Probability, Random Variables and Probability Distributions

Stat 120

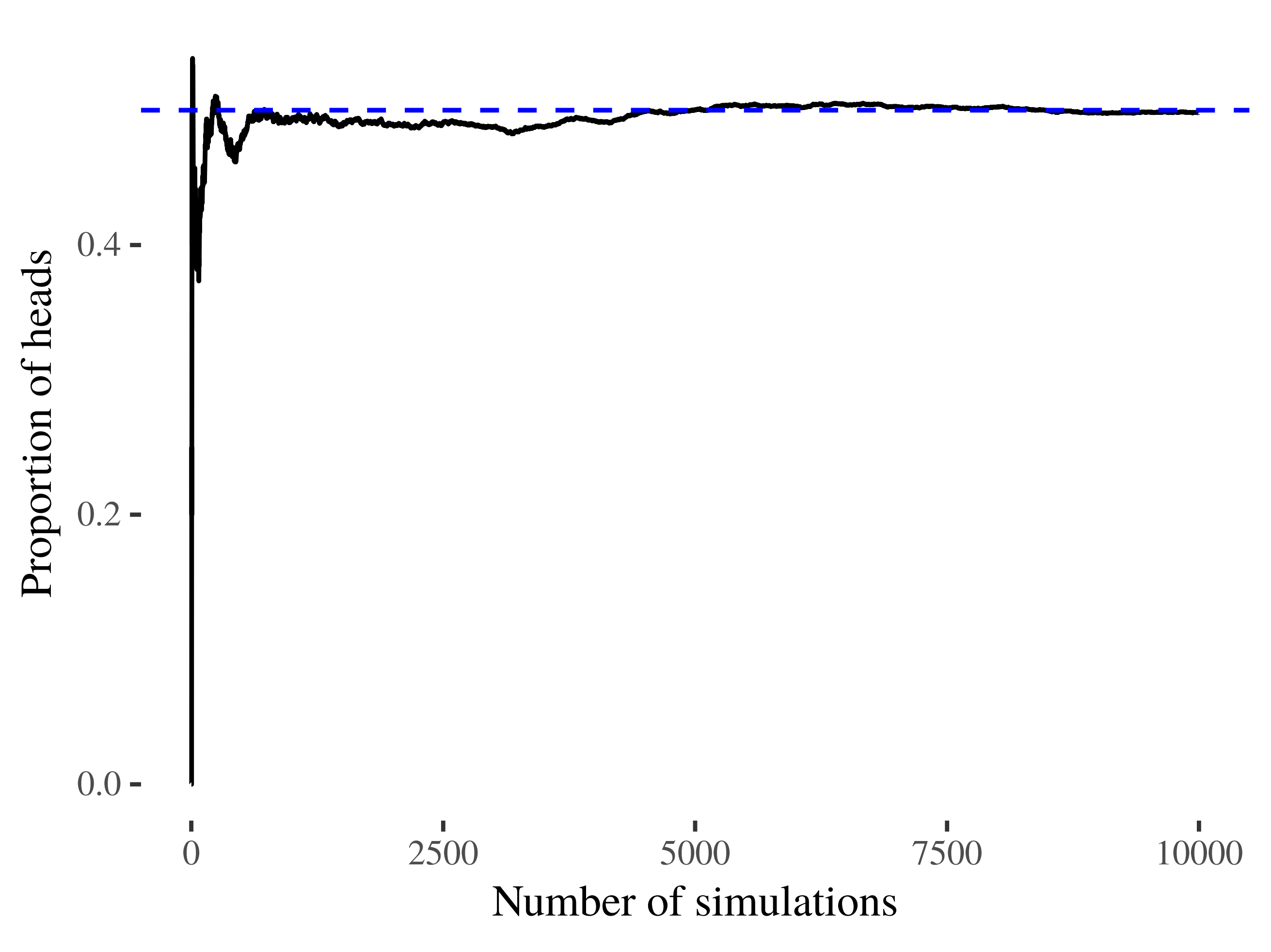
May 30 2022

Law of large numbers (Head = 1, Tail = 0)

```r
set.seed(123)  # for reproducibility
n <- 10000  # total simulations
x <- sample(c(0,1), n, replace = TRUE)
s <- cumsum(x)  # cumulative/running sum
p.hat <- s/(1:n)  # prop. heads in N simulations
results <- data.frame(x = x,
                        s = s,
                        p.hat = p.hat)

results <- results %>%
  mutate(n = row_number())

ggplot(results, aes(x = n, y = p.hat)) +
  geom_line() +
  geom_hline(yintercept = 0.5,
             col = "blue",
             linetype = "dashed")+
  labs(x = "Number of simulations",
       y = "Proportion of heads")
```



Law of large numbers

As more observations are collected, the proportion $\hat{p}_n$ of occurrences with a particular outcome converges to the probability p of that outcome.

What are random variables?

x	0	1	2
$P(X=x)$	0.25	0.50	0.25

Random Variable (RV)

- a variable whose value is a numerical outcome of a random process.
- **Notation:** $P(X=x)$ means the probability that RV X equals the number given by x.

Example: flip a coin twice

- $X = \#$ of Heads observed

$$P(X=0) = P(TT) = \frac{1}{4}$$

$$P(X=1) = P(HT \text{ or } TH) = \frac{2}{4}$$

$$P(X=2) = P(HH) = \frac{1}{4}$$

Discrete random variables

X is a discrete RV if you can list its possible values

Describe a distribution of a discrete RV with

- **Shape** : plot x-values vs. $P(X = x)$
- **Expected Value or Mean of X** :

$$E(X) = \mu_X = \sum_{\substack{\text{allvalues} \\ \text{of}}} xP(X = x)$$

- **Standard deviation and Variance of X** :

$$SD(X) = \sigma_X = \sqrt{\mathrm{Var}(X)}$$
$$\mathrm{Var}(X) = \sigma_X^2 = \sum_{\substack{\text{allvalues} \\ \text{of X}}} (X - \mu_X)^2 P(X = x)$$

Recall: Sample proportions

The **sample proportion** is

$$\hat{p} = \frac{X}{n}$$

The sample proportion is a **Random Variable!**

The **mean** and **SD** of this sample proportion are:

$$E(\hat{p}) = p$$

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Example: Blood testing

Context: You want to find a Type B blood donor, but you only have enough money to test 4 people.

- 11% of the population are Type B
- What is the **probability distribution** for the random variable?

Example: Blood testing

Y = number of people tested until you find Type B donor or run out of money

y	1	2	3	4
$P(Y = y)$	0.11	0.0979	0.0871	0.7050

$$
\begin{aligned}
P(Y = 1) &= 0.11, \\
P(Y = 2) &= (.89)(.11), \\
P(Y = 3) &= (.89)^2(.11), \\
P(Y = 4) &= (.89)^3(.11) + (.89)^4
\end{aligned}
$$

A special discrete model: The Binomial model: $Binom(n, p)$

- $X =$ number of "success" in n independent trials
- $p = P$(success) for each trial

Probability model:

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

The term $\binom{n}{x}$ (read "n choose x") counts the number of ways that we can see x successes and $\mathbf{n} - \mathbf{x}$ failures:

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

Expectation and SD:

$$\mu = E(X) = np \qquad \sigma = \mathrm{SD}(X) = \sqrt{np(1-p)}$$

Is it Binomial?

Check the following four conditions:

(1) The trials are **independent**.

(2) The number of trials, n, is **fixed**.

(3) Each trial outcome can be classified as a **success or failure**.

(4) The **probability of a success**, p, is the **same** for each trial.

Binomial or Not

(1) Count the number of heads in 2 flips of a coin

(2) Two baseball teams play a series of games until one of them wins a total of four games. You count the total number of games played.

(3) You play ten games of solitaire and count how many times you win.

(4) You collect a sample of 50 M&M candies and count the number of green ones

▶ Click for answer

Reese's Pieces

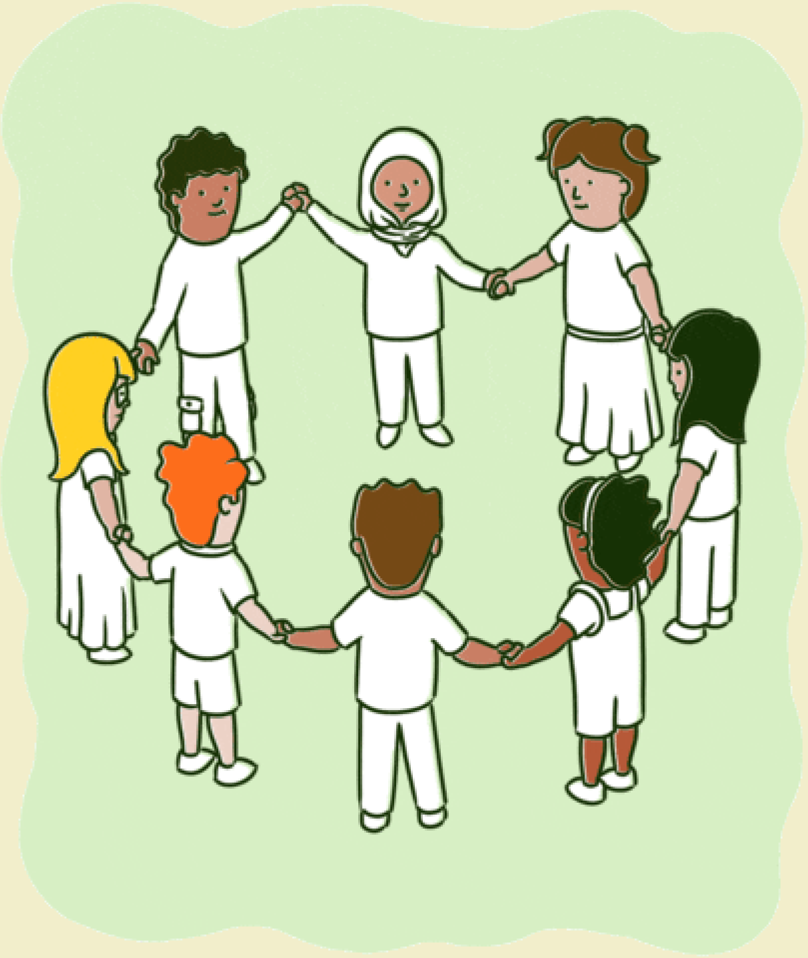
- Reese's Pieces candies have three colors: **orange, brown, and yellow**.
- Which color do you think has more candies in a package: **orange, brown or yellow**?

Suppose you wanted to draw 26 Reeses Pieces from a jar with 52% of the candies being orange.

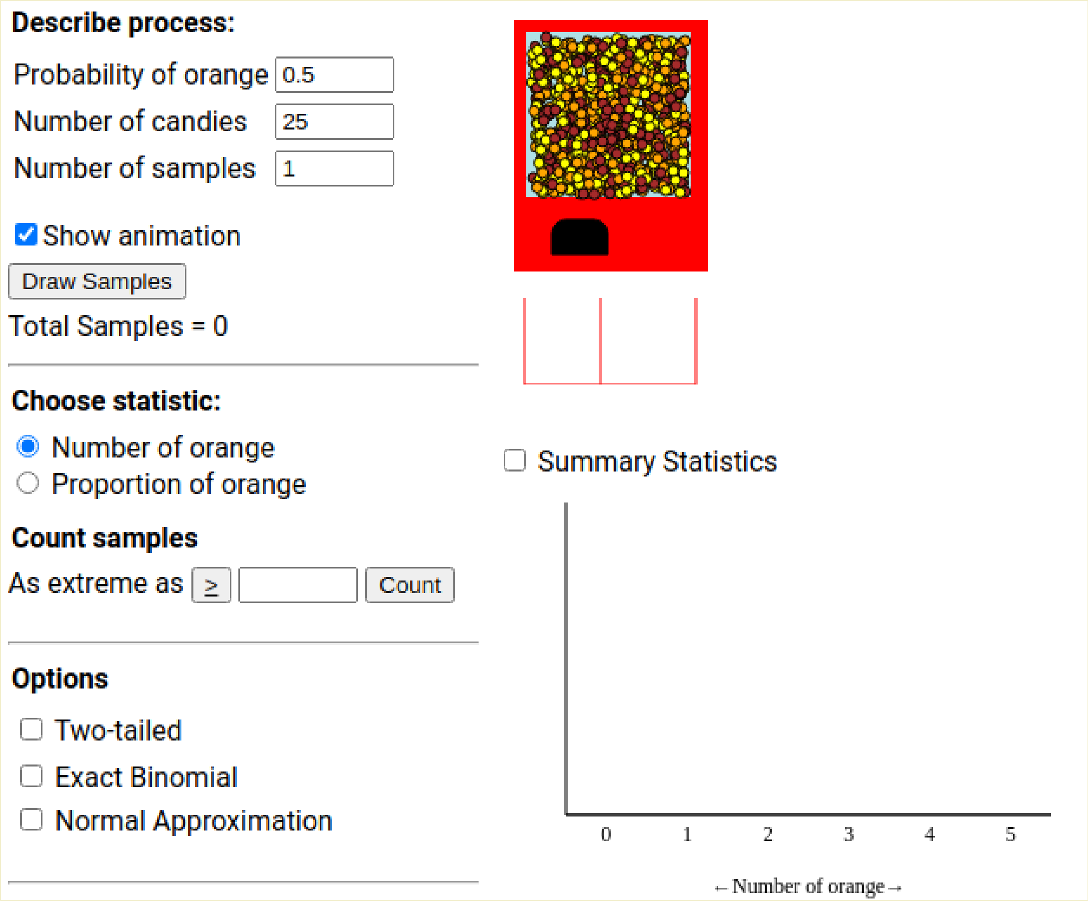
Let's go to a web applet and simulate the distribution of the proportion of orange candies!!

Your Turn 1

05:00



Click on the link below!



http://www.rossmanchance.com/applets/2021/oneprop/OneProp.htm?candy=1

Another example: Blood donor

Previously, $\mathbf{Y} =$ number of people tested until you find Type B donor or run out of money

- Does Y have a **binomial distribution**?

- No, you are counting something (# Type B) but you don't have a fixed number of trials (sample size)

Now, you are going to check the blood types of 4 people. Define the random variable: $X =$ the **number of people in your sample with Type B blood.**

- Does X have a **binomial distribution**?

- Yes, you are counting successes (Type B) with n=4 people sampled (trials), each with an p=11% of chance of success (Type B)

Blood donor

$$X \sim Binom(n = 4, p = 0.11)$$

$$P(X = x) = \binom{4}{x} 0.11^x (1 - 0.11)^{4-x}$$

x	0	1	2	3	4
$P(X = x)$	0.6274	0.3102	0.0575	0.0047	0.0002

$$\mu = E(X) = 4(0.11) = 0.44$$

$$\sigma = SD(X) = \sqrt{4(0.11)(1 - 0.11)} = 0.626$$

$$P(X = 0) = \binom{4}{0} 0.11^0 (1 - 0.11)^{4-0}$$
$$= 0.6274$$

$$P(X = 1) = \binom{4}{1} 0.11^1 (1 - 0.11)^{4-1}$$
$$= 0.3102$$

Linear Combinations of RVs

- Any function of a RV is itself a RV.
- Let X be a RV and a and b be constants.

$$E(aX \pm b) = aE(X) \pm b$$

$$V(aX \pm b) = a^2V(X) \quad SD(aX \pm b) = a \cdot SD(X)$$

- Let X and Y be any two RVs

$$E(X \pm Y) = E(X) \pm E(Y)$$

- Let X and Y be any two RVs that are independent of each other

$$V(X \pm Y) = V(X) + V(Y) \quad SD(X \pm Y) = \sqrt{V(X) + V(Y)}$$

Back to sample proportions

We can write the **sample proportion** as a function of a Binomial random variable X :

$$\hat{p} = \frac{X}{n}$$

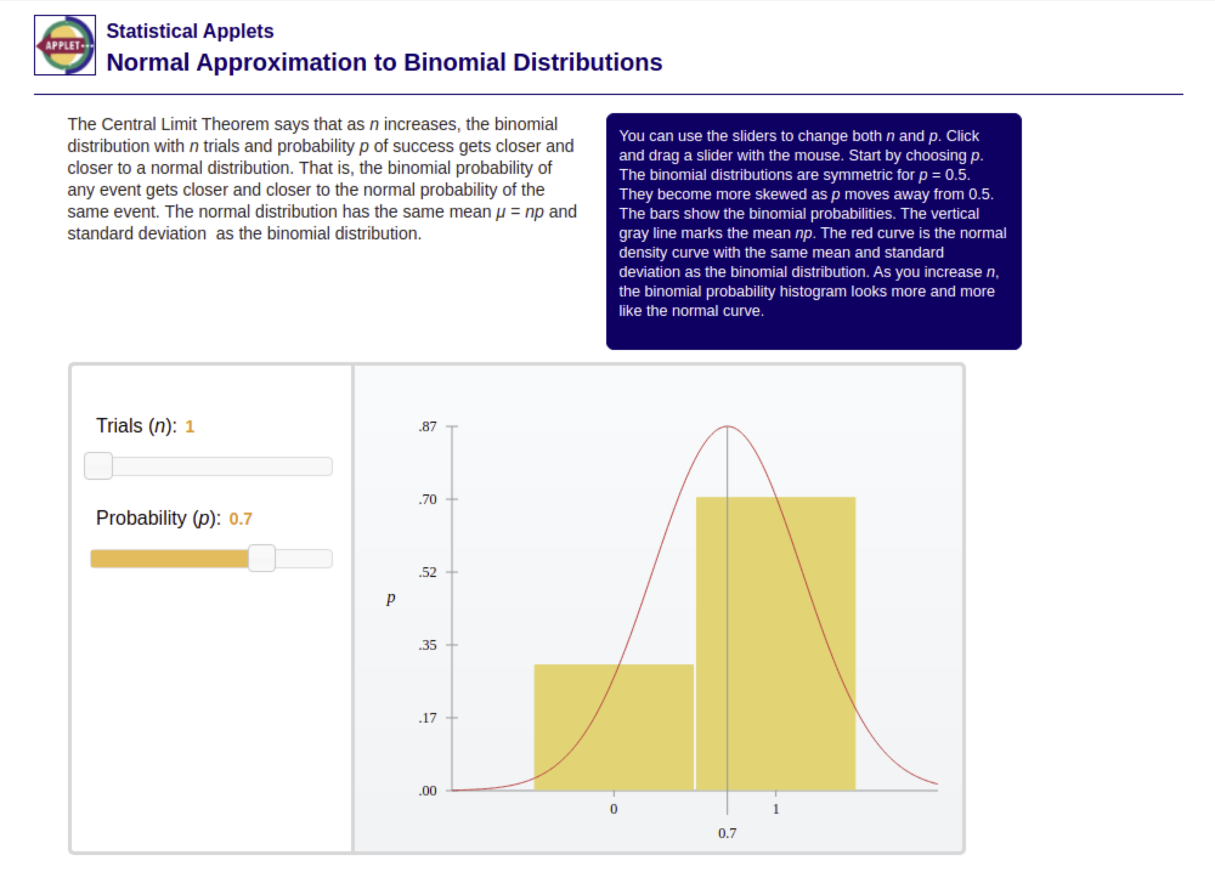
We learned earlier this term that a sample proportion behaves like a **normal distribution** when n is large (CLT).

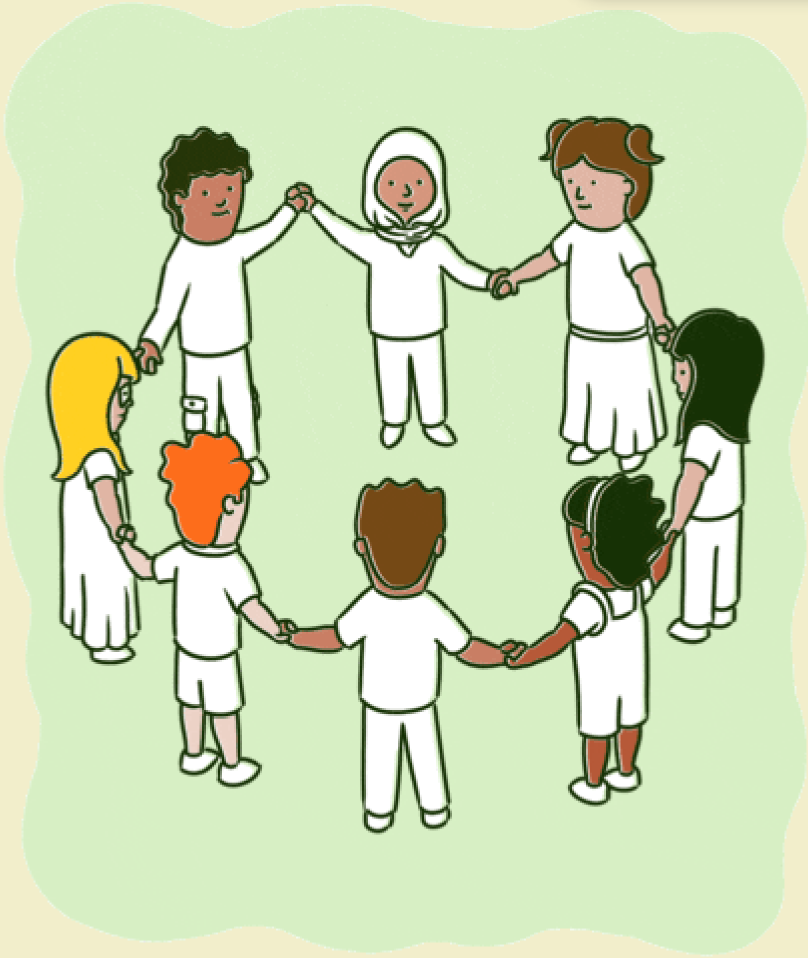
So... how are the binomial and normal distributions connected??

Your Turn 2

03:00

Click on the link below!





https://digitalfirst.bfwpub.com/stats_applet/stats_applet_2_cltbinom.html

Normal approximation for a Binomial RV

When n is large, a **Binomial** RV X can be modeled, approximately, by a **Normal** model with mean and SD

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

What is "large n"?

- expect **at least** 10 successes **and** at least 10 failures: $np \geq 10$ (expected successes) $n(1-p) \geq 10$ (expected failures)

Where might you see Binomial RVs again?

In **regression models** when your response variable is **categorical**, or a binomial count!

- **Logistic regression** models assumes that Y = response = **binomial** RV

$$p(X) = P(Success \mid X) = \text{ function of } X \text{ (explanatory vars)}$$

Example: What factors are related to success on the MN Comprehensive Assessment (MCA) reading test?

- Case = student
- Y = pass (1) or fail $(0) \sim \text{Binom}(n = 1, p)$
- $p(X) = P($ a student passes $|X)$ = function of earlier reading assessments (grades)

Where might you see Binomial RVs again?

Example: What factors are related to species extinction?

- Case = island
- $\mathrm{n} = \#$ animal species on island at the start of the study
- $Y = \#$ animals gone extinct over decade

$$Y \sim \mathrm{Binom}(n, p)$$

- $p(x) =$P(an animal goes extinct $|x) =$ function of island size, human population size, ...

Logistic regression: models the **log odds** of success as a linear function of X's :

$$\text{odds of success } = \tfrac{p}{1-p}$$
$$\log\left(\tfrac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Continuous random variables

X is a **continuous** RV if it takes on values in some interval of numbers

- E.g. a **random** number between 0 and 1
- E.g. a **Normal Random Variable** with mean μ_X and SD σ_X

Other **continuous distributions** you've seen this term

- t-distribution
- chi-square distribution
- F-distribution