## 5.3 t-test and outliers

The data for this example are arranged as follows. There are 200 sets of 10 random numbers, corresponding to 10% contamination by the wide Gaussian; followed by another 200 sets, with 20% contamination, and so on up to 50%. This gives 100 sets of pairs of 10 to compare.

One way of seeing the effect of the outliers is to look at the distribution of $t$ as a function of contamination, and sketch significance levels on the graphs from the tables of $t$. The next step up is to calculate for each dataset the significance level of some fixed presumed difference in the means as a function of contamination. To do this you need to be able to integrate the $t$-distribution to get the area above the value of $t$ that turns up for each dataset. Either way, you can see that the test gets less sensitive, as you would expect.

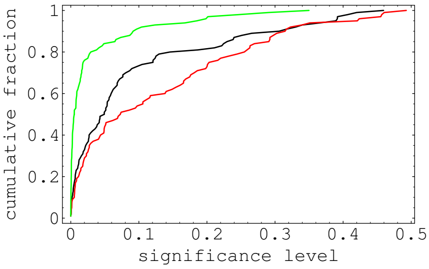

Figure 1: Distribution of one-tailed significance values for a difference in location of 1 unit. Black, no intrinsic difference, no outliers; red, 30% outliers but no intrinsic difference – sensitivity is reduced; green, 0.5 unit intrinsic difference and no outliers.

The Bayesian approach could, for example, just look at the chance that $\mu_1 - \mu_2$ exceeds a certain value, using the posterior distribution of the difference of means given $t$.

To check what happens when there is a real difference in the means, use the data in the other data file, arranged as described before.