Notes of Ch6: Sampling, ”introductory Statistics” by Wonnacott

Mahdi Sadeghzadeh Ghamsary

June 2019

1 Introduction

Main Question: what can we expect of a random sample from a known population?
previous examples were sampling without replacement. now assume we want too choose a number of a 80 million numbers with known population. the random variable from first selection is designed by X_1 and and other selection designed as well, then we have:

$$p(X_1) \equiv p(X_2) \equiv ... \equiv p(X_n) \tag{1}$$

this equality hold true if we sample with replacement, since second selection have same probability with first one. fortunately, this equation holds true for sampling without replacement.
we already know the distribution of X_1 is same as the distribution of population. However, if we sample without replacement, the X_2 is depend on X_1 and the population changes along with relative probabilities. as the conditional distribution of X_2 is depend on X_1, so if we know $p(X_1)$, then we have $p(X_2|X_1)$. but if we have no knowledge about X_1, there is no reason to differ distribution of X_2 from the distribution on X_1. in other word, if we have no knowledge of the first selection, the consideration about distribution of second selection doesn't change.
if the population is infinite, we have:

$$p(X_2|X_1) \approx p(X_2) \tag{2}$$

so we can say the last equation holds true for sampling without replacement in a large sample set. depending on last equation, the random variable X_2 is independent of the X_1 if the population is infinite.
so the sample of n numbers gives us n random variables $X_1, X_2, ..., X_n$, Each X_i has the same (marginal) distribution of population X.
the equation 1 holds in all cases, regardless the population size. but the independence of $X_1, X_2, ..., X_n$ is more complex issue. if the population is finite and the sampling is without replacement, the X_i is dependent, but for simplicity we shall they are independent in the rest of book.

2 Sample SUM

we define S, as sum of sample observations:

$$S = X_1 + X_2 + ... + X_n$$

so the expected value of S is:

$$E(X) = E(X_1) + E(X_2) + ... + E(X_n)$$

as mentioned in equation 1, the distribution of all variables are same, and this distribution is same as the population distribution (μ) too. can therefor be written:

$$\begin{aligned} E(S) &= \mu + \mu + ... + \mu \\ &= n\mu \end{aligned}$$

$$\Rightarrow \mu_S = n\mu \tag{3}$$

in the same way the variance of S can be calculated:

$$varS = var(X_1 + X_2 + ... + X_n)$$

we have :

$$var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$$

so for independent X and Y:

$$var(X + Y) = var(X) + var(Y)$$

now if $X_1, X_2, ..., X_n$ assumed independent variables, we have:

$$\begin{aligned} varS &= varX_1 + varX_2 + ... + varX_n \\ &= \sigma^2 + \sigma^2 + ... + \sigma^2 \end{aligned}$$

$$\Rightarrow varS = n\sigma^2 \tag{4}$$

3 Sample Mean

$$\begin{aligned} \bar{X} &= \frac{1}{n}(X_1 + X_2 + ... + X_n) \\ &= \frac{1}{n}S \end{aligned}$$

it's important to remember that $\bar{X}$,as well as S, is a random variable that fluctuate from sample to sample. it seems intuitively clear that $\bar{X}$ will fluctuate about the same central value as an individual observation, but with less deviation because of "averaging out":

$$\begin{aligned}\bar{X} &= \frac{1}{n}\mu_S \\ &= \frac{1}{n}(n\mu) \\ \Rightarrow \mu_{\bar{X}} &= \mu \qquad (5)\end{aligned}$$

and in a same way for variance:

$$\begin{aligned}\sigma^2_{\bar{X}} &= (\frac{1}{n})^2\sigma^2_S \\ &= \frac{1}{n^2}(n\sigma^2) \\ \Rightarrow \sigma^2_{\bar{X}} &= \frac{\sigma^2}{n} \qquad (6)\end{aligned}$$

4 the central limit theorem

checking the sample distribution($\bar{X}$):
a. when the population distribution is normal
Theorem. if X and Y are normal, then any linear combination $Z = aX + bY$ is also a normal random variable. so $\bar{X}$ is exactly normal.
b. when the population of distribution is not normal
the sample mean becomes normally distributed as n grows, no matter the parent population is.
The Central Limit Theorem. As the sample size n increases, the distribution of mean, $\bar{X}$, of a sample taken from practically any population approaches a normal distribution, (with mean μ and variance $\frac{\sigma^2}{n}$).

5 sampling from finite population, without replacement

we have already argued in the first section that all X_i in a sample of n observation $X_1, X_2, ..., X_n$ will have same marginal distribution whether or not we replace, so equation 3 still true. but the variance of sample mean depends on replacement. formally, if we sampling without replacement, the $X_1, X_2, ..., X_n$ are not independent, so we must modify equation 4:

$$var\bar{X} = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}(\frac{N-n}{N-1}) \qquad (7)$$

6 Bernoulli population

this is the simplest kind of probability distribution, lumped at only two values, 0 and 1. if we design the $p(X = 1)$ as π, then we have:

$$\begin{aligned} \mu &= 0 \times (1 - \pi) + 1 \times \pi \\ &= \pi \end{aligned}$$

and for variance:

$$\begin{aligned} \sigma^2 &= (0 - \pi)^2(1 - \pi) + (1 - \pi)^2(\pi) \\ &= \pi(1 - \pi) \end{aligned}$$

7 questions

1. why the distribution of random variables are same as μ?
2. how the equation 1 holds true for all cases but equation 2 doesn't?
3. what does the $\sigma^2_{\bar{X}}$ mean? and if it means the mean of variances of all possible samples, why it divided by n?