

Estimating a Proportion

Suppose that there is a population of N items, of which M have Type A and $N - M$ have Type B. Put

$$p = \frac{M}{N} = \text{proportion having Type A.}$$

We take a random sample of size n . Let X be the number of Type A in the sample.

If the sampling is done with replacement then $X \sim \text{Bin}(n, p)$. If the sampling is done without replacement then $X \sim \text{Hg}(n, M, N)$. Usually we assume that n is much smaller than M and $N - M$, so that X is approximately $\text{Bin}(n, p)$.

Let

$$Y = \frac{X}{n} = \text{sample proportion of Type A.}$$

Then

$$\mathbb{E}(Y) = \mathbb{E}\left(\frac{X}{n}\right) = \frac{1}{n} \mathbb{E}(X) = \frac{1}{n} \times np = p$$

and

$$\begin{aligned} \text{Var}(Y) &= \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \text{Var}(X) \\ &= \frac{1}{n^2} \times npq, \quad \text{where } q = 1 - p, \\ &= \frac{pq}{n}. \end{aligned}$$

We do not know the true value of p , so we estimate it using Y . We write

$$\begin{aligned} Y &= \text{estimator} = \text{a random variable} \\ y &= \text{the value of } Y \text{ from our sample} \\ \hat{p} &= \text{the estimated value of } p, \end{aligned}$$

so that $\hat{p} = y$.

Definition Let T be any estimator for p . The *error in estimation* is $T - p$. The *bias* of T is $\mathbb{E}(T - p)$, which is $\mathbb{E}(T) - p$. The estimator T is *unbiased* if its bias is zero; that is, $\mathbb{E}(T) = p$. The *mean squared error* of T , written $\text{MSE}(T)$, is $\mathbb{E}((T - p)^2)$.

Proposition 7 $\text{MSE}(T) = \text{Var}(T) + [\text{Bias}(T)]^2$.

Proof We have

$$\begin{aligned}
 \text{MSE}(T) &= \mathbb{E}[(T - p)^2], && \text{by definition,} \\
 &= \mathbb{E}(T^2 - 2Tp + p^2) \\
 &= \mathbb{E}(T^2) - 2p\mathbb{E}(T) + p^2 \\
 &= \mathbb{E}(T^2) - [\mathbb{E}(T)]^2 + [\mathbb{E}(T)]^2 - 2p\mathbb{E}(T) + p^2 \\
 &= \text{Var}(T) + (\mathbb{E}(T) - p)^2 \\
 &= \text{Var}(T) + [\text{Bias}(T)]^2. \quad \blacksquare
 \end{aligned}$$

It is usually best to use an unbiased estimator, if possible. If T is unbiased, then $\text{MSE}(T) = \text{Var}(T)$. If T_1 and T_2 are both unbiased estimators for p , then we say that T_1 is *better* than T_2 if $\text{Var}(T_1) < \text{Var}(T_2)$.

If Y is the sample proportion of Type A, then $\mathbb{E}(Y) = p$, so Y is an unbiased estimator of p . Also, $\text{Var}(Y) = pq/n$, so the standard deviation of Y is $\sqrt{pq/n}$.

So

$$\begin{aligned}
 \text{larger sample} &\equiv \text{larger } n \\
 &\rightarrow \text{smaller standard deviation} \\
 &\equiv \text{smaller MSE, because } Y \text{ is unbiased,} \\
 &\rightarrow \text{better estimate of } p.
 \end{aligned}$$

We do not know p , so we do not know the standard deviation, so we estimate it by

$$\sqrt{\frac{\hat{p}\hat{q}}{n}},$$

where $\hat{q} = 1 - \hat{p}$. This estimate is called the *standard error*, which is often abbreviated to S.E. or s.e.