Calculus-based Intro to
Probability & Statistics

Math 11: Discussion 8

TA: Thu Nguyen
Tue. 05/19/2020

Recap of discussion 8:

1. Review of Probability
2. Review of Statistical Inference
3. Prepare for Midterm 2!

Contents

8.1 Upcoming assignments

Assignments	Chapters	Deadlines
Homework	Ch. 18	Wed. 05/20
Quiz	Ch. 18	Thu. 05/21
Homework	Ch. 19	Fri. 05/22
Quiz	Ch. 19	Sat. 05/23

Note: Midterm 2 on Saturday, including everything up confidence interval and not including hypothesis testing.

Key concepts (not exhaustive):

1. *Probability distributions:* discrete, continuous, and customized
2. *Statistical inference:* Law of Large Number, Central Limit Theorem
3. *Confidence interval:* building and interpreting

8.2 Probability

Key concepts that you need to have a solid understanding of (or at least be comfortable talking about):

1. Probability rules:
   (a) 3 basic probability rules
   (b) Bayes' rules
   (c) Law of Total Probability

2. Random variables:
   (a) pmf, pdf
   (b) Expected value, Variance: definitions and properties

3. Random variable distributions:
   (a) Discrete: $Bern(p)$, $Geom(p)$, $Bino(n, p)$, $Pois(\lambda)$
   (b) Continuous: $Unif(a, b)$, $Exp(\lambda)$, $\mathcal{N}(\mu, \sigma)$

   - For each of them, be sure you know what they model and how they are different.
   - Be comfortable enough with each distribution so that you can handle combinations of them.
   - If time allows, be comfortable to talk about other concepts related to a distribution given a pdf:
     - median, IQR

Math 11: Discussion 8

Example 1. *Suppose pandemics happen at an average of* 4 *for every* 100 *years. Find the probability of*

(a) *There are* 2 *pandemics in the next* k *years.*

(b) *There are at least* 3 *pandemics in the next* k *years.*

(c) *The next pandemic is at least* k *years from now.*

(d) *There are no pandemics in the next* k *years given that there is no in the next* n *years.*

Solution:

Let X be the rv. denoting the number of pandemics in every k years, and Y denoting the wait time until the next pandemics:

$$X \sim Pois(.04k) \qquad \text{and} \qquad Y \sim Exp(.04k)$$

(a) There are 2 pandemics in the next k years.

$$\mathbb{P}(X=2) = e^{-\lambda}\frac{\lambda^k}{k!} = e^{-.04k}\frac{(.04k)^2}{2}$$

(b) There is at least 3 pandemic in the next k years.

$$\begin{aligned}\mathbb{P}(X \geq 3) &= 1 - \mathbb{P}(X=0) - \mathbb{P}(X=1) - \mathbb{P}(X=2)\\ &= 1 - e^{-.04k} - \big(e^{-.04k}\big)(.04k) - \big(e^{-.04k}\big)\frac{(.04k)^2}{2}\\ &= 1 - \big(e^{-.04k}\big)\left(1 + .04k + \frac{(.04k)^2}{2}\right)\end{aligned}$$

(c) The next pandemic is at least k years from now.

We note that we can either model this as the observing 0 pandemics in k years or that the wait time until the next pandemics in at least k years. Let's go with the wait time:

$$\mathbb{P}(Y \geq k) = e^{-\lambda k} = e^{-.04k}$$

(d) There are no pandemics in the next k years given that there is no in the next n years.

$$\mathbb{P}(Y \geq k \mid Y \geq n) = \mathbb{P}\big(Y \geq (k-n)\big) = e^{-.04(k-n)}$$

by the memorylessness property.

Example 2. *Continue on from the previous example. Suppose we know that the mean textbook cost is indeed* \$150*. However, among textbooks used in STEM classes, the mean is* \$250*. Suppose further that individual textbook cost assumes a normal distribution.*

(a) Suppose that 25% *of all textbooks are over* \$180*. Find the standard deviation of the cost of all textbooks.*

(b) Suppose that 15% *of STEM textbooks are below* \$200*. Find the standard deviation of the cost of STEM textbooks.*

Solution:

Let (μ_1, σ_1) be the mean and standard deviation of the cost of all textbooks, and (μ_2, σ_2) for that of STEM.

(a) We know that $\mu_1 = 150$, and we want to find σ_1. From the hypothesis:

$$\mathbb{P}(X_1 > 180) \;=\; \mathbb{P}(Z > {}^{30}/_{\sigma_1}) = .25$$

Looking up the z-table gives us the critical point of .67:

$$\frac{30}{\sigma_1} = .67 \quad\implies\quad \sigma_1 = 45$$

(b) We know that $\mu_2 = 250$, and we want to find σ_2. From the hypothesis:

$$\mathbb{P}(X_2 < 200) \;=\; \mathbb{P}(Z < {}^{-50}/_{\sigma_2}) = .15$$

Looking up the z-table gives us the critical point of -1.04:

$$\frac{-50}{\sigma_2} = -1.04 \quad\implies\quad \sigma_2 = 48$$

8.3 Statistical Inference

Key concepts that you need to have a solid understanding of (or at least be comfortable talking about):

1. From population to sample:
   (a) $\hat{p}$ and $\bar{x}$
   (b) Law of Large Number
   (c) Central Limit Theorem
   - Regarding CLT, be sure to check for the conditions.
   - Important! CLT applies to averages and not individual observations.
   - Make sure you know what each of the 2 theorems above provide us with.
2. From sample to population:
   (a) Confidence interval: building and interpreting
   - Qualitative data: z-score
   - Quantitative data:
     – Approximation: z-score
     – Exact distribution: t-score, if the data points $x_i \sim \mathcal{N}(\mu, \sigma)$

   (b) Margin of Error: control through sample size or the confidence level

Example 3. *Suppose among the new patients at a local hospital,* 64% *are there for Corona virus related reasons. Suppose we are going to randomly survey* 2500 *new patients.*

(a) Find the probability that there are between 1580 *and* 1640 *patients for Corona virus related reasons.*

(b) Suppose the original proportion is 64% *is not correct, and among the* 2500 *patients,* 1500 *are for Corona virus related reasons. Build and interpret a* 95% *confidence interval for the true proportion.*

Solution:

(a) Assume p be the (assumed) true proportion, then $p = .64$. We check for the conditions of Central Limit Theorem and all conditions are satisfied. We now apply the theorem to get

$$\hat{p} \sim \mathcal{N}(.64, .0096)$$

We note that observing the total number between 1580 and 1640 is the same as observing the proportion between .632 and .656, which implies that:

$$\begin{aligned}\mathbb{P}\Big(\text{total number } \in [1580, 1640]\Big) &= \mathbb{P}\big(.632 \le \hat{p} \le .656\big) \\ &= \mathbb{P}\left(\frac{-5}{6} \le Z \le \frac{5}{3}\right) = \Phi(5/3) - \Phi(-5/6) = .75\end{aligned}$$

(b) Now, we observe that $\hat{p} = \frac{1500}{2500} = .6$. By the Central Limit Theorem, we know that $\hat{p}$ approximately has a normal distribution. At 95% confidence, the critical value is $z^* = 1.96$:

$$95\% \text{ CI} = \left(\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = \Big(.6 \pm (1.96)(.0098)\Big) = \Big(6 \pm .019\Big)$$

Example 4. *Suppose we survey* 64 *textbooks randomly and find out that the average cost is* \$150 *with a standard deviation of* \$40*. Suppose further each textbook is iid and is normally distribution with the same mean and variance.*

Build a 90% *confidence interval for the cost of textbooks.*

Solution:

Let $\bar{x}$ be the observed average cost, then

$$n = 64, \qquad \bar{x} = 150, \qquad s = 40$$

Since the individual cost has a normal distribution, we will use the critical value from the t-distribution with 63 degrees of freedom. At 90% confidence, we have $t^*_{63} = 1.67$:

$$90\% \text{ CI} = \left(\bar{x} \pm t^*_{63} \cdot \frac{s}{\sqrt{n}}\right) = \Big(150 \pm (1.67)(5)\Big) = \Big(150 \pm 8.35\Big)$$