BIOSTAT 701

Introduction to Statistical Theory and Methods I

Lynn Lin

Preparation

- **Binomial distribution**: https://duke.zoom.us/rec/share/JYMmLRxdyQBvkOqdWuu07NR4qxtCvhcBNL0pvmaqMEtFWwPrVwOb-YZ1sufJRd22.qEXkqT8JrJFuF0vR?startTime=1648497149000

- **Multinomial and hypergeometric distributions**: https://duke.zoom.us/rec/share/TxkNcbBpJu1xbVlah-0jrwLiUh16yHLosLPh_rBpUfsQywFgaUjV8S4dr9OeM4IJ.szhYTgu6zhNwCCq1

Random variables

- Random variable is a variable that takes on different values determined by chance.
- **Qualitative RV**: The possible values vary in kind but not in numerical degree, such as gender. They are also called categorical or nominal variable.
- **Quantitative RV**
  - Discrete RV: When the random variable can assume only a countable, sometimes infinitely many number of values. E.g., Number of books in BIOSTAT701 students' backpack; Flipping a fair coin, the number of tails you get before the first Head 2.
  - Continuous RV: When the random variable can assume uncountable number of values in a line interval. Examples: Weight, time, height

Definition and properties

- The probability distribution for a discrete RV is the probability $p(k) = P(X = k)$ associated with each value of k.
- The probability associated with every value of k lies between 0 and 1 (and may be 0 or 1).
- The sum of probabilities for all values of k is equal to 1.
- The probabilities for a discrete random variable are additive.

PMF

- Formally, for a discrete RV X with possible values $x_1, \ldots, x_k$, a probability mass function (PMF) is a function such that
  - $f(x_i) = P(X = x_i)$
  - $f(x_i) \geq 0$
  - $\sum_{i=1}^{k} f(x_i) = 1$

CDF

- The cumulative distribution function (CDF) for a discrete RV X is denoted as $F(x)$:
  - $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$
  - $0 \leq F(x) \leq 1$
  - If $x \leq y$, then $F(x) \leq F(y)$

Question

- Suppose the PMF for X is P(X=0) = 0.886; P(X = 1) = 0.111; P(X=2) = 0.003
- Calculate F(0), F(1), F(2) and plot F(x)

Example

- A bag contains dollar bills, 1 bill of $2, 3 bills of $5, 4 bills of $10, 2 bills of $20. Let X = the value of a bill randomly picked from the bag. What is the probability distribution of X? What is $P(X > 5)$, and $P(5 \leq X \leq 10)$?

Mean of a discrete RV

- The mean or expected value of the discrete RV X is $\mu = E(X) = \sum x_i f(x_i)$
  - The mean is a weighted average of all possible outcomes of X
  - Roughly speaking, the expectation of a random variable is that the average of outcomes if you sample the random variable infinite many times.

Variance of a discrete RV

- The variance of X, denoted as σ^2 or $V(X)$, is
  $E(X-\mu)^2 = \sum (x_i - \mu)^2 f(x_i) = \sum x_i^2 f(x_i) - \mu^2$
  - The standard deviation of X is $\sigma = \sqrt{\sigma^2}$

The Bernoulli distribution

- The Bernoulli distribution is a discrete distribution for a random variable with two possible values 0 = failure and 1 = success
- If we have many Bernoulli trials, and we sum the number of successes across the trials, this becomes the binomial distribution (which we will talk about later)
- For example, imagine we observe a health outcome (1=outcome present, 0=outcome absent) on 100 patients. This is actually 100 Bernoulli trials! The total number of patients with the outcome follows a binomial distribution.
- For now we will just focus on the features of a single Bernoulli trial.

The Bernoulli distribution

- PMF can be written in different ways:
  - $P(X = x) = p$, if $x = 1$; $P(X = x) = 1 - p$, if $x = 0$;
  - $P(X = x) = p^x(1 - p)^{1-x}$ for $x \in \{0,1\}$

Simulating Bernoulli trials in R

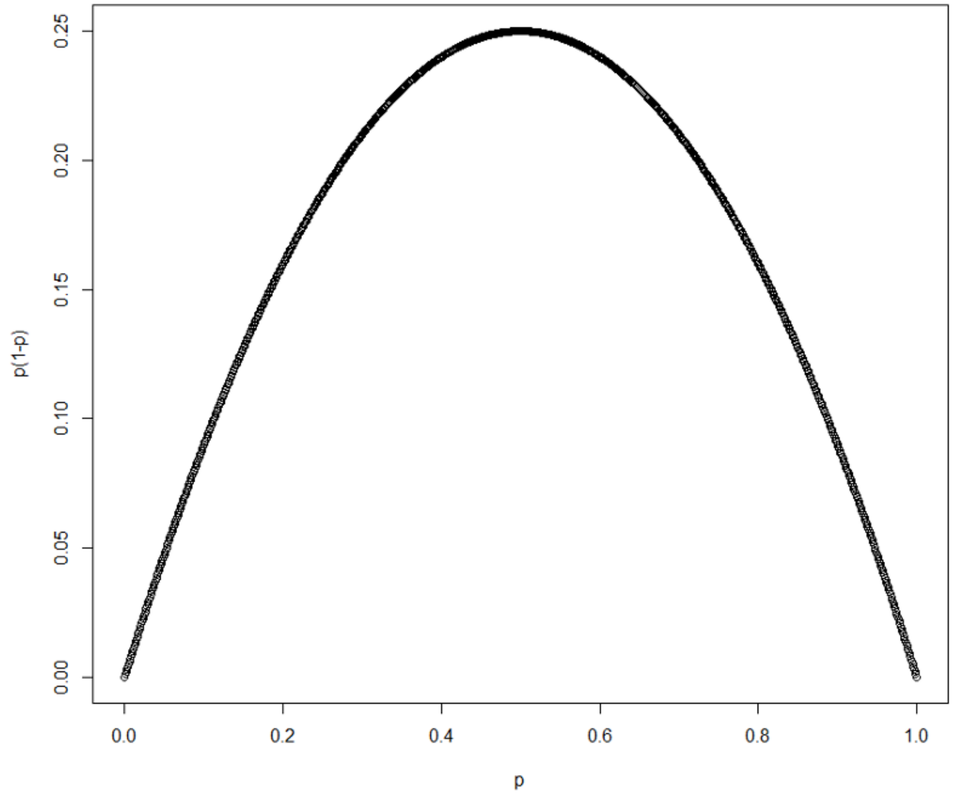
- We can use the rbinom() function in R to simulate a Bernoulli trial
- We can use the dbinom() function in R for PMF
- We can use the pbinom() function in R for Bernoulli CDF
- $$F(X) = P(X \leq x) = \begin{cases} 0 & x < 0 \\ 1 - p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Mean and variance

- For the Bernoulli distribution: $E(X) = p(1) + (1-p)(0) = p$
- $Var(X) = E(X^2) - E(X)^2 = p(1^2) + (1-p)(0^2) - p^2 = p(1-p)$
- But how to interpret variance of a Bernoulli random variable?

Variance

- Generally, variance gives a measure of how far from the mean the observations are (it might be easier to think about the SD)
- For the Bernoulli distribution, the variance depends on the mean
- The variance is highest at a mean of 0.5
- This is consistent with intuition because the closer the expected value is to 0.5 the more variability we would expect to see in a series of experiments




```
p = seq(0,1,.001)
varP = p*(1-p)
plot(p,varP,ylab="p(1-p)")
```

Two Bernoulli trials

- Consider 2 Bernoulli trials, what is the probability of observing the sequence {1,1}?
  - $p \times p = p^2$
  - This follows from conditional independence: $P(X_1 = 1 and X_2 = 1) = P(X_1 = 1)P(X_2 = 1 \mid X_1 = 1)$. But since the two trials are independent, we can replace $P(X_2 = 1 \mid X_1 = 1)$ by $P(X_2 = 1)$

Four Bernoulli trials

- Similarly, what is the probability of observing the sequence {1,1,1,1}?
  - p^4
  - We can label {1,1,1,1} as {4 successes in a row} and also as {exactly 4 successes}, since there is only one sequence which yields 4 successes.

Four Bernoulli trials

- Now consider the sequence {1,1,0,1}
  - $p \times p \times (1-p) \times p = p^3(1-p)$
- However, this event isn't the same thing as {exactly 3 successes}, since {0,1,1,1}, {1,0,1,1} and {1,1,1,0} also generate exactly 3 successes.
- In other words, the P{exactly 3 successes} = P{{0,111} or {1,0,1,1} or {1,1,0,1} or {1,1,1,0}}.
- These events are disjoint, so the probabilities add. Moreover, the component probabilities are identically $p^3(1-p)$, and so P{exactly 3 successes} = $4p^3(1-p)$.

Binomial distribution

- Binomial experiment satisfies the following four conditions:
  - The experiment consists of n identical trials.
  - Each trial results in one of the two outcomes, called a success and a failure.
  - The probability of success, denoted π, remains the same from trial to trial.
  - The n trials are independent. That is, the outcome of any trial does not affect the outcomes of the others.
- The binomial RV represents the probability of exactly X successes out of n trials.

Binomial distribution

- In general, if n is the number of trials, and π is the probability of success, then: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, where the possible values of k are 0, 1, 2, ..., n.
  - This is the PMF of the binomial distribution
- Thus, a random variable X which follows a binomial distribution with n trials and parameter π the probability of success can be denoted as $X \sim Bin(n, p)$.
- Note that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, and $n! = n \cdot (n-1) \cdot (n-2) \cdots 1$.

Note

- R has a function to calculate the value of the PMF for small to moderate values of n, and similarly for the CDF.
- When n is very large, it is usually better to use an approximation to the binomial distribution (as discussed later).

Summary

- Use the binomial distribution when the goal is to count the total number of successes (equivalently, the proportion of successes) out of a series of n Bernoulli trials with the same probability of success for each trial, where the order of the successes doesn't matter.
- For example, when the event in question is the presence or absence of a surgical complication, you only care about the number of people with complications, not their order within the dataset
- The binomial distribution can also be understood to be the sum of the {0,1} outcomes from n (independent) Bernoulli RVs.

Example

- Cross fertilizing a red and a white flowers produces red flower 25% of the time. Now we cross fertilize 5 pairs of red and white flowers and produce 5 offspring. Find the probability that (a) There will be no red flowered plants in the 5 offspring? (b) There will be 4 or more red flowered plants?

Mean and standard deviation of binomial distribution

- Suppose we have $X \sim Bin(n, p)$. We have
- $\mu = E(X) = np$
- $\sigma = \sqrt{Var(X)} = \sqrt{np(1-p)}$
- By the independence of the Bernoulli RVs

Example continued

- (c) Of the 5 cross-fertilized offspring, how many red flowered plants do you expect? (d) What is the standard deviation?

Multinomial distribution

- The multinomial distribution is a common distribution for characterizing categorical variables.
- Suppose a RV Z has k categories, we can code each category as an integer, thus, $Z \in \{1,2,...,k\}$
- Suppose that $P(Z = j) = p_j$, then the parameter $\{p_1, \ldots, p_k\}$ describes the entire distribution of Z with the constraint that $p_1 + p_2 + \cdots + p_K = 1$
- Further, suppose we generate $Z_1, \ldots, Z_n$ iid from the above distribution and let $X_j = \sum_{i=1}^{n} I(Z_i = j)$, i.e., the number of observations in the category j.

Multinomial distribution

- Then the random vector $X = (X_1, \ldots, X_k)$ is said to be from a multinomial distribution with parameter $\{p_1, \ldots, p_k\}$
  - $X \sim M_k(n; p_1, \ldots, p_k)$
- The PMF is $P(X = x) = P(X_1 = x_1, .., X_k = x_k) = \frac{n!}{x_1!, \ldots, x_k!} p_1^{x_1} \cdots p_k^{x_k}$
  - The multinomial coefficient $\frac{n!}{x_1!, \ldots, x_k!}$ is the number of possible ways to put n balls into k boxes.

Multinomial distribution

- The multinomial distribution is often used in the analysis of RxC "contingency tables", where R denotes the number of rows, C denotes the columns, and the interior of the table contains counts of the number of individuals falling within each cross-classification of the row and the column.

Population	Poor	Good	Excellent	Total
Drug A	25	30	45	100
Drug B	15	30	55	100
Overall	40	60	100	200

- Q: How many individuals are reported in the table?

Multinomial distribution

- Each row (except the “Total”) can be viewed as a random vector from a multinomial distribution.
  - E.g., the first row (25,30,45) can be viewed as a random draw from a multinomial distribution $M_3(n = 100; p_1, \ldots, p_3)$.
  - The second row can be viewed as the other random draw from the same distribution.

Population	Poor	Good	Excellent	Total
Drug A	25	30	45	100
Drug B	15	30	55	100
Overall	40	60	100	200

Hypergeometric distribution

- The hypergeometric distribution is a discrete probability distribution that describes the probability of successes in draws, without replacement.
- In general, given a set of N objects contains
  - K objects classified as successes
  - N-K objected classified as failures
- A sample of size n objects is selected randomly (without replacement) from the N objects, where $K \leq N$ and $n \leq N$.

Hypergeometric distribution

- Let the RV X denote the number of successes in the sample. Then X is a hypergeometric RV and
- $$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$
  - $\max\{0, n-(N-K)\} \leq x \leq \min\{n, K\}$

Geometric distribution

- In a series of independent Bernoulli trials with constant probability p of a success, let the RV X denote the number of trials until the first success.
- Then X is a geometric RV with parameter $0 < p < 1$
  - $f(x) = (1-p)^{x-1}p$; for $x = 1,2,...$
  - Note that the PMF is defined for $x \geq 1$ indicates that the trial that resulted in the first success is included in the total number of trials required to obtain a success.

Geometric distribution

- Alternatively, we can let the RV Y denote the number of trials before the first success, i.e., the number of failures that must occur prior to a success happening
- Then Y is a geometric RV with parameter $0 < p < 1$
  - $f(x) = (1-p)^x p$; for $x = 0,1,2,...$

Negative binomial distribution

- A generalization of a geometric distribution in which the RV X is the number of Bernoullis trials required until r successes occur results in the negative binomial distribution, with parameters $0 < p < 1$ and $r = 1,2,...$
  - $f(x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r$; for $x = r, r+1, r+2,...$

Poisson distribution

- The Poisson is a distribution which is often applied to counts.
  - E.g., if you have a count such as the number of events occurring within a certain time interval or a certain spatial area, then the Poisson distribution should come into consideration.
- It is also a special case of the binomial distribution with n approaching infinity and p approaching zero.
  - E.g., $Bin(n = 10^4, p = 10^{-6})$ is hard to compute
  - When values are that extreme, we can use approximates to make computation feasible

Poisson distribution

- Define $\lambda = np$
- Then, we can rewrite the Binomial PMF as

$$P(X = x) = \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^{x}\left(1 - \frac{\lambda}{n}\right)^{n-x} = \frac{n(n-1)\cdots(n-x-1)}{n^x}\frac{\lambda^x}{x!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^x}$$

  - $\frac{n(n-1)\cdots(n-x-1)}{n^x} \approx 1$
  - $(1-\lambda/n)^n \approx e^{-\lambda}$
  - $(1-\lambda/n)^x \approx 1$
- $P(X = x) = \frac{\lambda^x}{x!}e^{-\lambda}$