Lecture 3

★ What are we doing here?

- Our goal is to extract "knowledge" from the "data".
  - "Knowledge" → a quantitative summary of data behavior.
  - "Data" → results of (repeated) measurements.

- The data, x, with measurements $\{x_i\}$ where $i = 1, 2, \ldots N$, are viewed as realizations of a random variable, X.

- The most important problem in data mining is how to estimate the pdf $h(x)$ from which the values of our measurements, x, are drawn.

- The integral of a pdf like $h(x)$, is called the cumulative distribution function (CDF),

$$\boxed{H(x) = \int_{-\infty}^{x} h(x')\,dx'}$$

- We assume that all distributions are properly normalized, i.e.

$$H(\infty) = \int_{-\infty}^{\infty} h(x')\,dx' = 1$$

- Our challenge is that given the data, x, we need to model the "true" or "population" pdf $h(x)$ using a data-derived, "empirical", pdf $f(x)$.

- This is because we never observe all of the true pdf $h(x)$, but only measure a finite number of samples from $h(x)$.

- An additional complication is that our measurements are associated with uncertainties,

$$e(x) = p(x \mid \mu, I)$$

where "μ" is the true value, and "I" represents all the information about the error distribution.

- For the commonly used Gaussian error distribution,

$$p(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where $I \equiv \sigma$, the standard deviation of the Gaussian captures your knowledge of the error distribution.

Descriptive statistics:

- Any arbitrary distribution function can be described by:

- Location parameters
- Scale (or width) parameters
- Shape parameters (typically dimensionless)

– These descriptive statistics:
- when based on the "true" distribution $h(x)$, are called "population statistics."
- when based on the finite-sized data set, or $f(x)$, are called "sample statistics".

– Some important examples of descriptive statistics, also referred to by their "moments":

- Expectation value (arithmetic mean):

$$\mu \equiv E(x) = \int_{-\infty}^{\infty} x\, h(x)\, dx$$

- Variance:

$$V \equiv \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 h(x)\, dx$$

- Standard deviation:

$$\sigma = \sqrt{V}$$

- Mode:

$$\left(\frac{dh(x)}{dx}\right)_{x_m} = 0$$

where,

$x_m \rightarrow$ mode

- P% quantiles (p is called a percentile), q_p:

$$\frac{P}{100} = \int_{-\infty}^{q_p} h(x)\,dx$$

→ The most frequently used quantiles are the median, q_{50}, & first and third quartiles q_{25} & q_{75}.

- In general, the higher the moment of the statistic, the more difficult it is to estimate it from a smaller sample set.

<u>Data-based estimates of descriptive statistics</u>

- When calculated from data, these are called "sample statistics".

- Ignoring the uncertainties on the data (for now), the integrals from the above equations can be replaced by a summation with an appropriate constant of proportionality.
- For example,

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

and

$$s = \sqrt{\frac{1}{\ \ } \sum^{N} (x_i - \bar{x})^2}$$

$\sqrt{\frac{1}{N-1} \sum_{i=1}}$

- The symbols for the mean ($\bar{x}$) and standard deviation (s) are different to highlight that these are "estimators" to the "true" values defined as μ and σ.

– Estimators like these typically have a variance (V) and bias, judged by their mean squared error (MSE),

$$MSE = V + bias^2$$

- Estimators whose V & bias vanish as $N \to \infty$ are called "consistent estimators".
- You can have unbiased estimators which are not consistent.

– When N is sufficiently large and if the variance of $h(x)$ is finite, then from the central limit theorem, we can write:

- Uncertainty on the mean estimator:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$$ (aka "standard error of the mean")

- Uncertainty on the standard deviation:

$$\sigma_s = \frac{s}{\sqrt{2(N-1)}} = \frac{1}{\sqrt{2}} \sqrt{\frac{N}{N-1}}\, \sigma_{\bar{x}}$$

– Different estimators are compared based on

Different estimators are compared based on their "efficiency", which is the size of the dataset required to attain a certain accuracy (defined as estimator/uncertainty of estimator).
eg: Mean is more efficient than median by factor of $\pi/2$.

- A parameter with the minimum variance bound (MVB) attainable is called minimum variance (unbiased) estimator (MVUE).

- For real data, we also need an estimator that is "robust" to outliers.
  - An example is the median and interquartile range.
  - The standard error for any given quantile is

$$\sigma_{q_p} = \frac{1}{h_p} \sqrt{\frac{p(1-p)}{N}}$$

where h_p is the value of the pdf at the p^{th} percentile.

- Important distributions that are frequently used (see text for exact formulae):

  - Uniform distribution
  - Gaussian distribution
  - χ^2 distribution
  - Exponential distribution
  - Student-t distribution
  - Discrete variables: Binomial distribution

→ Poisson distribution (special case of binomial distribution).