Lecture 12: Populations, Random Samples, and Statistics

Mathematical Statistics I, MATH 60061/70061

Tuesday October 19, 2021

Reference: Casella & Berger, 5.1-5.2

Populations and random samples

In statistics we often think of having a **population**, or a population distribution $f(x)$.

The random variables $X_1, \ldots, X_n$ are called a **random sample** of size n from a population with distribution $f(x)$ if

1. $X_1, \ldots, X_n$ are independent.
2. The marginal PDF or PMF of each X_i is the same function $f(x)$.

Alternatively, $X_1, \ldots, X_n$ are called **independent and identically distributed** (iid) random variables with PDF or PMF $f(x)$.

A random sample is viewed as sampling from an *infinite population* or from a *finite population with replacement* so that X_i's are independent.

Distribution of a random sample

- The joint PDF or PMF of a random sample $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n) = f(x_1)f(x_2) \cdot \cdots \cdot f(x_n) = \prod_{i=1}^{n} f(x_i).$$

- If the population PDF/PMF is a member of a **parametric family** with PDF/PMF given by $f(x \mid \theta)$, then the joint PDF/PMF is

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

Sampling without replacement from a finite population

Sometimes we consider sampling *without replacement* from a finite population $\{x_1, ..., x_N\}$. For example, a survey of n persons from a population of size N. This is sometimes called **simple random sampling**.

When sampling without replacement, $X_1, ..., X_n$ cannot be a random sample.

Let x and y be distinct elements of $\{x_1, ..., x_N\}$, then

- $P(X_2 = y \mid X_1 = y) = 0$
- $P(X_2 = y \mid X_1 = x) = 1/(N-1)$

So X_1 and X_2 are *not* independent.

Sampling without replacement from a finite population

In a simple random sample, X_i's are *dependent*, but each of them has the *same* marginal distribution. By the law of total probability

$$P(X_2 = x) = \sum_{i=1}^{N} P(X_2 = x \mid X_1 = x_i) P(X_1 = x_i),$$

where for one value of the index, say k, $x = x_k$

$$P(X_2 = x \mid X_1 = x_k) = 0,$$

and for all other $j \neq k$,

$$P(X_2 = x \mid X_1 = x_j) = 1/(N-1).$$

Thus,

$$P(X_2 = x) = (N-1)\left(\frac{1}{N-1}\frac{1}{N}\right) = \frac{1}{N}.$$

The dependence becomes weak when the population size N is much larger than the sample size n.

Statistics

Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a **statistic**. The probability distribution of a statistic Y is called the **sampling distribution** of Y.

A statistic T cannot be a function of a parameter, and it must only depend on the data. Also, T must be defined for all possible data values.

Some important statistics

- Sample mean

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Sample variance

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

- Sample standard deviation

$$S = \sqrt{S^2}$$

Properties of observed sample mean and variance

Let $x_1, \ldots, x_n$ be any numbers and $\bar{x} = (x_1 + \cdots + x_n)/n$. Then

1. $(n-1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$
2. $\min_a \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2\bar{x}x_i) \\
&= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i \\
&= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\end{aligned}
$$

$\min_a \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2$

Write $\sum_{i=1}^n (x_i - \bar{x})^2$ as $\sum_{i=1}^n ((x_i - a) - (\bar{x} - a))^2$. Let $y_i = x_i - a$ and apply the first result:

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \sum_{i=1}^n y_i^2 - n\bar{y}^2 \\ &= \sum_{i=1}^n (x_i - a)^2 - n(\bar{x} - a)^2.\end{aligned}$$

So

$$\begin{aligned}\sum_{i=1}^n (x_i - a)^2 &= \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - a)^2 \\ &\geq \sum_{i=1}^n (x_i - \bar{x})^2,\end{aligned}$$

and the lower bound is attained when $a = \bar{x}$.

Sum of random variables

Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $E(g(X_1))$ and $\mathrm{Var}(g(X_1))$ exist. Then

$$E\left(\sum_{i=1}^{n} g(X_i)\right) = \sum_{i=1}^{n} E(g(X_i)) = n \cdot E(g(X_1))$$

and

$$\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = \sum_{i=1}^{n} \mathrm{Var}(g(X_i)) = n \cdot \mathrm{Var}(g(X_1)).$$

(Functions of independent random variables are independent.)

Sample mean and sample variance

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then

1. $E(\bar{X}) = \mu$,
2. $\mathrm{Var}(\bar{X}) = \sigma^2/n$,
3. $E(S^2) = \sigma^2$.

The first two properties can be proved using linearity and the independence property of variance.

Expectation of sample variance

We have $\sum (X_i - \mu)^2 = \sum (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$. Therefore

$$
\begin{aligned}
E(S^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \\
&= \frac{1}{n-1}\left(E\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - nE(\bar{X} - \mu)^2\right) \\
&= \frac{1}{n-1}(n\sigma^2 - n\sigma^2/n) \\
&= \sigma^2.
\end{aligned}
$$

Expectation of sample variance

We have $\sum(X_i-\mu)^2 = \sum(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2$. Therefore

$$\begin{aligned}
E(S^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right) \\
&= \frac{1}{n-1}\left(E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) - nE(\bar{X}-\mu)^2\right) \\
&= \frac{1}{n-1}(n\sigma^2 - n\sigma^2/n) \\
&= \sigma^2.
\end{aligned}$$

So, the statistic $\bar{X}$ is an **unbiased** estimator of μ, and S^2 is an **unbiased** estimator of σ^2.

Question: Is S also an unbiased estimator of σ?