Biostatistics

Th1

Descriptive Statistics

Carina Silva
(*carina.silva@estesl.ipl.pt*)
Mestrado em Farmácia

ESCOLA SUPERIOR DE
TECNOLOGIA DA SAÚDE
DE LISBOA
INSTITUTO POLITÉCNICO DE LISBOA

Contents

Introduction

Why use statistical analysis?

Common statistical pitfalls

Contents

Contents

Introduction

Why use statistical analysis?

Common statistical pitfalls

Basic Concepts

Describing and summarizing data

Descriptive Statistics

Plots

Contents

Introduction

Why use statistical analysis?

Common statistical pitfalls

Basic Concepts

Describing and summarizing data

Descriptive Statistics

Plots

Multiple choice questions

Introduction

Traditional scientific research consists on four interrelated stages:

- Problem definition.

Introduction

Traditional scientific research consists on four interrelated stages:

- Problem definition.
- Data gathering.

Introduction

Traditional scientific research consists on four interrelated stages:

- Problem definition.
- Data gathering.
- Data analysis.

Introduction

Traditional scientific research consists on four interrelated stages:

- Problem definition.
- Data gathering.
- Data analysis.
- Data interpretation and conclusions.

Frequently Asked Questions

- What kind of study design should I use?

Frequently Asked Questions

- What kind of study design should I use?
- How many replicates should I use?

Frequently Asked Questions

- What kind of study design should I use?
- How many replicates should I use?
- Should I use SD or SEM?

Frequently Asked Questions

- What kind of study design should I use?
- How many replicates should I use?
- Should I use SD or SEM?
- Can I use inferential statistical analysis?

Frequently Asked Questions

- What kind of study design should I use?
- How many replicates should I use?
- Should I use SD or SEM?
- Can I use inferential statistical analysis?
- How to report statistics in a paper?

Frequently Asked Questions

- What kind of study design should I use?
- How many replicates should I use?
- Should I use SD or SEM?
- Can I use inferential statistical analysis?
- How to report statistics in a paper?
- ...

Common statistical pitfalls

- Today, statistics is widely accepted as a powerful tool in the scientific research process, with great increase in the use of statistical methods documented in a wide range of non-statistical journals.

Common statistical pitfalls

- Today, statistics is widely accepted as a powerful tool in the scientific research process, with great increase in the use of statistical methods documented in a wide range of non-statistical journals.
- Nevertheless, there is a wide consensus that standards are generally low, as a large proportion of published research contains statistical errors and shortcomings.

Common statistical pitfalls

- Today, statistics is widely accepted as a powerful tool in the scientific research process, with great increase in the use of statistical methods documented in a wide range of non-statistical journals.
- Nevertheless, there is a wide consensus that standards are generally low, as a large proportion of published research contains statistical errors and shortcomings.
- The problem is a serious one, since the inappropriate use of statistical analysis may lead to incorrect conclusions and then a waste of valuable resources are made.

Common statistical pitfalls

How to use descriptive analysis:

- If using arithmetic means and standard deviations, it should be evident, that the data is at least approximately normally distributed and not skewed. Otherwise these measures can not be used meaningfully for describing the data.

Common statistical pitfalls

How to use descriptive analysis:

- If using arithmetic means and standard deviations, it should be evident, that the data is at least approximately normally distributed and not skewed. Otherwise these measures can not be used meaningfully for describing the data.
- In every case, standard deviations should preferably be reported in parentheses [ie, mean (SD)] than using mean ± SD expressions, as the latter specification can be confused with a 95% confidence interval by the reader.

Common statistical pitfalls

How to use descriptive analysis:

- If using arithmetic means and standard deviations, it should be evident, that the data is at least approximately normally distributed and not skewed. Otherwise these measures can not be used meaningfully for describing the data.
- In every case, standard deviations should preferably be reported in parentheses [ie, mean (SD)] than using mean ± SD expressions, as the latter specification can be confused with a 95% confidence interval by the reader.
- For skewed data, as often the case in biological and medical research, giving medians, quartiles or ranges is more suitable, although one has to be aware that the range is sensitive to outliers and hence sometimes may be unsuitable as a summary statistic.

Common statistical pitfalls

- If consecutively applying nonparametric tests for statistical data analysis, one should avoid giving means and standard deviations, as these parameters are, by definition, not tested by a nonparametric test and hence do not make sense for describing the data under investigation. Here medians, ranges or interquartile-ranges should have preference.

Common statistical pitfalls

- If consecutively applying nonparametric tests for statistical data analysis, one should avoid giving means and standard deviations, as these parameters are, by definition, not tested by a nonparametric test and hence do not make sense for describing the data under investigation. Here medians, ranges or interquartile-ranges should have preference.
- The standard error of the mean (SEM), although commonly and erroneously used for statistical description (perhaps because it makes data seem less variable), is not a descriptive statistic, but rather an inferential method used for statistical estimation.

Common statistical pitfalls

Interpretation:

- If claiming significance of effects, one has to ensure, that a statistical significance test has been employed.

Common statistical pitfalls

Interpretation:

- If claiming significance of effects, one has to ensure, that a statistical significance test has been employed.
- If results do not exhibit statistical significance, it is crucial to be careful in drawing conclusions, as lack of statistical significance does not invariably mean there was no effect or no difference at all.

Common statistical pitfalls

Interpretation:

- If claiming significance of effects, one has to ensure, that a statistical significance test has been employed.
- If results do not exhibit statistical significance, it is crucial to be careful in drawing conclusions, as lack of statistical significance does not invariably mean there was no effect or no difference at all.
- For example, perhaps merely the sample size was too small to safeguard statistical significance.

P-value the Holy Grail of experimental science

As any scientist can tell you, the holy grail of an experiment is a low **p-value**, a statistical measure that tells whether your findings are indicative of an actual effect, not just randomness and chance.

P-value the Holy Grail of experimental science

A **p-value** is a measure of how much evidence we have against the null hypothesis. (The null hypothesis, traditionally represented by the symbol H0, represents the hypothesis of no change or no effect.)

The term *statistically significant* pervades the published literature. But how many times did you see an appropriate report of the results? Usually, $p < 0.05$ or worse $p > 0.05$ or $p = NS$ meaning "not statistically significant".

Basic Concepts

Statistical Analysis typically involves using data from a sample to make inferences about the characteristics of a population.

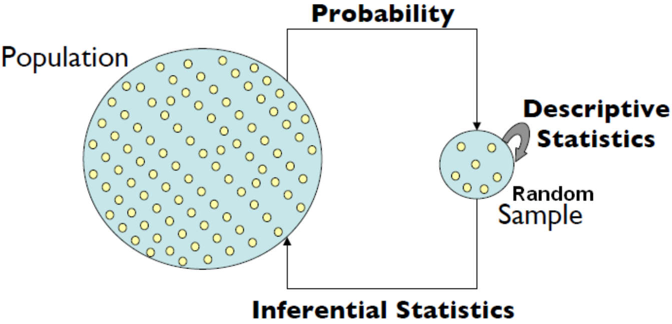
- **Population**: can be defined as including all items with the characteristic one wishes to understand (also called universe).

Basic Concepts

Statistical Analysis typically involves using data from a sample to make inferences about the characteristics of a population.

- **Population**: can be defined as including all items with the characteristic one wishes to understand (also called universe).
- **Sample**: Subset of the population, often chosen randomly and preferably representative of the population as a whole.

The ”central dogma” of inferential statistics



Exercises

Identify which sentences represent a population or a sample.

- Amount of active drug in all 20mg Prozac capsules manufactured in June 1996.
- Prior myocardial infarction status (yes or no) among 150 males aged 45 to 64 years.
- Bioavailabilities of a drugs oral dose (relative to i. v. dose) in all healthy subjects under identical conditions.
- CD4 counts of every Portuguese diagnosed with AIDS as of January 1, 1996.
- Test results (positive or negative) among 50 pregnant women taking a home pregnancy test.

Basic Concepts

- **Variable**: is a characteristic or condition that changes or has different values for different individuals.

Basic Concepts

- **Variable**: is a characteristic or condition that changes or has different values for different individuals.
- **Random Variable**: is a function that assigns a real number, $X(s)$, to each outcome s in a sample. Random Variables are represented by capital letters (e.g. X_1, X_2).

Basic Concepts

- **Variable**: is a characteristic or condition that changes or has different values for different individuals.
- **Random Variable**: is a function that assigns a real number, $X(s)$, to each outcome s in a sample. Random Variables are represented by capital letters (e.g. X_1, X_2).
- **Observation** (or case): Is a concretization of a variable. For example, the weight of a randomly chosen rat is such an observation. (Observations are represented by lowercase, e.g. x_1, x_2.)

Basic Concepts

- **Variable**: is a characteristic or condition that changes or has different values for different individuals.
- **Random Variable**: is a function that assigns a real number, $X(s)$, to each outcome s in a sample. Random Variables are represented by capital letters (e.g. X_1, X_2).
- **Observation** (or case): Is a concretization of a variable. For example, the weight of a randomly chosen rat is such an observation. (Observations are represented by lowercase, e.g. x_1, x_2.)
- **Parameter**: is a numeric quantity, usually unknown, that describes a certain population characteristic. For example, the population mean, μ, is a parameter that is often used to indicate the average value of a quantity. (Usually are represented by Greek letters, e.g. μ, σ.)

Basic Concepts

- **Estimator/Statistic**: is a statistic (that is, a function of the data) that is used to infer the value of an unknown parameter in a statistical model. Suppose there is a fixed parameter θ that needs to be estimated. An estimator of θ is usually denoted by the symbol $\hat{\theta}$. If X is used to denote a random variable corresponding to the observed data, the estimator (itself treated as a random variable) is symbolized as a function of that random variable, $\hat{\theta}(X)$. The estimate for a particular observed data set (i.e. for $X = x$) is then $\hat{\theta}(x)$, which is a fixed value.

Basic Concepts

- **Estimate/Statistic**: An estimate is an indication of the value of an unknown quantity based on observed data. More formally, an estimate is the particular value of an estimator that is obtained from a particular sample of data and used to indicate the value of a parameter. (e.g. $\bar{x}$ is an estimate of μ).

Basic Concepts

The following definitions are vital in understanding descriptive statistics:

Types of Data

Qualitative or Categorical

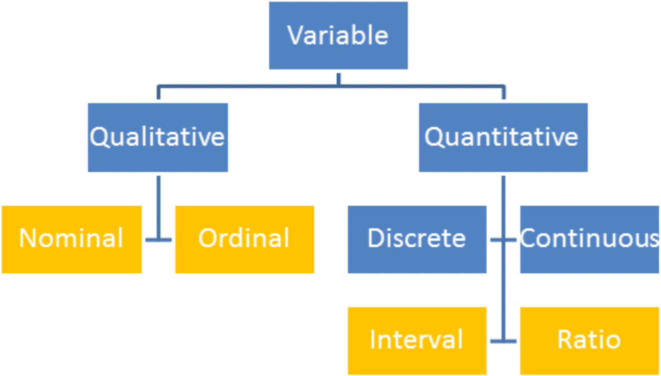
- Nominal (race, blood type, etc.)

- Ordinal (the order or rank of the categories is meaningful. For example, staff members may be asked to indicate their satisfaction with a training course on an ordinal scale ranging from "poor" to "excellent".)

Quantitative

- Discrete (number of planets orbiting a distant star. Could even be countably infinite.)

- Continuous (your age, not rounded off; weight)


Variable
Qualitative
Quantitative
Nominal
Ordinal
Discrete
Continuous
Interval
Ratio

Exercises

Classify the following variables:

- CD4 count represents numbers of CD4 lymphocytes per liter of peripheral blood.
- The amount of active drug in a 20mg Prozac capsule.
- Prior myocardial infarction status classified in yes or no. And if it is classified as number of prior MI's?
- In one of the first truly random trials in Britain, patients with pulmonary tuberculosis received either streptomycin or no drug (Medical Research Council, 1948). Patients were classified after six months into one of the following six categories: considerable improvement, moderate/slight improvement, no material change, moderate/slight deterioration, considerable deterioration, or death.

Exercises

Go to ▸ Link

Do exercises 1 and 2 from descriptive exercises folder.

Descriptive statistics are most often used to examine:

- **central tendency** (location) of data, i.e. where data tend to fall, as measured by the mean, median, and mode.

Descriptive statistics are most often used to examine:

- **central tendency** (location) of data, i.e. where data tend to fall, as measured by the mean, median, and mode.
- **dispersion** (variability) of data, i.e. how spread out data are, as measured by the variance and its square root, the standard deviation.

Descriptive statistics are most often used to examine:

- **central tendency** (location) of data, i.e. where data tend to fall, as measured by the mean, median, and mode.
- **dispersion** (variability) of data, i.e. how spread out data are, as measured by the variance and its square root, the standard deviation.
- **skew** (symmetry) of data, i.e. how concentrated data are at the low or high end of the scale, as measured by the skew index.

Descriptive statistics are most often used to examine:

- **central tendency** (location) of data, i.e. where data tend to fall, as measured by the mean, median, and mode.
- **dispersion** (variability) of data, i.e. how spread out data are, as measured by the variance and its square root, the standard deviation.
- **skew** (symmetry) of data, i.e. how concentrated data are at the low or high end of the scale, as measured by the skew index.
- **kurtosis**(peakedness) of data, i.e. how concentrated data are around a single value, as measured by the kurtosis index.

Measures of central tendency

Population Mean:

Let X denote a variable of interest (for example, blood pressure), the population mean of X is denoted by μ ($E(X) = \mu$).

Sample Mean:

For a sample with n observations, $(x_1, x_2, ..., x_n)$, the sample mean is given by:

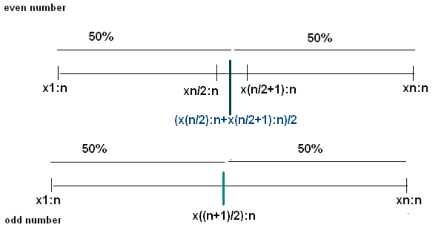
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Also the sample mean, $\bar{X}$ (see the difference on previous notation for sample mean estimate), is an estimator of the population mean μ ($\hat{\theta}(X) = \bar{X}$). When we calculate a sample mean we are only using data from a subset of the population. As a result, $\bar{x}$ simply represents our "best guess" of the true value of the population mean μ.

Measures of central tendency

Median:

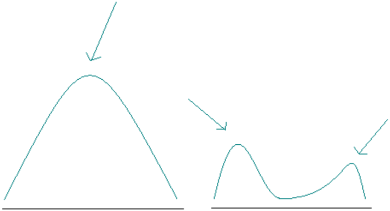
The median is the value in the middle when observations are arranged in ascending order. With an odd number of observations, the median is the middle value. With an even number of observations the median is the average of the two middle numbers. Median is often used in statistics because this value represents the exact middle of the data better than the mean.



Measures of central tendency

Mode:

The mode is the value that occurs most frequently in the data. The mode is a useful way to summarize qualitative data. For example, the mode eye color is brown, implying the most common eye color is brown.



Measures of Variability

Population Variance:

Let X denote a variable of interest the population variance of X denoted σ^2 ($VAR(X) = \sigma^2$).

Sample Variance:

For a sample with n observations, the sample variance is given by:

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

Just like $\bar{X}$ which is an estimator of the population mean μ, the sample variance, S^2, is an estimator of the population variance σ^2 ($\hat{\theta}(X) = S^2$).
Note that the variance illustrates how values of X are distributed or spread around the mean. The larger the dispersion or spread around the mean, the larger the variance.

Measures of Variability

Standard Deviation:

The standard deviation is the square root of the variance. Thus we have:

Population Standard Deviation: $\sigma = \sqrt{\sigma^2}$
Sample Standard Deviation: $s = \sqrt{s^2}$

The sample standard deviation is an estimator of the population standard deviation σ. Typically, the standard deviation is a more useful statistic than the variance because it is measured in the same units as the original data (the variance is measured in squared units).

Measures of Variability (Cont.)

Another measure of variability usually used is the **Standard Error**. Every estimator has, e.g., the **Standard Error of the Mean** is given by:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

and the **Estimated Standard Error of the Mean** is given by:

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}}$$

The **Standard Deviation** evaluates how much variability there is in a sample and the **Standard Error** evaluates how much expected variability there is in the estimated parameter.

Measures of Variability

Coefficient of variation measures the relative dispersion:

$$v = \frac{s}{\bar{x}} \times 100\%$$

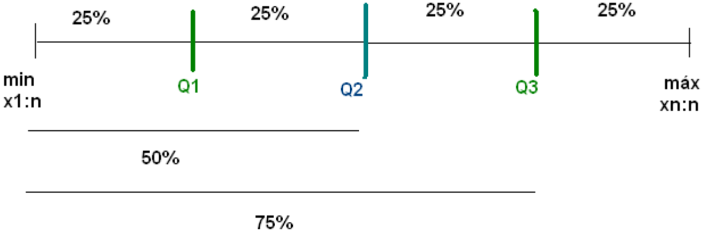
Describes the variance of two data sets better than the standard deviation.

Range measures the distance between the lowest and highest values $(x_{n:n} - x_{1:n})$ in the data set and generally describes how spread out data are.

Percentiles measure the percentage of data points which lie below a certain value when the values are ordered.

Commonly used percentiles

- First (lower) decile = 10th percentile
- First (lower) quartile, Q_1, = 25th percentile
- Second (middle) quartile, Q_2, = 50th percentile
- Third quartile, Q_3, = 75th percentile
- Ninth (upper) decile = 90th percentile



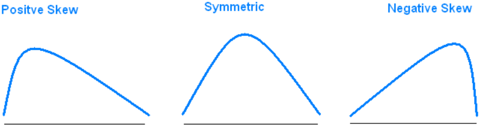
Quantiles

The quantiles are values which divide the distribution such that there is a given proportion of observations below the quantile.

- For example, the median is a quantile. The median is the central value of the distribution, such that half the points are less than or equal to it and half are greater than or equal to it.
- The *p* quantile, (also known as the 100p%-percentile) is the point in the data where 100p% is less, and 100(1-p)% is larger.
- What does it mean ”the 0.45 quantile of the weight of rats distribution is 0.2Kg”.

Measures of Skewness

More conceptually, skew defines the relative positions of the mean, median, and mode. If a distribution is **skewed to the right** (positive skew), the mean lies to the right of both the mode (most frequent value and hump in the curve) and median (middle value). That is, mode $<$ median $<$ mean. But, if the distribution is **skewed left** (negative skew), the mean lies to the left of the median and the mode. That is, mean $<$ median $<$ mode.



Measures of skewness

Skew indicates the degree of symmetry in a data set. The more skewed the distribution, the higher the variability of the measures.

$$skew = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)s^3}$$

- $skew > 0$ positive skew.
- $skew = 0$ symmetric distribution.
- $skew < 0$ negative skew.

In older textbooks skewness is defined by $\frac{m_3}{(m_2)^{3/2}}$, where $m_r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^r}{n}$ is the sample moment of order r.

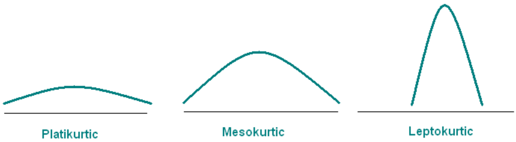
Measures of Kurtosis

Measures of kurtosis describe how concentrated data are around a single value, usually the mean. Thus, kurtosis assesses how peaked or flat is the data distribution. The more peaked or flat the distribution, the less normally distributed is the data.

$$k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)s^4} - 3$$

Measures of Kurtosis

Mesokurtic distributions are, like the normal bell curve, neither peaked nor flat (k=0).
Platykurtic distributions are flatter than the normal bell curve (k<0).
Leptokurtic distributions are more peaked than the normal bell curve (k>0).



Exercises

Go to ▸ Link

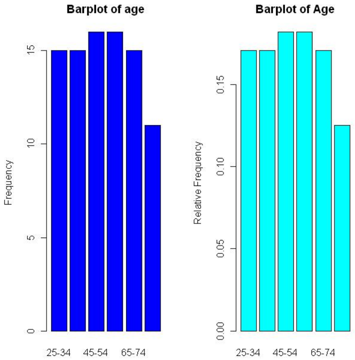
Do exercise 3.

Barplot

Numerical data can be graphically illustrated using a **Barplot** (or bar chart). A bar chart draws a bar with a height proportional to the count in the table. The height could be given by the frequency, or the proportion. The graph will look the same, but the scales may be different.

Barplot

The bars are not contiguous (touching one another) nor do the areas of the strips have a meaning; rather, the heights of the rectangles are proportional to the frequency.



Error bars

- Figures with error bars can, if used properly, give information describing the data (descriptive statistics), or information about what conclusions, or inferences, are justified (inferential statistics).
- These two basic categories of error bars are depicted in exactly the same way, but are actually fundamentally different.

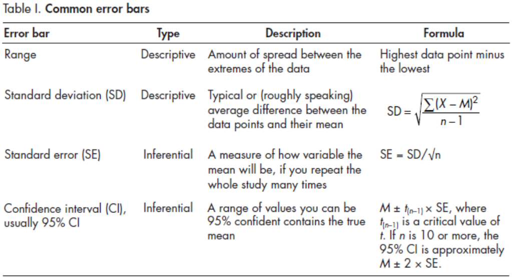


Table I. **Common error bars**

Error bar	Type	Description	Formula
Range	Descriptive	Amount of spread between the extremes of the data	Highest data point minus the lowest
Standard deviation (SD)	Descriptive	Typical or (roughly speaking) average difference between the data points and their mean	$SD = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$
Standard error (SE)	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times	$SE = SD/\sqrt{n}$
Confidence interval (CI), usually 95% CI	Inferential	A range of values you can be 95% confident contains the true mean	$M \pm t_{(n-1)} \times SE$, where $t_{(n-1)}$ is a critical value of t. If n is 10 or more, the 95% CI is approximately $M \pm 2 \times SE$.

Figure: Source: Cumming *et al.* (2007)

Descriptive Error bars

Range and standard deviation (SD) are used for descriptive error bars because they show how the data are spread.

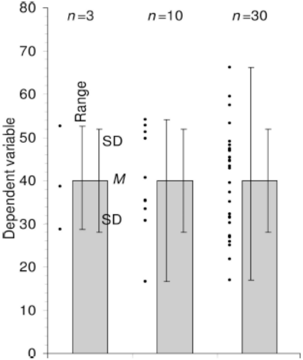


Figure 1. **Descriptive error bars.** Means with error bars for three cases: n = 3, n = 10, and n = 30. The small black dots are data points, and the column denotes the data mean *M*. The bars on the left of each column show range, and the bars on the right show standard deviation (SD). *M* and SD are the same for every case, but notice how much the range increases with n. Note also that although the range error bars encompass all of the experimental results, they do not necessarily cover all the results that could possibly occur. SD error bars include about two thirds of the sample, and 2 x SD error bars would encompass roughly 95% of the sample.

Figure: Source: Cumming *et al.* (2007)

Descriptive Error bars

- Descriptive error bars can also be used to see whether a single result fits within the normal range.

For example, if you wished to see if a red blood cell count was normal, you could see whether it was within 2SD of the mean of the population as a whole. Less than 5% of all red blood cell counts are more than 2SD from the mean, so if the count in question is more than 2SD from the mean, you might consider it to be abnormal.

Empirical rule

- Approximately 68% of the measurements are in the interval $[\bar{x} - s; \bar{x} + s]$,
- Approximately 95% of the measurements are in the interval $[\bar{x} - 2s; \bar{x} + 2s]$,
- Approximately 99.7% of the measurements are in the interval $[\bar{x} - 3s; \bar{x} + 3s]$.

Sometimes, this approximation can be a bit crude, depending upon the sample size and skewness of the sample, but its surprisingly accurate.

Inferential Error bars

- In experimental biology it is more common to be interested in comparing samples from two groups, to see if they are different.

For example, you might be comparing wild–type mice with mutant mice, or drug with placebo, or experimental results with controls. To make inferences from the data (i.e., to make a judgment whether the groups are significantly different, or whether the differences might just be due to random fluctuation or chance), a different type of error bar can be used. These are **standard error** (SE) bars and **confidence intervals** (CIs).

Inferential Error bars

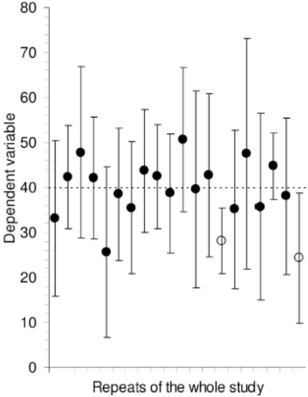


Figure 2. **Confidence intervals.** Means and 95% CIs for 20 independent sets of results, each of size $n = 10$, from a population with mean μ = 40 (marked by the dotted line). In the long run we expect 95% of such CIs to capture μ; here 18 do so (large black dots) and 2 do not (open circles). Successive CIs vary considerably, not only in position relative to μ, but also in length. The variation from CI to CI would be less for larger sets of results, for example $n = 30$ or more, but variation in position and in CI length would be even greater for smaller samples, for example $n = 3$.

Figure: Source: Cumming *et al.*, (2007)

Error bars

- **Rule 1:** when showing error bars, always describe in the figure legends what they are.
- **Rule 2:** the value of n (i.e., the sample size, or the number of independently performed experiments) must be stated in the figure legend. It is essential that n (the number of independent results) is carefully distinguished from the number of replicates, which refers to repetition of measurement on one individual in a single condition, or multiple measurements of the same or identical samples.

Error bars: replicates

Consider trying to determine whether deletion of a gene in mice affects tail length. We could choose one mutant mouse and one wild type, and perform 20 replicate measurements of each of their tails. We could calculate the means, SDs, and SEs of the replicate measurements, but these would not permit us to answer the central question of whether gene deletion affects tail length, because n would equal 1 for each genotype, no matter how often each tail was measured.

For replicates, $n = 1$, and it is therefore inappropriate to show error bars or statistics.

Error bars: replicates

Rule 3: error bars and statistics should only be shown for independently repeated experiments, and never for replicates. If a representative experiment is shown, it should not have error bars or P-values, because in such an experiment, $n = 1$.

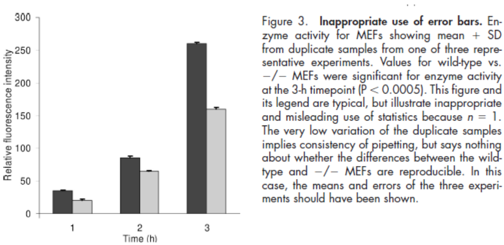


Figure 3. **Inappropriate use of error bars.** Enzyme activity for MEFs showing mean + SD from duplicate samples from one of three representative experiments. Values for wild-type vs. −/− MEFs were significant for enzyme activity at the 3-h timepoint ($P < 0.0005$). This figure and its legend are typical, but illustrate inappropriate and misleading use of statistics because $n = 1$. The very low variation of the duplicate samples implies consistency of pipetting, but says nothing about whether the differences between the wild-type and −/− MEFs are reproducible. In this case, the means and errors of the three experiments should have been shown.

Figure: Source: Cumming *et al.*, (2007)

Error bars

Rule 4: because experimental biologists are usually trying to compare experimental results with controls, it is usually appropriate to show inferential error bars, such as SE or CI, rather than SD. However, if n is very small (for example $n = 3$), rather than showing error bars and statistics, it is better to simply plot the individual data points.

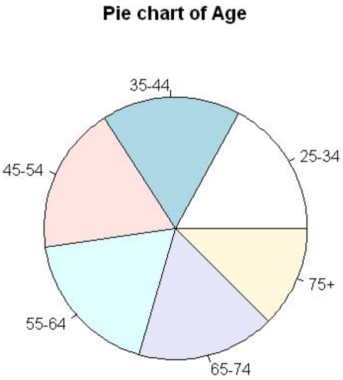
Error Bars: some considerations

When first seeing a figure with error bars, ask yourself:
”What is *n*?
”Are they independent experiments, or just replicates?” and,
”What kind of error bars are they?”

If the figure legend gives you satisfactory answers to these questions, you can interpret the data, but remember that error bars and other statistics can only be a guide: you also need to use your biological understanding to appreciate the meaning of the numbers shown in any figure.

Pie chart

The same data can be studied with pie charts. Each class corresponds to each slice, and the slice areas are proportional to the frequencies.

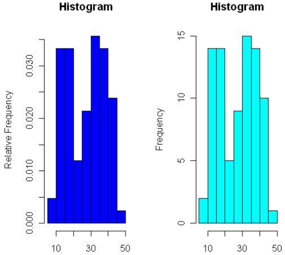


Histogram

Numerical data can be graphically illustrated using a histogram. A histogram shows the ranges of the variable of interest on the horizontal axis and either the frequency or relative frequency of each range on the vertical axis.

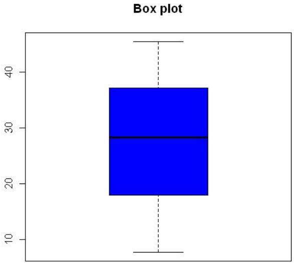
Histogram

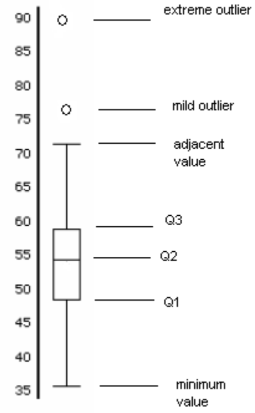
For histograms the bars are contiguous and the areas of the rectangles are proportional to the frequencies. As such, notice that the areas for a histogram sum to one. This has direct and important applications to probability (and percentages and proportions).



Boxplot

A box plot provides an excellent visual summary of many important aspects of a distribution. The box stretches from the lower hinge (defined as the 25th percentile) to the upper hinge (the 75th percentile) and therefore contains the middle half of the scores in the distribution.





Interquartile Range: $IQR = Q_3 - Q_1$

Inner fences: $\{Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)\}$

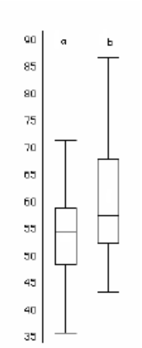
Outer fences: $\{Q_1 - 3(IQR), Q_3 + 3(IQR)\}$

This is a pictorial display that provides the main descriptive measures of the measurement set:
L - the largest measurement inside the inner fences
Q_3 - The upper quartile
Q_2 - The median
Q_1 - The lower quartile
S - The smallest measurement inside the inner fences

A *potential (mild) outlier* is a value located at a distance of more than $1.5IQR$ from the box. An *(extreme) outlier* is a value located at a distance of more than $3IQR$ from the box.

Parallel Boxplot

It is often useful to compare data from two or more groups by viewing box plots from the groups side by side.



Now we may compare mean, median (central location) and spread (variation like IQR) of two data sets.

Exercises

Go to ▸ Link

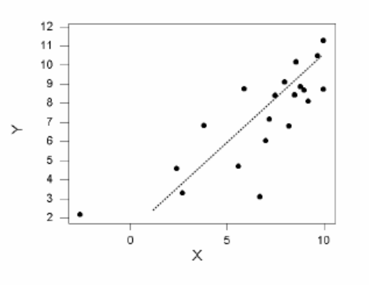
Do exercises 4 and 5.

Scatter Diagrams

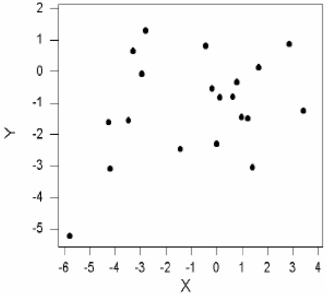
Often we are interested in the relationships between two quantitative variables.

Typical Patterns

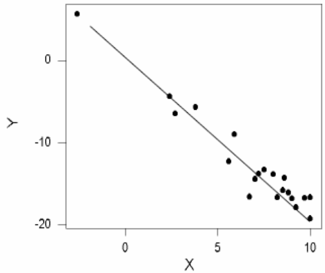
Positive linear relationship: if X increases then Y increases and vice versa.



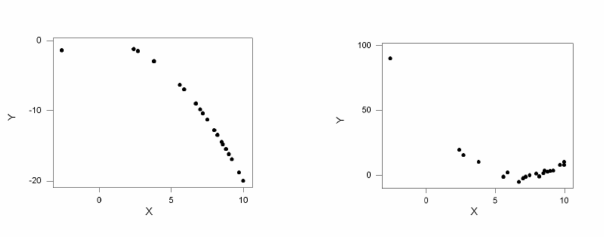
No relationship



Negative linear relationship: If X increases then Y decreases and vice versa.



Nonlinear (concave) relationship (sometimes not so easy to identify)



Measures of association

Two numerical measures are presented, for the description of linear relationship between two variables depicted in the scatter diagram.

- Covariance - is there any pattern to the way two variables move together?
- Correlation coefficient - how strong is the linear relationship between two variables?

Covariance

Population covariance$=COV(X,Y)=\frac{\sum_{i=1}^{N}(X_i-\mu_x)(Y_i-\mu_y)}{N}$

Sample covariance$=cov(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

- If the two variables move the same direction (either both increase or both decrease), the covariance is a large positive number.
- If the two variables move in two opposite directions (one increases when the other one decreases), the covariance is a large negative number.
- If the two variables are linearly unrelated, the covariance will be close to zero.

The coefficient of correlation

Population coefficient correlation: $\rho = \frac{COV(X,Y)}{\sigma_x \sigma_y}$

Sample coefficient of correlation: $r = \frac{cov(X,Y)}{s_x s_y}$

This coefficient answers the question: How strong is the association between X and Y.

- or r = +1 Strong positive linear relationship
- or r = 0 No linear relationship
- or r = -1 Strong negative linear relationship

Note: Correlation measures linear association only. A zero correlation does not necessarily mean there is no relationship between X and Y.

Correlation Coefficients

- Pearson coefficient: normally distributed data.
- Spearman coefficient (rank correlation):for ordered categorical variables at least; when normality assumption is not verified.

Misuses of correlation coefficient

- It is incorrect to calculate a simple correlation coefficient for data which include more than one observation on some or all of the subjects, because such observations are not independent.

Misuses of correlation coefficient

- It is incorrect to calculate a simple correlation coefficient for data which include more than one observation on some or all of the subjects, because such observations are not independent.
- Correlation is inappropriate for comparing alternative methods of measurements of the same variable, because it assesses association not agreement.

Misuses of correlation coefficient

- It is incorrect to calculate a simple correlation coefficient for data which include more than one observation on some or all of the subjects, because such observations are not independent.
- Correlation is inappropriate for comparing alternative methods of measurements of the same variable, because it assesses association not agreement.
- Regression and correlation are separate techniques serving different purposes and need not automatically accompany each other.

Exercises

Go to ▸ Link

Do exercise 6.

Multiple choice questions

Multiple choice questions

Each statement is either true or false.

The mean of a large sample of size n

1. Is always greater tan the median.
2. Is calculated from the formula $\frac{\sum x}{n}$.
3. Estimates the population mean with greater precision than the mean of a small sample.
4. Increases as the sample size increases.
5. Is always greater than the standard deviation.

Multiple choice questions

The following are measures of the spread of a distribution.

1. Interquantile range.
2. Standard deviation.
3. Range.
4. Median.
5. Mode

Multiple choice questions

As the size of a random sample increases:

1. The standard deviation decreases.
2. The standard error of the mean decreases.
3. The mean decreases.
4. The accuracy of the parameter estimates increases.

Multiple choice questions

A correlation coefficient

1. Always lies in the range 0-1.
2. Can be use to predict one variable from another.
3. Could be used to summarize the relatioship between haemoglobin concentration and blood group in a sample of hospital patients.
4. Is a measure of the extent to which two continuous variables are linearly related.

Bibliography

- Ewens, W.J. and Grant, G.R. (2001). *Statistical Methods in Bioinformatics: An Introduction.* Springer, New York.
- Krijnen, W. P. (2009). Applied Statistics for Bioinformatics using R.
- Cummimg, G., Fidler, F. and Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of Cell Biology*, 177(1):7-11.