Discrete Probability Distributions

Intro to discrete ~~variables~~ Random variables

Flip a coin three times

$\Omega = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$

$A_0 = \{\text{no heads}\} = \{TTT\}$ ← $\{X = 0\}$

$A_1 = \{\text{one head}\} = \{HTT, THT, TTH\}$ ← $\{X = 1\}$

$A_2 = \{\text{two heads}\} = \{HHT, HTH, THH\}$ ← $\{X = 2\}$

$A_3 = \{\text{three heads}\} = \{HHH\}$ ← $\{X = 3\}$

⇐ X represents # of heads

Random variables.

A numerically valued map X of an outcome ω from a sample space Ω to the real line $\mathbb{R}$

$$X : \Omega \to \mathbb{R} : \omega \to X(\omega)$$

is called a **random variable (r.v.)**, and usually determined by an experiment. We conventionally denote random variables by uppercase letters X, Y, Z, U, V, etc., from the end of the alphabet. In particular, a **discrete random variable** is a random variable that can take values on a finite set $\{a_1, a_2, \ldots, a_n\}$ of real numbers (usually integers), or on a countably infinite set $\{a_1, a_2, a_3, \ldots\}$. The statement such as "$X = a_i$" is an event since

$$\{\omega : X(\omega) = a_i\}$$

$\{X=1\} = \{\omega \in \Omega : X(\omega) = 1\} = \{HTT, THT, TTH\}$

is a subset of a sample space Ω.

Frequency function

We can consider the probability of the event $\{X = a_i\}$, denoted by $P(X = a_i)$. The function

$$p(a_i) := P(X = a_i)$$

over the possible values of X, say $a_1, a_2, \ldots$, is called a **frequency function**, or a **probability mass function**. The frequency function p must satisfy

$$\sum_i p(a_i) = 1,$$

where the sum is over the possible values of X. The frequency function will completely describe the probabilistic nature of random variable.

Joint distributions of discrete random variables.

When two discrete random variables X and Y are obtained in the same experiment, we can define their **joint frequency function** by

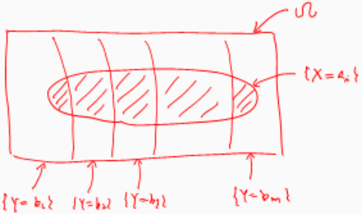
$$p(a_i, b_j) = P(X = a_i, Y = b_j) = P(\{X=a_i\} \cap \{Y=b_j\})$$

where $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ are the possible values of X and Y, respectively. The **marginal frequency function** of X, denoted by p_X, can be calculated by

$$p_X(a_i) = P(X = a_i) = \sum_j p(a_i, b_j),$$

$= P(\{X=a_i\})$

where the sum is over the possible values $b_1, b_2, \ldots$ of Y. Similarly, the marginal frequency function $p_Y(b_j) = \sum_i p(a_i, b_j)$ of Y is given by summing over the possible values $a_1, a_2, \ldots$ of X.



$$\underbrace{P(\{X=a_i\})}_{P_X(a_i)} = \underbrace{P(\{X=a_i\}\cap\{Y=b_1\})}_{P(a_i,\, b_1)} + \underbrace{P(\{X=a_i\}\cap\{Y=b_2\})}_{P(a_i,\, b_2)} + \cdots + \underbrace{P(\{X=a_i\}\cap\{Y=b_m\})}_{P(a_i,\, b_m)}$$

$$P_X(a_i) = \sum_j P(a_i, b_j)$$

Example

An experiment consists of throwing a fair coin three times. Let X be the number of heads, and let Y be the number of heads before the first tail.

1. List the sample space Ω. $= \{HHH, \dots, TTT\}$ ← 8 outcomes
2. Describe the events $\{X = 0\}$, $\{Y = 0\}$, and $\{X = 0, Y = 0\}$.
3. Find the frequency function p for X and Y. And compute the joint frequency $p(0, 0)$.

2. $\{X=0\} = \{TTT\}$, $\{Y=0\} = \{TTT, THH, TTH, THT\}$, $\{X=0, Y=0\} = \{TTT\}$

3. $p_X(a) = P(\{X=a\})$: $p_X(0) = P(\{X=0\}) = P(\{TTT\}) = \frac{1}{8}$

$p_Y(b) = P(\{Y=b\})$: $p_Y(0) = P(\{Y=0\}) = P(\{TTT, THH, TTH, THT\}) = \frac{4}{8} = \frac{1}{2}$

$p(a,b) = P(X=a, Y=b) = P(\{X=a\} \cap \{Y=b\})$: $p(0,0) = P(\{X=0\} \cap \{Y=0\}) = P(\{TTT\}) = \frac{1}{8}$

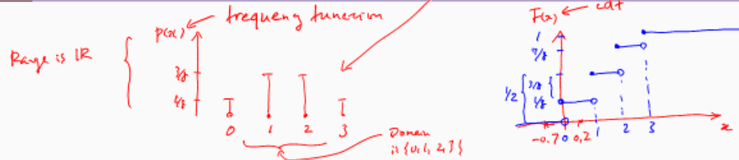
Example

An experiment consists of throwing a fair coin three times. Let X be the number of heads, and let Y be the number of heads before the first tail.

1. List the sample space Ω.
2. Describe the events $\{X = 0\}$, $\{Y = 0\}$, and $\{X = 0, Y = 0\}$.
3. Find the frequency function p for X and Y. And compute the joint frequency $p(0, 0)$.

1. $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$
2. $\{X = 0\} = \{TTT\}$, $\{Y = 0\} = \{TTT, THH, THT, TTH\}$, and $\{X = 0, Y = 0\} = \{TTT\}$,
3. $p_X(0) = \frac{1}{8}$; $p_X(1) = \frac{3}{8}$; $p_X(2) = \frac{3}{8}$; $p_X(3) = \frac{1}{8}$.
   $p_y(0) = \frac{1}{2}$; $p_y(1) = \frac{1}{4}$; $p_y(2) = \frac{1}{8}$; $p_y(3) = \frac{1}{8}$.
   $p(0, 0) = P(X = 0, Y = 0) = \frac{1}{8}$.

Cumulative distribution function.



Another useful function is the **cumulative distribution function (cdf)**, and it is defined by

$$F(x) = P(X \leq x), \qquad -\infty < x < \infty.$$

The cdf of a discrete r.v. is a nondecreasing step function. It jumps wherever $p(x) > 0$, and the jump at a_i is $p(a_i)$. cdf's are usually denoted by uppercase letters, while frequency functions are usually denoted by lowercase letters.

Independent random variables.

If $P(A \cap B) = P(A)P(B)$ then A and B are independent

How about X and Y? If the joint frequency function $p(x,y)$ satisfies

$$p(a,b) = P(\{X=a\} \cap \{Y=b\}) = P(\{X=a\})\,P(\{Y=b\}) = p_X(a)\,p_Y(b)$$

for every pair (a,b) then X and Y are independent

Let X and Y be discrete random variables with joint frequency function $p(x, y)$. Then X and Y are said to be **independent**, if they satisfy

$$p(x, y) = p_X(x)p_Y(y)$$

for all possible values of (x, y).

They are marginal frequency functions

$p_X(x) = \sum_y p(x,y)$ and $p_Y(y) = \sum_x p(x,y)$

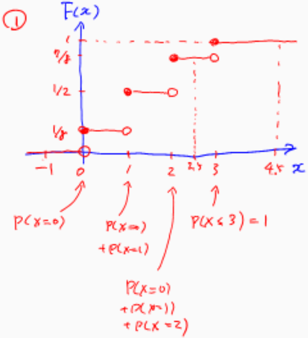
Example

Continue the same experiment of throwing a fair coin three times. Let X be the number of heads, and let Y be the number of heads before the first tail.

$P(X \leq x)$ — Random variable — variable

1. Find the cdf $F(x)$ for X at $x = -1, 0, 1, 2, 2.5, 3, 4.5$.
2. Are X and Y independent?

①



$F(-1) = 0$, $F(0) = \frac{1}{8}$, $F(1) = \frac{1}{2}$, $F(2) = \frac{7}{8}$, $F(2.5) = \frac{7}{8}$

$F(3) = 1$, $F(4.5) = 1$

② $p(0,0) = \frac{1}{8}$, $p_X(0) = \frac{1}{8}$, $p_Y(0) = \frac{1}{2}$

$p(0,0) = \frac{1}{8} \neq \frac{1}{16} = p_X(0)\, p_Y(0)$

They are not independent.

Example

Continue the same experiment of throwing a fair coin three times. Let X be the number of heads, and let Y be the number of heads before the first tail.

1. Find the cdf $F(x)$ for X at $x = -1, 0, 1, 2, 2.5, 3, 4.5$.
2. Are X and Y independent?

1. $F(-1) = 0$, $F(0) = \frac{1}{8}$, $F(1) = \frac{1}{2}$, $F(2) = F(2.5) = \frac{7}{8}$, and $F(3) = F(4.5) = 1$.
2. Since $p_X(0) = \frac{1}{8}$, $p_Y(0) = \frac{1}{2}$, and $p(0,0) = \frac{1}{8}$, we find that $p(0,0) \neq p_X(0)p_Y(0)$. Thus, X and Y are not independent.

Bernoulli Trials and Binomial Distributions

When does equally likely outcome fail?

1. Unfair coin $P(H) = 0.52$ $P(T) = 0.48$

2. Toss a die twice and you win a game if the sum is higher than 10.

$$X = \begin{cases} 1 & \text{if you win} \\ 0 & \text{if you lose} \end{cases}$$

$\{X = 1\} = \{(5,6), (6,6), (6,5)\}$

$P(X=1) = \frac{3}{36}$

$X = 0$ or $X = 1$ are not equally likely.

} Bernoulli trial

Bernoulli trials.

A Bernoulli random variable X takes value only on 0 and 1. It is determined by the **parameter** p (which represents the probability that $X = 1$), and the frequency function is given by

$$p(1) = p$$

$$p(0) = 1 - p$$

If A is the event that an experiment results in a "success," then the **indicator random variable**, denoted by I_A, takes the value 1 if A occurs and the value 0 otherwise.

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A; \\ 0 & \text{otherwise (i.e., } \omega \notin A). \end{cases}$$

Then I_A is a Bernoulli random variable with "success" probability $p = P(A)$. We will call such experiment a **Bernoulli trial**.

Consider n independent Bernoulli trials: $X_1, X_2, \dots, X_n$

X = # of successful Bernoulli trials

Find the frequency function $p(k) = P(X = k)$.

For example, consider $n = 5$ and $k = 2$.

What $\to X = 2$? $\quad X_1, X_2, X_3, X_4, X_5$

Exactly two successes

1	2	3	4	5
1	1	0	0	0
1	0	1	0	0
⋮				
0	0	0	1	1

$P(X_1=1, X_2=1, X_3=0, X_4=0, X_5=0)$

$= P(X_1=1)P(X_2=1)P(X_3=0)P(X_4=0)P(X_5=0)$

$= (p)(p)(1-p)(1-p)(1-p) = p^2(1-p)^3$

How many patterns? $\Updownarrow$ Choose 2 success attempt from 1, 2, 3, 4, 5. $\quad \binom{5}{2}$

$P(X = 2)$

$= p^2(1-p)^3\binom{5}{2} = \binom{5}{2}p^2(1-p)^3 = p(2)$

$\Downarrow$ In general for n and k

$$P(X = k) = p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$X_1, \ldots, X_n$

If we have n independent Bernoulli trials, each with a success probability p, then the probability that there will be exactly k successes is given by

$$P(X = k) = p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

The above frequency function $p(k)$ is called a **binomial distribution** with parameter (n, p).

X = # of successes is called binomial random variable with (n, p)

success probability

# of Bernoulli trials

Example

Five fair coins are flipped independently. Find the frequency function of the number of heads obtained.

$X = \#$ of heads.

$n=5$, $p=\frac{1}{2}$

$$p(k) = P(X=k) = \binom{5}{k} p^k (1-p)^{5-k} = \binom{5}{k} \left(\frac{1}{2}\right)^5$$

Example

Five fair coins are flipped independently. Find the frequency function of the number of heads obtained.

The number X of heads represents a binomial random variable with parameter $n = 5$ and $p = \frac{1}{2}$. Thus, we obtain

$$p(k) = \binom{5}{k} \left(\frac{1}{2}\right)^5$$

Example

A company has known that their screws is defective with probability 0.01. They sell the screws in packages of 10 and are planning a money-back guarantee

p   n

1. at most one of the 10 screws is defective, and they replace it if a customer find more than one defective screws, or
2. they replace it even if there is only one defective.

For each of the money-back guarantee plans above what proportion of packages sold must be replaced?

Consider a Bernoulli trial : $X_i = \begin{cases} 1 & \text{if the i-th screw is defective} \\ 0 & \text{otherwise} \end{cases}$

$X_1, \cdots, X_{10}$

⇓

X = # of defective screws.

$P(k) = P(X=k) = \binom{10}{k}(0.01)^k(0.99)^{10-k}$

1. $P(X \geq 2) = 1 - P(X \leq 1) = 1 - p(0) - p(1)$

2. $P(X \geq 1) = 1 - P(X=0) = 1 - p(0)$

Example

A company has known that their screws is defective with probability 0.01. They sell the screws in packages of 10, and are planning a money-back guarantee

1. at most one of the 10 screws is defective, and they replace it if a customer find more than one defective screws, or
2. they replace it even if there is only one defective.

For each of the money-back guarantee plans above what proportion of packages sold must be replaced?

The number X of defective screws in a package represents a binomial random variable with parameter $p = 0.01$ and $n = 10$.

1. $P(X \geq 2) = 1 - p(0) - p(1) =$ $1 - \binom{10}{0}(0.01)^0(0.99)^{10} - \binom{10}{1}(0.01)^1(0.99)^9 \approx 0.004$
2. $P(X \geq 1) = 1 - p(0) = 1 - \binom{10}{0}(0.01)^0(0.99)^{10} \approx 0.096$

Relation between Bernoulli trials and binomial random variable.

$$X_i = \begin{cases} 1 & \text{if success} \\ 0 & \text{if failure} \end{cases}$$

Y = # of successes $= X_1 + X_2 + \cdots + X_n$

A binomial random variable can be expressed in terms of n Bernoulli random variables. If $X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables with success probability p, then the sum of those random variables

$$Y = \sum_{i=1}^{n} X_i$$

is **distributed as** the binomial distribution with parameter (n, p).

Sum of independent binomial random variables.

Theorem

If X and Y are independent binomial random variables with respective parameters (n, p) and (m, p), then the sum $X + Y$ is distributed as the binomial distribution with parameter $(n + m, p)$.

X = Binomial random variable, Y = Binomial random variable

$X = Z_1 + Z_2 + \cdots + Z_n$ (sum of Bernoulli trials)

$Y = Z_{n+1} + Z_{n+2} + \cdots + Z_{n+m}$ (m Bernoulli trials)

$$X + Y = Z_1 + Z_2 + \cdots + Z_n + Z_{n+1} + \cdots + Z_{n+m}$$

$(n+m)$ Bernoulli trials

$$P(X+Y=k) = p(k) = \binom{n+m}{k} p^k (1-p)^{n+m-k}$$

Sum of independent binomial random variables.

Theorem

If X and Y are independent binomial random variables with respective parameters (n, p) and (m, p), then the sum $X + Y$ is distributed as the binomial distribution with parameter $(n + m, p)$.

Observe that we can express $X = \sum_{i=1}^{n} Z_i$ and $Y = \sum_{i=n+1}^{n+m} Z_i$ in terms of independent Bernoulli random variables Z_i's with success probability p. Then the resulting sum $X + Y = \sum_{i=1}^{n+m} Z_i$ must be a binomial random variable with parameter $(n + m, p)$.

Expectations

Expectation.

Let X be a discrete random variable whose possible values are $a_1, a_2, \ldots$, and let $p(x)$ is the frequency function of X. Then the **expectation** (**expected value** or **mean**) of the random variable X is given by

$$E[X] = \sum_i a_i p(a_i).$$

We often denote the expected value $E(X)$ of X by μ or μ_X. For a function g, we can define the expectation of function of random variable by

$$E[g(X)] = \sum_i g(a_i) p(a_i).$$

Variance.

The **variance** of a random variable X, denoted by $\mathrm{Var}(X)$ or σ^2, is the expected value of "the squared difference between the random variable and its expected value $E(X)$," and can be defined as

$$\mathrm{Var}(X) := E[(X - E(X))^2] = E[X^2] - (E[X])^2.$$

The square-root $\sqrt{\mathrm{Var}(X)}$ of the variance $\mathrm{Var}(X)$ is called the **standard error (SE)** (or **standard deviation (SD)**) of the random variable X.

Example

A random variable X takes on values 0, 1, 2 with the respective probabilities $P(X = 0) = 0.2$, $P(X = 1) = 0.5$, $P(X = 2) = 0.3$. Compute

1. $E[X]$
2. $E[X^2]$
3. $\mathrm{Var}(X)$ and SD of X

Example

A random variable X takes on values 0, 1, 2 with the respective probabilities $P(X = 0) = 0.2$, $P(X = 1) = 0.5$, $P(X = 2) = 0.3$. Compute

1. $E[X]$
2. $E[X^2]$
3. $\mathrm{Var}(X)$ and SD of X

1. $E[X] = (0)(0.2) + (1)(0.5) + (2)(0.3) = 1.1$
2. $E[X^2] = (0)^2(0.2) + (1)^2(0.5) + (2)^2(0.3) = 1.7$
3. $\mathrm{Var}(X) = E[(X - 1.1)^2] =$ $(-1.1)^2(0.2) + (-0.1)^2(0.5) + (0.9)^2(0.3) = 0.49$. Also, using $\mathrm{Var}(X) = E[X^2] - (E[X])^2$ we can calculate $\mathrm{Var}(X) = (1.7) - (1.1)^2 = 0.49$.
   Then we obtain the SD of $\sqrt{0.49} = 0.7$.

Expectation for two variables.

Suppose that we have two random variables X and Y, and that $p(x, y)$ is their joint frequency function. Then the expectation of function $g(X, Y)$ of the two random variables X and Y is defined by

$$E[g(X, Y)] = \sum_{i,j} g(a_i, b_j) p(a_i, b_j),$$

where the sum is over all the possible values of (X, Y).

Properties of expectation.

One can think of the expectation $E(X)$ as "an operation on a random variable X" which returns the average value for X.

1. Let a be a constant, and let X be a random variable having the frequency function $p(x)$. Then we can show that

$$E[a + X] = \sum_x (a + x)p(x) = a + \sum_x x\, p(x) = a + E[X].$$

2. Let a and b be scalars, and let X and Y be random variables having the joint frequency function $p(x, y)$ and the respective marginal density functions $p_X(x)$ and $p_Y(y)$.

$$\begin{aligned} E[aX + bY] &= \sum_{x,y} (ax + by)p(x, y) \\ &= a \sum_x x\, p_X(x) + b \sum_y y\, p_Y(y) = aE[X] + bE[Y]. \end{aligned}$$

Linearity property of expectation.

Let a and $b_1, \ldots, b_n$ be scalars, and let $X_1, \ldots, X_n$ be random variables. By applying the properties of expectation repeatedly, we can obtain

$$E\left[a + \sum_{i=1}^{n} b_i X_i\right] = a + E\left[\sum_{i=1}^{n} b_i X_i\right] = a + b_1 E[X_1] + E\left[\sum_{i=2}^{n} b_i X_i\right]$$
$$= \cdots = a + \sum_{i=1}^{n} b_i E[X_i].$$

It is also useful to observe the above property as that of "linear operator."

Expectations of Bernoulli and binomial random variables.

Let X be a Bernoulli random variable with success probability p. Then the expectation of X becomes

$$E[X] = 0 \times (1 - p) + 1 \times p = p.$$

Now let Y be a binomial random variable with parameter (n, p). Recall that Y can be expressed as the sum $\sum_{i=1}^{n} X_i$ of independent Bernoulli random variables $X_1, \ldots, X_n$ with success probability p. Thus, by using property (c) of expectation we obtain

$$E[Y] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = np.$$

Expectations with Independent Random Variables

Expectation for two independent random variables.

Suppose that X and Y are independent random variables. Then the joint frequency function $p(x, y)$ of X and Y can be expressed as

$$p(x, y) = p_X(x)p_Y(y).$$

And the expectation of the function of the form $g_1(X) \times g_2(Y)$ is given by

$$\begin{aligned} E[g_1(X)g_2(Y)] &= \sum_{x,y} g_1(x)g_2(y)\, p(x, y) \\ &= \left[\sum_x g_1(x)p_X(x)\right] \times \left[\sum_y g_2(y)p_Y(y)\right] = E[g_1(X)] \times E[g_2(Y)]. \end{aligned}$$

Covariance and correlation.

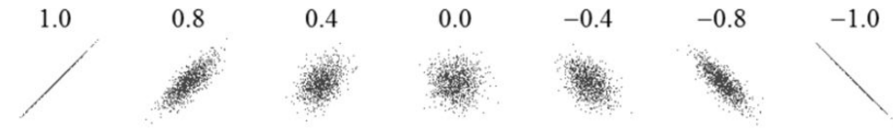
Suppose that we have two random variables X and Y. Then the **covariance** of two random variables X and Y can be defined as

$$\mathrm{Cov}(X,Y) := E((X-\mu_x)(Y-\mu_y)) = E(XY) - E(X)\times E(Y),$$

where $\mu_x = E(X)$ and $\mu_y = E(Y)$. Then the **correlation coefficient**

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

measures the strength of the dependence of the two random variables. The value ρ ranges from -1 to 1, and the relationship with the joint distribution $p(x,y)$ on xy-plane is visualized below.



Properties of variance and covariance.

1. If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$ by observing that $E[XY] = E[X] \cdot E[Y]$.
2. In contrast to the expectation, the variance is *not* a linear operator. For two random variables X and Y, we have

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y). \tag{3.1}$$

Moreover, if X and Y are independent, by observing that $\mathrm{Cov}(X, Y) = 0$ in (3.1), we obtain $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. In general, we have

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n).$$

if $X_1, \ldots, X_n$ are independent random variables.

Example

The joint frequency function $p(x, y)$ of two discrete random variables, X and Y, is given by

		$X = -1$	$X = 0$	$X = 1$
Y	-1	0	1/2	0
	1	1/4	0	1/4

1. Find the marginal frequency function for X and Y.
2. Find $E[X]$ and $E[Y]$.
3. Find $\mathrm{Cov}(X, Y)$.
4. Are X and Y independent?

Solution.

1. $p_X(-1) = \frac{1}{4}$; $p_X(0) = \frac{1}{2}$; $p_X(1) = \frac{1}{4}$.
   $p_Y(-1) = \frac{1}{2}$; $p_Y(1) = \frac{1}{2}$.
2. $E[X] = (-1)\left(\frac{1}{4}\right) + (0)\left(\frac{1}{2}\right) + (1)\left(\frac{1}{4}\right) = 0$
   $E[Y] = (-1)\left(\frac{1}{2}\right) + (1)\left(\frac{1}{2}\right) = 0$
3. $E[XY] = (-1)(1)\left(\frac{1}{4}\right) + (0)(-1)\left(\frac{1}{2}\right) + (1)(1)\left(\frac{1}{4}\right) = 0$
   Thus, we obtain $\mathrm{Cov}(X, Y) = E[XY] - E[X]Y[Y] = 0$.
4. No, X and Y are not independent, because
   $p(-1, -1) = 0 \neq p_X(-1)p_Y(-1) = \frac{1}{8}$.

Variances of Bernoulli and binomial random variables.

Let X be a Bernoulli random variable with success probability p. Then the expectation $E[X]$ is p, and the variance of X is

$$\mathrm{Var}(X) = (0 - p)^2 \times (1 - p) + (1 - p)^2 \times p = p(1 - p).$$

Since a binomial random variable Y is the sum $\sum_{i=1}^{n} X_i$ of independent Bernoulli random variables, we obtain

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1 - p).$$

Hypergeometric Distribution

Example

A committee of 5 students is to be formed from a group of 8 women and 12 men.

1. How many different outcomes can a 5-member committee be formed?
2. How many different outcomes can we form a committee consisting of 2 women and 3 men?
3. What is the probability that a committee consists of 2 women and 3 men?

Example

A committee of 5 students is to be formed from a group of 8 women and 12 men.

1. How many different outcomes can a 5-member committee be formed?
2. How many different outcomes can we form a committee consisting of 2 women and 3 men?
3. What is the probability that a committee consists of 2 women and 3 men?

1. $\binom{20}{5} = \frac{20 \times 19 \times 18 \times 17 \times 16}{5!} = 15504$
2. $\binom{8}{2} \times \binom{12}{3} = 6160$
3. $\frac{6160}{15504} \approx 0.40$

Hypergeometric distribution.

Consider the collection of N subjects, of which m belongs to one particular class (say, "tagged" subjects), and $(N - m)$ to another class (say, "non-tagged"). Now a sample of size r is chosen randomly from the collection of N subjects. Then the number X of "tagged" subjects selected has the frequency function

$$p(k) = \frac{\binom{m}{k}\binom{N-m}{r-k}}{\binom{N}{r}}$$

where $0 \leq k \leq r$ must also satisfy $k \geq r - (N - m)$ and $k \leq m$. The above frequency function $p(k)$ is called a **hypergeometric** distribution with parameter (N, m, r).

Example

A lot, consisting of 50 items, is inspected. Suppose that the lot contains 5 defective items. In the following two inspection procedures, What is the probability that no defective items are found in the inspection?

1. Four items are randomly selected and tested.
2. Ten items are randomly selected and tested.

Example

A lot, consisting of 50 items, is inspected. Suppose that the lot contains 5 defective items. In the following two inspection procedures, What is the probability that no defective items are found in the inspection?

1. Four items are randomly selected and tested.
2. Ten items are randomly selected and tested.

The number X of defective items in the inspection has a hypergeometric distribution with $N = 50$, $m = 5$.

1. Here we choose $r = 4$, and calculate $P(X = 0) = \frac{\binom{5}{0}\binom{45}{4}}{\binom{50}{4}} \approx 0.65$
2. Here we choose $r = 10$, and calculate $P(X = 0) = \frac{\binom{5}{0}\binom{45}{10}}{\binom{50}{10}} \approx 0.31$

Relation between Bernoulli trials and Hypergeometric distribution.

Let A_i be the event that a "tagged" subject is found at the i-th selection.

$$X_i = \begin{cases} 1 & \text{if } A_i \text{ occurs;} \\ 0 & \text{otherwise.} \end{cases}$$

Then X_i is a Bernoulli trial with "success" probability $p = P(A_i) = \frac{m}{N}$.
Then the number Y of "tagged" subjects in a sample of size r can be expressed in terms of r Bernoulli random variables.

$$Y = \sum_{i=1}^{r} X_i$$

is **distributed as** a hypergeometric distribution with parameter (N, m, r).

Example

An urn contains 10 red balls and 20 blue balls, and balls are drawn one at a time without replacement. Let A_i be the event that a red ball is drawn at the i-th attempt.

1. Find $P(A_1)$
2. Calculate $P(A_2)$.
3. Find $P(A_i)$ in general.

Example

An urn contains 10 red balls and 20 blue balls, and balls are drawn one at a time without replacement. Let A_i be the event that a red ball is drawn at the i-th attempt.

1. Find $P(A_1)$
2. Calculate $P(A_2)$.
3. Find $P(A_i)$ in general.

1. $P(A_1) = \frac{10}{30} = \frac{1}{3}$.
2. If we draw 2 balls then we have 10×29 outcomes in which the second ball is red. Thus, $P(A_2) = \frac{10 \times 29}{30 \times 29} = \frac{1}{3}$.
3. In general we have $10 \times 29 \times 28 \times \cdots \times (31 - i)$ outcomes in which i-th ball is red, and obtain

$$P(A_i) = \frac{10 \times 29 \times 28 \times \cdots \times (31 - i)}{30 \times 29 \times 28 \times \cdots \times (31 - i)} = \frac{1}{3}.$$

Expectation of hypergeometric random variable.

Let X_i be a Bernoulli trial of finding a "tagged" subject in the i-th selection. Then the expectation of X becomes

$$E[X_i] = \frac{m}{N}.$$

Now let Y be a hypergeometric random variable with parameter (N, m, r). Recall that Y can be expressed as the sum $\sum_{i=1}^{r} X_i$ of the above Bernoulli trials. We can easily calculate

$$E[Y] = E\left[\sum_{i=1}^{r} X_i\right] = \sum_{i=1}^{r} E[X_i] = \frac{mr}{N}$$

Dependence of Bernoulli trials in Hypergeometric distribution.

Suppose that $i \neq j$. Then we can find that

$$X_i X_j = \begin{cases} 1 & \text{if } A_i \text{ and } A_j \text{ occurs;} \\ 0 & \text{otherwise.} \end{cases}$$

is again a Bernoulli trial with "success" probability $p = P(A_i \cap A_j) = \frac{m(m-1)}{N(N-1)}$. Since $E[X_i] = E[X_j] = \frac{m}{N}$ and $E[X_i X_j] = \frac{m(m-1)}{N(N-1)}$, we can calculate

$$\mathrm{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = \frac{m(m-N)}{N^2(N-1)} < 0$$

Therefore, X_i and X_j are dependent, and negatively correlated.

Variance of hypergeometric random variable.

The variance of Bernoulli trial with success probability $\frac{m}{N}$ is given by

$$\mathrm{Var}(X_i) = \left(\frac{m}{N}\right)\left(1 - \frac{m}{N}\right) = \frac{m(N-m)}{N^2}$$

Together with $\mathrm{Cov}(X_i, X_j) = \frac{m(m-N)}{N^2(N-1)}$, we can calculate

$$\begin{aligned}
\mathrm{Var}(Y) = \mathrm{Var}\left(\sum_{i=1}^{r} X_i\right) &= \sum_{i=1}^{r} \mathrm{Var}(X_i) + 2\sum_{j=2}^{r}\sum_{i=1}^{j-1} \mathrm{Cov}(X_i, X_j) \\
&= r \times \frac{m(N-m)}{N^2} + r(r-1) \times \frac{m(m-N)}{N^2(N-1)} \\
&= \frac{mr(N-m)(N-r)}{N^2(N-1)}
\end{aligned}$$

Assignment No.3

SS: Murray R. Spiegel, John Schiller, and R. Alu Srinivasan, *Probability and Statistics 4th ed.* McGraw-Hill.
Chapter 2: Random Variables, Discrete Probability Distributions, Distribution Functions for Discrete Random Variables, Joint Distributions: Discrte Case, Independent Random Variables.
Chapter 3: Definition of Mathematical Expectation, Functions of Random Variables, Some Theorems on Expectation, Variance and Standard Deviation, Some Theorems on Variance, Covariance, Correlation Coefficient.
Chapter 4: Binomial Distribution, Some Properties of Binomial Distribution, Hypergeometric Distribution.

TH: Elliot A. Tanis and Robert V. Hogg, *A Brief Course in Mathematical Statistics*. Prentice Hall.
Section 2.1–2.6.

WM: Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, and Keying Ye, *Probability & Statistics for Engineers & Scientists*, 9th ed. Prentice Hall.

Problem

Let $p(k)$, $k = -1, 0, 1$, *be the frequency function for random variable* X. *Suppose that* $p(0) = \frac{1}{4}$, *and that* $p(-1)$ *and* $p(1)$ *are unknown.*

1. *Show that* $E[X^2]$ *does not depend on the unknown values* $p(-1)$ *and* $p(1)$.
2. *If* $E[X] = \frac{1}{4}$, *then find the values* $p(-1)$ *and* $p(1)$.

Problem

The joint frequency function $p(x, y)$ of two discrete random variables, X and Y, is given by

		$X=1$	$X=2$	$X=3$
Y	1	c	$3c$	$2c$
	2	c	c	$2c$

1. *Find the constant c.*
2. *Find $E[X]$ and $E[XY]$.*
3. *Are X and Y independent?*

Problem

A pair (X, Y) of discrete random variables has the joint frequency function

$$p(x, y) = \frac{xy}{18}, \; x = 1, 2, 3 \text{ and } y = 1, 2.$$

1. *Find* $P(X + Y = 3)$.
2. *Find the marginal frequency function for* X *and* Y.
3. *Find* $E[Y]$ *and* $\mathrm{Var}(Y)$.
4. *Are* X *and* Y *independent? Justify your answer.*

Problem

1. *Determine the constant c so that $p(x)$ is a frequency function if $p(x) = cx$, $x = 1, 2, 3, 4, 5, 6$.*
2. *Similarly find c if $p(x) = c\left(\frac{2}{3}\right)^x$, $x = 1, 2, 3, 4, \ldots$.*

Problem

A study shows that 40% of college students binge drink. Let X be the number of students who binge drink out of sample size $n = 12$.

1. *Find the mean and standard deviation of X.*
2. *Do you agree that the probability that X is 5 or less is higher than 50%? Justify your answer.*

Problem

Let X and Y be independent random variables.

1. *Show that* $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$.
2. *If* $E[X] = 1, E[Y] = 2$ *and* $\mathrm{Var}(X) = 4, \mathrm{Var}(Y) = 9$ *then find the mean and the variance of* $Z = 3X - 2Y$.

Problem

In a lot of 20 light bulbs, there are 9 bad bulbs. Let X be the number of defective bulbs found in the inspection. Find the frequency function $p(k)$, and identify the range for k in the following inspection procedures.

1. *An inspector inspects 5 bulbs selected at random and without replacement.*
2. *An inspector inspects 15 bulbs selected at random and without replacement.*

Computer project.

A researcher visits a study area and capture $m = 30$ individuals at the beginning, mark them with a yellow tag, and then release them back into the environment. Next time the researcher returns and captures another sample of $r = 20$ individuals. Some of the individuals in this second sample have been marked during the initial visit, known as "recaptures." Let X be the number of recaptures. Suppose that $N = 200$ is the population size, and simulate $n = 1000$ observations.

1. What is the frequency function for X? Can you find $E[X]$ and $\mathrm{Var}(X)$?
2. Draw the relative frequency histogram from the simulation, and the probability histogram from the frequency function for X.
3. Calculate `mean()` and `sd()` from the observations, and create a table to compare them with $E[X]$ and $\sqrt{\mathrm{Var}(X)}$.

Computer project, continued.

Suppose that the population size N is unknown. Then how can we guess N from the values m, r, and the observed number X of recaptures? You will investigate the performance of the following strategies:

I. Estimate it by $N = \dfrac{m \times r}{X + e}$ with small value $e = 0.25$.

II. Estimate it by $N = \dfrac{m \times r}{X}$ only when $X \geq 1$.

III. Estimate it by $N = \dfrac{(m+1) \times (r+1)}{X+1}$

Assuming the population size $N = 200$, continue the simulation of sampling X. But this time use it to calculate the estimation of N for each strategy. Which strategy do you recommend in order to predict the unknown population size N?

Answers

Problem 1.

1. $E[X^2] = p(-1) + p(1) = 1 - p(0) = \frac{3}{4}$
2. $E[X] = -p(-1) + p(1) = \frac{1}{4}$. Together with $p(-1) + p(1) = \frac{3}{4}$, we obtain $p(-1) = \frac{1}{4}$ and $p(1) = \frac{1}{2}$.

Problem 2.

1. $\sum_{x=1}^{3}\sum_{y=1}^{2} p(x,y) = 10c = 1$ implies that $c = \frac{1}{10}$.
2. $E[X] = (1)\left(\frac{1}{5}\right) + (2)\left(\frac{2}{5}\right) + (3)\left(\frac{2}{5}\right) = \frac{11}{5}$
   $E[Y] = (1)\left(\frac{3}{5}\right) + (2)\left(\frac{2}{5}\right) = \frac{7}{5}$
   $E[XY] = (1)(1)\left(\frac{1}{10}\right) + (2)(1)\left(\frac{3}{10}\right) + (3)(1)\left(\frac{2}{10}\right) + (1)(2)\left(\frac{1}{10}\right) + (2)(2)\left(\frac{1}{10}\right) + (3)(2)\left(\frac{2}{10}\right) = \frac{31}{10}$
3. X and Y are not independent because $p(1,1) = \frac{1}{10} \neq p_X(1)p_Y(1) = \frac{3}{25}$. Or, you can find it by calculating

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = \frac{1}{50}$$

Problem 3.

1. $P(X+Y=3) = p(1,2) + p(2,1) = \frac{2}{9}$
2. $p_X(y) = \frac{x}{6}$ for $x = 1, 2, 3.$
   $p_Y(y) = \frac{y}{3}$ for $y = 1, 2.$
3. $E[Y] = (1)\left(\frac{1}{3}\right) + (2)\left(\frac{2}{3}\right) = \frac{5}{3}$; $E[Y^2] = (1)^2\left(\frac{1}{3}\right) + (2)^2\left(\frac{2}{3}\right) = 3$
   $\mathrm{Var}(Y) = E[Y^2] - (E[Y])^2 = \frac{2}{9}$
4. Yes, since the joint frequency function satisfies
   $p(x,y) = p_X(x)p_Y(y)$ for all $x = 1, 2, 3$ and $y = 1, 2.$

Problem 4.

1. $\sum_{x=1}^{6} p(x) = c \times \frac{6(6+1)}{2} = 21c = 1$ Thus, we obtain $c = \frac{1}{21}$
2. $\sum_{x=1}^{\infty} p(x) = c \times \frac{(2/3)}{1-(2/3)} = 2c = 1$ Thus, we obtain $c = \frac{1}{2}$

Problem 5.

1. The mean is $np = (12)(0.4) = 4.8$, and the standard deviation is $\sqrt{(12)(0.4)(0.6)} \approx 1.7$.
2. Yes, because $P(X \leq 5) \approx 0.665$.

Problem 6.

1. $\mathrm{Var}(aX) = E[(aX - E[aX])^2] = E[(aX - aE[X])^2] =$
   $E[a^2(X - E[X])^2] = a^2 E[(X - E[X])^2] = a^2 \mathrm{Var}(X)$.
2. $E[Z] = E[3X - 2Y] = (3)(1) - (2)(2) = -1$
   $\mathrm{Var}(Z) = \mathrm{Var}(3X + (-2)Y) = \mathrm{Var}(3X) + \mathrm{Var}((-2)Y) =$
   $(3)^2(4) + (-2)^2(9) = 72$

Problem 7.

X has a hypergeometric distribution with $N = 20$ and $m = 9$.

1. Here we choose $r = 5$. Thus,

$$p(k) = \frac{\binom{9}{k}\binom{11}{5-k}}{\binom{20}{5}}$$

   where k takes a value in the range of $0 \leq k \leq 5$.

2. Here we choose $r = 15$. Thus,

$$p(k) = \frac{\binom{9}{k}\binom{11}{15-k}}{\binom{20}{15}}$$

   where k takes a value in the range of $4 \leq k \leq 9$.