Unit 2 Foundations of Inference

Parametric Statistical Models

Advantages of Modeling Assumptions

1/1 point (graded)

As in the video above, a population consists of n individuals labeled $1, 2, \ldots, n$. Let X_i denote the number of siblings of individual i. We assume that $X_1, \ldots, X_n$ are **i.i.d.** (independent and identically distributed) as some random variable X. You are deciding between using one of two possible different models for the random variable X:

Model 1: X is distributed as $\text{Poiss}(\lambda)$ for some unknown $\lambda > 0$.

Model 2: X takes values in $\{1, 2, 3, 4, 5, 6, \geq 7\}$, and for $i = 1, 2, \ldots 7$, we let p_i denote the (unknown) probability that $X = i$. Here "≥ 7" is a placeholder for when the number of siblings is at least 7. For example, we do not distinguish between an individual having 7 siblings or 10 siblings in this model.

Which one of the following **best** describes an advantage of using a Poisson distribution (Model 1) over the distribution in Model 2 to model X?

- It allows us to model the data continuously.
- It allows individuals to have an arbitrarily large number of siblings.
- ✔ It reduces the amount of unknowns needed for modeling.

Solution:

Option 1 requires us to find the value of one unknown, λ, to specify the distribution of X. With Option 2, it is required to find 7 unknowns (all of the p_i's) to specify the distribution. Option 1 requires less information and is hence a simpler modeling task.
The first choice, "It allows us to model the data continuously.", is incorrect because the Poisson distribution is a discrete model, so it does not model the distribution continuously. Note that our data is discrete, so it makes sense to model this data with a discrete distribution. Both distributions in Option 1 and 2 are discrete.
The second choice, "It allows individuals to have an arbitrarily large number of children.", is a disadvantage of selecting Option 1 because we would never expect an individual to have, say, 200 siblings. But the Poisson model allows this to happen!
Remark: While the focus of this class is not on modeling, it is good to keep the following principle in mind: some models may perform better than others, but there is no such thing as *THE correct* model. The task of a statistician is to use reasonable assumptions to find a tractable model that gives useful approximations to a given data set.

Modelling a Binary Data Set

1/1 point (graded)

You would like to determine the percentage of coffee drinkers in your university, and collected the following binary data set from random students on campus, 1 for coffee drinker and 0 for otherwise:

$$0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1.$$

Let Y_i denote the i'th number in this list. You decide to model this data set under the following assumptions:

- $Y_1, \ldots, Y_n$ are **identically distributed** as some random variable Y.
- $Y_1, \ldots, Y_n$ are **independent**.
- Y_i only takes the value 0 or 1.

Under these assumptions, how many unknowns are needed to specify the distribution of Y?

1 ✔ **Answer:** 1

Solution:

A random variable that takes values only 0 or 1 is necessarily a Bernoulli random variable. Hence, only the mean (*i.e.* the probability that $Y_i = 1$) is needed to specify the distribution.

Approximating the unknown parameter

1/1 point (graded)

As above, let $Y_1, \dots, Y_n$ denote the i'th number in the binary data set.

Recall that $Y_1, \dots, Y_n$ are assumed to be independent and identically distributed (**i.i.d.**) as some distribution Y. In the future, we will abbreviate this assumption with the notation $Y_1, \dots, Y_n \overset{iid}{\sim} Y$.

Which of the following converges to $\mathbb{E}[Y_i] = \mathbb{E}[Y]$ as $n \to \infty$?
(Choose all that apply.)

- [x] $\frac{\text{total number of 1's}}{n}$
- [ ] Y_n
- [ ] $\text{Median}(Y_1, \dots, Y_n)$
- [x] $\frac{1}{n}\sum_{i=1}^{n} Y_i$

✔

Solution:

Note that $\frac{\text{total number of 1's}}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$: these two expression are equal. By the law of large numbers, both converges to $\mathbb{E}[Y]$ $(= \mathbb{E}[Y_i])$ as $n \to \infty$.
Remark: In this problem, we did not stress the type of convergence. For this example of Bernoulli random variables, the conclusion holds for both convergence in probability (weak convergence) and convergence almost surely (strong convergence). You are encouraged to review the types of convergence in Chapter 1.

A Basic Statistical Model: Sample space

1/1 point (graded)

You have a coin that either lands heads, which you denote by 1, or tails, which you denote by 0. Let X be a random variable representing this coin flip, with an (unknown) distribution. You run a **statistical experiment** consisting of n iid tosses of the coin and record your data set as $X_1, X_2, X_3, \dots X_n$.

(It makes sense to assume the coin tosses $X_1, \dots, X_n$ as identically distributed, since we always toss the same coin; and as independent, since these tosses do not affect each other.)

We now construct a **statistical model** $\left(E, \{P_\theta\}_{\theta \in \Theta}\right)$ associated with this experiment, where

$$\begin{aligned} E &\text{ is a sample space for } X\text{, i.e. a set that contains all possible outcomes of } X, \\ \{\mathbb{P}_\theta\}_{\theta \in \Theta} &\text{ is a family of probability distributions on } E, \\ \Theta &\text{ is a parameter set, i.e. a set consisting of some possible values of } \theta. \end{aligned}$$

What is the **smallest sample space** for X? We can use this as the sample space E in our statistical model.
(Below, $[0, 1]$ denotes the closed interval between 0 and 1. In contrast, $\{0, 1\}$ denotes the set with two elements, 0 and 1.)

- (•) $\{0, 1\}$
- ( ) $[0, 1]$
- ( ) $\mathbb{R}$
- ( ) $\mathbb{R}^2$

✔

Solution:

Here the coin is either heads (denoted by 1) or tails (denoted by 0), so $\{0, 1\}$ is the smallest sample space of X. The remaining choices are valid, but not the smallest, sample spaces of X.

A Basic Statistical Model: Family of distributions and Parameter set

2/2 points (graded)

Continuing from the previous problem, which of the following is the smallest family of probability distributions that the distribution of X belongs to? We can use this family as $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ in our statistical model.

- (●) Bernoulli
- ( ) Poisson
- ( ) Binomial

✔

The distribution of X is a member of the family with some unknown parameter θ. According to the information given about the experiment, which of the following represents the set of all possible values of the parameter θ? We can use this set as the parameter set Θ in our statistical model.

- ( ) $\{0, 1\}$
- ( ) $\{0, 1/2, 1\}$
- (●) $[0, 1]$
- ( ) $\mathbb{R}$

✔

Solution:

1. Since the (smallest) sample space of X is $\{0, 1\}$, X follows a Bernoulli distribution.

2. The first and second choices, $\{0, 1\}$ and $\{0, 1/2, 1\}$, place too many restrictions on the distribution of X. Also, be sure to not confuse the space where the parameter θ lives with the sample space, where the random variable X lives! The fourth choice, $\mathbb{R}$, allows for values of θ that do not make sense according to modeling X as $\text{Ber}(\theta)$. For example, there is no such thing as $\text{Ber}(-1/2)$.
   We are not given any assumptions on the distribution of the coin, so we need to allow θ to take all possible values that make sense according to our modeling assumption. Since θ represents the probability that $X = 1$, we must have $0 \leq \theta \leq 1$. Hence, the third choice, $[0, 1]$, is correct.

Using this problem and the previous one, we can construct the statistical model $(\{0, 1\}, \{\text{Ber}(\theta)\}_{\theta\in[0,1]})$ for the distribution of the RV X representing the outcome of the coin flip.

Review: Sample Spaces of Distributions

4/4 points (graded)

Recall that a **sample space** of a random variable X is a set that contains all possible outcomes of X.

Note that the sample space of X is *not unique*. For example, if $X \sim \text{Ber}(p)$, then both $\{0, 1\}$ and $\mathbb{R}$ can serve as a sample space. However, in general, we associate a random variable with its smallest possible sample space (which would be $\{0, 1\}$ if $X \sim \text{Ber}(p)$).

Find the **smallest sample space** for each of the following random variables.

$X_1 \sim \text{Poiss}(\lambda)$, a **Poisson** random variable with parameter λ:

- ◯ $\{0, 1\}$
- ◉ $\{x \in \mathbb{Z} : x \geq 0\}$
- ◯ $[0, \infty)$
- ◯ $(-\infty, \infty)$

✔

$X_2 \sim \mathcal{N}(0, 1)$, a **standard Gaussian (or normal)** random variable with mean 0 and variance 1:

- ◯ $\{0, 1\}$
- ◯ $\{x \in \mathbb{Z} : x \geq 0\}$
- ◯ $[0, \infty)$
- ◉ $(-\infty, \infty)$

$X_3 \sim \exp(\lambda)$, an **exponential** random variable with parameter $\lambda > 0$:

- ◯ $\{0, 1\}$
- ◯ $\{x \in \mathbb{Z} : x \geq 0\}$
- ◉ $[0, \infty)$
- ◯ $(-\infty, \infty)$

✔

$X_4 \sim \mathcal{I}(Y > 0)$ where Y is standard Gaussian and $\mathcal{I}$ is the **indicator function**.

Recall the definition of the indicator function is:

$$\mathcal{I}(Y > 0) = \begin{cases} 1 & \text{if } Y > 0 \\ 0 & \text{if } Y \leq 0. \end{cases}$$

- ◉ $\{0, 1\}$
- ◯ $\{x \in \mathbb{Z} : x \geq 0\}$
- ◯ $[0, \infty)$
- ◯ $(-\infty, \infty)$

✔

Solution:

- A Poisson random variable is discrete and can take values on all non-negative integers.
- Gaussian random variables can take any real value.
- The Exponential distribution is continuous and is restricted to all non-negative real values.
- The final random variable is an indicator, so it must take values in $\{0, 1\}$. Note that X_4 is in fact Bernoulli.

Z means all integers

Statistical Model Definition Concept check

1/1 point (graded)

Which of the following is a statistical model?

- ◯ $\left(\{1\}, (\text{Ber}\,(p))_{p\in(0,1)}\right)$
- ◉ $\left(\{0,1\}, (\text{Ber}\,(p))_{p\in(0.2,0.4)}\right)$
- ◯ Both of the above
- ◯ None of the above

✔

Solution:

Solution in video below.
The set $\{1\}$ is not the sample space of the distribution $\text{Ber}\,(p)$, so the first choice $\left(\{1\}, (\text{Ber}\,(p))_{p\in(0,1)}\right)$ is not a statistical model. On the other hand, $\left(\{0,1\}, (\text{Ber}\,(p))_{p\in(0.2,0.4)}\right)$ is a valid statistical model.
Remark: In the model $\left(\{0,1\}, (\text{Ber}\,(p))_{p\in(0.2,0.4)}\right)$, the parameter p is restricted to be in the interval $(0.2, 0.4)$. Such a restriction is perfectly valid, and can be useful for performing modeling tasks.

A Non-Example of a Statistical Model

0 points possible (ungraded)

(This problem is strictly pedagogical and is ungraded.)

Let $\mathcal{U}\,([0,a])$ denote the uniform distribution on the interval $[0,a]$. Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{U}\,([0,a])$ for some unknown $a > 0$. Which one of the following is *not* a statistical model associated with this statistical experiment?

- ◉ $\left([0,a], (\mathcal{U}\,([0,a]))_{a>0}\right)$ ✔
- ◯ $\left(\mathbb{R}_+, (\mathcal{U}\,([0,a]))_{a>0}\right)$
- ◯ Neither choice above is a statistical model.

Solution:

See video below.
The first choice $\left([0,a], (\mathcal{U}\,([0,a]))_{a>0}\right)$ is not a statistical model because the sample space, as written, depends on an unknown parameter a.
The second choice $\left(\mathbb{R}_+, (\mathcal{U}\,([0,a]))_{a>0}\right)$ is a statistical model because for any value of a, the random variables $X_1, \ldots, X_n$ will have sample space contained in the interval $[0,\infty) = \mathbb{R}_+$.

Sample space should NOT depend on the parameter otherwise it's not a sample space

8. Exercises on Statistical models

Bookmark this page

A Parametric Model for Rock Samples

2/2 points (graded)

You are testing out a new scale that measures weights. To do so, you collect a particular rock and take n measurements, using the same rock each time. Let X_i denote the ith measurement of a particular rock.

Based on prior knowledge, you expect your data $X_1, \ldots, X_n$ to consist of **i.i.d.** (independent and identically distributed) samples from a Gaussian distribution with unknown mean $\mu > 0$. The scale that you are using to weigh the sample comes with a guarantee from the manufacturer that the variance of your data set will be 0.23. Given this information, your goal is to write down a statistical model $(E, \{P_\theta\}_{\theta \in \Theta})$ for this statistical experiment.

Which of the following is (are)

(1) a **formally** valid statistical model(s)? (Choose all that apply.)

- [x] $((-\infty, \infty), \{N(\mu, 0.23)\}_{\mu>0})$
- [x] $((-\infty, \infty), \{N(\mu, 0.23)\}_{\mu \in \mathbb{R}})$
- [x] $((-\infty, \infty), \{N(\mu, \sigma^2)\}_{\mu, \sigma^2>0})$
- [x] $((-\infty, \infty), \{N(\mu, \sigma^2)\}_{\mu \in \mathbb{R}, \sigma^2>0})$

(2) the statistical model that **best** incorporates all known information?

- [x] $((-\infty, \infty), \{N(\mu, 0.23)\}_{\mu>0})$
- [ ] $((-\infty, \infty), \{N(\mu, 0.23)\}_{\mu \in \mathbb{R}})$
- [ ] $((-\infty, \infty), \{N(\mu, \sigma^2)\}_{\mu, \sigma^2>0})$
- [ ] $((-\infty, \infty), \{N(\mu, \sigma^2)\}_{\mu \in \mathbb{R}, \sigma^2>0})$

Solution:

1. All of the above choices are valid statistical models. The sample space of a Gaussian is $(-\infty, \infty)$. In the first and second choice, $((-\infty, \infty), \{N(\mu, 0.23)\}_{\mu>0})$ and $((-\infty, \infty), \{N(\mu, 0.23)\}_{\mu \in \mathbb{R}})$ use the mean μ to parametrize the Gaussian. In the third and fourth choice, $((-\infty, \infty), \{N(\mu, \sigma^2)\}_{\mu, \sigma^2>0})$ and $((-\infty, \infty), \{N(\mu, \sigma^2)\}_{\mu \in \mathbb{R}, \sigma^2>0})$, parametrize the Gaussian distribution by the mean and variance. Since these choices restrict $\sigma^2 > 0$, both are valid statistical models.

2. For the purposes of modeling, in general it is best to choose the statistical model that incorporates all known information about the sample. Usually this reduces the amount of unknowns in the model or the size of the parameter space. Since we are given that the data is Gaussian, the variance is 0.23, and the mean μ, is positive, it makes sense to incorporate this information into the model. Note that choice 1 $((-\infty, \infty), \{N(\mu, 0.23)\}_{\mu>0})$ uses everything that we are given in the problem set-up, so it is likely the best choice of statistical model in this scenario.

Parametric vs. Nonparametric Models

0/1 point (graded)

A statistical model $(E, \{P_\theta\}_{\theta \in \Theta})$ is **parametric** if all parameters $\theta \in \Theta$ can be specified by a **finite** number of unknowns. Equivalently, this means that Θ is a subset of $\mathbb{R}^m$. In particular, if $\Theta \subset \mathbb{R}^m$, then P_θ is uniquely specified by the m entries of the vector θ.

Which of the following statistical models are parametric?
(Choose all that apply.)

- [x] $E = \{x \in \mathbb{Z} : x \geq 0\}$;
  $\{P_\theta\}_{\theta \in \Theta}$ is the set of all probability distributions with the sample space $\{x \in \mathbb{Z} : x \geq 0\}$.

- [ ] $E = \{0, 1\}$;
  $\{P_\theta\}_{\theta \in [0,1]} = \{\text{Ber}(\theta)\}_{\theta \in [0,1]}$. ✔

- [x] $E = (-\infty, \infty)$;
  $\{P_{\sigma^2}\}_{\sigma^2 \in (0,\infty)}$ is the set of all centered (mean 0) Gaussian distributions $N(0, \sigma^2)$ where $\sigma^2 > 0$. ✔

- [ ] $E = \{1, 2, 3, 4\}$;
  $\{P_{(p_1,p_2,p_3,p_4)}\}_{(p_1,p_2,p_3,p_4) \in S}$ is defined in terms of
  - S: the set of all $(p_1, p_2, p_3, p_4) \in \mathbb{R}^4$ such that $0 \leq p_i \leq 1$ for all $i = 1, \ldots, 4$ and $\sum_{i=1}^{4} p_i = 1$;
  - $P_{(p_1,p_2,p_3,p_4)}$: the distribution defined by setting the probability of outcome i to be p_i.

  ✔

- [x] $E = (-\infty, \infty)$;
  $\{P_{(\mu,\sigma^2)}\}_{(\mu,\sigma^2) \in \mathbb{R} \times (0,\infty)}$ is the set of all Gaussian distributions $N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. ✔

- [x] $E = (0, \infty)$;
  $\{P_\theta\}_{\theta \in [0,\infty)} = \{\mathcal{U}([0, \theta])\}_{\theta \in (0,\infty)}$ is the set of all uniform distributions on the interval $[0, \theta]$ with $\theta > 0$. ✔

- [x] $E = [0, 1]$;
  $\{P_\theta\}_{\theta \in \Theta}$ is the set of all probability distributions given by a probability density function $f : \mathbb{R} \to \mathbb{R}$ with f continuous and $\int_0^1 f(x)\, dx = 1$.

Solution:

- "$E = \{x \in \mathbb{Z} : x \geq 0\}$ and $\{P_\theta\}_{\theta \in \Theta}$ is the set of all probability distributions with sample space $\{x \in \mathbb{Z} : x \geq 0\}$.", specifying the distribution requires us to know the probability of the outcome i for all $i \in \mathbb{Z}$ such that $i \geq 0$. An infinite amount of information (or unknowns) is required, so this statistical model is non-parametric.

- "$E = \{0, 1\}$ and $\{P_\theta\}_{\theta \in [0,1]} = \{\text{Ber}(\theta)\}_{\theta \in [0,1]}$.",
  "$E = (-\infty, \infty)$ and $\{P_{\sigma^2}\}_{\sigma^2 \in (0,\infty)}$ is the set of all centered (mean 0) Gaussian distributions...", and
  "$E = (-\infty, \infty)$ and $\{P_\theta\}_{\theta \in [0,\infty)} = \{\mathcal{U}([0, \theta])\}_{\theta \in [0,\infty)}$...," respectively, all require just a single unknown to specify the distribution. These models are parametric.

- The choice "$E = \{1, 2, 3, 4\}$ and $\{P_{(p_1,p_2,p_3,p_4)}\}_{(p_1,p_2,p_3,p_4) \in S}$..." requires three unknowns to specify the distribution (once p_1, p_2, and p_3 are specified, p_4 is uniquely determined). This model is parametric. It would remain parametric even if one said, "there are four unknowns, p_1, p_2, p_3, p_4".

- The choice "$E = (-\infty, \infty)$ and $\{P_{(\mu,\sigma^2)}\}_{(\mu,\sigma^2) \in \mathbb{R} \times (0,\infty)}$ is the set of all Gaussian distributions $N(\mu, \sigma^2)$..." requires only the specification of the mean and variance, so it is also parametric.

- Similarly, in the last choice "$E = [0, 1]$ and $\{P_\theta\}_{\theta \in \Theta}$ is the set of all probability distributions given by a probability density function $f : \mathbb{R} \to \mathbb{R}$...", the space of continuous density functions cannot be specified by a finite amount of information; you would need to know the values of the function on an infinite subset of $[0, 1]$ to be able to uniquely determine it. Hence, this statistical model is also non-parametric.

Statistical Model for a Censored Exponential

0/1 point (graded)

Let X denote an exponential random variable with unknown parameter $\lambda > 0$. Let $Y = \mathcal{I}(X > 5)$, the indicator that X is larger than 5.

Recall the definition of the indicator function here is

$$\mathcal{I}(X > 5) = \begin{cases} 1 & \text{if } X > 5 \\ 0 & \text{if } X \leq 5. \end{cases}$$

We think of Y as a **censored** version of the Exponential random variable X: we cannot directly observe X, but we are able to gather some information about it (in this case, whether or not X is larger than 5.)

Observe that Y is a Bernoulli random variable. Thus, the statistical model for Y can be written $(\{0, 1\}, \{\text{Ber}(f(\lambda))\}_{\lambda>0})$ for some function f of λ. What is $f(\lambda)$?

(Type **lambda** for λ. Use the help button below for help with formula input).

$f(\lambda) =$ lambda^2 ✖ **Answer:** e^(-5*lambda)

λ^2

STANDARD NOTATION

Solution:

Note that $Y = 1$ if and only if $X > 5$. Hence, we need to compute the probability that $X > 5$. Recall that the density of $\text{Exp}(\lambda)$ is given by $\lambda e^{-\lambda x}$. We just need to compute

$$P(X > 5) = \int_5^\infty \lambda e^{-\lambda x}\, dx = e^{-5\lambda}.$$

We conclude that if $X \sim \text{Exp}(\lambda)$, then $Y \sim \text{Ber}(e^{-5\lambda})$. Hence, $f(\lambda) = e^{-5\lambda}$.

Linear regression as a statistical model I

1/2 points (graded)

Consider the linear regression model introduced in the slides and lecture, restated below:

Linear regression model : $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ are i.i.d from the linear regression model $Y_i = \beta^\top X_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$ for an unknown $\beta \in \mathbb{R}^d$ and $X_i \sim \mathcal{N}_d(0, I_d)$ independent of ε_i.

Suppose that $\beta = \mathbf{1} \in \mathbb{R}^d$, which denotes the d-dimensional vector with all entries equal to 1.

What is the mean of Y_1?

$\mathbb{E}[Y_1] =$ 0 ✔ **Answer:** 0

What is the variance of Y_1? (Express your answer in terms of d.)

$\text{Var}(Y_1) =$ d ✖ **Answer:** d+1

d

Solution:

By definition of the model and setting $\beta = \mathbf{1}$, we have

$$Y_1 = \beta^T X_1 + \varepsilon_1 = \mathbf{1}^T X_1 + \varepsilon_1 = \varepsilon_1 + \sum_{j=1}^{d} X_{1,j}.$$

where $X_{i,j}$ denotes the j'th coordinate of $X_i \sim \mathcal{N}(0, I_d)$. By linearity of expectation,

$$\mathbb{E}[Y_1] = \mathbb{E}[\varepsilon_1] + \sum_{j=1}^{d} \mathbb{E}[X_{1,j}] = 0$$

Next we compute the variance. Since $X_{1,1}, \ldots, X_{1,d}, \varepsilon_i$ are mutually independent, the variance is additive:

$$\mathrm{Var}[Y_1] = \mathrm{Var}[\varepsilon_1] + \sum_{j=1}^{d} \mathrm{Var}[X_{1,j}] = d + 1$$

because $X_{1,1}, \ldots, X_{1,d}, \varepsilon_1 \overset{iid}{\sim} N(0, 1)$.

Linear regression as a statistical model II

2/2 points (graded)

Recall the linear regression model as introduced above in the previous question. This model is parametric, although it is not written in the standard notation previously introduced for parametric statistical models. In this problem, you will explicitly write the linear regression model as a parametric statistical model.

We will represent the linear regression model as an ordered pair $(E, \{P_\beta\}_{\beta \in \Theta})$. Here E denotes the sample space associated to the distribution P_β, where P_β is defined as follows for $\beta \in \mathbb{R}^d$:

The random ordered pair $(X, Y) \subset \mathbb{R}^d \times \mathbb{R}$ is distributed as P_β if:

- $X \sim N(0, I_d)$,
- $Y \sim \beta^T X + \varepsilon$, *where* $\varepsilon \sim N(0, 1)$ *and* ε *is independent of* X.

The set Θ in the ordered pair $(E, \{P_\beta\}_{\beta \in \Theta})$ denotes the parameter space for this model.

The sample space for the linear regression model can be written $E = \mathbb{R}^k$ for some integer k. What is k? (Express your answer in terms of d.)

Hint: You should use the fact that $\mathbb{R}^{m+n} = \mathbb{R}^m \times \mathbb{R}^n$ for all integers $m, n \geq 0$.

$k =$ d+1 ✔ **Answer:** d+1

$d + 1$

The parameter space for the model can be written as $\Theta = \mathbb{R}^j$ for some integer j. What is j? (Express your answer in terms of d.)

$j =$ d ✔ **Answer:** d

d

Solution:

The statistical experiment is given by the iid sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Where $X_i \sim N(0, I_d)$ and $Y_i = \beta^T X_i + \varepsilon_i$ for $\varepsilon_i \sim N(0, 1)$ and some true parameter $\beta \in \mathbb{R}^d$. In particular, $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. Therefore, $(X_i, Y_i) \in \mathbb{R}^{d+1}$, so indeed $E = R^{d+1}$ is the sample space for this model. We conclude that $k = d + 1$.

This model is parametrized by the vector $\beta \in \mathbb{R}^d$. That is, specifying the value of β uniquely determines the distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$. Hence, the parameter is β, and the parameter space is $\Theta = \mathbb{R}^d$. We conclude that $j = d$.

Preparation: Injectivity

1/1 point (graded)

The notation $f : S \to T$ denotes that f is a function, also called a **map** , defined on all of a set S and whose outputs lie in a set T. A function $f : S \to T$ is **injective** if for all $x, y \in S$, $f(x) = f(y)$ implies that $x = y$.

Alternatively: a function is injective if we can **uniquely** recover some input x based on an output $f(x)$.

Which of the following functions are injective? (Choose all that apply.)

- [x] $f_1 : \mathbb{R} \to \mathbb{R}$, given by $f_1(x) = x$.
- [ ] $f_2 : \mathbb{R} \to \mathbb{R}$, given by $f_2(x) = x^2$.
- [ ] $f_3 : \mathbb{R} \to \mathbb{R}$, given by $f_3(x) = \sin(x)$.
- [x] $f_4 : [0, 1] \to \{\text{probability distributions on } \{0, 1\}\}$, given by $f_4(p) = \text{Ber}(p)$.

✔

Solution:

The first choice $f_1(x) = x$ is the identity function, so if $f_1(x) = f_1(y)$, then $x = y$ by definition of f_1. So f_1 is injective.
The second choice $f_2(x) = x^2$ is not injective because, for example, both $+1$ and -1 map to the same value, 1, after applying f_2. In general, if $f_2(x) = c$ for some constant $c > 0$, then there are two possible choices for x: either $x = \sqrt{c}$ or $x = -\sqrt{c}$.
The third choice $f_3(x) = \sin(x)$ is not injective. In fact, there are infinitely many points x such that $f_3(x) = 0$. Recall from trigonometry that all values in the set $\{2\pi x : x \in \mathbb{Z}\}$ will map to 0 after applying f_3.
The fourth choice $f_4(p) = \text{Ber}(p)$ is injective: if $p \in [0, 1]$, then $f_4(p) = \text{Ber}(p)$, so that p specifies the probability that $X \sim \text{Ber}(p)$ is equal to 1. Since a distribution on $\{0, 1\}$ is uniquely determined by $P(X = 1)$, the map f_4 is injective.

Identifiability of Statistical Models

1/1 point (graded)

Let $\{P_\theta\}_{\theta\in\Theta}$ denote a family of distributions that depends on an unknown parameter $\theta \in \Theta$.

Recall that the parameter θ is **identifiable** if the map $\theta \mapsto P_\theta$ is injective. Here, the notation $\theta \mapsto P_\theta$ denotes a function that takes as input $\theta \in \Theta$ and outputs a probability distribution P_θ. In other words, if $\theta \neq \theta'$ (and both in Θ), then $P_\theta \neq P_{\theta'}$

Which of the following families of distributions has an identifiable parameter? (Choose all that apply.)

- [x] $\{\text{Ber}(p)\}_{p\in[0,1]}$
- [ ] $\{\text{Ber}(p^2)\}_{p\in[-1,1]}$
- [x] $\{\text{Ber}(\sin(p))\}_{p\in[0,\frac{\pi}{2}]}$
- [ ] $\{\text{Ber}(\sin(p))\}_{p\in[0,\pi]}$

✔

Solution:

Remark: A family of distributions $\{\text{Ber}(f(p))\}_{p\in S}$ (here $S \subset \mathbb{R}$ is a set where the parameter p lives) has the parameter p identified if and only if the function $f(p)$ is injective.
The function $f(p) = p$ is injective on the interval $[0, 1]$, so the first choice $\{\text{Ber}(p)\}_{p\in[0,1]}$ is correct. However, the function $f(p) = p^2$ on the interval $[-1, 1]$ is not injective, so the second choice $\{\text{Ber}(p^2)\}_{p\in[-1,1]}$ is incorrect.
Let's look more carefully at the last two choices, $\{\text{Ber}(\sin(p))\}_{p\in[0,\frac{\pi}{2}]}$ and $\{\text{Ber}(\sin(p))\}_{p\in[0,\pi]}$. Observe that the function $f(p) = \sin(p)$ *is* injective on the interval $[0, \frac{\pi}{2}]$ but *is not* injective on the interval $[0, \pi]$. Hence, $\{\text{Ber}(\sin(p))\}_{p\in[0,\frac{\pi}{2}]}$ has an identified parameter, but $\{\text{Ber}(\sin(p))\}_{p\in[0,\pi]}$ does not have an identified parameter.

11. Identifiability exercises

Bookmark this page

Identifiability of Statistical Models 2

1/1 point (graded)

Let $X_i = Y_i^2$ where $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{U}([0, a])$ for some unknown parameter a. We observe the i.i.d. samples $X_1, \ldots, X_n$, but not the Y_i's themselves.
Hint: Compute the cdf of X_i.
Is the parameter a identifiable from the common distribution the X_i's?

- (•) Yes
- ( ) No

✔

Solution:

Write $X_i \sim X$ and note that X is supported on the interval $[0, a^2]$. Let us compute the CDF of X in terms of a.

$$\mathbf{P}(X \le t) = \mathbf{P}(Y \le \sqrt{t}) = \min\left(\int_0^{\sqrt{t}} \frac{1}{a}\, dy, 1\right) = \min\left(\frac{\sqrt{t}}{a}, 1\right).$$

For different values of a, the CDF of X are different; hence a is identifiable.

Identifiability of Statistical Models 3

1/1 point (graded)

Let $X_i = \mathcal{I}(Y_i \geq a/2)$ where $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{U}([0, a])$ for some unknown parameter a. We observe the independent samples $X_1, \ldots, X_n$ but not the Y_i's themselves.

Is the parameter a identifiable from the common distribution of the X_i's?

- ○ Yes
- ◉ No ✔

Solution:

Note that X is a Bernoulli random variable with parameter $p := P\left(\mathcal{I}\left(Y_i \geq \frac{a}{2}\right) = 1\right) = P\left(Y_i \geq \frac{a}{2}\right)$.
For any choice of a, we have by the distribution of Y_i that $p = P(Y_i \geq a/2) = 1/2$. Hence, for any choice of a, the random variable X is distributed as $\text{Ber}(1/2)$. The parameter a is not identifiable.

Review of terminology

0/1 point (graded)

You have access to samples $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}$. You construct a statistical model $((-\infty, \infty), \{P_\theta\}_{\theta \in \mathbb{R}})$ for this statistical experiment. Imagine that somehow you are able to figure out the true distribution $\mathbb{P}$ and you realize that $\mathbb{P} = P_{\theta^*}$ for some particular parameter value $\theta^* \in \mathbb{R}$.

Your goal is to uncover the true parameter θ^*. Which assumptions below (individually, each on its own) are sufficient to recover the true parameter θ^* from the distribution?
(Choose all that apply.)

- ☐ There is another value $\theta' \in \mathbb{R}$ such that $\theta' \neq \theta^*$ but P_{θ^*} and $P_{\theta'}$ are the same distribution.
- ☑ The given statistical model $((-\infty, \infty), \{P_\theta\}_{\theta \in \mathbb{R}})$ is well-specified.
- ☑ The parameter θ is identifiable for the given statistical model. ✔

✖

Solution:

The third choice, "The parameter θ is identified for the given statistical model.", is correct. If θ is identified, then the map $\theta \mapsto P_\theta$ is injective. Hence, given the output P_{θ^*}, which is the true distribution, we can uniquely recover the true parameter θ^*.

The first choice, "There is another value $\theta' \in \mathbb{R}$ such that $\theta' \neq \theta^*$ but P_{θ^*} and $P_{\theta'}$ are the same distribution.", is incorrect because this implies that the parameter θ is *not* identified. This implies that by only knowing the distribution P_{θ^*}, we have no way of saying if θ' or θ^* is the true parameter.

Recall that a statistical model $(E, \{P_\theta\}_{\theta \in \Theta})$ associated to a statistical experiment $X_1, \ldots, X_n$ is **well-specified** if there exists $\theta^* \in \Theta$ such that $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta^*}$. Note that the problem statement implies that our model is well-specified. However, this assumption is not enough to be able to recover the true parameter θ^* from the distribution P_{θ^*} because the parameter θ may not be identified.

Parametric Estimation and Confidence Intervals

Which Statistics are Estimators?

1/1 point (graded)

Let $X_1, \dots, X_n \overset{iid}{\sim} P_\theta$ where the distribution P_θ depends on an unknown parameter $\theta \in \mathbb{R}$. Which of the following statistics are considered **estimators** ?
(Choose all that apply.)

- [ ] θ
- [x] 4.2
- [x] $\sum_{i=1}^{n} i^2 X_i^i$
- [x] $\frac{1}{n}\sum_{i=1}^{n} X_i$
- [ ] $\frac{1}{n}\sum_{i=1}^{n} X_i - \theta$

✔

Solution:

Recall that a statistic, loosely speaking, is a function of the data that can be easily computed. Recall that an estimator is a special kind of statistic, specifically one that does not depend on the unknown parameter θ.

Note that the first and last choices, θ and $\frac{1}{n}\sum_{i=1}^{n} X_i - \theta$, both have some explicit dependence of θ, so they cannot be estimators.

On the other hand, the remaining expressions 4.2, $\sum_{i=1}^{n} i^2 X_i^i$, and $\frac{1}{n}\sum_{i=1}^{n} X_i$ only depend on $X_1, \dots, X_n$ and known constants (but not θ), so they are indeed estimators.
Remark: The second estimator, 4.2, is potentially a very poor choice as it does not depend on the data set. But formally (that is according to the definition), it is still considered an estimator and it has very low variance!

Consistency of an Estimator

1/1 point (graded)

An estimator $\hat{\theta}_n$ is **weakly consistent** if $\lim_{n\to\infty} \hat{\theta}_n = \theta$, where the convergence is in probability.

Suppose that in the previous problem the unknown parameter θ is the common mean of $X_1, \ldots, X_n$. Assume that $\theta \neq 4.2$. Which of the following is a weakly consistent estimator for θ? (Choose all that apply.)

- [ ] θ
- [ ] 4.2
- [ ] $\sum_{i=1}^{n} i^2 X_i^i$
- [x] $\frac{1}{n}\sum_{i=1}^{n} X_i$
- [ ] $\frac{1}{n}\sum_{i=1}^{n} X_i - \theta$

✔

Solution:

By the weak law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}[X_1] = \theta$ in probability, so this is the correct choice.

From the previous question, the first and last choice, θ and $\frac{1}{n}\sum_{i=1}^{n} X_i - \theta$, are not even estimators, so these options are incorrect. Since $\theta \neq 4.2$, this estimator cannot be consistent. Finally, there is no guarantee that $\sum_{i=1}^{n} i^2 X_i^i$ converges to θ. In fact, for many choices of distribution, this statistic will diverge to ∞.

Quantifying Consistency

0/1 point (graded)

Note: The problem statement has been changed to asking about convergence in probability instead of almost surely. Attempts will be reset.

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber(p)}$. Let $\overline{X}_n$ be the estimator given by $\frac{1}{n}\sum_{i=1}^{n} X_i$.

What is the smallest constant c such that

$$n^c\left(\overline{X}_n - p\right) = n^c\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p\right)$$

does **not** converge to 0 in probability as $n \to \infty$?

0 ✖ **Answer:** .5

Solution:

Let $\sigma = \sqrt{p(1-p)}$ denote the common standard deviation of $X_1, \ldots, X_n$. By the central limit theorem,

$$\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - p\right) = \frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p\right) \to N(0,1)$$

where the convergence is in distribution. As a result, we see that for $c < 1/2$,

$$n^c\left(\overline{X}_n - p\right) = \frac{\sigma}{n^{1/2-c}}\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - p\right) \approx \frac{\sigma}{n^{1/2-c}}N(0,1) \to 0$$

in probability as $n \to \infty$. Hence, $c = 1/2$ is the smallest possible value of c such that

$$n^c\left(\overline{X}_n - p\right) = n^c\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p\right)$$

does **not** converge to 0 in probability as $n \to \infty$.

Remark: As defined in the third video in this section, this implies that the estimator $\overline{X}_n$ is $\sqrt{n}$-consistent. This means that the estimator $\overline{X}_n$ converges to the true parameter at a relatively fast rate, so this gives us something stronger than just consistency.

The Expectation of the Average

1/1 point (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{U}([a, a+1])$ where a is an unknown parameter. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote the sample mean. In terms of a, what is $\mathbb{E}[\overline{X}_n]$?

$\mathbb{E}\left[\overline{X}_n\right] =$ a+0.5 ✔ **Answer:** a+1/2

$a + 0.5$

Solution:

Note that since the X_i's are identically distributed, by linearity of expectation,

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[X_1] = a + \frac{1}{2}.$$

Computing Bias

1/1 point (graded)

Recall: Let $\hat{\theta}_n$ denote an estimator for a true parameter θ. Here n specifies the sample size. The **bias** of $\hat{\theta}_n$ is defined to be

$$\mathbb{E}\left[\hat{\theta}_n\right] - \theta.$$

Let $X_1, \ldots, X_n$ be defined as in the previous question. Compute the bias of the estimator $\overline{X}_n$ with respect to the parameter a.

0.5 ✔ **Answer:** .5

Solution:

The bias is given by $\mathbb{E}\left[\overline{X}_n\right] - a = 1/2$, where we applied the previous part. Note that this implies that $\overline{X}_n - \frac{1}{2}$ is an unbiased estimator.

(Optional) Expectation of nonlinear functions and Jensen's Inequality

0 points possible (ungraded)

Let X be a positive random variable with expectation λ. How does $\mu = \mathbb{E}\left[\frac{1}{X}\right]$ compare to $\frac{1}{\lambda}$?

- ○ In general, μ and λ are not comparable
- ◉ $\mu \geq \frac{1}{\lambda}$
- ○ $\mu \leq \frac{1}{\lambda}$

✔

Solution:

Note that the function $x \mapsto \frac{1}{x}$ is a convex function on $(0, \infty)$, hence we can use Jensen's inequality that implies

$$\mathbb{E}\left[f(X)\right] \geq f\left(\mathbb{E}\left[X\right]\right)$$

for all convex functions f to conclude

$$\mu = \mathbb{E}\left[\frac{1}{X}\right] \geq \frac{1}{\mathbb{E}\left[X\right]} = \frac{1}{\lambda}.$$

Variance of the Sample Mean

1/1 point (graded)

Again, let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{U}([a, a+1])$ where a is an unknown parameter. In terms of n, what is the variance of the estimator $\overline{X}_n$?

$\text{Var}\left[\overline{X}_n\right] =$ (1/12)/n ✔ **Answer:** 1/(12*n)

$\frac{\frac{1}{12}}{n}$

Solution:

Since $X_1, \ldots, X_n$ are independent, the variance is additive. Hence,

$$\text{Var}\left(\overline{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left(X_i\right) = \frac{1}{n}\text{Var}\left(X_1\right)$$

Note that we used the fact that the X_i's are identically distributed. Next,

$$\text{Var}\left(X_1\right) = \mathbb{E}\left[X_1^2\right] - \left(\mathbb{E}\left[X_1\right]\right)^2 = \int_a^{a+1} x^2\,dx - \left(a+\frac{1}{2}\right)^2 = a^2 + a + 1/3 - a^2 - a - 1/4 = 1/12.$$

Hence,

$$\text{Var}\left(\overline{X}_n\right) = \frac{1}{n}\text{Var}\left(X_1\right) = \frac{1}{12n}.$$

Find the Quadratic Risk

0/1 point (graded)

Let $\hat{\theta}_n$ denote an estimator for a parameter θ. The **quadratic risk** of $\hat{\theta}_n$ is defined to be

$$\mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right].$$

As in the previous problem on variance, let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{U}([a, a+1])$ where a is an unknown parameter. What is the quadratic risk of the estimator $\overline{X}_n - \frac{1}{2}$?

Quadratic risk : 1/(12*n)+ 0.5^2 ✖ **Answer:** 1/(12*n)

$\frac{1}{12\cdot n} + 0.5^2$

Solution:

Recall that

$$\text{quadratic risk} = \text{variance} + \text{bias}^2.$$

We showed in a previous question that this estimator is unbiased. Also note that $\text{Var}\left(\overline{X}_n\right) = \text{Var}\left(\overline{X}_n - \frac{1}{2}\right) = \frac{1}{12n}$. Hence, the quadratic risk is also $\frac{1}{12n}$.

Properties of Estimators

0/1 point (graded)

Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ denote an estimator for a parameter θ. Here n denotes the sample size. Which of the following properties of $\hat{\theta}_n$ would (completely by itslef) ensure that $\hat{\theta}_n$ converges in probability to θ as $n \to \infty$? (Choose all that apply.)

- [x] $\hat{\theta}_n$ is consistent. ✔
- [x] $\hat{\theta}_n$ is unbiased.
- [x] The quadratic risk of $\hat{\theta}_n$ goes to 0 as $n \to \infty$. ✔
- [x] The variance of $\hat{\theta}_n$ goes to 0 as $n \to \infty$.

✖

Solution:

The first choice is correct, because by definition, consistency implies that the estimator $\hat{\theta}_n \to \theta$ as $n \to \infty$. The third choice, "The quadratic risk of $\hat{\theta}_n$ goes to 0 as $n \to \infty$.", is correct because if the quadratic risk $\mathbb{E}\left[(\hat{\theta}_n - \theta)^2\right] \to 0$ then $\hat{\theta}_n \to \theta$ in L^2. By the properties of convergence, this implies that $\hat{\theta}_n \to \theta$ in probability.

Recall: Refer to Chapter 1 to review the relationship between the different types of convergence.

The second choice, "$\hat{\theta}_n$ is unbiased.", is incorrect. We give an example that shows that this choice is incorrect. Note that if $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$, then $\hat{\theta}_n := X_1$ is an unbiased estimator for μ because $\mathbb{E}[X_1] = \mu$. However, it is not consistent: $X_1 - \mu$ does not tend to 0 as $n \to \infty$.

Using this same example, we can also see that the fourth choice "The variance of $\hat{\theta}_n$ goes to 0 as $n \to \infty$." is incorrect. The estimator $\hat{\theta}_n := 0$ (as an estimator for μ) has variance 0 for all n. If $\mu \neq 0$, then $\hat{\theta}_n - \mu = -\mu$, which is constant for all n and does not converge to 0.

7. Exercise: Strengths and Weaknesses of Estimators

Bookmark this page

You observe samples $X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(\theta)$ where $\theta \in (0, 1)$ is an unknown parameter. Suppose that n is much larger than 1 so we have access to many samples from the specified distribution. Consider three candidate estimators for θ.

- $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 0.5
- X_1.

In the next three questions, you will consider potential strengths and weaknesses of these estimators.

In this particular section (just for now), the phrase “efficiently computable" refers to the existence of an explicit formula. More precisely, here we say that an estimator $\hat{\theta}_n$ is efficiently computable if there's a formula that takes as input $X_1, \ldots, X_n$ whose output is $\hat{\theta}_n$. (Not all estimators are efficiently computable, in this sense of the word. We may encounter examples of such estimators in a later unit.)

Strengths and Weaknesses of Estimators I

1/1 point (graded)

Which of the following is a potential **disadvantage** of using $\hat{\theta}_n = .5$ as an estimator for θ? (Choose all that apply.)

- [x] Unless $\theta = .5$, this estimator is biased.
- [x] Unless $\theta = .5$, this estimator is not consistent.
- [x] This estimator does not use any of the samples.
- [ ] This estimator is efficiently computable.

✔

Solution:

- Biased estimators inherently introduce errors in parameter estimation. Hence, having non-zero bias is a potential disadvantage of an estimator, so the first choice is correct.
- If the estimator does not converge to the true parameter as the sample size n grows very large, this is also a potential disadvantage. Estimators that are not consistent have inherently limited accuracy in approximating the true parameter, regardless of how many samples are taken. Hence, the second choice, "Unless $\theta = .5$, this estimator is not consistent," is also correct.
- The third choice, "This estimator does not depend on the sample, which is the only information that is given to us related to the distribution.", is correct. We haven't used any information given to us so we can't expect to learn anything new about the true parameter with this estimator.
- The fourth choice is not a disadvantage. The given estimator is indeed efficiently computable, and this is in general an *advantage* of certain estimators. In contrast, there are estimators that often don't have an explicit formula (for example, we may encounter such estimators in a later unit on Generalized Linear Models), and often requires approximate computation via a computer program.

Strengths and Weaknesses of Estimators II

1/1 point (graded)

Which of the following is a potential **disadvantage** of using $\hat{\theta}_n = X_1$ as an estimator for θ? (Choose all that apply.)

- [ ] This estimator is unbiased.
- [x] This estimator is not consistent.
- [x] This estimator uses only information given by only one sample, even though we have access to many samples.
- [x] The quadratic risk of this estimator does not tend to 0 as the sample size $n \to \infty$.

✔

Solution:

- The first option is incorrect: unbiasedness is in general an advantage of an estimator, and the estimator X_1 is unbiased because $\mathbb{E}[X_1] = \theta$.
- Since X_1 is Bernoulli, it is either 0 or 1, it is not equal to the true parameter θ which is assumed to lie in $(0, 1)$. Hence, $\hat{\theta}_n = X_1$ is not consistent. This is a potential disadvantage of the estimator, so the second choice, "This estimator is not consistent.", is correct. Intuitively, inconsistent estimators inherently have limited accuracy in approximating the true parameter, no matter how many samples are collected. Hence, inconsistency is in general a disadvantage.
- As a statistician, it is generally best to use all information that is given about a distribution for parameter estimation. One sample (which we know is either 0 or 1) does not tell us much about the underlying distribution. Hence, it is a potential disadvantage that the samples $X_2, \ldots, X_n$ were not used to construct the estimator $\hat{\theta}_n = X_1$. The third choice, "This estimator uses only information given by one sample, even though we have access to many samples.", is thus correct.

- Note that $\text{Var}[X_1] = \theta(1-\theta)$. Since

$$\text{quadratic risk} = \text{variance} + \text{bias}^2.$$

and X is unbiased, the quadratic risk is equal to the variance: $\theta(1-\theta)$. This estimator does not tend to θ as $n \to \infty$. This is a potential disadvantage, because if the quadratic risk does not go to 0, then $\hat{\theta}_n$ does not converge to θ in L^2. Intuitively, an estimator that does not converge in L^2 inherently has limited accuracy no matter how many samples are collected. Thus the fourth choice, "The quadratic risk of this estimator does not tend to 0 as the sample size $n \to \infty$.", is correct.

Remark: Convergence in L^2 is, mathematically speaking, a stronger guarantee than convergence in probability. That is, if $\hat{\theta}_n \overset{L^2}{\to} \theta$, then also, $\hat{\theta}_n \overset{\text{prob}}{\to} \theta$. Refer to Chapter 1 to review the different notions of convergence and their properties.

Which of the following are potential **advantages** of using $\hat{\theta}_n = \overline{X}_n$ as an estimator for θ? (Choose all that apply.)

- [x] This estimator is unbiased.
- [x] This estimator is consistent.
- [x] This estimator is efficiently computable.
- [x] The quadratic risk of this estimator tends to 0 as the sample size $n \to \infty$.

✔

Solution:

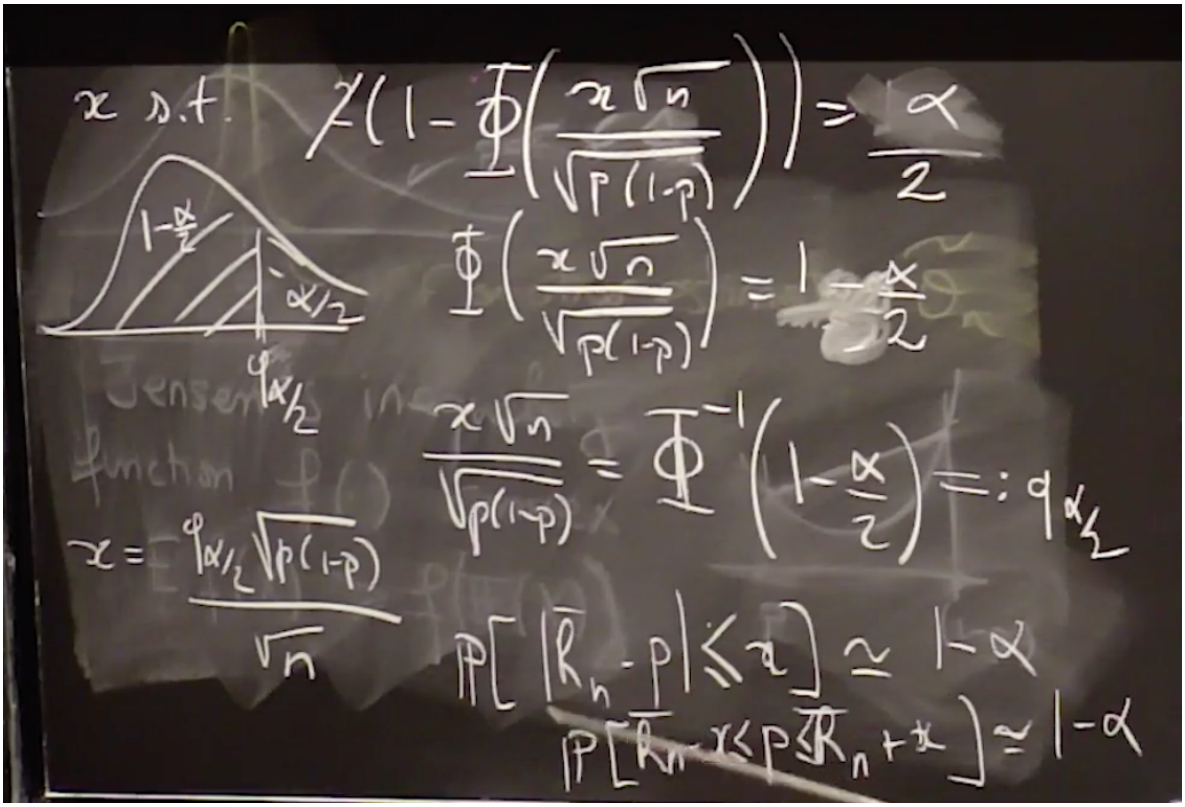
All of the above properties are potential advantages.

- Unbiased estimators avoid inherent inaccuracies of approximation that result from biasedness, so the first choice is correct.
- Consistent estimators become better and better approximations to the true parameter as the sample size increases. This is an advantage so the second choice is correct.
- Averages can be computed in linear time from the data, so is a computationally efficient estimator. In general, for applications, we need to work with estimators that can be computed efficiently, so the third choice is correct.
- The estimator is unbiased, so its variance is the same as its quadratic risk. By independence and the iid assumption,

$$\text{Var}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_1) = \frac{\theta(1-\theta)}{n}$$

which tends to 0 as $n \to \infty$. Thus, $\hat{\theta}_n \overset{L^2}{\sim} \theta$, which ensures consistency. Hence, the last choice, "The quadratic risk of this estimator tends to 0 as the sample size $n \to \infty$." , is an advantage of this estimator.

Confidence intervals
goes with slide 21



Random or Deterministic?

4/4 points (graded)

As in the video above, let $R_1, \ldots, R_n \overset{iid}{\sim} \text{Ber}(p)$ for some unknown parameter p. We estimate p using the estimator $\hat{p} = \overline{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$.

For a fixed number α, after applying the CLT (and doing some algebra), we obtained

$$\lim_{n\to\infty} \mathbf{P}\left(\left[\overline{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, \overline{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}\right] \ni p\right) = 1-\alpha.$$

Which of the quantities in the equation above is random and which is deterministic?
(Choose one for each column.)

$\overline{R}_n$:	n :	$q_{\alpha/2}$:	p :
(●) random	( ) random	( ) random	( ) random
( ) deterministic	(●) deterministic	(●) deterministic	(●) deterministic
✔	✔	✔	✔

Solution:

- $\overline{R}_n = \frac{\sum_{i=1}^{n} R_i}{n}$ is function of the random variables R_i, and hence is random.
- n is the sample size, a deterministic number.
- $q_{\alpha/2}$ is a number given a fixed α, hence deterministic.
- p is the unknown parameter, a number, hence deterministic.

Remark 1: Once we substitute a realization for $\overline{R}_n$ (e.g. from data), the expression $\left[\overline{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, \overline{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}\right] \ni p$ becomes deterministic since all involved quantities are deterministic.

Remark 2: The unknown parameter p is deterministic in the classical (frequentist) approach. In the course *6.431x, Probability-the Science of Uncertainty and Data*, we have seen that in the Bayesian approach, p is modeled as a random variable. We will revisit Bayesian statistics from a different perspective later in this course.

Conservative bound

0/1 point (graded)

As in the video above, let $R_1, \ldots, R_n \overset{iid}{\sim} \text{Ber}(p)$ for some unknown parameter p. We estimate p using the estimator $\hat{p} = \overline{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$.

Recall that by the central limit theorem, for any p, $(0 < p < 1)$:

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\sqrt{n}\frac{\overline{R}_n - p}{\sigma_p}\right| < q_{\alpha/2}\right) = \lim_{n\to\infty} \mathbf{P}\left(\overline{R}_n - q_{\alpha/2}\frac{\sigma_p}{\sqrt{n}} < p < \overline{R}_n + q_{\alpha/2}\frac{\sigma_p}{\sqrt{n}}\right) = 1-\alpha$$

where $\sigma_p = \sqrt{p(1-p)}$.

To construct a confidence interval, we need to replace σ_p above by a number c that does not depend on the unknown parameter p.

Which of the following conditions on c will guarantee that for all p in $(0, 1)$,

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\sqrt{n}\frac{\overline{R}_n - p}{c}\right| < q_{\alpha/2}\right) \geq 1-\alpha?$$

(Choose all that apply.)

- [ ] $c \geq \sigma_p$ for **all** p ✔
- [x] $c \geq \sigma_p$ for **some** p
- [ ] $c = \max_p (\sigma_p)$ ✔
- [x] $c \leq \sigma_p$ for **all** p
- [ ] $c \leq \sigma_p$ for **some** p
- [x] $c = \min_p (\sigma_p)$

✖

Solution:

Any number c such that

$$\left(\overline{R}_n - q_{\alpha/2}\frac{c}{\sqrt{n}},\ \overline{R}_n + q_{\alpha/2}\frac{c}{\sqrt{n}}\right) \supseteq \left(\overline{R}_n - q_{\alpha/2}\frac{\sigma_p}{\sqrt{n}},\ \overline{R}_n + q_{\alpha/2}\frac{\sigma_p}{\sqrt{n}}\right) \qquad \text{for all } p$$

will give the required probability for all p. Hence any $c \geq \max_p (\sigma_p)$ works.

Note: In this example, since $\sigma_p = \sqrt{p(1-p)}$, $\max_p (\sigma_p) = \max_p \left(\sqrt{p(1-p)}\right) = 1/2$.

Solving for a Confidence Interval: Algebra

2/2 points (graded)

In the problems on this page, we will continue building the confidence interval of asymptotical level **95%** by solving for p as in the video.

Recall that $R_1, \ldots, R_n \overset{iid}{\sim} \text{Ber}(p)$ for some unknown parameter p, and we estimate p using the estimator $\hat{p} = \overline{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$.

As in the method using a conservative bound, our starting point is the result of the central limit theorem:

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{p(1-p)}}\right| < q_{\alpha/2}\right) = 1 - \alpha.$$

In this second method, we solve for values of p that satisfy the inequality $\left|\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{p(1-p)}}\right| < q_{\alpha/2}$.

To do this, we manipulate $\left|\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{p(1-p)}}\right| < q_{\alpha/2}$ into an inequality involving a quadratic function $Ap^2 + Bp + C$ where $A > 0$, B, C depend on n, $q_{\alpha/2}$, and $\overline{R}_n$. Which of the following is the correct inequality?
(We will use find the values of A, B, and C in the next problem.)

- (●) $Ap^2 + Bp + C < 0$ where $A > 0$.
- ( ) $Ap^2 + Bp + C > 0$ where $A > 0$.

✔

Let p_1 and p_2 with $0 < p_1 < p_2 < 1$ be the two roots of the quadratic function $Ap^2 + Bp + C$. What values of p satisfy the correct inequality above?

- ◯ $(p < p_1) \cup (p > p_2)$
- ◉ $p_1 < p < p_2$
- ◯ $0 < p < p_1$
- ◯ $p_2 < p < 1$
- ◯ $0 < p < 1$

✔

Solution:

$$\left|\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{p(1-p)}}\right| < q_{\alpha/2} \implies \left(\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{p(1-p)}}\right)^2 < q_{\alpha/2}^2$$
$$\implies \left(\overline{R}_n - p\right)^2 < \frac{p(1-p)\,q_{\alpha/2}^2}{n}$$
$$\implies p^2\left(1 + \frac{q_{\alpha/2}^2}{n}\right) - p\left(2\overline{R}_n + \frac{q_{\alpha/2}^2}{n}\right) + \left(\overline{R}_n\right)^2 < 0$$

Hence, the inequality is of the form $Ap^2 + Bp + C<0$ for some $A > 0$.
The quadratic function $Ap^2 + Bp + C < 0$ where $A > 0$ is convex, so the parabola opens up, and the region in which the parabola is below the x-axis is the interval between the two roots. Given $0 < p_1 < p_2 < 1$, the region is $p_1 < p < p_2$.

Solving for a Confidence Interval: Numerical Descriptions

0.4/2 points (graded)

Continuing from above, enter numerical values for $A > 0$, B, C such that the inequality in the previous problem is equivalent to $\left|\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{p(1-p)}}\right| < q_{\alpha/2}$ for the case when the sample size is $n = 100$, and the observed value of $\overline{R}_n$ is 0.645.

Carry out the computations with the goal of computing a confidence interval of p at asymptotic level 95%. **Note:** Because polynomials differing by only an overall rescaling constant yield the same roots, use $C = \left(\overline{R}_n\right)^2$ here as in the previous problem.

(If necessary, round your answers to the nearest four decimal places (10^{-4}).

$0 < A =$ 1.0271 ✖ **Answer:** 1+(1.96^2)/100

$B =$ -1.3171 ✖ **Answer:** -(2*0.645+1.96^2/100)

$C =$ 0.4160 ✔ **Answer:** 0.645^2

Now, as indicated previously, use the above values (**rounded to the nearest** 10^{-4}) to compute a confidence interval $\mathcal{I}_{\text{solve}}$ of p of asymptotic level 95%.
If necessary, round your endpoints to the nearest two decimal places (10^{-2}).

$p \in [$ ✖ **Answer:** 0.5473323 , ✖ **Answer:** 0.7319435]

Solution:

Recall from the previous problem that

$$\left|\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{p\,(1-p)}}\right| < q_{\alpha/2} \implies p^2\left(1+\frac{q_{\alpha/2}^2}{n}\right) - p\left(2\overline{R}_n + \frac{q_{\alpha/2}^2}{n}\right) + \left(\overline{R}_n\right)^2 < 0.$$

Plugging $n = 100$, $\overline{R}_n = 0.645$, and $q_{\alpha/2} = q_{0.025} = 1.96$ into the inequality above gives

$$p^2\left(1+\frac{1.96^2}{100}\right) - p\left(2\,(0.645) + \frac{\mathbf{1}.96^2}{100}\right) + 0.645^2 < 0.$$

The quadratic formula gives the roots $\dfrac{-B \pm \sqrt{B^2 - 4AC}}{2A}$, which are

$$\begin{aligned} p_1 &= 0.5473323 \\ p_2 &= 0.7319435. \end{aligned}$$

This gives the confidence interval $[p_1, p_2] \approx [0.55, 0.73]$.

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670

Convergences of different quantities

0/3 points (graded)

As in lecture, recall that $R_1, \ldots, R_n \overset{iid}{\sim} \text{Ber}(p)$ for some unknown parameter p, and we estimate p using the estimator $\hat{p} = \overline{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$.

As in the methods before, our starting point is the following result of the central limit theorem:

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{p(1-p)}}\right| < q_{\alpha/2}\right) = 1 - \alpha.$$

Choose the correct convergence statement for each quantity below:
(Choose all that apply for each column.)
Note: In the third and fourth choices below, "is approximated by (in distribution)", means that the CDFs are close; i.e. $\lim_{n\to\infty} F_n(x) - G_n(x) \to 0$, where F_n is the CDF of the RV in the question and G_n is the CDF of the normal distribution with mean p and the written variance, e.g. $\mathcal{N}(p, \frac{p(1-p)}{n})$.

$\overline{R}_n$:	$\sqrt{n}\left(\overline{R}_n - p\right)$:	$\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{p(1-p)}}$:
☐ $\xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$	☐ $\xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$	☑ $\xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$
☐ $\xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, p(1-p))$	☑ $\xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, p(1-p))$	☐ $\xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, p(1-p))$
☑ is approximated by (in distribution) $\mathcal{N}(p, \frac{p(1-p)}{n})$	☐ is approximated by (in distribution) $\mathcal{N}(p, \frac{p(1-p)}{n})$	☐ is approximated by (in distribution) $\mathcal{N}(p, \frac{p(1-p)}{n})$
☐ is approximated by (in distribution) $\mathcal{N}(p, \frac{p(1-p)}{\sqrt{n}})$	☐ is approximated by (in distribution) $\mathcal{N}(p, \frac{p(1-p)}{\sqrt{n}})$	☐ is approximated by (in distribution) $\mathcal{N}(p, \frac{p(1-p)}{\sqrt{n}})$
☑ $\xrightarrow[n\to\infty]{(\mathbf{P})} p$	☐ $\xrightarrow[n\to\infty]{(\mathbf{P})} p$	☐ $\xrightarrow[n\to\infty]{(\mathbf{P})} p$
☐ $\xrightarrow[n\to\infty]{(\mathbf{P})} 1$	☐ $\xrightarrow[n\to\infty]{(\mathbf{P})} 1$	☐ $\xrightarrow[n\to\infty]{(\mathbf{P})} 1$
☐ $\xrightarrow[n\to\infty]{(\mathbf{P})} \sqrt{p(1-p)}$	☐ $\xrightarrow[n\to\infty]{(\mathbf{P})} \sqrt{p(1-p)}$	☐ $\xrightarrow[n\to\infty]{(\mathbf{P})} \sqrt{p(1-p)}$

$\sqrt{\overline{R}_n\left(1-\overline{R}_n\right)}$:	$\frac{\sqrt{\overline{R}_n\left(1-\overline{R}_n\right)}}{\sqrt{p\left(1-p\right)}}$:	$\left(\sqrt{n}\frac{\overline{R}_n-p}{\sqrt{p\left(1-p\right)}}\right)\left(\frac{\sqrt{p\left(1-p\right)}}{\sqrt{\overline{R}_n\left(1-\overline{R}_n\right)}}\right)$:
☐ $\xrightarrow[n\longrightarrow\infty]{(d)}\mathcal{N}\left(0,1\right)$	☐ $\xrightarrow[n\longrightarrow\infty]{(d)}\mathcal{N}\left(0,1\right)$	☑ $\xrightarrow[n\longrightarrow\infty]{(d)}\mathcal{N}\left(0,1\right)$
☐ $\xrightarrow[n\longrightarrow\infty]{(d)}\mathcal{N}\left(0,p\left(1-p\right)\right)$	☐ $\xrightarrow[n\longrightarrow\infty]{(d)}\mathcal{N}\left(0,p\left(1-p\right)\right)$	☐ $\xrightarrow[n\longrightarrow\infty]{(d)}\mathcal{N}\left(0,p\left(1-p\right)\right)$
☐ is approximated by (in distribution) $\mathcal{N}\left(p,\frac{p(1-p)}{n}\right)$	☐ is approximated by (in distribution) $\mathcal{N}\left(p,\frac{p(1-p)}{n}\right)$	☐ is approximated by (in distribution) $\mathcal{N}\left(p,\frac{p(1-p)}{n}\right)$
☐ is approximated by (in distribution) $\mathcal{N}\left(p,\frac{p(1-p)}{\sqrt{n}}\right)$	☐ is approximated by (in distribution) $\mathcal{N}\left(p,\frac{p(1-p)}{\sqrt{n}}\right)$	☐ is approximated by (in distribution) $\mathcal{N}\left(p,\frac{p(1-p)}{\sqrt{n}}\right)$
☐ $\xrightarrow[n\longrightarrow\infty]{(\mathbf{P})}p$	☐ $\xrightarrow[n\longrightarrow\infty]{(\mathbf{P})}p$	☐ $\xrightarrow[n\longrightarrow\infty]{(\mathbf{P})}p$
☐ $\xrightarrow[n\longrightarrow\infty]{(\mathbf{P})}1$	☑ $\xrightarrow[n\longrightarrow\infty]{(\mathbf{P})}1$	☐ $\xrightarrow[n\longrightarrow\infty]{(\mathbf{P})}1$
☑ $\xrightarrow[n\longrightarrow\infty]{(\mathbf{P})}\sqrt{p\left(1-p\right)}$	☐ $\xrightarrow[n\longrightarrow\infty]{(\mathbf{P})}\sqrt{p\left(1-p\right)}$	☐ $\xrightarrow[n\longrightarrow\infty]{(\mathbf{P})}\sqrt{p\left(1-p\right)}$
✔	✔	

Solution:

1. $\sqrt{\overline{R}_n\left(1-\overline{R}_n\right)}\xrightarrow[n\longrightarrow\infty]{\mathbf{P}}\sqrt{p\left(1-p\right)}$ by the continuous mapping theorem.
2. $\frac{\sqrt{\overline{R}_n\left(1-\overline{R}_n\right)}}{\sqrt{p\left(1-p\right)}}\xrightarrow[n\longrightarrow\infty]{\mathbf{P}}1$ since constant multiple of sequences that converge in probability still converge in probability.
3. $\left(\sqrt{n}\frac{\overline{R}_n-p}{\sqrt{p\left(1-p\right)}}\right)\left(\frac{\sqrt{p\left(1-p\right)}}{\sqrt{\overline{R}_n\left(1-\overline{R}_n\right)}}\right)\xrightarrow[n\longrightarrow\infty]{(d)}\mathcal{N}\left(0,1\right)$ by Slutsky.

For slide 25

Video Note:

In the video below, the roots of the equation

$$1.03p^2 - 1.32098p + 0.416025 = 0$$

should be $p_1 = 0.558$ and $p_2 = 0.724$ instead of what is written. This leads to the confidence interval

$$I_{\text{solve}} = [0.56, 0.72]$$

which is centered around 0.64.

Homework

1. Statistical Models and Identifiability

Bookmark this page

For each of the following examples, define a statistical model and check whether the parameter of interest is identifiable. Follow the definitions closely; it is helpful to consider the following: What is Θ and P_θ? What would it mean for the model to be identifiable?

(a)

3/4 points (graded)

1. One observes n i.i.d. Poisson random variables with unknown parameter λ.

- (●) λ identifiable
- ( ) λ not identifiable

✔

2. One observes n i.i.d. exponential random variables with parameter λ, which is unknown but a priori known to be no larger than 10.

- (●) λ identifiable
- ( ) λ not identifiable

✔

3. One observes n i.i.d. uniform random variables in the interval $[0, \theta]$, where θ is unknown.

- (selected) θ identifiable
- θ not identifiable

✔

4. One observes n i.i.d. Gaussian random variables with unknown parameters μ, σ^2.

- (μ, σ^2) identifiable
- (selected) (μ, σ^2) not identifiable

✖

1. One observes the sign of n i.i.d. Gaussian random variables with unknown parameters μ, σ^2.

- (μ, σ^2) identifiable
- (selected) (μ, σ^2) not identifiable

✔

2. *StatGen* is a statistical procedure to test the relevance of genes. When well calibrated, it outputs the (random) proportion of active genes in a (random) cell. We want to estimate the distribution of this proportion. To that end, we take n iid cells and submit them to *StatGen*. We model the output of *StatGen* as n random variables $X_1, \dots, X_n$ that have uniform distribution on $[0, \theta]$ for some unknown theta.

- θ identifiable
- (selected) θ not identifiable

✖

3. The US Census Bureau is interested in finding out the average commute time of Bostonians. To that end, it randomly selects n individuals, with replacement, among the people who work and live in the Boston area, and asks to each if their commute time is at least 20 minutes. The commute time of a random person is assumed to follow an exponential distribution with parameter λ.

- (selected) λ identifiable
- λ not identifiable

4. Willy Wonka's contains 67 identical machines. Each machine has a lifetime that is modeled as an exponential random variable with some unknown parameter λ. After a certain time $T = 500$ days, one has observed the lifetimes of all machines that have stopped working before T. The parameter of interest is λ.

- (selected) λ identifiable
- λ not identifiable

✔

2. Biased and unbiased estimation for variance of Bernoulli variables

Bookmark this page

(a)

2/2 points (graded)

Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli random variables, with unknown parameter $p \in (0, 1)$. The aim of this exercise is to estimate the common variance of the X_i.

First, recall what $\mathrm{Var}(X_i)$ is for Bernoulli random variables.

$\mathrm{Var}(X_i)$ = p*(1-p) ✔

$p \cdot (1 - p)$

Let $\overline{X}_n$ be the sample average of the X_i,

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

We are interested in finding an estimator for $\mathrm{Var}(X_i)$, and propose to use

$$\hat{V} = \overline{X}_n (1 - \overline{X}_n).$$

Check the correct statement that applies to $\hat{V}$:

- ○ $\hat{V}$ is not consistent because $\mathrm{Var}(X_i)$ is not linear in p
- ◉ $\hat{V}$ is consistent because of the Law of Large Numbers and Continuous Mapping Theorem

(b)

0/2 points (graded)

Now, we are interested in the bias of $\hat{V}$. Compute:

$\mathbb{E}[\hat{V}] - \text{Var}(X_i) =$ p/n - p*(1-p) ✖

$\frac{p}{n} - p \cdot (1-p)$

Using this, find an unbiased estimator $\hat{V}'$ for $p(1-p)$ if $n \geq 2$.

Write `barX_n` for $\overline{X}_n$.

$\hat{V}' =$ barX_n* n/(n-1) ✖

$\overline{X}_n \cdot \frac{n}{n-1}$

STANDARD NOTATION

Delta Method and Confidence Intervals

Confidence Interval Concept Check 1

1/1 point (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$, where θ is an unknown parameter. You construct a **confidence interval** $\mathcal{I} = [L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$ for θ.

Complete the next sentence with one of the options below. The confidence interval $\mathcal{I}$ is ...

- (●) A random object
- ( ) A deterministic object

✔

Solution:

As defined, a confidence interval $\mathcal{I} = [L, U]$ for an unknown parameter θ is a *random* interval such that the expressions for its endpoints L, U do **not** depend on θ.

Remark 1: Let's write $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ for the endpoints of the random interval $\mathcal{I}$. Note that a and b are functions of the random sample that do not depend on θ. In practice, one uses data (e.g. realizations $x_1, \ldots, x_n$ of the i.i.d. observations $X_1, \ldots, X_n$) to compute the *realization* $\mathcal{I}_{\text{real}}$ of the confidence interval $\mathcal{I}$:

$$\mathcal{I}_{\text{real}} := [L(x_1, \ldots, x_n), U(x_1, \ldots, x_n)].$$

Remark 2: For this concept, it is important to distinguish the random variable $\mathcal{I}$ (the confidence interval) from its realization $\mathcal{I}_{\text{real}}$, which can be formed only after collecting data.

Confidence Interval Concept Check 2

0/1 point (graded)

Recall that a **realization** of a random variable X is the value that it takes when we observe X. For example, if $X \sim \text{Ber}(1/2)$ and we observe the event $X = 1$, then $x = 1$ is the realization (observed value) of the random variable X.

Let $\mathcal{I}, \mathcal{J}$ be some 95% and 98% asymptotic confidence intervals respectively for the unknown parameter p. Which of the following statements is true?

- ( ) Any realization of $\mathcal{I}$ is a **subinterval** of any realization of $\mathcal{J}$.
- (●) Any realization of $\mathcal{J}$ is a **subinterval** of any realization of $\mathcal{I}$.
- ( ) None of the above ✔

✖

slide 26

A larger confidence interval would be larger (in terms of the error margin) i.e. if we are 100% confident then we're basically talking about the whole space
But the confidence intervals could be with different methods or staggered so it doesn't necessarily mean that one is within the other

Confidence Interval Concept Check 3

0/1 point (graded)

In a new experiment consisting of 150 couples, 75 couples are observed to turn their heads to the left and the remaining 75 couples turned their heads to the right when kissing. Let p denote the (unknown) parameter which specifies the probability that a couple turns their head to the right.

Which of the following statements are correct regarding this experiment? You are given that **exactly one** but not both of choices 3 and 4 is correct. Also, assume that the given confidence intervals are an instance of a random interval computed upon observing the given data.
(Choose all that apply.)

- [x] $[0, 0.5]$ is a 50% asymptotic confidence interval for p. ✔
- [ ] $[0.5, 1]$ is a 50% asymptotic confidence interval for p. ✔
- [ ] $[0.466, 0.533]$ is a 50% asymptotic confidence interval for p. ✔
- [ ] $[0.48, 0.52]$ is a 50% asymptotic confidence interval for p.

Solution:

See the next video for presented solution.

The first three answer choices are correct, and the final choice is incorrect.

Let $R_1, R_2, \ldots, R_{150} \overset{iid}{\sim} \text{Ber}(p)$ denote the sampled response (without loss of generality, assume that $R_i = 1$ encodes that the i-th couple turns their heads to the right, and $R_i = 0$ encodes that the couple turns their heads to the left.) Let $P = \text{Ber}(p)$ denote the common distribution of $R_1, \ldots R_{150}$.

Consider the sample mean $\overline{R}_n$. By the central limit theorem,

$$\sqrt{n}\left(\frac{\overline{R}_n - p}{\sqrt{p(1-p)}}\right) \underset{(d)}{\longrightarrow} N(0,1).$$

Now we examine the answer choices in order.

1. Consider the interval $[0, 0.5]$. Since $\overline{R}_n = 0.5$, this interval is a realization of the (random) confidence interval $\mathcal{I} = (0, \overline{R}_n)$. We compute that

$$P(\mathcal{I} \ni p) = P(p \leq \overline{R}_n) = P(\overline{R}_n - p \geq 0) = P\left(\frac{\overline{R}_n - p}{\sqrt{p(1-p)}} \geq 0\right).$$

   This probability goes to 0.5 as $n \to \infty$. Therefore, the realization $[0, 0.5]$ of the random interval $[0, \overline{R}_n]$ is an asymptotic confidence interval of level 0.5.

2. The interval $[0.5, 1]$ is a realization of the (random) confidence interval $\mathcal{I} = (\overline{R}_n, 1)$. We see that

$$P(\mathcal{I} \ni p) = P(\overline{R}_n \leq p \leq 1) = P(\overline{R}_n - p \leq 0).$$

   By the reasoning in the previous part, we must also have that $\lim_{n\to\infty} P(\mathcal{I} \ni p) = 1/2$.

3. Given that either choice 3 or choice 4 is correct but not both, it must be that the wider of the 2 intervals $[0.466, 0.533]$ is a 50% asymptotic confidence interval for p. Otherwise, the narrower interval $[0.48, 0.52]$ being a 50% asymptotic confidence interval for p implies the same for the wider interval.

Confidence Interval Concept Check 4

0/1 point (graded)

If $[0.34, 0.57]$ is a **realization** of a (non-asymptotic, for some fixed n) 95% confidence interval for an unknown parameter p, then which of the following is true?

The probability that the unknown parameter p is in this interval is

- ○ ≥ 0.025
- ○ ≥ 0.05
- ◉ ≥ 0.95
- ○ None of the above, because p and $[0.34, 0.57]$ are both deterministic. ✔

✖

Solution:

Given some unknown but fixed parameter $\theta \in \mathbb{R}$ for a parametric model and random variables $X_1, \ldots, X_n$ distributed i.i.d. P_θ, recall that the non-asymptotic 95% Confidence Interval of p is an interval $\mathcal{I} = \mathcal{I}(X_1, \ldots, X_n)$ such that $\Pr(\mathcal{I} \ni \theta) \geq 0.95$. It is important to note that there is randomness here, given by the randomness of $\mathcal{I}$.

A realization of a random variable is *deterministic*. The interval $[0.34, 0.57]$ either contains the parameter p, or it doesn't. In other words, the expression $\Pr([0.34, 0.57] \ni p)$ is equal to 1 or 0. Hence, the correct choice is "None of the above, because p and $[0.34, 0.57]$ are both deterministic.".

Confidence Interval Concept Check 5

2/2 points (graded)

Based on some data gathered by your company, you produce a (realization of a) confidence interval $[0.34, 0.57]$ that has (asymptotic) level 95%. Upon presenting your data and confidence interval to your employers, they ask you two questions:

Can the interval $[0.34, 0.57]$ also be used as a (realization of a) confidence interval of (asymptotic) level 98 % ?

- ◯ Yes
- ◉ No

✔

Can the interval $[0.34, 0.57]$ also be used as a (realization of a) confidence interval of (asymptotic) level 90 % ?

- ◉ Yes
- ◯ No

✔

Solution:

A confidence interval $\mathcal{I}$ at level 95% for the parameter p satisfies

$$\mathbf{P}\left[\mathcal{I} \ni p\right] \geq 0.95 \geq 0.90.$$

By definition, $\mathcal{I}$ is also a confidence interval of (asymptotic) level 90%.

However, a confidence interval at level 95% may be too small to also be a confidence interval at level 98%. Hence, the first statement is not true in general: the answer to the first question is "No."

In *Concept Check 2* I and J were **some** 95% and 98% asymptotic confidence intervals. **Their bounds will depend on the method** we use (Conservative bound, Plug-in, I-Solve) and **our sample**. In the exercise, we were not told how we got these bounds. Hence, the relationship between them is unclear. Although, one interval will contain another if we use the same method and sample for both I and J. This is explained in this video (1:17 - 2:18).

In *Concept Check 5* we are **given the realization of a confidence interval**. We are asked if this realization **can be used** as a realization of a confidence interval of 90% (or 98%) asymptotic level.

We can rephrase the question as *"Can the given bounds [0.34, 0.57] be a realization for a 90% (or 98%) confidence interval?"*.

Let's say we use the same method and sample to construct 90% and 98% confidence intervals, as we use for the 95% confidence interval. We will obtain 2 intervals, one of them will be narrower and one of them will be wider than the given realization [0.34, 0.57] of a 95% confidence interval. I'll leave it for the reader to figure out which is which.

Let's say that we got [0.3565, 0.5469] and [0.3353, 0.5715] (*these numbers are made up and presented in random order*).

This means that:

- We **can** use [0.34, 0.57] as the bounds for __% confidence interval (*because we will get tighter bounds [0.3565, 0.5469] if we use the same method and sample*)
- We **can NOT** use [0.34, 0.57] as the bounds for __% confidence interval (*because we will get wider bounds [0.3353, 0.5715] if we use the same method and sample*)

Does this clarify things for you?

Review: Exponential Random Variables

3/3 points (graded)

Let $X \sim \exp(\lambda)$ for some $\lambda > 0$. Which of the following is the (smallest possible) sample space for X?

- ◯ $\mathbb{N}$
- ◯ $\mathbb{Z}$
- ◉ $[0, \infty)$
- ◯ $(-\infty, \infty)$

✔

Which of the following is the probability density function (pdf) for X? (Assume that $x > 0$).

- ◉ $\lambda e^{-\lambda x}$
- ◯ $\frac{1}{\lambda} e^{-\lambda x}$
- ◯ $\lambda e^{\lambda x}$
- ◯ $\lambda e^{-\lambda x^2}$

✔

What is $\mathbb{E}[X]$?
(By now, you may simply memorize this and not rederive it everytime.)

$\mathbb{E}[X] =$ 1/lambda ✔ **Answer:** 1/lambda

Solution:

- An exponential random variable takes values on all positive real numbers. Therefore, the smallest possible sample space for X is given by $[0, \infty)$.
- By definition, the density of an exponential random variable is given by the function $x \mapsto \lambda e^{-\lambda x}$.
- For completeness, we use the formula for the density to compute the mean of an exponential random variable. By definition and integration by parts,

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x\lambda e^{-\lambda x}\, dx \\
&= -xe^{-\lambda x}\Big|_\infty^0 + \int_0^\infty e^{-\lambda x}\, dx \\
&= 0 - \frac{1}{\lambda} e^{-\lambda x}\Big|_\infty^0 \\
&= \frac{1}{\lambda}.
\end{aligned}$$

Memoryless Property of Exponential Random Variables

0/2 points (graded)

Let $X \sim \exp(1)$. What is $\mathbf{P}(X > 3)$?

$\mathbf{P}(X > 3) =$ 0.0497 **Answer:** exp(-3)

Let $t > 0$. What is $\mathbf{P}(X > t + 3 | X > t)$?

$\mathbf{P}(X > t + 3 | X > t) =$ 0.0497 **Answer:** exp(-3)

Solution:

The density of $\exp(1)$ is given by e^{-x}. Therefore,

$$\mathbf{P}(X > 3) = \int_3^\infty e^{-x}\, dx = -e^{-x}\Big|_3^\infty = e^{-3}.$$

Next, by the memoryless property of the exponential distribution, for any $s, t > 0$, it holds that

$$\mathbf{P}(X > s + t | X > t) = \mathbf{P}(X > s).$$

Apply the above equality with $s = 3$ shows that

$$\mathbf{P}(X > 3 + t | X > t) = \mathbf{P}(X > 3) = \exp(-3).$$

Consistency and Biasedness

2/4 points (graded)

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \exp(\lambda)$. Let $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ denote the sample mean of the data set.

To which value does $\overline{X}_n$ converge (both a.s. and in probability) as $n \to \infty$?
(Choose all that apply)

- [x] $\mathbb{E}[X_i]$
- [ ] $\frac{1}{\mathbb{E}[X_i]}$
- [ ] $\mathbb{E}\left[\frac{1}{X_i}\right]$
- [ ] λ
- [x] $\frac{1}{\lambda}$

✔

To which value does $\frac{1}{\overline{X}_n}$ converge (both a.s. and in probability) as $n \to \infty$? (Choose all that apply)

- [ ] $\mathbb{E}[X_i]$
- [x] $\frac{1}{\mathbb{E}[X_i]}$
- [ ] $\mathbb{E}\left[\frac{1}{X_i}\right]$
- [x] λ
- [ ] $\frac{1}{\lambda}$

✔

Which of the following is the bias of $\frac{1}{\overline{X}_n}$ as an estimator of λ? (Choose all that apply.)

- [ ] $\mathbb{E}\left[\frac{1}{\overline{X}_n}\right] - \lambda$ ✔
- [ ] $\mathbb{E}\left[\frac{1}{\overline{X}_n}\right] - \frac{1}{\mathbb{E}[X_i]}$ ✔
- [ ] $\mathbb{E}\left[\frac{1}{\overline{X}_n}\right] - \frac{1}{\mathbb{E}[\overline{X}_n]}$ ✔
- [x] $\frac{1}{\mathbb{E}[X_i]} - \lambda$
- [x] $\frac{1}{\mathbb{E}[X_i]} - \frac{1}{\mathbb{E}[\overline{X}_n]}$

✖

1

Which of the following are properties of $\frac{1}{\overline{X}_n}$ as an estimator of λ? (Choose all that apply.)

- [ ] consistent ✔
- [x] unbiased

✖

Solution:

- By the (strong/weak) law of large numbers

$$\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow[n\to\infty]{a.s/\mathbf{P}} \mathbb{E}[X_i] = \frac{1}{\lambda}.$$

- On the other hand, by the continuous mapping theorem

$$\frac{1}{\overline{X}_n} \xrightarrow[n\to\infty]{a.s/\mathbf{P}} \frac{1}{\mathbb{E}[X_i]} = \lambda.$$

- Hence, we can answer the last part immediately: $\frac{1}{\overline{X}_n}$ is a consistent estimator of λ.

- However,

$$\mathbb{E}\left[\frac{1}{\overline{X}_n}\right] \neq \frac{1}{\mathbb{E}\left[\overline{X}_n\right]} = \lambda.$$

So the bias of $\frac{1}{\overline{X}_n}$ as an estimator of $\lambda = \frac{1}{\mathbb{E}[X_i]} = \frac{1}{\mathbb{E}\left[\overline{X}_n\right]}$ is

$$\text{Bias} = \mathbb{E}\left[\frac{1}{\overline{X}_n}\right] - \frac{1}{\mathbb{E}\left[\overline{X}_n\right]}.$$

Remark: Since the function $\frac{1}{x}$ is convex (by the shape of its graph or by $\left(\frac{1}{x}\right)'' = \frac{2}{x^3} > 0$), Jensen's inequality gives $\mathbb{E}\left[\frac{1}{\overline{X}_n}\right] > \frac{1}{\mathbb{E}\left[\overline{X}_n\right]}$ and hence the bias is greater than zero.

The **Central Limit Theorem** states that if $X_1, \ldots, X_n$ are i.i.d. and

$$\mathbb{E}[X_1] = \mu < \infty \; ; \qquad \text{Var}(X_1) = \sigma^2 < \infty,$$

then

$$\sqrt{n}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right] \xrightarrow[n\to\infty]{(d)} W \qquad \text{where } W \sim \mathcal{N}(0, ?).$$

What is $\text{Var}(W)$? (Express your answer in terms of n, μ and σ).

$\text{Var}(W) =$ sigma^2 ✔ **Answer:** sigma^2

σ^2

STANDARD NOTATION

Solution:

For any n,

$$\text{Var}\sqrt{n}\left(\overline{X}_n - \mu\right) = n\text{Var}\left(\overline{X}_n\right) = \text{Var}(X_i) = \sigma^2.$$

The central limit theorem states as $n \to \infty$, the distribution of $\sqrt{n}\left(\overline{X}_n - \mu\right)$ becomes Gaussian with the variance above (and mean 0); that is,

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2).$$

Note: The variance of W is called the **asymptotic variance** of $\overline{X}_n$, even though it equals the variance of $\sqrt{n}\overline{X}_n$.

8. The One-Dimensional Delta Method

Bookmark this page

Applying Linear Functions to a Random Sequence

2/3 points (graded)

Let $(Z_n)_{n\geq 1}$ be a sequence of random variables such that

$$\sqrt{n}(Z_n - \theta) \xrightarrow[n\to\infty]{(d)} Z$$

for some $\theta \in \mathbb{R}$ and some random variable Z.

Let $g(x) = 5x$ and define another sequence by $Y_n = g(Z_n)$.

The sequence $\sqrt{n}(Y_n - g(\theta))$ converges. In terms of Z, what random variable does it converge to?

$$\sqrt{n}(Y_n - g(\theta)) \xrightarrow[n\to\infty]{(d)} \quad Y.$$

(Answer in terms of Z)

$Y =$ 5*Z ✔

$5 \cdot Z$

What theorem did we invoke to compute Y?
(There can be more than 1 acceptable answers.)

- ◯ Laws of large number
- ◯ Central Limit theorem
- ◯ Slutsky theorem ✔
- ◉ Continuous mapping theorem

✔

If $\mathsf{Var}(Z) = \sigma^2$, what is $\mathsf{Var}(Y)$? This is the asymptotic variance of $(Y_n)_{n\geq 1}$.
(Answer in terms of σ^2.)

$\mathsf{Var}(Y) =$ 25*sigma^2 ✔ **Answer:** 25*sigma^2

$25 \cdot \sigma^2$

Solution:

1. $$\begin{aligned}\sqrt{n}(Y_n - g(\theta)) &= \sqrt{n}(g(Z_n) - g(\theta)) &&= \sqrt{n}(5Z_n - 5\theta) \\ &&&= 5\left(\sqrt{n}(Z_n - \theta)\right) \xrightarrow[n\to\infty]{(d)} 5Z\end{aligned}$$

   by the continuous mapping theorem because $5\left(\sqrt{n}(Z_n - \theta)\right)$ is a linear and hence continuous function of Z_n in the last step. Alternatively, since we were given that $\sqrt{n}(Z_n - \theta) \xrightarrow[n\to\infty]{(d)} Z$, and 5 converges trivially in probability to itself, we can also use Slutsky theorem to conclude.

2. Since $Y = 5Z$, $\mathsf{Var}(Y) = 25\mathsf{Var}(Z) = 25\sigma^2$.

slide 33

An Estimator for the Mean of an Exponential Random Variable

1/1 point (graded)

In the next two problems, we will repeat the computation in lecture.

Let $X_1, \ldots, X_n \sim \exp(\lambda)$ where $\lambda > 0$.

Since $\mathbb{E}[X] = \frac{1}{\lambda}$, by the central limit theorem,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{\lambda}\right) \xrightarrow[n\to\infty]{(d)} N(0, \sigma^2).$$

What is σ^2 in terms of λ?

$\sigma^2 =$ lambda^(-2)

✔ **Answer:** 1/lambda^2

λ^{-2}

Applying the Delta Method to an Exponential Random Variable

1/1 point (graded)

As above, let $X_1, \ldots, X_n \sim \exp(\lambda)$ where $\lambda > 0$. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote the sample mean. By the CLT, we know that

$$\sqrt{n}\left(\overline{X}_n - \frac{1}{\lambda}\right) \xrightarrow[n\to\infty]{(d)} N(0, \sigma^2)$$

for some value of σ^2 that depends on λ, which you computed in the problem above.

If we set g to be

$$\begin{aligned} g: \mathbb{R} &\to \mathbb{R} \\ x &\mapsto 1/x, \end{aligned}$$

then by the Delta method,

$$\sqrt{n}\left(g(\overline{X}_n) - g\left(\frac{1}{\lambda}\right)\right) \xrightarrow[n\to\infty]{(d)} N(0, \tau^2).$$

where τ^2 is the asymptotic variance and can be expressed in terms of λ.

What is the asymptotic variance τ^2?
(Choose all that apply.)

- [ ] $g'(\lambda)\,\mathrm{Var}X$
- [ ] $g'(\lambda)\frac{1}{\lambda^2}$
- [x] $g'(E[X])^2\mathrm{Var}X$
- [x] $g'\left(\frac{1}{\lambda}\right)^2\frac{1}{\lambda^2}$
- [ ] $\frac{1}{\lambda^2}$
- [x] λ^2

✔

Solution:

By the previous problem, we have

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n}X_i - 1/\lambda) \xrightarrow{(d)} \mathcal{N}\left(0, \frac{1}{\lambda^2}\right).$$

To apply the Delta method, first observe that $g'(x) = -1/x^2$ and that g is continuously differentiable for $x > 0$. By the Delta method,

$$\begin{aligned}\sqrt{n}\left(g(\overline{X}_n) - g(\mathbb{E}[X])\right) = \sqrt{n}\left(\frac{1}{\overline{X}_n} - \frac{1}{1/\lambda}\right) &\xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, g'(\mathbb{E}[X])^2\mathrm{Var}X\right)\\ &= \mathcal{N}\left(0, \left(g'\left(\frac{1}{\lambda}\right)\right)^2\left(\frac{1}{\lambda^2}\right)\right)\\ &= \mathcal{N}(0, \lambda^2).\end{aligned}$$

Since $g'(x) = -1/x^2$, $g'\left(\frac{1}{\lambda}\right) = -\lambda^2$, and hence the asymptotic variance of $g(\overline{X}_n)$ evaluates to λ^2.

Warning: It's very important that we apply g' to the value $1/\lambda$, and not λ. We start with a consistent estimator, namely $\overline{X}_n$, whose limit is $\mathbb{E}[X] = 1/\lambda$, and the Delta method asks us to apply g' to the limit of that consistent estimator. Be careful about this, as it is a common source of errors.

When does the delta method apply?

0/1 point (graded)

Let $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} X$. The distribution of X depends on a **positive** parameter θ, which is a function of the mean μ, i.e $\theta = g(\mu)$. You estimate θ by the estimator $\widehat{\theta} = g(\overline{X}_n)$.

For which function g can the delta method be applied? Remember that $\theta > 0$.
(Choose all that apply.)

- [x] $g(x) = x^3$ ✔
- [x] $g(x) = \sqrt{x}$ ✔
- [ ] $g(x) = \ln(x)$ ✔
- [ ] $g(x) = \begin{cases} x & \text{if } x \leq 1 \\ 2x - 1 & \text{if } x > 1 \end{cases}$
- [x] $g(x) = \frac{1}{x-1}$ ✔

✖

Solution:

For the Delta method to apply, g' exists and is continuous at $\mathbb{E}[X] = g^{-1}(\theta)$. Since θ and $\mu = \mathbb{E}[X]$ are unknown, for the Delta method to apply, we need to make sure g is continuously differentiable at all possible values of $\mathbb{E}[X]$ given that $\theta > 0$. Let us first go through the correct choices:

1. $g(x) = x^3$ is continuously differentiable everywhere.
2. $g(x) = \sqrt{x}$ is continously differentiable for all $x > 0$. Given any $\theta > 0$, $\mu = g^{-1}(\theta) = \theta^2 > 0$. So for all possible values of $\mathbb{E}[X]$, g satisfies the requirement; hence Delta method applies.
3. Similarly, $g(x) = \ln x$ is continously differentiable for all $x > 0$. Given any $\theta > 0$, $\mu = g^{-1}(\theta) = e^{\theta} > 0$. Again, Delta method applies.
4. $g(x) = \frac{1}{x-1}$ is continously differentiable everywhere except at $x = 1$. However, inverting $\theta = g(\mu) = \frac{1}{\mu - 1}$ gives $\mu = \frac{1}{\theta} + 1$, so $\mu \neq 1$ for all $\theta > 0$. Hence the Delta method applies.

Here is the incorrect choice: $g(x) = \begin{cases} x & \text{if } x \leq 1 \\ 2x - 1 & \text{if } x \geq 1 \end{cases}$" is a 1-to-1 piecewise linear function and is continuously differentiable everywhere except at $x = 1$. Observe that $g(1) = 1$, hence when $\theta = 1$, $\mu = 1$. There is a possible value of μ when $g'(\mu)$ does not exist, so the Delta method does not apply.