1 Base R

1. The `starwars` dataset contains the characteristics of 87 different Star Wars characters.

```
glimpse(starwars)

Rows: 87
Columns: 14
$ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "Leia...
$ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, 180...
$ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.0, ...
$ hair_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "brown"...
$ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "light"...
$ eye_color  <chr> "blue", "yellow", "red", "yellow", "brown", "blue", "blu...
(plus 8 more variables)
```

What will output from the lines of code given in parts (a)-(d)? **Select ONE each**.

(a) (3 points) `starwars[1,]`

- ◯ A vector of length 14
- ◯ A vector of length 87
- ◯ An error
- ● A dataframe with 1 row and 14 columns
- ◯ A dataframe with 87 rows and 1 column

(b) (3 points) `starwars[,4]`

- ◯ A vector of length 14
- ◯ A vector of length 87
- ◯ An error
- ◯ A dataframe with 1 row and 14 columns
- ● A dataframe with 87 rows and 1 column

(c) (3 points) `starwars$hair_color`

- ◯ A vector of length 14
- ● A vector of length 87
- ◯ An error
- ◯ A dataframe with 1 row and 14 columns
- ◯ A dataframe with 87 rows and 1 column

(d) (3 points) `starwars$4`

- ◯ A vector of length 14
- ◯ A vector of length 87
- ● An error
- ◯ A dataframe with 1 row and 14 columns
- ◯ A dataframe with 87 rows and 1 column

(e) (4 points) Some of the values of the `mass` column are `NA` Two of the lines of code below returns the average mass (of characters who actually have a mass) and the other two return NA. Briefly explain which are which.

- (A): `mean(starwars$mass != NA)`
- (B): `mean(!is.na(starwars$mass))`
- (C): `mean(starwars$mass, na.rm = T)`
- (D): `mean(starwars$mass > 0)`

Solution: Code (A) and (D) will return `NA` because the != and < operators do not work when one or more side is `NA`. The correct way to check whether an item is `NA` is with `!is.na()` like in code (B). Or, we can explicitly ignore `NA` values within the `mean()` function like in (C).

2. Define a character vector `colors` of length 5 and a character vector `names` of length 4.

```
colors <- c("red", "blue", "yellow", "green", "purple")
names <- c("Inky", "Blinky", "Pinky", "Clyde")
```

(a) (4 points) Identify the length of the vector that will output from the following lines of code, and briefly explain your answer. If the code will instead return an error, explain why.

i. `c(colors, names)`

Solution: The output will be a vector of length 9, since the `c()` command will combine the two vectors into a single one.

ii. `colors + names`

Solution: The output will be an error, because the + operator works for numeric objects, and the contents of our vectors are character, not numeric.

iii. `paste(colors, names)`

Solution: The output will have length 5. The `paste` command will connect the character strings together in an element-wise way. So the first elements of each will be pasted together, the second elements of each will be pasted together, etc. The output will be length 5 because that is the longer of the two input vectors.

iv. `paste(colors, "")`

Solution: The output will be a character vector of length 5 and will be exactly the same as the original `colors` vector. The empty string is appended to each individual entry of `colors`.

(b) (3 points) We would like to identify all of the colors that either start with the letter "g" or are at least five letters long. Which line of R code can be used to do this? **Select ONE**.

- ◯ `(nchar(colors) >= 5) & (substr(colors, 1, 1) == "g")`
- ◯ `(length(colors) >= 5) & (substr(colors, 1, 1) == "g")`
- ● `(nchar(colors) >= 5) | (substr(colors, 1, 1) == "g")`
- ◯ `(length(colors) >= 5) | (substr(colors, 1, 1) == "g")`

3. Use the following lines of R code to answer the questions below. Assume that at the beginning, the R environment is empty. Each line will attempt to run once, in order, even if a previous line returned an error.

```
1 rent <- 1200
2 utilities <- 50
3 utilities_adj <- utilities + "10"
4
5 total <- rent + utilities
6 total_adj <- rent + utilities_adj
```

(a) (3 points) What is the value of the variable `total`? **Select ONE**.

- ◯ 1200
- ● 1250
- ◯ 1260
- ◯ undefined

(b) (3 points) What is the value of the variable `total_adj`? **Select ONE**.

- ◯ 1200
- ◯ 1250
- ◯ 1260
- ● undefined

(c) (5 points) Which line(s) of code will return an error? **Select ALL that apply**.

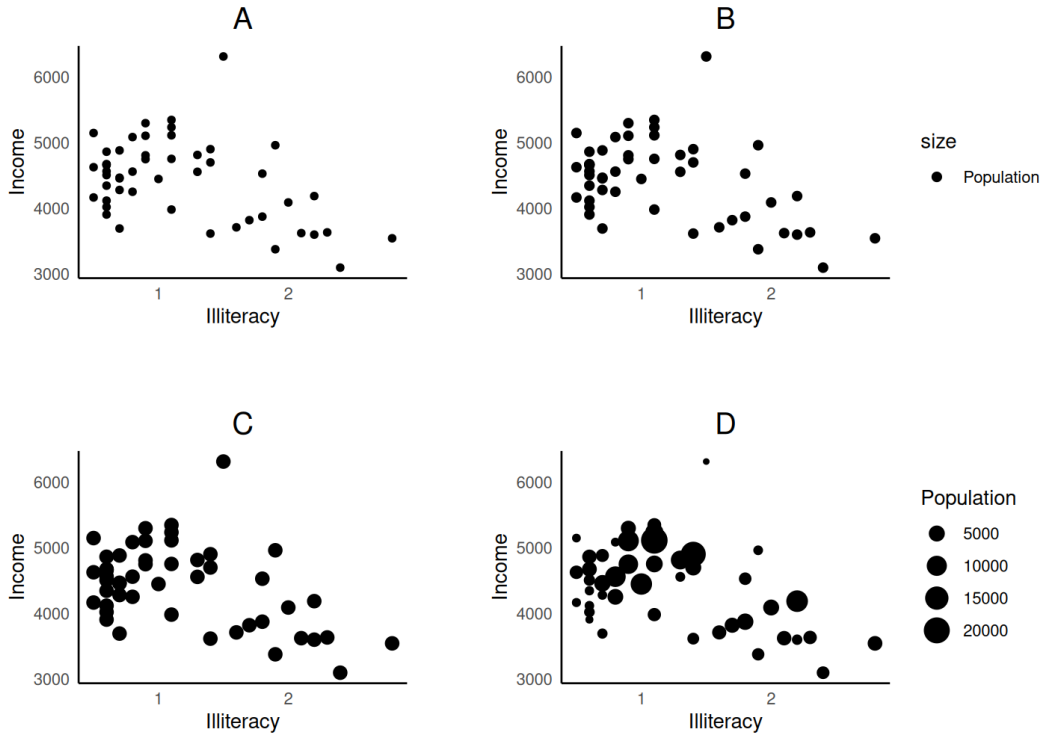
- ◯ 1
- ◯ 2
- ● 3
- ◯ 5
- ● 6

(d) (4 points) Explain why no numbers are printed in the console from running this code.

Solution: We don't expect to see any value print while running this code, because we are only assigning variables with '<-' and never actually viewing or printing the value of any of these variables. The value of the variable in the environment will change, but it will not print.

2 ggplot

1. Below are four different `geom_point()` plots of illiteracy rate versus income for US states in the `states` dataframe. Some plots also show the states' population.



(a) (4 points) Match the four plots (A, B, C, D) to the four different `geom_point()` calls below.

A `geom_point(aes(x = Illiteracy, y = Income))`

D `geom_point(aes(x = Illiteracy, y = Income, size = Population))`

B `geom_point(aes(x = Illiteracy, y = Income, size = "Population"))`

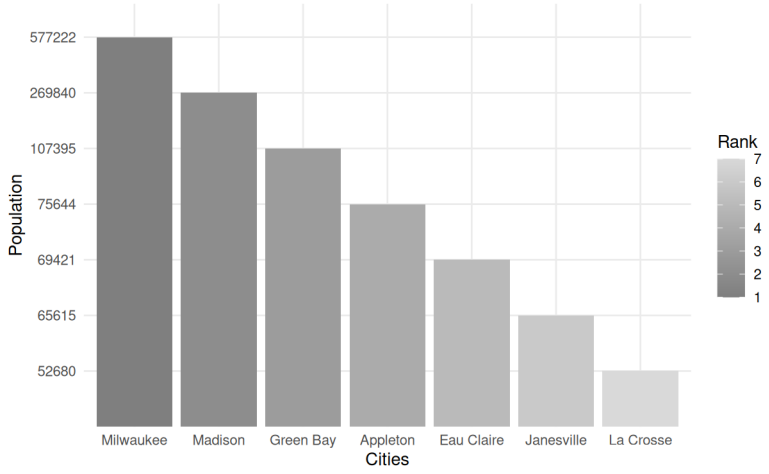
C `geom_point(aes(x = Illiteracy, y = Income), size = 3)`

(b) (3 points) The following code returns an error when trying to run `geom_smooth()`. How should the error be fixed? **Select ONE**.

```
states %>%
  ggplot() +
  geom_point(aes(x = Income, y = Illiteracy)) +
  geom_smooth()
```

- ◯ `geom_smooth()` should be added before `geom_point()`.
- ◯ `x = Income` and `y = Illiteracy` should be moved outside of `aes()`.
- ● The `aes()` mapping should be moved inside of `ggplot()`.
- ◯ None of the above; the code should not return any error.

2. The `cities_WI` dataset has 7 rows and 3 columns, containing the name, population, and population rank 1-7 of Wisconsin's seven largest cities. We want to re-create the following bar plot:



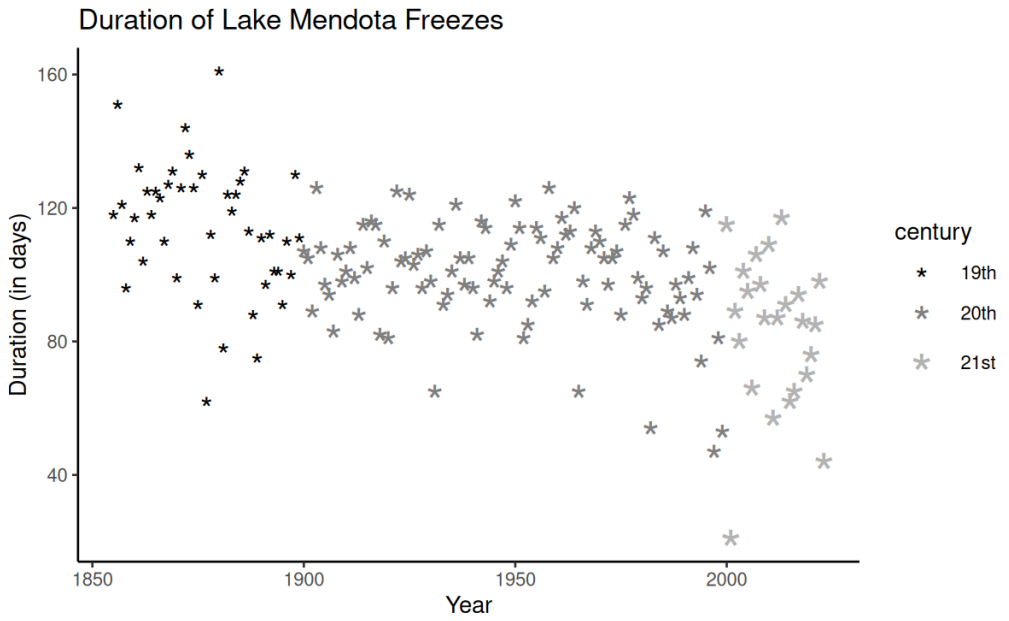
(a) (3 points) Given that we know the population of the 7 cities, should the plot be created with `geom_bar()` or `geom_col()`? Briefly explain your answer.

Solution: The plot must have been made with `geom_col()`. We use `geom_col()` when we have both an x and y column and can therefore set the height of the bars (population in this case) manually. With `geom_bar()` we only provide one variable and the heights of the bars are calculated automatically based on counts of observations.

(b) (3 points) How should we write the call to `ggplot()` to make it so the bar interiors are colored based on the cities' `Rank`? The other plot aesthetics are intentionally left out. **Select ONE**.

- ● `ggplot(aes(..., fill = Rank))`
- ○ `ggplot(aes(..., col = Rank))`
- ○ `ggplot(aes(...), fill = Rank)`
- ○ `ggplot(aes(...), col = Rank)`

3. The scatter plot below was created with the `mendota` data from lecture. It shows the duration of Lake Mendota freezes over time, and also indicates the century each point belongs to.



(a) (5 points) For each aesthetic used in the plot, indicate whether it is a variable or constant aesthetic, or if it is not used in the plot. **Select ONE each**.

`x` :	● Variable	○ Constant
`y` :	● Variable	○ Constant
`color` :	● Variable	○ Constant
`shape` :	○ Variable	● Constant
`size` :	● Variable	○ Constant

(b) (3 points) Which two aesthetics correspond to the same variable in the data? What information is being conveyed by those two aesthetics?

Solution: `color` and `size` are both being used to refer to the century each data point belongs to.

(c) (3 points) Which line of code below create a red line parallel to the x-axis representing the mean duration? **Select ONE**.

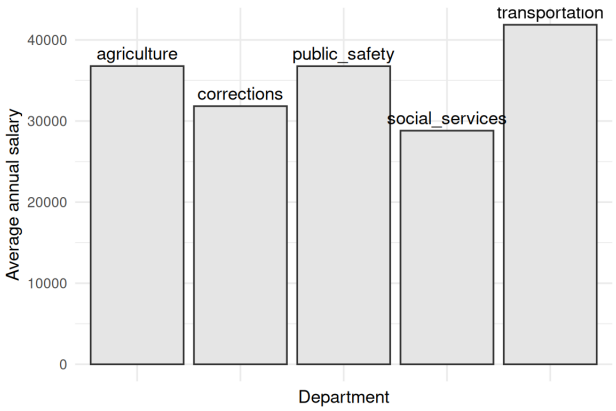
● `geom_hline(yintercept = mean(mendota$duration), color = "red")`
○ `geom_vline(yintercept = mean(duration), color = "red")`
○ `geom_line(yintercept = mean(mendota$duration), color = "red")`
○ `geom_hline(yintercept = mean(duration), color = "red")`
○ `geom_vline(xintercept = mean(mendota$duration), color = red)`
○ `geom_line(xintercept = mean(duration), color = red)`

4. The fictitious dataset `state_employees` contains columns for the ID, department, number of years of service, and annual salary for 100000 state employees.

```
state_employees %>%
  sample_n(5)
```

	employee_id	department	years_service	salary
	<int>	<chr>	<dbl>	<dbl>
1	10783	agriculture	3	23660
2	69923	public_safety	7	33917
3	35260	corrections	8	28847
4	3336	agriculture	4	25772
5	45582	social_services	5	26632

The graph below shows the average annual salary for the five departments in the dataset.



(a) (6 points) In the plot above, which aesthetics are variable aesthetics? **Select ALL that apply**.

- [x] `x`
- [x] `y`
- [ ] `color`
- [ ] `fill`
- [x] `text`
- [ ] `line`

(b) (3 points) Was the plot above made using `geom_bar()` or `geom_col()`? Briefly explain how you can tell.

Solution: This plot must have been made with `geom_col()` after calculating the average annual salary since we need to provide both an x and a y to specify the height of the bars. `geom_bar()` would give a count of each department which is not what we want in this case.

5. Your zoologist friend is studying three species of zebra. She has gathered data on the physical characteristics of 160 of these zebras and asks for your help analyzing it. Below are 5 example rows from full dataset.

	length	height	sex	species
	<dbl>	<dbl>	<chr>	<chr>
1	237	142	F	mountain
2	245	145	M	plains
3	216	128	F	mountain
4	235	139	F	plains
5	282	152	M	grevy's

She sends you the following message about how she wants to graph the data: "I want to make a scatter plot of each zebra's body length and shoulder height, with a single smooth line showing the trend. I want the color of the points to be different for the different species, and I want all of the points to be slightly transparent in case they overlap".

(a) (6 points) Which `ggplot2` graph types should you add to the plot to fulfill her request? **Select ALL that apply**.

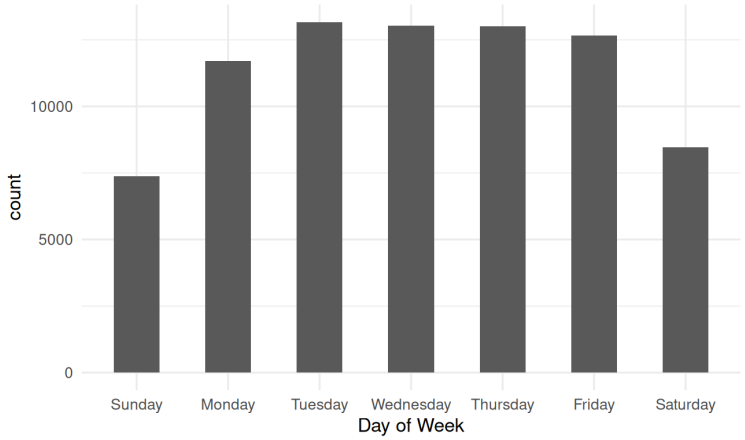
- ◯ `geom_density()`
- ● `geom_point()`
- ◯ `geom_line()`
- ● `geom_smooth()`
- ◯ `geom_histogram()`
- ◯ `geom_boxplot()`

(b) (4 points) For each of the following aesthetics in your plot, identify whether they would need to be global variable, local variable, or constant. **Select ONE each**.

`x` :	● Global	◯ Local	◯ Constant
`y` :	● Global	◯ Local	◯ Constant
`color` :	◯ Global	● Local	◯ Constant
`alpha` :	◯ Global	◯ Local	● Constant

6. The dataframe `births` (with 79,421 rows and 1 column) records the day of the week for 79,421 births. The code below (some details omitted) creates a graph to count the number of births for each day.

```
births %>%
    group_by(day_of_week) %>%
    summarize(count = n()) %>%
    ggplot(aes(x = day_of_week, y = count)) +
        geom_col()
```



(a) (3 points) Explain why we need to create a new variable called `count`. Identify a different geom that could produce the same plot without needing to tidy the data or manually create this new variable.

Solution: Since `geom_col()` requires both an x and a y variable, we need to pre-calculate the counts so that we can set the bar height manually. We could instead use `geom_bar()` which only takes a categorical x variable and automatically computes the counts.

(b) (4 points) You want to make the following changes to the plot: Add labels at the top of each bar displaying the count for that day, and add a line showing the count of the tallest bar. Which geoms do you need to add to your plot? **Select ALL that apply**.

- ◯ `geom_point()`
- ● `geom_text()`
- ● `geom_hline()`
- ◯ `geom_vline()`

(c) (3 points) You want to color the bar areas pastel blue. Which aesthetics do you need to add to your plot? **Select ONE**.

- ◯ `color` as a variable aesthetic
- ◯ `color` as a constant aeesthetic
- ◯ `fill` as a variable aesthetic
- ● `fill` as a constant aesthetic

3 dplyr

1. The `birthwt` dataset records birth weight and risk factors for 189 pregnancies. We are interested in columns `low` (0 or 1, indicating low birth weight), `smoke` (0 or 1, indicating mother's smoking status), and `bwt` (birth weight of newborn in grams).

```
> birthwt %>% select(low, smoke, bwt) %>% glimpse()
Rows: 189
Columns: 3
$ low   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ smoke <int> 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, ...
$ bwt   <int> 2523, 2551, 2557, 2594, 2600, 2622, 2637, 2637, 2663, 2665, 2722, ...
```

Use these function names as well as the column names to fill in the blanks below. You might not need to use every function/column, and you might need to use some more than once.

- `select`
- `filter`
- `group_by`
- `summarize`
- `count`
- `min`
- `slice_min`
- `median`
- `mean`

(a) (3 points) Fill in the blanks to count the number of rows for every combination of low birth weight (low vs normal) and mother's smoking status (yes vs no).

```
birthwt %>%
 group_by (low,   smoke   ) %>%
 summarize (numRows = n())
```

(b) (5 points) Fill in the blanks to find the average and median weight of the 50 smallest newborns in the data.

```
birthwt %>%
 slice_min (   bwt   , n = 50) %>%
 summarize (avgWt =   mean   (bwt), medWt =   median   (bwt))
```

2. The `cabbages` dataset gives the cultivar type, date planted, head weight (in kg) and vitamin C content of 60 cabbages.

```
> glimpse(cabbages)
Rows: 60
Columns: 4
$ Cult   <fct> c39, c39, c39, c39, c39, c39, c39, c39, c39, c39, c39, c39, c39,  ...
$ Date   <fct> d16, d16, d16, d16, d16, d16, d16, d16, d16, d16, d20, d20, d20,  ...
$ HeadWt <dbl> 2.5, 2.2, 3.1, 4.3, 2.5, 4.3, 3.8, 4.3, 1.7, 3.1, 3.0, 2.8, 2.8,  ...
$ VitC   <int> 51, 55, 45, 42, 53, 50, 50, 52, 56, 49, 65, 52, 41, 51, 41, 45,  ...
```

Use these function names to fill in the blanks below. You might not need to use every function, and you might need to use some more than once.

- `relocate`
- `select`
- `group_by`
- `summarize`
- `max`
- `slice_max`
- `n`
- `count`
- `mean`

(a) (3 points) Fill in the blanks to remove the `VitC` column, then print the largest cabbage for each of the cultivar types.

```
cabbages %>%
  select(-VitC) %>%
 group_by(Cult) %>%
 slice_max(HeadWt)
```

(b) (4 points) Fill in the blanks to count the number of cabbages for each cultivar-date combination, and calculate the average vitamin C content for each cultivar-date combination.

```
cabbages %>%
 group_by(Cult, Date) %>%
 summarize(counts = n(),
  avgVitC = mean(VitC))
```

3. Your zoologist friend is studying three species of zebra. She has gathered data on the physical characteristics of 160 of these zebras and asks for your help analyzing it. Below are 5 example rows from full dataset.

	length <dbl>	height <dbl>	sex <chr>	species <chr>
1	237	142	F	mountain
2	245	145	M	plains
3	216	128	F	mountain
4	235	139	F	plains
5	282	152	M	grevy's

(a) (3 points) You want to re-structure the data so that all of the measurements are in a single column called *measurement*, with a column called `type` labeling what is being measured.

	sex <chr>	species <chr>	type <chr>	measurement <dbl>
1	M	grevy's	length	273
2	M	grevy's	height	146
3	M	grevy's	length	294
4	M	grevy's	height	157
5	M	grevy's	length	279
6	M	grevy's	height	157

Which line of R code should be used? **Select ONE**.

- ● `pivot_longer(length:height, names_to = "type", values_to = "measurement")`
- ○ `pivot_wider(length:height, names_to = "type", values_to = "measurement")`
- ○ `pivot_longer(names_from = "type", values_from = "measurement")`
- ○ `pivot_wider(names_from = "type", values_from = "measurement")`

(b) (4 points) Consider the following code which creates a new dataframe, `zebra_summary`.

```
zebra_summary <- zebras %>%
  group_by(species) %>%
  summarize(avgLength = mean(length),
            avgHeight = mean(height))
```

Which of the following statments are true about `zebra_summary`? **Select ALL that apply**.

- ● `zebra_summary` has three rows.
- ○ `zebra_summary` has two columns.
- ● `zebra_summary` has a column called “species".
- ○ `zebra_summary` has a column called “sex".

4. A dataframe `lions` has 32 rows, each representing a unique lion.

   It has two numeric columns: `age`, each lion's age in decimal years, and `proportion.black`, the percentage of that lion's nose which is black. It has a logical column `adult`, which is `TRUE` if `age` is greater than 3 and `FALSE` otherwise.

   (a) (4 points) Consider the following code which creates a new dataframe, `lions_summary`.

```
lions_summary <- lions %>%
 group_by(adult) %>%
 summarize(averageAge = mean(age),
 averagePropBlack = mean(proportion.black))
```

   Which of the following statements are true about `lions_summary`? **Select ALL that apply**.

   - [x] `lions_summary` has 2 rows.
   - [ ] `lions_summary` has 2 columns.
   - [x] `lions_summary` has a column called "`adult`".
   - [ ] If there were any `NA` values in the `age` column of `lions`, all the values of `averageAge` in `lions_summary` will be `NA`.

   (b) (6 points) Which lines of code can be used to keep only the rows of `lions` whose age is greater than 3? **Select ALL that apply**.

   - [ ] `lion %>% select(adult)`
   - [ ] `lion %>% select(age)`
   - [ ] `lion %>% select(age > 3)`
   - [x] `lion %>% filter(adult)`
   - [ ] `lion %>% filter(age)`
   - [x] `lion %>% filter(age > 3)`

   (c) (3 points) The dataframe `lion_traits` has 25 rows, each representing a unique lion. It has three columns: `id`, a number that uniquely identifies each lion, `sex`, a categorical variable for whether the lion is male or female, and `length`, the length of the lion's body in meters.

   How many rows and columns will result from the following join operation? Briefly explain your answers. If the code returns an error, explain why.

```
left_join(lions, lion_traits)
```

   Solution: The code will return an error. There are no columns in common between the two datasets, so there is nothing to join by.

5. The fictitious dataset `state_employees` contains columns for the ID, department, number of years of service, and annual salary for 100000 state employees.

```
state_employees %>%
  sample_n(5)
```

	employee_id	department	years_service	salary
	<int>	<chr>	<dbl>	<dbl>
1	10783	agriculture	3	23660
2	69923	public_safety	7	33917
3	35260	corrections	8	28847
4	3336	agriculture	4	25772
5	45582	social_services	5	26632

(a) (4 points) Consider the following code which creates a new dataframe, `employee_summary`.

```
employee_summary <- state_employees %>%
 filter(department != "agriculture") %>%
 group_by(department) %>%
 summarize(avg_salary = mean(salary), employees = n())
```

Which of the following statements are TRUE about `employee_summary`? **Select ALL that apply**.

- [x] `employee_summary` has 4 rows.
- [ ] `employee_summary` has 5 rows.
- [x] `employee_summary` has a column called "`department`".
- [ ] If were any `NA` values in the `salary` column of the original data, all the values of `avg_salary` in `employee_summary` will be `NA`.

(b) (4 points) Continue working with the 100000 rows of employee salary data. Another dataset called `employment_type` has 74033 rows and two columns: numeric `employee_id` and categorical `type` which labels each employee as either full-time or part-time. The 74033 employee IDs are all also part of the larger `state_employees` dataset.

How many rows and columns result from following join? Which columns will contain `NA` values?

```
left_join(state_employees, employment_type)
```

Solution: The joined data will have 100000 rows, just like the original `state_employees` data. There will be 5 columns: the four original columns from `state_employees` and the column for employment type from the `employment_type` dataframe.

This `type` column will have `NA` values for the rows that appear in `state_employees` but not in `employment_type`.

6. (5 points) Researchers perform two experiments. In the first experiment, researchers record the partcipants' names and 8 other variables, which are stored in `df1`. In the second experiment, researchers record participants' names and 6 other variables, which are stored in `df2`. The name column is the only variable that appears in both dataframes.

   `df1` has 15 rows and `df2` has 25 rows (one row per participant). 5 people participated in both experiments, so they appear in both dataframes.

   Give the number of rows and the number of columns that result from the following join operations.

Join	Number of rows	Number of columns
`inner_join(df1, df2)`	5	15
`left_join(df1, df2)`	15	15
`right_join(df1, df2)`	25	15
`full_join(df1, df2)`	35	15
`anti_join(df1, df2)`	10	9

7. (5 points) A company is creating a video game where the player fights increasingly difficult monsters through different layers of a dungeon. One team creates a dataset `monster_stats` with the combat statistics of different monsters. The column `name` gives the monster's name, `level` gives the difficulty level of the monster, and `atk` gives the monster's attack.

	name	level	atk
	<chr>	<dbl>	<dbl>
1	skeleton	1	30
2	ooze	2	45
3	haunt	4	75

A different team creates a dataset `monster_locations` that describes where different monsters appear. The column `name` gives the monster's name, `level` gives the layer of the dungeon where it lives, and `enviro` gives the monster's preferred environment.

	name	level	enviro
	<chr>	<dbl>	<chr>
1	chicken	0	mild
2	spider	1	humid
3	skeleton	1	arid
4	ooze	3	humid

The following R code is used to join the two dataframes based on the `name` column.

```
full_join(monster_stats, monster_locations, join_by(name))
```

Using the grid below, re-create the dataframe that would result from the call to `full_join()`. Use the top shaded row for the column names and the un-shaded cells for the data values. Be sure to account for any missing data. The column names and the ordering of rows/columns do not need to be exact to get full credit, but the contents of the data must be accurate.

	name	level.x	level.y	atk	enviro
1	skeleton	1	1	30	arid
2	ooze	2	3	45	humid
3	haunt	4	NA	75	NA
4	chicken	NA	0	NA	mild
5	spider	NA	1	NA	humid

8. The dataset `students` gives the name, lecture section, homework, quiz, and exam grades of several students in a class. Below are 5 example rows from the full dataset.

	name	section	hw	quiz	exam
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	John	1	90	95	85
2	Arya	1	87	92	83
3	Fred	2	93	99	84
4	Riley	2	85	89	75
5	Tyler	2	67	70	61

(a) (4 points) We want to make a new column that combines `hw`, `quiz`, and `exam` into an overall final score, then converts that number into a categorical letter grade based on a scale.

For example, a student who scores between 92-100 should get an "A", a student who scores between 88-92 should get an "AB", and so on. Which `dplyr` commands should we use to accomplish this goal? **Select ALL that apply**.

- ● `mutate()`
- ○ `as.factor()`
- ○ `relocate()`
- ● `case.when()`

(b) (3 points) Starting from the original `students` data, we want to create a dataframe that has a column for "assignment type" and a column for "value" rather than three separate `hw, quiz, exam` columns. Which R code can be used to create this data?. **Select ONE**.

	name	section	assign_type	value
	<chr>	<dbl>	<chr>	<dbl>
1	John	1	hw	90
2	John	1	quiz	95
3	John	1	exam	85
4	Arya	1	hw	87

- ○ `students %>% summarize(assign_type = c(hw, quiz, exam), value = n())`
- ○ `students %>% pivot_wider(names_from = assign_type, values_from = value)`
- ● `students %>% pivot_longer(c(hw, quiz, exam), names_to = assign_type)`
- ○ `students %>% count(c(hw, quiz, exam))`

9. (5 points) In the dataframe `cakes`, a bakery records the flavor and occasion for each cake order they receive each week. For example, this week they received 4 orders for chocolate birthday cakes.

```
> cakes
# A tibble: 8 × 3
  flavor    occasion num_ordered
  <chr>     <chr>          <dbl>
1 chocolate birthday           4
2 vanilla   birthday           3
3 vanilla   wedding            1
4 redvelvet wedding            1
5 chocolate wedding            2
6 carrot    birthday           2
7 vanilla   other              4
8 redvelvet other              2
```

The bakery wants to structure the data such that there is a separate column for each occasion type, rather than a single "occasion" column.

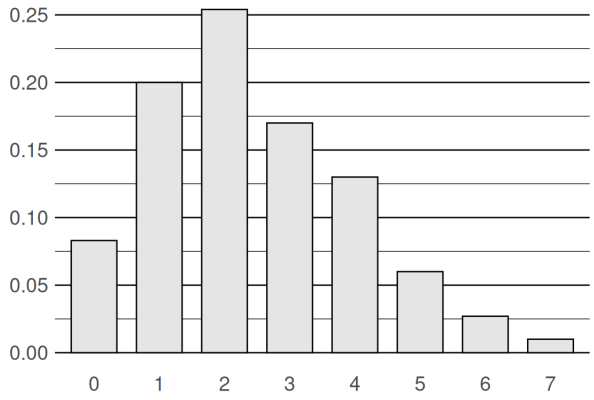
Using the grid below, re-create the dataframe that would result from the call to `pivot_wider`. Use the top shaded row for the column names and the un-shaded cells for the data values. Be sure to account for any missing data.

```
pivot_wider(cakes, names_from = "occasion", values_from = "num_ordered")
```

	flavor	birthday	wedding	other
1	chocolate	4	2	NA
2	vanilla	3	1	4
3	redvelvet	NA	1	2
4	carrot	2	NA	NA

4 Probability

1. The number of goals in a game in the FIFA World Cup is given by the random variable X. The distribution of X is given in the plot below.



(a) (3 points) Find the 25th, 50th and the 75th percentile of X.

Solution: The percentiles are based on the cumulative probabilities. $P(X \leq 0)$ is about 0.08, $P(X \leq 1)$ is about $0.08 + 0.2 = 0.28$, $P(X \leq 2)$ is about $0.28 + 0.25 = 0.53$, $P(X \leq 3)$ is about $0.53 + 0.17 = 0.7$, and $P(X \leq 4)$ is about $0.7 + 0.13 = 0.83$.

So, the 25th percentile is 1, the 50th percentile is 2, and the 75th percentile is 4. These are the values of X that capture at least this much cumulative probability.

(b) (3 points) Is $E(X)$ equal to, smaller than, or larger than the 50th percentile of X? Explain how you can tell without knowing $E(X)$.

Solution: $E(X)$ must be larger than the 50th percentile (the median) because of the skew in the distribution of X. Most of the values are on the left side of the graph, around 1 or 2, but there is a tail of bigger values going to the right. The mean will be pulled higher because of this right tail.

(c) (3 points) Which option is closest to the probability of scoring 5 or more goals in a FIFA World Cup game? **Select ONE**.

- ● 0.11
- ○ 0.02
- ○ 0.5
- ○ 0.96

2. The distribution of the number of children in a household (X) in a certain neighborhood is given below.

x	0	1	2	3
$P(X = x)$	?	0.4	0.25	0.15

It can be shown that $E(X) = 1.35$ and $SD(X) = 0.96$.

(a) (3 points) Which of the following statements is the most specific accurate description of X? **Select ONE**.

- ● X is a discrete RV.
- ○ X is a binomial RV.
- ○ X is a continuous RV.
- ○ X is a normal RV.

(b) (3 points) What is the median number of children per household? **Select ONE**.

- ○ 0
- ● 1
- ○ 2
- ○ Cannot be determined from the information given

(c) (3 points) Consider taking a random sample of four households from this neighborhood. What is the distribution of $\bar{X}_4$, the average number of children per household?

- ○ $\bar{X} \dot{\sim} N(1.35, 0.96)$
- ○ $\bar{X} \dot{\sim} N(1.35, 0.96\sqrt{4})$
- ○ $\bar{X} \dot{\sim} N(1.35, \frac{0.96}{\sqrt{4}})$
- ● None of the above

3. Researchers have planted wildflowers in several plots in an attempt to study local insect wildlife. Each day, the number of monarch butterflies in each plot is observed. Suppose the number of monarch butterflies per plot is given by random variable X with the following partial probability distribution:

x	0	1	2	3
$P(X = x)$	?	?	0.2	0.1

The average number of monarch butterflies per plot is $E(X) = 0.8$.

(a) (4 points) Find $P(X = 0)$ and $P(X = 1)$.

Solution: From the definition of expected value,

$$\begin{aligned} E(X) &= \sum x \cdot P(x) \\ 0.8 &= 0 \cdot P(0) + 1 \cdot P(1) + 2 \cdot 0.2 + 3 \cdot 0.1 \\ 0.8 &= P(1) + 0.4 + 0.3 \\ 0.1 &= P(1) \end{aligned}$$

Since the probabilities must sum to 1, we have $P(0) = 1 - (0.1 + 0.2 + 0.1) = 0.6$.

(b) (3 points) Write an expression for $\text{Var}(X)$. You do not need to simplify your answer and can leave it as an expression.

Solution:

$$\begin{aligned} \text{Var}(X) &= \sum (x - E(X))^2 \cdot P(x) \\ &= (0 - 0.8)^2 \cdot 0.6 \ + \ (1 - 0.8)^2 \cdot 0.1 \ + \ (2 - 0.8)^2 \cdot 0.2 \ + \ (3 - 0.8)^2 \cdot 0.1 \end{aligned}$$

4. The number of credits that a UW-Madison course is worth follows this distribution, with some numbers removed. (Not real data)

x	1	2	3	4
$P(X = x)$	?	?	0.4	0.4

(a) (3 points) Suppose $P(X = 1)$ and $P(X = 2)$ are equal. What are $P(X = 1)$ and $P(X = 2)$?

Solution: Probabilities must sum to 100%. 1 - 0.4 - 0.4 = 0.2, both are equal so they are both 0.1.

(b) (3 points) Write out a mathematical expression for $E[X]$. You do not need to simplify your answer and can leave it as an expression.

Solution:

$$E(X) = 1(0.1) + 2(0.1) + 3(0.4) + 4(0.4)$$

(c) (3 points) Write out a mathematical expression for $Var[X]$. You can write $E[X]$ instead of the full expression above where needed. You do not need to simplify your answer and can leave it as an expression.

Solution: Var(X) $= [(1 - \text{E[X]})^2 \cdot 0.1] + [(2 - E[X])^2 \cdot 0.1] + [(3 - E[X])^2 \cdot 0.4] + [(4 - E[X])^2 \cdot 0.4]$

5. Based on historic data, a carnival believes that a player has a 7% chance of winning at a carnival basketball game. They model the number of winners (out of 350 total players) as Binom(350, 0.07).

   (a) (5 points) Identify the four assumptions that are being made for this binomial model, and briefly discuss how each one applies to this situation. Name one assumption that might not be well met, and explain.

   Solution:

   - B - the outcomes are binary, which is true for a game (each player either wins or loses)
   - I - the players are independent of each other, which is probably true
   - N - there is a fixed number of players, which is true ($n = 350$)
   - S - each player is equally likely (7% chance) to win.

   Out of all of the assumptions, the same probability assumption may not be well met. Some players might be more or less skilled at the game and therefore have a different probability of winning.

   (b) Assuming the binomial model is accurate, fill in the three blanks of `pbinom()` below to calculate the probability that 30 or more players out of 350 win basketball.

```
1 - pbinom(___29___, ___350___, ___0.07___)
```

   (c) Consider performing the calculation in (b) with a normal approximation. Fill in the four blanks of `pnorm()` below to approximate the probability that 30 or more players out of 350 win basketball.

   You don't need to simplify your answers for the blanks.

```
pnorm(___30___, _350*0.07_, _sqrt(350*0.07*0.93)_,
                    _lower.tail = F_)
```

6. Four different bird species, with different levels of rarity, live in a conservation area. Scientists regularly walk through the area and record the count of unique of bird species they observe (out of 4). A distribution for X, the number of species observed, is given below.

x	0	1	2	3	4
$P(X = x)$	0.2	0.4	0.25	0.1	0.05

(a) (5 points) For each of the four BINS assumptions, comment on whether the assumption is met or not met for X. Is X a binomial random variable?

Solution: Here, each "trial" is a species of bird.

- B - Each trial is binary, since we either see or do not see each species.
- I - We can safely assume that each bird species is independent.
- N - There are four trials/species.
- S - The probability of success is not the same, since some species are more or less common.

Since the same probability assumption is violated, X is not a binomial random variable.

(b) (3 points) What is the 75th percentile of X?

Solution: This 75th percentile is the smallest value x such that $P(X \leq x)$ is at least 0.75. We have $P(X \leq 1) = 0.6$ and $P(X \leq 2) = 0.85$ so the 75th percentile is 2.

(c) (3 points) Write an expression to calculate $E(X)$. You do not need to simplify your answer and can leave it as an expression.

Solution:

$$E(X) = 0(0.2) + 1(0.4) + 2(0.25) + 3(0.1) + 4(0.05)$$

7. Define four different independent binomial random variables.

$$X_1 \sim \text{Binom}(5, 0.6), \qquad X_2 \sim \text{Binom}(25, 0.4), \qquad X_3 \sim \text{Binom}(100, 0.1), \qquad X_4 \sim \text{Binom}(100, 0.4)$$

(a) (4 points) Fill in the blanks below with “less than", “greater than", or “equal to".

- The expectation of X_1 is _____greater than_____ 2.5.
- The smallest possible value of X_1 is _____equal to_____ the smallest possible value of of X_4.
- The expectation of X_3 is _____less than_____ the expectation of X_4.
- The standard deviation of X_3 is _____less than_____ the standard deviation of X_4.

(b) (3 points) Which of the resulting combinations below is also a binomial random variable? **Select ONE**.

- ○ $Y_1 = X_1 + X_2$
- ○ $Y_2 = X_2 + X_3$
- ● $Y_3 = X_2 + X_4$
- ○ $Y_4 = X_3 + X_4$

(c) (3 points) Which of the four variables is *least* well approximated as $N\left(np, \sqrt{np(1-p)}\right)$? **Select ONE**.

- ● X_1
- ○ X_2
- ○ X_3
- ○ X_4

8. 32 men competed in the 2024 Paris Qualification Round for shot put. If an athlete threw the ball at least 21.35 meters, they qualified for the final.

   (a) (5 points) Let X be the number of athletes who qualify for the final. Determine whether X meets the assumptions of a binomial random variable by briefly explaining why each individual assumption is or is not met.

   Solution: B: The trials are binary because each athlete either qualifies or does not qualify.

   I: The trials are independent because the result of one athlete does not affect the performance of the other athletes.

   N: There are a fixed number of trials ($n = 32$).

   S: The trials do not have the same probability, because the athletes have different skill levels and do not have the same probability of qualifying.

   Because the same probability assumption is not met, X is not binomial.

   (b) (4 points) Regardless of your answer to the previous question, assume that X, the number of athletes who qualify for the final, follows a binomial distribution, with some number of trials n and some probability of success p.

   Which lines of R code below can be used to calculate the probability that 10 or more athletes qualify for the final? **Select ALL that apply**.

   ◯ `pbinom(9, n, p)`
   ◯ `1 - pbinom(10, n, p)`
   ● `1 - pbinom(9, n, p)`
   ◯ `pbinom(10, n, p)`

9. (a) (3 points) For the three random variables listed below, identify its distribution, which will be either binomial or normal. Be sure to write which *specific* binomial or normal RV it follows.

   i. A multiple-choice test has 15 questions, each of which has five choices. A specific unprepared student takes the test and has a 80% chance of getting a question wrong. Let X_1 denote the number of answers that the student gets right in the test. Find the distribution of X_1.

   Solution: $X_1 \sim Binom(15, 0.2)$

   ii. Suppose that grade points of undergraduate students at a university have a distribution with a mean of 2.63 and a variance of 0.15. Let X_2 denote the average grade points of 100 undergraduate students. Find the distribution of X_2.

   Solution: $X_2 \sim N\left(2.63, \sqrt{\frac{0.15}{100}}\right)$

   iii. An experimental medication was given to 3000 patients with a certain medical condition. It is previously known that 2.6% people show severe side effects to a component in the medication. Let X_3 denote the number of patients who develop severe side effects. Find the distribution of X_3.

   Solution: $X_3 \sim Binom(3000, 0.026)$

(b) (3 points) Out of the variables binomially distributed in (a), which of those can be most accurately approximated by a normal distribution? What will be the normal approximation for that variable?

Solution: The RV X_3 in (a.iii) is approximately normal since we have a very large sample size n, even though the probability p is close to 0. The normal approximation is

$$X_3 \dot{\sim} N\left(3000(0.026), \sqrt{3000(0.026)(1-0.026)}\right)$$

10. Let $M \sim N(10, 1)$ denote the shoe sizes for males in the U.S. The area under the curve between shoe size 8 and 12 is 95%, i.e. $P(8 \leq M \leq 12) = 0.95$.

    Let $F \sim N(8, 0.8)$ denote the shoe sizes for females in the U.S.

    (a) (3 points) Which is taller - the height of the peak of the M bell-curve or the height of the peak of the F bell-curve? Briefly explain your answer.

    Solution: The height of F must be taller than the height of M because F has a smaller standard deviation, and therefore the bell curve for F is more concentrated around 8 than the bell curve for M is concentrated around 10.

    (b) (4 points) Find the points lo and hi such that the area under the curve for F is 95%. In other words, $P(lo \leq F \leq hi) = 0.95$. You do not need to simplify your answers and can leave them as an expression.

    Solution: We are given the percentiles that cover 95% of the area under the curve for M. If we convert these into z-scores, we get

    $$z_{lo} = \frac{8-10}{1} = -2, \qquad z_{hi} = \frac{12-10}{2} = 2$$

    (two standard deviations below and two standard deviations above the mean). Then we can convert these into values on F.

    $$lo = 0.8(-2) + 8 = 6.4, \qquad hi = 0.8(2) + 8 = 9.6$$

    So $P(6.4 \leq F \leq 9.6) = 0.95$.

    (c) (3 points) We want to confirm that the values `lo` and `hi` from part (a) actually do have 95% of the area between them. Which lines of R code return the area between `lo` and `hi`? **Select ALL that apply**.

    - ◯ `2*pnorm(lo, 8, 0.8)`
    - ◯ `2*pnorm(hi, 8, 0.8)`
    - ● `1 - 2*pnorm(lo, 8, 0.8)`
    - ◯ `1 - 2*pnorm(hi, 8, 0.8)`
    - ◯ `pnorm(lo, 8, 0.8) - pnorm(hi, 8, 0.8)`
    - ● `pnorm(hi, 8, 0.8) - pnorm(lo, 8, 0.8)`

11. Consider a standard normal variable $Z \sim N(0,1)$. The table below lists several lower-tail probabilities of Z.

Z-score	$P(Z \leq -1.5)$	$P(Z \leq -1)$	$P(Z \leq -0.5)$	$P(Z \leq 0.5)$	$P(Z \leq 1)$	$P(Z \leq 1.5)$
`pnorm`	`pnorm(-1.5)`	`pnorm(-1)`	`pnorm(-0.5)`	`pnorm(0.5)`	`pnorm(1)`	`pnorm(1.5)`
Probability	0.067	0.159	0.309	0.691	0.841	0.933

Use the table of probabilities to answer the questions below.

(a) (3 points) What is $P(-1 \leq Z \leq 0.5)$? **Select ONE**.

- ◯ 0.691
- ● $0.691 - 0.159$
- ◯ $0.691 + 0.159$
- ◯ $1 - (0.691 + 0.159)$

(b) (3 points) What is $P(Z \leq -1 \text{ OR } Z > 1)$? **Select ONE**.

- ◯ 0
- ◯ 0.159
- ● 2×0.159
- ◯ 0.691
- ◯ 2×0.691
- ◯ 1

(c) (3 points) What value is closest to the 95th percentile of Z? **Select ONE**.

- ◯ -1.5
- ◯ -1
- ◯ -0.5
- ◯ 0.5
- ◯ 1
- ● 1.5

12. Assume that the monthly electricity consumptions of all households in a certain region is approximately normally distributed with a mean of 1200 kilowatt-hours and a standard deviation of 110 kilowatt-hours.

(a) (3 points) John Smith recieved a notice informing him that his household electricity consumption lies 2 standard deviation above the mean. What is John Smith's household electricity consumption? You do not need to simplify your answer and can leave it as an expression.

Solution:

$$x = 1200 + 2(110)$$

(b) (3 points) Find the value of average electricity consumption in 100 houses such that it lies 2 standard deviation below the mean of the *sampling distribution of sample means*. You do not need to simplify your answer and can leave it as an expression.

Solution: The mean electricity consumption of 100 housees is $\bar{X} \sim N(1200, \frac{110}{\sqrt{100}}) = N(1200, 11)$. A value of the sample mean two SDs below average is

$$\bar{x} = 1200 - 2(11)$$

(c) (4 points) Using the *d*, *p* or *q* function for the normal distribution, fill in the blanks below such that you get R code for the interquartile range of the monthly electricity consumptions of all households in that region.

The interquartile range is the difference between the 25th percentile and the 75th percentile of the data/distribution, or the range of the middle 50% of the data/distribution.

```
IQR = ___qnorm___(___0.75___, 1200, 110) -
      ___qnorm___(___0.25___, 1200, 110)
```

13. You are interested in comparing the performance of Student A at University A with Student B at University B.

Grade point averages at University A are normally distributed with mean 2.8 and standard deviation 0.4. Grade point averages at University B are normally distributed with mean 2.5 and standard deviation 0.5.

(a) (3 points) Student A has a GPA of 3.2 at University A. Which R code below calculates the GPA of Student B who has the same percentile GPA at University B? **Select ONE**.

- ● textttpnorm(3.2, 2.8, 0.4) %>% qnorm(2.5, 0.5)
- ○ `pnorm(3.2, 2.5, 0.5) %>% qnorm(2.8, 0.4)`
- ○ `qnorm(3.2, 2.8, 0.4) %>% pnorm(2.5, 0.5)`
- ○ `qnorm(3.2, 2.5, 0.5) %>% pnorm(2.8, 0.4)`

(b) (3 points) What GPA score from University A corresponds to a z-score of -1? **Select ONE**.

- ○ -0.4
- ○ 0.4
- ● 2.4
- ○ 2.8

(c) (3 points) Consider taking a random sample of five students from University A. What is the distribution of $\bar{X}$, the average of their GPA's? **Select ONE**.

- ○ $\bar{X} \dot{\sim} N(2.8, 0.4)$
- ○ $\bar{X} \dot{\sim} N(2.8, 0.4\sqrt{5})$
- ● $\bar{X} \dot{\sim} N(2.8, \frac{0.4}{\sqrt{5}})$
- ○ None of the above

14. At a certain hospital, the number of liters of blood donated per week is given by normal variable D with $\mu_D = 51, \sigma_D = 7$. The number of liters of blood required for transfusions per week is given by normal variable R with $\mu_R = 56, \sigma_R = 4$. You can assume R and D are independent.

(a) (4 points) The number of liters required this week is $r = 48$. What number of donated liters has the same percentile as $r = 48$? In other words, if $P(R \leq 48) = p$, find d such that $P(D \leq d) = p$.

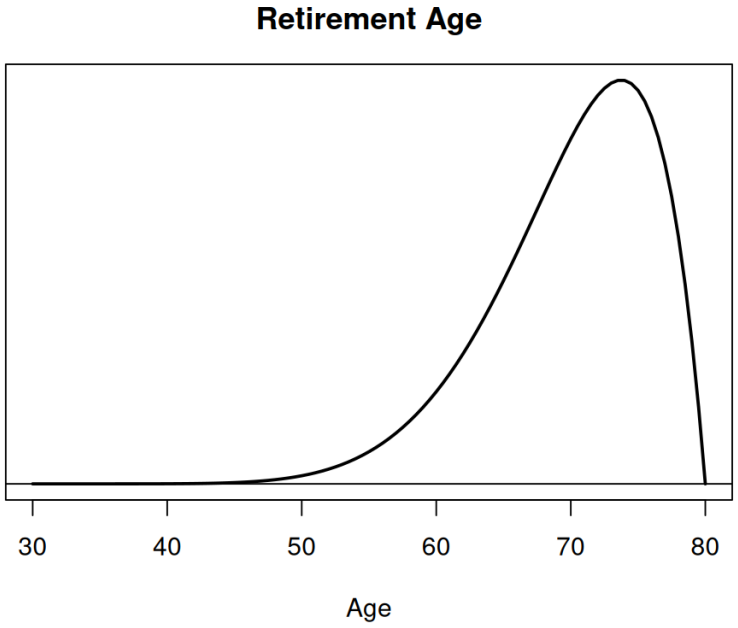
Solution: The value $r = 48$ is exactly 2 standard deviations below the mean of $\mu_R = 56$, since $56-2(4) = 48$.

The corresponding value of d will be two standard deviations below the mean $\mu_D = 51$. This is $d = 51 - 2(7) = 37$.

(b) (3 points) Identify the sampling distribution for $\bar{D}_{52}$, the average liters of blood donated per week over a 52-week year.

$$\bar{D}_{52} \sim \underline{\quad \text{N} \quad}(\underline{\quad 51 \quad}, \underline{\quad 7/\sqrt{52} \quad})$$

15. The continuous random variable X, shown below, represents the population of retirement ages in the US. X is defined on [30, 80] and has has expected value 64 and variance 92.16.



(a) (3 points) **Select ONE**. X will realize to a value greater than 64:

- ◯ less than half the time
- ◯ close to half the time
- ● more than half the time

(b) (3 points) **Select ONE**. The sample mean $\bar{X}_{75}$ is calculated by taking the average of 75 random draws from X. $\bar{X}_{75}$ will realize to a value greater than 64:

- ◯ less than half the time
- ● close to half the time
- ◯ more than half the time

(c) (5 points) Fill in the blanks to write R code to approximate $P(\bar{X}_{75} \geq 62)$. Your answer should be a function with four arguments.

`pnorm( 62 , 64 , sqrt(92.16/75) , lower.tail = F )`

16. The plot below shows the densities of three different distributions, named Solid, Dashed, and Dotted. One of these distributions is $N(0,1)$, standard normal, another one is $t(5)$, the t distribution with 5 degrees of freedom, and the final one is $D(0,1)$, some other distribution with mean 0 and standard deviation 1.

(a) (3 points) Identify each distribution by the corresponding line type.

$N(0,1)$:	● Solid	○ Dashed	○ Dotted
$t(5)$:	○ Solid	○ Dashed	● Dotted
$D(0,1)$:	○ Solid	● Dashed	○ Dotted

(b) (3 points) Rank the three distributions using their line type names (Solid, Dashed, Dotted) in order from that with the smallest to the largest values of the 0.975 quantile.

(Smallest) Dashed < Solid < Dotted (Largest)

(c) (3 points) A random sample of size $n = 400$ is drawn from distribution $D(0,1)$ and we calculate the sample mean

$$\bar{X} = \frac{1}{400}\sum_{i=1}^{400} X_i.$$

What are the mean and standard deviation of the distribution of $\bar{X}$? **Select ONE**.

○ $\mu_{\bar{X}} = 0, \quad \sigma_{\bar{X}} = 1$

○ $\mu_{\bar{X}} = 0, \quad \sigma_{\bar{X}} = \frac{1}{400}$

○ $\mu_{\bar{X}} = 400, \quad \sigma_{\bar{X}} = 400$

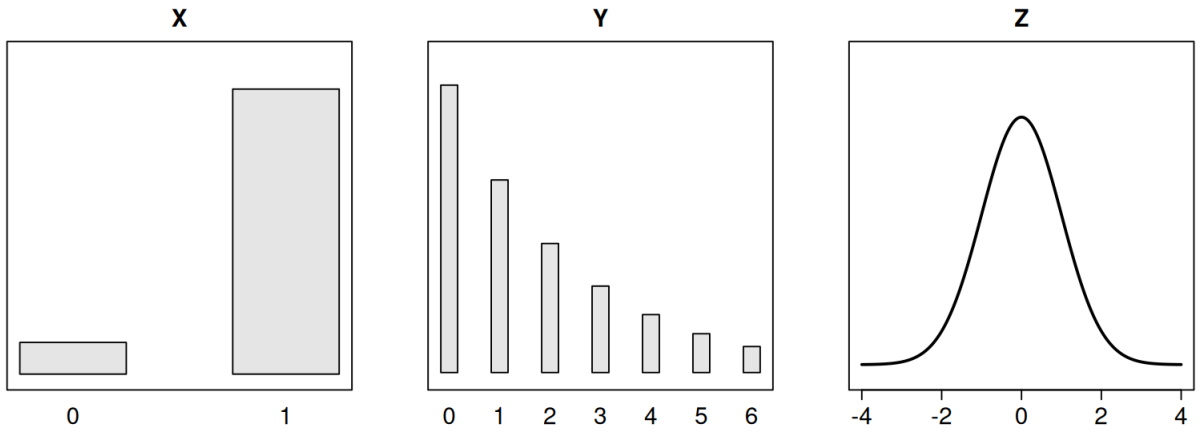
● $\mu_{\bar{X}} = 0, \quad \sigma_{\bar{X}} = \frac{1}{\sqrt{400}}$

(d) (3 points) Which of the three distributions will have a shape closest to that of the distribution of $\bar{X}$? (Note that the scale will be different.) **Select ONE**.

● $N(0,1)$ ○ $t(4)$ ○ $D(0,1)$

5 Inference

1. Consider the following random variables: X is discrete and can take on values 0 or 1, Y is a different skewed discrete variable, and Z is normal.



(a) (5 points) Which of the following will have a normal or approximately normal distribution? **Select ALL that apply**.

- ● The average of 200 samples from X.
- ● The average of 100 samples from Y.
- ○ The median of 100 samples from Y.
- ● The average of 1 sample from Z.
- ● The sum of 200 samples from X.

A researcher who does not know the true distribution of X wants to estimate $p = P(X = 1)$. They report a 95% confidence interval for p as $(0.845, 0.921)$.

(b) (3 points) Which line of R code correctly calculates the critical value / quantile score for the 95% CI? **Select ONE**.

- ○ `qnorm(0.9)`
- ○ `qnorm(0.95)`
- ● `qnorm(0.975)`
- ○ `qnorm(0.99)`

(c) (4 points) Suppose the researcher instead set $\alpha = 0.02$ to determine the confidence level. Assuming the data does not change, how does this interval compare to the 95% CI above? Fill in the blanks with "a smaller", "the same", or "a larger".

The new CI would have ___a larger___ margin of error compared to the 95% CI.

The new CI would have ___the same___ point estimate compared to the 95% CI.

2. A doctor records the cholesterol level of 20 of her patients (in mg/dL). She reports a 95% confidence interval for the average cholesterol:

$$(147.29,\ 215.71)$$

(a) (3 points) Identify the point estimate and margin of error of this CI. You do not need to simplify your answer.

Solution: The point estimate is the midpoint of the interval:

$$\text{PE} = \frac{147.29 + 215.71}{2}$$

To get the upper bound, we add the margin of error to the point estimate, and to get the lower bound, we subtract the margin of error from the point estimate. So, the MOE is the half-width of the interval. Here are three ways to calculate this:

$$\text{MOE} = \text{PE} - 147.29 = 215.71 - \text{PE} = \frac{215.71 - 147.29}{2}$$

(b) (3 points) Which of the following statements is the best interpretation of this CI? **Select ONE**.

- ◯ If we repeated the CI procedure across multiple samples, 95% of the intervals would cover the cholesterol level for the patients in the sample.
- ● If we repeated the CI procedure across multiple samples, 95% of the intervals would cover the mean cholesterol level for all patients.
- ◯ 95% of the 20 patients in this sample have a cholesterol level between 147.29 and 215.71.
- ◯ 95% of all patients have a cholesterol level between 147.29 and 215.71.

(c) (3 points) The doctor wants to be especially sure that her confidence interval correctly covers the true mean cholesterol of her patients. What change should she make to the 95% interval above? **Select ONE**.

- ◯ Choose a smaller confidence level, which gives a narrower interval.
- ◯ Choose a larger confidence level, which gives a narrower interval.
- ◯ Choose a smaller confidence level, which gives a wider interval.
- ● Choose a larger confidence level, which gives a wider interval.

3. A 95% confidence interval for the true proportion of left-handed Americans is reported as (0.086, 0.112).

   (a) (3 points) Which line of R code correctly calculates the critical value / quantile score for the margin of a 95% CI? **Select ONE**.

   ◯ `qnorm(0.05)`
   ◯ `qnorm(0.95)`
   ◯ `qnorm(0.9)`
   ● `qnorm(0.975)`

   (b) (4 points) Which of the following statements are true about the CI? **Select ALL that apply**.

   ◯ A 98% CI on the same data would be narrower than the 95% CI.
   ● A 98% CI on the same data would be wider than the 95% CI.
   ● A 98% CI would have the same point estimate as the 95% CI.
   ◯ A 98% CI would have a different point estimate as the 95% CI.

   (c) (3 points) A hypothesis test of $H_0 : p = 0.1$ versus $H_A : p \neq 0.1$ (where p is the true proportion of left-handed Americans) is performed. The test resulted in a p-value smaller than the chosen significance level $\alpha = 0.05$. Which of the following statements is the best conclusion of this test? **Select ONE**.

   ◯ The true proportion of left-handed Americans is 0.1.
   ◯ The true proportion of left-handed Americans is not 0.1.
   ◯ We do not have evidence that the true proportion of left-handed Americans is different from 0.1.
   ● We have evidence that the true proportion of left-handed Americans is different from 0.1.

4. We are interested in the true percentage, p, of UW-Madison students who know that Bucky Badger's full first name is Buckingham. We ask 100 students, and construct a 90% confidence interval for p as (0.50, 0.70).

   (a) (3 points) Which of the following correctly calculates the critical value for a 90% CI? **Select ONE**.

   - ◯ `qnorm(0.9)`
   - ● `qnorm(0.95)`
   - ◯ `qnorm(0.975)`
   - ◯ `dbinom(2, 3, 0.4)`

   (b) (4 points) Which of the following statements are TRUE about this interval? **Select ALL that apply**.

   - ● A 95% confidence interval would have a larger margin of error than this one.
   - ◯ A 95% confidence interval would have the same margin of error as this one.
   - ● If we had asked 500 students AND decreased confidence level to 80%, the resulting interval would be narrower than this one.
   - ◯ If we had asked 500 students AND decreased confidence level to 80%, we do not know if the resulting interval would be narrower or wider than this one.

   (c) (3 points) Which of the following conclusions can be made from this confidence interval? **Select ONE**.

   - ◯ At least half of all UW-Madison students know this fact.
   - ◯ p is below 0.8.
   - ● 60% of the students sampled knew this fact.
   - ◯ We are 95% confident that 60% of all UW students know this fact.

5. Carabiners made at a factory must have a breaking strength of at least 4500 lbs ($H_0 : \mu \geq 4500$). A manager thinks that the mean breaking strength may be less than 4500 lbs ($H_A : \mu < 4500$) so he takes a sample of carabiners to test his hypothesis with $\alpha = 0.1$. If he has evidence that $\mu < 4500$, he will halt carabiner production.

   (a) (5 points) Classify each scenario as a type I error, type II error, or correct decision. **Select ONE each.**

   i. The true mean breaking strength is 4500 lbs. The manager allows production to continue.

   ○ Type I error ○ Type II error ● Correct decision

   ii. The true mean breaking strength is 4500 lbs. The manager halts production.

   ● Type I error ○ Type II error ○ Correct decision

   iii. The true mean breaking strength is 4300 lbs. The manager allows production to continue.

   ○ Type I error ● Type II error ○ Correct decision

   iv. The true mean breaking strength is 4300 lbs. The manager halts production.

   ○ Type I error ○ Type II error ● Correct decision

   v. The true mean breaking strength is 4800 lbs. The manager allows production to continue.

   ○ Type I error ○ Type II error ● Correct decision

   (b) (3 points) Explain whether the manager should make α larger or smaller to avoid mistakenly selling carabiners that are not strong enough.

   Solution: The manager should make α larger in order to avoid making a type II error. A type II, or false negative error, would mean that the alternative is true (the breaking strength is too low) but the manager's test failed to detect this. By making α larger, he makes it easier to reject the null in favor of the alternative.

6. A manager at a campus cafeteria is deciding what pizza toppings to offer. He thinks that if over 25% of his customers like pineapple on pizza, he will offer pineapple as a topping.

   He takes a sample of 45 customers and finds that 14 of them like pineapple on pizza.

   (a) (3 points) Identify null and alternative hypotheses to answer the manager's question. They should be in terms of p, the proportion of customers who like pineapple on pizza.

   Solution:

   $$H_0 : p = 0.25 \text{ or } p \leq 0.25$$

   $$H_A : p > 0.25$$

   (b) (3 points) The manager will model the number of customers who like pineapple as Binom(45, p). Which line of R code below correctly calculates the p-value for the test? **Select ONE**.

   - ◯ `pbinom(14, 45, 0.25)`
   - ◯ `1 - pbinom(14, 45, 0.25)`
   - ◯ `pbinom(14, 45, 14/45)`
   - ● `1 - pbinom(13, 45, 0.25)`
   - ◯ `1 - pbinom(15, 45, 0.25)`
   - ◯ `1 - pbinom(15, 45, 14/45)`

   (c) (5 points) The observed p-value is greater than the manager's chosen significance level $\alpha = 0.05$. Which of the following statements are true? **Select ALL that apply**.

   - ◯ We are certain that 25% or less of all customers like pineapple on pizza.
   - ● It is possible that more than 25% of all customers like pineapple on pizza.
   - ● We do not have evidence that more than 25% of all customers like pineapple on pizza.
   - ◯ We could have made a Type I error.
   - ● We could have made a Type II error.

7. A spinner at a carnival will either land on “Lose" or “Win" when spun. Assume spins are independent and each one has the same probability of landing on “Win".

   (a) (3 points) The carnival claims the spinner has a win probability of $p = 0.4$. Assuming this is correct, which lines of R code below can be used to find the probability of winning exactly twice in 3 spins? **Select ALL that apply**.

   - ◯ `0.6*0.6`
   - ◯ `0.6*0.6*0.4`
   - ● `3*0.6*0.4*0.4`
   - ● `dbinom(2, 3, 0.4)`
   - ◯ `dbinom(2, 3, 0.6)`
   - ● `dbinom(1, 3, 0.6)`
   - ◯ `pbinom(2, 3, 0.4)`
   - ◯ `pbinom(2, 3, 0.6)`
   - ◯ `pbinom(1, 3, 0.4)`

   (b) (3 points) You suspect that the true win probability p is actually less than 0.4. Write a null and alternative hypothesis to test the value of p.

   Solution: We are specifically looking for evidence that p is less than 0.4, so that must be our alternative.

   $$H_0 : p = 0.4 \quad \text{versus} \quad H_A : p < 0.4$$

   or

   $$H_0 : p \geq 0.4 \quad \text{versus} \quad H_A : p < 0.4$$

   (c) (3 points) You decide to complete the test with significance level $\alpha = 0.05$. You spin the spinner 10 times and win 2 times, giving a p-value of $P\Big(Binom(10, 0.4) \leq 2\Big) = 0.167$. What is the most accurate conclusion of the test? **Select ONE**.

   - ● Because our p-value is greater than α, we do not have evidence of $p < 0.4$.
   - ◯ Because our p-value is greater than α, we have evidence of $p < 0.4$.
   - ◯ We have evidence that the true $p = 0.4$.
   - ◯ We have evidence that the true $p = 0.2$.

6 Proportions

1. Consider trying to estimate the proportion of UW-Madison students that will graduate at the end of the semester, p. You ask 50 random students, and 7 of them will graduate at the end of the semester.

   We are specifically interested in using a confidence interval or hypothesis test to compare p to 0.15.

   (a) (4 points) Which of the following statements are true? **Select ALL that apply**.

   - ◯ The center of a confidence interval for p would be 0.15.
   - ● The center of a confidence interval for p would be 7/50.
   - ● The standard error of a confidence interval for p would be approximate.
   - ◯ As we increase the confidence level toward 100%, our interval would widen and approach [0, 100].

   (b) (3 points) Consider performing a hypothesis test of whether **greater than 15%** of students will graduate at the end of the semester. Write appropriate hypotheses and identify the null distribution for this test.

   Solution: The hypotheses are

   $$H_0 : p = 0.15 \quad \text{versus} \quad H_A : p > 0.15$$

   where p is the true proportion of students graduating at the end of the semester.

   We can do this test with either a binomial or a normal null. So, the null would be $Binom(50, 0.15)$ or $N(0, 1)$.

2. (4 points) We are interested in comparing the proportion of individuals with bachelor's degrees among adult men 25 and older in Minnesota and Wisconsin. Let p_M and p_W represent these two population proportions. In random samples of $n_M = 300$ Minnesota men and $n_W = 400$ Wisconsin men 25 and older, the number of individuals with bachelor's degrees is $x_M = 90$ and $x_W = 100$.

   (a) (4 points) Without simplification, write a numerical expression using the provided data for the standard error of the test statistic Z in a hypothesis test for the equality of proportions p_M and p_W.

$$Z = \frac{\hat{p}_M - \hat{p}_W}{\text{SE}} \sim N(0,1)$$

Solution: We need to calculate a pooled proportion since we are testing for an equality of proportions. This is $\hat{p} = \frac{90+100}{300+400} = \frac{190}{700}$. The standard error for the difference in proportions is

$$\hat{se}(\hat{p}_M - \hat{p}_W) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_m} + \frac{1}{n_W}\right)} = \sqrt{\frac{190}{700}\left(1 - \frac{190}{700}\right)\left(\frac{1}{300} - \frac{1}{400}\right)}$$

   (b) (5 points) Suppose that the test statistic from the previous problem has value $z = 1.47$ with two-sided p-value 0.14. Which of the following statements are true? **Select ALL that apply**.

- [x] In the sample data, the proportion of Minnesota men with a bachelor's degree is greater than the proportion of Wisconsin men with a bachelor's degree.
- [ ] The test is statistically significant at the 5% level.
- [ ] There is proof beyond a reasonable doubt that Minnesota has a higher proportion or individuals with college degrees among men aged 25 and older than Wisconsin does.
- [x] There is insufficient evidence to conclude that there is a difference in the proportions of individuals with college degrees among men aged 25 and older between Minnesota and Wisconsin, using a 5% significance level.
- [x] A 95% confidence interval for $p_M p_W$ will contain the value 0.

3. (5 points) The HR director at a company plans to ask a sample of employees about whether they are satisfied with their job. He will use the sample to make a 95% CI for p, the true proportion of employees that are satisfied.

   Assume the director is using standard error $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and is using critical value 1.96. How big of a sample size n does he need to take so that the margin of error of the interval will be **at most** 0.02?

   Write an expression and then solve for n. You do not have to report the specific value of n but n should be the only unknown.

Solution: The standard error depends on $\hat{p}$, and it is largest when $\hat{p} = 0.5$. So, the required n is given by:

$$0.02 = 1.96\sqrt{\frac{0.5 \times 0.5}{n}}$$

$$\left(\frac{0.02}{1.96}\right)^2 = \frac{0.25}{n}$$

$$n = \frac{0.25}{(\frac{0.02}{1.96})^2}$$

4. 30 study participants were divided into groups of 15 and randomly selected to watch either a funny video or a sad video. Afterwards, each participant was asked if they wanted to give away half of their compensation money to a charity unrelated to the video they watched.

   Let p_{funny} be the true proportion of people who view the funny video who choose to give money away. Let p_{sad} be the true proportion of people who view the sad video who choosen to give money away. A Z test of hypotheses

   $$H_0 : p_{funny} = p_{sad} \quad \text{versus} \quad H_A : p_{funny} > p_{sad}$$

   resulted in a p-value of 0.8.

   (a) (4 points) Report the p-value for the following pairs of hypotheses.

   - $H_0 : p_{funny} = p_{sad}$ versus $H_A : p_{funny} < p_{sad}$

     Solution: The p-value is $1 - 0.8 = 0.2$.

   - $H_0 : p_{funny} = p_{sad}$ versus $H_A : p_{funny} \neq p_{sad}$

     Solution: The p-value is $2 \times 0.2 = 0.4$.

   (b) (4 points) Which of the following statements about the study are true? **Select ALL that apply**.

   - ◯ More participants who watched the funny video decided to give money away.
   - ● More participants who watched the sad video decided to give money away.
   - ● We do not have evidence at the 5% level that the groups have a different probability of giing money away.
   - ◯ We have evidence at the 5% level that the groups have a different probability of giving money away.

5. From 1999 to 2008, the U.S. Mint released a huge quantity of special commemorative quarters (25 cent coins), each of which had a unique design on the back celebrating one of the fifty states. Let p be the true proportion of all the commemorative quarters that have a Wisconsin design on the back. Consider the following hypothesis test:

$$H_0 : p = 1/50 \text{ vs. } H_\alpha : p > 1/50$$

You randomly collect 200 of these commemorative quarters and find that 11 have Wisconsin designs on the back.

In conducting this test, you may assume that any relevant assumptions are met and do not have to mention nor explain them.

(a) (3 points) What is the observed value of the test statistic for this test?

Solution: The test statistic is the count of "successes", which are Wisconsin quarters. The observed test statistic in this case is 11.

(b) (3 points) Assuming the null hypothesis is true, what is the sampling distribution of the test statistic?

$X \sim$ Binom (200 , 1/50)

(c) (3 points) Which line of R code correctly calculates the p-value for this test? **Select ONE**.

- ◯ `pbinom(11, 200, 1/50)`
- ◯ `pbinom(11-1, 200, 1/50)`
- ◯ `1 - pbinom(11, 200, 1/50)`
- ● `1 - pbinom(11-1, 200, 1/50)`

(d) (4 points) Assume the p-value is 0.003. **Select ALL** correct statements or justifiable conclusions.

- ● The result is statistically significant at the $\alpha = 0.05$ level.
- ● There is strong evidence that the proportion of special commemorative state coins with Wisconsin on the back is greater than 1/50.
- ◯ Based on the same data, a 95% confidence interval for the true proportion of Wisconsin quarters would contain the value 1/50.
- ◯ The p-value from a two-sided test with the same data would also be equal to 0.003.

6. A 95% confidence interval for the difference in population proportions $p_1 - p_2$ is

$$-0.1 < p_1 - p_2 < 0.2.$$

(a) (3 points) What is true about the p-value of a two-sided hypothesis test of $H_0 : p_1 = p_2$ versus $H_a : p_1 \neq p_2$? **Select ONE**.

- ◯ The p-value is less than 0.05.
- ◯ The p-value is equal to 0.05.
- ● The p-value is greater than 0.05.

(b) (3 points) Suppose $\hat{p}_1, \hat{p}_2$ and n_1, n_2 are the observed proportion of successes and the number of trials from each group, and $\hat{p}$ is the common proportion of successes across both groups. What is the point estimate for the above interval? **Select ONE**.

- ◯ 1
- ◯ $\hat{p}_2$
- ◯ $\hat{p}$
- ● $\hat{p}_1 - \hat{p}_2$

(c) (3 points) What is the standard error for the above interval? **Select ONE**.

- ◯ $\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$
- ◯ $\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} - \frac{1}{n_2})}$
- ● $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
- ◯ $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} - \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

7 Means

1. The output of R `t.test()` on two sets of data, `shopOne` and `shopTwo`, is given below.

```
        Welch Two Sample t-test

data:  shopOne and shopTwo
t = 0.019454, df = 91.984, p-value = 0.4923
alternative hypothesis: true difference in means is greater than -10
99 percent confidence interval:
 -12.20095       Inf
sample estimates:
mean of x mean of y
 139.7905  149.7723
```

(a) (4 points) Write a numerical expression for the standard error for the difference of means of Shop One and Shop Two. You do not need to solve or simplify.

Solution: From the test statistic formula, we have

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{\text{se}}$$

$$0.019454 = \frac{139.7905 - 149.7723 + 10}{\text{se}}$$

$$se = \frac{139.7905 - 149.7723 + 10}{0.019454}$$

(b) (6 points) How can we write the pair of hypotheses is being tested in the `t.test()` output above? **Select ALL that apply**.

- ○ $H_0 : \mu_X - \mu_Y \leq -10$ and $H_\alpha : \mu_X - \mu_Y \geq -10$
- ● $H_0 : \mu_X - \mu_Y \leq -10$ and $H_\alpha : \mu_X - \mu_Y > -10$
- ○ $H_0 : \mu_X - \mu_Y \neq -10$ and $H_\alpha : \mu_X - \mu_Y = -10$
- ● $H_0 : \mu_X - \mu_Y = -10$ and $H_\alpha : \mu_X - \mu_Y > -10$
- ○ $H_0 : \mu_X - \mu_Y = -10$ and $H_\alpha : \mu_X - \mu_Y \neq -10$
- ○ $H_0 : \mu_X - \mu_Y = -10$ and $H_\alpha : \mu_X - \mu_Y \geq -10$

(c) (3 points) Draw the correct statistical conclusion for the above hypothesis test. You answer should reference the `t.test()` output.

Solution: We have a large p-value, which means we have a non-significant result. We do not have evidence that the difference $\mu_X - \mu_Y$ is greater than -10.

2. A grocery store is concerned that chocolate bars from a supplier weigh less than the advertised 55 grams.

   (a) (3 points) What are the null and alternative hypotheses that can be used to test μ, the true average weight of the chocolate bars? **Select ONE**.

   ◯ $H_0 : \mu = 55$ versus $H_A : \mu \neq 55$
   ◯ $H_0 : \mu = 55$ versus $H_A : \mu > 55$
   ● $H_0 : \mu = 55$ versus $H_A : \mu < 55$
   ◯ $H_0 : \mu \neq 55$ versus $H_A : \mu = 55$
   ◯ $H_0 : \mu > 55$ versus $H_A : \mu = 55$
   ◯ $H_0 : \mu < 55$ versus $H_A : \mu = 55$

   (b) (4 points) The store takes a random sample of 40 of these chocolate bars and measure an average weight of 53 grams with standard deviation 2.5 grams.

   Write a numerical expression that calculates the test statistic for your hypotheses. Do not simplify, evaluate, or round. Also identify the T null distribution for this test.

   Solution: The test statistic is

   $$t_{obs} = \frac{53 - 55}{2.5/\sqrt{40}}$$

   and the null distribution is a T with $n - 1$ or 39 degrees of freedom.

   (c) (3 points) The store takes a different sample of 40 chocolate bars from a different brand and weighs them. The weights of the original 40 observations are in the R vector `brand1` and the weights of the new 40 observations are in the R vector `brand2`. Which R code can be used to test whether there is a difference in average weight between the two brands? **Select ONE**.

   ● `t.test(brand1, brand2)`
   ◯ `t.test(brand2, mu = 55)`
   ◯ `t.test(brand1, brand2, alternative = "greater")`
   ◯ `t.test(brand1, brand2, alternative = "less")`
   ◯ `t.test(brand1, brand2, paired = T)`

3. Identify each of the following scenarios as having independent or paired data.

   (a) (2 points) The performance of 200 students on the practice SAT and actual SAT exam, where student has their score recorded for both exams.

   ○ Independent ● Paired

   (b) (2 points) The SAT performance of 100 students from a 9:00 AM class and the SAT performance of 100 students from a 1:00 PM class.

   ● Independent ○ Paired

   (c) (2 points) The SAT performance of 100 pairs of siblings who go to the same school. Each of the 200 individuals has their own score recorded.

   ○ Independent ● Paired

4. (4 points) What lines of R code can be used to perform a paired T test on x and y? **Select ALL that apply**.

   ○ `t.test(x, y)`
   ● `t.test(x - y)`
   ● `t.test(x, y, paired = T)`
   ○ `t.test(x - y, paired = T)`

5. In a large bookstore, you are trying to determine if the average number of pages of a book in the sci-fi section is *different* from the average number of pages of a book in the mystery section.

   - Let μ_S be the average page length of all books in the sci-fi section.
   - Let μ_M be the average page length of all books in the mystery section.

   (a) (3 points) State the appropriate null and alternative hypotheses for this test.

   Solution: We want to know if the average is different, which implies a two-sided test for a difference in means.

   $$H_0 : \mu_S - \mu_M = 0 \quad \text{versus} \quad H_A : \mu_S - \mu_M \neq 0$$

   (b) (3 points) You take a random sample of 30 sci-fi books and find the average page length to be 236.4 with standard deviation 23.8. You take a random sample of 40 mystery books and find the average page length to be 231.3 with standard deviation 52.6.

   Write a numerical expression that calculates the test statistic for your hypotheses. Do not simplify, evaluate, or round.

   Solution: We are using the Welch T test statistic for a difference in means.

   $$t_{obs} = \frac{\bar{x}_S - \bar{x}_M - 0}{\sqrt{\frac{s_S^2}{n_S} + \frac{s_M^2}{n_M}}} = \frac{236.4 - 231.3 - 0}{\sqrt{\frac{23.8^2}{30} + \frac{52.6^2}{40}}}$$

   (c) (4 points) Consider the following output of `t.test()` on this data. **Select ALL** true statements or justifiable conclusions.

```
        Welch Two Sample t-test

data:  scifi and mystery
t = 0.54033, df = 57.381, p-value = 0.5911
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.72092  23.86401
sample estimates:
mean of x mean of y
 236.3856  231.3140
```

   - ◯ There is strong evidence that the population mean length of sci fi books is larger than that of mystery books.
   - ● The sample mean length of sci fi books is greater than the sample mean of mystery books.
   - ◯ The two-sided hypothesis test is significant at the $\alpha = 0.05$ level.
   - ● There is a lack of evidence that the population mean lengths of either book genre differ from each other.

6. You ask a group of 30 people from Country A to each report the number of pets they own. You then ask a group of 40 people from Country B to each report the number of pets they own.

   Let μ_A be the true average value that all members of country A would give upon this request, and similarly for μ. You are interested in the quantity $\mu_A - \mu_B$.

   (a) (3 points) Which of the following statements are true about how to approach this problem through statistical inference? **Select ONE**.

   - ◯ Two-sample means inference is impossible to conduct because the sample sizes are different.
   - ◯ Two-sample means inference is impossible to conduct because the random variables are technically discrete.
   - ● Two-sample inference is more appropriate than paired-sample inference in this case.
   - ◯ Paired-sample inference is more appropriate than two-sample inference in this case.

   (b) (3 points) Assume the correct value of the test statistic from problem 14 is -4.12. If we are trying to find evidence that people from country A own fewer pets on average than in country B, which R code correctly calculates the p-value for this test? **Select ONE**.

   - ● `pt(-4.12, df = W)`
   - ◯ `2*pt(-4.12, df = W)`
   - ◯ `2*pt(abs(-4.12), df = W)`
   - ◯ `1 - pt(-50, df = W)`

8 Linear Regression

1. The `trees` has 31 observations with the Girth, Volume and Height of cherry trees. Below is the coefficient table of a linear regression model for volume vs girth on the `trees` data fit with `lm()`.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Use the R functions `pt(q, df, lower.tail)` and `qt(p, df, lower.tail)` to answer parts (a) and (b) below.

(a) (3 points) uild a 90% confidence interval for the slope β_1. Use one of the R functions above to write the expression for the critical value.

Solution: The estimated slope is 5.0659 with standard error 0.2474. We need to use the 95th percentile on the $T(n-2)$ distribution as our critical value. So, the CI is given by

$$5.0659 \pm \texttt{qt(0.95, df = 31-2)} \times 0.2474$$

(b) (4 points) Based on the output of the table, identify the test statistic for testing $H_A : \beta_1 < 0$. Use one of the R functions above to write the expression for the correct p-value for this test.

Solution: The test statistic is 20.48, the same as the test statistic given in the output for testing $\beta_1 \neq 0$. The p-value for our one-sided test is the area below our test statistic on the null $T(n-2)$ distribution. This is

```
pt(20.48, df = 31-2)
```

(c) (3 points) **Select ONE**. The p-value in (b) is approximately equal to:

- ◯ $< 2 \times 10^{-16}$
- ◯ $< 1 \times 10^{-16}$
- ◯ 0.05
- ● 1

2. The dataset `skincancer` records the skin cancer mortality (`Mort`) in deaths per 10 million and the latitute (`Lat`) in degrees of the 48 contiguous US states. For example, Florida has latitude 28.0, and Wisconsin has latitude 44.5.

   Below is a partial `summary` output for the linear model for skin cancer mortality (Y) versus latitude (X).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  388.312     23.767  16.338  < 2e-16 ***
Lat           -5.966      0.597  -9.994 4.15e-13 ***
---
```

   (a) (3 points) Let β_1 be the true linear relationship between skin cancer mortality and latitude. Identify the test statistic and degrees of freedom for the following hypotheses:

$$H_0 : \beta_1 = -4 \quad \text{versus} \quad H_A : \beta_1 \neq -4$$

   Solution: The test statistic is given by

$$t_{obs} = \frac{-5.966 - 4}{0.597}$$

   with $n - 2 = 46$ degrees of freedom.

   (b) (3 points) **Select ONE**. We know that that the p-value for the test in (a) must be:

   - ◯ Less than 4.15×10^{-13}
   - ◯ Equal to 4.15×10^{-13}
   - ● Greater than 4.15×10^{-13}

   (c) (4 points) A 95% confidence interval for the average mortality at a latitude of 35 is reported as (171.7, 187.3). Which of the following statments are true about a 95% prediction interval for mortality rate at $x = 35$? **Select ALL that apply**.

   - ● The center of the PI is the center of (171.7, 187.3).
   - ◯ The center of the PI is different from the center of (171.7, 187.3).
   - ◯ The PI is narrower than (171.7, 187.3).
   - ● The PI is wider than (171.7, 187.3).

3. A regression line is fit on 50 (x, y) pairs. Summaries of the data and the regression line are given below.

	X	Y
Mean	26.8	80.1
Range	[10.5, 39.8]	[29.8, 123.4]

Regression line: $\hat{y}_i = 4.06 + 2.84x_i$

(a) (3 points) 95% prediction interval for y at $x = 50$ is reported as (129.15, 162.25). Why would this result be considered extrapolation? **Select ONE**.

- ● The value 50 is outside of the range of the original X values.
- ○ The lower bound of the interval is outside the range of the original Y values.
- ○ The interval fails to cover the mean of the original Y values.
- ○ None of the above; this is not considered extrapolation.

(b) (3 points) A test for the linear regression slope against null hypothesis $H_0 : \beta_1 = 0$ results in a test statistic of $t_{obs} = 20.8$. What is the standard error of the estimated slope?

Solution: Using the formula for a slope test statistic, we have

$$\text{Test stat} = \frac{\text{Estimated slope} - \text{Null}}{\text{Standard error}}$$

$$20.8 = \frac{2.84 - 0}{se}$$

$$se = \frac{2.84}{20.8}$$

(c) (3 points) Based on the value of test statistic, which of the following conclusions is most reasonable? **Select ONE**.

- ○ X and Y have a negative relationship.
- ● X and Y have a positive relationship.
- ○ X and Y have no linear relationship relationship.
- ○ The value of X causes a significant change in the value of Y.

4. 74 individuals were asked to give their annual income and rate their life satisfaction (out of 100). A linear model is fit on life satisfaction (y) in terms of income (x).

   The predicted satisfaction for an annual income of $x^* = \$70,000$ is $\hat{y} = 61.2$ and a 95% prediction interval at $x^* = \$70,000$ is given by

   $$(38.4, 84.0)$$

   (a) (4 points) Which of these changes will result in a narrower interval? **Select ALL that apply**.

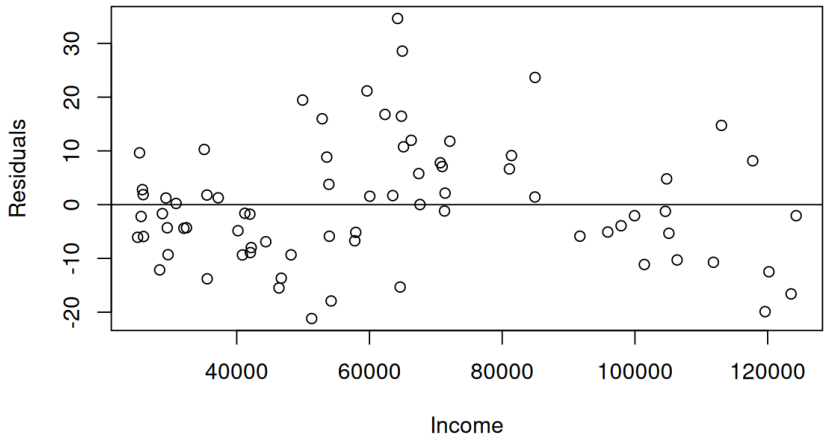
   - [x] Using a confidence level of 90%.
   - [ ] Using a confidence level of 98%.
   - [x] Building a confidence interval instead of a prediction interval.
   - [x] Building a prediction interval at $\bar{x} = 62900$ instead of $x = 70000$.

   (b) (3 points) The y-intercept of the linear model is 22.8. Write a numerical expression for the slope of the linear model. You do not need to solve or simplify.

   Solution: We know that the linear model passes through (70000, 61.2) and also passess through (0, 22.8). So, the slope of the model is given by

   $$\hat{\beta}_1 = \frac{61.2 - 22.8}{70000 - 0}$$

   (c) (3 points) A scatterplot of income (x) versus the model residuals is given below.



   Based on the residual plot, which of the linear modeling assumptions is most severely violated for this data? **Select ONE**.

   - [x] X and Y have a linear relationship.
   - [ ] The residuals have mean 0.
   - [ ] The residuals have constant variance.
   - [ ] None of the above; the linear model is well fit.

5. Data is collected on the duration each winter that a lake's surface is frozen over a period of 103 consecutive years. The correlation coefficient between the variables is $r = -0.4$.

   The year variable has a mean $\bar{x} = 1950$ and standard deviation $s_x = 30$. The freeze duration variable has $\bar{y} = 90$ and standard deviation $s_y = 17$. Consider fitting a simple linear regression model to this data.

   (a) (3 points) Write a numerical expression sing only numbers and arithmetic symbols for the predicted freeze duration in the year You do not need to simplify your answer and can leave it as an expression.

   Solution:

   $$\hat{\beta}_1 = -0.4\left(\frac{17}{30}\right), \qquad \hat{\beta}_0 = 90 - \left(-0.4\left(\frac{17}{30}\right)\right) \times 1950$$

   $$\text{Predicted 1890 duration} = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1(1890)$$

   (b) (3 points) A 90% confidence interval for the freeze duration of the true regression line at $x = 1890$ has the form:

   $$\hat{y}_1 \pm c \times SE_C$$

   A 90% prediction interval for the duration that the lake was frozen in the single year 1890 has the form

   $$\hat{y}_1 \pm c \times SE_P$$

   where c is a critical value from some distribution and SE_C, SE_P are some positive values. Which R code calculates the critical value c? **Select ONE**.

◯ `qnorm(0.95)`	● `qt(0.95, df=101)`	◯ `qt(0.95, df=102)`
◯ `qnorm(0.9)`	◯ `qt(0.9, df=101)`	◯ `qt(0.9, df=102)`

   (c) (3 points) Which quantity, SE_C or SE_P, is larger? Briefly explain your answer.

   Solution: A confidence interval for the regression line is narrower than the prediction interval at the same x. Since they have the same critical value, we must have $SE_C < SE_P$. The prediction standard error also takes into account the additional variability from a new data point.