3 Discrete Random Variables and Probability Distributions

3.1 Random Variables

Definition: A **random variable** is a mapping from the sample space, $\mathcal{S}$, to the real line.

Example: Toss a coin once. Determine the sample space.

Let X be the number of heads in **one** toss of the coin.

The random variable X is defined by

Here, X is a **Bernoulli random variable**, defined below.

□

Definition: A random variable is **Bernoulli** if its only possible values are **0** and **1**.

Example: The possible outcomes for the JMU tennis team are in the sample space $\mathcal{S}$ ={win, lose, draw}.

□

Remark: A **Bernoulli** random variable, X, may be viewed as having sample space $\mathcal{S}$ ={success, failure}, such that $\begin{cases} X(\text{failure}) & = & 0 \\ X(\text{success}) & = & 1 \end{cases}$.

Two types of **random variables**, X, include

(a) **discrete** – Possible values can be listed in a finite or countably infinite sequence.

(b) **continuous** – Possible values contain an interval on the real line, and $P(X = c) = 0$ for all constants c.

3.2 Probability Distributions for Discrete Random Variables

Definition: Let X be a discrete random variable. The **probability mass function (pmf)** is $p(x) = P(X = x)$.

Definition: The **probability distribution** of a **discrete** random variable X consists of the possible values of X along with their associated probabilities.

Example: *Computer crashing.* Let X be the number of times that a particular computer crashes in a month. Suppose that

$$\begin{cases} p(0) &= P(X=0) &= 0.3 \\ p(1) &= P(X=1) &= 0.4 \\ p(2) &= P(X=2) &= 0.2 \\ p(3) &= P(X=3) &= 0.1 \end{cases}$$

(a) Determine the **probability distribution** of X.

x	0	1	2	3	sum
$p(x) = P(X = x)$	0.3	0.4	0.2	0.1	1

(b) Display the **probability distribution** of X in a **histogram**.

(c) Display the **probability distribution** of X in a **line graph**.

(d) Restate this example using the perspective of *marbles in an urn*.

Definition: The **cumulative distribution function**, denoted **cdf**, of a discrete random variable X is

$$F(x) = P(X \leq x) = \sum_{y: y \leq x} p(y), \text{ for all } x \in \Re.$$

The **cdf** often is called the **distribution function**.

Example: *Revisit computer crashing.*

For any (discrete, continuous, etc.) **cumulative distribution function** $F(x)$,

(1) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.

(2) $F(x)$ is monotonic nondecreasing.

(3) $F(x)$ is continuous from the right; i.e.,

$$\lim_{x \downarrow a} F(x) = F(a), \text{ for all } a \in \Re.$$

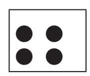
These three properties above also guarantee that $F(x)$ is a cdf.

A cdf $F(x)$ has **left** limits; i.e.,
$\lim_{x \uparrow a} F(x)$ exists for all $a \in \Re$.

Determine $P(X = x)$ in terms of any cdf $F(x)$.

Recall the Bernoulli distribution:

$$p(x) = \begin{cases} 1-p, & \text{if } x = 0 \\ p, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases} \qquad \text{for } 0 \leq p \leq 1.$$

Example: Let X be the number of times  is obtained when rolling an unbiased die once.

(a) Determine the **probability mass function** of X.

(b) Display the **probability mass function** of X in a **histogram** and a **line graph**.

(c) Determine the **cumulative distribution function** of X.

(d) Graph the **cumulative distribution function** of X.

3.3 Expected Values of Discrete Distributions

Determine the **expected value** or **mean** of a **discrete** random variable.

The Expected Value of X

Example: *Revisit computer crashing.* Suppose we have the population $\{0, 0, 0, 1, 1, 1, 1, 2, 2, 3\}$. Determine the mean, μ, of this population.

Now, let X be a random variable with the following **probability mass function**:

$$\begin{cases} p(0) = P(X = 0) = 0.3 \\ p(1) = P(X = 1) = 0.4 \\ p(2) = P(X = 2) = 0.2 \\ p(3) = P(X = 3) = 0.1 \end{cases}$$

Definition: The **expected value** of X, the **mean value** of X, the **population mean** of X, is

$$EX = \mu_x = \sum_x x \; p(x).$$

The Expected Value of a Function

Example: *Revisit computer crashing.* Let $h(x)$ be the cost of maintaining a computer per month as a function of the number of crashes.

Suppose that the systems administrator charges \$10 to restore a crashed computer, and charges \$7 as a monthly base fee.

Determine the mean monthly maintenance cost.

In general,

$$Eh(X) = \sum_x h(x)\ p(x).$$

Write $h(x)$ as a formula.

In general, if a and b are constants, then

$$E(aX + b) = aEX + b.$$

□

The **variance** of a **discrete** random variable X is

$$\sigma^2 = \sigma_x^2 = E(X - \mu_x)^2 = \sum_x (x - \mu)^2\ p(x).$$

Variance measures the **spread** in the distribution.

Example: Derive the " shortcut formula for σ^2 " on p. 105;
i.e., prove that $\sigma^2 = EX^2 - \mu^2$ (for $\sigma^2 < \infty$).

□

The **standard deviation** of *any* random variable X is $\sigma = \sqrt{\sigma^2}$.

Example: *Revisit computer crashing.* Compute the **variance** and **standard deviation** of X, where X is the number of computer crashes in a month.

x	0	1	2	3
$p(x)$	0.3	0.4	0.2	0.1

□

Suppose $h(x) = ax + b$, for constants a and b. Compute the **variance** and **standard deviation** of $h(x)$ in terms of σ_x^2.

□

Example: *Revisit computer crashing.* Compute the **variance** and **standard deviation** of $h(X)$, the monthly maintenance cost.

□

Example: Let X be the **old** salary at some company. Let $\mu_x = \$50,000$ and $\sigma_x = \$30,000$ be the mean and standard deviation, respectively, of X. Let Y be the **new** salary at some company.

(a) Suppose the **new** salary is the **old** salary plus a \$1000 bonus.

(b) Suppose the **new** salary is based on doubling the **new** salary.

□

Exercise 3.31, p. 117: *Slightly modified.* Recall a Bernoulli random variable X with distribution:

$$p(x) = \begin{cases} 1-p, & \text{if } x = 0 \\ p, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases} \qquad \text{for } 0 \leq p \leq 1.$$

(a) Compute the **mean** of X.

(b) Compute EX^2.

(c) Compute the **variance** of X.

(d) Compute the **standard deviation** of X.

(e) Compute EX^{79}.

$\square$

3.4 The Binomial Probability Distribution

Let $Y_1, \ldots, Y_n$ be independent Bernoulli(p) random variables; i.e., $p = P(\text{success})$, where n is fixed.

Let $X = \sum_{i=1}^{n} Y_i = (\text{number of successes})$

Then, $X \sim \text{Binomial}(n, p)$.

Hence, whenever n, the number of trials, is fixed in advance, and X is the sum of n independent Bernoulli(p) random variables, then $X \sim \text{Binomial}(n, p)$.

Goal: Develope the formula for the **probability mass function** of X, which is a Binomial(n, p) random variable.

Example: Toss a **biased** coin **5** times, where $P(\text{heads}) = 0.4$.

(a) Determine $P(H_1 H_2 T_3 T_4 T_5)$.

(b) How many ways can we order H_1, H_2, T_3, T_4, and T_5? *Hint:* Think of these as 5 distinct cards.

(c) How many redundancies exist per ordering (including the original ordering)? *Hint:* $\{H_1, H_2, T_3, T_4, T_5\}$ is equivalent to $\{H_2, H_1, T_5, T_3, T_4\}$.

(d) Compute $P(X = 2)$.

□

In general, if $X \sim \text{Binomial}(n, p)$, then

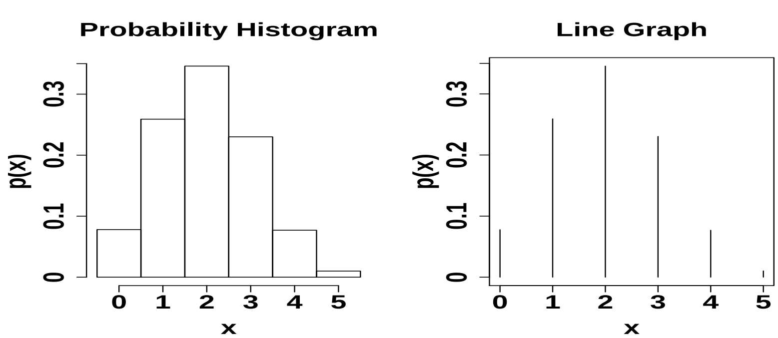
$$P(X = x) = \binom{n}{x} p^x \ (1 - p)^{n-x}, \ \ x = 0, 1, 2, \ldots, n.$$

Example: *Revisit.* Toss a **biased** coin **5** times, where $P(\text{heads}) = 0.4$. Let X be the number of heads.

(a) Determine the **probability mass function** of X.

x	$p(x)$
0	0.0778
1	0.2592
2	0.3456
3	0.2304
4	0.0768
5	0.0102
sum	1

(b) Display the **probability mass function** of X in a **histogram** and a **line graph**.



(c) Determine the **mean** of X.

(d) Determine the **variance** of X.

(e) Determine the **standard deviation** of X.

Exercise 3.58, p. 127: Show that $EX = np$ when $X \sim \text{Binomial}(n, p)$, for all integers $n \geq 1$.

Intuitively, if we toss this biased coin ($p = 0.4$) 100 times, how many heads do we expect *on average*?

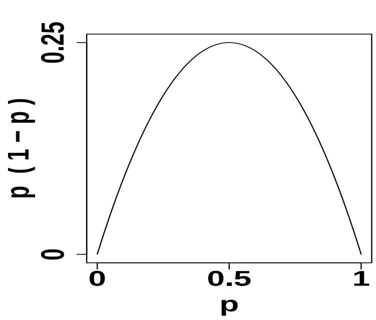
If $X \sim \text{Binomial}(n, p)$, then
$\mu_x = np$, $\sigma_x^2 = np(1-p)$, and $\sigma_x = \sqrt{np(1-p)}$.

Example: *Revisit.* Toss a **biased** coin **5** times, where $P(\text{heads}) = 0.4$. Let X be the number of heads. Compute the **mean**, **variance**, and **standard deviation** of X, using these new formulas.

Exercise 3.56, p. 127: Let $X \sim \text{Binomial}(n, p)$.

(a) For fixed n, are there values of p ($0 \leq p \leq 1$) for which $\text{Var}(X) = 0$?

(b) For fixed n, what value of p maximizes $\text{Var}(X)$?



□

Exercise 3.57, p. 127:

(a) Show that $b(x;\ n,\ 1-p) = b(n-x;\ n,\ p)$.

(b) Skip this part.

(c) What does part (a) imply about the necessity of including values of p greater than 0.5 in Appendix Table A.1?

□

3.5 Hypergeometric and Negative Binomial Distributions

Hypergeometric Distribution – Already discussed somewhat in section 2.3.

Examples include:

(a) Exercise 2.33, p. 74 (buses with cracks)

(b) Exercise 2.38, p. 74 (light bulbs)

Scenario:

(a) N = size of population

(b) M = number of successes in the population

(c) n = sample size

(d) Sample **withOUT** replacement, unlike *binomial*.

(e) Then, $\boldsymbol{X}$, the sample number of successes, is **hypergeometric**.

In sample surveys, the sampling often is done withOUT replacement.

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad (3.15)$$

where x is an integer such that $0 \le x \le M$, $0 \le n - x \le N - M$, and $0 \le x \le n$.

Let $p = M/N$

What is EX?

$$\text{Var}(X) = \left(\frac{N-n}{N-1}\right) np(1-p)$$

Example: Suppose $n = N$. Determine $\text{Var}(X)$.

Example: Suppose $n \ll N$. What is the **correction factor** *approximately*?

(a) The variance of the **hypergeometric** distribution is close to what?

(b) Sampling **withOUT** replacement is close to what?

(c) **Hypergeometric** probabilities are close to what?

□

Example: Suppose that **45** of the **100** United States Senators are Democrats. Let $\boldsymbol{X}$ be the number of Democrats in a simple random sample of size **5**.

(1) Sample **withOUT** replacement.

(a) Name the distribution of X.

(b) Determine the **mean** of X.

(c) Determine the **variance** of X.

(d) Determine the **standard deviation** of X.

(e) Determine $P(X = 2)$.

(2) Sample **WITH** replacement.

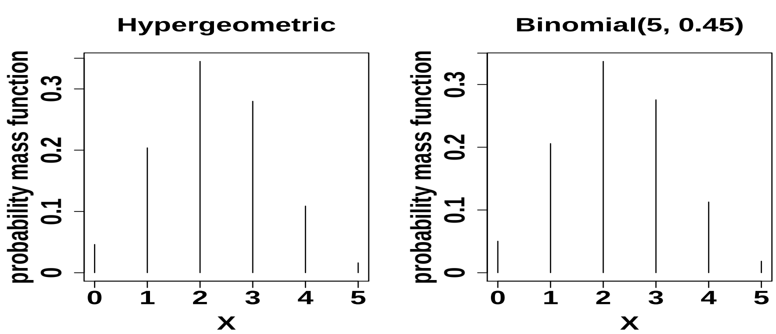
(a) Name the distribution of X.

(b) Determine the **mean** of X.

(c) Determine the **variance** of X.

(d) Determine the **standard deviation** of X.

(e) Determine $P(X = 2)$.



□

Negative Binomial Distribution

Consider the sequence of independent Bernoulli(p) trials.

Suppose we continue sampling until we observe $\boldsymbol{r}$ Bernoulli **successes**.

Let $\boldsymbol{X}$ be the number of **failures** which precede the $\boldsymbol{r}$th success.

Then, $\boldsymbol{X}$ is **negative binomial** with parameters $\boldsymbol{r}$ and $\boldsymbol{p}$.

Example: Toss a coin until you achieve **3** heads, where $p = P(\text{heads}) = 0.4$.

Let X be the number of tails which precede the third heads.

(a) Determine the distribution of X.

(b) Determine $P(X = 5)$.

□

The general formula for the **probability mass function** of a **negative binomial** random variable with parameters r and p is:

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

When $r = 1$, then X is also called a **geometric** random variable with parameter p.

Example: Toss a coin, where $P(\text{heads}) = 0.25 = 1/4$.

(a) On average, how many times do we need to toss the coin to obtain the **first** heads?

(b) Determine the average number of **tails** preceding the **first** heads.

(c) On average, how many times do we need to toss the coin to obtain the **second** heads?

(d) Determine the average number of **tails** preceding the **second** heads.

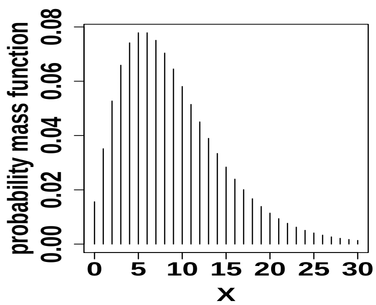
□

If $\boldsymbol{X}$ is **negative binomial** with parameters $\boldsymbol{r}$ and $\boldsymbol{p}$, then

(a) $EX = r\ (1-p)/p$, and

(b) $\text{Var}(X) = r\ (1-p)/p^2$.

Example: Toss a coin, where $P(\text{heads}) = 0.25$, until **3** heads are observed. Let X be the number of tails observed until the **3**rd heads is observed.



(a) Determine the probability that the coin is tossed **9** times.

(b) Determine the **mean** of X.

(c) Determine the expected number of coin tosses.

(d) Determine the **variance** of X.

(e) Determine the **standard deviation** of X.

(f) Determine the **standard deviation** of the number of coin tosses.

□

3.6 The Poisson Probability Distribution

Consider the **Binomial($\boldsymbol{n, p}$)** distribution, such that $\boldsymbol{n}$ is **huge**, $\boldsymbol{p}$ is small, but $\boldsymbol{np}$ is **moderate** (i.e., neither huge nor small).

Example: *Radioactive decay.* Consider a radioactive substance containing 3,000,000 atoms, such that decaying atoms are **independent** of each other, and $p = P(\text{A particular atom decays in the next day}) = 1/1,000,000$.
Compute the **mean** number of atomic decays in the next day.

□

Consider letting $n \to \infty$ and $p \to 0$ such that $np \to \lambda$, a positive constant, where $X \sim \text{Binomial}(n,\ p)$.
In this limit, $X \sim \text{Poisson}(\lambda)$.

Derive the formula for the **probability mass function** of a Poisson(λ) distribution.

□

If $X \sim \text{Poisson}(\lambda)$ for $\lambda > 0$, then

$$P(X = x) = \frac{1}{x!}\, \lambda^x\, e^{-\lambda}, \text{ for } x = 0, 1, 2, \ldots$$

You need NOT memorize this formula.

Prove that the above formula is a valid **probability mass function**.

Exercise 3.84, p. 138: **Prove** that the **mean** of a Poisson(λ) distribution is λ.

□

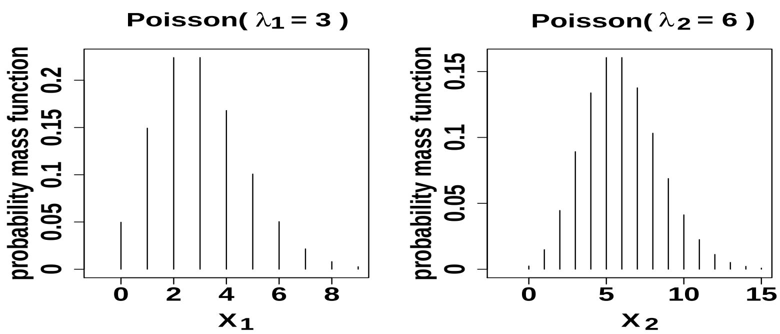
Similarly, it can be shown that the **variance** of a Poisson(λ) distribution is also λ.

Hence, $\mu_x = \sigma_x^2 = \lambda$, if $X \sim \text{Poisson}(\lambda)$.

Example: *Revisit radioactive decay.* Consider a radioactive substance containing 3,000,000 atoms, such decaying atoms are **independent** of each other, and
$p = P(\text{A particular atom decays in the next day}) = 1/1,000,000.$

(a) Let X_1 be the number of decays in **one** day. Determine the probability that **at least one** atom decays in the **next day**.

(b) Let X_2 be the number of decays in **two** days. Determine the probability that **at least one** atom decays in the next **two** days.



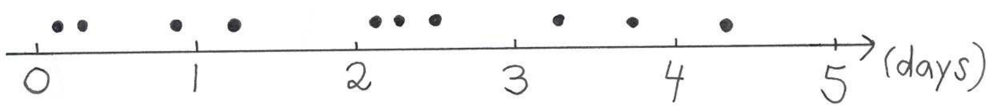
□

Remark: Typically, when a **Binomial(n, p)** distribution is reasonably approximated by a **Poisson(λ)** distribution, n and p are difficult to determine (or estimate), but λ can be estimated from the data. **How?**

The Poisson Process is explained by the following:

(a) The probability of a **success** (such as a radioactive decay) in the next day is **independent** of its past.

(b) The **mean** of a Poisson process based on **two** days is **twice** as large as the **mean** of the same Poisson process based on **one** day.



Example: Consider the number of recombinations (breaks) in DNA (chromosome pairs) when DNA strands are passed to offspring.

□