Types of Variables

In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken.

- Identify the main research question of the study.
- Who are the subjects in this study, and how many are included?
- The study found that students who were identified as upper-class took more candy than others. How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

Solution

- The research question is "Is there a difference between the unethical behaviors of people who identify themselves as having low and high social-class rank?"
- The cases are 129 University of California at Berkeley undergraduates.
- Two variables: (1) social-class rank (categorical) and (2) number of candies taken (numerical, discrete).

Random Sampling

Suppose we want to estimate household size, where a "household" is defined as people living together in the same dwelling, and sharing living accommodations. If we select students at random at an elementary school and ask them what their family size is, will this be a good measure of household size? Or will our average be biased? If so, will it overestimate or underestimate the true value?

Solution

Yes, the estimate will be biased, and it will tend to overestimate the true family size. Notice that families without children cannot be sampled using the sampling strategy. Additionally, if a family has two children, it is twice as likely to be sampled than a family with one child since there are twice as many children included in the sample (on average).

Summary Statistics

Facebook data indicate that 50% of Facebook users have 100 or more friends, and that the average friend count of users is 190. What do these findings suggest about the shape of the distribution of number of friends of Facebook users?

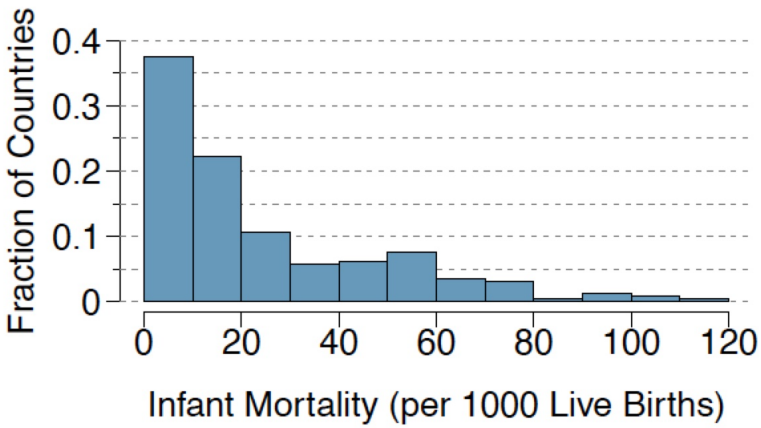
Solution

The statement 50% of Facebook users have over 100 friends" means that the median number of friends is 100, which is lower than the mean number of friends (190), which suggests a right skewed distribution for the number of friends of Facebook users.

Summary Statistics

The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014.

1. Estimate Q1, the median, and Q3 from the histogram.
2. Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning.



Solution

First we eyeball the heights of the bars in the relative frequency histogram (as shown on the right). Remember that these relative frequencies must add up to 1. Then, find out which bin the 25th, 50th and 75th percentiles fall in and estimate Q1, median (Q2), and Q3 as the midpoint of those bins, respectively.

- $Q1 \approx 5$
- $Q2 \approx 15$
- $Q3 \approx 35$

Since the distribution is right skewed, we would expect the mean to be higher than the median.

Conditional Probability

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table displays the distribution of health status of respondents to this survey (excellent, very good, good, fair, poor) and whether or not they have health insurance.

		Health Status				
		Excellent	Very good	Good	Fair	Poor
Health Coverage	No	0.0230	0.0364	0.0427	0.0192	0.0050
	Yes	0.2099	0.3123	0.2410	0.0817	0.0289

1. Are being in excellent health and having health coverage mutually exclusive?
2. What is the probability that a randomly chosen individual has excellent health?
3. What is the probability that a randomly chosen individual has excellent health given that he has health coverage?
4. What is the probability that a randomly chosen individual has excellent health given that he doesn't have health coverage?
5. Do having excellent health and having health coverage appear to be independent?

Solution

1. No, there are individuals who are both excellent in health and have health coverage.
2. $\mathrm{P}(\text{ excellent health }) = 0.2329$
3. $\mathrm{P}(\text{excellent health} \mid \text{health coverage}) = 0.2099/0.8738 = 0.24$
4. $\mathrm{P}(\text{excellent health} \mid \text{no health coverage}) = 0.0230/0.1262 = 0.18$
5. No, because the probability that a person has excellent health varies between the two health coverage categories (24% vs 18%). That is, knowing something about someone's health coverage provides useful information in predicting whether the person has excellent health, which means the variables are not independent.

Bayes Rule

A genetic test is used to determine if people have a predisposition for thrombosis, which is the formation of a blood clot inside a blood vessel that obstructs the flow of blood through the circulatory system. It is believed that 3% of people actually have this predisposition. The genetic test is 99% accurate if a person actually has the predisposition, meaning that the probability of a positive test result when a person actually has the predisposition is 0.99. The test is 98% accurate if a person does not have the predisposition. What is the probability that a randomly selected person who tests positive for the predisposition by the test actually has the predisposition?

Solution

$$P(\text{ pre } \mid \text{ positive }) = \frac{P(\text{ pre and positive })}{P(\text{ positive })} = \frac{0.0297}{0.0297 + 0.0194}$$
$$= 0.6049$$

Expectations, Variances

3.44 Scooping ice cream. Ice cream usually comes in 1.5 quart boxes (48 fluid ounces), and ice cream scoops hold about 2 ounces. However, there is some variability in the amount of ice cream in a box as well as the amount of ice cream scooped out. We represent the amount of ice cream in the box as X and the amount scooped out as Y. Suppose these random variables have the following means, standard deviations, and variances:

	mean	SD	variance
X	48	1	1
Y	2	0.25	0.0625

(a) An entire box of ice cream, plus 3 scoops from a second box is served at a party. How much ice cream do you expect to have been served at this party? What is the standard deviation of the amount of ice cream served?

(b) How much ice cream would you expect to be left in the box after scooping out one scoop of ice cream? That is, find the expected value of $X - Y$. What is the standard deviation of the amount left in the box?

(c) Using the context of this exercise, explain why we add variances when we subtract one random variable from another.

Solution

1. $E(X + Y_1 + Y_2 + Y_3) = E(X) + 3 * E(Y) = 48 + 3 \times 2 = 54$
   $V(X + Y_1 + Y_2 + Y_3) = V(X) + 3 \times V(Y) =$
   $1 + 3 \times 0.0625 = 1.1875$ $SD(X + Y_1 + Y_2 + Y_3) =$
   $\sqrt{V(X + Y_1 + Y_2 + Y_3)} = \sqrt{1.1875} \approx 1.09$
2. $E(X - Y) = E(X) - E(Y) = 48 - 2 = 46$
   $V(X - Y) = V(X) + V(Y) = 1 + 0.0625 = 1.0625$
   $SD(X - Y) = \sqrt{V(X - Y)} = \sqrt{1.0625} \approx 1.03$
3. Initially we do not know exactly how much ice cream is in the box. Then we scoop out an unknown amount. We should now be even more unsure about the amount of ice cream that is left in the box.

RVs and their Distributions

The American Community Survey estimates that 47.1% of women ages 15 years and over are married.

- We randomly select three women between these ages. What is the probability that the third woman selected is the only one who is married?
- What is the probability that all three randomly selected women are married?
- On average, how many women would you expect to sample before selecting a married woman? What is the standard deviation?
- If the proportion of married women was actually 30%, how many women would you expect to sample before selecting a married woman? What is the standard deviation?
- Based on your answers to parts (c) and (d), how does decreasing the probability of an event affect the mean and standard deviation of the wait time until success?

Solution

- Since we are asked for the probability of a certain number of trials until the first success we use a geometric distribution with p = 0:471. Let X be the trial at which the first married woman is selected. Then, $P($ first married woman is the 3^{rd} selected $) = (1 - 0.471) \times (1 - 0.471) \times 0.471 = 0.1318$.
- $\mathrm{P}(\,$ all three women are married $) = 0.471^3 = 0.1045$.
- Use the mean and standard deviation of a geometric distribution with p = 0.471: $\mu = \frac{1}{0.471} = 2.12, \quad \sigma = \sqrt{\frac{(1-p)}{p^2}} = \sqrt{\frac{(1-0.471)}{0.471^2}} = \sqrt{2.38} \approx 1.54$
- Use the mean and standard deviation of a geometric distribution with p = 0.30.
  $\mu = \frac{1}{0.30} = 3.33, \quad \sigma = \sqrt{\frac{(1-p)}{p^2}} = \sqrt{\frac{0.70}{0.30^2}} \approx 2.79$
- When p is smaller, i.e. the event is rarer, the expected number of trials before a success and the standard deviation are higher.

RVs and their distributions

Occasionally an airline will lose a bag. Suppose a small airline has found it can reasonably model the number of bags lost each weekday using a Poisson model with a mean of 2.2 bags. (a) What is the probability that the airline will lose no bags next Monday?
(b) What is the probability that the airline will lose less than 2 bags on next Monday?
(c) What is the probability that the airline will lose 0,1 , or 2 bags on Monday and Tuesday?

Solution

(a) $\frac{\lambda^k \times e^{-\lambda}}{k!} = \frac{2.2^0 \times e^{-2.2}}{0!} = 0.1108$

(b)

$$P(0) + P(1) + P(2) = \frac{2.2^0 \times e^{-2.2}}{0!} + \frac{2.2^1 \times e^{-2.2}}{1!} + \frac{2.2^2 \times e^{-2.2}}{2!}$$
$$= 0.1108 + 0.2438 + 0.2681$$
$$= 0.6227$$

CLT

Suppose that the proportion of defective items in a large manufactured lot is 0.1. Use the central limit theorem to determine the smallest sample of items that must be taken from the lot in order for the probability to be at least 0.99 that the proportion of defective items in the sample will be less than 0.13?

Plus

Exercises in class
Find MLEs
Normal Distributions
Distribution of the Sample Mean