4

Hypothesis testing using the likelihood ratio test

We started the book with the one-sample t-test. There, we had the following procedure:

- Given independent and identically distributed data y, define a null hypothesis: $H_0 : \mu = \mu_0$
- Compute the sample mean $\bar{y}$ and the standard error SE
- Reject the null hypothesis if the absolute value of $\bar{y}/SE$ is larger than 2.

Here, we turn to a closely related test: the *likelihood ratio test statistic*.

4.1 The likelihood ratio test: The theory

Suppose that $X_1, \dots, X_n$ are independent and normally distributed with mean μ and standard deviation σ (assume for simplicity that σ is known).

Let the null hypothesis be $H_0 : \mu = \mu_0$ and the alternative be $H_1 : \mu \neq \mu_0$. Here, μ_0 is a number, such as 0.

The likelihood of the data y can be computed under the null model, in which $\mu = \mu_0$, and under the alternative model, in which μ has some specific alternative value. To make this concrete, imagine 10 data points being generated from a Normal(0,1).

```
y<-rnorm(10)
```

We can compute the joint likelihood under a null hypothesis that $\mu = 0$:

```
likNULL<-prod(dnorm(y,mean=0,sd=1))
likNULL
```

```
## [1] 9.151e-06
```

On the log scale, we would need to add the log likelihoods of each data point:

```
loglikNULL<-sum(dnorm(y,mean=0,sd=1,log=TRUE))
loglikNULL
```

```
## [1] -11.6
```

Similarly, we can compute the log likelihood with μ equal to the maximum likelihood estimate of μ, the sample mean.

```
loglikALT<-sum(dnorm(y,mean=mean(y),sd=1,log=TRUE))
loglikALT
```

```
## [1] -11.59
```

Essentially, the likelihood ratio test compares the ratio of likelihoods of the two models; on the log scale, the difference in log likelihood is taken. This is because a ratio on the log scale becomes a difference. The likelihood ratio test then chooses the model with the higher log likelihood, provided that the higher likelihood is high enough (we will just make this more precise).

One can specify the test in general terms as follows. Suppose that the likelihood is with respect to some parameter θ. We can evaluate the likelihood at μ_0, the null hypothesis value of the parameter, and evaluate the likelihood using the maximum likelihood estimate

$\widehat{\theta}$ of the parameter, as we did above. The likelihood ratio can then be written as follows:

$$\Lambda = \frac{max_{\theta\in\omega_0}(lik(\theta))}{max_{\theta\in\omega_1)}(lik(\theta))} \tag{4.1}$$

where, $\omega_0 = \{\mu_0\}$ and $\omega_1 = \{\forall\mu \mid \mu \neq \mu_0\}$. The function max just selects the maximum value of any choices of parameter values; in the case of the null hypothesis there is only one value, μ_0. In the case of the alternative model, the maximum likelihood estimate $\widehat{\theta}$ is the maximum value.

Now, assuming that the data are coming from a normal distribution, the numerator of the likelihood ratio statistic is:

$$lik(\theta = \mu_0) = \frac{1}{(\sigma\sqrt{2\pi})^n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right) \tag{4.2}$$

For the denominator, the MLE $\bar{X}$ is taken as μ:

$$lik(\theta = \bar{X}) = \frac{1}{(\sigma\sqrt{2\pi})^n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \tag{4.3}$$

The likelihood ratio statistic is then:

$$\Lambda = \frac{lik(\theta = \mu_0)}{lik(\theta = \bar{X})} = \frac{\frac{1}{(\sigma\sqrt{2\pi})^n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right)}{\frac{1}{(\sigma\sqrt{2\pi})^n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)} \tag{4.4}$$

Canceling out common terms:

$$\Lambda = \frac{exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right)}{exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)} \tag{4.5}$$

Taking logs:

$$\begin{aligned}\log\Lambda &= \left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_0)^2\right) - \left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\\ &= -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}(X_i-\bar{X})^2\right)\end{aligned} \tag{4.6}$$

Now, it is a standard algebraic trick to rewrite $\sum_{i=1}^{n}(X_i-\mu_0)^2$ as a sum of two terms:

$$\sum_{i=1}^{n}(X_i-\mu_0)^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu_0)^2 \tag{4.7}$$

If we rearrange terms, we obtain:

$$\sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}(X_i-\bar{X})^2 = n(\bar{X}-\mu_0)^2 \tag{4.8}$$

Now, we just established above that $\log\Lambda$ is:

$$\log\Lambda = -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}(X_i-\bar{X})^2\right) \tag{4.9}$$

Consider the term in the brackets:

$$(\sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}(X_i-\bar{X})^2) \tag{4.10}$$

This can be rewritten as:

$$n(\bar{X}-\mu_0)^2 \tag{4.11}$$

Rewriting in this way gives us:

$$\log \Lambda = -\frac{1}{2\sigma^2} n(\bar{X} - \mu_0)^2 \tag{4.12}$$

Rearranging terms:

$$-2\log \Lambda = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2} \tag{4.13}$$

Or even more transparently:

$$-2\log \Lambda = \frac{(\bar{X} - \mu_0)^2}{\frac{\sigma^2}{n}} \tag{4.14}$$

This should remind you of the t-test! Basically, just like in the t-test, what this is saying is that we reject the null when $\mid \bar{X} - \mu_0 \mid$, or negative two times the difference in log likelihood, is large!

Now we will define what it means for $-2\log\Lambda$ to be large. There is a theorem in statistics that states that for large n, the distribution of $-2\log\Lambda$ approaches the chi-squared distribution, with degrees of freedom corresponding to the difference in the number of parameters between the two models being compared.

We will define the *likelihood ratio test statistic*, call it LRT, as follows. Here, $Lik(\theta)$ refers to the likelihood given some value θ for the parameter, and $logLik(\theta)$ refers to the log likelihood.

$$\begin{aligned} LRT &= -2 \times (Lik(\theta_0)/Lik(\theta_1)) \\ \log LRT &= -2 \times \{logLik(\theta_0) - logLik(\theta_1)\} \end{aligned} \tag{4.15}$$

where θ_1 and θ_0 are the estimates of θ under the alternative and null hypotheses, respectively. The likelihood ratio test rejects H_0 if $\log LRT$ is sufficiently large. As the sample size approaches infinity, $\log LRT$ approaches the chi-squared distribution:

$$\log LRT \rightarrow \chi^2_r \text{ as } n \rightarrow \infty \tag{4.16}$$

Here, r is called the degrees of freedom and is the difference in the number of parameters under the null and alternative hypotheses.

The above result is called *Wilks' theorem*. The proof of Wilks' theorem is fairly involved but you can find it in Lehmann's textbook *Testing Statistical Hypotheses*.

Note that sometimes you will see the form:

$$\log LRT = 2\{logLik(\theta_1) - logLik(\theta_0)\} \tag{4.17}$$

It should be clear that both statements are saying the same thing; in the second case, we are just subtracting the null hypothesis log likelihood from the alternative hypothesis log likelihood, so the negative sign disappears.

That's the theory. Let's see how the likelihood ratio test works for (a) simulated data, and (b) our running example, the English relative clause data from Grodner and Gibson (2005).

4.2 A practical example using simulated data

A practical example will make the usage of this test clear. Let's just simulate data from a linear model:

```
x<-1:10
y<- 10 + 20*x+rnorm(10,sd=10)
```

Here, the null hypothesis that the slope is 0 is false (it has value 20). Now, we fit a null hypothesis model, without a slope:

```
## null hypothesis model:
m0<-lm(y~1)
```

We will compare this model's log likelihood with that of the alternative model, which includes an estimate of the slope:

```
## alternative hypothesis model:
m1<-lm(y~x)
```

The difference in log likelihood, multiplied with -2, is:

```
LogLRT<- -2*(logLik(m0)-logLik(m1))
## observed value:
LogLRT[1]
```

```
## [1] 34.49
```

The difference in the number of parameters in the two models is one, so $\log LRT$ has the distribution χ_1^2. Is the observed value 34.49 unexpected under this distribution? We can calculate the probability of obtaining the likelihood ratio statistic we observed above, or a value more extreme, given the χ_1^2 distribution.

```
pchisq(LogLRT[1],df=1,lower.tail=FALSE)
```

```
## [1] 4.286e-09
```

Just like the critical t-value in the t-test, the critical chi-squared value here is:

```
## critical value:
qchisq(0.95,df=1)
```

```
## [1] 3.841
```

If minus two times the observed difference in log likelihood is larger than this critical value, we reject the null hypothesis.

Note that in the likelihood test above, we are comparing one nested model against another: the null hypothesis model is nested inside the alternative hypothesis model. What this means is that the alternative hypothesis model contains all the parameters in the null hypothesis model (i.e., the intercept) plus another one (the slope).

4.3 A real-life example: The English relative clause data

The likelihood ratio test is also the way that hypothesis testing is done with the linear mixed model. Here is how it works. Let's look again at the Grodner and Gibson (2005) English relative clause data. The null hypothesis here refers to the slope parameter. When we have the sum contrast coding, the intercept β_0 refers to the grand mean, and the slope β_1 is the amount by which subject and object relative clause mean reading times deviate from the grand mean. Testing the null hypothesis that β_1 is 0 amounts to testing whether there is any difference in means between the two relative clause types. This becomes clear if we consider the following.

Let object relatives be coded as +1 and subject relatives as −1. Then, the mean reading time μ_{or} for object relatives in the linear mixed model is:

$$\mu_{or} = \beta_0 + \beta_1 \tag{4.18}$$

Similarly, the mean reading time μ_{sr} for subject relatives is:

$$\mu_{sr} = \beta_0 - \beta_1 \tag{4.19}$$

If the null hypothesis is that $\mu_{or} - \mu_{sr} = 0$, then this amounts to saying that:

$$(\beta_0 + \beta_1) - (\beta_0 - \beta_1) = 0 \tag{4.20}$$

Removing the brackets gives us:

$$\beta_0 + \beta_1 - \beta_0 + \beta_1 = 0 \tag{4.21}$$

This yields the equation:

$$2\beta_1 = 0 \tag{4.22}$$

Dividing both sides of the equation by 2, we get the null hypothesis that $\beta_1 = 0$.

Incidentally, if we had rescaled the contrast coding to be not ±1 but $\pm 1/2$, the parameter β_1 would represent exactly the difference between the two means, and null hypothesis in equation (4.22) would have come out to be $\beta_1 = 0$. This is why it is sometimes better to recode the contrasts as $\pm 1/2$ rather than ±1. See Schad et al. (2020a) for details; we will discuss this in the contrast coding chapter as well.

Let's load the data, set up the contrast coding, and fit the null versus the alternative models. We will fit varying intercept and varying slopes for subject and item, without correlations for items. We don't attempt to fit a model with by-items varying intercepts and slopes with a correlation because we would get a singularity in the variance covariance matrix.

```r
gg05e1<-read.table("data/grodnergibsonE1crit.txt",header=TRUE)

gg05e1$so <- ifelse(gg05e1$condition=="objgap",1,-1)
gg05e1$logrt<-log(gg05e1$rawRT)

library(lme4)
m0<-lmer(logrt~1 + (1+so|subject)+(1+so||item),gg05e1)
m1<-lmer(logrt~1 + so + (1+so|subject)+(1+so||item),gg05e1)
```

Notice that we keep all random effects in the null model. We say that the null model is nested inside the full model.

Next, we compare the two models' log likelihoods. There is a function in the `lme4` package that achieves that: the `anova` function:

```
anova(m0,m1)
```

```
## refitting model(s) with ML (instead of REML)

## Data: gg05e1
## Models:
## m0: logrt ~ 1 + (1 + so | subject) + ((1 | item) + (0 + so | item))
## m1: logrt ~ 1 + so + (1 + so | subject) + ((1 | item) + (0 + so |
## m1:     item))
##    npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
## m0    7 707 739   -347      693
## m1    8 703 739   -343      687  6.15  1      0.013
```

You can confirm from the output that the `Chisq` value shown is minus two times the difference in log likelihood of the two models. The p-value is computed using the chi-squared distribution with one degree of freedom because in the two models the difference in the number of parameters is one:

```
round(pchisq(6.15,df=1,lower.tail=FALSE),3)
```

```
## [1] 0.013
```

It is common in the psycholinguistics literature to use the t-value from the linear mixed model output to conduct the hypothesis test on the slope:

```
summary(m1)$coefficients
```

```
##             Estimate Std. Error t value
## (Intercept)  5.88306    0.05176 113.669
## so           0.06202    0.02422   2.561
```

The most general method for hypothesis testing is the likelihood ratio test shown above. One can use the t-test output from the linear mixed model for hypothesis testing, but this should be done only when the data are balanced. If there is lack of balance (e.g., missing data for whatever reason), the likelihood ratio test is the best way to proceed. In any case, when we talk about the evidence against the null hypothesis, the likelihood ratio test is the only reasonable way to talk about what evidence we have. See Royall (1997) for more discussion of this point. The essence of Royall's point is that the most reasonable way to talk about the evidence in favor of a particular model is with reference to, i.e., relative to, a baseline model.

One can also use the likelihood ratio test to evaluate whether a variance component should be included or not. For example, is the correlation parameter justified for the subjects random effects? Recall that we had a correlation of 0.58. Is this statistically significant? One can test this in the following way:

```
m1<-lmer(logrt~1 + so + (1+so|subject)+(1+so||item),gg05e1)
m1NoCorr<-lmer(logrt~1 + so + (1+so||subject)+(1+so||item),gg05e1)
anova(m1,m1NoCorr)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: gg05e1
## Models:
## m1NoCorr: logrt ~ 1 + so + ((1 | subject) + (0 + so | subject)) + ((1 |
## m1NoCorr:     item) + (0 + so | item))
## m1: logrt ~ 1 + so + (1 + so | subject) + ((1 | item) + (0 + so |
## m1:     item))
##          npar AIC BIC logLik deviance Chisq Df
## m1NoCorr    7 710 741   -348      696
## m1          8 703 739   -343      687   8.7  1
##          Pr(>Chisq)
## m1NoCorr
## m1           0.0032
```

The test indicates that we can reject the null hypothesis that the correlation parameter is 0. We will return to this parameter in the chapter on simulation.

4.4 Exercises

4.4.1 Chinese relative clauses

Load the following two data-sets:

```
gibsonwu<-read.table("data/gibsonwucrit.txt",header=TRUE)
gibsonwu2<-read.table("data/gibsonwu2012datarepeat.txt",header=TRUE)
```

The data are taken from two experiments that investigate (inter alia) the effect of relative clause type on reading time in Chinese. The data are from Gibson and Wu (2013) and Vasishth et al. (2013) respectively. The second data-set is a direct replication attempt of the first.

Chinese relative clauses are interesting theoretically because they are prenominal: the relative clause appears before the head noun.

As iscussed in Gibson and Wu (2013), the consequence of Chinese relative clauses being prenominal is that the distance between the gap in relative clause and the head noun is larger in subject relatives than object relatives. Hsiao and Gibson (2003) were the first to suggest that the larger distance in subject relatives leads to longer reading time at the head noun. Under this view, the prediction is that subject relatives are harder to process than object relatives. If this is true, this is interesting because in most other languages that have been studied, subject relatives are easier to process than object relatives; so Chinese will be a very unusual exception cross-linguistically.

The data provided are for the critical region (the head noun). The experiment method is self-paced reading, so we have reading times in milliseconds.

The research hypothesis is whether the difference in reading times between object and subject relative clauses is negative. For both data-sets, investigate this question by (a) fitting a paired t-test (by-subjects and by items), (b) fitting the most complex linear mixed model you can to the data and then interpreting the t-value, and (c) the likelihood ratio test. What can we conclude about the research question?

4.4.2 Agreement attraction in comprehension

Load the following data:

```
datE1<-read.table("data/dillonE1.txt",header=TRUE)
```

The data are taken from an experiment that investigate (inter alia) the effect of number similarity between a noun and the auxiliary verb in sentences like the following. There are two levels to a factor called Int(erference): low and high.

(a) low: The key to the cabinet *are* on the table
(b) high: The key to the *cabinets are* on the table

Here, in (b), the auxiliary verb *are* is predicted to be read faster than in (a), because the plural marking on the noun *cabinets* leads the reader to think that the sentence is grammatical. (Note that both sentences are ungrammatical.) This phenomenon, where the high condition is read faster than the low condition, is called **agreement attraction**.

The data provided are for the critical region (the auxiliary verb *are*). The experiment method is eyetracking; we have total reading times in milliseconds.

The research question is whether the difference in reading times between high and low conditions is negative.

- First, figure out which linear mixed model is appropriate for these data (varying intercepts only? varying intercepts and slopes? with or without correlations?).

- Then, carry out a statistical test using (a) the paired t-test (using the t.test function), (b) the t-test of the linear mixed model, and (c) the likelihood ratio test. What is your conclusion? Is there evidence for agreement attraction in the data?

4.4.3 The grammaticality illusion

Load the following data-sets:

```
english<-read.table("data/embeddingenglish.txt",header=TRUE)
dutch<-read.table("data/embeddingdutch.txt",header=TRUE)
```

In an offline accuracy rating study on English double center-embedding constructions, Gibson and Thomas (1999) found that grammatical constructions (e.g., example a below) were no less acceptable than ungrammatical constructions (e.g., example b) where a middle verb phrase (e.g., was cleaning every week) was missing.

(a) The apartment that the maid who the service had sent over was cleaning every week was well decorated.

(b) *The apartment that the maid who the service had sent over — was well decorated

Based on these results from English, Gibson and Thomas (1999) proposed that working-memory overload leads the comprehender to forget the prediction of the upcoming verb phrase (VP), which reduces working-memory load. This came to be known as the *VP-forgetting hypothesis*. The prediction is that in the word immediately following the final verb, the grammatical condition (which is coded as +1 in the data-frames) should be harder to read than the ungrammatical condition (which is coded as -1).

The data provided above test this hypothesis using self-paced reading for English (Vasishth et al., 2011), and for Dutch (Frank et al., 2015). The data provided are for the critical region (the noun phrase, labeled NP1, following the final verb). We have reading times in log milliseconds.

Is there support for the VP-forgetting hypothesis cross-linguistically, from English and Dutch?