DMD Formula Sheet

Gari Jenkins

Variables : ***Discrete*** (restricted, separated values ie no. of breakdowns) ; ***Variable*** (any value ie height, weight)

Why analyse data : 'raw data' ⇒ information ⇒ support decision making

Frequency Tables : sub-range ~ 'class'; width ~ class interval ; most likely class ~ modal class **but loss of information**

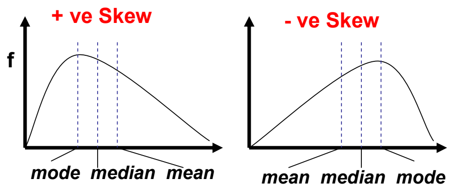
Histograms : represent frequency table ; **Area** not hight should be compared.

Measures of Central Tendancy (or average) ~ a ***value which is typical of a set of data***

Arithmetic Mean	The Median	The Mode
$\boxed{\text{Mean} = \sum x / n}$ . . . of sample μ = Mean of Population ***can be distorted by extreme values***	Middle observation when the observations have been placed in ascending order	The most frequently occurring value in a set. ***One value must dominate if meaningful***

SKEW

For normal distribution *:* ***Mean = Median = Mode***



Measure of Dispersion

➔ ***Range*** = *largest value – smallest value*

➔ ***Standard Deviation (SD)***

$\boxed{\sigma = \sqrt{\dfrac{\sum (x-\bar{x})^2}{n}}}$ for a **population** or $\boxed{s = \sqrt{\dfrac{\sum (x-\bar{x})^2}{n-1}}}$ for the **sample**

Subtract mean from all observations ~ square all the deviations ~ sum the squared deviations ~ divide by n or (n-1) number of observations ~ square root

➔ ***Variance*** = $(\textbf{SD})^2$

Assessing Risk

Measuring Risk ~ probability is measured on a scale from 0 (impossible) to 1 (certain)

- ***Classic / a priori approach*** ~ p(A) = no. of outcomes which represent occurrence of event A / total no. of possible outcomes
- ***Empirical / relative frequency approach*** **~** p(A) = no. of times event a occurred / total no. of observations
- ***Subjective approach*** **~** person uses judgement to assess probability **BUT** availability bias / misperception of randomness

Mutually Exclusive Events **~** if the occurrence of one event means that the simultaneous occurrence of other event is impossible

p(A or B) = p(A) + p(B) . . . . . if not mutually exclusive ⇒ **p(A or B) = p(A) + p(B) – p(A and B)** **ADDITION RULE**

Complementary Events **~** set of outcomes not belonging to an event: **p(A not occurring) = 1 – p(A)**

Independent Events **~** 2 events are independent if the probability of one occuring is not affected by the other

Conditional Probability ~ p(A|B)~ the probability of A occuring given that B has occurred

Multiplication Rule **~** is the probability of one event **and** another event occurring

p(A and B) = p(A) x p(B) . . . . . if not mutually exclusive ⇒ **p(A and B) = p(A) x p(B|A)** **MULTIPLICATION RULE**

Always use a probability tree to solve / clarify awkward problems

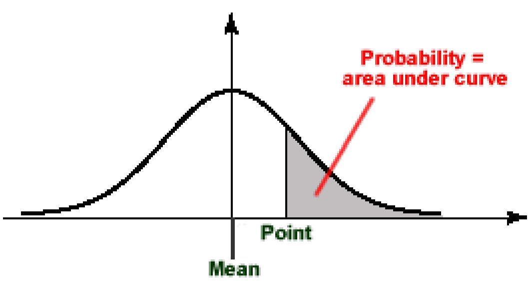
Normal Distribution

Under normal conditions, probabilities follow well defined patterns ~ the normal or Gaussian Bell curve is most important
To apply a normal distribution, you MUST know : its **MEAN** …….its **Standard Deviation**

Z = the number of Standard Deviations of a point from the Mean

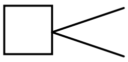
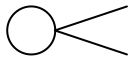
$$\boxed{Z = \frac{\text{Point} - \text{Mean}}{\text{Standard Deviation}}}$$

Then use the **Z** / ρ table to work out the single tail probability



Modelling Decisions ~ Decision Trees

Expected Value is a '***long run average result'*** ; **calculated by multiplying each outcome by its probability of occurrence & summing**

Decision Node:  **Either / Or** **Chance Node:**  **multiply & sum**

Limitations to this Expected Monetary Value (EMV)

- ➔ Expected value represents the average payoff if the decision were repeated several times. It is reasonable for one-off?
- ➔ EMV criterion assumes the decision involves only one objective ~ might be several which conflict ie social
- ➔ Assumes the decision maker is risk neutral ~ might not be !
- ➔ Probabilities and payoffs are only guestimates.

Hypothesis Testing

Overview ~ **1. formulate hypothesis 2. take probability sample from population 3. decide whether hypothesis is rejected**

Null Hypothesis (Ho) ~ Hypothesis to be tested. **Must be able to calculate probability of obtaining observed data if Ho true**

Alternative Hypothesis (H_1) ~ Hypothesis against which Ho tested. Either **two- tailed (<>) or one-tailed (<) or (>)**

p-value ~ probability of obtaining observed results if Ho (Null Hypothesis) true **= p(obtaining observed results|Ho is true)**

Level of Significance ~ What p-value be to reject Ho ? **0.05** ie if p-value < 0.05 – ***Ho is Rejected***

Sampling Theory ~ when you take a sample, how likely is it that it is a good estimate of the population ?

Conditions:

- ➔ ***Population must be normally distributed and SD of population must be known, or***
- ➔ ***Sample must be >30***

A large sample is more likely to give an accurate estimation of the population mean than a small one. The 'tightness' or Standard Deviation of the sample is called the Standard Error of the Mean ~ **STEM**

Factors that affect size of STEM:

- Size of the sample ~ **↑ sample size : ↓ STEM**
- Variability of the population ~ **↓ population SD : ↓ STEM**

$$\text{STEM} = \frac{\text{Population Standard Deviation}}{\sqrt{\text{Sample size}}}$$

Relating this to: 1. Hypothesis Testing

Using **Z values** as before, we can test whether sample is representative of the population ie **if p-value < 0.05** ~ **REJECT**

Relating this to: 2. Quality Control

Use **Statistical Quality Control** (SQC) to monitor one feature of process output ~ **to see if population mean has shifted**

- **95%** of area under curve falls within **1.96 SDs ~ used for upper/lower Warning Limits = +/- 1.96 x STEM**
- **99.8%** of area under curve falls within **3.09 SDs ~ used for upper/lower Action Limits = +/- 3.09 x STEM**

1. Select sample & plot 2. If within Warning Limits ~ ok 3. If > Warning Limits, take another sample. If still above, take action otherwise ~ ok. 4. If > Action Limits, take action

Confidence Intervals & Sample Size ~ use sample results to provide estimates about the population

Assume :

- ➔ **sample is simple random sample**
- ➔ **size ≥ 30**
- ➔ **selecting from very large population**

But to work out STEM, we must know what the population SD !!
Since sample size ≥ 30, we can assume **sample SD ≈ population SD**

$$\Rightarrow \text{STEM} \cong \frac{\text{Sample Standard Deviation}}{\sqrt{\text{Sample size}}}$$ ∴**95%** probability of sample mean falling within 1.96 STEMs

95% confidence Interval is: Sample Mean +/- 1.96 STEM

99.8% confidence Interval is: Sample Mean +/- 1.96 STEM

Given the maximum error, we can work out **Sample Size = [1.96 population standard deviation / Max error]^2**

Analysing associations between variables : Correlation & Regression

1. Correlation ~ measure of the ***strength*** of association between 2 variables.

Plot onto a scatter diagram : ***Product- Moment Correlation Coefficient*** **(r)** is measure of ***strength*** of linear correlation

-1 (perfectly negative correlation) ----------- **0** --------- (perfectly positive correlation) **+1**

r unchanged if:

- ➔ constant added observation of 1 or both variables
- ➔ constant is multiplied by observation of 1 or more variable

r is ***UNIT FREE***

Interpreting Correlation

- ➔ apparent strong association between 2 variables may be ***CHANCE*** / ***bad luck*** etc. Need to test whether the observed correlation is **statistically significant**:

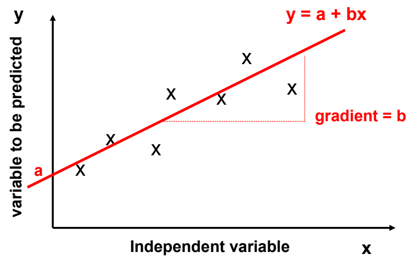
  Ho Pop Correlation is zero (no correlation) ; H_1 Pop Correlation is not zero

Using **p-value** (computer generated), if **< 0.05,** we can reject Ho – there does appear to be correlation

- ➔ dont assume high correlation proves causal link : coincidence / causal link with 'hidden' 3rd variable
- ➔ product – moment correlation only measures ***linear*** association & assumes both variables are ***normally distributed***

2. Two Variable Regression ~ used to describe the ***nature*** of the association between 2 variables.

Can be used to understand the relationship between variables & make forecasts. Essentially involves fitting a line through the scatter points on the scatter diagram.



Interpret a ~ if none of x, then a is the residual amount

Interpret b ~ each unit of x increases y by b

Can predict by substituting into the regression line. ***Interpolation*** (within the observed range) ; ***Extrapolation*** (outside range) - **less reliable since we have no evidence that the linear relationship we have identified will apply**.

Measuring how well the line fits the data ~ '***goodness of fit'*** can be measured by the **coefficient of determination** = $\mathbf{r^2}$
Note – the square of the correlation coefficient & has values **0** ---------- **1** (perfect fit ~ 100%)

Interpret r^2

- ➔ measures how well the regression line fits the scatter points ; closer to 100% the better
- ➔ shows the % of variation in the dependent variable that can be explained by variation in the other variable. Other factors can account for the difference between r^2 and 100%

Note: a high value doesn't guarantee that we have the best regression model. We need **significance test:**

p-values:
Ho Population Slope : $\beta = 0$ . . .if $p < 0.05$, we can reject and say they are statistically significant ie there is a relationship
Ho Population Intercept ; $\alpha = 0$ . . same again, but if $p > 0.05$, we cannot reject the hypothesis that population intercept is zero, BUT it is a good idea to leave the intercept in the equation since its removal can distort measures such as r^2

Assumtions underpinning significance test:

- for given value of x, errors associated are normally distributed
- for each value of x, the SD of errors is the same
- errors associated with any 2 observations are independent

Limitations of the 2 Variable Regression Analysis:

- we assume that past relationship will apply in the future
- assumed relationship is linear
- extrapolation based on above can be risky
- only 1 variable has been used to make predictions; increase additional independent variables (multiple regression)
- certain observations may unduly influence estimate of best fit

3. Multiple Regression Analysis

We can obtain more accurate forecasts if we include more independent variables in regression model.

Goodness of fit can be measured by **coefficient of multiple determination R^2** ***(R is coefficient of multiple correlation)***

BUT : R^2 always ***increases*** (or fails to decrease) as the no. of independent variables increases (even if the new variables have no relationship with the dependent variable. If we just concentrate on R^2 we might be encouraged to add useless variables. To counteract this, the **adjusted** R^2 is often used ie penalises it for the no. of variables in the model.

Significance Tests

Ho = $\beta_1 = \beta_2 = \ldots \beta_{n \text{ number of variables}} = 0$; **F-statistic used to test hypothesis (that none of the variables are related)**

Most packages give the p-value automatically.We can do this also for **each** variable & use **T-statistic** & p-value to test

In multiple regression ~ problem of **MULTICOLLINEARITY** **(where some or all of the variables are highly correlated)**

- ➔ leads to estimated coefficients having the wrong sign
- ➔ leads to p-values for T-test that are misleading

RULE OF THUMB : Problem if correlation between 2 independent variables > 0.7

Dealing with Multicolinearity : Combine the correlated variables into a single 'super variable **or** Use only one variable from the highly correlated variables (but model may lose predictive power) **or** Ignore p-value for T-tests & use judgement

Including Qualitative Variables – Dummy Variables

Nominal variables eg location / gender / age etc ~ use dummy variables to represent (**either 0 or 1**). But if we can predict the other variables given from one, you again get a problem with multicollinearity. Therefore, we use ***n-1 variables***, where n is the number of possible variables (the absent one then becomes the benchmark)

Forecasting Methods $\boxed{\textbf{Forecast Error = Actual Value – Forecast}}$

Mean Error	Mean Squared Error	Mean Absolute Percentage Error MAPE
= ∑ (forecast errors) / n measure of **BIAS** not accuracy	Square errors, add 'em up, ÷n Used for ***comparison*** but it penalises large errors heavily	Add up absolute %errors, ÷n Widely used but if actual value are small, MAPE can be large ∴USE **MdAPE** (median of APE)

Naïve 1 Forecasts ~ simply use last observation for forecast

Simple Exponential Smoothing (SES) ~ used for flat trends: ***smoothing constant*** α (0 to 1)

Stability (not overreact to freak figures) ~ low value α; Sensitivity (respond to trends) ~ high value α

α determined by using values 0.1; 0.2 etc & calculating MSE. Plot α vs. MSE to get α with lowest MSE error