BIOS6611 Homework 2: Discrete Distributions, Expected Value, and Properties of Estimators

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.

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]

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

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 $p \leq 0.01$.

Continuous Distributions: Exponential and Normal

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

[1] Def. 4.2 Rosner

[2] Def. 4.4 Rosner

[3] Equation 4.5 Rosner

[4] Eq. 4.10 Rosner

[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]

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

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):

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

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.

Expected Value and Variance

The **expected value** represents the "average" value of the random variable. The **expected value of a discrete random variable** is defined as[8]:

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

The **expected value of a continuous random variable** is defined as[9]:

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

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[10]:

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

The **variance of a continuous random variable** is defined by[11]:

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

The variance is related to the expected value through the following equation[12]:

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

[6] Def B.3 C&H

[7] Def B.21 C&H

[8] Def 4.5 Rosner

[9] Def A.3 C&H

[10] Def 4.6 Rosner

[11] Def A.3 C&H

[12] Prop A.2 C&H

Properties of Estimators: Bias, Consistency and Efficiency

Say we are interested in estimating the mean height of a population. Since we cannot collect data on the entire population, we collect a sample, or subset of the population. Then, we use the mean height of the sample to estimate the mean height of the population. In this case, we call the sample mean an **estimator** of the population mean.

We desire that all estimators share certain properties. For example, we like an estimator ($\hat{\theta}$) to be unbiased, that is, equal to the parameter (θ) it is estimating, on average. **Bias** is defined as:

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Some estimators may be biased when the sample size is small. However, when the sample size becomes infinite, we desire that this bias disappears. This property is called **consistency**. A sequence of estimators ($W_n = W_n(X_1, X_2, ..., X_n)$) is consistent for θ iff:[13]

$$lim_{n \to \infty} P_\theta(|W_n - \theta| < \epsilon) = 1$$

for every $\epsilon > 0$.

Another important property of estimators is **efficiency**. Efficiency depends on the variance of the estimator as the sample size becomes infinite.

The efficiency of two estimators can be compared using the **asymptotic relative efficiency (ARE)** equation. If two estimators (V_n, W_n) are consistent and their variances converge in distribution, the ARE of V_n with respect to (wrt) W_n is given by[14]:

$$ARE(V_n, W_n) = Var(W_n)/Var(V_n)$$

[13] Def 10.1.1 C&B

[14] Def 10.1.16 C&B

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?

Exercise 2: Expected Value and Variance

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.

[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)
```

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?[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]

Exercise 3: Properties of Estimators: Bias, Consistency, and Efficiency

Drs. Bob and Billy are friends with a shared interest in heights. They learned that the average height for the population of adult male patients at the University of Colorado Hospital (UCH) is 70 inches. Dr. Bob hypothesizes that the median height of adult male patients that arrive at the hospital the next day will be close to the population average. Assume the heights of adult male patients seen at the UCH follow a normal distribution with mean=70 inches and variance=15 inches^2.

3a. Assume that 100 patients are seen the next day. If Dr. Bob calculates the median height for the adult male patients seen on that day, what is the **bias** of his median estimate wrt the population mean? (Hint: Simulate a normal distribution of size n=100)

3b. Although improbable, now assume the number of patients seen in a day increases. Increase the sample size from 100 to 100,000, by 100 person increments. Calculate the bias for the varying sample sizes. Plot the results. What does this say about the **consistency** of the median estimate wrt the population mean?

3c. How does the variance of the data wrt the median estimator change as the sample size increases?[20]

3d. Dr. Billy bets Dr. Bob that the sample mean is more **efficient** (i.e. less variable about the population mean) than the sample median. To compare the relative efficiency of estimators, simulate 10,000 normal distributions with sample size n=1000, population mean=70

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

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

[20] $Var(X) = \sum_{\forall i}(x_i - median)^2/n$, where $\forall i$ means "for all" values of i

inches, and variance=15 inches^2. Calculate the median and mean for each simulation. Then compare the variance of the set of sample medians to the variance of the set of sample means. Using the results of your simulation, which estimator is more efficient?

3e. Extra Credit: What is the Cramer-Rao Lower Bound, and why does it relate to this exercise?[21]

[21] Hint: Feel free to Google or refer to a textbook for help answering this question.