Answers:

1. D
2. B
3. A
4. D
5. B
6. B
7. C
8. A
9. C
10. A
11. B
12. B
13. D
14. A
15. C
16. B
17. D
18. D
19. C
20. C
21. B
22. A
23. D
24. D
25. C
26. A
27. E
28. B
29. A
30. A

Solutions:

1. Categorical variables place individuals into groups. Age is quantitative because it takes on numerical values.

2. A skewed-left distribution has a long tail to the left, and most of the data lie to the right. The median is more resistant and is not as affected by low-value outliers.

3. The mean of the data is $\frac{5+6+7+11+15+16+17}{7}=\frac{77}{7}=11$, and the standard deviation would be found by dividing by $n=7$ instead of $n-1=6$ since the problem states that the data represent the population. Therefore, the standard deviation is
$$\sqrt{\frac{(5-11)^2+(6-11)^2+(7-11)^2+(11-11)^2+(15-11)^2+(16-11)^2+(17-11)^2}{7}}$$
$$=\sqrt{\frac{36+25+16+0+16+25+36}{7}}=\sqrt{\frac{154}{7}}=\sqrt{22}=4.690.$$

4. A and B are standard properties of normal density curves, and C is true also because the areas underneath normal density curves is always related to the standard deviation of the data. Therefore, all of the statements are true.

5. On the calculator, $\text{normalcdf}(-1.4,1.4,0,1)=0.838486...$

6. On the calculator, $\text{normalcdf}(130,10^{99},120,20)=0.3085...$, so to the nearest percentage, the answer is 31%.

7. The closer r gets to 1 or -1, the closer the data approach a linear distribution. The closer to 0, from either direction, the data become less linear.

8. $\text{residual}=\text{observed}-\text{predicted}$, so the residual for this data point is $8-(1.5+0.75\cdot 3)=8-3.75=4.25$.

9. Negative r values indicate an inverse relationship, so A is true. B is also true in that correlation coefficients are always between -1 and 1. Therefore, C would have to be false.

10. I is true because correlation does not imply causation is true for any variables. II is false because although averaging over a long time period smoothes out the day-to-day variations, this isn't necessarily always the case. III is false since there are multiple ways to regress (linear, quadratic, etc.).

11. A stratum is a division of the population into individuals with similar attributes, so it can't be the entire population; hence, I is false. A census attempts to contact every individual in the entire population, so II is true. A sample doesn't consist of the entire population, so III is false.

12. Making the standard deviation smaller would improve the accuracy, and the standard deviation of a sample is the standard deviation of the population divided by the square root of the number of items in the sample. Thus, increasing the sample size does improve the accuracy, and using smaller sample sizes makes the estimate less accurate. Avoiding bias also improves accuracy, so the answer is B.

13. Undercoverage occurs when some groups in the population are left out when choosing the sample, so this would cause bias. Nonresponse occurs when an individual chosen for the sample cannot be contacted or does not cooperate, so this would cause bias. Poorly worded questions may give misleading results due to the fact that all questions need to be worded the same for all participants, so this would cause bias. Thus, all three answers cause bias.

14. Some probability sampling designs give each member of the population an equal chance to be selected. The use of change to select the sample is the essential principle of statistical sampling, so A is correct. B is incorrect because we may not be able to accurately interview each household, or, for example, how might we include homeless people. This really depends on the population. C is incorrect because censuses are not sampling methods.

15. $P(\text{tail}) = 1/2$ and $P(\text{odd}) = 3/6 = 1/2 \Rightarrow P(\text{tail and odd}) = 1/2 \cdot 1/2 = 1/4$ since the two events are independent.

16. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A) \cdot P(B) = 0.25 + 0.5 - 0.25 \cdot 0.5 = 0.625$

17. For A, $B \subseteq A$, so they are not disjoint. For B, both sets contain the number 21, for example, so they are not disjoint. For C, both sets contain 2, for example, so they are not disjoint. For D, the two sets are disjoint because no number is simultaneously less than 5 and greater than 10. Therefore, the answer is D.

18. There are two outcomes of each toss of the coin, and there are four coins, so the total number of outcomes is $2^4 = 16$.

19. $P(Y) = P(X \cup Y) + P(X \cap Y) - P(X) = 0.25 + 0.15 - 0.2 = 0.2 \Rightarrow P(Y^C) = 1 - 0.2 = 0.8$

20. Because the table gives a probability distribution, we must have $1 = k + 2k + 3k + 4k = 10k \Rightarrow k = 1/10$. Therefore, $P(X > 2) = 4k = 4 \cdot 1/10 = 2/5 = 0.4$.

21. $\bar{x} = 10(0.2) + 20(0.2) + 40(0.2) + 60(0.4) = 2 + 4 + 8 + 24 = 38$

22. $\sigma = \sqrt{(10-38)^2(0.2) + (20-38)^2(0.2) + (40-38)^2(0.2) + (60-38)^2(0.4)} = \sqrt{416}$ $= 20.396078...$, so rounded to the nearest ten-thousandth, the answer is 20.3961.

23. All three are standard assumptions of binomial distributions.

24. $\mu = \dfrac{90 \cdot 1 + 10 \cdot 0}{100} = 0.9$, so $\sigma = \sqrt{\dfrac{90(1-0.9)^2 + 10(0-0.9)^2}{100}} = \sqrt{0.09} = 0.3$

25. A level C confidence interval for μ is $\bar{x} = z^* \pm \dfrac{\sigma}{\sqrt{n}}$, where z^* is the value with area C between $-z^*$ to z^* under the standard normal curve. Therefore, decreasing the confidence level makes the confidence interval get smaller., so I is false. Decreasing the sample size would make the error larger, making the width of the confidence interval larger, so II is true. Increasing the standard deviation also makes the error larger, so III is true as well.

26. This is a two-sided test since the alternative hypothesis features a $\neq$ sign. Therefore, on the calculator, $1 - \text{normalcdf}(-1.6, 1.6, 0, 1) = 0.1095985...$, so rounded to the nearest ten-thousandth, the p-value is 0.1096.

27. You have rejected the null hypothesis and accepted the alternative hypothesis when the alternative hypothesis is true. No error has been commited.

28. The chi-square distribution is skewed-right, so it is not symmetric, making I true and III false. For II, p-values come from the chi-square distribution with $r - 1$ degrees of freedom. If $r = 2$, we are comparing just 2 proportions. The chi-square statistic is just the square of the Z statistic, and the p-value for chi-square is exactly the same as the 2-sided p-value for Z, so II is true.

29. Using the calculator, enter stats $\rightarrow$ tests $\rightarrow$ 4:2-SampTTest, and enter the information given for the means, standard deviations, and sample sizes. If Slow jazz is used as the data with subscript 1, select $\mu_1 > \mu_2$ as the hypothesis and pool the

data. The result gives $t = 2.828$ with $p = 0.003208...$, so the p-value, rounded to the nearest ten-thousandth, is 0.0032.

30. There are 84 seedlings total, so the expected value of the number of green seedlings is $84(.25) = 21$; for yellow-green, the expected number is $84(.5) = 42$; and for albino, the expected number is $84(.25) = 21$. The chi-square statistic is

$$\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{(22-21)^2}{21} + \frac{(50-42)^2}{42} + \frac{(12-21)^2}{21} = \frac{1}{21} + \frac{64}{42} + \frac{81}{21} = \frac{114}{21}$$

Therefore, using the calculator, $\chi^2\text{cdf}\left(114/21, 10^{99}, 2\right) = 0.06625225...$, so the p-value, rounded to the nearest hundred-thousandth, is 0.06625.