DSE 210: Probability and Statistics using Python

Worksheet 12 — Sampling

1. A box contains 9 red marbles and 1 blue marbles. Nine hundred random draws are made from this box, with replacement. What is distribution of the number of red marbles seen, roughly?

2. Suppose that in the world at large, 1% of people are left-handed. A sample of 200 people is chosen at random. Give a 99% confidence interval for the number of them that are left-handed.

3. A dartboard is partitioned into 20 wedges of equal size, numbered 1 through 20. Half the wedges are painted red, and the other half are painted black. Suppose 100 darts are thrown at the board, and land at uniformly random locations on it.

   (a) Let X_i be the number of darts that fall in wedge i. What are $\mathbb{E}(X_i)$ and $\text{var}(X_i)$?

   (b) Using a normal approximation, give an upper bound on X_i that holds with 95% confidence.

   Let Z_r be the number of darts that fall on red wedges, let Z_b be the number of darts that fall on black wedges, and let $Z = |Z_r - Z_b|$ be the absolute value of their difference. We would like to get a 99% confidence interval for Z. To do this, define

   $$Y_i = \begin{cases} 1 & \text{if } i\text{th dart falls in red region} \\ -1 & \text{if } i\text{th dart falls in black region} \end{cases}$$

   and notice that $Z_r - Z_b$ can be written as $Y_1 + Y_2 + \cdots + Y_{100}$, the sum of independent random variables.

   (c) What are $\mathbb{E}(Y_i)$ and $\text{var}(Y_i)$?

   (d) Using the central limit theorem, we can assert that $Z_r - Z_b$ is approximately a normal distribution. What are the parameters of this distribution?

   (e) Give a 99% confidence interval for Z.

4. Suppose colorblindness appears in 1% of people. How large must a sample be in order for the probability of it containing at least one colorblind person to be at least 95%?

5. You have hired a polling agency to determine what fraction of San Diegans like sushi. Unknown to the agency, the actual fraction is exactly 0.5.

   The agency is going to poll a random subset of the population and return the observed fraction of sushi-lovers. How far off would you expect their estimate to be (i.e. what standard deviation) if:

   (a) they poll 100 people?

   (b) they poll 2500 people?

6. A sample is taken to find the fraction of females in a certain population. Find a sample size so that this fraction is estimated within 0.01 with confidence at least 99%.

7. A survey organization wants to take a simple random sample in order to estimate the percentage of people who have seen Downton Abbey. To keep the costs down, they want to take as small a sample as possible. But their client will only tolerate chance errors of 1% or so in the estimate. Should they use a sample of size 100, or 2500, or 10000? An auxiliary source of information suggests the population percentage will be in the range 20% to 40%.

8. In a certain city, there are 100,000 people age 18 to 24. A random sample of 500 of these people is drawn, of whom 194 turn out to be currently enrolled in college. Estimate the percentage of all persons age 18 to 24 in the city who are enrolled in college. Give a 95.5% confidence interval for your estimate.

9. A survey research company uses random sampling to estimate the fraction of residents of Austin, Texas, who watch Spanish-language television. They are satisfied with the estimate they get using a sample size of 1,000 people.

   They then want to also estimate this fraction for Dallas, which has similar demographics to Austin, but twice the population. What sample size would be suitable for Dallas?

10. The National Assessment of Educational Progress tests nationwide samples of 17-year olds in school. In 1992, the students in a random sample of size 1000 averaged 307 on the math component of the test; the standard deviation of the scores was about 30.

    Estimate the nationwide average score on the math test. What is the standard deviation of this estimate?

11. A box contains many pieces of papers with numbers on them. 100 random draws are made from the box, with replacement, and the sum of the draws is 297.

    (a) Can you estimate the average of the numbers in the box?

    (b) Can you give a confidence interval for your estimate, based on the information so far?

12. A lake contains an unknown number of fish. 1000 of them are caught, marked with red spots, and then returned to the water. Later, a random subset of 100 fish are caught from the lake, and it is found that Z of them have red spots.

    (a) How would you estimate the number of fish in the lake, in terms of Z?

    (b) Let F be the true number of fish in the lake. In terms of F, give a normal approximation to the distribution of Z. with what parameters?

    (c) If you had to give a 95% confidence interval for the number of fish in the lake, what would it be?