Model Checking

All models are wrong, some models are useful.

G. Box

The main point of assessing goodness of fit is to decide if the model makes sense, rather than to try and pin down the "true" model. We shall try to answer two questions:

Do the inferences from the model make sense? We want a model that is able to produce inferences that are compatible with knowledge about the problem at hand. This corresponds to an external check.

Is the model consistent with the data? This provides an internal check. The idea is that the model should be able to generate predictive samples that are compatible with the observed data.

Predictive Posterior

The predictive posterior is given by

$$p(z|y) = \int_{\Theta} p(z|y,\theta)p(\theta|y)d\theta$$

given $\theta^{(i)}$, a sample from $p(\theta|y)$, we can obtain a sample of $p(z|y)$ by sampling from $p(z|y,\theta^{(i)})$.

A *leave-one-out* analysis can be performed by sampling from $p(z|y_{-i})$, where y_{-i} denotes the sampled with y_i deleted, comparing to y_i and repeating for all i.

Some questions: Why just "one out"? Do we do this for single observations or for clusters? No easy answers ...

SAT example

Assumptions of the model (SAT-V example),

- *Normality of* $\bar{y}_{.j}|\theta_j, \sigma_j^2$ *with* σ_j^2 *assumed known.* The design and analysis were such that the assumptions seem justifiable in this case.
- *Exchangeability of the prior of the* θ_j*'s.* There is no desire to include in the model features such as
  - the effect in school A is probably larger than the effect in school B
  - the effects in schools A and B are more similar than in schools A and C
- *Normality of* $\theta_j|\mu, \tau$*.* Why not Cauchy or asymmetrically distributed?
- *Uniformity of the hyperprior distribution of* (μ, τ)*.*

- *Comparing the posterior to substantive knowledge.*

	95% P.I.	$E(\theta_j \mid y)$
A	(-2, 35)	12
B	(-4,19)	8
C	(-12,20)	7
D	(-6,21)	8
E	(-9,16)	5
F	(-11,19)	6
G	(-2,28)	11
H	(-9,27)	9

  The estimated treatment effects range from 5 to 12 points, which are plausible values. The extreme values also seem plausible.

- *Posterior predictive distribution.* We simulate the posterior predictive distribution of a hypothetical replication.

If we have, say 500 draws from $p(\theta, \mu, \tau|y)$, we can simulate a hypothetical replicated dataset, $y^{rep} = (y_1^{rep}, \ldots, y_8^{rep})$, by drawing each y_j^{rep} from a normal distribution with mean θ_j and variance σ_j^2.

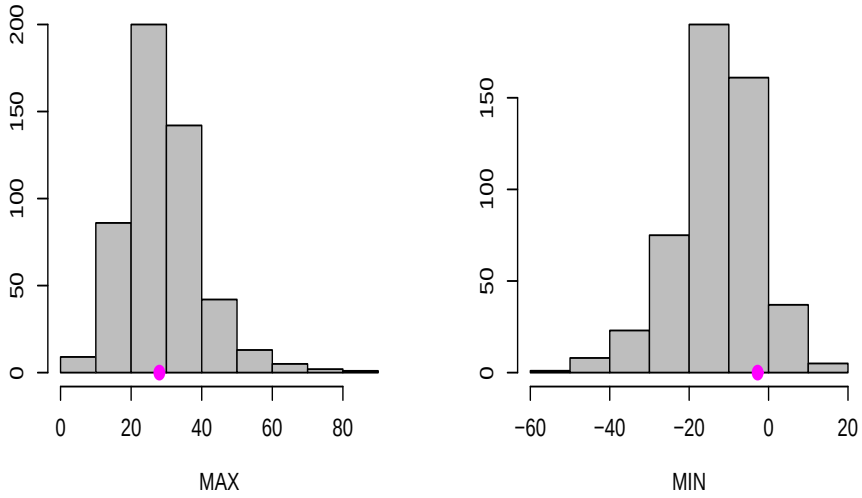
```
for (i in 1:500){
sample.theta[i,]<-conditional.theta(ybar,sample.mu[i],
sample.tau[i],sigma)
for (j in 1:8){
y.future[i,j]<-rnorm(1,sample.theta[i,j],sigma[j])}}
> min(y.future)
[1] -51.00044
> max(y.future)
[1] 82.02397
```

The model generated values for each school are plausible values.

Does the model fit the data? We will examine the posterior distributions for $\max_j y_j$, $\min_j y_j$, $mean(y_j)$ and $sd(y_j)$.

The observed min, max, weighted mean and SD of the eight observations are: -2.75, 28.39, 7.9 and 10.51 respectively.

```
y.max<-apply(y.future,1,max)
length(y.max[y.max<=28.39])/500
#[1] 0.502
y.min<-apply(y.future,1,min)
length(y.min[y.min>=-2.75)]/500
#[1] 0.158
y.mean<-apply(y.future,1,mean)
length(y.mean[y.mean<=7.9])/500
#[1] 0.506
```


0
50
100
150
200
0
20
40
60
80
MAX
0
50
100
150
−60
−40
−20
0
20
MIN
0
50
100
150
−10
0
10
20
30
MEAN
0
50
100
150
200
0
10
20
30
40
SD