PHP2510

EXAM #1 WEEKS 1-6 ANSWER KEY

1. [1 point] Oncologists are studying a rare, slow-growing form of soft-tissue sarcoma. Once a tumor is identified (Stage I), a primary concern is the time it takes to progress to Stage II, which requires a more aggressive treatment plan. Identify the most reasonable distribution (with it’s caveat) below.

A. Normal, but the symmetry should be checked
B. Poisson, but the mean = variance property should be checked
C. Binomial, but we need to know n (number of patients in the study)
D. Exponential, but the memoryless property should be checked
E. Geometric, but patients must visit the doctor at regular intervals

Exponential is the most natural choice for a time-to-event (waiting) outcome because it is continuous with support of $x \geq 0$*. We mentioned this often in lectures.*

2. [1 point] When is the Normal distribution an accurate approximation of the Binomial distribution?

A. Per the central limit theorem, for sufficiently large n
B. For sufficiently large values of n*p
C. For sufficiently large values of n*p and n*(1-p)
D. When the Binomial is a sum of iid Bernoulli random variables
E. Always

If p is very near 0 or 1, the symmetry of a Normal isn't reasonable (there is no mass to one side), even for somewhat large n. For example, at n=1000 and p=0.001, the binomial has pretty much all its mass at 0,1,2, and 3, so it’s not smooth nor Normal. See CHIHARA page 88 for more details.

3. [1 point] Explain in words the difference between a Normal distribution and a t distribution. **The t-distribution is similarly shaped, but has fatter tails.**

See SPEEGLE, page 141.

4. [1 point] You have a dataset named `vaccine_data` with information about the vaccination status of various patients, including columns for `vaccine_name`, `patient_age`, and `country`. You want to calculate the average age of patients for each `vaccine_name` and store the result in a new data frame. Which of the following code snippets would correctly accomplish this task?

A. vaccine_data %>% group_by(vaccine_name) %>% summarize(avg_age = mean(patient_age, na.rm = TRUE))
B. vaccine_data %>% count(vaccine_name, avg_age = mean(patient_age, na.rm = TRUE))

C. vaccine_data %>% filter(vaccine_name) %>% mutate(avg_age = mean(patient_age, na.rm = TRUE))`
D. vaccine_data %>% summarize(avg_age = mean(patient_age, na.rm = TRUE)) %>% group_by(vaccine_name)

The group_by statement is needed and it needs to come before summarizing. We learned this with the tidyverse handout.

5. [1 point] How would you amend the code above to instead calculate the average age per vaccine_name-country pair? **Add country into the group_by statement**

We learned this with the tidyverse handout.

6. [1 point] **True** or False: the central limit theorem result happens faster (i.e. with fewer samples) if the population distribution is symmetric and bell-shaped.

The CLT isn't needed at all to have Normality of the sample mean if the population is Normal itself (mixtures of independent Normals are Normal themselves). As we get further from that distribution in the population, we need larger n for the sample mean to converge. We discussed this in lecture.

7. [1 point]. A test for a rare condition with 1% prevalence has a sensitivity of 99%. What specificity is required so that the positive predictive value is at least 50% (or greater)?

A. 50%
B. 90%
C. 99%
D. 99.01%
E. Not enough information is given to calculate

This can be solved using standard probability formulae.
PPV is defined as: P(D=+|T=+) = P(D=+, T=+)/P(T=+)
Numerator: P(D=+, T=+) = P(T=+|D=+) P(D=+) = 0.99*0.01*
Denominator: P(T=+) = P(T=+, D=+) + P(T=+, D=-) = P(T=+|D=+) P(D=+) + P(T=+|D=-)* P(D=-) = 0.99*0.01 + (1-sens)(1-0.01)*
Plugging this into the PPV formula, and solving for sensitivity when PPV=0.5 (or greater), we get sens = 0.99 (or greater)

8. [5 points] You want to know when patients are likely to start to show symptoms after being exposed to a novel infection disease. When little is known about that the incubation period (the time between exposure and showing symptoms), scientists may be willing to assume that it behaves like a U[0, θ] distribution. You want to know θ ("theta") so that you can advise how long patients should be quarantined. You interview 40 people (who have all been exposed to a known carrier on a known date, and then eventually exhibited symptoms) and collect how long it took each to show symptoms. You have two ideas for how to estimate θ:

A. 2 times the sample average
B. The largest value observed

Compare these two estimators in terms of biasedness (2 points each). Explain your findings (1 point).

Extra Credit [+2 points] Which estimator has a smaller variance?

For this problem, use the following facts:

- X ~ U[a,b] random variable has the following properties:
    - pdf f(x) = 1 / (b-a)
    - cdf F(x) = (x-a) / (b-a)
    - E[X] = (a+b)/2
    - $V[X] = (b-a)^2/12$.
- For any estimator of θ, say Y, the bias is defined as: E[Y] - θ
- Generally, f(x) = F′(x) dx (the pdf is the derivative of the cdf). Use this fact when trying to calculate the expected value of the part B estimator (i.e. in order to figure out the pdf, figure out the cdf first)

X_i~iid U[0, θ], i = 1, …, 40

For (A):

$Y = 2\bar{X}$, where $\bar{X} = \sum_{i=1}^{40} X_i \,/\, 40$

$E[2\bar{X}] = 2E[\bar{X}] = 2 * \sum_{i=1}^{40} E[X_i] \,/\, 40 = 2 * (\theta/2) = \theta$

option A is unbiased (by the linearity of expectations)

For (B):

Y = max(X_i); To calculate the expectation, we need the pdf; Per the hint, we start with the cdf:

P(Y < y) = P(max(X_i) < y) = P(X_1 < y, X_2 < y, …, X_40 < y) = P(X_i < y)^40 = F(y)^40 = (y/θ)^40

Then taking the derivative, pdf $f(y) = (40/\theta^{40})*y^{39}$.

$E[Y] = \int_0^\theta y * f(y)\,dy = \int_0^\theta 40 * y^{40}/\theta^{40}\,dy = 40/\theta^{40} * \int_0^\theta y^{40}dy = 40y^{41}/(\theta^{40} * 41)\,|_0^\theta =$ 40*θ/41

So Y is biased (it is too small by a factor of 1/41)

A is unbiased because it is the correct linear function of the sample mean to recover theta in expectation (via linearity of expectations). B is biased because we never see the true theoretical maximum and what we observe is always too small (unlike the average, we never “even” out)

EXTRA CREDIT:

A: $V[2*\bar{X}] = 4*V[\bar{X}] = 4*V[X]/40 = (1/10) * (\theta^2/12) = \theta^2/120$

B: For V(Y), we can use E[Y^2] - E^2[Y], where the second term is known from the bias estimation. To get the first term, we calculate $\int_0^{\theta} y^2 * f(y)\, dy$ *and get (40/42) * (θ^2), so the entire expression is* θ^2 ** (40/42 - $(40/41)^2$)*

Additional Notes of Interest:

- *(A) is called the "method of moments" estimator – we align the sample mean with the theoretical expectation*
- *Using a calculator, it is clear that (B) has smaller variance than (A)*
- *An adjustment of (B) to remove the bias – namely (n+1)/n * max(x_i) – is the uniform minimum variance unbiased estimator in the general case of n samples, and therefore the truly preferred option*
- *This problem shows that while* $\overline{X}$ *generally has useful properties, there are settings where it is not the preferred estimator*