Descriptive Statistics

ratio vs interval

Ratio	Interval	Ordinal	Nominal
• absolute zero • geometric/harmonic mean	arbitrary zero arithmetic mean		
ax	$ax+b$	f(old)=new order preserving monotonic	
geometric mean $\left[\sqrt[n]{x_1 x_2 \cdots x_n}\right.$ $\left.\dfrac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}\right]$ harmonic mean	mean, SD, Pearson's corr., t-test, F-test	median, percentile, rank, run test, sign test	mode, entropy, χ^2 test

cross sectional: several variables at one instance

time series: one var across several instances

longitudinal/panel: combine above

4 moments
- mean
- variance
- skewness
- kurtosis

measures of shape

skewness

measure of symmetry (or its lack)

Pearson's moment coefficient of skewness (g_1)

$$\boxed{g_1 = \frac{\sum_{i=1}^{n} (X_i-\bar{X})^3/n}{\sigma^3}}$$

$g_1 \to 0$ when data symmetrical

$g_1 \to +ve$ positively skewed

$g_1 \to -ve$ negatively skewed

symmetric : skewness $\in (-0.5, 0.5)$

moderate skew : $\in (-1,-0.5) \cup (0.5,1)$

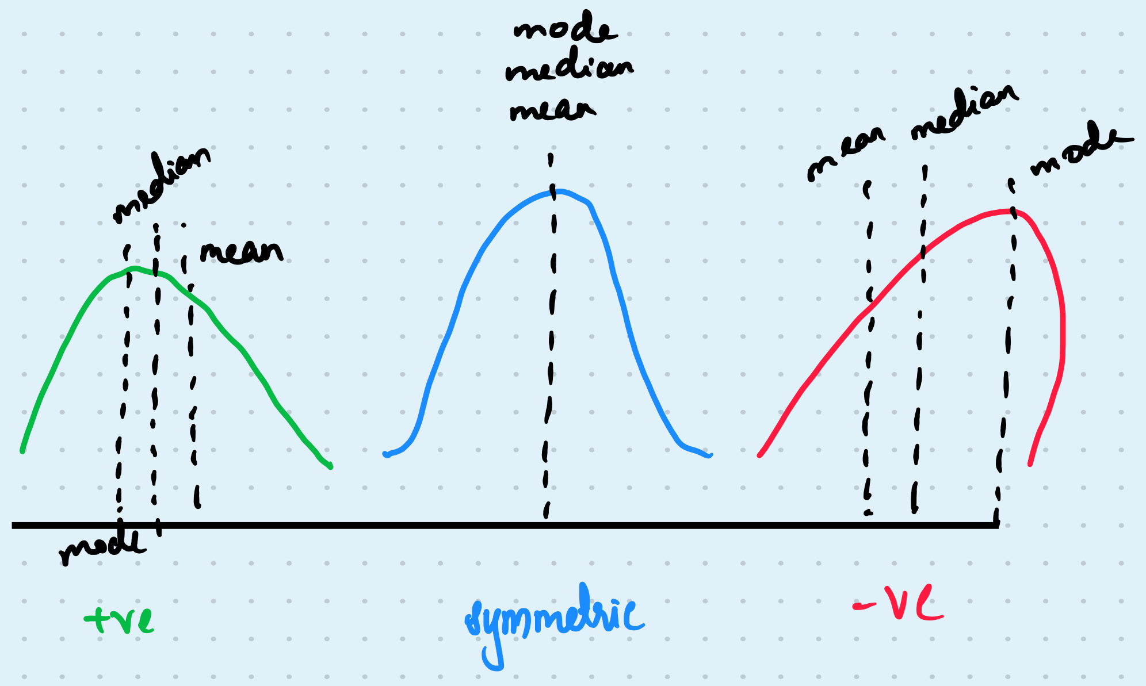
highly skew : $\in (-\infty,-1) \cup (1,\infty)$

Skewness with sample of n observations (G_1)

$$\boxed{G_1 = \frac{\sqrt{n(n-1)}}{(n-2)} g_1}$$

(Joanes & Gill, 1998)

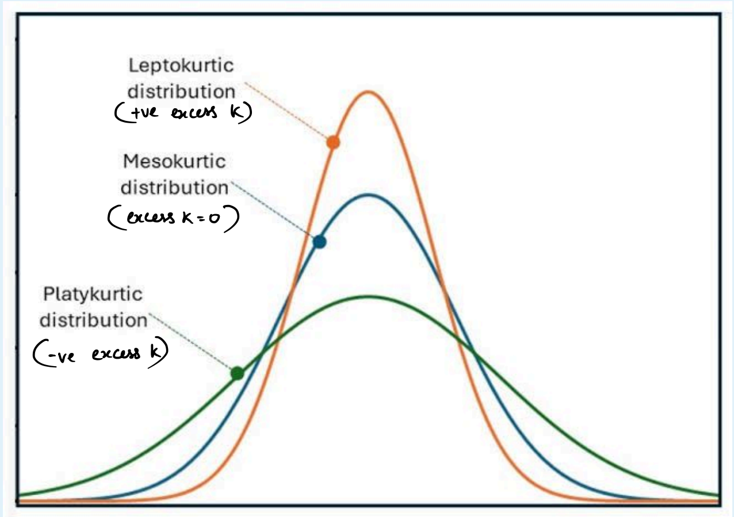
$G_1 \to 1$ as n increases



kurtosis

aimed at the tail ; checks whether tail is heavy or light

$$\boxed{\text{Kurtosis} = \frac{\sum_{i=1}^{4} (X_i-\bar{X})^4/n}{\sigma^4}}$$



platykurtic: low frequency of outliers

leptokurtic: high frequency of outliers

$$\boxed{\text{excess kurtosis} = \text{kurtosis} - 3}$$

kurtosis only captures information from outliers thanks to higher power

exploratory data analytics (EDA)

preliminary exploration

summary statistics

part of EDA , summarise data

central tendency

mean $\to \sum(x_i - \bar{x}) = 0$

median : less influenced by outliers, no need for entire dataset

mode

quantile

let $x_1 \cdots x_n$

quantiles : data points that divide the dataset into equal sized parts

k^{th} q-quantile $\Rightarrow$ k/q of set before, $(q-k)/q$ of set after

q-1 quantile points exist

x-percentile $\Rightarrow$ x% of set before, $(100-x)$% of set after

$$\boxed{Q_k = \frac{n+1}{q} \times k}$$

Say value at position 5 is 21 and value at position 6 is 22

Value at position 5.1 ≈

$21 + 0.1(22-21) = 21 + (0.1 \times 1)$
$= 21.1$

measures of variation

identify outliers, how close records are to the mean

feature w/ low variability → unlikely to have statistical significance w/ target variable

Coefficient of variation

$$\boxed{CV = \frac{\sigma}{\bar{x}}}$$

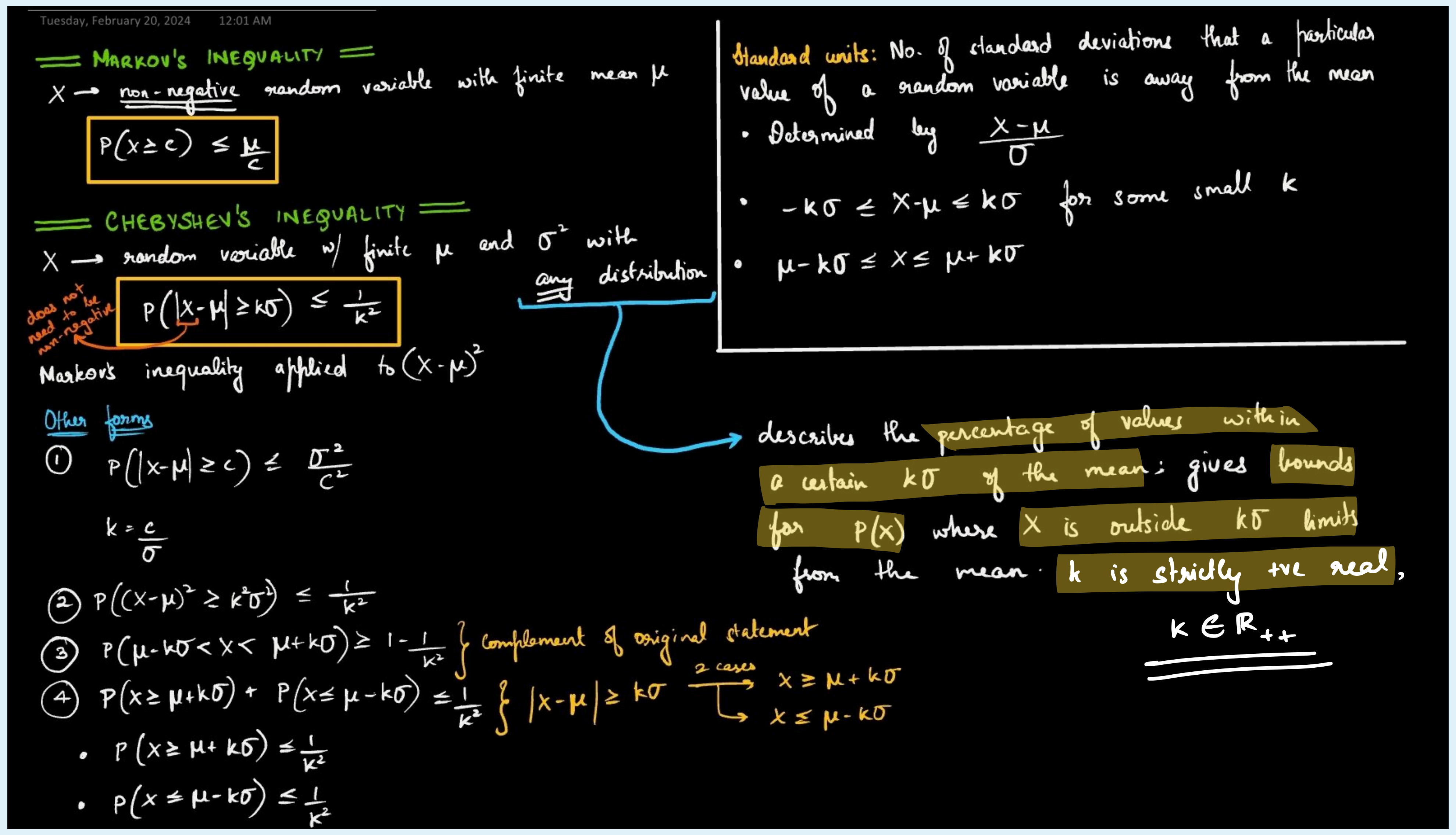
$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{n}$$

$$\sigma^2_{sml} = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{n-1}$$

IQR

$IQR = Q_3 - Q_1$

Outliers $\in (-\infty, Q_1 - 1.5(IQR))$
$\cup$
$(Q_3 + 1.5(IQR), \infty)$



Tuesday, February 20, 2024 12:01 AM

== MARKOV'S INEQUALITY ==

X → non-negative random variable with finite mean μ

$$\boxed{P(x \ge c) \le \frac{\mu}{c}}$$

== CHEBYSHEV'S INEQUALITY ==

X → random variable w/ finite μ and σ^2 with any distribution

$$\boxed{P(|X-\mu| \ge k\sigma) \le \frac{1}{k^2}}$$

does not need to be non-negative

Markov's inequality applied to $(X-\mu)^2$

Other forms

① $P(|x-\mu| \ge c) \le \frac{\sigma^2}{c^2}$

$k = \frac{c}{\sigma}$

② $P((x-\mu)^2 \ge k^2\sigma^2) \le \frac{1}{k^2}$

③ $P(\mu - k\sigma < X < \mu + k\sigma) \ge 1 - \frac{1}{k^2}$ } complement of original statement

④ $P(x \ge \mu + k\sigma) + P(x \le \mu - k\sigma) \le \frac{1}{k^2}$ } $|x-\mu| \ge k\sigma \xrightarrow{\text{2 cases}} x \ge \mu + k\sigma$, $x \le \mu - k\sigma$

- $P(x \ge \mu + k\sigma) \le \frac{1}{k^2}$
- $P(x \le \mu - k\sigma) \le \frac{1}{k^2}$

Standard units: No. of standard deviations that a particular value of a random variable is away from the mean

- Determined by $\frac{X-\mu}{\sigma}$
- $-k\sigma \le X-\mu \le k\sigma$ for some small k
- $\mu - k\sigma \le X \le \mu + k\sigma$

describes the percentage of values within a certain kσ of the mean ; gives bounds for P(x) where X is outside kσ limits from the mean . k is strictly +ve real,

$k \in \mathbb{R}_{++}$