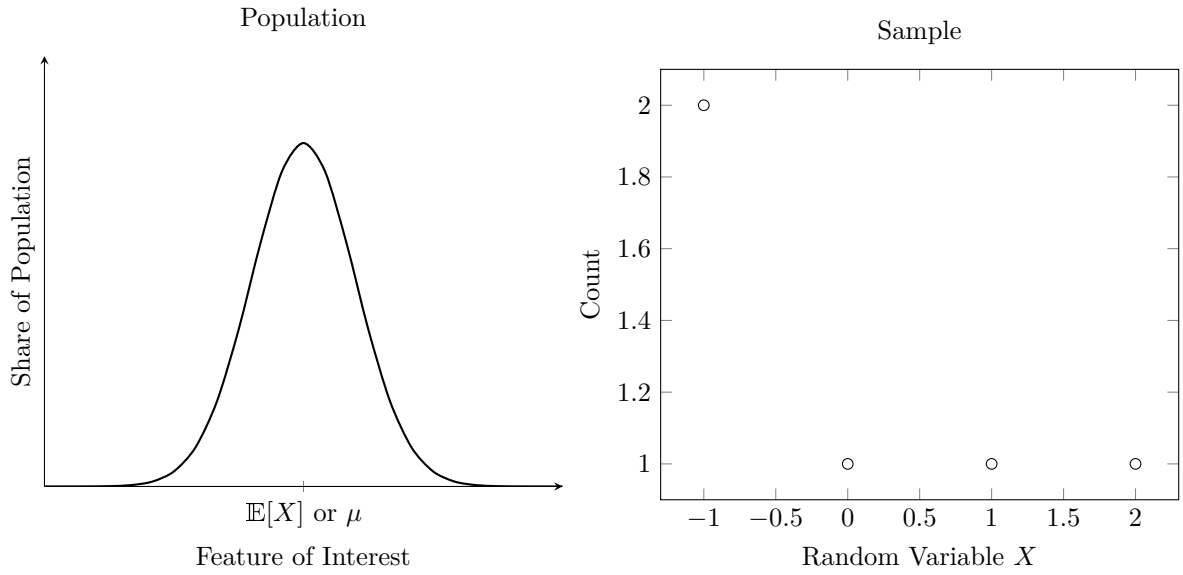

Figure 1:

What is the sample mean, our estimator for $\mathbb{E}[X]$ or μ?

$$\begin{aligned}\overline{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \text{ (Aronow and Miller 2019, 97)} \\ &= \frac{1}{5}\big((-1) + (-1) + 0 + 1 + 2\big) \\ &= 0.2\end{aligned}$$

```
X<-c(-1,-1,0,1,2)
mean(X) #(-1+-1+0+1+2)/5 X_bar
```

```
## [1] 0.2
```

Digression: How do we know $\overline{X}$ is unbiased (i.e., $\mathbb{E}[\overline{X}] - \mathbb{E}[X] = 0$)? From (Aronow and Miller 2019, 97):

$$\begin{aligned}
\mathbb{E}[\overline{X}] &= \mathbb{E}[\frac{1}{n}(X_1 + X_2 + ... + X_n)] \\
&= \frac{1}{n}\mathbb{E}[X_1 + X_2 + ... + X_n] \\
&= \frac{1}{n}\big(\mathbb{E}[X_1] + \mathbb{E}[X_2] + ... + \mathbb{E}[X_n]\big) \\
&= \frac{1}{n}\big(\mathbb{E}[X] + \mathbb{E}[X] + ... + \mathbb{E}[X]\big) \\
&= \frac{1}{n} n \mathbb{E}[X] \\
&= \mathbb{E}[X]
\end{aligned}$$

Note that this means that *in expectation*, $\overline{X}$ is unbiased for $\mathbb{E}[X]$. Unbiasedness is a "repeated sample property" because doing something in expectation means doing it many, many times (infinitely many times, actually). Thus, in finite samples there is always the chance that an unbiased estimator is in fact biased. This is to say that we cannot "prove" an estimator is unbiased via simulation since we cannot run a simulation an infinite number of times. However, in combination with our analytically derived knowledge that an estimator is unbiased, averaging across many repeated simulations can help confirm our derivations even if there is a small amount of bias in the results.

What is the (sampling) variance of our estimator?

$$\begin{aligned}
\hat{V}[\overline{X}] &= \frac{\hat{V}[X]}{n} \text{ (Aronow and Miller 2019, 114)} \\
&= \frac{1}{n}\Big(\frac{n}{n-1}(\overline{X^2} - \overline{X}^2)\Big) \\
&= \frac{1}{n-1}\big(\overline{X^2} - \overline{X}^2\big) \\
&= \frac{1}{n-1}\left(\frac{1}{n}\Big((-1)^2 + (-1)^2 + 0^2 + 1^2 + 2^2\Big) - (0.2)^2\right) \\
&= \frac{1}{4}\big(\frac{1}{5}(7) - 0.04\big) \\
&= 0.34
\end{aligned}$$

```
#R by default returns the estimated population variance:
var(X) #((-1-0.2)^2+(-1-0.2)^2+(0-0.2)^2+(1-0.2)^2+(2-0.2)^2)/(5-1) #V_hat(X)
#or
(mean(X^2)-mean(X)^2)*5/(5-1) #V_hat(X)

#we need to divide by n to find the estimated variance of the sample mean:
((mean(X^2)-mean(X)^2)*5/(5-1))/5 #V_hat(X_bar)
#or
var(X)/5 #V_hat(X_bar)
```

What is the 95% confidence interval for our estimator?

$$
\begin{aligned}
CI_{1-\alpha}(\overline{X}) &= \overline{X} \pm \sqrt{\hat{V}[\overline{X}]} \cdot z_{-\frac{\alpha}{2}} \text{ (Aronow and Miller 2019, 125)} \\
CI_{0.95}(\overline{X}) &= \overline{X} \pm \sqrt{\hat{V}[\overline{X}]} \cdot z_{-\frac{0.05}{2}} \\
&= 0.2 \pm \sqrt{0.34} \cdot 1.96 \\
&= 0.2 \pm 1.14 \\
&= \{-0.94, 1.34\}
\end{aligned}
$$

```
CI_lower<-mean(X)-(sqrt(var(X)/5)*1.96) #1.96 approX. qnorm(0.025)
CI_upper<-mean(X)+(sqrt(var(X)/5)*1.96)
```

What does this confidence interval mean? It means that if we did this for an infinitely large sample, the random interval $\{-0.94, 1.34\}$ will contain $\mathbb{E}[X]$ with at least 95% probability (Aronow and Miller 2019, 124).

If conducting a hypothesis test, what would our test statistic be? We would use the t-statistic, or t-stat, which is based on Student t's distribution but can be approximated by the Normal. Assume that our null hypothesis (H_0) for the true population mean (μ or $\mathbb{E}[X]$) is 0.

$$
\begin{aligned}
t &= \frac{\overline{X} - \mu_{H_0}}{\sqrt{\hat{V}[\overline{X}]}} \text{ (Aronow and Miller 2019, 129)} \\
&= \frac{0.2 - 0}{\sqrt{0.34}} \\
&= 0.343
\end{aligned}
$$

```
t<-(mean(X)-0)/(sqrt(var(X)/5))
```

What is the probability (p-value) that we would have observed a sample mean (and corresponding t-stat) this extreme if the true mean was really 0? Assuming a two-tailed hypothesis test:

$$
\begin{aligned}
p &= 2\big(1 - \Phi(|t|)\big) \text{ (Aronow and Miller 2019, 130)} \\
&= 2\big(1 - \Phi(0.343)\big) \\
&= 0.73
\end{aligned}
$$

```
2*pnorm(abs(t),lower.tail=FALSE)
```

```
## [1] 0.7316006
```

What does all of this mean? It means that assuming asymptotically valid confidence intervals, t-stats, and p-values, there is very little reason to believe that the true mean is not zero based on our single estimated sample mean. If the true mean really is zero, then observing the sample we did would have been very likely (with a probability or p-value of 0.73). Thus, we cannot reject the null hypothesis that the true mean is zero.

Bibliography

Aronow, Peter B., and Benjamin T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press.