Probability
Lec 1

Data → facts, and figures
Dataset → Data collected in a particular study
Categorical Data → Data that can be grouped into categories (Qualitative)

- Primary Data: Data collected by oneself // Raw Data
- Secondary Data: Processed Data e.g: news, TV

- Statistics: Numerical facts e.g "Statistics of Sports"
- population: Set of all elements of interest in study → Survey to collect data on entire population → census
- Sample: Subset of population → Survey to collect data on sample → Sample Survey
  → Difficult to work on population, so sample is used
- Observation: set of measurements obtained for a particular element

parameter (measure) calc from pop // property

- N = pop size → μ = pop mean
- n = sample size → $\bar{x}$ = sample mean

statistic (measure) cal from sample

- Elements: Entities on which data are collected
- Sample Size: Num of Elements

- Variable: Characteristic of interest for the elements
  - Qualitative: Variable that can not be expressed numerically
  - Quantitative: can be expressed numerically

Quantitative
- Discrete → can take countable data, has "gaps" (integers, nums)
- Continuous → measurable data (take int as fraction)

Elements (Nominal, Ratio, Ratio, Ratio, Ordinal) / observation / variable

Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
American Century Intl. Disc	IE	14.37	30.53	1.41	3-Star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-Star
American Century Ultra	DE	24.94	10.88	0.99	3-Star
Artisan Small Cap	DE	16.92	15.67	1.18	3-Star
Brown Cap Small	DE	35.73	15.85	1.20	4-Star
DFA U.S. Micro Cap	DE	13.47	17.23	0.53	3-Star
Fidelity Contrafund	DE	73.11	17.99	0.89	5-Star
Fidelity Overseas	IE	48.39	23.46	0.90	4-Star
Fidelity Sel Electronics	DE	45.60	13.50	0.89	3-Star
Fidelity Sh-Term Bond	FI	8.60	2.76	0.45	3-Star
Gabelli Asset AAA	DE	49.81	16.70	1.36	4-Star
Kalmar Gr Val Sm Cp	DE	15.30	15.31	1.32	5-Star
Marsico 21st Century	DE	17.44	15.16	1.31	5-Star
Mathews Pacific Tiger	IE	27.86	32.70	1.16	3-Star
Oakmark I	DE	40.37	9.51	1.05	2-Star
PIMCO Emerg Mkts Bd D	FI	10.68	13.57	1.25	3-Star
RS Value A	DE	26.27	23.68	1.36	4-Star
T. Rowe Price Latin Am.	IE	53.89	51.10	1.24	4-Star
T. Rowe Price Mid Val	DE	22.46	16.91	0.80	4-Star
Thornburg Value A	DE	37.53	15.46	1.27	4-Star
USAA Income	FI	12.10	4.31	0.62	3-Star
Vanguard Equity-Inc	DE	24.42	13.41	0.29	4-Star
Vanguard Sht-Tm TE	FI	15.68	2.37	0.16	3-Star
Vanguard Sm Cp Idx	DE	32.58	17.01	0.23	3-Star
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4-Star

Source: Morningstar Funds500 (2008).

- 25 elements
- 5 variables

Statistics
- Descriptive summarization // μ, std, var(x)
- Inferential (Statistical Inference)
  - Estimation of parameters on base of sample — $\bar{x}$ is an estimator // point estimator
  - Hypothesis of parameters

Lec 2

- Measurement Scales:

① Nominal → Data consists of labels/names used to identify an attribute of the element e.g: FundType
→ numeric/non-numeric codes may be used e.g: 1→ coke, 2→ Sprite, 3→ Fanta

② Ordinal → Data has properties of nominal scale and order/rank of data is meaningful
e.g: Morning-star rank → 1-star // describes the rank, + has an order 1st, 2nd, 3rd

③ Interval → Data has properties of ordinal scale
→ interval b/w values expressed in terms of fixed unit of measure
→ Always numeric
e.g: SAT Scores → can be ranked in terms of performance
→ Differences b/w intervals are meaningful → st 1 (620-550) 70 more points than st 2.

④ Ratio → Data has properties of interval scale + ratio of 2 values is meaningful
→ zero value indicates that nothing exists for the variable at zero point
e.g: cost, distance, height

- Interval / Ratio → Quantitative
- Nominal / Ordinal → Qualitative

- Study Design:

	Variable of interest: Profit
time	Profit
2000	$175k
2001	⋮
⋮	⋮
2010	

① Time Series: Data collected over several periods of time
↳ trends + forecast + multivariate time series (more than 2 interested var)
+ univariate // // (only 1 interested var)

② Crossectional Study / Design: Data collected at the same/approximately the same point in time

	Temp	Humidity
city 1	⋮	⋮
city 2	⋮	⋮
city 3	⋮	⋮

- Sampling error: processing sample instead of population
- Non-sampling error: Mistakes that happen during data collection, recording, analysis

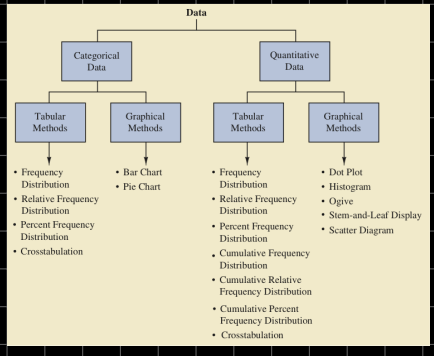
- Underestimation: prediction/calculation too low — 45 min required for a task, but prediction of 30 min for completion
- Overestimation: // too high — 45 min, but 70 min prediction

Lec 3 Summarization of Qualitative variable

① Relative and % relative Frequency distribution

Name	f	R.F	PRF(%)	C.F	(%)
	19	⋮	⋮	19	
	8	⋮	⋮	19+8	27
	5	⋮	⋮	19+8+5	32
total ( )					

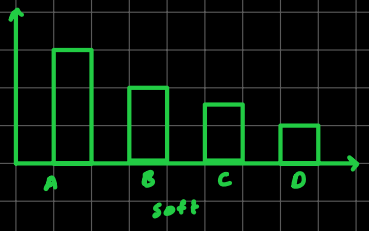
→ Minimum
→ Maximum
→ Most Preferred
→ Outliers
→ Minimum Preferred



- R.F = Relative frequency = $\frac{\text{frequency of the class}}{n}$ — n = total frequency
- PRF = % relative frequency = // × 100

- C.F = Cummulative frequency / percentile = prev + next; order does not matter

② Visualization
- Bar chart / Pie chart → Qualitative
  x-axis: categories → trend with time
  y-axis: frequency
  note of scaling



time: x-axis
var of interest: y-axis

- Pie chart / Sector diagram / Pizza Sliced diagram

- mean = measure for central location / arg
- median = // // // // / middle value when sorted in ascending order — $\frac{n+1}{2}$
- mode = most occurring value
- Range = Largest − Smallest value / measure of variability — dependance on extreme values
- Interquartile Range (IQR) = Q3 − Q1 / // // // — not dependant on extreme values

- Variance = $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$, $\frac{\sum f(x_i - \bar{x})^2}{n-1}$, $\frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)$
- 6 sf $s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$

- Standard Deviation = $\sqrt{\frac{\sum (x_i - \bar{x})}{n-1}}$, $\sqrt{\frac{\sum f(x_i - \bar{x})}{n-1}}$, $\sqrt{\frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)}$

Summarization of Quantitative Variables

Q- The following scores represent the final exam scores for an elementary statistics course:

23, 60, 79, 32, 57, 74, 52, 70, 82, 36, 80, 77, 81, 95, 41, 65, 92, 85, 55, 76, 52, 10, 64, 75, 78, 64, 84, 98, 81, 67, 41, 71, 83, 54, 76, 82, 88, 62, 74, 43, 60, 78, 89, 17, 48, 84, 90, 15, 79, 34, 67, 80, 69, 74, 63, 80, 85, 61, 25, 72

① Create classes + interrels — 5 sy km aur 20 say ziadah nhi if generalizing

- Range: span of data
- $k = \lceil 1 + 3.3 \log n \rceil$ e.g $k = 6.867 \approx 7$ → N.o of classes
- $\left\lceil \frac{R}{k} \right\rceil$ = interval ; R = range

→ class boundary creation (S1)

→ Construct Group frequency
→ Calculate r.f, prf, CF
→ Draw r.f histogram, and comment on distribution of data.
→ Calc. avg score and std.

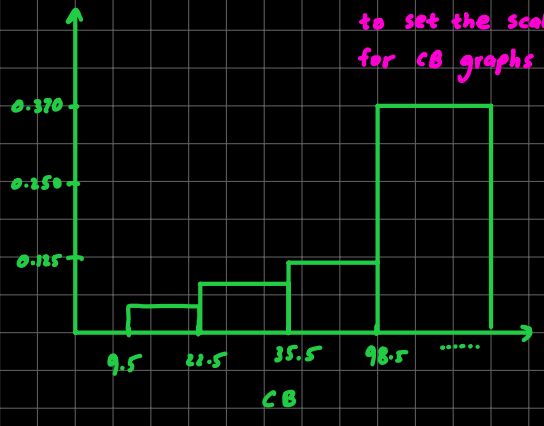
Lec 4

	-5	+5		
	10 - 22		- .05	.05+
	23 - 35		10.5	22.5

- Histogram → Overall commenting on data
- class limits have units of original data
- class boundaries → Horizontal / vertical diff, find common point
  lower → sub
  upper → add

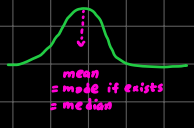
- Create CB

Do not need to set the scale for CB graphs

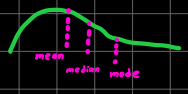


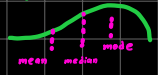
- Behaviours:

① Symmetric / Normal: Pattern like a Bell shaped curve.
Same freq for equidistant points



→ Skewness:
+vely skewed: mean > median > mode
-vely skewed: mode > median > mean





② $S_k = \frac{\text{mean} - \text{mode}}{SD} = \frac{3(\text{mean} - \text{median})}{SD}$

- S_k -ve : -vely skewed
- S_k +ve : +vely skewed
- $S_k = 0$: Symmetric

.85 1.62 sth moderate/ high according to value

Range: -3 —— 3

③ Histogram

- $\bar{x} = \frac{\Sigma x}{n} = \frac{\Sigma fx}{\Sigma f}$ (group mean)

- * 65.83 → class limits matching

Lec 5

Countable

- Discrete variables → classes not recommended → act as a qualitative variable.
  - → range and intervals too small
  - → make ungrouped freq.
  - → class intervals (grouped freq)
- Walpole Ex 1.17, pg # 31

Q. calc. 10% trimmed mean for smokers
Q. Make dot-plot of datasets A and B on the same line
Q. Calculate coefficient of variation (CV) and comment on consistency of the datasets

- Trimmed mean:

① Sort
② "Round" the num of values
③ Take mean after discarding values

- std = dispersion from mean
- trimmed mean for removal of outliers

- Dot-plot:
  - plot each point on a number line
  - make key
  - Dispersion, interval density
  - Plot mean

→ if overall, then

else,

- $CV = \frac{\sigma}{\mu} \times 100$] pop
- $= \frac{s}{\bar{x}} \times 100$] sample

- Shows how large the standard deviation is from the mean
- std has same unit of measurement

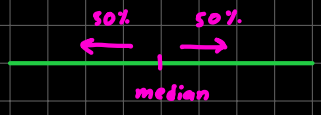
- Relative comparison + unitless
- std = abs measure of dispersion
- CV = relative measure //
- ideal when comparing variablity of variables having different standard deviation and different mean

- CV ↑ less consistent
- CV ↓ more consistent

Heaps → median

Lec 6

- median - cuts data into 2 halves

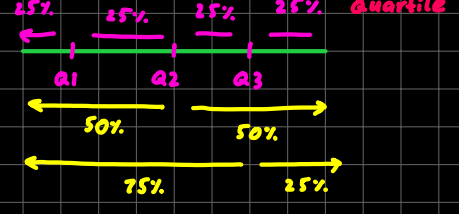


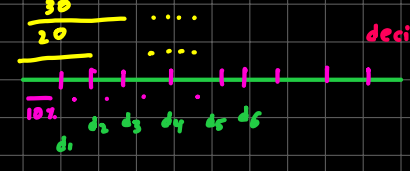
Percentile Points:

① Sort the data
② Calculate index
↳ $i = \frac{p}{100} \times n$; 1–99
③ If i is an integer, then p^{th} percentile will be the average of values placed at i^{th} and $(i+1)^{th}$ position

→ If i is non integer, round up, and pick the value at that position

→ groups data in segments e.g:





Sorted data
Q_2 = median
$d_5 = Q_2$ = median
⋮

Q- considering data of monthly starting salaries of 12 business graduates

3450
3550
3650
3480
3355
3310
3490
3730
3540
3920
3520
3480

a) Calc five-point summary of the data and IQR → $Q_3 - Q_1$
b) Make a box-plot and detect if there exists any outlier, also comment on the shape of the distribution
c) Calc z-scores and detect any outlier if exists

measure of relative location of a value from mean

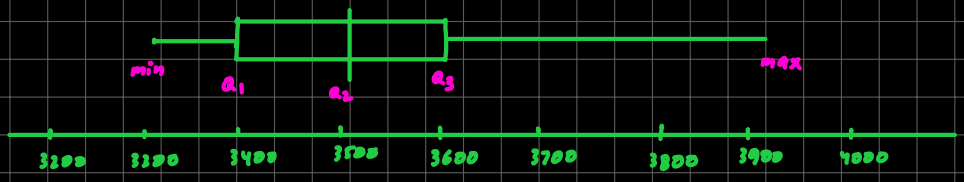
$\frac{IQR}{2}$ = semi IQR

IQR → middle span

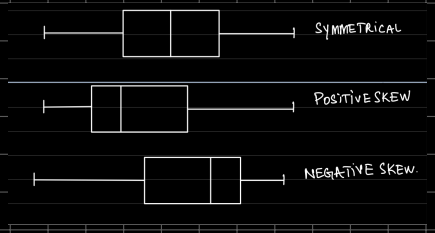
a) Minimum point
Maximum point
Quartiles points

① Apply procedure for finding percentiles

b)



→ take diff to confirm the larger box. $Q_3 - Q_2$, $Q_2 - Q_1$
right box large → +vely skewed
left box large → -vely skewed



lower fence = $Q_1 - 1.5\,(IQR)$
Upper fence = $Q_3 + 1.5\,(IQR)$

c) $z = \frac{x - \bar{x}}{s}$; $-3 \text{ —— } 3$

if any point exists outside the range, then the x value corresponding to it is the outlier.

- $z = -1.5$, x_i is 1.5 standard deviations below the mean
- $z = 1.5$, // above the mean

lec 7

- anomalies = outliers
- weights = frequency
- cross tabulation / joint frequency table

* combining 2 individual tabs

2 Qualitative

size ↓ / Fuel →	Premium	Regular	Diesel	total →
compact				→
midsize	frequency of each			→
large				→
total ↓	↓	↓	↓	□ Grand total

Column wise total
marginal frequency / partial

2 Quantitative

→ median if deviated data

	City MPG			
Cylinder	≤ 15	16 - 20	21-25	total
1-5	0	0	4	4
6 - 10	2	3	0	5
11- 15	1	0	0	1
total	3	3	4	10

1 Qualitative, 1 Quantitative

	Cylinder			
size	1-5	6-10	11-15	total
compact	2	2	0	4
midsize	2	2	0	4
large	0	1	1	2
total	4	5	1	10