STATISTICS

Marco Caserta
marco.caserta@ie.edu

IE University

Random variables

- A random variable is a numeric quantity whose value depends on the outcome of a random event
  - We use a capital letter, like X, to denote a random variable
  - The values of a random variable are denoted with a lower case letter, in this case x
  - For example, $P(X = x)$
- There are two types of random variables:
  - Discrete random variables often take only integer values
    - Example: Number of credit hours, Difference in number of credit hours this term vs last
  - Continuous random variables take real (decimal) values
    - Example: Cost of books this term, Difference in cost of books this term vs last

Expectation

- We are often interested in the average outcome of a random variable.
- We call this the EXPECTED VALUE (mean), and it is a weighted average of the possible outcomes

$$\mu = E(X) = \sum_{i=1}^{k} x \, P(X = x_i)$$

- Do not miss the following TED video: Dan Gilbert on “Why We Make Bad Decisions.”

Expected value of a discrete random variable

In a game of cards you win \$1 if you draw a heart, \$5 if you draw an ace (including the ace of hearts), \$10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

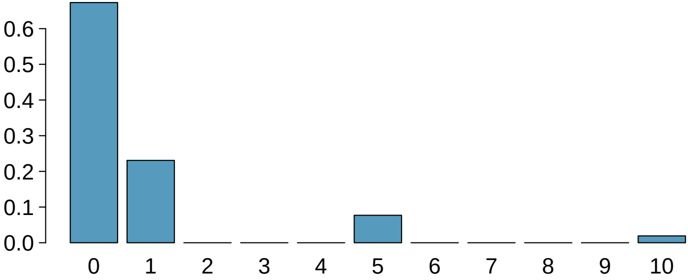
Expected value of a discrete random variable

In a game of cards you win \$1 if you draw a heart, \$5 if you draw an ace (including the ace of hearts), \$10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

Event	X	$P(X)$	$X\,P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

Expected value of a discrete random variable (cont.)

Below is a visual representation of the probability distribution of winnings from this game:



Variability

We are also often interested in the variability in the values of a random variable.

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} (x_i - E(X))^2 P(X = x_i)$$

$$\sigma = SD(X) = \sqrt{Var(X)}$$

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

X	$P(X)$	$X\ P(X)$	$(X - E(X))^2$	$P(X)\ (X - E(X))^2$
1	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \times 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \times \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \times 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \times \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \times 84.0889 = 1.6242$
0	$\frac{35}{52}$	$0 \times \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \times 0.6561 = 0.4416$
		$E(X) = 0.81$		

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

X	$P(X)$	$X\ P(X)$	$(X - E(X))^2$	$P(X)\ (X - E(X))^2$
1	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \times 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \times \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \times 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \times \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \times 84.0889 = 1.6242$
0	$\frac{35}{52}$	$0 \times \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \times 0.6561 = 0.4416$
		$E(X) = 0.81$		$V(X) = 3.4246$

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

X	$P(X)$	$X\,P(X)$	$(X - E(X))^2$	$P(X)\,(X - E(X))^2$
1	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \times 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \times \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \times 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \times \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \times 84.0889 = 1.6242$
0	$\frac{35}{52}$	$0 \times \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \times 0.6561 = 0.4416$
		$E(X) = 0.81$		$V(X) = 3.4246$
				$SD(X) = \sqrt{3.4246} = 1.85$

Linear combinations

- A LINEAR COMBINATION of random variables X and Y is given by

$$aX + bY$$

where a and b are some fixed numbers.

Linear combinations

- A linear combination of random variables X and Y is given by

$$aX + bY$$

  where a and b are some fixed numbers.
- The average value of a linear combination of random variables is given by

$$E(aX + bY) = a \times E(X) + b \times E(Y)$$

- Thus, if we consider a linear function of a random variable $a + bX$, we have:

$$E(a + bX) = a + b \times E(X)$$

Calculating the expectation of a linear combination

On average you take 10 minutes for each statistics homework problem and 15 minutes for each chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and physics homework for the week?

Calculating the expectation of a linear combination

On average you take 10 minutes for each statistics homework problem and 15 minutes for each chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and physics homework for the week?

$$\begin{aligned} E(5S + 4C) &= 5 \times E(S) + 4 \times E(C) \\ &= 5 \times 10 + 4 \times 15 \\ &= 50 + 60 \\ &= 110\ min \end{aligned}$$

Linear combinations

- The variability of a linear combination of two independent random variables is calculated as

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

Linear combinations

- The variability of a linear combination of two independent random variables is calculated as

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

- The standard deviation of the linear combination is the square root of the variance.

Linear combinations

- The variability of a linear combination of two independent random variables is calculated as

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

- The standard deviation of the linear combination is the square root of the variance.
- Thus, if we consider a linear function of a random variable $a + bX$, we have:

$$V(a + bX) = b^2 \times V(X)$$

Linear combinations

- The variability of a linear combination of two independent random variables is calculated as

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

- The standard deviation of the linear combination is the square root of the variance.
- Thus, if we consider a linear function of a random variable $a + bX$, we have:

$$V(a + bX) = b^2 \times V(X)$$

Note: If the random variables are not independent, the variance calculation gets a little more complicated and will be presented later.

Calculating the variance of a linear combination

The standard deviation of the time you take for each statistics homework problem is 1.5 minutes, and it is 2 minutes for each chemistry problem. What is the standard deviation of the time you expect to spend on statistics and physics homework for the week if you have 5 statistics and 4 chemistry homework problems assigned?

Calculating the variance of a linear combination

The standard deviation of the time you take for each statistics homework problem is 1.5 minutes, and it is 2 minutes for each chemistry problem. What is the standard deviation of the time you expect to spend on statistics and physics homework for the week if you have 5 statistics and 4 chemistry homework problems assigned?

$$\begin{aligned} V(5S + 4C) &= 5^2 \times V(S) + 4^2 \times V(C) \\ &= 25 \times 1.5^2 + 16 \times 2^2 \\ &= 56.25 + 64 \\ &= 120.25 \end{aligned}$$

Practice

A casino game costs \$5 to play. If you draw first a red card, then you get to draw a second card. If the second card is the ace of hearts, you win \$500. If not, you don't win anything, i.e. lose your \$5. What is your expected profits/losses from playing this game? Remember: profit/loss = winnings - cost.

(A) A loss of 10¢

(B) A loss of 25¢

(C) A loss of 30¢

(D) A profit of 5¢

Practice

A casino game costs \$5 to play. If you draw first a red card, then you get to draw a second card. If the second card is the ace of hearts, you win \$500. If not, you don't win anything, i.e. lose your \$5. What is your expected profits/losses from playing this game? Remember: profit/loss = winnings - cost.

(A) A loss of 10¢

(B) A loss of 25¢

(C) *A loss of 30¢*

(D) A profit of 5¢

Event	*Win*	*Profit: X*	$P(X)$	$X \times P(X)$
Red, A♥	*500*	*500 - 5 = 495*	$\frac{25}{52} \times \frac{1}{51} = 0.0094$	$495 \times 0.0094 = 4.653$
Other	*0*	*0 - 5 = -5*	$1 - 0.0094 = 0.9906$	$-5 \times 0.9906 = -4.953$
				$E(X) = -0.3$

Fair game

A FAIR game is defined as a game that costs as much as its expected payout, i.e. expected profit is 0.

Fair game

A FAIR game is defined as a game that costs as much as its expected payout, i.e. expected profit is 0.

Do you think casino games in Vegas cost more or less than their expected payouts?

Fair game

A FAIR game is defined as a game that costs as much as its expected payout, i.e. expected profit is 0.

Do you think casino games in Vegas cost more or less than their expected payouts?

If those games cost less than their expected payouts, it would mean that the casinos would be losing money on average, and hence they wouldn't be able to pay for all this ⇒

Simplifying random variables

Random variables do not work like normal algebraic variables:

$$X + X \neq 2X$$

Simplifying random variables

Random variables do not work like normal algebraic variables:

$$X + X \neq 2X$$

$$\begin{aligned} E(X+X) &= E(X) + E(X) \\ &= 2E(X) \end{aligned}$$

$$E(2X) = 2E(X)$$

$$\begin{aligned} Var(X+X) &= Var(X) + Var(X) \text{ (assuming independence)} \\ &= 2\ Var(X) \end{aligned}$$

$$\begin{aligned} Var(2X) &= 2^2\ Var(X) \\ &= 4\ Var(X) \end{aligned}$$

Simplifying random variables

Random variables do not work like normal algebraic variables:

$$X + X \neq 2X$$

$$\begin{aligned} E(X+X) &= E(X) + E(X) \\ &= 2E(X) \end{aligned}$$

$$\begin{aligned} Var(X+X) &= Var(X) + Var(X) \text{ (assuming independence)} \\ &= 2\ Var(X) \end{aligned}$$

$$E(2X) = 2E(X)$$

$$\begin{aligned} Var(2X) &= 2^2\ Var(X) \\ &= 4\ Var(X) \end{aligned}$$

Note: $E(X + X) = E(2X)$ but $Var(X + X) \neq Var(2X)$.

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average $2,154 with a standard deviation of $132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average \$2,154 with a standard deviation of \$132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average \$2,154 with a standard deviation of \$132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

$$E(X_1 + X_2 + X_3 + X_4 + X_5) = E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5)$$

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average $2,154 with a standard deviation of $132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

$$\begin{aligned} E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\ &= 5 \times E(X) = 5 \times 2,154 = \$10,770 \end{aligned}$$

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average \$2,154 with a standard deviation of \$132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

$$\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 2,154 = \$10,770 \\
Var(X_1 + X_2 + X_3 + X_4 + X_5) &= Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var(X_5)
\end{aligned}$$

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average \$2,154 with a standard deviation of \$132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

$$\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 2,154 = \$10,770 \\
Var(X_1 + X_2 + X_3 + X_4 + X_5) &= Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var(X_5) \\
&= 5 \times V(X) = 5 \times 132^2 = \$87,120
\end{aligned}$$

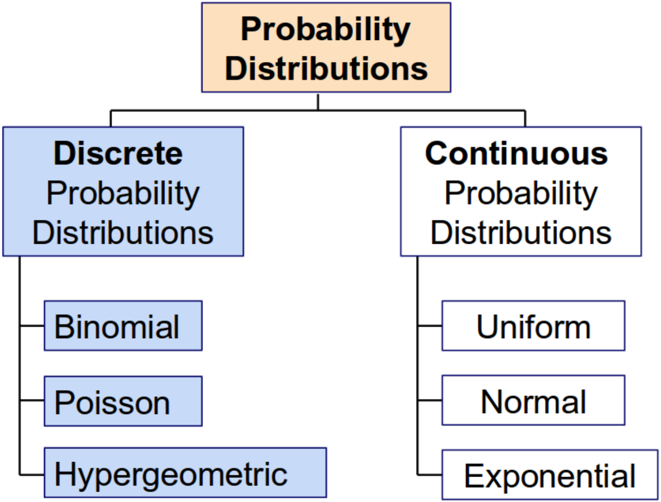
Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average \$2,154 with a standard deviation of \$132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

Note that we have 5 cars each with the given annual maintenance cost $(X_1 + X_2 + X_3 + X_4 + X_5)$, not one car that had 5 times the given annual maintenance cost $(5X)$.

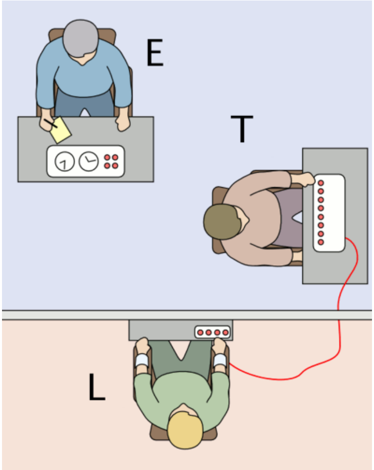
$$
\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 2,154 = \$10,770 \\
Var(X_1 + X_2 + X_3 + X_4 + X_5) &= Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var(X_5) \\
&= 5 \times V(X) = 5 \times 132^2 = \$87,120 \\
SD(X_1 + X_2 + X_3 + X_4 + X_5) &= \sqrt{87,120} = 295.16
\end{aligned}
$$

Probability Distributions



MILGRAM EXPERIMENT

- Stanley Milgram, a Yale University psychologist, conducted a series of experiments on obedience to authority starting in 1963.
- Experimenter (E) orders the teacher (T), the subject of the experiment, to give severe electric shocks to a learner (L) each time the learner answers a question incorrectly.
- The learner is actually an actor, and the electric shocks are not real, but a prerecorded sound is played each time the teacher administers an electric shock.



Milgram experiment (cont.)

- These experiments measured the willingness of study participants to obey an authority figure who instructed them to perform acts that conflicted with their personal conscience.
- Milgram found that about 65% of people would obey authority and give such shocks.
- Over the years, additional research suggested this number is approximately consistent across communities and time.

Bernouilli random variables

- Each person in Milgram's experiment can be thought of as a trial.
- A person is labeled a success if she refuses to administer a severe shock, and failure if she administers such shock.
- Since only 35% of people refused to administer a shock, probability of success is $p = 0.35$.
- When an individual trial has only two possible outcomes, it is called a Bernoulli random variable.

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Let's call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock":

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Let's call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock":

#1: $\underset{\text{(A) refuse}}{\underline{0.35}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Let's call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock":

$$\#1: \quad \underset{\text{(A) refuse}}{\underline{0.35}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\#2: \quad \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) refuse}}{\underline{0.35}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Let's call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock":

$$\#1: \underset{\text{(A) refuse}}{\underline{0.35}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\#2: \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) refuse}}{\underline{0.35}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\#3: \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) refuse}}{\underline{0.35}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Let's call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock":

$$\#1: \underset{\text{(A) refuse}}{\underline{0.35}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\#2: \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) refuse}}{\underline{0.35}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\#3: \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) refuse}}{\underline{0.35}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\#4: \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) refuse}}{\underline{0.35}} = 0.0961$$

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Let's call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock":

#1: $\underset{\text{(A) refuse}}{\underline{0.35}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$

#2: $\underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) refuse}}{\underline{0.35}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$

#3: $\underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) refuse}}{\underline{0.35}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$

#4: $\underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) refuse}}{\underline{0.35}} = 0.0961$

The probability of exactly one 1 of 4 people refusing to administer the shock is the sum of all of these probabilities.

$$0.0961 + 0.0961 + 0.0961 + 0.0961 = 4 \times 0.0961 = 0.3844$$

Binomial distribution

The question from the prior slide asked for the probability of given number of successes, k, in a given number of trials, n, ($k = 1$ success in $n = 4$ trials), and we calculated this probability as

$$\#\ of\ scenarios \times P(single\ scenario)$$

Binomial distribution

The question from the prior slide asked for the probability of given number of successes, k, in a given number of trials, n, ($k = 1$ success in $n = 4$ trials), and we calculated this probability as

$$\#\ of\ scenarios \times P(single\ scenario)$$

- $\#\ of\ scenarios$: there is a less tedious way to figure this out, we'll get to that shortly...

Binomial distribution

The question from the prior slide asked for the probability of given number of successes, k, in a given number of trials, n, ($k = 1$ success in $n = 4$ trials), and we calculated this probability as

$$\# \; of \; scenarios \times P(single \; scenario)$$

- $\# \; of \; scenarios$: there is a less tedious way to figure this out, we'll get to that shortly...
- $P(single \; scenario) = p^k \; (1-p)^{(n-k)}$

  probability of success to the power of number of successes, probability of failure to the power of number of failures

Binomial distribution

The question from the prior slide asked for the probability of given number of successes, k, in a given number of trials, n, ($k = 1$ success in $n = 4$ trials), and we calculated this probability as

$$\# \, of \; scenarios \times P(single \; scenario)$$

- $\# \, of \; scenarios$: there is a less tedious way to figure this out, we'll get to that shortly...
- $P(single \; scenario) = p^k \; (1 - p)^{(n-k)}$

  probability of success to the power of number of successes, probability of failure to the power of number of failures

The Binomial distribution describes the probability of having exactly k successes in n independent Bernouilli trials with probability of success p.

Counting the # of scenarios

Earlier we wrote out all possible scenarios that fit the condition of exactly one person refusing to administer the shock. If n was larger and/or k was different than 1, for example, $n = 9$ and $k = 2$:

Counting the # of scenarios

Earlier we wrote out all possible scenarios that fit the condition of exactly one person refusing to administer the shock. If n was larger and/or k was different than 1, for example, $n = 9$ and $k = 2$:

RRSSSSSSS

Counting the # of scenarios

Earlier we wrote out all possible scenarios that fit the condition of exactly one person refusing to administer the shock. If n was larger and/or k was different than 1, for example, $n = 9$ and $k = 2$:

RRSSSSSSS
SRRSSSSSS

Counting the # of scenarios

Earlier we wrote out all possible scenarios that fit the condition of exactly one person refusing to administer the shock. If n was larger and/or k was different than 1, for example, $n = 9$ and $k = 2$:

RRSSSSSSS
SRRSSSSSS
SSRRSSSSS
...
SSRSSRSSS
...
SSSSSSSRR

writing out all possible scenarios would be incredibly tedious and prone to errors.

Calculating the # of scenarios

Choose function

The choose function is useful for calculating the number of ways to choose k successes in n trials.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Calculating the # of scenarios

Choose function

The choose function is useful for calculating the number of ways to choose k successes in n trials.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- $k = 1$, $n = 4$: $\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4\times3\times2\times1}{1\times(3\times2\times1)} = 4$

Calculating the # of scenarios

Choose function

The choose function is useful for calculating the number of ways to choose k successes in n trials.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- $k = 1$, $n = 4$: $\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4\times3\times2\times1}{1\times(3\times2\times1)} = 4$
- $k = 2$, $n = 9$: $\binom{9}{2} = \frac{9!}{2!(9-1)!} = \frac{9\times8\times7!}{2\times1\times7!} = \frac{72}{2} = 36$

Properties of the choose function

Which of the following is false?

(A) There are n ways of getting 1 success in n trials, $\binom{n}{1} = n$.

(B) There is only 1 way of getting n successes in n trials, $\binom{n}{n} = 1$.

(C) There is only 1 way of getting n failures in n trials, $\binom{n}{0} = 1$.

(D) There are $n - 1$ ways of getting $n - 1$ successes in n trials, $\binom{n}{n-1} = n - 1$.

Properties of the choose function

Which of the following is false?

(A) There are n ways of getting 1 success in n trials, $\binom{n}{1} = n$.

(B) There is only 1 way of getting n successes in n trials, $\binom{n}{n} = 1$.

(C) There is only 1 way of getting n failures in n trials, $\binom{n}{0} = 1$.

(D) *There are $n-1$ ways of getting $n-1$ successes in n trials, $\binom{n}{n-1} = n-1$.*

Binomial distribution (cont.)

Binomial probabilities

If p represents probability of success, $(1 - p)$ represents probability of failure, n represents number of independent trials, and k represents number of successes

$$P(k \ successes \ in \ n \ trials) = \binom{n}{k} p^k \ (1-p)^{(n-k)}$$

Assumptions:

- There are several trials, each with only two possible outcomes
- The probability of success in each trial is always the same
- Trials are independent

Binomial distribution: Mean and Variance

- Let X be the number of successes in n independent trials, each with probability of success p. Then, X follows a binomial distribution with *mean*:

$$\mu = E[X] = np$$

and variance:

$$\sigma^2 = E\left[(X-\mu)^2\right] = np(1-p)$$

Which of the following is not a condition that needs to be met for the binomial distribution to be applicable?

(A) the trials must be independent

(B) the number of trials, n, must be fixed

(C) each trial outcome must be classified as a *success* or a *failure*

(D) the number of desired successes, k, must be greater than the number of trials

(E) the probability of success, p, must be the same for each trial

Which of the following is not a condition that needs to be met for the binomial distribution to be applicable?

(A) the trials must be independent

(B) the number of trials, n, must be fixed

(C) each trial outcome must be classified as a *success* or a *failure*

(D) *the number of desired successes, k, must be greater than the number of trials*

(E) the probability of success, p, must be the same for each trial

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(A) pretty high

(B) pretty low

Gallup: http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx, January 23, 2013.

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(A) pretty high

(B) *pretty low*

Gallup: http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx, January 23, 2013.

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(A) $0.262^8 \times 0.738^2$

(B) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(C) $\binom{10}{8} \times 0.262^8 \times 0.738^2$

(D) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(A) $0.262^8 \times 0.738^2$

(B) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(C) $\binom{10}{8} \times 0.262^8 \times 0.738^2 = 45 \times 0.262^8 \times 0.738^2 = 0.0005$

(D) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

The birthday problem

What is the probability that 2 randomly chosen people share a birthday?

The birthday problem

What is the probability that 2 randomly chosen people share a birthday?

Pretty low, $\frac{1}{365} \approx 0.0027$.

The birthday problem

What is the probability that 2 randomly chosen people share a birthday?

Pretty low, $\frac{1}{365} \approx 0.0027$.

What is the probability that at least 2 people out of 366 people share a birthday?

The birthday problem

What is the probability that 2 randomly chosen people share a birthday?

Pretty low, $\frac{1}{365} \approx 0.0027$.

What is the probability that at least 2 people out of 366 people share a birthday?

Exactly 1! (Excluding the possibility of a leap year birthday.)

The birthday problem (cont.)

What is the probability that at least 2 people (1 match) out of 121 people share a birthday?

The birthday problem (cont.)

What is the probability that at least 2 people (1 match) out of 121 people share a birthday?

Somewhat complicated to calculate, but we can think of it as the complement of the probability that there are no matches in 121 people.

$$P(no\ matches) = 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{120}{365}\right)$$

The birthday problem (cont.)

What is the probability that at least 2 people (1 match) out of 121 people share a birthday?

Somewhat complicated to calculate, but we can think of it as the complement of the probability that there are no matches in 121 people.

$$
\begin{aligned}
P(\textit{no matches}) &= 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{120}{365}\right) \\
&= \frac{365 \times 364 \times \cdots \times 245}{365^{121}}
\end{aligned}
$$

The birthday problem (cont.)

What is the probability that at least 2 people (1 match) out of 121 people share a birthday?

Somewhat complicated to calculate, but we can think of it as the complement of the probability that there are no matches in 121 people.

$$
\begin{aligned}
P(no\ matches) &= 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{120}{365}\right) \\
&= \frac{365 \times 364 \times \cdots \times 245}{365^{121}} \\
&= \frac{365!}{365^{121} \times (365 - 121)!}
\end{aligned}
$$

The birthday problem (cont.)

What is the probability that at least 2 people (1 match) out of 121 people share a birthday?

Somewhat complicated to calculate, but we can think of it as the complement of the probability that there are no matches in 121 people.

$$\begin{aligned}
P(no\ matches) &= 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{120}{365}\right) \\
&= \frac{365 \times 364 \times \cdots \times 245}{365^{121}} \\
&= \frac{365!}{365^{121} \times (365 - 121)!} \\
&= \frac{121! \times \binom{365}{121}}{365^{121}}
\end{aligned}$$

The birthday problem (cont.)

What is the probability that at least 2 people (1 match) out of 121 people share a birthday?

Somewhat complicated to calculate, but we can think of it as the complement of the probability that there are no matches in 121 people.

$$
\begin{aligned}
P(no\ matches) &= 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{120}{365}\right) \\
&= \frac{365 \times 364 \times \cdots \times 245}{365^{121}} \\
&= \frac{365!}{365^{121} \times (365 - 121)!} \\
&= \frac{121! \times \binom{365}{121}}{365^{121}} \approx 0
\end{aligned}
$$

The birthday problem (cont.)

What is the probability that at least 2 people (1 match) out of 121 people share a birthday?

Somewhat complicated to calculate, but we can think of it as the complement of the probability that there are no matches in 121 people.

$$\begin{aligned}
P(\textit{no matches}) &= 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{120}{365}\right) \\
&= \frac{365 \times 364 \times \cdots \times 245}{365^{121}} \\
&= \frac{365!}{365^{121} \times (365 - 121)!} \\
&= \frac{121! \times \binom{365}{121}}{365^{121}} \approx 0 \\
P(\textit{at least } 1 \textit{ match}) &\approx 1
\end{aligned}$$

An analysis of Facebook users

A recent study found that “Facebook users get more than they give”. For example:

- 40% of Facebook users in our sample made a friend request, but 63% received at least one request
- Users in our sample pressed the like button next to friends’ content an average of 14 times, but had their content “liked” an average of 20 times
- Users sent 9 personal messages, but received 12
- 12% of users tagged a friend in a photo, but 35% were themselves tagged in a photo

Any guesses for how this pattern can be explained?

http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

An analysis of Facebook users

A recent study found that "Facebook users get more than they give". For example:

- 40% of Facebook users in our sample made a friend request, but 63% received at least one request
- Users in our sample pressed the like button next to friends' content an average of 14 times, but had their content "liked" an average of 20 times
- Users sent 9 personal messages, but received 12
- 12% of users tagged a friend in a photo, but 35% were themselves tagged in a photo

Any guesses for how this pattern can be explained?

Power users contribute much more content than the typical user.

http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

This study also found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

We are given that $n = 245, p = 0.25$, and we are asked for the probability $P(K \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

This study also found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

We are given that $n = 245, p = 0.25$, and we are asked for the probability $P(K \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$$\begin{aligned} P(X \geq 70) &= P(K = 70 \text{ or } K = 71 \text{ or } K = 72 \text{ or } \cdots \text{ or } K = 245) \\ &= P(K = 70) + P(K = 71) + P(K = 72) + \cdots + P(K = 245) \end{aligned}$$

This study also found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

We are given that $n = 245, p = 0.25$, and we are asked for the probability $P(K \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$$
\begin{aligned}
P(X \geq 70) &= P(K = 70 \text{ or } K = 71 \text{ or } K = 72 \text{ or } \cdots \text{ or } K = 245) \\
&= P(K = 70) + P(K = 71) + P(K = 72) + \cdots + P(K = 245)
\end{aligned}
$$

This seems like an awful lot of work...

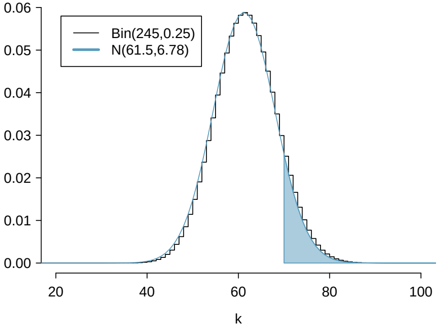
Normal approximation to the binomial

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

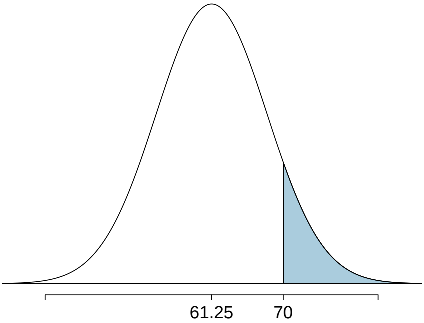
$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.

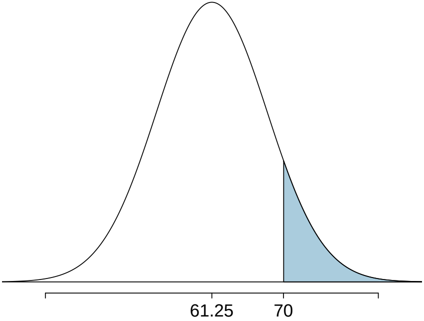


What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?

What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?

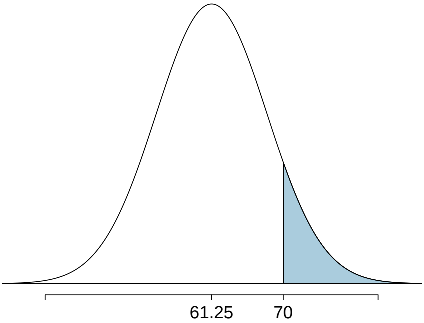


What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

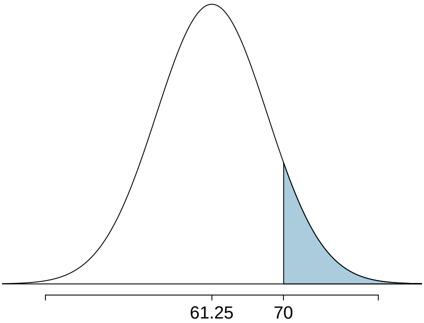
What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

	Second decimal place of Z				
Z	0.05	0.06	0.07	0.08	0.09
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	0.9015

What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

	Second decimal place of Z				
Z	0.05	0.06	0.07	0.08	0.09
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	0.9015

$$P(Z > 1.29) = 1 - 0.9015 = 0.0985$$

Poisson distribution

- The Poisson distribution is often useful for estimating the number of rare events in a large population over a short unit of time for a fixed population if the individuals within the population are independent.
- The rate for a Poisson distribution is the average number of occurrences in a mostly-fixed population per unit of time, and is typically denoted by λ.
- Using the rate, we can describe the probability of observing exactly k rare events in a single unit of time.

Poisson distribution

p(observe k rare events) = $\frac{\lambda^k e^{-\lambda}}{k!}$, where:

- k may take a value 0, 1, 2, and so on, and $k!$ represents k-factorial.
- The letter $e \approx 2.718$ is the base of the natural logarithm.
- The mean and standard deviation of this distribution are λ and $\sqrt{\lambda}$, respectively.

Poisson distribution

Assume an interval is divided into a very large number of equal subintervals where the probability of the occurrence of an event in any subinterval is very small.
Poisson distribution assumptions:

- The probability of the occurrence of an event is constant for all subintervals.
- There can be no more than one occurrence in each subinterval.
- Occurrences are independent; that is, an occurrence in one interval does not influence the probability of an occurrence in another interval.

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

Given $\lambda = 2$.

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

Given $\lambda = 2$.

$$P(\text{only 1 failure in a week}) = \frac{2^1 \times e^{-2}}{1!}$$

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

Given $\lambda = 2$.

$$\begin{aligned} P(\text{only 1 failure in a week}) &= \frac{2^1 \times e^{-2}}{1!} \\ &= \frac{2 \times e^{-2}}{1} \end{aligned}$$

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

Given $\lambda = 2$.

$$
\begin{aligned}
P(\text{only 1 failure in a week}) &= \frac{2^1 \times e^{-2}}{1!} \\
&= \frac{2 \times e^{-2}}{1} \\
&= 0.27
\end{aligned}
$$

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given day the electricity fails three times.

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given day the electricity fails three times.

We are given the weekly failure rate, but to answer this question we need to first calculate the average rate of failure on a given day: $\lambda_{day} = \frac{2}{7} = 0.2857$. Note that we are assuming that the probability of power failure is the same on any day of the week, i.e. we assume independence.

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given day the electricity fails three times.

We are given the weekly failure rate, but to answer this question we need to first calculate the average rate of failure on a given day: $\lambda_{day} = \frac{2}{7} = 0.2857$. Note that we are assuming that the probability of power failure is the same on any day of the week, i.e. we assume independence.

$$P(\text{3 failures on a given day}) = \frac{0.2857^1 \times e^{-0.2857}}{3!}$$

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given day the electricity fails three times.

We are given the weekly failure rate, but to answer this question we need to first calculate the average rate of failure on a given day: $\lambda_{day} = \frac{2}{7} = 0.2857$. Note that we are assuming that the probability of power failure is the same on any day of the week, i.e. we assume independence.

$$\begin{aligned} P(\text{3 failures on a given day}) &= \frac{0.2857^1 \times e^{-0.2857}}{3!} \\ &= \frac{0.2857 \times e^{-0.2857}}{6} \end{aligned}$$

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given day the electricity fails three times.

We are given the weekly failure rate, but to answer this question we need to first calculate the average rate of failure on a given day: $\lambda_{day} = \frac{2}{7} = 0.2857$. Note that we are assuming that the probability of power failure is the same on any day of the week, i.e. we assume independence.

$$
\begin{aligned}
P(\text{3 failures on a given day}) &= \frac{0.2857^1 \times e^{-0.2857}}{3!} \\
&= \frac{0.2857 \times e^{-0.2857}}{6} \\
&= 0.0358
\end{aligned}
$$

Is it Poisson?

- A random variable may follow a Poisson distribution if the event being considered is rare, the population is large, and the events occur independently of each other .
- However we can think of situations where the events are not really independent. For example, if we are interested in the probability of a certain number of weddings over one summer, we should take into consideration that weekends are more popular for weddings.
- In this case, a Poisson model may sometimes still be reasonable if we allow it to have a different rate for different times; we could model the rate as higher on weekends than on weekdays.
- The idea of modeling rates for a Poisson distribution against a second variable (day of the week) forms the foundation of some more advanced methods called GENERALIZED LINEAR MODELS. There are beyond the scope of this course.

Hypergeometric Distribution

- "n" trials in a sample taken from a finite population of size N
- Sample taken without replacement
- Outcomes of trials are dependent

Goal: Concerned with finding the probability of "x" successes in the sample where there are "S" successes in the population.

- Sampling with replacement (or very large population) $\Rightarrow$ Binomial distribution
- Sampling without replacement (prob changes with each selection) $\Rightarrow$ Hypergeometric distribution

Hypergeometric Distribution: The Logic

1. The number of possible ways that x successes can be selected from the sample out of S successes in the population is:
$$C_x^S = \binom{S}{x} = \frac{S!}{x!(S-x)!}$$

2. The number of possible ways that $n-x$ nonsuccesses can be selected from the sample out of $N-S$ nonsuccesses in the population is:
$$C_{n-x}^{N-S} = \binom{N-S}{n-x} = \frac{(N-S)!}{(n-x)!(N-S-n+x)!}$$

3. The total number of different samples of size n that can be obtained from a population of size N is:
$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

Hypergeometric Distribution

$$p(x) = \frac{C_x^S C_{n-x}^{N-S}}{C_n^N}$$

Example

A company receives shipment of 20 items. Because inspection of each individual item is expensive, it has a policy of checking a random sample of 6 items from such a shipment, and if no more than 1 sample item is defective, the remainder will not be checked. What is the probability that a shipment of 5 defective items will not be subjected to additional checking?

- "defective" is "success" in this example
- $N = 20$, "S=5", $n = 6$
- We need to find $p(x \leq 1) = p(x = 0) + p(x = 1)$

$$\begin{aligned}
p(0) &= \frac{\frac{5!}{0!5!} \times \frac{15!}{6!9!}}{\frac{20!}{6!14!}} = 0.129 \\
p(1) &= \frac{\frac{5!}{1!4!} \times \frac{15!}{5!10!}}{\frac{20!}{6!14!}} = 0.387
\end{aligned}$$

Joint Probability Distribution

- A joint probability function is used to express the probability that X takes the specific value x and simultaneously Y takes the value y, as a function of x and y:

$$p(x, y) = (X = x \cap Y = y)$$

- Marginal Probability Distribution:
  - $p(x) = \sum_y p(x, y)$
  - $p(y) = \sum_x p(x, y)$
- Properties of the joint distribution:
  - $0 \leq p(x, y) \leq 1$
  - sum of the joint $p(x, y)$ over all pair of values must be equal to 1

Conditional Probability Distribution

- The conditional probability of X, given $Y = y$ is:

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

- The conditional probability of Y, given $X = x$ is:

$$p(y|x) = \frac{p(x, y)}{p(x)}$$

- Two jointly distributed random variables X and Y are said to be INDEPENDENT if and only if:

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

Conditional Mean and Variance

- The conditional mean is:

$$\mu_{Y|X} = E[Y|X] = \sum_y (y|x)p(y|x)$$

- The conditional variance is:

$$\sigma^2_{Y|X} = E[(Y - \mu_{Y|X})^2|X] = \sum_y \left[(y - \mu_{y|x})^2|x\right] p(y|x)$$

Covariance and Correlation

- Let X and Y be two discrete random variables with mean μ_x and μ_y, respectively;
- The expected value of $(x-\mu_x)(y-\mu_y)$ is called covariance between X and Y:

$$Cov(X,Y) = E[(x-\mu_x)(y-\mu_y)] = \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y)$$

- The correlation between X and Y is:

$$\rho = Corr(X,Y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

- Remember: $-1 \leq \rho \leq 1$

Application: Portfolio Analysis

- Let random variable X be the price for stock A and random variable Y be the price for stock B
- The market value, W, for the portfolio is given by the linear function:

$$W = aX + bY$$

where a is the number of shares of stock A, and b is the number of shares of stock B

Application: Portfolio Analysis

- The MEAN value of W is:

$$\mu_W = E[W] = E[aX + bY] = a\mu_x + b\mu_y$$

- The VARIANCE of W is:

$$\sigma_W^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2abCov(x, y)$$

- An alternative formula for the variance (using the correlation) is:

$$\sigma_W^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2abCorr(x, y)\sigma_x\sigma_y$$

Portfolio Analysis: Example

p(x,y)	Economic Condition	Passive Fund (X)	Aggressive Fund (Y)
0.2	Recession	-25€	-200€
0.5	Stable	50€	60€
0.3	Expanding	100€	350€

- Compute $\mu_X, \mu_Y, \sigma_X, \sigma_Y, Cov(X, Y), \rho$
- Suppose the portfolio is composed of 40% X and 60% Y. What is the average return and the variability of the portfolio?
- Interpret your findings (See textbook, pg. 186)