Session 2: Solutions for Exercises and Practical

Exercise 2.1.1

a) $\mathrm{E}(\bar{Y}) = \mathrm{E}(\frac{1}{n}\sum Y_i) = \frac{1}{n}\sum \mathrm{E}(Y_i) = \frac{1}{n}n\mu = \mu$

$\mathrm{Var}(\bar{Y}) = \mathrm{Var}(\frac{1}{n}\sum Y_i)$ = (since Y_i independent) $\frac{1}{n^2}\sum \mathrm{Var}(Y_i) = \frac{1}{n^2}n\mathrm{Var}(Y_i) = \frac{\sigma^2}{n}$

b) $Z = \dfrac{\bar{Y}-\mu}{\sqrt{\mathrm{Var}(\bar{Y})}}$. Z is a linear transformation (using fixed constants) of a Normally distributed random variable, so Z is Normally distributed.

$\mathrm{E}(Z) = \dfrac{1}{\sqrt{\mathrm{Var}(\bar{Y})}}\mathrm{E}[\bar{Y}-\mu] = \dfrac{1}{\sqrt{\mathrm{Var}(\bar{Y})}}[\mu-\mu] = 0.$

$\mathrm{Var}(Z) = \dfrac{1}{\mathrm{Var}(\bar{Y})}\mathrm{Var}[\bar{Y}-\mu] = \dfrac{1}{\mathrm{Var}(\bar{Y})}\mathrm{Var}[\bar{Y}] = 1.$

So $Z \sim N(0,1)$.

Exercise 2.4.1

$$V_\mu = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)^2 \Rightarrow \mathrm{E}(V_\mu) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(Y_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \sigma^2.$$

Practical

Question 1

a),b) **Theory**: From section 2.4, we know that estimator (1) and (4) are unbiased, while (3) underestimates σ^2. Since (2) is slightly larger than unbiased (1), it will overestimate the true variance.
The following is a typical output, approximately confirming the above

```
. sim_V,n(50) reps(100000) mu(12) sig_sq(16)

Sample size:50. Y_1,..., Y_50 ~N(12,16).

From 100000 repeated samples of the four estimators:
mean of Vmu (V1) =      16.0145     variance of V1 = 10.2601  mse of V1 = 10.2603
mean of Vmu_1 (V2) =    16.3413     variance of V2 = 10.6832  mse of V2 = 10.7997
mean of Vn (V3) =       15.6935     variance of V3 = 10.0516  mse of V3 = 10.1455
mean of Vn_1 (V4) =     16.0138     variance of V4 = 10.4661  mse of V4 = 10.4663
```

Notice that the MSE is almost exactly equal to the variance for the unbiased estimators, as we would expect. You may notice that the mean MSE of V_3 is smaller than that of V_4. [Indeed, as a further exercise, if you're mathematical, you may like to prove that the denominator (n+1) has, under normality, the *smallest* MSE of estimators of this form for σ^2.] However, the bias of V_3, (and of the 1/(n+1) estimator) underestimates the parameter, and for dispersion parameters underestimation is less desirable (less cautious) than overestimation, since it leads to underestimates of uncertainty, so the smaller MSE does not compensate for this undesirable property (see eg Casella & Berger p332).

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

$\mathrm{Var}(S^2) = 2\sigma^4/(n-1) = 2*16^2/49 \approx 10.45$. This is quite close to the observed variance of V4 above.

Question 2

Theory

Let $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ (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: $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$.}

From Var(S) definition: $\mathrm{Var}(S) = \mathrm{E}(S^2) - [\mathrm{E}(S)]^2$
$\Rightarrow [\mathrm{E}(S)]^2 = \mathrm{E}(S^2) - \mathrm{Var}(S)$
Then, since $\mathrm{E}(S^2) = \sigma^2$,
$[\mathrm{E}(S)]^2 = \sigma^2 - \mathrm{Var}(S)$
$\Rightarrow \mathrm{E}[S] = \sqrt{\sigma^2 - \mathrm{Var}(S)}$
So we expect S to underestimate σ; but as $n \to \infty$, $\mathrm{Var}(S) \to 0$, so $\mathrm{E}[S] \to \sigma$.

Generally, for a linear transformation, just as the arithmetic mean of $f(x_i)$ = f(arithmetic mean of x_i), so $\mathrm{E}(f(X)) = f(\mathrm{E}(x))$; but for a non-linear function, such as the sqrt(x) above, or $1/x$, this does not hold: the mean of $1/x_i$ is not 1/mean(x_i).

Repeated sampling

Use the command `sim_S` to:

a) confirm the existence of the bias in S ;

Here is an example with 100,000 repeated samples:

```
. sim_S,n(50) reps(100000) mu(20) sig_sq(16)

Sample size:50. Y_1,..., Y_50 ~N(20,16).

From 100000 repeated samples:
mean of S_sq =          16.0006     variance of S_sq = 10.4143
mean of S =             3.9798      variance of S = 0.1620
```

Note that although the observed bias for S^2 is only 0.0006, for S it is -0.0202.

b) It can be shown that *for a Normal population* $\mathrm{E}(S) \approx \sigma[1-(1/4(n-1))]$ and hence that the bias is $-\sigma/4(n-1)$; investigate this with `sim_S`.

In the output above (for (a)) the theoretical bias is -1/49 = -0.0204, which is very close to the observed bias.

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%.

i) For 1% bias, $[\sigma/4(n-1)]/\sigma = 0.01 \Rightarrow 1/4(n-1) = 1/100 \Rightarrow n=26$; this is confirmed by the approximate 1% bias exhibited below:

```
. sim_S,n(26) reps(100000) mu(20) sig_sq(16)

Sample size:26. Y_1,..., Y_26 ~N(20,16).

From 100000 repeated samples:
mean of S_sq =          15.9983     variance of S_sq = 20.3726
mean of S =             3.9602      variance of S = 0.3152
```

ii) For 5% bias, $1/4(n-1) = 1/20 \Rightarrow n=6$; note the approximate 5% bias below:

```
. sim_S,n(6) reps(100000) mu(20) sig_sq(16)

Sample size:6. Y_1,..., Y_6 ~N(20,16).

From 100000 repeated samples:
mean of S_sq =          15.9789     variance of S_sq = 101.8274
mean of S =             3.8044      variance of S = 1.5056
```

d) It can be shown that *for a Normal population* $\text{Var}(S) \approx \sigma^2/2(n-1)$. Investigate this with `sim_S`.

For $\sigma^2 = 16$, n=50 $\Rightarrow$ $\text{Var}(S) \approx 0.1633$; n=26 $\Rightarrow$ $\text{Var}(S) \approx 0.32$; n=6 $\Rightarrow$ $\text{Var}(S) \approx 1.6$; these are approximately confirmed by the outputs above at (a) and (c).

Question 3
Theory

Consider two random variables $X_1, X_2 \overset{iid}{\sim} (\mu, \sigma^2)$ and $U = a\,X_1 + (1-a)\,X_2$, where a is a fixed constant. Determine algebraically:

a) Whether U is biased for μ.

$E(U) = aE(X_1) + (1-a)E(X_2) = a\mu + (1-a)\,\mu = \mu$. So U is unbiased for μ.

b) $\text{Var}(U) = a^2\text{Var}(X_1) + (1-a)^2\text{Var}(X_2) = \sigma^2(2a^2 - 2a + 1)$
since independent random variables so $\text{Cov}(X_1, X_2)=0$.

c) The value of a for which U is efficient.

We need the value of a for which $\text{Var}(U)$ is a minimum. We thus need to find the minimum of $f(a) = 2a^2 - 2a + 1$. We thus solve for $\frac{df}{da} = 0$:

$\frac{df}{da} = 4a - 2$, and $4a - 2 = 0 \Rightarrow a = ½$, at which $\text{Var}(U) = ½\sigma^2$.

d) The relative efficiency of U when $a = ½$ compared to when $a = 1/3$.

The relative efficiency is $\text{Var}(U|\ a = 1/3)/\ \text{Var}(U|\ a = ½) = 10/9$ (just using the expression in (b) above to calculate).

e) The value of a for which $U = \overline{X}$.

Since $\overline{X} = ½\,X_1 + ½\,X_2 \Rightarrow U = \overline{X}$ at $a = ½$.

Note: expressions of the form $U = a\,X_1 + (1-a)\,X_2$ are an example of **weighting**, a very common technique in statistics. There are a number of reasons we might want to give different weights to X_1, X_2, and construct a weighted average rather than the simple (equally weighted) average, where $a = (1-a) = ½$; although the weighted averages are unbiased, the simple average has the least variance.

Repeated sampling ("simulation")

You probably won't get exactly these numbers for the observed results, because different (pseudo)random draws will have taken place to obtain these. But your observed results should be 'similar'.

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	*200*	*199.8027*	*50*	*48.0671*	*1*
0.33	*200*	*200.1643*	*55.5800*	*56.9618*	*1.19*
0.25	*200*	*199.8581*	*62.5*	*60.81*	*1.27*

a) Table above suggests unbiased U.
b) Observed variances are quite close to theoretical value.
c) as $|a-0.5|$ increases you should find the observed variance increasing. And at a=0.5 the observed variance should be close to $½\sigma^2$.
d) The theoretical relative efficiency should be 10/9 = 1.11; on these 1000 samples we observe a relative efficiency of 1.19.

Question 4 [Further exercise]

Again, you probably won't get exactly the same results, but yours should be quite close to these.

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

	observed		theoretically expected	
Estimator	mean	variance	mean	variance
sample mean	*20.0067*	*0.1563*	*20*	*0.16*
sample median	*20.0207*	*0.2296*	*20*	*≈1.571*0.16 =0.25*

You should find that as you increase the number of repeated samples, the observed means and the variance of the mean approach the theoretical values. The variance of the sample median should require a large sample to give the relative efficiency of the sample mean close to 1.57.