# Data Science: Probability

Book to use : <https://rafalab.github.io/dsbook/probability.html#discrete-probability>

## Section 1: Discrete Probability

### [Introduction to Discrete Probability](https://courses.edx.org/courses/course-v1:HarvardX+PH125.3x+1T2020/course/#block-v1:HarvardX+PH125.3x+1T2020+type@sequential+block@1976a4117b50412ab72109415c674198)

## Discrete Probability

### Key points

* The probability of an event is the proportion of times the event occurs when we repeat the experiment independently under the same conditions.

Pr(A)=probability of event A

* An event is defined as an outcome that can occur when when something happens by chance.
* We can determine probabilities related to discrete variables (picking a red bead, choosing 48 Democrats and 52 Republicans from 100 likely voters) and continuous variables (height over 6 feet).

## Monte Carlo Simulations

### Key points

* Monte Carlo simulations model the probability of different outcomes by repeating a random process a large enough number of times that the results are similar to what would be observed if the process were repeated forever.
* The sample() function draws random outcomes from a set of options.
* The replicate() function repeats lines of code a set number of times. It is used with sample() and similar functions to run Monte Carlo simulations.

### Video code

Note that your exact outcome values from the Monte Carlo simulation will differ because the sampling is random.

beads <- rep(c("red", "blue"), times = c(2,3)) # create an urn with 2 red, 3 blue

beads # view beads object

sample(beads, 1) # sample 1 bead at random

B <- 10000 # number of times to draw 1 bead

events <- replicate(B, sample(beads, 1)) # draw 1 bead, B times

tab <- table(events) # make a table of outcome counts

tab # view count table

prop.table(tab) # view table of outcome proportions

## Setting the Random Seed

### The set.seed() function

Before we continue, we will briefly explain the following important line of code:

set.seed(1986)

Throughout this book, we use random number generators. This implies that many of the results presented can actually change by chance, which then suggests that a frozen version of the book may show a different result than what you obtain when you try to code as shown in the book. This is actually fine since the results are random and change from time to time. However, if you want to to ensure that results are exactly the same every time you run them, you can set R’s random number generation seed to a specific number. Above we set it to 1986. We want to avoid using the same seed every time. A popular way to pick the seed is the year - month - day. For example, we picked 1986 on December 20, 2018:  2018 − 12 − 20 = 1986.

You can learn more about setting the seed by looking at the documentation:

?set.seed

In the exercises, we may ask you to set the seed to assure that the results you obtain are exactly what we expect them to be.

### Important note on seeds in R 3.5 and R 3.6

R was recently updated to version 3.6 in early 2019. In this update, the default method for setting the seed changed. This means that exercises, videos, textbook excerpts and other code you encounter online may yield a different result based on your version of R.

If you are running R 3.6, you can revert to the original seed setting behavior by adding the argument sample.kind="Rounding". For example:

set.seed(1)

set.seed(1, sample.kind="Rounding") # will make R 3.6 generate a seed as in R 3.5

Using the sample.kind="Rounding" argument will generate a message:

non-uniform 'Rounding' sampler used

This is not a warning or a cause for alarm - it is a confirmation that R is using the alternate seed generation method, and you should expect to receive this message in your console.

## Using the mean Function for Probability

### An important application of the mean() function

In R, applying the mean() function to a logical vector returns the proportion of elements that are TRUE. It is very common to use the mean() function in this way to calculate probabilities and we will do so throughout the course.

Suppose you have the vector beads from a previous video:

beads <- rep(c("red", "blue"), times = c(2,3))

beads

[1] "red" "red" "blue" "blue" "blue"

To find the probability of drawing a blue bead at random, you can run:

mean(beads == "blue")

[1] 0.6

This code is broken down into steps inside R. First, R evaluates the logical statement beads == "blue", which generates the vector:

FALSE FALSE TRUE TRUE TRUE

When the mean function is applied, R coerces the logical values to numeric values, changing TRUE to 1 and FALSE to 0:

0 0 1 1 1

The mean of the zeros and ones thus gives the proportion of TRUE values. As we have learned and will continue to see, probabilities are directly related to the proportion of events that satisfy a requirement.

## Probability Distributions

### Key points

* The probability distribution for a variable describes the probability of observing each possible outcome.
* For discrete categorical variables, the probability distribution is defined by the proportions for each group.

## Independence

### Errors and clarifications

At 4:32 in the video, it should be 16 out of 51, not 12 out of 51. While the audio is incorrect, the transcript and the text on screen both now have the correct number.

Note that this calculation only applies to getting blackjack in the order "Ace, Face/10". We consider the full probability including the possible order "Face/10, Ace" later when discussing the addition rule.

### Key points

* Conditional probabilities compute the probability that an event occurs given information about dependent events. For example, the probability of drawing a second king given that the first draw is a king is:

Pr(Card 2 is a king∣Card 1 is a king)=3/51

* If two events A and B are independent, Pr(A∣B)=Pr(A).
* To determine the probability of multiple events occurring, we use the multiplication rule.

### Equations

The multiplication rule for independent events is:

Pr(A and B and C)=Pr(A)×Pr(B)×Pr(C)

The multiplication rule for dependent events considers the conditional probability of both events occurring:

Pr(A and B)=Pr(A)×Pr(B∣A)

We can expand the multiplication rule for dependent events to more than 2 events:

Pr(A and B and C)=Pr(A)×Pr(B∣A)×Pr(C∣A and B)

## Assessment: Introduction to Discrete Probability

### **Probability of cyan**

1/1 point (graded)

One ball will be drawn at random from a box containing: 3 cyan balls, 5 magenta balls, and 7 yellow balls.

What is the probability that the ball will be cyan?  correct

0.2 Loading

Submit

You have used 1 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Probability of not cyan**

1/1 point (graded)

One ball will be drawn at random from a box containing: 3 cyan balls, 5 magenta balls, and 7 yellow balls.

What is the probability that the ball will not be cyan?  correct

0.8 Loading

Submit

You have used 1 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Sampling without replacement**

1/1 point (graded)

Instead of taking just one draw, consider taking two draws. You take the second draw without returning the first draw to the box. We call this sampling without replacement.

What is the probability that the first draw is cyan and that the second draw is not cyan?

Provide at least 3 significant digits.

  correct

0.171 Loading

Submit

You have used 1 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Sampling with replacement**

1/1 point (graded)

Now repeat the experiment, but this time, after taking the first draw and recording the color, return it back to the box and shake the box. We call this sampling with replacement.

What is the probability that the first draw is cyan and that the second draw is not cyan?  correct

0.16 Loading

Submit

You have used 1 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

### [Combinations and Permutations](https://courses.edx.org/courses/course-v1:HarvardX+PH125.3x+1T2020/course/#block-v1:HarvardX+PH125.3x+1T2020+type@sequential+block@b1ea2c318d8e495eaa7d51942b6626f6)

## Combinations and Permutations

### Key points

* paste() joins two strings and inserts a space in between.
* expand.grid() gives the combinations of 2 vectors or lists.
* permutations(n,r) from the **gtools** package lists the different ways that r items can be selected from a set of n options when order matters.
* combinations(n,r) from the **gtools** package lists the different ways that r items can be selected from a set of n options when order does not matter.

### Code: Introducing paste() and expand.grid()

# joining strings with paste

number <- "Three"

suit <- "Hearts"

paste(number, suit)

# joining vectors element-wise with paste

paste(letters[1:5], as.character(1:5))

# generating combinations of 2 vectors with expand.grid

expand.grid(pants = c("blue", "black"), shirt = c("white", "grey", "plaid"))

### Code: Generating a deck of cards

suits <- c("Diamonds", "Clubs", "Hearts", "Spades")

numbers <- c("Ace", "Deuce", "Three", "Four", "Five", "Six", "Seven", "Eight", "Nine", "Ten", "Jack", "Queen", "King")

deck <- expand.grid(number = numbers, suit = suits)

deck <- paste(deck$number, deck$suit)

# probability of drawing a king

kings <- paste("King", suits)

mean(deck %in% kings)

### Code: Permutations and combinations

Correction: The code shown does not generate all 7 digit phone numbers because phone numbers can have repeated digits. It generates all possible 7 digit numbers without repeats.

library(gtools)

permutations(5,2) # ways to choose 2 numbers in order from 1:5

all\_phone\_numbers <- permutations(10, 7, v = 0:9)

n <- nrow(all\_phone\_numbers)

index <- sample(n, 5)

all\_phone\_numbers[index,]

permutations(3,2) # order matters

combinations(3,2) # order does not matter

### Code: Probability of drawing a second king given that one king is drawn

hands <- permutations(52,2, v = deck)

first\_card <- hands[,1]

second\_card <- hands[,2]

sum(first\_card %in% kings)

sum(first\_card %in% kings & second\_card %in% kings) / sum(first\_card %in% kings)

### Code: Probability of a natural 21 in blackjack

aces <- paste("Ace", suits)

facecard <- c("King", "Queen", "Jack", "Ten")

facecard <- expand.grid(number = facecard, suit = suits)

facecard <- paste(facecard$number, facecard$suit)

hands <- combinations(52, 2, v=deck) # all possible hands

# probability of a natural 21 given that the ace is listed first in `combinations`

mean(hands[,1] %in% aces & hands[,2] %in% facecard)

# probability of a natural 21 checking for both ace first and ace second

mean((hands[,1] %in% aces & hands[,2] %in% facecard)|(hands[,2] %in% aces & hands[,1] %in% facecard))

### Code: Monte Carlo simulation of natural 21 in blackjack

Note that your exact values will differ because the process is random and the seed is not set.

# code for one hand of blackjack

hand <- sample(deck, 2)

hand

# code for B=10,000 hands of blackjack

B <- 10000

results <- replicate(B, {

hand <- sample(deck, 2)

(hand[1] %in% aces & hand[2] %in% facecard) | (hand[2] %in% aces & hand[1] %in% facecard)

})

mean(results)

## The Birthday Problem

### Key points

* duplicated() takes a vector and returns a vector of the same length with TRUE for any elements that have appeared previously in that vector.
* We can compute the probability of shared birthdays in a group of people by modeling birthdays as random draws from the numbers 1 through 365. We can then use this sampling model of birthdays to run a Monte Carlo simulation to estimate the probability of shared birthdays.

### Code: The birthday problem

# checking for duplicated bdays in one 50 person group

n <- 50

bdays <- sample(1:365, n, replace = TRUE) # generate n random birthdays

any(duplicated(bdays)) # check if any birthdays are duplicated

# Monte Carlo simulation with B=10000 replicates

B <- 10000

results <- replicate(B, { # returns vector of B logical values

bdays <- sample(1:365, n, replace = TRUE)

any(duplicated(bdays))

})

mean(results) # calculates proportion of groups with duplicated bdays

## sapply

### Key points

* Some functions automatically apply element-wise to vectors, such as sqrt() and \*.
* However, other functions do not operate element-wise by default. This includes functions we define ourselves.
* The function sapply(x, f) allows any other function f to be applied element-wise to the vector x.
* The probability of an event happening is 1 minus the probability of that event not happening:

Pr(event)=1−Pr(no event)

* We can compute the probability of shared birthdays mathematically:

 Pr(shared birthdays)=1−Pr(no shared birthdays)=1−(1×364365×363365×...×365−n+1365)

### Code: Function for birthday problem Monte Carlo simulations

Note that the function body of compute\_prob() is the code that we wrote in the previous video. If we write this code as a function, we can use sapply() to apply this function to several values of n.

# function to calculate probability of shared bdays across n people

compute\_prob <- function(n, B = 10000) {

same\_day <- replicate(B, {

bdays <- sample(1:365, n, replace = TRUE)

any(duplicated(bdays))

})

mean(same\_day)

}

n <- seq(1, 60)

### Code: Element-wise operation over vectors and sapply

x <- 1:10

sqrt(x) # sqrt operates on each element of the vector

y <- 1:10

x\*y # \* operates element-wise on both vectors

compute\_prob(n) # does not iterate over the vector n without sapply

x <- 1:10

sapply(x, sqrt) # this is equivalent to sqrt(x)

prob <- sapply(n, compute\_prob) # element-wise application of compute\_prob to n

plot(n, prob)

### Code: Computing birthday problem probabilities with sapply

# function for computing exact probability of shared birthdays for any n

exact\_prob <- function(n){

prob\_unique <- seq(365, 365-n+1)/365 # vector of fractions for mult. rule

1 - prod(prob\_unique) # calculate prob of no shared birthdays and subtract from 1

}

# applying function element-wise to vector of n values

eprob <- sapply(n, exact\_prob)

# plotting Monte Carlo results and exact probabilities on same graph

plot(n, prob) # plot Monte Carlo results

lines(n, eprob, col = "red") # add line for exact prob

## How Many Monte Carlo Experiments are Enough?

### Key points

* The larger the number of Monte Carlo replicates B, the more accurate the estimate.
* Determining the appropriate size for B can require advanced statistics.
* One practical approach is to try many sizes for B and look for sizes that provide stable estimates.

### Code: Estimating a practical value of B

This code runs Monte Carlo simulations to estimate the probability of shared birthdays using several B values and plots the results. When B is large enough that the estimated probability stays stable, then we have selected a useful value of B.

B <- 10^seq(1, 5, len = 100) # defines vector of many B values

compute\_prob <- function(B, n = 22){ # function to run Monte Carlo simulation with each B

same\_day <- replicate(B, {

bdays <- sample(1:365, n, replace = TRUE)

any(duplicated(bdays))

})

mean(same\_day)

}

prob <- sapply(B, compute\_prob) # apply compute\_prob to many values of B

plot(log10(B), prob, type = "l") # plot a line graph of estimates

### [Addition Rule and Monty Hall](https://courses.edx.org/courses/course-v1:HarvardX+PH125.3x+1T2020/course/#block-v1:HarvardX+PH125.3x+1T2020+type@sequential+block@f213f32e527043ec8c0718cf398fae49)

## The Addition Rule

### Key points

* The addition rule states that the probability of event A or event B happening is the probability of event A plus the probability of event B minus the probability of both events A and B happening together.

Pr(A or B)=Pr(A)+Pr(B)−Pr(A and B)

* Note that (A or B) is equivalent to (A|B).

### Example: The addition rule for a natural 21 in blackjack

We apply the addition rule where A = drawing an ace then a facecard and B = drawing a facecard then an ace. Note that in this case, both events A and B cannot happen at the same time, so Pr(A and B)=0.

Pr(ace then facecard)=4/52×16/51

Pr(facecard then ace)=16/52×4/51

Pr(ace then facecard | facecard then ace)=4/52×16/51+16/52×4/51=0.0483

## The Monty Hall Problem

### Key points

* Monte Carlo simulations can be used to simulate random outcomes, which makes them useful when exploring ambiguous or less intuitive problems like the Monty Hall problem.
* In the Monty Hall problem, contestants choose one of three doors that may contain a prize. Then, one of the doors that was not chosen by the contestant and does not contain a prize is revealed. The contestant can then choose whether to stick with the original choice or switch to the remaining unopened door.
* Although it may seem intuitively like the contestant has a 1 in 2 chance of winning regardless of whether they stick or switch, Monte Carlo simulations demonstrate that the actual probability of winning is 1 in 3 with the stick strategy and 2 in 3 with the switch strategy.
* For more on the Monty Hall problem, you can [watch a detailed explanation here](https://www.khanacademy.org/math/precalculus/prob-comb/dependent-events-precalc/v/monty-hall-problem) or r[ead an explanation here](https://en.wikipedia.org/wiki/Monty_Hall_problem).

### Code: Monte Carlo simulation of stick strategy

B <- 10000

stick <- replicate(B, {

doors <- as.character(1:3)

prize <- sample(c("car","goat","goat")) # puts prizes in random order

prize\_door <- doors[prize == "car"] # note which door has prize

my\_pick <- sample(doors, 1) # note which door is chosen

show <- sample(doors[!doors %in% c(my\_pick, prize\_door)],1) # open door with no prize that isn't chosen

stick <- my\_pick # stick with original door

stick == prize\_door # test whether the original door has the prize

})

mean(stick) # probability of choosing prize door when sticking

### Code: Monte Carlo simulation of switch strategy

switch <- replicate(B, {

doors <- as.character(1:3)

prize <- sample(c("car","goat","goat")) # puts prizes in random order

prize\_door <- doors[prize == "car"] # note which door has prize

my\_pick <- sample(doors, 1) # note which door is chosen first

show <- sample(doors[!doors %in% c(my\_pick, prize\_door)], 1) # open door with no prize that isn't chosen

switch <- doors[!doors%in%c(my\_pick, show)] # switch to the door that wasn't chosen first or opened

switch == prize\_door # test whether the switched door has the prize

})

mean(switch) # probability of choosing prize door when switching

* 1. [Assessment: Discrete Probability](https://courses.edx.org/courses/course-v1:HarvardX+PH125.3x+1T2020/course/#block-v1:HarvardX+PH125.3x+1T2020+type@sequential+block@5fb5c4bd901649c78789d9bd467d163a)

## Introduction

library(gtools)

library(tidyverse)

## Question 1: Olympic running

In the 200m dash finals in the Olympics, 8 runners compete for 3 medals (order matters). In the 2012 Olympics, 3 of the 8 runners were from Jamaica and the other 5 were from different countries. The three medals were all won by Jamaica (Usain Bolt, Yohan Blake, and Warren Weir).

Use the information above to help you answer the following four questions.

### **Question 1a**

1.0/1.0 point (graded)

How many different ways can the 3 medals be distributed across 8 runners?  correct

336 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1b**

1.0/1.0 point (graded)

How many different ways can the three medals be distributed among the 3 runners from Jamaica?  correct

6 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1c**

1.0/1.0 point (graded)

What is the probability that all 3 medals are won by Jamaica?  correct

0.017 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1d**

1.0/1.0 point (graded)

Run a Monte Carlo simulation on this vector representing the countries of the 8 runners in this race:

runners <- c("Jamaica", "Jamaica", "Jamaica", "USA", "Ecuador", "Netherlands", "France", "South Africa")

For each iteration of the Monte Carlo simulation, within a replicate() loop, select 3 runners representing the 3 medalists and check whether they are all from Jamaica. Repeat this simulation 10,000 times. Set the seed to 1 before running the loop.

Calculate the probability that all the runners are from Jamaica.  correct

0.016 Loading

Submit

You have used 3 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Question 2: Restaurant management

A restaurant manager wants to advertise that his lunch special offers enough choices to eat different meals every day of the year. He doesn't think his current special actually allows that number of choices, but wants to change his special if needed to allow at least 365 choices.

A meal at the restaurant includes 1 entree, 2 sides, and 1 drink. He currently offers a choice of 1 entree from a list of 6 options, a choice of 2 different sides from a list of 6 options, and a choice of 1 drink from a list of 2 options.

### **Question 2a**

1.0/1.0 point (graded)

How many meal combinations are possible with the current menu?  correct

180 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 2b**

1.0/1.0 point (graded)

The manager has one additional drink he could add to the special.

How many combinations are possible if he expands his original special to 3 drink options?  correct

270 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 2c**

1.0/1.0 point (graded)

The manager decides to add the third drink but needs to expand the number of options. The manager would prefer not to change his menu further and wants to know if he can meet his goal by letting customers choose more sides.

How many meal combinations are there if customers can choose from 6 entrees, 3 drinks, and select 3 sides from the current 6 options?  correct

360 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 2d**

1.0/1.0 point (graded)

The manager is concerned that customers may not want 3 sides with their meal. He is willing to increase the number of entree choices instead, but if he adds too many expensive options it could eat into profits. He wants to know how many entree choices he would have to offer in order to meet his goal.

- Write a function that takes a number of entree choices and returns the number of meal combinations possible given that number of entree options, 3 drink choices, and a selection of 2 sides from 6 options.

- Use sapply() to apply the function to entree option counts ranging from 1 to 12.

What is the minimum number of entree options required in order to generate more than 365 combinations?  correct

9 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 2e**

1.0/1.0 point (graded)

The manager isn't sure he can afford to put that many entree choices on the lunch menu and thinks it would be cheaper for him to expand the number of sides. He wants to know how many sides he would have to offer to meet his goal of at least 365 combinations.

- Write a function that takes a number of side choices and returns the number of meal combinations possible given 6 entree choices, 3 drink choices, and a selection of 2 sides from the specified number of side choices.

- Use sapply() to apply the function to side counts ranging from 2 to 12.

What is the minimum number of side options required in order to generate more than 365 combinations?  correct

7 Loading

Submit

You have used 3 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Questions 3 and 4: Esophageal cancer and alcohol/tobacco use, part 1

Case-control studies help determine whether certain exposures are associated with outcomes such as developing cancer. The built-in dataset esoph contains data from a case-control study in France comparing people with esophageal cancer (cases, counted in ncases) to people without esophageal cancer (controls, counted in ncontrols) that are carefully matched on a variety of demographic and medical characteristics. The study compares alcohol intake in grams per day (alcgp) and tobacco intake in grams per day (tobgp) across cases and controls grouped by age range (agegp).

The dataset is available in base R and can be called with the variable name esoph:

head(esoph)

You will be using this dataset to answer the following four multi-part questions (Questions 3-6).

You may wish to use the **tidyverse** package:

library(tidyverse)

The following three parts have you explore some basic characteristics of the dataset.

Each row contains one group of the experiment. Each group has a different combination of age, alcohol consumption, and tobacco consumption. The number of cancer cases and number of controls (individuals without cancer) are reported for each group.

### **Question 3a**

1.0/1.0 point (graded)

How many groups are in the study?  correct

88 Loading

Submit

You have used 5 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3b**

1.0/1.0 point (graded)

How many cases are there?

Save this value as all\_cases for later problems.

  correct

200 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3c**

1.0/1.0 point (graded)

How many controls are there?

Save this value as all\_controls for later problems.

  correct

975 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

The following four parts ask you to explore some probabilities within this dataset related to alcohol and tobacco consumption.

### **Question 4a**

1.0/1.0 point (graded)

What is the probability that a subject in the highest alcohol consumption group is a cancer case?  correct

0.401 Loading

Submit

You have used 4 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 4b**

1.0/1.0 point (graded)

What is the probability that a subject in the lowest alcohol consumption group is a cancer case?  correct

0.065 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 4c**

1.0/1.0 point (graded)

Given that a person is a case, what is the probability that they smoke 10g or more a day?  correct

0.61 Loading

Submit

You have used 6 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 4d**

1.0/1.0 point (graded)

Given that a person is a control, what is the probability that they smoke 10g or more a day?  correct

0.461 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Questions 5 and 6: Esophageal cancer and alcohol/tobacco use, part 2

### **Question 5a**

1.0/1.0 point (graded)

For cases, what is the probability of being in the highest alcohol group?  correct

0.225 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 5b**

1.0/1.0 point (graded)

For cases, what is the probability of being in the highest tobacco group?  correct

0.155 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 5c**

1.0/1.0 point (graded)

For cases, what is the probability of being in the highest alcohol group **and** the highest tobacco group?  correct

0.05 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 5d**

1.0/1.0 point (graded)

For cases, what is the probability of being in the highest alcohol group **or** the highest tobacco group?  correct

0.33 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

The following six parts look at probabilities related to alcohol and tobacco consumption among the controls and also compare the cases and the controls.

### **Question 6a**

1.0/1.0 point (graded)

For controls, what is the probability of being in the highest alcohol group?  correct

0.068 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 6b**

1.0/1.0 point (graded)

How many times more likely are cases than controls to be in the highest alcohol group?  correct

3.274 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 6c**

1.0/1.0 point (graded)

For controls, what is the probability of being in the highest tobacco group?  correct

0.084 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 6d**

1.0/1.0 point (graded)

For controls, what is the probability of being in the highest alcohol group **and** the highest tobacco group?  correct

0.013 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 6e**

1.0/1.0 point (graded)

For controls, what is the probability of being in the highest alcohol group **or** the highest tobacco group?  correct

0.139 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 6f**

1.0/1.0 point (graded)

How many times more likely are cases than controls to be in the highest alcohol group **or** the highest tobacco group?  correct

2.365809 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Section 2: Continuous probability

### [2.1 Continuous Probability](https://courses.edx.org/courses/course-v1:HarvardX+PH125.3x+1T2020/course/#block-v1:HarvardX+PH125.3x+1T2020+type@sequential+block@a90cf02b5930452b94f4f18f1a3f91ef)

## Continuous Probability

### Key points

* The cumulative distribution function (CDF) is a distribution function for continuous data x that reports the proportion of the data below a for all values of a:

F(a)=Pr(x≤a)

* The CDF is the probability distribution function for continuous variables. For example, to determine the probability that a male student is taller than 70.5 inches given a vector of male heights x, we can use the CDF:

Pr(x>70.5)=1−Pr(x≤70.5)=1−F(70.5)

* The probability that an observation is in between two values a,b is F(b)−F(a).

### Code: Cumulative distribution function

Define x as male heights from the **dslabs** heights dataset:

library(tidyverse)

library(dslabs)

data(heights)

x <- heights %>% filter(sex=="Male") %>% pull(height)

Given a vector **x**, we can define a function for computing the CDF of **x** using:

F <- function(a) mean(x <= a)

1 - F(70) # probability of male taller than 70 inches

## Theoretical Distribution

### Key points

* pnorm(a, avg, s) gives the value of the cumulative distribution function F(a) for the normal distribution defined by average avg and standard deviation s.
* We say that a random quantity is normally distributed with average avg and standard deviation s if the approximation pnorm(a, avg, s) holds for all values of a.
* If we are willing to use the normal approximation for height, we can estimate the distribution simply from the mean and standard deviation of our values.
* If we treat the height data as discrete rather than categorical, we see that the data are not very useful because integer values are more common than expected due to rounding. This is called discretization.
* With rounded data, the normal approximation is particularly useful when computing probabilities of intervals of length 1 that include exactly one integer.

### Code: Using pnorm() to calculate probabilities

Given male heights x:

library(tidyverse)

library(dslabs)

data(heights)

x <- heights %>% filter(sex=="Male") %>% pull(height)

We can estimate the probability that a male is taller than 70.5 inches using:

1 - pnorm(70.5, mean(x), sd(x))

### Code: Discretization and the normal approximation

# plot distribution of exact heights in data

plot(prop.table(table(x)), xlab = "a = Height in inches", ylab = "Pr(x = a)")

# probabilities in actual data over length 1 ranges containing an integer

mean(x <= 68.5) - mean(x <= 67.5)

mean(x <= 69.5) - mean(x <= 68.5)

mean(x <= 70.5) - mean(x <= 69.5)

# probabilities in normal approximation match well

pnorm(68.5, mean(x), sd(x)) - pnorm(67.5, mean(x), sd(x))

pnorm(69.5, mean(x), sd(x)) - pnorm(68.5, mean(x), sd(x))

pnorm(70.5, mean(x), sd(x)) - pnorm(69.5, mean(x), sd(x))

# probabilities in actual data over other ranges don't match normal approx as well

mean(x <= 70.9) - mean(x <= 70.1)

pnorm(70.9, mean(x), sd(x)) - pnorm(70.1, mean(x), sd(x))

## Probability Density

### Key points

* The probability of a single value is not defined for a continuous distribution.
* The quantity with the most similar interpretation to the probability of a single value is the probability density function f(x).
* The probability density f(x) is defined such that the integral of f(x) over a range gives the CDF of that range.

F(a)=Pr(X≤a)=∫a−∞f(x)dx

* In R, the probability density function for the normal distribution is given by dnorm(). We will see uses of dnorm() in the future.
* Note that dnorm() gives the density function and pnorm() gives the distribution function, which is the integral of the density function.

## Plotting the Probability Density

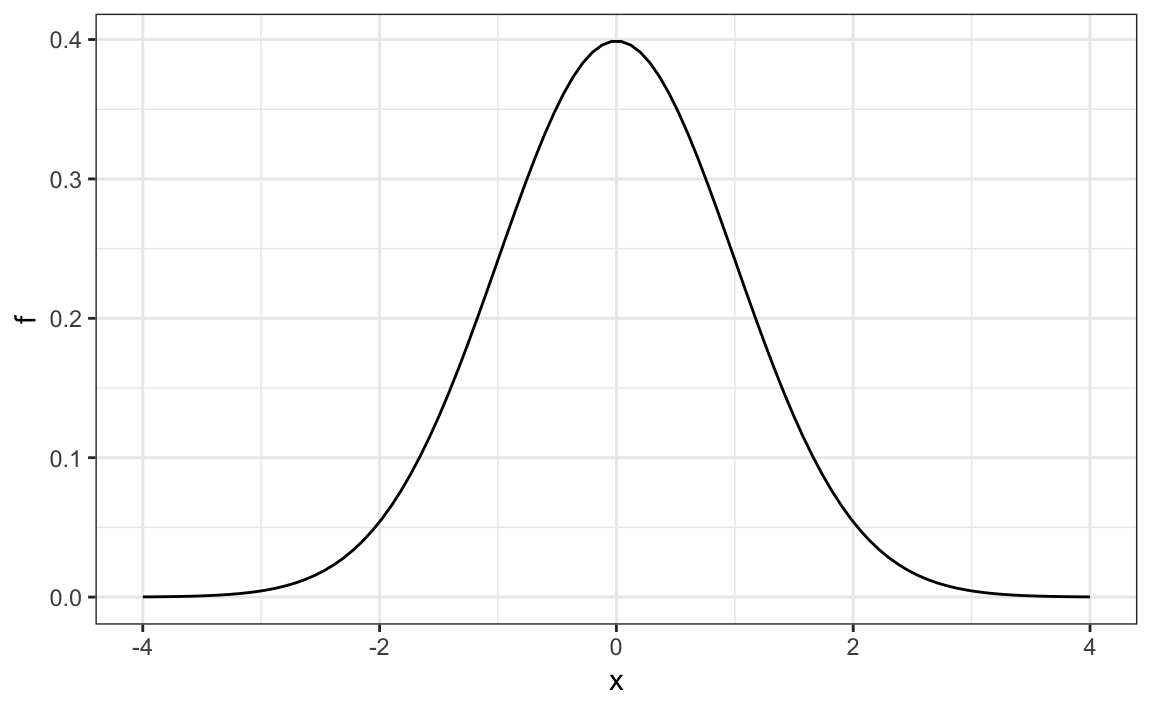
### Plotting the probability density for the normal distribution

We can use dnorm() to plot the density curve for the normal distribution. dnorm(z) gives the probability density f(z) of a certain z-score, so we can draw a curve by calculating the density over a range of possible values of z.

First, we generate a series of z-scores covering the typical range of the normal distribution. Since we know 99.7% of observations will be within −3≤z≤3, we can use a value of z slightly larger than 3 and this will cover most likely values of the normal distribution. Then, we calculate f(z), which is dnorm() of the series of z-scores. Last, we plot z against f(z).

library(tidyverse)  
x <- seq(-4, 4, length = 100)  
data.frame(x, f = dnorm(x)) %>%  
 ggplot(aes(x, f)) +  
 geom\_line()

Here is the resulting plot:



Note that dnorm() gives densities for the standard normal distribution by default. Probabilities for alternative normal distributions with mean mu and standard deviation sigma can be evaluated with:

dnorm(z, mu, sigma)

## Monte Carlo Simulations

### Key points

* rnorm(n, avg, s) generates n random numbers from the normal distribution with average avg and standard deviation s.
* By generating random numbers from the normal distribution, we can simulate height data with similar properties to our dataset. Here we generate simulated height data using the normal distribution.

### Code: Generating normally distributed random numbers

# define x as male heights from dslabs data

library(tidyverse)

library(dslabs)

data(heights)

x <- heights %>% filter(sex=="Male") %>% pull(height)

# generate simulated height data using normal distribution - both datasets should have n observations

n <- length(x)

avg <- mean(x)

s <- sd(x)

simulated\_heights <- rnorm(n, avg, s)

# plot distribution of simulated\_heights

data.frame(simulated\_heights = simulated\_heights) %>%

ggplot(aes(simulated\_heights)) +

geom\_histogram(color="black", binwidth = 2)

### Code: Monte Carlo simulation of tallest person over 7 feet

B <- 10000

tallest <- replicate(B, {

simulated\_data <- rnorm(800, avg, s) # generate 800 normally distributed random heights

max(simulated\_data) # determine the tallest height

})

mean(tallest >= 7\*12) # proportion of times that tallest person exceeded 7 feet (84 inches)

## Other Continuous Distributions

### Key points

* You may encounter other continuous distributions (Student t, chi-squared, exponential, gamma, beta, etc.).
* R provides functions for density (d), quantile (q), probability distribution (p) and random number generation (r) for many of these distributions.
* Each distribution has a matching abbreviation (for example, norm() or t()) that is paired with the related function abbreviations (**d**, **p**, **q**, **r**) to create appropriate functions.
* For example, use rt() to generate random numbers for a Monte Carlo simulation using the Student t distribution.

### Code: Plotting the normal distribution with dnorm

Use d to plot the density function of a continuous distribution. Here is the density function for the normal distribution (abbreviation norm()):

x <- seq(-4, 4, length.out = 100)

data.frame(x, f = dnorm(x)) %>%

ggplot(aes(x,f)) +

geom\_line()

### [2.2 Assessment: Continuous Probability](https://courses.edx.org/courses/course-v1:HarvardX+PH125.3x+1T2020/course/#block-v1:HarvardX+PH125.3x+1T2020+type@sequential+block@bf6b2a8baf1e403b8117fd03f7263d66)

## Questions 1 and 2: ACT scores, part 1

The ACT is a standardized college admissions test used in the United States. The four multi-part questions in this assessment all involve simulating some ACT test scores and answering probability questions about them.

For the three year period 2016-2018, [ACT standardized test scores](http://www.act.org/content/act/en/products-and-services/the-act/scores/national-ranks.html) were approximately normally distributed with a mean of 20.9 and standard deviation of 5.7. (Real ACT scores are integers between 1 and 36, but we will ignore this detail and use continuous values instead.)

First we'll simulate an ACT test score dataset and answer some questions about it.

Set the seed to 16, then use rnorm() to generate a normal distribution of 10000 tests with a mean of 20.9 and standard deviation of 5.7. Save these values as act\_scores. You'll be using this dataset throughout these four multi-part questions.

**(IMPORTANT NOTE! If you use R 3.6 or later, you will need to use the command format set.seed(x, sample.kind = "Rounding") instead of set.seed(x). Your R version will be printed at the top of the Console window when you start RStudio.)**

### **Question 1a**

1.0/1.0 point (graded)

What is the mean of act\_scores?  correct

20.84012 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1b**

1.0/1.0 point (graded)

What is the standard deviation of act\_scores?  correct

5.675237 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1c**

1.0/1.0 point (graded)

A perfect score is 36 or greater (the maximum reported score is 36).

In act\_scores, how many perfect scores are there out of 10,000 simulated tests?  correct

41 Loading

Submit

You have used 5 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1d**

1.0/1.0 point (graded)

In act\_scores, what is the probability of an ACT score greater than 30?  correct

0.052 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1e**

1.0/1.0 point (graded)

In act\_scores, what is the probability of an ACT score less than or equal to 10?  correct

0.028 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

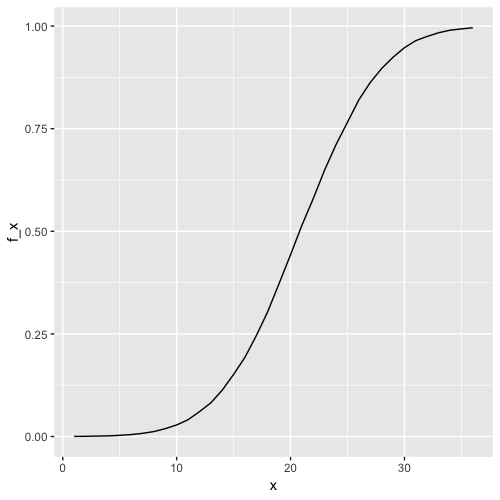
SaveSave Your Answer Show Answer

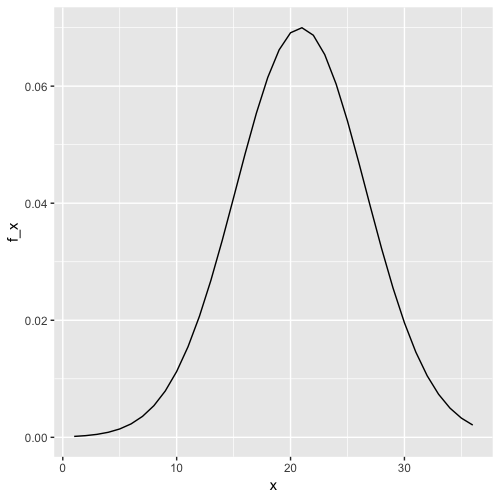
### **Question 2**

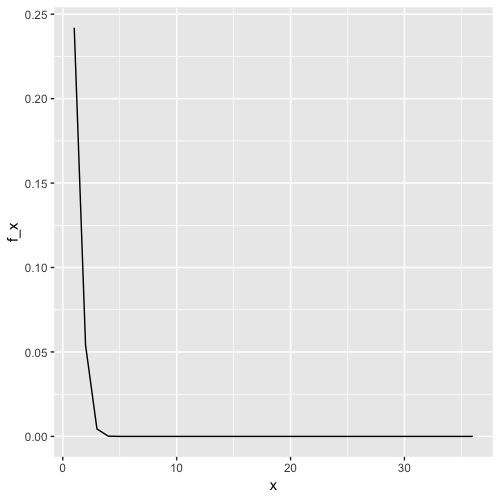
1/1 point (graded)

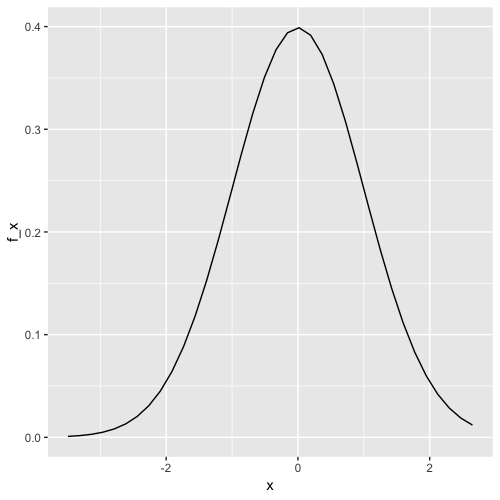
Set x equal to the sequence of integers 1 to 36. Use dnorm to determine the value of the probability density function over x given a mean of 20.9 and standard deviation of 5.7; save the result as f\_x. Plot x against f\_x.

Which of the following plots is correct?









correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Questions 3 and 4: ACT scores, part 2

Convert act\_scores to Z-scores. Recall from [Data Visualization](https://www.edx.org/course/r-data-visualization-2) (the second course in this series) that to standardize values (convert values into Z-scores, that is, values distributed with a mean of 0 and standard deviation of 1), you must subtract the mean and then divide by the standard deviation. Use the mean and standard deviation of act\_scores, not the original values used to generate random test scores.

### **Question 3a**

1.0/1.0 point (graded)

What is the probability of a Z-score greater than 2 (2 standard deviations above the mean)?  correct

0.0227 Loading

Submit

You have used 9 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3b**

1.0/1.0 point (graded)

What ACT score value corresponds to 2 standard deviations above the mean (Z = 2)?  correct

32.3 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3c**

1.0/1.0 point (graded)

A Z-score of 2 corresponds roughly to the 97.5th percentile.

Use qnorm() to determine the 97.5th percentile of normally distributed data with the mean and standard deviation observed in act\_scores.

What is the 97.5th percentile of act\_scores?  correct

32.07179 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

In this 4-part question, you will write a function to create a CDF for ACT scores.

Write a function that takes a value and produces the probability of an ACT score less than or equal to that value (the CDF). Apply this function to the range 1 to 36.

### **Question 4a**

1.0/1.0 point (graded)

What is the minimum integer score such that the probability of that score or lower is at least .95?

Your answer should be an integer 1-36.

  correct

31 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 4b**

1.0/1.0 point (graded)

Use qnorm() to determine the expected 95th percentile, the value for which the probability of receiving that score or lower is 0.95, given a mean score of 20.9 and standard deviation of 5.7.

What is the expected 95th percentile of ACT scores?  correct

30.27 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 4c**

1.0/1.0 point (graded)

As discussed in the Data Visualization course, we can use quantile() to determine sample quantiles from the data.

Make a vector containing the quantiles for p <- seq(0.01, 0.99, 0.01), the 1st through 99th percentiles of the act\_scores data. Save these as sample\_quantiles.

In what percentile is a score of 26?

Your answer should be an integer (i.e. 60), not a percent or fraction. Note that a score between the 98th and 99th percentile should be considered the 98th percentile, for example, and that quantile numbers are used as names for the vector sample\_quantiles.

  correct

82 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

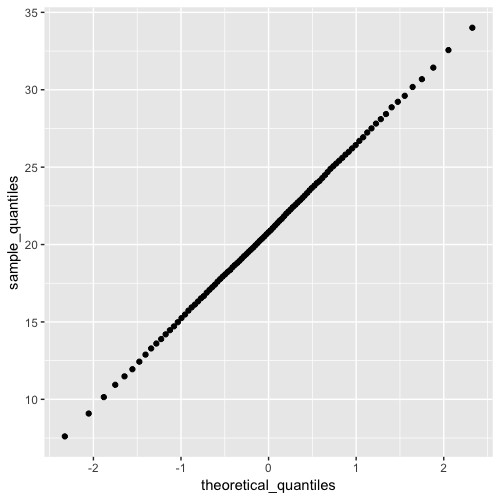
SaveSave Your Answer Show Answer

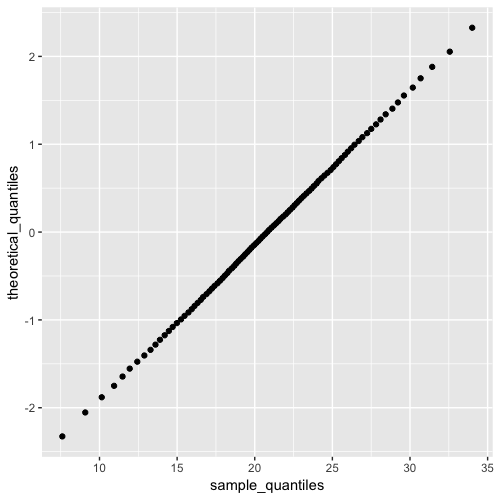
### **Question 4d**

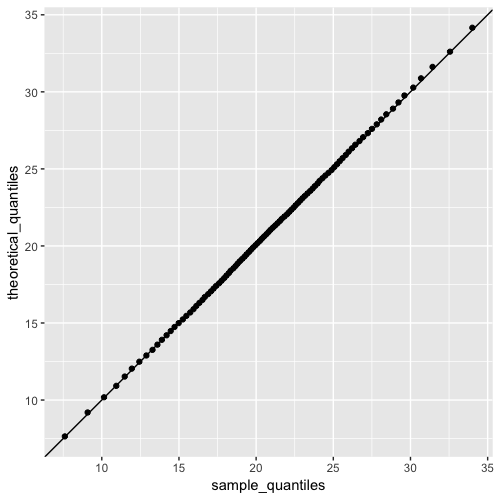
1.0/1.0 point (graded)

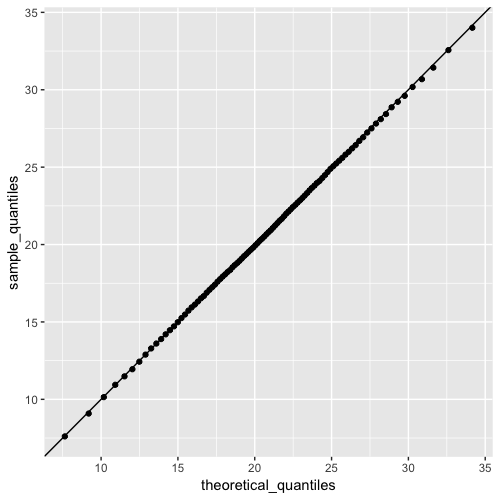
Make a corresponding set of theoretical quantiles using qnorm() over the interval p <- seq(0.01, 0.99, 0.01) with mean 20.9 and standard deviation 5.7. Save these as theoretical\_quantiles. Make a QQ-plot graphing sample\_quantiles on the y-axis versus theoretical\_quantiles on the x-axis.

Which of the following graphs is correct?









correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Section 3: Random Variables, Sampling Models and Central Limit theorem

### [3.1 Random Variables and Sampling Models](https://courses.edx.org/courses/course-v1:HarvardX+PH125.3x+1T2020/course/#block-v1:HarvardX+PH125.3x+1T2020+type@sequential+block@3901333598734983b099cb1cd0075a5e)

## Random Variables

### Key points

* Random variables are numeric outcomes resulting from random processes.
* Statistical inference offers a framework for quantifying uncertainty due to randomness.

### Code: Modeling a random variable

# define random variable x to be 1 if blue, 0 otherwise

beads <- rep(c("red", "blue"), times = c(2, 3))  
x <- ifelse(sample(beads, 1) == "blue", 1, 0)

# demonstrate that the random variable is different every time

ifelse(sample(beads, 1) == "blue", 1, 0)

ifelse(sample(beads, 1) == "blue", 1, 0)

ifelse(sample(beads, 1) == "blue", 1, 0)

## Sampling Models

### Key points

* A sampling model models the random behavior of a process as the sampling of draws from an urn.
* The **probability distribution of a random variable** is the probability of the observed value falling in any given interval.
* We can define a CDF F(a)=Pr(S≤a) to answer questions related to the probability of S being in any interval.
* The average of many draws of a random variable is called its **expected value**.
* The standard deviation of many draws of a random variable is called its **standard error**.

### Monte Carlo simulation: Chance of casino losing money on roulette

We build a sampling model for the random variable S that represents the casino's total winnings.

# sampling model 1: define urn, then sample

color <- rep(c("Black", "Red", "Green"), c(18, 18, 2)) # define the urn for the sampling model

n <- 1000

X <- sample(ifelse(color == "Red", -1, 1), n, replace = TRUE)

X[1:10]

# sampling model 2: define urn inside sample function by noting probabilities

x <- sample(c(-1, 1), n, replace = TRUE, prob = c(9/19, 10/19)) # 1000 independent draws

S <- sum(x) # total winnings = sum of draws

S

We use the sampling model to run a Monte Carlo simulation and use the results to estimate the probability of the casino losing money.

n <- 1000 # number of roulette players

B <- 10000 # number of Monte Carlo experiments

S <- replicate(B, {

X <- sample(c(-1,1), n, replace = TRUE, prob = c(9/19, 10/19)) # simulate 1000 spins

sum(X) # determine total profit

})

mean(S < 0) # probability of the casino losing money

We can plot a histogram of the observed values of S as well as the normal density curve based on the mean and standard deviation of S.

library(tidyverse)

s <- seq(min(S), max(S), length = 100) # sequence of 100 values across range of S

normal\_density <- data.frame(s = s, f = dnorm(s, mean(S), sd(S))) # generate normal density for S

data.frame (S = S) %>% # make data frame of S for histogram

ggplot(aes(S, ..density..)) +

geom\_histogram(color = "black", binwidth = 10) +

ylab("Probability") +

geom\_line(data = normal\_density, mapping = aes(s, f), color = "blue")

## Distributions versus Probability Distributions

### Key points

* A random variable X has a probability distribution function F(a) that defines Pr(X≤a) over all values of a.
* Any list of numbers has a distribution. The probability distribution function of a random variable is defined mathematically and does not depend on a list of numbers.
* The results of a Monte Carlo simulation with a large enough number of observations will approximate the probability distribution of X.
* If a random variable is defined as draws from an urn:
  + - The probability distribution function of the random variable is defined as the distribution of the list of values in the urn.
    - The expected value of the random variable is the average of values in the urn.
    - The standard error of one draw of the random variable is the standard deviation of the values of the urn.

## Notation for Random Variables

### Key points

* Capital letters denote random variables (X) and lowercase letters denote observed values (x).
* In the notation Pr(X=x), we are asking how frequently the random variable X is equal to the value x. For example, if x=6, this statement becomes Pr(X=6).

## Central Limit Theorem

### Key points

* The Central Limit Theorem (CLT) says that the distribution of the sum of a random variable is approximated by a normal distribution.
* The expected value of a random variable, E[X]=μ, is the average of the values in the urn. This represents the expectation of one draw.
* The standard error of one draw of a random variable is the standard deviation of the values in the urn.
* The expected value of the sum of draws is the number of draws times the expected value of the random variable.
* The standard error of the sum of independent draws of a random variable is the square root of the number of draws times the standard deviation of the urn.

### Equations

These equations apply to the case where there are only two outcomes, a and b with proportions p and 1−p respectively. The general principles above also apply to random variables with more than two outcomes.

Expected value of a random variable:

ap+b(1−p)

Expected value of the sum of n draws of a random variable:

N\*(ap+b(1−p))

Standard deviation of an urn with two values:

∣b–a∣\*sqrt(p(1−p))

Standard error of the sum of n draws of a random variable:

Sqrt(n)\*∣b–a∣\*sqrt(p(1−p))

[3.2 The Central Limit Theorem Continued](https://courses.edx.org/courses/course-v1:HarvardX+PH125.3x+1T2020/course/#block-v1:HarvardX+PH125.3x+1T2020+type@sequential+block@2e738fee07c94db2912c38e66353352c)

## Averages and Proportions

### Key points

* Random variable times a constant

The expected value of a random variable multiplied by a constant is that constant times its original expected value:

E[aX]=aμ

The standard error of a random variable multiplied by a constant is that constant times its original standard error:

SE[aX]=aσ

* Average of multiple draws of a random variable

The expected value of the average of multiple draws from an urn is the expected value of the urn (μ).

The standard deviation of the average of multiple draws from an urn is the standard deviation of the urn divided by the square root of the number of draws (σ/n−−√).

* The sum of multiple draws of a random variable

The expected value of the sum of *n* draws of a random variable is n times its original expected value:

E[nX]=nμ

The standard error of the sum of *n* draws of random variable is √n times its original standard error:

SE[nX]=sqrt(n)σ

* The sum of multiple different random variables

The expected value of the sum of different random variables is the sum of the individual expected values for each random variable:

E[X1+X2+⋯+Xn]=μ1+μ2+⋯+μn

The standard error of the sum of different random variables is the square root of the sum of squares of the individual standard errors:

SE[X1+X2+⋯+Xn]=sqrt(σ21+σ22+⋯+σ2n)

* Transformation of random variables

If X is a normally distributed random variable and a and b are non-random constants, then aX+b is also a normally distributed random variable.

## Law of Large Numbers

### Key points

* The law of large numbers states that as n increases, the standard error of the average of a random variable decreases. In other words, when n is large, the average of the draws converges to the average of the urn.
* The law of large numbers is also known as the law of averages.
* The law of averages only applies when n is very large and events are independent. It is often misused to make predictions about an event being "due" because it has happened less frequently than expected in a small sample size.

## How Large is Large in CLT?

### Key points

* The sample size required for the Central Limit Theorem and Law of Large Numbers to apply differs based on the probability of success.
  + - If the probability of success is high, then relatively few observations are needed.
    - As the probability of success decreases, more observations are needed.
* If the probability of success is extremely low, such as winning a lottery, then the Central Limit Theorem may not apply even with extremely large sample sizes. The normal distribution is not a good approximation in these cases, and other distributions such as the Poisson distribution (not discussed in these courses) may be more appropriate.

[3.3 Assessment: Random Variables, Sampling Models, and the Central Limit Theorem](https://courses.edx.org/courses/course-v1:HarvardX+PH125.3x+1T2020/course/#block-v1:HarvardX+PH125.3x+1T2020+type@sequential+block@6fdf5ba84cc74f039a5a9723d475af59)

## Questions 1 and 2: SAT testing

The SAT is a standardized college admissions test used in the United States. The following two multi-part questions will ask you some questions about SAT testing.

This is a 6-part question asking you to determine some probabilities of what happens when a student guessed for all of their answers on the SAT. Use the information below to inform your answers for the following questions.

An old version of the SAT college entrance exam had a -0.25 point penalty for every incorrect answer and awarded 1 point for a correct answer. The quantitative test consisted of 44 multiple-choice questions each with 5 answer choices. Suppose a student chooses answers by guessing for all questions on the test.

### **Question 1a**

1.0/1.0 point (graded)

What is the probability of guessing correctly for one question?  correct

0.2 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1b**

1.0/1.0 point (graded)

What is the expected value of points for guessing on one question?  correct

0 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1c**

1.0/1.0 point (graded)

What is the expected score of guessing on all 44 questions?  correct

0 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1d**

1.0/1.0 point (graded)

What is the standard error of guessing on all 44 questions?  correct

3.316 Loading

Submit

You have used 4 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1e**

1.0/1.0 point (graded)

Use the Central Limit Theorem to determine the probability that a guessing student scores 8 points or higher on the test.  correct

0.0079 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 1f**

1.0/1.0 point (graded)

Set the seed to 21, then run a Monte Carlo simulation of 10,000 students guessing on the test.

(IMPORTANT! If you use R 3.6 or later, you will need to use the command set.seed(x, sample.kind = "Rounding") instead of set.seed(x). Your R version will be printed at the top of the Console window when you start RStudio.)

What is the probability that a guessing student scores 8 points or higher?  correct

0.008 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

The SAT was recently changed to reduce the number of multiple choice options from 5 to 4 and also to eliminate the penalty for guessing.

In this two-part question, you'll explore how that affected the expected values for the test.

### **Question 2a**

1.0/1.0 point (graded)

Suppose that the number of multiple choice options is 4 and that there is no penalty for guessing - that is, an incorrect question gives a score of 0.

What is the expected value of the score when guessing on this new test?  correct

11 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 2b**

1.0/1.0 point (graded)

Consider a range of correct answer probabilities p <- seq(0.25, 0.95, 0.05) representing a range of student skills.

What is the lowest p such that the probability of scoring over 35 exceeds 80%?  correct

0.85 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Question 3: Betting on Roulette

A casino offers a House Special bet on roulette, which is a bet on five pockets (00, 0, 1, 2, 3) out of 38 total pockets. The bet pays out 6 to 1. In other words, a losing bet yields -$1 and a successful bet yields $6. A gambler wants to know the chance of losing money if he places 500 bets on the roulette House Special.

The following 7-part question asks you to do some calculations related to this scenario.

### **Question 3a**

1.0/1.0 point (graded)

What is the expected value of the payout for one bet?  correct

−0.07894737 Loading

Submit

You have used 3 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3b**

1.0/1.0 point (graded)

What is the standard error of the payout for one bet?  correct

2.366 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3c**

1.0/1.0 point (graded)

What is the expected value of the average payout over 500 bets?

Remember there is a difference between expected value of the average and expected value of the sum.

  correct

−0.07894737 Loading

Submit

You have used 4 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3d**

1.0/1.0 point (graded)

What is the standard error of the average payout over 500 bets?

Remember there is a difference between the standard error of the average and standard error of the sum.

  correct

0.1058209 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3e**

1.0/1.0 point (graded)

What is the expected value of the sum of 500 bets?  correct

−39.47368 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3f**

1.0/1.0 point (graded)

What is the standard error of the sum of 500 bets?  correct

52.91045 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3g**

1.0/1.0 point (graded)

Use pnorm() with the expected value of the sum and standard error of the sum to calculate the probability of losing money over 500 bets, Pr(X≤0).  correct

0.7721805 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Section 4: The Big Short

## The Big Short: Interest Rates Explained

### Correction

At 2:35, the displayed results of the code are incorrect. Here are the correct values:

n\*(p\*loss\_per\_foreclosure + (1-p)\*0)

[1] -4e+06

sqrt(n)\*abs(loss\_per\_foreclosure)\*sqrt(p\*(1-p))

[1] 885438

### Key points

* Interest rates for loans are set using the probability of loan defaults to calculate a rate that minimizes the probability of losing money.
* We can define the outcome of loans as a random variable. We can also define the sum of outcomes of many loans as a random variable.
* The Central Limit Theorem can be applied to fit a normal distribution to the sum of profits over many loans. We can use properties of the normal distribution to calculate the interest rate needed to ensure a certain probability of losing money for a given probability of default.

### Code: Interest rate sampling model

n <- 1000

loss\_per\_foreclosure <- -200000

p <- 0.02

defaults <- sample( c(0,1), n, prob=c(1-p, p), replace = TRUE)

sum(defaults \* loss\_per\_foreclosure)

### Code: Interest rate Monte Carlo simulation

B <- 10000

losses <- replicate(B, {

defaults <- sample( c(0,1), n, prob=c(1-p, p), replace = TRUE)

sum(defaults \* loss\_per\_foreclosure)

})

### Code: Plotting expected losses

library(tidyverse)

data.frame(losses\_in\_millions = losses/10^6) %>%

ggplot(aes(losses\_in\_millions)) +

geom\_histogram(binwidth = 0.6, col = "black")

### Code: Expected value and standard error of the sum of 1,000 loans

n\*(p\*loss\_per\_foreclosure + (1-p)\*0) # expected value

sqrt(n)\*abs(loss\_per\_foreclosure)\*sqrt(p\*(1-p)) # standard error

### Code: Calculating interest rates for expected value of 0

We can calculate the amount x to add to each loan so that the expected value is 0 using the equation lp+x(1−p)=0. Note that this equation is the definition of expected value given a loss per foreclosure l with foreclosure probability p and profit x if there is no foreclosure (probability 1−p).

We solve for x=−lp1−p and calculate x:

x = - loss\_per\_foreclosure\*p/(1-p)

x

On a $180,000 loan, this equals an interest rate of:

x/180000

### Equations: Calculating interest rate for 1% probability of losing money

We want to calculate the value of x for which Pr(S<0)=0.01. The expected value E[S] of the sum of n=1000 loans given our definitions of x, l and p is:

μS=(lp+x(1−p))∗n

And the standard error of the sum of n loans, SE[S], is:

σS=∣x−l∣\*sqrt(np(1−p))

Because we know the definition of a Z-score is Z=x−μσ, we know that Pr(S<0)=Pr(Z<−μσ). Thus, Pr(S<0)=0.01 equals:

Pr(Z<(−{lp+x(1−p)}n)/((x−l)\*sqrt(np(1−p))))=0.01

z<-qnorm(0.01) gives us the value of z for which Pr(Z≤z)=0.01, meaning:

z=(−{lp+x(1−p)}n)/((x−l)\*sqrt(np(1−p)))

Solving for x gives:

x=−l\*(np−z\*sqrt(np(1−p)))/(n(1−p)+z\*sqrt(np(1−p)))

### Code: Calculating interest rate for 1% probability of losing money

l <- loss\_per\_foreclosure

z <- qnorm(0.01)

x <- -l\*( n\*p - z\*sqrt(n\*p\*(1-p)))/ ( n\*(1-p) + z\*sqrt(n\*p\*(1-p)))\x

x/180000 # interest rate

loss\_per\_foreclosure\*p + x\*(1-p) # expected value of the profit per loan

n\*(loss\_per\_foreclosure\*p + x\*(1-p)) # expected value of the profit over n loans

### Code: Monte Carlo simulation for 1% probability of losing money

Note that your results will vary from the video because the seed is not set.

B <- 100000

profit <- replicate(B, {

draws <- sample( c(x, loss\_per\_foreclosure), n,

prob=c(1-p, p), replace = TRUE)

sum(draws)

})

mean(profit) # expected value of the profit over n loans

mean(profit<0) # probability of losing money

## The Big Short

### Key points

* The Central Limit Theorem states that the sum of independent draws of a random variable follows a normal distribution. However, when the draws are not independent, this assumption does not hold.
* If an event changes the probability of default for all borrowers, then the probability of the bank losing money changes.
* Monte Carlo simulations can be used to model the effects of unknown changes in the probability of default.

### Code: Expected value with higher default rate and interest rate

p <- .04

loss\_per\_foreclosure <- -200000

r <- 0.05

x <- r\*180000

loss\_per\_foreclosure\*p + x\*(1-p)

### Equations: Probability of losing money

We can define our desired probability of losing money, z, as:

Pr(S<0)=Pr(Z<−E[S]/SE[S])=Pr(Z<z)

If μ is the expected value of the urn (one loan) and σ is the standard deviation of the urn (one loan), then E[S]=nμ and SE[S]=sqrt(n)\*σ.

As in the previous video, we define the probability of losing money z=0.01. In the first equation, we can see that:

z=−E[S]/SE[S]

It follows that:

z=−nμ/sqrt(n)σ=−√nμ/σ

To find the value of n for which z is less than or equal to our desired value, we take z≤−sqrt(n)μ/σ and solve for n:

N ≥ z\*z\*σ\* σ /μ\* μ

### Code: Calculating number of loans for desired probability of losing money

The number of loans required is:

z <- qnorm(0.01)

l <- loss\_per\_foreclosure

n <- ceiling((z^2\*(x-l)^2\*p\*(1-p))/(l\*p + x\*(1-p))^2)

n # number of loans required

n\*(loss\_per\_foreclosure\*p + x \* (1-p)) # expected profit over n loans

### Code: Monte Carlo simulation with known default probability

This Monte Carlo simulation estimates the expected profit given a known probability of default p=0.04. Note that your results will differ from the video because the seed is not set.

B <- 10000

p <- 0.04

x <- 0.05 \* 180000

profit <- replicate(B, {

draws <- sample( c(x, loss\_per\_foreclosure), n,

prob=c(1-p, p), replace = TRUE)

sum(draws)

})

mean(profit)

### Code: Monte Carlo simulation with unknown default probability

This Monte Carlo simulation estimates the expected profit given an unknown probability of default 0.03≤p≤0.05, modeling the situation where an event changes the probability of default for all borrowers simultaneously. Note that your results will differ from the video because the seed is not set.

p <- 0.04

x <- 0.05\*180000

profit <- replicate(B, {

new\_p <- 0.04 + sample(seq(-0.01, 0.01, length = 100), 1)

draws <- sample( c(x, loss\_per\_foreclosure), n,

prob=c(1-new\_p, new\_p), replace = TRUE)

sum(draws)

})

mean(profit) # expected profit

mean(profit < 0) # probability of losing money

mean(profit < -10000000) # probability of losing over $10 million