Unit 2 Foundations of Inference

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

(Choose all that apply.)

(1)	(2)
☑ $((-\infty, \infty), \{N(\mu, 0.23)\}_{\mu>0})$	☑ $((-\infty, \infty), \{N(\mu, 0.23)\}_{\mu>0})$
☑ $((-\infty, \infty), \{N(\mu, 0.23)\}_{\mu\in\mathbb{R}})$	☐ $((-\infty, \infty), \{N(\mu, 0.23)\}_{\mu\in\mathbb{R}})$
☑ $((-\infty, \infty), \{N(\mu, \sigma^2)\}_{\mu,\sigma^2>0})$	☐ $((-\infty, \infty), \{N(\mu, \sigma^2)\}_{\mu,\sigma^2>0})$
☑ $((-\infty, \infty), \{N(\mu, \sigma^2)\}_{\mu\in\mathbb{R},\sigma^2>0})$	☐ $((-\infty, \infty), \{N(\mu, \sigma^2)\}_{\mu\in\mathbb{R},\sigma^2>0})$
✔	✔

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 $Y_i = \beta^\top X_i + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$ 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.

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

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.