ECE 3075A
Random Signals

Lecture 19

Introduction to Statistics & Sampling - II

School of Electrical and Computer Engineering
Georgia Institute of Technology
Fall, 2003

Exercise 4-2.1

An endless production line is turning out solid-state diodes and every 100th diode is tested for reverse current I_{-1} and forward current I_1 at diode voltages of –1 and 1, respectively.

1. If I_{-1} has a true mean value of 10^{-6} and a variance of 10^{-12}, how many diodes must be tested to obtain a sample mean whose standard deviation is 5% of the true mean?
2. If I_1 has a true mean value of 0.1 and a variance of 0.0025, how many diodes must be tested to obtain a sample mean whose standard deviation is 2% of the true mean?

$$\frac{\sigma_{I_{-1}}^2}{n_{I_{-1}}} = \frac{10^{-12}}{n_{I_{-1}}} = (10^{-6}\times 0.05)^2, \quad \therefore n_{I_{-1}} = 400$$

$$\frac{\sigma_{I_1}^2}{n_{I_1}} = \frac{0.0025}{n_{I_1}} = (0.1\times 0.02)^2, \quad \therefore n_{I_1} = 625$$

With n = 625 $\quad \text{var}(\hat{\bar{I}}_{-1,n}) = 10^{-12}/625 = 16\times 10^{-16} \quad \therefore \text{std}(\hat{\bar{I}}_{-1,n}) = 4\times 10^{-8}$

$$\text{std}(\hat{\bar{I}}_{1,n}) = 2\times 10^{-3}, \text{ stays unchanged}$$

Sample Variance

- Do not confuse variance of sample mean with sample variance.
- Sample mean is an estimate of the mean of the population.
- A good estimate of the population mean does not readily represent good knowledge of the statistical properties of the population. Need to know some higher order statistics.
- What would be a reasonable estimate of the **variance of the population**?

With $X_i, i = 1, 2, \cdots, n$ being the r.v.s in the sample, the sample variance is defined as

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

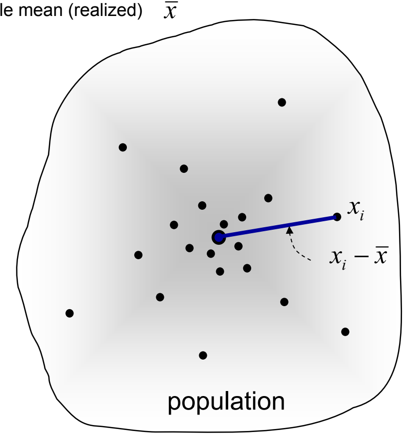
which is the average of the square of the difference between each r.v. and the sample mean.

Sample Mean and Sample Variance

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

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

- Sample item (sampled data) x_i
- Sample mean (realized) $\bar{x}$



Sample variance can be considered as the average **distance** between each random variable in the sample and the sample mean. It is thus an indication how dispersive the population is.

Centroid & Minimization of Sample Variance

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

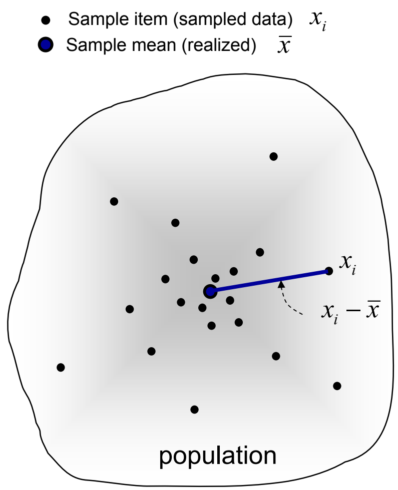
$$\frac{d}{dx}D(x) = \frac{d}{dx}\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(-2)(x_i - x) = 0$$

At $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$, D is minimum.

Note, $\left.\frac{d^2}{dx^2}D(x)\right|_{x=\bar{x}} = \frac{2}{n} > 0$

$\bar{x}$ is also called the centroid of $\{x_i\}$.



Expectation of Sample Variance

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

$$= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - \frac{2X_i}{n}\sum_{j=1}^{n}X_j + \frac{1}{n^2}\sum_{j=1}^{n}X_j\sum_{k=1}^{n}X_k\right)$$

$$E[S^2] = \left(\frac{1}{n}\sum_{i=1}^{n}E[X_i^2]\right) - \left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E[X_iX_j]\right)$$

$$= \overline{X^2} - \frac{1}{n^2}\left[(n^2 - n)\bar{X}^2 + n\overline{X^2}\right] = \left(\frac{n-1}{n}\right)\overline{X^2} - \left(\frac{n-1}{n}\right)\bar{X}^2 = \frac{n-1}{n}\sigma^2$$

The expected value of the sample variance is not the true variance, and thus a biased estimate. We make the following adjustment

$$\tilde{S}^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\bar{X}})^2, \quad \therefore\ E[\tilde{S}^2] = \frac{n}{n-1}E[S^2] = \sigma^2$$

More on Sample Variance

When the population is not large, $E[S^2] = \frac{N}{N-1}\frac{n-1}{n}\sigma^2$

We can define the following to remove the bias:

$$\tilde{S}^2 = \frac{N-1}{N}\frac{n}{n-1}S^2 = \frac{N-1}{N}\frac{n}{n-1}\sum_{i=1}^{n}(X_i - \hat{\bar{X}})^2 \quad \rightarrow E[\tilde{S}^2] = \sigma^2$$

The variance of the sample variance is

$$\mathrm{var}(S^2) = \frac{\mu_4 - (\sigma^2)^2}{n}$$

where $\mu_4 = E\left[(X - \bar{X})^4\right]$ the 4th central moments of the population

And $\mathrm{var}(\tilde{S}^2) = \frac{n(\mu_4 - \sigma^4)}{(n-1)^2}$

Distributions of Parameter Estimates

- Without knowing the true distribution, we may still want to have some idea about the distribution of the parameter estimates.
- If $X_i, i = 1, 2, \cdots, n$ are Gaussian and independent with mean $\bar{X}$ and variance σ^2, then the sample mean is also Gaussian and the normalized random variable $Z = \frac{\hat{\bar{X}} - \bar{X}}{\sigma/\sqrt{n}}$ is Gaussian with zero mean and unit variance. This true regardless of the size of the sample.
- If $X_i, i = 1, 2, \cdots, n$ are not Gaussian, then Z is asymptotically Gaussian as $n \rightarrow \infty$ according to the central limit theorem. As a rule of thumb, the asymptotic result is reached when $n \geq 30$.
- When n is not large enough to ensure normality, then the following normalized random variable Z has a Student's t distribution: $Z = \frac{\hat{\bar{X}} - \bar{X}}{\tilde{S}/\sqrt{n}} = \frac{\hat{\bar{X}} - \bar{X}}{S/\sqrt{n-1}}$

