STAT2012 - Practial 5

User: jchan

August 19, 2011

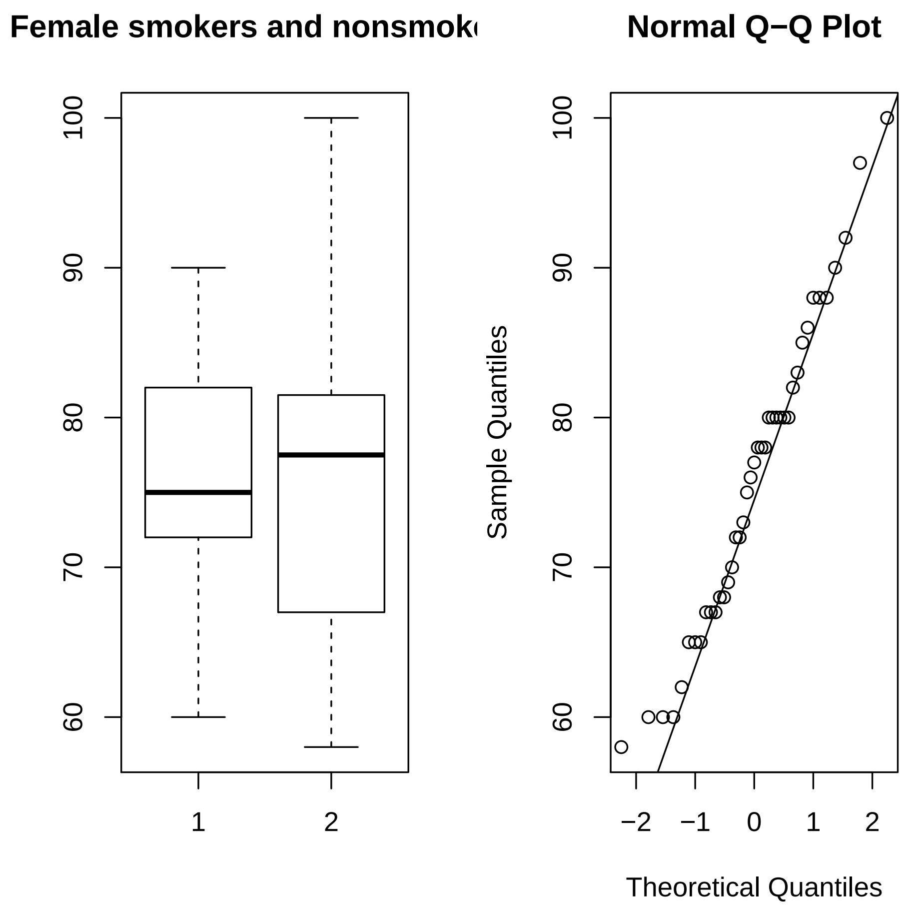
Question 1

```
> survey = read.csv("http://www.maths.usyd.edu.au/u/UG/IM/STAT2012/r/survey.csv")
> attach(survey)
> pulse.sf = pulse[smoke == 1 & sex == 2]
> pulse.sf

[1] 73 67 72 82 90 60 88 75 80

> pulse.nf = pulse[smoke == 2 & sex == 2]
> pulse.nf

 [1]  97  88  83  77  60  78  80  70  67  60  80 100  69  80  68  58  88  80  80
[20]  78  67  72  78  65  86  65  76  92  68  85  65  62
```

Question 1 (a) $H_0 : \mu_x = \mu_y$ against $H_1 : \mu_x \neq \mu_y$

Question 1 (b)

```
> par(mfrow = c(1, 2))
> boxplot(pulse.sf, pulse.nf)
> title("Female smokers and nonsmokers")
> qqnorm(c(pulse.sf, pulse.nf))
> qqline(c(pulse.sf, pulse.nf))
```



As the spread of pulse measures for female nonsmokers seems to be higher, the assumption of equal variance may not be satisfied. However the qq-plot looks quite close to a straight line indicating that the normality condition is approximately satisfied.

Question 1 (c)

```
> t.test(pulse.sf, pulse.nf, alternative = "two.sided", mu = 0,
+     paired = F, var.equal = T)

        Two Sample t-test

data:  pulse.sf and pulse.nf
t = 0.1608, df = 39, p-value = 0.8731
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.479609  8.771276
sample estimates:
```

```
mean of x mean of y
 76.33333  75.68750
```

The test statistic is 0.1608 and the p-value is 0.8731.

Since the p-value is greater than 0.05, we accept H_0 and conclude that the data are consistent with H_0 that the pulse among female smokers and nonsmokers are the same.

The confidence interval is (-7.4796, 8.7713). Since it contains 0, the conclusion is the same as accepting H_0.

Question 1 (d)

```
> t.test(pulse.sf, pulse.nf, alternative = "two.sided", mu = 0,
+     paired = F, var.equal = F)

        Welch Two Sample t-test

data:  pulse.sf and pulse.nf
t = 0.1716, df = 14.188, p-value = 0.8662
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.417387  8.709054
sample estimates:
mean of x mean of y
 76.33333  75.68750
```

The test statistic is 0.1716 and the p-value is 0.8662. The test statistic and p-value are similar to the former test with the equality of variance assumption. However the degree of freedom is quite different. It is 14.19 rather than 39 with the equality of variance assumption.

Question 2 (a)

```
> rank = rank(c(pulse.sf, pulse.nf))
> m = length(pulse.sf)
> n = length(pulse.nf)
> rankA = rank(c(pulse.sf, pulse.nf))[1:m]
> rankA

[1] 18.0 10.0 16.5 31.0 38.0  3.0 36.0 19.0 27.5

> rankB = rank(c(pulse.sf, pulse.nf))[(m + 1):(m + n)]
> rankB
```

```
 [1] 40.0 36.0 32.0 21.0  3.0 23.0 27.5 15.0 10.0  3.0 27.5 41.0 14.0 27.5 12.5
[16]  1.0 36.0 27.5 27.5 23.0 10.0 16.5 23.0  7.0 34.0  7.0 20.0 39.0 12.5 33.0
[31]  7.0  5.0
```

```
> sum(rankB)
```

```
[1] 662
```

```
> w = sum(rankA)
> w
```

```
[1] 199
```

There are ties and the sample sizes are large too. Normal approximation should be used.

Question 2 (b)

```
> wilcox.test(pulse.sf, pulse.nf, alternative = "two.sided", mu = 0,
+     exact = F, correct = F)

        Wilcoxon rank sum test

data:  pulse.sf and pulse.nf
W = 154, p-value = 0.7522
alternative hypothesis: true location shift is not equal to 0
```

```
> w0 = w - m * (m + 1)/2
> w0
```

```
[1] 154
```

The test statistic is $W_0 = 199 - 45 = 154$ and the p-value is 0.7522. Since the p-value is greater than 0.05, we accept H_0 and conclude that the pulse for female smokers and nonsmokers are about the same.

Question 2 (c)

```
> EW = m * (m + n + 1)/2
> EW
```

```
[1] 189
```

```
> sumsqrank = sum(rank^2)
> sumsqrank
```

```
[1] 23792.5

> g = (m + n) * ((m + n + 1)^2)/4
> varW = m * n * (sumsqrank - g)/((m + n) * (m + n - 1))
> varW

[1] 1002.995

> z0 = (w - EW)/sqrt(varW)
> z0

[1] 0.3157553

> p = 2 * (1 - pnorm(z0))
> p

[1] 0.7521883
```

The p-value agrees with that from Q2(b).