Section 3

Foundations of Statistical Inference (PLSC 503)

7 Feb. 2019

In this problem set we will investigate random sampling. We will work with the same population in all three exercises. The population consists of 100 people, and we are interested in their age distribution. A data set with the ages in the population is on Canvas (see `pset2-data.RData`).

Exercise 1

1. Let X_p be a random variable describing the age distribution in the population. What is the PMF of X_p? Hint: use `table(age_pop)` in `R`.
2. X_p is a random variable, but is it randomly generated? Briefly motivate your answer.
3. You sample four units from the population at random with equal probability *with* replacement. Let X_1, X_2, X_3 and X_4 denote the ages of the four sampled units. Answer the following questions analytically (i.e., with pen and paper).
   (a) What is the marginal distribuiton (PMF) of X_1?
   (b) What is the expectation of X_1?
   (c) What is the variance of X_1?
   (d) What is the marginal distribution of X_2?
   (e) What is the joint distribution of X_3 and X_4?
   (f) Are X_1, X_2, X_3 and X_4 IID?
4. Consider defining a new random variable as $X_m = (X_1 + X_2 + X_3 + X_4)/4$. Answer the following questions analytically (i.e., with pen and paper).
   (a) Is X_m a sample statistic? Briefly motivate your answer.
   (b) X_m is a natural estimator for some aspect of the population distribution. Which aspect?
   (c) What is the expectation of X_m?
   (d) hat is the variance of X_m?
5. True or false?
   (a) X_p and X_1 have the same marginal distributions.
   (b) X_p and X_m have the same marginal distributions.
   (c) X_m and X_1 have the same marginal distributions.

Solution:

1.

```r
# Read in dataset
load("pset2-data.RData")

# Tabulate
table(age_pop)
```

```
## age_pop
## 55 56 57 58 59
## 11 46 33  7  3
```

We could write it like this,

$$f_{X_p}(X_p = x) = \begin{cases} 0.11 & : x = 55 \\ 0.46 & : x = 56 \\ 0.33 & : x = 57 \\ 0.07 & : x = 58 \\ 0.03 & : x = 59 \\ 0 & : o.w. \end{cases}$$

2. No. X_p describes the age distribution for some finite population of 100 units. This population of units is fixed, not randomly generated.

3a. X_1 is a single draw from X_p. Thus, the PMF is the same as X_p.

3b. $\mathbb{E}[X_1] = \sum_x xf(x) = 55 \cdot 0.11 + 56 \cdot 0.46 + 57 \cdot 0.33 + 58 \cdot 0.07 + 59 \cdot 0.03 = 56.45$

3c.

$$\begin{aligned} \text{Var}[X_1] &= \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 \\ &= \left[(55)^2 \cdot 0.11 + (56)^2 \cdot 0.46 + (57)^2 \cdot 0.33 + (58))^2 \cdot 0.07 + (59)^2 \cdot 0.03\right] - [56.45]^2 \\ &= 3187.39 - 3186.603 = 0.787 \end{aligned}$$

3d. This is the same as X_p and X_1 since we are sampling with replacement. We can use the following notation $X_p \stackrel{d}{=} X_2$ to say "X_p and X_1 have the same distribution".

3e. Sampling with replacement implies IID, so $g_{X_3,X_4}(X_3 = x, X_4 = y) = f_{X_3}(X_3 = x)f_{X_4}(X_4 = y) = f_{X_p}(X_p = x)f_{X_p}(X_p = y)$.

3f. Yes, sampling with replacement implies IID.

4a. Yes. X_m is a function of the sample data, i.e. $X_m(X_1, \ldots, X_4)$, so it is a "sample statistic".

4b. X_m is an estimator for the population mean. You could justify this by referencing the "plug-in principle" as described in Aronow & Miller.

4c. Let's do the generic version for N (set $N = 4$ for 4 draws).

$$\begin{aligned} \mathbb{E}[X_m] &= \mathbb{E}\left[\frac{X_1 + \cdots + X_N}{N}\right] \\ &= \frac{1}{N}\left(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_N]\right) \\ &= \frac{N\mathbb{E}[X_p]}{N} \\ &= \mathbb{E}[X_p] \end{aligned}$$

where the 2nd and 3rd lines follow from the IID assumption.

4d. Again I'm solving for generic N.

$$
\begin{aligned}
\mathrm{Var}\left[X_m\right] &= \mathrm{Var}\left[\frac{X_1 + \cdots + X_N}{N}\right] \\
&= \frac{1}{N^2}\left(\mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_N]\right) \\
&= \frac{N}{N^2}\mathrm{Var}[X_p] \\
&= \frac{\mathrm{Var}[X_p]}{N}
\end{aligned}
$$

Where the 3rd line follows from the IID assumption.

5a. True. We already showed this in 3(a).

5b. False. X_m depends on X_p, but they do not have the same marginal distribution.

5c. False. Same logic as above; also note that if this were true it would yield a contradiction given the answers of 5a and 5b.

Exercise 2

Exercise 3