W1211 Introduction to Statistics
Lecture 9

Wei Wang

Oct 3, 2012

Binomial RV

- The binomial random variable X associated with a binomial experiment consisting of n trials is defined as

  X = the number of successes among the n trials.

- The pmf of a binomial rv X depends on the two parameters n and p, we denote the pmf by $b(x; n, p)$. The cdf will be denoted by

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

  Note that x can only take values in {0,1,…, n}.

Example

<u>Ex.</u> Roll a ten-sided die four times. What is the probability of getting exactly one three?

S = rolling a three.

F = rolling something other than a three.

P(S) = p = 0.1 and P(F) = 1-p = 0.9

Let X = the number of threes, then X is Bin(4, 0.1) and we want to calculate P (X=1). There are four possible ways of rolling a three: SFFF, FSFF, FFSF, FFFS.

P(SFFF) = P(S)P(F)P(F)P(F) = $(1-p)^3 p = (.9)^3(.1)$ = 0.0729

Similarly, P(FSFF) = P(FFSF) = P(FFFS) = 0.0729.

P(X=1) = P(SFFF) + P(FSFF) +P(FFSF) + P(FFFS)

= 4(0.0729) = 0.2916

Binomial pmf

- From the previous example, we see that
  P(X=1) = $b(1; 4, p) = 4(1-p)^3 p$
  = {# of outcomes with X=1} • {prob. of any particular outcome with X=1}

- Thus more generally, we have
  $b(x; n, p)$ = {# of outcomes with X=x} • {prob. of any particular outcome with X=x}

- The pmf of a binomial rv is

$$b(x;n,p) = \begin{cases} \boxed{\binom{n}{x}}\boxed{p^x(1-p)^{n-x}} & x = 0,1,2,3,\ldots,n \\ 0 & \text{otherwise} \end{cases}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Example

<u>Ex.</u> (Ten-sided die cont.) Use binomial pmf to verify P(X=1) we have calculated.

$$\mathrm{P}(\mathrm{X}=1)=\binom{4}{1}(1/10)^1(9/10)^3=\frac{4!}{1!3!}(1/10)^1(9/10)^3=0.2916$$

What is the probability of getting less than two 3's in four rolls?

$$\begin{aligned}\mathrm{P}(\mathrm{X}<2) &= \mathrm{P}(\mathrm{X}=0)+\mathrm{P}(\mathrm{X}=1)\\ &= \binom{4}{0}(1/10)^0(9/10)^4+\binom{4}{1}(1/10)^1(9/10)^3\\ &= 0.6561+0.2916=0.9477\end{aligned}$$

- Try using `dbinom(); pbinom()` to calculate the things above.

Example

Ex. Suppose we are searching for new apartments in the city, and our goal is to find an apartment among the top 5% (based on some criteria). Our strategy is to randomly sample 20 apartments from the pool, and choose the best out of these 20. What is the probability that we will accomplish our goal?

Mean and Variance of Binomial

- Proposition:

If $X \sim \text{Bin}(n, p)$, then $E(X) = np$, $\text{Var}(X) = np(1 - p) = npq$, and $\sigma_X = \sqrt{npq}$ (where $q = 1 - p$).

We'll show an easy proof in chapter 5.

Hypergeometric Distribution

Ex. (Socks example cont.) Suppose there are 50 colored socks in the drawer, of which 16 are red and the other 34 are blue. We are going to randomly draw 10 sock out of the drawer without replacement. What is the probability that we will have exactly 2 blue socks?

- As pointed out in the socks example, when we have a *finite* or *small* population, and we sample without replacement, the binomial approximation will not be appropriate.

- Notice that any subset of 10 socks in this example is equally likely to be chosen.

- Again, we use X = the number of successes (blue socks) in the sample we draw, then X is said to have the hypergeometric distribution.

Parameters

- It is easy to see that the probability distribution of X depends on three parameters:

  n = sample size (10 in socks example).

  M = total number of successes in the population (34 in socks example).

  N = total number of individuals in the population (50 in socks example).

  We wish to obtain P(X=x) = $h(x; n, M, N)$.

- P(X=2) = h(2; 10, 34, 50) = {# of outcomes with X=2} / {# of possible outcomes}.

- Thus we have

$$h(2; 10, 34, 50) = \frac{\overbrace{\boxed{\binom{34}{2}}}^{\text{\# of ways of selecting 2 blue socks}}\overbrace{\boxed{\binom{16}{8}}}^{\text{\# of ways of selecting 8 red socks}}}{\boxed{\binom{50}{10}}}$$

- To compute one can use R command: `choose(n,k)`,

  pmf: `dhyper(x,M,N-M,n)`, cdf: `phyper(x,M,N-M,n)`.

Hypergeometric pmf and statistics

- If X is the number of successes in a completely random sample of size n drawn from a population consisting of M successes and $(N - M)$ failures, then the distribution of X is given by

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

  for x an integer satisfying $\max(0, n - N + M) \le x \le \min(n, M)$.

- Proposition:

  $$\boxed{\text{If } X \sim \text{hypergeometric with pmf } h(x; n, M, N)\text{, then } E(X) = n(M/N),\quad Var(X) = (N - n)/(N - 1)\, n\, (M/N)\, (1 - M/N)\, .}$$

Connection with Binomial

- From the proposition, notice that if we let *p*=*M*/*N*, we get

$$\begin{aligned} \mathrm{E(X)} &= np \\ \mathrm{Var(X)} &= \left(\frac{N-n}{N-1}\right) \cdot np(1-p) \end{aligned}$$

- Notice that if we fix n, and let N be sufficiently large, Var(X) → $np(1 - p)$ which is the variance of a binomial rv. This is the reason why we can use a binomial model to approximate hypergeometric when population is large.

- $\left(\frac{N-n}{N-1}\right)$ is often called finite population correction factor.

Poisson Distribution

- Poisson Distribution is for describing outcomes that come in the form of count data, e.g., visits to a particular website during a time interval
- But unlike Binomial or Hypergeometric Distribution, there is no simple experiment that Poisson Distribution is based on.
- A random variable X is said to have Poisson Distribution with parameter $\mu (> 0)$ if the pmf of X is

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

Poisson Distribution PMF

- Verify the pmf is a valid pmf

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

Poisson Distribution PMF

- Verify the pmf is a valid pmf

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

- Recall from Calculus

$$e^{\mu} = 1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + \frac{\mu^4}{4!} + \cdots$$

Poisson Distribution PMF

- Verify the pmf is a valid pmf

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

- Recall from Calculus

$$e^{\mu} = 1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + \frac{\mu^4}{4!} + \cdots$$

- So

$$p(0;\mu) + p(1;\mu) + p(2;\mu) + \cdots = e^{\mu} \times e^{-\mu} = 1$$

Example

- Let X denote the number of creatures of a particular type captured in a trap during a given time period. Suppose that X has a Poisson distribution with =4.5, so on average traps will contain 4.5 creatures. Then the probability that a trap contains exactly five creatures is

$$P(X = 5) = \frac{e^{-4.5}(4.5)^5}{5!} = 0.1708$$

Example

- Let X denote the number of creatures of a particular type captured in a trap during a given time period. Suppose that X has a Poisson distribution with =4.5, so on average traps will contain 4.5 creatures. Then the probability that a trap contains exactly five creatures is
$$P(X = 5) = \frac{e^{-4.5}(4.5)^5}{5!} = 0.1708$$
- The probability that the a trap has at most five creatures is
$$P(X \leq 5) = \sum_{x=0}^{5} \frac{e^{-4.5}(4.5)^x}{x!} = .7029$$

Poisson Distribution as a Limit

- 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 $b(x; n; p) \to p(x; \mu)$.
- So in any binomial experiment in which n is large and p is small, , then Binomial can be approximated by Poisson Distribution with parameter $\mu = np$.

Example

- A typesetter, on the average makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?

Example

- A typesetter, on the average makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?
- X is the number of errors in five pages

$$X \sim Bin(1500, 1/500)$$

Example

- A typesetter, on the average makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?
- X is the number of errors in five pages

$$X \sim Bin(1500, 1/500)$$

- Exact solution

$$P(X \leq 2) = \sum_{x=0}^{2} \binom{1500}{x} \left(\frac{1}{500}\right)^{x} \left(\frac{499}{500}\right)^{1500-x} = .4230$$

Example

- A typesetter, on the average makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?
- X is the number of errors in five pages

$$X \sim Bin(1500, 1/500)$$

- Exact solution

$$P(X \leq 2) = \sum_{x=0}^{2} \binom{1500}{x} (\frac{1}{500})^x (\frac{499}{500})^{1500-x} = .4230$$

- With Poisson Approximation $\mu = np = 3$

$$P(X \leq 2) \approx e^{-3} + 3e^{-3} + \frac{3^2 e^{-3}}{2} = .4232$$

Mean and Variance of Poisson Distribution

- If X has a Poisson Distribution with parameter μ, then $E(X) = Var(X) = \mu$.
- It can be derived directly from the pmf, or through the Binomial limit argument.
- If X is $b(x; n; p)$, then

$$E(X) = np \to \mu, Var(X) = np(1 - p) \to \mu$$