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?

Solution:

$\mu = p = 0.9$
$\sigma^2 = p(1-p) = 0.09$
Distribution of red marbles $N(n\mu, n\sigma^2) = N(810, 81)$

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

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

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 and notice that $Z_r - Z_b$ can be written as $Y_1 + Y_2 + \ + Y_{100}$, the sum of independent random variables.
(c) What are $E(Y_i)$ and $var(Y_i)$?
(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.

Solution:

(a) Let X_i be the number of darts in wedge i.
$E[X_i] = np = 5$
$Var(X_i) = np(1-p) = 4.75$

(b) X_i is approximated by $N(5, 4.75)$
95% confidence interval: $[5 - 2\sqrt{4.75}, 5 + 2\sqrt{4.75}] = [0.6, 9.4]$, hence upper bound is 9.4.

(c)$E[Y_i] = 0$
$Var(Y_i) = 1$

(d)$Z_r - Z_b = Y_1 + Y_2 + \ + Y_{100}$
$E[Z_r - Z_b] = E[Y_1] + E[Y_2] + ... + E[Y_{100}] = 0$
$Var(Z_r - Z_b) = Var(Y_1 + Y_2 + ... + Y_{100}) = Var(Y_1) + Var(Y_2) + ... + Var(Y_{100}) = 100$
Hence $Z_r - Z_b$ is approximated by $N(0, 100)$

(e)99% Confidence interval for $Z_r - Z_b$ is [-30, 30]
Hence 99% Confidence interval for $Z = |Z_r - Z_b|$ is [0, 30]

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?

Solution:

(a)$\mu = 0.5$
$\sigma^2 = 0.5 * 0.5$
$\hat{\sigma} = \sigma/\sqrt{100} = 0.05$

(b)$\mu = 0.5$
$\sigma^2 = 0.5 * 0.5$
$\hat{\sigma} = \sigma/\sqrt{2500} = 0.01$

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%

Solution:

When n=100, $\mu = 0.2$
within 99% confidence interval, $\hat{\mu}$ is in $[0.097, 0.303]$
When n=100, $\mu = 0.4$
within 99% confidence interval, $\hat{\mu}$ is in $[0.274, 0.526]$

When n=2500, $\mu = 0.2$
within 99% confidence interval, $\hat{\mu}$ is in $[0.179, 0.221]$
When n=2500, $\mu = 0.4$
within 99% confidence interval, $\hat{\mu}$ is in $[0.375, 0.425]$

When n=10000, $\mu = 0.2$
within 99% confidence interval, $\hat{\mu}$ is in $[0.189, 0.210]$
When n=10000, $\mu = 0.4$
within 99% confidence interval, $\hat{\mu}$ is in $[0.387, 0.413]$

Clearly, when sample size is 10000, the estimated sample mean is the most close to 19% to 21% or 39% to 41%.

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.

Solution:

$\mu = p = 194/500 = 0.388$
$\sigma^2 = p(1-p)$
$N(\mu, \sigma/\sqrt{500})$
For a 95% confidence interval, the offset should within 2 times of standard deviation, hence
$[\mu - 2 * \hat{\sigma}, \mu + 2 * \hat{\sigma}] = [0.34442, 0.43158]$

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?

Solution:

The sample mean is approximately normal distributed, the mean and variance of this normal distribution can be computed in terms of the population mean and variance, and the size of sample, the population size doesn't matter.

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?

Solution:

Due to law of large number, nationwide average score can be estimated by sample mean, which is 307.
$\sigma = 30$
$\hat{\sigma} = \sigma/\sqrt{1000} = 0.9487$