Solutions to the Super Six 2.0

Scoring guidelines for the original questions can be found on AP Central. The solutions for the original questions have been copied directly from the Model Solutions provided by CollegeBoard. The remaining solutions (and questions) have been written by me. The solutions are not always written as scoring guidelines might require. The purpose of these materials is to provide a cumulative review using a relatively small number of questions. For helping students align to the scoring guidelines, use lots of FRAPPY's!

The source of the questions is as follows:

1. 2022 #6
2. 2023 #1
3. 2022 #1
4. 2022 #2
5. 2022 #3
6. 2023 #3

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	$\frac{51}{139} \approx 0.3669$	$\frac{33}{68} \approx 0.4853$
Successful Treatment	$\frac{88}{139} \approx 0.6331$	$\frac{35}{68} \approx 0.5147$

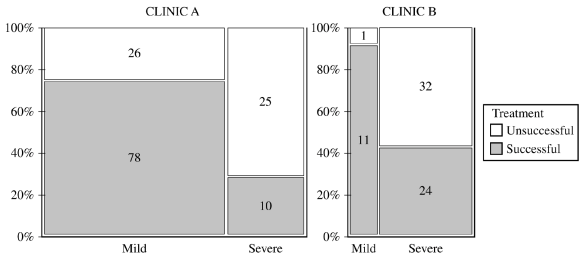
(ii) Based on the relative frequency table in part (a-i), which clinic is more successful in treating allergy sufferers? Justify your answer.

Clinic A appears to be more successful in treating allergy sufferers than Clinic B. Clinic A was successful for 63.3% of the allergy sufferers it treated, while Clinic B was successful for only 51.5% of the allergy sufferers it treated.

(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.

No. This is an observational study; it is not a randomized experiment. Cause and effect can only be established with a well-designed, randomized experiment. There may be other variables, besides where a patient was treated, that affect the success rates for treating allergies. For example, Clinic A may mostly treat mild allergy cases that are easy to treat successfully, while Clinic B may mostly treat severe allergy cases that are more difficult to treat successfully.

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: *More successful in treating mild 1. allergies. Clinic A successfully treated 75.0% of mild allergy sufferers, while Clinic A successfully treated only 28.6% of severe allergy sufferers.*

- Clinic B: *More successful in treating mild allergies. Clinic B successfully treated 91.7% of mild allergy sufferers, while Clinic B successfully treated only 42.9% of severe allergy sufferers.*

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

- Clinic A: *More likely to treat mild allergy sufferers than severe allergy sufferers. Of the 139 allergy sufferers treated at Clinic A, 104 (74.8%) suffered from mild allergies, while only 35 (25.2%) suffered from severe allergies.*

- Clinic B: *More likely to treat severe allergy sufferers than mild allergy sufferers. Of the 68 allergy sufferers treated at Clinic B, 56 (82.4%) suffered from severe allergies, while only 12 (17.6%) suffered from mild allergies.*

(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.

The more successful clinic identified in part (a-ii) is Clinic A, which is different from the physician's conclusion that Clinic B is better when taking allergy severity into account. This happens because for both clinics the success rate is much higher for mild allergy sufferers (75.0 percent versus 28.6 percent for Clinic A, and 91.7 percent versus 42.9 percent for Clinic B). Clinic A treats mostly mild allergy sufferers OR (74.8 percent of its patients) while Clinic B treats mostly severe allergy sufferers (82.4 percent of its patients). Therefore, when combining the results across allergy severity categories to obtain the table from part (a-i), the facts that Clinic A treats a larger proportion of mild allergy sufferers and mild allergy sufferers have a higher success rate make it appear as if Clinic A is better overall.

(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.

Number the patients from 1 to n. Use a random number generator to select 200 unique random numbers from 1 to n. Contact the patients with those numbers for your sample.

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

Using a random number generator, select a random number from one to ten. Start on the list with that patient (e.g., patient #7). Then select every 20th patient after that for your 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.

To account for the variation created by these differences, a stratified random sample should be selected. That is, the researcher should randomly select some patients from both socio-economic levels to comprise the sample.

(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.

This a cluster sample. If the clusters of homes adequately represent the entire city, this method would make it easier. However, there are probably concerns that the five clusters chosen would leave out certain subgroups of the city and not be representative.

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

(i) had a successful treatment? *123/207 = 59.4%*
(ii) had a successful treatment given that they were seen at Clinic A? *88/139 = 63.3%*
(iii) had a successful treatment or were seen at Clinic A? *(123 + 139 – 88)/207 = 84.1%*
(iv) had a successful treatment and were seen at Clinic A? *88/207 = 42.5%*

(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.

P(successful) = 59.4%; P(successful | Clinic A) = 63.3%; These percentages are fairly close, so we can say that clinic and success are independent. Note: If you said that they are about 4% different so they are not independent, I would mark that as correct! Remember we need chi-square to truly measure independence.

(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.

This is a convenience sample. It will undercover patients who come later in day. It is hard to know how those patients might be different, but suppose that the people who work all day outside have to come

later in the day after their work. They might have worse allergies and be more difficult to treat. This would lead in an over-estimate of the success rate, making it look better than it actually is.

(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.

This is a voluntary response survey. Only people with very strong emotions will fill out the survey, probably people with unsuccessful treatment. This will lead to a (probably very severe) under-estimate of the success rate of the treatment.

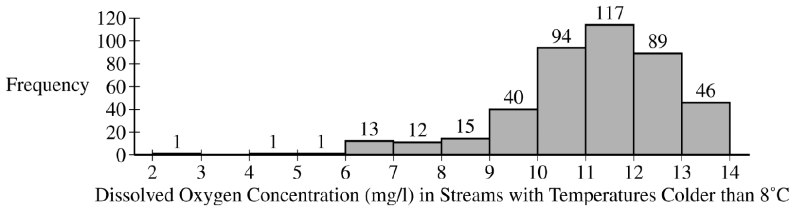
(m) A physician notes that the patients at Clinic B tend to have a higher socio-economic status than those at Clinic A. Explain how this might create a confounding variable.

Having a higher socio-economic status might lead to patients having access to other, extra treatments beyond just those provided by the clinic. Those extra treatments might lead to a higher success rate. In the end we would be uncertain if the higher success rate was caused by the choice of clinic or by the extra treatments that some patients could afford.

(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.

The embarrassment could lead to a disproportionate number of unsuccessful patients declining to answer the survey (nonresponse bias). It also could lead to patients whose treatment was unsuccessful not telling the truth and stating that the treatment was successful (response bias) because they don't want to admit that they did not follow the instructions. In both cases, the parameter is under-estimated.

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).

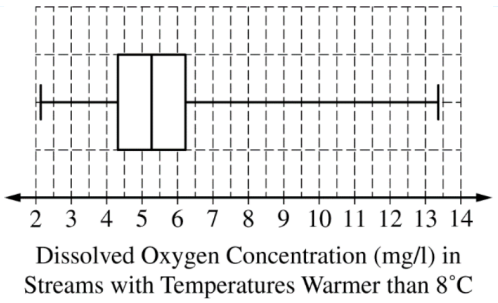


(a) The researchers constructed the histogram shown for the dissolved oxygen concentration in streams from the sample with water temperatures colder than 8°C. Based on the histogram, describe the distribution of dissolved oxygen concentration in streams with water temperatures colder than 8°C.

The histogram of dissolved oxygen concentration in Alaskan streams with water temperature colder than 8°C is unimodal and skewed left with a median between 11 and 12 mg/l . The first quartile is in the bin from 10-11 mg/l and the third quartile is in the bin from 12-13 mg/l , so the IQR is approximately 2 mg/l . There do not appear to be any high outliers, but there are several potential low outliers because the values in the 2-3, 4-5, and 5-6 bins are all certainly more than 1.5 IQR below the first quartile.

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.



(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.

If the researchers' belief is correct, then streams with water temperature colder than 8°C are healthier for wildlife.

The distribution of dissolved oxygen 2. concentration for colder streams has a higher center because its median (between 11 mg/l and 3. 12 mg/l) is larger than the median for warmer streams (5.43 mg/l).

The shape of the distribution of dissolved oxygen concentration for colder streams is different from the shape of the distribution for warmer streams. The distribution of values of dissolved oxygen concentration for colder streams is skewed to the left but the distribution of values for warmer streams is skewed to the right.

Both distributions have a similar spread because they both have similar IQR values — approximately 2 mg/l for the colder streams and 1.73 mg/l for the warmer streams.

(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.

*IQR = 1.73; Q3 + 1.5*IQR = 8.715; Q1 – 1.5*IQR = 1.795; The minimum of 2.1 is above the lower fence, so there are no low outliers. The max of 13.45 is above the upper fence, so there is at least one high outlier.*

*5.54 ± 2*1.64 = 2.26, 8.82; Because the mean and standard deviation are not resistant and these data are clearly right skewed, this is not the best way to test for outliers. And indeed, this rule states that the minimum would be an outlier, which does not seem appropriate.*

(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.

We would round to the nearest tenth, a key might look like: 9|4 = 9.4 mg/l.

(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?

Because the interval of 11 to 12 has the highest frequency, this line on the cumulative frequency plot would have the steepest slope. Because there are no observations between 3 and 4, this interval would have a horizontal line.

(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.

Because the data is left skewed, the mean would be lower than the median dissolved oxygen concentrate.

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

43 observations are below 9 mg/l. 43/429 ≈ 10th percentile.

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

The dissolved oxygen concentration for streams warmer than 8° C is typically 1.64 mg/l from the mean.

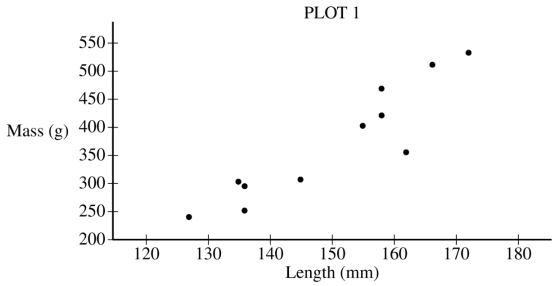
(j) Would you describe these data to be continuous or discrete? Explain.

It is continuous because the data can be collected with any take on any value. It doesn't, for example need to be a whole number or a number to the nearest tenth or hundredth.

(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.

- *min = 35.78° F*
- *Q1 ≈ 39.9° F*
- *Med ≈ 41.77° F*
- *Q3 ≈ 43.02° F*
- *Max ≈ 56.21 ° F*
- *Mean ≈ 41.97° F*
- *SD ≈ 2.95 ° F*
- *IQR ≈ 3.11° F*
- *Range ≈ 20.43° F*

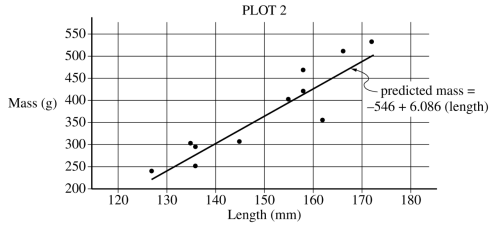
3. A biologist gathered data on the length, in millimeters (mm), and the mass, in grams (g), for 11 bullfrogs. The data are shown in Plot 1.



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

The scatterplot reveals a strong, positive, roughly linear association between the mass and length of bullfrogs. There are no points that seriously deviate from the straight-line pattern of the points in the plot.

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.

The value of the slope of the least-squares regression line is 6.086. This value indicates that the predicted mass of a bullfrog increases by 6.086 grams for each additional millimeter of length.

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

The coefficient of determination is $r^2 \approx 0.819$. This value indicates that 81.9% of the variation in bullfrog mass can be explained by variation in bullfrog length as described by the least-squares line.

(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?

The largest residual in absolute value belongs to the bullfrog with length 162 millimeters and mass 356 grams.

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

The least-squares regression line overestimates the mass of the bullfrog with length 162 millimeters. Plot 2 shows that the point for the bullfrog with length 162 millimeters is below the least-squares regression line.

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

$\sqrt{(0.819)} = 0.905 \rightarrow$ There is a strong, positive, linear relationship between bullfrog length and weight.

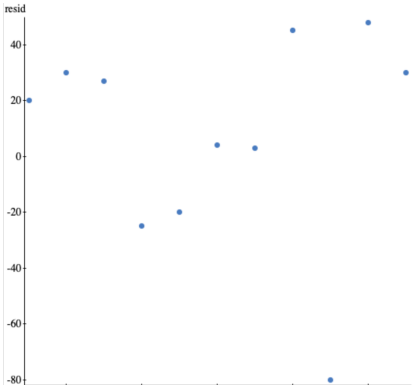
(f) If interpreted in context, the y-intercept is an extrapolation. Explain.

The y-intercept is -546 grams. It makes no sense for a bullfrog to have negative weight. The collected data is between 120 and 180 mm. So making a prediction for 0 mm is an extrapolation that produced a nonsensical result.

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

*predicted mass = -546 + 6.086*140 = 306.04 g. Residual ≈ 320 – 306 = 14 g. This bullfrog weighs 14 grams more than predicted for its length.*

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



The graph looks random, with no pattern, as we hoped.

(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.

Yes, high leverage points are extreme in x-direction and often exert a large influence on the regression, especially the slope.

(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?

The predicted length ≈ 367 mm and this point is the mean length. The (mean, mean) point is always on the regression line.

(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?

$\frac{6}{9} * \frac{5}{8} = 41.7\%$

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: *New drug, placebo*
- Experimental units: *The 72 people who receive the new drug or placebo.*
- Response variable: *Improvement in acne severity*

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.

Improvement scores will vary due to many factors, including initial acne severity, what treatment is received, and other variables such as diet and genetics. Because the pairs of twins are similar in initial acne severity, pairing allows for the variation in improvement scores due to the treatment received to be distinguished from variation due to initial acne severity, unlike in a completely randomized design. Consequently, using the matched-pairs design will provide a more precise estimate of the mean difference in improvement in acne severity for the new drug compared to the placebo and make it easier to find convincing evidence that the new drug is better, if it really is better.

(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.

For each pair of twins, label one person as twin A and label the other person as twin B. For each pair of twins, toss a coin. If the coin lands on heads, twin A gets the placebo and twin B gets the active drug. If the coin lands on tails, twin A gets the active drug and twin B gets the placebo.

OR

Label the members of each pair of twins as"Twin 1" and "Twin 2." Using a random number generator, generate an integer from 1 to 2. Give the drug to the twin whose number is selected and the placebo to the twin whose number is not selected. Repeat for all pairs of twins.

OR

Label 1 notecard "A" and another notecard "B." For each pair of twins, shuffle the cards and give one card to each twin. The twin who gets "A" receives the drug and the twin who gets "B" receives the placebo.

(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.

Because teenagers probably have different habits and hormones that effects their acne differently than middle-aged people, this would create variability in the improvement of acne. To reduce this variation, blocking by age would evenly distribute the age groups to the treatments and reduce this source of variation.

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

Yes, using a placebo that looks like the medication and is labeled by an assistant, neither the subjects nor the evaluators would need to know which subject is receiving which treatment.

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

Yes, 36 people are receiving each treatment.

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

The drug is the factor, two levels, the real medication and the placebo.

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

(i) What does "statistically significant" mean?

That the difference in the improvement in acne severity is a bigger difference than would happen by chance variation.

(ii) Can the researcher conclude the medication caused the improvement?

Yes, this is an experiment so a causal conclusion is appropriate.

(iii) Who can this conclusion be applied to?

The results can be applied to people similar to our volunteers.

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

The placebo serves as a baseline for comparison. The volunteers know they are going to have their acne severity evaluated. So, for one example, they may be more conscientious about their washing habits, which might improve their acne. The placebo group will enable the researcher to see if the drug itself makes a difference.

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.

Random variable A, which represents the amount of shampoo in a randomly selected bottle, follows a normal distribution with mean 1. 0.6 liter and standard deviation 0.04 liter. Then, the probability that a randomly selected bottle is underfilled is

$P(x < 0.5)$, *using normal cdf with* $\mu = 0.60$ & $\sigma = 0.04$; $\rightarrow 0.0062$

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.

The random variable of interest, X, is the number of underfilled bottles in a box of 10 bottles. The distribution of X is binomial with parameters n = 10 and p = 0.0062.

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

The crate will be rejected by the warehouse if two or more underfilled bottles are found in the box. The probability of that is

$P(X \geq 2) = 1 - P(X \leq 1)$; binomial distribution; n = 10; p = 0.0062; $\rightarrow$ *0.0017. Note: you can show the binomial formula, but you don't need to if you describe the distribution this thoroughly.*

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.

For the adjusted programming of the filling machine,
the probability of an underfilled bottle is

$P(x < 0.5)$, *using normal cdf with* $\mu = 0.56$ *&* $\sigma = 0.03$; $\rightarrow 0.0228$

Because the probability of an underfilled bottle is greater for the adjusted programming, this would result in more rejected shipments. The company should continue with the original machine programming.

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

68% of the bottles will contain between 0.56 liters and 0.64 liters.
95% of the bottles will contain between 0.52 liters and 0.68 liters.
99.7% of the bottles will contain between 0.48 liters and 0.72 liters.

(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?

Using the given mean, standard deviation, and inverse normal (draw it!) $\rightarrow$ *0.662 liters.*

(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.

z = (0.65 – 0.60)/0.04 = 1.25; a bottle with 0.65 liters is 1.25 standard deviations above the mean.

(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?

Using inverse normal, $\mu = 0$, $\sigma = 1$, *we find z-score = -2.054. Then* $-2.054 = (0.5 - 0.6)/\sigma \rightarrow \sigma = 0.049$

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.

The proportion of bath fizzies containing \$1 is equal to the P(X = \$1) = 1 – (0.2 + 0.05 + 0.05 + 0.01 + 0.01) = 0.68.

(ii) Calculate the proportion of bath fizzies that contain at least \$10.

The proportion of bath fizzies that contain at least \$10 is equal to the P(X ≥ \$10) = 0.05 + 0.05 + 0.01 + 0.01 = 0.12.

(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.

Given a bath fizzy contains at least \$10, then the probability that it contains \$100 = 0.01/0.12 = 8.3%.

(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.

The expected value of the distribution of X is E(X) = 1(0.68) +5(0.2) +10(0.05) + 20(0.05) + 50(0.01) + 100(0.01) = \$4.68.

The expected value is the mean of the cash prizes that result from the long run of many, many trials of randomly selecting bath fizzies and determining the amount each contains.

(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.

The expected value of the distribution of X in euros is 4.68(0.89) ≈ 4.17 euros.

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

The events are mutually exclusive and discrete. The variable can only take on one of six different values and it cannot take on more than value at a time.

(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?

$(0.68)^3 = 31.4\%$

(ii) What is the probability that you don't receive a prize of more than $1 until your 7th purchase?

$(0.68)^6 * 0.32 = 3.2\%$

(iii) What is the probability that if you buy five fizzies, at least one contains a prize of more than $1?

$1 - (0.68)^5 = 85.5\%$

(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?

*Binomial! n = 10; p =0.32; μ = 10*0.32 = 3.2 times; σ = √(10*0.32*0.68) = 1.48 times.*

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

Geometric! p = 32%; μ = 1/0.32 = 3.125 times; σ = √(0.68)/0.32 = 2.58 times.

(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.

According to the law of large numbers, the observed probabilities will match the theoretical probabilities in the long run. So, in the short run, it is not surprising if in the short run you don't observe less likely events.

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

$\sigma = \sqrt{(1 - 4.68)^2 * 0.68 + (5 - 4.68)^2 * 0.20 + \cdots} = 11.57$;
The typical distance of a prize amount from the mean is $11.57.

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

$\mu = 4.68 + 4.68 + 4.68 = \$14.04$;
$\sigma = \sqrt{11.57^2 + 11.57^2 + 11.57^2} = \$20.04$;

(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?

$\mu = 4.69 + 7.50 = \$12.19$;
$\sigma = \sqrt{11.57^2 + 23.50^2} = \$26.19$;
In order for the standard deviation to be correct, the events need to be independent.

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

$\mu = 4.69 - 7.50 = -\$2.81$; that is, the Christmas special pays out an average of \$2.81 more than the regular prizes.
$\sigma = \sqrt{11.57^2 + 23.50^2} = \$26.19$; the standard deviation of a difference is the same as the sum!