Probability Theory

Applications for Data Science
Module 3 Discrete Random Variables

Anne Dougherty

February 21, 2021

TABLE OF CONTENTS

Random Variables

At the end of this module, students should be able to

- Define a discrete random variable and give examples of a probability mass function and a cumulative distribution function.

- Calculate probabilities of Bernoulli, **Binomial**, Geometric, and **Negative Binomial** random variables.

- Calculate the expectation and variance of a discrete rv.

Binomial random variable

Key ideas:
- n Bernoulli trials
- same prob of success in each trial
- indep. Bernoulli trials

Examples:

- Suppose you toss a fair coin 12 times. What is the probability that you'll get 5 heads? $p = 1/2$
- Suppose you pick a random sample of 25 circuit boards used in the manufacture of a particular cell phone. You know that the long run percentage of defective boards is 5%. What is the probability that 3 or more boards are defective? $p = .05$
- Suppose 40% of online purchasers of a particular book would like a new copy and 60% want a used copy. What is the probability that amongst 100 random purchasers, 50 or more used books are sold? $p = .60$

These three situations, and many more, can be modeled by a binomial random variable.

Properties of a binomial random variable;

- Experiment is n trials (n is fixed in advance)
- Trials are identical and result in a success or a failure (i.e. Bernoulli trials) with $P(\text{success}) = p$ and $P(\text{failure}) = 1 - p$.
- Trials are independent (outcome of one trial does not influence any other)

If X is the number of successes in the n independent and identical trials, X is a binomial random variable.
Notation: $X \sim Bin(n, p)$

Find the pmf, expectation, and variance for a binomial random variable, $X \sim Bin(n,p)$.

What is the sample space for a binomial experiment?

$S = \left\{ (x_1, x_2, \ldots, x_n) \mid x_i = \begin{cases} 1 & \text{if success on } i^{th} \text{ trial} \\ 0 & \text{if failure} \end{cases} \right\}$ $\quad |S| = 2^n$

but each element does not have equal prob.

$P(X=0) = P(\{00\cdots0\}) = (1-p)^n$

$P(X=1) = P(\{10\cdots0, 010\cdots0, \ldots, 00\cdots01\}) = np(1-p)^{n-1}$

$P(X=2) = P(\{\underbrace{110\cdots0}_{2\text{ 1's } \& \text{ } n-2 \text{ 0's}}, \ldots\}) = \binom{n}{2} p^2 (1-p)^{n-2}$

$P(X=k) = P(\{\underbrace{\_\,\_\,\_\,\_\,\_\,\_\,\_}_{\substack{k \text{ 1's} \\ n-k \text{ 0's}}}\}) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$

Observe: $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$

binomial thm

Definition: The expected value of a discrete random variable, $E(X)$, is given by

$$E(X) = \sum_{k} kP(X = k)$$

$X \sim Bin(n, p)$

$$E(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np$$

Recall: Bern (p), expected value is p.

Definition: The **variance** of a random variable is given by $\sigma_X^2 = V(X) = E[(X - E(X))^2]$.

Computational formula: $V(X) = E(X^2) - (E(X))^2$.

For $X \sim Bin(n,p)$

$$V(X) = \sum_k (k - E(X))^2 P(X=k)$$
$$= \sum_{k=0}^{n} (k-np)^2 \binom{n}{k} p^k (1-p)^{n-k}$$
$$= \underbrace{np(1-p)}_{\text{variance of Bern}(p)}$$

$X_1, X_2, \ldots, X_n$ as our indep. Bern(p) r.v.,

$$\text{Bin}(n,p) \longrightarrow X = \sum_{k=1}^{n} X_k$$

Negative binomial random variable

key ideas
indep Bernoulli trials
until r successes
count # of failures until r successes.

Examples:

- Suppose you toss a fair coin until you obtain 5 heads. How many tails before the fifth head?
- Suppose you randomly choose circuit boards until you find 3 defectives. You know that the long run percentage of defective boards is 5%. How many must you examine?
- Suppose 40% of online purchasers of a particular book would like a new copy and 60% want a used copy. How many new books are sold before the fiftieth used book?

These three situations can be modeled by a **negative** binomial random variable.

Definition: Repeat independent Bernoulli trials until a total of r successes is obtained. The negative binomial random variable Y counts the number of failures before the r^{th} success. Notation: $Y \sim NB(r, p)$.

- The number of successes r is fixed in advance.
- Trials are identical and result in a success or a failure (i.e. Bernoulli trials) with $P(\text{success}) = p$ and $P(\text{failure}) = 1 - p$.
- Trials are independent (outcome of one trial does not influence any other)

Compare to $X \sim Bin(n, p)$: X is the number of successes in the n independent and identical trials and n is fixed in advance.

Example: A physician wishes to recruit 5 people to participate in a medical study. Let $p = .2$ be the probability that a randomly selected person agrees to participate. What is the probability that 15 people must be asked before 5 are found who agree to participate.

Y is the number of failures before the 5 people are found.

$$S = \left\{ (x_1, x_2, x_3, \dots) \;\middle|\; x_i = \begin{cases} 1 & \text{if success on } i^{th} \text{ trial} \\ 0 & \text{if failure} \end{cases} \text{ and } \sum x_i = 5 \right\}$$

$$P(Y=0) = P(\{11111\}) = (.2)^5$$

$$P(Y=1) = P(\{011111, 101111, 110111, 111011, 111101\}) = \binom{5}{4}(.2)^5(.8)$$

$$P(Y=2) = \binom{6}{4}(.2)^5(.8)^2$$

$$\underbrace{\_\ \_\ \_\ \_\ \_\ \_}_{6 \text{ spots},\ 2\ 0\text{'s},\ 4\ 1\text{'s}} \;\Big|\; \underline{1}$$

$$P(Y=k) = \binom{k+5-1}{4}(.2)^5(.8)^k$$

$$\_\ \_\ \_\ \_\ \_\ \_ \cdots \_ \;\Big|\; 1$$

$k+5-1$ 4 1's, k 0's

$$P(Y=15) = \binom{19}{4}(.2)^5(.8)^{15}$$

$Y \sim NB(r,p)$

pmf $P(Y=k) = \binom{k+r-1}{r-1} p^r (1-p)^k, \quad k = 0, 1, 2, \ldots$

<u>Show:</u> $\sum_{k=0}^{\infty} P(Y=k) = 1$

$$E(Y) = \frac{r(1-p)}{p}$$

$$V(Y) = \frac{r(1-p)}{p^2}$$

Relationship between geometric r.v. & $NB(r,p)$

$X \sim Geom(p) \leftarrow$ repeat indep, identical Bernoulli trials until the first success

$Y \sim NB(1,p) \leftarrow$ counts the number of failures before 1^{st} success

<u>Note:</u> $Y \sim X-1$

$$E(Y) = E(X-1) = E(X) - 1 = \frac{1}{p} - 1 = \frac{1-p}{p}$$

<u>NB</u>(r,p) $\underline{\qquad\quad}|\underline{\quad}|\underline{\qquad}|$

↑ r^{th} success