TOPIC 2

Population and sample, their characteristics. Estimation of the true values of measured value

Within the physical sciences there are many problems which may have an exact answer, but in the life sciences many of the questions asked may not have a fixed answer.

Biostatistics is the study of statistics as applied to biological areas. Biological laboratory experiments, medical research (including clinical research), and health services research all use statistical methods. Many other biological disciplines rely on statistical methodology.

Technological advances continually make new disease prevention and treatment possibilities available for health care. Consequently, a substantial body of medical research explores alternative methods for treating diseases or injuries. Because outcomes vary from one patient to another, researchers use statistical methods to quantify uncertainty in the outcomes, summarize and make sense of data, and compare the effectiveness of different treatments.

A course in introductory biostatistics is often required for professional students in public health, dentistry, nursing, and medicine, and for graduate students in nursing and other biomedical sciences, a requirement that is often considered a roadblock.

In this chapter we'll study basic notion of biostatistics.

2.1. Important definition

Probability of an event can be expressed as a ratio of the number of likely outcomes to the number of possible outcomes. It is denoted by *p* and must be between 0 and 1. Probability of an event not occurring is equal *q=1-p*.

Types of data. Choice of statistical technique depend on the type of data. Data will be from one of 4 scales of measurement: ***nominal*** (is divided into qualitative categories or groups – male/female, black/white), ***ordinal*** (data can be placed in a meaningful order – $1^{-st}/2^{-nd}/3^{rd}$), ***interval*** (have equidistant points between each of the scale elements – temperature scale), ***ratio*** (The factor which clearly defines a ratio scale is that it has a true zero point - the Centigrade scale has a zero point but it is an arbitrary one. The Fahrenheit scale has its equivalent point at -32^{o}.) Data may also be characterized as ***discrete*** (can take only certain values and none in between – the number of syringes used in a clinic on any given day may increase or decrease only by units of one) or ***continuous*** (may take any value, most biomedical variables are continuous – blood pressure).

A **variable** is a quantity that may vary from object to object.

A **sample** (or data set) is a collection of values of one or more variables. A member of the sample is called an element.

The **sample space** or **population** is the set of all possible values of a variable.

The **size of a sample** is the number of elements in the sample.

A **statistic** is a numerical characteristic of a sample.

A **parameter** is a numerical characteristic of a population.

2.2. Data Distributions

Any individual data value belongs to or originates from some population of values that has certain properties that are collectively designated as a **distribution**.

Distributions describe the frequency of occurrence of individual data values about some specified central value: a **mean or median** as defined below. Distributions are characterized by a **probability distribution function**, which is expressed in terms of distribution parameters that relate to the central tendency and the dispersion of values about this central value.

The **normal or Gaussian distribution** has a typical bell-shaped symmetrical frequency of occurrence below and above the mean. **Nonnormal distributions** have a nonsymmetric occurrence frequency, and such distributions can be made to approach normality by an averaging process.

2.3. Characterizing distributions (descriptive statistics)

A database that is presumed to have a normal distribution may be characterized by two types of statistical parameters: one that establishes its central value (these measures include mean, mode and median) and one that characterizes the spread or dispersion of values around the central value.

Two important characteristics of any distribution are the "center" and the variability.

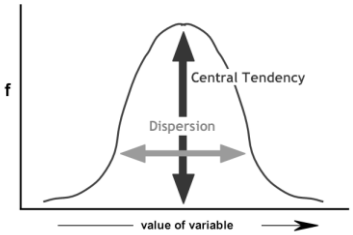

Fig. 2.1 Characteristics of any distribution

The **mean** (frequently known as the **arithmetic average**), **variance**, and **standard deviation** can be realized in two ways: 1) as a true parameter value based on extensive measurement or other knowledge of the **entire population**, in which case these parameters are designated by the symbols μ, σ^2 and σ respectively or 2) as estimates of the true values based on **samples from the population**, in which case they are designated by the symbols $\bar{x}$, s^2 and s respectively.

2.3.1. Mean

The arithmetic mean is the sum of the individual values in a data set divided by the number of values in the data set. We can compute a mean of both a finite population and a sample. For the mean of a finite population (denoted by the symbol μ), we sum the individual observations in the entire population and divide by the population size, *N*. When data are based on a sample, to calculate the sample mean (denoted by the symbol ($\bar{x}$) we sum the individual observations in the sample and divide by the number of elements in the sample, *n*. The sample mean is the sample analog to the mean of a finite population. Formulas for the population and sample means are shown below:

Population mean (μ):

$\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$, where x_i – individual value from a finite population of size *N*.

Sample mean ($\bar{x}$):

$\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$, where x_i - individual value of samples of size *n* .

Example . Calculation of Mean (*n=5*)

Index (i)	*x*
1	70
2	80
3	95
4	100
5	125
Σ	470

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{470}{5} = 94$$

Sample mean for a frequency distribution

$\bar{x} = \dfrac{\sum_{i=1}^{k} x_i f_i}{n}$, f_i- frequency of data x_i

Example . Calculation of mean for a frequency distribution (*n=10*).

Index (i)	*x*	*f*
1	70	2
2	80	2
3	95	1
4	100	3
5	125	2
Σ		10

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i f_i}{n} = \frac{70 \cdot 2 + 80 \cdot 2 + 95 + 100 \cdot 3 + 125 \cdot 2}{10} = 94.5$$

2.3.2. Mode

The **mode (Mo)** is the observed value that occurs with the greatest frequency. It is found by simple inspection of the frequency distribution.

For example, the distribution consisting of the elements 6, 9, 9, 5, 8, then mode would be 9.

2.3.3. Median

The **median** is the figure that divides the frequency distribution in half when all the scores are listed in order. When a distribution has an odd number of elements, the median is therefore the middle one; when it has an even number of elements, the median lies halfway between the two middle scores (i.e. it is the average or mean of the two middle scores).

For example, in a distribution consisting of the elements 6, 9, 15, 17, 24, the median would be 15. If the distribution were 6, 9, 15, 17, 24, 29, the median would be 16 (the average of 15 and 17).

2.3.4. Variance

When the mean *x* of a set of measurements has been obtained, it is usually matter of considerable interest to measure the degree of variation or dispersion around this mean. Are the *x*'s all rather close to *x*, or are some of them dispersed widely in each direction?

Common measures of dispersion, used more frequently because of their desirable mathematical properties, are the interrelated measures variance and standard deviation. Instead of using the absolute value of the deviations about the mean, both the variance and standard deviation use squared deviations about the mean, defined for the its observation as $(x_i-\mu)^2$.

Formula

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

which is called the deviation score method, calculates the ***population variance*** (σ^2) for a finite population. For infinite populations we cannot calculate the population parameters such as the mean and variance. These parameters of the population distribution must be approximated through sample estimates. Based on random samples we will draw inferences about the possible values for these parameters.

The ***sample variance*** is defined as the sum of the squared deviations from the mean, divided by *n-1*.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}.$$

The use of the divisor (*n - 1*) instead of *n* is clearly not very important when *n* is large. It is more important for small values of *n*.

The **variance** is measured in the square of the units in which the *x*'s are measured. ***For example***, if *x* is the time in seconds, the variance is measured in seconds squared (sec^2). It is convenient, therefore, to have a measure of variation expressed in the same units as the *x*'s, and this can be done easily by taking the square root of the variance. This quantity is the ***standard deviation.***

Example.

Calculation of Population Variance

Suppose we have a small finite population (*N = 5*), with the following blood sugar values: 70, 80, 95, 100 and 125.

Solution

x_i	$x_i-\mu$	$(x_i-\mu)^2$
70	-24	576
80	-14	196
95	1	1
100	6	36
125	31	961
Σ	0	1 770

$$\mu = \frac{\sum_{i=1}^{5} x_i}{N} = \frac{470}{5} = 94;$$

$$\sigma^2 = \frac{\sum_{i=1}^{5}(x_i-\mu)^2}{N} = \frac{1770}{5} = 354.$$

Example.

Calculation of sample variance

Blood Cholesterol Measurements for a Sample of 10 Persons 276, 304, 316, 188, 214, 252,333, 271, 245, 198.

Person	x_i	$x_i-\bar{o}$	$(x_i-\bar{o})^2$
1	276	16,3	265,69
2	304	44,3	1962,49
3	316	56,3	3169,69
4	188	-71,7	5140,89
5	214	-45,7	2088,49
6	252	-7,7	59,29
7	333	73,3	5372,89
8	271	11,3	127,69
9	245	-14,7	216,09
10	198	-61,7	3806,89
Σ	2597	0	22210,1

Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2597}{10} = 259,7.$$

Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{22210,1}{9} = 2467,789.$$

2.3.5 Standard deviation

This is the most commonly used measure of the spread or dispersion of data around the mean.

A related term is the ***population standard deviation (σ),*** which is the square root of the variance:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}.$$

The ***sample standard deviation*** is defined as the square root of the *sample variance (s^2):*

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

Sample standard deviation for a frequency distribution is:

$$s = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{n-1}}.$$

2.3.6. Degrees of freedom

The ***degrees of freedom*** is the number of independent differences available for estimating the variance or standard deviation from a set of data; it is one less than the total number of data values *n*, since one degree of freedom (of the total degrees equal to *n*) is used to estimate the mean. The degrees of freedom (*df*) for sample variance and standard deviation is (*n - 1*).

2.3.7. Relative standard deviation. Coefficient of variation

Although the standard deviation of analytical data may not vary much over limited ranges of such data, it usually depends on the magnitude of such data: the larger the figures, the larger *s*. Therefore, for comparison of variations (e.g. precision) it is often more convenient to use the *relative standard deviation* (RSD) than the standard deviation itself. The RSD is expressed as a fraction, but more usually as a percentage and is then called *coefficient of variation* (CV).

Often, however, these terms are confused.

Sample: $RSD = \frac{s}{\bar{x}}$

$CV = \frac{s}{\bar{x}} \cdot 100\%.$

It is an index, a dimensionless quantity because the standard deviation is expressed in the same units as the mean and could be used to compare the difference in variation between two types of measurements.

Population: $RSD = \frac{\sigma}{\mu}$

$CV = \frac{\sigma}{\mu} \cdot 100\%.$

2.3.8. The Sampling Distribution of the Mean

If a random sample of size *n* is drawn from a normal population having a mean μ and variance σ^2, then the mean of the *n* values in the sample $\bar{x}_n$ is a random variable whose distribution has the mean μ and a variance of σ^2/n or a standard deviation of $\sigma/\sqrt{n}$, which is frequently called the **standard error of the mean**. Note that this parameter, which establishes the reliability of the mean $\bar{x}_n$ decreases as the square root of *n*; it is necessary to quadruple the values *n* in order to hake the standard error of the mean.

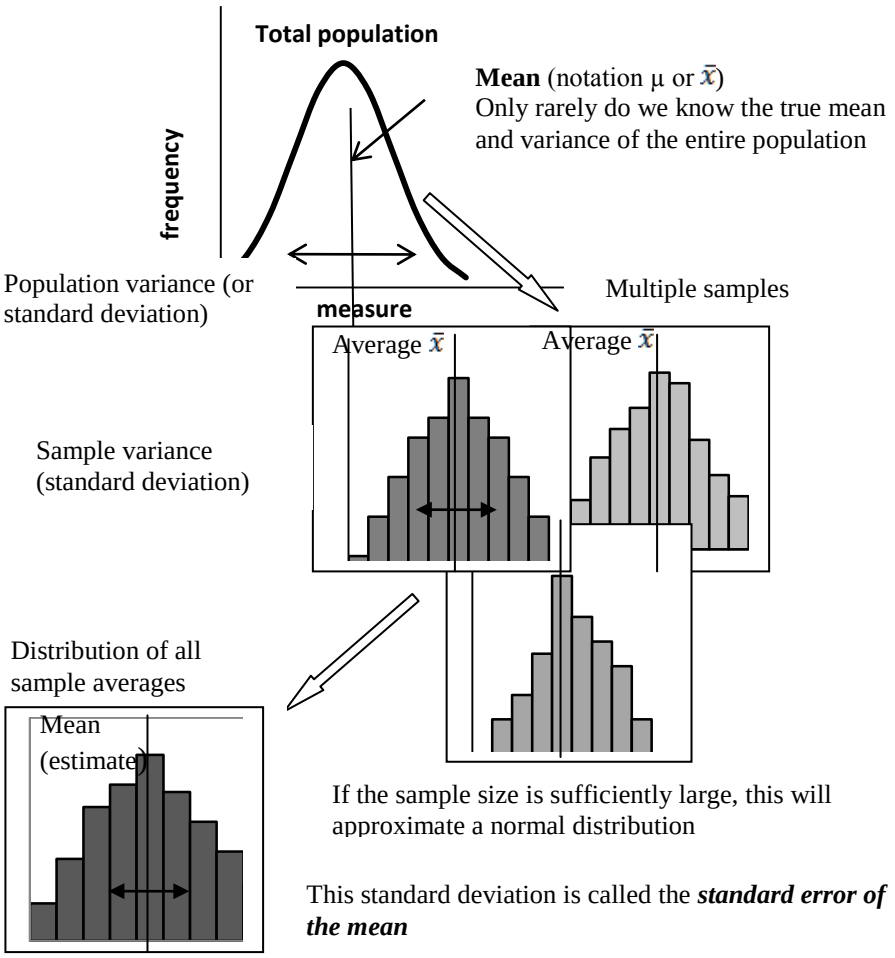

Fig. 2.2. Visualizing multiple distributions

In fig. 2.2 we see that we can sample a large population multiple times. Each of our samples has an average and a variance associated with it. If we aggregated all of the averages from these samples, we can create a sampling distribution of the mean for the population. The standard deviation of this distribution is more properly known as the **standard error of the mean** (SEM or, simply, the standard error, SE).

2.4. Normal distribution

When information from a large population is examined it will be found that there will be many deviations from the mean. Both positive and negative deviations will occur with nearly the same frequency. Also small deviations will occur more frequently than large deviations.

The normal distribution was discovered first by the French mathematician Albert DeMoivre in the 1730s. Gauss found that the normal distribution with a mean of zero was often a useful model for characterizing measurement errors. In the 1890s in England, Sir Francis Galton found applications for the normal distribution in medicine.

The normal distribution is determined by two parameters: the ***mean*** and the ***variance.***

The fact that the mean and the variance of the normal distribution are the natural parameters for the normal distribution explains why they are sometimes preferred as measures of location and scale.

The normal distribution has ***three main characteristics***.

First, its probability density is bell-shaped, with a single mode at the center. As the tails of the normal distribution extend to $\pm\infty$, the distribution decreases in height and remains positive. It is symmetric in shape about μ, which is both its mean and mode. ***For a normal distribution the mean, median, and mode are all equal to one another***.

Another parameter, σ, along with the mean, completes the characterization of the normal distribution. The relationship between σ and the area under the normal curve provides the ***second*** main characteristic of the normal distribution. The parameter σ is the standard deviation of the distribution. Its square is the variance of the distribution.

For a normal distribution, 68.26% of the probability distribution falls in the interval from $\boldsymbol{\mu - \sigma}$ to $\boldsymbol{\mu + \sigma}$. The wider interval from $\boldsymbol{\mu - 2\sigma}$ to $\boldsymbol{\mu + 2\sigma}$ contains 95.45% of the distribution. Finally, the interval from $\boldsymbol{\mu - 3\sigma}$ to $\boldsymbol{\mu + 3\sigma}$ contains 99.73% of the distribution, nearly 100% of the distribution. The fact that nearly all observations from a normal distribution fall within $\pm 3\sigma$ of the mean explains why the ***three-sigma limits*** are used so often in practice.

Third, a complete mathematical description of the normal distribution can be found in the equation for its density. The ***probability density function f(x)*** for a normal distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

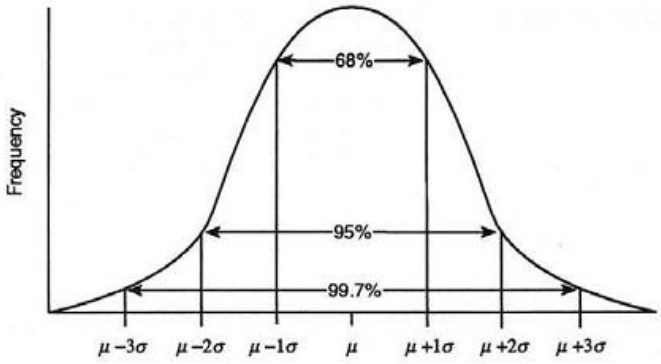

Fig. 2.3. The standard normal distribution

Note. Because these proportions hold true for every normal distribution, they should be memorized.

Example.

Therefore, if population's resting heart is normally distributed with a mean μ of 70 and a standard deviation *S* of 10, the proportion of the population that has a resting heart rate between certain limits can be stated.

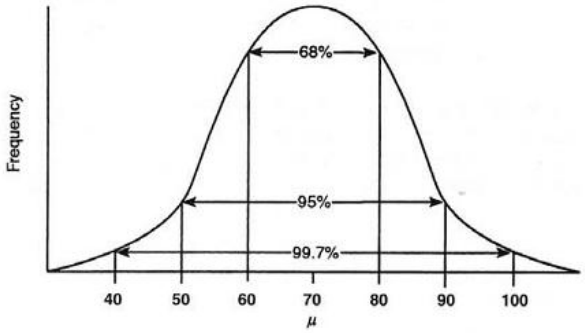

Fig. 2.4. Normal distribution of a resting heart rate

Fig. 2.4 shows, because 68% of the distribution lies within approximately ±1 standard deviations of the mean, 68% of the population will have a resting heart rate between 60 and 80 beats/min.

Similarly, 95% of the population will have a heart rate between approximately 70 ± (2 x 10) = 50 and 90 beats/min (i.e., within 2 standard deviations of the mean).

The unique bell shape of the normal distribution curve may be characterized by an equation called the **normal probability density function**, which gives the probability of finding a given distribution value as a function of that value. To avoid having a separate equation for each measured parameter with its unique units, the function is adjusted to make the area under the distribution curve equal to one or unity, and this adjusted equation is called the **standard normal distribution**.

The results of such calculations are given as tabular values.

2.5. Student's distribution (t-distribution)

Student's distribution arises when the population standard deviation is unknown and has to be estimated from the data.

Student's t-distribution is the probability distribution of the ratio

$$t = \frac{\bar{x}_n - \mu}{s/\sqrt{n}}.$$

t is a random variable.

Student's distribution is often referred to just as the ***t*- distribution**.

The *t*-distribution is similar in shape to the normal, it has an expected mean of zero, but its variance depends on the degrees of freedom *df* associated with *s*, which is equal to (*n - 1*). As *n* approaches infinity, the *t*-distribution approaches the normal distribution and its variance approaches 1.

Tables of *t* values, designated as t_γ, at selected *df* (degrees of freedom) are given for various probabilities - 0.95,

0.99, 0.999 etc. and usually called a ***critical t or t_γ.***

As illustrated above, the *t*-distribution has many properties which differentiate it from the standard normal distribution.

1. The Student *t*-distribution is different for different sample sizes.
2. The Student *t*-distribution is generally bell-shaped, but with smaller sample sizes shows increased variability (flatter). In other words, the distribution is less peaked than a normal distribution and with thicker tails. As the sample size increases, the distribution approaches a normal distribution. For $n > 30$, the differences are negligible.
3. The mean is zero (much like the standard normal distribution).
4. The distribution is symmetrical about the mean.
5. The variance is greater than one, but approaches one from above as the sample size increases ($\sigma^2=1$ for the standard normal distribution).
6. The population standard deviation is unknown.
7. The population is essentially normal (unimodal and basically symmetric).

2.6. Confidence limits of a measurement

Whenever a mean is calculated there should be an estimate of variability with it, since to appreciate the mean fully we need to know how confident we can be that the population's true mean lies close to this value. If the SEM is given, we can estimate a **confidence interval** for the mean. A **confidence interval** gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. We saw for normal distribution that there is a 68% chance of finding the true mean within one standard error of the mean. This range is therefore called the 68% confidence interval.

In practice 90%, 95%, and 99% intervals are often used, with ***95% being the most commonly used.***

The end points of the confidence interval are referred to as **confidence limits.** Interval estimates are often desirable because the estimate of the mean varies from sample to sample. Instead of a single estimate for the mean, a confidence interval generates a l*ower* and *upper* limit for the mean. The interval estimate gives an indication of how much uncertainty there is in our estimate of the true mean.

The narrower the interval, the more precise is our estimate.

Confidence limits are defined as:

$$\bar{x} \pm t_\gamma \cdot \frac{s}{\sqrt{n}}$$

where

$\bar{x}$ - mean of subsamples

t_γ - critical value of the *t*-distribution with $n - 1$ degrees of freedom

s - standard deviation of mean of subsamples

n - number of subsamples

The critical values for *t* are tabulated.

Table 1. Critical values for *t*

degrees of freedom	Confidence		
	0.95	**0.99**	**0.999**
1	12.70620	63,65674	636,6192
2	4.30265	9,92484	31,5991
3	3.18245	5,84091	12,9240
4	2.77645	4,60409	8,6103
5	2.57058	4,03214	6,8688
6	2.44691	3,70743	5,9588
7	2.36462	3,49948	5,4079
8	2.30600	3,35539	5,0413
9	2.26216	3,24984	4,7809
10	2.22814	3,16927	4,5869
11	2.20099	3,10581	4,4370
12	2.17881	3,05454	4,3178
13	2.16037	3,01228	4,2208
14	2.14479	2,97684	4,1405
15	2.13145	2,94671	4,0728
16	2.11991	2,92078	4,0150
17	2.10982	2,89823	3,9651
18	2.10092	2,87844	3,9216
19	2.09302	2,86093	3,8834
20	2.08596	2,84534	3,8495
21	2.07961	2,83136	3,8193
22	2.07387	2,81876	3,7921
23	2.06866	2,80734	3,7676
24	2.06390	2,79694	3,7454
25	2.05954	2,78744	3,7251
26	2.05553	2,77871	3,7066

27	2.05183	2,77068	3,6896
28	2.04841	2,76326	3,6739
29	2.04523	2,75639	3,6594
30	2.04227	2,75000	3,6460
∞	1.95996	2,57583	3,2905

To find the applicable value, the number of *degrees of freedom* has to be established by: $df = n - 1$.

Example

Suppose that we conduct a survey of 19 millionaires to find out what percent of their income the average millionaire donates to charity. We discover that the mean percent is 15 with a standard deviation of 5 percent. Find a 95% confidence interval for the mean percent.

Solution

We use the formula:

$$\bar{x} \pm t_\gamma \cdot \frac{s}{\sqrt{n}}$$

We get

$$15 \pm t_\gamma \cdot \frac{5}{\sqrt{19}}$$

Since n = 19, there are 18 degrees of freedom. Using the table (Critical values for *t)*, we have that

$t_\gamma = 2.10$

Hence the margin of error is

$$\pm 2.10 \cdot \frac{5}{\sqrt{19}} = \pm 2.4$$

We can conclude with 95% confidence that the millionaires donate between 12.6% and 17.4% of their income to charity.

Exercises

Independent work in the class

1. Find the medians of the following data sets: {8, 7, 3, 5, 3}; {7, 8, 3, 6, 10, 10}.

2. A sample of data was selected from a population: {195, 179, 205, 213, 179,216, 185, 211}. Calculate variance and standard deviations.

3. In a sample of 25 experimental subjects, the mean score on a post experimental measure of aggression was 55 with a standard deviation of 5. Construct a 95% confidence interval for the population mean.

4. Suppose that a sample of pulse rate gives a mean of 71.3, with a standard deviation that can be assumed to be 9.4. How many patients should be sampled to obtain a 95% confidence interval for the mean that has half-width 1.2 beats per minute?

5. The standard hemoglobin reading for normal healthy adult males is 15 g/100ml. The standard deviation is about 2.5 g/100 ml. For a group of 26 male construction workers, the sample mean was 16 g/100 ml.

a) Construct a 95% (99%) confidence interval for the male construction workers. What is your interpretation of this interval relative to the normal adult male population?

b) What would the confidence interval have been if the above results were obtained based on 19 construction workers?

c) Repeat b for 14 construction workers.

d) Do fixed-level confidence intervals shrink or widen as the sample size increases (all other factors remaining the same)? Explain your answer.

e) What is the half-width of the confidence interval that you would obtain for 14 workers?

6. The mean diastolic blood pressure for 225 randomly selected individuals is 75 mmHg with a standard deviation of 12.0 mmHg. Construct a 99% confidence interval for the mean.

Homework

Exercises

1. Assume you have the following datasets for a sample: {3, 3, 3, 3, 3}; {5, 7, 9,11}; {4, 7, 8}; {33, 49}

a) Compute *s* and s^2*;*

b) Describe the results you obtained.

2. The following cholesterol levels of 10 people were measured in mg/dl: {260, 150, 165, 201, 212, 243, 219, 227, 210, 240}. For this sample:

a) Calculate the mean and median.

b) Calculate the variance and standard deviation.

c) Calculate the coefficient of variation.

3. Suppose we randomly select 20 students enrolled in an introductory course in biostatistics and measure their resting heart rates. We obtain a mean of 66.9 (s = 9.02). Calculate a 95% confidence interval for the population mean and give an interpretation of the interval you obtain.

4. Suppose the sample size is 16 and the mean score is 55 with a standard deviation of 5. Construct a 99% confidence interval for the population mean.

5. Suppose you want to construct a 95% confidence interval for mean aggression scores, and you can assume that the standard deviation of the estimate is 5. How many experimental subjects do you need for the half-width of the interval to be no larger than 0.4?
6. The mean weight of 100 men in a particular heart study is 61 kg with a standard deviation of 7.9 kg. Construct a 95% confidence interval for the mean.
7. The mean diastolic blood pressure for 225 randomly selected individuals is 75 mmHg with a standard deviation of 12.0 mmHg. Construct a 95% confidence interval for the mean.

Mathematical treatment of direct and indirect measurements

3.1. Basics of measurements

Measurement - assignment of numerals to represent physical properties.
There are two types of measurements for data

- *Qualitative* - non-numerical or verbally descriptive also have 2 types:
  - Nominal - no order or rank (for example: list);
  - Ordinal - allows for ranking but differences between data is meaningless (for example: alphabetical list).
- *Quantitative* - numerical ranking also have 2 types:
  - Interval - meaningless comparison (for example: calendar);
  - Ratio - based on fixed or natural zero point (for example weight, pressure, Kelvin).

Categories of measurement

- *Direct Measurement:* A process of obtaining the measurement of some entity by reading a measuring tool (for example*:* a ruler for length, a scale for weight, or a protractor for angle size, etc).
- *Indirect Measurement:* method of measurement in which the value of a quantity is obtained from measurements made by direct methods of measurement of other quantities linked to the measured by a known ***relationship ;***is a technique that uses proportions to find a measurement when direct measurement is not possible to obtain or is danger (for example*:* if the length and width of a rectangle are multiplied to find the area of that rectangle, then the area is an indirect measurement).

Error analysis

Experience shown that no measurement, however carefully made, can be completely free of uncertainty. **Error analysis** is the study and evaluation of uncertainty in measurement.

In science, we use the term "error" as being interchangeable with "uncertainty." As such, errors are not mistakes.

Error – normal random variation not a mistake.

If you have a no changing parameter and you measure this repeatedly the measurement will not always be precisely the same but will cluster around a mean X_o. The deviation around X_o – error term where you can assume your measurement is X_o as long is deviation is small.

Error is the collective noun for any departure of the result from the "true" value.

In general, the result of any measurement of physical quantity must include both the value itself (best value) and its error (uncertainty). The result is usually quoted in the form

$$x = x_{best} \pm \Delta x$$

where x_{best} is the best estimate of what we believe is a true value of the physical quantity and Δx is the estimate of absolute error (uncertainty).

The uncertainty discussed up is sometimes called the **absolute uncertainty, the absolute error,** or just **the error.**

To characterize the quality of a measurement we define the **relative uncertainty (the relative error), as the ratio of the uncertainty to the measurement itself**:

$$\delta = \frac{\Delta x}{x_{best}} \cdot 100\%.$$

Measurement errors may be classified as either ***random*** or ***systematic***, depending on how the measurement was obtained**.**

An instrument could cause a random error in one situation and a systematic error in another.

Random Error refers to the spread in the values of a physical quantity from one measurement of the quantity to the next, caused by random fluctuations in the measured value. This type of error also affects the *precision* of the experiment.

Systematic Error refers to an error which is present for every measurement of a given quantity; it may be caused by a bias on the part of the experimenter, a miscalibrated or even faulty measuring instrument, etc. Systematic errors affect the *accuracy* of the experiment.

Random errors are statistical fluctuations (in either direction) in the measured data due to the precision limitations of the measurement device. *Random errors can be evaluated through statistical analysis and can be reduced by averaging over a large number of observations*.

Systematic errors are reproducible inaccuracies that are consistently in the same direction. These errors are difficult to detect and cannot be analyzed statistically. If a systematic error is identified when calibrating against a standard, applying a correction or correction factor to compensate for the effect can reduce the bias. Unlike random errors, *systematic errors cannot be detected or reduced by increasing the number of observations*.

Accuracy and Precision

When one considers the quality of a measurement there are two aspects to consider. The first is if one were to repeat the measurement, how close would new results be to the old, i.e., how reproducible is the measurement? Scientists refer to this as the ***precision of the measurement***.

Secondly, a measurement is considered "good" if it agrees with the true value. This is known as the ***accuracy of the measurement***. But there is a potential problem in that one needs to know the "true value" to determine the accuracy.

A good measurement must be close to the "true value" and be reproducible.

Accuracy and Precision

These two words do not mean the same thing.

"Accuracy" deals with how close is a measured value to an accepted or "true value.

"Precision" deals with how reproducible is a given measurement.

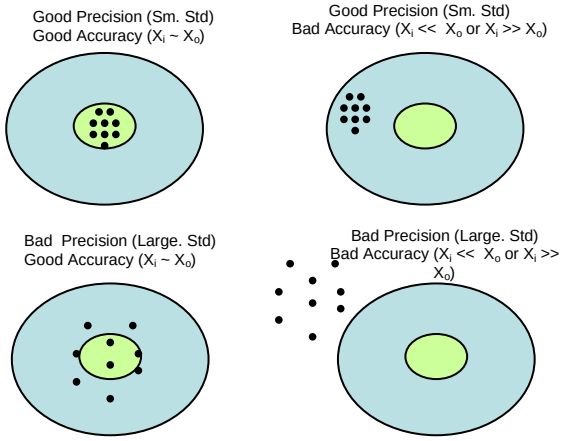

Fig. 3.1

Tactics to decrease error on practical measurements:

1. Make Measurements several Times.
2. Make Measurements on Several Instruments.
3. Make successive Measurements on different parts of instruments (different parts of ruler).

3.2. Error analysis for direct measurement

A better way to estimate uncertainty is to make multiple measurements of the same quantity and *analyze the dataset using statistical functions*.

Suppose we make n measurements of a quantity x and get the values $x_1, x_2,..., x_n$.*(the values of each measurement being denoted by x_i, where i takes on the values from 1 to n)*

Estimates of direct measured value.

Either because of variability in the quantity x or because of inherent and unavoidable random errors in our measuring procedure, not all values of the individual measurements x_i will be the same. Our best estimate of the "***true***" value of x is then given by the ***average*** or ***mean*** value of x:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Measured value x_i (5 measurements):
71, 72, 72, 73, 71.

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5}(71+72+72+73+71) = 71.8.$$

The best estimate of the uncertainty in the individual values x_i is the sample ***standard deviation*** s (or SD), defined as:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} d_i^{\,2}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The term d_i in this equation, called the deviation, is simply the difference between the i^{th} measurement x_i and the mean value x.

If the deviations are all very small, then our measurements are all close together and are said to be precise.

If the deviations of a measurement were averaged, the result would be zero because of high and low values would cancel each other.

This is why the standard deviation is found by first squaring the deviations, then averaging these positive squares (using *n-1* rather than *n*), and finally taking the square root of the result.

For the five previous measurements, the standard deviation *s* is found to be:

Measurement number, i	*Measured value,* x_i	*Deviation* $d = x_i - \bar{x}$
1	71	-0.8
2	72	0.2
3	72	0.2
4	73	1.2
5	71	-0.8
	$\bar{x} = 71.8$	$\bar{d} = 0$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{4}(0.64 + 0.04 + 0.04 + 1.44 + 0.64)} \approx 0.84$$

When we report the average value of *n* measurements, the uncertainty we should associate with this average value is the ***standard deviation of the mean***, often called the ***standard error*** (SE).

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The standard error is smaller than the standard deviation by a factor of $\frac{1}{\sqrt{n}}$.

While the standard deviation indicates the amount of variation of the data about the mean, the standard error expresses how much the mean of n measurements would be expected to vary if the entire n measurements were repeated again.

For our measurements standard deviation of the mean is

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0.84}{\sqrt{5}} \approx 0.38.$$

Thus, our final answer to the question "what is our true measured value of *x*?" is

$$x = \bar{x} \pm \frac{s}{\sqrt{n}} = 71.8 \pm 0.38.$$

Since uncertainty is related statistically to the width of a bell-shaped normal distribution, we can clarify the meaning of such a report by associating a confidence level with the uncertainty:

$$\mu = \bar{x} \pm t_\gamma \cdot \frac{s}{\sqrt{n}} = \bar{x} \pm \Delta x$$

The confidence is defined to be the probability, stated as a percentage, that the ***"true"*** mean value actually falls within the limits mean ± uncertainty.

$$\bar{x} - t_\gamma \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_\gamma \cdot \frac{s}{\sqrt{n}}$$

$$\bar{x} - \Delta x \le \mu \le \bar{x} + \Delta x$$

t_γ - coefficient from the standard normal distribution table at given degree of freedom (*df=n-1*) (is tabulated.)

Δx is ***absolute error of measurement.***

95% confidence limits in our example is

$$71.8 \pm 2.775 \cdot 0.38 = 71.8 \pm 1.05$$

Absolute error

$\Delta x = 1.05.$

Relative error

$$\delta = \frac{1.05}{71.8} \cdot 100\% \approx 1.46\%.$$

***Example* 1.** For the determination of the clay content in the particle-size analysis, a semi-automatic pipette installation is used with a 20 mL pipette. This volume is approximate and the operation involves the opening and closing of taps. Therefore, the pipette has to be calibrated, i.e. both the accuracy (trueness) and precision have to be established.

A tenfold measurement of the volume yielded the following set of data (in mL):

19.441	19.812	19.829	19.828	19.742
19.797	19.937	19.847	19.885	19.804

Solution

The mean is 19.842 mL and the standard deviation 0.0627 mL. According to table for $n = 10$ is $t_\gamma = 2.26$ *(df = 9)* . pipette volume = $19.842 \pm 2.26\ (0.0627/\sqrt{10}\ )$ = **19.84 ± 0.04 mL**

Note that the pipette has a systematic deviation from 20 mL as this is outside the found confidence interval.

Example 2. For samples of 10 paired measurements, the mean difference $\left(\bar{d}\right)$ is 21.0, and the variance $\left(s^2\right)$ is 250. What are the 95% confidence limits for the difference between the population means $\left(\mu_1 - \mu_2\right)$?

Solution

Standard error of the mean difference is $s_{\bar{d}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{250}{10}} = 5.0$

Degrees of freedom according to table for $n = 10$ is $t_\gamma = 2.262$ *(df=9,γ=0.95)* .

Confidence limits are

$$\bar{d} \pm t_\gamma s_{\bar{d}} = 21.0 \pm 2.262 \cdot 5.0 = 21.0 \pm 11.3$$

[9.7; 32.3]

3.3. Error analysis for indirect measurement

In the majority of experiments the quantity of interest is not measured directly, but must be calculated from other quantities. Such measurements are called *indirect*. As you know by now, the quantities measured directly are not exact and have errors associated with them. While we calculate the parameter of interest from the directly measured values, it is said that the errors of the direct measurements propagate. This section describes how to calculate errors in case of indirect measurements.

Suppose we have measured the value of a quantity x with an uncertainty, which we denote Δx. In order to test a theoretical formula, suppose that we need to calculate y as function of x; i. e., $y = f(x)$. We want to know the uncertainty in y due to the uncertainty in the value of x. The answer comes from the differential calculus: if $y = f(x)$ and Δx is small, then

$$\Delta y \approx \frac{dy}{dx}\Delta x = \frac{df}{dx}\Delta x$$

In the case where f depends on two or more variables ($f(x, y)$), we have:

$$\Delta f \approx df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy$$

Taking the square and the average, we get the law of propagation of uncertainty:

$$(df)^2 = \left(\frac{\partial f}{\partial x}\right)^2 (dx)^2 + \left(\frac{\partial f}{\partial y}\right)^2 (dy)^2 + 2\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)\overline{dxdy}$$

If the measurements of x and y are uncorrelated, then $\overline{dxdy} = 0$, and using the definition of s, we get:

$$s_{\bar{f}} = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 (s_{\bar{x}})^2 + \left(\frac{\partial f}{\partial y}\right)^2 (s_{\bar{y}})^2}\ .$$

Example

$f=x+y$

(x,y are measured direct)

Standard deviation of the mean of ***f*** is

$$s_{\bar{f}} = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 (s_{\bar{x}})^2 + \left(\frac{\partial f}{\partial y}\right)^2 (s_{\bar{y}})^2} = \sqrt{\left(\frac{\partial (x+y)}{\partial x}\right)^2 (s_{\bar{x}})^2 + \left(\frac{\partial (x+y)}{\partial y}\right)^2 (s_{\bar{y}})^2} = \quad \text{as} \left(\frac{\partial (x+y)}{\partial x} = 1 \quad \frac{\partial (x+y)}{\partial y} = 1\right)$$

$$= \sqrt{(s_{\bar{x}})^2 + (s_{\bar{y}})^2}$$

Example 1.

The height of a cone is *H*=30cm, the radius of its base *R*=10cm. How will the volume of the cone change if we increase *H* by 3 mm and diminish *R* by 1 mm?

Solution: The volume of the cone is

$$V = \frac{1}{3}\pi R^2 H$$

We replace the change in volume approximately by the differential

$$\Delta V \approx dV = \frac{1}{3}\pi\left(\frac{\partial V}{\partial R}dR + \frac{\partial V}{\partial H}dH\right) = \frac{1}{3}\pi\left(2RHdR + R^2 dH\right) =$$

$$= \frac{1}{3}\pi\left(-2\cdot 10\cdot 30\cdot 0.1 + 100\cdot 0.3\right) = -10\pi \approx -31.4\ cm^2$$

Example 2.

A cylinder radius and height have been measured direct in 5 measurements. Its mean radius and height are: $\bar{r}$ =3cm; $\bar{h}$ =10cm. Standard errors of means of radius and height are respectively: $s_{\bar{r}}$ =0.01 and $s_{\bar{h}}$ =0.05). Find confidence limits of a measurement for area and volume of cylinder (95 % confidence interval).

Solution.

The volume and surface area of the cylinder are given by

$$V = \pi r^2 h \qquad S = 2\pi r^2 + 2\pi r h$$

The mean of volume

$$\bar{V} = \pi \bar{r}^2 \bar{h} = 282.6$$

I. Standard error of the mean of volume is

$$s_{\bar{v}} = \sqrt{\left(\frac{\partial V}{\partial r}s_{\bar{r}}\right)^2 + \left(\frac{\partial V}{\partial h}s_{\bar{h}}\right)^2} = \sqrt{(2\pi\bar{r}h)^2 s_{\bar{r}}^2 + (\pi\bar{r}^2)^2 s_{\bar{h}}^2} =$$

$$= \sqrt{3.55 + 1.99} = 2.35$$

Degrees of freedom according to table for *n* = 5 is t_γ = 2.775*(df = 4)* .

Absolute error

$$\Delta V = t_\gamma \cdot s_{\bar{v}} = 2.775\cdot 2.35 = 6.52$$

Value of volume is

$$V = \bar{V} \pm \Delta\bar{V} = 282.6 \pm 6.52\ \text{cm}^3$$

Relative error

$$\delta = \frac{\Delta\bar{V}}{\bar{V}} = \frac{6.52}{282.6}\cdot 100\% = 2.31\%$$

The mean of surface

$$\bar{S} = 2\pi\bar{r}^2 + 2\pi\bar{r}\bar{h} = 56.52 + 188.4 = 244.92$$

II. Standard error of the mean of area is

$$s_{\bar{s}} = \sqrt{\left(\frac{\partial S}{\partial r}s_{\bar{r}}\right)^2 + \left(\frac{\partial S}{\partial h}s_{\bar{h}}\right)^2} = \sqrt{(4\pi r + 2\pi h)^2 s_{\bar{r}}^2 + (2\pi r)^2 s_{\bar{h}}^2} =$$

$$= \sqrt{1.01 + 0.89} = 1.38$$

Degrees of freedom according to table for *n* = 5 is t_γ = 2.775*(df = 4)* .

Absolute error

$$\Delta\bar{S} = t\cdot s_{\bar{s}} = 2.775\cdot 1.38 = 3.83$$

Value of surface is

$$S = \bar{S} \pm \Delta\bar{S} = 244.92 \pm 3.83\ \text{cm}^2.$$

Relative error

$$\delta = \frac{\Delta\bar{S}}{\bar{S}} = \frac{3.83}{244.92}\cdot 100\% = 1.56\%.$$

Exercises

Independent work in the class

1. (Direct measurement)

In a quality control test conducted by a local factory, a random sample of 16 plastic bearings being produced by an injection machine was taken and their diameter measured in cm.

1.004	1.001	0.998	1.002	1.001	1.000	1.010	1.003

0.999	1.006	1.000	0.998	1.003	1.005	1.001	1.001

a) find the sample mean $\bar{x}$ and sample variance s^2 of diameter.
b) find a 95 % confidence interval for the man diameter μ
c) find a 99 % confidence interval for the man diameter μ

***2. (Indirect measurement*)**

5 tablet were weighted (*m*), then were measured their thickness (*h*) and diameter (*d*). Results are in table.

m,g	0.338	0.390	0.387	0.389	0.388
h,cm	0.53	0.54	0.53	0.55	0.54
d, cm	0.92	0.92	0.93	0.93	0.91

1. Find the density of tablet $\rho = \frac{4m}{\pi d^2 h}$.
2. Find a 95 % confidence limits for the man density.

Homework

Exercises

1. A solid body's density is calculated by a formula: $\rho = \frac{m}{V}$. Where *V* is the volume of the body and m is the mass of the body. A cylinder's density is calculated by a formula: $\rho = \frac{4m}{\pi d^2 l}$. Where *d* is the diameter of the cylinder and *l* is its length. Find relative error for density, if d_1 = 0,01282; d_2 = 0,01283; d_3 = 0,01283; d_4 = 0,01283; d_5 = 0,01281, $m \pm \Delta m = 0{,}08994 \pm 0{,}00002, kg;\ l \pm \Delta l = 0{,}0547 \pm 0{,}00005, m;\ \bar{\rho} = 5790$ kg/m^3.

2. We have an ideal gas equation, *pV = nRT*. We perform an isothermal experiment (T = const.) and measure the change of volume (V) with pressure (p), so we need to find the error ΔV. Our 'experimental' equation takes the form: *pV=nRT*, where n and R are constants (and as a rule we do not consider any errors for physical constants, therefore $\Delta n = 0$ and $\Delta R = 0$). We have, however, errors from measuring temperature, ΔT, and pressure, Δp. Obtain ΔV.

3. Quinine sulfate concentrations in tablets (m = 0.25 g) detected by means of spectrophotometer analysis ($\lambda = 234nm$) are: 99.9%; 99.8%; 99.6%; 99.1%; 99.2%; 99.2%. Find mean, absolute and relative errors (95 % confidence interval).

4. Resistance R_i measurement gives results: 6.270 ohm; 6.273;6.277; 6.271; 6.276; 6.272; 6.278; 6.275; 6.277; 6.274. Find relative error for resistance (95 % confidence interval).

5. We can find ethanol viscosity coefficient using formula $\eta = \mathrm{At}/t_0$, where A=0.001 Pa·s, *t* and t_0 the flowing time of ethanol and reference liquids (water) measured with the same viscometer. 5 measurement gives: t-6.2; 6.4; 6.4; 6.2; 6.3 s; t_0 – 4.1; 4.1; 4.0; 4.0; 3.8 s. Find standard error of the mean of viscosity coefficient and absolute and relative errors for ethanol viscosity coefficient (95 % confidence interval).

6. The volume of cylinder is given by formula $V = \frac{\pi h d^2}{4}$**.** A cylinder radius and height have been measured direct in 10 measurements. Its mean diameter and height are: $\bar{d}$ =30 mm; $\bar{h}$ =50 mm. Standard errors of means of diameter and height are respectively: $s_{\bar{d}}$ =0.1 mm and $s_{\bar{h}}$ =0.1 mm). Find standard errors of means of cylinder volume. Obtain absolute and relative errors (95 % confidence interval).

7. 10 measurement of lungs capillary diameter gives results: 2.83 mcm; 2.82; 2.81; 2.85; 2.87; 2.86; 2.83; 2.85; 2.83; 2.84 mcm. Find standard errors of means of capillary diameter. Obtain absolute and relative errors (95 % confidence interval).

8. Solution concentration can be calculated using formula $C_x = C_0(d_0/d_x)$, where d_0, d_x layers thickness. 5 measurements give the following results: d_o =5.7 mm, $s_{\bar{d}_0}$ =0.15 mm, d_x =8.5 mm, $s_{\bar{d}_x}$ =0.18 mm. Solution concentration C_0=2%. Find absolute and relative solution concentration errors (95 % confidence interval).

9. Microanalytical method show oxygen concentration 9.29;%, 9.38; 9.35; 9.43; 9.53; 9.48; 90.61; 9.68%. Find a 95 % confidence limits for the concentration. Calculate absolute and relative concentration errors.

10. 10 equals probes give sodium concentration: 1%; 1.05; 1.1; 0.99; 0.97; 0.98; 1.08; 1.07; 1.01; 1.03%. Find a 95 % confidence limits for the concentration.