Statistics Cheat Sheet

Inferential Statistics – Samples and Random Sampling

Inferential Statistics: The part of statistics that deals with the ability to extend findings from a sample to make inferences about a population

Representativeness: A sample needs to be representative of the population it is drawn from in order for findings in the sample to be applicable as inferences about the population. A sample is representative if the attributes of the sample in terms of summary statistics are very close to the population summary statistics of attributes (where population is available).

Representative samples are selected using **Random Sampling** methods, including **Simple Random Sampling** (every member of the population has an equal chance of being selected), and **Stratified Random Sampling** (The population is subdivided into different strata depending on important attributes, and then Simple Random Sampling is applied to each strata or group to select the sample)

Inferential Statistics – Random Variables, Probability Distributions, and p-values

Random Variables: Variables whose **outcomes cannot be predicted with 100% certainty** until the event is completed. Examples include # of customers at an ATM on any given day, # of accident cases brought to the Emergency section of a particular hospital in one week, # of insurance claims filed per month

Probability Distributions: The range of possible outcomes of a random variable (on the X-Axis) and the associated probability of each outcome, based on historical data (frequency counts) on the Y-Axis, is a Probability Distribution.

Commonly used distributions

Distribution	Type	Description	Excel Formula to calculate p-value	Parameters in the formula
Bernoulli	Discrete	Random variable can have only two outcomes (Ex: Yes/No, Default/No Default)	Usually sum of Bernoulli used via Binomial - see below	Usually sum of Bernoulli used via Binomial - see below
Binomial	Discrete	Sum of outcomes of a Bernoulli distribution (Ex: # of defaults in 200 policies, # of wins in 20 matches)	BINOM.DIST(x,n,p,True/False)	x = # of successes n = total number of trials/events p = probability of success (based on historical data) True - use if you need cumulative probability (p < = x) False - use if you need point probability (p = x)
Poisson	Discrete	Random variable can have an infinite number of possible outcomes, but only discrete - that is, from 0 to ∞ (Ex: # of customers walking into a store in a day, # of hospital admissions in a day etc)	POISSON.DIST(x,mean,True/False)	x = outcome of interest mean = observed average of outcome in historical data True - use if you need cumulative probability (p < = x) False - use if you need point probability (p = x)
Normal	Continuous	Random variable can have an infine number of outcomes - that are continuous (Ex: Time spent on a website, spend in a store)	NORMAL.DIST(x,mean, std dev, TRUE)	x = outcome of interest mean = average in historical data std dev = std deviation in historical data TRUE - Always use TRUE for continuous distributions to get cumulative probability (p < = x)