Chapter 3: Distributions of Random Variables (Part 1)

Day 6 topics:
Section 3.1: Random variables
Section 3.2: Binomial distribution

3.1. **Random variables.**

Definition 3.1. *A **random variable (r.v.)** assigns numerical values to the outcome of a random phenomenon.*

Notation:
A random variable is usually denoted with a capital letter such as X, Y, or Z.

Example 3.2. ***Data points***
Suppose you have a dataset of size 3 ($n = 3$) with $x_1 = 5, x_2 = 3$, and $x_3 = 6$.

- *Each data point is the outcome of a random phenomenon*
- *Each data point is a numerical value*
- *The data points are examples of values of a sequence of random variables X_1, X_2, and X_3*
- *For datasets, we almost always assume the data points came from random variables that are independent and have the same **distribution**.*
- *To calculate the likelihood of data, we need to know the distribution of the random variable that models the data.*
- *First, let's remind ourselves how to calculate them mean and variance of a dataset:*
  - *What is the mean of the data points?*

$$\bar{x} = 14/3 = 4.66667$$

  - *What are the variance and standard deviation of the data points?*

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$
$$s^2 = 2.3333$$
$$s = 1.527525 \sim 1.53$$

Example 3.3. ***Rolling a die***
Suppose you roll a fair die. Let the random variable (r.v.) X be the outcome of the roll, i.e. the value of the face showing on the die.

(1) *What is the probability distribution of the r.v. X?*

(2) *What is the expected outcome of the r.v. X?*

- $$7/2 = 3.5$$
- *Not a possible outcome!*
- *Do not round.*

(3) *Now suppose the 6-sided die is not fair. How would we calculate the expected outcome?*

x	$\mathbb{P}(X = x)$	$x\mathbb{P}(X = x)$
1	0.10	0.1
2	0.20	0.4
3	0.05	0.15
4	0.05	0.2
5	0.25	1.25
6	0.35	2.10
sum	1	$\mu = 4.2$

(From Textbook § 2.1.5)

Definition 3.4. *A* ***probability distribution*** *consists of all disjoint outcomes and their associated probabilities.*

Rules for a probability distribution
A probability distribution is a list of all possible outcomes and their associated probabilities that satisfies three rules:

(1) The outcomes listed must be disjoint.
(2) Each probability must be between 0 and 1.
(3) The probabilities must total to 1.

Probability distributions are usually either **discrete** or **continuous**, depending on whether the random variable is discrete or continuous.

(Back to Textbook § 3.1.1)

Definition 3.5. *A* ***discrete*** *r.v. X takes on a finite number of values or countably infinite number of possible values.*

Definition 3.6. *A* ***continuous*** *r.v. X can take on any real value in an interval of values or unions of intervals.*

Class notes:
Sketch histogram of discrete values 1, 2, ..., 6 with continuous approximation
X could be values from a die or a waiting time in minutes

§ 3.1.2 Expectation

- We call the mean of a random variable its **expected value**
- The expected value is calculated as a weighted average

Definition 3.7. ***Expected value*** *of a discrete random variable*
If X takes on outcomes x_1, ..., x_k with probabilities $P(X = x_1)$, ..., $P(X = x_k)$, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$\begin{aligned}\mu = E(X) &= x_1 P(X = x_1) + \cdots + x_k P(X = x_k) \\ &= \sum_{i=1}^{k} x_i P(X = x_i).\end{aligned}$$

§ 3.1.3 Variability of random variables

- Just like with data, the variability of a r.v. is described with its variance or standard deviation.
- Squared deviations from the mean are weighted by their respective probabilities

Definition 3.8. ***Variance*** *of a discrete random variable*
If X takes on outcomes x_1, ..., x_k with probabilities $P(X = x_1), \ldots, P(X = x_k)$ and expected value $\mu = E(X)$, then the variance of X, denoted by Var(X) or σ^2, is

$$\begin{aligned}Var(X) &= (x_1 - \mu)^2 P(X = x_1) + \cdots + (x_k - \mu)^2 P(X = x_k) \\ &= \sum_{i=1}^{k} (x_i - \mu)^2 P(X = x_i).\end{aligned}$$

Definition 3.9. ***Standard deviation*** *of a discrete random variable*
The standard deviation of X, labeled $SD(X)$ or σ, is

$$SD(X) = \sigma = \sqrt{\sigma^2}$$

Example 3.10. ***Rolling a fair die: variance***
Suppose you roll a fair 6-sided die. Let the random variable (r.v.) X be the outcome of the roll, i.e. the value of the face showing on the die.
Find the variance and standard deviation of X.

$$\mu = 3.5$$

x	$\mathbb{P}(X = x)$	$x - \mu$	$(x - \mu)^2$	$\mathbb{P}(X = x)(x - \mu)^2$ *R*
1	*1/6*	*-2.5*	*6.25*	*1.04166667*
2	*1/6*	*-1.5*	*2.25*	*0.3750*
3	*1/6*	*-0.5*	*0.25*	*0.04166667*
4	*1/6*	*0.5*	*0.25*	*0.04166667*
5	*1/6*	*1.5*	*2.25*	*0.3750*
6	*1/6*	*2.5*	*6.25*	*1.04166667*
sum	*1*			$\sigma^2 = 2.916667$

$$\sigma = 1.707825$$

R commands

x	$\mathbb{P}(X = x)$	$x - \mu$	$(x - \mu)^2$	$\mathbb{P}(X = x)(x - \mu)^2$ *R*
$x < -1 : 6$	$px < -1/6$	$x - mu$	$(x - mu)^2$	$px * (x - mu)^2$
sum	*1*			$sum(px * (x - mu)^2)$

Example 3.11. ***Vaccinated people testing positive for Covid-19***
About 25% of people that test positive for Covid-19 are vaccinated for Covid-19. Define the r.v. X to be 1 if someone that tests positive is vaccinated and 0 if they are not vaccinated.

(1) *Make a table for the probability distribution for the r.v. X*

(2) *What is the expected value of X?*

$$\mathbb{E}[X] = p = 0.25$$

(3) *What is the variance of X?*

$$Var[X] = pq = 0.25 \cdot 0.75 = 0.1875$$
$$SD[X] = 0.4330127$$

§ 3.1.4 Linear combinations of random variables

Seems abstract, but actually fundamental to basic statistical theory - using it all the time!

Definition 3.12. ***Linear combinations*** *of random variables.*
If X and Y are random variables and a and b are constants, then

$$aX + bY$$

is a linear combination of the random variables.

Theorem 3.13. ***Expected value of a linear combination*** *of random variables.*
If X and Y are random variables and a and b are constants, then

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y) = a\mu_X + b\mu_Y$$

Example 3.14. ***Expected money for rolling 3 dice***
Let the random variables X_1, X_2, X_3 be the values shown on 3 fair 6-sided dice rolls.
Suppose you are given in dollars the amount of the first roll, plus twice the value of the second roll, plus 4 times the value of the third roll.
How much money do you expect to get?

$$\mathbb{E}(X_1 + 2X_2 + 4X_3) = \mathbb{E}(X_1) + 2\mathbb{E}(X_2) + 4\mathbb{E}(X_3) = 7 \cdot 3.5 = 24.5$$

Expected value of the sample mean

Theorem 3.15. ***Variance of a linear combination*** *of random variables.*
If X and Y are INDEPENDENT random variables and a and b are constants, then

$$\mathit{Var}(aX + bY) = a^2\,\mathit{Var}(X) + b^2\,\mathit{Var}(Y)$$

Example 3.16. ***Variance of money for rolling 3 dice***
Let the random variables X_1, X_2, X_3 be the values shown on 3 fair 6-sided dice rolls. Suppose you are given in dollars the amount of the first roll, plus twice the value of the second roll, plus 4 times the value of the third roll.
What are the variance and standard deviation of the amount you get from the 3 rolls?

$$\mathit{Var}(X_1 + 2X_2 + 4X_3) = \mathit{Var}(X_1) + 4\,\mathit{Var}(X_2) + 16\,\mathit{Var}(X_3) = 21 \cdot 17.5 = 367.5$$

$$\mathit{SD}(X_1 + 2X_2 + 4X_3) = \sqrt{367.5} = 19.17029$$

Variance of the sample mean

Example 3.17. ***Vaccinated people testing positive for Covid-19 (revisited)***
About 25% of people that test positive for Covid-19 are vaccinated for Covid-19. Define the r.v. X_i to be 1 if someone that tests positive is vaccinated and 0 if they are not vaccinated.
Suppose 3 people have tested positive for Covid-19 (independently of each other). Let T denote the number of people that are vaccinated amongst the 3 that tested positive.

(1) *Using the r.v.'s X_i, write a mathematical equation for calculating T.*

$$T = \sum_{i=1}^{3} X_i$$

(2) *What is the expected value of T?*

$$\mathbb{E}[T] = 3 \cdot p = 3 \cdot 0.25 = 0.75$$

(3) *What is the variance of T?*

$$Var[T] = npq = 3 \cdot 0.25 \cdot 0.75 = 3 \cdot 0.1875 = 0.5625$$
$$SD[T] = 0.5625$$

(4) *What is the probability distribution of T?*

x	$\mathbb{P}(X = x)$
0	0.75^3 = *0.422*
1	$3 \cdot 0.25 \cdot 0.75^2$ = *0.422*
2	$3 \cdot 0.25^2 \cdot 0.75$ = *0.141*
3	0.25^3 = *0.016*

3.2. **Binomial distribution.**

- Many situations involve modeling independent random events that have 2 possible outcomes (binary), such as
  - Repeatedly flipping a coin
  - Whether a person that tested positive with Covid-19 is vaccinated or not
- Repeated events are referred to as **trials**
- The 2 possible outcomes are referred to as **successes** and **failures**.
- We denote the probability of a success as p.
- We denote the probability of a failure as $q = 1 - p$.

3.2.1. *Bernoulli distribution.*

Definition 3.18. ***Bernoulli*** *random variable.*
If X is a random variable that takes value 1 with probability of success p and 0 with probability $1 - p$, then X is a Bernoulli random variable.

probability table; ONE trial

- We call the probability of success p the **parameter** of the Bernoulli distribution.
- Each value of p identifies a specific Bernoulli distribution out of the **family** of Bernoulli r.v.'s where p is any value between 0 and 1 (inclusive).
- If a r.v. X is modeled by a Bernoulli distribution, then we write in shorthand

$$X \sim Bern(p)$$

Theorem 3.19. *Mean and SD of a Bernoulli r.v.*
If X is a Bernoulli r.v. with probability of success p, then

$$\mathbb{E}(X) = p$$
$$Var(X) = p(1 - p)$$
$$SD(X) = \sqrt{p(1 - p)}$$

Covid example

3.2.2. *Binomial distribution.*

Recall Example 3.17 from Day 7:

- About 25% of people that test positive for Covid-19 are vaccinated for Covid-19. p
- Define the r.v. X_i to be 1 if someone that tests positive is vaccinated and 0 if they are not vaccinated. S, F
- Suppose 3 people have tested positive for Covid-19 (independently of each other). fixed n; independent trials
- Let T denote the number of people that are vaccinated amongst the 3 that tested positive. total number S

The random variable T above is an example of a Binomial random variable.

In general, a random variable X is **Binomial** if the following hold:

(1) The trials are independent.
(2) The number of trials, n, is fixed.
(3) Each trial outcome can be classified as a *success* or *failure*.
(4) The probability of a success, p, is the same for each trial.
(5) The r.v. X is the total number of successes in the n trials.

Definition 3.20. *Distribution of a* ***Binomial*** *random variable.*
Let X be the total number of successes in n independent trials, each with probability p of a success.
Then probability of observing exactly k successes in n independent trials is

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

- The parameters of a binomial distribution are p and n.
- If a r.v. X is modeled by a binomial distribution, then we write in shorthand

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

Theorem 3.21. *Mean and SD of a Binomial r.v.*
If X is a binomial r.v. with probability of success p, then

$$\mu = \mathbb{E}(X) = np$$
$$\sigma^2 = Var(X) = np(1-p)$$
$$\sigma = SD(X) = \sqrt{np(1-p)}$$

Covid example 3.17

Example 3.22. ***Vaccinated people testing positive for Covid-19 (revisited)*** *About 25% of people that test positive for Covid-19 are vaccinated for Covid-19. Suppose 10 people have tested positive for Covid-19 (independently of each other). Let X denote the number of people that are vaccinated amongst the 10 that tested positive.*

(1) *What is the expected value of X?*

$$\mathbb{E}[X] = 10 \cdot p = 10 \cdot 0.25 = 2.5$$

(2) *What is the SD of X?*

$$Var[X] = npq = 10 \cdot 0.25 \cdot 0.75 = 10 \cdot 0.1875 = 1.875$$
$$SD[X] = 1.369$$

(3) *What is the probability that exactly 4 of the 10 people that tested positive are vaccinated?*

$$P(X = 4) = \binom{10}{4} 0.25^k (0.75)^{10-k} = \frac{10!}{k!(10-k)!} p^k (1-p)^{10-k} = 0.145998$$

```
dbinom(x, size, prob)
dbinom(x= 4, size=10, prob = .25)= 0.145998
```

(4) *What is the probability that at most 3 of the 10 people that tested positive are vaccinated?*

$$P(X \leq 3) =$$

$P(X \leq q) =$

```
pbinom(q, size, prob, lower.tail = TRUE)
pbinom(3, size=10, prob = .25)= 0.2502823
```

(5) *What is the probability that at least 5 of the 10 people that tested positive are vaccinated?*

$$P(X \geq 5) = P(X > 4) = 0.078$$
$$P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.922$$

$P(X > q) =$

```
pbinom(q, size, prob, lower.tail = FALSE)
pbinom(4, size=10, prob = .25, lower.tail = FALSE)
= 0.07812691
```

$P(X \leq q) =$

```
pbinom(4, size=10, prob = .25, lower.tail = TRUE)
= 0.9218731
```