1. To compare success rates for treating allergies at two clinics that specialize in treating allergy sufferers, researchers selected random samples of patient records from the two clinics. The following table summarizes the data.

	Clinic A	Clinic B	**Total**
Unsuccessful treatment	51	33	84
Successful treatment	88	35	123
Total	139	68	207

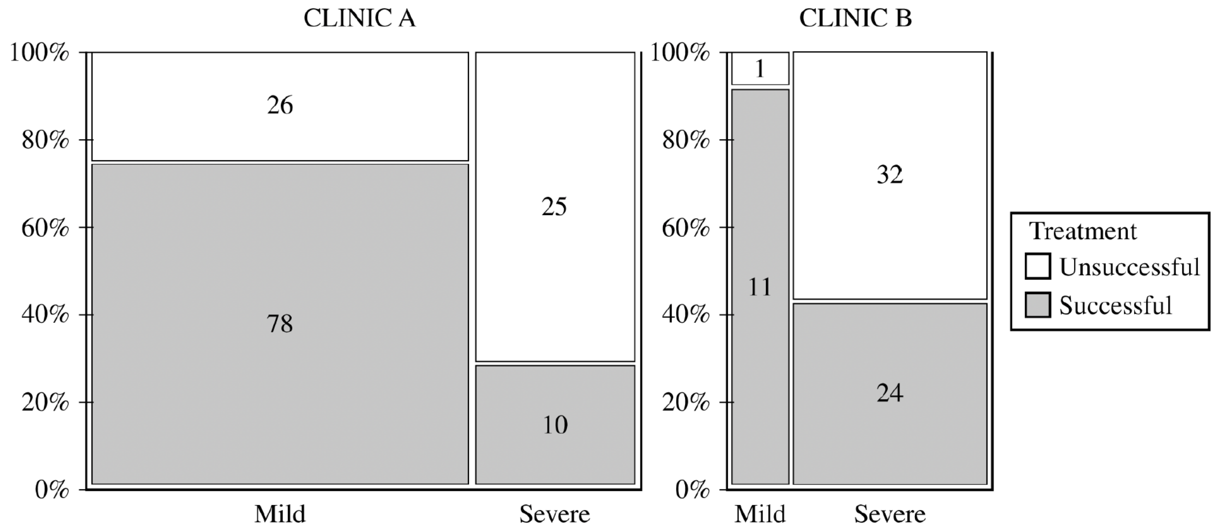
(a) (i) Complete the following table by recording the relative frequencies of successful and unsuccessful treatments at each clinic.

	Clinic A	Clinic B
Unsuccessful treatment		
Successful treatment		

(ii) Based on the relative frequency table in part (a-i), which clinic is more successful in treating allergy sufferers? Justify your answer.

(b) Based on the design of the study, would a statistically significant result allow the researchers to conclude that receiving treatments at the clinic you selected in part (a-ii) causes a higher percentage of successful treatments than at the other clinic? Explain your answer.

A physician who worked at both clinics believed that it was important to separate the patients in the study by severity of the patient's allergy (severe or mild). The physician constructed the following mosaic plot. The values in the mosaic plot represent the number of patients who were either successfully treated or unsuccessfully treated in each allergy severity group within each clinic. For example, the value 78 represents the number of patients successfully treated in the mild group within Clinic A.



Based on the mosaic plot, the physician concluded the following:

For mild allergy sufferers, **Clinic B** was more successful in treating allergies.

For severe allergy sufferers, **Clinic B** was more successful in treating allergies.

(c) (i) For each clinic, which allergy severity is treated more successfully? Justify your answer.

- Clinic A:
- Clinic B:

(ii) For each clinic, which allergy severity is more likely to be treated? Justify your answer.

- Clinic A:
- Clinic B:

(d) Using your answers from part (c), give a reasonable explanation of why the more successful clinic identified in part (a-ii) is the same as or different from the physician's conclusion that Clinic B is more successful in treating both severe and mild allergies.

(e) In a further study, a researcher decides to take a simple random sample of the patients seen by Clinic A. Using a list of the all the patients, describe a method for selecting a sample of size 200.

(f) Using the same list, describe how to conduct a systematic random sample.

(g) The researcher notes that the patients at Clinic B come from two very different neighborhoods—one that has a higher socio-economic level and one that is lower. Describe and name the sampling method the researchers would use to take this source of variation into account.

(h) The researcher has received a grant to study the health insurance status for potential patients in the city. They draw a grid on the city, where each square of the grid contains approximately 50 homes. They randomly select five squares and use those homes to collect their sample. What is the name of this sampling method? Describe any advantages and disadvantages of this method.

(i) Using the original data (see part a), what is the probability that a randomly selected patient…

(i) had a successful treatment?
(ii) had a successful treatment given that they were seen at Clinic A?
(iii) had a successful treatment or were seen at Clinic A?
(iv) had a successful treatment and were seen at Clinic A?

(j) Are the events "a successful treatment" and "seen at Clinic A" independent? Use the two probabilities from part (i) that correctly answer this question.

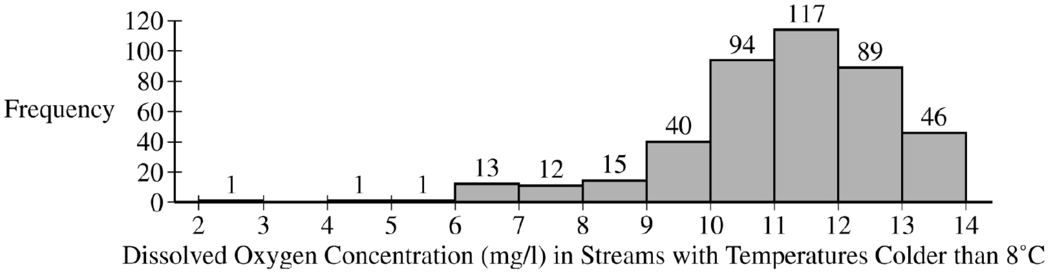
(k) In yet another survey, the researcher interviews the first 50 patients in a given day. What is the name of this method and why might it lead to a biased result? Include the concept of undercoverage in your discussion.

(l) Last survey! The researcher mails a survey to all the patients in the clinic database, asking them to fill out an online form. Describe this method and its probable bias.

(m) A physician notes that the patients at Clinic B tend to have a higher socio-economic status than those at Clinic A. Explain how this might create a confounding variable.

(n) Patients who fail to follow their doctor's instructions may be embarrassed to admit that the treatment was unsuccessful. Describe how this might both create a nonresponse and a response bias.

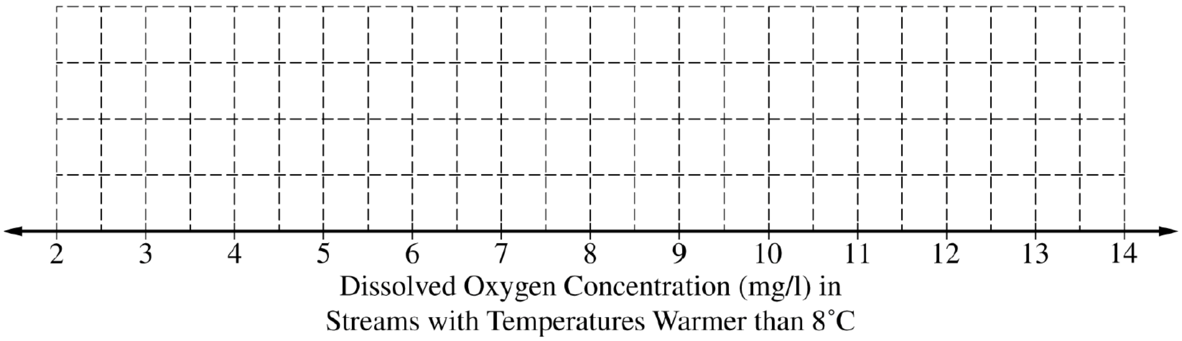
2. As part of a study on the chemistry of Alaskan streams, researchers took water samples from many streams with temperatures colder than 8°C and from many streams with temperatures warmer than 8°C. For each sample, the researchers measured the dissolved oxygen concentration, in milligrams per liter (mg/l).



(a) The researchers constructed the histogram shown for the dissolved oxygen concentration in streams from the sample with water temperatures colder than 8°C. Based on the histogram, describe the distribution of dissolved oxygen concentration in streams with water temperatures colder than 8°C.

Min	Q1	Median	Q3	Max	Mean	Std. Dev.
2.10	4.39	5.43	6.12	13.45	5.54	1.64

(b) The researchers computed the summary statistics shown in the table for the dissolved oxygen concentration in streams from the sample with water temperatures warmer than 8°C. Use the summary statistics to construct a box plot for the dissolved oxygen concentration in streams with water temperatures warmer than 8°C. Do not indicate outliers.


2
3
4
5
6
7
8
9
10
11
12
13
14
Dissolved Oxygen Concentration (mg/l) in
Streams with Temperatures Warmer than 8˚C

(c) The researchers believe that streams with higher dissolved oxygen concentration are generally healthier for wildlife. Which streams are generally healthier for wildlife, those with water temperature colder than 8°C or those with water temperature warmer than 8°C? Using characteristics of the distribution of dissolved oxygen concentration for temperatures colder than 8°C and characteristics of the distribution of dissolved oxygen concentration for temperatures warmer than 8°C, justify your answer.

(d) Find both outlier fences for part (b) and justify whether or not there are outliers for the sample of with water temperatures warmer than 8° C. Also use the two standard deviation rule to check for outliers. Compare these two results and discuss any difference between them.

(e) If the data from part (a), colder than 8° C, were made into a stemplot, how would you round the data? Write a key that the stemplot would use.

(f) If the data from part (a), colder than 8° C, were made into a cumulative frequency plot (using the same interval width on the histogram), which interval on the graph would have the steepest slope? Would the graph be horizontal for an interval? If so, where?

(g) By examining the histogram on part (a), would expect that mean of the data colder than 8° C by greater than, smaller than, or about the same as the median? Explain.

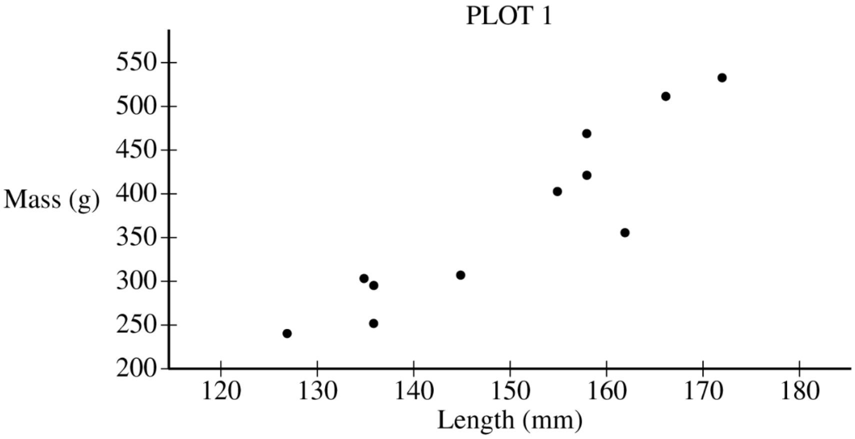
(h) What is the approximate percentile of a stream colder than 8° C if it has a reading of 9mg/l? (hint: there are 429 data points)

(i) For streams warmer than 8° C, the standard deviation is 1.64 mg/l. Interpret this value in context.

(j) Would you describe these data to be continuous or discrete? Explain.

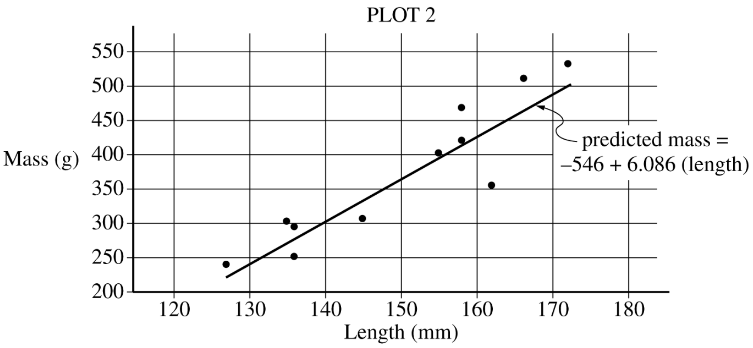
(k) Convert the summary statistics in part (b) to °F, using the formula F = 1.8C + 32. Also find the range and IQR in °F.

3. A biologist gathered data on the length, in millimeters (mm), and the mass, in grams (g), for 11 bullfrogs. The data are shown in Plot 1.



(a) Based on the scatterplot, describe the relationship between mass and length, in context.

From the data, the biologist calculated the least-squares regression line for predicting mass from length. The least-squares regression line is shown in Plot 2.



(b) Identify and interpret the slope of the least-squares regression line in context.

(c) Interpret the coefficient of determination of the least-squares regression line, $r^2 \approx 0.819$, in context.

(d) From Plot 2, consider the residuals of the 11 bullfrogs.

(i) Based on the plot, approximately what is the length and mass of the bullfrog with the largest absolute value residual?

(ii) Does the least-squares regression line overestimate or underestimate the mass of the bullfrog identified in part (d-i)? Explain your answer.

(e) Find and interpret the correlation coefficient in context.

(f) If interpreted in context, the y-intercept is an extrapolation. Explain.

(g) A bullfrog that is 140 mm long is found that weighs 320 grams. Find and interpret this bullfrog's residual.

(h) By just using rough estimates from Plot 2, sketch a residual plot. Does it look as we hoped it would? Explain.

(i) Would a bullfrog that is much longer than the rest of the bullfrogs have influence on the slope of the regression equation? Explain, including the correct name of such a point.

(j) The mean length for the 11 bullfrogs is approximately 150 mm. What is the predicted mass for this bullfrog? What is special about this point?

(k) Five of the bullfrogs are male and six are female. If the researcher randomly selects two of them, what is the probability that they are both female?

4. A dermatologist will conduct an experiment to investigate the effectiveness of a new drug to treat acne. The dermatologist has recruited 36 pairs of identical twins. Each person in the experiment has acne and each person in the experiment will receive either the new drug or a placebo. After each person in the experiment uses either the new drug or the placebo for 2 weeks, the dermatologist will evaluate the improvement in acne severity for each person on a scale from 0 (no improvement) to 100 (complete cure).

(a) Identify the treatments, experimental units, and response variable of the experiment.

- Treatments:
- Experimental units:
- Response variable:

Each twin in the experiment has a severity of acne similar to that of the other twin. However, the severity of acne differs from one twin pair to another.

(b) For the dermatologist's experiment, describe a statistical advantage of using a matched-pairs design where twins are paired rather than using a completely randomized design.

(c) For the dermatologist's experiment, describe how the treatments can be randomly assigned to people using a matched-pairs design in which twins are paired.

(d) Suppose instead of 36 twins, the researcher had 100 volunteers. Sixty of the volunteers were teenagers and the rest were middle-aged. Describe why a block design would be appropriate and how this design would be implemented.

(e) Could this experiment be carried out in a double-blind manner? Explain.

(f) Does the original experiment have replication? Explain.

(g) What is the factor? How many levels are there?

(h) When the experiment is complete, the researcher states that the medication was statistically significantly effective.

(i) What does "statistically significant" mean?
(ii) Can the researcher conclude the medication caused the improvement?
(iii) Who can this conclusion be applied to?

(i) Describe why a placebo needs to be used as a form of control.

5. A machine at a manufacturing company is programmed to fill shampoo bottles such that the amount of shampoo in each bottle is normally distributed with mean 0.60 liter and standard deviation 0.04 liter. Let the random variable A represent the amount of shampoo, in liters, that is inserted into a bottle by the filling machine.

(a) A bottle is considered to be underfilled if it has less than 0.50 liter of shampoo. Determine the probability that a randomly selected bottle of shampoo will be underfilled. Show your work.

After the bottles are filled, they are placed in boxes of 10 bottles per box. After the bottles are placed in the boxes, several boxes are placed in a crate for shipping to a beauty supply warehouse. The manufacturing company's contract with the beauty supply warehouse states that one box will be randomly selected from a crate. If 2 or more bottles in the selected box are underfilled, the entire crate will be rejected and sent back to the manufacturing company.

(b) The beauty supply warehouse manager is interested in the probability that a crate shipped to the warehouse will be rejected. Assume that the amounts of shampoo in the bottles are independent of each other.

(i) Define the random variable of interest for the warehouse manager and state how the random variable is distributed.

(ii) Determine the probability that a crate will be rejected by the warehouse manager. Show your work.

To reduce the number of crates rejected by the beauty supply warehouse manager, the manufacturing company is considering adjusting the programming of the filling machine so that the amount of shampoo in each bottle is normally distributed with mean 0.56 liter and standard deviation 0.03 liter.

(c) Would you recommend that the manufacturing company use the original programming of the filling machine or the adjusted programming of the filling machine? Provide a statistical justification for your choice.

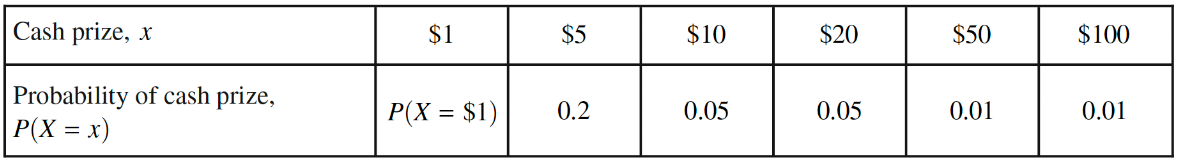
(d) Using the information from part (a) ($\mu = 0.60$, $\sigma = 0.04$), describe the distribution using the empirical rule.

(e) Using the information from part (a) ($\mu = 0.60$, $\sigma = 0.04$), how many liters of shampoo are in the 6% most overfilled bottles?

(f) Using the information from part (a) ($\mu = 0.60$, $\sigma = 0.04$), find the z-score for a bottle with 0.65 liters and interpret the value in context.

(g) Suppose the manager decided to keep the mean at 0.60 liters but is wants to no more than 2% of the bottles to be underfilled. What standard deviation satisfy this requirement?

6. Bath fizzies are mineral tablets that dissolve and create bubbles when added to bathwater. In order to increase sales, the Fizzy Bath Company has produced a new line of bath fizzies that have a cash prize in every bath fizzy. Let the random variable, X, represent the dollar value of the cash prize in a bath fizzy. The probability distribution of X is shown in the table.



Cash prize, x	\$1	\$5	\$10	\$20	\$50	\$100
Probability of cash prize, $P(X = x)$	$P(X = \$1)$	0.2	0.05	0.05	0.01	0.01

(a) Based on the probability distribution of X, answer the following. Show your work.
(i) Calculate the proportion of bath fizzies that contain $1.
(ii) Calculate the proportion of bath fizzies that contain at least $10.

(b) Based on the probability distribution of X, calculate the probability that a randomly selected bath fizzy contains $100, given that it contains at least $10. Show your work.

(c) Based on the probability distribution of X, calculate and interpret the expected value of the distribution of the cash prize in the bath fizzies. Show your work.

(d) The Fizzy Bath Company would like to sell the bath fizzies in France, where the currency is euros. Suppose the conversion rate for dollars to euros is 1 dollar = 0.89 euros. Using your expected value from part (c), calculate the expected value, in euros, of the distribution of the cash prize in the bath fizzies. Show your work.

(e) Would you describe these data as continuous or discrete? Explain. Are the events in this scenario mutually exclusive? Explain.

(f) Suppose you buy a number of fizzies and that each purchase is independent and follows the distribution of *X*.
(i) What is probability that you buy three fizzies they all have a $1 prize?
(ii) What is the probability that you don't receive a prize of more than $1 until your 7th purchase?
(iii) What is the probability that if you buy five fizzies, at least one contains a prize of more than $1?

(g) Now suppose you buy 10 fizzies and count the number of times of you win a prize more than $1. What are the mean and the standard deviation of this variable?

(h) If you buy fizzies until you win a prize more than $1, what are the mean and standard deviation of this distribution?

(i) A friend of yours buys the bath fizzies five times and only wins the $1 prize every time. They insist the game is rigged. Explain why your friend is incorrect according to the law of large numbers.

(j) Find and interpret the standard deviation of *X* (from part a).

(k) Find the mean and standard deviation of the sum of your prizes if you bought three fizzies.

(l) Suppose the company runs a Christmas special with a more generous payout. Call the variable *Y*, where the mean is $7.50 and the standard deviation is $23.50.

(i) If you buy one fizzie from each distribution, what is the mean and standard deviation of your total prize? What assumption is necessary for these answers to be true?

(ii) If you buy one of each, what is the mean and standard deviation of the difference of the payout?