Stat 230

Homework Problem Set 3

Due May 29 12:30 pm EST.

Please justify your answers and show the steps that lead you to your answer. Without a proper explanation, even a perfectly correct answer will receive a low score. Doubts regarding the problem sets/ Problem setup can be posted on Piazza. Simplify the expressions whenever you can.

Pr. 1 Suppose that 75 percent of the people in a certain (large) metropolitan area live in the city and 25 percent of the people live in the suburbs. If 1200 people attending a certain concert represent a random sample from the metropolitan area, what is the probability that the number of people from the suburbs attending the concert will be fewer than 270?

Pr. 2 As a result of comprehensive testing, drug A is known to be effective to 80% of the cases in which it is used. A new drug B arrives, and it helps 85 patients in a sample of 100. Is the new drug B more effective than the known one (A)?

To make progress, assume the drugs are equally efficient, and estimate the probability of seeing at least 85 cured cases. If the probability is very small, it would be unlikely that the known medicine could get such good outcomes in the same setup.

Remark 0.1. *This type of thinking is quite common is statistics.*

Pr. 3 There are 2,000 insects in the garden. An insecticide that claims it kills 99% of the insects when applied. The insecticide is applied to the garden.

a) What is the probability distribution for the number of surviving insects?

b) Find an expression for the probability that fewer than 30 insects survive?

c) Find a numerical value to the probability from b) using an appropriate approximation.

Pr. 4 A manufacturer receives a lot of 100 parts from a vendor. The lot will be unacceptable if more than 5 parts are defective. The manufacturer decides to randomly sample the 100 lots by opening k boxes, and accept if they find no defective parts.

a) What is the distribution of the defective parts in the sample that the manufacturer picks, if we are given there are R defective parts in the lot? Provide its name with parameters or a pmf.

b) How many boxes does the manufacturer need to open to ensure that the probability that they accept an unacceptable lot is less than 0.10?

c) How does the answer in a) change if the manufacturer alters their strategy so that they decide to accept if they see at most 1 defective part (instead of 0)?

Remark 0.2. *The earlier version of this problem b) asked for the probability of them not accepting an unaccetable lot. That was a mistake, the problem now reads correctly.*

Pr. 5 [Confidence intervals and Power analysis]

A biologist wants to study the prevalence of disease D in a large fish population S by the means of a catch and drop survey. Let n be the number of fish the biologists catches, and let p the (unknown) proportion of fish that carry the disease. All the fish are equally likely to be caught, regardless of if they have the disease or not. The biologist can tell just by looking at the fish if they have the disease or not.

a) What is the distribution of the number of the disease carrying fish that the biologist catches?

b) After catching a lot of fish, the biologists can estimate the true proportion p via

$$\hat{p} := \frac{\#\text{number of caught fish that have D}}{\#\text{total number of fish caught}}.$$

In terms of c, n, p and Φ, what is the (approximate) probability P that $\hat{p}$ deviates from p at most c? The associated interval

$$-c \leq \hat{p} - p \leq c$$

is called the P-*confidence interval* of $\hat{p}$. With such an interval the biologist is $100 \cdot P$ percent confident that p is within c distance of $\hat{p}$.

Hint: Formulate the confidence interval in terms of normal approximation to the number of diseased fish caught. To make the expression simpler, you can ignore the continuation correction.

c) How many fish does the biologist need to catch to be guaranteed to be at least 99% confident that the true proportion p of the disease carrying fish is $\hat{p} \pm 0.01$?

Hint: Use the fact that binomial distribution has greatest variance when $p = 0.5$.