Lecture 19, Mar 1, 2023

Sampling

- Often we can't measure the entire population, so we instead examine a subset
- How representative is this subset to the population?
- A sample could be thought of as actual data $x_1, \cdots, x_n$, with the assumption that each x_i is a realization of an independent random variable X_i
  - Since each of these measurements is done separately, they are separate random variables, but we assume that they all have the same distribution as the population

Definition

The *sample mean* (or realized/empirical sample mean) is defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The *random variable* of the mean is defined as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- $\bar{X}$ itself is a random variable
- Assuming all X_i are independent identically distributed (IID) with mean μ, then we can find $E[\bar{X}]$
  - $$\begin{aligned} E[\bar{X}] &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu \end{aligned}$$

Definition

The *sample variance* is defined as

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

The *sample variance random variable* is defined as

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

- Why $n-1$ in the denominator?
  - The mean itself $\bar{x}$ is a function of the data points, so it is not independent; normally it would not have an effect on the variance, but in the case of sample variance, it has a contribution, which we eliminate via the $n-1$ in the denominator
  - Assume $\mathrm{var}(X_i) = \sigma^2$ for all i (that is, each RV has the same variance); we would like to get $E[S^2] = \sigma^2$

  - $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
    $= \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - 2X_i\bar{X} + \bar{X}^2$
    $= \frac{1}{n-1}\left(n\bar{X}^2 - 2n\bar{X}^2 + \sum_{i=1}^{n} X_i^2\right)$
    $= \frac{1}{n-1}\left(-n\bar{X}^2 + \sum_{i=1}^{n} X^2\right)$
  - $E[S^2] = \frac{1}{n-1}\left(-nE[\bar{X}^2] + E\left[\sum_{i=1}^{n} X_i^2\right]\right)$
    $= \frac{1}{n-1}\left(-nE[\bar{X}^2] + \sum_{i=1}^{n} E[X_i^2]\right)$
    $= \frac{1}{n-1}\left(-n\left(E[\bar{X}]^2 + \mathrm{var}(\bar{X})\right) + \sum_{i=1}^{n} \mu^2 + \sigma^2\right)$
    $= \frac{1}{n-1}\left((-n\mu^2 + \mathrm{var}(\bar{X})) + n(\mu^2 + \sigma^2)\right)$
    $= \frac{1}{n-1}\left(-n\,\mathrm{var}(\bar{X}) + n\sigma^2\right)$
    $= \frac{1}{n-1}\left(-n\,\mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + n\sigma^2\right)$
    $= \frac{1}{n-1}\left(-n\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) + n\sigma^2\right)$
    $= \frac{1}{n-1}\left(-\frac{1}{n}n\sigma^2 + n\sigma^2\right)$
    $= \frac{1}{n-1}\left((n-1)\sigma^2\right)$
    $= \sigma^2$
- We call this an *unbiased estimator* of the variance

Histograms

- In the case of a discrete RV, the x axis is the possible outcomes, the y axis is the number of times each outcome is observed
  - As the number of samples increases, the histogram divided by the sample size approaches the PMF
- In the case of a continuous RV we make bins to contain ranges of observations
  - As the number of samples increases and the bin size approaches 0, the histogram divided by the sample size approaches the PDF