Section 5.2. (cont.)

Reminders:

i) $E(aX + bY) = aE(X) + bE(Y)$
ii) $E(X^2) = Var(X) + E(X)^2$ (see page 60, 2.3.1)
iii) $Var(aX + b) = a^2 Var(X)$ (see page 60)
iv) If X and Y are independent, $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$.

(Word to wise: good material for quiz questions here.)

Suppose $X_1, \dots, X_n$ *iid*. Suppose μ is the expected value of the common distribution of the X_i and σ^2 is its variance. The two statistics of greatest importance are:

- The sample mean: $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$,
- The sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$.

The expected value and variance of the statistic $\overline{X}$.

- $E(\overline{X}) = \frac{1}{n}(E(X_1) + \cdots + E(X_n)) = \mu$ (Theorem 5.2.6.a, p. 213-4).
- $Var(\overline{X}) = \frac{1}{n^2}(Var(X_1) + \cdots + Var(X_n)) = \sigma^2/n$ (Theorem 5.2.6.b, p. 213-4).

To deduce information about S^2 (see next page), we will use a fact of arithmetic:

Theorem 5.2.4. *For any numbers $x_1, \dots, x_n$, let $\overline{x}$ be their average.*
(a) The minimum value of $q(a) := \sum_{i=1}^n (x_i - a)^2$ is obtained when $a = \overline{x}$, and
(b) $q(\overline{x}) := \sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n x_i^2 - n\overline{x}^2$.

Proof. We calculate:

$$\begin{aligned}
\sum_{i=1}^n (x_i - a)^2 &= \sum_{i=1}^n (x_i - \overline{x} + \overline{x} - a)^2 \\
&= \sum_{i=1}^n (x_i - \overline{x})^2 + 2\sum_{i=1}^n (x_i - \overline{x})(\overline{x} - a) + \sum_{i=1}^n (\overline{x} - a)^2 \\
&= \sum_{i=1}^n (x_i - \overline{x})^2 + \sum_{i=1}^n (\overline{x} - a)^2.
\end{aligned}$$

Now, (a) follows since a sum of squares is always positive, so the last line is minimized when $a = \overline{x}$. For (b), set $a = 0$; our calculation shows:

$$\sum_{i=1}^n (x_i - 0)^2 = \sum_{i=1}^n (x_i - \overline{x})^2 + \sum_{i=1}^n (\overline{x} - 0)^2.$$

The expected value of the statistic S^2 (Theorem 5.2.6.c, p. 213-4).

$$\begin{aligned}
E(S^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right) \\
&= \frac{1}{n-1}\left(nE(X_1^2)-nE(\overline{X}^2)\right) \\
&= \frac{1}{n-1}\left(n(\sigma^2+\mu^2)-n(\frac{\sigma^2}{n}+\mu^2)\right) \\
&= \frac{1}{n-1}\left(n\sigma^2-\sigma^2\right) \\
&= \sigma^2.
\end{aligned}$$

Comment. This result explains the curious coefficient $\frac{1}{n-1}$ in the expression for the sample variance. This makes S^2 an unbiased estimator of σ^2. A related statistic (which might be called "the variance within the sample") is $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$. This is seldom used, because it gives an estimate for σ^2 that is systematically incorrect.

Homework. Exercise 5.5, page 256. Hint: $P(\overline{X}\le x)=P(X_1+\cdots+X_n\le nx)$.