Chapter 17

Relative Distributions and Smooth Tests of Goodness-of-Fit

In §16.2.2, we saw how to use the quantile function to turn uniformly-distributed random numbers into random numbers with basically arbitrary distributions. In this chapter, we will look at two closely-related data-analysis tools which go the other way, trying to turn data into uniformly-distributed numbers. One of these, the **smooth test**, turns a lot of problems into ones of testing a uniform distribution. Another, the **relative distribution**, gives us a way of comparing whole distributions, rather than specific statistics (like the expectation or the variance).

17.1 Smooth Tests of Goodness of Fit

17.1.1 From Continuous CDFs to Uniform Distributions

Suppose that X has probability density function f, and that f is continuous. The corresponding cumulative distribution function F is then continuous and strictly increasing (on the support of f). Since F is a fixed function, we can ask what the probability distribution of $F(X)$ is. Clearly,

$$\Pr(F(X) \leq 0) = 0 \quad (17.1)$$

$$\Pr(F(X) \leq 1) = 1 \quad (17.2)$$

Since F is continuous and strictly increasing, it has an inverse, the quantile function Q, which is also continuous and strictly increasing. Then, for $0 \leq a \leq 1$,

$$\begin{aligned} \Pr(F(X) \leq a) &= \Pr(Q(F(X)) \leq Q(a)) & (17.3) \\ &= \Pr(X \leq Q(a)) & (17.4) \\ &= F(Q(a)) = a & (17.5) \end{aligned}$$

Thus, when F is continuous and strictly-increasing, $F(X)$ is uniformly distributed on the unit interval,

$$F(X) \sim \mathrm{Unif}(0,1) \tag{17.6}$$

If the distribution of X is F, but we guess that it has some other distribution, with CDF F_0, then this trick will not work. $F_0(X)$ will still be in the unit interval, but it won't be uniformly distributed:

This only works if X really is distributed according to F. If instead X were distributed according, say, F_0, then $F(X)$ will still be in the unit interval, but it will not be uniformly distributed:

$$\begin{aligned} \Pr(F_0(X) \leq a) &= \Pr(X \leq Q_0(a)) & (17.7) \\ &= F(Q_0(a)) \neq a & (17.8) \end{aligned}$$

because $F_0 \neq Q^{-1}$.

Putting this together, we see that when X has a continuous distribution, $F(X) \sim \mathrm{Unif}(0,1)$ if and only if F is the cumulative distribution function for X. This means that we can reduce the problem of testing whether $X \sim F$ to that of testing whether $F(X)$ is uniform. We need to work out *one* testing problem, rather than many different testing problems for many different distributions.

17.1.2 Testing Uniformity

Now we have a random variable, say Y, which lives on the unit interval $[0,1]$, and we want to test whether it is uniformly distributed. There are several different ways we could do this. One frequently-used strategy is to use the Kolmogorov-Smirnov test: calculate the K-S distance,

$$d_{KS} = \max_{a \in [0,1]} \left| \widehat{F}_{n,Y}(a) - a \right| \tag{17.9}$$

where $\widehat{F}_{n,Y}(a)$ is the empirical CDF of Y, and look up the appropriate p-value for the K-S test. One could use any other one-sample non-parametric test here, like Cramér-von Mises or Anderson-Darling[1] All of these tests can work quite well in the right circumstances, and they have the advantage of requiring little additional work over and above typing `ks.test` or the like.

17.1.3 Neyman's Smooth Test

There are however two disadvantages of just applying off-the-shelf tests to check uniformity. One is that it turns out that they often do not have very high power. The other, which is in some ways even more serious, is that rejecting the null hypothesis of uniformity doesn't tell you *how* uniformity fails — it doesn't suggest any sort of natural alternative.

[1] You could even use a χ^2 test, but this would be dumb. Because the χ^2 test requires discrete data, using it means binning continuous values, thereby destroying information, to no good purpose.

As you can guess from my having brought up these points, there is a test which avoids both difficulties, called **Neyman's smooth test**. It works by embedding the uniform distribution on the unit interval in a larger class of alternatives, and then testing the null of uniformity against those alternatives.

The alternatives all have pdfs of the form

$$g(y;\theta) \equiv \begin{cases} \frac{e^{\sum_{j=1}^{d}\theta_j h_j(y)}}{z(\theta)} & 0 \leq y \leq 1 \\ 0 & \text{elsewhere} \end{cases} \tag{17.10}$$

where the h_j are carefully chosen functions (see below), and the **normalizing factor** or **partition function** $z(\theta)$ just makes sure the density integrates to 1:

$$z(\theta) \equiv \int_0^1 e^{\sum_{j=1}^{d}\theta_j h_j(y)} dy \tag{17.11}$$

No matter what functions we pick for the h_j, uniformity corresponds to the choice $\theta = 0$, since then the density is just 1. As we move θ slightly away from 0, the density departs *smoothly* from uniformity; hence the name of the test.

To ensure that everything works out, we need to put some requirements on the functions h_j: they need to be **orthogonal** to each other and to the constant function,

$$\int_0^1 h_j(y)dy = 0 \tag{17.12}$$

$$\int_0^1 h_j(y)h_k(y)dy = 0 \tag{17.13}$$

and **normalized** in magnitude,

$$\int_0^1 h_j^2(y)dy = 1 \tag{17.14}$$

Further details, while practically important, do not matter for the general idea of the test, so I'll put them off to §17.1.3.

We can estimate θ by maximum likelihood. Because uniformity corresponds to $\theta = 0$, we can test the hypothesis that $\theta = 0$ against the alternative that $\theta \neq 0$ with a likelihood ratio test. Writing $\ell(\hat{\theta})$ for the log-likelihood under the MLE, and $\ell(0)$ for the log-likelihood under the null, by general results on the likelihood-ratio (Appendix B), under the null, as $n \to \infty$,

$$2(\ell(\hat{\theta}) - \ell(0)) \rightsquigarrow \chi_d^2 \tag{17.15}$$

In fact, $\ell(0) = 0$ (why?), so we only need to calculate the log-likelihood under the alternative, and reject uniformity when, and only when, that log-likelihood is large.

Alternatively, and this was Neyman's original recommendation and what is usually meant by his "smooth test", we can calculate the sample mean of each of the h_j,

$$\overline{h_j} = \frac{1}{n}\sum_{i=1}^{n} h_j(y_i) \tag{17.16}$$

and form the test statistic

$$\Psi^2 = n\sum_{j=1}^{n} \overline{h_j}^2 \tag{17.17}$$

which also has a χ^2_d distribution under the null.[2]

It can be shown that Neyman's smooth test has, in a certain sense, optimal power against smooth alternatives like this — see Rayner and Best (1989) or Bera and Ghosh (2002) for the gory details. More importantly, for data analysis, when we reject the null hypothesis of uniformity, we have a ready-made alternative to fall back on, namely $g(y;\hat{\theta})$.

To make all this work, we have to pick some "basis functions" h_j, and we need to decide how many of them we want to use, d.

Choice of Function Basis

Neyman's original proposal was to use **orthonormal polynomials** for basis functions: h_j would be a polynomial of degree j, which was orthogonal to all the ones before it,

$$\int_0^1 h_j(y)h_k(y)dy = 0\ \forall k < j \tag{17.18}$$

including the constant "polynomial" $h_0(y) = 1$, and normalized to size 1,

$$\int_0^1 h_j^2(y)dy = 1 \tag{17.19}$$

Since there are $j+1$ coefficients in a polynomial of degree j, and this gives $j+1$ equations, the polynomial is uniquely determined. In fact, there are recursively formulas which let you find the coefficients of h_j from those of the previous polynomials[3]. Figure 17.1 shows the first few of these polynomials, and their exponentiated versions (which are what appear in Eq. 17.10).

[2] To appreciate what's going on, notice that $\overline{h_j} \rightarrow 0$ under the null, by the law of large numbers. (This is where being orthogonal to the constant function $h_0(y) = 1$ comes in.) Multiplying $\overline{h_j}^2$ by n corresponds to looking at $\sqrt{n}\overline{h_j}$, which should, by the central limit theorem, be a Gaussian; the variance of this Gaussian is 1. (This is where normalizing each h_j comes in.) Finally, $\sqrt{n}\overline{h_j}$ and $\sqrt{n}\overline{h_k}$ are uncorrelated. (This is where the mutual orthogonality of the h_j comes in.) Thus, the Ψ^2 statistic is a sum of d uncorrelated standard Gaussians, which has a χ^2_d distribution.

[3] In fact, the polynomials Neyman proposed to use are, as he knew, the "Legendre polynomials", though many math books (and Wikipedia) give the version of those defined on $[-1,1]$, rather than on $[0,1]$. If l_j is the polynomial on $[-1,1]$, then $h_j(y) = l_j(2(y-0.5))$.

Experience has shown that the specific choice of basis functions doesn't matter as much as ensuring that they are orthonormal. One could, for instance, use $h_j(y) = c_j \cos 2\pi j y$, where c_j is a normalizing constant[4].

Choice of Number of Basis Functions

As we make d in Eq. 17.10, we include more and more distributions in the alternative to the null hypothesis of uniformity. In fact, since any smooth function on $[0,1]$ can be approximated arbitrarily closely by sufficiently-high order polynomials[5], as we let $d \rightarrow \infty$ we eventually get *all* continuous distributions, other than uniformity, as part of the alternative. However, using a large value of d means estimating a lot of parameters, which means we are at risk of over-fitting. What to do?

Neyman's original advice was to guess a particular value of d before looking at the data and stick to it. (He thought $d = 4$ would usually be enough.) More modern approaches try to adaptively pick a good value of d. We could attempt this through cross-validation based on the log-likelihood, but what's usually done, in implemented software, is to pick d to maximize Schwarz's information criterion:

$$d^* = \operatorname*{argmax}_{d} \frac{1}{n}\ell(\widehat{\theta}^{(d)}) - \frac{d}{2}\frac{\log n}{n} \tag{17.20}$$

which imposes an extra penalty for each parameter (d), with the size of the penalty depending on how much data we have, and getting relatively harsher as n grows[6]. So in a **data-driven smooth test** (Kallenberg and Ledwina, 1997), we pick d^* using Eq. 17.20, and then compute the test statistic using d^*.

Unfortunately, since d^* is random (through the data), the nice asymptotic theory which says that the test statistic is χ^2_d under the null hypothesis no longer applies. However, this is why we have bootstrapping: by simulating from the null hypothesis, which remember is just Unif$(0,1)$, and treating the simulation output like real data we can work out the sampling distribution as accurately as we need. This sampling distribution then gives us our p-values.

Application: Combining p-Values

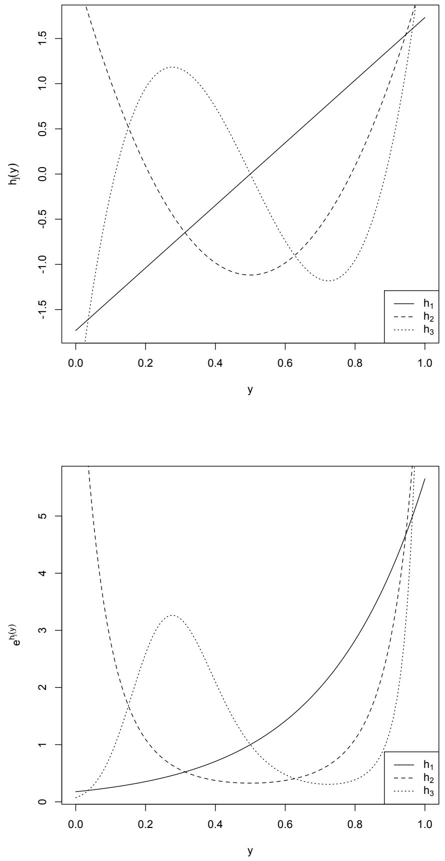
One useful property of p-values is that they are always uniformly distributed on $[0,1]$ under the null hypothesis[7]. Suppose we have conducted a bunch of tests of the same null hypothesis — these might be different clinical trials of the same drug, or

[4] If this makes you think of Fourier analysis, you're right.

[5] This may be obvious, but making it precise (what do we mean by "smooth" and "arbitrarily close"?) is the "Stone-Weierstrauss theorem". There is nothing magic about polynomials here; we could also use sines and cosines, or many other function bases.

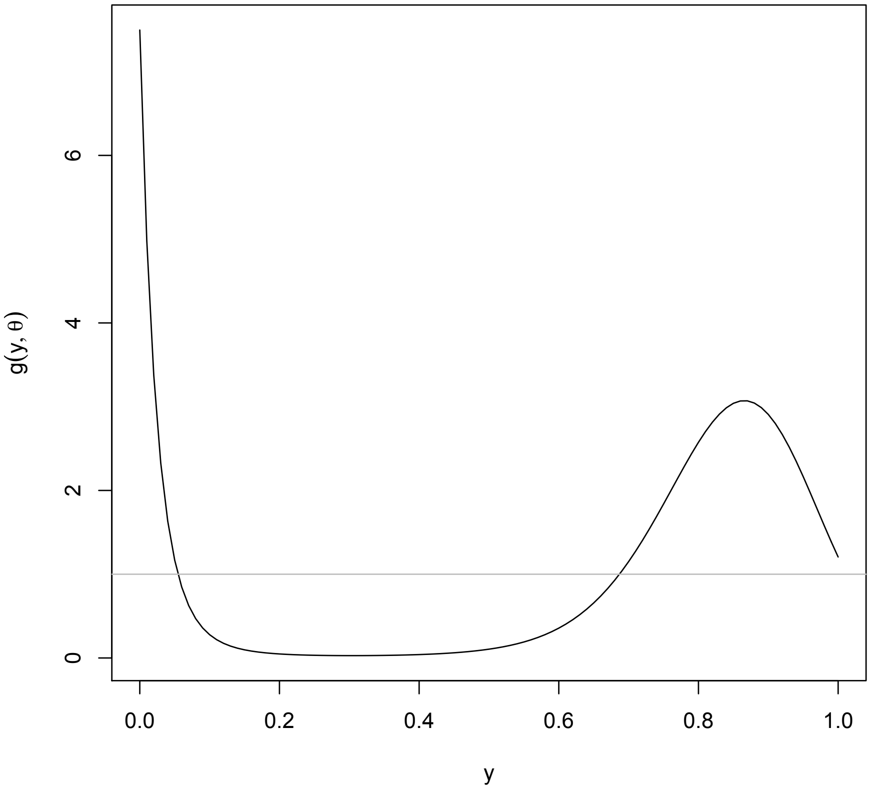
[6] It is common in the literature to see the criterion written out multiplied through by n, or even by $2n$. Also, it is often called the "Bayesian information criterion", or BIC. This is an unfortunate name, because, despite what Schwarz (1978) thought, it really has nothing at all to do with Bayes's rule or even Bayesian statistics. It's best thought of as a fast, but very crude and not always very accurate, approximation to cross-validation. If you want to know more, Claeskens and Hjort (2008) is probably the best reference.

[7] Unless someone has messed up a calculation, that is.



```
h1 <- function(y) { sqrt(12)*(y-0.5) }
h2 <- function(y) { sqrt(5)*(6*(y-0.5)^2-0.5) }
h3 <- function(y) { sqrt(7)*(20*(y-0.5)^3 - 3*(y-0.5)) }
curve(h1(x),ylab=expression(h[j](y)),xlab="y")
curve(h2(x),add=TRUE,lty="dashed")
curve(h3(x),add=TRUE,lty="dotted")
legend(legend=c(expression(h[1]),expression(h[2]),expression(h[3])),
  lty=c("solid","dashed","dotted"),x="bottomright")
curve(exp(h1(x)),ylab=expression(e^h[j](y)),xlab="y")
curve(exp(h2(x)),add=TRUE,lty="dashed")
curve(exp(h3(x)),add=TRUE,lty="dotted")
legend(legend=c(expression(h[1]),expression(h[2]),expression(h[3])),
  lty=c("solid","dashed","dotted"),x="bottomright")
```

Figure 17.1: Upper panel: the first three of the basis functions for Neyman's smooth tests, h_1, h_2 and h_3. Each h_j is a polynomial of order j which is orthogonal to the others, in the sense that $\int_0^1 h_j(y)h_k(y)dy = 0$ when $j \neq k$, but normalized in size, $\int_0^1 h_j^2(y)dy = 1$. The lower panel shows $e^{h_j(y)}$, to give an indication of how the functions contribute to the probability density in Eq. 17.10.



```
x <- (1:1e6)/1e6
z <- sum(exp(h1(x)+h2(x)-h3(x)))/1e6
curve(exp(h1(x)+h2(x)-h3(x))/z,xlab="y",ylab=expression(g(y,theta)))
abline(h=1,col="grey")
```

Figure 17.2: Illustration of a smooth alternative density: using the same basis functions as before, with $\theta_1 = 1$, $\theta_2 = 1$, $\theta_3 = -1$. The first two lines of the R calculate the normalizing constant $z(\theta)$ by a simple numerical integral. The grey line shows the uniform density.

attempts to replicate some surprising effect in separate laboratories[8]. If the tests are independent, then the p-values should be IID and uniform. It would seem like we should be able to combine these into some over-all p-value. This is *precisely* what Neyman's smooth test of uniformity lets us do.

Density Estimation by Series Expansion

As an aside, notice what we have done. By using a large enough d, as I said, densities which look like Eq. 17.10 can come as close as we like to any smooth density on $[0,1]$. And now we have at least two ways of picking d: by cross-validation, or by the Schwarz information criterion (Eq. 17.20). If we let $d \rightarrow \infty$ as $n \rightarrow \infty$, then we have a way of approximating any density on the unit interval, without knowing what it was to start with, or assuming a particular parametric form for it. That is, we have a way of doing non-parametric density estimation, at least on $[0,1]$, without using kernels.

If you want to estimate a density on $(-\infty,\infty)$ instead of on $[0,1]$, you can do so by using a transformation, e.g., the inverse logit. This is the opposite of what you did in the homework, where you used a transformation to take $[0,1]$ to $(-\infty,\infty)$ so you could use kernel density estimation.

17.1.4 Smooth Tests of Non-Uniform Parametric Families

Remember that we went into all these details about testing uniformity because we want to test whether X is distributed according to some continuous distribution with CDF F. From §17.1.1, if we define $Y = F(X)$, then $X \sim F$ is equivalent to $Y \sim \text{Unif}(0,1)$, so we have a two-step procedure for testing whether $X \sim F$:

1. Use the CDF F to transform the data, $y_i = F(x_i)$
2. Test whether the transformed data y_i are uniform

Let's think about what the alternatives considered in the test look like. For y, the alternative densities are (to repeat Eq. 17.10)

$$g(y;\theta) \equiv \begin{cases} \frac{e^{\sum_{j=1}^{d}\theta_j h_j(y)}}{z(\theta)} & 0 \leq y \leq 1 \\ 0 & \text{elsewhere} \end{cases} \tag{17.21}$$

Since $X = F^{-1}(Y)$, this implies a density for X:

$$\begin{aligned} g_X(x;\theta) &= \frac{e^{\sum_{j=1}^{d}\theta_j h_j(F(x))}}{z(\theta)} \frac{dF}{dx} && (17.22) \\ &= \frac{e^{\sum_{j=1}^{d}\theta_j h_j(F(x))}}{z(\theta)} f(x) && (17.23) \end{aligned}$$

[8] These are typical examples of **meta-analysis**, trying to combine the results of many different data analyses (without just going back to the original data).

where f is the pdf corresponding to the CDF F. (Why do we not have to worry about setting this to zero outside some range?) Just like $g(\cdot;\theta)$ is a modulation or distortion of the uniform density, $g_X(\cdot;\theta)$ is a modulation or distortion of $f(\cdot)$. If and when we reject the density f, $g_X(\cdot;\theta)$ is available to us as an alternative.

Even if $h_j(y)$ is a polynomial in y, $h_j(F(x))$ will not (in general) be a polynomial in x, but it remains true that

$$\int_{-\infty}^{\infty} h_j(F(x))h_k(F(x))f(x)dx = \delta_{jk} \tag{17.24}$$

Figure 17.3 illustrates what happens to the basis functions, and to particular alternatives.

When it comes to the actual smooth test, we can either use the likelihood ratio, or we can calculate

$$\overline{h_j} = \frac{1}{n}\sum_{i=1}^{n} h_j(y_i) = \frac{1}{n}\sum_{i=1}^{n} h_j(F(x_i)) \tag{17.25}$$

leading as before to the test statistic

$$\Psi^2 = n\sum_{j=1}^{n}\overline{h_j}^2 \tag{17.26}$$

The distribution of the test statistics is unchanged under the null hypothesis, i.e., still χ^2_d if d is fixed. (There are still d degrees of freedom, because we are still fixing d parameters from distributions of the form Eq. 17.23.) If d is chosen from the data, we still need to bootstrap, but can do so just as before.

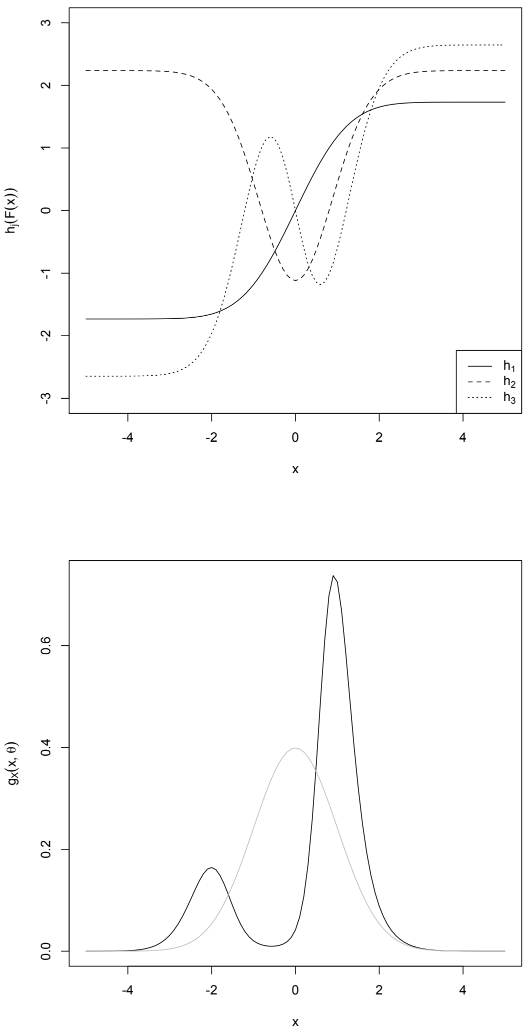
Estimated Parameters

So far, the discussion has assumed that F is fixed and won't change with the data. This is often not very realistic. Rather, F comes from some parametrized family of distributions, with parameter (say) β, i.e., $F(\cdot;\beta)$ is a different CDF for each value of β. For Gaussians, for instance, β is a vector consisting of the mean and variance (or mean and standard deviation). Let's assume that there are always corresponding densities, $f(\cdot;\beta)$, and these are always continuous.

We don't know β so we have to estimate it. After estimating, we'd like to test whether the model really matches the data. It would be convenient if we could do the following:

1. Get estimate $\widehat{\beta}$ from $x_1, x_2, \ldots x_n$
2. Calculate $y_i = F(x_i;\widehat{\beta})$
3. Apply a smooth test of uniformity to $y_1, y_2, \ldots y_n$

That is, it would be convenient if we could just *ignore* the fact that we had to estimate β.



```
curve(h1(pnorm(x)),xlab="x",ylab=expression(h[j](F(x))),from=-5,to=5,
  ylim=c(-3,3))
curve(h2(pnorm(x)),add=TRUE,lty="dashed")
curve(h3(pnorm(x)),add=TRUE,lty="dotted")
legend(legend=c(expression(h[1]),expression(h[2]),expression(h[3])),
  lty=c("solid","dashed","dotted"),x="bottomright")
curve(dnorm(x)*exp(h1(pnorm(x))+h2(pnorm(x))-h3(pnorm(x)))/z,xlab="x",
  ylab=expression(g[X](x,theta)),from=-5,to=5)
curve(dnorm(x),add=TRUE,col="grey")
```

Figure 17.3: Upper panel: the basis functions from Figure 17.1 composed with the standard Gaussian CDF. Lower panel: the alternative to the standard Gaussian corresponding to the alternative to the uniform distribution plotted in Figure 17.2, i.e., $\theta_1 = \theta_2 = 1$, $\theta_3 = -1$. The grey curve is the standard Gaussian density, corresponding to the flat line in Figure 17.2.

We can do this if $\widehat{\beta}$ is the maximum likelihood estimate. To understand this, think about the family of alternative distributions we're now considering in the test. Substituting into Eq. 17.23, they are

$$g_X(x;\beta,\theta) = \frac{e^{\sum_{j=1}^{d}\theta_j h_j(F(x;\beta))}}{z(\theta)} f(x;\beta) \quad (17.27)$$

The null hypothesis that $X \sim F(\cdot;\beta)$ for some β is thus corresponds to $X \sim G_X(\cdot;\beta,0)$ — we are still fixing d parameters in the larger family. And, generally speaking, when we fix d parameters in a parametric model, we get a χ^2_d distribution in the log-likelihood ratio test (Appendix B). If d is not fixed but data-driven, then, again, we need to bootstrap.

17.1.5 Implementation in R

The main implementation of smooth tests available in R is the `ddst` package (Biecek and Ledwina, 2010), standing for "data-driven smooth tests". It provides a `ddst.uniform.test`, which we could use for any family where we can calculate the CDF. But it also provides functions for directly testing several families of distributions, notably Gaussians (`ddst.norm.test`) and exponentials (`ddst.exp.test`).

Some Examples

Let's give `ddst.norm.test` some Gaussian data and see what happens.

```
> r <- rnorm(100)
> ddst.norm.test(r)

Data Driven Smooth Test for Normality

data:  r,   base: ddst.base.legendre,   c: 100
WT* = 0.6183, n. coord = 1
```

This reminds us what the data was, tells us that the test used Legendre polynomials (as opposed to cosines), that d was selected to be 1, and that the value of the test statistic was 0.6183. (The c setting has to do with the order-selection penalty, and is basically ignorable for most users.) These numbers are all attributes of the returned object.

What is missing is the p-value, because this is computationally expensive to calculate. (You can control how many bootstraps it uses, but the default is 1000.)

```
> ddst.norm.test(r,compute.p=TRUE)

Data Driven Smooth Test for Normality

data:  r,   base: ddst.base.legendre,   c: 100
WT* = 0.6183, n. coord = 1, p-value = 0.476
```

So the p-value is 0.476, giving us no reason to reject a Gaussian distribution when we're looking at numbers from the standard Gaussian. If we ignored the fact that d was selected from the data and plugged into the corresponding chi_d^2 distribution, we'd get a p-value of

```
> pchisq(0.6183,df=1,lower.tail=FALSE)
[1] 0.4316797
```

which to say a relative error of about 10%.

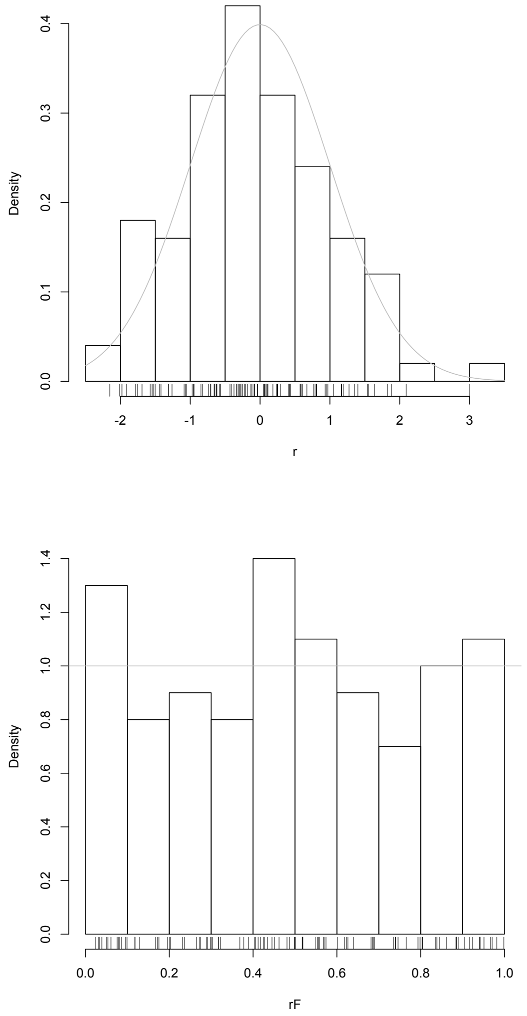
What if we give some non-Gaussian data? Say, the same amount of data from a t distribution with 5 degrees of freedom?

```
> ng <- rt(100,df=5)
> ddst.norm.test(ng,compute.p=TRUE)

Data Driven Smooth Test for Normality

data:  ng,   base: ddst.base.legendre,   c: 100
WT* = 16.5623, n. coord = 2, p-value = 0.007
```

Of course, it won't *always* reject, because the we're only looking at 100 samples, and the t distribution isn't *that* different from a Gaussian. Still, when I repeat this experiment many times, we get quite respectable power at the standard 5% size:

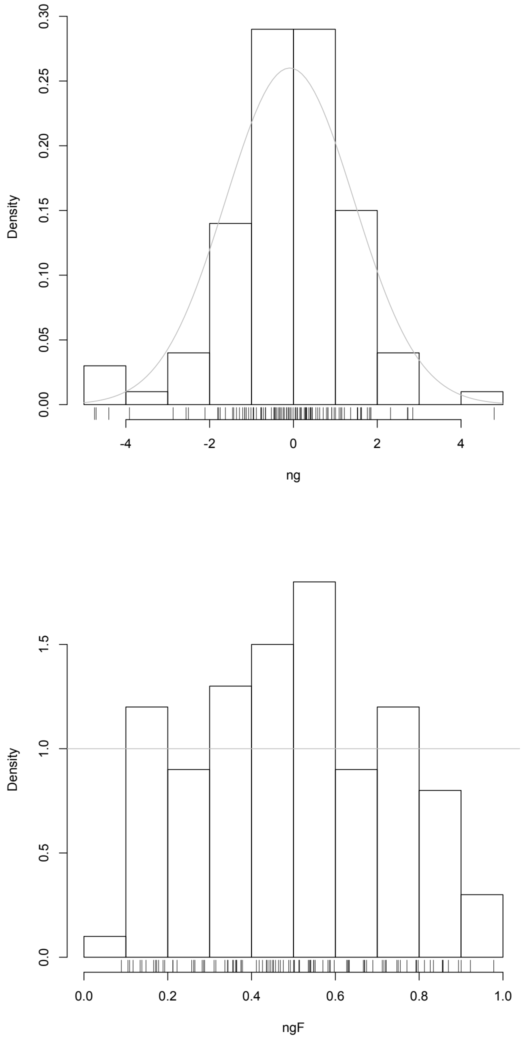
```
> mean(replicate(100,
+ ddst.norm.test(rt(100,df=5),compute.p=TRUE)$p.value)<0.05)
[1] 0.51
```

See Exercise 3 for a small project of `ddst.exp.test` to check a Pareto distribution.



```
plot(hist(r),freq=FALSE,main="")
rug(r)
curve(dnorm(x),add=TRUE,col="grey")
rF <- pnorm(r,mean=mean(r),sd=sd(r))
plot(hist(rF),freq=FALSE,main="")
rug(rF)
abline(h=1,col="grey")
```

Figure 17.4: Upper panel: histogram of the 100 random values from the standard Gaussian used in the text (exact values marked along the horizontal axis), plus the true density in grey. Lower panel: transforming the data according to the Gaussian *fitted to the data* by maximum likelihood.



```
plot(hist(ng),freq=FALSE,main="")
rug(ng)
curve(dnorm(x,mean=mean(ng),sd=sd(ng)),add=TRUE,col="grey")
ngF <- pnorm(r,mean=mean(ng),sd=sd(ng))
plot(hist(ngF),freq=FALSE,main="")
rug(ngF)
abline(h=1,col="grey")
```

Figure 17.5: Treating the draw from the t distribution discussed in the text the same as the Gaussian sample in Figure 17.4.

17.1.6 Conditional Distributions and Calibration

Suppose that we are not interested in the marginal distribution of X, but rather its conditional distribution given some other variable or variables C (for "covariates"). If the conditional density $f(x|C=c)$ is continuous in x for every c, then it is easy to argue, in parallel with §17.1.1, that $F(X|C=c)$, the conditional CDF, should $\sim$ Unif$(0,1)$. So, as long as we use the conditional CDF to transform X, we can apply smooth tests as before.

One important use of this is regression residuals. Suppose X is the target variable of a regression, with C being the predictor variables[9], and we have some parametric distribution in mind for the noise (Gaussian, say), with the noise ϵ being independent of C. Then the model is $X = r(C)+\epsilon$, so looking at the conditional CDF of X given Z is equivalent to looking at the at unconditional CDF of the residuals. We can then actually *test* whether the residuals are Gaussian, rather than just squinting at a Q-Q plot. We could also do this by applying a K-S test to the transformed residuals, but everything that was said above in favor of smooth tests would still apply.

Notice, by the way, that by applying the CDF transformation to the residuals, we are checking whether the model is properly calibrated, i.e., whether events it says happen with probability p actually have a frequency close to p. We do need to impose assumptions about the distribution of the noise to check calibration for a regression model, since if we *just* predict expected values, we say nothing about how often any particular range of values should happen.

Later, when we look at graphical models and at time series, we will see several other important situations where a statistical model is really about conditional distributions, and so can be checked by looking at conditional CDF transformations. It seems to be somewhat more common to apply K-S tests than smooth tests after the conditional CDF transformation (e.g., Bai 2003), but I think this is just because smooth tests are not as widely known and used as they should be.

[9] I know you're used to X being the predictor and Y being the target.

17.2 Relative Distributions

So far, I have been talking about how we can test whether our data follows some hypothesized distribution, or family of distributions, by using the fact that $F(X)$ is uniform if and only if X has CDF F. If the values of $F(x_i)$ are close enough to being uniform, the true CDF has to be pretty close to F (with high confidence); if they are far from uniform, the true CDF has to be far from F (again with high confidence).

In many situations, however, we already know (or are at least pretty sure) that X doesn't have some distribution, say F_0, and what we are interested in is *how* X fails to follow it; we want, in other words, to compare the distribution of X to some reference distribution F_0. For instance:

1. We are trying a new medical procedure, and we want to compare the distribution of outcomes for patients who got the treatment to those who did not.

2. We want to compare the distribution of some social outcome across two categories at the same time. (For instance, we might compare income, or lifespan, for men and for women.)

3. We might want to compare members of the same category at different times, or in different locations. (We might compare the income distribution of American men in 1980 to that of 2010, or the lifespans of American men in 2010 to those of Canadian men.)

4. We might compare our actual population to the distribution predicted by a model we know to be too simple (or just approximate) to try to learn what it is missing.

You learned how to do comparisons of simple summaries of distributions in baby stats. (For instance, you learned how to compare group means by doing t-tests.) While these certainly have their places, they can miss an awful lot. For example, a few years ago now an anesthesiologist came to the CMU statistics department for help evaluating a new pain-management procedure, which was supposed to reduce how many pain-killers patients recovering from surgery needed. Under both the old procedure and the new one, the distribution was strongly bimodal, with some patients needing very little by way of pain-killers, many needing much more, and a few needing an awful lot of drugs. Simply looking at the change in the mean amount of drugs taken, or even the changes in the mean and the variance, would have told us very little about whether things were any better[10].

In this example, the **reference distribution**, F_0, is given by the distribution of drug demand for patients on the old pain-management protocol. The new or **comparison** sample, $x_1, \ldots x_n$, are realizations of a random variable X, representing the demand for pain-killers under the new protocol. X follows the **comparison distribution** F, which is presumably not the same as F_0; how does it differ?

[10] I am omitting some details, and not providing a reference because the study is still, so far as I know, unpublished.

The idea of the **relative distribution** is to characterize the change in distributions by using F_0 to transform X into $[0,1]$, and then looking at how it departs from uniformity. The **relative data**, or **grades**, are

$$r_i = F_0(x_i) \tag{17.28}$$

Simply put, we take the comparison data points and see where they fall in the reference distribution.

What is the cumulative distribution function of the relative data? Let's look at this first at the population level, where we have F_0 (the reference distribution) and F (the comparison distribution), rather than just samples. Let's call the CDF of the relative data G:

$$\begin{aligned} G(a) &\equiv \Pr(R \leq a) && (17.29)\\ &= \Pr(F_0(X) \leq a) && (17.30)\\ &= \Pr(X \leq Q_0(a)) && (17.31)\\ &= F(Q_0(a)) && (17.32) \end{aligned}$$

where remember $Q_0 = F_0^{-1}$ is the quantile function of the reference distribution. This in turn implies a probability density function of the relative data:

$$\begin{aligned} g(y) &\equiv \left.\frac{dG}{da}\right|_{a=y} && (17.33)\\ &= \left.\frac{dF}{du}\right|_{u=Q_0(y)} \left.\frac{dF_0^{-1}}{da}\right|_{a=y} && (17.34)\\ &= f(Q_0(y))\frac{1}{f_0(Q_0(y))} = \frac{f(Q_0(y))}{f_0(Q_0(y))} && (17.35) \end{aligned}$$

This only applies when $y \in [0,1]$; elsewhere, $g(y)$ is straightforwardly 0.

When $g(y) > 1$, we have $f(Q_0(y)) > f_0(Q_0(y))$ — that is, values around $Q_0(y)$ are relatively more probable in the comparison distribution than in the reference distribution. Likewise, when $g(y) < 1$, the comparison distribution puts less weight on values around $Q_0(y)$ than does the reference distribution. If the comparison distribution was exactly the same as the reference distribution, we would, of course, get $g(y) = 1$ everywhere.

One very important property of the relative distribution is that it is invariant under monotone transformations. That is, suppose instead of looking at X, we looked at $h(X)$ for some monotonic function h. (An obvious example would be change of units, but we might also take logs or powers.) Summary statistics like differences in means are generally not even *equi*-variant[11]. But it is easy to check (Exercise 4) that

[11] Remember that a statistic, say δ, is a function of the data, $\delta(x_1, x_2, \ldots x_n)$. The statistic is *invariant* under a transformation h if $\delta(h(x_1), h(x_2), \ldots h(x_n)) = \delta(x_1, x_2, \ldots x_n)$ — the transformation does not change the statistic. The statistic is *equivariant* if it "changes along with" the transformation, $\delta(h(x_1), h(x_2), \ldots h(x_n)) = h(\delta(x_1, x_2, \ldots x_n))$. Maximum likelihood estimates are equivariant. Statistics like the mean are equivariant under linear and affine transformations (but not others).

the relative distribution of $h(X)$ is the same as the relative distribution of X. This expresses the idea that the difference between the reference and comparison distributions is independent of our choice of a coordinate system for X.

17.2.1 Estimating the Relative Distribution

In some situations, the reference distribution can come from a theoretical model, but the comparison distribution is unknown, though we have samples. Estimating the relative density g is then extremely similar to what we had to do in the last section for hypothesis testing. Non-parametric estimation of g can thus proceed either through fitting series expansions like Eq. 17.10 (with a data-driven choice of d, as above), or through using a fixed, data-independent transformation to map $[0, 1]$ to $(-\infty, \infty)$ and using kernel density estimation[12].

If, on the other hand, neither the reference nor the comparison distribution is fully known, but we have samples from both, estimating the relative distribution involves estimating Q_0, the quantile function of the reference distribution. This is typically estimated as just the empirical quantile function, but in principle one could use, say, kernel smoothing to get at Q_0. Once we have an estimate for it, though, we have reduced the problem of estimating g to the case considered in the previous paragraph.

Uncertainty in the estimate of the relative density g is, as usual, most easily assessed through the bootstrap. Be careful to include the uncertainty in estimates of Q_0 as well, if the reference quantiles have to be estimated. One can, however, also use asymptotic approximations (Handcock and Morris, 1999, §9.6).

17.2.2 R Implementation and Examples

Relative distribution methods were introduced by Handcock and Morris (1998, 1999), who also wrote an R package, `reldist`, which is by far the easiest way to work with relative distributions. Rather than explain abstractly how this works, we'll turn immediately to examples.

Example: Conservative versus Liberal Brains

In the homework, we have looked at the data from Kanai *et al.* (2011), which record the volumes of two parts of the brain, the amygdala and the anterior cingulate cortex (ACC), adjusted for body size, sex, etc., and political orientation on a five-point ordinal scale, with 1 being the most conservative and 5 the most liberal[13]. The subjects being British university students, the lowest score for political orientation recorded was 2, and so we will look at relative distributions between those students and the rest of the sample. That is, we take the conservatives as the comparison sample, and the rest as the reference sample[14].

[12] We saw how to do this in the homework

[13] I am grateful to Dr. Kanai for graciously sharing the data.

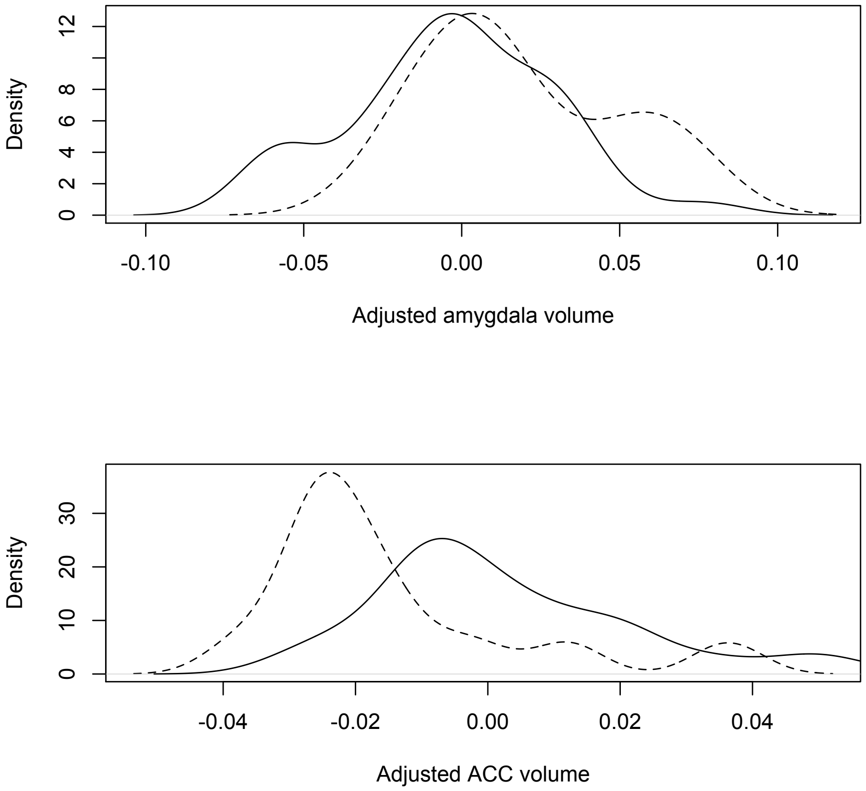
[14] This implies no value judgment about conservatives being "weird", but rather reflects the fact that there are many fewer of them than of non-conservatives in this data.

Having loaded the data into the data frame `n90`, we can look at simple density estimates for the two classes and the two variables (Figure 17.6). This indicates that conservative subjects tend to have relatively larger amygdalas and relatively smaller ACCs, though with very considerable overlap. (We are not looking at the uncertainty here at all.)

Enough preliminaries; let's fine the relative distribution. Figure 17.7).

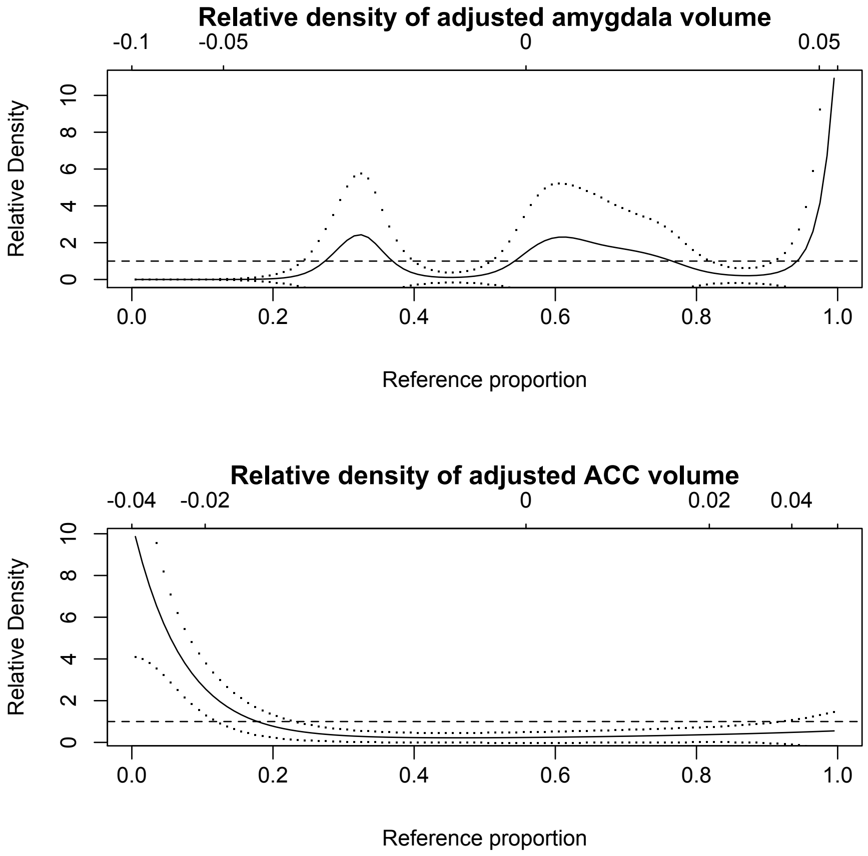
```
library(reldist)
acc.rel <- reldist(y=n90$acc[n90$orientation<3],
  yo=n90$acc[n90$orientation>2],ci=TRUE,
  yolabs=pretty(n90$acc[n90$orientation>2]),
  main="Relative density of adjusted ACC volume")
```

The first argument is the comparison sample; the second is the reference sample. The labeling of the horizontal axis is in terms of the quantiles of the reference distribution; I convert this back to the original units with the optional `yolabs` argument`The function pretty() is a built-in routine for coming up with reasonable axis tick-marks from a vector.  See help(pretty)`.. The dots show a pointwise 95%-confidence band, but based on asymptotic approximations which should not be taken seriously when there are only 77 reference samples and just 13 comparison samples.



```
par(mfrow=c(2,1))
plot(density(n90$amygdala[n90$orientation>2]),main="",
  xlab="Adjusted amygdala volume")
lines(density(n90$amygdala[n90$orientation<3]),lty="dashed")
plot(density(n90$acc[n90$orientation<3]),lty="dashed",main="",
  xlab="Adjusted ACC volume")
lines(density(n90$acc[n90$orientation>2]))
```

Figure 17.6: Estimated densities for the (adjusted) volume of the amygdala (upper panel) and ACC (lower panel) in non-conservative (solid lines) and conservative (dashed) students.



```
par(mfrow=c(2,1))
reldist(y=n90$amygdala[n90$orientation<3],
  yo=n90$amygdala[n90$orientation>2],ci=TRUE,
  yolabs=pretty(n90$amygdala[n90$orientation>2]),
  main="Relative density of adjusted amygdala volume")
reldist(y=n90$acc[n90$orientation<3],
  yo=n90$acc[n90$orientation>2],ci=TRUE,
  yolabs=pretty(n90$acc[n90$orientation>2]),
  main="Relative density of adjusted ACC volume")
```

Figure 17.7: Relative distribution of adjusted brain-region volumes, contrasting conservative subjects (comparison samples) to non-conservative subjects (reference samples). Dots indicate 95% confidence limits, but these are based on asymptotic approximations which don't work here. (The supposed lower limit for the relative density of the amygdala is almost always negative!) The dashed lines mark a relative density of 1, which would be

```
library(np)
data(oecdpanel)
in.oecd <- oecdpanel$oecd==1
reldist(y=oecdpanel$growth[in.oecd],
  yo=oecdpanel$growth[!in.oecd],
  yolabs=pretty(oecdpanel$growth[!in.oecd]),
  ci=TRUE,ylim=c(0,3))
```

Figure 17.8: Relative distribution of the per-capita GDP growth rates of OECD-member countries compared to those of non-OECD countries.

Example: Economic Growth Rates

For a second example, let's return to the OECD data on economic growth featured in Chapter 15. We want to know how the economic growth rates of countries which are already economically developed compares to the growth rates of developing and undeveloped countries. I approximate "is a developed country" by "is a membership of the OECD", as in §15.5.1. I will take the non-developed countries as the reference distribution and the OECD members as the comparison group, mostly because there are more of the former and they are more diverse.

The basic commands now go as before (aside from loading the data from a different library):

```
library(np)
data(oecdpanel)
in.oecd <- oecdpanel$oecd==1
reldist(y=oecdpanel$growth[in.oecd],
  yo=oecdpanel$growth[!in.oecd],
  yolabs=pretty(oecdpanel$growth[!in.oecd]))
```

Examining the resulting plot (Figure 17.8), the relative distribution is unimodal, peaking around the 60th percentile of the reference distribution, a growth rate of about 2.5% per year. The relative distribution drops below 1 at both low (negative) or high ($> 0.05\%$) growth rates — developed countries, at least over the period of this data, tend to grow steadily and within a fairly narrow band, without so much of both the positive and negative extremes of non-developed countries[15]

It's also worth illustrating how to use `reldist` for comparison to a theoretical CDF. A *very* primitive, or better yet nihilistic, model of economic growth would say that the factors causing economies to grow or shrink are so many, and so various, and so complicated that there is no hope of tracking them systematic, but rather that we should regard them as effectively random. As we know from introduction probability, the average of many small independent forces has a Gaussian distribution; so

[15] It's easy to tell a story for why the distribution of growth rates for poor countries is so wide. Some poor countries grow very slowly or even shrink because they suffer from poor institutions, corruption, war, lack of resources, technological backwardness, etc.; some poor countries grow very quickly if they over-come or escape these obstacles and can quickly make us of technologies developed elsewhere. Nobody has a particular good story for why the growth rates of all developed countries are so similar.

we'll just assume that each country grows (or shrinks) by some independent Gaussian amount every year.

Doing this just means applying the cumulative distribution function of the model's distribution to the values from our comparison sample, as in Figure 17.9. The result does not look too different from Figure 17.8. (This does not mean that the nihilistic model of economic growth is right.)

17.2.3 Adjusting for Covariates

Another nice use of relative distributions is in adjusting for covariates or predictors more flexibly than is easy to do with regression. Suppose that we have measurements of *two* variables, X and Z. In general, when we move from the reference population to the comparison population, both variables will change their marginal distributions. If the marginal distribution of Z changes, and the conditional distribution of X given Z did not, then the marginal distribution of X would change. It is often informative to know how the change in the distribution of X compares to what would be anticipated just from the change in Z:

- The two populations might be male and female workers in the same industry, with X income and Z (say) education, or some measure of qualifications.

- The two populations might be students at two different schools, or taught in two different ways, with X their test scores at the end of the year, and Z some measure of prior knowledge.

Write the conditional density of X given Z in the reference population as $f_0(x|z)$. Then, just from the definitions of conditional and marginal probability,

$$f_0(x) = \int f_0(x|z) f_0(z) dz \tag{17.36}$$

If the distribution of the covariate Z is instead taken from the comparison population, we get a *different* distribution for x,

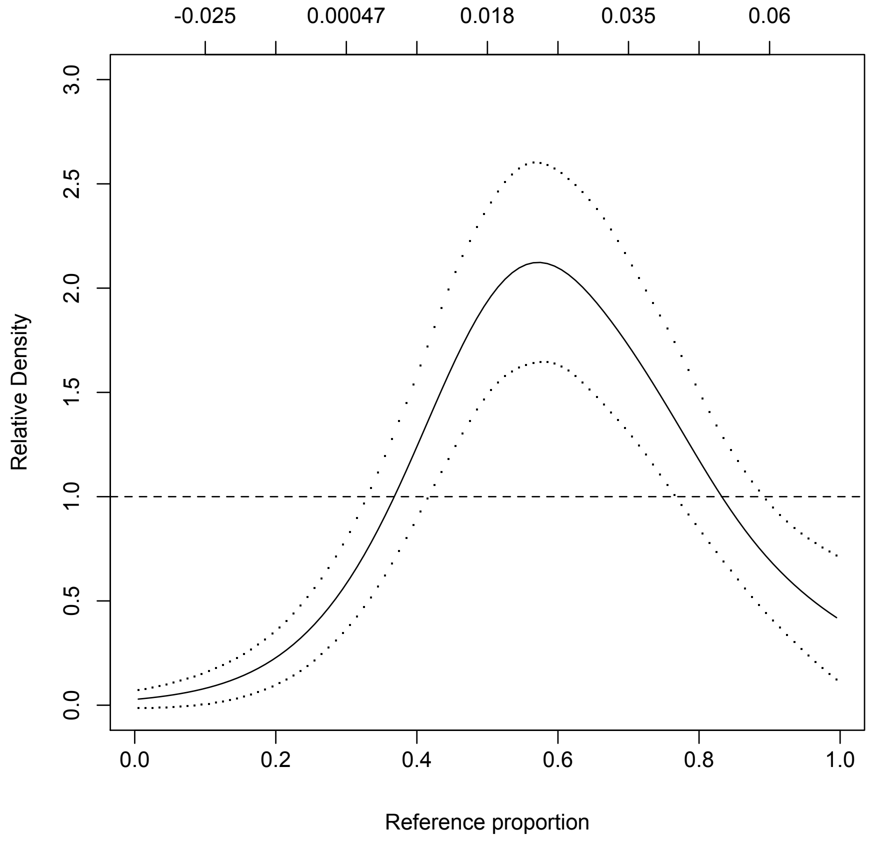
$$f_{0C}(x) = \int f_0(x|z) f(z) dz \tag{17.37}$$

with the C standing for "covariate" or "compensated", depending on who you talk to. This is the distribution we would have seen for X if the distribution of X shifted but the relation between X and Z did not.

Before, we looked at the relative distribution of the comparison distribution F to the reference distribution F_0, which had the density (Eq. 17.35) $g(y) = f(Q_0(y))/f_0(Q_0(y))$. Notice that

$$\frac{f(Q_0(y))}{f_0(Q_0(y))} = \frac{f_{0C}(Q_0(y))}{f_0(Q_0(y))} \frac{f(Q_0(y))}{f_{0C}(Q_0(y))} \tag{17.38}$$

The first ratio on the right-hand side the relative density of F_{0C} compared to f_0; the second ratio is the relative density of F compared to F_{0C}.



```
growth.mean <- mean(oecdpanel$growth[!in.oecd])
growth.sd <- sd(oecdpanel$growth[!in.oecd])
r = pnorm(oecdpanel$growth[in.oecd],growth.mean,growth.sd)
reldist(y=r,ci=TRUE,ylim=c(0,3))
top.ticks <- (1:9)/10
top.tick.values <- signif(qnorm(top.ticks,growth.mean,growth.sd),2)
axis(side=3,at=top.ticks,labels=top.tick.values)
```

Figure 17.9: Distribution of the growth rates of developed countries, relative to a Gaussian fitted to all growth rates.

I have written everything as though Z were just a scalar, but it could be a vector, so we can adjust for multiple covariates at once. Also, it is important to emphasize that there is no implication that Z is in any sense the cause of X here (though such adjustments are often more *interesting* when that's true).

Example: Adjusting Growth Rates

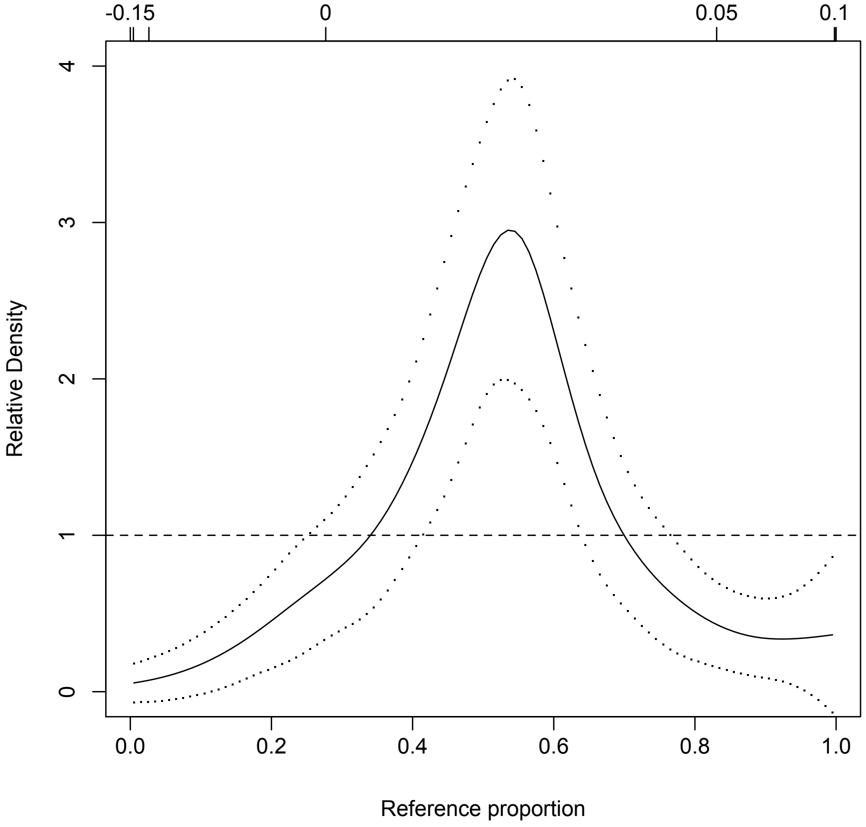
It will be easier to see how this works with an example. The `oecdpanel` data set also includes a variable called `humancap`, which is the log of the average number of years of education of people over the age of 15[16] How do the growth rates of developed countries compare to those of undeveloped countries once we adjust for education?

As Figure 17.10 shows, after adjusting for education levels, the relative density shifts somewhat to the left, with its peak peaked closer to the median of the reference distribution. That is, some of the higher-than-usual growth of the developed countries can be explained away by their (unusually high: Figure 17.11) levels of education. But the relative density is now even *more* sharply peaked than it was before.

Again, it would be rash to read too much causality into this. It could be that education promotes economic growth[17], or it could be that education is a luxury of rich societies, which grow faster than average for other reasons.

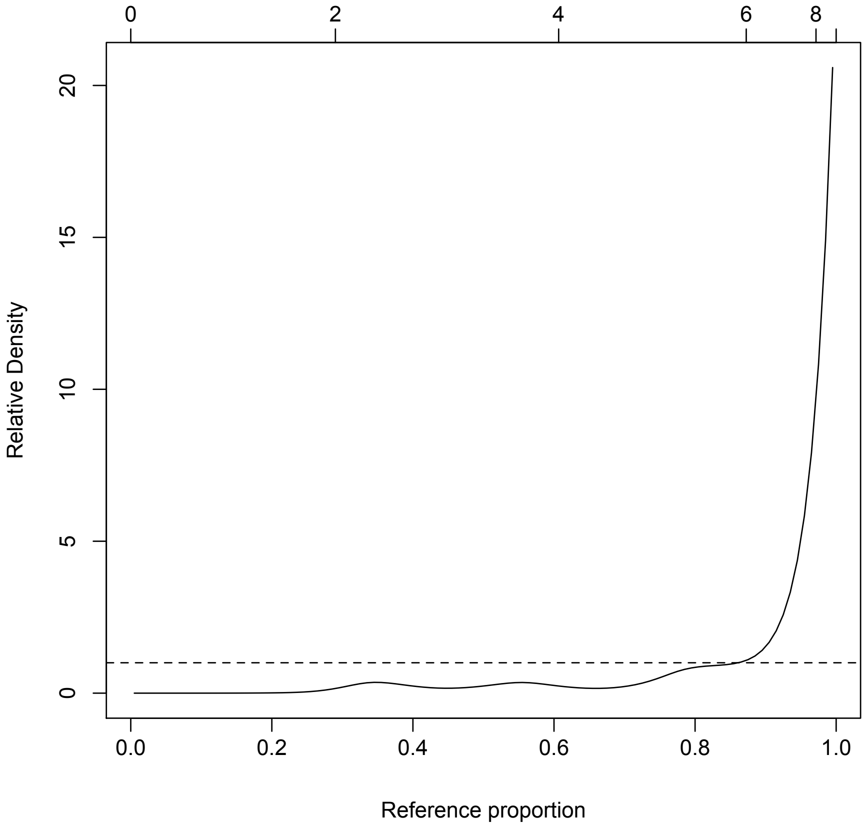
[16] If you look at `help(oecdpanel)`, it calls this variable "average secondary school enrollment rate", but that's clearly wrong, and examining the original papers referenced there shows the correct meaning of the variable. I am not sure why it was logged. (Incidentally, `humancap` stands for "human capital". Whether education is best thought of in this way, or indeed whether years of schooling are a good measure of human capital, are hard questions which we fortunately do not have to answer.)

[17] Certainly it's convenient for a teacher to think so.



```
reldist(y=oecdpanel$growth[in.oecd],
  yo=oecdpanel$growth[!in.oecd],
  yolabs=pretty(oecdpanel$growth[!in.oecd]),
  z=oecdpanel$humancap[in.oecd],
  zo=oecdpanel$humancap[!in.oecd],
  decomp="covariate",
  ci=TRUE,ylim=c(0,4))
```

Figure 17.10: Relative distribution of per-capita GDP growth rates after adjusting for education (`humancap`).



```
reldist(y=exp(oecdpanel$humancap[in.oecd]),
  yo=exp(oecdpanel$humancap[!in.oecd]),
  yolabs=pretty(exp(oecdpanel$humancap[!in.oecd])))
```

Figure 17.11: Relative distribution of years of education, comparing OECD countries to non-OECD countries.

17.3 Further Reading

On smooth tests of goodness of fit, see Bera and Ghosh (2002) (a pleasantly *enthusiastic* paper) and Rayner and Best (1989). The `ddst` package is ultimately based on Kallenberg and Ledwina (1997). On relative distributions, see Handcock and Morris (1998) (an expository paper aimed at social scientists) and Handcock and Morris (1999) (a more comprehensive book with technical details).

17.4 Exercises

To think through, not to hand in.

1. §17.1.3 asserts that one could use cosines orthonormal basis functions in a Neyman test, with $h_j(x) = c_j \cos 2\pi j x$. Find an expression for the normalizing constant c_j such that these functions satisfy Eq. 17.18 and Eq. 17.19.

2. Prove Eq. 17.24. *Hint:* change of variables. Also, prove that

$$\int_{-\infty}^{\infty} f(x) exp^{\sum_{j=1}^{d} \theta_j h_j(F(x))} dx = \int_{0}^{1} exp^{\sum_{j=1}^{d} \theta_j h_j(y)} dy = z(\theta) \qquad (17.39)$$

3. If $X \sim$ Pareto(α, x_0), then $\log X/x_0 \sim$ Exp(α) — the log of a power-law distributed variable has an exponential distribution. Using the `wealth.dat` data from Chapter 5 and `ddst.exp.test`, test whether net worths over \$3 × 10^8 follow a Pareto distribution.

4. Let $T = h(X)$ for some fixed and strictly monotonic function h. Prove that the relative density of T is the same as the relative density of X. *Hint:* find the density of T under both the reference and comparison distribution in terms of f_0, f and h.