Term Test 1: Practice Problems : Solutions

Problem 1. We consider a population of N units with bernoulli random variables $\{0, 1\}$ data values. Suppose we choose a simple random sample without replacement of n units from this population. Let $p = \frac{1}{N}\sum_{i=1}^{N} y_i$ be the population proportion, and $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - p)^2$ denotes the population variance. Let $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{p})^2$ be the sample proportion and the sample variance, respectively.

(a) Is $\widehat{p}$ unbiased estimator fo p? Justify.

$$E(\hat{p}) = E\left(\frac{1}{n}\sum y_i\right) = \frac{1}{n}\sum E(y_i) = \frac{1}{n}\sum p = \frac{1}{n}(np) = p$$

Yes $\hat{p}$ is unbiased for p.

(b) Show that σ^2 can be writen as $\sigma^2 = p(1-p)$.

$$\sigma^2 = \frac{1}{N}\left[\sum y_i^2 - 2p\sum y_i + \sum p^2\right] = \frac{1}{N}\left[\sum y_i - 2p\sum y_i + Np^2\right]$$
$$= \frac{1}{N}\left[Np - 2pNp + Np^2\right] = \frac{1}{N}N(p - 2p^2 + p^2) = p(1-p)$$

(c) Show that s^2 can be writen as $s^2 = \frac{n}{n-1}\widehat{p}(1-\widehat{p})$.

$$s^2 = \frac{1}{n-1}\left[\sum y_i^2 - 2\hat{p}\sum y_i + \sum \hat{p}^2\right] = \frac{1}{n-1}\left[n\hat{p} - 2n\hat{p}^2 + n\hat{p}^2\right]$$
$$= \frac{1}{n-1}n(\hat{p} - \hat{p}^2) = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

(d) Is the sample variance s^2 unbiased estimator for σ^2? Find $E(s^2)$

Under SRS without replacement, we have ~~seen~~ ~~to~~ demonstrate in class that when estimating μ, the sample variance s^2 is biased Estimator for σ^2 and $E[s^2] = \frac{N}{N-1}\sigma^2$, which is also true when $\mu = p$

(e) Compute $V(\widehat{p})$. Given the result in (d), find an unbiased estimator for $V(\widehat{p})$ in term of $\widehat{p}$.

① from the estimation of μ, we have demonstrated that using SRS without Rep,

$$V(\bar{y}) = \frac{N-n}{N-1}\frac{\sigma^2}{n}. \text{ For } \mu = p, \quad V(\hat{p}) = \frac{N-n}{N-1}\frac{\sigma^2}{n}.$$

② Based on (d), an unbiased estimator for σ^2 is $\frac{N-1}{N}s^2$

thus, unbiased Estimator for $V(\hat{p})$ is $\frac{N-n}{N-1}\cdot\frac{N-1}{N}\frac{s^2}{n} = \frac{N-n}{N}\frac{s^2}{n}$ ⊛

Using $s^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$, ⇒ ⊛ become $\hat{V}(\hat{p}) = \frac{N-n}{N}\frac{1}{n}\frac{n}{n-1}\hat{p}(1-\hat{p})$

Estimat of $V(\hat{p})$ ⇒ $\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}$

Problem 2. To estimate the proportion of voters in favor of a controversial proposition, a simple random sample of XXXXX eligible voters was contacted and questioned. Of these, 552 reported that they favored the proposition. The study also reports a margin of error $\pm$ 3%, 19 out of 20. The number of eligible voters in the population is approximately 1,800,000.

(a) What "a margin of error $\pm$ 3%, 19 out of 20" means ?. ⟹ means 95% Margin of Error is 0.03, e.g For at least 95% of all random samples, the Bound on error [the amount by which the sample proportion $\hat{p}$ is expected to differ from the true population proportion p] is 0.03.

(b) Complet the study by computing the missing sample size.

n for estimating p: we use $n = \dfrac{Npq}{(N-1)D + pq}$

p unknown, we use $p = 0.5$ to get conservative n.

$\Rightarrow n = \dfrac{1800000 \times 0.5 \times 0.5}{(1800000-1)\frac{0.03^2}{4} + 0.5^2} = 1110.4$

⟹ $n = 1111$ eligible voter should be contacted.

(c) Using (b), estimate the population proportion in favor.

$$\hat{p} = \frac{552}{1111} = 0.497$$

(d) Give a 95% confidence interval for population proportion.

a 95% C.I for p is $(\hat{p} - ME, \hat{p} + ME)$

$$(0.497 - 0.03, 0.497 + 0.03)$$

$$\boxed{(0.467, 0.527)}$$

Interpretation: we are 95% confidence that the true prop p lies between 0.467 and 0.527

Problem 3. Consider the following data from a simple random sample of size n = 4 from a population of size N = 250, in which y is the variable of interest and x is an auxiliary variable. The population mean of the x's is 3.9.

(a) Suggest two types of estimators for estimating the mean of y. Summarize some properties of each estimator.

Ratio $\mu_y = \frac{\bar{y}}{\bar{x}} \mu_x$ (biased)

simple: $\bar{y} = \frac{1}{n} \sum y_i$ (unbiased)

Regression $\mu_{reg} = \bar{y} + b(\mu_x - \bar{x})$ (biased)

(b) The data obtained were recorded in variables y and x, analyzed in R and produced results presented below (see next page). Based on the R output, answer the following questions.

From R output

(i) Estimate mean of y using the simple estimator, and estimate the variance of the estimator.

line 8 → $\hat{\mu} = \bar{y} = \frac{590}{4} = 147.5$

line 14 $\hat{V}(\hat{\mu}) = (1 - \frac{n}{N}) \frac{s_y^2}{n} = \boxed{416.501}$

(ii) Estimate mean of y using the ratio estimator, and estimate the variance of the estimator.

lines ~~8-12~~ $\hat{\mu}_r = 177 = \frac{\sum y}{\sum x} \cdot \mu_x = 177$

line 15 $\hat{V}(\hat{\mu}_r) = (1 - \frac{n}{N}) \frac{s_r^2}{n} = 117.71 = \boxed{117.71}$

(iii) Estimate mean of y using the regression estimator, and estimate the variance of the estimator.

line 24 $\hat{\mu}_{reg} = \bar{y} + b(\mu_x - \bar{x}) = 171.72$

line 26 $\hat{V}(\hat{\mu}_{reg}) = (1 - \frac{n}{N}) \frac{MSE}{n} = \boxed{154.3}$

(iv) Estimate mean of y using the difference estimator, and estimate the variance of the estimator.

lines 3-8-9 $\hat{\mu}_D = \bar{y} + (\mu_x - \bar{x}) = \frac{590}{4} + (3.9 - \frac{13}{4}) = 590.65$

$\hat{V}(\hat{\mu}_D) = (1 - \frac{n}{N}) \frac{1}{n} \frac{\sum (d_i - \bar{d})^2}{n-1}$ (not provided in the R output)

(v) Based on the data, which estimator appears preferable in this situation?

We would prefere Rotor estimator which provides small variance.

R output:

```
> # Population
> N<-250
3 > mu_x<-3.9
> # SRS of size n=4
> n<-4
> y<-c(150, 100, 200, 140)
> x<-c(4,2,4,3)
8 > ysum<-sum(y); ysum
[1] 590
9 > xsum<-sum(x); xsum
[1] 13
> s2_y<-var(y); s2_y
[1] 1691.667
> r<-(ysum/xsum); r
[1] 45.38462
12 > (ysum/xsum)*mu_x
[1] 177
> s2_r<-var(y-r*x); s2_r
[1] 478.501
14 > (1 - n/N)*(s2_y/n)
[1] 416.15
15 > (1 - n/N)*(s2_r/n)
[1] 117.7112
> (1 - n/N)*(1/mu_x)^2*(s2_r/n)
[1] 7.739069
> (1 - n/N)*(1/mu_x)*(s2_r/n)
[1] 30.18237
> cor(x,y)
[1] 0.8676399
> # regression of y on x
> fitreg<-lm(y~x)
> coef(fitreg)
(Intercept)           x
   26.36364    37.27273
> yhat<-fitted(fitreg)
> ehat<-y-yhat
24 > mean(y) + coef(fitreg)[2]*(mu_x-mean(x))
       x
171.7273
> MSE<-sum( ehat^2 )/(n-2)
26 > (1 - n/N)*(MSE/n)
[1] 154.3091
> (1 - n/N)*(MSE/(n-1))
[1] 205.7455
> (1 - n/N)*(MSE/(n-2))
[1] 308.6182
```