repeated? How do our data compare to typical results of an experiment like ours?" In order to understand whether the results of a study are consistent with a scientific model, we need to be able to think about the effect that inherent variability has on the data that we collect.

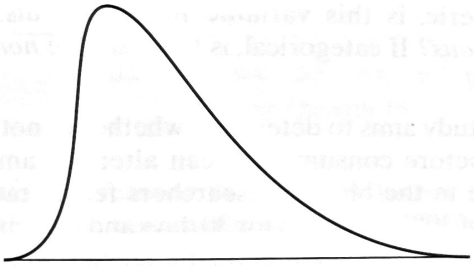
Typically, we settle on a particular statistic, such as the sample mean, as a useful representation of the variable we are studying. The sampling distribution of our statistic tells us what to expect under a given model. By knowing what kinds of results can be expected, we can make an inference about the model. In particular, we can determine if our data suggest that our model is wrong. This will become clearer when we put these ideas into practice in the chapters ahead. For now, our goal is to have a good working understanding of variability, of distributions in general, and of sampling distributions of statistics in particular.

## Unit I Summary Exercises

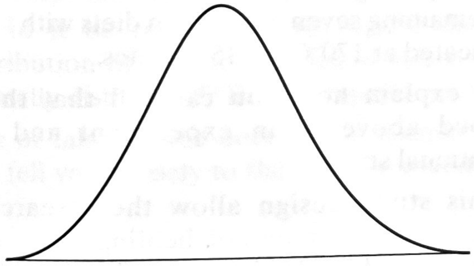
**I.1** Precipitation, measured in inches, for the month of March in Minneapolis, Minnesota, was recorded for 25 consecutive years. The values ranged from 0.3 up to 4.7, with a mean of 1.7 and an SD of 1.1.

(a) Which of the following is a rough histogram for the data? Explain your choice.

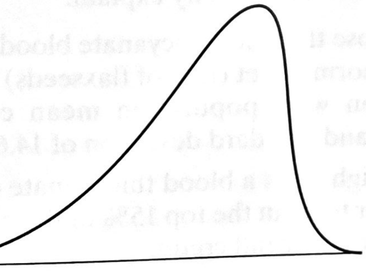
(I)



(II)



(III)



(b) The median is____________________ the mean. Choose one and explain your choice.

(i) greater than

(ii) equal to

(iii) less than

**I.2** Here is a list of life expectancies in 12 South American countries:

62, 64, 65, 66, 70, 71, 72, 73, 73, 74, 75, 75

The mean of these data is 70, and the SD is 4.6. (You do *not* need to verify this.) Without doing any calculations, which data point had the largest contribution to the SD? That is, if we could remove one of the data points, which data point should we remove if our goal is to make the SD of the remaining 11 points as small as possible? Why?

**I.3** In a group of 18 patients, there were 8 men and 10 women. Suppose we were to choose two of them, at random and without replacement. What is the probability that they would be the same sex?

**I.4** A researcher took a random sample of 20 mice and found that 5 of the 20 mice (25%) weighed more than 26 gm. *In the context of this setting*, explain what is meant by the sampling distribution of a percentage.

**I.5** Tree diameters for a certain species of tree are normally distributed with a mean of 20 cm and a standard deviation of 5 cm.

(a) What is the probability that the diameter of a randomly chosen tree will be between 16 cm and 23 cm?

(b) Suppose we take a sample of $n = 5$ trees. What is the probability that the average of the 5 diameters will be between 16 cm and 23 cm?

**I.6** Consider a hypothetical population of dogs in which there are four possible weights, all of which are equally likely: 40, 50, 65, or 70 pounds. A sample of size $n = 2$ is drawn from this population. We are interested in the sampling distribution of the total weight of the two dogs selected. How many possible values are there for the total?

**I.7** In a study of distance runners, the mean weight was 63.1 kg. Weights followed a normal distribution. Also, 75% of the weights were between 58.1 and 68.1 kg. The standard deviation of weights is

(i) less than 5 kg

(ii) equal to 5 kg

(iii) more than 5 kg

Choose one of these and explain your choice.

**I.8** Researchers wanted to compare two drugs, formoterol and salbutamol, in aerosol solution, for the treatment of patients who suffer from exercise-induced asthma. Patients were to take a drug, do some exercise, and then have their "forced expiratory volume" measured. There were 30 subjects available.[1]

(a) Should this be an experiment or an observational study? Why?

(b) Within the context of this setting, what is the placebo effect?

**I.9** Heights of American women ages 18–24 follow a normal distribution with an average of 64.3 inches. (Assume that measurements are made to the nearest 0.1 inch.) Moreover, 50% of the heights are between 62.5 inches and 66.1 inches. What is the standard deviation of heights?

**I.10** A certain cross between sweetpea plants will produce progeny that are either purple flowered or white flowered; the probability of a purple flowered plant is $p = 9/16$. Suppose 100 progeny are to be examined. Use the normal approximation to the binomial distribution to find the probability that at least 54 of them will be purple flowered.

**I.11** For each of the following cases (a and b),

(i) state whether the study should be observational or experimental and why.

(ii) state whether blinding should be used. If the study should be run blind or double-blind, who should be blinded and why?

(a) An investigation of whether taking an aspirin every day decreases the chance of having a stroke.

(b) An investigation of whether attending religious services regularly reduces blood pressure.

**I.12** For each of the following situations state whether or not a binomial would be an appropriate probability model for the variable $Y$ and explain why.

(a) Seeds of the garden pea (*Pisum sativum*) are either yellow or green. A certain cross between pea plants produces progeny that are in the ratio 3 yellow:1 green. Suppose your goal is to get 3 yellow, but you don't care how many green you get. You sample, one at a time, until you have exactly 3 progeny that are yellow. Let $Y$ be the number of progeny you have to observe in order to get 3 yellow. Is $Y$ a binomial random variable? Why or why not?

(b) Some people exercise every day, some exercise occasionally, and some never exercise. Suppose you take a random sample of 45 people and ask each of them how often they exercise. Let $Y$ be the number of people, out of 45, who exercise every day. Is $Y$ a binomial random variable? Why or why not?

Problems **I.13–I.16** refer to the following Flaxseed and cyanide case study

The *Journal of Nutrition and Food Science* contained an article entitled "Flaxseed (Linum usitatissimum L.) consumption and blood thiocyanate concentration in rats".[2] The questions below are motivated from this study.

*Flaxseed is a nutrient rich seed but contains cyanogenic glycosides, which can release hydrogen cyanide (HCN) into the body after consumption. This study aims to determine the cyanogenic content of raw and heated (170 °C, 15 min.) flaxseed as well as its effect on the blood thiocyanate (SCN2) concentration, a derivate of HCN, in rats.*

**I.13** One variable studied was the amount of thiocyanate (mg/L) in the blood after rats consumed raw or heated flaxseeds.

(a) Is this variable *numeric* or *categorical*?

(b) If numeric, is this variable inherently *discrete* or *continuous*? If categorical, is this variable *nominal* or *ordinal*?

**I.14** This study aims to determine whether or not heating flaxseeds before consumption can alter the amount of thiocyanate in the blood. Researchers fed 14 rats a diet consisting of 30% flaxseeds for 30 days and then measured the amount of thiocyanate in the blood. Seven of the rats were randomly assigned to eat diets with raw flaxseeds, while the remaining seven were given diets with flaxseeds that were heated at 170°C for 15 minutes.

(a) Briefly explain how you can tell that the study described above is an experiment and not an observational study.

(b) Will this study design allow the researchers to investigate whether or not heating the seeds can actually affect blood thiocyanate levels in flaxseed rich diets? Very briefly explain.

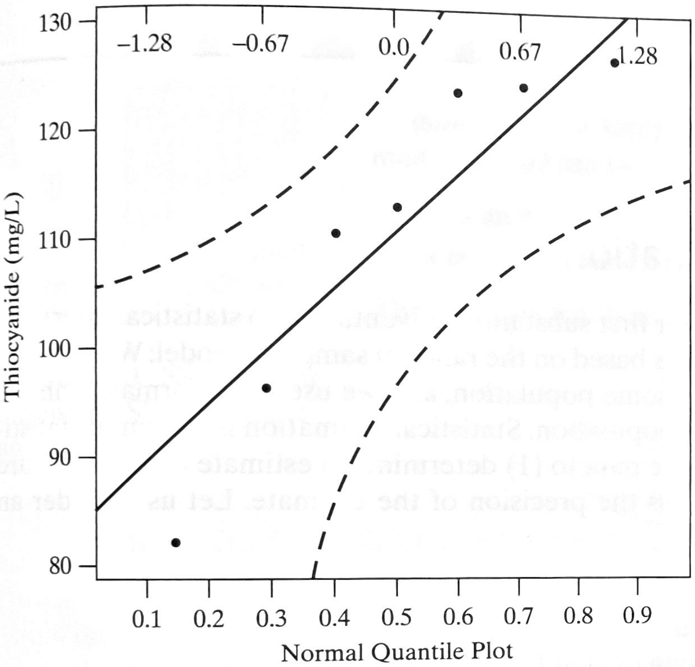
**I.15** Suppose that the thiocyanate blood concentration in rats fed a normal diet (free of flaxseeds) follows a normal distribution with population mean concentration of 53.3 mg/L and standard deviation of 14.6 mg/L.

(a) How high must a blood thiocyanate concentration be in order to be in the top 15% of the population? Show your work for full credit.

(b) The value you just computed in part (a) is called the ________ percentile. (Fill in the blank.)

(c) Would it be unusual to obtain a random sample of seven rats for which their sample mean thiocyanate concentration is more than 70.4 mg/L? Justify your answer, showing and briefly discussing a well-labeled illustration or appropriate computations.

**I.16** The following graph is a normal probability plot of the blood thiocyanide concentration for the seven rats eating the raw 30% flaxseed diet. Statistical software reports the Shapiro-Wilk's normality test $P$-value is 0.2675.

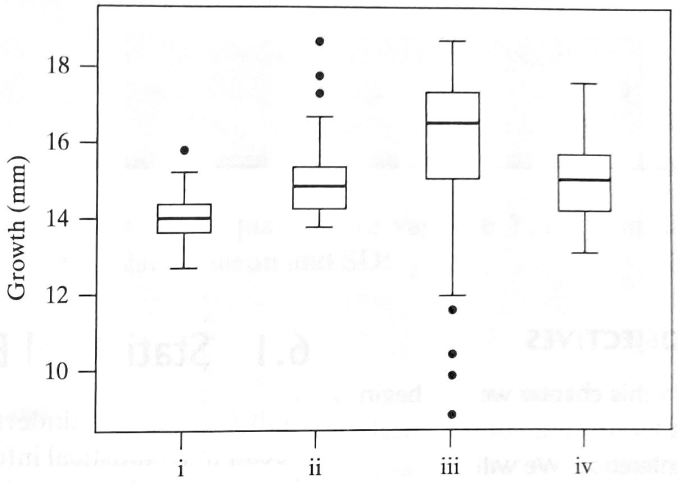


(a) Would it be reasonable to regard the population from which these data came as normal? Briefly explain using the graph and Shapiro-Wilk's test $P$-value to support your answer.

(b) Would it be reasonable to regard the sampling distribution of $\overline{Y}$ (for samples of size $n = 7$) to be normally distributed? Briefly explain.

(c) True or false? If the dots on the normal probability plot fell very closely to the line, we would have good evidence that blood thiocyanide concentrations are normally distributed.

**I.17** Consider the following four boxplots (i, ii, iii, iv).



(a) Among the four sets of data, which is most strongly skewed to the left?

(b) Among the four sets of data, which has the smallest standard deviation?

(c) Among the four sets of data, which has the lowest third quartile?

(d) The interquartile range of boxplot iii is approximately ______ (2, 5, 7, 10, 17). Choose a value.

**I.18** Bill lengths of a population of male blue jays have a normal distribution with mean 25.4 mm and standard deviation 0.8 mm. A bill is considered to be "short" if it is shorter than 24.0 mm. Suppose that a researcher has a large collection of these male blue jays and takes measurements each day on 10 of the birds. What is the percentage of days on which at least 1 blue jay out of the 10 has a short bill? Hint: Start by finding the probablity of a short bill for one blue jay.