1 Supplementary file: The distribution of sample means

The expected value of the sample variance

- OK, first we want to know how the variance of your samples (s^2) is related to the population variance σ^2.
- Remember that the variance of a sample is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

- We can rewrite this as:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i^2 - 2 \cdot x_i + \bar{x})^2}{n}$$
$$= \frac{\sum_{i=1}^{n} x_i^2 - 2 \cdot \bar{x} \cdot \sum_{i=1}^{n} x_i + \bar{x}^2}{n}$$

The expected value of the sample variance (2)

- Further rewriting: Since $\sum_{i=1}^{n} x_i = n \cdot \bar{x}$:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - 2 \cdot \bar{x} \cdot \sum_{i=1}^{n} x_i + \bar{x}^2}{n}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - 2 \cdot \bar{x} \cdot n \cdot \bar{x} + \bar{x}^2}{n}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$$

The expected value of the sample variance (3)

- Now we can calculate the expected value of s^2:

$$
\begin{aligned}
E(S^2) &= E\left(\frac{\sum\limits_{i=1}^{n} X_i^2}{n} - \bar{X}^2\right) \\
&= E\left(\frac{\sum\limits_{i=1}^{n} X_i^2}{n}\right) - E(\bar{X}^2) \\
&= \frac{\sum\limits_{i=1}^{n} E(X_i^2)}{n} - E(\bar{X}^2) = \frac{n \cdot E(X_i^2)}{n} - E(\bar{X}^2)
\end{aligned}
$$

The expected value of the sample variance (4)

- And $\frac{n \cdot E(X_i^2)}{n} - E(\bar{X}^2)$ of course simplifies to $E(X_i^2) - E(\bar{X}^2)$
- So, now we have to figure out what $E(X_i^2)$ and $E(\bar{X}^2)$ are.
- The "easiest" (I know, right?) way to do this is to start with the population variance σ^2 and the variance of the sample means $\sigma_{\bar{x}}^2$

The expected value of the sample variance (5)

- We can define the population variance as $\sigma^2 = E(X_i - \mu)^2$, the expected value of the squared deviations of X from the population mean μ
- Let's rewrite this:

$$
\begin{aligned}
\sigma^2 = E(X_i - \mu)^2 &= E(X_i^2 - 2X_i\mu + \mu^2) \\
&= E(X_i^2) - E(2X_i\mu) + E(\mu^2) \\
&= E(X_i^2) - 2\mu E(X_i) + \mu^2
\end{aligned}
$$

since μ is a constant (and μ^2 is too, of course).

The expected value of the sample variance (6)

Continuing from previous slide: - We already determined that $\mu = E(X)$, so:

$$\begin{aligned}\sigma^2 &= E(X_i^2) - 2\mu E(X_i) + \mu^2 \\ &= E(X_i^2) - 2\mu^2 + \mu^2 = E(X_i^2) - \mu^2\end{aligned}$$

- Solving for $E(X_i^2)$:

$$\begin{aligned}\sigma^2 &= E(X_i^2) - \mu^2 \\ &\Leftrightarrow E(X_i^2) = \sigma^2 + \mu^2\end{aligned}$$

- OK, so now we know that the expected value of a squared random variable is equal to the sum of the population variance σ^2 and the square of the population mean μ^2.

The expected value of the sample variance (7)

- Next up: the variance of sample means $\sigma_{\bar{x}}^2$
  - This is the square of the *standard error* of the mean $\sigma_{\bar{x}}$
- We can define the variance of sample means as $\sigma_{\bar{x}}^2 = E(\bar{X} - \mu)^2$, i.e. the expected value of the squared deviations of the sample means from the true population mean
- We can rewrite this just like we did for the sample variance (this works exactly the same as before; if you are bored, you can skip the next two slides).

The expected value of the sample variance (7a)

- We can define the variance of the sample means as $\sigma_{\bar{x}}^2 = E(\bar{X} - \mu)^2$
- Let's rewrite this:

$$\begin{aligned}
\sigma^2 &= E(\bar{X} - \mu)^2 = E(\bar{X}^2 - 2\bar{X}\mu + \mu^2) \\
&= E(\bar{X}^2) - E(2\bar{X}\mu) + E(\mu^2) \\
&= E(\bar{X}^2) - 2\mu E(\bar{X}) + \mu^2
\end{aligned}$$

since μ is a constant (and μ^2 is too, of course).

The expected value of the sample variance (7b)

Continuing from previous slide: - We already determined that $\mu = E(\bar{X})$, so:

$$\begin{aligned}\sigma_{\bar{x}}^2 &= E(\bar{X}^2) - 2\mu E(\bar{X}) + \mu^2 \\ &= E(\bar{X}^2) - 2\mu^2 + \mu^2 = E(\bar{X}^2) - \mu^2\end{aligned}$$

- Solving for $E(\bar{X}^2)$:

$$\begin{aligned}\sigma_{\bar{x}}^2 &= E(\bar{X}^2) - \mu^2 \\ &\Leftrightarrow E(\bar{X}^2) = \sigma_{\bar{x}}^2 + \mu^2\end{aligned}$$

- OK, so now we know that the expected value of the squared mean of a random variable is equal to the sum of the variance of the sample means $\sigma_{\bar{x}}^2$ and the square of the population mean μ^2.

The expected value of the sample variance (8)

- Plugging $E(X_i^2) = \sigma^2 + \mu^2$ and $E(\bar{X}^2) = \sigma_{\bar{x}}^2 + \mu^2$ into our term for the expected value of the sample variance:

$$\begin{aligned} E(S^2) &= E(X_i^2) - E(\bar{X}^2) = \sigma^2 + \mu^2 - (\sigma_{\bar{x}}^2 + \mu^2) \\ &= \sigma^2 - \sigma_{\bar{x}}^2 \end{aligned}$$

- In words: the expected value of the sample variance is equal to the population variance minus the variance of the sample means.
  - This means that the sample variance *systematically* underestimates the population variance
  - The sample variance is *NOT* an unbiased estimator of the population variance.

The expected value of the variance of sample means

- We start with the relationship we just figured out: $E(\sigma^2_{\bar{x}}) = \sigma^2_{\bar{x}} = E(\bar{X}^2) - \mu^2$ (since $E(\bar{X}^2)$ and μ^2 are both constants).
- We can rewrite $\bar{X}^2$ as:

$$\begin{aligned}
\bar{X}^2 &= \frac{(X_1 + X_2 + \cdots + X_n)^2}{n^2} \\
&= \frac{1}{n^2} \cdot \Bigg(X_1^2 + X_2^2 + \cdots + X_n^2 \\
&\quad + 2 \sum_{i=1} \sum_{j=i+1} X_i \cdot X_j \Bigg)
\end{aligned}$$

The expected value of the variance of sample means (2)

- If (and only if!) $X_1, X_2, \ldots, X_n$ are independent (i.e. the value of X_1 doesn't depend on the value of X_2, or X_3, etc.), we can write the expected value of the final term of this expression as:

$$E\left(2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_i \cdot X_j\right) = n \cdot *(n-1) \cdot E(X_i) \cdot E(X_j)$$
$$= n \cdot (n-1) \cdot \mu^2$$

- Since $E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$ and $E(X_i) = \mu$

The expected value of the variance of sample means (3)

- With that, we can rewrite $E(\bar{X^2})$ as

$$E(\bar{X^2}) = \frac{1}{n^2} \cdot \Big(E(X_1)^2 + E(X_2)^2 + \dots \\ + E(X_n)^2) + n \cdot (n-1) \cdot \mu^2\Big)$$

- But we know already (through our hard work earlier) that the expected value of the square of X_i is $E(X_i^2) = \sigma^2 + \mu^2$.
- So we can replace $E(X_1)^2 + E(X_2)^2 + \dots + E(X_n)^2$ with $n \cdot (\sigma^2 + \mu^2) = n \cdot \sigma^2 + n \cdot \mu^2$.

The expected value of the variance of sample means (4)

- Let's do that now:

$$E(\bar{X}^2) = \frac{1}{n^2} \cdot \left(n \cdot \sigma^2 + n \cdot \mu^2 + n \cdot (n-1) \cdot \mu^2\right)$$
$$= \frac{\sigma^2}{n} + \frac{n \cdot \mu^2 + n * 2 \cdot \mu^2 - n \cdot \mu^2}{n^2} = \frac{\sigma^2}{n} + \mu^2$$

- Plugging this into our previous equation $\sigma^2_{\bar{x}} = E(\bar{X}^2) - \mu^2$ we get:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n} + \mu^2 - \mu^2 = \frac{\sigma^2}{n}$$

- If we take the square root of this, we *FINALLY* get the **standard error of the mean**:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Correcting the bias in the expected value of the sample variance

- Before we actually use our hard-earned $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$, a quick detour:
- Remember that the expected value of the sample variance was biased by the variance of the sample mean, i.e. $E(S^2) = \sigma^2 - \sigma^2_{\bar{x}}$?
- Now we know what the variance of the sample mean is, so let's plug it in:

$$\begin{aligned} E(S^2) = \sigma^2 - \sigma^2_{\bar{x}} = \sigma^2 - \frac{\sigma^2}{n} = \frac{n \cdot \sigma^2 - \sigma^2}{n} \\ = \sigma^2 \cdot \frac{n-1}{n} \end{aligned}$$

Correcting the bias in the expected value of the sample variance (2)

- We just found out that the expected value of the sample variance $E(s^2)$ underestimates the true population variance σ^2 by a factor of $\frac{n-1}{n}$.
- That means we can apply a correction factor to the sample variance so that it becomes an *unbiased* estimator of the population variance:

$$
\begin{aligned}
s_{n-1}^2 &= s^2 / \frac{n-1}{n} = s^2 \cdot \frac{n}{n-1} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{n} \cdot \frac{n}{n-1} \\
&= \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{n-1}
\end{aligned}
$$

Correcting the bias in the expected value of the sample variance (3)

- Most statistical software will use this corrected formula for computing the sample variance: $s^2_{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- If you want a more intuitive explanation of what is going on here, watch the videos at EasyStats: `http://easystats.org/`