Biostatistics [SBE304] (Fall 2019)
Tutorial 2

Discrete Random Variables and Probability Distributions

R (probability distributions, data manipulation and basic visualization)

Prof. Ayman M. Eldeib   Asem Alaa

Thursday 3$^{\text{rd}}$ October, 2019

1   Tutorial facts

The problems in this tutorial comprises:

(A) Programming Works:

- ❯_ **6** programming in-class demos.
- ⟨⟩ **3** programming homework.
- 🖥 **1** self-practicing programming works.

(B) Problem Set:

- 🗪 **4** problems to be solved in-class.
- 🏠 **8** problems homework.
- 🧑‍🏫 **6** self-practicing problems.

Join this GitHub assignment page to create a repository for your submissions: `https://classroom.github.com/a/qoNe7Vv6`

2   Discrete Random Variables and Probability Distributions

2.1   Pre-class reading

1. Lecture notes of "Discrete Random Variables" by Prof. Ayman M. Eldeib

2. Chapter 3 (pp. 42-65) of Montgomery's textbook

3. Chapter 1: Generative Models for Discrete Data from Modern Statistics for Modern Biology *by Susan Holmes and Wolfgang Huber*

2.2   Chapter overview

2.2.1   Discrete RVs

Probabilities assigned to various outcomes in $\mathcal{S}$ in turn determine probabilities associated with the values of any particular rv X.

Probability Mass Fxn/Probability Distribution,(pmf):

$$p(x) = P(X = x)$$

The conditions $p(x) \geq 0$ and $\sum_{\text{all possible x}} p(x) = 1$ are required for any pmf.

Cumulative Distribution Function   (To compute the probability that the observed value of X will be at most some given x) **Cumulative Distribution Function(cdf):** $F(x)$ of a discrete rv variable X with pmf $p(x)$ is defined for every number x by

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y)$$

For any number x, $F(x)$ is the probability that the observed value of X will be at most x.

For discrete rv, the graph of $F(x)$ will be a step function- jump at every possible value of X and flat btwn possible values.

For any two number a and b with $a \leq b$:

$P(a \leq X \leq b) = F(b) - F(a^-)$

$P(a < X \leq b) = F(b) - F(a)$

$P(a \leq X \leq a) = F(a) - F(a^-) = p(a)$

$P(a < X < b) = F(b^-) - F(a)$

(where a^- is the largest possible X value strictly less than a)

Taking $a = b$ yields $P(X = a) = F(a) - F(a - 1)$ as desired. **Expected value or Mean Value**

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

Describes where the probability distribution is centered and is just a weighted average of the possible values of X given their distribution.

The Expected Value of a Function: Sometimes interest will focus on the expected value of some function $h(x)$ rather than on just $E(x)$.

If the RV X has a set of possible values D and pmf $p(x)$, then the expected value of any function $h(x)$, denoted by $E[h(X)]$ or $\mu_{h(X)}$ is computed by

$$E[h(X)] = \sum_{D} h(x) \cdot p(x)$$

Properties of Expected Value:

$$E(aX + b) = a \cdot E(X) + b$$

Variance of X: Let X have pmf $p(x)$ and expected value μ. Then the $V(X)$ or σ_X^2 is

$$V(X) = \sum_{D} (x - \mu)^2 \cdot p(x) = E[(X - \mu)^2]$$

The standard deviation (SD) of X is $\sigma = \sqrt{\sigma}$

Alternatively,

$$V(X) = \sigma^2 = [\sum_{D} x^2 \cdot p(x)] - \mu^2 = E(X^2) - [E(X)]^2$$

Properties of Variance

1. $V(aX + b) = a^2 \cdot \sigma^2$

2. In particular, $\sigma_{aX} = |a| \cdot \sigma_x$

3. $\sigma_{X+b} = \sigma_X$

2.2.2 Discrete Distributions

parameter: Suppose p(x) depends on a quantity that can be assigned any one of a number of possible values, with each different value determining a different probability distribution. Such a quantity is called a parameter of distribution. The collection of all probability distributions for different values of the parameter is called a family of probability distributions.

The Binomial Probability Distribution

1. The experiment consists of $n \; trials$ where n is fixed

2. Each trial can result in either success (S) or failure (F)

3. The trials are independent

4. The probability of success $P(S)$ is constant for all trials

Note that in general if the sampling is without replacement, the experiment will not yield independent trials. However, if the sample size (number of trials) n is at most 5% of the population, then the experiment can be analyzed as though it were exactly a binomial experiment.

Binomial rv X: = no of S's among the n trials

pmf of a Binomial RV:,

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x} \quad : x = 0, 1, 2, \ldots$$

cdf for Binomal RV: Values in Tble A.1

$$B(x; n, p) = P(X \leq x) = \sum_{y=0}^{x} b(y; n, p)$$

Mean & Variance of X If $X \sim Bin(n, p)$ then

$$E(X) = np \quad V(X) = npq$$

Hypergeometric Probability Distribution The Hypergeometric Distribution The assumptions leading to the hypergeometric distribution are as follows:

1. The population or set to be sampled consists of N individuals, objects, or elements (a finite population).

2. Each individual can be characterized as a success (S) or a failure (F), and there are M successes in the population.

3. A sample of n individuals is selected without replacement in such a way that each subset of size n is equally likely to be chosen.

The RV of interest is X = the number of S's in the sample. The probability distribution of X depends on the parameters n, M, and N, so we wish to obtain $P(X = x) = h(x; n, M, N)$.

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Negative Binomial Distribution

1. The experiment consists of independent trials

2. Each trial can result in either Success(S) or Failure(F)

3. The probability of success is constant from trial to trial

4. The experiment continues until a total of r successes have been observed, where r is a specified integer

RV Y: = the no of trials before the rth success.
Negative Binomial rv: $X = Y - r$ the number of failures that precede the rth success. In contrast to the binomial rv, the number of successes is fixed while the number of trials is random.
pmf of the negative binomial rv : with parameters r = number of S's and $p = P(S)$ is

$$nb(x; r, p) = \binom{x + r - 1}{r - 1} p^r (1 - p)^x \qquad x = 0, 1, 2, \ldots$$

Mean & Variance of negative binomial rv X: with pmf $nb(x; r, p)$

$$E(X) = \frac{r(1 - p)}{p} \quad V(X) = \frac{r(1 - p)}{p^2}$$

Geometric Distribution **RV X:** = the no of trials before the 1st success.
pmf of the geometric rv :
$$p(x) = q^{x-1} p$$
$$E(X) = \sum x q^{x-1} p = 1/p$$

The Poisson Probability Distribution Useful for modeling rare events
1) independent: no of events in an interval is independent of no of events in another interval
2) Rare: no 2 events at once
3) Constant Rate: average events/unit time is constant ($\mu > 0$)
RV X= no of occurrence in unit time interval
Possion distribution/ Poisson pmf: of a random variable X with parameter $\mu > 0$ where

$$p(x; \mu) = \frac{e^{-\mu} \cdot \mu^x}{x!} \quad x = 0, 1, 2, \ldots$$

Binomial Approximation: Suppose that in the binomial pmf $b(x; n, p)$, we let $n \to \infty$ and $p \to 0$ in such a way that np approaches a value $\mu > 0$. Then $b(x; n, p) \to p(x; \mu)$.
That is to say that in any binomial experiment in which n(the number of trials) is large and p(the probability of success) is small, then $b(x; n, p) \approx p(x; \mu)$, where $\mu = np$.
Mean and Variance of X: If X has probability distribution with parameter μ, then $E(X) = V(X) = \mu$

2.3   Problem Set

1. PROBLEM

Roulette winnings   In the game of roulette, a wheel is spun and you place bets on where it will stop. One popular bet is that it will stop on a red slot; such a bet has an $\frac{18}{38}$ chance of winning. If it stops on red, you double the money you bet. If not, you lose the money you bet. Suppose you play 3 times, each time with a $1 bet. Let Y represent the total amount won or lost. Write a probability model for Y.

1. SOLUTION

2. PROBLEM

Multiple choice quiz   In a multiple choice quiz there are 5 questions and 4 choices. Robin has not studied for the quiz at all, and decides to randomly guess the answers. What is the probability that

1. the first question she gets right is the 3rd question?

2. she gets exactly 3 or exactly 4 questions right?

3. she gets the majority of the questions right?

2. SOLUTION

3. PROBLEM

Species hot spots.   "Hot spots" are species-rich geographical areas. A Nature (Sept. 1993) study estimated the probability of a bird species in Great Britain inhabiting a butterfly hot spot at .70. Consider a random sample of 4 British bird species selected from a total of 10 tagged species. Assume that 7 of the 10 tagged species inhabit a butterfly hot spot.

1. What is the probability that exactly half of the 4 bird species sampled inhabit a butterfly hot spot?

2. What is the probability that at least 1 of the 4 bird species sampled inhabits a butterfly hot spot?

3. SOLUTION

4. PROBLEM

 What would be the most appropriate probability distribution for each of the following random variables:

1. Whether a tumor is benign or malignant.

2. Number of people with a malignant tumor out of 10 patients with tumor.

3. Size of tumors.

4. Number of people diagnosed with malignant tumor in California every year.

4. SOLUTION

5. Problem

Pollution control regulations.   A task force established by the Environmental Protection Agency was scheduled to investigate 20 industrial firms to check for violations of pollution control regulations. However, budget cutbacks have drastically reduced the size of the task force, and they will be able to investigate only 3 of the 20 firms. If it is known that 5 of the firms are actually operating in violation of regulations, find the probability that

1. None of the three sampled firms will be found in violation of regulations.

2. All three firms investigated will be found in violation of regulations.

3. At least 1 of the 3 firms will be operating in violation of pollution control regulations.

5. Solution

6. Problem

Distribution of slugs   The distributional pattern of pulmonate slugs inhabiting Libya was studied in the AIUB Journal of Science and Engineering (Aug. 2003). The number of slugs of a certain species found in the survey area was modeled using the negative binomial distribution. Assume that the probability of observing a slug of a certain species (say, Milax rusticus) in the survey area is .2. Let Y represent the number of slugs that must be collected in order to obtain a sample of 10 Milax rusticus slugs.

1. Give the probability distribution for Y as a formula.

2. What is the expected value of Y? Interpret this value.

3. Find $P(Y = 25)$.

6. Solution

7. Problem

Reflection of neutron particles   Recall that particles released into an evacuated duct collide with the inner duct wall and are either scattered (reflected) with probability .16 or absorbed with probability .84.

1. If 4 particles are released into the duct, what is the probability that all 4 will be absorbed by the inner duct wall? Exactly 3 of the 4?

2. If 20 particles are released into the duct, what is the prob- ability that at least 10 will be reflected by the inner duct wall? Exactly 10?

7. Solution

Mailrooms contaminated with anthrax. During autumn 2001, there was a highly publicized outbreak of anthrax cases among U.S. Postal Service workers. In Chance (Spring 2002), research statisticians discussed the problem of sampling mailrooms for the presence of anthrax spores. Let Y equal the number of mailrooms contaminated with anthrax spores in a random sample of n mailrooms selected from a population of N mailrooms. The researchers showed that Y has a hypergeometric probability distribution. Let r equal the number of contaminated mailrooms in the population. Suppose $N = 100$, $n = 3$, and $r = 20$.

1. Find $p(0)$.

2. Find $p(1)$.

3. Find $p(2)$.

4. Find $p(3)$.

8. Solution

9. Problem

Reverse cocaine sting. An article in The American Statistician (May 1991) described the use of probability in a reverse cocaine sting. Police in a midsize Florida city seized 496 foil packets in a cocaine bust. To convict the drug traffickers, police had to prove that the packets contained genuine cocaine. Consequently, the police lab randomly selected and chemically tested 4 of the packets; all 4 tested positive for cocaine. This result led to a conviction of the traffickers.

1. Of the 496 foil packets confiscated, suppose 331 contain genuine cocaine and 165 contain an inert (legal) powder. Find the probability that 4 randomly selected packets will test positive for cocaine.

2. Police used the 492 remaining foil packets (i.e., those not tested) in a reverse sting operation. Two of the 492 packets were randomly selected and sold by undercover officers to a buyer. Between the sale and the arrest, however, the buyer disposed of the evidence. Given that 4 of the original 496 packets tested positive for cocaine, what is the probability that the 2 packets sold in the reverse sting did not contain cocaine? Assume the information provided in part a is correct.

3. The American Statistician article demonstrates that the conditional probability, part b, is maximized when the original 496 packets consist of 331 packets containing genuine cocaine and 165 containing inert powder. Recalculate the probability, part b, assuming that 400 of the original 496 packets contain cocaine.

9. Solution

10. Problem

When to replace a maintenance system. An article in the Journal of Quality of Maintenance Engineering (Vol. 19, 2013) studied the problem of finding the optimal replacement policy for a maintenance system. Consider a system that is tested every 12 hours. The test will determine whether there are any flaws in the system. Assume the probability of no flaw being detected is .85. If a flaw (failure) is detected, the system is repaired. Following the 5th failed test, the system is completely replaced. Now let Y represent the number of tests until the system needs to be replaced.

1. Give the probability distribution for Y as a formula. What is the name of this distribution?

2. Find the probability that the system needs to be replaced after 8 total tests.

Extinct New Zealand birds. Refer to the Evolutionary Ecology Research (July 2003) study of the patterns of extinction in the New Zealand bird population. Of the 132 bird species saved in the **NZBIRDS** file, 38 are extinct. Suppose you randomly select 10 of the 132 bird species (without replacement) and record the extinct status of each.

1. What is the probability that exactly 5 of the 10 species you select are extinct?

2. What is the probability that at most 1 species is extinct?

11. Solution

Elevator passenger arrivals. A study of the arrival process of people using elevators at a multi-level office building was conducted and the results reported in Building Services Engineering Research and Technology (Oct., 2012). Suppose that at one particular time of day, elevator passengers arrive in batches of size 1 or 2 (i.e., either 1 or 2 people arriving at the same time to use the elevator). The researchers assumed that the number of batches, N, arriving over a specific time period follows a Poisson process with mean l = 1.1. Now let X N represent the number of passengers (either 1 or 2) in batch N and assume the batch size has a Bernoulli distribution with $p = P(X_N = 1) = 0.4$ and $q = P(X_N = 2) = 0.6$. Then, the total number of passengers arriving over a specific time period is $Y = \sum_{i=1}^{N} X_i$. The researchers showed that if $X_1, X_2, ..., X_N$ are independent and identically distributed random variables and also independent of N, then Y follows a compound Poisson distribution.

1. Find $P(Y = 0)$, i.e., the probability of no arrivals during the time period. [Hint: $Y = 0$ only when $N = 0$.]

2. Find P1Y = 12, i.e., the probability of only 1 arrival during the time period. [Hint: Y = 1 only when N = 1 and X 1 = 1.]

3. Find $P(Y = 2)$, i.e., the probability of 2 arrivals during the time period. [Hint: $Y = 2$ when $N = 1$ and $X_1 = 2$, or, when $N = 2$ and $X_1 + X_2 = 2$. Also, use the fact that the sum of Bernoulli random variables is a binomial random variable.]

4. Find $P(Y = 3)$, i.e., the probability of 3 arrivals during the time period. [Hint: $Y = 3$ when $N = 2$ and $X_1 + X_2 = 3$, or, when $N = 3$ and $X_1 + X_2 + X_3 = 3$.]

12. Solution

Use of road intersection. The random variable Y, the number of cars that arrive at an intersection during a specified period of time, often possesses (approximately) a Poisson probability distribution. When the mean arrival rate λ is known, the Poisson probability distribution can be used to aid a traffic engineer in the design of a traffic control system. Suppose you estimate that the mean number of arrivals per minute at the intersection is one car per minute.

1. What is the probability that in a given minute, the number of arrivals will equal three or more?

2. Can you assure the engineer that the number of arrivals will rarely exceed three per minute?

13. Solution

---

14. Problem

Tapeworms in fish. The negative binomial distribution was used to model the distribution of parasites (tapeworms) found in several species of Mediterranean fish (Journal of Fish Biology, Aug. 1990). Assume the event of interest is whether a parasite is found in the digestive tract of brill fish, and let Y be the number of brill that must be sampled until a parasitic infection is found. The researchers estimate the probability of an infected fish at 0.544. Use this information to estimate the following probabilities:

1. $P(Y = 3)$

2. $P(Y \leq 2)$

3. $P(Y > 2)$

14. Solution

---

15. Problem

The number of messages that arrive at a Web site is a Poisson random variable with a mean of five messages per hour.

1. What is the probability that five messages are received in 1.0 hour?

2. What is the probability that 10 messages are received in 1.5 hours?

3. What is the probability that fewer than two messages are received in 0.5 hour?

4. Determine the length of an interval of time such that the probability that no messages arrive during this interval is 0.90.

15. Solution

---

16. Problem

The probability that your call to a service line is answered in less than 30 seconds is 0.75. Assume that your calls are independent.

1. If you call 10 times, what is the probability that exactly nine of your calls are answered within 30 seconds?

2. If you call 20 times, what is the probability that at least 16 calls are answered in less than 30 seconds?

3. If you call 20 times, what is the mean number of calls that are answered in less than 30 seconds?

4. What is the probability that you must call four times to obtain the first answer in less than 30 seconds?

5. What is the mean number of calls until you are answered in less than 30 seconds?

16. Solution

---

17. Problem

Mercury in seafood. An issue of Consumer Reports found widespread contamination and mislabeling of seafood in supermarkets in New York City and Chicago. The study revealed one alarming statistic: 40% of the swordfish pieces available for sale had a level of mercury above the Food and Drug Administration (FDA) maximum amount. For a random sample of three swordfish pieces, find the probability that

1. All three swordfish pieces have mercury levels above the FDA maximum.

2. Exactly one swordfish piece has a mercury level above the FDA maximum.

3. At most one swordfish piece has a mercury level above the FDA maximum.

17. Solution

🏠   **18. Problem**

Suppose that a healthcare provider selects 20 patients randomly (without replacement) from among 500 to evaluate adherence to a medication schedule. Suppose that 10% of the 500 patients fail to adhere with the schedule. Determine the following:

1. Probability that exactly 10% of the patients in the sample fail to adhere.

2. Probability that fewer than 10% of the patients in the sample fail to adhere.

3. Probability that more than 10% of the patients in the sample fail to adhere.

4. Mean and variance of the number of patients in the sample who fail to adhere.

18. Solution

---

3 ⓡ Programming Language

3.1 ⓡ Programming Problems

>_   **1. Problem**

Pollutant Mean (deferred from week 1) The data in the directory **data/pollution_data/specdata** of the demo repository (clone or pull) [1] contains measurements of several pollutant substances in air. Using R script, implement a function that loads multiple **csv** files into a single table and computes the mean value of a given **pollutant**.

The demo file **pollutantMean.R** will be uploaded/updated on the following directory **week1** of the demo repository.

---

[1] https://github.com/sbme-tutorials/biostatistics-sbe304.git

2. PROBLEM

Propability Distribution Functions   From the manual pages, identify the difference between the following functions:

- `dbinom`, `pbinom`, `rbinom`;

- `dnbinom`, `pnbinom`, `rnbinom`;

- `dhyper`, `phyper`, `rhyper`;

- `dgeom`, `pgeom`, `rgeom`;

- `dpois`, `ppois`, `rpois`,

then explain the use case of each function.

3. PROBLEM

In this chapter we have concentrated on discrete random variables, where the probabilities are concentrated on a countable set of values. How would you calculate the probability mass at the value $X = 2$ for a binomial $b(10, 0.3)$ with dbinom? Use dbinom to compute the cumulative distribution at the value 2, corresponding to $P(X \leq 2)$, and check your answer with another R function.

4. PROBLEM

Use `?Distributions` in R to get a list of available distributions. Make plots of the probability mass or density functions for various distributions (using the functions named `dXXXX`), and list five distributions that are not discrete.

5. PROBLEM

Generate 100 instances of a Poisson(3) random variable. What is the mean? What is the variance as computed by the R function `var`?

6. PROBLEM

University admissions   Suppose a university announced that it admitted 2,500 students for the following year's freshman class. However, the university has dorm room spots for only 1,786 freshman students. If there is a 70% chance that an admitted student will decide to accept the offer and attend this university, what is the approximate probability that the university will not have enough dormitory room spots for the freshman class?

7. PROBLEM

Survey response rate   Pew Research reported that the typical response rate to their surveys is only 9%. If for a particular survey 15,000 households are contacted, what is the probability that at least 1,500 will agree to respond?

8. PROBLEM

Mutations along the genome of HIV (Human Immunodeficiency Virus) occur at random with a rate of 5×10^{-4} per nucleotide per replication cycle. This means that after one cycle, the number of mutations in a genome of about $10^4 = 10,000$ nucleotides will follow a Poisson distribution with rate 5.

1. What is the probability mass distribution of observing $0 : 12$ mutations in a genome of $n = 10^4$ nucleotides, when the probability is $p = 5 \times 10^{-4}$ per nuceotide? Is it similar when modeled by the binomial $b(n, p)$ distribution and by the Poisson $(\lambda = np)$ distribution ?

2. Simulate a mutation process along $10,000$ positions with a mutation rate of 5×10^{-4} and count the number of mutations. Repeat this many times (e.g 300000) and plot the distribution with the `barplot` function.

>_   9. PROBLEM

Whenever we note that we keep needing a certain sequence of commands, it's good to put them into a function. The function body contains the instructions that we want to do over and over again, the function arguments take those things that we may want to vary. Write a function to compute the probability of having a maximum as big as `m` when looking across `n` Poisson variables with rate `lambda`.

📄   10. PROBLEM

C. elegans genome nucleotide frequency: Is the mitochondrial sequence of C. elegans consistent with a model of equally likely nucleotides?

1. Explore the nucleotide frequencies of chromosome M by using a dedicated function in the Biostrings package from Bioconductor.

2. Test whether the *C. elegans* data is consistent with the uniform model (all nucleotide fequencies the same) using a simulation. Hint: This is our opportunity to use Bioconductor for the first time. Since Bioconductor's package management is more tightly controlled than CRAN's, we need to use a special install function (from the BiocManager package) to install Bioconductor packages.

   ```
   if (!requireNamespace("BiocManager", quietly = TRUE))
       install.packages("BiocManager")
   BiocManager::install(c("Biostrings", "BSgenome.Celegans.UCSC.ce2"))
   ```

   After that, we can load the genome sequence package as we load any other R packages.