**The role of statistics in veterinary clinical research**

## Learning Objectives

At the completion of this session, you will be able to

1. Describe the principles of estimation and hypothesis testing.
2. Appropriately use and interpret a p-value and/or confidence intervals from a given study.
3. Appropriately use the terms: census, sample, survey, and surveillance.
4. Utilize appropriately the various common sampling techniques in veterinary research and clinical application.

## Census, Sampling, Survey and Surveillance

### Definitions

1. *Census*  a collection of information on every individual in the population or all the members of a group with certain specified attributes.
2. *Sample*  collection of information upon a subset of individuals in which the results are inferred to be representative of a larger population.
3. *Survey:* an investigation in which information is systematically collected to estimate the occurrence of an event in the population, but in which the experimental method is NOT used. This means no intervention or manipulation of the population is used to obtain this information.
4. *Surveillance:* an ongoing scrutiny, generally using methods distinguished by their practicality, uniformity, and frequently their rapidity rather than complete accuracy. In essence, it is a monitoring of certain events that is used to detect a change in trend or distribution in order to initiate investigative or control measures. A survey is not a surveillance, but could become one if it is continued to monitor the population initially investigated. For example, a surveillance can be used to measure a change in infection rate between seasons, geographical regions, etc.

Two examples of surveillance organizations are:

* 1. *NAHMS (National Animal Health Monitoring System):*

1. supervised by the USDA
2. monitors disease prevalence and cost of production
   1. *MCI (Market Cattle Identification):*
3. also supervised by the USDA
4. collects serum samples from every adult cow slaughtered and tests for antibodies to Brucella. The surveillance is followed back to the herd of origin. This system requires extremely competent animal identification methods.

### Types of Sampling

1. *Probability sampling:* is a random access to every individual. Every individual in the population has a known chance of being sampled (i.e. 1/10, 1/1000, etc.). Inference of the sample is applied to the rest of the population. The degree of bias depends on how the sample was taken and this will determine if the sample truly represents the rest of the population.
2. *Non-probability sampling:* this is done on the basis of convenience and the sample is usually not representative of the population under investigation. For example, if the investigation was to determine prevalence of a certain disease among deer and only those deer easily caught were sampled, this may not be representative because maybe those deer that are easily caught are that way because they are ill. Another example would be in a survey to measure prevalence of Heartworm in Colorado, the investigator would only ask those veterinarians that he knows. This may lead to a bias, for example if those veterinarians were located in the western slope area where heartworm prevalence is higher. There is no design to this method of sampling. The problem with this type of sampling comes when the results are applied to the entire population. This type of sampling may work and may actually be necessary at the beginning of an investigation because it may answer an initial question (e.g. Is there heartworm at all?).

#### Types of Probability Sampling

1. *Simple random sampling:* this is the ideal situation. Every individual will have an equal chance to appear in the sample. This type of sample can be done correctly in several ways.

For example:

* 1. Assign each individual a number and using tables (computer or book), select five numbers. This will not guarantee a representative sample, but it will decrease bias and give a better chance for a representative sample.
  2. Pull names from a hat.

The disadvantage of this is that a list of every individual in the population is needed. This could prove to be a difficult task.

1. *Simple stratified sampling:* the population is divided into strata (subgroups) according to certain criteria that are important to the investigation. For example, in the Heartworm study dogs could be divided into large and small size because large dogs have a higher incidence of the disease. Then a random sample is performed among each strata. The problem of this method is that each stratum needs to be equal in size to the others and this is not likely to happen.
2. *Proportional stratified sampling:* this takes into account the problem of strata of unequal size. The sample among strata is obtained with regard to the contribution of the strata to the size of the total population. For example, if large dogs contributed 50%, medium dogs 30%, and small dogs 20%, a sample of 300 dogs would include: 150 large dogs, 90 medium dogs and 60 small dogs. Then a simple random sampling can be done among them. This kind of sampling is the most commonly used.
3. *Cluster sampling:* the unit of sampling will be a group of individuals rather than a single individual. For example, if there are three dogs in a kennel cage, this would represent one dog unit. If any one of these dogs were positive for Heartworm, then the unit would be considered positive. Every animal in the unit must be surveyed.
4. *Multistage sampling:* this is when more than one of the above methods is incorporated into the investigation design. For example, in the heartworm survey:
   1. a letter is sent to all vets asking whether or not they wish to participate.
   2. those that respond are used in the survey.
   3. the sample is clustered by clinic and every dog that comes in for a period of six months must be surveyed.
   4. the clusters are stratified by size of dog and region.

This chart indicates how these various methods could be used:

Population Characteristics and Sampling Techniques Appropriate

for Each Population Type

|  |  |  |
| --- | --- | --- |
| Population  Characteristic | Population Type | Appropriate  Sampling Technique |
| Population is generally a homogeneous mass of individual units. | Number of breeding bitches of a particular breed housed in a specific kennel from which random samples are selected for testing the presence or absence of a disease in the vaginal swab. | Simple Random |
| Population consists of definite strata, each of which is distinctly different, but the units within the stratum are as homogeneous as possible. | A particular bull breeding farm in which the total population consists of three breeds (strata), each with equal numbers of bulls. A sample is needed to evaluate the libido among bulls on the farm. | Simple Stratified |
| Population contains definite strata with differing characteristics. Each strata has a proportionate ratio in terms of number of members of every other strata. | A county in which the total dairy population consists of farms with three different size herds. | Proportional Stratified |
| Population consists of clusters whose characteristics are similar, yet whose unit characteristics are as heterogeneous as possible. | A survey of small animal wards in a teaching hospital to evaluate the presence or absence of antibiotic resistant bacterial spp. All wards are similar in atmosphere, purpose, design, etc. Yet the patients differ widely in individual characteristics: species, breed, sex, reason for hospitalization, and so forth. | Cluster Sampling |

## Concepts of Statistics Used in Veterinary Medicine

Statistical inference is the process whereby one draws conclusions regarding a population from the results observed in a sample taken from that population. There are two categories of statistical inference: estimation and hypothesis testing. Estimation is concerned with estimating the specific value of an unknown population parameter while hypothesis testing is concerned with making a decision about a hypothesized value of an unknown population parameter. In either case, we first need some background concerning something called the *standard error*.

### The standard deviation of the mean (standard error)

Suppose we defined a population to be all 100 dairy cows on a farm, and we took repeated random samples consisting of 20 cows from the herd and calculated the mean body weight of each sample. We would find that the estimated mean body weight of each 20 cow sample would vary around the true (unknown) population mean body weight of the whole herd. We would also note, that after consulting with a statistician, that these sample means follow a t distribution. A t distribution is like a standard normal distribution with slightly fatter tails and a lower center. The statistician would also tell us some other interesting facts about what we did.

First, if we had chosen a larger sample size, perhaps 30 cows, then the distribution of the sample means would be approximated by the standard normal distribution (Z). Secondly, our experiment was in fact demonstrating the principle of the Central Limit Theorem (CLT). The CLT states that whenever n is moderately large, the mean has approximately a normal distribution regardless of the distribution of the underlying variable. So even if the body weights of the 100 cows in the herd were not normally distributed, our sample means would be.

An estimate of the average variation or standard deviation of the sample means is called the standard error (SE). It is estimated as the sample standard deviation (S) divided by the square root of the number of observations in the sample (n):



Example: the mean body weight of a sample of 20 dairy cows was 650 kg with a standard deviation of 40 kg. The standard error of the mean estimate is = 40/4.472 = 8.94. Therefore the best estimate of the mean body weight of this dairy herd is 640 kg with a standard error of 8.94 kg.

As the number of observations in the sample increases the variability of the mean decreases i.e., the standard error gets smaller. The SE provides a measure of how far from the true population value the estimate is likely to be. Most often the estimate will be within one standard error of the mean and is unlikely to be more than 2 SE's away from it.

The standard error and standard deviation are commonly confused, which is understandable given that the standard error is the standard deviation of the sample means. The standard error is used to describe the *preciseness* of our estimate, while the standard deviation is used to describe the *variability* of the population or distribution.

#### The standard error of a proportion

Binomial or dichotomous data are often viewed in terms of proportions - for example, the proportion of individuals who have a particular condition in a given population. An estimate (p) of the true population proportion can be obtained simply by counting the number of events in a sample:

p = r/n

where r = the number of events

n = the number of observations in the sample

The standard error of this proportion is given by:

, where q = 1 - p

4

For example, suppose we did not know the true first service conception rate in a particular dairy herd. We could estimate it by observing the number of pregnancies which result from breeding, the 20 recently freshened cows. Suppose 11 cows became pregnant. An estimate of the true herd first service conception rate (p) is therefore 11/20 = 0.55 (or 55%). The standard error of this estimate is 

### Estimation

When we wish to estimate unknown population parameters such as the mean body weight of cows () or the variance of the body weights of cows (s2), we take a random sample of the population and calculate the sample mean () and the sample variance (S2). These estimates are called point estimates. They represent estimates of the true population parameters and, as such, have a certain degree of inherent variability associated with them. After calculating a point estimate, we would like to know how good an approximation of the true population value this estimate is (i.e., what is the precision of the estimate?). A confidence interval (CI) is a way of quantifying the precision of the estimate. A CI consists of a lower and upper limit on either side of the point estimate. It is calculated using the following format:

Point estimate + percentile of x Standard error of

of parameter the distribution the estimate

Example: To calculate a 95% CI for our estimate of the body weight of a herd of cows, we first find that the appropriate percentile value of the t distribution [with 19 (n - 1) degrees of freedom] is 2.093. We know that the standard error of the mean is = 8.94. Thus the 95% CI for the mean body weight is 650 kg + (2.093 x 8.94) = 650 kg + 18.71 = 631.29, 668.71.

The interpretation of the CI is critical. This 95% CI means that in repeated sets of samples, 95% of such intervals would be expected to contain the true value of the population (herd) mean. So, if we were to repeat the sampling of the herd many times, there would be a 95% chance that the CI of 631.29 to 668.71 would include the true (unknown) value of the mean body weight of the herd. As we shall see below, calculation of CIs is also very useful when performing hypothesis testing.

The exact level of confidence is explicitly stated, for example, a 99% CI or a 95% CI. A 99% CI for this same estimate would be 650 kg + (2.861 x 8.94) = 650 kg + 25.58 = 624.42, 675.58 (the 99th percentile of the t distribution with 19 degrees of freedom is 2.861). This interval is wider that the 95% limit, as we would expect, since the mean is more likely to be included.

### Hypothesis Testing - an example using the t test

Estimation using CIs and hypothesis testing are closely related. In estimating a CI, we use the sample data to estimate what we think is a likely set of values for the population parameter of interest. In hypothesis testing we use our sample data to test whether our estimated value for the parameter is different enough from a hypothesized value to conclude that a true difference exists. Hypothesis testing actually centers around rejecting or not rejecting the null hypothesis. The null hypothesis is a statement which you are trying to refute using your data. This is best explained by an example:

We are interested in determining whether BST (growth hormone) affects the body weight of adult dairy cows. We perform an experiment where we randomly assigned half our dairy herd to receive BST (growth hormone) and the other half to receive a placebo (e.g., saline).

To test the effect of BST we first formulate a **null hypothesis** that states that there is no difference in the mean body weight of the two groups of cows:

Ho = the mean body weights of the treatment and control groups do not differ.

If we reject this null hypothesis, then we accept our **alternative hypothesis**, which formally stated is:

Ha = the mean body weights of the treatment and control groups do differ.

Another important concept is the **p value**. The p value quantifies exactly how unusual the observed result from our experiment would be *if the null hypothesis were true*. The formal definition of a p value is:

The p value is the probability of obtaining a value of the test statistic at least as large as the one observed, *given that the null hypothesis is true*.

So, if the observed result is very unlikely, given that the null hypothesis is true, we would get a very small p value (e.g., a probability of 0.001) and we would reject the null hypothesis in favor of the alternative. In other words, if we find that the cows receiving BST gained a lot of weight, say an average of 50 kg, compared to the control (saline) group, our test would have a very small p value associated with it. This would say that if the mean body weights of the two groups of cows really did not differ, the probability of observing a difference of 50 kg is very unlikely. In this case, we would decide to reject the null hypothesis and conclude that the alternative hypothesis was correct i.e., that the mean body weights of the treated and control cows really do differ.

Prior to actually performing the test we need to define a descriptive level of significance or an **alpha value** which forms the decision rule for rejecting or not rejecting our null hypothesis. Defining a descriptive level of significance or alpha level is simply deciding how unlikely our result has to be before we decide to reject the null hypothesis. Frequently an alpha level of < 0.05 is chosen, although if one wanted to be very stringent a level of < 0.01 could be specified.

Let us calculate a hypothesis (or significance) test for our experiment. We first set an alpha level of < 0.05. The results obtained from the experiment were as follows:

Treated group: n= 50, mean body weight () = 700, ST = 38

Control group: n= 50, mean body weight () = 650, SC = 42

This particular two-sample significance test is performed using a T statistic:



5

where Sp refers to the pooled standard deviation of the two groups - in this case equal to 40.

The denominator of the T statistic is in fact the pooled standard error of the mean difference between the treatment and control groups, which in this example equals 8.

T = (700 - 650) = 6.25

8

By referring this value to a t table with n - 2 degrees of freedom (= 98) we find that the p value associated with this result is < 0.001. Thus a difference of 50 kg between the two groups, if the null hypothesis was true, is so unlikely that we reject the null hypothesis in favor of the alternative.

If we were to calculate a 95% CI around the observed treatment difference of 50 kg, we would obtain values of 34.12 and 64.12. Because this interval does not contain 0, we can conclude with 95% confidence that there is a significant change in body weight with BST treatment. Thus, calculating a 95% CI is equivalent to performing the above significance test at an alpha level of 0.05.

### Hypothesis Testing - an example using the chi-squared test

Frequently we want to perform a hypothesis test on data which are either qualitative or binomial (i.e., a proportion). In this situation the chi-square test is an appropriate method of testing whether a relationship or association exists between two variables.

For example, in our BST experiment on the 100 dairy cows, we were also interested in knowing whether fertility was affected by use of the hormone. During the 12-month period following treatment with either BST hormone or the placebo, the following data were collected concerning the subsequent fertility of the 100 cows.

Table of Observed frequencies

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Pregnancy | Status |  |
|  |  | Yes | No | Total |
| Treatment | BST | 40 | 10 | 50 |
|  | Placebo | 30 | 20 | 50 |
|  | Total | 70 | 30 | 100 |

The null and alternative hypotheses are defined as:

Ho = there is no association between BST treatment and subsequent fertility.

Ha = there is an association of some type between BST treatment and subsequent fertility.

To perform the test, we find for each cell in the table the frequency that we would expect to occur, if the null hypothesis were true. We use the row and column totals (called the marginal totals) to do this. The first row of the table represents the 50 cows that received BST. The probability of a cow being in the first row is therefore one half. If there was no association between BST and fertility, we would expect each column of the table to have the same proportion of its members (i.e., 1/2) in the first row. So, we would expect 35 of the 70 cows in the first column to be in the first row (i.e., 70 x 50/100) and 15 of the 30 cows in the second column to be in the first row (i.e., 30 x 50/100). The general formula used to calculate expected frequencies is therefore:

Expected frequency = Column total x Row total

Grand total

Under the null hypothesis of no association, the following table of expected frequencies was produced:

Table of Expected frequencies

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Pregnancy | Status |  |
|  |  | Yes | No | Total |
| Treatment | BST | 35 | 15 | 50 |
|  | Placebo | 35 | 15 | 50 |
|  | Total | 70 | 30 | 100 |

We now compare the observed frequencies with the expected frequencies using the chi-square test. If the two variables are not associated, the observed and expected frequencies in each cell should be close. The chi-square statistic is:



6

For our BST experiment:







When the null hypothesis is true the test statistic is distributed as a chi-square with degrees of freedom equal to:

(number of rows - 1) x (number of columns - 1)

So, the chi-square test in this example has one degree of freedom. Referring the value of 4.762 with one degree of freedom to a chi-square table, we find the p value to be < 0.05 (the critical value for p < 0.05, for a 1 degree chi-square test is 3.84). We therefore reject the null hypothesis and conclude that there is an association between BST treatment and subsequent fertility. In this instance, BST was associated with better fertility.

### Power and Error rates

The following table summarizes the decisions that result in hypothesis testing:

Outcomes of hypothesis testing

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Truth |  |
|  |  | Different | Not Different |
|  | Difference | Power | Type I error |
| Decision | (Reject Ho) | (1-) | () |
|  |  |  |  |
|  | No Difference | Type II error | Confidence |
|  | (accept Ho) | () | (1-) |

If the true state of nature is that the null hypothesis is really true and the decision is made to accept it, then a correct decision has been made. However, if the null hypothesis was rejected, we have made a false positive decision by accepting the alternative hypothesis. This is called a *type I error* and occurs with a probability of alpha (). The probability of correctly accepting the null hypothesis as true is therefore 1 -

If the alternative hypothesis is really true, then we can either make a correct decision by rejecting the null hypothesis or a wrong decision by failing to reject the null hypothesis - a false negative result. A false negative result is called a *type II error* and occurs with a probability of beta (ß). The probability of correctly accepting the alternative hypothesis is (1 - ß) which is commonly called the power of the test. This is a measure of how likely your experiment is to find a real difference in your data, if a real difference actually exists.

For a fixed sample size, a and ß are inversely related. If one guards against making a type I error by choosing a small a or p value, then ß will be correspondingly large and the power of the test will be reduced (the probability of making a type II error increases). Conversely, if ß is reduced to avoid making a type II error, a is increased so the risk of a type I error is greater.

Ideally in research design we would like both a and ß to be small. Usually alpha is set by convention e.g., a = < 0.05. The resulting power of a test (1 - ß) for a given a, is dependent on the number of observations in the experiment. Sample size calculations can be made to determine the number of observations required to achieve a given level of power (again, refer to your local statistician for help).

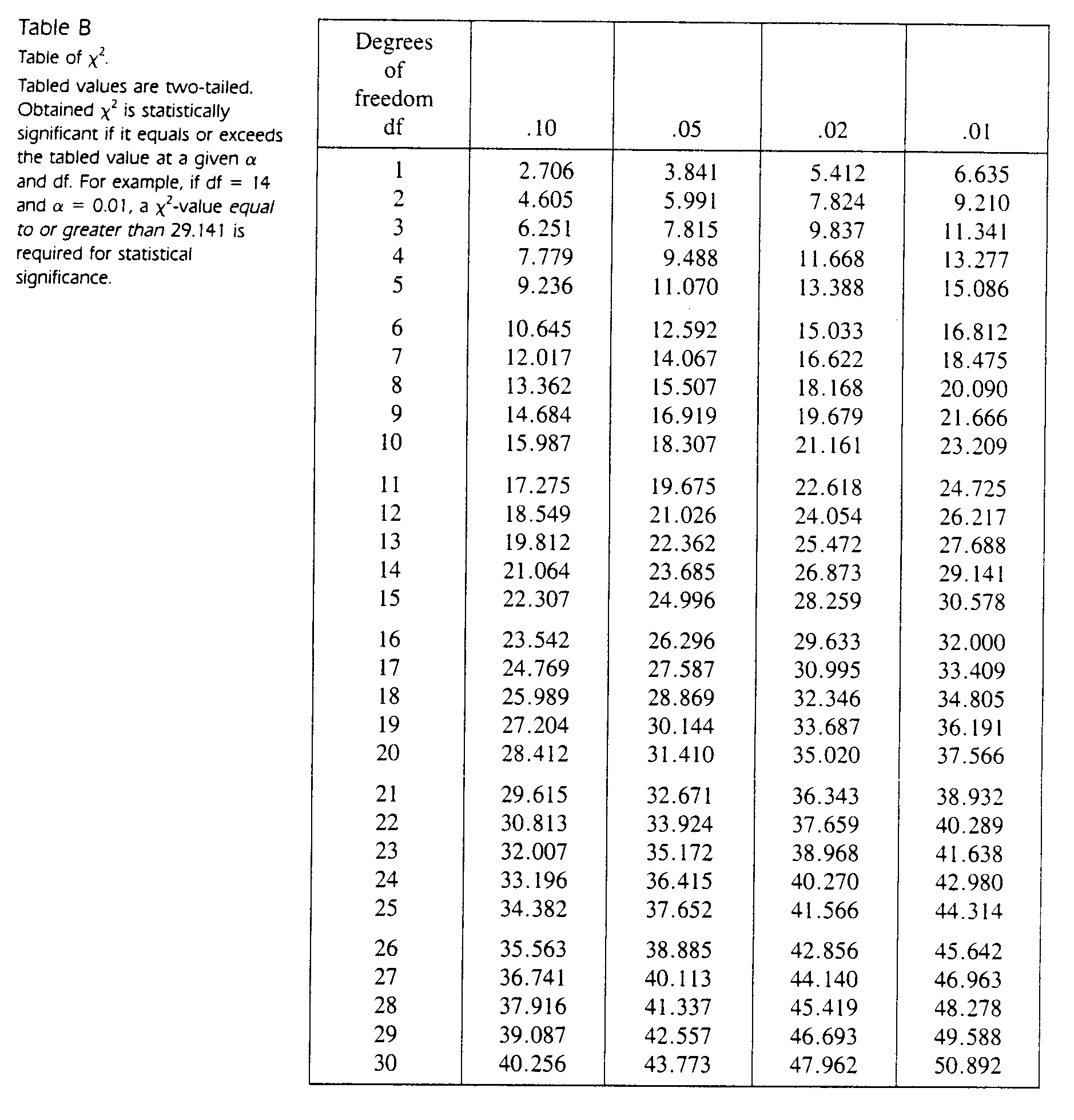
### Experts only

The best way to understand the above table is to apply the concepts of sensitivity, specificity and predictive values which are used to describe the value of diagnostic test information. If we regard the columns indicating the "true state of nature" as the disease status, the rows indicating the "hypothesis accepted" as the test information, and define the Ho as negative and Ha as positive, then we can show the following:

The type I error is the false positive rate that occurs with a probability equal to  or the p value. The type II error is the false negative rate and occurs with a probability equal to ß.

The specificity (SP = TN/TN + FP) is the probability of correctly accepting the null hypothesis, which occurs with a probability of 1 - . Because  is usually set at 0.05, the SP will always be high (0.95). When SP is high we maximize predictive value positive (TP/TP + FP), thus when we find a positive result (by rejecting the null hypothesis) we can usually be very sure that the alternative hypothesis is really true.

The sensitivity (SE = TP/TP + FN) is the probability of correctly accepting the alternative hypothesis, which occurs with a probability of 1 - ß. The SE is therefore the power of the test. Because ß is not set at a given level it tends to vary widely and can often be high, producing a low SE or **power**. When SE is low the predictive value negative (TN/TN + FN) tends to be low. So, if we find a negative result (by accepting the null hypothesis) we often cannot be sure if the null hypothesis is really true. This is what we mean when we say a test or experiment has low power; we are concerned that our negative result may be wrong because the test may have failed to detect a difference that was really present.



**How to present data and information?**

The presentation of data in the form of tables, graphs and charts is an important part of the process of data analysis and report writing. Although results can be expressed within the text of a report, data are usually more digestible if they are presented in the form of a table or graphical display. Graphs and charts can quickly convey to the reader the essential points or trends in the data - remember, a picture can say a thousand words! Graphs and charts are particularly useful when data are being presented to an audience, because information has to be conveyed in a limited time period. Graphs and charts can, however, be misleading especially if the author uses subtle alterations of scale such as excluding a 0 reference point on one of the axes.

There are some general common sense recommendations to follow when presenting data:

i) The presentation should be as simple as possible. Avoid the trap of adding too much information. It is not the aim to include all the information you have but only a summary of the essential feature(s) you are tying to illustrate. A good rule of thumb is to only present one idea or to have only one purpose for each graph or chart you create.

ii) The presentation should be self-explanatory. A chart or graph is not serving its purpose if the reader cannot comprehend the legends or has to refer to the text in order to understand it. There is a careful balance between too much information which makes the graph or chart too complicated and too little information that makes the chart difficult to comprehend or worse misleading.

iii) The title should be clear, and concise indicating what?, when?, and where? the data were obtained.

iv) Codes, legends and labels should be clear and concise, following standard formats if possible.

v) The use of footnotes is advised to explain essential features of the data that are critical for the correct interpretation of the graph or chart.

**I. Tables**

There are no real hard and fast rules with regard to table design. Tables are a standard method of presenting qualitative or categorical data, but they can also be used to summarize quantitative data (see Table 1). The simplest table is the two-column frequency table. The first column indicates the grouping of the data, while the second column lists the frequencies or count for each group. For example:

Table 1. Estimated cattle population size by breed

in Jefferson County for 1992

Breed Estimated Population

Hereford 27500

Angus 16250

Charolais 7100

Tables may be used to cross-classify individuals by two or more qualitative variables to form a **frequency** or **contingency table**. For example, in Table 2 the same Jefferson County cattle population has now been cross-classified by breed and sex. The number or frequency of individuals in each "cell" of the table is presented, as well as the percentage of the total population that each sex-breed combination represents. For example, there are 100 Charolais males which represent 2% of the total population.

Table 2. Breed and sex distribution of cattle in

Jefferson County for 1992

Male Female

Hereford 14200 (27%) 13300 (26%) │ 27500 (53%)

│

Angus 11100 (21%) 5150 (10%) │ 16250 (31%)

│

Charolais 100 (2%) 7000 (14%) │ 7100 (16%)

─────────────────────────────────────┼──────────

│

Total 25400 (50%) 25450 (50%) │ 50850 (100%)

Contingency tables can be made more complex by adding more variables or "layers". For example we could expand Table 2 to include vaccination status, and whether the animal reacted to a tuberculin test. This would produce a 4-dimensional contingency table which includes the variables breed, sex, vaccine status, and T.B. status (positive or negative). Contingency tables are analyzed using an array of statistical techniques different from quantitative data. Simple relationships between two variables can be explored with a chi-square test. Binomial variables i.e., those that have only two outcomes (0,1) are often presented in the form of a table e.g., male or female in Table 2. Binomial data can be explored using the binomial model or, if there is a large number of observations, an approximation to the normal distribution may be used (see later).

**II. Graphs**

Graphs are a useful method to display **quantitative** data. The standard graph uses two rectangular co-ordinates (called the x and y axes). The independent variable is usually plotted on the horizontal x axis, while the response or outcome variable is plotted on the vertical y axis. The outcome variable is usually a quantitative measure such as a frequency (count) or a percentage.

There are several different types of graphs:

i) Histogram - A **frequency** histogram and a **relative frequency** histogram are two very useful display graphs applicable to quantitative (measured or counted) data. Specifically, they are used to depict a **frequency distribution** of a quantitative variable. A frequency distribution contains the counts or frequencies of all the values of a variable observed in the data. For example, the following is a frequency distribution describing the parity (a **discrete quantitative variable**) of a herd of 125 milking cows.

An example of a frequency distribution for a discrete quantitative variable (parity)

Parity Frequency Relative Cumulative

frequency(%) frequency

1 39 31.2 39

2 34 27.2 73

3 22 17.6 95

4 18 14.4 113

>5 12 9.6 125

───────────────────────────────────────

Total 125 100.0 125

───────────────────────────────────────

The frequency column represents the absolute number of observations or counts for each parity. The relative frequency column represents the proportion or percentage of the total observations that a particular group or class interval represents. This frequency distribution shows that there are 39 first calf heifers which represent 31% of the herd, while cows that have had at least 4 calves represent 24% of the herd.

The following is a frequency distribution for a **continuous quantitative variable**; the birth weight of 40 calves born on a dairy farm. The frequency distribution describes the **relative distribution** of the data, which as we will learn later, is a central issue in generating descriptive statistics and performing tests of statistical inference.

An example of a frequency distribution for a continuous quantitative variable (the birth weight of 40 calves)

Birth Wt. Frequency Relative Cumulative

(lbs) frequency(%) frequency

65 1 2.5 1

70 1 2.5 2

72 1 2.5 3

76 2 5.0 5

78 3 7.5 8

79 2 5.0 10

83 4 10.0 14

84 6 15.0 20

85 8 20.0 28

87 4 10.0 32

90 3 7.5 35

92 1 2.5 36

95 2 5.0 38

98 1 2.5 39

102 1 2.5 40

───────────────────────────────────────----

Total 40 100.0 40

───────────────────────────────────────

A histogram, unlike a bar chart, contains no scale breaks (gaps) along the x axis. The independent variable is divided into a number of sub-intervals or groups called **class intervals** which are plotted on the horizontal x axis. Generally between 5 and 20 class intervals are used depending on the range of the data and the total number of observations. The class intervals are always the same width.

In a **frequency histogram**, the y axis is used to record the absolute number of observations or counts for each class interval. The following example uses the parity data presented above, in this case, each parity (1 - > 5) acts as a natural class interval on the x axis:

Graph 1 - Example of a frequency histogram:

A frequency histogram representing the distribution of a dairy herd by parity



7

In a **relative frequency histogram** the relative frequency, (expressed either as a percentage or a proportion) for each class interval is plotted on the y axis. For the calf birth weight example, the data ranged from 65 (the smallest observation) to 102 (the largest observation). Eight class intervals, each 5 units (lbs) wide would appear to be most logical (see graph 2):

An alternative method for creating a frequency distribution is to use a stem-and-leaf plot (see below). This is a very simple and effective method of presenting data which can be easily performed by hand.

The area of the histogram (calculated by multiplying the height by the width of each class interval) represents all of the data collected (a concept termed "the area under the curve"). This is an important point, as the histogram can be used to determine what proportion of the data a certain value or a range of values makes up. Examining the shape of a frequency histogram also provides useful information. The distribution shown in graph 2 is roughly symmetrical about a central value (85 lbs). The symmetrical, unimodel nature of the data resembles the normal distribution (see later). By contrast the parity data represented in graph 1 is asymmetrical. Most frequency distributions encountered in biological data are either symmetrical or skewed to the right.

Graph 2 - Example of a relative frequency histogram:

A relative frequency histogram representing the distribution of 40 calf birth weights.



8

ii) Arithmetic scale line graph - This type of graph is particularly useful when data are recorded at intervals in time. Such data are called a **time series**. The scale of each axis is constant - an equal distance represents an equal quantity. A simple line is drawn to demonstrate the relationship between the two variables. An example is graph 3 which depicts the total number of new lambs born each week during a 12-week lambing season.

iii) Frequency polygon

A frequency polygon is used when you want to present more than one set of data in terms of a frequency distribution on one graph. The frequency polygon is constructed from a histogram by simply using a line to join up the mid-points of the top of each column for each class interval. The area under the curve still represents the total number of observations.

Graph 3 - Example of a line graph

Total number of lambs born each week on the Smith farm

during the 1992 lambing season.



9

Graph 4 - Example of a frequency polygon

The number of lambs born each week on two different farms during the 1992 lambing season.



10

iv) Scatter Diagram

This graph is extremely useful to explore the relationship or association between two continuous variables. Several sets of paired data are plotted on the graph and the resulting pattern of points is indicative of a possible relationship. For example, during the clinical examination of a number of beef calves suffering from viral pneumonia the heart rate and respiratory rate were recorded for each calf. The following scatter diagram indicates that there seems to be some sort of relationship between the two variables. (N.B. The Pearson correlation co-efficient (r) could be used to quantify the strength of the linear relationship between these two variables).

Graph 5 - Example of a scatter diagram

Comparison of the observed heart rates and respiratory rates for 20 beef calves examined with clinical signs of respiratory distress



11

**III. Stem-and-leaf plot**

Stem and leaf plots represent a simple and extremely effective method by which a frequency distribution can be constructed by hand. Stem and leaf plots clearly show the range of the data, where the observations are concentrated and the shape of the distribution. The following example illustrates how to create such a plot. Data representing the live weights (kg) of 24 pigs prior to slaughter were obtained from a large commercial operation:

61.2 71.8 66.7 65.3 60.2 64.3

62.3 65.3 68.3 63.9 69.7 66.0

67.6 69.5 63.7 71.1 59.6 67.9

65.6 68.4 65.9 65.2 62.5 68.3

Examining the raw data we see that it ranges from 59.6 to 71.8 kgs. We decide to create 13 class intervals each 1 kg in width that will include all the data i.e., 59.0 - 59.9, 60.0 - 60.9,....,71.0 - 71.9. The leading two digits of the class intervals represent the stems of the plot while the last digit represents the leaves. The stems represent the row or class interval in which the observation is placed, while the leaves represent the ordering of the observations within each class interval or stem. The above data were used to create the following stem and leaf plot.

59 │6

60 │2

61 │2

62 │3 5

63 │7 9

64 │3

65 │2 3 3 6 9

66 │0 7

67 │6 9

68 │3 4 3

69 │5 7

70 │

71 │1 8

Notice how this simple plot gives immediate information on the range of data, the frequency distribution of the data as well as the presence of extreme or outlier observations. Note also that all of the original data values can be reproduced from the plot if necessary.

**IV. Charts**

It is often convenient to present data pictorially. Information can be conveyed much more quickly by a diagram than by a table of numbers. Charts and diagrams can help a reader or audience quickly get the salient point of an analysis or report. Unfortunately, because diagrams can be very misleading they should be treated as a compliment to numbers, not a replacement. There are several types of charts which are used to illustrate data, the two most common being the pie chart and bar chart:

i) Pie Chart

A pie chart is used to display the frequency distribution of a **qualitative** variable (whereas the histogram displays the frequency distribution of a quantitative variable). The relative frequency of each group or category is proportional to the number of degrees or angle of the pie. Each sector therefore represents the proportion of the total number of observations that belong to that particular category. Pie charts work best when a relatively small number of categories are used e.g., 5 or 6. Pie charts are easier to read if the segments are ordered by size.

Graph 6 - Example of a pie chart

Total annual beef production (in 1000 tons) by feedlot for Jefferson County (1991)



12

ii) Bar Chart

Histograms and pie charts depict the distribution of a single variable. A **bar chart** or bar diagram shows the relationship between two or more variables, usually one being quantitative and the other qualitative or a quantitative variable which has been grouped, such as time or age in years (see graph 7). The bars which represent the different groups are shaded, hatched or colored and are always the same width.

A bar chart must not be confused with a histogram. A bar chart can be distinguished from the histogram by the presence of scale breaks or gaps between its columns or bars (a frequency histogram has no scale breaks or gaps along its horizontal axis). A scale is not used along the horizontal axis of a bar chart (compare with the frequency histogram). Bar charts may be constructed with either horizontal or vertical bars. In the latter case, the vertical axis depicts the frequency while the different classes or groups are arranged on the horizontal axis, clearly separated by spaces.

Graph 7 - Example of a vertical bar chart

Comparison of the percentage morbidity by age group for beef cattle affected with respiratory disease in two feedlots in Jefferson County.



13