STAT 390 A
Statistical Methods in Engineering and Science

Week 2 Lectures – Part 2 – Spring 2023

Discrete Random Variables and Probability Distributions

Alexander Giessing

Department of Statistics

University of Washington

April 7, 2023

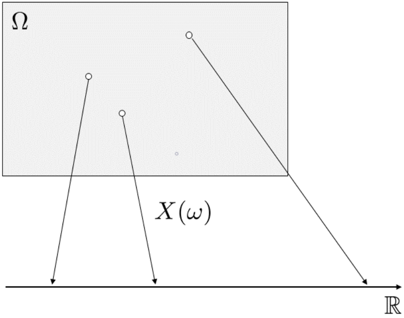
Outline

What are Random Variables?

Often we are not directly interested in the random outcome $\omega \in \Omega$ of an experiment but rather some numerical aspects of the random outcome.

Random Variable

Let Ω be a sample space. A random variable is a function $X : \Omega \to \mathbb{R}$.



Examples of Random Variables

- *Tossing a coin three times.*
  - sample space $\Omega = \{H, T\} \times \{H, T\} \times \{H, T\}$.
  - Let X be the be the number of heads.

Outcomes	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
Notation	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
$X(\omega)$	3	2	2	2	1	1	1	0

- *Random dialing. Consider a scammer who picks and calls telephone numbers at random in a certain area.*
  - sample space $\Omega = \{\text{all allowable phone numbers in that area}\}$.
  - Let $X(\omega)$ be 1 if call ω is picked up and 0 o/w.

Examples of Random Variables (Cont.)

- *Rare disease. To measure the prevalence of a disease, we sample the population until we get a certain number of cases.*
  - sample space $\Omega = \{S, FS, FFS, SFFS, \ldots\}$.
  - Let X be the number of samples required to obtain the first case. Then, $X(\omega)$ is the number of letters in ω.

- *Spatial data. Let X be the current temperature at a random location (defined by latitude and longitude).*
  - sample space $\Omega = [0; 180] \times [0; 360]$.
  - $X(\omega) =$ current temperature at location ω.

Types of Random Variables

Range of a Random Variable

The set of all possible values of a random variable X is called the range of X.

Discrete Random Variable

A random variable is discrete if its range is countable.

Continuous Random Variable

A random variable is continuous if its range is uncountable.

- Special cases:
  - A r.v. whose range comprises only finitely many elements is discrete.
  - A r.v. whose range is an interval of the real line is continuous.
  - A r.v. is of mixed type if it is continuous and discrete in different parts of its range.

Important Notation: Uppercase vs. Lowercase

- The notation $X(\omega) = x$ means that x is the value that the function (i.e. random variable) X assigns to the outcome ω.
- We say that x is a **realization** or **observed value** of $X(\omega)$.
- We often drop ω and write X instead of $X(\omega)$.
- We use random variables to (implicitly) define events (subsets of Ω) as follows:

$$\{X = a\} := \{\omega \in \Omega : X(\omega) = a\}.$$

Example: Let X count the number of heads in three consecutive coin tosses. Then, $\{X = 2\} = \{HHT, HTH, THH\}$.

Outline

Probability Mass Function (PMF)

Once a discrete random vriable X is introduced, the sample space is no longer important. It suffices to list the possible values of X and their corresponding probabilities.

Probability Mass Function

Let X be a discrete random variable. The probability mass function of X is the function $p : \mathbb{R} \to [0, 1]$ defined by

$$p(a) := P(X = a) \quad \text{for} \quad -\infty < a < \infty.$$

- Let X be a discrete r.v. with range $R_X = \{a_1, a_2, \ldots a_m\}$. Then,

$$p(a_i) > 0, \quad p(a_1) + p(a_2) + \ldots + p(a_m) = 1, \quad \text{and} \quad p(a) = 0 \quad \forall a \notin R_X.$$

Examples of PMFs

Tossing a coin three times. Let X be the be the number of heads.

Outcomes	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
Notation	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
$X(\omega)$	3	2	2	2	1	1	1	0

Since consecutive coin tosses are independent, we have

$$\begin{aligned}
p(3) &= P(\{HHH\}) = (1/2)^3 = 1/8 \\
p(2) &= P(\{HHT, HTH, THH\}) = 3 \times (1/2)^3 = 3/8 \\
p(1) &= P(\{HTT, THT, TTH\}) = 3 \times (1/2)^3 = 3/8 \\
p(0) &= P(\{TTT\}) = (1/2)^3 = 1/8,
\end{aligned}$$

and $p(a) = 0$ for all $a \notin \{0, 1, 2, 3\}$.

Examples of PMFs (Cont.)

To measure the prevalence of a rare disease, we sample the population until we get a certain number of cases. Let p be the prevalence probability of the disease and $q = 1 - p$. Denote by X the number of samples until we obtain the first case.

Suppose that we sample people independently, i.e. whether or not a person has the diesease does not affect which person we sample next. Then,

$$\begin{aligned}
p(1) &= P(X = 1) = P(S) = p \\
p(2) &= P(X = 2) = P(SF) = qp \\
p(3) &= P(X = 3) = P(SFF) = q^2 p \\
&\dots \\
p(n) &= P(X = n) = P(S \underbrace{F \dots F}_{(n-1)\text{ times}}) = q^{n-1} p \\
&\dots,
\end{aligned}$$

and $p(n) = 0$ for all $n \notin \mathbb{N}$.

Cumulative Distribution Function (CDF)

Cumulative Distribution Function (Discrete Random Variable)

Let X be a discrete random variable. The cumulative distribution function of X is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(a) := P(X \leq a) = \sum_{x \leq a} p(x) \quad \text{for} \quad -\infty < a < \infty.$$

$F(a)$ is the probability that the observed value of X is at most a.

- Let X be a discrete r.v. with range $R_X = \{a_1, a_2, \ldots\}$. Then,

$$F(a) = \sum_{a_i \leq a} p(a_i).$$

Properties of CDfs

1. F is monotone increasing, i.e.

$$F(a) \leq F(b) \quad \text{for all } a \leq b.$$

2. F is bounded, i.e. $0 \leq F(a) \leq 1$. Moreover,

$$\lim_{a \to +\infty} F(a) = \lim_{a \to +\infty} P(X \leq a) = 1,$$
$$\lim_{a \to -\infty} F(a) = \lim_{a \to -\infty} P(X \leq a) = 0.$$

3. F is right-continuous, i.e.

$$\lim_{\epsilon \to 0^+} F(a + \epsilon) = F(a).$$

- Every cdf satisfies properties 1, 2, and 3, and, conversely, any function F satisfying properties 1,2 and 3 is a cdf of some random variable.

Examples of CDFs

Tossing a coin three times. Let X be the be the number of heads.

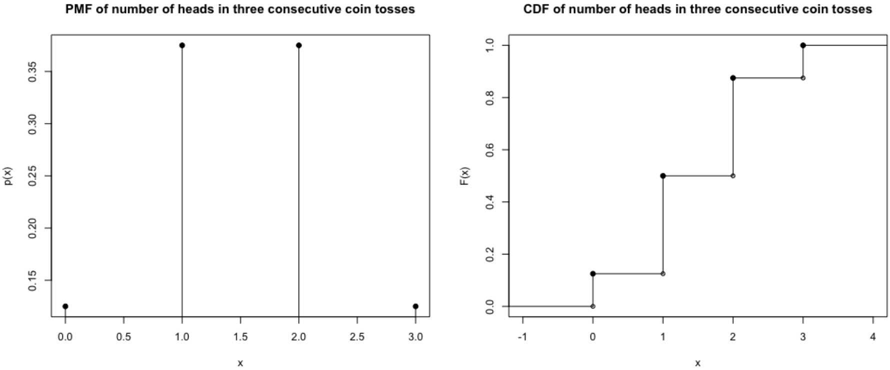
The pmf of X is given by

a	0	1	2	3
$p(a)$	$1/8$	$3/8$	$3/8$	$1/8$

The cdf of X is given by

$$F(a) = \begin{cases} 0 & \text{if } a < 0, \\ 1/8 & \text{if } 0 \leq a < 1, \\ 1/2 & \text{if } 1 \leq a < 2, \\ 7/8 & \text{if } 2 \leq a < 3, \\ 1 & \text{if } a \geq 3. \end{cases}$$

Examples of CDFs (Cont.)



- Hight of each step in cdf is equal to the value of pmf at that step.
- The black circles indicate that the end point is included, the white circles that the end point is excluded, i.e. the cdf is right-continuous.

Examples of CDFs (Cont.)

Rare disease. Let p be the prevalence probability of the disease and $q = 1 - p$. Denote by X the number of samples until we obtain the first case.

The pmf of X is given by

$$p(x) = \begin{cases} q^{x-1}p & \text{if } x \in \mathbb{N}, \\ 0 & \text{o/w}, \end{cases}$$

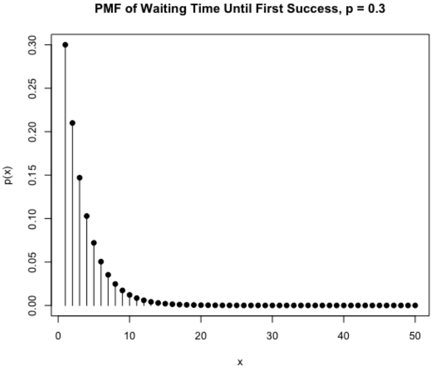
and the cdf of X is given by

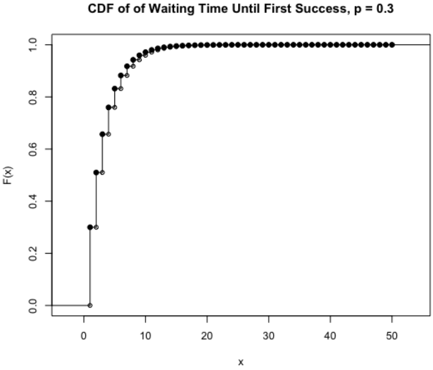
$$F(x) = \sum_{k=1}^{\lfloor x \rfloor} q^{k-1}p = p\frac{1 - q^{\lfloor x \rfloor}}{1 - q} = 1 - q^{\lfloor x \rfloor},$$

for all $x \in \mathbb{R}$.

(Recall that $\lfloor x \rfloor$ denotes the largest integer smaller than $x \in \mathbb{R}$.)

Examples of CDFs (Cont.)





R Code for Preceeding Plots

```
## Coin Tosses
> x <- c(0,1,2,3)
> pmf <- c(1/8, 3/8, 3/8, 1/8)
> cdf <- stepfun(x, c(0, 1/8, 1/2, 7/8, 1), right=TRUE)
> par(mfrow = c(1,2), cex=0.8)
> plot(x, pmf, type="h",main="PMF of number of heads in three
+ consecutive coin tosses",xlab="x",ylab="p(x)")
> points(x,pmf,pch =19)
> plot(cdf, main="CDF of number of heads in three consecutive
+ coin tosses", xlab = "x", ylab = "F(x)")
> points(x, c(1/8, 1/2, 7/8, 1),pch =19)
```

R Code for Preceeding Plots (Cont.)

```
## Rare Disease
> p <- 0.3 3 prevalence probability
> q <- 1-p
> x <- 1:50
> pmf <- q^(x-1)*p
> x.floor <- floor(x)
> cdf <- stepfun(x, c(0, 1-q^x.floor), right=TRUE)
> par(mfrow = c(1,2), cex=0.8)
> plot(x, pmf, type="h",main="PMF of Waiting Time Until First
+ Success, p = 0.3",xlab="x",ylab="p(x)")
> points(x,pmf,pch =19)
> plot(cdf, main="CDF of of Waiting Time Until First Success,
+ p = 0.3", xlab = "x", ylab = "F(x)")
> points(x, 1- q^(x.floor), pch =19)
```

Outline

1. What are Random Variables?
2. Discrete Random Variables
3. Expectation and Variance of Discrete Random Variables

Expected Gains and Losses of a Simple Gamble

Suppose you are in a casino to play roulette. The American roulette has the numbers 0, 00, and 1–36. If you bet \$1 on an even number, your chance to win is 18/38 and to loose 20/38.

- If you played 38,000 times, you would expect to win about 18,000 times and loose about 20,000 times. *(Recall that this is the frequency interpretation of probabilities.)*

- Thus,

$$\begin{aligned}\text{expected payoff per game} &= \frac{\$1 \times 18,000 - \$1 \times 20,000}{38,000} \\ &= \underbrace{\$1 \times \frac{18}{38} + (-\$1) \times \frac{20}{38}}_{\text{sum over value} \times \text{probability}} \approx -\$0.0526.\end{aligned}$$

  That is, on average, you would lose 5 cents per game.

Expectation/ Expected Value/ Mean

Expectation (Discrete Random Variable)

The expectation of a discrete random variable X taking values $a_1, a_2, \ldots$ and with probability mass function p is defined as

$$\mathrm{E}[X] := \sum_i a_i P(X = a_i) = \sum_i a_i p(a_i).$$

- We often write μ for $\mathrm{E}[X]$.
- We also call $\mathrm{E}[X]$ the expected value or mean of X.
- Since the expectation is fully determined by the probability distribution of X, we also speak of the expectation or mean of the distribution.
- The expectation of X is **not the value that is likely to be observed** or the value that should be expected. **It is the long-run average.**

Examples of Expected Values

- *Tossing a coin three times. Let X be the number of heads.*
  The pmf of X is given by

a	0	1	2	3
$p(a)$	1/8	3/8	3/8	1/8

  The expected number of heads shown in three consecutive coin tosses is

$$\mathrm{E}[X] = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{3}{2}.$$

- *Rare disease with prevalence probability p. Let $q = 1 - p$. and denote by X the number of samples until we obtain the first case.*
  The pmf of X is given by

$$p(x) = \begin{cases} q^{x-1}p & \text{if } x \in \mathbb{N}, \\ 0 & \text{o/w}, \end{cases}$$

  The expected number of samples drawn until the first case is

$$\mathrm{E}[X] = \sum_{n=1}^{\infty} n \times q^{n-1} p = \frac{d}{dq}\left(\sum_{i=1}^{\infty} q^n\right) p = \frac{d}{dq}\left(\frac{q}{1-q}\right) p = \frac{1}{p}.$$

Examples of Expected Values (Cont.)

For a rare disease with 1% prevalence rate, the following group testing is used: Pull the blood sample of 10 people together. If the result is negative, all of them are negative. If the result is positive, test each of them individually. If each test costs $1, what is the expected cost?

$$P(\{\text{none have the disease}\}) = 0.99^{10} = 0.9044,$$

$$P(\{\text{at least one has the disease}\}) = 1 - 0.99^{10} = 0.0956.$$

- Let X be the cost of the group testing. Then,

$$\mathrm{E}[X] = \$1 \times 0.9044 + \$11 \times 0.0956 \approx \$1.956.$$

- Compare this with the cost of 10 of the naive testing procedure that tests all 10 people individually!

Expectation of Functions of Random Variables

- Let X be a random variable and $g : \mathbb{R} \rightarrow \mathbb{R}$. Then, $Y = g(X)$ is a random variable taking values in $b_1, b_2, \ldots$ (say) and with pmf and expectation,

$$P(Y = b) = \sum_{a_i : g(a_i) = b} P(X = a_i) \qquad \text{and} \qquad \mathrm{E}[Y] = \sum_i b_i P(Y = b_i).$$

- The following result simplifies calculating $\mathrm{E}[Y]$ dramatically:

CHANGE-OF-VARIABLE FORMULA (DISCRETE RANDOM VARIABLE)

Let X be a discrete random variable taking values $a_1, a_2, \ldots$ and with probability mass function p. For any function $g : \mathbb{R} \rightarrow \mathbb{R}$ we have,

$$\mathrm{E}[g(X)] = \sum_i g(a_i) P(X = a_i) = \sum_i g(a_i) p(a_i).$$

- **Linearity of Expectation:** $\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$ for all $a, b \in \mathbb{R}$.

Example: Functions of Random Variables

Tossing a coin three times. Let X be the number of heads and $Y = (X-2)^2$.

- Range of X and Y are

X	0	1	2	3
Y	4	1	0	1

- The pmf of Y is given by

b	0	1	4
$p(b)$	3/8	1/2	1/8

- Expected value of Y can be computed in two ways:

$$\mathrm{E}[Y] = 0 \times \frac{3}{8} + 1 \times \frac{1}{2} + 4 \times \frac{1}{8} = 1.$$

$$\mathrm{E}[Y] = \mathrm{E}[(X-2)^2] = 4 \times \frac{1}{8} + 1 \times \frac{3}{8} + 0 \times \frac{3}{8} + 1 \times \frac{1}{8} = 1.$$

Variance and Standard Deviation

Variance and Standard Deviation

Let X be a random variable. Tve variance and standard deviation of X are defined as, respectively,

$$\mathrm{Var}(X) := \mathrm{E}[(X - \mathrm{E}[X])^2] \quad \text{and} \quad \mathrm{SD}(X) := \sqrt{\mathrm{Var}(X)}.$$

- We often write σ^2 for $\mathrm{Var}(X)$ and σ for $\mathrm{SD}(X)$.
- Variance and standard deviation are measures of the spread of X (around its mean).
- **Shortcut formula:** $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$.
- **Properties:** $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$ and $\mathrm{SD}(aX + b) = |a|\mathrm{SD}(X)$ for all $a, b \in \mathbb{R}$.