Introduction to mathematical Statistics Lecture6

Binomial Distribution (Review)

Def. Binomial Experiment:

1) It consists of n trials
2) Each trial results in 1 of 2 possible outcomes, "S" or "F"
3) The probability of getting a certain outcome, say "S", remains the same, from trial to trial, say P("S")=p
4) These trials are independent, that is the outcomes from the previous trials will not affect the outcomes of the up-coming trials

Def. Binomial Distribution:

Let X denotes the total # of "S" among the n trials then X~B(n, p)

Its probability density function (pdf), also referred to as the probability mass function (p.m.f.) since X is a discrete random variable is

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0,1,\cdots,n$$

In addition, we have:

$E(X) = np$

$Var(X) = np(1-p)$

Quiz 2, question 1. An exam consists of 10 multiple choice questions. Each question has 4 possible choices. Only 1 is correct. Jeff did not study for the exam. So he just guesses at the right answer for each question (pure guess, not an educated guess). What is his chance of passing the exam? That is, to make at least 6 correct answers.

Answer: Yes, this is a binomial experiment with n=10, p=0.25, "S"=choose the right answer for each question.

Let X be the total # of "S"

P(pass)=P(X$\geq$6)=P(X=6 or X=7 or X=8 or X=9 or X=10)

= P(X=6)+ P(X=7)+ P(X=8)+ P(X=9)+ P(X=10)

$$=\sum_{k=6}^{10}\binom{10}{k}0.25^k 0.75^{10-k} =$$

Quiz 2, question 2. Let X~B(n, p), please derive the m.g.f. of X. Please show the entire derivation for full credit.

m.g.f. of X

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tx} P(X = x)$$

$$= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \sum_{x=0}^{n} \binom{n}{x} (e^t p)^x (1-p)^{n-x}$$

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

Theorem. $(a+b)^n = \sum_{x=0}^{n}\binom{n}{x}a^x b^{n-x}$

Quiz 2, question 3. Let $X_1 \sim B(n_1, p)$ and $X_2 \sim B(n_2, p)$. Furthermore, X_1 and X_2 are independent. Please derive the distribution of $Y = X_1 + X_2$

Solution. $M_Y(t) = E(e^{tY}) = E\left(e^{t(X_1+X_2)}\right) = E(e^{tX_1+tX_2}) = E(e^{tX_1} \cdot e^{tX_2}) = E(e^{tX_1}) \cdot E(e^{tX_2})$

$$= M_{X_1} \cdot M_{X_2} = [pe^t + (1-p)]^{n_1} \cdot [pe^t + (1-p)]^{n_2}$$

$$= [pe^t + (1-p)]^{n_1+n_2}$$

$$\therefore Y \sim B(n_1 + n_2, p)$$

Looking forward: Data Analysis for Categorical Data

- Simplest form : binary data

$$X = \begin{cases} 0 \\ 1 \end{cases}, \quad X = \begin{cases} 'S' \\ 'F' \end{cases}, \quad \cdots$$

Exercise. Let us consider a study of all families with 3 children. (Without loss of generality, consider only single birth.) What is the probability that a family randomly selected from this study has exactly 1 boy?

Solution.

[a] $n = 3$, $'S' = \text{boy}$

$$P('S') = 1/2$$

$$X = \text{total number of boys}$$

$$X \sim B(n = 3, p = 1/2)$$

$$P(X = 1) = \binom{3}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^2 = \frac{3}{8}$$

[b] $P(\text{exactly one boy}) = P\big((B \cap G \cap G) \cup (G \cap B \cap B) \cup (G \cap G \cap B)\big)$

$$= P(B \cap G \cap G) + P(G \cap B \cap B) + P(G \cap G \cap B)$$

$$= P(B)P(G)P(G) + P(G)P(B)P(B) + P(G)P(G)P(B)$$

$$= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

Bernoulli Distribution

- $X \sim \text{Bernoulli}(p)$
- When n=1,

$$P(X = 'S') = p \; ; \; P(X = 'F') = 1 - p$$

$$X = \text{total number of <S> in 1 trial} = \begin{cases} 0, 1-p \\ 1, p \end{cases}$$

$$P(X = x) = p^x (1-p)^{1-x}, x = 0,1$$

- Let $X_i \sim Bernoulli(p), i = 1, \cdots, n$. Furthermore, $X_i's$ are all independent.

  Let $X = \sum_{i=1}^{n} X_i$. Then, $X \sim B(n, p)$

Example. We are concerned. – Many Stony Brook University students appear to be smokers. A random sample of n=200 students was selected to estimate the percentage of smokers. Suppose the total number of smokers were found to be X among this sample.

Solution. $\hat{p} = \frac{X}{n}, X \sim B(n, p)$

$\hat{p}$: the point estimator of p

Confidence Interval of p

How to construct the confidence interval?

Method 1. (exact) Use $X \sim B(n, p)$

Method 2. (approximate - asymptotic statistics) Use normal distribution. (***Central Limit Theorem***)

Central Limit Theorem

Let $X_i, i = 1, \cdots, n$, be a random sample from a population with mean μ and variance σ^2. Under some regularity conditions (*that hold for all problems studied in AMS312), we have

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \to N(0,1) \text{ when n goes to infinity}$$

or

When n is "large enough",

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)$$

Example.

$$\text{Let } X_i = \begin{cases} 0, \text{otherwise} \\ 1, \text{if the ith student in the sample is a smoker} \end{cases}$$

$$\text{Total number of smokers}, X = \sum_{i=1}^{n} X_i$$

$$\hat{p} = \frac{X}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

<u>Large Sample Distribution of the Sample Proportion:</u>

$X_i \sim \text{Bernoulli}(p)$

$\mu = 1 \cdot p = p$

$\sigma^2 = 1 \cdot p \cdot (1-p) = p(1-p)$

By CLT,

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1) \text{ when n is "large"}$$

Order Statistics

Let $X_1, X_2, \cdots, X_n$ be a random sample from a population with p.d.f. $f(x)$. Then,

$$X_{(1)} = \min(X_1, X_2, \cdots, X_n)$$
$$X_{(n)} = \max(X_1, X_2, \cdots, X_n)$$

and $X_{(1)} \le X_{(2)} \le \cdots \le X_{(k)} \le \cdots \le X_{(n)}$

p.d.f.'s for $X_{(1)}$ and $X_{(n)}$

W.L.O.G. (Without Loss of Generality), let's assume X is continuous.

$$P(X_{(1)} \succ x) = P(X_1 \succ x, X_2 \succ x, \cdots, X_n \succ x)$$

$$= \prod_{i=1}^{n} P(X_i \succ x) = \prod_{i=1}^{n} [1 - F_{X_i}(x)]$$

$$F_{X_{(1)}}(x) = 1 - \prod_{i=1}^{n} [1 - F_{X_i}(x)]$$

$$f_{X_{(1)}}(x) = -\frac{d}{dx} \prod_{i=1}^{n} [1 - F_{X_i}(x)] = -\frac{d}{dx} \prod_{i=1}^{n} [1 - F(x)]^n = n \cdot [1 - F(x)]^{n-1} \cdot f(x)$$

$$F_{X_{(n)}}(x) = P(X_{(n)} \le x) = P(X_1 \le x, X_2 \le x, \cdots, X_n \le x)$$

$$= \prod_{i=1}^{n} F_{X_i}(x) = [F(x)]^n$$

$$f_{X_{(n)}}(x) = n \cdot [F(x)]^{n-1} \cdot f(x)$$

Example. Let $X_i \overset{i.i.d.}{\sim} \exp(\lambda).\ i = 1, \cdots, n$

Please 1. Derive the p.d.f. of $X_{(1)}$

2. Derive the p.d.f. of $X_{(n)}$

Solutions.

1. $X_{(1)} = \min(X_1, X_2, \cdots, X_n)$

$$P(X_{(1)} \succ x) = \prod_{i=1}^{n} P(X_i \succ x)$$

$$= \prod_{i=1}^{n} (1 - F(x))$$

$$= \left(1 - F(x)\right)^n$$

$$F_{X_{(1)}}(x) = 1 - \left(1 - F(x)\right)^n$$

$$f_{X_{(1)}}(x) = n\left(1 - F(x)\right)^{n-1} f(x)$$

$f(x) = \lambda e^{-\lambda x},\ x \succ 0$ (exponential distribution)

$$F(x) = \int_0^x f(u)du = \int_0^x \lambda e^{-\lambda u} du$$

$$= \left[-e^{-\lambda u}\right]_0^x = 1 - e^{-\lambda x}$$

$$f_{X_{(1)}}(x) = n \cdot \lambda e^{-\lambda x} \cdot \left[1 - (1 - e^{-\lambda x})\right]^{n-1}$$

$$= n \cdot \lambda e^{-\lambda x} \cdot \left(e^{-\lambda x}\right)^{n-1}$$

$$= n\lambda \left(e^{-\lambda x}\right)^n,\ x \succ 0$$

2. $X_{(n)} = \max(X_1, X_2, \cdots, X_n)$

$$F_{X_{(n)}}(x) = P(X_{(n)} \le x) = \prod_{i=1}^{n} P(X_i \le x) = \left[F(x)\right]^n$$

$$f_{X_{(n)}}(x) = n\left[F(x)\right]^{n-1} f(x)$$

$$f_{X_{(n)}}(x) = n\left[1 - e^{-\lambda x}\right]^{n-1} \lambda e^{-\lambda x},\ x \succ 0$$

Homework: Please read the entire Chapter 4.