S3 Revision Notes

Combinations of random variables

Key Words: random variables, expectation, variance

- *For the random variables X and Y*
  - $E(aX \pm bY) = aE(X) \pm bE(Y)$
  - $Var(aX \pm bY) = a^2Var(X) + b^2Var(Y)$ *(Independent variables!)*
- *If* $X \sim N(\mu_1, {\sigma_1}^2)$ *and* $Y \sim N(\mu_2, {\sigma_2}^2)$*then:*
  - $aX \pm bY \sim N(a\mu_1 \pm b\mu_2, a^2{\sigma_1}^2 + b^2{\sigma_2}^2)$
  - $X_1 + X_2 + \cdots + X_n + Y_1 + Y_2 + \cdots + Y_m \sim N(n\mu_1 \pm m\mu_2, n{\sigma_1}^2 + m{\sigma_2}^2)$ *(Note no square!)*

Sampling

Key Words: census, sampling, simple random sampling, lottery sampling, systematic sampling, stratified sampling, sampling without replacement, quota sampling, primary and secondary data

- *Census: observes every member of a population*
  - *Used if: population is small or if large accuracy required*
  - *ADV: accurate (of full picture)*
  - *DIS: time consuming, expensive, cannot be used when testing involves destroying articles, info is difficult to process for large volume*
- *Sampling: observes part of a population*
  - *Cheap but doesn't give full picture (size depends on accuracy required)*
- *Simple random sampling: Every member of the population must have an equal chance of being selected*
  - *ADV: free from biased, easy, fair*
  - *DIS: not suitable when sample size is large*
- *Lottery sampling: randomly drawn numbers from a container (from a sampling frame)*
  - *ADV: random, easy, each ticket has known chance of selection*
  - *DIS: not suitable for large pop, sampling frame needed (ordered list of sample selected)*
- *Systematic Sampling: ordered list and selected every nth member from the list (randomly select 1st number)*
  - *Use when: the population is too large for simple random number sampling*
  - *ADV: simple, suitable for large samples*
  - *DIS: only random if ordered list is randomly listed, can have bias*
- *Stratified sampling: taking a proportion of the sampling strata relative to the size of the strata in the population. E.g if sample of 50 taken from 500 take 1/10th of each strata*
  - *Use when: sample is large, population divides naturally into mutually exclusive groups*
  - *ADV: more accurate estimates (than simple random when clear strata)*
  - *DIS: if bias within strata then will be bias in the sample, strata need to be clearly defined*
- *All above involved sampling without replacement, sampling with replacement is called unrestricted random sampling (items can be selected more than once)*
- *Quota Sampling: First decide on groups into which the population is divided and a number from each group to be interviewed to form quotas. Then go out and interview and enter each result into the relevant quota. If someone refuses to answer or belongs to a quota which is already full then ignore that persons reply and continue interviewing until all quotas are full.*
  - *Used when: it is not possible to use random methods - for example when the whole population is not known (homeless in a big city)*
  - *ADV: quick, cheap, easy admin*
  - *DIS: not possible to estimate sampling errors (as not random), human error, non-responses are ignored, human bias*
- *Primary data: collected 1st hand and used by the person collecting it*
  - *ADV: collection is known, accuracy is known, exact data collected is needed*
  - *DIS: costly in time and effort*
- *Secondary data: use someone else's data*
  - *ADV: cheap, large quantities available*
  - *DIS: collection methods and accuracy not known, not in ideal form, may have bias*

Estimation, Confidence intervals, Testing, χ^2, Spearman's rank

Key words: statistic, unbiased and bias estimators, standard error, Central Limit Theorem, confidence intervals, χ^2, degrees of freedom, spearman's rank, hypothesis testing

- *A statistic is a random variable consisting only of the sample items* X_i *(no other quantities)*
- *If a statistic is unbiased then e.g*
  - $E(X) = \mu$, *or* $E(mode) = expeceted\ value\ of\ the\ mode$
- *If* T *is a biased estimator of* θ *then: (where* x *is the bias)*
  - $\theta = E(T) - x$
- *Standard error is the standard deviation of a sample distribution:* $\frac{\sigma}{\sqrt{n}}$
- *Central Limit Theorem states*
  - *If a sample of* $X_1, X_2, \ldots, X_n$ *has a population mean,* μ*, and variance,* σ^2*, then, for large n,* $X \approx \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- *Confidence intervals provide a range in which the mean is likely to lie*
  - *95% confidence limits are* $\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$
  - *90% confidence limits are* $\bar{x} \pm 1.6449 \times \frac{\sigma}{\sqrt{n}}$ *(other limits found by tables)*
  - *If* μ *lies outside the range of the CI limits then there is sufficient evidence to reject/question*
- *χ^2-distribution*
  - *Used to test 2 lists of frequencies (observed against expected)*
  - *If frequency of a column is* < 5 *then combine (collect) two columns until* ≥ 5
  - $\chi^2 = \sum \frac{(O-E)^2}{E}$
- *When testing χ^2against a particular distribution calculate the expected values as follows:*
  - *Uniform:* np *(e.g. dice 120 roles if fair each expected 20 times)*
  - *Continuous uniform:* np *(be aware to account for class boundaries as continuous)*
  - *Normal: account for class boundaries as continuous, then (using tables and standardizing) calculate* $P(X = x_1) = P(X < x_1)$ *then for* $P(X = x_2) = P(X < x_2) - P(X < x_1)$ *etc...*
  - *Poisson:* $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$
  - *Binomial:* $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
  - *Contingency tables: For* $E(AX) = P(A) \times P(X)$

	A	B
X		
Y		

- *Degrees of freedom, v, = number of cells (after grouping if necessary) – number of linear combinations connecting the frequencies*
- *Spearman's Rank Correlation Coefficient*
  - $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$
- *Hypothesis testing (H_0: null, H_1: alternative)Types of testing:*
  - *Between means!* $H_0: \mu_x = \mu_y$ ***accept*** H_0 *if between critical regions (not in critical regions)*
  - *Contingency Tables!* $H_0: no\ relationship$ ***accept*** H_0 *if* $\chi^2\ calculated < tables$
  - *Distribution Testing!* $H_0: x\ is\ a\ suitable\ model$ ***accept*** H_0 *if* $\chi^2\ calculated < tables$
  - *Spearman's Rank!*$H_0: \rho = 0$ ***accept*** H_0 *if* $\rho_s\ calculated < tables$
  - *Confidence Intervals! If* μ *lies outside* $\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$ *then evidence to question* μ