From the textbook *The Statistical Sleuth*

Mean [20]: In your text the word "mean" denotes a population mean (μ) while the work "average" denotes a sample average ($\bar{Y}$).

Variance [20]: The variance is a measure of the dispersion, or spread, of the members of a population. The population variance is denoted σ^2 and its square root (s^2) is called the standard deviation.

Standard deviation [20] for a sample is calculated as $s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{(n-1)}}$, the square root of the variance.

Sample means have there own variance and standard deviation [33], which depends on the sample size. The standard deviation of the means is called the standard error and it is calculated as the sample standard deviation divided by the square root of n. $s_{\bar{Y}} = s/\sqrt{n}$.

Parameter (population summary number, denoted by Greek letters) versus Statistic (sample summary number). All hypotheses are stated in terms of population parameters.

Idea behind a statistical test

All statistical tests depend on some statistic following a known statistical distribution. The more common distributions used in statistical test are the Z, t, Chi square (χ^2) and F distribution.

Z distribution [34] : used to test a mean against an hypothesized value (H_0: $\mu = \mu_0$) or the difference between two means against an hypothesized value (H_0: $\mu_1 - \mu_2 = \delta$), where δ is often 0. Use of the Z distribution is appropriate when the variance (σ^2) is known or the sample size is very large. The Z distribution is called the standard normal distribution.

Chi square distribution [559] : used to test a variance (σ^2) against an hypothesized value (H_0: $\sigma^2 = \sigma_0^2$). The variance (σ^2) is estimated from a sample (S^2).

t distribution [34] : used to test a mean against an hypothesized value (H_0: $\mu = \mu_0$) or the difference between two means against an hypothesized value (H_0: $\mu_1 - \mu_2 = \delta$), where δ is often 0. Use of the t distribution is appropriate for any sample size and when the variance (σ^2) is unknown and estimated from the sample (S^2). The t-distribution is the ratio of a normal distribution ($\bar{Y}$ in the numerator) and a chi square distribution ($S_{\bar{Y}}$ in the denominator).

F distribution [125] : used to test two variance estimates (S^2_1 and S^2_2 estimating σ^2_1 and σ^2_2) for equality (H_0: $\sigma_1^2 = \sigma_2^2$).

For a test to be valid the distribution must hold under the null hypothesis. The distribution will not be the same if the null hypothesis is not true, but this is not important for the test of hypothesis.

For the t distribution [44] the test statistic is either $t = \frac{\bar{Y} - \mu_0}{S_{\bar{Y}}}$ or $t = \frac{\left(\bar{Y}_1 - \bar{Y}_2\right) - \delta}{S_{\bar{Y}_1 - \bar{Y}_2}}$. If the null hypothesis is true, then the t test statistic should be one of the normal bell shaped curves of the t distribution (there is one for each possible degrees of freedom). The t-test statistic is calculated and values that would be "unusual" under the null hypothesis would cause rejection of the null hypothesis. Commonly, t values that would occur only 5% of the time under the null hypothesis are considered unusual.

There are some assumptions that also must hold for this to be true. The assumptions are (1) the variable of interest (Y_i) is normally distributed and (2) that the error term is independent.

For the second case that tests between two means, $t = \frac{\left(\bar{Y}_1 - \bar{Y}_2\right) - \delta}{S_{\bar{Y}_1 - \bar{Y}_2}}$, if we calculate a combined variance from the two variances (one for each mean) then there is an additional assumption, (3) that the variances are the same.

Also for the second case the variance [39] is $S_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. However, the degrees of freedom for this variance are difficult to know unless the variances are equal and can be pooled. So, we first test the variances for equality using the F test of equal variances. If they are equal we pool the variances, weighting by the degrees of freedom. The calculation is $S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$.

Using the pooled variance the calculation for the standard error is $S_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s_1^2 \frac{1}{n_1} + \frac{1}{n_2}}$.

In this process we have made several assumptions. First, we have assumed that the data are normally distributed. This assumption is needed to use the t-test, which is based on a normal distribution. The second assumption is that the variances are homogeneous. If the two variances are not the same then we should not pool them into a single estimate.

Central limit theorem [33, 60]: states that if the summed variables have a finite variance then they will be approximately normally distributed. Many real processes do in fact have distributions with finite variance which results in the common use of the normal distribution.

P values [13, 46] are simply the probability of observing a given value, or larger value, of a statistic. For example, if we calculate a t value and it is 1.7, what is the probability that we get a value of 1.7 or larger for a t test? For a two-tailed t test, what is the probability that we get an absolute value of 1.7 or larger for a t test? To obtain this probability refer to tables of the t distribution.

Confidence intervals [45] are intervals that will include the true population parameter 100(1-α)% of the time. For normally distributed variables like means we use the t distribution to construct confidence intervals. Variances follow a Chi square distribution, so confidence intervals for variances are calculated differently.

For a mean, or a mean difference, the confidence interval is calculated as "the parameter estimate ± $t_{\alpha/2}$ * the standard error of the estimate."

For a mean the parameter estimate is $\bar{Y}$ and the standard error is $s/\sqrt{n}$. The confidence interval is given by $\bar{Y} \pm t_{\alpha/2} s_{\bar{Y}}$ and is usually expressed as $P\left(\bar{Y} - t_{\alpha/2} s_{\bar{Y}} \le \mu \le \bar{Y} + t_{\alpha/2} s_{\bar{Y}}\right) = 1 - \alpha$.

For the difference between two means the parameter estimate is $\bar{Y}_1 - \bar{Y}_2$ and the standard error is $s_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$, where s_p^2 is the pooled variance estimate. The confidence interval is given by $(\bar{Y}_1 - \bar{Y}_2) \pm t_{\alpha/2} s_{\bar{Y}_1 - \bar{Y}_2}$ and is expressed as $P\left((\bar{Y}_1 - \bar{Y}_2) - t_{\alpha/2} s_{\bar{Y}_1 - \bar{Y}_2} \le \mu \le (\bar{Y}_1 - \bar{Y}_2) + t_{\alpha/2} s_{\bar{Y}_1 - \bar{Y}_2}\right) = 1 - \alpha$.

Similar calculations are used to calculate confidence intervals for treatment means [45] in Analysis of Variance, difference between treatment means [45] in ANOVA and for confidence intervals on regression coefficients (intercepts and slopes) for linear regression [186]. All of these are usually normally distributed and the confidence interval calculated as "parameter estimate ± $t_{\alpha/2}$ * standard error of the estimate."

An Introduction to SAS® Programming

SAS programs consists of two major type of steps

- The DATA step – used to create or modify a SAS dataset – [Contents > SAS Products > Base SAS > SAS Language concepts > Data Step Concepts]
  - SAS dataset – a file containing a collection of similar information
  - Can be visualized as a two-dimensional array (table) that looks like spreadsheet (e.g. like EXCEL)
  - Observation – each row represents the information for a single item
  - Variable – each column contains one type of item
  - In addition to data values, variable names and types, lengths, labels and formats are stored in this file
  - Source of data for the data step
    - Convert a raw data file to a SAS data set. This can be either stored in a separate file or included in the SAS program itself
    - The SAS datasets are initially created from some source of data. The SAS dataset, once created, can be stored as a "permanent" file or recreated each time SAS is run on that dataset.
    - When the SAS dataset is created we can assign formats and labels
    - When the SAS dataset is created we can perform modifications to the data, transformations and calculations
- The PROC step – (PROC is from PROCedure) used to process SAS data sets
  - Allows File manipulation - SORTS
  - Report preparation – PRINT and tabulation procedures
  - Analysis – MEANS, FREQ, various statistical analyses
  - Graphics – CHART, PLOT, TIMEPLOT
  - Utilities – CONTENTS, DATASETS, FORMAT, file import and compare utilities

SAS Display Manager (in SAS help see "Using SAS Software in Your Operating Environment > Using SAS in Windows > Running programs in the SAS Windowing environment"). Here there are descriptions of the SAS interface, including graphics of SAS windows and menu options.

- Program editor window (= Editor) - type, edit, save and submit SAS programs
- Log window (= Log) – Displays the SAS log (notes and messages produced when programs run)
- Output window (= Output) – Displays output from SAS program runs

General rules about SAS programs

SAS statements can begin on any line.

SAS statements can be continued on another line as long as no word is split.

More than one statement can be written on a single line.

At least one blank must separate each word or item in a SAS statement, except for mathematical operators (e.g. +, -, *, /, =).

Each SAS statement must end with a semicolon, ";".

In PC SAS it is a good practice to end each DATA step or PROC step with a "RUN;" statement.

Statements surrounded by a pair of "/*" to start and "*/" to end can be included anywhere in a SAS statement that a blank would appear. The enclosed section is a comment. This can also be used to turn any segment of a SAS program into an inactive comment section.

The order of many statements is not important. This is true in both DATA steps and PROC steps. There are some logical exceptions. For example, you cannot process data until after an "INPUT" statement.

Introduction to the SAS DATA step – creating a SAS data set from raw data.

The DATA step starts with the word DATA followed by a data-set-name.

There are some limits on what can be used as the data-set-name. The SAS help (9.1.3) states that "A SAS name can be up to eight characters long. The first character must be a letter (A,B,C,...,Z) or underscore (_). Subsequent characters can be letters, numbers (0 to 9), or underscores. Note that no blanks are allowed. Two names (_N_ and _ERROR_) are reserved by the SAS System."

Names longer than eight characters are acceptable in SAS 9.1.3, up to 32 characters.

Lengths of variables and names in SAS

Maximum Length of SAS Names	
SAS Application	**Max Length**
Arrays	32
CALL routines	16
Catalog entries	32
DATA step statement labels	32
DATA step variable labels	256
DATA step variables	32
DATA step windows	32
Engines	8
Filerefs	8
Formats, character	31
Formats, numeric	32
Functions	16
Generation data sets	28
Informats, character	30
Informats, numeric	31
Librefs	8
Macro variables	32
Macro windows	32
Macros	32
Members of SAS data libraries (SAS data sets, views, catalogs, indexes) except for generation data sets	32
Passwords	8
Procedure names (first 8 characters must be unique, and may not begin with "SAS")	16
SCL variables	32