Extended homework exercise

For SNPs that have no true association with the outcome, the p-values for the association test should have a uniform distribution between 0 and 1. In a genome-wide association study, while we hope for some true associations, nearly all the SNPs will not have any association with the outcome, so nearly all the p-values should come from a uniform distribution. When a large fraction of p-values depart from the expected uniform distribution it usually indicates a problem with the data, either poor data quality or confounding by population substructure.

A standard quality-control measure for GWAS analyses is to compare the p-values to a uniform distribution with a quantile-quantile plot. We plot $-\log_{10}p$, sorted from smallest to largest, against the expected distribution of $-\log_{10}p$ if the p-values had a uniform distribution. In R

```
qqplot(-log10(ppoints(length(pvalues))), -log10(pvalues))
```

should lie, for most of the SNPs, along the diagonal line given by

```
abline(0,1)
```

A numerical summary of the departure from the uniform distribution is the so-called 'genomic control coefficient' λ. If `beta` and `se` are the coefficient estimates and standard errors, respectively, then

```
lambda <- median( (beta/se)^2)/ 0.4549
```

measures the departure of the median p-value from its expected position. We expect to see lambda close to 1, and this is typically the case in moderate-sized GWAS analyses.

However, lambda behaves differently in SNP x environment interaction analyses, and spuriously high values of lambda can appear as the usual standard error calculations fail to take account of the randomly-changing observed distribution of the environmental variable. [Voorman et al, PLoS One, May 2011]

You will demonstrate this phenomenon.

To generate a single data set of size 1000, simulate

```
environ <- rpois(1000, 10)
outcome <- environ*(environ+rnorm(1000))
```

and a SNP variable with minor allele frequency 0.3

```
gene <- rbinom(1000, 2, 0.3)
```

The linear model for the main effect of SNP is

```
maineffect <- lm(outcome~gene)
```

and to extract the main effect's coefficient estimate (`beta`), its estimated standard errors (`se`) and corresponding p-value, look at the final row of

```
coef(summary(maineffect))
```

For the interaction between SNP and environment, we similarly fit

```
interact <- lm(outcome~gene*environ)
```

and extract what we need from the last row of `coef(summary(interact))`

1. In a GWAS study there would be a single environmental variable, a single outcome phenotype variable, and many SNPs. For this exercise we will assume 5000 independent SNPs. Simulate a single GWAS study with 5000 independent SNPs, and
(a) estimate the main effect of each SNP with the outcome, draw the quantile-quantile plot, and compute lambda.
(b) estimate the interaction between each SNP and environment, draw the quantile-quantile plot, and compute lambda

2. Simulate a reasonable number of repetitions of this study. You will not be able to draw hundreds of quantile-quantile plots, but you will be able to compute `lambda` for the main effects and interaction analyses. Summarize the results.

3. For enthusiastic people who know more about regression modeling, you might try one of the following extra exercises:
(a) show that if the relationship between outcome and environmental variable is simulated to be exactly linear the spurious data-quality problem goes away
(b) use the model-robust standard errors provided by the `vcovHC()` function in the `sandwich` package instead of the model-based standard errors provided by `lm()`, and show that the problem goes away.
(c) Read Voorman et al, PLoS One. 2011; 6(5): e19416.