STA101 MIDTERM - SELECTED PROBLEMS KEY

Summer I, 2021, Duke University

Part 1: True-or-False

4 (False)

If we set the significance level at $\alpha = 0.05$ and find a p-value of 0.04, then in our hypothesis testing framework, we would believe that we have enough evidence to reject the null hypothesis in favor of the alternative hypothesis. However, this doesn't actually make the null hypothesis false; it may be true, and in this case we would be committing a Type I error.

6 (False)

Recall that for a causal effect to be identified, we need to have a *randomized* experiment. In this case, the scientist is not randomizing patients between the hospitals; the patients already have chosen which hospital they are going to. This sort of self-selection prevents this from being a truly randomized experiment.

8 (False)

The calculation expression $\mathbb{P}(\text{three 3-pointers in a row}) = (0.429)^3$ is only correct if we know that the probability that Steph Curry makes a basket is independent from shot to shot. However, this is not given in the problem.

Part 2: Fill in the Blank

16 ($\approx 55.8\%$)

Let's start with some notation. We'll denote an individual who is HIV positive by HIVP, and a person who is HIV negative by HIVN. Similarly, if an individual takes a rapid test and gets a positive result, we'll denote this event by TP; if they get a negative result, we'll denote this even by TN.

Next, we record the information we're given using our notation:

- HIV Prevalence: $\mathbb{P}(HIVP) = 0.06$ and $\mathbb{P}(HIVN) = 0.94$
- Test Accuracy: $\mathbb{P}(TP \mid HIVP) = 0.99$, so $\mathbb{P}(TN \mid HIVP) = 0.01$. Similarly, $\mathbb{P}(TN \mid HIVN) = 0.95$, so $\mathbb{P}(TP \mid HIVN = 0.05)$

Finally, we use Bayes' rule:

$$\begin{aligned}\mathbb{P}(HIVP \mid TP) &= \frac{\mathbb{P}(TP \mid HIVP)\mathbb{P}(HIVP)}{\mathbb{P}(TP \mid HIVP)\mathbb{P}(HIVP) + \mathbb{P}(TP \mid HIVN)\mathbb{P}(HIVN)} \\ &= \frac{(0.99)(0.06)}{(0.99)(0.06) + (0.05)(0.94)} \\ &\approx 55.8\%\end{aligned}$$

Given a positive rapid test, the probability that an individual is actually HIV positive is about 55.8%.

Part 3: Multiple Choices

18 (1:a, 2:b, 3:c, 4:d)

Bimodal refers to a distribution with 2 clearly defined peaks in its histogram;this must be (a) since no other distribution has this property. Right-skewed refers to a distribution with more larger values than smaller values, especially if there are large outliers. This has to be (c) since no other distribution has this property. While it is possible to argue that (d) is multi-modal (though this isn't really clear from the plot), the description uniform is much more appropriate for (d) than for (b), so we assign plot (b) to multi-modal and plot (d) to uniform. To summarize:

- Bimodal: (a)
- Multi-modal: (b)
- Right-skewed: (c)
- Uniform: (d)

19 (A)

Bar plots (B) and histograms (C) are used to describe a single variable, so they would not be appropriate for describing the relationship between two variables. Dot plots (D) are used to described the relationship between to numeric variables, but species is categorical, so this would also not be appropriate. Box plots (A) are useful in visualizing the relationship between a categorical and numeric variable, so this is the answer.

20 (B)

To use the CLT for proportions, we need to check two conditions:

- Independence
- Success-Failure Condition

The first condition immediately makes (D) false. The CLT tells us which conditions we need for $\widehat{p}$ to be approximately normal when n is large, so answer (C) cannot be right because it doesn't also assume the above conditions hold. We check if there are enough success and failures via $np > 10$ *and* $n(1-p) > 10$; it's not enough for one of them to be larger than 10, so (A) is false. However, to check this condition, we need to estimate p (we don't know p after all), so we can instead check if $n\widehat{p} > 10$ and $n(1-\widehat{p}) > 10$, so (B) is true.

24 (C)

The political scientist is not necessarily interested in any causal relationship, so an observational study seems more appropriate. This leaves us with (A), (C), or (E), which differ in their sampling strategies. We know that the political scientist wants to have an even representation of countries among 5 continents, so a simple random sample (E) would not be appropriate since that does not take into account the continents in sampling. It seems stratified sampling (C) is most appropriate because we can break the population across continents as strata, and then perform simple random sampling within each continent. If we considered each continent a cluster (A), then we would select only some continents and ignore others, and that's not the desired strategy.