Week1

- Types of data: Categorical, Numerical. Categorical has two subtypes: Ordinal, Nominal. Numerical has two subtypes: Interval and Ratio.
- Typically all measured quantities are numerical, and all counted quantities are categorical.
- Umbrella principles of effective visualization – Know Purpose, Ensure Integrity, Maximize data/minimize non-data ink, Annotate data (preferred over legend)
- Information display steps – (a) Defining message (b) Choosing form (c) Creating design
- Choose a table representation when
    - displaying a complete (partial) dataset,
    - data is too wide to display on a graph,
    - explain how numbers are derived or calculated.
- Choose a chart representation when
    - Comparing a slice of information
    - Show cause vs effect
    - Show change over time
    - Show patterns of data distribution (eg. normal curve)
- Appropriate graph types for conveying messages

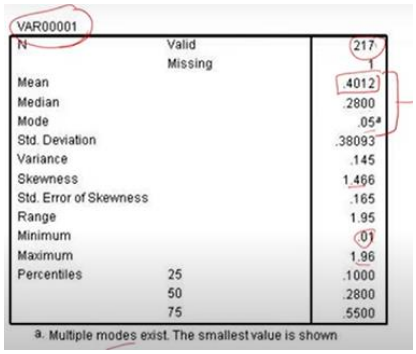
Message	*Chart type*
Components of one item	Pie chart
Components of multiple items	100% column/stacked column chart
Item comparison	Bar chart
Change over time	Column/line chart
Frequency/distribution	Histogram
Outliers	Box plot
Correlation	Paired bar, Scatter chart

- A concise definition of a dashboard: A visual display of the most important information needed to achieve one or more objectives that has been consolidated on a single screen so it can be monitored and understood at a glance.
- Common Issues plots : Axes not starting from 0, percentages in pie chart not adding up, Inappropriate chart type, 3D pie charts

Week2

- PMF (discrete) represent probability masses for each value of a random variable.
- PDF (continuous) represent probability density for a random variable. It's not possible to find the probability for a particular value of a continuous random variable, but for a small range of values.
- Trace-driven simulation is limited to provided data values. Of course, the disadvantage is, for values outside the historically observed data, prediction isn't possible.
- Alternatively, you might choose to "fit" a theoretical distribution, or even an empirical distribution (that you've built). Theoretical distributions are preferred over empirical distributions, due to the limitations of available data in the latter.

- To define a distribution, we need to define its density/distribution function and estimate its parameters (mean, std. deviation etc)
- While defining density functions, make sure that for $x < a_0$, $y = 0$ and for $x > a_k$, , $y = 1$, where a_0 and a_k are arbitrary values.
- Selecting the appropriate distribution to fit given data also depends on the domain.
- To "fit" a theoretical distribution, follow these steps.
    - Get the descriptive statistics of the data as follows.

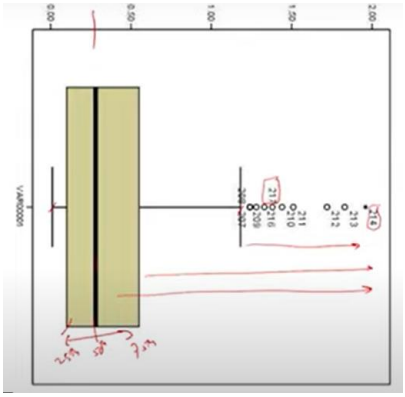


VAR00001

N	Valid	217
	Missing	1
Mean		.4012
Median		.2800
Mode		.05[a]
Std. Deviation		.38093
Variance		.145
Skewness		1.466
Std. Error of Skewness		.165
Range		1.95
Minimum		.01
Maximum		1.96
Percentiles	25	.1000
	50	.2800
	75	.5500

a. Multiple modes exist. The smallest value is shown

    - Is the distribution symmetric?
        - If the mean, median and mode are not equal to each other, then it is not symmetric. We can eliminate symmetric distributions like normal and uniform at this point. However, it's not necessary that the distribution is symmetric, if the mean and median are equal.
    - What's the support of the distribution?
        - If the support of the distribution includes negative X-axis, we could expect at least some values in the sample-set to negative. If the minimum value is greater than 0, it could mean that the support is restricted to positive X-axis.
    - What could be the shape of the distribution?
        - Plot histogram and observe the shape.
        - If the skewness is positive, the right tail of the distribution is bigger than the left tail. Else, the left tail is bigger than right tail. Note that skewness of a normal distribution is 0, and that of a exponential distribution is 2.
        - For a std. normal distribution, kurtosis is 3. An increased kurtosis (>3) can be visualized as a thin "bell" with a high peak whereas a decreased kurtosis corresponds to a broadening of the peak and "thickening" of the tails.
    - Are there outliers?
        - Plot box-plot and see if there're datapoints beyond the whiskers. This will also agree with the skewness. If the distribution is right skewed, median will be placed on the left of the distribution.



- Other metrics?
  - Find co-efficient of variation (cv = std. dev/mean). If cv = 1, then distribution could be exponential. If cv > 1, the distribution could be lognormal. Note that this applies only to continuous distributions, and only when mean is non-zero. In the case of discrete distributions, cv is called Lexis ratio.
  - If mean and variance are equal (or approximately so), then the distribution could be Poisson.
  - If the standard deviation is low, and the intervals between the quartiles are more or less same, then the distribution could be uniform.

- Once a distribution has been "fit", we should check how good is the fit – using a chi-square/Kolmogorov-Smirnov (KS) test or a probability plot.
- There are two kinds of probability plots = P-P plot and Q-Q plot.
  - Q-Q plot is plotted taking equivalent quantiles from model and sample distributions that have equal probabilities. If the equivalent quantiles are equal then the Q-Q plot will be a 45^{o} line with an intercept 0. Q-Q plot can amplify the differences between the tails of the model and sample distributions.
  - P-P plot is plotted taking equivalent probabilities from model and sample distributions at the same quantiles. If the equivalent probabilities are equal then the P-P plot will be a 45^{o} line with an intercept 0. This is valid for both continuous and discrete datasets. P-P plot can amplify the differences in the middle portion of the model and sample distributions.
- Chi-square tests are based on a null/alternative hypothesis. Null hypothesis is that the datapoints are from the model distribution. Alternative hypothesis is that the datapoints are not from the model distribution.
  - If the tabulated chi-squared statistic (scipy.stats.chi2.ppf) is more than the calculated value (stats.chisquare(obs_freq, expec_freq)), then we will fail to reject null hypothesis. Alternatively, as the calculated value approaches zero, there is more evidence that the null hypothesis is true.
  - In terms of p-value, if p-value is greater than (1 – confidence level), we fail to reject null hypothesis.

Week3

- Consider a B-school which shortlisted 1200 candidates (960 men and 240 women) for its post-graduate management program. Out of these, 324 candidates were given offer letters for admission. The data is included here:

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

Contingency table

- After reviewing the record, a women's forum raised the issue of gender discrimination on the basis that 288 male candidates were offered admission against only 36 female candidates.

- This can be proved or disproved using conditional probability.
- Calculate the joint and marginal probabilities as below.

- In terms of probabilities, the previous table can now be rewritten as:

	Male	Female	Total
Offers made	0.24	0.03	0.27
Not offered	0.56	0.17	0.73
Total	0.8	0.2	1.0

- Joint probabilities (of what?) appear in the main body of the table (e.g. 0.24, 0.03).
- Marginal probabilities (of what?) appear in the margin of the table (e.g. 0.8, 0.2).

- From the above table, we have P(A ∩ M) = 288/1200 = 0.24, P(M) = 960/1200 = 0.8 and P(A | M) = 288/960 = 0.3. Thus, we can write
  - P(A | M) * P(M) = P(A ∩ M).
- On similar grounds, P(F | M) = 0.15
- Since P(F | M) < P(A | M), offering admissions appears to be a biased towards the candidate being male.
- In other words, Conditional Probability = Joint Probability/Marginal Probability. This is the Baye's formula.
- Assuming two suppliers S1 and S2 supplying products to a factory,

- We are interested in the posterior probability that a particular supplier is guilty of supplying bad quality product *given that* we have bad quality raw material at our doorstep – $Pr(S1|B)$ or $Pr(S2|B)$.
- This is an application of Baye's theorem – finding posterior probability given some initial facts and numbers.

Thus, the probability that S1 is the supplier, given that the product is of bad quality is given by the following formula,

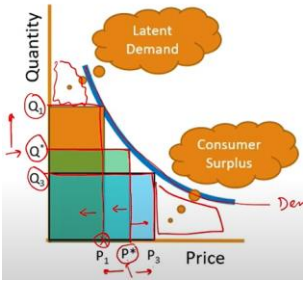
$$\Pr(S1 \mid B) = \frac{\Pr(S1) * \Pr(B \mid S1)}{\Pr(S1) * \Pr(B \mid S1) + \Pr(S2) * \Pr(B \mid S2)}.$$

- It's not possible to draw inferences purely based on statistical behavior on samples. We need do hypothesis tests.
- In hypothesis tests, two tests must be conducted, one each for the null hypothesis (H_0) and alternative hypothesis (H_A)
  - Let f_o be the observed frequencies
  - Let f_e be the expected frequencies, if the null hypothesis were true. (obtained by multiplying row total and column total, divided by the total sample size)

- Chi-square statistic can now be calculated using the formula $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$
- When H0 is true, f_o and f_e tend to be closer, and hence the calculated chi-square statistic is relatively small. Larger the calculated chi-square statistic is, greater is the evidence against H_0
- Degree of freedom is given by the product of #rows and #columns

Week4

- Demand response curve (shown below) deals with a single product in a single market, whereas a typical demand curve in Economics deals with the entire market.

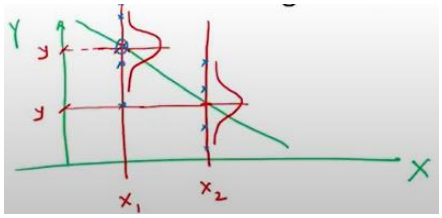


- Latent demand lies towards the top left of the demand-response curve, and gets realized as you reduce the price.
- Consumer surplus lies towards right bottom of the demand-response curve and gets realized (reduces) as you increase the price.
- Revenue-maximizing price is not necessarily same as Profit-maximizing price, and may even be conflicting objectives.
- Properties of demand-response curve are: Non-negative (first quadrant), Continuous, Differentiable, Downward sloping
- There are two types of demand-response curve.
  1. Linear curve
     - Slope of the curve is defined as $\partial(p_1, p_2) = \frac{D(p_2) - D(p_1)}{p_2 - p_1}$
     - Elasticity of the curve is defined as $\epsilon(p_1, p_2) = -\frac{[d(p_2) - d(p_1)]/d(p_1)}{(p_2 - p_1)/p_1}$
     - Elasticity of 2 means a 10% reduction in price leads to a 20% increase in demand (sales).
     - Elasticity of certain items are very less (Eg. Salt), whereas that of certain others are very high (Eg. Movies)
     - In certain cases, the short-term elasticity could differ from the long-term elasticity. Thus, short-term elasticity of airline travel is less, but the long-term elasticity is high. In the case of a two-wheeler, the short-term elasticity is high, but long-term elasticity is lower.
     - $D(p) = D_0 - m * p$ is the simplest price response curve. This is a linear curve. D_0 is called the market price and represents the demand when the price is 0.
     - The price at which the demand is 0 is called satiating price. $P_S = D_0/m$
     - Elasticity can also be represented as $\epsilon = \frac{m*p}{D_0 - m*p}$. Note that elasticity depends on the price at the specific point.
     - When p = 0, elasticity is 0. When p is P_S, elasticity is also infinite.

- Thus, keep in mind slope and elasticity are not same. While slope remains for a linear equations, elasticity can keep changing with price (X-axis)

2. Non-linear curve (constant elasticity curve)
   - We can keep elasticity unchanged, with a non-linear demand curve $D = Cp^{-\epsilon}$, where C is a constant given as demand when p = 1
   - In the case of a non-linear curve, we can't find the market by putting up 0 price; neither can we find the satiating price, since the demand never reduces to 0. It's asymptotic either way.
   - Revenue = p * D = $Cp^{(1-\epsilon)}$
   - When elasticity is less than 1 (inelastic), the revenue can be increased by simply increasing the prices.
   - When elasticity is greater than 1 (elastic), the revenue can only be increased by setting price closer to 0.

- We can use a simple linear regression (SLR) to find the slope and intercept of a linear response curve. We can perhaps use SLR even in the case of a non-linear curve with constant-elasticity.
- Use Excel's scatter chart and plot the trend of price-response curve; the trend-line should give the slope m and intercept D_0
- SLR plots $\mu_{y|x}$ as $\beta_0 + \beta_1 x$, where β_0 is equal to D_0, and β_1 is equal to m. Deviations of actual values from the predicted values are called errors. Expected value of sum of squared errors is 0.
- In order to use SLR, we assume that error terms are independent of each other, of equal variance and normally distributed.
- $y = \beta_0 + \beta_1 x + \epsilon \text{ where } \epsilon \sim N(0, \sigma_\epsilon^2)$ is t he general representation of the linear model.
- Following picture represents this graphically.

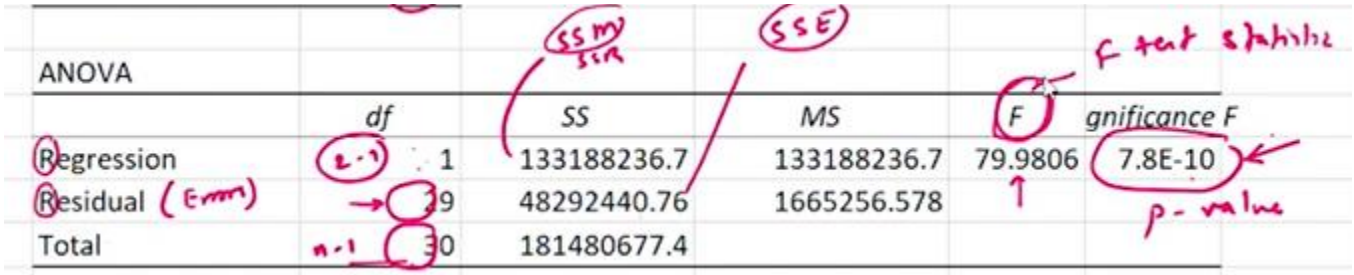


- Excel's data analysis pack gives the following statistics when SLR is used. Here's the interpretation of these values.

Regression Statistics	
Multiple R	0.85668
R Square	0.7339
Adjusted R Square	0.72472
Standard Error	1290.45
Observations	31

   - Multiple-R signifies the correlation between the price and demand (CORREL).
   - R-Square is obtained by squaring Multiple-R.
   - Standard error is σ_ϵ (the square-root of the variance mentioned earlier).

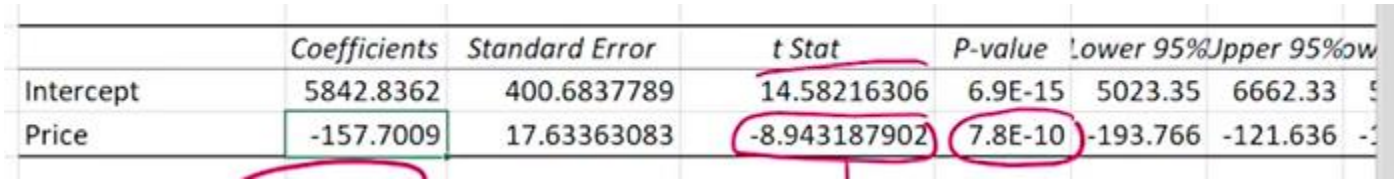
- R-square value is also called coefficient of determination and represents the power of the explanatory variable to explain dependent variable. In our case, R-square value of 0.7339 means that price values can explain ~73% of demand.
- R-square value can also be found as SS_Regression/Total_Regression, or 1 – (SS_Residual/Total_Regression)



ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	133188236.7	133188236.7	79.9806	7.8E-10
Residual	29	48292440.76	1665256.578		
Total	30	181480677.4			

- df_Regression=1, since there are two parameters for a linear regression (bias and weight).
- df_Total=30, since the #observations=31.
- df_Residual=29 (df_Total – df_Regression)
- SS_Regression is SSM (Sum of squares of model).
- SS_Residual is SSE (Sum of square of errors).
- MS_Regression is SSM/df_Regression.
- MS_Residual is SSE/df_Residual. This quantity is equal to variance. Standard_Error = sqrt(MS_Residual)

Now, we must take a hypothesis between x and y. H_0 = x and y are independent; H_A = x and y are dependent.
F indicates the test statistic. p-value (7.8E-10) is lesser than 0.05, so at 95% confidence level, we can reject the hypothesis and hence conclude the demand and price are dependent.



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ow*
Intercept	5842.8362	400.6837789	14.58216306	6.9E-15	5023.35	6662.33	
Price	-157.7009	17.63363083	-8.943187902	7.8E-10	-193.766	-121.636	

Above table indicates the slope and the intercept of the linear equation. t_Stat and P-Value are what a local hypothesis test of these values yields. $H_0 = \beta_0$ (or β_1) is zero; β_0 (or β_1) is non-zero. Since the p-values in both cases is lesser than 0.05, so at 95% confidence level, we can reject both hypothesis and hence conclude that β_0 (and β_1) are both non-zero.