BIOS6611 Homework 2: Data Distributions, Expected Value, and the Central Limit Theorem

Contents

Discrete Distributions: Binomial and Poisson

A random variable for which there exists a discrete set of numeric values is a **discrete random variable**.[1] Discrete random variables are assigned probabilities by a mathematical relationship called a **probability mass function (PMF)** or **probability distribution**.[2] Examples of discrete probablilty distributions include the binomial and Poisson.

[1] Def. 4.2 Rosner

[2] Def. 4.4 Rosner

The **binomial** distribution is characterized by the number of independent trials (n), where each trial has two possible outcomes (success or failure), the number of successes (k), and the probability of success (p). The binomial PMF is defined:[3]

[3] Equation 4.5 Rosner

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n$$

The **Poisson** distribution is characterized by the rate parameter (λ), and the number of events (k) occurring in a time period (t). The Poisson PMF is defined:

$$Pr(X = k) = e^{-\lambda} \lambda^k / k!, \qquad k = 1, 2, ...$$

When the sample size is large and the probability is small, the Poisson distribution is a valid approximation to the binomial distribution, where $\lambda = np$.[4] How large is large, and how small is small? Rosner provides a conservative rule to use the approximation when $n \geq 100$ and $p \leq 0.01$.

[4] Eq. 4.10 Rosner

Continuous Distributions: Exponential and Normal

A random variable whose possible values cannot be enumerated is a **continuous random variable**.[5] Continuous random variables are

[5] Def. 4.3 Rosner

assigned probabilities by a mathematical relationship called a **probability density function (PDF)**. Examples of continuous probability distributions include the exponential and normal (also sometimes called Gaussian).

The **exponential** distribution is characterized by a rate parameter (λ). The exponential PDF is defined:[6]

[6] Def B.3 C&H

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \lambda > 0$$

The exponential is the only continuous distribution with the **memoryless property**[7] (i.e. the distribution of "waiting time" until a certain event does not depend on how much time has already elapsed):

[7] Def B.21 C&H

$$P(X \geq t + h | X > h) = P(X \geq t)$$

The **normal** (also called **Gaussian**) distribution is also a continuous distribution. It is the most common distribution used in statistics, and will be discussed further in the next homework.

The Central Limit Theorem

The Central Limit Theorem (CLT) is a surprising result that applies to *any* population distribution with a finite mean (that is, when $E(X) = \mu$ exists). At large sample sizes ("asymptotically"), the sampling distribution of the mean converges in distribution to a normal distribution.

Or, more mathematically, for a population with mean μ and standard deviation σ, the CLT says that as $n \to \infty$, $\bar{X} \to N(\mu, \frac{\sigma}{n})$.[8]

[8] This distribution strongly resembles the sampling distribution of the mean that we discussed in the previous section. The distinction is that $\bar{X} \sim N(\mu, \frac{\sigma}{n})$ holds true regardless of sample size for a normally distributed population. In contrast, this statement is only true at large sample sizes for non-normal populations.

A common point of confusion: The CLT is a statement about the sampling distribution of the mean ($\bar{X}$), not the distribution of the sample (X-values). The distribution of the sample becomes more similar to the distribution of the population as n increases.

The CLT is *extremely* useful because at high sample size we now can apply techniques developed for statistical inference on normally distributed populations–even if the population distribution is extremely skewed or we don't know the underlying distribution.

What sample size is "large"? Large is heuristically defined as $n \geq 30$. This rule-of-thumb is generally well-known, but due to compuing advances, there might be better alternatives if your worried that your sample size is too small[9].

[9] Namely, non-parametric methods or bootstrap / permutation methods. BIOS6611 students will read Hesterberg, Tim. "It's Time To Retire the 'n >= 30' rule." *Proceedings of the Joint Statistical Meetings*, Alexandria, VA (2008) let in the semester.

Expected Value and Variance (Reminder from RLab1)

The **expected value** represents the "average" value of the random variable. The **expected value of a discrete random variable** is defined as[10]:

[10] Def 4.5 Rosner

$$E(X) = \mu = \sum_{i=1}^{R} x_i Pr(X = x_i)$$

The **expected value of a continuous random variable** is defined as[11]:

[11] Def A.3 C&H

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx$$

The **variance** represents the spread of a random variable, relative to the expected value. The **variance of a discrete random variable**, denoted by Var(X), is defined by[12]:

[12] Def 4.6 Rosner

$$Var(X) = \sigma^2 = \sum_{i=1}^{R} (x_i - \mu)^2 Pr(X = x_i)$$

The **variance of a continuous random variable** is defined by[13]:

[13] Def A.3 C&H

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

The variance is related to the expected value through the following equation[14]:

[14] Prop A.2 C&H

$$Var(X) = E[X^2] - E[X]^2$$

References:

- C&B = Casella, G., & Berger, R. L. (2001). Statistical Inference (2nd edition). Australia; Pacific Grove, CA: Duxbury Press.
- C&H = Chihara, L. M., & Hesterberg, T. C. (2011). Mathematical Statistics with Resampling and R (1 edition). Hoboken, N.J: Wiley.
- Rosner = Rosner, B. (2010). Fundamentals of Biostatistics (7 edition). Boston: Duxbury Press.

Exercises

For full credit, show your work and code for all problems, unless noted otherwise.

Exercise 1: Discrete Distributions: Binomial and Poisson

Your classmate is backpacking in Patagonia. While there, she discovers that 2.5% of the people she meets in the region are affected by pulmonary sarcoidosis. Wondering whether this sample prevalence is unusually high, she begins calculating the probability of her sample prevalence using a binomial distribution. However, this quickly becomes too computationally intensive. She wonders whether she could use the Poisson approximation instead to reduce the computational burden. Needing help, she writes you for advice.

1a. Calculate the probability that 2.5% of Patagonians have the disease, assuming a sample size of 120 and population prevalence of 1%. Use both the exact binomial probability and the Poisson approximation of it. Compare the two.

1b. Allow your sample size to vary between 80 and 400 (by an increment of 40), while the population prevalence varies between 0.25% and 2.5% (by an increment of 0.25%).[15] The prevalence in your sample is still 2.5%. Calculate the difference between the exact binomial probability and the Poisson approximation of the binomial, under all combinations of parameters. Plot the results.[16]

1c. At what sample size and prevalence would you recommend that your friend use the Poisson approximation to the binomial? How does this compare to the general recommendation given by Rosner?

[15] Note: There are 90 different combinations here

[16] Hint:
```
n=seq(80,400,by=40)
p=seq(0.0025,.025,by=.0025)
np<-expand.grid(n=n,p=p)
```

Exercise 2: Expected Value and Variance for Exponential Distribution

Sally just started the Master's program in biostatistics at the University of Colorado Anschutz Medical Campus. She is originally from Iowa. While she loves Iowa and all of its cornfields, she desires to establish residency in Colorado, so that she will qualify for in state tuition the following year. Sally goes to the Division of Motor Vehicles with all of her forms. Before she enters the building, Sally wonders how long she should expect to wait in line before being helped. Assume the service times follow an exponential distribution with a rate of 3 people helped per hour.

2a. How long should Sally expect to wait in line? Use the definition of expected value and calculus to answer this problem.[17]

2b. What is the variation around this estimate? Use calculus to answer this problem.[18]

2c. Reproducibly simulate an Exponential(3) distribution of size 100,000. Calculate the mean and variance of your simulated distribution. How similar are these values to your answers above?

2d. Now suppose that Sally has been at the DMV for 10 minutes and has not been helped. Assume Sally is still just as oblivious about the number of people ahead of her as when she got there. How long should Sally expect to wait now?[19]

[17] Hint: Integration by parts necessary. You can also use some math software or a website like wolframalpha.com to calculuate the integration and check you answers.

[18] Hint: $Var(X) = E[X^2] - E[X]^2$

[19] Hint: Memoryless property of the exponential function.

Exercise 3: Ozone Status for Normal Approximation to the Binomial

Load ozone.csv into R. This is EPA data (epa.gov/outdoor-air-quality-data) that tabulates ozone levels from January 2017 to June 2017 for the Denver-Aurora-Lakewood area.

3a. Estimate the daily probability of "good" ozone levels.

3b. Making the assumption that the probability of ozone levels being rated "good" is constant from day-to-day, use the binomial distribution to calculate the "exact" probability that at least 5 of the next 7 days will have "good" ozone.

3c. Recalculate this probability using the normal approximation to the binomial.

3d. Do you think it's believable that the total days with "good" ozone status follows a binomial distribution? Justify your answer, either with domain knowledge, statistics intuition, or by using the data itself.
