Part II: Biostatistics

1. Statistical variable

The unit that is the basis of the study is called the statistical unit (= smallest sampling unit, e.g. animal, herd, region, . . .). A property of the given statistical unit that is the subject of the study is called a (statistical) variable. Variables are things that we measure, control, or manipulate in research. They differ in many respects, most notably in the role they are given in our research and in the type of measures that can be applied to them. It is an attribute that describes a person, place, thing, or idea. The value of the variable can "vary" from one entity to another. The following classification of variables is given solely for sake of completeness, because the distinction between the different categories is in reality not so unambiguous. Also, whether or not continuous variables exist in real life is more a matter of philosophy than hard science.

1.1. Quantitative vs. Qualitative Variables

1.1.1. Quantitative variables (measurement variables)

A quantitative variable is a variable that can be expressed numerically. Its 'level' can both be measured and expressed by a quantity, which we call the value of the variable. The different states of the variable can be expressed in a numerically ordered fashion. Quantitative variables can be further classified as **discrete** or **continuous**. If a variable can take on any value between its minimum value and its maximum value, it is called a continuous variable; otherwise, it is called a discrete/discontinuous variable.

1.1.1.1. Continuous variables

Continuous variables can (at least theoretically) take an infinite number of values between any two fixed points: e.g. there are (in theory) an infinite number of possible milk yields between 800 kg and 801 kg. In practice, the precision of the measurement device defines the number of possible values and thus the extent of 'continuousness' of the variable. Despite this limitation, a number of variables in animal science and health are considered continuous (e.g. volumes, weights, temperatures, time etc).

1.1.1.2. Discontinuous variables (discrete variables)

These are variables that can take only a limited number of fixed values, with no intermediate values possible. E.g. the number of ticks on an animal is limited to whole numbers. The number of ticks on an animal could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.3 ticks.
Note: quantitative variables can also be classified as interval and ratio variable based on their measurement scale. The meanings of interval and ratio variables are given below.

1.1.2. Qualitative variables

Qualitative variables take on values that are names or labels. Qualitative variables do not have the continuous nature that quantitative variables have. For example, gender is a qualitative variable. The two measurements on this variable, female and male are different in a qualitative way. Using a 0 and a 1 instead of names doesn’t change this fact. Another example of a qualitative variable is disease status (diseased or disease free using 0 for not disease and 1 for diseased instead of names doesn’t change this fact). Qualitative variables could be nominal or ordinal. Unlike ordinal variables nominal variables, assuming no ranking or ordering. In the nominal scale, numbers are used simply as names and have no real quantitative value. When the nominal variable has only two states (e.g. diseased or disease-free, female or male, dead or alive) it is referred to as a binomial variable. When it can assume more than two states it is called a multinomial variable. Numerals on sports uniforms are an example of nominal variable. Thus, 9 is *different* from 7, but that is all you can say. The person represented by 9 is not “more than” the person represented by 7, and certainly it would be meaningless to calculate a mean from the two scores. On a nominal scale, the numbers mean only that the categories are different. In fact, for a nominal scale variable, the numbers could be assigned to categories at random.

Some qualitative variables have the characteristic of *order:* e.g. resistance to a disease is expressed as low, medium and high and they are called ordinal variables. These three states can be recorded as 1, 2 and 3. However, this does not imply that medium animals are twice as resistant as low animals or that the difference (or ratio) between high and medium animals is the same as between medium and low animals. Only an order or rank is assumed. However, when these variables are used in risk analysis or risk factor analysis contexts, special care must be paid to be sure that ordering on its own suffices to do the analysis or whether limiting values between different states are required to allow comparison.

In the ordinal scale, the object with the number 3 has less or more of something than the object with the number 5. Finish places in a race are an example of an ordinal scale. The runners finish in rank order, with 1 assigned to the winner, 2 to the runner-up, and so on. Here, 1 means less time than 2.

1.2. Explanatory vs response variables

The purpose of statistical analyses is to quantify the probability that 2 variables, namely the response and the explanatory variables, are associated. The response or dependant variable is the variable that is considered to vary in function of variations of the explanatory variables. Explanatory variables also referred to as independent variables or predictors are the variables that are expected to have an effect on the response, whatever the effect of other variables. Explanatory variables can either be continuous or discrete. Classically, discrete variables with more than 2 categories (*n* categories) are transformed in (*n*-1) *dummy* binary variables (0 or 1).

Explanatory variable: variable controlled by the researcher; changes in this variable may produce changes in the dependent variable. It is also called predictor, independent, covariate, factor variable

Response variable: the observed variable that is expected to change as a result of changes in the independent variable in an experiment. It is also called outcome, dependent,

A key distinction between dependent/independent is whether a variable is being predicted by the remaining variables, or whether it is being used to make the prediction. The variable singled out to be predicted from the remaining variables is called dependent variable and variables used to make the prediction are called independent variables.

2. Hypothesis testing

A **statistical hypothesis** is an assumption about a population parameter. This assumption may or may not be true. **Hypothesis testing** refers to the formal procedures used by statisticians to accept or reject statistical hypotheses. In statistics, a result is called statistically significant if it has been predicted as unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the significance level. These tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance.

There are two types of statistical hypotheses.
Null hypothesis: the null hypothesis, denoted by H_0, is usually the hypothesis that sample observations result purely from chance. Ho, is theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug. We would write

Ho: there is no difference between the two drugs on average.

We give special consideration to the null hypothesis. This is due to the fact that the null hypothesis relates to the statement being tested, whereas the alternative hypothesis relates to the statement to be accepted if / when the null is rejected. The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either "Reject H_0 in favour of H_1" or "Do not reject H_0"; we never conclude "Reject H_1", or even "Accept H_1". If we conclude "Do not reject H_0", this does not necessarily mean that the null hypothesis is true; it only suggests that there is not sufficient evidence against H_0 in favour of H_1. Rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.

Alternative hypothesis: the alternative hypothesis, denoted by H_1 or H_a, is the hypothesis that sample observations are influenced by some non-random cause. For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect, on average, compared to that of the current drug. We would write

H_1: the two drugs have different effects, on average.

The alternative hypothesis might also be that the new drug is better, on average, than the current drug. In this case we would write

H_1: the new drug is better than the current drug, on average.

Hypothesis testing consists of four steps.

- ⇒ State the hypotheses: this involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.
- ⇒ Formulate an analysis plan: the analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.
- ⇒ Analyse sample data: find the value of the test statistic (mean score, proportion, t-score, z-score, etc.) described in the analysis plan.
- ⇒ Interpret results: apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.

The analysis plan includes decision rules for rejecting the null hypothesis. The decision rules could be described in two ways - with reference to a P-value or with reference to a region of acceptance.

- ⇒ P-value: the strength of evidence in support of a null hypothesis is measured by the **P-value**. Suppose the test statistic is equal to *S*. The P-value is the probability of observing a test statistic as extreme as *S*, assuming the null hypothesis is true. If the P-value is less than the significance level, we reject the null hypothesis.
- ⇒ Region of acceptance: the **region of acceptance** is a range of values for a test statistics for which the null hypothesis is not rejected. The region of acceptance is defined so that the chance of making a Type I error is equal to the significance level. The set of values outside the region of acceptance is called the **region of rejection**. If the test statistic falls within the region of rejection, the null hypothesis is rejected. In such cases, we say that the hypothesis has been rejected at α level of significance.

These approaches are equivalent. Some statistics texts use the P-value approach; others use the region of acceptance approach.

Two types of errors can result from a hypothesis test.

Type I error: type I error is made when the null hypothesis is incorrectly rejected, or when we conclude that the outcomes are different when in fact they are not. *The risk of falsely concluding that the null hypothesis is incorrect is the risk of committing a type I error*. This error is traditionally represented by α or by p (a value of 5% (0.05) is accepted as the maximum risk to incorrectly reject the null hypothesis).

Type II error: type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called Beta, and is often denoted by β. The probability of not committing a Type II error is called the *Power of the test.* A type II error can arise in many situations but the most important cause is a sample size that is too low. Beta depends upon sample size and it gets smaller as the sample size gets larger.

Summary of statistical errors

Table 4 shows the relationship between the reality, the hypotheses, and the conclusions of the statistical test and the possible errors.

Decision made on the basis of sample data		True situation in the population	
		H_0 correct	H_0 incorrect
	Reject H_0	1. Type I error(α)	3. Correct decision (1- β)
	Fail to reject (Retain) H_0	2. Correct decision (1- α)	4. Type II error (β)

3. Data analysis and interpretation

3.1. Basic statistical methods

The basic principle of statistical testing is to compare two or more averages, taking into account the variances and the numbers of observations of the various groups. Statistical testing usually makes a distinction between independent/explanatory/predictor and dependent/response/outcome variables. Independent variables are those variables that distinguish the different groups (characteristics of individuals or experimental design). Dependent variables are usually the variables of interest in the observational study or in experiment (production, mortality, parasite burden, etc).

Animal science and animal health studies often deal with:

- Continuous data (e.g.: milk production)
- Binary data (e.g.: diseased or not)
- Count data (e.g.: tick load on animals)
- Survival data (e.g.: disease-free survival time)

In research we often wishes to find out if there are differences between some groups i.e. prove a hypothesis. To test our hypothesis we perform various tests like the t-test, Chi-Square test, Analysis of Variance (ANOVA) and Regression analysis to mention a few. In the sections below we will discuss the statistical tests which are often used in veterinary and animal sciences.

3.1.1. Analysis of continuous data

3.1.1.1. The t-test

3.1.1.1.1. Independent sample t-test

The independent sample t-test assesses whether the means of two groups are statistically different from each other. This analysis is appropriate whenever you want to compare the means of two groups/independent sample. It is methods for assessing the association of a categorical predictor (binary predictor) variable with a continuous outcome/response variable. While using t-test we assume that the population from which sample has been taken is normal or approximately normal, sample is a random sample, observations are independent.

Statistical Analysis of the t-test

The formula for the t-test is a ratio. The top part of the ratio is just the difference between the two means or averages. The bottom part is a measure of the variability or dispersion of the scores. The top part of the formula is easy to compute -- just find the difference between the means. The bottom part is called the **standard error of the difference**. To compute it, we take the variance for each group and divide it by the number of sample in that group. We add these two values and then take their square root. The specific formula is given below

$$SE(\overline{X_T} - \overline{X_C}) = \sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}$$

Remember, that the variance is simply the square of the standard deviation.

The final formula for the t-test therefore becomes,

$$t = \frac{\overline{X_T} - \overline{X_C}}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$

The t-value will be positive if the first mean is larger than the second and negative if it is smaller. Once we compute the t-value we have to look it up in a table of significance to test whether the ratio is large enough to say that the difference between the groups is not likely to have been a chance finding. To test the significance, we need to set a risk level (called the alpha level). The alpha level is often set at 0.05. This means that five times out of a hundred we would find a statistically significant difference between the means even if there was none (i.e., by "chance").

We also need to determine the degrees of freedom (df) for the test. In the t-test, the degree of freedom is the sum of the animals or individuals in both groups minus 2. Given the alpha level, the df, and the t-value, you can look the t-value up in a standard table of significance to determine whether the t-value is large enough to be significant. If the calculated value of '*t*' is either equal to or exceeds the table value, we infer that the difference is significant, but if calculated value of *t* is less than the concerning table value of *t*, the difference is not treated as significant.

Example 1

An experiment was carried out to verify whether or not a dietary supplement had an effect on the milk production of dairy animals. The results obtained during the course of this experiment are shown in Table 1. Do these results suffice to allow us to conclude that the supplement increased the milk production? The null hypothesis is: the milk production obtained when fed the supplemented ration is equal to the milk production obtained when fed the basic ration.

Animal id	*Non-supplemented*	*Milk yield*	*Animal id*	*Supplemented*	*Milk yield*
1	*basic*	*735*	*31*	*Supplemented*	*847*
2	*basic*	*674*	*32*	*Supplemented*	*406*
3	*basic*	*844*	*33*	*Supplemented*	*1046*
4	*basic*	*768*	*34*	*Supplemented*	*1424*
5	*basic*	*750*	*35*	*Supplemented*	*883*
6	*basic*	*851*	*36*	*Supplemented*	*751*
7	*basic*	*960*	*37*	*Supplemented*	*925*
8	*basic*	*793*	*38*	*Supplemented*	*1130*
9	*basic*	*764*	*39*	*Supplemented*	*755*
10	*basic*	*765*	*40*	*Supplemented*	*903*
11	*basic*	*887*	*41*	*Supplemented*	*964*
12	*basic*	*1009*	*42*	*Supplemented*	*919*
13	*basic*	*877*	*43*	*Supplemented*	*1436*
14	*basic*	*878*	*44*	*Supplemented*	*947*
15	*basic*	*741*	*45*	*Supplemented*	*865*
16	*basic*	*712*	*46*	*Supplemented*	*1040*
17	*basic*	*763*	*47*	*Supplemented*	*1255*
18	*basic*	*1174*	*48*	*Supplemented*	*1216*
19	*basic*	*651*	*49*	*Supplemented*	*1047*
20	*basic*	*786*	*50*	*Supplemented*	*929*

21	*basic*	*760*	*51*	*Supplemented*	*927*
22	*basic*	*1043*	*52*	*Supplemented*	*1232*
23	*basic*	*708*	*53*	*Supplemented*	*913*
24	*basic*	*726*	*54*	*Supplemented*	*1139*
25	*basic*	*923*	*55*	*Supplemented*	*1120*
26	*basic*	*825*	*56*	*Supplemented*	*1072*
27	*basic*	*843*	*57*	*Supplemented*	*954*
28	*basic*	*586*	*58*	*Supplemented*	*700*
29	*basic*	*875*	*59*	*Supplemented*	*959*
30	*basic*	*670*	*60*	*Supplemented*	*1125*

Here our independent variable is the ration type (basic and supplement) and it has two levels and the response variable is milk yield and it is a continuous variable. T-test is appropriate to test this data. So, let’s use the independent sample t-test to test our hypothesis of no difference.

In STATA t-test is computed using a command “ttest”

```
. ttest milkyield, by(ration)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Suppleme |      30       994.3    38.34775    210.0393      915.87     1072.73
   basic |      30    811.3667     22.7707    124.7202    764.7954     857.938
---------+--------------------------------------------------------------------
combined |      60    902.8333    25.11244    194.5201    852.5835    953.0832
---------+--------------------------------------------------------------------
    diff |            182.9333    44.59882                 93.6591    272.2076
------------------------------------------------------------------------------
    diff = mean(Suppleme) - mean(basic)                           t =   4.1018
Ho: diff = 0                                     degrees of freedom =       58

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9999         Pr(|T| > |t|) = 0.0001          Pr(T > t) = 0.0001
```

Now let’s use the STATA output in example to discuss the t – value.

We obtained a t value: t = 4.1018 and Pr(|T| > |t|) =0.0001

Interpretation

This t value is significant at the 0.0001 level, which means that we would get a value of t as high as 4.1018 or higher only one time out of a thousand by chance if the null hypothesis is true. So we reject the null hypothesis of no difference, accept the alternate hypothesis, and conclude that the animals who fed on supplement produced more milk than those who fed on basic ration. In other words the difference between the milk production when fed the supplemented ration and the milk production when fed the basic ration is greater than zero (in this case it is 182.93).

3.1.1.1.2. *Paired sample t-test*

A paired t-test is used to compare two population means where you have two samples in which observations in one sample can be paired with observations in the other sample. Examples include measurements taken at two time points on the same individuals, or on other naturally linked pairs, as in a clinical trial where one eye is treated and the other serves as a control.

Another example could be a comparison of two different methods of measurement or two different treatments where the measurements/treatments are applied to the same subjects (e.g. blood pressure measurements using a stethoscope and a dynamap). In these cases, the two samples are not independent and failure to take account of the pairwise relationships wastes information and is potentially erroneous.

The paired t-test procedure first computes the pairwise differences for each individual or linked pair. In the first example, this is the change in the outcome from the first time point to the second, and in the second, the difference between the outcomes for the treated and control eyes. Then a t-test is used to assess whether the population mean of these paired differences differs from zero. An increase in power results because between-individual variability is eliminated in the first step. The paired t-test is also implemented using the ttest command in Stata.

Formula:

$$t = \frac{\bar{x} - \Delta}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ is the mean of the change scores, Δ is the hypothesized difference (0 if testing for equal means), *s* is the sample standard deviation of the differences, and *n* is the sample size. The number of degrees of freedom for the problem is $n – 1$.

Assumption of paired t test

The difference b/t the observation in each pair and the set of differences for all pairs is approximately normally distributed even though the original observation in the groups may not be.

Example 2

Nelson et al 1998 conducted a randomized cross- over trial of two diets in 11 insulin dependent diabetic dogs; they measured serum glucose as the variable indicating the quality of diabetic control. The diets contained either low insoluble fiber (LF) or high insoluble fiber (HF). Each dog was randomly allocated to receive a particular diet first. The dogs were adapted to the diet for two months and then fed on it for six months: evaluation was performed at six week intervals.

As the study ran over 16 months of each dog's life, we might expect the animal to change in its metabolic responses to diabetes during the course of the trial, irrespective of diet. This would reduce the value of a cross-over deign since there might be considerable variability in the within dog comparisons even without a change in diet. However, as the order in which the dogs received the diets was determined randomly, the result should not be biased.

The table below has been developed from the authors' summary results and gives the mean morning pre-prandial serum glucose concentrations (mmol/l) for each dog in each six month period.

Dog	low insoluble fiber (LF)	high insoluble fiber (HF)
1	9.44	9.28
2	17.61	8.67
3	8.89	6.28
4	16.94	12.67
5	10.39	6.67
6	11.78	7.28
7	15.06	15.39
8	7.06	5.61
9	19.56	11.94
10	8.22	5.11
11	23.17	17.33

The null hypothesis states that the true mean difference in the pre-prandial serum glucose levels between the low fiber and high fiber diets is zero; the two sided alternative hypothesis is that it is not zero.

```
. ttest lf == hf

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      lf |      11    13.46545     1.59857    5.301857    9.903619    17.02729
      hf |      11    9.657273    1.243816    4.125271    6.885878    12.42867
---------+--------------------------------------------------------------------
    diff |      11    3.808182    .8718391    2.891563    1.865603     5.75076
------------------------------------------------------------------------------
     mean(diff) = mean(lf - hf)                                   t =   4.3680
 Ho: mean(diff) = 0                              degrees of freedom =       10

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9993         Pr(|T| > |t|) = 0.0014          Pr(T > t) = 0.0007
```

Interpretation

The p–value 0.0014 is (two tailed); so the data is not consistent with the null hypothesis which we therefore reject. The mean of the pre-prandial serum glucose difference (LF-HF), estimated as 3.81 mmol/l is significantly different from zero, indicating that the high fiber diet significantly reduces fasting blood sugar.

3.1.1.1.3. Single sample t-test

The single sample t method tests a null hypothesis that the population mean is equal to a specified value.

Formula:

$$t = \frac{\bar{x} - \Delta}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ is the sample mean, Δ is a specified value to be tested, *s* is the sample standard deviation, and *n* is the size of the sample. Degrees of freedom for is *n* – 1

Assumption

The sample data is from normally distributed population of values and are representative of that population (random selection)

Example 3

The following table shows the daily live weight gain of a random sample of 36 growing pigs in a rearing unit. The rearing unit expects a mean daily weight gain of 607g for this stage of growth (weaning to 10 weeks of age) based on current performance indicators. Are these values consistent with a mean daily gain of around 607g?

Daily weight gain	
577	621
596	623
594	598
612	602
600	581
584	631
618	570
627	595
588	603
601	605
606	616
559	574
615	578
607	600
608	596
591	619
565	636
586	589

Ho is that the mean daily live weight gain is 607g. The alternative hypothesis is that it is not.

```
. ttest dailyweightgain == 607

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
dailyw~n |      36    599.1944    3.109336    18.65601    592.8822    605.5067
------------------------------------------------------------------------------
    mean = mean(dailyweightgain)                                  t =  -2.5104
Ho: mean = 607                                   degrees of freedom =       35

   Ha: mean < 607               Ha: mean != 607                 Ha: mean > 607
 Pr(T < t) = 0.0084         Pr(|T| > |t|) = 0.0168          Pr(T > t) = 0.9916
```

Interpretation: the p –value is 0.0168 (two tailed); so the data is inconsistent with the mean daily mean gain of 607g. So we reject H_o. The test revealed that the pigs have significantly poorer mean daily weight gain than expected. The confidence interval indicates that the true mean weight gain may even be as low as 593g per day.

3.1.1.2. Analysis of variance (ANOVA)

As stated earlier, the significance of the difference between the means of two samples can be judged through *t*-test, but the difficulty arises when we happen to examine the significance of the difference amongst more than two sample means at the same time. So, ANOVA can be seen as methods for assessing the association of a categorical predictor (with two or more than two levels) with a continuous outcome. Using this technique, one can draw inferences about whether the samples have been drawn from populations having the same mean.

Note: Variance is an important statistical measure and is described as the mean of the squares of deviations taken from the mean of the given series of data. It is a frequently used measure of variation. Its square root is known as standard deviation.

If we take only one factor and investigate the differences amongst its various categories having numerous possible values, we are said to use one-way ANOVA and in case we investigate two factors at the same time, then we use two-way ANOVA. In a two or more way ANOVA, the interaction (i.e., inter-relation between two independent variables/factors), if any, between two independent variables affecting a dependent variable can as well be studied for better decisions.

The Basic Principle of ANOVA

The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples. In terms of variation within the given population, it is assumed that the values of (X_{ij}) differ from the mean of this population only because of random effects i.e., there are influences on (X_{ij}) which are unexplainable, whereas in examining differences between populations we assume that the difference between the mean of the j^{th} population and the grand mean is attributable to what is called a 'specific factor' or what is technically described as treatment effect. Thus while using ANOVA; we assume that each of the samples is drawn from a normal population and that each of these populations has the same variance. We also assume that all factors other than the one or more being tested are effectively controlled. This, in other words, means that we assume the absence of many factors that might affect our conclusions concerning the factor(s) to be studied.

In short, we have to make two estimates of population variance: one based on between samples/groups variance and the other based on within samples/groups variance.

3.1.1.2.1. One-Way Analysis of Variance (ANOVA)

When we need to compare sample averages of trial with multiple treatments, or more generally across more than two independent samples; for this purpose, one-way analysis of variance (ANOVA) and the *F*-test take the place of the *t*-test.

Computational Formulas for One-Way ANOVA

Like any other test, the ANOVA test has its own test statistic. The statistic for ANOVA is called the F statistic, which we get from the F Test. The ANOVA F-statistic is a ratio of the Between Group Variaton divided by the Within Group Variation:

It has six steps

1. Calculate the total sum of squares (SST)

Total sum of squares is sum of squares of all observations

$$= \sum(obs\ - grand\ mean)^2 = \quad SST = \Sigma(Xi - \overline{X})^2$$

2. Calculate the *b/n groups'* sum of squares *SSB*

Between groups Sum of Squares (or Treatment Sum of Squares) – variation in the data between the different samples (or treatments).

$$=\sum(mean\ of\ groups\ - grand\ mean)2 * n\ of\ the\ group = \quad SSB = \Sigma\left(\overline{X}_k - \overline{X}\right)^2 * nk$$

3. SS Within groups

Within variation (or Error Sum of Squares) – variation in the data from each individual treatment.

$$SSW = \Sigma\left(X_i - \overline{X}_k\right)^2$$

$$= \sum(obs\ - mean\ of\ the\ group)^2 =$$

Total sum of squares = *SS B/n groups* + *SS Within groups*

4. Calculate the degrees of freedom

- ✓ dfb = k-1 and
- ✓ dfw = n – k

Where n is the total sample size and k is the number of groups

5. Construct the mean square estimates by dividing SSB and SSW by their degrees of freedom:

$$MS_w = SSW / dfw$$
$$MS_b = SSB / dfb$$

6. Find *F* ratio by Formula:

$$F = MS_b / MS_w$$

Example 4

A study was conducted to evaluate the efficacy of trypanocidal drugs. The efficacy was assessed using a change in PCV value. Two types of trypanocidal drugs were used. Fifteen tyrpanosoma infected animals were randomly assigned to one of the three treatments (five animals for each). The PCV value of the animals after the experiment was presented below. Did the use of the trypanocidal drugs improve the PCV value?

Treatments	*PCV*
Control	*22*
Control	*19*
Control	*15*
Control	*24*
Control	*18*
Diminazene A.	*26*
Diminazene A.	*35*
Diminazene A.	*40*
Diminazene A.	*23*
Diminazene A.	*30*
Trypamidium–Samorin	*25*
Trypamidium–Samorin	*44*
Trypamidium–Samorin	*31*
Trypamidium–Samorin	*21*
Trypamidium–Samorin	*27*

```
. oneway pcv trypanocidaldrug

                        Analysis of Variance
    Source              SS          df       MS            F     Prob > F
------------------------------------------------------------------------

Between groups      378.133333       2   189.066667      4.15     0.0428
 Within groups           547.2      12         45.6
------------------------------------------------------------------------
```

The F-test assesses the null hypothesis that the mean value of the outcome (PCV in our case) is the same across all the populations sampled from, against the alternative that the means differ in at least two of the populations. For example, the one-way ANOVA shown in analysis above the F-test for between groups (P = 0.0428), suggests that mean PCV differs between the treatment groups. This indicates that the trypanocidal drugs significantly influenced the PCV value.

3.1.1.2.2. Two-way ANOVA

The example above is referred to as the one-way ANOVA because you can divide all the scores in one way only, by the type of the trypanosomal drug to which animals were assigned. The trypanosomal drug group is called a “factor” and this factor has three levels, meaning there are three application (control, *Diminazene A.* and *Trypamidium–Samorin*) categories. There may, however, be another factor that classifies individuals, and in that case we would have a two-way, or a two-factor, ANOVA. The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors). The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.

EXAMPLE 5

In the experiment we used as an example one way ANOVA, animals were assigned to one of the three applications noted above, as well as to one of the two feeding management as indicated in table below. Now we are interested in comparing the three means that represent change in trypanosomal drug application and the two means that represent their feeding management.

Trypanocidal drug	feeding managment	PCV
Control	good	22
Control	poor	19
Control	poor	15
Control	good	24
Control	poor	18
Diminazene A	poor	26
Diminazene A	good	35
Diminazene A	good	40
Diminazene A	poor	23
Diminazene A	poor	30
Trypamidium–Samorin	poor	25
Trypamidium–Samorin	good	44
Trypamidium–Samorin	good	31
Trypamidium–Samorin	poor	21
Trypamidium–Samorin	poor	27

```
. anova pcv trypanocidaldrug##feedingmanagment

                           Number of obs =         15    R-squared     =  0.8368
                           Root MSE      =    4.09607    Adj R-squared =  0.7462

                  Source | Partial SS    df       MS           F     Prob > F
      -------------------+----------------------------------------------------
                   Model |  774.333333     5  154.866667       9.23     0.0024
                         |
             trypanoci~g |       407.4     2       203.7      12.14     0.0028
             feedingma~t |         360     1         360      21.46     0.0012
 trypanoci~g#feedingma~t |        36.2     2        18.1       1.08     0.3802
                         |
                Residual |         151     9  16.7777778
      -------------------+----------------------------------------------------
                   Total |  925.333333    14  66.0952381
```

Note: string variables may not be used as factor variables in two way ANOVA, so we should code it with numerical value.

Converting string variables with numeric values into numeric values
Gen pcv_n = real(pcv)   or
destring, replace
Converting string variables with non-numeric values into numeric values
encode trypanocidaldrug, generate(trypanocidaldrug2)

Interpretation: Here the p-value for the interaction term (0.38) is greater than 0.05 so it indicates that there is no interaction between trypanosomal drug and feeding management.

N.B. The *t*-test and ANOVA are basic tools for assessing the statistical significance of differences between the average values of a continuous outcome across two or more samples. They can be seen as methods for assessing the association of a categorical predictor – binary in the case of the *t*-test, with more than two levels in the case of ANOVA – with a continuous outcome. Both are based in statistical theory for normally distributed outcomes, but work well for many other types of data; and both turn out to be special cases of linear regression models.

3.1.2. Linear regression

Linear regression determines how the average value of the continuous response variable *y* varies with the value of the explanatory variable(s) *x*. Linear regression calculates an equation that minimizes the distance between the fitted line and all of the data points. Technically, ordinary least squares (OLS) regression minimizes the sum of the squared residuals. The average values of the response variable are assumed to lie on a “regression line” or “line of means.” In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.

Thus, regression analysis allows us (among other things) to obtain estimations of the value of response variable in function of the values of one or more explanatory variables. It is thus necessary to first of all determine which variable will be expressed as a function of the others: this variable becomes the dependent variable (traditionally denoted by *y*) and the other variables then become independent or explanatory variables (traditionally denoted by *x*). The linear model then becomes:

$$Y_{i} = \beta_0 + \beta_j X_{ij} + \epsilon i$$

When:

Y_i Value of the dependent variable for individual i

β_0 First parameter of the model (often called the constant/intercept of the model)

β_j Regression coefficient/slope of the independent variable j

Xij Value of independent variable j for individual i

$\beta_0 + \beta_j X_{ij}$ The deterministic part of the model, determined by the values of the various independent variables (also known as the regression line)

ϵi the random part of the model due to random fluctuations of the individual values of Yi around the value $\beta_0 + \beta_j X_{ij}$, whose actual value is determined by uncontrolled factors (or variable not included in the model). Values of this part must be distributed normally, with identical variance for all ϵi and the ϵi must be independent of each other. We estimate these errors by residuals; these are the difference between the observed (actual) value of the observation and the value predicted by the model.

How does Y change with one unit of X?

β_o is where the line crosses the *Y* axis. The quantity β_1 is the slope and it is the rate of change in *Y* for a unit change in *X* given that the other predictors held constant. If the slope is 0, it means we have a straight line parallel to the *x* axis. It also means that we cannot predict *Y* from a knowledge of *X* since there is no relationship between *Y* and *X*.

If there is only one independent variable in the model, it is referred to as simple linear regression.

3.1.2.1. Simple linear regression

Example 6

For 10 consecutive years, we have data on the number of pigs sold in a given country and the average price of pigs. We wonder if the volumes of sales (independent variable) influence the price per unit (dependent variable)?

Year	*Number of pigs sold (millions)*	*Price per pig($)*
2000	*73*	*18*
2001	*79*	*20*
2002	*80*	*17.8*
2003	*69*	*21.4*
2004	*66*	*21.6*
2005	*75*	*15*
2006	*78*	*14.4*
2007	*74*	*17.8*
2008	*74*	*19.6*
2009	*84*	*14.1*

The figure below shows the relationship between volume and unit price and indicates that there may in fact be a negative linear relationship (the price decreases as the volume increases) [in stata scatter plots are generated using a sign tax 'sc"].

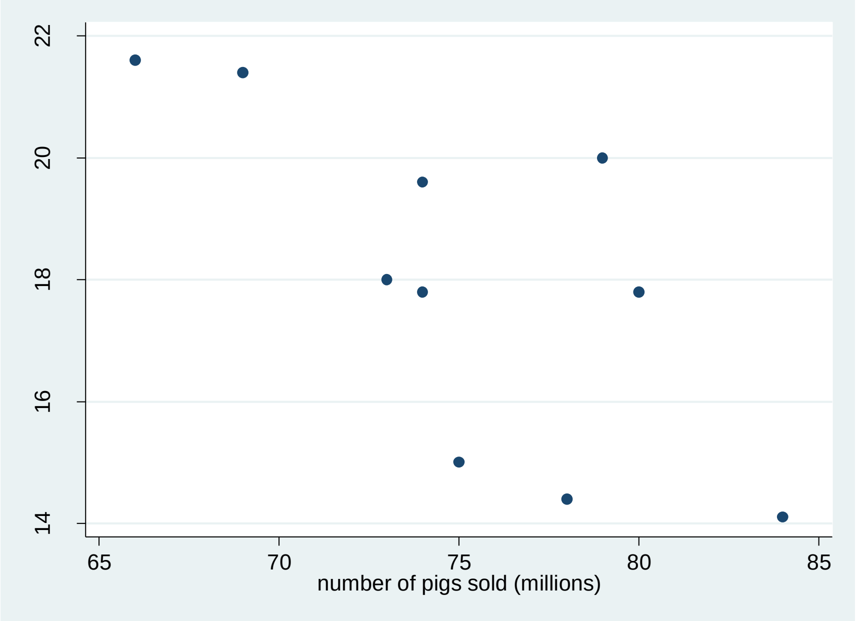


Figure 1: Scatterplot: price per pig in function of sales volume

Now lets analyse the above data set using linear regression. In stata linear regression is computed using a sign tax 'reg".

reg dep_variable indep_var

```
. reg  priceperpig numberofpigssoldmillions

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,      8) =    7.99
       Model |  34.1342349     1  34.1342349           Prob > F      =  0.0223
    Residual |  34.1867667     8  4.27334584           R-squared     =  0.4996
-------------+------------------------------           Adj R-squared =  0.4371
       Total |  68.3210017     9  7.59122241           Root MSE      =  2.0672

------------------------------------------------------------------------------
                 priceperpig |     Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------------------+------------------------------------------------------------------
numberofpigssoldmillions |   -.366877   .1298104    -2.83   0.022    -.6662202   -.0675337
                       _cons |   45.55915   9.783604     4.66   0.002     22.99812    68.12018
------------------------------------------------------------------------------
```

The information of interest to us is found in the lines starting with 'numberofpigssoldmillions' and (_cons). The column headed Coef. represents the two regression parameters.(_cons) (45.55915) stands for β_0, the constant of the regression. This constant is in theory the price per pig if nothing is sold (x = 0). Number (-.366877) stands for β_1, the regression coefficient/slope. This means every time the number of pigs sold increases by 1 the price decreases by $ 0.37. The negative sign indicates the negative trend of the regression line. The model P-value is significant (0.0223) and it indicate a very good fit of the model. The P-value of our explanatory variable (number of pigs sold) is also significant (0.022) and it indicates a strong association between the number of pigs sold and the price of the pigs.

The regression line for the price per unit in function of sales volume thus becomes:

Expected value {price} = 45.56 - 0.37*numberofpigssoldmillions

These expected values (i.e. regression line) can be calculated in stata using the sign tax 'predict'.

Predicted values

predict pred

18.78   16.58   16.21   20.24   21.35   18.04   16.94   18.41   18.41 and 14.74

Let's again use the scatter plot to express graph together the observed and the predicted values

Sign tax

sc pred priceperpig numberofpigssoldmillions , c(l)

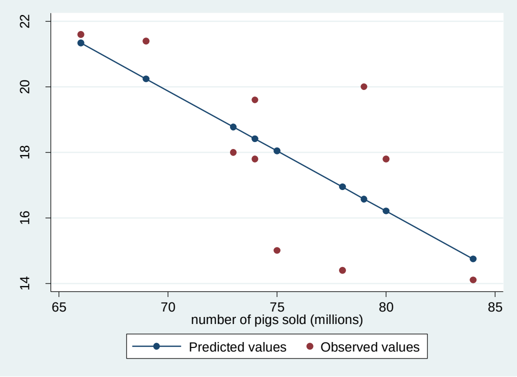


Figure 2: Scatterplot: price per pig (both observed and predicted) in function of sales volume

As it is clearly observed in the diagram the price of the pig is inversely related to the number of pigs sold, as the as the number of pigs sold increases the price per pig decreases.

3.1.2.2. Multiple linear regression

If we have two or more explanatory variable and a continuous response variable we use multiple linear regression.

Example 7

In the table below, the calving to conception interval in cows was recorded. It is believed that the older the cow the longer the interval. Using ‘age’ as a single continuous explanatory variable for the conception interval is just the same as using the quantity of pig sale as explanatory variable of the price in the example above.

id	calvcon	age	metritis	Ovar	id	calvcon	age	metritis	ovar
1	124	3	0	0	31	123	6	0	0
2	76	7	0	0	32	165	4	0	0
3	145	2	1	0	33	270	9	0	1
4	122	4	1	0	34	190	3	0	0
5	138	3	0	0	35	78	3	1	0
6	60	3	0	0	36	69	4	0	0
7	55	3	0	0	37	73	6	0	0
8	154	3	0	0	38	154	5	0	0
9	95	4	1	0	39	112	2	0	0
10	94	3	0	0	40	119	7	1	0
11	55	5	0	0	41	111	5	1	0
12	134	2	0	0	42	147	6	0	1
13	76	2	0	0	43	142	4	1	0
14	94	4	1	0	44	163	6	0	0
15	159	3	0	0	45	233	8	1	0
16	64	2	0	0	46	115	5	0	0
17	114	3	0	1	47	93	4	0	0
18	154	3	0	0	48	56	6	0	0
19	102	3	0	0	49	227	5	0	1
20	101	4	0	0	50	98	5	0	0
21	186	5	1	0	51	164	4	1	0
22	90	3	0	1	52	161	10	0	0
23	182	4	0	1	53	144	4	0	0
24	122	5	0	0	54	117	3	0	1
25	79	2	0	0	55	89	5	0	0
26	129	4	0	1	56	96	2	0	0
27	208	3	0	0	57	71	4	0	0
28	117	2	1	0	58	131	4	0	1
29	140	4	1	0	59	146	11	1	0
30	51	2	0	0	60	165	6	1	1

```
. reg  calvcon age

      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  1,    58) =    8.15
       Model |   16075.324     1   16075.324           Prob > F      =  0.0060
    Residual |  114368.609    58  1971.87258           R-squared     =  0.1232
-------------+------------------------------           Adj R-squared =  0.1081
       Total |  130443.933    59  2210.91412           Root MSE      =  44.406

------------------------------------------------------------------------------
     calvcon |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    8.47646   2.968753     2.86   0.006     2.533855    14.41907
       _cons |    87.8671   13.90357     6.32   0.000     60.03608    115.6981
------------------------------------------------------------------------------
```

The predicted values can be calculated:

Predicted averages (pred = coef(_cons) + coef(age)*age):

. predict pred

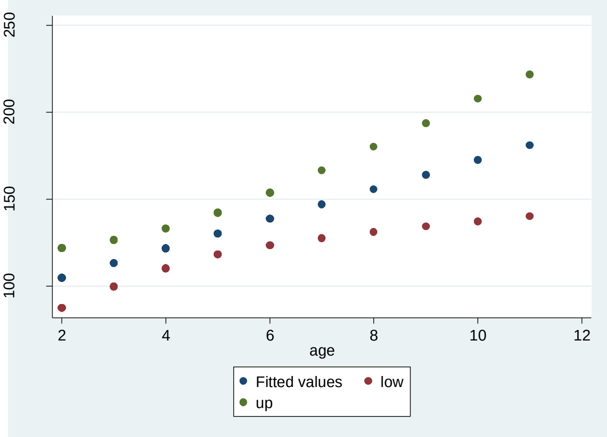
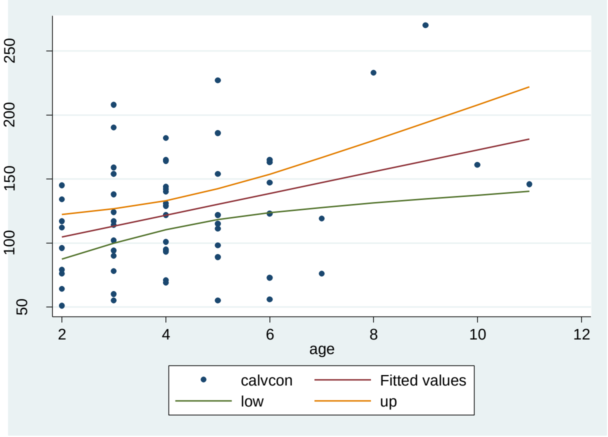
Predicted standard errors of the observations

. predict stdp, stdp

95% confidence interval

. gen low=pred-1.96*stdp

. gen up=pred+1.96*stdp

. sc pred low up age	. sc calvcon pred low up age, c(i L L L) m(o i i i)
	

It is also believed that the presence of metritis (coded as 0 or 1) and ovarian disease (coded as 0 or 1) have an effect on the interval. Introducing these explanatory variables in the model make 2 different major changes:

- the model becomes “multivariate”: there are several explanatory variables
- some of the explanatory variables are discrete (metritis and ovarian disease)

The generalization of the linear model allows the use of discrete explanatory variables but they should either be coded as 0 or 1 or declared as discrete using "i." as a prefix and "xi:" as an introduction to the command.

```
. xi:reg  calvcon age i.metritis i.ovar
i.metritis        _Imetritis_0-1      (naturally coded; _Imetritis_0 omitted)
i.ovar            _Iovar_0-1          (naturally coded; _Iovar_0 omitted)

      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  3,    56) =    5.50
       Model |  29694.0532     3  9898.01772           Prob > F      =  0.0022
    Residual |   100749.88    56    1799.105           R-squared     =  0.2276
-------------+------------------------------           Adj R-squared =  0.1863
       Total |  130443.933    59  2210.91412           Root MSE      =  42.416

------------------------------------------------------------------------------
     calvcon |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   7.030164   2.907865     2.42   0.019     1.205011    12.85532
_Imetritis_1 |   17.07149   13.05991     1.31   0.196    -9.090659    43.23364
    _Iovar_1 |   39.21718   15.00384     2.61   0.011     9.160887    69.27348
       _cons |    83.2339   13.39646     6.21   0.000     56.39755    110.0702
------------------------------------------------------------------------------
```

Interpretation:

The model P value (0.0022) is very low indicating a strong association between the response and the explanatory variables. The significance of individual variables can then be evaluated. They are all significantly associated to calving interval (P<0.05) except metritis (P=0.196). All have a positive correlation.

The coef. of age (7.03) indicate that when the age of the cow increases by one year the calving to conception interval increases by 7.03 days and 83.23 is the calving to conception interval of zero year age cow. So,

- *The calving interval of 5 year old cow = 83.23+7.03*5 = 451.3days*
- *Calving interval of 5 year old cow with metrities but no ovarian disease = 83.23+7.03*5 +17.07*1 = 468.37dyas*
- *Calving interval of 5 year old cow with ovarian disease but no metrities = 83.23+7.03*5 +39.21 = 490.51days*

3.1.3. Chi Square statistics

As a test of independence, x^2 test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. Do the numbers of individuals or objects that fall in each category differ significantly from the number you would expect? Is this difference between the expected and observed due to sampling variation, or is it a real difference? For instance, we may be interested in knowing whether a new medicine is effective in curing a given disease or not, x^2 test will help us in deciding this issue. In such a situation, we proceed with the null hypothesis that the two attributes (new medicine and cure of disease) are independent which means that new medicine is not effective in curing disease. On this basis we first calculate the expected frequencies and then work out the value of

x^2. If the calculated value of x^2 is less than the table value at a certain level of significance for given degrees of freedom, we conclude that null hypothesis stands which means that the two attributes are independent or not associated (i.e., the new medicine is not effective in curing the disease). But if the calculated value of x^2 is greater than its table value, our inference then would be that null hypothesis does not hold good which means the two attributes are associated and the association is not because of some chance factor but it exists in reality (i.e., the new medicine is effective in curing the disease). It may, however, be stated here that x^2 is not a measure of the degree of relationship or the form of relationship between two attributes, but is simply a technique of judging the significance of such association or relationship between two attributes.

Example 8

A researcher wants to test the effect of an anticoagulant drug on female patients with myocardial infarction. The researcher hope the drug lowers mortality, and set up his null hypothesis as follows:

- *H_o: There is no difference in mortality between the treated groups and the control group.*
- *H_a: The mortality in the treated group is lower than in the control group.*

The 2 × 2 contingency table in which each patient is classified as belonging to one of the four cells:

	Control	Treated	Total
Lived	89	223	312
Died	40	39	79
Total	129	262	391

The mortality in the control group is 40/129 = 31% and in the treated it is 39/262 = 15%. But could this difference have arisen by chance? We use the x^2 test to answer this question. What we are really asking is whether the two categories of classification (control vs. treated by lived vs. died) are independent of each other. If they are independent, what frequencies would we expect in each of the cells? And how different are our observed frequencies from the expected ones? How do we measure the size of the difference?

To determine the expected frequencies, consider the following:

	Control	*Treated*	*Total*
Lived	*a*	*b*	*a+b*
Died	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*N*

If the categories are independent, then the probability of a patient being both a control and living is P(control) × P(lived). The expected frequency of an event is equal to the probability of the event times the number of trials = N × P. So the expected number of patients who are both controls and live is

$$\mathrm{N} \times \mathrm{P(control\ and\ lived)} = \mathrm{N} \times \mathrm{P(control)} \times \mathrm{P(lived)} = \mathrm{N} \times \frac{(\mathrm{a}+\mathrm{c})}{\mathrm{N}} \times \frac{(\mathrm{a}+\mathrm{b})}{\mathrm{N}} = (\mathrm{a}+\mathrm{c}) \times \frac{(\mathrm{a}+\mathrm{b})}{\mathrm{N}}$$

In our case this yields the following table:

	Control	*Treated*	*Total*
Lived	*129*312/391=103*	*262*312/391=209*	*312*
Died	*129*79/391=26*	*262*79/391=53*	*79*
Total	*129*	*262*	*391*

Another way of looking at this is to say that since 80% of the patients in the total study lived (i.e., 312/391 = 80%), we would expect that 80% of the control patients and 80% of the treated patients would live. These expectations differ, as we see, from the observed frequencies noted earlier, that is, those patients treated did, in fact, have a lower mortality than those in the control group. Well, now that we have a table of observed frequencies and a table of expected values, how do we know just how different they are? Do they differ just by chance or is there some other factor that causes them to differ? To determine this, we *calculate its* x^2 *value. This is obtained by taking the observed value in each cell, subtracting from it the expected value in each cell, squaring this difference, and dividing by the expected value for each cell. When this is done for each cell, the four resulting quantities are added together to get* x^2 *value. Symbolically this formula is as follows:*

$$x^2 = \frac{(\mathrm{O_a} - \mathrm{e_a})2}{\mathrm{e_a}} + \frac{(\mathrm{O_b} - \mathrm{e_b})2}{\mathrm{e_b}} + \frac{(\mathrm{O_c} - \mathrm{e_c})2}{\mathrm{e_c}} + \frac{(\mathrm{O_d} - \mathrm{e_d})2}{\mathrm{e_d}}$$

When O *is observed frequency and* e *is expected frequency*
The particular value of x^2 *that we get for our example happens to be 13.94.*
From our knowledge of the distribution of values of x^2*, we know that if our null hypothesis is true, that is, if there is no difference in mortality between the control and treated group, then the probability that we get a value of* x^2 *as large or larger than 13.94 by chance alone is very, very low; in fact this probability is less than 0.005. Since it is not likely that we would get such a large value of* x^2 *by chance under the assumption of our null hypothesis, it must be that it has arisen not by chance but because our null hypothesis is incorrect. We, therefore, reject the null hypothesis at the .005 level of significance and accept the alternate hypothesis, that is, we conclude that among women with myocardial infarction the new drug does reduce mortality. The probability of obtaining these results by chance alone is less than 5/1000 (0.005). Therefore, the probability of rejecting the null hypothesis, when it is in fact true (type I error) is less than 0.005.*

N.B.: That value of x^2 that must be obtained from the data in order to be significant is called the *critical value.* The critical value of x^2 at the 0.05 level of significance for a 2 × 2 table is 3.84. This means that when we get a value of 3.84 *or greater* from a 2 × 2 table, we can say there is a significant difference between the two groups.

For a contingency table that has *r* rows and *c* columns, the x^2 test can be calculated as follows:

Example 9

Suppose you have the following categorical data set.

Trypanosome spp.	*District A*	*District B*	*District C*	*Total*
T. congolense	*31*	*14*	*45*	*90*
T. brucei	*2*	*5*	*53*	*60*
T. vivax	*53*	*45*	*2*	*100*
Totals	*86*	*64*	*100*	*250*

H_o*: the distribution of trypanosome spp. is not spatially different*
H_a*: the spatial distribution of the trypanosome spp. is different*
We use the above equation i.e. x^2 *= the sum of all the* $(f_o - f_e)^2 / f_e$
Here f_o *denotes the frequency of the observed data and* f_e *is the frequency of the expected values. Now we need to calculate the expected values for each cell in the table and we can do that using the row total times the column total divided by the grand total (N). For example, for cell a (i.e. the incidence of T. congolense in District A) the expected value would be 90*86/250.*

We could now set up the following table:

Observed	*Expected*	*|O-E|*	*(O—E)^2*	*(O—E)^2/E*
31	*30.96*	*0.04*	*0.0016*	*0.0000516*
14	*23.04*	*9.04*	*81.72*	*3.546*
45	*36*	*9*	*81*	*2.25*
2	*20.64*	*18.64*	*347.45*	*16.83*
5	*15.36*	*10.36*	*107.33*	*6.99*
53	*24*	*29*	*841*	*35.04*
53	*34.4*	*18.6*	*345.96*	*10.06*
45	*25.6*	*19.4*	*376.36*	*14.7*
2	*40*	*38*	*1444*	*36.1*
□^{2}				*125.51605*

x^2= 125.516

Degrees of Freedom = (c - 1)(r - 1) = 2(2) = 4

We reject Ho because 125.516 is greater than 9.488 (for alpha = 0.05)

Thus, we would reject the null hypothesis that there is no relationship between location and trypanosome spp. Our data tell us there is a statistically significant relationship between trypanosome spp. and location.

This could easily done in STATA as follows

```
. tab  district trysspp, chi2

           |             trysspp
  district |  T.  congo  T. brucei   T. vivax |     Total
-----------+----------------------------------+----------
District A |        31          2         53  |        86
District B |        14          5         45  |        64
District C |        45         53          2  |       100
-----------+----------------------------------+----------
     Total |        90         60        100  |       250

          Pearson chi2(4) = 125.5186   Pr = 0.000
```

3.1.4. Logistic regression

In epidemiology, data are often binomial (animals are infected or not, seropositive or not…). Logistic regression is appropriate when the dependent variable (outcome) is dichotomous (i.e., can be coded as 1 = event, 0 = no event), and when the question deals with the occurrence of the event of interest within a specified period time and the people/animals are all followed for that length of time. However, when follow-up time for people in the study differs, then survival analysis should be used.

Like continuous data, binomial data are characterised by their average (p) and their variance (p.(1-p)).

Using linear regression to analyse binomial data faces two major problems:

- The variance varies with p and, hence, the assumption of homoscedasticity is violated. Homoscedasticity refers to the fact that the variance of the outcome is constant at all level of the explanatory variable and within all combinations of the explanatory variable.
- The mean response should be constrained between 0 and 1 (proportions below 0 and over 1 are not possible but could be obtained using a linear regression)

The logit transformation meets these 2 concerns.

The logistic transformation

$$\text{logit}(p) = \ln\left[\frac{p}{1-p}\right]$$

logit(p) ranges between -∞ and ∞. Values below -7 or over 7 denote extreme proportions. A value of 0 corresponds to a proportion of 0.5. Proportions of 0 and 1 cannot be accommodated by the logistic transformation.

Table 1. Relation between logit(p) and p

p	1-p	var	logit
0	1	0	
0.1	0.9	0.09	-2.20
0.2	0.8	0.16	-1.39
0.3	0.7	0.21	-0.85
0.4	0.6	0.24	-0.41
0.5	0.5	0.25	0.00
0.6	0.4	0.24	0.41
0.7	0.3	0.21	0.85
0.8	0.2	0.16	1.39
0.9	0.1	0.09	2.20
1	0	0	

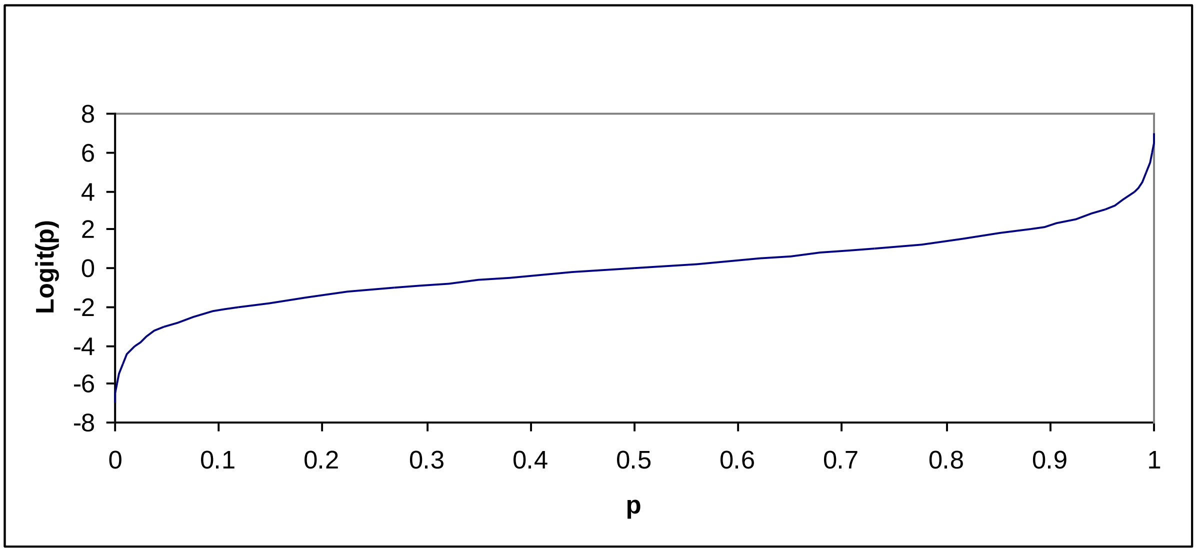


Fig. Relation between logit(p) and p

The inverse relationship is as follows:

$$p = \frac{e^{\log it(p)}}{1 + e^{\log it(p)}}$$

The logit can then be used in a linear regression as a response variable. Explanatory variables are used as for a linear regression.

logit(p) = a + bx

$$p = \frac{e^{(a+bx)}}{1 + e^{(a+bx)}}$$

Odds ratios are quite easily calculated from a logistic regression:

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \frac{e^{(a+b)}}{e^a} = \frac{e^a.e^b}{e^a} = e^b$$

since the explanatory variable x = 1 for p_1 and 0 for p_0

Fitting a logistic regression model

The calculation of the predictors is not as for a linear regression. It is based on iteration of estimates and the calculation of maximum likelihood. The set of values for the predictors that generate the highest likelihood to fit the data is retained as a model.

In Stata, the command is:

logit response_variable explanatory_variables

Example

A study was conducted to determine the influence of climatic variable on the distribution of *Rhipicephalus microplus*. Does the distribution of *R. microplus* associated with temperature? how?

Table 2

id	microplus	tmax	tmin	id	microplus	tmax	tmin
1	0	35	22	31	0	33	20
2	0	33	20	32	0	34	21
3	0	33	20	33	1	32	22
4	1	33	23	34	1	32	22
5	0	33	21	35	1	33	23
6	0	32	22	36	0	33	20
7	1	33	22	37	0	32	20
8	0	33	20	38	0	34	21
9	1	32	20	39	1	32	20
10	0	34	21	40	1	33	22
11	0	34	20	41	0	35	22
12	0	33	20	42	0	33	20
13	1	33	21	43	0	31	24
14	1	33	21	44	1	32	23
15	0	33	20	45	0	32	19
16	0	31	23	46	0	33	20
17	0	33	22	47	0	33	20
18	0	33	23	48	0	33	23
19	0	33	22	49	1	32	23
20	0	33	22	50	0	35	23
21	0	33	20	51	1	32	22
22	0	34	20	52	0	32	23
23	1	32	22	53	0	34	21
24	0	32	22	54	0	33	20
25	0	33	20	55	0	33	20
26	1	32	23	56	1	33	21
27	0	34	21	57	0	34	21
28	1	32	22	58	1	31	23
29	1	33	21	59	0	33	20
30	1	33	22	60	0	33	20

```
. logit  microplus  tmax tmin

Iteration 0:   log likelihood =  -38.19085
Iteration 1:   log likelihood = -31.300714
Iteration 2:   log likelihood = -31.185085
Iteration 3:   log likelihood = -31.184865
Iteration 4:   log likelihood = -31.184865

Logistic regression                               Number of obs   =         60
                                                  LR chi2(2)      =      14.01
                                                  Prob > chi2     =     0.0009
Log likelihood = -31.184865                       Pseudo R2       =     0.1834

------------------------------------------------------------------------------
   microplus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        tmax |  -.9301729   .4275883    -2.18   0.030    -1.768231   -.0921152
        tmin |   .5094865   .2683912     1.90   0.058    -.0165505    1.035524
       _cons |   18.91084   16.26246     1.16   0.245    -12.96301    50.78468
------------------------------------------------------------------------------
```

Interpretation of the coefficients and the p-value

tmax: coef = -0.93 and p-value = 0.030 – meaning the distribution of *R. microplus* is significantly associated with tmax hence the p-value is less than 0.05. So we reject our H_o at 0.03 significance level. The type I error we could make by rejecting our null hypothesis is only 3 in 100. The negative sign of the coefficient indicates that when tmax decreases it is more suitable for distribution of R. microplus

tmin: coef = 0.51 and p-value = 0.058 – here the p-value is greater than 0.05 so we don’t have sufficient evidence to reject our null hypothesis.

logit(*p*) = $\ln\left(\frac{p}{1-p}\right) = \_cons - 0.93 * t\max + 0.51 * t\min$

The predicted values can be estimated in Stata using a command ‘*predict’*. By default, ‘*predict’* predicts the proportion of successes. The ‘*xb’* value (linear estimation) and the observations’ standard errors can also be estimated using the ‘xb’ and ‘stdp’ options.

```
. predict p

. predict xb, xb

. predict stdp, stdp

. gen pred=exp(xb)/(1+exp(xb))

. gen lower=exp(xb-1.96*stdp)/(1+exp(xb-1.96*stdp))

. gen upper=exp(xb+1.96*stdp)/(1+exp(xb+1.96*stdp))

. list
```

id	microplus	tmax	tmin	p	xb	stdp	pred	lower	upper
1	0	35	22	0.0804307	-2.436509	1.061194	0.0804307	0.0108096	0.4117904
2	0	33	20	0.1686625	-1.595136	0.4882486	0.1686625	0.0722853	0.3456605
3	0	33	20	0.1686625	-1.595136	0.4882486	0.1686625	0.0722853	0.3456605
4	1	33	23	0.483337	-0.0666768	0.5557225	0.483337	0.2394144	0.7354669
5	0	33	21	0.2524383	-1.08565	0.3432191	0.2524383	0.1469959	0.3982073

more–

Odds ratio scan either be estimated by calculating the exponential of the coefficients or by using the ‘*or*’ option in the ‘*logit*’ command.

3.1.5. Non- parametric analysis

Many statistical methods require assumptions to be made about the format of the data to be analysed. For example, the paired t-test introduced above requires that the distribution of the differences be approximately normal, while the unpaired t-test requires an assumption of Normality to hold separately for both sets of observations. Fortunately, these assumptions are often valid in clinical data, and where they are not true

Of the raw data it is often possible to apply a suitable transformation. There are situations in which even transformed data may not satisfy the assumptions, however, and in these cases it may be inappropriate to use traditional (parametric) methods of analysis. (Methods such as the t-test are known as ‘parametric’ because they require estimation of the parameters that define the underlying distribution of the data; in the case of the t-test, for instance, these parameters are the mean and standard deviation that define the normal distribution.)

Nonparametric methods provide an alternative series of statistical methods that require no or very limited assumptions to be made about the data. The analyses in non-parametric tests are usually based on the ranks of the data, i.e. on observations when they are arranged in increasing (or decreasing) order, rather than on the raw data.

Table parametric tests and some non-parametric equivalent

Parametric test	**Non parametric test**
Single sample t-test	Sign test
Paired t-test	Sign test, Wilcoxon signed rank test
Independent sample t-test/two sample t-test	Wilcoxon signed rank test/ Mann-Whitney U-test
One way ANOVA	Kruskal-Wallis one way ANOVA
Two way ANOVA	Friedman two-way ANOVA
Pearson correlation	Spearman rank correlation