Inhalt

[**X01: DATA Step** 6](#_Toc111713644)

[**X02: PROC Step** 6](#_Toc111713645)

[**P01: Create library** 6](#_Toc111713646)

[**P02: Import DATA** 6](#_Toc111713647)

[Importing data from Excel file: 6](#_Toc111713648)

[Importing data from CSV file: 6](#_Toc111713649)

[Importing data from SPSS file: 6](#_Toc111713650)

[**P03: Create dataset** 7](#_Toc111713651)

[**P04: Variable definition** 7](#_Toc111713652)

[**P05: Attribute definition** 7](#_Toc111713653)

[**P06: Enter data** 7](#_Toc111713654)

[CARDS/DATALINES 7](#_Toc111713655)

[**P07: Sort data** 8](#_Toc111713656)

[**P08: Merge data** 8](#_Toc111713657)

[**P09: SQL** 8](#_Toc111713658)

[**X03: Data cleaning** 8](#_Toc111713659)

[**P10: Recode variable** 8](#_Toc111713660)

[**P11: SELECT cases** 9](#_Toc111713661)

[SELECT Statement: 9](#_Toc111713662)

[IF Statement: 9](#_Toc111713663)

[**P12: Calculate new variable** 9](#_Toc111713664)

[**X04: Descriptive statistics** 9](#_Toc111713665)

[**X05: Measurement types** 10](#_Toc111713666)

[**X05A: Categorical variables** 10](#_Toc111713667)

[Nominal 10](#_Toc111713668)

[Dichotomous 10](#_Toc111713669)

[Ordinal 10](#_Toc111713670)

[**X05B: Continuous variable** 10](#_Toc111713671)

[**X06: Descriptive statistics for continuous variables** 11](#_Toc111713672)

[**X06A: Central tendency** 11](#_Toc111713673)

[Mean (Arithmetic) 11](#_Toc111713674)

[Median 11](#_Toc111713675)

[Mode 11](#_Toc111713676)

[**X06B: Statistics for dispersion measures** 11](#_Toc111713677)

[Range 11](#_Toc111713678)

[Interquartile range 11](#_Toc111713679)

[Standard deviation and variance 11](#_Toc111713680)

[Standard error of mean (SEM) 11](#_Toc111713681)

[Coefficient of variation (CV): 11](#_Toc111713682)

[**X07: Graphic expression of descriptive statistics** 12](#_Toc111713683)

[**X07A: Graphic of central tendency** 12](#_Toc111713684)

[Bar chart 12](#_Toc111713685)

[Boxplot 12](#_Toc111713686)

[**X07B: Graphics of dispersion** 12](#_Toc111713687)

[Error bars 12](#_Toc111713688)

[Histogram 12](#_Toc111713689)

[**X08: Testing distribution and normality:** 12](#_Toc111713690)

[Statistics of distribution 12](#_Toc111713691)

[**X08A: SKEWNESS AND KURTOSIS** 13](#_Toc111713692)

[SKEWNESS 13](#_Toc111713693)

[KURTOSIS 13](#_Toc111713694)

[**X08B: Graphics of distribution** 13](#_Toc111713695)

[QQ-Plot 13](#_Toc111713696)

[PP-Plot 14](#_Toc111713697)

[Boxplot 14](#_Toc111713698)

[**X08C: Normality tests** 14](#_Toc111713699)

[**X09: Tests for Homogeneity of Variance** 14](#_Toc111713700)

[**X10: Frequency tables** 15](#_Toc111713701)

[**X11:Contingency tables (Cross tables)** 15](#_Toc111713702)

[**X12:Comparisons** 15](#_Toc111713703)

[**X13:Sample Types** 15](#_Toc111713704)

[**X14: Parametric test assumptions** 15](#_Toc111713705)

[**Continuous variable comparison** 15](#_Toc111713706)

[**X15: One sample median test** 15](#_Toc111713707)

[**X16: One sample t-test** 16](#_Toc111713708)

[**X17: Two independent samples t-test** 16](#_Toc111713709)

[**X18: One Way ANOVA** 16](#_Toc111713710)

[**X19: Post-Hoc tests for ANOVA** 16](#_Toc111713711)

[**X20: Mann-Whitney U or Wilcoxon rank sum test** 16](#_Toc111713712)

[**X21: Kruskal-Wallis test** 16](#_Toc111713713)

[**X22: Posthoc test adjustment** 17](#_Toc111713714)

[**X23: Paired sample t-test** 17](#_Toc111713715)

[**X24: One-way repeated measures ANOVA** 17](#_Toc111713716)

[**X25: Wilcoxon signed rank sum test** 17](#_Toc111713717)

[**X26: Friedman test** 17](#_Toc111713718)

[**Categorical variable comparisons** 17](#_Toc111713719)

[**X27: Chi-square** 17](#_Toc111713720)

[**X28: Fisher’s exact test** 17](#_Toc111713721)

[**X29: One sample Chi-square** 17](#_Toc111713722)

[**X30: binomial test** 18](#_Toc111713723)

[**X31: McNemar’s test** 18](#_Toc111713724)

[**X32: Cochran’s Q Test** 18](#_Toc111713725)

[**X33: Correlation** 18](#_Toc111713726)

[**X34: Pearson correlation** 18](#_Toc111713727)

[**X35: Spearman correlation** 18](#_Toc111713728)

[**X36: Kendall Tau correlation** 19](#_Toc111713729)

[**X37: Cronbach alpha** 19](#_Toc111713730)

[**X38: Partial correlation** 19](#_Toc111713731)

[**X39: Regresssion analysis** 19](#_Toc111713732)

[**X40: Logistic regression** 19](#_Toc111713733)

[**X41: Ordinal logistic regression** 19](#_Toc111713734)

[**X42: Linear regression** 19](#_Toc111713735)

[**X43: Cox regression** 20](#_Toc111713736)

[**X44: Regression assumption** 20](#_Toc111713737)

[**X45: Regression coefficient** 20](#_Toc111713738)

[**X46: Goodness of fit** 20](#_Toc111713739)

[**X47: Power and sample size determination** 20](#_Toc111713740)

[**X48: Sample size** 20](#_Toc111713741)

[**X49: Power** 21](#_Toc111713742)

[**S01 UNIVARIETE Procedure** 22](#_Toc111713743)

[Normality test with graphs (Histogram, QQ-Plot, PP-Plot) 22](#_Toc111713744)

[Testing Homogenity of Variance 22](#_Toc111713745)

[One sample median test 22](#_Toc111713746)

[**S02 MEANS Procedure** 22](#_Toc111713747)

[**S03 GLM Procedure** 22](#_Toc111713748)

[Homogenity of variances 22](#_Toc111713749)

[One Way ANOVA 22](#_Toc111713750)

[One-way ANOVA and Post-Hoc tests for ANOVA 23](#_Toc111713751)

[Repeated Measures ANOVA 23](#_Toc111713752)

[RUN; 23](#_Toc111713753)

[**S04: FREQ Procedure** 23](#_Toc111713754)

[Frequency tables 23](#_Toc111713755)

[Contingency tables 23](#_Toc111713756)

[Chi-square test and Fisher’s exact test 23](#_Toc111713757)

[One sample chi-square test 23](#_Toc111713758)

[Binomial test 24](#_Toc111713759)

[McNemar test 24](#_Toc111713760)

[Cochrane Q 24](#_Toc111713761)

[**S05: TABULATE Procedure** 24](#_Toc111713762)

[**S06: TTEST Procedure** 24](#_Toc111713763)

[One sample t-test 24](#_Toc111713764)

[Two independent samples t-test 24](#_Toc111713765)

[Paired samples t-test 24](#_Toc111713766)

[**S07: NPAR1WAY Procedure** 25](#_Toc111713767)

[Mann-Whitney U or Wilcoxon rank sum 25](#_Toc111713768)

[Kruskal Wallis test 25](#_Toc111713769)

[**S08: MULTTEST Procedure** 25](#_Toc111713770)

[**S08: CORR Procedure** 25](#_Toc111713771)

[Pearson, Spearman and Kendall Tau-b correlations 25](#_Toc111713772)

[Chronbach alpha 25](#_Toc111713773)

[Partial correlation 25](#_Toc111713774)

[**S09: LOGISTIC Procedure** 25](#_Toc111713775)

[Binary logistic regression 25](#_Toc111713776)

[Ordinal logistic regresion 26](#_Toc111713777)

[**S10: REG Procedure** 26](#_Toc111713778)

[Linear regression 26](#_Toc111713779)

[**S11: PHREG Procedure** 26](#_Toc111713780)

[Cox regression 26](#_Toc111713781)

[**S12: POWER Procedure** 26](#_Toc111713782)

[One sample mean comparison (ONESAMPLEMEAN) 26](#_Toc111713783)

[One sample mean estimation 26](#_Toc111713784)

[One sample proportion estimation (ONESAMPLEFREQ) 26](#_Toc111713785)

[Two sample mean comparison (TWOSAMPLEMEANS) 27](#_Toc111713786)

[One way ANOVA (ONEWAYANOVA) 27](#_Toc111713787)

[Paired means (PAIREDMEANS) 27](#_Toc111713788)

[Two sample frequencies (TWOSAMPLEFREQ) 28](#_Toc111713789)

[McNemar Exact Conditional Test (PAIREDFREQ) 29](#_Toc111713790)

[Survival (TWOSAMPLESURVIVAL) 29](#_Toc111713791)

[Multipl regression (MULTREG) 30](#_Toc111713792)

[Logistic regression (LOGISTIC) 31](#_Toc111713793)

# **X01: DATA Step**

You can use DATA steps to create, describe, or modify your data sets. With DATA steps, you can inform SAS how to read the data, and create or delete new variables and observations. You can also convert data in another format to a SAS dataset.

# **X02: PROC Step**

In PROC step, you can find a group of SAS statements that execute statistical procedures. It is possible to analyze data, to produce reports, or to manage SAS files in PROC steps.

# **X03: Data cleaning**

Data cleansing is a set of activities to detect and correct errors and inconsistencies in data sets. It consists of a series of activities with the goal of developing consistent data. These steps may include one or more of the following: 1. importing data 2. merging records 3. handling missing data, 4. standardization or normalization 5. eliminating duplicates 6. verification 7. finding and handling redundancies 8. exporting data.

Besides specific solutions, you can also use incidental instructions for data cleansing.

# **X04: Descriptive statistics**

Descriptive statistics is the summary of data for better understanding. That is, the goal is the clear presentation and preparation of empirical data. Tables, graphs and the determination of relevant key figures are used to obtain an overview of the entire data material. They are used to describe the central tendency, the dispersion and the distribution of a sample. In concrete terms, this means that these measures can be used, for example, to estimate how old a respondent is on average in his or her study or how high the age heterogeneity is.

# **X05: Measurement types**

The measurement types, is one of the most important theme in statistics. Scales can be classified in a hierarchy according to their level. That is, the higher they become, the greater the information content. It also increases the number of mathematical operators that can be applied to the data.

The scale level determines,

* which (mathematical) operations are permissible,
* which transformations can be performed,
* which information the corresponding characteristic provides,
* which interpretations allow expressions of the corresponding characteristic.

# **X05A: Categorical variables**

Categorical variables include a restricted number of categories or distinct groups.

## Nominal

Variables that are part of the nominal scale are,

* discrete
* qualitative
* divisible into categories.

Example: birth place.

## Dichotomous

The nominal variables that have only two categories or levels are dichotomous variables.

Example: "Yes" or "No" questions.

## Ordinal

Variables that are part of the ordinal scale are;

* discrete
* qualitative
* divisible into categories
* to be put into a RANK order.

Example: Likert type scale questions (1. Strongly disagree 2. Disagree 3. Neither agree nor disagree 4. Agree 5. Strongly agree).

# **X05B: Continuous variable**

Continuous variables are numeric variables with an infinite number of values. Between any two scale values, an infinite number of intervals can be defined.

Example: Serum cholesterol level or birth date are continuous variables.

# **X06: Descriptive statistics for continuous variables**

# **X06A: Central tendency**

A measure of central tendency is a single value that describe a set of data.

## Mean (Arithmetic)

The mean indicates the average of all values. The mean is equal to the sum of all the values in the data set divided by the number of values in the data set.

## Median

The median divides the data set exactly in the middle, so that there are as many values above it as below it.

When the number of elements in the sample (n) is odd, the median is equal to the value of the ((n+1)/2)th element. When n is even, the median is equal to the average of the (n/2)th and the (n/2+1)th element. To form the median, the data must at least meet the requirements of an ordinal scale.

## Mode

The modal value simply describes the value that occurs most frequently. You can calculate mode for all scale types, also for nominal scale data.

# **X06B: Statistics for dispersion measures**

Dispersion (or variability, spread) is the statistical parameter or characteristic that provides information about the variability of data. The widely used dispersion measures are range, interquartile range, standard deviation, standard error of mean, variance.

## Range

The range is the difference between largest and smallest data value, very sensitive to outliers.

## Interquartile range

Interquartile range is the difference between the 25th and 75th percentile. It is range in which the "middle" 50% of all values lie.

## Standard deviation and variance

Standard deviation (SD) is the square root of sum of squared deviation from the mean divided by the number of observations minus one. Variance is square of standard deviation.

## Standard error of mean (SEM)

Mathematically, the SEM is calculated by dividing the SD by the square root of the sample size.

## Coefficient of variation (CV):

All measures of dispersion described above depend on the unit of measurement. The standard deviation is SD/Mean). The CV is independent of the unit of measurement.

# **X07: Graphic expression of descriptive statistics**

# **X07A: Graphic of central tendency**

## Bar chart

A bar chart is suitable for the graphical representation of the frequency of expressions of arbitrarily scaled characteristics, whereby you can represent absolute or relative frequencies.

## Boxplot

The box plot (also box-whisker plot) is a diagram that is used for the graphical representation of a continuous or ordinal characteristic. It combines different scatter and position measures in one representation. Therefore, all values of the so-called five-point summary, i.e. the median, the two 1st and 3rd quartiles and the minimum and maximum values are displayed. Outliers and extreme values can be displayed.

# **X07B: Graphics of dispersion**

## Error bars

Error bars are used in the graphical representation of numeric data and are used to visualize the possible deviations of the measured values from the actual value of the considered measured variable, which are based errors.

## Histogram

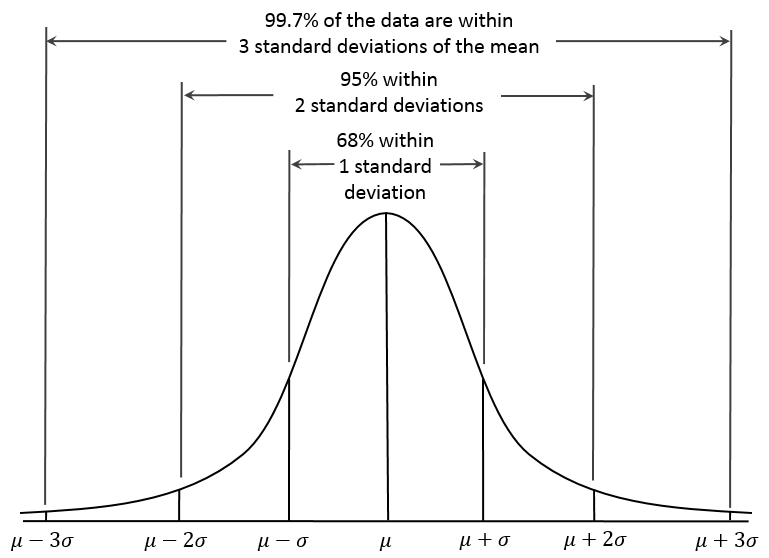
A histogram is a graphical representation of the frequency distribution of scale variables. It requires the division of the data into classes (bins), which can have a constant or various width. The area of column represents the relative or absolute frequencies of the class.

# **X08: Testing distribution and normality:**

The best known distribution is the normal distribution, familiar the nice bell-shaped curve. On the other hand, data following any probability distribution can also be valuable. It is possible to normalize data by conversion (such as Z-conversion, logarithmic or exponential conversion).

## Statistics of distribution

The normal distribution is a very common probability distribution for continuous variables. In the parametric approach to inferential statistics, values are assumed to be normally distributed.



The main characteristics of normal distribution are:

* The shape of the distribution is symmetrical around the mean,
* The mean, median and mode are same point,
* The area under the curve is 68% for ± 1SD, 95% for ± 2SD and 99.7% ±3 SD.

# **X08A: SKEWNESS AND KURTOSIS**

## SKEWNESS

Skewness is a measure of the asymmetry of the distribution. Skewness; If skewness;

* ≤ -1 or ≥1 highly skewed,
* -1 - -0.5 or 0.5 - 1 moderately skewed,
* -0.5 - 0.5 approximately symmetric.

## KURTOSIS

Kurtosis provides information about the height and tail length of the central peak.

* **Leptokurtic** - kurtosis > 3 (Very thin tails compared to the normal distribution)
* **Mesokurtic** - kurtosis = 3 (Similar tails to the normal distribution)
* **Platykurtic** - kurtosis < 3 (Very thick tails compared to the normal distribution)

# **X08B: Graphics of distribution**

## QQ-Plot

The quantile-quantile plot (QQ plot) is an explorative, graphical tool in which the quantiles of two statistical variables are plotted against each other in order to compare their distributions. In order to interpret normal distribution the QQ plot is used to graphically represent observed values and a normal distribution is present.

Quantile: Quantiles belong to the measures of location in statistics. They divide a certain amount of data so that one part p is less than or equal to and the other part 1-p is greater than or equal to the quantile. The 10% quantile or 0.1 quantile, for example, indicates that exactly 10 percent of the values in a distribution are below the quantile. The rest of the values are above it.

## PP-Plot

In probability-probability plots (PP plots), the observed distribution function is plotted against the theoretical distribution function.

## Boxplot

The boxplot or box-whisker plot is a diagram that clearly displays the most important location and dispersion measures. The minimum, the 1st quartile, the median, the 3rd quartile and the maximum are shown.

Outliers and extreme values can also be displayed in the box plot.

# **X08C: Normality tests**

The normality tests are may use together with graphical assessment of normality. The greatest shortfall of normality test is; for small sample sizes, normality tests have little power to reject the null hypothesis and the small samples most often pass normality tests. In the normality tests the null hypothesis is “sample distribution is normal.” The significant test results (p<0.05) mean the distribution is non-normal.

You can use the normality tests along with the graphical assessment of normality. The null hypothesis is "The distribution of the sample is normal." The significant test results (p<0.05) mean that the distribution is not normal.

The tests available in SAS:

**Kolmogorov-Smirnov (K-S) test with Lilliefors correction:** K-S test is its high sensitivity to extreme values. Lilliefors test provide better power than classical K-S via correction.

**Shapiro-Wilk test:** This test provides better power than the K-S test. Suitable for little samples.

**Anderson-Darling test:** You can use the Anderson-Darling test starting for a sample size of n≥8.

# **X09: Tests for Homogeneity of Variance**

In the t-test or ANOVA, one of the assumptions is the homogeneity of variance (HOV). Therefore, ANOVA is resistant to small deviations from the HOV assumption, you only need to worry about large deviations. Three common HOV tests:

**Bartlett’s Test:** Bartlett’s Test is the most common and powerful test for the HOV assessment, in case of all groups are normally distributed.

**Levene’s Test:** Levene’s Test is not sensitive to symmetric heavy-tailed distributions.

**Brown-Forsythe Test:** The Brown-Forsythe test is insensitive to skewed distributions and deviations from normality.

**Folded F:** The folded form of the F-statistic tests the hypothesis that the variances are equal, and it is automatically reported in the t-test in SAS.

# **X10: Frequency tables**

Frequency tables are created when you want to display the absolute and relative frequencies of the values of your variables. A frequency table represents how often each feature occurs.

# **X11:Contingency tables (Cross tables)**

Contingency tables (Crosstabs), are descriptive statistics to get an overview of at least two categorical variables. In the crosstab, it can be read how often the combination of the values of the characteristics occurs.

# **X12:Comparisons**

Statistical methods are all methods that help you to investigate your hypotheses. Which of these methods is useful for you also depends on your data. Therefore, you should examine your data in the first step.

In inferential statistics, the null hypothesis (H0) means that two possibilities are equal. Statistical tests can be performed to calculate the probability that the null hypothesis is true. For example,

H0 μ1= μ2 OR p1= p2 (Mean or proportion of first sample equal to second sample)

H1 μ1≠ μ2 OR p1≠ p2 (Mean or proportion of first sample not equal to second sample)

# **X13:Sample Types**

**Dependent samples:** If your values come from different measurements of the same sample or one sample influences the values in the other sample, then the samples are dependent (e.g. pre-post evaluation or right foot length-left foot length).

**Independent samples:** If your values come from subjects in different samples or one sample does not reveal any information about the values in the other sample, then the samples are independent (comparison treatment group with placebo group).

# **X14: Parametric test assumptions**

The main parametric test assumptions are:

Normality: your data should have a normal distribution

Homogeneity of variances: Your data from multiple groups should have the same variance

Linearity: for some specific tests, your data should have a linear relationship.

# **Continuous variable comparison**

# **X15: One sample median test**

This test allows you to check whether the median of the given variable or distribution is different from a given value. The following sign test is a simple nonparametric procedure at the same time. You should report Pr >= |M| values in Outputs.

# **X16: One sample t-test**

The one-sample t-test is a significance test that uses the mean of a sample to test whether the mean of a population is equal to (null hypothesis) a certain value.

# **X17: Two independent samples t-test**

The two-sample t-test (or student t-test) uses the mean values of two independent samples to test how the mean values of two populations different from the other. It is assumed that the data of the samples originate from 2 normally distributed population and equal variances.

# **X18: One Way ANOVA**

Analysis of variance (ANOVA) is the name given to a large group of data-analytical and structure-testing statistical methods that can be used for numerous applications. The variance of one or more target variables is explained by the influence of one influencing variables (factors). The simplest form of ANOVA tests the influence of a nominal variable on a continuous variable by comparing intra-group and inter-group variances.

# **X19: Post-Hoc tests for ANOVA**

Post-hoc tests are significance tests from statistics that help overcome the false discovery rate. If you have significant differences in a group of means using One-Way ANOVA or a Kruskal-Wallis test, you should use post-hoc tests for pairwise comparisons of means to obtain information about which means are significantly different from each other. Or, they provide information about which group means are not significantly different by group-wise comparisons.

The Tukey post-hoc test should be used when you would like to make pairwise comparisons between group means when the sample sizes for each group are equal.

If the sample sizes are not equal, you can use a modified version of the test known as the Tukey-Kramer test.

# **X20:** **Mann-Whitney U or Wilcoxon rank sum test**

The Mann Whitney U test or Wilcoxon rank sum test is a nonparametric alternative to the two-sample t-test (Student's t-test). Samples t-test (Student's t-test). Mann Whitney U test based on the order in which the observations from the two samples are available.

# **X21: Kruskal-Wallis test**

The Kruskal-Wallis test is a hypothesis test for multiple independent samples comparison. It used when the study group does not satisfy the parametric test assumption for the one-way ANOVA.

Since the Kruskal-Wallis test is a nonparametric, distribution-free procedure, the data used do not have to be normally distributed, unlike the ANOVA. The only requirement is that the data have at least an ordinal scale level.

# **X22: Posthoc test adjustment**

You can use PROC MULTTEST to solve the multiplicity problem. PROC MULTTEST can adjust the raw p-values of multiple tests using one of several adjustment methods. You must enter a p-value list to perform multiplicity adjustment.

# **X23: Paired sample t-test**

The paired samples t-test is a statistical test used to determine if there is a significant difference between two dependent groups.

# **X24: One-way repeated measures ANOVA**

Repeated measures designs usually include measuring the same subjects more than once. The idea behind this, keeping the subjects the same to better estimate variance and attribute possible effects. As a result, experimental designs with repeated measures also generally have higher statistical power.

# **X25: Wilcoxon signed rank sum test**

The Wilcoxon test or Wilcoxon signed-rank test checks whether two dependent samples differ significantly from each other. The Wilcoxon test is a non-parametric test and is therefore subject to significantly lower requirements than its parametric counterpart, the t-test for dependent samples.

# **X26: Friedman test**

The Friedman test is a non-parametric alternative to repeated measures ANOVA. It is used to determine whether there is a statistically significant difference between the means of three or more groups in which the same subjects appear in each group.

The Friedman test perform by vertically prepared data in SAS. If you have horizontally arranged data you should convert it vertical form and calculate ranks before analysis.

# **Categorical variable comparisons**

# **X27: Chi-square**

The Chi-square test is a statistical test procedure that can make statements about the relationship between nominal or ordinal variables. The Chi-square test is also a hypothesis test. By definition, because of the test examines the relationship between the variables in terms of independence, it is often also referred to as the chi-square goodness-of-fit test or chi-square independence test. The representation and calculation runs thereby mainly over the cross table.

# **X28: Fisher’s exact test**

Fisher's exact test is used to determine whether or not there is a significant association between two categorical variables. It is usually used as an alternative to the chi-square test of independence when one or more of the cell counts is less than 5.

# **X29: One sample Chi-square**

You can use a one sample Chi-square test (or Chi square goodness-of-fit test) to see if the observed proportions of a categorical variable differ from the assumed proportions.

# **X30: binomial test**

A one-sample binomial test allows you to test whether the proportion of successes in a two-level categorical dependent variable is significantly different from an assumed value. Commonly, it is tested whether the sample originates from the respective population.

# **X31: McNemar’s test**

As a member of the chi-square test, you can use the McNemar test to compare two related samples with regard to a dichotomous characteristic.

# **X32: Cochran’s Q Test**

The Cochran Q test is a nonparametric test procedure designed to determine whether two or more dependent samples differ significantly in the proportion of cases in each of two categories. Extension of the Mc Nemar test to more than 2 time points AND non-dichotomous variable.

# **X33: Correlation**

Correlation analysis and the correlation coefficient is the specific measure to quantify the strength of the linear relationship between two variables in a correlation analysis.

Correlation is a broad class of statistical test interest in the context of dependence, although in common usage it usually refers to how close two variables are to a linear relationship. It provides a method for measuring the strength of a linear relationship between two numerical variables.

Correlation coefficient serves as a measure of the strength of the correlation of the interval-scaled characteristics and takes on values between -1 and 1.

Below are rules of thumb for the correlation coefficient.

±0.91 to ±1.00 Very strong  
±0.71 to ±0.90 High  
±0.41 to ±0.70 Moderate  
±0.21 to ±0.40 Small   
±0.00 to ±0.20 Slight or negligible  
(Source: Hair Jnr, Money, Samouel, & Page, 2007)

# **X34: Pearson correlation**

Pearson correlation is a simple way to determine the linear relationship between two variables.

# **X35: Spearman correlation**

The Spearman correlation coefficient (rank correlation coefficient), which is also referred to as Spearman's Rho, is similar to the Pearson correlation coefficient a method to calculate correlations between variables. The correlation is calculated using previously assigned ranks, whereby the exact distance between the data points is not relevant.

# **X36: Kendall Tau correlation**

The Kendall-Tau correlation coefficient tests two variables for an nondirectional relationship. It works for two ordinal variables, two metric variables, or a mixture of both. It shows either a positive correlation, a negative correlation or no correlation.

# **X37: Cronbach alpha**

Related items can be summed to obtain a total score for each participant. The Cronbach's coefficient alpha estimates the reliability of this type of scale by determining the internal consistency of the test or the average correlation of items within the test.

Cronbach's alpha, like the correlation coefficient, can take values between -1 and 1.

# **X38: Partial correlation**

The partial correlation calculates the correlation between two variables excluding the effect of a third variable. The partial correlation says how strongly the 1st variable x correlates with the 2nd variable y, if the correlation of both variables with the variable z is calculated out.

# **X39: Regresssion analysis**

Regression is a statistical method that makes it possible to model relationships between a dependent variable and one or more independent variables.

Regression analysis is used to infer or predict the relationship between one or more variables and another variable.

# **X40: Logistic regression**

Logistic regression or logit model refers to regression analyses for (usually multiple) modeling of the distribution of dependent categorical variables. Unless logistic regressions are more specifically identified as multinomial or ordina logistic regressions, binomial logistic regression for dichotomous (binary) dependent variables is usually meant. Here, the independent variables can be nominal, ordinal or scale.

# **X41: Ordinal logistic regression**

The goal of ordinal regression is to predict probabilities for the occurrence of individual categories as a function of covariates.

Likert scales are an example of the application of ordinal regression.

# **X42: Linear regression**

In linear regression, you try to predict the values of one variable using one or more other variables. The variable to be predicted is called the dependent variable. The variables used to predict are called predictors or independent variables.

To predict the dependent variable, you look at the relationship between the predictors (independent variables) and the dependent variable. The closer the relationship, the better you can predict the dependent variable. As the name suggests, however, linear regression only looks at linear relationships.

# **X43: Cox regression**

Cox regression or Cox proportional hazards model is essentially a statistical regression model commonly used in medical research to study the association between patient survival and one or more predictor variables.

Cox proportional hazards regression analysis performs well for both quantitative predictor variables and categorical variables. In addition, the Cox regression model enhances survival analysis methods to simultaneously assess the effects of multiple risk factors on survival.

# **X44: Regression assumption**

When performing multiple regression, a number of preconditions must be tested in order to obtain valid results.

The main preconditions are:

* Normal distribution
* Homoskedasticity (homogenic dispersion)
* Linearity
* No strong multicollinearity
* Uncorrelatedness of the errors
* Scale properties
* No strong outliers

# **X45: Regression coefficient**

In regression analysis, regression coefficients represent the influence of different predictor variables in the regression equation. The regression weights beta indicate by how many units the value of the dependent variable increases or decreases if the predictor variable increases by 1.

# **X46: Goodness of fit**

The goodness of fit indicates "how well" an estimated model can explain a set of observations. Measures of goodness of fit allow a statement about the deviation between the theoretical values of the investigated random variables, which are expected or predicted on the basis of the model, and the actually measured values.

# **X47: Power and sample size determination**

In statistical hypothesis testing, you state a null hypothesis, that is, the claim that the effect does not exist, and attempt to gather evidence to reject it in favor of the null hypothesis. Evidence is collected in the form of sample data, and a statistical test is used to evaluate it.

Type I (alpha) error: Rejection null hypothesis when it is true.

Type II (beta) error: Acceptance, but the null hypothesis is false.

# **X48: Sample size**

Sample size planning is an important step in the planning of quantitative studies and it depends on the test procedures to be used.

In the context of clinical trials, however, a priori case number planning is mandatory.

A pre-study sample size calculation is based on assumptions, on rough ideas of how the data, possible correlations and assumed differences might look like. Expected means, scatter, and significances must be named and included in the calculation.

Results of previous studies serve as the basis for these assumed values. The more precise the idea of the expected values is, the more precisely the case numbers can be planned.

# **X49: Power**

Power can be defined as the probability of correctly rejecting a false null hypothesis.

When designing a study, you should set the power level in the same way as one would set the significance level. Often a statistical power of 80% is chosen, so that a true difference will not be detected in 20% of the cases.

# **P01: Create library**

To read or write SAS data sets, you must first assign a name (LIBNAME) to a library. Think of this process as assigning a nickname to a folder.

**LIBNAME** MyGitHub "c:\MyGitHub\";

# **P02: Import DATA**

If you have data in any supported format (Excel, CVS, SPSS etc) you can import them into SAS before analysis. You have to create and define a library before data importing.

## Importing data from Excel file:

**PROC** IMPORT DATAFILE="c:\MyGitHub\ExcelData.xlsx"

OUT=MyGitHub.ExcelData

DBMS=xlsx;

SHEET=”Sheet1”; /\* If you don’t point out a sheet SAS uses first sheet\*/

RUN;

## Importing data from CSV file:

**PROC** IMPORT DATAFILE="c:\MyGitHub\CSVData.csv"

OUT=MyGitHub.CSVData

DBMS=csv;

RUN;

## Importing data from SPSS file:

**PROC** IMPORT DATAFILE="c:\MyGitHub\SPSSData.sav"

OUT=MyGitHub.SPSSData

DBMS=sav;

RUN;

# **P03: Create dataset**

To enter data set you should define variable names first. If

**DATA** MyGitHub.TestData;

INPUT CaseID $ DOB Gender $

/\*To define character variable, you should use $ sign after variable name\*/

CARDS;

003 2011 female

004 2021 male;

RUN;

# **P04: Variable definition**

by INPUT statement you can define names of your variables that will be include in dataset. Data values must be separated by at least one space. List input defining a string variable, it requires to specify the variable names and a $ (dollar) sign.

# **P05: Attribute definition**

by ATTRIB statement we can define attributes of our variables that will be include in dataset.

|  |  |
| --- | --- |
| Attribute | Description |
| Name | 1to 32 characters and must begin with letter or underscore |
| Length | It refers number of bytes used to store variables |
| Format | Format is an instruction that how SAS **write** data values. |
| Informat | Informat is an instruction that how SAS **read** data values. |
| Label | You can assign labels to display more descriptive information about the variables. Valid labels must up to 256 characters |

**DATA** MyGitHub.LeadTestRev;

ATTRIB CaseID label = 'Case ID’ length = 8

YOB label = ‘Year of birth' length = **8**

BloodLeadLevel= ‘Lead Level’ length =**8**;

SET MyGitHub.LeadTest;

RUN;

# **P06: Enter data**

## CARDS/DATALINES

Use the CARDS or DATALINES statement with an INPUT statement to read data that you enter directly in the program, rather than data stored in an external file.

**DATA** MyGitHub.LeadTest;

INPUT CaseID $ DOB Gender $ City $ MothersJob $ BloodLeadLevel;

Age=**2022**-DOB;

CARDS;

001 2006 female Ankara working 0.51

002 2009 male Berlin working 3.31

003 2011 female Athens notworking 6.52;

RUN;

# **P07: Sort data**

You can sort SAS dataset observations by the values of one or more character or numeric variables into the original dataset or a new dataset using the SORT procedure.

**PROC** SORT DATA= MyGitHub.Lead OUT= MyGitHub.Lead\_Sorted;

BY CaseID descending;

RUN;

# **P08: Merge data**

If you have two dataset that have same order, you could use one-to-one merge. Otherwise using match-merging by using a key variable should be performed.

**DATA** MyGitHub.LeadMerged;

MERGE MyGitHub.Lead\_1 MyGitHub.Lead\_2;

BY CaseID;

RUN;

# **P09: SQL**

PROC SQL is a powerful SAS procedure that combines the functionality of DATA and PROC steps into a single step.

You can use PROC SQL to sort, merge, subdivide, join (merge) and concatenate data sets, create new variables and print the results or create a new table.

**PROC** **SQL**;

CREATE TABLE MyGitHub.LeadMerge\_LeftJoin AS

SELECT Lead\_1.\*, Lead\_2.PostLead

FROM MyGitHub.Lead\_1 Lead\_1 LEFT JOIN MyGitHub.Lead\_2 Lead\_2

ON MyGitHub.Lead\_1.CaseID= MyGitHub.Lead\_1.CaseID;

**QUIT**;

# **P10: Recode variable**

IF-THEN-ELSE rules can be used to classify or recode variables.

**DATA** MyGitHub.LeadClassified;

SET MyGitHub.LeadMerge;

IF BloodLeadLevel LE 1.99 THEN LeadClass=1;

ELSE IF BloodLeadLevel GE 2 AND BloodLeadLevel LE 3.99 THEN LeadClass=2;

ELSE LeadClass=3;

RUN;

|  |  |  |
| --- | --- | --- |
| **Symbolic** | **Alphabetic** | **Mean** |
| = | EQ | equals |
| ^= or ~= | NE | not equal |
| > | GT | greater than |
| < | LT | less than |
| >= | GE | greater than or equal |
| <= | LE | less than or equal |
| in | IN | selecting multiple values |
| & | AND | and |
| | | OR | or |

# **P11: SELECT cases**

You can select a sub-group of data by many way.

## SELECT Statement:

**DATA** MyGitHub.LeadLow;

SET MyGitHub.LeadMerge;

SELECT (BloodLeadLevel);

WHEN (LE 1.99) LeadClass=**1**;

OTHERWISE DELETE;

END;

RUN;

By SELECT statement you can perform recode also.

## IF Statement:

**DATA** MyGitHub.LeadLow;

SET MyGitHub.LeadClassified;

IF LeadClass=1;

RUN;

# **P12: Calculate new variable**

In SAS, you can easily calculate new variables by using arithmetic operators or standard functions.

**DATA** MyGitHub.LeadCalculated;

SET MyGitHub.LeadMerge;

LeadChange=PreLead-PostLead;

Age=2022-DOB;

BMI=Weight/(Height/100)^2;

RUN;

# **S01 UNIVARIETE Procedure**

## Normality test with graphs (Histogram, QQ-Plot, PP-Plot)

**PROC** UNIVARIATE DATA= MyGitHub.LeadCalculated normal plot;

VAR BloodLeadLevel;

CLASS City;

QQPLOT BloodLeadLevel /NORMAL (MU=EST SIGMA=EST COLOR=RED L=**1**);

PPPLOT BloodLeadLevel /NORMAL (MU=EST SIGMA=EST COLOR=RED L=**1**);

HISTOGRAM /normal;

**RUN;**

## Testing Homogenity of Variance

**PROC** **UNIVARIATE** DATA = MyGitHub.LeadCalculated;

VAR BloodLeadLevel;

HISTOGRAM BloodLeadLevel / NORMAL;

**RUN**;

## One sample median test

**PROC** **UNIVARIATE** DATA= MyGitHub.LeadCalculated LOCCOUNT Mu0 = **2**;

VAR BloodLeadLevel;

RUN;

# **S02 MEANS Procedure**

PROC MEANS DATA= MyGitHub.LeadCalculated;

VAR BloodLeadLevel;

CLASS City;

RUN;

# **S03 GLM Procedure**

## Homogenity of variances

**PROC** **GLM** DATA= MyGitHub.LeadCalculated;

CLASS City;

MODEL BloodLeadLevel=City;

MEANS City;

MEANS Weight\_Status / HOVTEST=LEVENE;/\*To perform Levene’s homogeneity of variance test\*/

RUN;

Proc GLM can also be used to compare more than 2 groups.

## One Way ANOVA

PROC ANOVA only for balanced ANOVA designs applicable.

**PROC** **ANOVA** DATA= MyGitHub.LeadCalculated;

CLASS City;

MODEL BloodLeadLevel=City;

MEANS City /BON /\* Bonferroni t-tests\*/

LSD /\* Unprotected t-tests\*/

TUKEY /\* Tukey studentized range\*/

SCHEFFE; /\* Scheffe contrasts\*/

**RUN**;

## One-way ANOVA and Post-Hoc tests for ANOVA

PROC GLM can perform any ANOVA/Regression/ANCOVA design.

**PROC** **GLM** DATA= MyGitHub.LeadCalculated;

CLASS City;

MODEL BloodLeadLevel=City;

MEANS City /BON /\* Bonferroni t-tests\*/

LSD /\* Unprotected t-tests\*/

TUKEY /\* Tukey studentized range\*/

SCHEFFE; /\* Scheffe contrasts\*/

RUN;

## Repeated Measures ANOVA

**PROC** **GLM** DATA = MyGitHub.LeadCalculated;

CLASS City Gender;

MODEL BloodLeadLevel BloodLeadLevel\_1 BloodLeadLevel\_2= City Gender City\*Gender/nouni; /\*By this option we will analyze effect of City, Gender and interaction of this two simultaneously\*/

REPEATED Lead **3** /PRINTE; /\*REPEATED statement, and Lead is the user-defined name for the single within-subjects factor\*/

/\* /PRINTE option requests that SAS print out Mauchly’s test of sphericity\*/

LSMEANS City Gender City\*Gender; /\*The LSMEANS statement requests the cell means associated with the main effects for city and gender/\*

# RUN;

# **S04: FREQ Procedure**

## Frequency tables

**PROC** **FREQ** DATA = MyGitHub.LeadCalculated;

VAR BloodLeadLevel City;

**RUN**;

## Contingency tables

**PROC** **FREQ** DATA= MyGitHub.LeadCalculated;

TABLES City Gender/ NOCOL NOPERCENT CHISQ FISHER;

/\*NOCOL means no display column percentages\*/

/\*NOPERCENT means no display total percentages\*/

/\*CHISQ means calculate chi-square test\*/

/\*FISHER means calculate Fisher’s exact test\*/

**RUN**;

## Chi-square test and Fisher’s exact test

**PROC** **FREQ** DATA=MyGitHub.LeadCalculated;

TABLES LeadLevelGroupped\*Gender / chisq fisher;

**RUN;**

## One sample chi-square test

**PROC** **FREQ** DATA=MyGitHub.LeadCalculated;

TABLES LeadLevelGroupped\*Gender / chisq fisher;

**RUN**;

## Binomial test

**PROC** **FREQ** DATA=MyGitHub.LeadCalculated;

TABLES Cities / chisq testp=(**25** **25** **25** **25**);

**RUN**;

## McNemar test

**PROC** **FREQ** DATA=MyGitHub.LeadCalculated;

TABLES Gender / binomial(p=**.5**);

exact binomial;

**RUN;**

## Cochrane Q

**PROC** **FREQ** DATA=MyGitHub.LeadCalculated;

TABLE LeadGrouped1\* LeadGrouped1 /agree EXACT mcnem;

**RUN**;

# **S05: TABULATE Procedure**

**PROC** **TABULATE** data= MyGitHub.LeadLow;

VAR City LeadClass; TABLE City\*LeadClass \* (N Rowpctn);

**RUN**;

# **S06: TTEST Procedure**

## One sample t-test

**PROC** **TTEST** DATA= MyGitHub.LeadCalculated H0 = **2**;

VAR BloodLeadLevel;

**RUN;**

## Two independent samples t-test

**PROC** **TTEST** DATA= MyGitHub.LeadCalculated;

VAR BloodLeadLevel;

CLASS Gender;

**RUN;**

## Paired samples t-test

**PROC** TTESTDATA= MyGitHub.LeadCalculated;

PAIREDBloodLeadLevel**\*** Post\_BloodLeadLevel**;**

RUN**;**

# **S07: NPAR1WAY Procedure**

## Mann-Whitney U or Wilcoxon rank sum

**PROC** NPAR1WAY DATA= MyGitHub.LeadCalculated;

CLASS Gender;

VAR BloodLeadLevel;

**RUN;**

## Kruskal Wallis test

**PROC** NPAR1WAY DATA= MyGitHub.LeadCalculated;

CLASS City;

VAR BloodLeadLevel;

**RUN;**

# **S08: MULTTEST Procedure**

**DATA** MyGitHub.DataForAdjustment;

INPUT Test $ pval;

CARDS;

test1 .09108

test2 .69122

test3 .00177

test4 .57181

test5 .03121

test6 .01413

;

# **S08: CORR Procedure**

## Pearson, Spearman and Kendall Tau-b correlations

**PROC** **CORR** DATA= MyGitHub.LeadCalculated pearson spearman kendall;

VAR BloodLeadLevel;

WITH Age;

**RUN**;

## Chronbach alpha

**PROC** **CORR** DATA= MyGitHub.LeadCalculated nomiss alpha;

VAR Question1 Question2 Question3;

**RUN;**

## Partial correlation

**PROC** **CORR** DATA= MyGitHub.LeadCalculated;

VAR BloodLeadLevel Hgb;

PARTIAL Age;

**RUN;**

# **S09: LOGISTIC Procedure**

## Binary logistic regression

**PROC** **LOGISTIC** DATA=MyGitHub.LeadCalculated;

CLASS Gender / param=ref ;

MODEL LeadGrouped1=Age Hgb Gender;

**RUN;**

## Ordinal logistic regresion

**PROC** **LOGISTIC** DATA=MyGitHub.LeadCalculated;

CLASS Gender(ref='1') City(ref='Berlin') / param=reference;

MODEL Severity = Gender City Age;

**RUN;**

# **S10: REG Procedure**

## Linear regression

**PROC** **REG** DATA= MyGitHub.LeadCalculated;

MODEL BloodLeadLevel= Hgb Age;

**RUN;**

# **S11: PHREG Procedure**

## Cox regression

**PROC** **PHREG** data= MyGitHub.LeadCalculated;

MODEL Recovery\*censor(**0**)= BloodLeadLevel Age;

**RUN;**

# **S12: POWER Procedure**

## One sample mean comparison (ONESAMPLEMEAN)

|  |  |  |
| --- | --- | --- |
|  | **Sample size** | **Power** |
| **Assumptions** | You assume that the serum glucose level will be 100 for 1st group and 125 for 2nd group. The pooled standard deviation will be 25. How many subjects do you need to achieve a power 80%? | You found the serum glucose level of our patient group is 100±25 and that the differences below 125 are not significant. Our study group was 50.  Calculate the power. |
| **Code** | **PROC** **POWER**;  ONESAMPLEMEANS TEST=t  NULLMEAN = **125**  MEAN = **100**  STDDEV = **25**  POWER = **.8**  NTOTAL = **.** ;  **RUN**; | **PROC** **POWER**;  ONESAMPLEMEANS TEST=t  NULLMEAN = **125**  MEAN = **100**  STDDEV = **25**  POWER = **.**  NTOTAL = **50** ;  **RUN**; |
| **Result** | 10 | >0.999 |

## One sample mean estimation

Assume that, the expected population standard deviation to be 10, and employing t-distribution to estimate a mean with 95% confidence and a precision of 2.

Use <https://statulator.com/SampleSize/ss1M.html> site.

## One sample proportion estimation (ONESAMPLEFREQ)

Assume that 50% of the subjects in the population have the factor of interest, for estimating the expected proportion with 5% absolute precision and 95% confidence.

Use <https://statulator.com/SampleSize/ss1P.html> site.

## Two sample mean comparison (TWOSAMPLEMEANS)

|  |  |  |
| --- | --- | --- |
|  | **Sample size** | **Power** |
| **Assumptions** | You assume that the serum glucose level will be 100 for 1st group and 125 for 2nd group. The pooled standard deviation will be 25. How many subjects do you need to achieve a power 80%? | You found the serum glucose level of our patient group is 100±25 for 1st group and 100±25 for 2nd group. Our total sample size was 34 and calculated p value is 0.01  Calculate the power. |
| **Code** | **PROC** **POWER**;  TWOSAMPLEMEANS TEST=diff  GROUPMEANS = **100** | **125**  STDDEV = **25**  NPERGROUP = **.**  POWER = **0.8**;  **RUN**; | **PROC** **POWER**;  TWOSAMPLEMEANS TEST=diff  GROUPMEANS = **100** | **125**  STDDEV = **25**  NPERGROUP = **17.**  POWER = **.**;  **RUN**; |
| **Result** | 17 | 0.807 |

## One way ANOVA (ONEWAYANOVA)

|  |  |  |
| --- | --- | --- |
|  | **Sample size** | **Power** |
| **Assumptions** | You assume that the serum glucose level will be 100 for 1st group and 110 for 2nd group and 120 for 3rd group. The pooled standard deviation will be 20. How many subjects do you need to achieve a 5% alpha error and a power 80%? | You found the serum glucose level of our patient group is 100±20 for 1st group and 110±20 for 2nd and 120±20 for 3rd group. Each group have 21 subjects. calculated p value is 0.05.  Calculate power. |
| **Code** | **PROC** **POWER** ;  ONEWAYANOVA  GROUPMEANS = **100** | **110** | **120**  STDDEV = **20**  ALPHA = **0.05**  NPERGROUP = **.**  POWER = **.8**;  **RUN**; | **PROC** **POWER** ;  ONEWAYANOVA  GROUPMEANS = **100** | **110** | **120**  STDDEV = **20**  ALPHA = **0.05**  NPERGROUP = **21**  POWER = **.**;  **RUN**; |
| **Result** | 21 | 0.815 |

## Paired means (PAIREDMEANS)

|  |  |  |
| --- | --- | --- |
|  | **Sample size** | **Power** |
| **Assumptions** | You assume that the serum glucose level will be decreased 20±20 between before and after measurement. The correlation coefficient between before and after results will be 0.7. How many subjects do you need to achieve a power 80%? | You found the serum glucose level the serum glucose level will be decreased 20±20 between before and after measurement. The correlation coefficient was found 0.7 and p value of paired t test was 0.05. You have 20 paired measurement.  Calculate the power. |
| **Code** | **PROC** **POWER**;  PAIREDMEANS TEST=diff  MEANDIFF = **20**  STD = **20**  CORR = **.7**  NPAIRS = **.**  POWER = **0.8**;  **RUN**; | **PROC** **POWER**;  PAIREDMEANS TEST=diff  MEANDIFF = **20**  STD = **20**  CORR = **.7**  NPAIRS = **20**  POWER = **.**;  **RUN**; |
| **Result** | 7 | >0.999 |

## Two sample frequencies (TWOSAMPLEFREQ)

|  |  |  |
| --- | --- | --- |
|  | **Sample size** | **Power** |
| **Assumptions** | You assume that the hyperglycemia proportion will be 20% for 1st group and 10% for second group. Null proportion difference assume as 0%. How many subjects do you need to achieve a power 80%? | You found the the hyperglycemia proportion will be 20% for 1st group and 10% for second group. You have 200 subjects in each group.  Calculate the power. |
| **Code (Chi square test)** | **PROC** **POWER**;  TWOSAMPLEFREQ TEST=pchi  GROUPPROPORTIONS = (**.2** **.1**)  NULLPROPORTIONDIFF = **0**  POWER = **.80**  NPERGROUP =**.**;  **RUN**; | **PROC** **POWER**;  TWOSAMPLEFREQ TEST=pchi  GROUPPROPORTIONS = (**.2** **.1**)  NULLPROPORTIONDIFF = **0**  POWER = **.**  NPERGROUP =**200**;  **RUN**; |
|  |  |  |
| **Result** | 199 | 0.802 |
| **Code (Fisher’s exact test)** | **PROC** **POWER**;  TWOSAMPLEFREQ TEST=fisher  GROUPPROPORTIONS = (**.2** **.10**)  POWER = **.8**  NPERGROUP = **.** ;  **RUN**; | **PROC** **POWER**;  TWOSAMPLEFREQ TEST=fisher  GROUPPROPORTIONS = (**.2** **.10**)  POWER = **.**  NPERGROUP = **200** ;  **RUN**; |
|  |  |  |
| **Result** | 215 | 0.768 |

## McNemar Exact Conditional Test (PAIREDFREQ)

|  |  |  |
| --- | --- | --- |
|  | **Sample size** | **Power** |
| **Assumptions** | You are planning to compare hyperglycemia proportion before and after treatment in a study. The discordance proportions assume that 15% and 25 %. How many subjects do you need to achieve a power 80%? | You found the hyperglycemia discordant proportion 15% and 25 %. You have 200 subjects in each group.  Calculate the power. |
| **Code** | **PROC** **POWER**;  PAIREDFREQ dist=normal METHOD=connor  DISCPROPORTIONS = **0.15** | **0.25**  NPAIRS = **.**  POWER = **.8**;  **RUN**; | **PROC** **POWER**;  TWOSAMPLEFREQ TEST=pchi  GROUPPROPORTIONS = (**.2** **.1**)  NULLPROPORTIONDIFF = **0**  POWER = **.**  NPERGROUP =**200**;  **RUN**; |
|  |  |  |
| **Result** | 312 | 0.610 |

## Survival (TWOSAMPLESURVIVAL)

|  |  |  |
| --- | --- | --- |
|  | **Sample size** | **Power** |
| **Assumptions** | Assume that you want to compare survival rates for two treatment procedures. You intend to use a log-rank test to compare the overall survival curves for the two treatments. In 1st group proportion of survival decrease 10% and 2nd group 5%. How many subjects do you need per group to achieve a power of 0.8, with a two sided significance level of 0.05. | You compared survival rates two treatment procedures. You performed a log-rank test to compare the overall survival curves for the two treatments. In 1st group proportion of survival decrease 10% and 2nd group 5%. The calculated p value is 0.05. You have 80 subjects each group.  Calculate the power. |
| **Code** | **proc** **power**;  TWOSAMPLESURVIVAL TEST=logrank  CURVE("Existing Treatment") = **1** : **0.95** **2** : **0.90** **3**:**0.85** **4**:**0.80** **5**:**0.75**  CURVE("Proposed Treatment") = **1** : **0.90** **2** : **0.80** **3**:**0.70** **4**:**0.60** **5**:**0.50**  GROUPSURVIVAL = "Existing Treatment" | "Proposed Treatment"  ACCRUALTIME = **2**  FOLLOWUPTIME = **3**  POWER = **0.80**  ALPHA=**0.05**  NPERGROUP = **.** ;  **RUN**; | **PROC** **POWER**;  TWOSAMPLESURVIVAL TEST=logrank  CURVE("Existing Treatment") = **1** : **0.95** **2** : **0.90** **3**:**0.85** **4**:**0.80** **5**:**0.75**  CURVE("Proposed Treatment") = **1** : **0.90** **2** : **0.80** **3**:**0.70** **4**:**0.60** **5**:**0.50**  GROUPSURVIVAL = "Existing Treatment" | "Proposed Treatment"  ACCRUALTIME = **2**  FOLLOWUPTIME = **3**  POWER = **.**  ALPHA=**0.05**  NPERGROUP = **80** ;  **RUN**; |
|  |  |  |
| **Result** | 81 | 0.798 |

## Multipl regression (MULTREG)

|  |  |  |
| --- | --- | --- |
|  | **Sample size** | **Power** |
| **Assumptions** | Assume that, you prepare a study and intend to use a multiple linear regression analysis. You will evaluate effect of plasma cholesterol and glucose level on systolic blood pressure. You will have been 2 predictors and control one of them. Estimated R2 is 0.49 and your assumed R2differences is 0.09. How many subjects do you need to achieve power of 0.8. | You performed a study and use multiple linear regression to analyze results. You evaluated effect of plasma cholesterol and glucose level on systolic blood pressure. You have 2 predictors and control one of them. Calculated R2 is 0.49 and R2differences is 0.09. Your study group consist 100 patients.  Calculate the power for gender. |
| **Code** | **PROC** **POWER**;  MULTREG  MODEL = fixed  NFULLPREDICTORS = **5**  NTESTPREDICTORS = **1**  RSQUAREFULL = **0.49**  RSQUAREDIFF = **0.09**  NTOTAL = **.**  POWER = **0.8** ;  **RUN**; | **PROC** **POWER**;  MULTREG  MODEL = fixed  NFULLPREDICTORS = **2**  NTESTPREDICTORS = **1**  RSQUAREFULL = **0.49**  RSQUAREDIFF = **0.09**  NTOTAL = **100**  POWER = **.** ;  **RUN**; |
| **Result** | 47 | 0,986 |

## Logistic regression (LOGISTIC)

|  |  |  |
| --- | --- | --- |
|  | **Sample size** | **Power** |
| **Assumptions** | Assume that, you prepare a study and intend to use a logistic regression analysis. You will evaluate effect of gender (a binomial variable, proportion of males 0.5) and plasma glucose level (a normally distributed scale variable estimated mean is 100 and standard deviation is 17). hospitalization. You assume that Odd’s ratio will be 2 and response probability will be 0.65. How many subjects do you need total to estimate gender effect on hospitalization and to achieve a power of 0.8, with a two sided significance level of 0.05. | You performed a study and use logistic regression analysis. You evaluated effect of gender (a binomial variable, proportion of males 0.5) and plasma glucose level (a normally distributed scale variable estimated mean is 100 and standard deviation is 17). hospitalization. You calculated Odd’s ratio as 2 with a two sided significance level of 0.05. Your total sample was 200 subject and 130 of them participated to the study(response rate was be 0.65).  Calculate the power for gender. |
| **Code** | **PROC** **POWER**;  LOGISTIC  ALPHA = **0.05**  VARDIST("Glucose") = normal(**100**, **15**)  VARDIST ("Gender")= binomial (**0.5**, **1**)  TESTPREDICTOR = "Glucose" "Gender"  TESTODDSRATIO = **2**  RESPONSEPROB = **0.65**  NTOTAL = **.**  POWER = **.8** ;  **RUN**; | **PROC** **POWER**;  LOGISTIC  ALPHA = **0.05**  VARDIST("Glucose") = NORMAL(**100**, **15**)  VARDIST ("Gender")= binomial (**0.5**, **1**)  TESTPREDICTOR = "Glucose" "Gender"  TESTODDSRATIO = **2**  RESPONSEPROB = **0.65**  NTOTAL = **200**  POWER = **.**;  **RUN**; |
|  |  |  |
| **Result (Gender)** | 291 | 0.642 |
|  |  |  |
| **Result (Glucose)** | 7 | >0.999 |