

## Solutions for Assignment 2

1. (from Montgomery) The following are the burning times (in minutes) of chemical flares of two different formulations. The design engineers are interested in both the mean and variance of the burning times. Thus the engineers want to test:

- $H_0$  (null hypothesis): There **are no** differences between the two different chemical formulations with regard to burning times.
- $H_1$  (alternative hypothesis): There **are** differences between the two different chemical formulations with regard to burning times.

(a) In order to test this hypothesis we need to consider what test statistics might be appropriate. The engineers already have suggested differences in the mean and variances as candidates of interest, thus we will use the following test statistics:

$$g_m(\mathbf{Y}_1, \mathbf{Y}_2) = |\bar{Y}_1 - \bar{Y}_2|$$
$$g_v(\mathbf{Y}_1, \mathbf{Y}_2) = |V(\mathbf{Y}_1) - V(\mathbf{Y}_2)|$$

```
> ## code in the data
> y1 <- c(65,82,81,67,57,59,66,75,82,70)
> y2 <- c(64,56,71,69,83,74,59,82,65,79)
> y <- c(y1,y2)
> x <- c(rep(1,10), rep(2,10))
>
> ## conduct the randomization test
> g.m <- real()
> g.v <- real()
>
> for(nsim in 1:5000){
+ xsim <- sample(x)
+ g.m[nsim] <- abs(mean(y[xsim==1])-mean(y[xsim==2]))
+ g.v[nsim] <- abs(var(y[xsim==1])-var(y[xsim==2]))
+ }
>
> g.m.obs <- abs(mean(y[x==1])-mean(y[x==2]))
> g.v.obs <- abs(var(y[x==1])-var(y[x==2]))
>
> p.value.m <- length(g.m[g.m>=g.m.obs])/length(g.m)
> p.value.v <- length(g.v[g.v>=g.v.obs])/length(g.v)
>
> p.value.m
[1] 0.9802
> p.value.v
[1] 0.9612
```

(b) Based upon the data and the two test statics employed, we are unable to reject the null hypothesis at the  $\alpha = 0.05$  level. Even though we did not find a difference based upon the particular test statistics, it always

important to know as many details of the experiment as possible. In particular, since we used a randomization test, we would like to know if the treatments were randomly assigned. For example, were the two different chemical formulations randomly assigned to flour casings (i.e. CRD). Were some tests done one day and others tests on other days, etc.

2. On the website, you will find data labeled ‘HW2.txt’. These data contain measurements on wheat yields from 30 agriculture plots (as in the notes). Two different fertilizer types (fertilizer A, fertilizer B) were randomly applied to the 30 plots in a **completely randomized design**. Thus fertilizer A was randomly applied to 15 plots and fertilizer B was randomly applied to 15 plots.

(a) The experiment was conducted based upon a completely randomized design. From the table, it can be seen that the means are somewhat similar, although  $\bar{y}_A$  is slightly smaller than  $\bar{y}_B$ . The third quartile for  $y_b$  also appears to be quite a bit higher than the similar statistic for  $y_A$ . In the figure as well as the table, it can be seen that the medians are also somewhat different. Additionally, from the boxplots we see that interquartile range for the two groups looks quite different. This difference in spread can also be seen in the histograms.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
yA	19.66	20.77	20.95	21.28	21.67	23.95
yB	10.21	20.69	23.10	22.90	27.15	28.19

(b) Conduct the following hypothesis test at the  $\alpha = 0.05$  using a randomization test. Remember to consider carefully the test statistics you might use.

- $H_0$  (null hypothesis): Fertilizer type does not affect yield.
- $H_1$  (alternative hypothesis): Fertilizer type does affect yield.

In order to test the null hypothesis, we want a test statistic such that:

$$g(\mathbf{y}_A, \mathbf{y}_B) \text{ is probably } \begin{cases} \text{small under } H_0 \\ \text{large under } H_1 \end{cases}$$

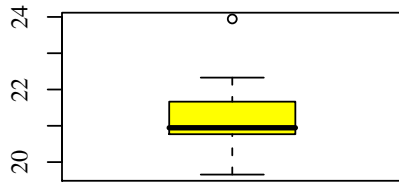
Based upon the exploratory data analysis, the IQR seems like a candidate. So I will use the following test statistic:

$$g(\mathbf{Y}_A, \mathbf{Y}_B) = |IQR(\mathbf{Y}_A) - IQR(\mathbf{Y}_B)|$$

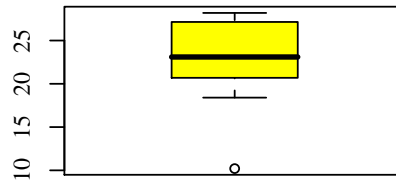
Now let’s conduct the randomization test:

```
> ## Use the absolute value of the difference of the IQR as test statistic
> ## conduct the randomization test
> y <- c(yA, yB)
```

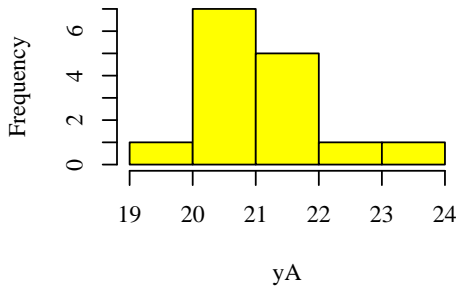
**Boxplot of yA**



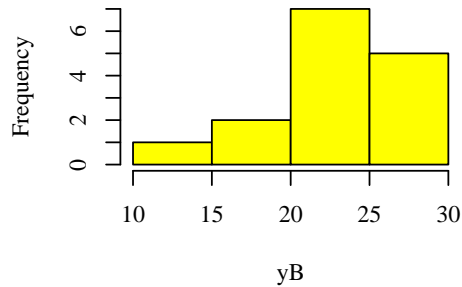
**Boxplot of yB**



**Histogram of yA**



**Histogram of yB**



```

> x <- c(rep("A", length(yA)), rep("B", length(yB)))
>
> g <- real()
>
> for(nsim in 1:5000){
+ xsim <- sample(x)
+ IQR.A.sim <- quantile(y[xsim=="A"], prob=0.75)-quantile(y[xsim=="A"], prob=0.25)
+ IQR.B.sim <- quantile(y[xsim=="B"], prob=0.75)-quantile(y[xsim=="B"], prob=0.25)
+
+ g[nsim] <- abs(IQR.A.sim-IQR.B.sim)
+ }
>
> IQR.A.obs <- quantile(y[x=="A"], prob=0.75)-quantile(y[x=="A"], prob=0.25)
> IQR.B.obs <- quantile(y[x=="B"], prob=0.75)-quantile(y[x=="B"], prob=0.25)
> g.obs <- abs(IQR.A.obs-IQR.B.obs)
>
> p.value<- length(g[g>=g.obs])/length(g)
> p.value
[1] 0.0012

```

Based upon the test statistic and the data, we reject the null hypothesis at the  $\alpha = 0.05$  level.

(c) Since we reject the null hypothesis at the  $\alpha = 0.05$  level, it seems that the difference in IQR between the two groups is not due to plot-to-plot

variation. Since a CRD was used, it is unlikely that there is any pre-experimental bias. However, we might like to know about uncontrollable factors that we might be able to measure, for example weather conditions over the growing season.

### 3. Does sample size matter for p-values?

To conduct the experiment for a sample size of  $n=10$ , I did the following:

- (a) Generate a sample of size  $n$  from a normal population  $(\mu_A, \sigma_A)$  and a sample of size  $n$  from a normal population  $(\mu_B, \sigma_B)$ .
- (b) Conduct the randomization test using  $g(\mathbf{Y}_A, \mathbf{Y}_B) = |\bar{Y}_B - \bar{Y}_A|$  and store the p-value.
- (c) Repeat steps (a) - (b) 100 times or more and average the p-values.
- (d) Now repeat steps (a)-(c) for  $n = \{50, 100, 1000\}$ .

The following R-code does does this process:

```
> n.size <- c(5, 50, 100, 1000)
> p.value.out <- real()
>
> for(i in 1:4){
+ print(i)
+ p.value.avg <- real()
+
+ for(j in 1:100){
+ ## sample size
+ n <- n.size[i]
+ yA <- rnorm(n, 5, 2)
+ yB <- rnorm(n, 6, 2)
+ y <- c(yA, yB)
+ x <- c(rep("A", n), rep("B", n))
+
+ ## randomization test
+ g <- real()
+ for(nsim in 1:5000){
+ xsim <- sample(x)
+ g[nsim] <- abs(mean(y[xsim=="A"])-mean(y[xsim=="B"]))
+ }
+
+ g.obs <- abs(mean(y[x=="A"])-mean(y[x=="B"]))
+
+ p.value <- length(g[g>=g.obs])/length(g)
+
+ p.value.avg[j] <- p.value
+ }
+ p.value.out[i] <- mean(p.value.avg)
+ }
[1] 1
[1] 2
[1] 3
[1] 4
> p.value.out
[1] 0.374484 0.089898 0.020934 0.000000
```

sample size	$n = 10$	$n = 50$	$n = 100$	$n = 1000$
p-value	0.374484	0.089898	0.020934	0.000000

From the table, we can see that p-value does in fact decrease as the sample size increases. This has led to many criticisms of this testing paradigm. One of the more famous is by Leonard Savage (1957):

“Null hypotheses of no difference are usually known to be false before the data are collected . . . when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science.”