

Describing Data - Graphically and Numerically The Normal Distributions

In the paper "Reproduction in Laboratory colonies of Bank Vole," the authors presented the results of a study of litter size. (A vole is a small rodent with a stout body, blunt nose, and short ears.) As each new litter was born, the number of babies was recorded, and the accompanying results were obtained.

1	4	4	5	5	6	6	7	7	8
2	4	5	5	5	6	6	7	7	8
2	4	5	5	6	6	6	7	7	8
3	4	5	5	6	6	6	7	8	8
3	4	5	5	6	6	7	7	8	9
3	4	5	5	6	6	7	7	8	9
3	4	5	5	6	6	7	7	8	9
3	4	5	5	6	6	7	7	8	10
3	4	5	5	6	6	7	7	8	10
4	4	5	5	6	6	7	7	8	11

The authors also kept track of the color of the first born in each litter. (B=brown, G=gray, W=white, and T=tan)

B	B	T	W	T	G	G	G	B	B
W	B	W	B	T	T	G	B	T	B
B	T	B	B	B	G	W	B	B	G
G	G	G	B	B	T	B	W	T	T
B	T	B	B	T	W	W	B	G	B
B	B	B	G	T	B	B	T	T	G
G	B	B	B	B	G	W	G	T	G
B	B	B	B	G	G	T	T	W	G
G	W	T	G	T	B	B	G	B	B
B	G	T	W	B	G	T	W	G	W

- Which variable, litter size or color, is categorical?
- Which variable is quantitative?
- Make a bar chart of the colors.
- Make a histogram of the litter sizes.
- Make a dotplot of the litter sizes.
- Are there any outliers in the histogram or dotplot?
- Describe the shape of the histogram (symmetric or skewed).
- Find the mean of the litter sizes.
- Is the mean resistant to outliers?
- Find the median of the litter sizes.
- Is the median resistant to outliers?
- Find the range of the litter sizes.

13. Find the 5-number summary of the litter sizes.

14. What is the interquartile range?

15. Make a boxplot of the litter sizes.

16. Find the variance of the litter sizes.

17. Find the standard deviation of the litter sizes.

18. Is standard deviation resistant to outliers?

23. If a curve is skewed to the right, the (mean or median) will be further to the right than the (mean or median).

24. What is the difference between \bar{x} and μ ?

25. What is the difference between s and σ ?

28. Sketch the graph of $N(266, 16)$, the distribution of pregnancy length from conception to birth for humans.

29. What is the 68-95-99.7 rule?

30. Using the empirical rule (the 68-95-99.7 rule), find the length of the longest 16% of all pregnancies. Sketch and shade a normal curve for this situation.

31. Using the empirical rule, find the length of the middle 99.7% of all pregnancies. Sketch and shade.

32. Using the empirical rule, find the length of the shortest 2.5% of all pregnancies. Sketch and shade.

33. Using the empirical rule, what percentile rank is a pregnancy of 218 days?

34. What percentile rank is a pregnancy of 298 days?

35. What percentile is a pregnancy of 250 days?

36. What is the percentile of a pregnancy of 266 days?

37. What z-score does a pregnancy of 279 days have?

38. What percent of humans have a pregnancy lasting less than 279 days? Sketch and shade a normal curve.

39. What z-score does a pregnancy of 257 days have?
40. What percent of humans have a pregnancy lasting less than 257 days? Sketch and shade.
41. What percent of humans have a pregnancy lasting longer than 280 days? Sketch and shade.
42. What percent of humans have a pregnancy lasting between 260 and 270 days? Sketch and shade.
43. Would you say pregnancy length is a continuous or discrete variable? Justify.
45. How long would a pregnancy have to last to be in the longest 10% of all pregnancies?
46. How short would a pregnancy be to be in the shortest 25% of all pregnancies?
47. How long would a pregnancy be to be in the middle fifth of all pregnancies?

49. Make a back-to-back stemplot of the following data:

Reading Scores

4th Graders	12	15	18	20	20	22	25	26	28	29
	31	32	35	35	35	36	37	39	40	42
7th Graders	1	12	15	18	18	20	23	23	24	25
	27	28	30	30	31	33	33	33	35	36

50. Make a comparison between 4th grade and 7th grade reading scores based on your stemplot.
51. What is the mode of each set of scores?

Exploring Bivariate Data

55. Graph the following hot dog data:

Calories	Sodium (milligrams)
108	149
130	350
132	345
135	360
138	360
140	375
144	380
145	390
150	400
163	415
167	400
172	420
176	450
180	500
184	505
195	500
200	515

56. What is the response variable?

57. What is the explanatory variable?

58. What is the direction of this scatterplot?

59. What is the form of this scatterplot?

60. What is the strength of this scatterplot?

61. Are there clusters? If so, where?

62. Are there outliers?

64. Calculate and interpret the correlation.

65. Calculate the correlation without the point (108, 149).

67. What two things does correlation tell us about a scatterplot?

68. If I change the units on sodium to grams instead of milligrams, what happens to the correlation?

69. What is the highest correlation possible?

70. What is the lowest correlation possible?

71. Correlation only applies to what type(s) of relationship(s)?

72. Is correlation resistant to outliers?

73. Does a high correlation indicate a strong cause-effect relationship?

74. Sketch a scatterplot with a correlation of about 0.8.

75. Sketch a scatterplot with a correlation of -0.5 .

76. Find the least-squares regression line (LSRL) for the calories-sodium data.
77. Draw the LSRL on your scatterplot.
78. What is the slope of this line, and what does it tell you in this context?
79. What is the y-intercept of this line, and what does it tell you in this context?
80. Predict the amount of sodium in a hot dog with 155 calories.

81. Predict the amount of sodium in a hot dog with 345 calories.

82. Why is the prediction in problem 80 acceptable but the prediction in problem 81 not?

85. The point (\bar{x}, \bar{y}) is always on the LSRL. Find this point, and label it on your scatterplot.
86. Find the standard deviation of the calories.
87. Find the standard deviation of the sodium.
88. Using the equations on your formula card, verify the slope and intercept of the LSRL.
89. Find the coefficient of determination for this data.
90. What does r^2 tell you about this data?

Producing Data

119. What is the difference between an observational study and an experiment?

120. What is a voluntary response sample?

121. How are a population and a sample related but different?

122. Why is convenience sampling biased?

123. SRS stands for what kind of sample? Name and define.

124. Discuss how to choose a SRS of 4 towns from this list:

Allendale
Gratiot

Bangor
Hillsdale Ionia

Chelsea
Joliet

Detour
Kentwood

Edmonton
Ludington

Fennville

125. What is a stratified random sample?

130. Why is the wording of questions important? Give an example.

133. If I test a drug at 100 mg, 200 mg, and 300 mg, I am testing one variable at three _____.

134. What is the placebo effect?

135. What is the purpose of a control group?

136. Give an example of when we may not want to use a placebo/control group.

139. What does double-blind mean, and why would we want an experiment to be double-blind?

Probability

146. What is independence?
147. You are going to flip a coin three times. What is the sample space for each flip?
148. You are going to flip a coin three times and note how many heads and tails you get. What is the sample space?
149. You are going to flip a coin three times and note what you get on each flip. What is the sample space?
150. Make a tree diagram for the three flips.
151. There are three ways I can drive from Fremont to Grand Rapids and four ways I can drive from Grand Rapids to my home. How many different ways can I drive from Fremont to my home through Grand Rapids?
152. How many different four-digit numbers can you make?
153. How many different four-digit numbers can you make without repeating digits?
154. What is an event in probability?
155. Any probability is a number between (and including) _____ and _____.
156. All possible outcomes together must have probability of _____.
157. If S is the sample space, $P(S) =$ _____.
158. What are complements? Give an example and draw a Venn diagram.

Use the following chart for questions 160-163:

M & M Color	Brown	Red	Yellow	Green	Orange	Blue
Probability	0.3	0.2	0.2	0.1	0.1	?

160. What is the probability that an M & M is blue?
161. What is the probability that an M & M is red or green?
162. What is the probability that an M & M is yellow and orange?
163. What is the probability that an M & M is not brown or blue?
164. Bre can beat Erica in tennis 9% of the time. Erica can swim faster than Bre 8% of the time. What is the probability that Bre would beat Erica in a tennis match and in a swimming race?
165. What assumption are you making in problem 164? Do you think this assumption is valid?
166. Using two dice, what is the probability that you would roll a sum of seven or eleven?
167. Using two dice, what is the probability that you would roll doubles?
168. Using two dice, what is the probability that you would roll a sum of 7 or 11 on the first roll and doubles on the second roll?

169. What assumption are you making in problem 168? Do you think this assumption is valid?

170. Using two dice, what is the probability that you would roll a sum of 7 or 11 that is also doubles?

175. Make a Venn diagram for the following situation:

- 45% of kids like Barney
- 25% of kids like Blue
- 55% of kids like Pooh
- 15% of kids like Blue and Pooh
- 25% of kids like Barney and Pooh
- 5% of kids like Barney, Blue, and Pooh
- 5% of kids like Blue but not Barney or Pooh

179. Make a probability histogram of the following grades on a four-point scale:

Grade	0	1	2	3	4
Probability	0.05	0.28	0.19	0.32	0.16

180. Using problem 179, what is $P(X > 2)$?

181. Using problem 179, what is $P(X \geq 2)$?

182. What is a uniform distribution? Draw a picture.

186. Normal distributions are (continuous or discrete).

187. Expected value is another name for _____.

188. Find the expected value of the grades in problem 179.

Sampling Distributions

1. The Helsinki Heart Study asks whether the anti-cholesterol drug gemfibrozil will reduce heart attacks. In planning such an experiment, the researchers must be confident that the sample sizes are large enough to enable them to observe enough heart attacks. The Helsinki study plans to give gemfibrozil to 2000 men and a placebo to another 2000 men. The probability of a heart attack during the 5-year period of the study for men this age is about 0.04. We can think of the study participants as an SRS from a large population, of which the proportion $p = 0.04$ will have heart attacks.
 - a. What is the expected number of heart attacks that the study will find in one group of 2000 men if the treatment doesn't change the probability of 0.04?

 - b. What is the probability that the group will suffer at least 75 heart attacks?

2. Children in kindergarten are sometimes given the Raven Progressive Matrices Test (RPMT) to assess their readiness for learning. Experience at Southward Elementary School suggests that the RPMT scores for its kindergarten pupils have a mean of 13.6 with a standard deviation of 3.1. The distribution is close to Normal. Mr. Brown has 22 children in his kindergarten class this year.
 - a. What is the probability that class's mean score will be less than 12.0?

 - b. Mr. Brown suspects that the class RPMT scores will be unusually low because the test was interrupted by a fire drill. He wants to find the level L such that there is only a probability of 0.05 that the mean score of his class fall below L . What is this value of L . (Hint: this requires you to find the z -score and then convert to the x -score.)

3. What is the Central Limit Theorem? How is the CLT used in sampling distributions?

Estimating with Confidence and Testing a Claim

4. Estimating Population Parameters.

- a. A random sample of 12 four-year old red pine trees was selected and the diameter (in) of each tree's main stem was measured. The resulting observations are as follows:

11.3 10.7 12.4 15.2 10.1 12.1 16.2 10.5 11.4 11.0 10.7 12.0

Find the estimate that can be used to estimate the true population mean.

Find the estimate that can be used to estimate the true population standard deviation.

Find the estimate that be used to estimate the true population proportion of trees whose diameter is greater than the average.

5. What is meant by the margin of error of a population parameter? What are the margin of errors for the following:

Population Mean

Population Proportion

6. What is the general form of all confidence intervals?

8. Retailers report that the use of cents-off coupons is increasing. The Scripps Howard News Service reported that proportion of all households that use coupons as 0.77. Suppose that this estimate was based on a random sample of 800 households.

a. Construct and interpret a 95% confidence interval for the true population proportion.

b. The manager of the retail store in reporting to his superiors claims that the true proportion of customers that use coupons is 80%. Carry out a hypothesis test of the manager's claim. Follow the four-step process.

c. Which type of error (Type I or Type II) could have been made based on your decision in part b? Describe that error in the context of the problem.