Default prior for psychological research

Erik van Zwet

March 23, 2019

Contents

Introduction

In principle, one can obtain prior information for almost any research question in the life sciences via a meta-analysis of the existing literature which is vast and easily accessible, for instance through PubMed. In practice, however, there are (at least) three practical obstacles. First, a meta-analysis is extra work and that is never popular. Second, the literature is not always reliable because of publication bias and such. Third, it is generally unclear what the scope of the meta-analysis should be.

Now, researchers often want to be "objective" or "non-informative". I believe this can be accomplished by performing a meta-analysis with a very wide scope. One might think that this would lead to very diffuse priors, but that turns out not to be the case.

Choosing a very wide scope to obtain prior information, means that the same meta-analysis can be (re-)used in many situations. As such, it can be used a default choice that requires no extra work.

The problem of publication bias in the literature remains, but there may be ways around it. For example, the Open Science Collaboration conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. All relevant information may be found in:

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015 August 28.

The data I use below are publicly available at https://osf.io/fgjvw/

All the p-values that were selected for replication were (essentially) under 0.05. However, it is well known that is a low standard of evidence. Therefore, I believe the underlying *effects* are still a fairly representative sample of effects that are studied in psychological research. I will use the (two-sided) p-values of the replications to estimate the distribution of these effects. This distribution could be used as a default prior.

Excluding the F-tests, there are 86 p-values.

```
setwd("M:\\papers\\default prior\\OSC")

d=read.csv("rpp_data.csv",header=TRUE)

# exclude missing replication p-values
```

```
d=d[!is.na(d$T_pval_USE..R.),]

# exclude F tests.
Ftests=c("F (2, 67) = 1.70","F(1.48, 41.458) = 1.401","F(18, 660) = 12.98",
          "F(2, 137) = 1.99", "F(2, 174) = 252.83","F(2, 72) = 1.97",
          "F(2, 99) = 2.601","F(2,122) = 0.24","F(2.33, 90) = .38",
          "F(21, 455304) = 261.93","F(3, 59) = 5.681","F(3,12)= 13.06",
          "F(4, 150) = .58")



d=d[!(d$Test.statistic..R. %in% Ftests),]
```

Estimating the prior

Suppose we have an unbiased, normally distributed estimator B of a regression coefficient β. Suppose the standard error se of the estimator is known. In other words, B has the normal distribution with mean β and known standard deviation se.

Defining $\theta = \beta/se$, it follows that the "Z-value" $Z = B/se$ has the normal distribution with mean θ and standard deviation 1.

We can obtain the (approximate) absolute value of Z by transforming the two-sided p-value

$$|Z| = |\Phi^{-1}(\text{pvalue})|.$$

We model the distribution of θ as a mixture of two zero-mean normals. We estimate the two standard deviations and the mixing probability by maximum likelihood. Note that since the prior is assumed to be symmetric, the absolute values of the Z-values are sufficient.

```
library(Rcgmin)

d$T_pval_USE..R.[which(d$X..Tails..R.==1)]=2*d$T_pval_USE..R.[which(d$X..Tails..R.==1)]
d$z=abs(qnorm(d$T_pval_USE..R./2))

loglik = function(par){
  lik=prod(par[3]*dnorm(d$z,mean=0,sd=sqrt(1+par[1]))+
             (1-par[3])*dnorm(d$z,mean=0,sd=sqrt(1+par[2])))
  -log(lik)
}
opt=Rcgmin(par=c(1,10,0.5), fn=loglik,lower=c(0,0,0),upper=c(100,100,1))

tau1=sqrt(opt$par[1])
tau2=sqrt(opt$par[2])
p=opt$par[3]
```
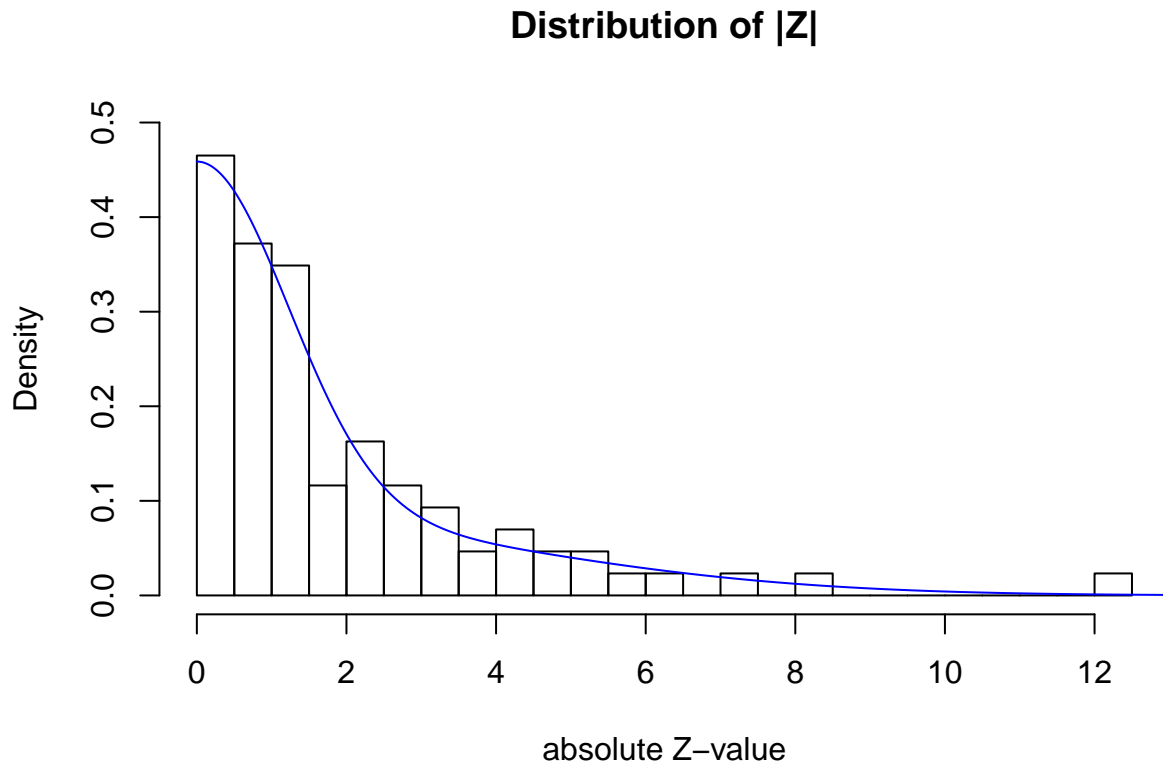
We estimate the standard deviation of the first component as $\hat{\tau}_1 = 0.68$ and of the second component as $\hat{\tau}_2 = 3.94$. The estimated probability of the first component is $p = 0.57$. Below, we show the fit.

```
z=seq(0,15,0.001)
fz=p*dnorm(z,0,sqrt(1+tau1^2)) + (1-p)*dnorm(z,0,sqrt(1+tau2^2))  # marginal of z
```

```
hist(d$z,main="Distribution of |Z|",20,prob=TRUE,xlab='absolute Z-value',ylim=c(0,0.5))
lines(z,2*fz,col='blue')
```
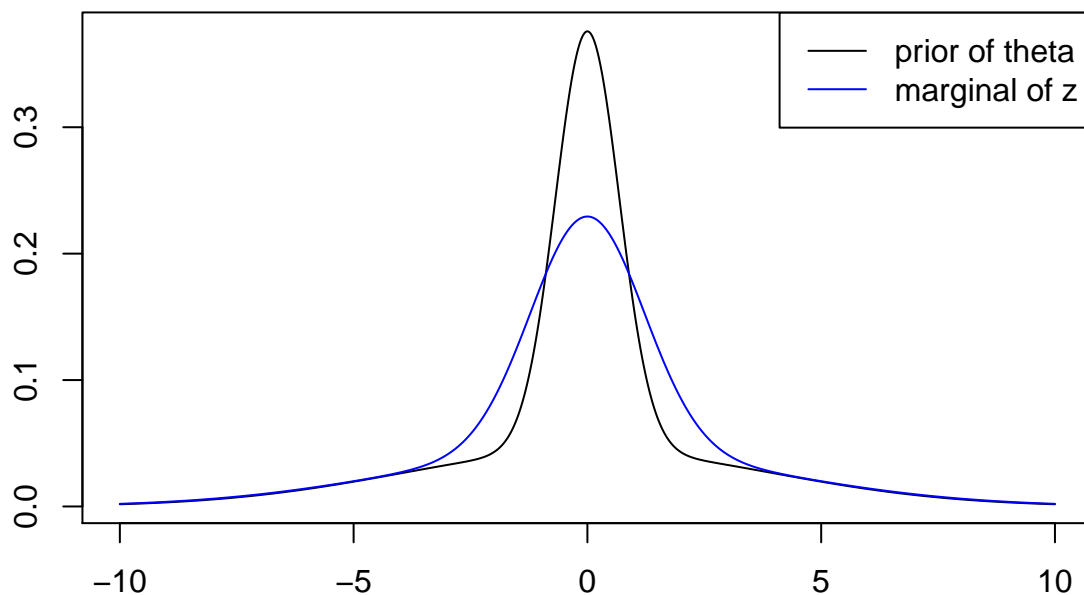
Distribution of |Z|

In the next Figure, we show the distribution of θ together with the (marginal) distribution of Z. Note that Z is the convolution of θ with the standard normal distribution.

```
z=round(seq(-10,10,0.01),2)
t=round(seq(-10,10,0.01),2)                                  # t stands for theta

ft=p*dnorm(t,0,tau1) + (1-p)*dnorm(t,0,tau2)                 # marginal of theta
plot(t,ft,type="l",xlab='',ylab='')

fz=p*dnorm(z,0,sqrt(1+tau1^2)) + (1-p)*dnorm(z,0,sqrt(1+tau2^2))  # marginal of z
lines(z,fz,type="l",col="blue")
legend("topright",lty=c(1,1),col=c("black","blue"),legend=c("prior of theta","marginal of z"))
```

Shrinkage

So, we have the following simple hierachical model:

- $P(D = 1) = 1 - P(D = 2) = 0.57$
- Given $D = 1$, θ has the normal distribution with mean 0 and standard deviation 0.68
- Given $D = 2$, θ has the normal distribution with mean 0 and standard deviation 3.94
- Given D and θ, Z has the normal distribution with mean θ and standard deviation 1.

In this model, it is easy to compute (explicitly) the posterior distribution of θ given $Z = z$. In the next Figure, we show the shrinkage of the mean by plotting $z/E(\theta \mid Z = z)$ as a function of z.

```
Et.z1=z*tau1^2/(tau1^2 +1)              # E(theta|Z=z,D=1)
Et.z2=z*tau2^2/(tau2^2 +1)              # E(theta|Z=z,D=2)
Vt.z1=  tau1^2/(tau1^2 +1)              # Var(theta|Z=z,D=1)
Vt.z2=  tau2^2/(tau2^2 +1)              # Var(theta|Z=z,D=2)
p1.z=p*dnorm(z,0,sqrt(1+tau1^2))/fz     # P(D=1 | Z=z) = P(Z=z | D=1)P(D=1)/f(z)
Et.z=p1.z*Et.z1 + (1-p1.z)*Et.z2        # E(theta | Z=z)

plot(z,z/Et.z,type="l",ylim=c(0,2.5),ylab='',main="Shrinkage factor: z / E(theta|Z=z)")
abline(h=1, lty=2)
```
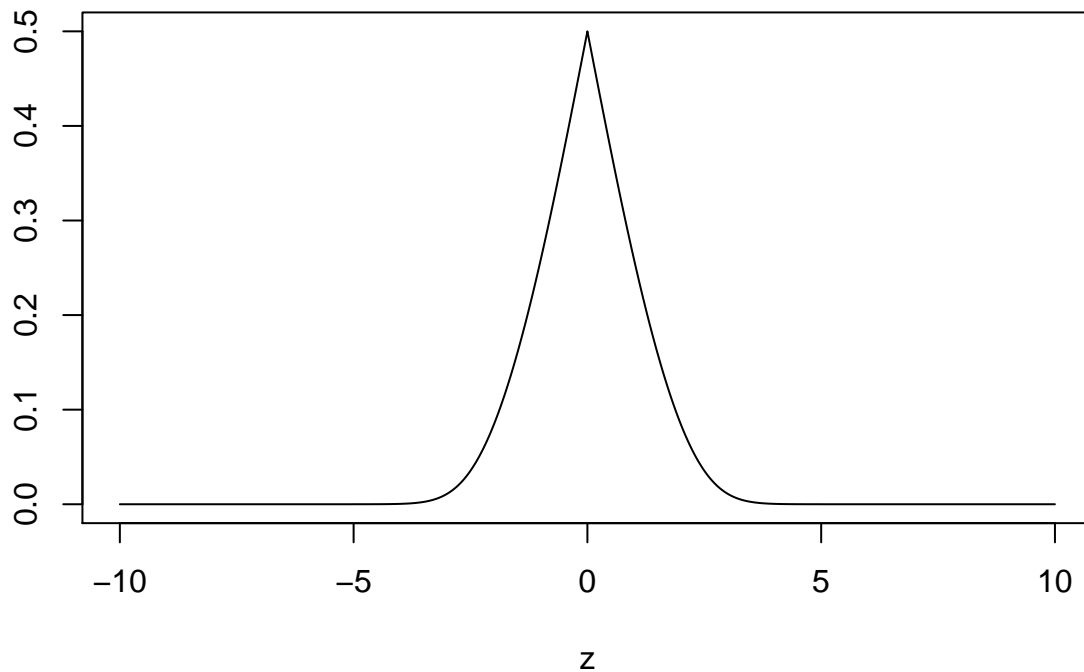
Shrinkage factor: z / E(theta|Z=z)
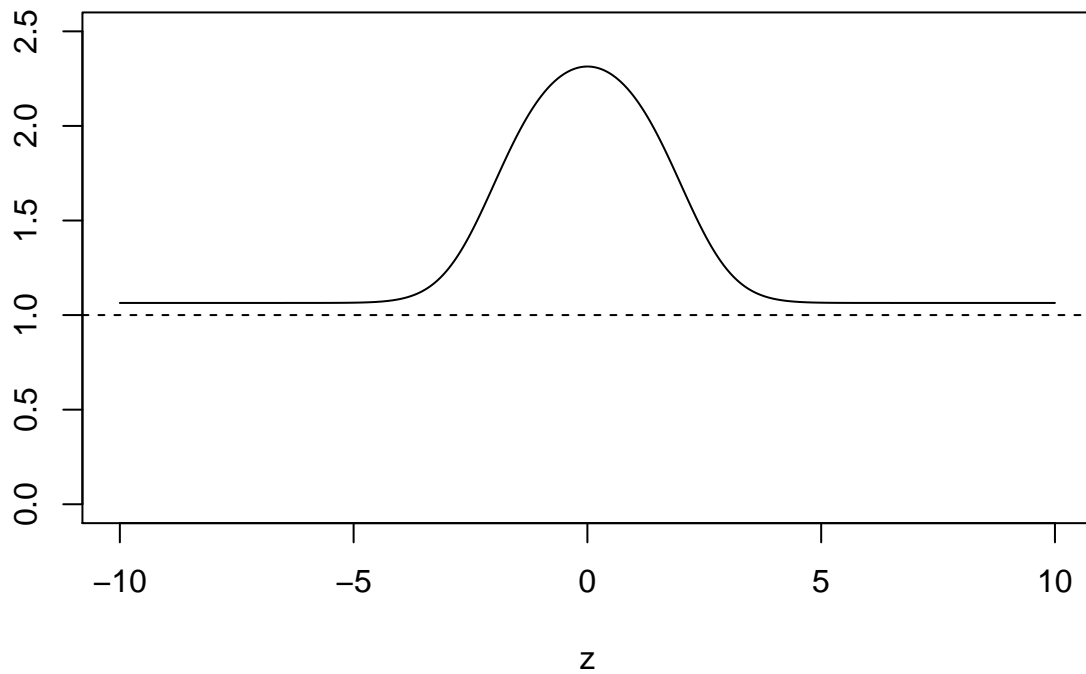
Type S error

A type S error occurs when the signs of B and β differ. We can compute the posterior probability of this event when $Z = z$.

```
ps.z=p1.z*pnorm(0,abs(Et.z1),sqrt(Vt.z1)) + (1-p1.z)*pnorm(0,abs(Et.z2),sqrt(Vt.z2))
plot(z,ps.z,type="l",ylab='',main="Posterior probability of a sign error")
```

Posterior probability of a sign error

Upon observing Z=1.96

In the final Figure, we show the posterior distribution of θ when we observe $Z = 1.96$. We compare this to the so-called confidence distribution. The confidence distribution coincides with the posterior that would result from using the flat (uniform) prior.

```
i=which(z == 1.96)
z[i]/Et.z[i]              # shrinkage 1.96 / E(theta | Z=1.96)

## [1] 1.712152

ps.z[i]                  # P(sign error | Z=1.96)

## [1] 0.08937591

ft.z=p1.z[i]*dnorm(t,Et.z1[i],sqrt(Vt.z1)) + (1-p1.z[i])*dnorm(t,Et.z2[i],sqrt(Vt.z2))
plot(t,ft.z,type="l",xlab="theta",ylab='',main="Upon observing Z=1.96")
lines(t,dnorm(t,z[i],1),col="blue")
legend(lty=c(1,1),"topleft",col=c("black","blue"),legend=c("Posterior density","Confidence density"))
```

Upon observing Z=1.96

Legend:
- Posterior density (black)
- Confidence density (blue)

x-axis: theta (from −10 to 10)
y-axis: 0.0 to 0.5