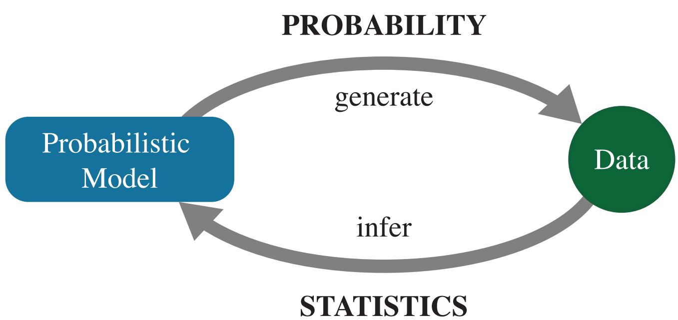
Part V

Statistical Inference

The focus until now in the book has been on *probability*. We can think of probability as defined by a probabilistic model, or distribution, which governs an "experiment," through which one *generates* samples, or events, from this distribution. One might ask questions about the probability of a certain event occurring, under the known probabilistic model.

We now turn our attention to *statistics*. In statistics, we go the other direction. We are given some data, and our goal is to *infer* the underlying probabilistic model that generated this data.



The figure above illustrates the difference in direction. While statistics and probability may sound different, they are actually closely linked. In particular, when a statistician is trying to "infer" (estimate) the underlying probabilistic model that generated some data, they might start by computing the probability that certain candidate models produced that data.

Because the data that we see is limited, either in quantity (there may only be a few samples) or in accuracy (the data may be somewhat noisy or corrupted), there is often some subjectivity involved in determining the best estimator for the underlying probabilistic model. In this sense, statistics is sometimes viewed as more of an art, where statisticians might argue with each other over which estimator is more "correct." We will see several examples of this in our study of statistical inference.

We start in Chapter 15 by discussing the most commonly used estimators, namely those for mean and variance. In Chapter 16 we move on to parameter estimation following the classical inference approach of maximum likelihood estimation. In Chapter 17 we continue looking at parameter estimation, but this time via the Bayesian inference approach, where we discuss maximum a posteriori estimators and minimum mean square error estimators. Along the way, we also touch on a few related topics like linear regression (see Section 16.7).

Although this is the main statistics part of the book, statistical topics come up throughout the book. In particular, the important topic of confidence intervals on estimators is deferred to Chapter 19, since it is better treated after a more in-depth discussion of tail probabilities.

15 Estimators for Mean and Variance

The general setting in statistics is that we observe some data and then try to infer some property of the underlying distribution behind this data. The underlying distribution behind the data is unknown and represented by random variable (r.v.) X. This chapter will briefly introduce the general concept of estimators, focusing on estimators for the mean and variance.

15.1 Point Estimation

Point estimation is an estimation method which outputs a single value. As an example of a point estimation, suppose we are trying to estimate the number of books that the average person reads each year. We sample n people at random from the pool of all people and ask them how many books they read annually. Let $X_1, X_2, \ldots, X_n$ represent the responses of the n people. We can assume that the pool of people is sufficiently large that it's reasonable to think of the X_i's as being independent and identically distributed (i.i.d.), where $X_i \sim X$ for all i. We would like to estimate $\theta = \mathbf{E}[X]$. A reasonable point estimator for θ is simply the average of the X_i's sampled.

Definition 15.1 *We write*

$$\hat{\theta}(X_1, X_2, \ldots, X_n)$$

to indicate an **estimator** *of the unknown value θ. Here $X_1, \ldots, X_n$ represent the sampled data and our estimator is a function of this data. Importantly, $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is a* **random variable***, since it is a function of random variables. We sometimes write $\hat{\theta}$ for short when the sample data is understood. We write*

$$\hat{\theta}(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n)$$

to indicate the **constant** *which represents our estimation of θ based on a specific instantiation of the data where $X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n$.*

15.2 Sample Mean

While $\hat{\theta}$ is the notation most commonly used for an estimator of θ, there are certain estimators, like the "sample mean," that come up so frequently that they have their own name.

Definition 15.2 (Mean estimator) *Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of r.v. X with unknown mean. The* **sample mean** *is a point estimator of $\theta = \mathbf{E}[X]$. It is denoted by $\overline{X}$ or by M_n, and defined by:*

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = M_n = \overline{X} \equiv \frac{X_1 + X_2 + \cdots + X_n}{n}. \tag{15.1}$$

The notation M_n is attractive because it specifies the number of samples, while the notation $\overline{X}$ is attractive because it specifies the underlying distribution whose mean we are estimating.

15.3 Desirable Properties of a Point Estimator

For any unknown parameter θ that we wish to estimate, there are often many possible estimators.

As a running example, throughout this section, let $X_1, X_2, \ldots, X_n$ be i.i.d. random samples from a distribution represented by r.v. X, with finite mean $\mathbf{E}[X]$ and finite variance σ^2.

In estimating $\theta = \mathbf{E}[X]$, consider two possible estimators:

$$\hat{\theta}_A = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$
$$\hat{\theta}_B = X_2.$$

What makes one estimator better than another? In this section we define some desirable properties of a point estimator.

Definition 15.3 *Let $\hat{\theta}(X_1, X_2, \ldots X_n)$ be a point estimator for θ. Then we define the* **bias** *of $\hat{\theta}$ by*

$$\mathbf{B}(\hat{\theta}) = \mathbf{E}\left[\hat{\theta}\right] - \theta.$$

If $\mathbf{B}(\hat{\theta}) = 0$, we say that $\hat{\theta}$ is an **unbiased estimator** *of θ.*

Clearly we would like our estimator to have zero bias.

Question: How do $\hat{\theta}_A$ and $\hat{\theta}_B$ compare with respect to bias?

Answer: They are both unbiased estimators.

Question: Nevertheless, why do you favor $\hat{\theta}_A$ over $\hat{\theta}_B$?

Answer: $\hat{\theta}_A$ feels less variable. This brings us to the second desirable property of an estimator, which is low mean squared error.

> **Definition 15.4** *The* **mean squared error (MSE)** *of an estimator $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is defined as:*
> $$\mathbf{MSE}(\hat{\theta}) = \mathbf{E}\left[\left(\hat{\theta} - \theta\right)^2\right].$$

> **Lemma 15.5** *If $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is an unbiased estimator, then*
> $$\mathbf{MSE}(\hat{\theta}) = \mathbf{Var}(\hat{\theta}).$$

Proof:

$$\mathbf{MSE}(\hat{\theta}) = \mathbf{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathbf{E}\left[\left(\hat{\theta} - \mathbf{E}\left[\hat{\theta}\right]\right)^2\right] = \mathbf{Var}(\hat{\theta}). \qquad \blacksquare$$

Question: How do $\hat{\theta}_A$ and $\hat{\theta}_B$ compare with respect to their MSE?

Answer: Using Lemma 15.5,

$$\mathbf{MSE}(\hat{\theta}_A) = \mathbf{Var}(\hat{\theta}_A) = \frac{1}{n^2} \cdot n\mathbf{Var}(X) = \frac{\mathbf{Var}(X)}{n}.$$

By contrast,

$$\mathbf{MSE}(\hat{\theta}_B) = \mathbf{Var}(\hat{\theta}_B) = \mathbf{Var}(X_2) = \mathbf{Var}(X).$$

Thus $\hat{\theta}_A$ has much lower MSE.

Finally, it is desirable that our estimator has the property that it becomes more accurate (closer to θ) as the sample size increases. We refer to this property as **consistency**.

> **Definition 15.6** *Let $\hat{\theta}_1(X_1), \hat{\theta}_2(X_1, X_2), \hat{\theta}_3(X_1, X_2, X_3), \ldots$ be a sequence of point estimators of θ, where $\hat{\theta}_n(X_1, X_2, \ldots, X_n)$ is a function of n i.i.d. samples. We say that r.v. $\hat{\theta}_n$ is a* **consistent estimator** *of θ if, $\forall \epsilon > 0$,*
> $$\lim_{n\to\infty} \mathbf{P}\left\{\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right\} = 0.$$

Lemma 15.7 *Let $\hat{\theta}_1(X_1), \hat{\theta}_2(X_1, X_2), \hat{\theta}_3(X_1, X_2, X_3), \ldots$ be a sequence of point estimators of θ, where $\hat{\theta}_n(X_1, X_2, \ldots, X_n)$ is a function of n i.i.d. samples. Assume that all the estimators have finite mean and variance. If*

$$\lim_{n\to\infty} \mathbf{MSE}(\hat{\theta}_n) = 0,$$

then $\hat{\theta}_n$ is a consistent estimator.

Proof: For any constant $\epsilon > 0$,

$$\begin{aligned}
\mathbf{P}\left\{\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right\} &= \mathbf{P}\left\{\left|\hat{\theta}_n - \theta\right|^2 \geq \epsilon^2\right\} \\
&\leq \frac{\mathbf{E}\left[\left|\hat{\theta}_n - \theta\right|^2\right]}{\epsilon^2} \qquad \text{by Markov's inequality (Theorem 5.16)} \\
&= \frac{\mathbf{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right]}{\epsilon^2} \\
&= \frac{\mathbf{MSE}(\hat{\theta}_n)}{\epsilon^2}.
\end{aligned}$$

Taking limits of both sides as $n \to \infty$, we have:

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right\} = \lim_{n\to\infty} \frac{\mathbf{MSE}(\hat{\theta}_n)}{\epsilon^2} = 0. \qquad \blacksquare$$

Question: In the proof of Lemma 15.7, why didn't we apply Chebyshev's inequality (Theorem 5.17)?

Answer: We don't know that $\theta = \mathbf{E}\left[\hat{\theta}_n\right]$, so we can't say that $\mathbf{MSE}(\hat{\theta}_n) = \mathbf{Var}(\hat{\theta}_n)$.

Question: Is $\hat{\theta}_A = \overline{X} = M_n$ a consistent estimator of $\mathbf{E}[X]$?

Answer: Yes. By Lemma 15.7, it suffices to show that

$$\lim_{n\to\infty} \mathbf{MSE}(M_n) = 0.$$

Given that we know that M_n is an unbiased estimator of $\mathbf{E}[X]$, Lemma 15.5 tells us that it suffices to show that

$$\lim_{n\to\infty} \mathbf{Var}(M_n) = 0.$$

But this latter fact is obviously true because $\mathbf{Var}(X)$ is finite and thus

$$\mathbf{Var}(M_n) = \frac{\mathbf{Var}(X)}{n} \to 0 \qquad \text{as } n \to \infty.$$

Hence, M_n is a consistent estimator.

15.4 An Estimator for Variance

Again let $X_1, X_2, \ldots, X_n$ denote n i.i.d. samples of an unknown distribution denoted by r.v. X, where $X_i \sim X$, and where $\mathbf{E}[X] = \mu$ and $\mathbf{Var}(X) = \sigma_X^2$ are finite. We have seen that $\overline{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ is a good estimator for $\mathbf{E}[X]$, satisfying all three desirable properties. We now turn to the question of a good estimator for $\mathbf{Var}(X)$.

There are two distinct cases to consider:

1. The case where we already know the mean and want to estimate $\theta = \mathbf{Var}(X)$.
2. The case where we do not know the mean and want to estimate $\theta = \mathbf{Var}(X)$.

It turns out that the best estimator is different for these two cases.

15.4.1 Estimating the Variance when the Mean is Known

Starting with the first case, suppose that μ is known. We can then define an estimator which computes the squared distance of each sample from μ and takes the average of these squared distances:

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = \overline{S^2} \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 . \tag{15.2}$$

Question: Is $\overline{S^2}$ as defined in (15.2) an unbiased estimator for $\theta = \mathbf{Var}(X)$?

Answer: Yes!

$$\mathbf{E}\left[\overline{S^2}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[(X_i-\mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}(X_i) = \mathbf{Var}(X).$$

15.4.2 Estimating the Variance when the Mean is Unknown

Now consider the second case, where μ is not known. This case is way more common but also trickier.

Question: Given that we don't know $\mu = \mathbf{E}[X]$, how can we replace μ in our definition of the estimator?

Answer: We can replace μ by $\overline{X} = \frac{X_1+X_2+\cdots+X_n}{n}$, which we already saw was a good estimator for $\mathbf{E}[X]$.

This leads us to an updated definition of our estimator, which now computes the squared distance of each sample from $\overline{X}$ and takes the average of these squared distances:

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = \overline{S^2} \equiv \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \tag{15.3}$$

Question: Is $\overline{S^2}$ as defined in (15.3) an unbiased estimator for $\theta = \mathbf{Var}(X)$?

Answer: Unfortunately, and surprisingly, the answer is no. In Exercise 15.4, you will prove that

$$\mathbf{E}\left[\overline{S^2}\right] = \frac{n-1}{n} \cdot \mathbf{Var}(X). \tag{15.4}$$

Question: Given (15.4), what is an unbiased estimator for $\theta = \mathbf{Var}(X)$ in the case where we don't know $\mathbf{E}[X]$?

Answer: We need to multiply $\overline{S^2}$ by $\frac{n}{n-1}$. The sample variance, defined next, does this.

Definition 15.8 (Variance Estimator) *Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of r.v. X with unknown mean and variance. The* **sample variance** *is a point estimator of $\theta = \mathbf{Var}(X)$. It is denoted by S^2 and defined by:*

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \tag{15.5}$$

Lemma 15.9 *The sample variance, S^2, from Definition 15.8 is an unbiased estimator of* $\mathbf{Var}(X)$.

Proof:

$$\mathbf{E}\left[S^2\right] \stackrel{(15.3)}{=} \frac{n}{n-1}\mathbf{E}\left[\overline{S^2}\right] \stackrel{(15.4)}{=} \frac{n}{n-1} \cdot \frac{n-1}{n} \cdot \mathbf{Var}(X) = \mathbf{Var}(X) \qquad \blacksquare$$

Question: The difference between the estimators $\overline{S^2}$ in (15.3) and S^2 in (15.5) is very slight. Does it really matter which we use?

Answer: Assuming that the number of samples, n, is large, in practice it shouldn't matter which of these two estimators we use.

15.5 Estimators Based on the Sample Mean

Simple estimators, like the sample mean, can sometimes be useful in estimating other, more complex quantities. We provide one example here and another in Exercise 15.6.

Example 15.10 (Estimating the number of tanks)

In World War II, the Allies were trying to estimate the number of German tanks. Each tank was assigned a serial number when it was created. When the Allies captured a tank, they would record its serial number.

Question: If the Allies captured the tanks with serial numbers shown in Figure 15.1, what is a good estimate for the total number of German tanks?

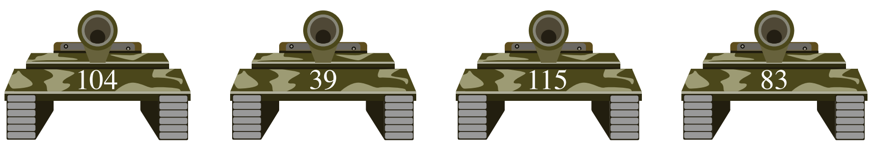


Figure 15.1 *Captured tanks with serial numbers shown.*

We are trying to estimate a *maximum*, call it θ, based on seeing n samples, $X_1, X_2, \ldots, X_n$, each of which are randomly picked *without replacement* from the integers $1, 2, \ldots, \theta$. Our goal is to determine $\hat{\theta}(X_1, X_2, \ldots, X_n)$.

Question: Are the n samples independent?

Answer: No. Once serial number k is seen, it will never be seen again.

There are many ways to estimate the max, θ. We will use the sample mean to estimate θ, by expressing the expectation of the sample mean as a function of θ.

$$\overline{X} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)$$
$$\mathbf{E}\left[\overline{X}\right] = \frac{1}{n}\left(\mathbf{E}\left[X_1\right] + \mathbf{E}\left[X_2\right] + \cdots + \mathbf{E}\left[X_n\right]\right).$$

Although the X_i's are not independent, they all have the same marginal distribution:

$$\mathbf{P}\left\{X_i = k\right\} = \frac{1}{\theta}, \qquad \text{where } 1 \le k \le \theta.$$

Hence,

$$\mathbf{E}\left[X_i\right] = \frac{1}{\theta} \cdot 1 + \frac{1}{\theta} \cdot 2 + \cdots + \frac{1}{\theta} \cdot \theta = \frac{\theta + 1}{2}.$$

But this implies

$$\mathbf{E}\left[\overline{X}\right] = \frac{\theta + 1}{2}. \tag{15.6}$$

Equivalently, we can write

$$\theta = 2\mathbf{E}\left[\overline{X}\right] - 1.$$

Hence, a reasonable estimator for θ could be

$$\hat{\theta}(X_1, X_2, \ldots, X_n) \equiv 2\overline{X} - 1. \tag{15.7}$$

Question: Is $\hat{\theta}$ from (15.7) an unbiased estimator?

Answer: Yes, by (15.6), we see that $\mathbf{E}\left[\hat{\theta}\right] = 2\mathbf{E}\left[\overline{X}\right] - 1 = \theta$.

Question: Is $\hat{\theta}$ from (15.7) a good estimator of θ?

Answer: Not necessarily. If the number of samples, n, is small, we could end up in the perverse situation where there is one very high sample, while most of the samples are far below the mean. In this case, our sample mean, $\overline{X}$, would be particularly low, so $\hat{\theta} = 2\overline{X} - 1$ might actually be smaller than the largest sample.

Now suppose we want to determine $\mathbf{MSE}(\hat{\theta})$. Since $\hat{\theta}$ is an unbiased estimator, by Lemma 15.5, $\mathbf{MSE}(\hat{\theta}) = \mathbf{Var}(\hat{\theta})$. Thus,

$$\begin{aligned}
\mathbf{MSE}(\hat{\theta}) = \mathbf{Var}(\hat{\theta}) &= \mathbf{Var}\left(2\overline{X} - 1\right) \\
&= \frac{4}{n^2}\mathbf{Var}(X_1 + X_2 + \cdots + X_n) \\
&= \frac{4}{n^2} \cdot \left(\sum_{i=1}^{n} \mathbf{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathbf{Cov}(X_i, X_j)\right) \quad \text{(by (5.11))} \\
&= \frac{4}{n} \cdot \left(\mathbf{Var}(X_1) + (n-1)\mathbf{Cov}(X_1, X_2)\right),
\end{aligned}$$

where the last line follows from the fact that all the X_i's have the same distribution, and all the pairs (X_i, X_j) have the same distribution.

From (5.13) and (5.14), we know that:

$$\mathbf{Var}(X_1) = \frac{(\theta - 1)(\theta + 1)}{12} \qquad \text{and} \qquad \mathbf{Cov}(X_1, X_2) = -\frac{\theta + 1}{12}.$$

Hence,

$$\mathbf{MSE}(\hat{\theta}) = \frac{4}{n} \cdot \left(\frac{(\theta - 1)(\theta + 1)}{12} - (n-1) \cdot \frac{\theta + 1}{12}\right) = \frac{1}{3n}(\theta + 1)(\theta - n).$$

So we see that the MSE of our estimate increases with the square of the highest value, θ, and decreases linearly with the number of samples, n.

15.6 Exercises

15.1 **Practice computing sample mean and sample variance**
The following 10 job sizes are measured: $5, 2, 6, 9, 1.5, 2.3, 7, 15, 8, 8.3$. What is the sample mean, $\overline{X}$? What is the sample variance, S^2?

15.2 **Accuracy of sample mean and sample variance**
Generate 30 instances of each of the following distributions – recall (13.2):
(i) $X \sim \text{Exp}(1)$
(ii) $X \sim \text{Exp}(.01)$
(iii)
$$X \sim \begin{cases} \text{Exp}\,(1) & \text{w/prob } 0.99 \\ \text{Exp}(.01) & \text{w/prob } 0.01 \end{cases}.$$
For each distribution, answer the following questions:
(a) What is the sample mean? Compare this with the true mean, $\mathbf{E}[X]$.
(b) What is the sample variance? Compare this with $\mathbf{Var}(X)$.
(c) For which distribution was the sample mean most (least) accurate? How about the sample variance? Provide some thoughts on why.
Now repeat the problem, generating 100 instances of each distribution.

15.3 **Variance–bias decomposition**
Given an estimator $\hat{\theta}(X_1, \ldots, X_n)$, prove that
$$\mathbf{MSE}(\hat{\theta}) = \mathbf{Var}(\hat{\theta}) + (\mathbf{B}(\hat{\theta}))^2, \tag{15.8}$$
where $\mathbf{B}(\hat{\theta}) \equiv \mathbf{E}\left[\hat{\theta}\right] - \theta$ is the bias of $\hat{\theta}$.

15.4 **Estimating variance is tricky**
Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of r.v. X with unknown finite mean and variance. Let $\overline{X}$ denote the sample mean. Define
$$\overline{S^2} \equiv \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$
Prove that $\overline{S^2}$ is *not* an unbiased estimator of $\mathbf{Var}(X)$. Follow these steps:
(a) Prove that $\mathbf{E}\left[\overline{S^2}\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Var}\left(X_i - \overline{X}\right)$.
(b) Show that $\mathbf{Var}\left(X_i - \overline{X}\right) = \frac{n-1}{n}\mathbf{Var}(X)$.
(c) Combine (a) and (b) to show that $\mathbf{E}\left[\overline{S^2}\right] = \frac{n-1}{n}\mathbf{Var}(X)$.

15.5 **Sample standard deviation**
Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of r.v. X with unknown finite mean and variance. Define the **sample standard deviation**, S, as

$$S = \sqrt{S^2},$$

where S^2 is the sample variance, given by (15.5). Is S an unbiased estimator of $\mathbf{std}(X)$? Prove your answer. [Hint 1: S is not a constant, so $\mathbf{Var}(S) > 0$.] [Hint 2: Use the fact that $\mathbf{E}\left[S^2\right] = \mathbf{Var}(X)$.]

15.6 **Arrivals at a web server: two estimators**
The arrival process of requests to a web server is well-modeled by a Poisson process with some average rate λ requests/minute. We're interested in

$$p_0 = \text{Fraction of minutes during which there are 0 requests.}$$

If we know λ, then we know from Chapter 12 that $p_0 = e^{-\lambda}$. But how can we estimate p_0 if we don't know λ? Let's suppose that we have sampled n minutes and let $X_1, X_2, \ldots, X_n$ denote the number of arrivals during each of the n minutes.

(a) One idea is to first define an estimator for λ, namely

$$\hat{\lambda}(X_1, \ldots, X_n) = \overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n),$$

and then define our estimator for p_0 to be

$$\hat{p_0}(X_1, \ldots, X_n) = e^{-\hat{\lambda}} = e^{-\overline{X}}.$$

Prove that $\hat{p_0}$ is a biased estimator of p_0. Follow these steps:
  (i) What does Jensen's inequality (Theorem 5.23) tell us about $\mathbf{E}\left[\hat{p_0}\right]$ as compared to p_0?
  (ii) Prove that $\mathbf{E}\left[\hat{p_0}\right] = e^{-n\lambda(1-e^{-1/n})}$. [Hint: Recall $X_i \sim \text{Poisson}(\lambda)$. What does this say about the distribution of $X_1 + X_2 + \cdots + X_n$?]
  (iii) Show that $\mathbf{E}\left[\hat{p_0}\right]$ converges to p_0 from above as $n \to \infty$.

(b) An alternative idea is to look at the average fraction of minutes with 0 arrivals and use that as our estimator. That is,

$$\hat{p_0}^{\text{alt}}(X_1, \ldots, X_n) = \frac{\text{number of } X_i \text{ equal to } 0}{n}.$$

Prove that $\hat{p_0}^{\text{alt}}$ is an unbiased estimator of p_0.

15.7 Acknowledgment

This chapter was written in collaboration with Weina Wang, who was a major contributor to the chapter contents and the exercises.