Session 2: Practical Exercise

The purpose of this practical is to use computer simulation to see how the results of repeated sampling from a population conform with theory, in terms of the bias and variance of sample estimators.

Before you start these question, copy all the files in the directory **U:\Download\Teach\Foundations\Inf/Practical2** into a sensible directory of your choice (which will be your working directory today). These files give you some 'home-made' Stata commands. Then open Stata and set it to your working directory using the `cd <path>` command. (Ask if you need help with any of these.)

Question 1

Theory

Suppose $Y_1,..., Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Consider the following estimators of σ^2 [we can assume μ is *known* for 1) and 2)]:

1) $V_1 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2$. 2) $V_2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu)^2$ 3) $V_3 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

4) $V_4 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Considering whether the correct degrees of freedom have been used (see lecture 2.4),
a) Which of these four are biased, and which are unbiased?
b) Which one will overestimate the true variance, and which one will underestimate the true variance?

Repeated sampling (*by simulated random sampling*)

Use the Stata command `sim_V` *to assess your answers*. You will need to type, for example,

```
sim_V,n(50) reps(100) mu(12) sig_sq(16)
```

[*Note that there are only three spaces in the above command! And remember that Stata is case sensitive*.]

This will draw 100 repeated samples $y_1,..., y_{50}$, from $Y_1,..., Y_{50} \overset{iid}{\sim} N(12, 16)$, and report the observed mean, variance and mean square error of the 100 sampled values of each of the four *estimators*. What do you notice about the mean square errors? Try increasing the number of reps to 1000 or even 10,000 (if you have the patience to wait): you should see convergence towards theoretical values.

c) Theoretically calculate the true variance of estimator 4) above ($=S^2$), when $\sigma^2 = 16$ and n=50 (see Appendix to lecture notes, p12). Check that the appropriate result of the simulation command approximately confirms your calculation.

Question 2

Theory

Let $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ (estimator 4 above). In the lecture (2.4) we prove that S^2 is unbiased for σ^2. Now show algebraically that, assuming S^2 is unbiased, S must be biased for σ. {Hint: you should be able to do this without getting inside the formula above: use instead the formula for the definition of variance: $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$.}

Repeated sampling by simulation
The command

```
sim_S,n(50) reps(1000) mu(20) sig_sq(16)
```

draws 1000 samples, each of size 50, from a $N(20,16)$ population, and reports the mean and variance of the 1000 realised S^2 and S values.

Whether or not you've managed the proof in the theory part above!, use the command `sim_S` to:
a) confirm the existence of the bias in S ;
b) It can be shown (under Normality assumptions) that $E(S) \approx \sigma[1-(1/4(n-1))]$ and hence that the bias is $-\sigma/4(n-1)$; investigate this with `sim_S`.
c) How serious is this bias in practice? Confirm the sample sizes for which you expect a bias of i) approx 1% of σ; ii) approx 5%.
d) It can be shown (under Normality assumptions) that $Var(S) \approx \sigma^2/2(n-1)$. Investigate this with `sim_S`.

Question 3
Theory

Consider two random variables X_1, $X_2 \overset{iid}{\sim} (\mu, \sigma^2)$; this means $E(X_i)= \mu$, $Var(X_i)= \sigma^2$, with no other distributional assumptions. Suppose we repeatedly sample this pair and calculate $U = aX_1 + (1-a)X_2$, where a is a fixed constant. Determine algebraically:
a) Whether U is biased for μ.
b) $Var(U)$.
c) The value of a for which U is efficient.
d) The relative efficiency of U when $a = ½$ compared to when $a = ⅓$.
e) The value of a for which $U = \bar{X}$ (the sample mean of X_1, X_2).

Repeated sampling by simulation
The Stata command `sim_U,a(0.2) reps(100)` does the following:

i) Samples X_1, X_2 100 times from a population with mean 200 and variance 100 (as it happens a Normally distributed population, for computational convenience); we can choose the number of repeated samples by substituting into `reps()`.

ii) Calculates U when $a = 0.2$: ie for each sampled pair x_1, x_2 the program calculates the appropriate value of u. We end up with 100 realisations of U; we can choose the value of a by substituting into `a()`.

iii) Calculates and displays the observed mean and variance of these 100 realisations of U (and also the theoretically expected values).

iv) The program also displays the observed mean and variance of the sampled values of X_1; and, for good measure, shows a histogram of these and of the sampled values of U.

If U is unbiased for μ, we expect the mean of our 100 samples to be close to μ, and to get closer as the number of samples drawn increases. We also expect the variance of our 100 samples to be close to the theoretical variance (for the chosen value of a),

To use the command you just type in, for example, the following:

```
sim_U,a(0.33) reps(10)
```

Now use the command to complete the following table:

value of *a*	expected mean of *U*	mean of 1000 samples of *U*	expected variance of *U*	variance of 1000 samples of *U*	observed relative efficiency of *a* = 0.5
0.5					
0.33					
0.25					

With the help of this table, and further use of the simulation command, assess your answers to 3a)-d) above. If you start with a small number of reps, eg <100, and increase up to 1000 or 10000 (depending on your patience), you should see the convergence towards theoretically expected values.

Question 4 [*Further exercise: you probably won't have time for this today*]

The command

```
sim_M,n(100) reps(1000) mu(20) sig_sq(16)
```

draws 1000 samples, each of size 100, from a $N(20,16)$ population, and reports the mean and variance of the 1000 realised sample means and sample medians; it also reports the ratio of the variances. Use this command and your knowledge of the theory to complete the following table:

For 1000 repeated samples of size 100 drawn from N(20,16):

	observed		theoretically expected	
Estimator	mean	variance	mean	variance
sample mean				
sample median				

Repeat for a larger sample size or a larger number of repeated samples (or both).