Fact Sheet - Statistics and Distributions

Given a population its mean value is denoted μ_X and its standard deviation is denoted by σ_X

Given a sample from a population the mean(average) of the sample values is denoted by $\overline{X}$ and the sample deviation is denoted by S where $S = (\sum_{i=1}^{N}(X_i - \overline{X})^2)/(N-1)$.

If sampling from finite populations without replacement use the *finite population correction factor* $f = \sqrt{\frac{P-N}{P-1}}$ where P is the size of the population and N the size of the sample, especially when the sample is 5 per cent or more of the population.

Sample Means - knowing population standard deviation σ_X

$$Z = (\overline{X} - \mu_X)/\frac{\sigma_X}{\sqrt{N}}$$

is $N(0,1)$- distributed.

Sample Proportions - knowing p the population proportion and p_s the sample proportion ($Np \geq 5$ and $Nq \geq 5$).

$$Z = (p_s - p)/(\sqrt{\frac{pq}{N}})$$

is $N(0,1)$-distributed - here $q = 1 - p$

Sample Means - knowing population standard deviation σ_X and using **finite population correction factor** f

$$Z = (\overline{X} - \mu_X)/\frac{\sigma_X . f}{\sqrt{N}}$$

is $N(0,1)$- distributed.

Sample Proportions - knowing p the population proportion and p_s the sample proportion ($Np \geq 5$ and $Nq \geq 5$) and using **finite population correction factor** f.

$$Z = (p_s - p)/(f\sqrt{\frac{pq}{N}})$$

is $N(0,1)$-distributed - here $q = 1 - p$

Hypothesis Testing

Hypothesis test for the mean - assuming σ_X is known.

$$Z = (\overline{X} - \mu_X)/(\frac{\sigma_X}{\sqrt{N}})$$

is $N(0,1)$, where N is the sample size.

Hypothesis text for the mean without knowledge of the population deviation - knowing $S = \sum_{i=1}^{N}(X_i - \overline{X})^2)/(N-1)$ instead.

$$t_{N-1} = (\overline{X} - \mu_X)/\frac{S}{\sqrt{N}}$$

is$t$-distributed, with $N-1$ degrees of freedom

Hypothesis Testing for Proportion of One Sample assuming sample is large enough ($Np \geq 5$ and $Nq \geq 5$)

$$Z = (p_S - p)/(\sqrt{\frac{pq}{N}})$$

is approximately $N(0,1)$-distributed. Here p_s stands for the sample proportion.

Hypothesis testing on differences between quantitative variables

Testing for the **difference between the means of two independent populations having equal known variances**.

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

is $N(0,1)$-distributed, assuming that the two populations have equal variances and sufficiently large sample sizes.

Testing for the **difference between the means of two independent normally distributed populations having equal variances** (though unknown). We assume we have only sample values $\overline{X_1}$, $\overline{X_2}$, S_1, S_2, sample averages and sample deviations respectively and sample sizes N_1, N_2, respectively.

Then

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{N_1} + \frac{1}{N_2})}}$$

where

$$S_P^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}$$

is $t_{N_1+N_2-2}$-distributed with $N_1 + N_2 - 2$ degrees of freedom

Testing a hypothesis about a **population variance**.
For a sample of size N the following statistic

$$\chi^2_{N-1} = \frac{(N-1)S^2}{\sigma_X^2}$$

is χ^2_{N-1}-distributed with $N - 1$ degrees of freedom Two-tailed test regarding σ_X^2 would be $H_0 : \sigma_X^2 = \sigma_0^2$

Testing **equality of variances** σ_1^2 and σ_2^2 from two independent populations.

The statistic

$$F_{N_1-1,N_2-1} = \frac{S_1^2}{S_2^2}$$

is F-distributed with $N_1 - 1, N_2 - 1$ degrees of freedom in the numerator and denominator respectively.

Two-tailed test - H_0: $\sigma_1^2 = \sigma_2^2$

Right hand rejection tail : $H_0 : \sigma_1^2 \leq \sigma_2^2$

Left hand rejection tail : $H_0 : \sigma_1^2 \geq \sigma_2^2$

Testing **differences between proportions from two independent populations** that have a certain characteristic- **for large sample sizes**. The statistic

$$Z = \frac{(p_{s_1} - p_{s_2}) - (p_1 - p_2)}{\sqrt{\overline{p}.\overline{q}(\frac{1}{N_1} + \frac{1}{N_2})}}$$

is $N(0,1)$, where $\overline{p} = \frac{X_1+X_2}{N_1+N_2}, \overline{q} = 1 - \overline{p}$ and p_{s_i} is the sample proportion from population i, and X_i is the number of elements in the i-th sample that have the characteristic of interest, and N_i is the size of sample i.

Testing for **differences between k Proportions**
For a contingency table with R rows and C columns of frequency of occurence information the statistic

$$\chi^2 = \sum\sum \frac{(f_O - f_t)^2}{f_t}$$

is chi-squared distributed with $(R-1)(C-1)$ degrees of freedom, where f_O is an observed frequency (an entry in the contingency table) and f_t is a theoretical frequency in the corresponding postion, computed under the assumption that there is no difference in proportions - using a pooled estimate.

Linear Regression Equation and Normal Equations

To fit the regression line given by the equation below

$$\hat{Y}_i = b_0 + b_1 X_i$$

solve for the parameters b_0 and b_1 by solving the normal equations below

$$\sum_{i=1}^{N} Y_i = nb_0 + b_1 \sum_{i=1}^{N} X_i$$

and

$$\sum_{i=1}^{N} X_i Y_i = b_0 \sum_{i=1}^{N} X_i + b_1 \sum_{i=1}^{N} X_i^2$$

Correlation

Sample correlation coefficient

$$r = \frac{\sum_{i=1}^{N} X_i Y_i - \frac{(\sum_{i=1}^{N}(X_i)(\sum_{i=1}^{N}(Y_i)}{N}}{\sqrt{\sum_{i=1}^{N} X_i^2 - \frac{(\sum_{i=1}^{N} X_i)^2}{N}}\sqrt{\sum_{i=1}^{N} Y_i^2 - \frac{(\sum_{i=1}^{N} Y_i)^2}{N}}}$$

Test Statistic for Determining if there is Correlation

H0: There is no correlation $\rho = 0$ - where $\rho =$ the population correlation coefficient

H1: There is correlation $\rho \neq 0$

Use the fact that the following statistic is

$$t_{n-2} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

t-distributed with $n-2$ degrees of freedom, where r is the sample correlation coefficient.

ANOVA - Analysis of Variance - One Way

Total Variation :

$$SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(X_{ij} - \overline{\overline{X}})^2$$

where c is the number of groups and n_j is the number of items in the jth group.
The Total variation is composed of the *Between-group variation (SSB)* and the *Within group variation*(SSW)
where $SSB = \sum_{j=1}^{c} n_j(\overline{X_j} - \overline{\overline{X}})^2$ and $\overline{X_j}$ is the sample mean of group j, and $\overline{\overline{X}}$ is the grand mean.
Then

$$F = \frac{(MSB)}{(MSW)} = \frac{\frac{SSB}{c-1}}{\frac{SSW}{n-c}}$$

is $F_{c-1,n-c}$-distributed.
Computational Formulae for ANOVA -

$$SSB = \sum_{j=1}^{c}\frac{T_j^2}{n_j} - \frac{(GT)^2}{n}$$

where GT is the grand total and

$$SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}X_{ij}^2 - \sum_{j=1}^{c}\frac{T_j^2}{n_j}$$

NONPARAMETRIC METHODS

Wald Wolfowitz One-Sample Runs Test for Randomness
The following is $N(0,1)$-distributed-

$$Z = \frac{U - (\frac{2n_1n_2}{n} + 1)}{\sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}}}$$

where U is the total number of runs, n_1 is the number of successes in sample, n_2 is the number of failures in sample and $n = n_1 + n_2$ is the sample size.

Wilcoxon Signed-Ranks Test For $n > 20$ the following test statistic is approximately $N(0,1)$-distributed:

$$Z = \frac{W - (\frac{n(n+1)}{4})}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

where $W = \sum_{i=1}^{n} R_i^{(+)}$ is the sum of positive ranks.

Wilcoxon Rank-Sum Test The following is approximately $N(0,1)$-distributed for large sample sizes ($n_1 \geq 10$ and $n_2 \geq 10$)

$$Z = \frac{T_1 - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}}$$

where T_1 is the sum of the ranks assigned to the n_1 observations in the first sample

Kruskal-Wallis Test for c Independent Samples

The following statistic is approximately χ^2-distributed with $(c-1)$ degrees of freedom (when the sample sizes in each group are greater then 5)

$$H = \left(\frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} \right) - 3(n+1)$$

where n is the total number of observations over the combined samples, $n = n_1 + n_2 + n_3 + ... + n_c$, n_j is the number of observations in the jth sample and T_j^2 is the square of the sum of the ranks assigned to the jth sample.

Friedman's Rank test for c Related Samples

The following statistic is approximately χ^2-distributed with $c-1$ degrees of freedom (when the number of blocks is greater than five)

$$F_R = \frac{12}{nc(c+1)} \sum_{j=1}^{c} R_j^2 - 3n(c+1)$$

where R_j^2 is the square of the rank total for group j, n is the number of independent blocks, and c is the number of groups.

Spearman's Rank-Correlation Procedure The following statistic is approximately $N(0,1)$-distributed when the sample size n is not very small

$$Z = r_s\sqrt{n-1}$$

where

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_{R_i}^2}{n(n^2-1)}$$

and $d_{Ri} = R_{X_i} - R_{Y_i}$, the rank difference scores