PSY 503: Foundations of Psychological Methods
Lecture 9: Random Variables II

Robin Gomila

Princeton

September 28, 2020

Uniform (discrete) distribution

Uniform (discrete) distribution

A random variable X follows a **uniform** distribution if each of the possible values of X has the same probability of occurrence.

As a result, a uniform discrete random variable X can be fully summarized using one parameter k, which corresponds to the number of possible values x that X can take on.

We write $X \sim Unif(k)$

PMF of $X \sim Unif(k)$

PMF of X is

$$P(X = x) = \begin{cases} \frac{1}{C} & \text{if } x \in Supp(X) \\ 0 & otherwise \end{cases} \tag{1}$$

Example: Rolling a die

- Let X be the outcome of a die roll.
- X can take on $k = 6$ different values: $\{1, 2, 3, 4, 5, 6\}$.

As a result, for any value of $X \in Supp(X)$: $P(X = x) = \frac{1}{6}$

Rolling a die: PMF using R

We can use the ddunif() function from the "extraDistr" package to calculate the probability of different values x_i and plot the PMF:

```r
library(extraDistr)

uniform_x <- 0:7
uniform_pmf <- ddunif(x = uniform_x,
                      min=1,
                      max=6)

round(uniform_pmf, 3)
```

```
## [1] 0.000 0.167 0.167 0.167 0.167 0.167 0.167 0.000
```

Uniform distribution in psychology studies

Uniform distribution in psychology studies

- Psychologists have used the properties of the uniform distribution to study honesty and lying

Uniform distribution in psychology studies

- Psychologists have used the properties of the uniform distribution to study honesty and lying
- Procedure
  - Invite participants to the lab
  - Ask them to roll a fair die *privately*
  - Report the outcome of the die roll

Uniform distribution in psychology studies

- Psychologists have used the properties of the uniform distribution to study honesty and lying
- Procedure
  - Invite participants to the lab
  - Ask them to roll a fair die *privately*
  - Report the outcome of the die roll
- Trick
  - Payoff structure
  - Make more money if die roll returned certain numbers (e.g., 5)

Uniform distribution in psychology studies

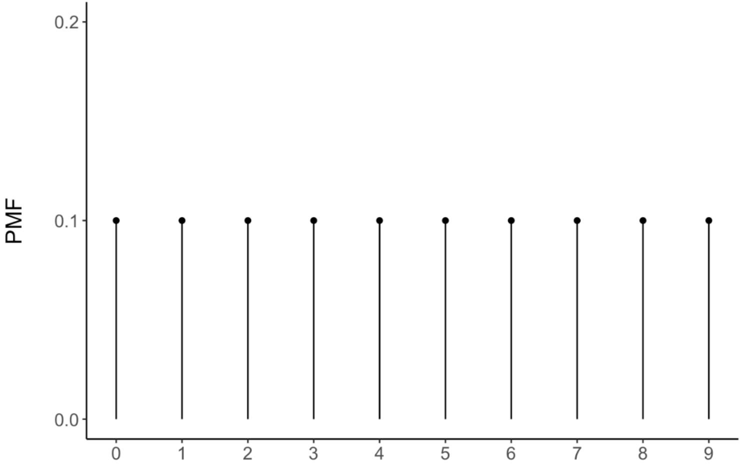
- Can't tell who lied
- Can tell if group lied, on average

Example: Election Fraud

- Uniform discrete distribution to study election fraud
- Examine the distribution of the last digit of the vote counts reported by the authorities
- A fair vote count is just as likely to end in any digit
- But people are bad are making up numbers: they tend to select some digits more frequently than others
- If last digit not uniformly distributed in some counties: red flag!

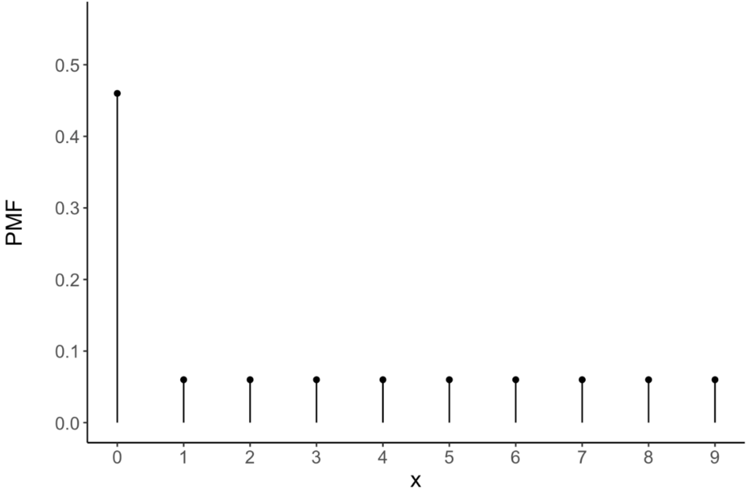
Election Fraud: Expected PMF of last digit

Let X be the last digit of the number of provincial vote counts at a given election. We expect, $X \sim Unif(k = 10)$ and $P(X = x) = \frac{1}{10}$. That is, the PMF of X should look like:



Election Fraud: Observed PMF of last digit

What would you conclude if instead, you observed:



Election fraud: Reference

Beber, B., & Scacco, A. (2012). What the numbers say: A digit-based test for election fraud. *Political analysis, 20*(2), 211-234.

Many distributions

We will encounter additional common parametric distributions of discrete random variables. Examples include the *Poisson distribution*, the *geometric distribution*, and *Benford's Law* (the distribution of first digits!!).

From PMFs to CDFs

- We have described the distribution of random variables using Probability Mass Functions (PMF)

From PMFs to CDFs

- We have described the distribution of random variables using Probability Mass Functions (PMF)
- Another useful function to describe random variables is called the *cumulative distribution function* (CDF)

Cumulative Distribution Function: Definition

- The CDF of a random variable is the function F such that $F(x) = P(X \leq x)$

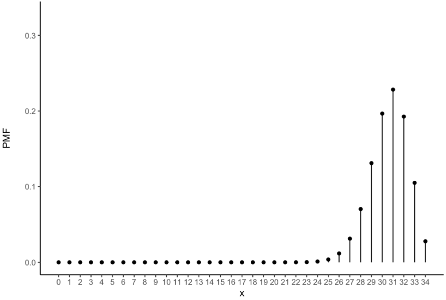
Cumulative Distribution Function: Definition

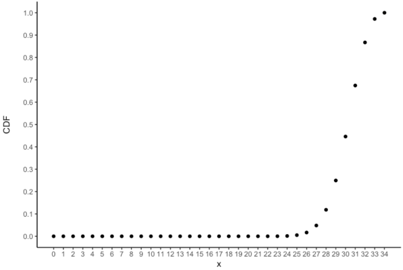
- The CDF of a random variable is the function F such that $F(x) = P(X \leq x)$
- PMF tells us the probability of each possible outcome
  - e.g., probability that exactly 30 tables show up if you accepted 34 reservations

Cumulative Distribution Function: Definition

- The CDF of a random variable is the function F such that $F(x) = P(X \leq x)$
- PMF tells us the probability of each possible outcome
  - e.g., probability that exactly 30 tables show up if you accepted 34 reservations
- CDF tells us the probability that an outcome below a specific outcome occurs
  - e.g., probability of less than 30 tables showing up if you accepted 34 reservations

PDF vs. CDF


PMF
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
x


CDF
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
x

Example: Trendy Dining in NYC

- How do we calculate $P(X > 30)$?
  - For X the number of "successes" from a series of $n = 34$ Bernoulli trials with probability of success $\theta = 0.90$

Example: Trendy Dining in NYC

- How do we calculate $P(X > 30)$?
  - For X the number of "successes" from a series of $n = 34$ Bernoulli trials with probability of success $\theta = 0.90$
- We could use the binomial equation

$$\frac{n!}{(n-x)!x!}\theta^x(1-\theta)^{n-x} \tag{2}$$

and calculate by hand:

$$P(X > 30) = P(X = 31) + P(X = 32) + P(X = 33) + P(X = 34)$$

Example: Trendy Dining in NYC

- How do we calculate $P(X > 30)$?
  - For X the number of "successes" from a series of $n = 34$ Bernoulli trials with probability of success $\theta = 0.90$
- We could use the binomial equation

$$\frac{n!}{(n-x)!x!}\theta^x(1-\theta)^{n-x} \tag{2}$$

and calculate by hand:

$$P(X > 30) = P(X = 31) + P(X = 32) + P(X = 33) + P(X = 34)$$

- Tedious!

Example: Trendy Dining in NYC

- Use the cumsum() function in R
- Generate the complete PMF of X, then use the cumsum() function

```
library(tidyverse)

x_vector <- 0:34
all_probs <- dbinom(x = x_vector,
                    size = 34,
                    prob = 0.9)

cdf_x <- cumsum(all_probs)
```

Example: Trendy Dining in NYC

X <int>	CDF <dbl>
25	0.005
26	0.017
27	0.048
28	0.119
29	0.250
30	0.446
31	0.674
32	0.867
33	0.972
34	1.000

1–10 of 10 rows

Example: Trendy Dining in NYC

X <int>	CDF <dbl>
25	0.005
26	0.017
27	0.048
28	0.119
29	0.250
30	0.446
31	0.674
32	0.867
33	0.972
34	1.000

1–10 of 10 rows

- We immediately see that $P(X \leq 30) = 0.446$, which implies that

$$P(X > 30) = 1 - 0.446 = 0.554$$

Summarizing Discrete Random Variables

- PMFs and CDFs are very useful tools to summarize information from rvs.
- Many other ways to summarize random variables!
    - e.g., mean, median, standard deviation, etc.

Arithmetic mean

- You have calculated the arithmetic mean plenty of times in your life
  - Add up a series of numbers (i.e., grades) and divide by the total number of grades

Arithmetic mean

- You have calculated the arithmetic mean plenty of times in your life
  - Add up a series of numbers (i.e., grades) and divide by the total number of grades
- Given a list of numbers x_1, x_2, ..., x_n, we define the *arithmetic mean* as:

$$\mu_x = \frac{1}{n} \sum_{j=1}^{n} x_j \tag{3}$$

Weighted mean

- Some numbers (e.g., grades) may have more weight than others

Weighted mean

- Some numbers (e.g., grades) may have more weight than others
- For a list of numbers x_1, x_2, ..., x_n and a list of weights p_1, p_2, ..., p_n, the *weighted mean* is defined as:

$$\text{weighted-mean}(x) = \sum_{j=1}^{n} x_j p_j \tag{4}$$

in which the weights are non-negative numbers that add up to 1

Weighted mean

- Some numbers (e.g., grades) may have more weight than others
- For a list of numbers x_1, x_2, ..., x_n and a list of weights p_1, p_2, ..., p_n, the *weighted mean* is defined as:

$$\text{weighted-mean}(x) = \sum_{j=1}^{n} x_j p_j \tag{4}$$

in which the weights are non-negative numbers that add up to 1

- Arithmetic mean is a special case

Expectation

Expectation of random variables

- Random variables are defined by their PMF / CDF
    - NOT by a series of numbers that we can add

Expectation of random variables

- Random variables are defined by their PMF / CDF
  - NOT by a series of numbers that we can add
- We talk about the **expectation** or **expected value** of a rv

Definition

- Let X be a discrete random variable.
- The *expectation* of X is defined by:

$$\mathbb{E}[X] = \sum_{j=1}^{n} x_j P(X = x_j) = \sum_{j=1}^{n} x_j p(x_j) \tag{5}$$

Definition

- Let X be a discrete random variable.
- The *expectation* of X is defined by:

$$\mathbb{E}[X] = \sum_{j=1}^{n} x_j P(X = x_j) = \sum_{j=1}^{n} x_j p(x_j) \tag{5}$$

- The expected value of a random variable is a function of its PMF

Expectation of Binomial distribution

- We have $X \sim Bin(34,\ 0.90)$

Expectation of Binomial distribution

- We have $X \sim Bin(34,\ 0.90)$
- What is the expectation of X?
  - If I draw a very large number of outcomes from this distribution, what will the mean value of these outcomes be?

Expectation of Binomial distribution

- Based on the PMF of X (see trendy dining slides)

$$\mathbb{E}[X] = \sum_{j=1}^{n} x_j P(X = x_j)$$

$$\begin{aligned} &= 0 \times 0.0 + 1 \times 0.0 + 2 \times 0.0 \ + ... + \ 23 \times 0.0 \\ &\quad + 24 \times 0.001 + 25 \times 0.004 + 26 \times 0.012 + 27 \times 0.031 \\ &\quad + 28 \times 0.070 + 29 \times 0.131 + 30 \times 0.197 + 31 \times 0.228 \\ &\quad + 32 \times 0.193 + 33 \times 0.105 + 34 \times 0.028 \end{aligned}$$

$$= 30.6$$

Expectation of rvs in R

- Using simulations
  - Draw a large number of observations from a distribution
  - Take the mean of these values

Expectation of rvs in R

```
binom3490_draws <- rbinom(n = 100000,
                          size = 34,
                          prob = .90)

mu_binom3490 <- mean(binom3490_draws)

round(mu_binom3490, 1)
```

```
## [1] 30.6
```

Expectation of rvs in R

```
binom3450_draws <- rbinom(n = 100000,
                          size = 34,
                          prob = .50)

mu_binom3450 <- mean(binom3450_draws)

round(mu_binom3450, 1)
```

```
## [1] 17
```

From expectation to PMF?

- Does the expectation of a rv tell us anything about its PMF?

From expectation to PMF?

- Does the expectation of a rv tell us anything about its PMF?
- NO!

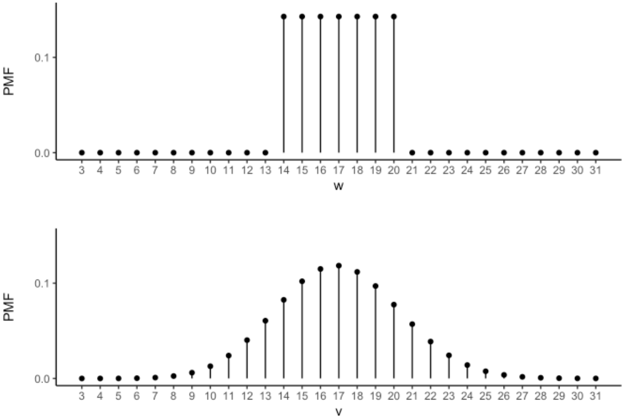
From expectation to PMF?

- Does the expectation of a rv tell us anything about its PMF?
- NO!
- Expectation is a **number** that informs us about the **centrality** of a rv
- Expectation tells you nothing about how often $X = 17$ will occur
  - Actually, 17 does not even have to be on the support of X

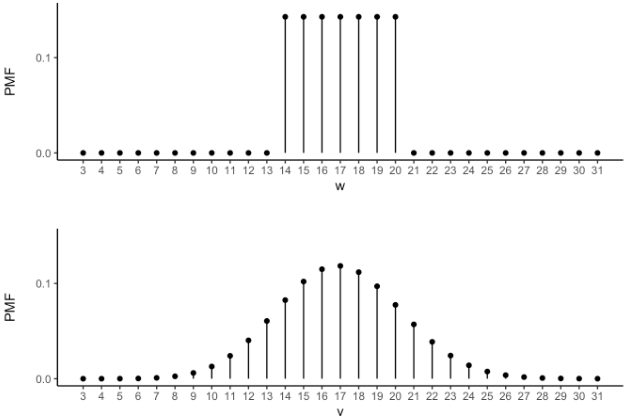
From expectation to PMF?

- Does the expectation of a rv tell us anything about its PMF?
- NO!
- Expectation is a **number** that informs us about the **centrality** of a rv
- Expectation tells you nothing about how often $X = 17$ will occur
  - Actually, 17 does not even have to be on the support of X
- Expectation does not tell you how often you will draw values very close or very far from 17

Illustration



Illustration



This is why rvs are often described with a measure of centrality and a measure of spread.

Properties if Expectations

$$\mathbb{E}[c] = c$$

Properties if Expectations

$$\mathbb{E}[c] = c$$

$$\mathbb{E}[X + c] = \mathbb{E}[X] + c$$

Properties if Expectations

$$\mathbb{E}[c] = c$$

$$\mathbb{E}[X + c] = \mathbb{E}[X] + c$$

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Properties if Expectations

$$\mathbb{E}[c] = c$$

$$\mathbb{E}[X + c] = \mathbb{E}[X] + c$$

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[cX] = c\mathbb{E}[X]$$

Variance of random variables

The variance tells us something about the average distance between X and $\mathbb{E}[X]$. For this reason, the variance of a random variable is defined as a function of its expectation.

Variance: Definition

$$\mathbb{V}[X] = \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big]$$

Variance: Definition

$$\mathbb{V}[X] = \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big]$$

- Often expressed in the following terms

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Proof

$$
\begin{aligned}
\mathbb{V}[X] &= \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big] \\
&= \mathbb{E}\Big[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\Big] \\
&= \mathbb{E}[X^2] - 2\mathbb{E}\Big[X\mathbb{E}[X]\Big] + \mathbb{E}\Big[\mathbb{E}[X]^2\Big] \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2
\end{aligned}
$$

Properties

$$\mathbb{V}[X + c] = \mathbb{V}[X]$$
$$\mathbb{V}[aX] = a^2\mathbb{V}[X]$$
$$\mathbb{V}[X] \geq 0$$

Standard Deviation

The standard deviation of a random variable X is defined as

$$\sigma_X = \sqrt{\mathbb{V}[X]}$$

Standard Deviation

The standard deviation of a random variable X is defined as

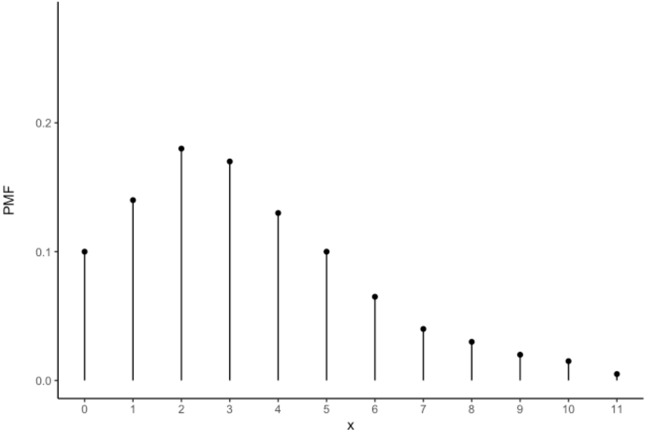
$$\sigma_X = \sqrt{\mathbb{V}[X]}$$

- Generally easier to interpret than the $\mathbb{V}[X]$
  - Same unit as the X.
  - σ_X corresponds to the average distance between X and $\mathbb{E}[X]$

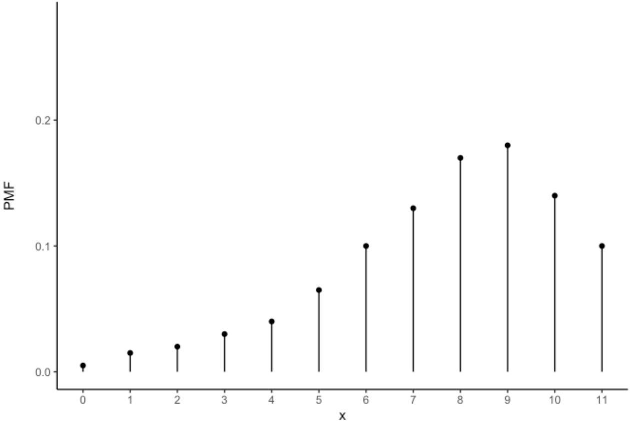
Skewness

- Distributions can be skewed
  - i.e., non-symmetrical
- If $\mathrm{Skew}[X] \neq 0$, the probability distribution of X is not symmetrical

Positively skewed rvs (or right skewed)



Negatively skewed rvs (or left skewed)



Formal definition

$$\text{Skew}[X] = \mathbb{E}\left[\frac{X - \mathbb{E}[X]}{\mathbb{V}[X]}\right]$$

Centrality and Skewness

- When rvs are skewed, expectation may not be the most relevant measure of centrality
- Expectation influenced by presence of extreme values at in the tail

Illustration: Students' grades

- Suppose that you are designing in an intervention aiming at improving students' grades in a particularly difficult course
  - Let X be students' grade at a given test with the following PMF:

$$p(x) = \begin{cases} \frac{2}{10} & \text{if } x = 38 \\ \frac{2}{10} & \text{if } x = 39 \\ \frac{4}{10} & \text{if } x = 40 \\ \frac{2}{10} & \text{if } x = 100 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Illustration: Students' grades

- Suppose that you are designing in an intervention aiming at improving students' grades in a particularly difficult course
  - Let X be students' grade at a given test with the following PMF:

$$p(x) = \begin{cases} \frac{2}{10} & \text{if } x = 38 \\ \frac{2}{10} & \text{if } x = 39 \\ \frac{4}{10} & \text{if } x = 40 \\ \frac{2}{10} & \text{if } x = 100 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

- 80% of the students obtain a grade between 38 and 40 out 100, whereas 20% of the students get 100
- $\mathbb{E}[X] = 52.5$: not as useful, even misleading!

Median

The *median* of a random variable X is a number such that, if we were to repeat the random phenomenon on which X is defined many many many times, 50% of the times we would observe an outcome smaller than the median and 50% we would observe an outcome larger than the median.

Mode

The *mode* is the most typical or common realization of a random variable. It corresponds to the "peak" of the probability distribution of a random variable.