STAT 3014/ 3914

Semester 2	Applied Statistics	2016

Week 3

Tutorial

1. A survey is carried out at a university to estimate the percentage of undergraduates living at home during the current semester. What is the population? and what is the parameter ?

Solution:

The population consists of all undergraduate registered in the **current semester**. The parameter is the percentage of these undergraduates living at home.

2. A survey of motor vehicles is conducted by randomly selecting a list of registration numbers from the master lists maintained by the State Government. We want to know the total number of kilometers travelled by the vehicle as well as the total distance travelled in 2012.

a) How would you collect the data (phone interview, mail out questionnaire or personal interview) and why?

b) What are the potential sources of non-sampling error?

Solution:

Probability sampling is used to produce a simple random sample from the master list.

a) Interview would give the best data as the interviewer could prompt the respondent to check car records etc. Mail out questionnaires are cheaper but they tend to have a low response rate. Phone interviews would be possible if the telephone contacts were available.

b) Some non-sampling errors are problems with recall; the current list of vehicles will not include some registered in 2001; process ignores unregistered vehicles and those registered in other states and territories.

3. Polls often conduct pre-election surveys by telephone. Could this bias the results ? How ? What if the sample is drawn from the telephone book ?

Solution:
Doing a survey by telephone could potentially introduce bias, because telephone subscribers are probably different to non-subscribers. However the percentage of non-subscribers is so small that this bias can usually be ignored. (Optional) How many household in Australia have telephones or cell phones ? Using phone books would introduce serious bias due to unlisted (private) numbers.

4. According to Census data, in 1950 the population of the U.S amounted to 151.3 million persons, and 13.4% of them were living in the West. In 2000, the population was 281.4 million, and 22.5% of them were living in the West. Is the difference in percentage statistically significant ? Is is practically significant? Or do these questions make sense ? Explain why ?

Solution:
The concept of statistical significance does not apply very well, because the data are for the whole population, rather than a sample. The difference is practically significant. The center of population is shifting to the West, and that makes a lot of difference to the economy and to the political balance of the country.

5. [Taken from Q10 in Extra reading] A coin will be tossed 100 times. You get to pick 11 numbers. If the number of heads turn out to be equal to the 11 numbers, you win a dollar. Which 11 numbers should you pick, and what is your chance (approximately) of winning? Explain.

Solution:
The likeliest number of heads is 50: pick that first. Your next two picks should be 49 and 51. Then 48 and 52. And so forth. You should pick 45 through 55. And your chance of winning is about 73%. You can approximate with a standard normal or work it out using the computer (see below).

```
sum(sapply(45:55, dbinom, 100, 0.5))

## [1] 0.728747

pbinom(55, 100, 0.5) - pbinom(44, 100, 0.5)

## [1] 0.728747
```

6. One public opinion poll uses a simple random sample of size 1,500 drawn from a town with a population of 25,000. Another poll uses a simple random sample of size 1,500 from a town with a population of 250,000. The polls are trying to estimate the percentage of voters who favor single-payer health issurance. Other things being equal:

a) the first poll is likely to be quite a bit more accurate than the second.

b) the second poll is likely to be quite a bit more accurate than the first.

c) there is not likely to be much difference in accuracy between the two polls.

Solution:
(c)

7. Enumerate all samples of size 2 that can be drawn without replacement from the set $(0, 3, 3, 4, 5)$ (population list).

a) Calculate the population mean, μ, and variance, σ^2.

b) Calculate the sample mean for each sample and numerically show that the variance of the sample means is the same as

$$\frac{S^2}{n}(1 - \frac{n}{N})$$

Solutions:
Students are welcome to use **R** for this.

a) List the 10 possible samples:

$$(0; 3), (0; 3), (0; 4), (0; 5), (3; 3), (3; 4), (3; 5), (3; 4), (3; 5), (4; 5)$$

b) Calculate the average for each sample and show that the average of the averages is 3, the population mean. Also show that the variance of the sample means is

$$\frac{S^2}{n}(1 - \frac{n}{N}) = \frac{3.5}{2}(1 - \frac{2}{5}) = 1.05$$

```
Y = c(0, 3, 3, 4, 5)
var(Y)

## [1] 3.5

z = combn(Y, 2)
z

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    0    0    0    0    3    3    3    3    3     4
## [2,]    3    3    4    5    3    4    5    4    5     5
```

```r
apply(z, 2, mean)
```

```
##  [1] 1.5 1.5 2.0 2.5 3.0 3.5 4.0 3.5 4.0 4.5
```

```r
mean(apply(z, 2, mean))
```

```
## [1] 3
```

```r
var(apply(z, 2, mean)) * (9/10)
```

```
## [1] 1.05
```

```r
(3.5 / 2) * (1 - 2/5)
```

```
## [1] 1.05
```

8. In a survey to determine the average height, μ, of a particular population of size $N = 1000$, a sample of n individuals is drawn and the heights of the selected people averaged to estimate μ. Suppose the standard deviation of the heights in the population is $\sigma = 6$ cm.

a) If the sample is drawn with replacement what sample size is required to ensure that with probability at least 0.9 the sample average is within 0.5 cm of μ?

b) If the sample is drawn without replacement how does the answer to (a) change ?

c) If we draw a sample of 40 people without replacement approximate the probability that the average height is at least 1 cm below μ.

Solution:

$N = 1000$, $\sigma = 6$ We want to estimate μ.

a) Using sampling with replacement

$$\begin{aligned} \Pr(|\bar{y} - \mu| \leq 0.5) &\geq 0.9 \\ \Pr(|Z| \leq \frac{0.5\sqrt{n}}{6}) &\geq 0.9 \\ \frac{\sqrt{n}}{12} &\geq 1.645 \\ n &\geq 389.7 \end{aligned}$$

b) Using sampling without replacement

$$\bar{y} \sim N(\mu, \frac{\sigma^2}{n}(\frac{N-n}{n-1}))$$

$$\frac{0.5}{\frac{6}{\sqrt{n}}\sqrt{\frac{1000-n}{999}}} > 1.645$$

and so $n > 280.6$

c) If $n = 40$ then

$$\bar{y} \sim N(\mu, \frac{6 \times 960}{40 \times 999})$$

$$\Pr(\bar{y} \leq \mu - 1) = 0.141$$