19) Unbiased Estimators

Defn 19.1: Let $(x_1, x_2, \ldots, x_n)$ be a dataset modelled by random variables $X_1, X_2, \ldots, X_n$ and let $t = h(x_1, x_2, \ldots, x_n)$ be an estimate for the value of a model parameter θ expressed as a function h evaluated on the dataset values $x_1, x_2, \ldots, x_n$.
Then the random variable $T = h(X_1, X_2, \ldots, X_n)$ is called an estimator.
Such an estimator is unbiased if $E[T] = \theta$ irrespective of value of θ.
Otherwise it is biased. The difference $E[T] - \theta$ is called the bias of T

Theorem 19.2: Suppose $X_1, X_2, \ldots, X_n$ is an iid sample from a distribution with expection $\mu < \infty$ and variance $\sigma^2 < \infty$ Then the sample mean

$$\boxed{\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)}$$

is an unbiased estimator for μ
The sample variance

$$\boxed{S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

is an unbiased estimator for σ^2.

proof: That $E[\bar{X}_n] = \mu$ we already know from chapter 13.

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} E[X_i]\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{\cancel{n}}\cancel{n}E[X_i]$$

$$= E[X_i]$$ as it is iid

$$= E[X] = \mu$$

$$E[S_n^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n} E[(X_i-\bar{X}_n)^2]\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} E[(X_i-\bar{X}_n)^2]$$ linearity of expectations Thm 10.2

We observe using Theorem 7.19

$$E[(X_i-\bar{X}_n)^2] = Var(X_i-\bar{X}_n) + (E[X_i-\bar{X}_n])^2$$

$$= Var(X_i-\bar{X}_n) + (E[X_i]-E[\bar{X}_n])$$

$$= Var(X_i-\bar{X}_n) + (\mu-\mu)^2$$

$$= \mathrm{Var}(X_i - \bar{X}_n) + 0$$

$$= \mathrm{Var}(X_i - \bar{X}_n)$$

$$\Rightarrow E[(X_i - X_n)^2] = \mathrm{Var}(X_i - \bar{X}_n)$$

To calculate $\mathrm{Var}(X_i - \bar{X}_n)$, we use the trick of writing

$$X_i - \bar{X}_n = X_i - \frac{1}{n}\sum_{j=1}^{n} X_j$$

$$= \frac{nX_i}{n} - \frac{1}{n}\sum_{j=1}^{n} X_j$$

$$= \frac{nX_i}{n} - \frac{1}{n}X_i - \frac{1}{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} X_j$$

$$= \frac{(n-1)X_i}{n} - \frac{1}{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} X_j$$

$$\Rightarrow \quad X_i - \bar{X}_n = \frac{n-1}{n} X_i - \frac{1}{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} X_j$$

We can use the fact that X_j is independant of X_i for $i \neq j$ so covariance is 0,

$$\mathrm{Var}(X_i - \bar{X}_n) = \mathrm{Var}\left(\frac{n-1}{n} X_i - \frac{1}{n} \sum_{j \neq i}^{n} X_j\right)$$

$$= \mathrm{Var}\left(\frac{n-1}{n} X_i\right) - \mathrm{Var}\left(\frac{1}{n} \sum_{j \neq i}^{n} X_j\right)$$

$$= \frac{(n-1)^2}{n^2} \mathrm{Var}(X_i) - \frac{1}{n^2} \mathrm{Var}\left(\sum_{j \neq i}^{n} X_j\right)$$

$$= \frac{(n-1)^2}{n^2} \mathrm{Var}(X_i) - \frac{1}{n^2} \sum_{j \neq i}^{n} \mathrm{Var}(X_j) \quad \text{by iid}$$

$$= \frac{(n-1)^2}{n^2} \mathrm{Var}(X_i) + \frac{1}{n^2} \cdot (n-1) \mathrm{Var}(X_j)$$

By iid sample $\mathrm{Var}(X_i) = \mathrm{Var}(X_j) = \sigma^2$. Hence

$$\mathrm{Var}(X_i - \bar{X}_n) = \frac{(n-1)^2}{n^2} \sigma^2 + \frac{1}{n^2} (n-1) \sigma^2$$

$$= \frac{n-1}{n} \sigma^2$$

For the third equality used transformation property of variance and Thm 7.25.

So

$$E[S_n^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} E[(X_i - \bar{X}_n)^2]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \mathrm{Var}(X_i - \bar{X}_n)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \frac{n-1}{n}\sigma^2$$

$$= \frac{1}{\cancel{n-1}} \cdot \cancel{n} \cdot \frac{\cancel{n-1}}{\cancel{n}}\sigma^2 = \sigma^2$$

$$\Rightarrow E[S_n^2] = \sigma^2$$ as required.

∎

Example: 17.1 (continued) An estimator for acceleration in the inclined plane experiment.

Galileo has given us a functional relationship between the distance travelled and time of travel:

$$x = \frac{1}{2} a t^2.$$

We can solve this for acceleration

$$a = \frac{2x}{t^2}$$

However each time the experiment is repeated, one will get different result x_i due to random errors. We modelled these observations by random variables

$$X_i = \frac{1}{2} a (t + U_i)^2 + V_i$$

where $U_i \sim N(0, \sigma_U^2)$ $V_i \sim N(0, \sigma_V^2)$.

These errors are all independant. We could try to estimate a by taking average of all the observations $x_1, x_2, \dots, x_n$:

$$a \approx \frac{2}{t^2} \frac{1}{n} \sum_{i=1}^{n} x_i$$

The corresponding <u>estimator</u> is

$$A = \frac{2}{t^2}\bar{X}_n$$

To check whether this estimator is unbiased, we calculate

$$E[A] = E\left[\frac{2}{t^2}\bar{X}_n\right] = \frac{2}{t^2}E[\bar{X}_n] = \frac{2}{t^2}E[X_i]$$

So we need expectation of X_i.

$$E[X_i] = E\left[\frac{1}{2}a(t+U_i)^2 + V_i\right]$$

$$= \frac{1}{2}aE[(t+U_i)^2] + E[V_i]$$

$\to 0$ as $V_i \sim N(0,\sigma_v^2)$

$E[V_i] = 0$

$$= \frac{1}{2}a\left(Var(t+U_i) + \left(E[t+U_i]\right)^2\right) + 0$$

$$= \frac{1}{2}a\left(Var(U_i) + \left(t + E[U_i]\right)\right)$$

$\to 0$ as $U_i \sim N(0,\sigma_v^2)$ by Thm 7.25 and linearity of expectation

$$= \frac{1}{2}a\left(Var(U_i) + t^2\right)$$

$\hookrightarrow \sigma_u^2$ as $U_i \sim N(0,\sigma_u^2)$

$\hookrightarrow Var(U_i)$

$$\Rightarrow E[X_i] = \frac{1}{2} a(\sigma_U^2 + t^2)$$

This gives

$$E[A] = \frac{2}{t^2} E[X_i] = a \cdot \frac{\sigma_U^2 + t^2}{t^2} \neq a$$

So the estimator is <u>unbiased</u>.
Taking the average is going to consistently under-estimate/overestimate the value of a.

Luckily we can fix by rescaling the estimator:

$$\bar{A} = \frac{t^2}{\sigma_U^2 + t^2} A = \frac{2}{\sigma_U^2 + t^2} \bar{X}_n$$

is unbiased.

Example 19.3: (R.A. Fisher 1925)

Leaves of maize plants can be divided into 4 types:

1) Starchy-green
2) starchy white
3) sugary green
4) Sugary white

In an experiment in which $n = 3839$ plants were grown, $n_1 = 1997$, $n_2 = 906$, $n_3 = 904$, $n_4 = 32$. These 4 numbers constitute our dataset.

We model the dataset with random variables N_1, N_2, N_3, N_4. According to genetic theory the types occur with probability

According to generic theory, the types occur with probabilities

$$p_1 = \frac{\theta + 2}{4} \qquad p_2 = p_3 = \frac{1-\theta}{4} \qquad p_4 = \frac{\theta}{4}$$

respectively where $0 < \theta < 1$

This implies that the number of plants of N_i of type i is binomially distributed with parameter p_i,

$$N_i \sim \mathrm{Bin}(n \ p_i)$$

However in this example the random variables are <u>not independant.</u>

Instead their joint distribution is the multinomial distribution,

$$(N_1, N_2, N_3, N_4) \sim \mathrm{Mult}(n, p_1, p_2, p_3, p_4)$$

The joint probability mass function is

$$p_{N_1,N_2,N_3,N_4}(n_1, n_2, n_3, n_4) = P(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_4 = n_4)$$
$$= \frac{n!}{n_1! n_2! n_3! n_4!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4}$$

The model parameter is θ

Given the parameters above, we need to find an estimator θ.

For example, lets take

$$T_4 = \frac{4}{n} N_4$$

$$E[T_4] = E\left[\frac{4}{n} N_4\right] = \frac{4}{n} E[N_4]$$

$$= \frac{4}{n}\left(n\frac{\theta}{4}\right) = \theta$$

Hence T_4 is an unbiased estimator.

On our dataset, this leads to the estimate for θ of

$$\theta \approx t_4 = \frac{4}{n} \cdot n_4 = \frac{4}{3839} \cdot 32 \approx 0.033$$

The next suggestion was to use an estimator

$$T_1 = \frac{4}{n} N_1 - 2$$

$$E[T_1] = E\left[\frac{4}{n} N_1 - 2\right] = \frac{4}{n} E[N_1] - 2$$

$$= \frac{4}{n}\left(n \cdot \frac{1}{4}(\theta + 2)\right) - 2 = \theta$$

Hence T_1 is an unbiased estimator.

On our dataset, this estimator leads to an estimate for θ of

$$\theta \approx t_1 = \frac{4}{n} n_1 - 2 = \frac{4}{3839} \cdot 1997 - 2 \approx 0.081$$

The values predicted by T_1 and T_4 are different. which one should we believe more?

↳ To decide this, we should consider which estimator we should expect to have a smaller error.

Defn 19.4: Let T be an estimator. The mean squared error of T is the number

$$\boxed{MSE(T) = E[(T-\theta)^2]}$$

Note that if estimator of T is unbiased, then the means square error of T is equal to the variance of T.

$$\boxed{MSE(T) = E[(T-E[T])^2] = Var(T)}$$

Example: 19.3 (continued) Calculating variances of 2 estimators T_1 and T_4:

$$\mathrm{Var}(T_4) = \mathrm{Var}\left(\frac{4}{n} N_4\right) = \frac{16}{n^2}\mathrm{Var}(N_4)$$

$$= \frac{16}{n^2} n p_4 (1-p_4) \quad \text{(var of binomial)}$$

$$= \frac{16}{n} \cdot \frac{\theta}{4}\left(1 - \frac{\theta}{4}\right)$$

$$= \frac{1}{n}\theta(4-\theta)$$

$$\mathrm{Var}(T_1) = \mathrm{Var}\left(\frac{4}{n} N_1 - 2\right)$$

$$= \frac{16}{n^2}\mathrm{Var}(N_1) \quad \text{(Thm 7.25)}$$

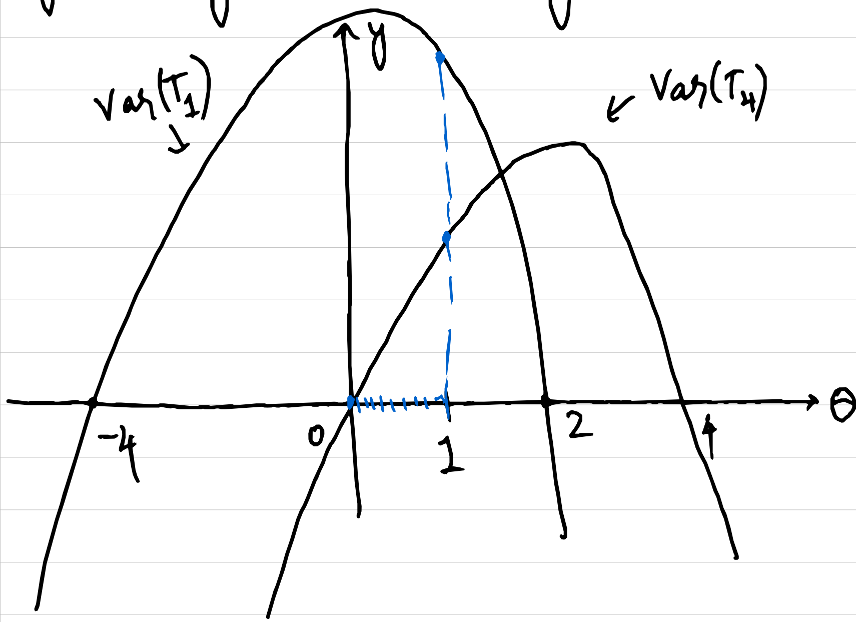
$$= \frac{16}{n^2} \cdot n p_1 (1-p_1)$$

$$= \frac{16}{n} \cdot \frac{\theta+2}{4} \cdot \left(1 - \frac{\theta+2}{4}\right) = \frac{1}{n}(\theta+4)(2-\theta)$$

So variances in each case is a quadratic function of θ.

We do not know a priori what value of θ is other than it lies between 0 and 1.

By plotting the variances against θ,



We can see that in interval $0 < \theta < 1$,

$$Var(T_4) < Var(T_1)$$

Hence T_4 is the better estimator as it has a smaller mean squared error

A more complicated estimator

$$T_{14} = \frac{(T_1 + T_4)}{2}$$

$$E[T_{14}] = E\left[\frac{T_1 + T_4}{2}\right] = \frac{1}{2} E[T_1 + T_4]$$

$$= \frac{1}{2}\left(E[T_1] + E[T_4]\right)$$

$$= \frac{1}{2}(\theta + \theta) = \theta$$

$\Rightarrow E[T_{14}] = \theta \Rightarrow T_{14}$ is unbiased.

To calculate the mean squared error, since T_{14} is unbiased, we need to calculate its variance.

$$Var(T_{14}) = Var\left[\frac{1}{2}(T_1 + T_4)\right]$$

$$= \frac{1}{4} Var(T_1 + T_4) \qquad \text{Thm 7.25}$$

$$= \frac{1}{4}\left[Var(T_1) + Var(T_4) + Cov(T_1, T_4)\right]$$

To calculate covariance $Cov(T_1, T_4)$, we need to use the joint distribution of N_1 and N_4.

Doing the calculation directly from the joint probability mass function will be tedious

So we use our trick of using indicator random variables: Introduce Y_{ai} so that

$$Y_{ai} = \begin{cases} 1 & \text{if ith leaf is of type a} \\ 0 & \text{otherwise} \end{cases}$$

Then

$$N_a = \sum_{i=1}^{n} Y_{ai}$$

For each leaf type a, the indicator random variable Y_{ai} form an iid sample from a Bernoulli distribution $Ber(p_a)$

Leaves from different plants are independant hence Y_{ai} is independant from Y_{bj} for all a and b if $i \neq j$
Each specific plant i can only have a single leaf type.

Each specific plant i can only have a single leaf type.

Hence if $Y_{ai}=1$ for some type a, then $Y_{bi}=0$ whenever $b\neq a$.

So $Y_{ai}Y_{bi}=0$ if $a\neq b$

With this more detailed specification of the dependance and independance among its variables we can now calculate the covariance.

We begin by calculating

$$E[N_1N_4] = E\left[\sum_{i=1}^{n} Y_{1i} \sum_{j=1}^{n} Y_{4j}\right]$$

$$= E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} Y_{1i}Y_{4j}\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} E[Y_{1i}Y_{4j}]$$

$$= \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} E[Y_{1i} Y_{4j}] + \sum_{i=1}^{n} E[Y_{1i} Y_{4j}]$$

$\Rightarrow$

$$E[N_1 N_4] = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} E[Y_{1i} Y_{4j}] + \sum_{i=1}^{n} E[Y_{1i} Y_{4j}]$$

Here we first used linearity of expectation and then we split up the double sum to where both variables refer to same plant ($i = j$) and those where they refer to different plants ($i \neq j$)

In the second summation, we can use that the ith leave cannot be at the same time of type 1 and type 4 hence

$$Y_{1i} Y_{4i} = 0$$

For the first summation; we can use independance of outcomes for the different plant

$$Y_{ai} \perp\!\!\!\perp Y_{bj} \Rightarrow E[Y_{ai} Y_{bj}] = E[Y_{ai}]E[Y_{bj}]$$

$$\forall i \neq j$$

So

$$E[N_1 N_4] = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} E[Y_{1i} Y_{4j}] + \sum_{i=1}^{n} E[Y_{1i} Y_{4j}]$$

$$= \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} E[Y_{1i}] E[Y_{4j}] + 0$$

$$= \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} p_1 p_4$$

$$= \sum_{i=1}^{n} (n-1) p_1 p_4 = n(n-1) p_1 p_4$$

Therefore

$$E[N_1 N_4] = n(n-1)\, p_1 p_4$$

We already know that $N_1 \sim Bin(n, p_1)$ and thus $E[N_1] = np_1$
But this also easy to calculate:

$$E[N_1] = E\left[\sum_{i=1}^{n} Y_{1i}\right] = \sum_{i=1}^{n} E[Y_{1i}]$$

$$= \sum_{i=1}^{n} p_1 = np_1$$

$$\Rightarrow E[N_1] = np_1$$

And similarly $E[N_4] = np_4$. This allows us to calculate covariance using Theorem 10.6

$$Cov(N_1, N_4) = E[N_1 N_4] - E[N_1]E[N_4]$$

$$= n(n-1)p_1 p_4 - np_1 np_4$$

$$= -np_1 p_4$$

Using this, we find the covariance

$$\mathrm{Cov}(T_1, T_4) = \mathrm{Cov}\left(\frac{4}{n} N_1 - 2, \frac{4}{n} N_4\right)$$

$$= \frac{16}{n^2} \mathrm{Cov}(N_1, N_4)$$

$$= \frac{-16}{n} \frac{\theta+2}{4} \frac{\theta}{4} = \frac{-1}{n} (\theta+2)\theta$$

The variance of estimator T_{14} is

$$\mathrm{Var}(T_{14}) = \frac{1}{4}\left(\mathrm{Var}(T_1) + \mathrm{Var}(T_4) + \mathrm{Cov}(T_1 T_4)\right)$$

$$= \frac{1}{4n}\left((\theta+2)(2-\theta) + \theta(4-\theta) - 2(\theta+2)\theta\right)$$

$$= \frac{1}{n}(1-\theta)(1+\theta)$$

$$\Rightarrow \mathrm{Var}(T_{14}) = \frac{1}{n}(1-\theta)(1+\theta)$$

✱ $\mathrm{Cov}\left(\frac{4}{n}N_1 - 2, \frac{4}{n}N_4\right)$

$$= E\left[\left(\frac{4}{n}N_1 - 2\right)\cdot\left(\frac{4}{n}N_4\right)\right] - E\left[\frac{4}{n}N_1 - 2\right]E\left[\frac{4}{n}N_4\right]$$

$$= E\left[\frac{16}{n^2}N_1N_4 - \frac{8}{n}N_4\right] - \left(\frac{4}{n}E[N_1] - 2\right)\frac{4}{n}E[N_4]$$

$$= \frac{16}{n^2}E[N_1N_4] - \cancel{\frac{8}{n}E[N_4]} - \left(\frac{16}{n^2}E[N_1]E[N_2] - \cancel{\frac{8}{n}E[N_4]}\right)$$

$$= \frac{16}{n^2}\left(E[N_1N_4] - E[N_1]E[N_4]\right)$$

$$= \frac{16}{n^2}\mathrm{Cov}(N_1, N_4)$$

In general:

$$\boxed{\mathrm{Cov}(rX + s, tY + u) = rt\,\mathrm{Cov}(X, Y)}$$

↳ proof given in WS7