Summarizing Measured Data

Nelson Fonseca
State University of Campinas

Statistical Concepts

- Mean

$$E[X] = \overline{X} = \int_{-\infty}^{\infty} x f_X(x)\,dx$$

– Second central moment => variance

$$\sigma_x^2 \overset{\Delta}{=} \overline{(X - \overline{X})^2} \overset{\Delta}{=} \overline{X^2} - (\overline{X})^2$$

– Standard deviation (central moment)

$$\sigma_x = \sqrt{\sigma_X^2}$$

– Coefficient of variation

$$C_X \overset{\Delta}{=} \frac{\sigma_X}{\overline{X}}$$

Statistical Concepts

- Covariance of two random variables X_1 and X_2

$$Cov\ (X_1, X_2) = E[(X_1 - E[X_1])\ (X_2 - E[X_2])]$$

$$var\ (X_1 + X_2) = var\ (X_1) + var\ (X_2) + 2Cov(X_1, X_2)$$

$$Corr\ (X_1, X_2) = Cov\ (X_1, X_2)\ /\ (\sigma_1\ \sigma_2)$$

Statistical Concepts

- Quantile – the *x* value at which the CDF takes a value α is called α-quantile or 100α-percentile (x_α)

$$P(x \le x_\alpha) = F(x_\alpha) = \alpha$$

Statistical Concepts

- Median – The 50-percentile (or 0.5-quantile) of a random variable

- Mode – The most likely value, x_i, that has the highest probability p_i or at which the pdf is maximum

Indices of Central Tendencies

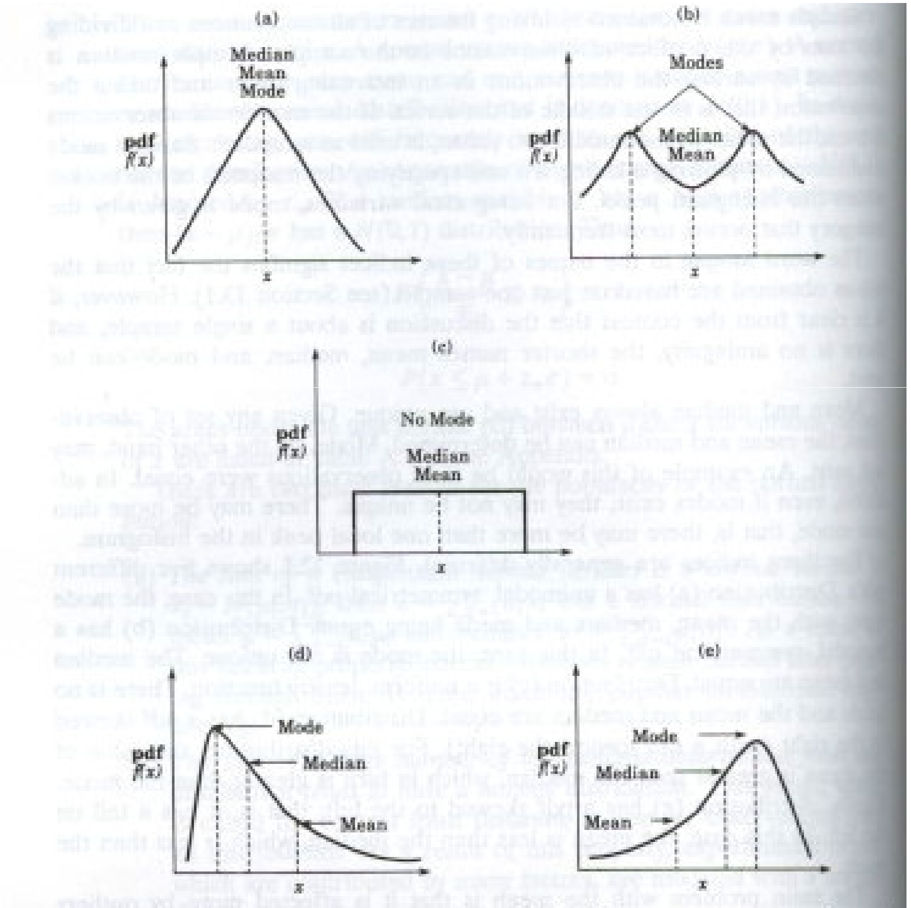

FIGURE 12.1 Five distributions showing relationships among mean, median, and mode.

Indices of Central Tendencies

- Mean:
  - total of all observation is of interest,
  - affected by outlier
  - usefulness depends on the number of samples, variance and skewness (ratio between maximum and minimum values)
- Median and Mode ignores the total information;
- Median and mean always exists, there can be more than one mode;

Mean

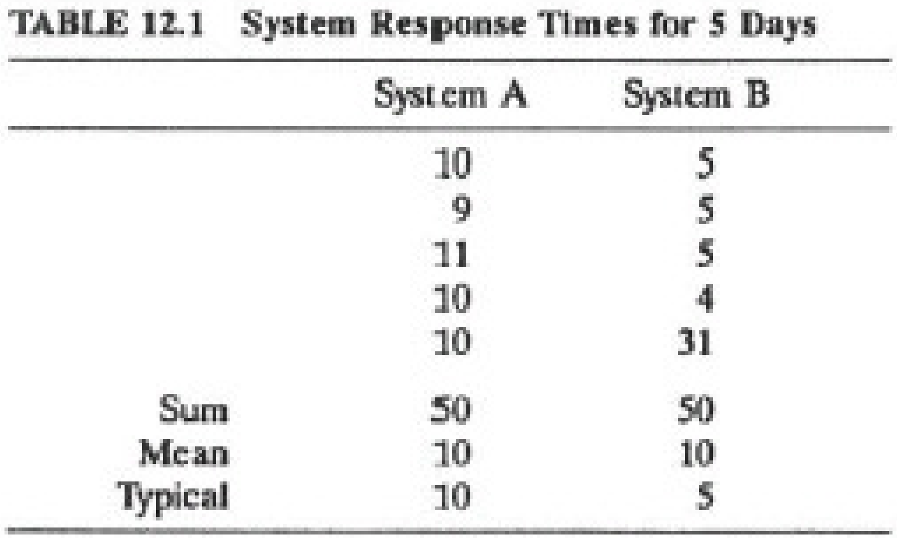

TABLE 12.1 System Response Times for 5 Days

	System A	System B
	10	5
	9	5
	11	5
	10	4
	10	31
Sum	50	50
Mean	10	10
Typical	10	5

Geometric Mean

- Cache hit ratio over several layers of caches
- Cache miss ratios
- Average error rate per hop on a multihop path in a network

$$\dot{x} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

Geometric Mean

- The geometric mean of a ratio is the ratio of the geometric means of the numerator and denominator (physical meaning). The choice of bases does not change the conclusion.

$$gm\left(\frac{x_1}{y_1},\frac{x_2}{y_2}\ldots\frac{x_n}{y_n}\right)=\frac{gm(x_1,x_2,\ldots x_n)}{gm(y_1,y_2,\ldots y_n)}=\frac{1}{gm\left(\frac{y_1}{x_1},\frac{y_2}{x_2},\ldots\frac{y_n}{x_n}\right)}$$

Geometric Mean

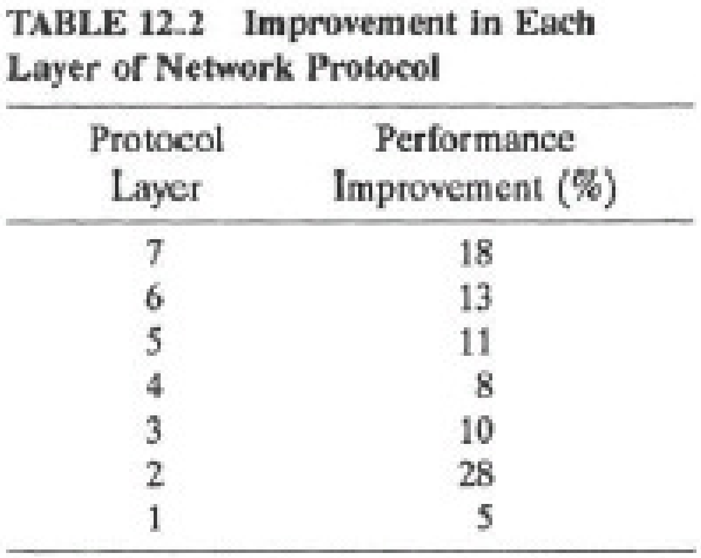

TABLE 12.2 Improvement in Each Layer of Network Protocol

Protocol Layer	Performance Improvement (%)
7	18
6	13
5	11
4	8
3	10
2	28
1	5

Variability

- 5-percentile and 95-percentile (fractile, quantile) – minimum and maximum
- Xth decile = 10X-percentile
- Xth quartile = 25xth quartile
- Median =second quartile
- Intequartile range (SIQR) = third – first quartile