Name:

Section:

Instructions:

- Write your name and section on this cover page.
- Turn off your cell phone and put it away.
- You **may** use a calculator. However, you **may not** use a calculator on your phone or any other device that connects to the internet.
- You have **50 minutes** to complete the exam.
- You are expected to obey the Honor Code while taking this test. You **may not** discuss the exam with any other students until the exams have been returned.
- You may ask the instructor for clarification during the exam. Students who violate the Honor Code will be referred to the Honor Code Council.
- If you witness others violating the Honor Code, you have a duty to report them to the Honor Code Council.
- Students must pledge to obey the Honor Code by signing below. **Unsigned exams will not be graded.**

I understand and agree to abide by the principles of the Honor Code of Mount Holyoke College.

Signature *Date*

Multiple Choice (circle the letter corresponding to your answer)

Use this table for questions 1 through 3. This table represents the first 8 observations from a sample of 54 cars from 1993. For each car, information on vehicle type (`type`), vehicle price in thousands of USD (`price`), vehicle mileage in city in miles per gallon (`mpgCity`), vehicle drive train (`driveTrain`), vehicle passenger capacity (`passengers`), and vehicle weight in pounds (`weight`) was collected.

ID	type	price	mpgCity	driveTrain	passengers	weight
1	small	15.90	25	front	5	2705
2	midsize	33.90	18	front	5	3560
3	midsize	37.70	19	front	6	3405
4	midsize	30.00	22	rear	4	3640
5	midsize	15.70	22	front	6	2880
6	large	20.80	19	front	6	3470
7	large	23.70	16	rear	6	4105
8	midsize	26.30	19	front	5	3495

1. Which of the following best describes the `driveTrain` variable?
   (a) categorical, ordinal
   (b) numerical, discrete
   (c) categorical, nominal
   (d) numerical, continuous
2. Which type of plot would be most useful for visualizing the relationship between `type` and `price`?
   (a) histogram
   (b) dot plot
   (c) scatterplot
   (d) side by side boxplot
3. Below are summary statistics for `price`. Which of the following is **true**?

Minimum	Q_1	Median	Mean	Q_3	Maximum
7.40	10.95	17.25	19.99	26.25	61.90

   (a) There is evidence that the distribution is left-skewed.
   (b) The maximum value of 61.90 would be identified as an outlier in a box plot.
   (c) A larger number of cars cost more than \$26,250 than cost less than \$10,950.
   (d) None of the above statements is true.
4. A political scientist is interested in the effect of government type on economic development. She wants to use a sample of 30 countries evenly represented among the Americas, Europe, Asia, and Africa to conduct her analysis. What type of

study should she use to ensure that countries are selected from each region of the world?

(a) Observational study - simple random sample

(b) Observational study - cluster

(c) Observational study - stratified

(d) Experiment

5. A researcher would like to study the effect of eating breakfast on a cognitive function. Volunteers are recruited through the study by posting flyers on campus. He randomly assigns subjects to two groups, one told to eat before participating in the study and one asked to eat breakfast following the study, however, he suspects whether or not the person typically eats breakfast affects this relationship (their typical breakfast habits). In order to address this, what is the **most reasonable** thing to do prior to assigning subjects to experimental groups?

(a) Cluster on typical breakfast habits.

(b) Randomly assign subjects to breakfast habits and then conduct an experiment.

(c) Sample from each strata, typical breakfast eater and not.

(d) Block on typical breakfast habits.

6. If a distribution is right-skewed, which of the following is true?

(a) $\bar{x} < median$

(b) $\bar{x} > median$

(c) $\bar{x} = median$

(d) None of the above

7. If $P(A|B) = P(A)$, then events A and B must be:

(a) disjoint

(b) independent

(c) complements

(d) dependent

8. If $P(A) = 1 - P(B)$, then events A and B must be:

(a) disjoint

(b) independent

(c) complements

(d) dependent

9. About 80% of college students consume alcohol to some degree. Of those students, 50% engage in binge drinking. In this statement, 50% is a _____ probability.

(a) marginal

(b) conditional

(c) joint

(d) none of the above

10. When a statistic is said to be robust, this means that

(a) it is impossible for the data to have any outliers.

(b) the statistic is greatly influenced by outliers.

(c) the statistic is not an outlier.

(d) the statistic is not greatly influenced by outliers.

11. The events that you and your study partner both earn A's in this course are

(a) dependent

(b) independent

(c) disjoint

(d) complements

12. Suppose 1,000 customers are randomly selected (from the population of customers) for a survey about their experiences in a department store, but only 217 people respond. This is an example of

(a) voluntary response bias

(b) nonresponse bias

(c) convenience sampling

(d) undercoverage

13. A probability distribution is a list of the possible outcomes and their corresponding probabilities that satisfies certain rules. The outcomes listed in a probability distribution **must be**

(a) dependent

(b) disjoint

(c) independent

(d) complements

14. The General Social Survey asked, "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours. In this scenario, 1.65 hours is a(n)

(a) observation

(b) variable

(c) sample statistic

(d) population parameter

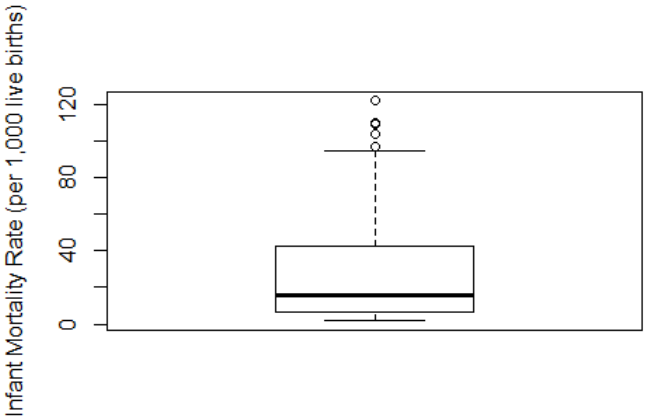
15. The boxplot and summary statistics below summarizes infant mortality rate (per 1,000 live births). Use it to answer: which of the following is **not** a true statement?

(a) The *median* $< \bar{x}$.

(b) The distribution is right-skewed.

(c) More than 50% of the observations are larger than 15.615.

(d) 25% of the observations are less than 6.505.



Minimum	Q_1	Median	Q_3	Maximum
1.800	6.505	15.615	42.140	121.630

	Aspirin (A)	Placebo (A^c)	Total
Fatal heart attack	10	26	36
Non-fatal heart attack	129	213	342
Stroke	119	98	217
None	10779	10697	21476
Total	11037	11034	22071

Table 1: Physician's Health Study, The New England Journal of Medicine, 1989

Short Answer

1. **Physician's Health Study.** In the 1980s, the Physician's Health Study examined 22,071 healthy male doctors aged 40-84 (at the start of the study). The participants were randomly assigned to either aspirin or placebo. An objective of the study was to determine whether taking aspirin reduces the risk of heart attacks. *Note, heart attack (HA) includes both fatal and non-fatal heart attacks.*

   Find the following probabilities using Table 1. Be sure to translate the words into probability notation and then find the numerical answer.

   (a) What is the probability that an individual who is taking aspirin (A) has a heart attack (HA)?

   (b) What is the probability that an individual who is taking placebo (A^c) has a heart attack (HA)?

   (c) What is the probability that an individual has a heart attack (HA)?

   (d) Using your answers in (a)-(b), do you think that aspirin reduces the risk of heart attacks?

   (e) Is your conclusion in (d) generalizable? If so, to what population? If not, why? (**I will not read more than 2 sentences.**)

2. **Hospital Safety.** We have two hospitals, A and B, and we want to know which hospital is safer for surgery. To assess safety, we collect information on the number of patients that have survived and died after surgery in each hospital (Table 2).

	Hospital A	Hospital B	Total
Died	63	16	79
Survived	2037	784	2821
Total	2100	800	2900

Table 2: Patient deaths and survivals after surgery by hospital

Notation:

- A is the event a patient is in Hospital A
- B is the event a patient is in Hospital B
- D is the event a patient dies after surgery
- D^c is the event a patient survives after surgery

For (a)-(d), show your work and use appropriate notation where needed. (3 pts each)

(a) $P(A)$

(b) $P(B)$

(c) $P(D|A)$

(d) $P(D|B)$

(e) If we look at $P(D|A)/P(D|B)$, we should find that it is greater than 1, indicating that patients at Hospital A are more likely to die after surgery than those at Hospital B. This suggests that Hospital B is safer, but Hospital A disagrees. **Identify one potential confounding variable that might account for why patients at Hospital A are more likely to die.** (1 pt)

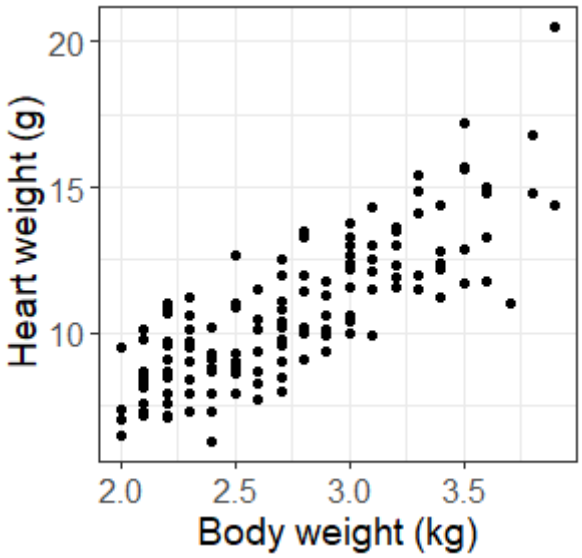
3. The following scatterplot and regression output is for predicting the heart weight (`Hwt`) in grams of cats from their body weight (`Bwt`) in kilograms. The coefficients are estimated using a dataset of 144 domestic cats.

```
Call:
lm(formula = Hwt ~ Bwt, data = cats)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5694 -0.9634 -0.0921  1.0426  5.1238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3567     0.6923  -0.515    0.607
Bwt           4.0341     0.2503  16.119   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.452 on 142 degrees of freedom
Multiple R-squared:  0.6466,Adjusted R-squared:  0.6441
F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```



(a) Identify the following:

i. Response variable:

ii. Predictor:

(b) What is the correlation between heart weight and body weight in domestic cats?

(c) Write out the linear model. Be sure to define any notation you use for variables.

(d) Interpret the following:

i. Intercept:

ii. Slope:

iii. R^2:

iv. Is it appropriate to interpret all of the above quantities? If there are any for which interpretation is inappropriate, explain why it is inappropriate.

(e) Suppose you have a cat that weighs 3.0 kg. What is the predicted weight of your cats heart?

(f) Suppose your cat's true heart weight is 11 g. What is the associated residual? Does the model overestimate or underestimate your cat's heart weight?

STAT 140 Midterm I Formula Sheet

- Sample statistics and outliers

  $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

  $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

  $IQR = Q_3 - Q_1$

  $Q_1 - 1.5 \times IQR$

  $Q_3 + 1.5 \times IQR$

- Probability rules

  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

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

  $P(A) = 1 - P(A^c)$

  $P(B|A) = \frac{P(A|B)P(B)}{P(A \text{ and } B)+P(A \text{ and } B^c)}$

- Linear regression

  $\hat{\beta}_1 = \frac{s_y}{s_x}R$

  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

  $R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$

  $\hat{\epsilon}_i = y_i - \hat{y}_i$