Data Visualization

How do you see data

Good Decisions are based on an accurate understanding of Good data.

Present data in precise, concise & understandable way

Data

Techniques > Tools

- Categorical
  - Eg Customer ratings
  - City
- Numerical
  - Discrete
    - Eg No. of consonant
    - Degrees para
  - Continuous
    - Time
    - Pressure

Core Principles on which visualization of data?

Nature of data dictates which ~~dicta~~ visualization technique.

Benifits of visual representation of data

- Communicate complex information concisely & powerfully.
- Create a "picture" for reasoning about & analyzing quantitive & conceptual information
- Makes Cognitive processing easier

- ▷ provides "content/ information rich" view at a glance
- ▷ Directs attention toward the content rather than methodology
- ▷ Describe, explore, summarize a set of numbers
- ▷ Convey messages about significance of the data.

4 principles of Effective visualization
→ know purpose → Ensure Integerity [VV Imp]
→ Maximize data ink : minimize non data ink
→ Show your data : annotate.

Dashboard
↳ A visual display of most important information needed to achieve one or more objectives that has been Consolidated on a single screen so it can be monitored & understood at a glance

Executing Your information display is a 3 step process

What I am trying to communicate?
What is the message?
How do I make the msg clear at a glance

Defining Message
⇓

Should I use text, table, graph or a diagram or a combination

Choosing form
⇓

What design principles lead to quick cognitive processing & effective message communication

Creating Design

Defining Message

Data Dashboard
↳ ~~How~~ A collection of related display serving the purpose

Probability Distribution

Def^n A statistical model that shows possible outcomes of a particular event or course of action as well as statistical liklihood of each event.

How to go abt data (ie fit a distribution

1. The data values themselves are used directly is si (simulation)

2. 'Fit' a theoretical distribution to the data

3. The data values could be used to define an 'empirical dist' function in some way

Empirical Distribution

- Using data we build our own distribution

Essential building blocks
Define ~~destiny~~ density / distribution function
Estimate parameters. (mean, std dev. etc)

The 3 approaches...

- Approach 1 is used to validate simulation model when comparing model output for an existing system with corresponding output for system itself

- Two Drawbacks : Simulation can only produce what happened historically, & there is seldom enough data to make all simulation runs.

- Approaches 2 & Approaches 3 avoid these short-comming so that any value between minimum & maximum can be generated so approaches 2 & 3 are prefered

If theoretical distributions can be found that fits observed data then App 2 is prefered over approach 3

Approach 3 v/s Approach 2
Empirical dist. may have some irregularities if small no. of data pts are available
App 2. smothens out data & may provide information on overall underlying distribution

In approach 3 its usually not possible to generate values outside observed simulation

If one wants to test performance of simulated system under extreme conditions that can't be done using approach 3

There may be compelling reasons in some situation for using a particular theoretical distribution. In that case to its better to get empirical support for that distribution from observed data

Clues from Summary Statistics

For summary statistics

→ For symmetric distribution mean & median should match. In the sample data, if these values are sufficiently close to each other we can think of symmetric dist. (eg. Normal)

→ Coefficient of variation (CV) [ratio of std dev & mean] for continuos distribution. The CV=1 for exponential dist. If histogram looks like a slightly right skewed curve with CV>1, then log normal could be a better approximation of the dist

→ Lexis Ratio: Same as CV for discrete distribution

→ Skewness (v): "measure of symmetry of a distribution". For normal distribution v=0 For v>0, the distribution is skewed towards right (exponential dist, v=2) & for v<0, distribution is skewed towards left

Parameter estimation

- Once distribution is guessed, next step is estimating parameters of the distribution

- Most common method of parameter estimation ML

Goodness of fit

- For the input data we have, assumed a probability distribution. we also have estimated parameters of same

Goodness can be checked by
→ Frequency comparison [a bit technical]
→ Probability Comparison (visual tool)
→ Goodness of fit test (statistical test of goodness)

QQ-plot - Quantile quantile plot
- Graph of the q_i- quantile of fitted distribution versus the q_i - quantile of distribution (sample)

$$x^M_{q_i} = \hat{F}^{-1}(q_i)$$
$$x^S_{q_i} = \tilde{F}^{-1}(q_i) = X_{(i)}, \; i = 1, 2, \ldots, n$$

If $F^{(x)}$ is correct distribution that is fitted for i

large sample size, then $F^{(x)}$ & $F_n(x)$ will be close together & the Q-Q plot will be approximately linear with intercept 0 & slope 1

For small same, even if $\hat{F}(x)$ is correct dist. there will be some departure from straight line

2. P-P plot (Probability - Probability plot)
A graph of model probability $\hat{F}(X_{(i)})$ against the sample probability $\tilde{F}_n(X_{(i)}) = q_i$; $i = 1, 2, \ldots, n$

- Its valid for both cts & discrete data sets
- If $\hat{F}(x)$ is correct distribution that is fitted for a large sample size, then $\hat{F}(x)$ & $F_n(x)$ will be close together & the P-P plot will be approximately linear with intercept 0 & slope 1

NOTE 'Q-Q plot will amplify differences between the tails of the model dist. & sample distribution
while P-P plot will amplify diff in middle portion of model & sample distribution

Goodness - of - fit test

A goodness of fit test is statistical hypothesis test that is used to asses formally whether observations $X_1, X_2, \ldots X_n$ are an independent sample from a particular distribution with function F

H_0: The X_i's are IIDRV with dist function F

· 2 famous tests

Chi Square test — Kolmogorov-Smirnov

Week 3

Determing & inferring association

- Categorical Variable [~~Dr~~

2 types

[ordinal data] arranged categorically (ordered) : Eg Rating

" " " " (unordered) : Eg Gender

- Eg B-School shortlist 1200 candidates (960 men & 240 women) for RCM 324 selected

[Contingency Table]

	Male	Female	Total
offer made	288	36	324
Not offer	672	204	876
Total	960	240	1200

Event

~~W~~ M → Randomly selecting male Candidate

(F) W → " " "

A → Admisson offer being made

(A^c) N → Randomly selected Cand. not offered.

(Joint Event) Pr. randomly observed candidate is male & ~~ob~~ given admisson

$288/1200 = P(M \cap A) = 0.24$

Pr. randomly observed Candidate is female & given ad.

$36/1200 = P(F \cap A) = 0.03$

$P(M \cap A^c) = 672/1200 = 0.56$

"ly

$P(\cancel{F} \cap A^c) = 204/1200 = 0.17$

Marginal Probab.

$P(M) = \frac{960}{1200} = 0.8$, $P(F) = \frac{240}{1200} = 0.2$

$P(A) = 0.27$, $P(A^c) = 0.73$

$= (324/1200)$ $= (876/1200)$

Is it ~~write~~ right?

(Condit^n Prob) $P(A|M) = \frac{P(A \cap M)}{P(M)} = \frac{288}{960} = \boxed{0.3}$

$$P(A|F) = \frac{36}{240} = 0.15$$

Bayes Rule

→ Given we have initial guesses about an event from which we calculate prior probabilities

→ Then from various sources like sample, product field res. etc we obtain more info. about events.

→ Given new updates we can update our prior beliefs by calculating revised probab. this is posterior probability

→ Bayes rule is used to calculate the posterior prob. if we have initial info & additional sample information

Chi-Square Test

Eg Brand Preference

City	Pref Brand Brand A	Brand B	Brand C	Total
Mum	279	73	225	577
Chennai	165	47	191	403
Total	444	120	416	980

Independent (Explanatory) city
Dependent (response) brand preference

Statistical independence is a symmetric property b/w variables (categorical)

Hypothesis
H_0: Categorical variables are independent
H_1: Categorical variables are not independent

Let f_o be the observed frequencies (from sample)
Let f_e be the expected frequency, if variables were independent

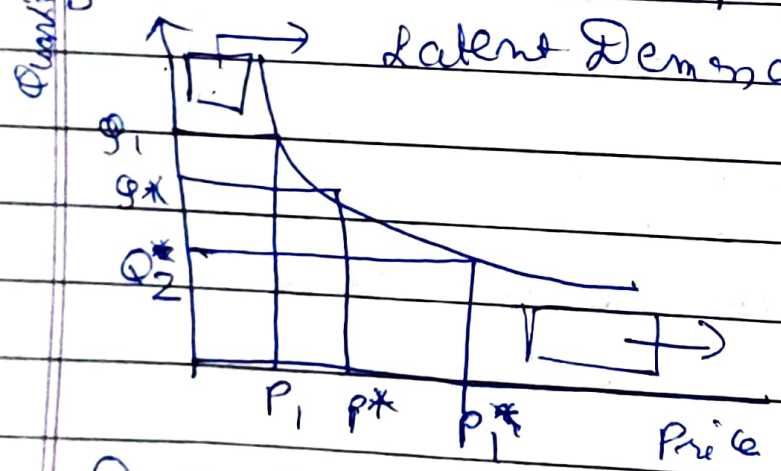
The expected freq of a cell equals product of row & col. total divided by sample size.

$$\chi^2 = \frac{\sum (f_o - f_e)^2}{f_e}$$

If $f_e \approx f_o$ · we get small numerator
∴ χ^2 is small
It Happens when H_0 is true
$df = (r-1) * (c-1)$

Week 4

Lecture 1 Demand Response Curve.



Benifit cons. get paying less price (Consumer surplus)

Diff. objec. to decide optimal price
- ~~Retail~~ Revenue Maximising Price
- Profit Maximising Price

- 4 properties
1. Non-negative
2. Continuous
3. Differentiable
4. Downward sloping (sloping.)

Fun 4
PAGE NO.:
DATE:

Lecture 2- Elasticity.

Slope

Measure how much demand changes in resp. to price

$$\partial(p_1, p_2) = \frac{\Delta D}{p_2 - p_1}$$

- If $p_1 > p_2 \Rightarrow D(p_2) > D(p_1)$. Hence slope is always negative

Slope can be used as local estimator of demand change for a small price change

Elasticity : Ratio of % change demand & % ΔP

$$\varepsilon(p_1, p_2) = \frac{[d(p_2) - d(p_1)] / d[p_1]}{(p_2 - p_1) / p_1}$$

$\boxed{\text{unitless quantity.}}$

Lecture 3- Linear Response Curve.

Simplest price response curve

$D(p) = D_0 - m * p$

D_0 (Demand at price = 0) → y intercept → Market size

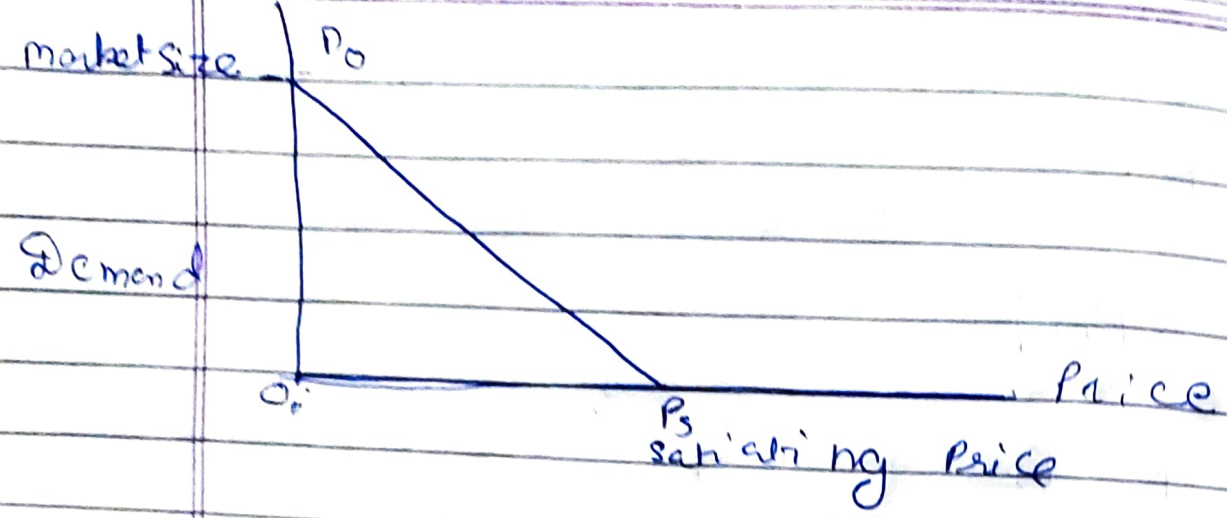
m is slope

The price at which demand = 0 is called satiating price, $\boxed{P_S = \frac{D_0}{m}}$

Elasticity $\varepsilon = \frac{mp}{D_0 - mp}$

as $p \to P_S$ $\varepsilon \to \infty$

$\varepsilon = 0$, when $p = 0$



Constant Elasticity Curve

- After algebraic transition, the constant elasticity curve is given by
$$D = cp^{-\varepsilon}$$
where c is constant [Demand at price 1]

- It is not guaranteed that demand is either finite or satiated ($D \to \infty$, as $p \to 0$. Also, $D \to 0$ for any p)

Revenue is $R = p * D = cp^{(1-\varepsilon)}$

~~Cobo~~ Constant Elasticity Curve
when $\varepsilon < 1$ (inelastic) revenue ↑, price
when $\varepsilon > 1$ (elastic) revenue ↑, price ↓

Lecture 4. Estimation Problem in Demand Response Curve

Proble. Can we estimate Demand response curve

Collect Data via Experiment.

Demand ii → Dependent Variable

prices → Explanatory variable

Linear Demand response curve

- A SLR [simple Linear Regression] model can help identify y-intercept & slope
- SLR can tell if linear relationship is a good fit for data available from market enperiement

Simple Linear Regression

Excel, Option, add-ins → Analysis Toolpak

Regression Statistics

Multiple R → Multiple Linear Regs

[Represent Correlation coefficients]

R^2 : (Multiple R)^2

Adjusted R square (make sense when > 1 explanatory Va)

Standard Err

Linear on Averge
Describes how conditional mean of Y depends on X

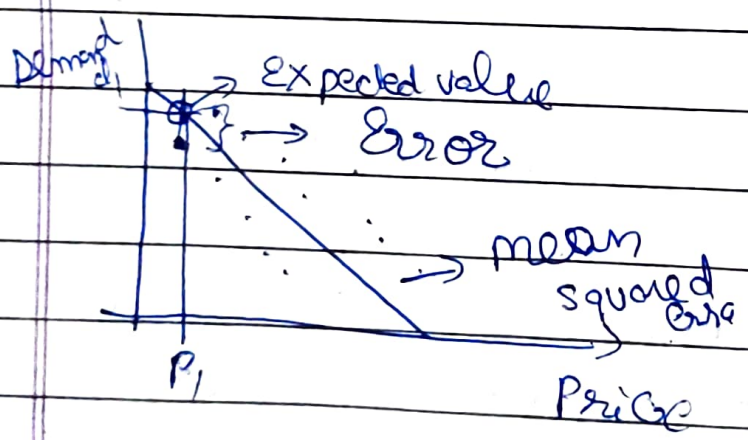
The SRM shows that Mean lies on a line with β_0 & slope β_1

$\mu_{Y|X} = E[Y|X=x] = \beta_0 + \beta_1 x$

↳ conditional mean of Y wrt X

Deviation of response from the mean around $\mu_{Y|X}$ are called errors

Error (denoted by ε, & $E[\varepsilon] = 0$



WE minimize (SSE) min Sum of Squared Error

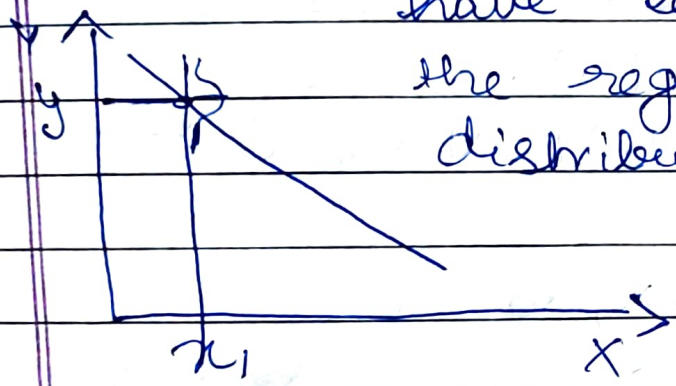
The SRM usually has 3 assumptions

1. Independent. Errors are independent of each other
2. Equal variance. All errors have same variance
$Var(\varepsilon) = \sigma_\varepsilon^2$
3. Normal. The Errors are normally distributed

Observed values of the response Y are linearly related to values of the explanatory variables X by the equation

$$y = \beta_0 + \beta_1 x + \varepsilon \quad \text{where } \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

The observations are independent of one another have equal variance around the regression line, & are normally distributed around regression line.



R^2 → Ability of regression to explain the variability in y

using prices as explanatory variable can explain 73% var. in y

~~the~~ R^2 is called Coefficient of determination.

$SS_m / SS_T = R^2$ [Because of Regression How much variability can be explained.