STATISTICAL MODELLING Part IIC, Michaelmas 2019
Practical 1: Introduction to R By courtesy of Dr. S. A. Bacallado and Dr. R. Shah

Solution to exercises

1. Let $Z \sim N(0, 1)$. Estimate $\mathbb{E}(Z|\{Z \geq 1\})$ and $\mathbb{E}(Z^6)$.

```
> z <- rnorm(1000000)
> mean(z[z >= 1])
[1] 1.527304
> mean(z^6)
[2] 15.09554
```

2. What is the upper 5% point of a χ^2_6 distribution?

```
> qchisq(0.05, df = 6, lower.tail = FALSE)
[1] 12.59159
```

3. Use R to solve

$$\begin{aligned} 3a + 4b - 2c + d &= 9 \\ 2a - b + 7c - 2d &= 13 \\ 6a + 2b - c + d &= 11 \\ a + 6b - 2c + 5d &= 27. \end{aligned}$$

```
> A <- matrix(c(3, 4, -2, 1,
>               2, -1, 7, -2,
>               6, 2, -1, 1,
>               1, 6, -2, 5),
>               nrow=4, ncol=4, byrow=TRUE)
> solve(A, c(9, 13, 11, 27))
[1] 1 2 3 4
```

4. Two lecturers mark the same Tripos question for two disjoint sets of students. They want to test whether their average marks are equal, but they are afraid the sample size is too small to apply the Central Limit Theorem, so one of them writes the following code.

```
> grades_1 <- c(10,11,14.5,15,15,18,12,19,18.5,19,20,13)
> grades_2 <- c(12,11,14.5,13,12,11,12,14.5,20,17)
> mean(grades_1)
> mean(grades_2)
> tstat <- mean(grades_1)-mean(grades_2)
> all_grades <- c(grades_1,grades_2)
> edtstat = rep(0, 10000)
> for (i in 1:10000) {
>   perm <- sample(all_grades)
>   edtstat[i] <- mean(perm[1:length(grades_1)])-mean(perm[-(1:length(grades_1))])
> }
> p_value <- mean(abs(tstat) <= abs(edtstat))
> p_value
```

How can you interpret the output of the last line? Search for the documentation of any function you have not encountered before.

Solution. (Note that I have cosmetically changed the last line of code with respect to the practical 1 .pdf to ease understanding)

This code iterates the following procedure: permute (why? check documentation for `sample` function) all the marks from both examiners, and form two groups of the same size as the two samples, then take the difference of their means. After the iterations, we have an empirical distribution of the test statistic under the null hypothesis of equal average marks, which the theory (of nonparametric bootstrap with replacement; this last concept is non-examinable) tells us is a valid approximation to the exact distribution of the test statistic under the null. The final line calculates how often in the iterations the absolute value of the difference in means is larger than or equal to the absolute value of the observed difference in means; in other words, it (approximately) calculates the probability under the null hypothesis of observing an absolute value of the test statistic larger than or equal to the absolute value of the observed test statistic, i.e. the (approximate) p-value for the two-sided test statistic. This is known as a permutation or exact test, and you can read a brief description of it here:

`https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests`