Categorical Data

Numerical Data

Summary Table for one variable

Contigency Table for two variables

Bar chart

Pie / Doughnuth chart

Pareto chart

Side by side chart

Doughnut chart

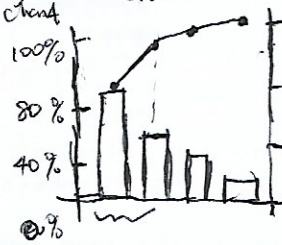
Ordered array

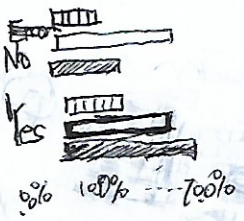
Stem and leaf display

Scatter plot

Timeseries

Fegency Distribution and cumulative Distrib

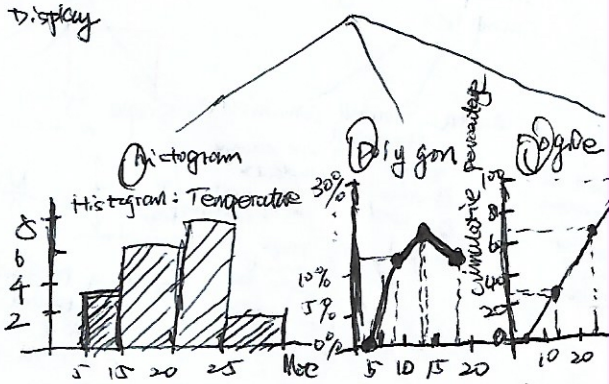




The Vital few

Large Medium Small

Separate the "vital few" from the "trivial many"



Central tendency

Mean Arithmetic — Affected by extreme values

Median : Median position: $\frac{n+1}{2}$ ⇒ even the average of two middle numbers

less sensitive than the mean to extreme value

Mode : Not affected ....

Weighted Mean : $\bar{X}_w = \frac{\sum w_i X_i}{\sum w_i}$: gives formal equal weight to all observations

Case	Grade	Grade
I	3	A
II	3	B
III	5	B

$\frac{(3\times4)+(3\times3)+(5\times3)}{3+3+5} =$

Geometric mean : measure the rate of change of a variable over time.

$$\bar{X}_G = (X_1 \times X_2 \times \cdots X_n)^{1/n}$$

Geometric mean rate of return:

$$\bar{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$$

investment from: \$100,000 → 50% decrease → 50000 → 100% increase → \$100,000

Geometric mean rate of return : $\bar{R}_G = [(1+(-.5))\times(1\times1)]^{\frac{1}{2}} - 1$

$= 0\%$

Arithmetic mean rate of return : $\bar{X} = \frac{(-0.5)+(1)}{2} = .025 = 25\%$

Quartiles

ranked value

Q1 25%

Q2 25%

Q3 25%

Q4 25%

$Q_1 = \frac{(n+1)}{4}$

$Q_2 = \frac{(n+1)}{2}$ Place

$Q_3 = \frac{3(n+1)}{4}$

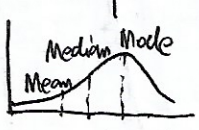
.5 need to ... /2

=Median

Xmin Q1 Q2 Q3 Xmax

25% 25% 25% 25%

IQR

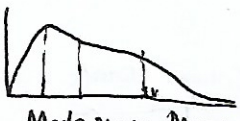


skewed Left

Mean < Median

Mean = Median = Mode

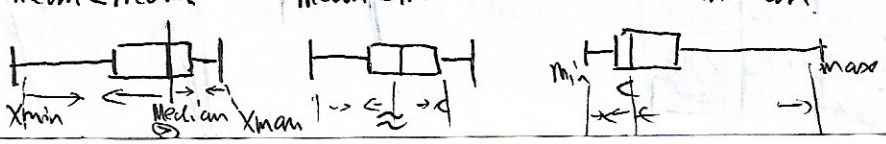
Mean = Median



Mode Median Mean

Mean > Median

mark the outlier with a star



DCOVA =
Define, Collect, Organize, Visualization, Analyse.

Categorical — Numerical

Nominal — ordinal (ranking), internal (ACT, SAT, C°, F°) — ratio (height, weight, Age salary)

: D

Samples

without regard to their probability / Non-probability samples

Judgment — based on judgement researcher must be confident

Quota — the desired sample is rare and hard to find

Snowball — ① use college students ② the first 100 12 customers ③ people on street simple one that happens to come your way.

convinience

Probability Samples — simple, may not be a good representation of the population.

Simple — ① ideal ② reduce likelihood of bias

Disadvan: ① hard to achieve ② accurate list of whole population ③ expensive

System matic — Regular intervals — $k = \frac{N\,(\text{population})}{n}$ — ④ Using formalize k=interval a=first member r=rth member

Adv: ① spread more ② easier than simple random.

Disadvan: may interact with some hidden pattern in the population

Stratified — two or more subgroups

① when the population is heterogeneous / dissimilar, where certain homogeneous or similar, sub-populations can be isolated (strata)
② Simple random sampling is mostly appropriate when the entire population from which the sample is taken is homogeneous.
③ ensure the representation of individuals across the entire population
④ better coverage
⑤ convinient
⑥ Difficulty in identifying appropriate stra.

Cluster — when the researcher cannot get ① a complete list of the members of a population. ② a list of subjects so widely scattered

vs. random samples:
① More cost effective, saving time
② less efficient
③ units close to each other may be very similar

survey errors:
① coverage error or selection bias: excluded from frame
② Nonresponse error or bias
③ Sampling error: random differences from sample to sample.
④ Measurement error: bad or leading question.

can be leveraged by survey designers

it can be ethical issue

Categorical — Numerical

One categorical variable

	Percentage
1	37%
2	29%
3	05%

summary table

Two categorical variable

Contigency Table

	No	Yes	Total
Small	170	20	190
Medium	100	40	140
Large	65	5	70
Total	335	65	400

↓

	No	Yes	Total
Small	42.5%	5.0%	47.5%
Medium	25.0%	10.0%	35.0%
Large	16.2%	1.2%	17.5%
Total	83.7%	16.2%	100.0%

Percentage of Row Totals

	No	Yes	Total
Small	89.47%	10.53%	100.0%
Medium	71.43%	28.57%	100.0%
Large	92.86%	7.14%	100.0%

Percentage of Column Totals

	No	Yes	Total
Small	50.75%	30.77%	47.50%
Medium	29.85%	61.54%	35.00%
Large	19.40%	7.69%	17.50%

ordered arrays / stem-and-leaf

Frequency distributions & cumulative Distributions

⑥

Class	Midpoints	Frequency
10 < X < 20	15	3
20 ≤ X < 30	25	6
30 ≤ X < 40	35	5
40 ≤ X < 50	45	4
50 ≤ X < 60	55	2
Total		20

Relative & percentage Frequency distribution

Cumulative Frequency distribution

Class	Frequency	Relative Frequency	Percentage	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
	3	.15	15%	3	15%	3	15%
	6	.30	30%	6	30%	9	45%
	5	.25	25%	5	25%	14	70%
	4	.20	20%	4	20%	18	90%
	2	.10	10%	2	10%	20	100%
Total	20	1.00	100%	Total	20	100%	20	100%