Lesson 5: Random variables and Binomial distribution

TB sections 3.1-3.2

Nicky Wakim

2024-10-14

Learning Objectives

1. Define random variables and how they map to probability distributions
2. Calculate the expected value and variance of discrete random variables
3. Calculate the expected value and variance of linear combinations of discrete random variables
4. Calculate probabilities for different events using a Binomial distribution

Where are we?

Data	Probability	Sampling Variability, and Statistical Inference
Collecting data	Probability rules	Sampling distributions
Categorical vs. Numeric	Independence, conditional	Central Limit Theorem
Summary statistics	Random variables and probability distributions	Confidence Intervals
Data visualization	Linear combinations	Hypothesis tests
	Binomial, Normal, and Poisson	

Inference for continuous data/outcomes

One sample t-test	3+ independent samples	Simple linear regression / correlation
2 sample tests: paired and independent	Power and sample size	Non-parametric tests

Inference for categorical data/outcomes

One proportion test	Fisher's exact test	Non-parametric tests
Chi-squared test	2 proportion test	Power and sample size

R

Basics	Reproducibility	Quarto	Packages	Data visualization	Data wrangling	R Projects	…

Learning Objectives

1. Define random variables and how they map to probability distributions
2. Calculate the expected value and variance of discrete random variables
3. Calculate the expected value and variance of linear combinations of discrete random variables
4. Calculate probabilities for different events using a Binomial distribution

Random variables

Random variable (RV or r.v.)

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

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

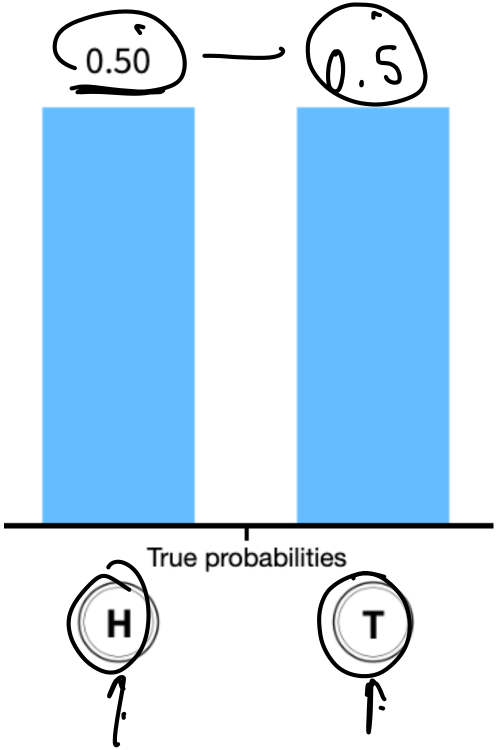
X X

From Lesson 3: Probability distributions

Probability distribution

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

- We've already seen one in our heads and tails example



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

X = # tails in 6 coin tosses

In the coin toss example...

- We can start to define the probability distribution
- Let's define the coin flip with the random variable X
  - Where $X = 1$ if we get a ~~heads~~ and $X = 0$ if we get a ~~tails~~

- We can create a table for the random variable and probabilities of each outcome:

Coin flip (x)	$x = 1$	$x = 0$
Probability ($P(X = x)$)	0.5	0.5

RV realized value of RV

$P(X = x)$

  - **Note:** I use X to refer to the random variable and x to refer to the realized value it takes
  - Then we write $P(X = x)$ to discuss the probability for each realized value (x) of the random variable (X)

$P(X=1) = 0.5$

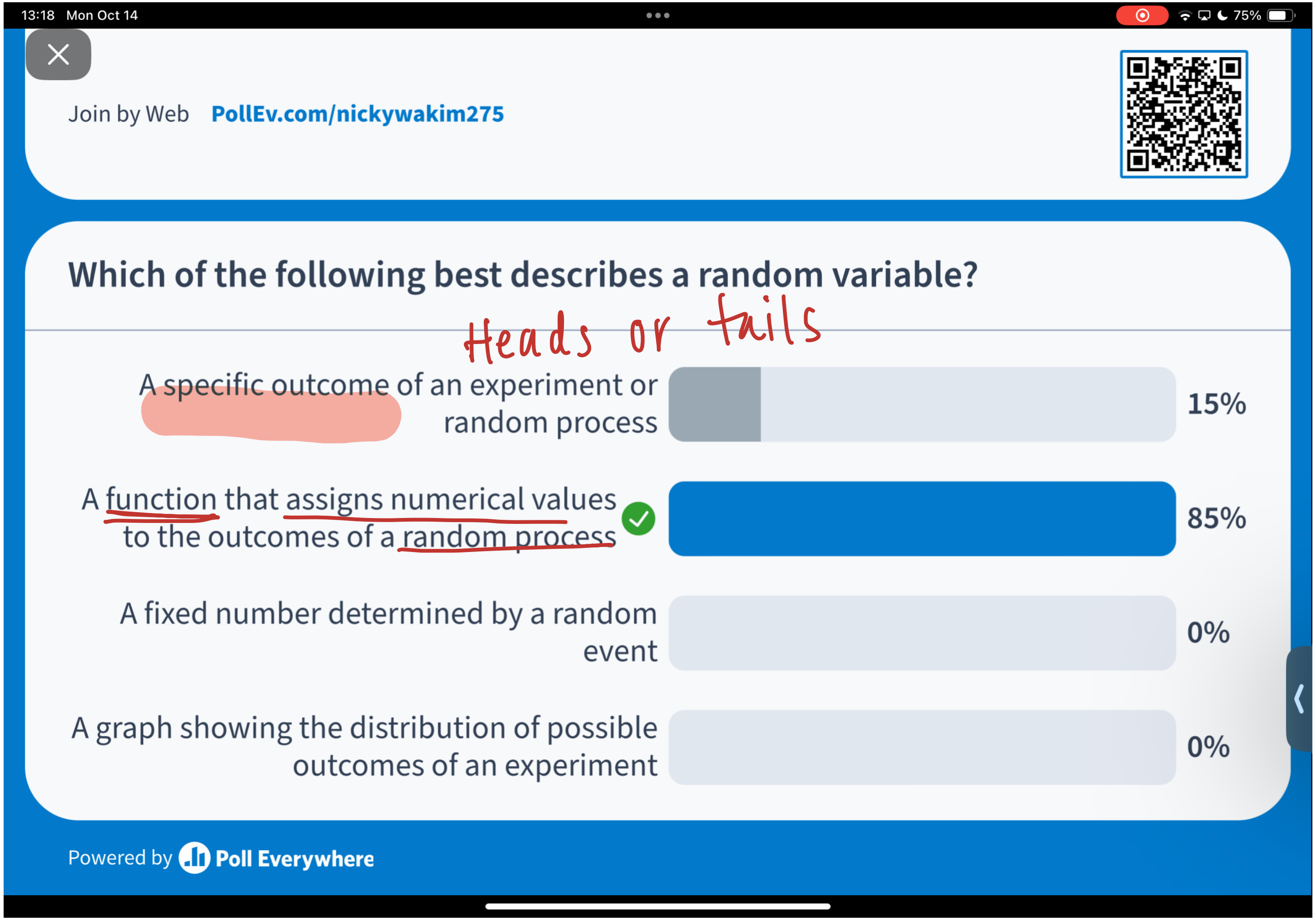
$P(X=0) = 0.5$

- Also note that the sum of the probabilities equal 1: $\sum_{x=0}^{1} P(X = x) = 1$

"sum"

$$\sum_{x=0}^{1} P(X = x) = P(X=0) + P(X=1) = 0.5 + 0.5 = 1$$

Poll Everywhere Question 1



Let's extend this to rolling a die

$\boxed{\cdot} \rightarrow 1/2$

Example 1: 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.

RV is function mapping roll to itself (identity fn)

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

x	1	2	3	4	5	6
$P(X=x)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

$$P(X=x) = \frac{1}{6} \quad \text{for } x = 1,2,3,4,5,6$$

Learning Objectives

1. Define random variables and how they map to probability distributions
2. Calculate the expected value and variance of discrete random variables
3. Calculate the expected value and variance of linear combinations of discrete random variables
4. Calculate probabilities for different events using a Binomial distribution

Discrete vs. continuous random variables

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

Discrete random variable

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

Continuous random variable

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

Think: # stars in sky

- Number of heads in a set of coin tosses
- Number of people who have had chicken pox in a random sample

↓ inf & countable

finite outcomes

$P(X = x)$ = real prob

Think:

- Height in a population → 3 m 3.01 m 3.0001 m
- Blood pressure in a population

↓

infinite & not countable

$P(X = x) = 0$

Expectation of random variables DISCRETE

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

Expected value of a *discrete* random variable

$x_1 = 1$, $x_2 = 0$ (H&T ex)

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:

$$\mu = E[X] = x_1 P(X = x_1) + x_2 P(X = x_2) + \ldots + x_k P(X = x_k)$$
$$= \sum_{i=1}^{k} x_i P(X = x_i)$$

Back to rolling a die

Example 1: Rolling a die

Let's go back to our fair die with RV X as the value of the face showing on the die.

2. What is the expected outcome of the RV X?

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

x	$\mathbb{P}(X = x)$
1	0.10
2	0.20
3	0.05
4	0.05
5	0.25
6	0.35

(2)

x	1	2	3	4	5	6
$P(X=x)$	1/6	1/6	1/6	1/6	1/6	1/6

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \frac{1}{6} + 6 \frac{1}{6}$$

(μ)

$$= 3.5$$

Not a possible outcome of X

(3)

$$\mu = E(X) = 1(0.1) + 2(0.2) + 3(0.05) + 4(0.05) + 5(0.25) + 6(0.35)$$

$$= 4.2$$

Variability of random variables

Just like with data, the variability of a r.v. is described with its variance or standard deviation

Variance of a discrete random variable

If X takes on outcomes $x_1, ..., x_k$ with probabilities $P(X = x_1), ..., P(X = x_k)$ and expected value $\mu = E(X)$, then the variance of X, denoted by $\text{Var}(X)$ or σ^2, is

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

Standard deviation of a discrete random variable

The standard deviation of X, labeled $SD(X)$ or σ, is

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

Back to rolling a die

Example 1: Rolling a die

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.

4. Find the variance and standard deviation of X.

x	$\mathbb{P}(X = x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

④ $$Var(X) = \sum_{i=1}^{6} (x_i - \underset{\downarrow\, 3.5}{\mu})^2 P(X = x_i)$$

$$= (1 - 3.5)^2 (1/6) + (2 - 3.5)^2\left(\tfrac{1}{6}\right) + (3 - 3.5)^2\left(\tfrac{1}{6}\right) + (4 - 3.5)^2\left(\tfrac{1}{6}\right) + (5 - 3.5)^2\left(\tfrac{1}{6}\right) + (6 - 3.5)^2\left(\tfrac{1}{6}\right)$$

$$\sigma^2 = Var(X) = 2.92$$

$$\sigma = SD(X) = \sqrt{Var(X)} = \sqrt{2.92} = 1.71$$

Learning Objectives

1. Define random variables and how they map to probability distributions
2. Calculate the expected value and variance of discrete random variables
3. Calculate the expected value and variance of linear combinations of discrete random variables
4. Calculate probabilities for different events using a Binomial distribution

Linear combinations of random variables

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.

a, b $\quad 3, 2$

$3X + 2Y \to$ weighted die

↳ one die

Theorem: Expected value of a linear combination of random variables

If X and Y are random variables and a and b are constants, then

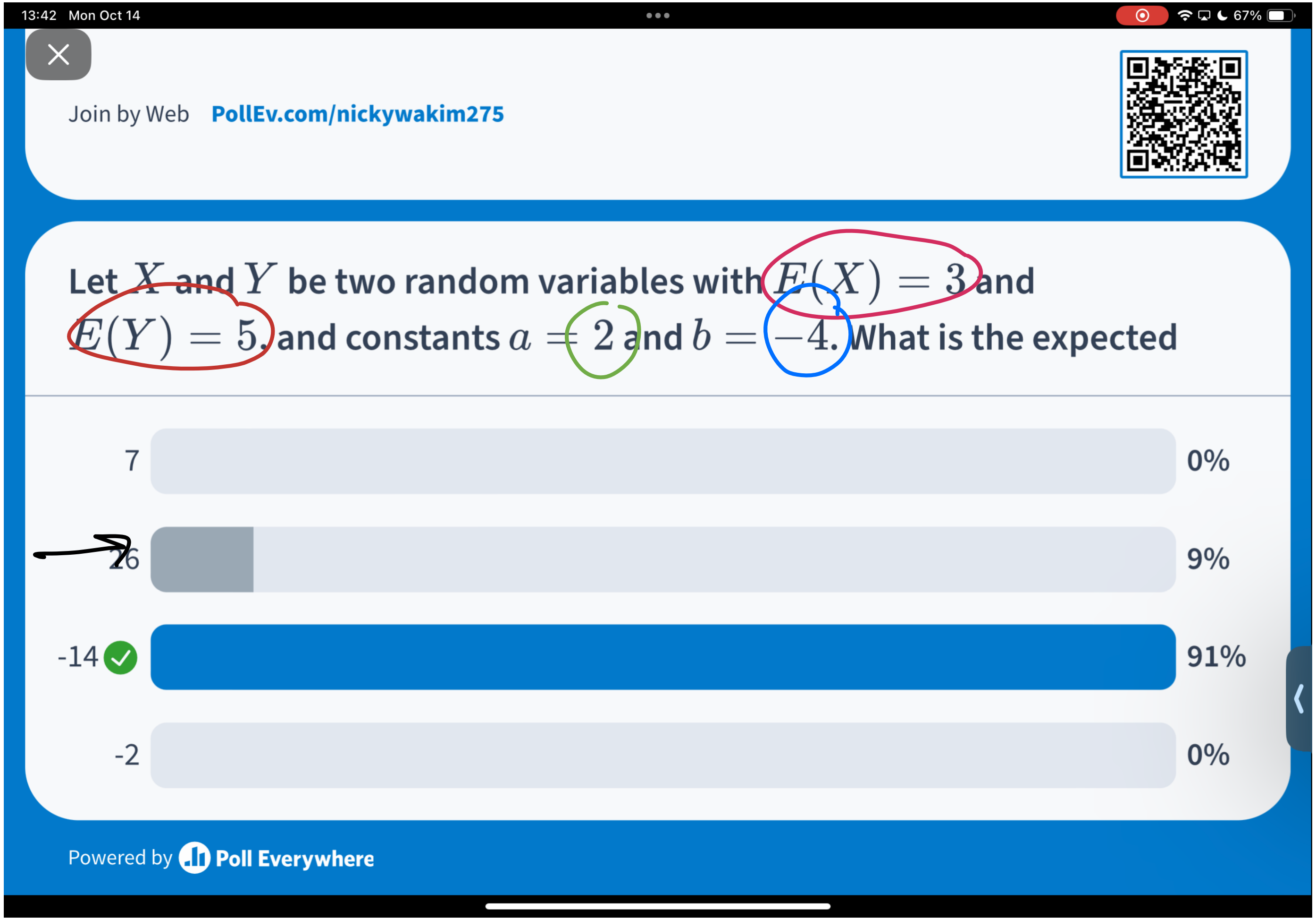
$$E(aX + bY) = aE(X) + bE(Y)$$

$E(2) = 2$

and

$$E(aX + b) = aE(X) + b$$

Poll Everywhere Question2



$$E(aX + bY)$$
$$= a E(X) + b E(Y)$$
$$= 2(3) - 4(5)$$
$$= 6 - 20$$
$$= -14$$

Keep rolling dice!

Example: Expected money for rolling 3 dice

Let the random variables X_1, X_2, X_3 be the values shown on rolls for 2 fair 6-sided dice and 1 unfair die (as described in our previous example). 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 unfair die roll. How much money do you expect to get?

X_1, X_2, X_3

linear combo: $M = 1 \cdot X_1 + 2X_2 + 4X_3$

$E(X_1) = E(X_2) = 3.5$

$E(X_3) = 4.2$

$$E(M) = E(1 \cdot X_1 + 2X_2 + 4X_3)$$
$$= E(X_1) + 2E(X_2) + 4E(X_3)$$
$$= 3.5 + 2(3.5) + 4(4.2)$$
$$= \$27.3$$

Variance of a linear combination

Theorem: Variance of a linear combination of random variables

If X and Y are **independent** random variables and a and b are constants, then

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

$$Var(aX + b) = a^2 Var(X)$$

$$var(b) = 0$$

Keep keep rolling dice!

rolls are independent!

Example: Expected money for rolling 3 dice

Let the random variables X_1, X_2, X_3 be the values shown on rolls for 2 fair 6-sided dice and 1 unfair die (as described in our previous example). 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 unfair die roll. What are the variance and standard deviation of the amount you get from the 3 rolls?

$$\mathrm{Var}(X_1 + 2X_2 + 4X_3)$$

$$= \mathrm{Var}(X_1) + 2^2\,\mathrm{Var}(X_2) + 4^2\,\underline{\mathrm{Var}(X_3)}$$

at home: calculate X_3 variance

Learning Objectives

1. Define random variables and how they map to probability distributions
2. Calculate the expected value and variance of discrete random variables
3. Calculate the expected value and variance of linear combinations of discrete random variables
4. Calculate probabilities for different events using a Binomial distribution

Binomial random variable

- **One specific type of discrete random variable** is a binomial random variable

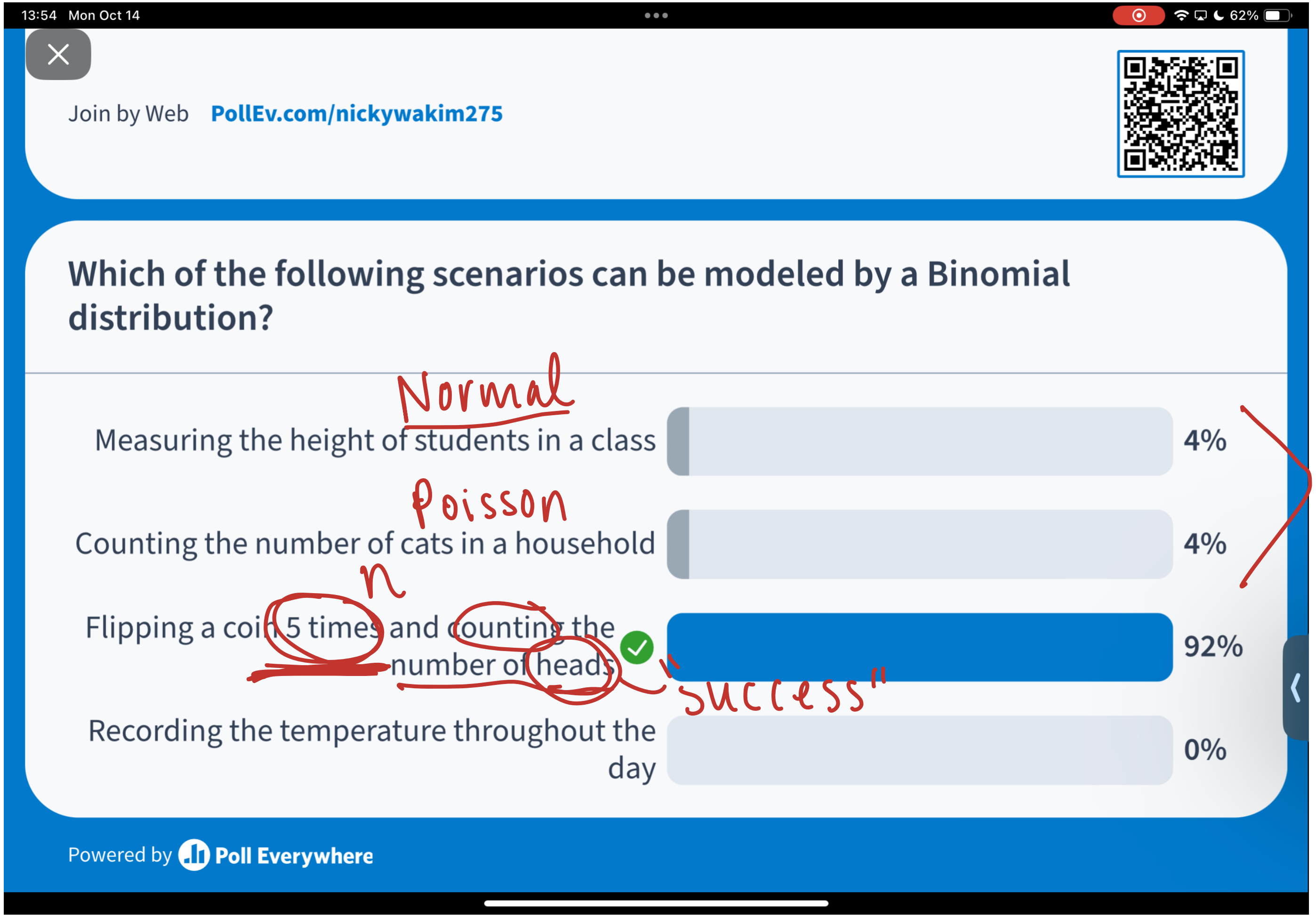
Binomial random variable

- X is a binomial random variable if it represents the number of successes in n independent replications (or trials) of an experiment where
  - Each replicate has two possible outcomes: either **success** or **failure**
  - The probability of success is p
  - The probability of failure is $q = 1 - p$

distributed as

- A binomial random variable takes on values $0, 1, 2, \ldots, n$.
- If a r.v. X is modeled by a Binomial distribution, then we write in shorthand $X \overset{\text{sim}}{\sim} \text{Binom}(n, p)$
- Quick example: The number of heads in 3 tosses of a fair coin is a binomial random variable with parameters $n = 3$ and $p = 0.5$.

Poll Everywhere Question 3



Bernoulli distribution

- When $n = 1$, aka we have a single trial, we give a different name to the random variable: Bernoulli

Bernoulli random variable

Bernoulli random variable. If X is a random variable that takes value 1 with probability of success p and 0 with probability $1 - p$ (or q), then X is a Bernoulli random variable.

- We call the probability of success p the **parameter** of the Bernoulli distribution.
- If a r.v. X is modeled by a Bernoulli distribution, then we write in shorthand $X \sim \text{Bernoulli}(p)$ or $X \sim \text{Bern}(p)$

$X \sim \text{Binom}(1, p)$

Mean and SD of a Bernoulli r.v.

If* X is a Bernoulli r.v. with probability of success p, then $E(X) = p$ and $\text{Var}(X) = p(1 - p)$

$E(X) = 0(1-p) + 1(p)$

$= p$

Relationship between Bernoulli and Binomial[1]

- The **Bernoulli distribution** is a special case of the Binomial distribution where $n = 1$
  - Specifically:

$$\text{Binomial}(1, p) = \text{Bernoulli}(p)$$

- To get a **Binomial distribution**, we simply extend the scenario from a **single** trial to **multiple** independent trials.
  - If we conduct n independent Bernoulli trials with the same success probability p, the total number of successes across these n trials will follow a Binomial distribution

- Quick example:
  - **Bernoulli**: If you flip a coin once, with probability $p = 0.5$ of landing heads, that is a Bernoulli trial.
  - **Binomial**: If you flip the coin 5 times, and you want to know how many times it will land heads, the number of heads will follow a Binomial distribution with parameters $n = 5$ and $p = 0.5$

Binomial distribution

H/T H/T H/T H/T H/T

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

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

n choose x $= \binom{n}{x} = \dfrac{n!}{x!(n-x)!}$

- 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

Mean and variance of a Binomial r.v

If X is a binomial r.v. with probability of success p, then $E(X) = np$ and $\mathrm{Var}(X) = np(1-p)$

Binomial distribution: R commands

R commands with their input and output:

R code	What does it return?
`rbinom()`	returns sample of random variables with specified binomial distribution
`dbinom()`	returns probability of getting certain number of successes
`pbinom()`	returns cumulative probability of getting certain number or less successes
`qbinom()`	returns number of successes corresponding to desired quantile

n, p

$P(X = x)$

$P(X \le x)$

90th

Binomial distribution example (1/5)

Vaccinated people testing positive for Covid-19

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 among the 10 that tested positive.

1. What is the expected value of X?
2. What is the SD of X?
3. What is the probability that exactly 4 of the 10 people that tested positive are vaccinated?
4. What is the probability that at most 3 of the 10 people that tested positive are vaccinated?
5. What is the probability that at least 5 of the 10 people that tested positive are vaccinated?

w/in positive test

$n = 10$

$p = 0.25$

X = # vaccinated

$X \sim \text{Binom}(\underset{(n)}{10}, \underset{(p)}{0.25})$

Binomial distribution example (2/5)

Vaccinated people testing positive for Covid-19

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 among the 10 that tested positive.

1. What is the expected value of X?

2. What is the SD of X?

① $E(X) = np = 10 \cdot 0.25 = 2.5$

Expected value of ppl vaxx is 2.5

② $Var(X) = np(1-p) = 10 \cdot 0.25(1-0.25) = 1.875$

$SD(X) = \sqrt{Var(X)} = \sqrt{1.875} = 1.369$

Binomial distribution example (3/5)

Vaccinated people testing positive for Covid-19

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 among the 10 that tested positive.

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

10 choose 4

$$P(X=4) = \binom{10}{4} 0.25^2 (1-0.25)^{10-4} = 0.146$$

```r
dbinom(x = 4, size = 10, prob = 0.25) # d for distribution
```

```
[1] 0.145998
```

- In general, for $P(X=k)$ we code: `dbinom(x = k, size = n, prob = p)`

4   # trials 10   p → 0.25

the probability that 4 are vaxx is 0.146.

Binomial distribution example (4/5)

Vaccinated people testing positive for Covid-19

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 among the 10 that tested positive.

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

$$
\begin{aligned}
P(X \leq 3) =& P(X=0) + P(X=1) + P(X=2) + P(X=3) \\
=& \binom{10}{0} 0.25^0 (0.75)^{10} + \binom{10}{1} 0.25^1 (0.75)^9 + \binom{10}{2} 0.25^2 (0.75)^8 + \binom{10}{3} 0.25^3 (0.75)^7 \\
=& 0.7758
\end{aligned}
$$

TRUE

```r
pbinom(q = 3, size = 10, prob = 0.25, lower.tail = T)
```

```
[1] 0.7758751
```

> lower.tail = FALSE

- In general, for $P(X \leq k)$ we code: `pbinom(q = k, size = n, prob = p)` with `lower.tail = T` as a default option

$\leq$

Binomial distribution example (5/5)

Vaccinated people testing positive for Covid-19

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 among the 10 that tested positive.

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

$$
\begin{aligned}
P(X \geq 5) &= P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) \\
&= \binom{10}{5} 0.25^5 (0.75)^5 + \binom{10}{6} 0.25^6 (0.75)^4 + \ldots + \binom{10}{10} 0.25^1 0 (0.75)^0 \\
&= \cancel{0.7759} \; 0.078
\end{aligned}
$$

```
pbinom(q = 4, size = 10, prob = 0.25, lower.tail = F) # switch to greater than!
```

```
[1] 0.07812691
```

$P(X > 4) = P(X \geq 5)$

```
1 - pbinom(q = 4, size = 10, prob = 0.25, lower.tail = T)
```

```
[1] 0.07812691
```

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

- In general, for $P(X > k)$ we code: `pbinom(q = k, size = n, prob = p, lower.tail = F)`

Resources!

- Seeing theory: random variables