Introduction to frequentist inference

Devan Allen McGranahan

September 30, 2016

Here we introduce frequentist statistical inference as it relates to other approaches and introduce distinctions between null hypothesis and alternate hypothesis signficance testing. The present chapter focuses on null hypothesis significance testing and begins by introducing Student's and Welch's t-tests for comparing the means of two groups. Linear regression is introduced as a solution to testing the relationship between two continuous variables. Assumptions of each model are discussed and relevant `R` functions are applied and interpreted.

Introduction

Statistical inference refers to the methods, theory, and practice of drawing conclusions about the parameters of broad populations based on data, gathered mostly from random samples of the population. There is tremendous diversity in approaches to statistics, even within ecology, which is actually quite narrow in the range of statistics used (although this is shifting). Broadly speaking, there are two main approaches to statistical inference[1]:

Statistics has a long, rich tradition; too long to summarize here and too rich to expect 100% accuracy in the following chapter. But the distinctions and definitions herein should help you get an idea of why we do the various things we do, and how they relate to each other.

- *Frequentist inference* emphasizes the frequency, or proportion, of the sample data to draw conclusions about the population. Under this approach, the correct conclusion about the population should have a high probabilty of being observed over repeated experimentation. Scientists can design experiments to increase the probability of reaching the correct conclusion by controlling all factors except for the variable of interest; under such conditions, the influence of unknown parameters can be ignored because they are either controlled for or are otherwise non-influential. There are two types of results from frequentist models:

  Significance testing: Eessentially a true/false conclusion about whether or not the null hypothesis can be rejected.

  Confidence interval: An estimated frequency that a parameter will occur within a certain range.

- Under *Bayesian inference*, one can assign probabilities to unknown parameters, which is one major way in which it differs from frequentist inference. Parameters in Bayesian models can be influenced by *prior* information about them such that the probability of a hypothesis being true is determined by a combination of the inherent likelihood of the hypothesis and agreement between the hypothesis

[1] A third, *fiducial inference*, barely got off the ground in the early 20th century.

and the prior information. And as new data are gathered, these priors can be updated. The result of Bayesian statistics is often a *probability distribution* for population parameters, defining a range and density distribution of potential values for that parameter.

Bayesian inference has a long and interesting history of its own. It is increasingly popular in ecology, made possible by important mathematical solutions in the early 1990s but larely driven by recent increases in personal computing power. But the steps remain complex and successful users have a lot of experience with data distributions and programming in addition to specific training on Bayesian methods.

All statistics used here are based on frequentist inference. This is not to say, however, that these statistics are homogeneous. As described above, frequentist inference can be divided into two approaches: *statistical hypothesis testing* and *confidence intervals*. The latter has advantages and we'll get there later in the course. For now we focus on statistical hypothesis testing[2], of which there are, again, two types: *null hypothesis testing* and *alternative hypothesis testing* (Table 1).

Approach	Originator	Description
Null hypothesis testing	Fischer	Dichotomous true/false result: Could the observed data be derived from chance if the null hypothesis is true? If observed data are significantly unlikely to be observed in the event of the null hypothesis being true, then the null hypothesis is rejected.
Alternative hypothesis testing	Nayman-Pearson	Two simple hypotheses are contrasted and evaluated by their *likelihood ratio*, or how many times more likely the data are under one model versus the other.

Table 1: A summary of two types of statistical hypothesis testing.

Null hypothesis significance testing

Despite critique both legitimate and misguided[3], we begin with null hypothesis significance testing, if not for the very least as a gentle introduction to the application of statistical models to data. The first such models an ecologist is likely to encounter is one of a small suite of models collectively (and probably too often interchangeably) referred to as "*t*-tests."

The t*-statistic and its tests*

t-tests are statistical tests in which the *test statistic*[4] follows a Student's *t*-distribution[5]. The reference to "Student" derives from the pen name of chemist W.S. Gosset, who first published the concept under the pseudonym Student in 1908 because his employer, the Guinness brewery in Dublin, forbade publication of findings. The name has stuck and is often used as a blanket descriptor of all *t*-tests, but it should technically be reserved for instances where the variance is

[2] A *statistical hypothesis* is a hypothesis tested by observing a process modeled from randomly-sampled data.

[3] These points will be developed later, but two camps of criticism for NHST include

- broad vilification of p-values, often based on frustration over widespread incorrect use or misinterpretation by researchers (hardly the fault of the test statistic)
- Inappropriate application of NHST models to ecological data given assumptions and underlying math related to error rate at low sample sizes; these critiques are more often than not probably fair.

[4] A *test statistic* is a quantity derived from sample data and used in a statistical hypothesis test. Each test statistic has a known distribution under the null hypothesis, which supports the calculation of a p-value.

[5] The shape of a t-distribution is similar to that of a normal distribution, in that it has two tails and greatest probability around its mean, but since the t-distribution describes just a sample, and not the whole population, its specific shape varies with sample size and approaches the normal distribution as sample size increases.

assumed to be equal among sample groups and is pooled in the calculation of the t-statistic.

Nomenclature aside, all t-statistics serve the purpose of determining if two sets of data are significantly different from each other. Because of the variability inherent in natural populations and phenomena, it isn't enough to just look at the average values of two groups and draw a conclusion as to whether they are different or not (Table 2)—we also need to consider the *variation* around the mean. This is the whole purpose of statistical analysis.

Sex	Mean	Variance	n
Male	113.4	13.82	10
Female	108.6	5.16	10

Table 2: Average mandible length (in mm) for 10 male and 10 female golden jackals. The males in this sample have longer mandibles, but how do we know if we can expect the pattern to hold for the broader population? *Answer: statistics!*

We use the t-test to ascertain whether two groups are expected to be different in the broad population by (1) calculating a t-statistic from our data and (2) determining the probability that our observation is due to chance. This second step depends on the statistical power of our sample, which relates mostly to the sample size, and is incorporated in the model as the *degrees of freedom*[6].

> If the probability of our observation being derived from chance is below an acceptable threshold—a common threshold is 5%, the source of the infamous 0.05 p-value—we reject the null hypothesis of "no difference" and conclude that male jackals do, in fact, tend to have longer mandibles than females.

DETERMINING DIFFERENCE IN MEANS BASED ON VARIABILITY SHOULD SOUND FAMILIAR, as it heralds a return to our trusy tools for modeling the distribution of data: moments. Calculating a t-statistic is as simple as plugging moments into an equation, although unlike our previous consideration of an entire data vector, here we must also consider moments within each group[7]:

$$t_{\text{Student's}} = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{2/n}} \tag{1}$$

Where

$$\begin{aligned} \bar{X} &= \text{group means,} \\ s_p &= \text{pooled standard deviation, and} \\ n &= \text{total sample size} \end{aligned}$$

The equation hints towards where this, and almost all statistical tests, are going: we're essentially calculating a ratio of *signal to*

[6] *Degrees of freedom* are the number of independent pieces of information that go into the estimate of a parameter, or test statistic. The maximum degrees of freedom for a model is sample size n-1 and are further reduced by a number of factors, such as terms being added to the model. The lower the degrees of freedom, the lower the power of the test statistic to distinguish an observed difference from chance.

[7] Because it determines which type of t-test we run: A proper Student's t-test (Eq. 1) assumes equal variance in both groups, but for a small penalty in degrees of freedom we can run Welch's t-test with unequal variance (Eq. 2).

noise– the signal is the difference in the means, the phenomenon we're interested in describing, while the noise is the variance in our sample. Based on the arrangement of the model terms—the difference in means being divided by variance—the effect of variance on the outcome becomes clear: too much variance and a small difference will be lost, and even a moderate to large difference could be overwhelmed such that only a large sample size—high degrees of freedom—will have enough statistical power to return a p-value that suggests rejecting the null hypothesis.

Handling variance assumptions Before we proceed with Student's t-test based on the test statistic calculated in Eq. 1, a quick look at moments in our data (Table 2) makes us question the assumption about equal variance. But just like our original question—how do we tell if the means are different?—how do we first determine if the variance is equal or not? The quick way is to look at the data (Fig. 1) and perform a variance test using function `var.test`:

```
> var.test(Length ~ Sex, jackal, ratio=1)

        F test to compare two variances

data:  Length by Sex
F = 2.681, num df = 9, denom df = 9, p-value = 0.1579
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.665931 10.793829
sample estimates:
ratio of variances
          2.681034
```

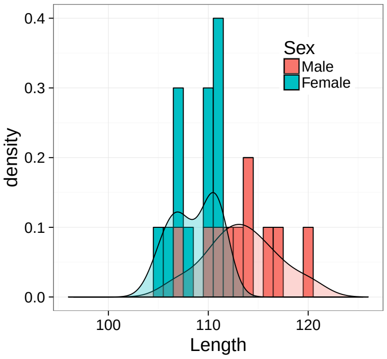

Figure 1: Distribution of male and female golden jackal mandible lengths, which seem normal enough but males have greater variance than females.

The results of the variance test are mixed[8]. Under this framework, if `p < 0.05`, we would reject the null hypothesis that the variance ratio is 1 in favor of the alternative hypothesis, which states the variance ratio is different than 1. Here, on one hand, `p > 0.05`, which suggests we should not reject the null hypothesis that the variance ratio is not different from one. But on the other hand we can *see* that the variance ratio is not 1, both visually in Fig. 1 and quantitatively in the test results: `ratio of variances = 2.68`, which $\neq 1$. A tell-tale sign that the test statistic here might not be terribly reliable is the confidence interval: sure it includes 1, the value of the ratio under the null hypothesis that we are testing for, but the range extends

[8] This can often be the case when trying to use statistical tests to support decisions on statistical tests. Everything has a certain amount of error, and if not watched carefully, error can accummulate. Diligence, scrutiny, and common sense about how data look with respect to assumptions is often better than trusting still more black-box tests.

from 0.67 to 10.80—an order of magnitude—a confidence interval that doesn't inspire much confidence.

In this case, it isn't worth debating whether the variance ratio is significantly different from 1, because a cousin of the Student's t-test in Eq. 1, Welch's t-test (Eq. 2), accommodates unequal variance by using the standard error of each group instead of pooled standard deviation:

$$t_{\text{Welch's}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2}$$

Where

$$\begin{aligned} \bar{X} &= \text{group means,} \\ s^2 &= \text{group variance, and} \\ n &= \text{group sample size} \end{aligned}$$

Mathematical equations can look intimidating but this one really isn't—we have all these data already, back in Table 2. We can plug them into a single line of code to calculate Welch's t-statistic:

```
>       j.sum
```

```
     Sex  Mean Variance  n
1   Male 113.4    13.82 10
2 Female 108.6     5.16 10
```

```
>       X1 <- j.sum[1,2]
>       X2 <- j.sum[2,2]
>       var1 <- j.sum[1,3]
>       var2 <- j.sum[2,3]
>       n1 <- j.sum[1,4]
>       n2 <- j.sum[2,4]
>       welch.t <- (X1-X2)/sqrt((var1/n1) + (var2/n2))
>       welch.t
```

```
[1] 3.48412
```

Performing the test As easy as that was, however, we're only halfway there—we still need to use the degrees of freedom to determine a p-value for our new t-statistic. In the old days, this would involve finding the appropriate table in the back of a statistics book by matching the α level—the significance level or acceptable probability that the effect is the result of chance—with the degrees of freedom, essentially the statistical power. Another option is to again use α and the degrees of freedom to calculate the *criticalt value*, which represents the point on the tail of the t distribution beyond which we reject the null hypothesis:

```
> qt(0.05, df=19, lower.tail=FALSE)

[1] 1.729133
```

The critical t value for 19 degrees of freedom at an acceptable probability of 5% that t is observed by chance is 1.73. Our observed t, 3.48, is further out on the tail (Fig. 2). Thus $p<0.05$, and we reject the null hypothesis.

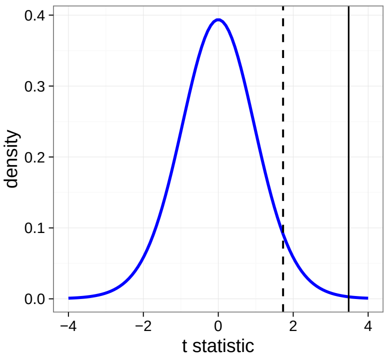

Figure 2: t distribution for α=0.05 at 19 degrees of freedom. Broken line represents the critical t value beyond which $p<0.05$; solid line represents our observed t, 3.48, so we can reject the null hypothesis.

Fortunately `R` has functions programmed with the equations to calculate a variety of test statistics and return exact p values:

```
> t.test(Length ~ Sex, jackal, var.equal=FALSE)

        Welch Two Sample t-test

data:  Length by Sex
t = 3.4843, df = 14.894, p-value = 0.00336
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.861895 7.738105
sample estimates:
  mean in group Male mean in group Female
               113.4                108.6
```

Let's go through these results:

- Right at the top —"`Welch Two Sample t-test`"—we see we're using Welch's t-statistic since we set `var.equal=FALSE`.
- Note that the t-statistic returned by `var.test` is very near the value we calculated "by hand," off only due to our rounding in `j.sum`
- `p < 0.05`, so we reject the null hypothesis that the difference in length of male and female golden jackals is 0.
- the 95% confidence interval includes our test statistic, which means that 95% of the time additional random draws from the golden jackal population will have a t-statistic in this range.

The difference between one-tailed and two-tailed tests Pay particular attention to how the alternative hypothesis is phrased: `true difference in means is not equal to 0`, which in other words means the alternative hypothesis is "the means are different." This is not the most specific hypothesis we could test—a more specific hypothesis would make an explicit prediction about which group will have the longer mandibles [9]. Doing so, however, would require us to have a good expectation of which group is larger—whether that expectation is based on knowledge of other species, some sort of ecological theory, or preliminary data—because testing such a hypothesis closes the door on getting a "positive" result if we are wrong. The door closes because by stating a specific expectation about which direction the relationship will go, we're limiting the test statistic to one side of the normal-ish t-distribution– hence, it is a *one-tailed* or *one-sided* test. Doing so concentrates our statistical power, which is why a one-sided test has a lower p-value than a two-sided test with the same test statistic and same degrees of freedom (Table 3)—assuming, of course, we set the direction correctly.

test	t	df	p
$\text{length}_{\text{male}} \neq \text{length}_{\text{female}}$	3.48	14.89	0.003
$\text{length}_{\text{male}} > \text{length}_{\text{female}}$	3.48	14.89	0.002
$\text{length}_{\text{male}} < \text{length}_{\text{female}}$	3.48	14.89	0.998

Table 3: Different alternative hypotheses (`tests`) yield much different p-values despite identical t-statistics and degrees of freedom. The different forms of alternative hypotheses are available through the `alternative=` argument in `t.test`, which can be defined as `'two.sided'`, `'less'`, or `'greater'`.

Linear regression

We've seen how the t-test makes simple work out of comparing mean values of a continuous variable across two groups, but we need more if we want to determine whether two continuous variables are related to each other. Such tests are called *regression*, and in some respects the same theory as the t-test applies—calculate ratios between the signal and the noise and determine the probability of the signal being due to chance.

Linear regression shares the same assumptions of the t-test and adds two more; the first is the assumption that the relationship between the two continuous variables can be described with a line. We want our model to draw this line such that it comes as close as possible to as many of the datapoints as possible. The more linear the relationship, the less distance between the average point and the line, and we can basically describe remaining error by summing all these distances. In fact, this begets the name of the basic approach to linear regression, *Ordinary Least Squares*: the model seeks the lowest (least) sum of error, which is squared to account for the fact that just as many points will fall below the line as above it, and without being squared would remain negative and subtract from total error.

The anatomy of a line Recall that we only need two bits of information to draw a line, the *intercept*—where the line crosses the Y

[9] This is, in fact, how Sir Ronald Fisher would have it: to give a strict true/false result and still be informative, the research hypothesis statement needs to be as explicit as possible.

axis—and the *slope*—essentially the angle the line makes as it moves away from the Y axis. In high school math we expressed these as

$$y = mx + b$$

Where

$$\begin{aligned} y &= \text{Y axis coordinate,} \\ m &= \text{slope of the line,} \\ x &= \text{X axis coordinate, and} \\ b &= \text{the intercept, or Y value when x=0} \end{aligned}$$

In statistics our notation is a bit fancier, but the same underlying equation is apparent:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{3}$$

Where

$$\begin{aligned} y_i &= \text{Y axis coordinate for point } i, \\ \beta_0 &= \text{the intercept, or Y value when x=0,} \\ \beta_1 &= \text{the slope, or the } \textit{regression coefficient}, \\ x_i &= \text{X axis coordinate for point } i, \text{ and} \\ \epsilon_i &= \text{an error term expressing the fact that given the variability} \\ &\quad \text{inherent in our data, we cannot draw a straight line.} \end{aligned}$$

Drawing our regression line is fairly simple– just as a line is defined by two bits of information, the slope and the intercept, it will always go through two points: it will cross the Y axis at the intercept, $\hat{\beta}_0$, and it will pass through the intersection of the mean of both variables. To calculate the intercept we must first have the regression coefficient, $\hat{\beta}_1$, which is straightforward, albeit potentially tedious, as it is basically comprised of several instances of taking the sum of every datapoint in X and Y minus the mean of the variable:

$$\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \tag{4}$$

The intercept comes from subtracting the product of the regression coefficient and the mean of x from y—taking Eq. 3 and solving for $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{5}$$

We can do this with one of our favorite datasets, `mtcars`:

```
> x <- mtcars$hp
> y <- mtcars$mpg
> reg.d <- data.frame(x, y, xdev=(x-mean(x)), ydev=(y-mean(y)),
+             xdevydev=((x-mean(x))*(y-mean(y))),
+             xdev2=(x-mean(x))^2,
+             ydev2=(y-mean(y))^2)
> SP <- sum(reg.d$xdevydev)
> SSx <- sum(reg.d$xdev2)
> SSy <- sum(reg.d$ydev2)
> (b1 <- SP / SSx)

[1] -0.06822828

> (b0 <- mean(y) - b1*mean(x))

[1] 30.09886
```

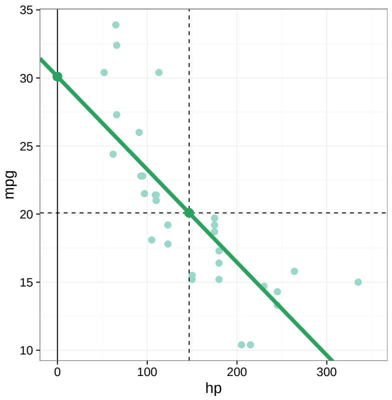

Figure 3: Fitting a regression line is remarkably easy: find the mean of both variables (intersection of broken lines) and calculate the y intercept following Eq. 5. The trick comes in determining how useful this model is for explaining variation in the data.

Thus the equation for the model describing the linear relationship between `mpg` and `hp` in the `mtcars` (Fig. 3) dataset is:

$$y_i = 30.099 - (0.068)x_i \tag{6}$$

Explaining variation The first question one might ask about our model should be: How well does it fit the data? One measurement is the *Coefficient of Determination*, or R^2. R^2 is the square of the Pearson's correlation coefficient, r (Eq. 7). The correlation coefficient r gives the sign of the association, and when squared, R^2 represents the proportion of variation explained by the statistical model.

$$r = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}} \tag{7}$$

Eq. 7 can be re-written using terms we used above as an expression as the ratio of signal:noise, or effect (SP) over variance (cumulative Sums of Squares):

$$r = \frac{\text{SP}}{\sqrt{\text{SS}_x \cdot \text{SS}_y}} \tag{8}$$

Where

$$\begin{aligned} \text{SP} &= \text{Sum of Products}, \\ \text{SS}_x &= \text{Sum of Squares for x}, \\ \text{SS}_y &= \text{Sum of Squares for y} \end{aligned}$$

Eq. 8 is easily coded into `R`:

Base `R` provides functions to calculate correlation coefficients (`cor`; default metric is `pearson`) and test their significance (`cor.test`).

```
> (r = SP /(sqrt(SSx*SSy)) )
```

```
[1] -0.7761684

> (r.sq = r^2)*100

[1] 60.24373
```

Thus, the linear model explains just over 60% of variation in the relationship between fuel economy and engine power in the `mtcars` dataset. But the question still remains: Is this relationship considered significant—*e.g.* is there a better-than-chance probability that repeated sampling will produce a similar result and give us reason to reject the null hypothesis, that there is no relationship between fuel economy and engine power?

Again, the null hypothesis significance test consists of

- calculating a t statistic (here, based on r)
- determining the critical t value

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} \tag{9}$$

Compare the numerator and denominator in Eq. 9 to Eq. 1. It is still effect/variance, signal:noise, but expressed here as the strength of the correlation conditioned by the degrees of freedom (n–2 for the two variables), divided by the *remaining, unexplained variance*, 1-R^2.

```
> (t.stat = (r*sqrt(30)) / sqrt(1-r.sq)    )

[1] -6.742389

> qt(0.05, df=30, lower.tail=TRUE)

[1] -1.697261
```

p<0.05, so we reject the null hypothesis that there is no linear correlation between fuel economy and engine power (Fig. 4).

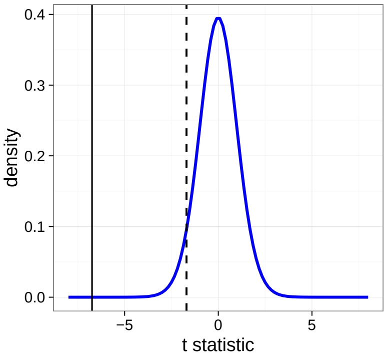

Figure 4: t distribution for α=0.05 at 30 degrees of freedom. Broken line represents the critical t value beyond which p<0.05; solid line represents our observed t, -6.74, so we reject the null hypothesis.

Fitting linear models with `lm` R provides a convenient function for fitting linear models: `lm`, which automates all of the test statistics calculated above. It operates on the formula interface `y` $\sim$ `x` and has two print methods; we begin with `summary`:

```
> lm1 <- lm(mpg ~ hp, mtcars)
> summary(lm1)

Call:
lm(formula = mpg ~ hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7121 -2.1122 -0.8854  1.5819  8.2360 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,       Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Let's go through these results:

- `Call` reminds us which model was fit.
- `Residuals` reports quantiles for the *residual error*, or the remaining variance not explained by the linear relationship.
- `Coefficients` are the meat of the linear regression. We calculated the estimates ourselves, above: `Intercept` = β_0, the Y intercept; `hp` = β_1, the slope of the regression line.
- Notice that `t value` is exactly what we calculated above, -6.742, and we're given a precise p value– `Pr(>|t|)`. Three asteriks indicate the p value falls between 0 and 0.001; this column helps one quickly scan `summary` results and assess significance. *Test statistics for the intercept term simply report whether the Y intercept is significantly different from 0, which it is in this model.*
- Along the bottom, not in table form, are results of another test we'll discuss shortly. For now, notice that the `Multiple R-squared` value = 0.6024, which we calculated above.

Generally speaking, we ignore `Adjusted R-squared`. Its intentions are good, but we have a better way of doing what it does—self-penalizing the addition of terms to the linear model—with information-theoretic model selection, to be discussed soon.

Testing model assumptions In addition to assumptions of normality and independence of samples, just like Student's t-test, linear regression assumes *homogeneity of variance*, *i.e.* variance remains constant as the mean increases. Violation of this assumption is called *heteroscedasticity*.[10]

Here we discuss two ways to check the important assumptions of linear regression. First we make a Q-Q plot of the residuals. We like to see them fit as closely as possible along a nice line, and at the very least, keep the confidence interval straddling the trendline (Fig. 5).

```
> library(car)
> qqPlot(lm1)
```

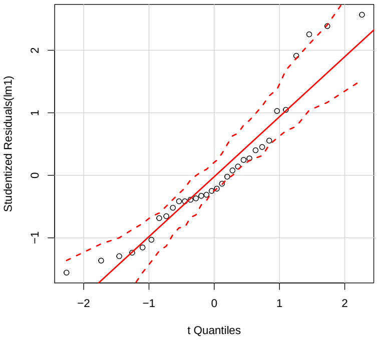

Figure 5: A Q-Q plot of the residuals as a visual inspection of equal variance.

A second approach is to use the Global Validation of Linear Models Assumptions function `gvlma` from the `gvlma` package. It reports

[10] Unfortunately there isn't a Welch's version of linear regression, although transformations like Box-Cox and often log transformations to improve normality also improve variance.

whether several assumptions are met or not; many models will fail the link function test and thus sink the global test, but we recommend focusing on Skewness—the third moment of data behind mean and variance, *skewness* is a fancy statistian word for whether the response variable is normally distributed—and Heteroscedasticity. *Kurtosis* is the fourth data moment and reflects the roundedness of the data distribution—whether it is steep, round, or flat.

```
> library(gvlma)
> summary(gvlma(lm1))

Call:
lm(formula = mpg ~ hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7121 -2.1122 -0.8854  1.5819  8.2360 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,	Adjusted R-squared:  0.5892 
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05 

Call:
 gvlma(x = lm1) 

                       Value   p-value                   Decision
Global Stat        18.845445 0.0008428 Assumptions NOT satisfied!
Skewness            2.977924 0.0844075    Assumptions acceptable.
Kurtosis            0.005681 0.9399169    Assumptions acceptable.
Link Function      12.369182 0.0004365 Assumptions NOT satisfied!
Heteroscedasticity  3.492659 0.0616415    Assumptions acceptable.
```

This model passes for skewness and heteroscedasticity, and since the residual Q-Q plot looks good (Fig. 5), we can wrap up this linear regression confident that we've fit a useful model.

Extending the linear model

It turns out the linear model—and in R its workhorse function, lm—forms the basis for many, if not most, of the tests we perform within the framework of frequentist inference. We discuss two extensions here: multiple regression and general analysis of variance (ANOVA) using the linear model.

Multiple linear regression

So far we've discussed linear regression as a single response variable tested against a single predictor variable. But often we want to test two or more predictor variables simultaneously and it is easy to do so within the linear regression framework by adding additional terms to the linear model. *Multiple regression* simply refers to testing one continuous reponse variable against one or more predictor variables p:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \qquad (10)$$

This is easily accomplished through the formula interface in lm:

```
> lm2 <- lm(mpg ~ hp + am, mtcars)
> summary(lm2)

Call:
lm(formula = mpg ~ hp + am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3843 -2.2642  0.1366  1.6968  5.8657

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.584914   1.425094  18.655  < 2e-16 ***
hp          -0.058888   0.007857  -7.495 2.92e-08 ***
ammanual     5.277085   1.079541   4.888 3.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.909 on 29 degrees of freedom
Multiple R-squared:  0.782,	Adjusted R-squared:  0.767
F-statistic: 52.02 on 2 and 29 DF,  p-value: 2.55e-10
```

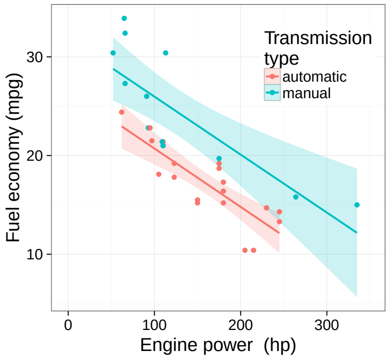

Figure 6: Multiple regression tests two terms with the same model. Here, a good alternative hypothesis is that at any given horsepower cars with manual transmissions will have better fuel economy, but across both transmission types fuel economy declines as engine power increases.

We see the `Coefficients` section of the results includes the new row `ammanual`, which refers to the additional variable we added to make our multiple regression model. The p values indicate the term is significant, but the `ammanual` is confusing. To better understand these results we turn away from the `summary` print option and towards `anova`.

Analysis of Variance

Conceptual background of ANOVA Analysis of variance—referred to so often by its acronym ANOVA that it is almost a word of its own and many journals don't even require the acronym to be defined—is a collection of statistical models that generalize the t-test to compare two or more groups by partitioning observed variance among terms defined as model parameters. Rather than a t-test, ANOVA is based on an F-test: F is the test statistic, follows a F distribution, and is defined as the ratio of variance *between* groups to variance *within* groups:

$$\begin{aligned} F &= \frac{\text{variance between groups}}{\text{variance within groups}} \\ &= \frac{\text{MS}_{\text{Treatments}}}{\text{MS}_{\text{Error}}} = \frac{\text{SS}_{\text{Treatments}}/(I-1)}{\text{SS}_{\text{Error}}/(n_T - I)} \end{aligned} \tag{11}$$

Where

$$\begin{aligned} \text{MS} &= \text{Mean Squares,} \\ \text{SS} &= \text{Sum of Squares,} \\ I &= \text{Number of groups (treatments),} \\ n_T &= \text{Total number of observations (sample size)} \end{aligned}$$

> *If there is more difference from group to group than there is among members within the groups, i.e. the groups form distinct clusters, the groups are considered significantly different.*

The ANOVA table ANOVA results are extracted from objects of class `lm` with the `anova` command. This produces an ANOVA table:

`R` also includes a wrapper function, `aov`, which can be used in place of `lm`. The ANOVA table is accessed via `summary`, *e.g.* `summary(aov(lm2))`. As far as we can tell this has no more than one use, but it is important, and will be discussed below.

```
> anova(lm2)

Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)
hp         1 678.37  678.37  80.153 7.627e-10 ***
am         1 202.24  202.24  23.895 3.460e-05 ***
```

```
Residuals 29 245.44    8.46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, instead of estimates of linear model coefficients and t-statistics, we have Sums of Squares (`Sum Sq`), Means of Squares (`Mean Sq`), and F-statistics. The Sums of Squares represent the amount of variance partitioned to each model term; unexplained variance is called Residual error and essentially constitutes ϵ in Eq. 10. Note that, as in Eq. 11, the Means of Squares are simply the Sums of Squares of a term divided by that term's degrees of freedom; each term gets one degree of freedom and the remainder are assigned to residual error. This penalizes residual error tremendously when Means of Squares are calculated, as residual error is divided by a much larger number than the test terms, and gives the signal a boost when divided by the noise to determine the F-statistic.

Recall those lines of text at the bottom of the `summary(lm2)` printout. You'll recognize the 2 and 29 degrees of freedom, and the F-statistic there, 52.02, is the mean of both F-statistics in the ANOVA table. When referring to the model fit, use $F_{2,29}$=50.02, p<0.001, R^2=0.78 from `summary(lm2)`; when referring to the significance of individual terms, use the results from `anova(lm2)`.

There are tradeoffs in adding terms to a multiple regression model. Each additional term is assigned one degree of freedom. On one hand, adding terms increases the influence of residual error on F by giving MS_{error} a smaller denominator, which reduces F. On the other hand, adding a "good" term will more than make up for the cost of one degree of freedom by absorbing variance in its own SS term, reducing SS_{error}, and giving MS_{error} a smaller numerator, which boosts F.

In the absence of continuous predictor Variables, ANOVA works on one or more categorical variables, as well:

```
> lm3 <- lm(mpg ~ am, mtcars)
> anova(lm3)

Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value   Pr(>F)
am         1 405.15  405.15   16.86 0.000285 ***
Residuals 30 720.90   24.03
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When given just more than a little thought, the use of a *linear* model, which evokes a regression line along a gradient, to test the difference between *group means* seems wacky. But it turns out that group-means comparison via ANOVA is just a special case of the linear model. Recall that:

1. In a linear model, the effect of term p is measured by the slope, β_p:

$$\textit{null hypothesis} = \text{no effect}, \ \beta_p = 0$$

$$\textit{alternative hypothesis} = \text{effect}, \ \beta_p \neq 0$$

2. Just two points are required to fit a line

Therefore, a flat line connecting two means would have zero slope, suggesting no difference between the means, while two different means would have some non-zero slope (Fig. 7). Whether or not those means are significantly different depends on variation around the means—the ratio of between-group variance to within-group variance (Eq. 11).

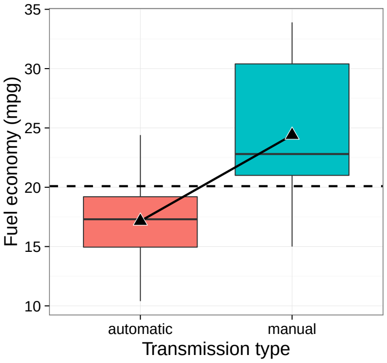

Figure 7: ANOVA tests the difference between two means using linear regression—testing whether the slope of the line connecting group means (solid line) is significantly different from the null hypothesis: no difference, or slope=0 (broken line).

But why not use the t-test? We already have a perfectly good way to compare difference in means, why add another? The reason is immediately clear when we consider (a) a term with *more than two groups*, and (b) more than one term. Testing, *e.g.*, three groups, would require a series of pairwise t-tests: Group A—Group B, Group A—Group C, Group B—Group C. Add additional groups and the number of tests increases geometrically, and the risk of getting a result due to chance increases. Furthermore, the statistical power of each test would be artificially reduced because it couldn't consider that the sample was in fact part of a larger dataset. ANOVA applies the theory of linear regression to again test for a significant deviation from a non-zero slope across any number of groups. But there are two different tests, and the way they work is not necessarily intuitive.

ANOVA on factors with three or more levels ANOVA—called by `anova(lm4)`—reports an overall statistic for predictor variable; in this example, the three-level variable `cyl` (Fig. 8):[11]

```
> lm4 <- lm(mpg ~ factor(cyl), mtcars)
> anova(lm4)

Analysis of Variance Table

Response: mpg
            Df Sum Sq Mean Sq F value    Pr(>F)
factor(cyl)  2 824.78  412.39  39.697 4.979e-09 ***
Residuals   29 301.26   10.39
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Meanwhile, `summary(lm4)` provides insight into how the linear model is adapted to the three groups:

```
> summary(lm4)
```

[11] At this point it is absolutely essential to pay attention to the `class` of the variable in `R`. When a variable like `cyl` is stored as `integer`, as is the default in `mtcars`, the default is for `lm` to treat it as a continuous variable. But we know engines have a discrete number of cylinders, and must either change the `class` in the `data.frame` or use `factor(cyl)` in all models.

```
Call:
lm(formula = mpg ~ factor(cyl), data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2636 -1.8357  0.0286  1.3893  7.2364

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   26.6636     0.9718  27.437  < 2e-16 ***
factor(cyl)6  -6.9208     1.5583  -4.441 0.000119 ***
factor(cyl)8 -11.5636     1.2986  -8.905 8.57e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.223 on 29 degrees of freedom
Multiple R-squared:  0.7325,	Adjusted R-squared:  0.714
F-statistic:  39.7 on 2 and 29 DF,  p-value: 4.979e-09
```

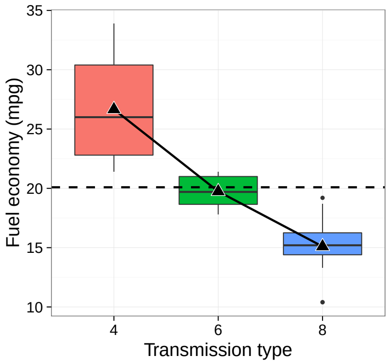

Figure 8: ANOVA tests the difference between two means using linear regression—testing whether the slopes of lines connecting group means (solid line) are significantly different from the null hypothesis: no difference, or slope=0 (broken line).

It isn't necessarily intuitive. Each group, or factor level, is added as a term in the linear model. In this way we can see the relative effect of each level but the output doesn't make it clear. Compare the coefficent estimates to group means; notice how the Intercept is equal to the mean fuel economy of 4-cylinder cars, and the coefficients of the next two terms—6- and 8-cylinder cars—are equal to both the magnitude and the sign of difference between their mean a fuel economy and that of 4-cylinder cars:

```
> cyl.means

cyl  4.00  6.00  8.0
mpg 26.66 19.74 15.1

> lm4$coefficients

 (Intercept) factor(cyl)6 factor(cyl)8
   26.663636    -6.920779   -11.563636
```

The negative values reflect the fact that fuel economy declines as the number of cylinders increases, which is consistent with the graph (Fig. 8), and ensures that the t-statistics have the appropriate sign and fall on the correct tail of the two-sided t-distribution. However the overall effect of `cyl` is not immediately clear; one needs the F statistic for that[12]. But the F-distribution is one-tailed and only positive, so looking at the F-statistic alone does not indicate whether the effect of cylinder number on fuel economy is positive or negative.

[12] Notice the same F, degrees of freedom, and p are returned by both `anova(lm4)` and `summary(lm4)`.

> *Interpreting ANOVA with categorical variables with multiple groups requires considerable attention to all of the information provided by the test results.In this way, the linear model constructs a line with negative slope.*

Post-hoc pairwise comparisons We mentioned above that pairwise comparisons—in the form of multiple t-tests among all groups—are undesirable, but pairwise comparisons are necessary to determine which group or groups are different from others once an overall effect has been determined by the ANOVA model. The solution is to conduct *post-hoc pairwise contrasts*, often with what is known as a Tukey test (or more formally, Tukey's Honest Significant Differences test).

The structure of the model is similar to that of a t-test except for the key feature that it corrects for *family-wise error rate*, which is the probability of getting a false positive given that several statistical comparisons are being made.

R provides a simple function for Tukey post-hoc comparisons, TukeyHSD, but we can't apply it directly to our linear model object of class lm; we need to use the alternate ANOVA wrapper, aov:

```
> TukeyHSD(aov(mpg ~ factor(cyl), mtcars))

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mpg ~ factor(cyl), data = mtcars)

$`factor(cyl)`
          diff        lwr        upr     p adj
6-4  -6.920779 -10.769350 -3.0722086 0.0003424
8-4 -11.563636 -14.770779 -8.3564942 0.0000000
8-6  -4.642857  -8.327583 -0.9581313 0.0112287
```

In these results, each row is a pairwise contrast between two groups/ factor levels and diff represents the value one gets from subtracting the mean of the second from the mean of the first. lwr and upr define the range of the 95% confidence interval for the difference in means, and p adj gives a p value that has been adjusted to be more strict given that multiple comparisons are being made. By presenting the sign of the differences in means, the results of TukeyHSD also provide similar information as the coefficients in summary(lm4) but in a more complete and clear way. Thus, in two lines of code, anova(lm(mpg ∼ factor(cyl), mtcars)) and TukeyHSD(aov(mpg ∼ factor(cyl), mtcars)) provide considerable information about the relationship of these variables.