Proportions, Differences of Means, and Sample Variance

Numerical Statistics Fall, 2021

Jake Underland

2021-10-14

Contents

The Sampling Distribution of Proportions

Bernoulli Distribution

Let $(X_1, X_2, \dots, X_n) \overset{iid}{\sim} Ber(p)$ drawn from an infinite ($\approx$ with replacement) population.

$$\begin{cases} P(X_i = 1) = p \\ P(X_i = 0) = 1 - p \end{cases} \quad (i = 1, 2, \dots, n)$$

The probabilities can be combined into one as

$$P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i} \text{ for } (x_i = 0, 1)$$

So, the joint probability can be expressed as:

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) &= \prod_{i=1}^{n} P(X_i = x_i) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{x_1+x_2+\cdots+x_n}(1-p)^{n-x_1-x_2-\cdots-x_n} \\ &= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \end{aligned}$$

Expectation and Variance:

$$\begin{aligned}
E(X) &= \sum_{x=0}^{1} P(X=x)x \\
&= 1 \times p + 0 \times (1-p) \\
&= p \\
Var(X) &= \sum_{x=0}^{1} P(X=x)x^2 - E(X)^2 \\
&= p - p^2 \\
&= p(1-p)
\end{aligned}$$

Binomial Distribution

Let $(X_1, X_2, \dots, X_n) \overset{iid}{\sim} Ber(p)$. Then, $Y = X_1 + X_2 + \cdots + X_n \sim Bin(n,p)$.

$$P(Y=x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \dots, n$$

Binomial Theorem:

$$(a+b)^n = \sum_{i=1}^{n} \binom{n}{x} a^x b^{n-x}$$

where the right hand side represents the probability of a particular value of x, or a particular combination of number of successes and number of failures.

Expectation and Variance:

$$\begin{aligned}
E(Y) &= E(\sum_{i=0}^{n} X_i) \text{ where } X_i \overset{iid}{\sim} Ber(p) \\
&= \sum_{i=0}^{n} E(X_i) \\
&= \sum_{i=0}^{n} p \\
&= np \\
Var(Y) &= Var(\sum_{i=0}^{n} X_i) \text{ where } X_i \overset{iid}{\sim} Ber(p) \\
&= \sum_{i=0}^{n} Var(X_i) \dots Independene \\
&= np(1-p)
\end{aligned}$$

Proportions

Suppose $(X_1, X_2, \dots, X_n) \overset{iid}{\sim} Ber(p)$. Then, the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is an unbiased estimator of the parameter p:

$$\begin{aligned} E(\bar{X}) &= E(\frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} p \\ &= p \; \square \end{aligned}$$

The variance of the sample mean $\bar{X}$ is:

$$\begin{aligned} Var(\bar{X}) &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \ldots Independence \\ &= \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) \\ &= \frac{p(1-p)}{n} \; \square \end{aligned}$$

Sampling Distribution of the Differences of Means

$$\begin{aligned} \Pi_1 &: \text{ A population with mean } \mu_1 \text{ and standard deviation } \sigma_1 \\ \Pi_2 &: \text{ A population with mean } \mu_2 \text{ and standard deviation } \sigma_2 \end{aligned}$$

We examine the following cases where the populations are infinite (sampling is with replacement) or populations are finite (sampling is without replacement).

Infinite Population (Sampling With Replacement)

A sample of size n_1 drawn from Π_1 and n_2 drawn from Π_2. Then,

$$\begin{aligned} \bar{X_1} &= \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \\ \bar{Y_2} &= \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i \end{aligned}$$

Where

$$\begin{aligned} E(X_1) &= \mu_1 \\ E(Y_2) &= \mu_2 \end{aligned}$$

and

$$\begin{aligned} Var(X_1) &= \frac{\sigma_1^2}{n_1} \\ Var(Y_2) &= \frac{\sigma_2^2}{n_2} \end{aligned}$$

Then, the expectation of $\bar{X_1} - \bar{Y_2}$ is

$$\mu_{\bar{X_1}-\bar{Y_2}} = E(\bar{X_1} - \bar{Y_2}) = E(\bar{X_1}) - E(\bar{Y_2}) = \mu_1 - \mu_2$$

The variance is

$$\sigma_{\bar{X}_1-\bar{Y}_2} = Var(\bar{X}_1 - \bar{Y}_2) = Var(\bar{X}_1) + Var(\bar{Y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Finite Population (Sampling Without Replacement)

A sample of size n_1 drawn from Π_1 and n_2 drawn from Π_2. Then,

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$$
$$\bar{Y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

Where

$$E(X_1) = \mu_1$$
$$E(Y_2) = \mu_2$$

and

$$Var(X_1) = \frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right)$$
$$Var(Y_2) = \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)$$

Then, the expectation of $\bar{X}_1 - \bar{Y}_2$ is

$$\mu_{\bar{X}_1-\bar{Y}_2} = E(\bar{X}_1 - \bar{Y}_2) = E(\bar{X}_1) - E(\bar{Y}_2) = \mu_1 - \mu_2$$

The variance is

$$\sigma_{\bar{X}_1-\bar{Y}_2} = Var(\bar{X}_1 - \bar{Y}_2) = Var(\bar{X}_1) + Var(\bar{Y}_2) = \frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)$$

The Sample Variance

Suppose the following:

$$\begin{aligned}
S^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\bar{X}} + \frac{1}{n}\sum_{i=1}^{n}\bar{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}^2 + \bar{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2
\end{aligned}$$

Then, we have a biased estimator of variance, as can be seen below:

$$
\begin{aligned}
E(S^2) &= E(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E\left[\frac{1}{n}\sum_{i=1}^{n} X_i \frac{1}{n}\sum_{j=1}^{n} X_j\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - \frac{1}{n^2} E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - \frac{1}{n^2} E\left[\underbrace{\sum_{i=1}^{n} X_i^2}_{n} + \underbrace{\sum_{i=1}^{n}\sum_{j\neq i}^{n} X_i X_j}_{n^2-n}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - \frac{1}{n^2}\sum_{i=1}^{n} E(X_i^2) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} E(X_i)E(X_j) \ldots Independence \\
&= \frac{n-1}{n} E(X_i^2) - \frac{n-1}{n} E(X_i)^2 \ldots Identical \\
&= \frac{n-1}{n}[E(X_i^2) - E(X_i)^2] \\
&= \frac{n-1}{n} Var(X_i)
\end{aligned}
$$

In order to obtain an unbiased estimator of variance,

$$
\begin{aligned}
E(S^2) &= \frac{n-1}{n} Var(X_i) \\
\implies E(\frac{n}{n-1}S^2) &= \frac{n}{n-1}E(S^2) = Var(X_i) \\
\frac{n}{n-1}S^2 &= \frac{n}{n-1} \cdot \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \ldots \square
\end{aligned}
$$

So, the sample variance is

$$
\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2
$$