STATISTICS

Valentin Amrhein

Could you ever "prove" a hypothesis?

No. If thousand experiments find support for a hypothesis, this does not mean it is necessarily true.

One single experiment that is eventually finding clear evidence against the hypothesis would make a revision of the hypothesis necessary.

Hypotheses cannot be proven, but they can be falsified.

The philosopher Karl Popper (1902-1994) pointed out that a good hypothesis is one that is capable of rejection.

A good hypothesis is a falsifiable hypothesis.

Thus, science often advances by falsifying hypotheses.

Which of the following hypotheses would be easier to falsify?

1. There are vultures on the Petersplatz.

2. There are no vultures on the Petersplatz.

If you go to the Petersplatz and do not find a vulture, does this mean hypothesis 1 is wrong?
No, because the vulture might just have hidden behind a tree.

However, the first time you see a vulture, hypothesis 2 is clearly wrong, so it is falsified and you can reject it.

Hypothesis 2 is a **null hypothesis** saying "nothing is happening". The good thing about null hypotheses is that they can be rejected.

In statistics, we reject the null hypothesis when our data show that the null hypothesis is sufficiently unlikely.

Consider the case you are tossing a coin, but you suspect the coin is marked (gezinkt).

Hypothesis: The coin is marked.

Null hypothesis: The coin is not marked.

We will reject the null hypothesis and assume that the coin is marked if we get "head" so often that the null hypothesis is sufficiently unlikely.

"Sufficiently unlikely" is usually defined as occurring with a probability of less then 5% ($P < 0.05$), which is also called **significant**.

Under the assumption that the null hypothesis "the coin is not marked" is true, the probability (P) of tossing a coin and getting head is 50% = 0.5

Getting several head in a row ("multiplication rule"):

0.5 * 0.5 = 0.25

0.5 * 0.5 * 0.5 = 0.125

0.5 * 0.5 * 0.5 * 0.5 = 0.0625

0.5 * 0.5 * 0.5 * 0.5 * 0.5 = 0.03125

Given that the null hypothesis is true, getting head 5 times in a row is so unlikely (P = 0.031) that we reject the null hypothesis and assume that the coin is marked.

P = Probability of the observed data (or data more extreme) **given that the null hypothesis is true**

= Probability of obtaining 5 (or more) head in a row "just by chance", although the coin was actually not marked.

Building models

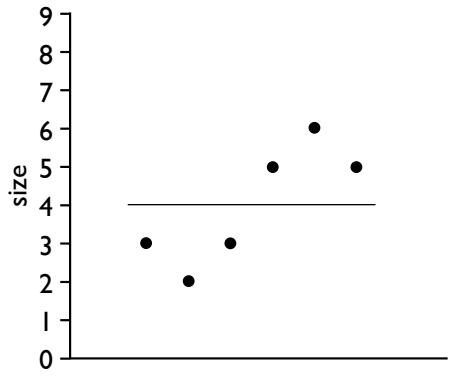
In statistics, we build models to test assumptions about the real world.

In the coin example, the model was that under the null hypothesis, the coin was perfectly shaped, so that head occurs in exactly 50% of cases, even if the coin is tossed one billion times.

Note that almost no coin will be that perfectly built, thus a perfect coin is clearly a model, not reality.

Null hypotheses are almost always models.

The mean is a model, too

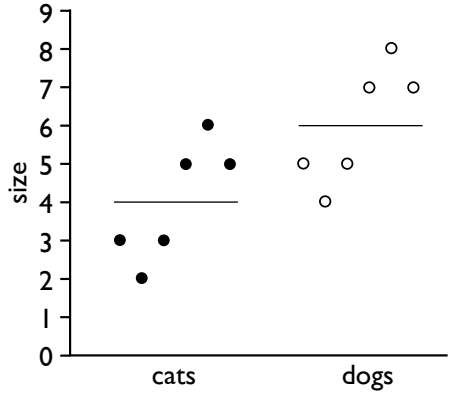


These are the sizes of 6 different cats. Mean cat size is 4, which is not reality, because no cat actually had size 4.

Reality was size 3 for Mimi, size 2 for Mausi, size 3 for Goldi, etc.

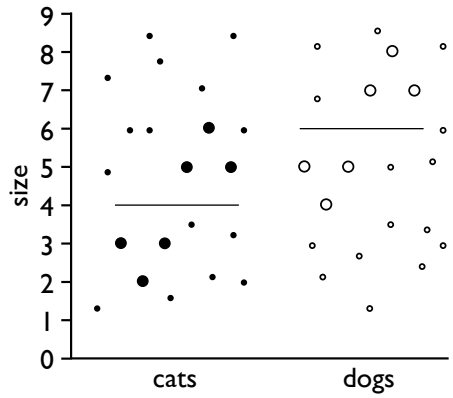
However, we can use the mean as a simplified model of *general* cat size, for example if we want to compare the sizes of cats and dogs.

Comparing two means



To study whether dogs are *generally* larger than cats, it would make little sense to compare the sizes of individual cats and dogs (note that the smallest dogs were smaller than the largest cat).

So let's take the mean as a model to compare pet sizes.
But can the apparently different means in our samples be taken as "prove" that dogs *in general* are larger than cats?



No. We have taken 6 cats and 6 dogs as random samples from larger populations of cats and dogs, and every new sample could turn out differently and make us doubt that cats and dogs differ in size.

The trick is, just as with tossing coins, to assume that cats and dogs are actually of the same size (null hypothesis), and to calculate the probability of finding two samples with means as different (or more different) as in our case.

If under the null hypothesis, the probability of finding two samples with means as different as ours is sufficiently low ($P < 0.05$), we will reject the null hypothesis and assume that cats and dogs are generally of different size.

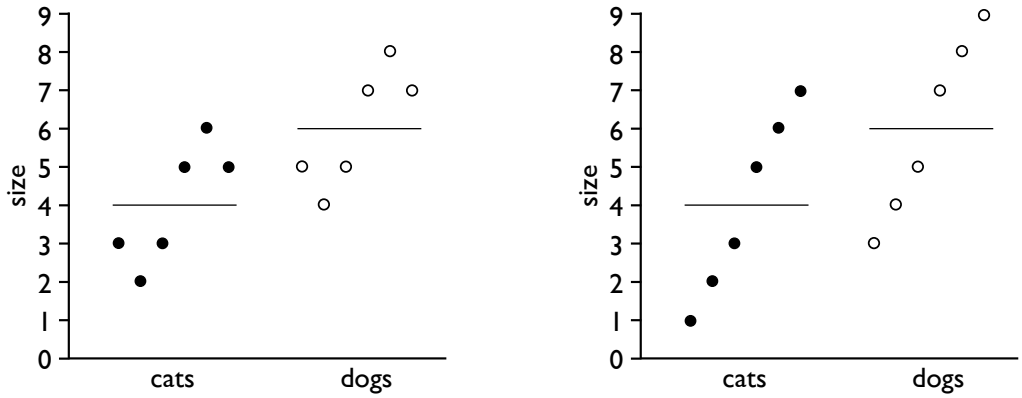
Unfortunately, there is no rule as simple as the multiplication rule in the coin example to calculate whether two means are sufficiently different.

Luckily, a smart guy called W.S. Gosset invented **Student's t** in 1908, which can be used to judge whether two means are different enough so that they are unlikely to have occurred given that the null hypothesis is true.

$t \approx$ difference between means / variation between data points

As the difference between means gets larger, the t value gets larger. Larger t values are associated with smaller P values, perhaps indicating a significant difference between means (if $P < 0.05$).

t ≈ difference between means / variation between data points



t = 2.24, P = 0.049

t = 1.46, P = 0.17

As the variation between data points gets larger, the t value gets smaller. Smaller t values are associated with larger P values. Although in the right figure, the difference between means is the same as in the left, it would be judged as non-significant (P > 0.05).

This makes sense, because with so much variation in the data, the means are actually not reliable enough to make firm conclusions.

Now we have almost all factors that determine how a test like the t-test will perform (the "power" of a test).

The first is the **effect size**, e.g. the difference between means. Large effect sizes are good.

The second is the **variance**, the variation between data points. Large variances are bad.

Note that we cannot control those two factors.
But we can control the third factor, which is the **sample size**.

Large sample sizes are good.
How large the sample size needs to be depends on the effect size you are looking for, on the variance, and on the particular test.

For a test like the t-test, a good rule of thumb is that
$n > 30$ is a good sample size
$10 < n < 20$ is a small sample size
$n < 10$ is a very small sample size.

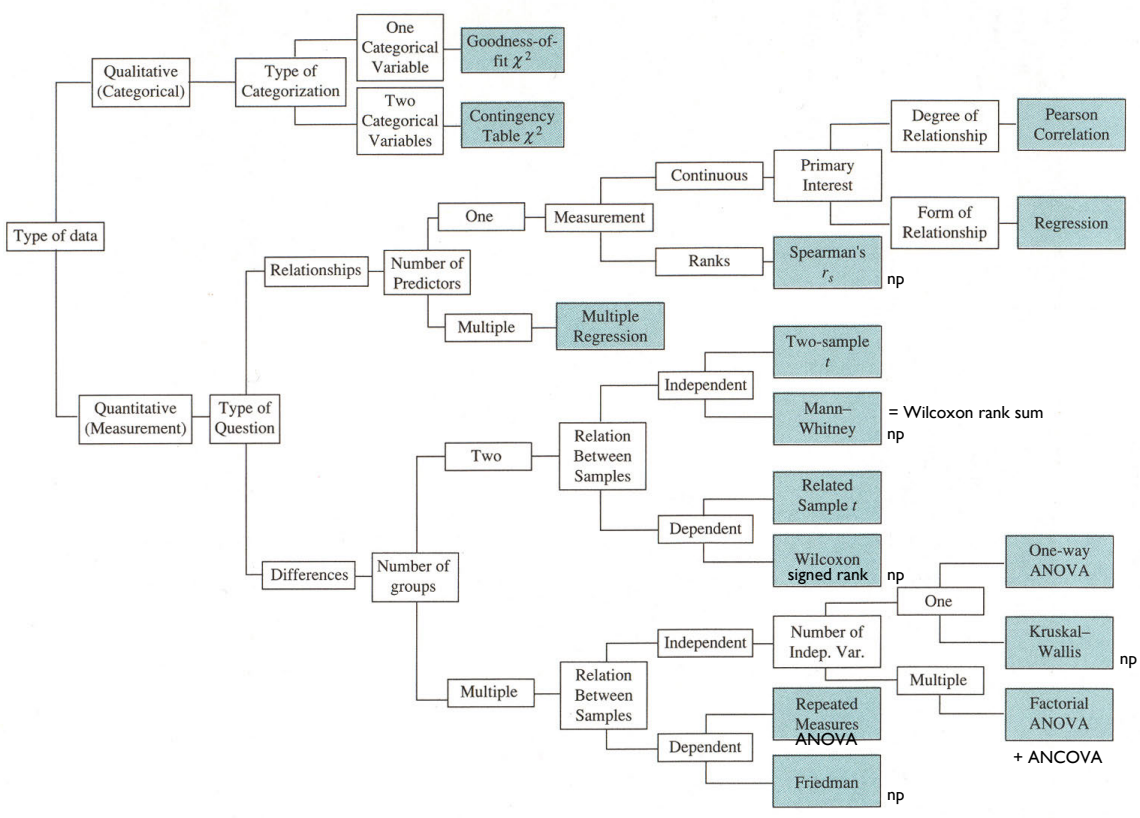
In principle, most statistical tests work in the same way as the t-test. We are interested in whether differences or relationships between samples are large enough to not have occurred just by chance.

We take our samples to calculate specific numbers according to rules that vary among tests. Those numbers have names like F, t, r, or U.

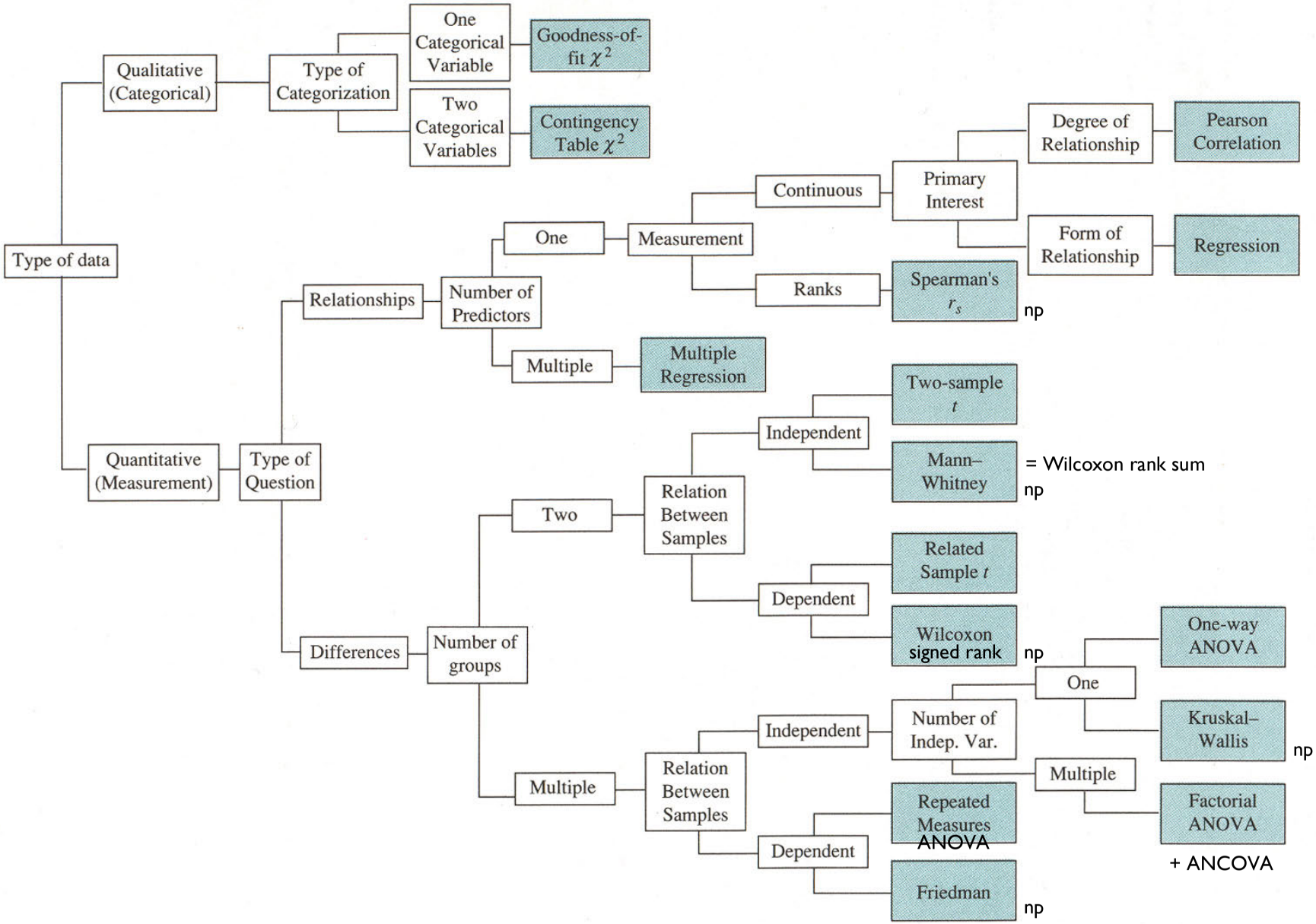
We then look in tables or let computer programs calculate whether the numbers are big enough so that we can reject the null hypothesis.

Knowing which test to use for what study question requires careful thinking, a good deal of practice, and the help of other people or of books and tables like the one following on the next page.

Always think about the test you are going to make BEFORE you do the experiment.

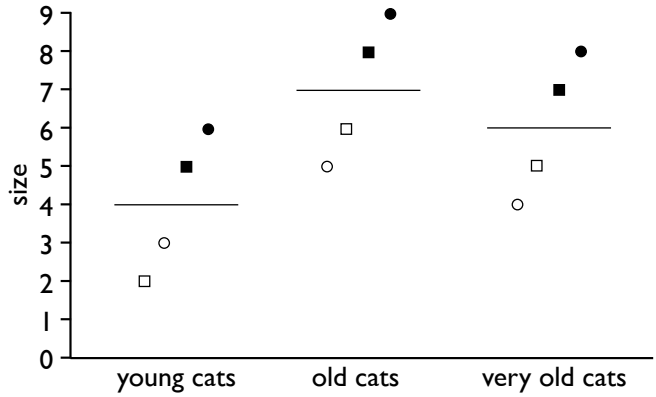


D.C. Howell (2004), Fundamental Statistics for the Behavioral Sciences. Modified by V. Amrhein. np = non-parametric test.



D.C. Howell (2004), Fundamental Statistics for the Behavioral Sciences. Modified by V. Amrhein. np = non-parametric test.

Comparing several means

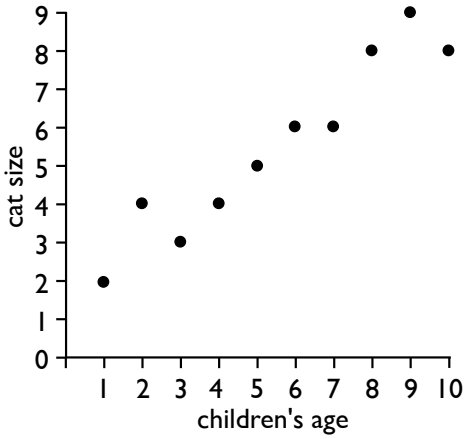


To study whether three (or more) groups of cats are generally of different size, we can use an **ANOVA** (analysis of variance).

Because the data are dependent (each cat was measured three times, at three different age classes), we use a repeated measures ANOVA.

In a paper, the results could be written like this:
The size of cats (n = 4) differed depending on the age class in which they were measured (repeated measures ANOVA, $F_{2,6} = 42$, $P < 0.001$). As judging from the figure, the cats were small when they were young, largest when they were old, but slightly smaller again when they got very old.

Studying relationships



To study whether cats are larger in households in which children are older, we can use a Pearson **correlation**.
In correlation and regression, the predictor variable is usually on the x-axis, the response variable on the y-axis. Do you think children's age really predicts cat size? Could it also be the other way round?
If your answer to the second question is 'no', use a **regression**.

In a paper, the results could be written like this:
Cats (n = 10) were larger in households with older children than in households with younger children (Pearson correlation, $r_8 = 0.96$, $P < 0.001$).

"The hardest part of any statistical work is getting started.
The truth is that there is no substitute for experience;
the way to know what to do is to have done it properly
lots of times before."

Michael J. Crawley (2005)

You probably could / should / have to
buy one of the excellent stats books available, see for example

www.camargue.unibas.ch/statistics.html