Contents

Section 1: R Commands

You are not expected to have all of these commands memorized, but you are expected to be able to know where to look. This chapter is a reference guide on how to do various tasks in R that we will encounter throughout the semester. This is not an exhaustive list, and does not supplement the online course notes for this chapter.

A Base R cheat sheet will be handed out in the first week. Insert this page into your course notes after this chapter.

Basics

Basic Math

```
3*pi^2 + 4*(3-log(5))^2
```

```
## [1] 37.34346
```

Combinatorics

```
factorial(5) # 5! = 5*4*3*2*!
```

```
## [1] 120
```

```
choose(5,3) # 5 choose 3
```

```
## [1] 10
```

Store a value into an object using the assignment operator `<-`

```
height <- 62
```

Print the result of a line of code

```
(height <- 62)
```

```
## [1] 62
```

Do math on objects

```
apples <- 5
oranges <- 4
(fruit <- apples + oranges)
```

```
## [1] 9
```

Vectors

Combine multiple numbers into a single vector object using the `c` operator.

```
(primes <- c(2,3,5,7,11,13,17,19,23,29))
```

```
##  [1]  2  3  5  7 11 13 17 19 23 29
```

Vector of the numbers 1 through 10

```
(first.ten <- 1:10)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

`seq(a, b, by=x)`: Create a sequence of numbers from `a`, to `b`, counting by `x`.

```
(odds <- seq(1, 10, by=2))
```

```
## [1] 1 3 5 7 9
```

`rep`: Repeat a sequence of numbers in varying patterns. See `? rep` for more info on the options.

```
rep(c(2,3), times=c(4,3))
```

```
## [1] 2 2 2 2 3 3 3
```

```
rep(c(2,3), each=2)
```

```
## [1] 2 2 3 3
```

```
rep(c(2, 3), length.out = 3)
```

```
## [1] 2 3 2
```

Character vectors

```
(letterz <- c("A", "b", "C", "d"))
```

```
## [1] "A" "b" "C" "d"
```

Indexing vectors

Extract values from a specific position in a vector.

Returns first and second number from `primes` and `letterz`

```
primes[1]
```

```
## [1] 2
```

```
letterz[2]
```

```
## [1] "b"
```

Returns first three numbers of a vector

```
primes[1:3]
```

```
## [1] 2 3 5
```

Operations on vectors

Element wise math operations

```
primes * first.ten # multiplication
```

```
##  [1]   2   6  15  28  55  78 119 152 207 290
```

```
primes - first.ten # subtraction
```

```
##  [1]  1  1  2  3  6  7 10 11 14 19
```

```
abs(primes - first.ten) # absolute value
```

```
##  [1]  1  1  2  3  6  7 10 11 14 19
```

```
pmin(primes, first.ten) # minimum value between each pairwise element. Also `pmax`
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

Functions that apply to the entire vector

```
sum(primes) # add up all values in the vector
```

```
## [1] 129
```

```
mean(primes) # average value in the vector
```

```
## [1] 12.9
```

```
min(primes) # min value within the vector. Also `max`
```

```
## [1] 2
```

Functions that tell you characteristics about the vector

```
(a <- rep(c(2,3), times=c(4,3)))
```

```
## [1] 2 2 2 2 3 3 3
```

```
length(a) # how many elements are in the vector
```

```
## [1] 7
```

```
unique(a) # which elements are unique? (remove duplicates)
```

```
## [1] 2 3
```

Boolean Operators

The standard comparison operators that will return either `TRUE` or `FALSE` are `=`, `!=`, `>`, `<`, `>=`, and `<=`.

Change a vector to TRUE and FALSE by writing a logical statement.

```
primes>6
```

```
##  [1] FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

Identify the numbers in primes that are greater than 6

```
primes[primes>6]
```

```
## [1]  7 11 13 17 19 23 29
```

Check if a value is contained inside a vector is using the `%in%` operator.

```
4 %in% primes
```

```
## [1] FALSE
```

AND and OR

Is 9 an odd prime?

```
(9 %in% odds) & (9 %in% primes)
```

```
## [1] FALSE
```

Is 9 an odd or a prime?

```
(9 %in% odds) | (9 %in% primes)
```

```
## [1] TRUE
```

Math on Boolean values

In the programming world, Boolean values resolve as 0 for FALSE and 1 for TRUE. This means we can do math on Boolean vectors. E.g. If we use the command `sum` then R automatically turns the vector into a numeric vector of 0s and 1s and then calculates the sum of the vector, which corresponds to the number of 1's.

The arithmetic average is calculated as $\frac{\sum(x)}{n}$. If x is a binary 0/1 data type, then the average is equivalent to the proportion of 1s.

Count the number of primes greater than 6

```
sum(primes>6)
```

```
## [1] 7
```

What proportion of primes are greater than 5?

```
mean(primes>5)
```

```
## [1] 0.7
```

Frequency and Proportion tables

Frequency table

```
get.numbers <- sample(1:10, 1000, replace=TRUE) # generate fake data
table(get.numbers) # create the table
```

```
## get.numbers
##   1   2   3   4   5   6   7   8   9  10 
## 112  97  95  93 104  92  94  97 103 113 
```

Proportions

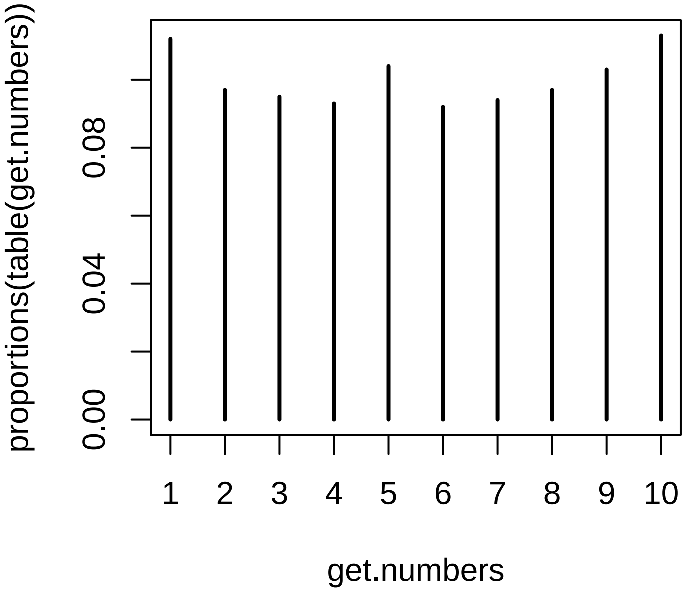
```
proportions(table(get.numbers))
```

```
## get.numbers
##     1     2     3     4     5     6     7     8     9    10 
## 0.112 0.097 0.095 0.093 0.104 0.092 0.094 0.097 0.103 0.113 
```

Plots

Discrete probability distribution: plot the table of proportions.

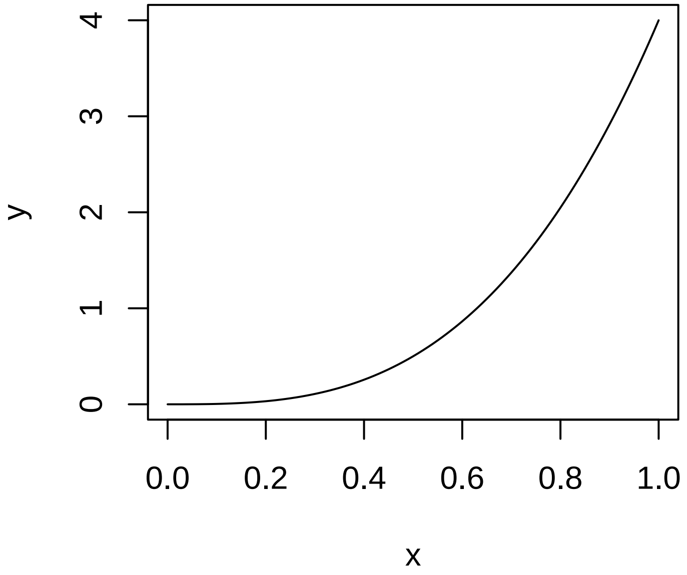
```
plot(proportions(table(get.numbers)))
```



Continuous probability distributions: we can plot functions directly, or create histograms and density curves from simulated values.

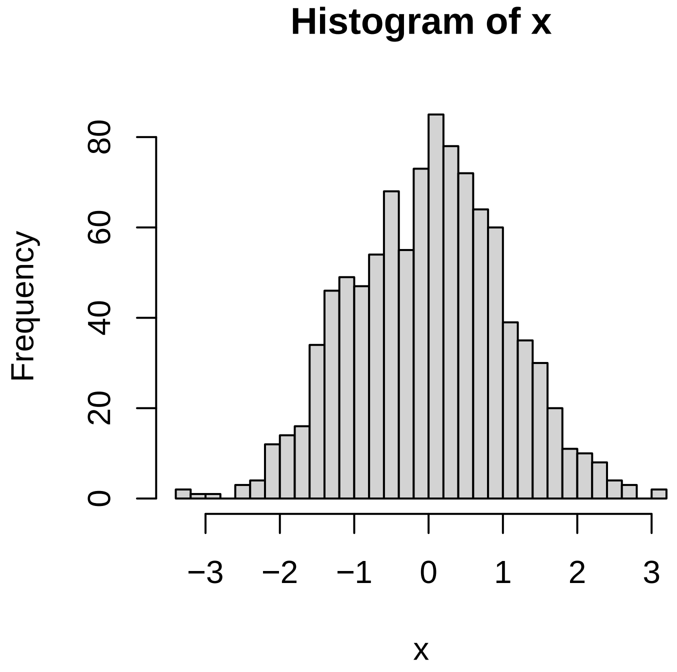
- Direct plotting of known functions

```
x <- seq(0,1, by=0.01) # create values in the domain
y <- 4*x^{3} # pdf
plot(x,y, type = 'l') # the lower case 'l' draws a line
```



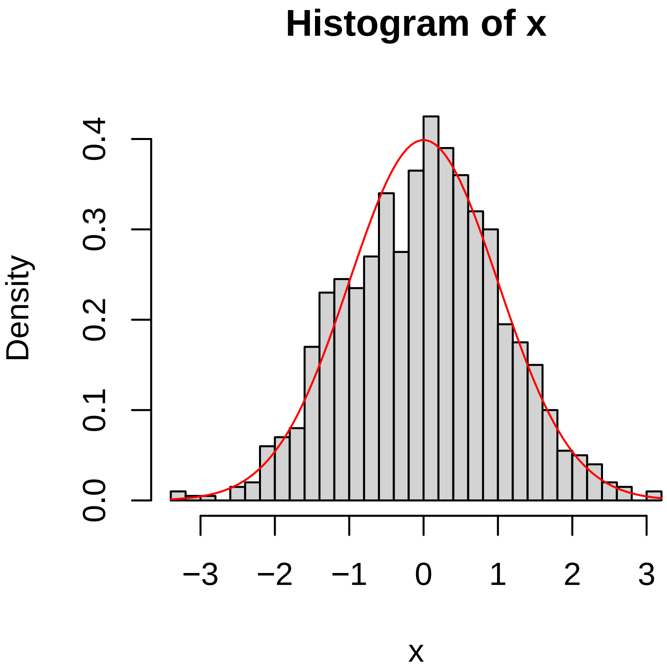
- simulating distributions

```
x <- rnorm(1000) # draw 1000 values from a standard normal distribution
hist(x, nclass=30) # create a frequency histogram with 30 bins
```



- Adding a known distributional curve over a histogram. Need to use `prob=TRUE` to change the y axis to a density so it's on the same scale as the curve.

```
hist(x, nclass=30, prob=TRUE)
curve(dnorm(x), add=TRUE, col="red") # note, this always stays x
```



Integration

To do finite integration you first define a function:

```
myfun <- function(x){x+5}
```

Then pass it to the `integrate` function.

```
(myint <- integrate(myfun, lower=0, upper=3))
```

```
## 19.5 with absolute error < 2.2e-13
```

The result of this integration can be accessed using `$value`

```
myint$value
```

```
## [1] 19.5
```

Simulation

Draw 10 samples from the numbers 1,2 or 3 with replacement.

```
sample(c(1,2,3), 10, replace=TRUE)
```

```
##  [1] 2 3 3 2 1 2 1 3 3 1
```

Conduct an experiment multiple times. Only the object last referenced will be saved out. E.g., `x` is not retained, only the value of `mean(x)`.

```
replicate(5, {
  x <- sample(c(1,2,3), 10, replace=TRUE)
  mean(x)
})
```

```
## [1] 1.9 1.9 1.8 2.0 1.8
```

Section 2.1: Probability Basics

A goal of Statistics is to describe the real world based on limited observations. Observations are influenced by random, and non-random conditions (e.g. the weather, what you ate for breakfast). Probability is a way to mathematically describe random events.

Vocabulary

- Experiment: Process that produces an observation
- Outcome: A possible observation
- Sample space: The set of All possible outcomes
- Event: Subsets of the sample space that describe a certain characteristic of the space
- Trial: a single running of The Experiment

Examples

1 - Roll a die and observe the number of dots on the face

2 - Stop a random person on the street and ask what month they are born

3 - Suppose a traffic light stays red for 90 seconds each cycle. When driving you arrive at the light and observe the amount of time that you are stopped until the light turns green.

① - Is an Experiment

Sample space

- Six possible outcomes

$S = \{1, 2, 3, 4, 5, 6\}$

- Event "Roll higher than a 3"

$E = \{4, 5, 6\}$

② 12 possible outcomes. = length of sample space

$S = \{$"Jan", "Feb", ..., "Nov", "Dec"$\}$

Event: "Born In summer"

③ Sample space Is A continuous Interval of Real #'s from $[0, 90]$

Event "You didn't have to stop"

$E = \{ 0 \}$

not $\emptyset$ empty set

You try it:

For each of the following problems, identify the sample space and the events described.

1. Observe eye color of a group of students.
   - Sample space:
   - Event student does not have blue eyes:
2. Number of credits a student can take:
   - Sample space:
   - Event student takes less than 9 credits:
3. Toss a coin and roll a die.
   - Sample space:
   - Event that you get tails:
4. A soccer team is in the playoffs. The team will play three games and will either win (`w`) or lose (`l`) each game (assume ties are not allowed).
   - Sample Space:
   - Event that at least 2 games are won:

Set Definitions

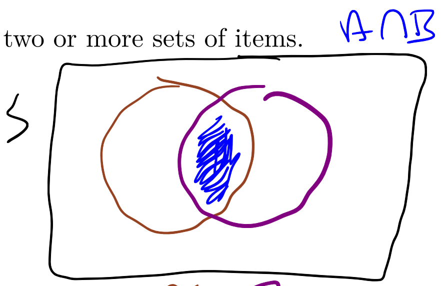
Let A and B be events in a sample space S. Complete the following definitions and write an example of each using context situations above.

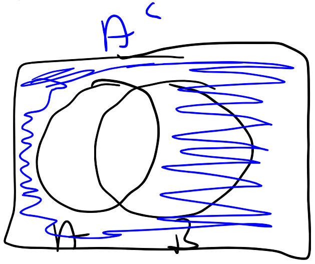
- $A \overset{\text{and}}{\cap} B$ set of outcomes that are In Both A & B at the same Time
- $A \underset{\text{or}}{\cup} B$ set of outcomes Either In A or B , or Both
- The *complement* of A is A^c (or $\bar{A}$) set of outcomes In S that are not In A.
- The symbol $\emptyset$ is The empty set, no set with outcomes
- A and B are **disjoint** or **mutually exclusive** if and only if $A \cap B = \phi$
- $A \cap B^c$ Elements in A and also not In B

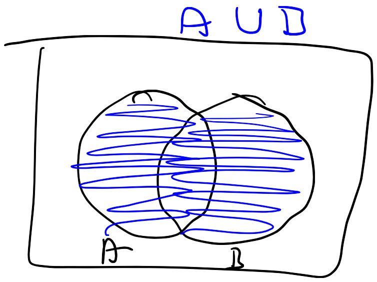
Note that element *and* outcome *can be used interchangeably.*

Venn Diagram

The most common kind of picture to make to describe sample spaces and events within sample spaces is a *Venn Diagram*. A Venn diagram uses overlapping circles or other shapes to illustrate the logical relationships between two or more sets of items.





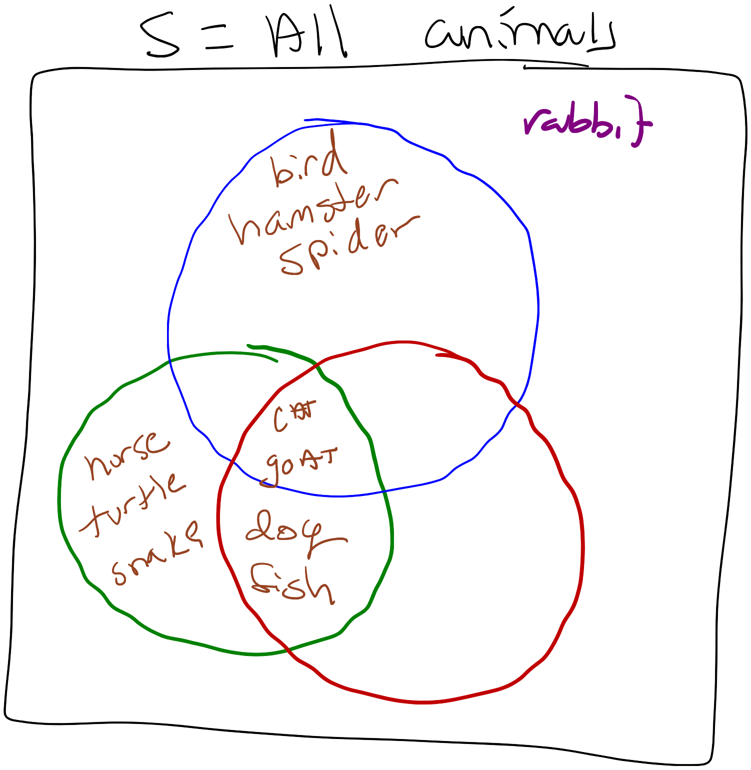


Example

Say 3 roommates are deciding on a pet. They use a Venn Diagram to determine which pet might be the best pick for them.

- Sidney prefers: cat, bird, hamster,spider, goat.
- Ralph prefers: dog, cat, fish,goat.
- Gilbert prefers: horse, cat, dog, turtle, snake,goat,fish

Create a Venn diagram that represents this example.



rabbit

What pet should they choose? CAT or GOAT

You try it:

A single card is drawn from a standard deck of cards. (Not sure what that looks like? See here: https://en.wikipedia.org/wiki/Standard_52-card_deck)

Let A be the event that an ace is selected, and let B be the event that a heart is drawn.

1. Define A and B using set notation. and Draw A venn Diagram

And shade where this Is At In The Venn Diagram

2. Write the event space, and what the following mean in context of a deck of cards.

- $A \cup B$

- $A \cap B$

- B^c

OPEN RSTUDIO ALSO

Set operations in R

We can also rely on R to perform union and intersection calculations. The following functions are used to compute intersections and unions. Each function can only take into consideration 2 vectors.

Base

- union: Either event, or both. (∪)
- intersect: Where do the two events overlap. (∩)
- setdiff: Where the two events do not overlap.

Example

Lets revisit the deck of cards problem from above: *A single card is drawn from a standard deck of cards. Let A be the event that an ace is selected, and let B be the event that a heart is drawn.*

First create the sample space and event vectors. I recommend that when you do this on your own you print the vector to ensure that what's being created is what is intended. Trust, but verify your code.

```
numbers <- rep(c(1:10, "J", "Q", "K", "A"), 4)
suits <- rep(c("H", "C", "D", "S"), each = 13)
deck <- paste0(numbers, suits) # Sample Space
aces <- c("AH", "AC", "AD", "AS") # Event A
hearts <- paste0(c(1:10, "J", "Q", "K", "A"), "H") # Event B
```

Then we can use R functions to find the following statements.

- $A \cup B$

```
(aces.and.hearts <- union(aces, hearts))
```

```
##  [1] "AH"  "AC"  "AD"  "AS"  "1H"  "2H"  "3H"  "4H"  "5H"  "6H"  "7H"  "8H"
## [13] "9H"  "10H" "JH"  "QH"  "KH"
```

- $A \cap B$

```
(ace.of.hearts <- intersect(aces, hearts))
```

```
## [1] "AH"
```

- B^c

```
(no.hearts <- setdiff(deck, hearts))
```

```
##  [1] "AC"  "1C"  "2C"  "3C"  "4C"  "5C"  "6C"  "7C"  "8C"  "9C"  "10C" "JC"
## [13] "QC"  "KD"  "AD"  "1D"  "2D"  "3D"  "4D"  "5D"  "6D"  "7D"  "8D"  "9D"
## [25] "10D" "JD"  "QS"  "KS"  "AS"  "1S"  "2S"  "3S"  "4S"  "5S"  "6S"  "7S"
## [37] "8S"  "9S"  "10S"
```

You try it:

Suppose that one card is to be selected from a deck of 20 cards that contains 10 red cards numbered from 1 to 10 and 10 blue cards numbered from 1 to 10. Let A be the event that a card with an even number is selected, let B be the event that a blue card is selected, and let C be the event that a card with a number less than 5 is selected.

Define the sample space and each event in R.

TACTIC: Define S, And each event A, B, C

Then use R to compute each of the following: 2 stages

- $(A \cap B) \cap C$

```
a.b <- intersect(A, B)
a.b.c <- intersect(a.b, C)
```

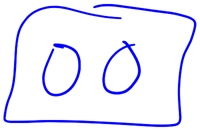
- $B \cup C^c$
- $A \cap (B \cup C)$.
- $A^c \cap B^c \cap C^c$

Definition of Probability

The probability of an event describes the proportion of time we expect the event to occur if we observed the event Generating process an infinite number of times.

Let S be a sample space. A valid *probability* of events A is a number $P(A)$ between 0 and 1 (inclusive), so $0 \leq P(A) \leq 1$, that satisfies the following *probability axioms*:

- The probability of the sample space $S = 1$
- Probabilities are countably additive. If A_1, A_2,...,A_n are disjoint then



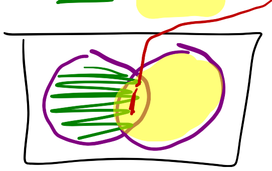
$$P(A_1 \cup A_2 \cup A_3 \dots A_n) = \sum_{i=1}^{n} P(A_i)$$

Probability rules

These are some important rules to memorize that come about as a result of the above axioms. Here are a few, there are more in the textbook.

Let A and B be events in the sample space S.

- The probability of ϕ is 0.
- If A and B are disjoint then $P(A \overset{\cup}{\cancel{\cap}} B) = P(A) + P(B)$.
- $P(A) = 1 - P(A^c)$.
- * $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Example

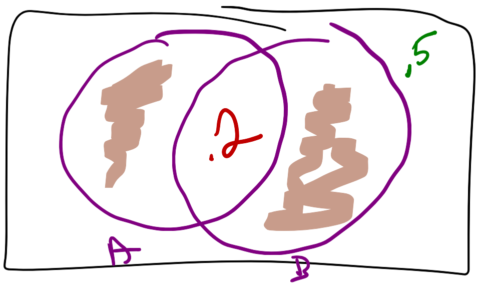
1. These rules allow you to manipulate equations to find unknown quantities based on known ones using algebra. Let's use these to show that $P\left(A \bigcap B\right) \geq 1 - P(A^C) - P(B^C)$ for any two events A and B defined on a sample space S.

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$
$$= 1 - P(A^c) + 1 - P(B^c) - P(A \cup B)$$
$$= \left[1 - P(A^c) - P(B^c)\right] + \underbrace{\left[1 - P(A \cup B)\right]}$$

Since $0 \leq P(A \cup B) \leq 1$ then $1 - P(A \cup B) \geq 0$

$$\therefore P(A \cap B) \geq \left[1 - P(A^c) - P(B^c)\right]$$

2. Events A and B are defined on a sample space S such that $P((A \cup B)^c) = 0.5$ and $P(A \cap B) = 0.2$. What is the probability that either A or B but not both will occur?



$$1 - .5 - .2 = .3$$

$$P\left[(A \cap B^c) \cup (A^c \cap B)\right]$$

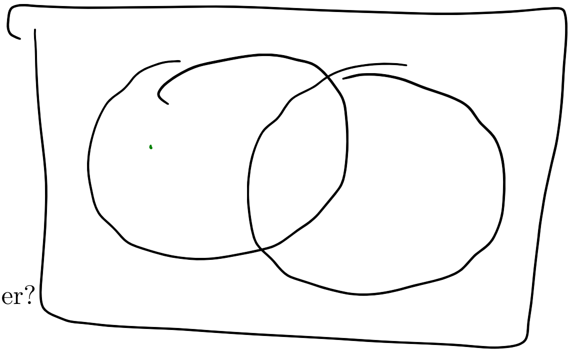
A = morning
B = Afternoon

$P(A) = .5$
$P(B) = .65$

You try it:

$P(A \cup B) = .85$

If 50 percent of the families in a certain city subscribe to the morning newspaper, 65 percent of the families subscribe to the afternoon newspaper, and 85 percent of the families subscribe to at least one of the two newspapers. Draw a Venn Diagram to represent this situation.



- What percentage of the families subscribe to both newspapers?
- What percentage of the families subscribe to only the afternoon paper?
- What percentage of the families don't subscribe to any paper?

Example

David Diez was interested in exploring the factors that contribute to an email being flagged as spam by Gmail's system. So they downloaded all their emails for a few months in 2012 and noted certain characteristics such as if it was flagged as spam (0 means no, and 1 means yes), and what size of a number it contained (none, small, or big). A *two-way table* of emails with these two characteristics are shown below.

```
     ##      Size of number
     ## Spam  none small  big  Sum
No   ##    0   400  2659  495 3554
Yes  ##    1   149   168   50  367
     ##   Sum  549  2827  545 3921 ← Number of emails
```

If you were to randomly select an email from this pool, calculate the following probabilities:

- It is flagged as spam $367/3921 = 0.09$
- It has a big number $545/3921 = 0.14$
- It is not flagged as spam and has a small number

$2659/3921 = 0.68$

You try it

The following data table describes the sex by species breakdown for 333 observed penguins on islands in the Palmer Archipelago, Antarctica.

```
##            Sex
## Species     female male Sum
##   Adelie        73   73 146
##   Chinstrap     34   34  68
##   Gentoo        58   61 119
##   Sum          165  168 333
```

If you were to select a penguin at random from these islands, what is the estimated probability that,

- the penguin is female

- the penguin is a Gentoo species

- the penguin is a male Chinstrap

Section 2.2: Simulation

A good way to start thinking about calculating probabilities is:

mean

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \quad x_i \begin{cases} \text{TRUE} & 1 \\ \text{FALSE} & 0 \end{cases}$$

$$\frac{\text{number of times an event can occurs}}{\text{size of sample space}}$$

There are two methods for calculating probabilities:

The *theoretical probability* can be solved mathematically or numerically. Sometimes the math needed to solve a problem is too complex, or intractable that solving it using tools such as algebra and calculus is impossible or relies on certain theories and understandings that you haven't encountered yet.

In this class we will explore probabilities both numerically (calculating theoretical probabilities), and also *estimating probabilities* using simulation. Simulation "simulates" a mathematical problem by using repeated sampling from a sample space and observing what occurs.

Example

Roll A Die & see what comes up

How would you find the probability of rolling a 4 on a six sided die?

Theoretical Probability: Count The Total possible ways The die could be 4 ÷ Total # of sides

* Count # of 4's over Total possible on 1 Roll

Estimated Probability:

Roll a Die a Bunch of times, Count the number of times a 4 Appears ÷ by The Total # of Rolls

Example using R

For Reference

Continuing with the deck of cards example, find the probability of selected events using R code.

How many Aces

- $P(A)$: `length(aces) / length(deck)`

Size of sample

```
## [1] 0.07142857
```

- $A \cup B$: `length(aces.and.hearts) / length(deck)`

```
## [1] 0.3035714
```

Simulations with `sample()`

The `R` function `sample` can be used to simulate sampling from a sample space. Think of it as putting names in a hat and individually drawing the names out of the hat. When you use the function `sample` you need to give a vector to sample from and also the sample size. The default is to sample **without replacement** with each item having equal probability of being selected.

What does sampling **without replacement** mean?

Once the outcome has been observed, It is Removed from the sample space before the next draw.

The same outcome can't be drawn more than once.

Example

Take a sample of 2 from the numbers 1 through 10 without replacement.

```
x <- 1:10   # Define the space
sample(x, size = 2) # Sample 2 items from the space
```

```
## [1] 3 2
```

You try it

Sample without replacement three months from the list of months

```
months <- c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug",
            "Sept","Oct","Nov","Dec")
sample(__________,size = _____)
```

The defaults can be changed. For instance, if you want to sample **with replacement** you would just add `replace=TRUE` to the command. This allows you to sample more values than the size of the sample space

Example

```
x <- 1:5
sample(x, size = 10, replace = TRUE)
```

```
##  [1] 5 1 4 4 2 2 5 5 1 3
```

You try it

Create a sample of 10 random days of the week.

```
days <- _______
sample(__________,size = _____, replace = _____)
```

Using simulation to compute probabilities

The goal of simulation is to compute probabilities of an event. We can do this in three steps:

1. Simulate an Experiment many times, storing the results in a vector of observations
2. Use a logical statement to test whether or not each element in the vector (observation) meets a criteria defined by the "event". Store these results In a new vector
3. Compute the proportion of TRUE's in the new vector to figure out the probability using the mean() function

Example

Suppose that two six-sided dice are rolled and the numbers appearing on the dice are added. Calculate the probabilities of the two events listed below using both theoretical methods and simulation.

- **Event D**: The sum of the two dice is 6.
- **Event E**: At least 1 die is a 2.

First we simulate the results from two die rolls, and the sample space that represents the sum of the two numbers added together.

1. Create two vectors, each length 10,000 by sampling the numbers from 1 to 6. These represent the rolls on each of two dice.

```r
die_1 <- sample(x=1:6, size=10000, replace=TRUE)
die_1[1:10] # look at what the first 10 rolls looks like for die 1
```

```
##  [1] 2 4 5 3 5 2 5 6 3 6
```

```r
die_2 <- sample(x=1:6, size=10000, replace=TRUE)
die_2[1:10] # look at what the first 10 rolls looks like for die 2
```

```
##  [1] 6 4 4 3 1 6 5 2 6 6
```

2. Add these two vectors of dice results together to create the sample space.

```
sum.of.2.dice <- die_1 + die_2
sum.of.2.dice[1:10] # confirm they add up as intended
```

```
##  [1]  8  8  9  6  6  8 10  8  9 12
```

Let's look at event D: *The sum of the two dice is 6.*

Theoretical Probability: Using the sample space defined in Example 2.8 in the textbook, there are 5 ways the sum of two dice equals 6, out of 36 total combinations. So $P(D) = 5/36 = 0.1389$

Simulation: Use a logical statement to identify if `sum.of.2.dice` is equal to 6, and compare the TRUE and FALSE results to the original values. (trust but verify)

```
D <- sum.of.2.dice == 6
D[1:10] #just checking
```

```
##  [1] FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
```

Now calculate $P(D)$ as:

```
sum(D)/length(D)
```

```
## [1] 0.138
```

```
mean(D) # exact same formula
```

```
## [1] 0.138
```

Let's look at event E: *At least 1 die is a 2.*

Theoretical: Again using the sample space diagrammed out in the textbook, there are 11 ways either die 1 or die 2 will roll a 2, and this is out of 36 total possibilities, so $P(E) = 11/36 = 0.306$

Simulation: Create our vector of TRUE and FALSES, and take the mean.

```
E <- die_1==2 | die_2==2
mean(E)
```

```
## [1] 0.3118
```

You try it

Roll three six-sided dice and add all the face up numbers. Use simulation to estimate the probability that the sum of the three dice is at least 10.

Using replicate to repeat experiments

What if we wanted to repeat an experiment several times? We can keep clicking the *run* button and record the results each time. However, the `replicate()` function will do this for us and much more efficiently. To use the function `replicate` we just wrap the simulation code in brackets and tell R how many times we want to repeat (or replicate) the experiment.

Example

Simulate rolling a dice 7 times and computing the sum of all rolls and recording if the sum is greater then 30.

```
dice <- sample(1:6, size=7,replace=TRUE)
sum(dice)>30
```

```
## [1] FALSE
```

Let's run this experiment 5 times.

```
replicate(n=5, {
  dice <- sample(1:6, size=7,replace=TRUE)
  sum(dice)>30
  })
```

```
## [1] FALSE FALSE FALSE FALSE FALSE
```

We will almost always want to replicate things a large number of times, say $n = 10000$. We then store the output in a vector.

```
results_dice <- replicate(n = 10000,{
  dice <- sample(1:6, size=7,replace=TRUE)
  sum(dice)>30
  })
```

Now, calculate the probability of rolling a die 7 times and getting a sum larger than 30.

```
mean(results_dice)
```

```
## [1] 0.0891
```

The following sequence is how you should approach writing code that uses the function `replicate`:

1.

2.

3.

4.

You try it

For both questions, compute the theoretical probability and then use simulation to confirm your results. Write all your work for both theoretical AND your simulation code in the space below.

1. If two die are rolled, what is the probability that the difference between the two numbers is less than 3?

2. A fair coin is repeatedly tossed ten times. Compute the probability that the last three coin tosses results in heads. (Hint, review Example 2.13 in the textbook for an example)

Additional notes.

Section 2.3: Conditional probability and independence

In this chapter we learn that we can update probabilities of an event happening if we know that certain events are observed. The updated probability of event A after we learn that event B has occurred is the **conditional probability** of A given B.

Example: Tulips

Suppose that we are given 20 tulip bulbs that are very similar in appearance and told that 8 tulips will bloom early, 12 will bloom late, 13 will be red, and 7 will be yellow. The following table summarizes information about the combination of features among these tulips:

	Early	Late	Sum
Red	5	8	13
Yellow	3	4	7
Sum	8	12	20

If one tulip bulb is selected at random, what is the probability that it will produce a red tulip?

Suppose that, under close examination, we know that it will be an early bulb. Given that it is an early bulb, what is the probability it is a red tulip?

Definition: Conditional Probability

Let A and B be events in the sample space S, with $P(B) \neq 0$. The **conditional probability of A given B** is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example

Suppose that $P(A) = .3, P(B) = .7$, and $P(A \cap B) = .2$. What is $P(A|B)$?

You try it

1. Suppose that $P\left(A\right)$=.7, $P\left(B\right) = .5$, and $P\left(A \bigcap B\right) = .2$. Find $P(A|B)$.

2. Find $P(A \cap B)$ if $P(A) = 0.2$, $P(B) = 0.4$, and $P(A|B) + P(B|A) = 0.75$.

Section 2.3.1 Independent Events

$P(A|B)$

In statistics we talk about independence a lot. If two events, A and B, are independent then knowing the outcome of B does not tell us any information about event A. Therefore, if A and B are independent events, then

$$\boxed{P(A|\underline{B}) = \underline{P(A)}} \quad \text{and} \quad \boxed{P(B|A) = P(B)}$$

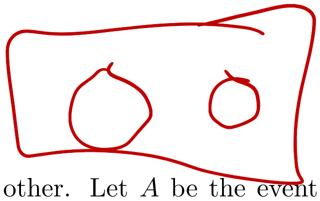
If learning the probability that B has occurred does not change the probability of A, then we say A and B are *independent.*

Give an example of two events that are independent.

$$\overbrace{P(A|B)}^{P(A)} = \frac{P(A\cap B)}{P(B)}$$

If two events A and B are independent then we can write $\boxed{P(A \cap B) \text{ as } P(A)P(B)}$

DISJOINT =0



Example: Machine failure

Suppose that two machines 1 and 2 in a factory are operated independently of each other. Let A be the event that machine 1 will become inoperative during a given 8-hour period; let B be the event that the machine 2 will become inoperative during the same period; and suppose that P(A)=1/3 and P(B)=1/4. We shall determine the probability that at least one of the machines will become inoperative during the given period.

Example: Graduation Requirements

School board officials are debating whether to require all high school seniors to take a proficiency exam before graduating. A student passing all three parts (math, language, and general) would be awarded a diploma; otherwise, they would receive only a certificate of attendance. A practice test given to this year's ninety-five hundred seniors resulted in the following failures: **Math: 3325; Language: 1900; General knowledge: 1425**

If "Student fails Math", "Student fails language", and "Student fails general knowledge" are independent events, what proportion of next year's seniors can be expected to fail to qualify for a diploma? Does independence seem reasonable here?

Example: Child Mortality

In a certain nation, statistics show that only two out of ten children born in the early 80s reached the age of 21. Assume the probability of child death is independent between children. If the same mortality rate is operative over the next generation, how many children does a person need to have if they wants to have at least a 75% probability that at least one child survives to adulthood?

You try it

Suppose that $P(A \bigcap B) = .2$, $P(A) = .6$, and $P(B) = .5$.

1. Are A and B mutually exclusive?

   no

2. Are A and B independent?

3. Find $P(A^C \cup B^C)$.

You try it:

Myra and Carlos are summer interns working as proofreaders for a local newspaper. Based on aptitude tests, Myra has a 50% chance of spotting a hyphenation error, while Carlos picks up on that same kind of mistake 80% of the time. Suppose the copy they are proofing contains a hyphenation error. What is the probability it goes undetected?

Simulating conditional probability

Simulating conditional probability is challenging. We will simulate the conditional probabilities by simulating $P(A \cap B)$ and either $P(A)$ or $P(B)$. We will then divide to get the conditional distribution of $P(B|A)$.

③ Define / Test events

Example ① Get dice ② roll dice

Two dice are rolled. Estimate the conditional probability that the sum of the dice is at most 4 given that at least one of the die is a 2. * Let A be that event that the sum of the dice is at most 4 and let B be the event that one of the die is a 2. Thus,we want $P(A|B)$

die <- 1:6

```
eventB <- replicate(10000,{
  dieroll <- sample(1:6,2,replace=TRUE)    or Sample (die, 2, replace=TRUE)
  # Define event B here
    2 %in% dieroll




  })

probB <- mean(eventB)

eventAB <- replicate(10000,{
  dieroll <- sample(1:6,2,replace=TRUE)
  # Define event A&B here





  })

probAB <-  ________________
(cond_prob <- _________/_________)
```

1. Create your sample space

2. Randomly draw from your sample space using sample()

3. Test the elements that came out of your sample against your event

4. Ensure that your final answer is a SINGLE TRUE or FALSE Boolean value.

2b) complete any remaining steps of the experiment as defined

Now, compute the theoretical probability. Does your calculation match what is given above?

Theorem 2.3 Law of Total Probability

Suppose that the events $A_1, A_2, \dots, A_k$ form a partition of the space S and $P(A_j) > 0$ for $j = 1, \dots, k$. Then, for every event B in S,

$$P(B) = \sum_{j=1}^{k} P(A_j)P(B|A_j)$$

Example: Voting preferences

The percentage of voters classified as Liberals in three different election districts are divided as follows: 21 % in the first district; 45% in the second district, and in the third district 75%. If a district is selected **at random** and a voter is selected at random from that district, what is the probability that she will be a Liberal?

You try it

In a certain study it was discovered that 15% of the participants were classified as *heavy smokers*, 30% as *light smokers*, and 55% as *nonsmokers*. In the five year study, 20% of the heavy smokers died, 10% of the light smokers died, and 4% of the nonsmokers died. What is the probability of death for this study?

Bayes' Rule and conditioning

Suppose that we are interested in which of several events $A_1, A_2, \dots, A_k$ will occur and that we will get to observe some other event B. If $P(B|A_i)$ is available for each i, then Bayes' theorem is a useful formula for computing the conditional probabilities of the A_i events given B. We will derive Bayes' Theorem in this section. Suppose that we have A_1, A_2, A_3 which form a partition of the sample space. There is another event we will call B that is contained in the same sample space.

Now suppose that we know the values of $P(A_j \cap B)$ and $P(A_j)$ for all j. We want to calculate $P(A_1|B)$. Given the formula we learned in this chapter for conditional probability, we can rewrite $P(A_1|B)$ as:

Remember, we don't know $P(B)$ but we do know $P(A_1 \cap B)$ and $P(A_1)$ so we can rewrite the denominator of the above formula giving us:

Now we know the denominator. Let's deal with the numerator. How can we rewrite the numerator so that we can use the information that we are given? Again, we can use the formulas for conditional probability. Thus, we have

and we can now calculate $P(A_1|B)$ because we know all the information on the right-hand side of the equation.

One can see from what we just did that Bayes' Rule is a simple statement about conditional probabilities. This simple rule forms the basis for Bayesian inference.

Theorem 2.4 Bayes' Rule

Of course, we can extend this rule for any number of A_i's.

Let $A_1, A_2, A_3, ..., A_k$ be a partition of the sample space S and let B be an event. Then

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}$$

Example: Disease test

Suppose that you are walking down the street and notice that the Department of Public Health is giving a free medical test for a certain disease. The test is 90 percent reliable in the following sense: If a person has the disease, there is a probability of .9 that the test will yield a positive response; whereas, if a person does not have the disease, there is a probability of .1 that the test will give a positive response.

Data indicate that your chances of having the disease are only 1 in 10,000. However, since the test costs you nothing, and is fast and harmless, you decide to stop and take the test. A few days later you learn that you had a positive response to the test. Now, what is the probability that you have the disease?

You try it

1. At a hospital's emergency room, patients are classified and 20% of them are critical, 30% are serious, and 50% are stable. Of the critical ones, 30% die; of the serious, 10% die; and the stable, 1% die. Given that a patient dies, what is the conditional probability that the patient was classified as critical.

2. In a certain city, 30% of the people are Conservatives, 50% are Liberals, and 20% are Independents. Records show that in a particular election, 65% of the Conservatives voted, 82% of the Liberals voted, and 50% of the Independents voted. If a person in the city is selected at random and it is learned that she did not vote in the last election, what is the probability that she is a liberal?

Additional notes.

Section 2.4: Counting Arguments

This section presents some common methods for counting the number of outcomes in a set. When there are a lot of outcomes in an experiment, it is convenient to have a method of determining how many outcomes there are in S.

Multiplication rule:

Suppose that an experiment has two parts or phases. In the first part there are n_1 outcomes and in the second part there are n_2 outcomes. The composite experiment which consists of both parts of the experiment then has $n_1 \times n_2$ possible outcomes.

Example: Despite all my rage...

Let $\boxed{E_1}$ denote the selection of a rat form a cage containing one female (F) rat and one male (M) rat. Let $\boxed{E_2}$ denote the administering of either drug A, drug B, or a placebo to the selected rat.

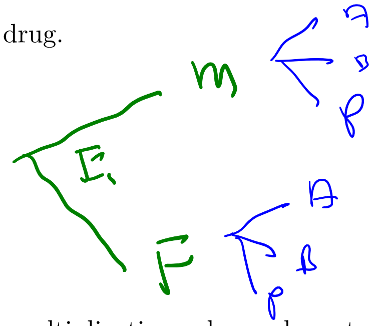
$n_1 = 2$

$n_2 = 3$

- How many possible outcomes are there? $2 \cdot 3 = 6$
- List the possible outcomes:

MA FA
MB FB
MP FP

Another way of illustrating the multiplication principle is with a *tree diagram*. The diagram shows that there are n_1=2 possibilities for the gender of the rat and that for each of these outcomes there are $n_2 = 3$ possibilities for the drug.



The multiplication rule can be extended to more than two experiments.

You try it

1. Each year starts on one of the $\overbrace{\text{seven days (Sunday through Saturday)}}^{E1}$. Each year is either $\overbrace{\text{a leap year (i.e., it}}^{E_2}$ includes February 29) or not. How many different calendars are possible for a year?

   $n_1 = 7$ $n_2 = 2$ $7 \cdot 2 = 14$

2. A chemical engineer wishes to observe the effects of temperature, pressure, and catalyst concentration on the yield resulting from a certain reaction. If she intends to include two different temperatures, three pressures, and two levels of catalyst, how many different runs must she make in order to observe each temperature-pressure catalyst combination exactly twice?

3. A restaurant offers a choice of four appetizers, fourteen entrees, six desserts, and five beverages. How many different meals are possible if a diner intends to order only three items, one from each menu? That is, you can't have two desserts and no entree. Similar-ish to SUNDAE

1^{st} choose what menu

2^{nd} choose Items from menu

Permutations

Sampling w/o Replacement

- An ordered arrangements of a countable set of objects is called a **permutation**.
- The number of permutations of n distinct objects is $n!$.
- The ! is a function called a **factorial** and is defined as

$$n! = n * (n-1) * (n-2) * \ldots * 1$$

R↓

```
factorial(3)
```

$3 \cdot 2 \cdot 1 = 6$

```
## [1] 6
```

Example

The "ice cream club" is hosting a make-your-own sundae at which the following are provided:

- Ice Cream flavors: Chocolate, Cookies-n-cream, Strawberry, Vanilla 4
- Toppings: Caramel, Hot Fudge, Marshmallow, M&Ms, Nuts, Strawberries 6

How many different sundaes are possible using one flavor of ice cream and three different toppings?

$$\underbrace{4}_{IC} \cdot (\underbrace{6 \cdot 5 \cdot 4}_{3\ \text{toppings}})$$

How many sundaes are possible using one flavor of ice cream and from $\boxed{0}$ to 6 toppings?

$$\uparrow 4 \Big(\overset{6\ \text{toppings}}{6!} \underset{\text{or}}{\overset{}{+}} \overset{5\ \text{toppings}}{\underset{\text{and}}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}} + 6 \cdot 5 \cdot 4 \cdot 3 + 6 \cdot 5 \cdot 4 + 6 \cdot 5 + 6 + \overset{0\ \text{toppings}}{1} \Big)$$

You try it

$4 \cdot 6! + 4(6 \cdot 5 \cdot 4 \cdot 3 \cdot 2) + \ldots\ldots + 4 \cdot 1$

There are 9 presidential candidates at a debate. How many different ways can candidates be lined up?

$9!$

factorial (9)

Combinations

$n \geq k$

If the order of objects is not important, then the number of ways of choosing k distinct objects from a set of n is given by

All possible orders

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

removing duplicate

Choose(n, k)

```
choose(3,2)
```

```
## [1] 3
```

$$\frac{3!}{2!\,(3-2)!} = \frac{3\cdot\cancel{2}\cdot\cancel{1}}{(\cancel{2}\cdot 1)(\cancel{1})} = 3$$

Example: $\alpha\beta\zeta$

The Alpha Beta Zeta sorority is trying to fill a pledge class of nine new members during fall rush. Among the twenty-five available candidates, fifteen have been judged marginally acceptable and ten highly desirable. How many ways can the pledge class be chosen to give a two-to-one ratio of highly desirable to marginally acceptable candidates?

Ignoring Groups

$n = 25$

$k = 9$

$\binom{25}{9}$

Choose$(25, 9)$

ways to choose 9 people out of 25

MA : 15 choose 3

HD : 10 choose 6

$$\binom{15}{3} \cdot \binom{10}{6} \qquad \begin{matrix} 25 \\ 9 \end{matrix}$$

Check does the top #'s add to n & do bottom #'s add to k

Choose$(15, 3)$ · Choose$(10, 6)$

You try it

For each of these, write the R code used to calculate the answer and the answer itself.

1. Among the 9 presidential candidates at a debate, 3 are republicans and 6 are democrats. How many different line ups are possible if the only ordering that matters is political party (not name)?

~~$3! \cdot 6!$~~ ·

b/c order doesn't matter

$$\binom{9}{3} \text{ or } \binom{9}{6} = 84$$

$$\frac{9!}{3!\,6!} = \frac{9 \cdot 8 \cdot 7}{3 \cdot 2 \cdot 1}$$

2. Nine students, five statistics majors and 4 computer science majors, interview for four summer internships sponsored by Google.

   a. In how many ways can Google choose a set of four interns?

   b. In how many can Google choose 2 stat majors and 2 computer science majors?

   c. How many sets of four can be picked such that not everyone in the set is the same major?

Combinatorial Probability

In the previous section our concern focused on counting the number of ways a given operation, or sequence of operations could be performed. In this section we want to calculate the probability that a certain combination will occur.

Example: Gender equality in promotions

Ten equally qualified marketing assistants are candidates for promotion to associate buyer; seven are men and three are women. If the company intends to promote four of the ten at random, what is the probability that exactly two of the four are women?

```
total     <-
two_women <-
(prob_two_women <-  )
```

Example: Urns and chips

An urn contains twenty chips, numbered 1 through 20. Two are drawn simultaneously. What is the probability that the numbers on the two chips will differ by more than 2? *Hint: Calculate the complement and subtract from one.*

```
total <- ___________
1 - ___________
```

You try it:

1. An apartment building has eight floors. If seven people get on the elevator on the first floor, what is the probability that they all want to get off on different floors? On the same floor?

2. If four dice are rolled, what is the probability that each of the four numbers that appear will be different?

Additional notes.

Section 3.1: Probability Mass Functions

Random Variables

Given a random experiment with an outcome sample space of S. A function that assigns one and only one real number to each element in S is called a **random variable**.

Example

1. Consider an experiment that is the single roll of a die, where the number of spots on the face up side of the die when rolled is observed.

- Outcome space:

- Space of the random variable X:

2. Dr. D has 2 dogs and 2 cats in her household, so $S = \{cat, dog\}$. Let Y be a random variable that denotes the type of animal. Y then maps each element in S to one and only one real number:

When the sample space only has two outcomes,

Distribution of a Random Variable:

When a probability distribution has been specified on the sample space of an experiment, we can determine a probability distribution for the possible values of each random variable X.

This section is focused on probability distributions for discrete distributions. It is said that X has a discrete distribution if X can only take the values of a finite number k different numbers $x_1, x_2, \dots, x_k$ or at most, an infinite sequence of different values $x_1, x_2, \dots$. Random variables that can take any value in an interval are called continuous and will be discussed in a later chapter. Working with discrete random variables requires **summation** while continuous random variables required **integration**.

Discrete variables are integers and usually represent a count of something while continuous variables take values in an interval of real numbers and often measure something.

Definition: PMF

A discrete random variable is a variable that takes integer values and is characterized by a *probability mass function (pmf)*. The pmf p of a random variable X is given by:

The above equation can be read as: the probability that the random variable X is equal to some value, x. Properties that the pmf satisfies:

1.

2.

The term *probability distribution* is a more generic term that describes the probabilities for each different value a random variable can take on. This holds for both discrete and continuous random variables. We will use the term probability distribution for all random variables, but the *pmf* is specific to discrete random variables, and *pdf* (chapter 4) is specific to continuous random variables.

Example

Consider a crooked dice where the cube is shortened in the one-six direction. This has the effect that 1's and 6's have a probability of 1/4 of being rolled, where the other faces each have a probability of 1/8.

- Define the random variable

- Write out the probability distribution.

- Is this a valid pmf? Explain.

You try it

Suppose your roll 2 dice. Let X be the sum of the two die. Write out the pmf. Don't forget you can refer to Example 2.8 in the textbook to visualize the sample space.

Using simulation to estimate discrete probability distributions.

In the cases we've encountered so far, the sample space and the values of the random variable have been discrete, that is, whole numbers. We will get into continuous random variables in the next chapter.

Example

Suppose your roll 2 dice. Let X be the sum of the two die. Use simulation to estimate the probability distribution.

```
die <- 1:6
d1  <- sample(die, 1000, replace = TRUE)
d2  <- sample(die, 1000, replace = TRUE)
sum.2d6   <- d1 + d2
```

The pmf of X is:

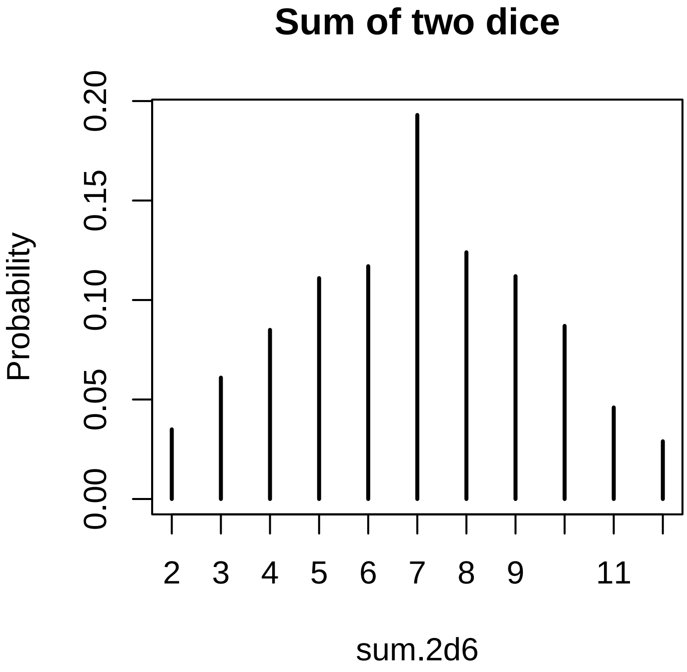
```
proportions(table(sum.2d6))
```

```
## sum.2d6
##     2     3     4     5     6     7     8     9    10    11    12
## 0.035 0.061 0.085 0.111 0.117 0.193 0.124 0.112 0.087 0.046 0.029
```

Plotting the pmf

We can use the function `plot` to plot the estimate of the pmf using the following code.

```
plot(proportions(table(sum.2d6)),
     main="Sum of two dice", ylab="Probability")
```



You try it

1. Three coins are tossed and the number of heads X is counted. Write out the theoretical pmf for X and confirm via simulation.

2. Seven balls number 1-7 are in an urn. Two balls are drawn from the urn without replacement and the sum of X of the numbers is computed. Estimate via simulation the pmf of X.

What are the least likely outcomes of X?

Challenge Example

Suppose you have a bag full of marbles; 50 are red and 50 are blue. You are standing on a number line, and you draw a marble out of the bag. If you get red, you go left one unit. If you get blue, you go right one unit. This is called a *random walk*. You draw marbles up to 100 times, each time moving left or right one unit. Let X be the number of marbles drawn from the bag until you return to 0 for the first time. The rv X is called the *first return time* since it is the number of steps it takes to return to your starting position.

Estimate the pmf of X.

Sampling from a known distribution

Sometimes you know or are given what the distribution of a random variable is, but have need to draw a random sample. We can still use the `sample()` function to do so, we just provide it a vector of probabilities to use.

Example: Blood types

In the United states, human blood comes in four types: O,A,B,AB. Take a sample of thirty blood types with the following probabilities: $P(O) = 0.45, P(A) = 0.4, P(B) = 0.11, P(AB) = 0.04$

```
bloodtypes <- c(_____,_____,_____,_____)
prob_bloodtypes <- c(_____,_____,_____,_____)
sample_blood <- sample(x =________________, size =_____, prob=________, replace=_____)
sample_blood[1:10] #quick peek to confirm
```

The estimated pmf is then:

```
proportions(table(sample_blood))
```

You try it

Suppose the proportion of M&Ms by color is: 14% yellow, 13% Red, 20% Orange, 12% Brown, 20% Green, and 21% Blue. Answer the following questions using simulation.

a. What is the probability that a randomly selected M&M is not green?

b. What is the probability that a randomly selected M&M is red, orange, or yellow?

Additional notes.

Section 3.2: Expected Value & Variance

NOTE: We are going out of order from the textbook in Chapter 3.

Expected Value (Speegle Ch 3.2)

Probability mass functions provide a global overview of a random variable's behavior. Many times we don't need to know everything about a variable. We often want to summarize the variable. One feature of a distribution which we might be interested in is the central tendency of a variable. One measure of central tendency is the *expected value* or *mean* of the observation. The term expected value and mean can be used interchangeably.

Definition: Expected Value

For a discrete random variable X with a pmf p, the *expected value* of X is

where the sum is taken over all possible values of the random variable X.

Example

Two books are assigned for a statistics class: a textbook costing \$137 and its corresponding study guide costing \$33. The university bookstore determined 20% of enrolled students do not buy either book, 55% buy the textbook only, and 25% buy both books, and these percentages are relatively constant from one term to another.

Let X be a random variable that denotes how much a single student will spend on their statistics book. The pmf is:

Interpret this value in context:

Confirm your results using simulation.

You try it:

A retirement portfolio's value increases by 18% during a financial boom and by 9% during normal times. It decreases by 12% during a recession. What is the expected return on this portfolio if each scenario is equally likely?

- Define a random variable.
- Write down the pdf.
- Calculate the theoretical expected value. Write your answer in a full sentence in context of the problem.
- Confirm using simulation.

Variance and standard deviation (Speegle Ch 3.5)

Although the mean is a useful descriptive statistic, it only gives us an idea of where the center of the distribution is located. For instance, the following table gives the monthly temperature of New York City and San Francisco:

months	J	F	M	A	M	J	J	A	S	O	N	D
NYC	32	34	42	53	63	72	77	76	68	57	48	37
SF	49	52	53	56	58	62	63	64	65	61	55	49

The mean temperature for San Francisco is about 57 degrees and the mean temperature for New York is around 55 degrees. So, there mean yearly temperature is about the same. Do you notice anything different about the two cities with regards to monthly temperatures?

To distinguish between 2 distributions with the similar means it might be useful to have a statistic that measures how spread out the distribution is. The variance and standard deviations are such measures.

Definition: Variance

Suppose X is a random variable with mean $\mu = E(X)$. The variance of X, denoted by Var(X), is defined as follows:

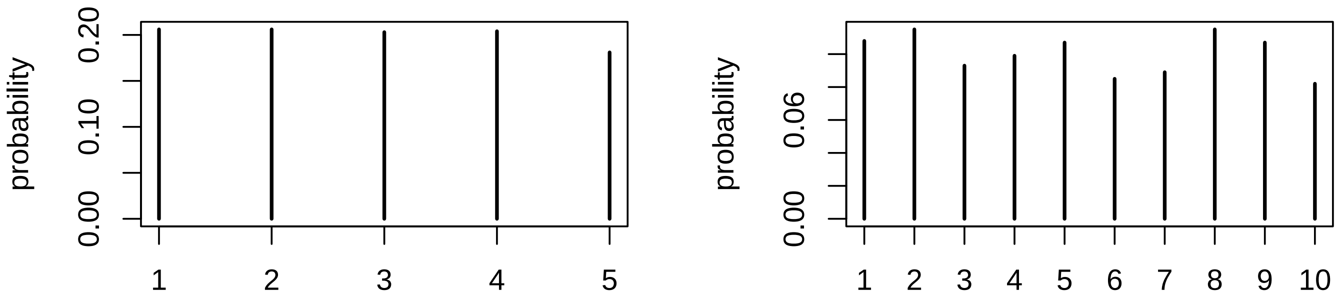
$$Var(X) = \sigma^2 = E[(X-\mu)^2] = \sum_{allk} (k-\mu)^2 * P(X=k)$$

The variance of a distribution provides

The standard deviation of a random variable X ($SD(X)$) is

Which of the two distributions below have the larger variance?

```
par(mfrow=c(1,2))
plot(proportions(table(sample(1:5, size=1000, replace=TRUE))), ylab="probability")
plot(proportions(table(sample(1:10, size=1000, replace=TRUE))), ylab="probability")
```



Example

Let’s return to the statistics book example and calculate $Var(X)$ and $SD(X)$. *Recap: The textbook costs $137, the study guide costing $33. 20% of students don’t buy either book, 55% buy the textbook only, and 25% buy both books.* Confirm your results using simulation.

You try it:

Return to the retirement portfolio question (*Recap: the value increases by 18% during a financial boom and by 9% during normal times, and decreases by 12% during a recession. Each scenario is equally likely*). Calculate the variance and standard deviation.

Section 3.3: Functions of random variables (Speegle 3.4)

There are many reasons why we might be more interested in looking at the distribution of a **function** of a random variable X than the actual variable X. One example would be that we're interested in the absolute distance the random variable is away from it's mean: $g(x) = |X - \mu|$, or we want to know the total gain or loss in a stock portfolio by adding up all the sum of the daily results.

The pmf of $g(X)$ can be computed as follows. For each y, the probability that $g(X) = y$ is given by $\sum p(x)$ where the sum is over all values of x such that $g(x) = y$.

Example

Let $X = -2, -1, 0, 1, 2$, all equally likely and $g(x) = X^2$. Find the pmf of $y = g(x)$, and $E(Y)$.

You try it

Using the random variable X in the above example, find the pmf and expected value of $Y = 2X + 1$.

Example

John travels to work five days a week. We will use X_1 to represent his travel time on Monday, X_2 to represent his travel time on Tuesday, and so on.

- Write an equation using $X_1, ... X_5$ that represents his travel time for the week, denoted by W.
- It takes John an average of 18 minutes each day to commute to work. What would you expect his average commute time to be for the week? Explain how you got to this answer?

What was a major assumption that we had to make to figure out this example?

Independent Random Variables (Speegle 3.5.1)

We say that two random variables are **independent** if the outcome of X does not give probabilistic information about the outcome of Y and vice versa.

Give an example of 2 variables that you think are **independent**:

Give an example of 2 variables that you think are **not independent**.

Theorem 3.8: Rules of Expectation

For random variables X and Y, and constants a, b, and c:

$$E[aX + bY] = aE[X] + bE[Y] \qquad \text{and} \qquad E[c] = c$$

Refer back to the commute time example. We intuitively reasoned that the expectation of the total time is equal to the sum of the expected individual times. This theorem generalizes and formalizes that statement to say that **the expectation of a sum of random variables is always the sum of the expectation for each random variable.**

Example

1. Find $E(2X + 5)$ if $E(X) = 4$

2. Find $E(2X + 5Y)$ if $E(X) = 4$ and $E(Y) = -2$

You try it

1. Let $E(X) = 2$. Find $E(3X - 1)$.

2. Find $E(2)$

3. Find $E(2X - 3Y)$ when $E(X) = -4$ and $E(Y) = 1$

Theorem 3.9: Alternative method to calculate variance

Now that we know some rules of expected value, we can use a simplified method to find the variance of a random variable.

Example

Let $X = -2, -1, 0, 1, 2$, all values equally likely. Find $E(X)$ and $Var(X)$.

You try it

Let $Y = 2x$ where $x = 0, 1, 2$ and $p(x) = .1, .5, .4$. Find $E(X)$ and $Var(X)$.

Theorem 3.10: Rules of Variance

1. Let X be a random variable and c a constant. Then

2. Let X and Y be independent random variables. Then

Example

Suppose that three random variables X_1, X_2, X_3 form a random sample from a distribution for which the mean is 5 and the variance is 3. Determine the value of $E(2X_1 - 3X_2 + X_3 - 4)$ and $\mathrm{Var}(2X_1 - 3X_2 + X_3 - 4)$.

You try it:

Marksmanship competition at a certain level requires each contestant to take ten shots with each of two different handguns. Final scores are computed by taking a weighted average of 4 times the number of bull-eyes made with the first gun plus 6 times the number gotten with the second gun. If Bertha has a 30% chance of hitting the bull's-eye with each shot from the first gun and a 40% chance with each shot from the second gun, what is the variance of her score?

Section 3.4: Named Discrete Distributions

Now that the foundations of random variables, probability distributions expectation and variance are under our belt, let's start to look at some special random variables that occur so commonly, or have such mathematically wonderful properties that they have specific names. We will look at 6 different types of discrete random variables. For each we will learn the following:

- How to define the random variable
- How to identify the parameters and write the distributional notation
- The formula for the pmf, and how to find theoretical probabilities
- Formulas for the theoretical mean and variance
- How to calculate all of the above using R commands (both theoretical, and via simulation)

A note on the R commands

In R, the common distributions are defined by their **root** name with 3 different prefixes:

- `d` to compute $P(X == x)$ e.g.: `dbinom, dgeom, dhyper, dnbinom`
- `p` to compute $P(X \leq x)$ e.g.: `pbinom, pgeom, phyper, pnbinom`
- `r` to randomly draw N samples from the specified distribution. e.g: `rbinom, rgeom, rhyper, rnbinom`

Bernoulli Distribution (Speegle 3.3)

Situation The simplest type of experiment is one in which there are only two outcomes (success/failure, live/die, true/false, yes/no etc.). When running simulations in Chapter 2, you wrote your experiment to get down to a single TRUE/FALSE. You were creating a Bernoulli random variable. This simple, yet fundamental random variable serves as the basis for the rest of the distributions in this chapter.

Random variable: Let X be a random variable that denotes the outcome from a Bernoulli trial with probability of success p. Specifically let $X = 1$ denote a success, and $X = 0$ denote a failure. (What is considered a *success* is entirely up to context. If you are interested in mortality rate for a certain disease, then "death" would be a success.)

Distributional Notation: $\boxed{X \sim Bernoulli(p)}$

pmf: $P(X = x) = p^x(1-p)^{1-x}$ ~~$x \geq$~~ 0 $x \in \{0,1\}$

Mean and variance: $E(X) = p$ $Var(X) = p(1-p)$

X	$P(x)$	
0	$p^0(1-p)^{1-x}$	$\to 1-p$
1	$p^1(1-p)^{1-1}$	$\to p$

R commands: There are no fancy named R commands for this distribution. You can simulate this random variable using `sample(c(0,1), prob=c(1-p, p))` directly, or through a Binomial random variable with n=1.

Example:

A beet seed has been planted, and will either germinate or not. The probability of germination is 0.8, and germination is considered a success.

$$P(X=x) = .8^x * (1-.8)^{1-x}$$

You try it:

Define 2 Bernoulli trials.

Binomial Distribution (Speegle 3.3.1)

Situation: If n independent random variables $X_1, ..., X_n$ all have the same Bernoulli distribution with probability of success p, then their sum is equal to the number of X_i's which equal 1, and the distribution of the sum is known as a Binomial distribution. Examples include:

- Toss 5 coins and count the number of heads
- The number of times in a week a person is late for work, whey they have a 10% chance of being late each day, independent of other days.

Random variable: Let X be a random variable that represents the number of "success" in a series of n independent Bernoulli trials each with probability success p.

Distributional Notation: $X \sim Binomial(n, p)$

pmf:

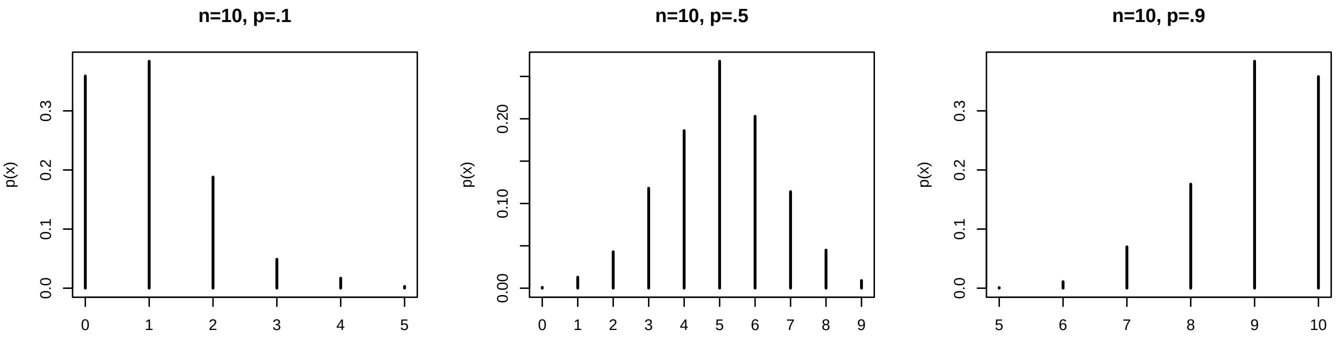
$$P(X = x) = \underbrace{\binom{n}{x}}_{\text{Any order}} \overset{\text{P(success)}}{p}^{\overset{\text{\# succ}}{x}} \overbrace{(1-p)}^{\text{P(FAIL)}}{}^{\overbrace{n-x}^{\text{\# Fail}}} \qquad x = 0, 1, 2, ..., n$$

Mean and variance: $E(X) = np \qquad Var(X) = np(1-p)$

R commands:

- `dbinom(x, size=n, prob=p)` to compute $\boxed{P(X == x)}$
- `pbinom(x, size=n, prob=p)` to compute $P(X \leq x)$
- `rbinom(N, size=n ,prob=p)` to randomly draw N samples from a $X \sim Binom(n, p)$ distribution.

Visualizing the shape of the distribution:



What happens to the distribution as p increases?

$$E(X) = \sum x \cdot P(x)$$

Example:

success

1. Plant 10 beet seeds, and assume that the germination of one seed is independent of the germination of another seed, and all seeds have a germination probability of p=.8. Let X be the number of seeds that germinated. 1) Write down the pmf, and 2) the distributional notation for X. Then compute the mean and variance both 3) theoretically, and 4) confirm using simulation.

$X \sim \text{Binomial}(10, .8)$

$\text{pmf} = \binom{10}{x}(.8)^x(.2)^{10-x}$

Theoretical $E(X) = n \cdot p = (10)(.8) = 8$

$Var(X) = n \cdot p \cdot (1-p) = (10)(.8)(.2) = 1.6$

Simulation

```
x <- rbinom(10000, 10, .8)
mean(x)
Var(X)
```

2. A coin for which the probability of heads is .6 is tossed nine times. Find the probability of obtaining 3 heads.

p n

$X \sim \text{Binomial}(9, .6)$

by hand using the pmf

$$P(X=3) = \binom{9}{3}(.6)^3(.4)^6$$

theoretical using R commands

```
dbinom(3, 9, .6)
```

using simulation

```
x <- rbinom(10000, 9, .6)
mean(x==3)
```

$P(X \geq 1)$ 0 1 2 3 . . . . .

$1 - P(X=0)$

3. 10 students are selected at random, each has a probability of 0.10 of being a Math major. What is the probability that at least one student is a math major?

by hand using the pmf

theoretical using R commands

using simulation

You try it: HW

1. A recent national study showed that approximately 45% of college students binge drink. Let X equal the number of students in a random sample of size $n = 12$ who binge drink. Calculate the following probabilities both by hand using the pmf, and R commands.

a. X is at most 2.

b. X is at least 1.

c. Use simulation to obtain the mean and variance of X.

2. ~~A certain electric system contains 10 components. Suppose that the probability that each individual will fail is .2 and that the components fail independently of each other. **Given that at least one of the components failed**, what is the probability that at least two of the components have failed? Write this out mathematically and simplify *before* you go to R.~~

Geometric Distribution (Speegle 3.3.2)

Situation Given a series of independent Bernoulli trials, we are accustomed to thinking of n and p as fixed, and x is considered the number of successes for a binomial distribution. Suppose that the problem is turned around though, and the question is asked, how many trials will be required in order to achieve the first success? Put this way, the number of trials is the random variable and number of successes is fixed.

- How many free throws can Stephen Curry make before he misses?
- The probability that a random person who smokes will develop a severe lung condition in their lifetime is about 0.3. How many people do you have to check on before you meet someone with a severe lung condition?

Random Variable: Let X be the number of failures before the first success in a Bernoulli process with probability of success p.

Distributional Notation: $X \sim Geom(p)$

pmf:

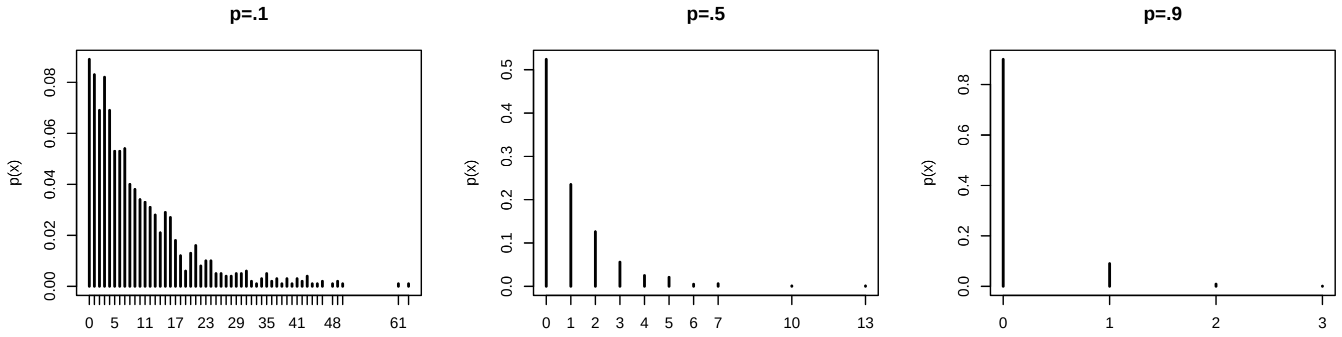
$$P(X=x) = \overbrace{(1-p)}^{P(\text{Fail})}{}^{\overset{\text{Failure}}{x}}\overset{P(\text{success})}{p} \qquad x = 0, 1, 2, \ldots$$

Mean and variance: $E(X) = \frac{1-p}{p}$ $\qquad Var(X) = \frac{1-p}{p^2}$

R commands:

- `dgeom(x,prob=p)` to compute $P(X == x)$
- `pgeom(x,prob=p)` to compute $P(X \leq x)$
- `rgeom(N,prob=p)` to randomly draw N samples from a $X \sim Geom(p)$ distribution.

Visualizing the shape of the distribution



What happens to the distribution as p increases?

Example

X = # shots before he misses

✓ $X \sim \text{Geometric}(.1)$

Professional basketball player Steve Nash was a 90% free throw shooter over his career. Answer the following questions using the formulas and also simulation.

✓ ✓ ✓ ✓ ✗ $X = 4$ $(.9)^4(.1)$

.9 .9 .9 .9 .1

a. If Steve Nash starts shooting free throws, how many would he expect to make before missing one?

Theoretical

$$E(X) = \frac{1-p}{p} = \frac{.9}{.1} = 9$$

Simulation

```
x <- rgeom(10000, .1)
mean(x)
```

b. What is the probability that he could make 20 in a row before he misses? $P(X=20)$

by hand using the pmf $\longrightarrow (.9)^{20}(.1)$

theoretical using R commands

```
dgeom(20, .1)
```

using simulation

```
mean(x ==20)
```

You try it:

Complete the following using both theoretical and simulation methods

1. The 2010 American Community Survey estimates that 47.1% of women ages 15 years and over are married. We randomly select three women ~~between these ages~~ over 15s.

a. What is the probability that the third women selected is the only one that is married?

N N M

b. On average, how many women would you expect to sample before selecting a married woman? What is the standard deviation?

2. A machine that produces a special type of transistor has a 2% defective rate. The production is considered a random process where each transistor is independent of the others.

$X = \#$ of good transistors

a. What is the probability that the 10th transistor produced is the first with a defect?

$P(X=9)$

b. What is the probability that the first failure occurs after the 4th transistor was produced?

✓ ✓ ✓ ✓ ? ? ?

$P(X > 4)$

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

$1 - [P(X=0) + P(X=1) + \dots + P(X=4)]$

$= 1 - \text{pgeom}(4\ .1\ .2)$

✓ ✓ ✗ ✓ ✗ Bernulli trials # of successes = Binomial

✗ ✗ ✓ # of failures before the 1st success : Geometric

✓ ✗ ✗ ✗ ✓ # of failures before the nth success

Negative Binomial Distribution (Speegle 3.6.2)

Situation A random variable with a negative binomial distribution originates from a context much like the one that yields the geometric distribution. Again, we focus on independent and identical trials, each of which results in one of two outcomes, success or failure. The probability of success, p, stays constant for each trial. The geometric case handles the number of cases until the first success occurs. What if we are interested in knowing the number of trials until the second, third, fourth, etc success occurs. Examples:

- How many people do you have to meet at college before you meet the 4th person from your hometown?
- Tire Mart has a lot of really cheap tires, but 20% of them are defective. How many tires do you need to go through to find 4 new tires?

context of problem

Random Variable: Let X denotes the number of failures before the nth success, with probability of success p.

Distributional Notation: $X \sim NegBin(n, p)$

pmf:

n successes + failures

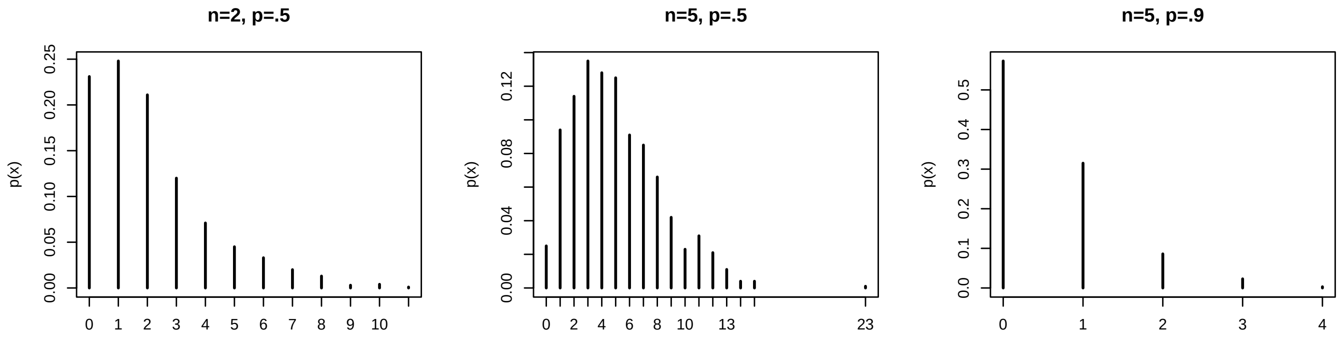
$$P(X = x) = \binom{x+n-1}{\underline{x}} \boxed{p^n} \boxed{(1-p)^x} \qquad x = 0, 1, 2, \ldots$$

We can think of the negative binomial distribution as the sum of r geometric distributions. This simplifes the **Mean and variance:** $E(X) = \boxed{\frac{n(1-p)}{p}}$ $\qquad Var(X) = \boxed{\frac{n(1-p)}{p^2}}$

R commands:

- `dnbinom(x, n, prob=p)` to compute $P(X == x)$
- `pnbinom(x, n, prob=p)` to compute $P(X \leq x)$
- `rnbinom(N, n, prob=p)` to randomly draw N samples from a $X \sim NegBin(n, p)$ distribution.

Visualizing the shape of the distribution:



What happens to the distribution as n and p change?

Example: Oil!

A geological study indicates that an exploratory oil well drilled in a particular region should strike oil with probability 0.2. Write down the pmf, and then calculate the mean and variance.

dry

1 Let X denote the number of dry wells drilled before the third oil strike (oil is found on that well)

$$X \sim \text{NegBin}(3, .2) \quad 2$$

pmf

$$P(X=x) = \binom{x+3-1}{x}(.2)^3(.8)^x$$

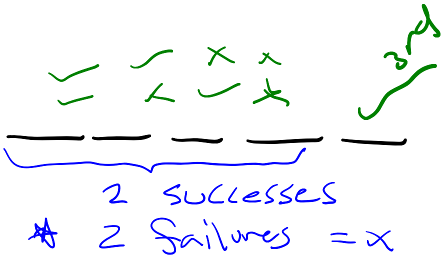
Theoretical

$$3 \quad E(x) = \frac{3(.8)}{.2} = \frac{n(1-p)}{p}$$

Simulation

```
x <- rnbinom(10000, 3, .2)
mean(x)
```

Find the probability that the third oil strike comes on the fifth well drilled.



by hand using the pmf

$$P(X=2) = \binom{4}{2}(.2)^3(.8)^2 =$$

R choose(4, 2)

theoretical using R commands

```
dnbinom(2, 3, .2)
```

using simulation

Using the x drawn from before

```
mean(x==2)
```

1) Define X as a sentence
2) Write X in distributional notation X ~ ….
3) write what you are asked to find in math notation e.g.
E(X) or P(X = #)

You try it:

Ten percent of the engines manufactured on an assembly line are defective.

- If engines are randomly selected one at a time and tested, what is the probability that the first non defective engine will be found on the second trial?

  Let X be...

- What is the probability that the third non defective engine will be found on the fifth trial?

  Let Y be ...

- Find the mean and variance of the number of trials on which the first non defective engine is found.

- Find the mean and variance of the number of failures until the third non defective engine is found.

Poisson Distribution (Speegle 3.6.1)

Situation: A Poisson process is one where events occur at random times during a fixed time period. The events occur independently from each other, but with a constant average rate over that time period. Examples include

- Number of calls per hour at a call center
- Number of hits on a webpage in a day
- Number of meteor strikes on the surface of the moon annually

Random Variable: Let X be the number of events occurring in a Poisson process with rate λ over one unit of time (e.g. per year, per second, per day).

Distributional Notation: $X \sim Poisson(\lambda)$

pmf:

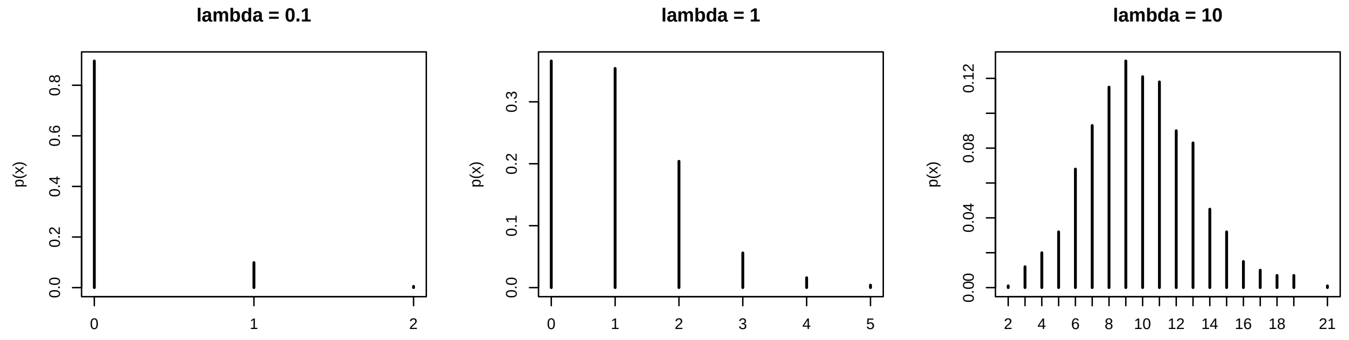
$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

Mean and variance: $E(X) = Var(X) = \lambda$

R commands:

- `dpois(x,lambda)` to compute $P(X == x)$
- `ppois(x,lambda)` to compute $P(X \leq x)$
- `rpois(N,lambda)` to randomly draw N samples from a $X \sim Poisson(\lambda)$ distribution.

Visualizing the shape of the distribution



What happens to the distribution as λ increases?

Example

The Taurids meteor shower is visible on clear nights in the Fall and can have visible meteor rates around five per hour. What is the probability that a viewer will observe exactly eight meteors in two hours?

by hand using the pmf

theoretical using R commands

using simulation

You try it:

Suppose a typist makes typos at a rate of 3 typos per 10 pages. What is the probability that they will make at most one typo on a five page document?

Hypergeometric Distribution (Speegle 3.6.3)

Situation: The hypergeometric distribution is a series of Bernoulli trials that are dependent. This occurs when we are *sampling without replacement* from a finite population. Examples include:

- Capture some fish, tag them & release them. Then come back later and fish some more, counting how many tagged ones you catch again.
- Create a bipartisan committee of 10 senators, and count the number of Republicans chosen.

Random Variable: Let X denote the number of success out of a sample size of k when drawing without replacement from a pool where there are a total of m successes and n failures available.

Distributional Notation: $X \sim Hypergeometric(m+n, n, k)$

pmf:

$$P(X = x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$$

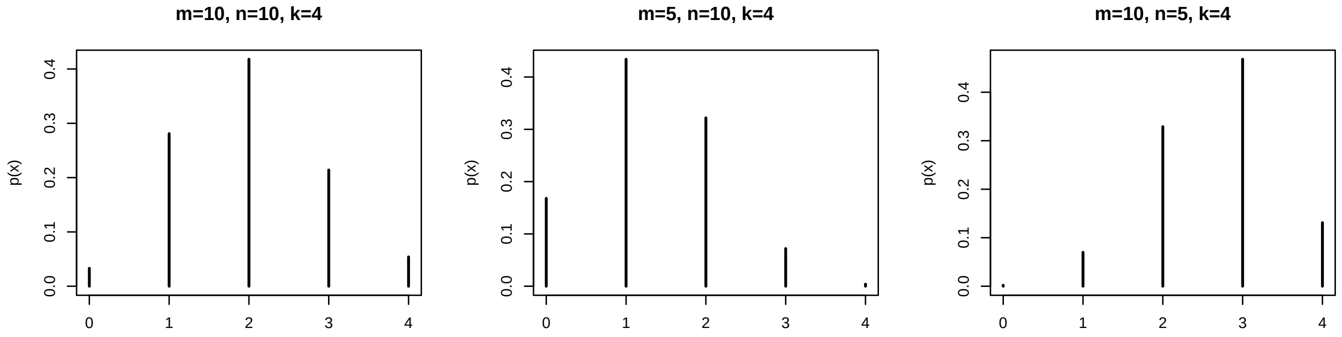
Mean and variance:

$$E[X] = k\left(\frac{m}{m+n}\right) \qquad V(X) = k\left(\frac{m}{m+n}\right)\left(\frac{n}{m+n}\right)\left(\frac{m+n-k}{m+n-1}\right)$$

R commands:

- `dhyper(x, m, n, k)` to compute $P(X == x)$
- `phyper(x, m, n, k)` to compute $P(X \leq x)$
- `rhyper(N, m, n, k)` to randomly draw N samples from a $X \sim Hypergeometric(m+n, n, k)$ distribution.

Visualizing the shape of the distribution



What happens to the distribution as the parameters change?

Example

1. An urn contains nine chips, five red and four white. Three are drawn out at random without replacement. Let X denote the number of red chips in the sample. Identify the parameters, and find $E(X)$ and $Var(X)$.

Theoretical

Simulation

2. In a small pond there are 50 fish, 10 of which have been tagged. If a fisherman's catch consists of 7 fish, selected at random and without replacement, and X denotes the number of tagged fish, what is the probability that exactly 2 tagged fish are caught?

by hand using the pmf

theoretical using R commands

using simulation

You try it:

1. Suppose that there are 3 defective items in a lot of 50 items. A sample of size 10 is taken at random and without replacement. Let X denote the number of defective items in the sample. Find the probability that the sample contains

a. Exactly 1 defective item.

b. At most 1 defective item.

2. A display case contains thirty-five diamonds, of which ten are real diamonds and twenty-five are fake diamonds. A burglar removes four gems at random, one at a time and without replacement. What is the probability that the last gem she steals is the second real diamond in the set of four?

Additional notes.

Section 4.1: Probability density functions

Random variables that can assume every value in an interval have continuous distributions. A continuous distribution can also be characterized by its probability density function (p.d.f.).

Many experiments or observations of random phenomenon do not have integers as outcomes, but instead are measurements selected from an interval of numbers. For example, you could find the length of time that it takes when waiting in line at the grocery store or the weight of a bag of potato chips advertised at 1 oz. If the measurements could come from an interval of possible outcomes, we call them continuous-type data.

Definition 4.1: Probability distribution function (pdf)

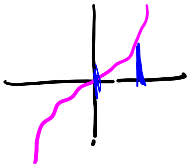
A probability density function (pdf) is a function f such that:

1. $f(x) \geq 0$ for all x
2. $\int f(x)dx = 1$ over the domain of support.

Example

Suppose $f_Y(y) = 4y^3$, $0 \leq y \leq 1$. Is this a valid probability distribution? Why? yes

1)



✓ $f(y) \geq 0$

2) $\int_0^1 4y^3 \, dy = 1$

Definition 4.3: Continuous random variable

A *continuous* random variable X is a random variable described by a pdf in the sense that…

$$P(a \leq x \leq b) = \int_a^b f_X(x)\, dx$$

whenever $a \leq b$, including the cases $a = -\infty$ or $b = \infty$.

Example

Let X be a random variable with the following p.d.f.

$$f(x) = \{ \begin{array}{cc} \frac{2}{3}x^{-1/3} & for 0 < x < 1, \\ 0 & otherwise. \end{array}$$

Compute P($X \leq 8/27$).

$$\int_0^{8/27} \frac{2}{3} x^{-1/3} dx = x^{2/3} \Big|_0^{8/27} = \frac{4}{9}$$

```
f.y
~~integrand~~ <- function(~~y~~x) { 2/3*x^(-1/3)__}
integrate(i~~ntegrand~~,lower= _0___ ,upper= 8/27_____)
```

You try it

Do both by hand, and using R. Don't forget to write your code down in these notes.

a. Suppose $f_Y(y) = 4y^3$, $0 \leq y \leq 1$. Find $P(0 \leq Y \leq \frac{1}{2})$.

b. For the random variable Y with pdf $f(y) = \frac{2}{3} + \frac{2}{3}y$ for $0 \leq y \leq 1$, find $P(\frac{3}{4} \leq Y \leq 1)$.

5/16

Definition 4.4: Cumulative distribution function (cdf)

The *cumulative distribution function (cdf)* associated with X (either discrete or continuous) is the function $F(x) = \underline{P(X \leq x)}$, or written out in terms of the pdf's and cdf's.

$$\boxed{F(x)} = \underline{P(X \leq x)} = \int_{-\infty}^{x} f(t)dt$$

pnbinom

for continuous variables and

$$F(x) = P(X \leq x) = \sum_{n=-\infty}^{x} p(n)$$

for discrete variables.

Theorem 4.1

Let X be a continuous random variable with pdf $\underline{f}$ and cdf $\underline{F}$.

1. $\frac{d}{dx} F = f$
2. $P(a \leq X \leq b) = F(b) - F(a)$
3. $1 - P(x \geq a) = 1 - F(a) = \int_a^\infty f(x)dx$

Example

Find the cdf for the random variable Y for the following pdf: $f_Y(y) = 4y^3$ for $0 \leq y \leq 1$. Calculate $\underline{P(0 \leq Y \leq 1/2)}$ using $F_Y(y)$.

$$F(y) = P(Y \leq y) = \int_0^y 4t^3\,dt = y^4$$

$$P(0 \leq Y \leq 1/2) = F(\tfrac{1}{2}) - F(0)$$

$$(\tfrac{1}{2})^4$$

Example

A random variable Y has CDF as follows:

$$F(y) = \begin{cases} 0 & \text{for } y < 0 \\ ln(y) & 1 \leq y \leq e \\ 1 & e < y \end{cases}$$

a. Find $P(Y < 2)$. $\ln(2)$

$\log(2)$ (R)

b. Find $P(2 < Y \leq 2\frac{1}{2})$ $\ln(2.5) - \ln(2)$

↑

c. Find $P(2 < Y < 2\frac{1}{2})$ SAME

d. Find $f(y)$ $f(y) = \frac{d}{dy} F(y) = \frac{d}{dy} \ln(y) = \frac{1}{y}$

by hand

use R to integrate

You try it

a. The cdf for a random variable Y is defined by $F(y) = 0$ for $y < 0$, $F(y) = 4y^3 - 3y^4$ for _0 $\leq y \leq 1$; and $F(y) = 1$ for $y > 1$. Find $P(\frac{1}{4} < Y < \frac{3}{4})$.

b. Suppose $F(y) = \frac{1}{12}\left(y^3 + y^2\right)$, $0 \leq y \leq 2$. Find $f(y)$.

c. In a certain country, the distribution of a family's disposable income, Y, is described by the pdf $f(y) = ye^{-y}$, $y \geq 0$. Find $F(y)$.

Additional notes.

Section 4.2 Expected value of a continuous random variable

If a random variable has a continuous distribution for which the p.d.f. is f, then the expectation $E(X)$ is defined as

Example: Jail time

Suppose X is the random variable that represents the prison sentence in years for persons convicted of grand theft auto and assume that X has a p.d.f. of $f(x) = \frac{1}{9}x^2$ for $0 < x < 3$. What is the **average** length of time these people spend in jail? Calculate this by hand, and using R.

You try it

Find the expected value for the following p.d.f.; $f(x) = 2x$ for $0 < x < 1$.

As in the discrete case, we can also define functions of random variables.

Theorem 4.2

Let X be a continuous random variable and let g be a function.

$$E[g(X)] = \int g(x)f(x)dx$$

Example

Let Y have probability density function $f_Y(y) = 2(1-y), 0 \leq y \leq 1$. Suppose that $W = Y^2$, in which case

$$f_W(w) = \frac{1}{\sqrt{w}} - 1, 0 \leq w \leq 1$$

Find $E(W)$ a) using $f(w)$ directly, and b) using theorem 4.2. Confirm both using R.

Example

Suppose that the p.d.f. of a random variable X with a continuous distribution is $f(x) = 2x$ for $0 < x < 1$. Find the expectation of $1/X$. Do this by hand, confirming your results using R.

You try it

Grades on the last test were not very good. Their distribution is as follows:

$$f(y) = \frac{1}{5000}(100 - y) \qquad \text{for } 0 \leq y \leq 100$$

As a way of curving the results, the professor announces that he will replace each person's grade, Y, with a new grade $g(Y)$ where $g(Y) = 10\sqrt{Y}$. Will the professor's strategy be successful in raising the class average above 60? Write down the equation, then use R to calculate the value.

Additional notes.

Section 4.3: Variance and Standard Deviation

Just like for discrete variables, the variance and the standard deviation measures the spread of the random variable around its mean. The formulas remain unchanged for the continuous random variable.

Example

Find the variance of the random variable Y, where

$$f_Y(y) = 3(1 - y)^2, 0 < y < 1$$

You try it

1. An exponential random variable has the following pdf: $f_Y(y) = \lambda e^{-\lambda y}$, $y \geq 0$. Show that the variance of Y is $1/\lambda^2$.

2. A random variable Y is described by the pdf $f_Y(y) = 2y$ for $0 \leq y \leq 1$. What is the standard deviation of $3Y + 2$

Section 4.4 Normal random variables

Situation The most widely used continuous distribution is the normal distribution, a distribution with the familiar "bell" shape. Many characteristics in nature exhibit this shape:

- heights of humans, trees, wombats
- failure of mechanical parts due to wear and tear
- random noise in electrical circuits

Random variable: Let X be a random variable from a Normal distribution defined by the parameters μ for the mean, and σ^2 for the variance.

Distributional Notation: $X \sim N(\mu, \sigma^2)$.

pmf:

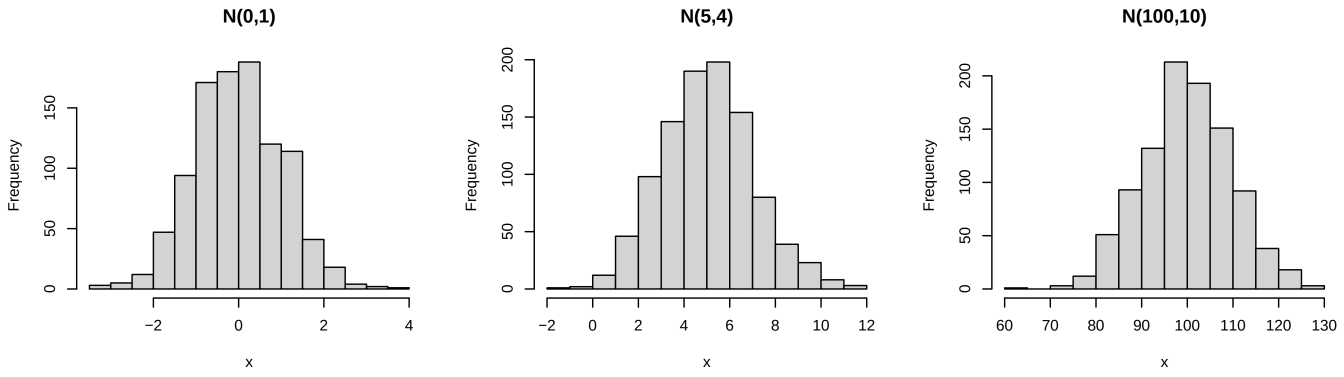
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\frac{(x-\mu)}{2\sigma}\right]^2}$$

Mean and variance: $E(X) = \mu \qquad Var(X) = \sigma^2$

R commands: Note that R uses the *standard deviation*, NOT the variance.

- `dnorm(x, mu, sd)` to compute $P(X == x)$
- `pnorm(x, mu, sd)` to compute $P(X \leq x)$ (the cdf)
- `rnorm(N, mu, sd)` to randomly draw N samples from a $X \sim N(\mu, \sigma^2)$ distribution.

Visualizing the shape of the distribution



How does the distribution change when μ and σ^2 change?

The normal distribution is an extremely important distribution and we will discuss some of its properties in this section. There are three main reasons why the normal distribution is so important:

1.

2.

3.

Standard Normal Random Variable

The *standard normal random variable* Z is a special case of the Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. The PDF then simplifies to

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Example

Use R to evaluate the following integrals under the Standard Normal Z distribution. In each case, draw a diagram of $f_Z(z)$ and shade the area that corresponds to the integral, then use R to calculate the area under the distribution curve.

1. $\frac{1}{\sqrt{2\pi}} \int_{-.44}^{1.33} e^{-z^2/2}$

2. $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{.94} e^{-z^2/2}$

Example

Use `pnorm` to calculate the theoretical probability for each question, and confirm via simulation using `rnorm`.

1. $P(Z > 1.3)$

2. $P(-0.15 < Z < 1.5)$

3. $P(Z < -2)$

You try it

Use R for all steps, do not do these by hand or using Z tables. Use the `integrate`, `pnorm` and `rnorm` functions.

1. $\frac{1}{\sqrt{2\pi}} \int_{-1}^{2} e^{-z^2/2}$

2. $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{2.1} e^{-z^2/2}$

3. $P(Z < 0.9)$

4. $P(1.1 < Z < 2.5)$

5. $P(Z > 0.9)$

Example

A normally distributed population of lemming body weights has a mean of 63.5 g and a standard deviation 12.2 g.

1. Define the random variable.

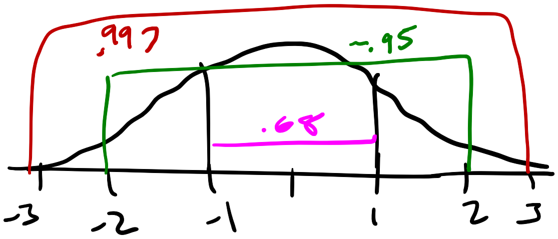
2. Draw a picture of the distribution

For each question below, write the question in math notation, sketch a picture, calculate the theoretical probability using `pnorm`, and simulate the probability using `rnorm`.

1. What proportion of this population is 78.0 g or larger?

2. What is the probability of choosing at random from this population a weight smaller than 41 g?

3. What is the probability of choosing at random from this population a weight between 60 and 70 g?

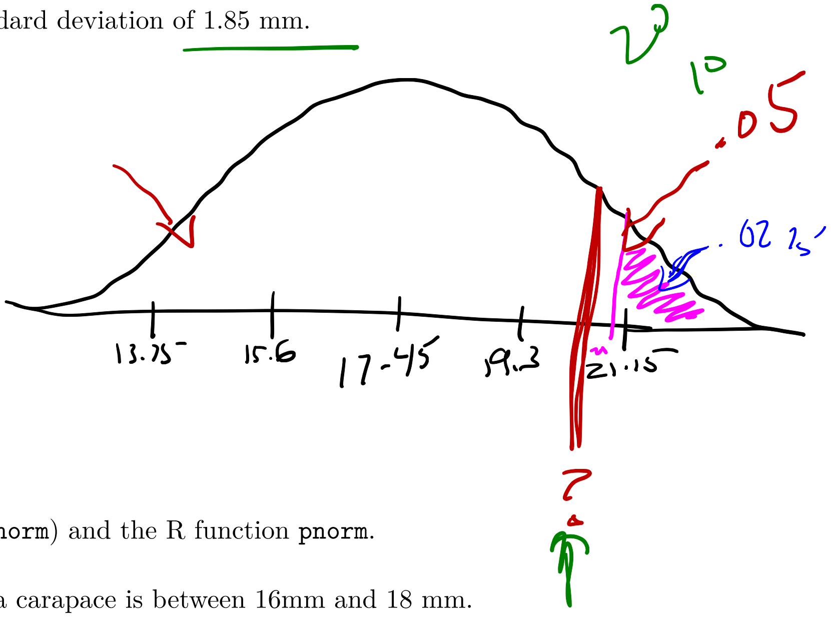


You try it

According to a recent study, the carapace length for adult males of a certain species of tarantula are normally distributed with a mean of 17.45 mm and a standard deviation of 1.85 mm.

Define the random variable.

Draw a picture of the distribution



Answer these questions using both simulation (`rnorm`) and the R function `pnorm`.

1. What is the probability that the length of a carapace is between 16mm and 18 mm.

2. Would a tarantula that had a carapace longer than 21 mm be unusual? $P(X \geq 21)$

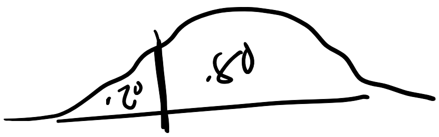
$1 - \text{pnorm}(21, \underset{\mu}{17.45}, 1.85) = .0275$

2.75%

What criteria did you use to determine what would be unusual? ≈ 5%

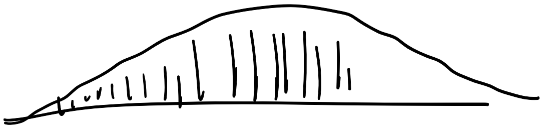
Inverse Normal

$P(X \le t) = .20$



It is often of interest to calculate a quantile of the normal distribution. For instance, maybe we want to know the score you would need on the SAT exam to be in the *top 10th percentile*. For this sort of problem we would want to use the `qnorm(p,mu,sigma)` where p is the area to the left of a certain value of interest.

A *quantile* divides the range of a probability distribution into intervals of equal probability.



Example

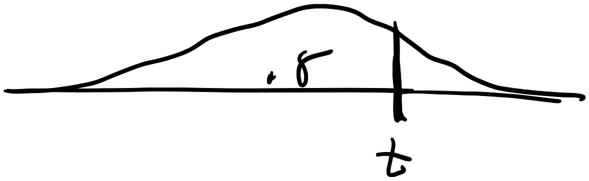
Let's look at those lemmings again. Recall the weight of a lemming can be described as $X \sim N(63.5, 12.1^2)$. What lemming weight corresponds to the 80th percentile?

1

- Translate question into mathematical notation $P(X \le t) = .8$

2

- Draw picture



3

- Find quantile x using `qnorm` qnorm(.8, 63.5, 12.1) = 73.68

You try it

Reconsider the tarantula example above. What carapace length corresponds to the **top 20th percentile**.

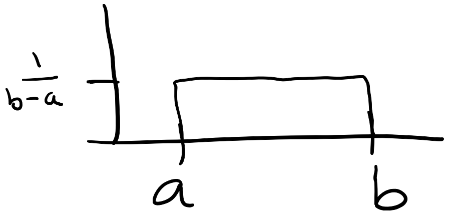
Section 4.5 Uniform and Exponential Random Variables

Uniform (Speegle 4.5.1)

Situation: Uniform random variables can be either discrete or continuous, and describe a distribution where all outcomes are equally likely. Examples include

* the result of a die roll * pseudo random number generator * round off error in measurements

pdf: $f(x) = \boxed{\frac{1}{b-a}}$ $\boxed{a \le x \le b}$



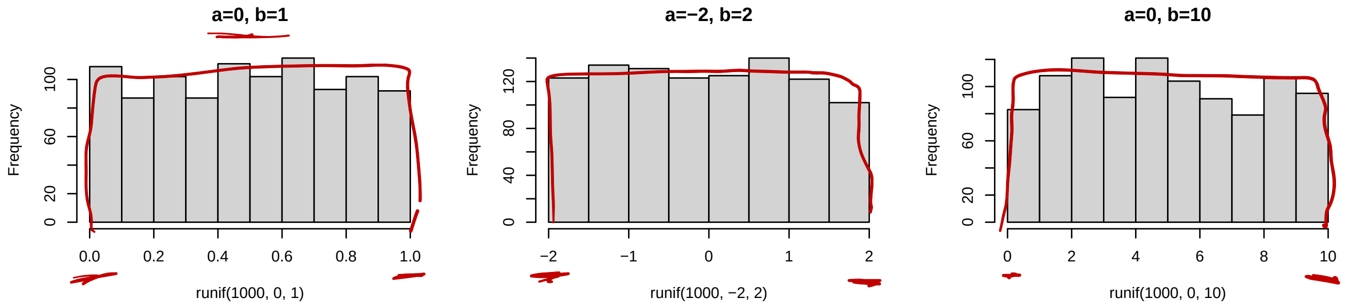
Distributional Notation: $X \sim Unif(\underline{a,b})$

Mean and variance: $E[X] = \frac{b+a}{2}$ $\quad Var(X) = \frac{(b-a)^2}{12}$

R commands:

- `dunif(x, a, b)` to compute $P(X == x)$ for discrete only
- `punif(x, a, b)` to compute $P(X \le x)$ (the cdf)
- `runif(N, a, b)` to randomly draw N samples from a $X \sim Unuf(a, b)$ distribution.

Visualizing the shape of the distribution



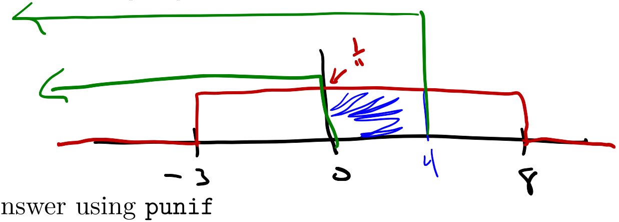
What happens to the distribution as you change a and b?

Example

$X \sim Unif(-3, 8)$

Suppose a random variable X has a uniform distribution on the interval [-3,8], then the p.d.f. of X is

$$f(x) = \begin{cases} 1/11 & -3 \leq x \leq 8 \\ 0 & \text{otherwise} \end{cases}$$



Calculate $P(0 \leq X \leq 4)$. Do this by hand, and confirm your answer using `punif`

$$\int_0^4 \frac{1}{11}\,dx = \frac{1}{11} x \Big|_0^4 = \frac{4}{11}$$

```
punif(4, -3, 8) -
  punif(0, -3, 8)

mean(x > 0 & x < 4)
```

Calculate the mean and standard deviation. Do this by hand, and confirm your answer using simulation.

$$E(X) = \frac{b+a}{2} = \frac{8-3}{2} = \frac{5}{2}$$

$$SD(x) = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(8+3)^2}{12}} = 3.17$$

```
x <- runif(10000, -3, 8)
mean(x)
sd(x)
```

You try it

Suppose that a random variable X has a uniform distribution on the interval [-4,10]. Write down the pdf of X, find the mean, standard deviation and the value of $P(-1 < X < 6)$ both theoretically and using simulation.

Exponential Random Variables (Speegle 4.5.2)

Situation: Exponential random variables measure the waiting time until the first event occurs in a Poisson process.

- The waiting time until an electronic component fails could be exponential
- The time between customers in a store.

Distributional Notation: Let $X \sim Exp(\lambda)$ be an exponential random variable with rate λ.

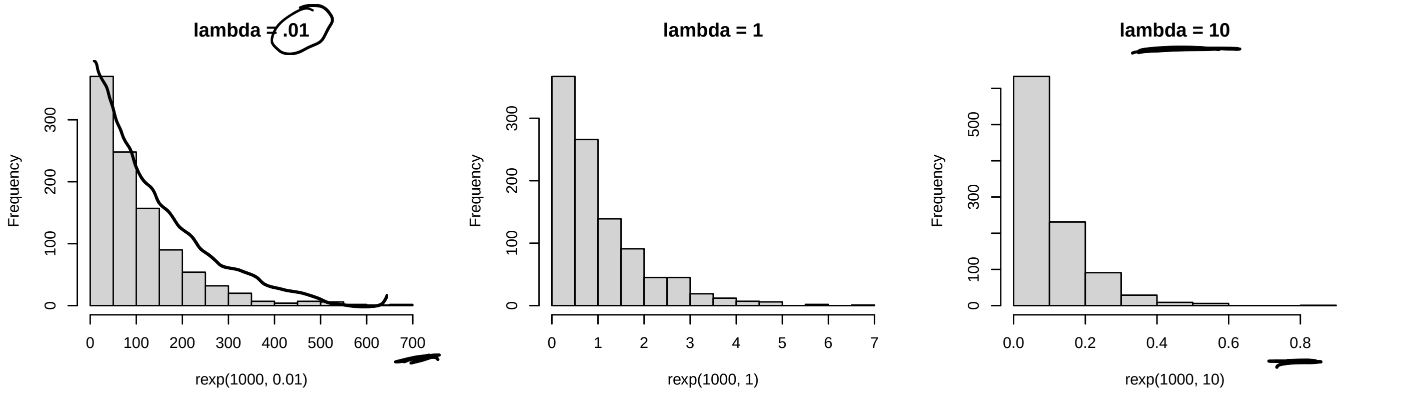
pdf: $f(x) = \lambda e^{-\lambda y}$ $y \geq 0$

Mean and variance: $E[X] = \frac{1}{\lambda}$ $Var(X) = \frac{1}{\lambda^2}$

R commands:

- d~~unif(~~x, lambda) to compute $P(X == x)$
- pexp p~~unif(~~x, lambda) to compute $P(X \leq x)$ (the cdf)
- r~~unif(~~N, lambda) to randomly draw N samples from a $X \sim Exp(\lambda)$ distribution. rexp

Visualizing the shape of the distribution

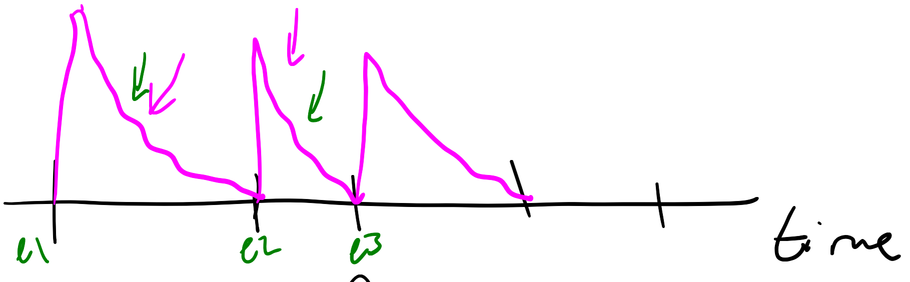


What happens to the distribution as you change lambda?

"memoryless"

memoryless

time between events ~ Exponential



events Randomly happen, Independent as part of a Poisson Process

Example

Suppose the time to failure (in years) for a particular component is distributed as an exponential random variable with rate $\lambda = 1/5$. For better performance, the system has two components installed, and the system will work as long as either components installed, and the system will work as long as either component is functional. Assume the time to failure for the two components is independent. What is the probability that the system will fail before 10 years has passed?

Let X1 be the time until component 1 fails
Let X2 be the time until component 2 fails

X1 ~ Exp(1/5)
X2 ~ Exp(1/5)

- Problem setup

  System fail in < 10 years if X1 < 10 & X2 < 10

  P(system fails) = P(X1 < 10)* P(X2 < 10) ⇒ B/C X_1 & X_2 are Independent!

- Theoretical Probability

$$\int_0^{10} \frac{1}{5} e^{-\frac{1}{5}x_1} dx_1 \cdot \int_0^{10} \frac{1}{5} e^{-\frac{1}{5}x_2} dx_2$$

$$pexp(10, \tfrac{1}{5})^2 = .747$$

- Estimated Probability using Simulation

```
time.to.failure <- rexp(10000, 1/5)

mean(time.to.failure < 10)^2
```

You try it

Customers arrive at a teller's window at a uniform rate 5 per hour. Let X be the length in minutes of time that the teller has to wait until they see their first customer after starting their shift.

a) Calculate the mean and standard deviation of the wait time in minutes using both theoretical formulas and simulation.

b) Find the probability that the teller waits less than 10 minutes for their first customer. Use both theoretical formulas (`pexp`) and simulation (`rexp`).

Additional notes.

Section 5.3 Estimating Continuous Distributions

Sections 5.1 and 5.2 were incorporated into chapter 3 and 4.

Transferring a variable from one scale to another is a problem that is familiar to us. For instance, supposed that we know the temperature in degrees Fahrenheit . We can easily transfer the temperature to degrees Celsius using the formula $C = \frac{5}{9}(F - 32)$. The following function in R will do this.

```r
degrees_F <- 85
(degrees_C <- 5/9*(degrees_F-32))
```

```
## [1] 29.44444
```

An analogous question arises in connection with random variables. Suppose that X is a random variable with a certain probability distribution and Y is another random variable that is some function of X, say $aX + b$ where a and b are constants. What can we say about the distribution of Y?

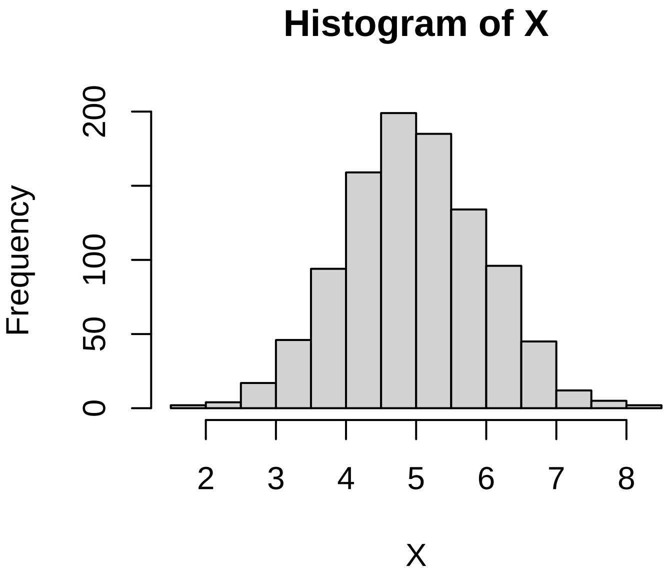
In this section, we will use simulation to answer questions about the distribution of random variables that are transformations or combinations of different random variables.

Example

Suppose X is a normally distributed random variable with mean of 5 and standard deviation of 1. We can easily simulate drawing random samples from this distribution using the function `rnorm`.

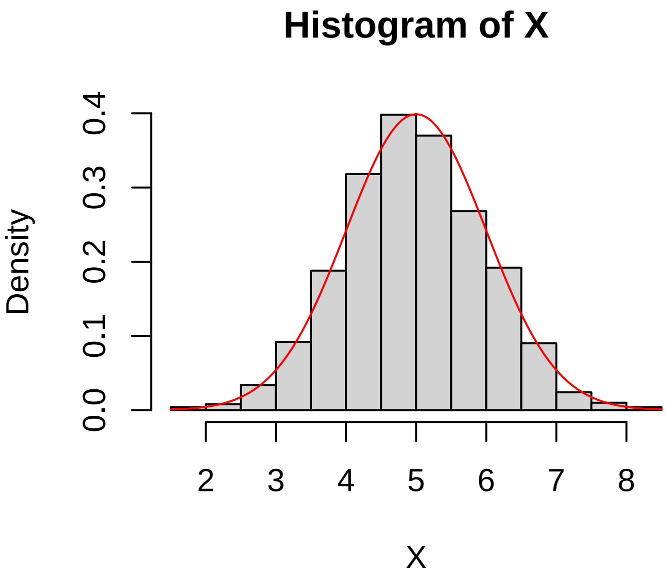
If we viewed a histogram of the X we would expect the histogram to be bell-shaped, centered at around 5. Because the standard deviation is equal to 1 we would expect that almost all values will be between 2 and 8 ($\mu \pm 3\sigma$).

```
X <- rnorm(1000,5,1)
hist(X)
```



For distributions where the pdf is continuous, we may also use a *density estimation*. The height of the density estimation is a weighted sum of the distances to all of the data points in the sample. We can overlay the density curve over the histogram. This will result in a smooth curve and will give us an idea of how a certain distribution “fits” with the simulated data.

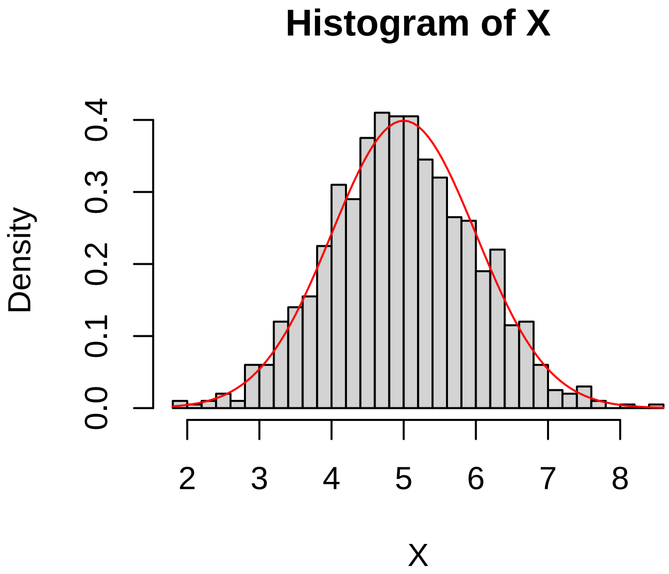
```
hist(X,prob=TRUE)
curve(dnorm(x,5,1),add=TRUE,col="red")
```



Not surprisingly, we can see that the red curve matches the curve in the histogram.

We can change the number of bins to get a closer look at the fit of the simulated data to the normal distribution.

```
hist(X,prob=TRUE,nclass=40)
curve(dnorm(x,5,1),add=TRUE,col="red")
```



Although not perfect, it would be hard to argue that the data that we simulated didn't follow a normal distribution with mean of 5 and standard deviation of 1. Of course since we created the data we know that it does arise from N(5,1). In practice, we don't usually know the exact distribution that data arise from so we want methods to decide whether the data fit with a particular distribution. There are specific tests that we could do called *Goodness of fit tests*, however, simulation to check to see whether a random variable follows a specific distribution.

Now suppose that we have a different random variable Y where $Y = 10X + 7$. What can we say about the distribution of Y? Let's consider what we already know about the distribution.

We know that $E(X) = 5$ and $SD(X) = 1$. Compute $E(Y)$ and $SD(Y)$.

Now that we know the mean and standard deviation of Y we might be interested in the shape of the distribution of Y. We know that X is normally distributed. What might be your best guess as to the distribution of Y?

We can use simulation to see if Y follows a normal distribution centered at 57 with standard deviation of 10. We will do this in 3 steps:

1.

2.

3.

It appears that if we take a linear transformation of a random variable X, say $Y = aX + b$ where a and b are constants and where X follows a normal distribution with mean μ and standard deviation σ, we end up with a random variable Y that also follows a normal distribution with mean $a\mu + b$ and variance $a^2\sigma^2$. More concisely, if $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ where a and b are constants, then $Y \sim N(a\mu + b, a^2\sigma^2)$.

You try it

Let $X \sim N(10, 2)$ and $Y = 3x - 5$. What is the distribution of Y? Show all work.

- Find the theoretical mean and variance:

- Use the 3 steps above to simulate a distribution of Y.

Of course, we are generalizing our example to all linear transformations of normal variables. It turns out that we can prove mathematically that the above result holds fairly easily. Our simulation didn't prove that the above result is true, however, it lead us to believe that perhaps the result holds for all linear transformations of a single variable X.

Transformations of combinations of random variables.

Example

Suppose that Z_1 and Z_2 are independent standard normal variables. Simulate the pdf of $Z_1 - Z_2$ and check whether or not $Z_1 - Z_2$ follows a $N(0, \sqrt{2})$.

Theorem 5.2

Let $X_1, ..., X_n$ be mutually independent normal random variables with means $\mu_1, ..., \mu_n$ and standard deviation $\sigma_1, ..., \sigma_n$. Then the random variable $\sum a_i X_i$ is a normal random variable with mean $\sum a_i \mu_i$ and standard deviation $\sqrt{\sum a_i^2 \sigma_i^2}$.

You try it

1. Let $X \sim N(4, 2)$ and $Y \sim N(-2, 1)$. Find and plot the distribution of $Z = 2x - 4Y$.

2. Let X and Y be independent uniform random variables on the interval [0,1].

a. Guess and write down a hypothesized distribution for $Z = X + Y$.

b. Simulate the pdf of Z and plot it. Sketch the plot in these notes, including a detailed x-axis. Is it as you expected?

c. Describe it's shape. Is this reasonable from a geometric perspective? Explain why.

d. Using the theoretical mean and variance, and the named distribution that best matches the shape above (regardless of your answer in part a), plot a distribution curve over the histogram and describe how well the simulation fits the hypothesized distribution

Exercise

Suppose we have 2 random variables where X_1 and X_2 follows an exponential distribution with $\lambda = 3$. Use simulation to show that $X_1 + X_2$ follows a $\Gamma \sim (2, 3)$.

It turns out that if we have the sum of n exponential random variables with parameter λ, we get……

Section 5.4 The Central Limit Theorem

One of the goals for the field of Statistics is to make inference (a conclusion) about the underlying behavior of a characteristic, based on limited information. For example, researchers may be interested in how many days per week high-school aged people are physically active. If we were to measure the number of active days for every single high school aged person in the entire world (the population) and calculate the mean, we would have our answer. However that is completely infeasible and impractical. We can however, take a representative, random sample of youth and calculate their sample mean to estimate this population value. What are some pros and cons to this approach?

If we want to estimate a population parameter such as the mean, using an estimate calculated on a sample, we need to know how these estimates behave. This section takes pieces of knowledge we've seen before, along with a few new propositions to give us the *Central Limit Theorem* that describes the behavior of the sample mean.

Definition 5.18: iid

First lets remind ourselves of the definition of independence from section 3.3 and write that definition down again.

Example

I want to understand study habits of students, specifically the number of hours they study for an exam. If I were to put all the names of students in this class into a hat and draw 10, would you consider the results from those students to be independent of each other? Why or why not?

What if I put the names of all students at Chico state into the hat and drew 10?

What if the hat contained the names of all college students in the United States?

So we can say that if the population of interest is large enough....

The term *independently and identically distributed (iid)* is an important concept in statistics. If random variables $X_1, X_2, \cdots, X_n$ are all mutually independent and all have the same distribution, then they are called *iid*. When you sample from any of the named distribution functions like `rnorm` or `rexp`, you are drawing samples from iid random variables.

Definition 5.19: Sampling Distribution

The word *statistic* is a generic term used to describe a numeric summary of a sample of data. When we calculate the mean or variance from a sample, this is called a *sample statistic*. E.g.

```
x <- rnorm(1000)
mean(x) # sample mean
```

```
## [1] 0.02791229
```

We know that each time we draw a random sample, we will generate a slightly different sample statistic.

```
replicate(5, {
  mean(rnorm(1000))
})
```

```
## [1]  0.021823621  0.009488512 -0.002086100  0.016092975 -0.008093415
```

The distribution of these sample statistics is called the *sampling distribution*. Knowing the behavior of the sampling distribution is key to making conclusions based on data.

Proposition 5.1

If $X_1, X_2, \cdots X_n$ are iid with mean μ and variance σ^2, then the sample mean $\bar{X} = \frac{1}{n}\sum_i X_i$ has the following mean and variance:

$$E[\bar{X}] = \mu \qquad Var(\bar{X}) = \frac{\sigma^2}{n}$$

Now recall that the standard normal random variable Z is calculated as $Z = \frac{x-\mu}{\sigma}$. We *standardize* the variable x by subtracting by it's mean and dividing by it's standard deviation.

Theorem 5.3: The Central Limit Theorem (CLT)

If $X_1, X_2, \cdots X_n$ are iid with mean μ and variance σ^2, then

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \to Z \qquad \text{as } n \to \infty$$

where Z is the Standard normal random variable.

Example

Let $X_1, X_2, \cdots X_n$ be independent Poisson random variables with rate 2. Assume that $n = 30$ is considered a large enough sample for the CLT to hold. Then the CLT says

Let's look at this via simulation

You try it

Let $X_1, X_2, \cdots X_n$ be independent exponential random variables with rate $1/3$. What is the distribution of the mean from a sample of $n = 50$? Figure this out both theoretically and confirm your results using simulation.

Usefulness of the CLT in practice

In summary, the CLT says,

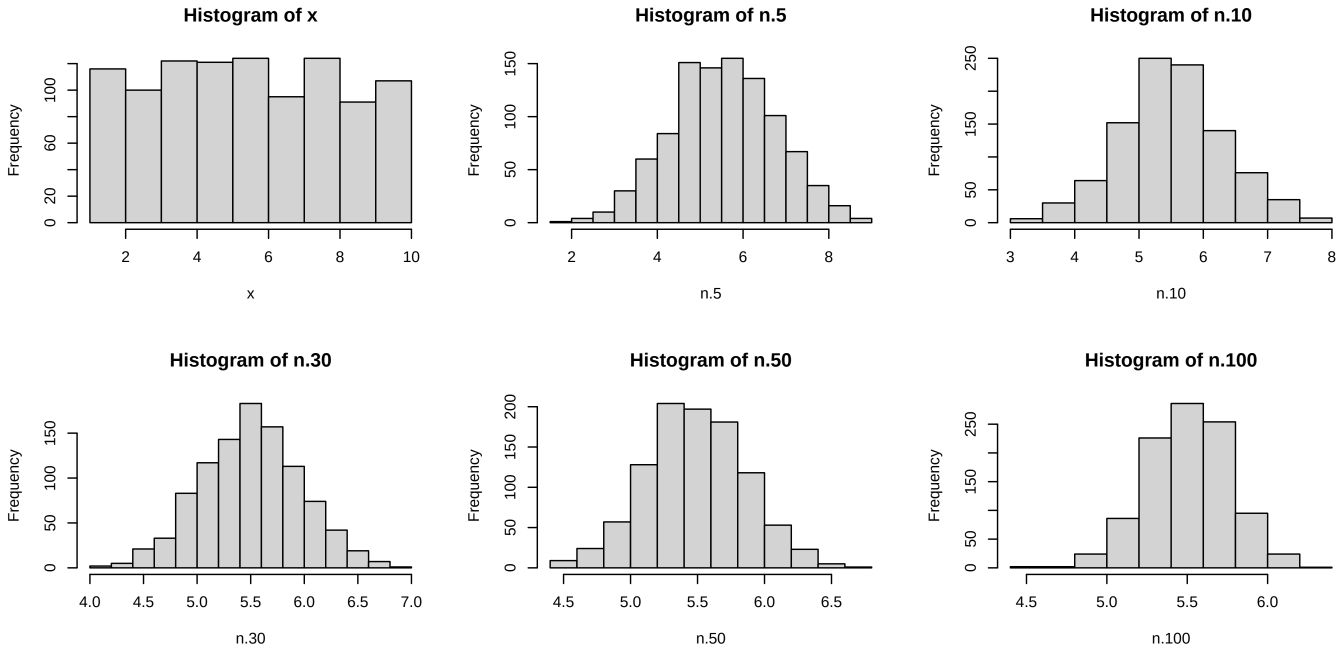
but how large is large?

Example

Let $X_1, X_2, \cdots X_n$ be iid Uniform random variables on the range [1, 10]. What is the distribution of the mean at varying sample sizes?

```r
x <- runif(1000, 1, 10)
n.5   <- replicate(1000, {mean(runif(5,   1, 10)) })
n.10  <- replicate(1000, {mean(runif(10,  1, 10)) })
n.30  <- replicate(1000, {mean(runif(30,  1, 10)) })
n.50  <- replicate(1000, {mean(runif(50,  1, 10)) })
n.100 <- replicate(1000, {mean(runif(100, 1, 10)) })
```

```r
par(mfrow=c(2,3))
hist(x);hist(n.5);hist(n.10); hist(n.30); hist(n.50);hist(n.100)
```

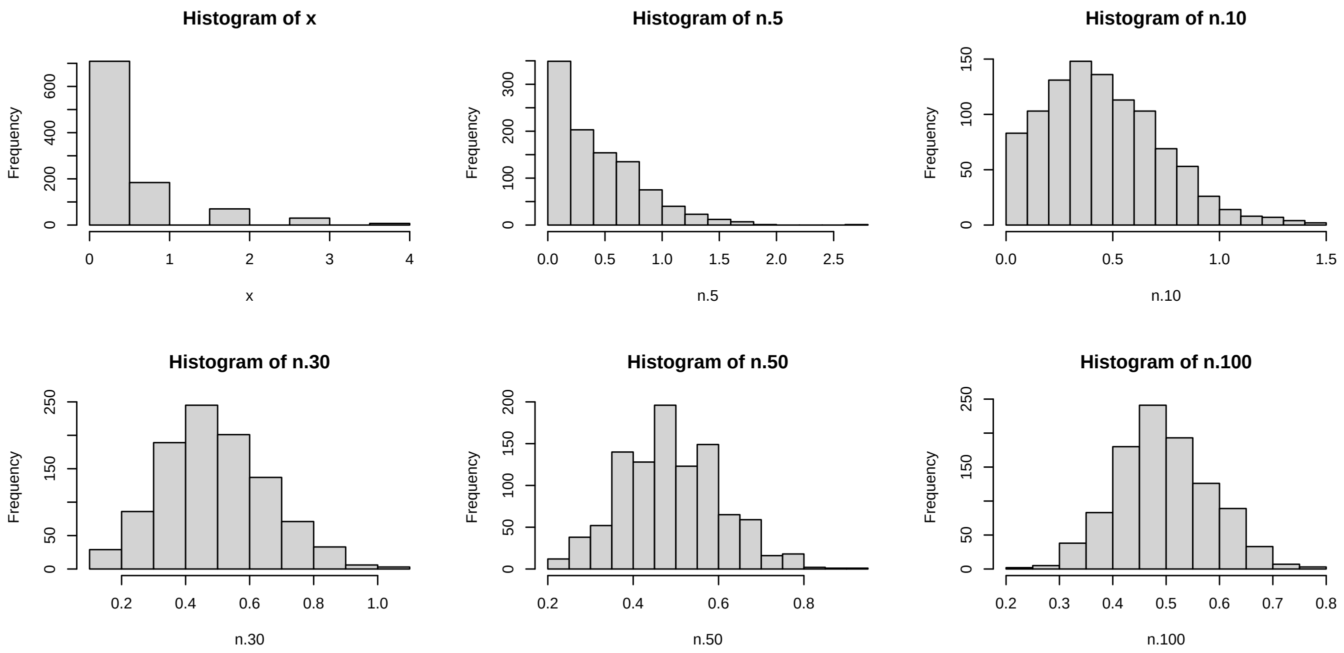


For this distribution,

But what about something that is heavily skewed? Let's look at a Zero-inflated Poisson distribution. This occurs when there is a low probability of an event happening to begin with, but when it does there is a poisson distribution for the number of events. E.g. number of cigarettes each day a person smokes (in 2022).

```r
create.zip <- function(nsamp){
  ifelse(rbinom(nsamp, size = 1, prob = .5) > 0, 0, rpois(nsamp, lambda = 1))
}
x <- create.zip(1000)
n.5   <- replicate(1000, {mean(create.zip(5))})
n.10  <- replicate(1000, {mean(create.zip(10))})
n.30  <- replicate(1000, {mean(create.zip(30))})
n.50  <- replicate(1000, {mean(create.zip(50))})
n.100 <- replicate(1000, {mean(create.zip(100))})
```

```r
par(mfrow=c(2,3))
hist(x);hist(n.5);hist(n.10); hist(n.30); hist(n.50);hist(n.100)
```



You try it

Roughly what is the smallest sample size (n) do we need to say that the mean of n Exponential random variables with rate 1/2 converge to a Normal distribution? Use Proposition 5.1 to figure out what the mean and sd of the sampling distribution should be, and plot a Normal density curve over your final answer to confirm.

Concluding remarks