STAT 516: Some discrete distributions

Lecture 11: Binomial and related distributions

Prof. Michael Levine

October 1, 2015

Bernoulli distribution: a model

- A **Bernoulli experiment** is the basic unit
- The two outcomes must be mutually exclusive and exhaustive
- The pmf for X that has Bernoulli distribution is

$$p(x) = p^x(1-p)^{1-x}$$

  where $x \in \{0, 1\}$
- Thus, $\mathbb{E}\,X = p$ and $Var\,(X) = \sigma^2 = p(1-p)$

Binomial distribution: a model

- A number of successes in a sequence of n **Bernoulli trials** X has the **binomial distribution**
- If n is fixed, the number of ways to choose x successes out of n is $\binom{n}{x} = \frac{n!}{x!(n-x)!}$
- The **binomial pmf** of X is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

  for $x = 0, 1, 2, \ldots, n$ and 0 elsewhere
- Verify directly that $p(x)$ is a true pmf
- n and p are the **parameters** of the binomial distribution

Mgf of binomial distribution

- Verify directly that

$$M(t) = [(1-p) + pe^t]^n$$

- By differentiating, find that

$$\mathbb{E}(X) = \mu = np$$

and

$$Var(X) = \sigma^2 = np(1-p)$$

An alternative way of working with a binomial distribution

- Represent $X = \sum_{i=1}^{n} X_i$ where each $X_i \sim b(1, p)$ and independent
- Check that the expectation and the variance can be easily obtained this way!

Example I

- Guessing at random multiple choice exam answers...e.g. take $X \sim b(20, 0.5)$
- Compute $P($ every answer is wrong$)$ and $P($ every answer is right$)$
- You can use R commands $dbinom(k, n, p)$ for the pmf $P(X = k)$ and $pbinom(k, n, p)$ for the cdf $P(X \leq k)$

Example II

- Let $Y \sim b(n, 1/3)$ so $P(Y \geq 1) = 1 - \left(\frac{2}{3}\right)^n$
- What is the smallest value of n s.t. $P(Y \geq 1) > 0.8$? Can you find an exact integer?

Important result

- A sum of independent $X_i \sim b(n_i, p)$ is still binomial:

$$Y = \sum_{i=1}^{n} X_i \sim b\left(\sum_{i=1}^{m} n_i, p\right)$$

Negative binomial distribution

- X is the number of failures in a sequence of Bernoulli trials to achieve r successes
- The pmf is

$$p(y) = \binom{y+r-1}{r-1} p^r (1-p)^y$$

for $y = 0, 1, 2, \ldots$

- When $r = 1$ it is a **geometric distribution**
- As an exercise, check that the mgf of the negative binomial distribution is

$$M(t) = p^r [1 - (1-p)e^t]^{-r}$$

- When $r = 1$, this becomes exactly the mgf of the geometric distribution: $M(t) = p[1 - (1-p)e^t]^{-1}$

An alternative version of the negative binomial distribution

- Sometime X is assumed to be the overall number of *trials* (and not just failures) it takes to achieve r successes
- In that case, the pmf is clearly

$$p(y) = \binom{y-1}{r-1} p^{r-1}(1-p)^{y-r}$$

- Check that in this case the moment generating function of X is

$$M(t) = (pe^t)^r(1-(1-p)e^t)^{-r}$$

Example

- A couple plans to have children until they have two boys; let X be the number of children they have eventually
- Use the second definition of the negative binomial by modeling $X \sim NB(r, p)$ with $r = 2$ and $p = 0.5$. Its pmf is $p(x) = (x-1)(0.5)^x$ for $x = 2, 3, \ldots$
- For example, the probability that such a couple will have at least one girl is $P(X \geq 3) = 1 - P(X = 2) = 0.75$
- Check that $P(X \geq 6) = 1 - P(X \leq 5) \approx 0.19$

Example

- How many people do you have to meet before finding somebody who shares your birthday?
- That number X has a geometric distribution with $p = \frac{1}{365}$ if all birthdays are equally likely and the birthdays of the people you meet are independent
- The pmf of X is $p(x) = p(1-p)^{x-1}$
- Check that for any given k $P(X > k) = (1-p)^k$; thus, the probability that you will have to meet at least 1000 people in your pursuit is $\left(\frac{364}{365}\right)^{1000} = 0.064$

An important property of the geometric distribution

- Geometric distribution is **memoryless**: for any two positive integers m and n

$$P(X > m + n | X > n) = P(X > m)$$

- We will see later that this property is shared with the exponential distribution which is the continuous analog of the geometric distribution!

Some basic facts about the geometric and negative binomial distributions

- Geometric X: $\mathbb{E}\,X = \frac{1}{p}$, $Var\,X = \frac{1-p}{p^2}$
- Negative binomial X: $\mathbb{E}\,X = \frac{r}{p}$, $Var\,X = \frac{r(1-p)}{p^2}$
- Note that the result for the negative binomial can be easily obtained by using the following representation:

$$X = X_1 + X_2 + \cdots + X_r$$

where X_i is the geometric random variable measuring the number of additional trials needed to obtain ith success after $i-1$ success was achieved

Hypergeometric Distribution

- A typical application is the acceptance sampling
- Let N be the total and D the number of defectives
- A sampling without replacement gives the number of defectives in a sample size n

$$p(x) = \frac{\binom{N-D}{n-x}\binom{D}{x}}{\binom{N}{n}}$$

for $x = 0, 1, 2, \ldots, n$

- X has a **hypergeometric distribution** with parameters (N, D, n)

Example

- Capture-recapture sampling is a classical example
- The population consists of N animals; D captured, tagged and released
- In the second sample size n, the number of tagged $X \sim Hypergeo(n, D, N)$
- Typically, the true population size is estimated as $N = \frac{nD}{X}$
- Can be problematic in practice if $X = 0$ or if the tagged animals cluster together after release
- Corrected version called *Petersen's estimator* is used in wildlife biology, when taking a census, by governments for estimating tax fraud and/or the number of people afflicted with a particular infection

Basic properties of the hypergeometric distribution

- Check that its $\mathbb{E}\, X = n\frac{D}{N}$ and $Var\, X = n\frac{D}{N}\frac{N-D}{N}\frac{N-n}{N-1}$
- The expectation is the same as that of the binomial distribution with $p = \frac{D}{N}$
- The variance is different from the binomial by the factor $\frac{N-n}{N-1}$ which is called a **population correction factor**
- Can show that if $\frac{D}{N} \to p$ for some $0 < p < 1$ as $N \to \infty$ the hypergeometric pmf

$$p(x) = \frac{\binom{N-D}{n-x}\binom{D}{x}}{\binom{N}{n}} \to \binom{n}{x} p^x (1-p)^{n-x}$$

Multinomial distribution

- Choose a number n and k probabilities $p_1, \ldots, p_k$
- **Multinomial distribution** is a generalization of the binomial with the pmf

$$\frac{n!}{x_1!x_2!\ldots x_{k-1}!x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

  - a joint pmf of $X_1, X_2, \ldots, X_{k-1}$
- Must be $\sum_{i=1}^{k} p_i = 1$ and $\sum_{i=1}^{m} x_i = n$
- Notation: $X \sim Mult(n, p_1, \ldots, p_k)$

Example

- Let the fair die be rolled 30 times.
- What is the probability that each face is obtained exactly 5 times?
- Consider $X \sim Mult(30, p_1, \ldots, p_6)$ where each $p_i = \frac{1}{6}$
- Find

$$P(X_1 = 5, X_2 = 5, \cdots, X_6 = 5) = \frac{30!}{(5!)^6} \left(\frac{1}{6}\right)^5 \cdots \left(\frac{1}{6}\right)^5 = .0004$$

Marginal distributions of multinomial distribution

- **Trinomial distribution is a special case** with

$$p(x,y) = \frac{n!}{x!y!(n-x-y)!} p_1^x p_2^y p_3^{n-x-y}$$

- Its mgf is

$$M(t_1,t_2) = (p_1 e^{t_1} + p_2 e^{t_2} + p_3)^n$$

- The marginal mgfs are $M(t_1,0) = [(1-p_1) + p_1 e^{t_1}]^n$ and $M(t_2,0) = [(1-p_2) + p_2 e^{t_2}]^n$ so that $X_i \sim b(n,p_i)$, $i = 1,2$
- Clearly, X_1 and X_2 are not independent!!

Marginal distributions of multinomial distribution

- Can easily verify that $Y|x \sim b[n-x, p_2/(1-p_1)]$ and

$$\mathbb{E}(Y|x) = (n-x)\left(\frac{p_2}{1-p_1}\right)$$

- Recall that the squared correlation coefficient is the product of the two coefficients of x and y in the respective conditional means! Thus,

$$\rho = -\sqrt{\frac{p_1 p_2}{(1-p_1)(1-p_2)}}$$

Example

- let X_6 be the number of sixes in these 30 rolls
- The marginal distribution is binomial...so $X_6 \sim bin\left(30, \frac{1}{6}\right)$
- Thus, $P(X_6 \geq 5) = 1 - P(X_6 \leq 4) = .5757$