Calculus-based Intro to
Probability & Statistics

Math 11: Discussion 6

TA: Thu Nguyen
Tue. 05/05/2020

Contents

6.1 Upcoming assignments

Assignments	Chapters	Deadlines
Homework	Ch. 10	Mon. 05/11
Quiz	Ch. 10	Tue. 05/12
Homework	Ch. 16	Wed. 05/13
Quiz	Ch. 16	Thu. 05/14
Lab 5		Fri. 05/08

6.2 Probability Revisited

Example 1. *Let X be a random variable with a pdf:*

$$f_X(x) \;=\; \begin{cases} c\; x^{1/2} & \textit{for } 0 \le x \le 1 \\ 0 & \textit{otherwise} \end{cases}$$

(a) Find c.

(b) Find $\mathbb{E}[X]$ and $Var[X]$.

(c) Find median and the IQR of X.

Solution:

(a) We recall one of the 3 fundamental rules of probability:

$$\int_{x\in\Omega} f_X(x)\,\mathrm{d}x \;=\; 1$$

In this case, we have:

$$\int_0^1 c\; x^{1/2}\,\mathrm{d}x \;=\; 1 \quad\Longrightarrow c\left(\frac{2}{3}\;x^{3/2}\right)\Bigg|_0^1 \;=\; 1 \quad\Longrightarrow c\;\frac{2}{3} \;=\; 1 \quad\Longrightarrow c \;=\; \frac{3}{2}$$

(b) We now have the complete pdf: $f_X(x) = \frac{3}{2}x^{1/2}$. From the definition of expected value:

$$\mathbb{E}[X] \;=\; \int_0^1 x f_X(x)\,\mathrm{d}x \;=\; \int_0^1 \frac{3}{2}x^{3/2}\,\mathrm{d}x \;=\; \frac{3}{2}\cdot\frac{2}{5}\;x^{5/2}\Big|_0^1 \;=\; \frac{3}{5}$$

To calculate the variance, recall the other formula: $Var[X] = \mathbb{E}[X^2] - \mathbb{E}^2[X]$:

$$Var[X] \;=\; \int_0^1 \frac{3}{2}x^{5/2}\,\mathrm{d}x - \left(\frac{3}{5}\right)^2 \;=\; \frac{3}{2}\cdot\frac{2}{7}\Big|_0^1 - \frac{9}{25} \;=\; \frac{3}{7} - \frac{9}{25} \;=\; \frac{12}{175}$$

(c) Let x_2 be the *median* or Q_2, we have:

$$\int_0^{x_2} \frac{3}{2}x^{1/2}\,\mathrm{d}x \;=\; \frac{1}{2} \quad\Longrightarrow\quad \frac{3}{2}\cdot\frac{2}{3}\;x_2^{3/2} \;=\; \frac{1}{2} \quad\Longrightarrow\quad x_2 \;=\; \left(\frac{1}{2}\right)^{2/3} \;=\; \frac{1}{\sqrt[3]{4}}$$

We can calculate x_1 for Q_1 and and x_3 for Q_3 similarly:

$$x_1^{3/2} \;=\; \frac{1}{4} \quad\Longrightarrow x_1 \;=\; \frac{1}{\sqrt[3]{16}} \qquad \text{and} \qquad x_3^{3/2} \;=\; \frac{3}{4} \quad\Longrightarrow x_1 \;=\; \sqrt[3]{\frac{9}{16}}$$

Hence, the IQR is:

$$IQR \;=\; x_3 - x_1 \;=\; \sqrt[3]{\frac{9}{16}} - \frac{1}{\sqrt[3]{16}}$$

Math 11: Discussion 6

Example 2. *Let X be a random variable with a pdf:*

$$f_X(x) \;=\; \begin{cases} \frac{1}{2}\, x^{1/2} & \textit{for } a \le x \le b \\ 0 & \textit{otherwise} \end{cases}$$

(a) If $a = 1$, what is b?

(b) If $b = 4$, what is a?

Solution:

By the same fundamental rule of

$$\int_{x \in \Omega} f_X(x)\, \mathrm{d}x \;=\; 1$$

In this case, we have:

$$\int_a^b \frac{1}{2}\, x^{1/2}\, \mathrm{d}x \;=\; 1 \quad\Longrightarrow\quad \frac{1}{2} \cdot \frac{2}{3}\, x^{3/2} \Big|_a^b \;=\; 1 \quad\Longrightarrow\quad b^{3/2} - a^{3/2} \;=\; 3$$

(a) If $a = 1$, then $a^{3/2} = 1$, we have:

$$b^{3/2} \;=\; 3 + 1 = 4 \quad\Longrightarrow\quad b \;=\; \sqrt[3]{4}$$

(b) If $b = 4$, then $b^{3/2} = 2^3 = 8$, we have:

$$a^{3/2} \;=\; 8 - 3 = 5 \quad\Longrightarrow\quad a \;=\; \sqrt[3]{25}$$

Example 3. *Let's look at the rents in San Francisco and San Diego. Suppose that rents in San Francisco have a normal distribution with mean of* \$3,200 *and a standard deviation of* \$800*, while rents in San Diego are lower, at a normal distribution with mean of* \$2,400 *and a standard deviation of* \$600*. Assume that the rents are independent across two cities.*

Suppose we are looking at a random rent in SF and SD. Let's calculate the probability of:

(a) *The rent in SF is over* \$4,000*.*

(b) *The rent in SD is between* \$2,100 *and* \$2,700*.*

(c) *The rent is SF is cheaper than that in SD.*

(d) *The rent is SF is two times more expensive than that in SD.*

Solution:

First, let us define the random variables to simplify the question. Let X and Y denote the rent in San Francisco and San Diego respectively. We have the distributions:

$$X \;\sim\; \mathcal{N}(3,200, 800) \qquad \text{and} \qquad Y \;\sim\; \mathcal{N}(2,400, 600)$$

Recall that if $Z \sim \mathcal{N}(0,1)$ is the *standard* normal r.v., then if $X \sim \mathcal{N}(\mu, \sigma)$, we have:

$$Z \;=\; \frac{X-\mu}{\sigma}$$

Note that for the purpose of doing computation on normal r.v., most of the times we will have to standardize the current random variable (X) to Z. In particular, all tables are for the standard normal random variable.

(a) The rent in SF is over \$4,000, which is to find $\mathbb{P}(X \geq 4,000)$:

$$\mathbb{P}(X \geq 4,000) \;=\; \mathbb{P}\left(\frac{X-3,200}{800} \geq \frac{4,000-3,200}{800}\right) \;=\; \mathbb{P}(Z \geq 1) \;=\; 1-\mathbb{P}(Z \leq 1) \;=\; .16$$

(b) The rent in SD is between \$2,100 and \$2,700, which is to find $\mathbb{P}(2,100 \leq Y \leq 2,700)$:

$$\begin{aligned}\mathbb{P}(2,100 \leq Y \leq 2,400) \;&=\; \mathbb{P}\left(\frac{2,100-2,400}{600} \leq \frac{Y-2,400}{600} \geq \frac{2,700-2,400}{600}\right)\\ &=\; \mathbb{P}(-.5 \leq Z \leq .5)\\ &=\; \Phi(.5)-\Phi(-.5) \qquad\qquad (\text{where } \Phi(z)=\mathbb{P}(Z \leq z))\\ &=\; .38\end{aligned}$$

(c) The rent is SF is cheaper than that in SD, which is to find $\mathbb{P}(X < Y)$, which is to find $\mathbb{P}(X - Y < 0)$.

We first need the distribution of $X - Y$. Let $W = X - Y$, we have:

$$\begin{aligned} \mathbb{E}[W] &= \mathbb{E}[X] - \mathbb{E}[Y] = 800 \\ Var[W] &= Var[X] + Var[Y] = 800^2 + 600^2 = 1000^2 && \text{(since } X \text{ and } Y \text{ are independent)} \end{aligned}$$

Here, you can take it for granted the fact that linear combinations of independent normal random variables are also normal. Hence, $W \sim \mathcal{N}(800, 1000)$. We now have:

$$\mathbb{P}(X - Y < 0) \;=\; \mathbb{P}\left(\frac{W - 800}{1000} < \frac{-800}{1000}\right) \;=\; \mathbb{P}(Z < -.8) \;=\; .21$$

(d) The rent is SF is two times more expensive than that in SD, which is to find $\mathbb{P}(X - 2Y > 0)$.

Similar to part (c), let $W = X - 2Y$:

$$\begin{aligned} \mathbb{E}[W] &= \mathbb{E}[X] - \mathbb{E}[2Y] = -1,600 \\ Var[W] &= Var[X] + Var[2Y] = 800^2 + 4(600^2) = 1,442^2 && \text{(since } X \text{ and } Y \text{ are independent)} \end{aligned}$$

Hence, $W \sim \mathcal{N}(-1,600, 1,442)$. We now have:

$$\mathbb{P}(X - 2Y > 0) \;=\; \mathbb{P}\left(\frac{W + 1,600}{1,442} > \frac{1,600}{1,442}\right) \;=\; \mathbb{P}(Z > 1.11) \;=\; .13$$

6.3 Statistical Inference

Motivation: Suppose we want to describe statistics about a population. However, as usually observed in real life, there is a multitude of reasons that obtaining the population statistics is practically impossible, instead, what we have are samples of the population. Our objective is to infer the population statistics from the sample statistics.

Now, suppose that we have a sample of $\{x_1, x_2, \dots, x_n\}$, where each of x_i is *iid* (independent and identically distributed) sampled from the population. There are 2 types of statistics:

proportion (qualitative data)	mean (quantitative data)
$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i=1}$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

where $\mathbb{1}_{x_i=1}$ is an indicator r.v. of if $x_i = 1$ (a success). In other words, we are taking the naive estimates that p is the proportion of successes in our sample, and $\bar{x}$ the average of our sample data.

Importantly, we have the *expected value* and the *variance* of those estimates:

Estimates	*Expectation*	*Variance*	
$\hat{p}$	$\mathbb{E}[\hat{p}] = p$	$Var[\hat{p}] = \frac{p(1-p)}{n}$	$\hat{p} \sim \left(p, \sqrt{\frac{p(1-p)}{n}}\right)$
$\bar{x}$	$\mathbb{E}[\bar{x}] = \mu$	$Var[\bar{x}] = \frac{\sigma^2}{n}$	$\bar{x} \sim \left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Table 1: Expectation and Variance of the estimates $\hat{p}$ and $\bar{x}$.

where the last column is a concise expression of the *expectation* and *standard deviation* (note that though they may look similar to the notation for the *Normal* distribution, they are not and do not infer that the estimates are normally distributed).

6.3.1 Law of Large Number

As shown in table 1, the expected values of the estimates are the true (population) statistics. Indeed, as the sample size gets larger, the sample estimates will converge to the true statistics:

$$\lim_{n\to\infty} \hat{p} = p \qquad \text{and} \qquad \lim_{n\to\infty} \bar{x} = \mu$$

6.3.2 Central Limit Theorem

Idea: the sampling distribution of *any mean* becomes more *Normal* as the sample size increases.

The precise statement is:

- Suppose we have a sample of $\{x_1, x_2, \ldots, x_n\}$, where each x_i is *iid*. Regardless of the true distribution of the r.v. X, the sample mean is approximately *Normal* distributed with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \frac{Var[X]}{n}$ as n goes to infinity.

In other words, if we let $S_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, then:

$$S_n \xrightarrow[n\to\infty]{\mathbb{P}} \mathcal{N}\Big(\ \mu, \frac{\sigma}{\sqrt{n}}\ \Big)$$

where $\xrightarrow[n\to\infty]{\mathbb{P}}$ means as n goes to infinity, the distribution of S_n converges to.

Remarks:

1. In determining how *large* n needs to be before CLT applies, in general:
   (a) for *qualitative data*, we want both $np \geq 10$ and $n(1-p) \geq 10$
   (b) for *quantitative data*, we want $n \geq 30$; and larger the more severe the skewness of the data
2. CLT applies to the *mean*, and not any individual observations.