The robustness/efficiency tradeoff

The contaminated normal distribution is a model for data with outliers. The expected value of the distribution is zero, so a good estimator of the expected value should be close to zero as often as possible. Since the contaminated normal distribution is symmetric, it also makes sense to consider the sample median as an estimate of the population mean.

The following simulation looks at the variances of the sample mean and the sample median for iid samples of size 50 from a contaminated normal distribution. A range of values for α, the proportion of outliers, is considered.

The results I get are

	α					
	0	0.01	0.05	0.1	0.2	0.3
median	0.03	0.031	0.033	0.038	0.045	0.062
mean	0.02	0.039	0.108	0.229	0.426	0.622

As α increases, the variance of both the sample median and the sample mean increase. This is because when there are more outliers, there is less information in the data and hence it is more difficult to get a precise estimate. However the increase in the sample median's variance is much less than that of the sample mean. Therefore the sample median is *robust* – it performs well under a wide variety of data models. The sample mean is not robust, since it performs well for normal data but much less well for certain other distributions (i.e. for contaminated normal distributions with $\alpha \geq 0.05$).

On the other hand, the sample mean is more *efficient* than the sample median when the data are normal. This can be viewed as a tradeoff between robustness (good performance under a variety of data models) and efficiency (nearly optimal performance under one or two data models of particular importance). It is a tradeoff because the most efficient estimators tend not to be robust, and vice versa.

Note that this tradeoff is different from the bias/variance tradeoff. Biased estimators are not necesarrily more or less robust than unbiased estimators.

```
## Storage for the sample median and sample mean variance estimates.
VMD <- NULL VME <- NULL

## The variance of component 1.
tau2 <- 100

## The variance of component 2.
sig2 <- 1

## alpha is the expected proportion of draws from the outlier component.
for (alpha in c(0,0.01,0.05,0.1,0.2,0.3))
{
  ## A 50x1000 array of iid true/false values in which each entry has
  ## probability alpha of being true.
  A <- array((runif(50*1000) < alpha), c(50, 1000))

  ## The component 1 draws.
  B <- array(sqrt(tau2)*rnorm(50*1000), c(50, 1000))

  ## The component 2 draws.
  C <- array(sqrt(sig2)*rnorm(50*1000), c(50, 1000))

  ## These are the contaminated normal draws.
  X <- A*B + (1-A)*C

  ## The sample medians.
  MD <- apply(X, 2, median)

  ## The sample means.
  ME <- colMeans(X)

  ## Save the variance estimates.
  VMD <- c(VMD, var(MD))
  VME <- c(VME, var(ME))
}
```