MATH 134A Review: The point estimate for the population variance

Let $y_1, \ldots, y_N$ be data collected from individuals within a population; here N is the population size. Then

$$\frac{y_1 + \cdots + y_N}{N} =: \mu$$

is the population mean. Also

$$\frac{(y_1 - \mu)^2 + \cdots + (y_N - \mu)^2}{N} =: \sigma^2$$

is the population variance.

Let $x_1, \ldots, x_n$ be data collected from individuals within a sample; here n is the sample size. Then

$$\frac{x_1 + \cdots + x_n}{n} =: \bar{x}$$

is the point estimate for the population mean. Also

$$\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} =: s^2$$

is the point estimate for the population variance. Why does the last equation make sense? Why divide by $(n-1)$? The answer/derivation can be found below.

Facts to Know

Let X be a (discrete) random variable with probability distribution function $p(x) = \mathbb{P}(X = x)$.

- The expectation of X is $\mathbb{E}(X) \coloneqq \sum_x xp(x)$.
- The variance of X is $\mathrm{Var}(X) \coloneqq \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

Let X and Y be (discrete) random variables.

- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
- $\mathbb{E}(cX) = c\mathbb{E}(X)$.
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \impliedby X$ and Y are independent.
- $\mathrm{Var}(cX) = |c|^2 \mathrm{Var}(X)$.

Derivation

Let $X_1, \ldots, X_n$ be independent and identically distributed (discrete) random variables. For example, X_i ($i = 1, \ldots, n$) can be the data value of an individual from a population of size $N > n$. Let μ be the expectation of $X_1, \ldots, X_n$ and let σ^2 be the variance of $X_1, \ldots, X_n$.

Theorem. *Define* $\overline{X} \coloneqq \frac{X_1 + \cdots + X_n}{n}$. *Then*

$$\mathbb{E}(\overline{X}) = \mu \qquad \textit{and} \qquad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

and

$$\mathbb{E}(\frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}) = \sigma^2.$$

Proof. We leave it as an exercise to prove $\mathbb{E}(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$. Observe

$$\mathbb{E}[(X_i)^2] = \mathrm{Var}(X_i) + [\mathbb{E}(X_i)]^2 = \sigma^2 + \mu^2$$

$$\mathbb{E}[(\overline{X})^2] = \mathrm{Var}(\overline{X}) + [\mathbb{E}(\overline{X})]^2 = \frac{\sigma^2}{n} + \mu^2$$

$$\begin{aligned}
\mathbb{E}(\frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}) &= \mathbb{E}(\frac{1}{n-1} \sum_{i=1}^{n} \big((X_i - \overline{X})^2\big)) \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}\big((X_i - \overline{X})^2\big) \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}\big((X_i)^2 - 2X_i\overline{X} + (\overline{X})^2\big) \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \big(\mathbb{E}[(X_i)^2] - 2\mathbb{E}[X_i\overline{X}] + \mathbb{E}[(\overline{X})^2]\big) \\
&= \frac{1}{n-1} \big(\textstyle\sum_{i=1}^{n} \mathbb{E}[(X_i)^2] - 2\sum_{i=1}^{n} \mathbb{E}[X_i\overline{X}] + \sum_{i=1}^{n} \mathbb{E}[(\overline{X})^2]\big) \\
&= \frac{1}{n-1} \big(n\mathbb{E}[(X_i)^2] - 2\mathbb{E}[\textstyle\sum_{i=1}^{n} X_i\overline{X}] + n\mathbb{E}[(\overline{X})^2]\big) \\
&= \frac{1}{n-1} \big(n\mathbb{E}[(X_i)^2] - 2\mathbb{E}[(n\overline{X})\overline{X}] + n\mathbb{E}[(\overline{X})^2]\big) \\
&= \frac{1}{n-1} \big(n\mathbb{E}[(X_i)^2] - 2n\mathbb{E}[(\overline{X})^2] + n\mathbb{E}[(\overline{X})^2]\big) \\
&= \frac{1}{n-1} \big(n\mathbb{E}[(X_i)^2] - n\mathbb{E}[(\overline{X})^2]\big) \\
&= \frac{1}{n-1} \big(n(\sigma^2 + \mu^2) - n(\tfrac{\sigma^2}{n} + \mu^2)\big) = \tfrac{1}{n-1}\big(n\sigma^2 - \sigma^2\big) = \sigma^2.
\end{aligned}$$