Data Summarization

UW ECE 657A - Background Topic

Mark Crowley

Lecture Outline

1. Summarizing Data
   - Central Tendency
   - Measures of Dispersion
2. Pearson Correlation Coefficient
3. Cross Correlation

Lecture Outline

Summarizing Data

We have data we need to find patterns in it.

- Simplest pattern is a summary of the data.

Summarizing A Single Variable

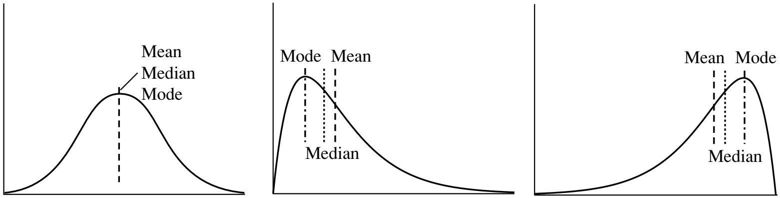
- Given a univariate sample $X_1, \ldots, X_n$ (could be Real, Natural, Integers)
- Goal: Summarize the variable compactly with a few numbers:
  - We want to summarize properties like spread, variation, range. Anything that can provide a summary statistic for the variable.
- Average : simplest and most common and estimate of central tendency.

$$\underline{\texttt{mean(x)}}) = \mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - **Pro:** If the samples come from a normal distribution then the average is the optimal estimate.
  - **Con:** Sensitive to outliers. (could be noise, data entry error, actual outliers)

Summarizing A Single Variable

- **Median:** If the samples are sorted then the median is the value that splits the list into half
- **Mode:** is the most common value in the list of samples (data can be bimodal or more)
- **Skew:** (third moment) high skew means the bulk of the data is at one end. Result: *Median* will be a better measure than mean.
- **Kurtosis:** (fourth moment) A measure of the heaviness of the tail of the distribution with respect to a set of points with a normal/Gaussian distribution and the same variance.



(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Central Moments of a Set of Points

Mean(1), Variance(2), Skew(3) and Kurtosis(4) are unified by a single type of calculation on the n data points.

$$\mu_k \approx \int_{-\infty}^{\infty} (x-c)^n f(x) dx$$

$$\mu_k \approx \frac{1}{n-k+1} \sum_{i=1}^{n} (X_i - \mu_{k-1})^k$$

The 3rd and 4th moments are usually normazlied by s^k just as Standard Deviation is.

Types of Mean Functions

- **Trimmed Mean:** ignoring small percentage of highest and lowest values
- **Geometric Mean:**

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \leq \texttt{Mean} \tag{1}$$

$$= \exp\left[\frac{1}{n}\sum_{i=1}^{n} \log x_i\right] \tag{2}$$

  - Arithmetic mean of logarithm transformed x
  - Good for positive values and output of growth rates
  - Most appropriate for ranking normalized results (different normalization can alter ordering for arithmetic or hamonic means)

Types of Mean Functions

- **Harmonic mean:** average of *rates*

$$H = \frac{n}{1/x_1 + 1/x_2 + \cdots + 1/x_n}$$

  - It is the reciprocal of arithmetic mean of the reciprocals of the sample points.
  - Appropriate for values that are inversely proportional to time such as “speedup”.

Mean Examples (in Matlab)

Data: `X=[1,1,1,1,1,1,100]`

- $n = 7$
- Mean=`sum(X)/n`=106/7=15.4
- Median=`median(X)`=1
- Mode=`Mode(X)`=1
- Trimmed mean(25%)=1
- Geometric Mean=1.9307
- Harmonic mean=1.1647

Measures of Dispersion: Variance and Deviation

- measure the spread of the data range
- **Standard Deviation:**

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

  - **Pro:** Same units as the data
  - **Con:** Sensitive to outliers
  - **matlab:**`std(x)`
- **Variance:**

$$\textbf{matlab:}\underline{\texttt{var(x)}} = \sigma^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  -

Variance and Deviation

- **Mean Absolute Deviation (MAD)**

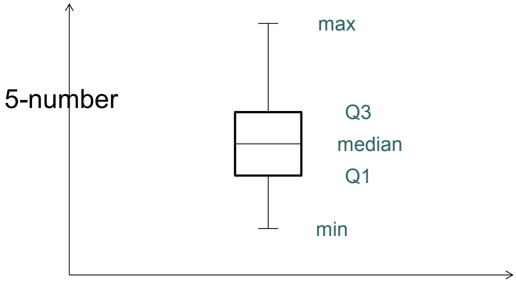
$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

  - Less sensitive to outliers than STD
  - **matlab:**`mad(x)`
- **Interquartile Range (IQR):** Difference between 75th (Q3) and 25th (Q1) percentile of data

Deviation Examples

Data: `X=[1,1,1,1,1,1,100]`

- $n = 7$
- Range=range(X)/n=99
- Std=std(X)=37.42
- MAD=mad(X)=24.24
- IQR=0



Lecture Outline

Pearson Correlation Coefficient (PCC)

The **Pearson Correlation Coefficient (PCC)** is slightly more complicated way to analyse the relation between two attributes.

- PCC measures of how strongly one attribute implies another

$$r = cov(v_1, v_2)/s_1 s_2$$
$$cov(v_1, v_2) = \frac{1}{n}\{(v_1 - \bar{v_1})(v_2 - \bar{v_2})^T\}$$

- **Interpretation:**
  - $-1 \leq r \leq 1$
  - -1 corresponds to negative correlation
  - $+1$ corresponds to positive correlation
  - Variance is a special case of covariance where $v_1 = v_2$
  - $r \neq 0$ implies dependency
- Independence implies covariance or correlation $=0$
- However, in general covariance or r$=$0 doesn't necessarily imply independence

PCC Examples

$$r = cov(v_1, v_2)/s_1 s_2$$

$$cov(v_1, v_2) = \frac{1}{n}\{(v_1 - \bar{v_1})(v_2 - \bar{v_2})^T\}$$

$$X = (2, 1, 3)$$

$$\bar{X} = 2 \quad S_X^2 = \frac{2}{3}$$

$$X - \bar{X} = (0, -1, 1)$$

$$Y = (1, 3, 2)$$

$$\bar{Y} = 2 \quad S_Y^2 = \frac{2}{3}$$

$$Y - \bar{Y} = (-1, 1, 0)$$

$$r = \left(\frac{1}{3}\right)\left(\frac{-1}{2/3}\right) = -0.5$$

PCC Examples

X=(2,1,3)	Y=(1,3,2)	r= -0.5	weak negative correlation
X=(2,1,2)	Y=(1,3,1)	r= -1	strong negative correlation
X=(2,1,2)	Y=(4,2,4)	r= 1	strong positive correlation
X=(2,1,2)	Y=(5,6,7)	r= 0	independent

Table: Some PCC exmaples

Lecture Outline

Cross Correlation

- Between two time series: association between values in the same time series separated by some lag $v_1(i), v_2(i)$
- Measures similarity between them by applying a time lag to one of them.
- It can be used to find repeated pattern or periodic nature so it can be used for prediction.
- Correlation coefficient r
- **Autocorrelation:** cross-correlation between two values at different points in time in the **same time series** (also called autocovariance)
  - series separated by some lag $v_1(i), v_1(i + lag)$
  - it can be used to find repeated pattern or periodic nature so it can be used for prediction.

$$R(s,t) = \frac{E[(X_t - \bar{x})(X_s - \bar{x})]}{\sigma_t \sigma_s}$$