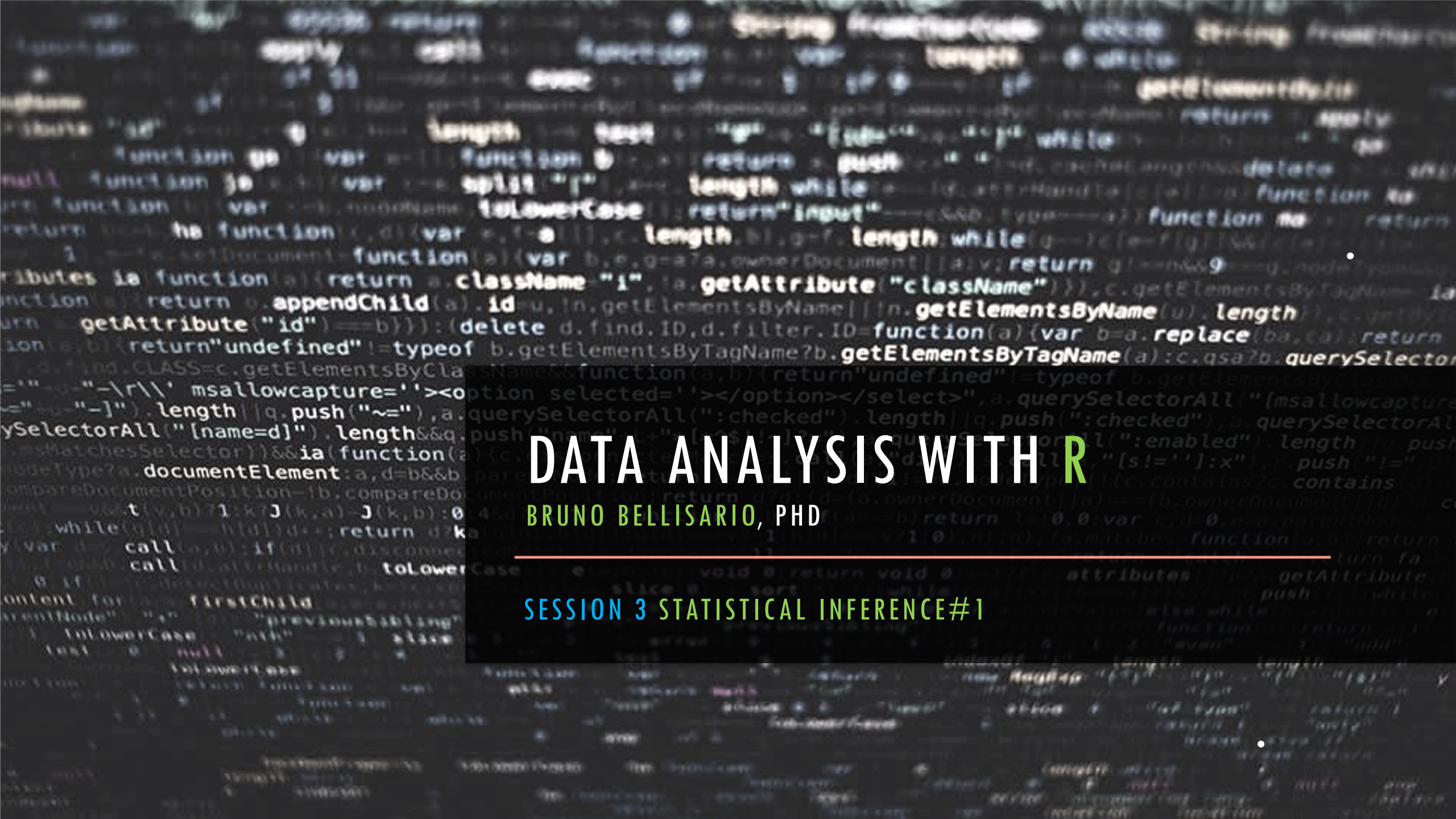


DATA ANALYSIS WITH R

BRUNO BELLISARIO, PHD

SESSION 3 STATISTICAL INFERENCE#1

A PROBABILISTIC VIEW OF REALITY

→To analyze a given fragment of reality relevant for the specific purpose at hand, one usually needs to collect some data. Data may come from past experiences and observations, or may result from some controlled processes, such as laboratory or field experiments. The data are then used to hypothesize about the laws (often called mechanisms) that govern the fragment of reality of interest.

→We are interested in laws expressed in probabilistic terms: They specify directly, or allow us to compute, the chances of some events to occur. Knowledge of these chances is, in most cases, the best one can get regarding prediction and decisions.

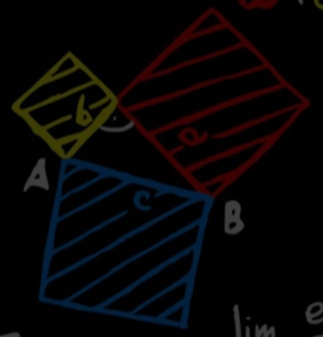
→An application of probability to the situation analyzed requires a few initial steps, in which the elements of the real situation are interpreted as abstract concepts of probability theory. Such interpretation is often referred to as building a probabilistic model of the situation at hand. How well this is done is crucial to the success of application.

A PROBABILISTIC VIEW OF REALITY

→One of the main concepts is that of an experiment-a term used in a sense somewhat broader than usual.

→It means any process, possibly under partial control, that we may observe and whose behavior in the future is not totally determined because it is influenced, at least in part, by chance.

STATISTICAL INFERENCE DEFINED



→Make inference is usually related to using statistical tests to draw some conclusion about the data.

→Hypothesis testing is important for determining if there are statistically significant effects. However, we should not place undo emphasis on p-values. Instead, we should realize that p-values are affected by sample size, and that a low p-value does not necessarily suggest a large effect or a practically meaningful effect.

→Important thing is to understand the limitations of relying on p-values, and why it is important to assess the size of effects and weigh practical considerations.

HYPOTHESIS TESTING: THE NULL AND ALTERNATIVE HYPOTHESES

→The null hypothesis always describes the case where e.g. two groups are not different or there is no correlation between two variables, etc.

→The alternative hypothesis is the contrary of the null hypothesis, and so describes the cases where there is a difference among groups or a correlation between two variables, etc.

→Notice that the definitions of null hypothesis and alternative hypothesis have nothing to do with what you want to find or don't want to find, or what is interesting or not interesting, or what you expect to find or what you don’t expect to find.

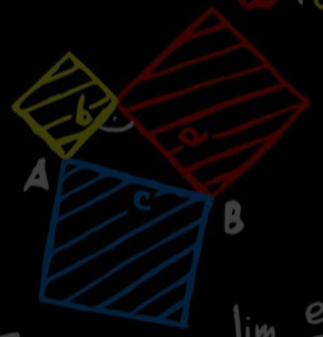
P-VALUE DEFINITION

→The p-value for the given data will be determined by conducting the statistical test.

→This p-value is then compared to a pre-determined value alpha. Most commonly, an alpha value of 0.05 is used, but there is nothing magic about this value.

→If the p-value for the test is less than alpha, we reject the null hypothesis.

→If the p-value is greater than or equal to alpha, we fail to reject the null hypothesis.



THEORY AND PRACTICE OF USING P-VALUES

→Given the assumption that the null hypothesis is true, the p-value is defined as the probability of obtaining a result equal to or more extreme than what was actually observed in the data.

→Technically, the p-value says nothing about the alternative hypothesis.

→But logically, if the null hypothesis is rejected, then its logical complement, the alternative hypothesis, is supported.

THE 0.05 ALPHA VALUE IS NOT DOGMA

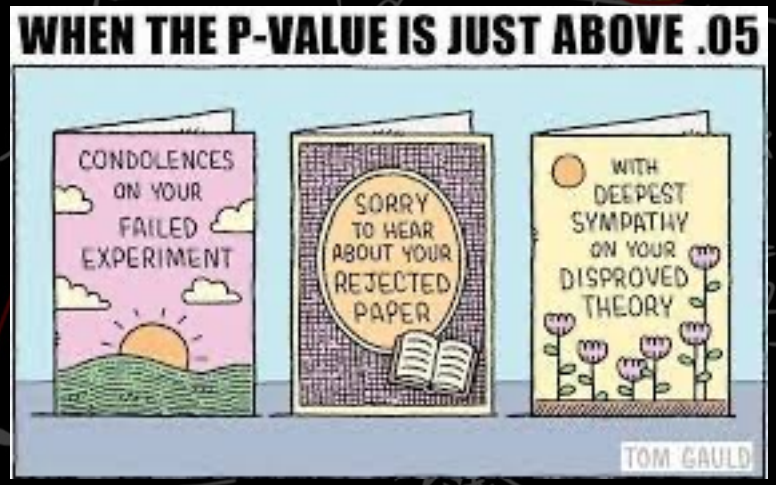
→The level of alpha is traditionally set at 0.05 in some disciplines, though there is sometimes reason to choose a different value.

→In theory, as a researcher, you would determine the alpha level you feel is appropriate. That is, the probability of making a Type I error when the null hypothesis is in fact true.

→Choosing a different alpha value will rarely go without question. It is best to keep with the 0.05 level unless you have good justification for another value or are in a discipline where other values are routinely used.

THE 0.05 ALPHA VALUE IS NOT DOGMA

P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥0.1	



THE 0.05 ALPHA VALUE IS NOT DOGMA

BIOLOGY LETTERS

royalsocietypublishing.org/journal/rsbl

Review

Cite this article: Halsey LG. 2019 The reign of the *p*-value is over: what alternative analyses could we employ to fill the power vacuum? *Biol. Lett.* **15**: 20190174.
http://dx.doi.org/10.1098/rsbl.2019.0174

Received: 4 March 2019
Accepted: 1 May 2019

Subject Areas:
bioinformatics

Keywords:
AIC, Bayesian, confidence intervals, effect size, statistical analysis

Author for correspondence:
Lewis G. Halsey
e-mail: l.halsey@roehampton.ac.uk

Animal behaviour

The reign of the *p*-value is over: what alternative analyses could we employ to fill the power vacuum?

Lewis G. Halsey

University of Roehampton, London SW15 4JD, UK

LGH, 0000-0002-0786-7585

The *p*-value has long been the figurehead of statistical analysis in biology, but its position is under threat. *p* is now widely recognized as providing quite limited information about our data, and as being easily misinterpreted. Many biologists are aware of *p*'s frailties, but less clear about how they might change the way they analyse their data in response. This article highlights and summarizes four broad statistical approaches that augment or replace the *p*-value, and that are relatively straightforward to apply. First, you can augment your *p*-value with information about how confident you are in it, how likely it is that you will get a similar *p*-value in a replicate study, or the probability that a statistically significant finding is in fact a false positive. Second, you can enhance the information provided by frequentist statistics with a focus on effect sizes and a quantified confidence that those effect sizes are accurate. Third, you can augment or substitute *p*-values with the Bayes factor to inform on the relative levels of evidence for the null and alternative hypotheses; this approach is particularly appropriate for studies where you wish to keep collecting data until clear evidence for or against your hypothesis has accrued. Finally, specifically where you are using multiple variables to predict an outcome through model building, Akaike information criteria can take the place of the *p*-value, providing quantified information on what model is best. Hopefully, this quick-and-easy guide to some simple yet powerful statistical options will support biologists in adopting new approaches where they feel that the *p*-value alone is not doing their data justice.

ERRORS IN INFERENCE

→For the most part, the statistical tests we use are based on probability, and our data could always be the result of chance. We can make two kinds of errors in testing the null hypothesis:

→Type I Error: Occurs when the null hypothesis really is true but, based on our decision rule, we reject the null hypothesis. In this case, our result is a false positive; we think there is an effect when really there isn't.

→Type II Error: Occurs when the null hypothesis is false but, based on our decision rule, we fail to reject the null hypothesis. In this case, our result is a false negative.

ERRORS IN INFERENCE

Type I and Type II Error

Type I Error (false-positive)

Type II Error (false-negative)

STATISTICAL POWER

- The statistical power of a test is a measure of the ability of the test to detect a real effect.
- It is related to the effect size, the sample size, and our chosen alpha level.
- The effect size is a measure of how unfair a coin is, how strong the association is between two variables, or how large the difference is among groups.
- As the effect size increases, or, as the number of observations we collect increases, or, as the alpha level increases, the power of the test increases.

EFFECT SIZES AND PRACTICAL IMPORTANCE

→It is important to remember to not let p-values be the only guide for drawing conclusions. It is equally important to look at the size of the effects you are measuring.

→It should be remembered that p-values do not indicate the size of the effect being studied. It shouldn't be assumed that a small p-value indicates a large difference between groups, or vice-versa.

→It should also be remembered that p-values are affected by sample size. For a given effect size and variability in the data, as the sample size increases, the p-value is likely to decrease. For large data sets, small effects can result in significant p-values.

EFFECT SIZE STATISTICS

- One way to account for the effect of sample size on our statistical tests is to consider effect size statistics. These statistics reflect the size of the effect in a standardized way and are unaffected by sample size.
- An appropriate effect size statistic for a t-test is Cohen's d. It takes the difference in means between the two groups and divides by the pooled standard deviation of the groups.
- Cohen's d equals zero if the means are the same and increases to infinity as the difference in means increases relative to the standard deviation.
- Effect size statistics are standardized so that they are not affected by the units of measurements of the data. This makes them interpretable across different situations, or if the reader is not familiar with the units of measurement in the original data.

THE PROBLEM OF MULTIPLE P-VALUES

- One concept that will be important is that when there are multiple tests producing multiple p-values, that there is an inflation of the Type I error rate. That is, there is a higher chance of making false-positive errors.
- This simply follows mathematically from the definition of alpha. If we allow a probability of 0.05, or 5% chance, of making a Type I error for any one test, as we do more and more tests, the chances that at least one of them having a false positive becomes greater and greater.
- One way we deal with the problem of multiple p-values in statistical analyses is to adjust p-values when we do a series of tests together (for example, if we are comparing the means of multiple groups).

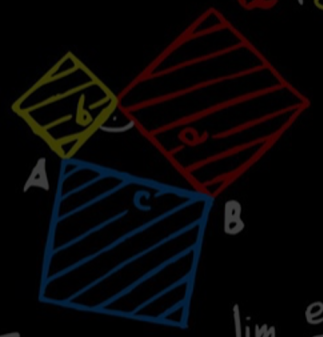
WHAT IS THE HYPOTHESIS?

- The most important consideration in choosing a statistical test is determining what hypothesis you want to test.
- More generally, what question are you are trying to answer.
- Often people have a notion about the purpose of the research they are conducting but haven't formulated a specific hypothesis.
- It is possible to begin with exploratory data analysis, to see what interesting secrets the data wish to say. But ultimately, choosing a statistical test relies on having in mind a specific hypothesis to test.

PLAN YOUR EXPERIMENTAL DESIGN BEFORE YOU COLLECT DATA

One of the most common mistakes people make in doing research is collecting a bunch of data without having thought through what questions they are trying to answer, what specific hypotheses they want to test, and what statistical tests they can use to test these hypotheses.

INDEPENDENT AND PAIRED VALUES



→An assumption of many statistical tests is that observations are independent of one another. This means that the value for one observation is unlikely to be influenced by the value of another observation.

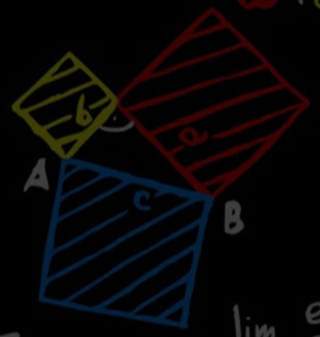
→However, dependent samples commonly arise in a few situations. One is repeated measures, in which the same subject is measured on multiple dates.

→A second is when we are taking multiple measurements of the same individual.

→A related concept is that of blocks. If observations can be broken into meaningful groups where values are likely to be different, this should be considered.

PARAMETRIC STATISTICS

- Parametric statistical tests include t-test, analysis of variance, and linear regression.
- Used when the dependent variable is an interval/ratio data variable.
- One advantage of using parametric statistical tests is that your audience will likely be familiar with the techniques and interpretation of the results. These tests are also often more flexible and more powerful than their nonparametric analogues.
- Their major drawback is that all parametric tests assume something about the distribution of the underlying data. If these assumptions are violated, the resultant test statistics will not be valid, and the tests will not be as powerful as for cases when assumptions are met.



COUNT DATA MAY NOT BE APPROPRIATE FOR COMMON PARAMETRIC TESTS

A frequent error is to use common parametric models and tests with count data for the dependent variable. Instead, count data could be analysed either by using tests for nominal data or by using regression methods appropriate for count data. These include Poisson regression, negative binomial regression, and zero-inflated Poisson regression.

COUNT DATA MAY NOT BE APPROPRIATE FOR COMMON PARAMETRIC TESTS

Count data or other discrete data can be used in cases where counts are used as a type of measurement of some property of subjects, provided that 1) the distribution of data or residuals from the analysis approximately meet test assumptions; and 2) there are few or no counts at or close to zero, or close to a maximum, if one exists.

This kind of count data will sometimes need to be transformed to meet the assumptions of parametric analysis. However, if there are many counts at or near zero, transformation is unlikely to help. It is usually not worth the effort to attempt to force count data to meet the assumptions of parametric analysis with transformations, since there are more appropriate methods available.

ASSUMPTIONS IN PARAMETRIC STATISTICS

→All statistical tests assume that the data captured in the sample are randomly chosen from the population.

→Observations should be independent of one another, except when the analysis takes non-independence into account. The independence of observation is often assumed from good experimental design.

→The distribution of the data should be conditionally normal in distribution. That is, the data are normally distributed once the effects of the variables in the model are taken into account. Practically speaking, this means that the residuals from the analysis should be normally distributed. This will usually be assessed with a histogram of residuals or with quantile–quantile plot.

ASSUMPTIONS IN PARAMETRIC STATISTICS

Tests for assessing if data is normally distributed

There are specific methods for testing normality, but these should be used in conjunction with either a histogram or a Q-Q plot. The Kolmogorov-Smirnov and the Shapiro-Wilk's W test whether the underlying distribution is normal. Both tests are sensitive to outliers and are influenced by sample size:

→For smaller samples, non-normality is less likely to be detected, but the Shapiro-Wilk test should be preferred as it is generally more sensitive

→For larger samples (i.e., more than one hundred), the normality tests are overly conservative, and the assumption of normality might be rejected too easily

→Null hypothesis: The data is normally distributed. If p> 0.05, normality can be assumed.

1-SAMPLE T-TEST

One-sample tests are not used too often but are useful to compare a set of values to a given default value.

Appropriate data

→One-sample data

→Data are interval/ratio, and are continuous

→Data are normally distributed

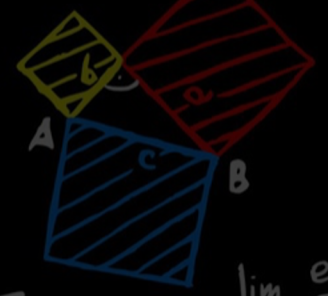
→Moderate skewness is permissible if the data distribution is unimodal without outliers

Hypotheses

→Null hypothesis: The means of the populations from which the data were sampled for each group are equal.

→Alternative hypothesis (two-sided): The means of the populations from which the data were sampled for each group are not equal.

2-SAMPLE T-TEST



The two-sample t-test is a commonly used test that compares the means of two samples.

Appropriate data

→Two-sample data

→Dependent variable is interval/ratio, and is continuous

→Independent variable is a factor with two levels

→Data for each population are normally distributed

→Samples need to have the same variance. Independent observations

Hypotheses

→Null hypothesis: The means of the populations from which the data were sampled for each group are equal.

→Alternative hypothesis (two-sided): The means of the populations from which the data were sampled for each group are not equal.