Proof of Sample Variance & Sample Covariance

Unbiased Estimator

Suppose we have a statistical model,parameterized by a real number θ, giving rise to a probability distribution for observed data, and a statistic $\hat{\theta}$ which serves as an estimator of θ based on any observed data x. That is, we assume that our data follow some unknown distribution $P(x \mid \theta)$(where θ is a fixed constant that is part of this distribution, but is unknown), and then we construct some estimator $\hat{\theta}$maps observed data to values that we hope are close to θ. The **bias** of $\hat{\theta}$ relative to θ is defined as

$$Bias_\theta[\,\hat{\theta}\,] = \mathrm{E}_{x|\theta}[\,\hat{\theta}\,] - \theta = \mathrm{E}_{x|\theta}[\,\hat{\theta} - \theta\,]$$

where $E_{x|\theta}$ denotes over the distribution $P(x \mid \theta)$, i.e. averaging over all possible observations$x$. The second equation follows since θ is measurable with respect to the conditional distribution $P(x \mid \theta)$

An estimator is said to be **unbiased** if its bias is equal to zero for all values of parameter θ.

In a simulation experiment concerning the properties of an estimator, the bias of the estimator may be assessed using the mean signed difference.

Sample Variance & Sample Covariance

The **sample variance** of a random variable demonstrates two aspects of estimator bias:

Firstly, the naive estimator is biased, which can be corrected by a scale factor;

Second, **the unbiased estimator** is not optimal in terms of mean squared error(MSE), which can be minimized by using a different scale factor, resulting in a biased estimator with lower MSE than the unbiased estimator.

Concretely, the naive estimator sums the squared deviations and divides by n, which is biased. Dividing instead by $n-1$ yields an unbiased estimator.

Conversely, MSE can be minimized by dividing by a different number (depending on distribution), but this results in a biased estimator. This number is always larger than $n-1$, so this is known as a shrinkage estimator, as it "shrinks" the unbiased estimator towards zero; for the normal distribution the optimal value is $n+1$.

Distribution

Varaiables X, Y follow different and certain distributions with

Expectation

$$E(X) = \mu_X, E(Y) = \mu_Y$$

Variance

$$\sigma_X^2, \sigma_Y^2$$

Covariance

$$Cov(X,Y) = \sigma_{XY}$$

Sample

Independent and identical distribution (I.I.D) random variables

$$X_i, Y_i \quad i = 1, 2, \ldots, n$$

precondition: X_i, Y_j are only correlated when $i = j$,that is

$$Cov(X_i, Y_j) = \begin{cases} Cov(X_i, Y_i) & i = j \\ 0 & i \neq j \end{cases}$$

Sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

For biased estimator is

Variance

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Covariance

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

For unbiased estimator is

Sample variance

$$\hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Sample covariance

$$\hat{\sigma}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

Biased Sample Variance's expectation

$$
\begin{aligned}
E[S_X^2] &= E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X + \mu_X - \overline{X})^2] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}[(X_i - \mu_X)^2 + (\mu_X - \overline{X})^2 + 2(X_i - \mu_X)(\mu_X - \overline{X})]] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)^2 + (\mu_X - \overline{X})^2 + 2(\mu_X - \overline{X})\frac{1}{n}\sum_{i=1}^{n}[(X_i - \mu_X)]] \\
&= \frac{1}{n}\sum_{i=1}^{n}E[(X_i - \mu_X)^2] + E[(\mu_X - \overline{X})^2 + 2(\mu_X - \overline{X})(\overline{X} - \mu_X)] \\
&= \frac{1}{n} \times n\sigma_X^2 - E[(\mu_X - \overline{X})^2] \\
&= \sigma_X^2 - E[(\overline{X} - \mu_X)^2] \\
&= \sigma_X^2 - Var(\overline{X})
\end{aligned}
$$

$$
\begin{aligned}
\because X_i\ (i = 1, 2, \ldots, n)\ \text{ is I. I. D} \\
Var(\overline{X}) &= Var(\frac{1}{n}\sum_{i=1}^{n}X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}Var(X_i) \\
&= \frac{1}{n}\sigma_X^2 \\
\therefore E[S_X^2] &= \frac{n-1}{n}\sigma_X^2
\end{aligned}
$$

To be a Unbiased Estimator,required

$$
E[\hat{\sigma}_X^2] - \sigma_X^2 = 0
$$

$\therefore$ unbiased estimator is

$$
\hat{\sigma}_X^2 = \frac{n}{n-1}E[S_X^2] = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2
$$

Biased Sample Convariance's expectation

$$
\begin{aligned}
& E[S_{XY}] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}[(X_i-\overline{X})(Y_i-\overline{Y})]] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}[(X_i-\mu_X+\mu_X-\overline{X})(Y_i-\mu_Y+\mu_Y-\overline{Y})]] \\
&= E[\frac{1}{n}\sum_{i=1}^{n}[(X_i-\mu_X)(Y_i-\mu_Y)+(\mu_X-\overline{X})(\mu_Y-\overline{Y})+(X_i-\mu_X)(\mu_Y-\overline{Y})+(Y_i-\mu_Y)(\mu_X-\overline{X})]] \\
&= \frac{1}{n}\sum_{i=1}^{n}E[(X_i-\mu_X)(Y_i-\mu_Y)]+ \\
&\quad E[(\mu_X-\overline{X})(\mu_Y-\overline{Y})+\frac{1}{n}\sum_{i=1}^{n}[(X_i-\mu_X)(\mu_Y-\overline{Y})+(Y_i-\mu_Y)(\mu_X-\overline{X})]] \\
&= \sigma_{XY}+E[(\mu_X-\overline{X})(\mu_Y-\overline{Y})+(\mu_Y-\overline{Y})\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_X)+(\mu_X-\overline{X})\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_Y)] \\
&= \sigma_{XY}+E[(\mu_X-\overline{X})(\mu_Y-\overline{Y})+(\mu_Y-\overline{Y})(\overline{X}-\mu_X)+(\mu_X-\overline{X})(\overline{Y}-\mu_Y)] \\
&= \sigma_{XY}-E[(\overline{X}-\mu_X)(\overline{Y}-\mu_Y)] \\
&= \sigma_{XY}-Cov(\overline{X},\overline{Y})
\end{aligned}
$$

$$
\begin{aligned}
Cov(\overline{X},\overline{Y}) &= E[(\overline{X}-\mu_X)(\overline{Y}-\mu_Y)] \\
&= E[(\frac{1}{n}\sum_{i=1}^{n}X_i-\mu_X)(\frac{1}{n}\sum_{i=1}^{n}Y_i-\mu_Y)] \\
&= \frac{1}{n^2}E[\sum_{i=1}^{n}(X_i-\mu_X)\sum_{i=1}^{n}(Y_i-\mu_Y)] \\
&= \frac{1}{n^2}E[\sum_{i=1}^{n}(X_i-\mu_X)(Y_i-\mu_Y)+\sum_{\substack{i=1,j=1,\\ i\neq j}}^{n}(X_i-\mu_X)(Y_j-\mu_Y)] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}Cov(X_i,Y_i)+\sum_{\substack{i=1,j=1,\\ i\neq j}}^{n}Cov(X_i,Y_j)
\end{aligned}
$$

$$
\begin{gathered}
\because Cov(X_i,Y_j)=0,(i,j=1,2,\ldots,n,i\neq j) \\
\therefore Cov(\overline{X},\overline{Y})=\frac{1}{n^2}\sum_{i=1}^{n}Cov(X_i,Y_i)=\frac{1}{n}\sigma_{XY} \\
\therefore E[S_{XY}]=\frac{n-1}{n}\sigma_{XY}
\end{gathered}
$$

$\therefore$ similarly,unbiased estimator is

$$
\hat{\sigma}_{XY}=\frac{n}{n-1}E[S_{XY}]=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})
$$

Refference

Another Proof of $E(\sigma_X^2)$

$$
\begin{aligned}
&E[\sigma_y^2] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum j = 1^n y_j\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[y_i^2 - \frac{2}{n} y_i \sum_{j=1}^{n} y_j + \frac{1}{n^2}\sum_{j=1}^{n} y_j \sum_{k=1}^{n} y_k\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}E[y_i^2] - \frac{2}{n}\sum_{j\neq i} E[y_i y_j] + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j}^{n} E[y_j y_k] + \frac{1}{n^2}\sum_{j=1}^{n} E[y_j^2]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}(\sigma^2+\mu^2) - \frac{2}{n}(n-1)\mu^2 + \frac{1}{n^2}n(n-1)\mu^2 + \frac{1}{n}(\sigma^2+\mu^2)\right] \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

Cite

- https://en.wikipedia.org/wiki/Bias_of_an_estimator
- https://en.wikipedia.org/wiki/Variance#Sample_variance
- https://en.wikipedia.org/wiki/Sample_mean_and_covariance
- https://en.wikipedia.org/wiki/Covariance