

1. (10 points total)

Identify the scale (eg – discrete/continuous and nominal/ordinal or interval/ratio)for the following variables.

1a (1 point). Calories consumed during the day

Answer: Continuous, ratio

1b (1 point) Marital status

Answer: Discrete, nominal

1c (1 point) Perceived health status reported as “poor”, “fair”, “good” or “excellent”

Answer: Discrete, ordinal

1d (1 point) Blood type

Answer: Discrete, nominal

1e (1 point) IQ score

Answer: Continuous, interval. I will also take discrete, ordinal

For the following health interview survey items, propose how the variable should be defined for analysis purposes. Specifically, suggest a variable name and coding scheme for its measurement. (Eg for age: **variable name: AGE coding scheme: years**)

1f (1 point) Weight

Answer: variable name: WEIGHT coding scheme: lbs, or kilograms

1g (1 point) Height

Answer: variable name: HEIGHT coding scheme: inches, feet, or centimeters, or meters

1h (1 point) Family Income

Answer: variable name: FINCOME coding scheme: \$/year

1i (1 point) Unemployment

Answer: variable name: EMPLOY coding scheme: Yes/No

1j (1 point) Number of stays in a mental hospital

Answer: variable name: STAYS coding scheme: Times

2. (10 points total)

The following table summarizes data obtained from a survey of 4390 persons regarding their awareness of suicide. The question asked of each study participant was: “How many people have you ever known who were victims of suicide?”

Number of Victims	Frequency
0	3944
1	279
2	97
3	40
4 or more	30
Total	4390

2a. (2 points).

Calculate the sample mean number of victims reported. In doing your calculations, give a score of 4.5 to the response “4 or more”.

Answer: Sample mean $\bar{X} = \frac{0 \times 3944 + 1 \times 279 + 2 \times 97 + 3 \times 40 + 4.5 \times 30}{4390} \approx 0.166$. **Thus, the sample mean number of victims reported is 0.166.**

2b. (2 points).

Calculate the sample median number of victims reported.

Answer: The data are already given in value order, thus, the median is the value of middle number. Since here the total number 4390 is even, the median should be (2195th value + 2196th value)/2 = (0 + 0)/2 = 0.

2c. (2 points)

If 1744 observations shift from a value of 0 to a value of “4 or more”, how does the sample mean change? That is, what is the new value of the sample mean?

Answer:

New sample mean $\bar{X}_{(new)} = \frac{0 \times (3944 - 1744) + 1 \times 279 + 2 \times 97 + 3 \times 40 + 4.5 \times (30 + 1744)}{4390} \approx 1.953$

2d. (2 points)

If 1744 observations shift from a value of 0 to a value of “4 or more”, how does the sample median change? That is, what is the new value of the sample median?

Answer: If 1744 observations shift from a value of 0 to a value of “4 or more”, there are still 3944-1744=2200 observations with value of 0. Thus, the new median = (2195th value + 2196th value)/2 = (0 + 0)/2 = 0.

2e. (2 points)

Consider a study of $n=60$ subjects in which subjects are asked “How many serious motor vehicle accidents have you had in the past 5 years?” You are asked to summarize the responses obtained and you need to choose between reporting the sample mean or the sample median. Which statistic would you choose to report and why?

Answer: Generally, the median is the preferred summary for skewed data as this is a better representation of the majority and is not influenced by extremes. However, I will accept “mean” as the answer if you are reasoning that you very much want to know about the extremes in this instance.

3. (10 points total)

3a. (2 points).

True or False:

The sample mean, sample median and sample mode can never be all the same.

Answer: False. When the distribution is perfectly symmetric, they will all be the same

3b. (2 points).

True or False:

The sample mean is always one of the data points.

Answer: False. When the sample size is even, the median cannot be an actual data point.

3c. (2 points).

True or False:

When the sample size “n” is odd, the sample median is one of the data points.

Answer: True.

3d. (2 points).

According to a newspaper article, the mean wage for a sports player of a particular sport in the U.K. was 676,000 pounds. Suppose it is known that the distribution of wages is skewed to the right. True or False: The median salary is greater than 676,000 pounds.**Answer: False. When the data are positively skewed, the mean > median**

3e. (2 points)

According to a recent National Center for Health Statistics Survey, among males aged 25-34 years: 2% have heights equal to 64 inches or less, 8% have heights equal to 66 inches or less, 27% have heights equal to 68 inches or less, 39% have heights equal to 69 inches or less, 54% have heights equal to 70 inches or less, 68% have heights equal to 71 inches or less, 80% have heights equal to 72 inches or less, 93% have heights equal to 74 inches or less, and 98% have heights equal to 76 inches or less. Which category of height has the median height? Explain your answer.

Answer: The category of “have heights more than 69 inches, but equal to 70 inches or less” has the median height. The median is the value at 50%th percentile. In this sample, “39% have heights equal to 69 inches or less while 54% have heights equal to 70 inches or less”. Thus the 50%th percentile has to be somewhere between 69 inches and 70 inches.

4. (10 points total)

A census survey taken by the U.S. Bureau of the Census reports that median earnings in the past 12 months was \$32,168 for women and \$41,975 for men. It also reported that mean earnings in the past 12 months was \$39,890 for women and \$56,724 for men.

4a. (5 points).

Does this data suggest that the distribution of income for each gender is symmetric, positive skewed (right skewed) or negative skewed (left skewed)? Explain your reasoning.

Answer: For each gender, the distribution is positive skewed (right skewed), because the mean is greater than the median.

4b. (5 points)

The results were obtained from a survey of 73.8 million women and 83.4 million men. Calculate the overall sample mean income.

Answer:

Overall sample mean income =

$$\frac{\text{mean income of women} \times \text{sample size of women} + \text{mean income of men} \times \text{sample size of men}}{\text{sample size of women} + \text{sample size of men}}$$

$$= \frac{39890 \times 73.8 + 56724 \times 83.4}{73.8 + 83.4} = 48821.02$$

Thus, the overall sample mean income is \$48,821.02.

5. (10 points total)

A local arts and crafts store has found, after many years of experience, 20% of people who enter the store actually buy something.

Suppose 3 people enter the store.

5a. (5 points).

How many possible outcomes are possible for the sale or non-sale of items to the three potential customers? *Hint: In developing your answer, you are actually specifying the sample space for the outcome of sales and non-sales to the three potential customers.*

Answer: For each person, there are two possible outcomes: buy something, or not buy anything. Thus, for the 3 persons, there are $2 \times 2 \times 2 = 2^3 = 8$ possible outcomes.

5b. (5 points).

What is the probability of at least one sale?

Answer:

$$\begin{aligned}
 & \mathbf{P[\text{At least one sale}]} \\
 &= \mathbf{P[\text{One sale}] + P[\text{Two sales}] + P[\text{Three sales}]} \\
 &= \mathbf{1 - P[0 \text{ sale}]} \\
 &= \mathbf{1 - P[1^{\text{st}} \text{ person didn't buy}] \times P[2^{\text{nd}} \text{ person didn't buy}] \times P[3^{\text{rd}} \text{ person didn't buy}]} \\
 &= \mathbf{1 - (1 - P[1^{\text{st}} \text{ person did buy}] \times (1 - P[2^{\text{nd}} \text{ person did buy}]) \times (1 - P[3^{\text{rd}} \text{ person did buy}])} \\
 &= \mathbf{1 - (1 - 20\%) \times (1 - 20\%) \times (1 - 20\%)} \\
 &= \mathbf{1 - 0.8^3} \\
 &= \mathbf{0.488}
 \end{aligned}$$

Thus, the probability of at least one sale is 0.488.

6. (10 points)

The following table summarizes results on an ELISA test used to aid in diagnosing Anthrax. These results are based on tests performed on 12 Anthrax cases, and 18 persons who did not have Anthrax. Use this table to answer the following questions.

ELISA for poly-D-glutamic acid capsule

	Anthrax Case	No Anthrax	Total
Test Positive	11	2	13
Test Negative	1	16	17
Total	12	18	30

6a. (2 points)

What is the sensitivity of the test?

Answer: Sensitivity = $P(\text{Test Positive} \mid \text{Anthrax}) = 11 / 12 \approx 91.7\%$.

6b. (2 points)

What is the specificity of the test?

Answer: Specificity = $P(\text{Test Negative} \mid \text{No Anthrax}) = 16 / 18 \approx 88.9\%$.

6c. (3 points)

Anthrax is relatively rare in the US population. Assume that the prevalence of Anthrax is 1 in 1,000,000. Now, suppose a simple random sample of 1,000,000 people are tested for Anthrax. Among the 1,000,000 people tested, what proportion would you expect to have positive test results?

Answer:

$$\begin{aligned}
 & \mathbf{P(\text{Test Positive})} \\
 &= \mathbf{P(\text{Test Positive and Anthrax}) + P(\text{Test Positive and Non Anthrax})} \\
 &= \mathbf{P(\text{Test Positive} \mid \text{Anthrax}) P(\text{Anthrax}) + P(\text{Test Positive} \mid \text{Non Anthrax}) P(\text{Non Anthrax})} \\
 &= \mathbf{P(\text{Test Positive} \mid \text{Anthrax}) P(\text{Anthrax}) + [1 - P(\text{Test Negative} \mid \text{Non Anthrax})] [1 - P(\text{Anthrax})]} \\
 &= \mathbf{\text{Sensitivity} \times \text{Prevalence} + [1 - \text{Specificity}] \times [1 - \text{Prevalence}]} \\
 &= \mathbf{11/12 \times 1/1,000,000 + [1 - 16/18] \times [1 - 1/1,000,000]} \\
 &= \mathbf{0.1111119}
 \end{aligned}$$

6d. (3 points)

How many false positive results would you expect as a result of the 1,000,000 tests?

Answer:

$$\begin{aligned} & \# \text{ of false positive} \\ &= P(\text{Test Positive} \mid \text{Not Anthrax}) \times P(\text{Not Anthrax}) \times 1,000,000 \\ &= [1 - \text{Specificity}] \times [1 - \text{Prevalence}] \times 1,000,000 \\ &= [1 - 16/18] \times [1 - 1/1,000,000] \times 1,000,000 \\ &= 111,111 \end{aligned}$$

7. (10 points total)

Suppose you take your temperature once a week for twelve weeks and the values are 98.2, 99.0, 98.5, 99.1, 98.6, 97.9, 98.2, 99.3, 98.1, 98.5, 99.0, and 98.8.

7a. (5 points)

Compute the sample range, sample variance, and sample standard deviation.

Answer:

Sample range: [97.9, 99.3]

Sample mean: $\bar{X} = \frac{98.2+99.0+98.5+99.1+98.6+97.9+98.2+99.3+98.1+98.5+99.0+98.8}{12} = 98.6$

Sample variance: $Var[X] = \frac{1}{12-1} \sum_{i=1}^{12} (X_i - \bar{X})^2 \approx 0.1982$

Sample standard deviation: $s = \sqrt{Var[X]} = \sqrt{0.1982} \approx 0.4452$

7b. (5 points)

What proportion of the twelve observations are within ± 0.68 standard deviations of the mean?

Hint – In developing your answer, you will need to compute the sample mean.

Answer:

The interval of mean ± 0.68 standard deviations is: $98.6 \pm 0.68*0.4452 = [98.3, 98.9]$.

There are 4 observations (98.5, 98.5, 98.6, 98.8) that are within ± 0.68 standard deviations of the mean. Thus, the required proportion is $4/12 = 33.3\%$.

8. (10 points total)

8a. (5 points)

Perhaps you've heard the saying "if you observe for long enough, a monkey will eventually write a Shakespeare play just by chance". **Hah!** Consider the setting of putting a typewriter in the hands of a monkey. Assume the typewriter has 50 keys and the monkey types key after key at random. What is the probability that the first seven keys that the monkey types is the seven character title *macbeth*?

Answer:

$$\begin{aligned}
 & \mathbf{P(\text{first seven characters are "macbeth"})} \\
 &= \mathbf{P(\text{first character is "m"}) P(\text{second character is "a"}) \dots P(7^{\text{th}} \text{ character is "h"})} \\
 &= \mathbf{1/50 \times 1/50 \times \dots \times 1/50} \\
 &= \mathbf{(1/50)^7}
 \end{aligned}$$

8b. (5 points)

Consider a jury trial setting in which the probability that a defendant is convicted, given he or she is actually guilty, is 0.95 and that the probability that a defendant is acquitted, given that he or she is actually innocent is 0.95. Suppose further that 90% of all defendants are actually guilty. If it is known only that a defendant is convicted, what is the probability that he or she is actually innocent?

Answer:

$$\begin{aligned}
 & \mathbf{P(\text{Convicted} | \text{Guilty}) = 0.95} \\
 & \mathbf{P(\text{Acquitted} | \text{Innocent}) = 0.95} \\
 & \mathbf{P(\text{Guilty}) = 90\% = 0.9} \\
 & \mathbf{P(\text{Innocent} | \text{Convicted})} \\
 &= \mathbf{1 - P(\text{Guilty} | \text{Convicted})} \\
 &= \mathbf{1 - \frac{P(\text{Guilty and Convicted})}{P(\text{Convicted})}} \\
 &= \mathbf{1 - \frac{P(\text{Convicted} | \text{Guilty}) P(\text{Guilty})}{P(\text{Convicted} | \text{Guilty}) P(\text{Guilty}) + P(\text{Convicted} | \text{Innocent}) P(\text{Innocent})}} \\
 &= \mathbf{1 - \frac{P(\text{Convicted} | \text{Guilty}) P(\text{Guilty})}{P(\text{Convicted} | \text{Guilty}) P(\text{Guilty}) + [1 - P(\text{Acquitted} | \text{Innocent})] [1 - P(\text{Guilty})]}} \\
 &= \mathbf{1 - \frac{0.95 \times 0.9}{0.95 \times 0.9 + [1 - 0.95] \times [1 - 0.9]}} \\
 & \approx \mathbf{0.0058 = 0.58\%}
 \end{aligned}$$

That is, if it is known only that a defendant is convicted, the probability that he or she is actually innocent is 0.0058.

9. (10 points total)

9a. (5 points)

Suppose the scores on a difficult exam have a sample mean of 67 and a standard deviation of 20. Suppose the instructor decides to boost each student's score by 10 points; that is, the instructor adds 10 points to each student's score. What are the values of the new sample mean and sample standard deviation?

Answer:**The new sample mean = the old sample mean + 10 = 77****The new sample standard deviation = the old sample standard deviation = 20**

9b. (5 points)

Next, suppose that the sample mean income for a group of workers is \$59,000 with a standard deviation of \$15,000. You are hired to report this information at a meeting in the U.K. If one British pound equals \$2.25, what is the value of the sample mean and sample standard deviation in units of British pounds?

Answer:**The sample mean in unit of British pounds = the sample mean in unit of US dollars * British pounds per US dollar = $59,000 * 1 / 2.25 = 26,222.22$ British pounds****The sample standard deviation in units of British pounds = the sample standard deviation in unit of US dollars * British pounds per US dollar = $15,000 * 1 / 2.25 = 6,666.67$ British pounds**

10. (10 points total)

The following are hypothetical data from a study of the relationship between cigarette smoking and the risk of low birth weight. Suppose preliminary analyses included the classification of infants into categories according to the number of cigarettes smoked per day by the mother. In particular, suppose the probabilities of the different smoking levels are the following:

Cigarettes per day	Probability
Zero	0.9082
1-5	0.0855
6-20	0.0059
> 20	0.0004
Total	1.0000

Next, suppose the following *conditional* probabilities are known.

Condition: Mother smokes	Conditional Probability of low birth weight infant
Zero cigarettes/day	0.035
1-5	0.379
6-20	0.813
> 20	0.540

10a. (5 points)

Calculate the probability of a low birth weight infant.

Answer:

$$\begin{aligned}
 & \mathbf{P(\text{Low birth weight infant})} \\
 & = \mathbf{P(\text{Low birth weight infant} \mid \text{Smoke 0 cigarettes/day}) P(\text{Smoke 0 cigarettes/day}) + P(\text{Low birth weight infant} \mid \text{Smoke 1-5 cigarettes/day}) P(\text{Smoke 1-5 cigarettes/day}) + P(\text{Low birth weight infant} \mid \text{Smoke 6-20 cigarettes/day}) P(\text{Smoke 6-20 cigarettes/day}) + P(\text{Low birth weight infant} \mid \text{Smoke } > 20 \text{ cigarettes/day}) P(\text{Smoke } > 20 \text{ cigarettes/day})} \\
 & = \mathbf{0.035 * 0.9082 + 0.379 * 0.0855 + 0.813 * 0.0059 + 0.540 * 0.0004} \\
 & \approx \mathbf{0.0692}
 \end{aligned}$$

Thus, the probability of a low birth weight infant is 0.0692.

10b. (5 points)

Are the events “>20 cigarettes per day” and “low birth weight infant” independent? Explain your reasoning using the rules of probability developed in class.

Answer: They are NOT independent.

According to the rules of probability, if A and B are independent, then $P(A|B) = P(A)$.

Here, $P(\text{Low birth weight infant} \mid > 20 \text{ cigarettes per day}) = 0.540$, which is much larger than $P(\text{Low birth weight infant}) = 0.0692$.