

4. Statistical modelling

An example of a statistical model

[Start of transcript. Skip to the end.](#)



So why would you want to do statistical modeling?

So let's look at a very simple example.

This is data that I collected a few years back from people just

like you, slightly dated sense of fashion now.

And when we did this, we observed n independent copies,

x_1 to x_n .

And I say copies of x , because I just

视频

[下载视频文件](#)

字幕

[下载 SubRip \(.srt\) file](#)

[下载 Text \(.txt\) file](#)

Advantages of Modeling Assumptions

1/1 point (graded)

As in the video above, a population consists of n individuals labeled $1, 2, \dots, n$. Let X_i denote the number of siblings of individual i . We assume that X_1, \dots, 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 **Poiss** (λ) for some unknown $\lambda > 0$.

Model 2: X takes values in $\{1, 2, 3, 4, 5, 6, \geq 7\}$, and for $i = 1, 2, \dots, 7$, 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 \mathbf{X} . With Option 2, it is required to find **7** unknowns (all of the $\mathbf{p_i}$'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.

提交

你已经尝试了1次（总共可以尝试2次）

i Answers are displayed within the problem

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 $\mathbf{Y_i}$ denote the \mathbf{i} 'th number in this list. You decide to model this data set under the following assumptions:

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

Under these assumptions, how many unknowns are needed to specify the distribution of \mathbf{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 $\mathbf{Y_i = 1}$) is needed to specify the distribution.

提交

你已经尝试了1次（总共可以尝试3次）

i Answers are displayed within the problem

Approximating the unknown parameter

1/1 point (graded)
As above, let $\mathbf{Y_1, \dots, Y_n}$ denote the \mathbf{i} 'th number in the binary data set.

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

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

- ☒ total number of 1's
 \mathbf{n} ✔
- ☐ $\mathbf{Y_n}$

☐ Median (Y_1, \dots, Y_n)

☒ $\frac{1}{n} \sum_{i=1}^n Y_i$ ✓




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 $n \rightarrow \infty$.
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.

提交

你已经尝试了1次（总共可以尝试2次）

 Answers are displayed within the problem

In the video above, Prof Rigollet mentioned that the number of car accidents that a person may encounter in a year follows a Poisson distribution. Why is that so? Thinking back to the course 6.431x, *Probability–the Science of Data and Uncertainty*, what kind of random process can we model the occurence of car accidents by?

讨论

显示讨论

主题： Unit 2 Foundation of Inference:Lecture 3: Parametric Statistical Models / 4. Statistical modelling

认证证书是什么？