ECE 3075A
Random Signals

Lecture 11
Introduction to Statistics & Sampling

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

Probability Theory & Statistics

- Probability Theory
  - A mathematical framework for dealing with uncertain aspects of the physical reality, the truth of which may otherwise take infinite effort to characterize without it. (Think about the example of radio noise that impairs reception, or how long two light bulbs connected in series or in parallel are expected to produce light.)
  - The framework is built upon set theory and set operations (Borel field and σ-algebra), and when extended to random variables it is practically supported by the Reimann integral (from the more general Lebesgue integral).

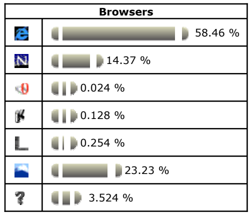
$$\int_{\Lambda(x)} g(x)\,d\mu(x) \quad \rightarrow \quad \int_{\Lambda(x)} g(x)\,f(x)\,dx$$

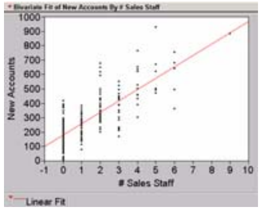
Probability Theory & Statistics

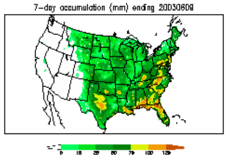
- Statistics
  - The science of assembling, sorting, tabulating, and analyzing **data** or a set of facts, in order to find the relationship or implicit regularity among them for other use.
    - Descriptive statistics: collecting, grouping, and presenting data for easy understanding and assimilation – tie to instinct and perception (or contrast to them)
    - Inductive Statistics: inferring some aspect of the truth about the origin or the environment where the data came from.

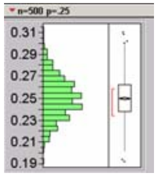
Graphic Methods of Everyday Statistics

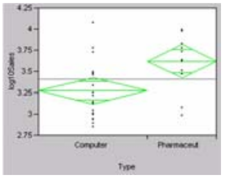
- In statistics, we use data and some derived results to convey a message; graphic representation of the data or results often makes interpretation easy – a picture is worth a thousand words or millions of data items. Examples:

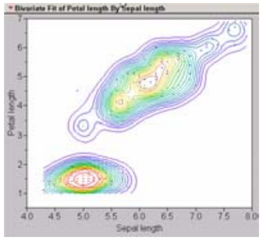












Scope of Statistics Here

- **Sampling theory**: deals with problems in selecting data set that is manageably small but enough to allow meaningful inference
- **Estimation theory**: concerned with figuring out the value of key parameters from data
- **Hypothesis testing**: to verify if an assertion is true based on the evidence provided in the data set
- **Regression or curve fitting**: to find mathematical expressions that best represent or characterize the data
- **Analysis of variance**: to assess the significance of variation in data and how they relate to reality

Sampling

- A population is a general collection of data whose statistically behavior is being studied. A population has size denoted by N; N may be infinitely large.
- A sample, or random sample, is part of the population that has been selected at random for the study; its size, i.e., the sample size, denoted by n, $n << N$, is the number of items or pieces of data selected for study. A sample is denoted by $\{x_i\}_{i=1}^n$.

The average of the data in the sample is called the **sample mean**.

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

We can also view the sampling as a realization of an experiment involving random variables, $X_1, X_2, X_3, \cdots, X_n$, which leads to a new r.v.

$$\hat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad E\left[\hat{\bar{X}}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \bar{X}$$

Sample Mean

- Therefore, the mean of the sample mean is the true mean of the population.
- We can use the sample mean as an estimate of the mean of the population; and since its expected value is identical to the true mean, it is an **unbiased estimator** – that is, it does not contain statistical deviation (for any sample, the deviation may be there, but the statistical average of the deviation is zero) from the true expected value.
- To determine if the sample mean is a good estimator, we need to evaluate its variance – to know how much the sample mean value will fluctuate. An estimator that fluctuate more is not as good as one that fluctuates less, provided that they're all unbiased.

Variance of Sample Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad E\left[\hat{\bar{X}}\right] = \bar{X}$$

$$\mathrm{var}(\hat{\bar{X}}) = E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right] - (\bar{X})^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_i X_j] - (\bar{X})^2$$

$$E[X_i X_j] = \begin{cases} \overline{X^2}, & i = j \\ \bar{X}^2, & i \neq j \end{cases} \qquad \mathrm{var}(\hat{\bar{X}}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_i X_j] - (\bar{X})^2$$

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

That is, the fluctuation of the value of the sample mean is inverse proportional to the sample size; the more items included in the sample, the better the sample mean as an estimate of the mean of the population.

Distribution of Sample Mean

$$\hat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad E\left[\hat{\bar{X}}\right] = \bar{X}, \quad \mathrm{var}\left(\hat{\bar{X}}\right) = \frac{\sigma_X^2}{n}$$

If n is large, $f_{\hat{\bar{X}}}(x) \approx \mathrm{N}\,(\bar{X}, \frac{\sigma_X^2}{n})$ from the central limit theorem

Example: Let $\bar{X} = 10,$ and $\sigma_X^2 = 9$

We select a sample of size n to estimate $\bar{X}$ and hope that the estimate has a standard deviation that is only 1% of the true mean. $\mathrm{var}\left(\hat{\bar{X}}\right) = \frac{\sigma_X^2}{n} = \frac{9}{n} = (0.01 \times 10)^2 = 0.01.$ Therefore, $n = 900$

We can also ask the question: What is the probability that the sample mean is within 1% range of the true mean?

$$\Pr\{9.9 \le \hat{\bar{X}} \le 10.1\} = F_{\hat{\bar{X}}}(10.1) - F_{\hat{\bar{X}}}(9.9) = \phi\left(\frac{10.1-10}{10*0.01}\right) - \phi\left(\frac{9.9-10}{10*0.01}\right)$$

$$= \phi(1) - \phi(-1) = 2\phi(1) - 1 = 2 \times 0.8413 - 1 = 0.6826$$

Weak Law & Strong Law of Large Numbers

Random variables $X_1, X_2, X_3, \cdots, X_n$ are identically distributed (with same mean and same finite variance, of course) and are at least pair-wise statistically independent, then, the sample mean satisfies

Weak Law

$$\lim_{n\to\infty} \Pr\left\{\left|\hat{\bar{X}}_n - \bar{X}\right| < \varepsilon\right\} = 1 \qquad \text{for any } \varepsilon > 0$$

Strong Law

$$\Pr\left\{\lim_{n\to\infty} \hat{\bar{X}}_n = \bar{X}\right\} = 1$$

Sampling without Replacement

- When population is large, $N \to \infty$, selecting sample with and without replacing the item back into the population does not seriously affect the statistical results.
- When population is not as large, sampling with and without replacement would lead to different statistical result. In particular, the variance of the sample mean is now

$$\mathrm{var}(\hat{\bar{X}}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \quad \Leftarrow \text{sampling without replacement}$$

as opposed to

$$\mathrm{var}(\hat{\bar{X}}) = \frac{\sigma^2}{n} \quad \Leftarrow \text{sampling with replacement or } N \to \infty$$

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

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

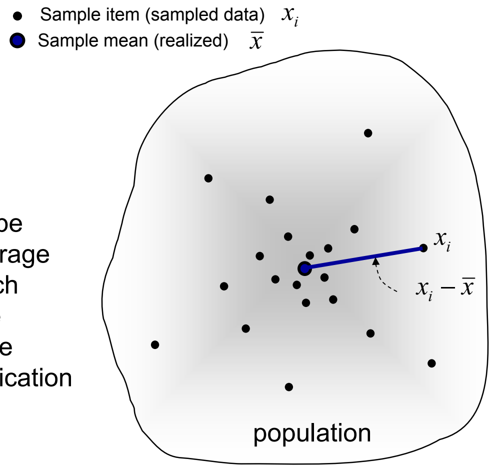
$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\bar{X}})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2$$

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) $\bar{x}$
x_i
$x_i - \bar{x}$
population

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

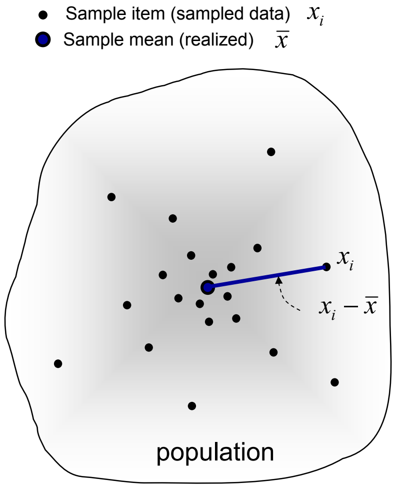
$$\frac{d}{dx}D(x) = \frac{d}{dx}\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2$$

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

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 $\{x_i\}$.


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

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 $n \to \infty$ 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: $$Z = \frac{\hat{\bar{X}} - \bar{X}}{\tilde{S}/\sqrt{n}} = \frac{\hat{\bar{X}} - \bar{X}}{S/\sqrt{n-1}}$$