

# STAT2012 - Practical 4

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
```

Question 1 (a)  $H_0 : \mu = 70$  against  $H_1 : \mu > 70$  or  $H_0 : p+ = 0.5$  against  $H_1 : p+ > 0.5$

## Question 1 (b)

```
> mu0 = 70
> d = pulse.sf - mu0
> d
```

```
[1] 3 -3 2 12 20 -10 18 5 10
```

```
> n = length(d[d != 0])
> n
```

```
[1] 9
```

```
> x = length(d[d > 0])
> x
```

```
[1] 7
```

The number  $n$  of nonzero differences is 9 and the number  $x$  of positive differences is 7.

### Question 1 (c)

```
> binom.test(x, n, 0.5, alt = "greater", 0.95)
```

```
Exact binomial test
```

```
data: x and n
number of successes = 7, number of trials = 9, p-value = 0.08984
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.4503584 1.0000000
sample estimates:
probability of success
      0.7777778
```

The test statistic is 7 and the  $p$ -value is 0.08984. 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 students who smoke is 70.

**Question 1 (d)** The  $p$ -value of the t-test is lower because the t-test is more powerful, i.e. it has a higher chance of rejecting  $H_0$  even if  $H_0$  is false. However it is based on the more restrictive normality assumption for the data. Hence the t-test is sensitive to outliers and gives valid result only if the normality assumption is satisfied. On the other hand, sign test is less powerful, i.e. lower chance of rejecting  $H_0$  even if  $H_0$  is false because it uses only the information of 'sign'. It can be applied to more general situation because the data is assumed to follow a symmetric distribution not necessary a normal distribution.

### Question 2 (a)

```
> r = rank(abs(d))
> r
```

```
[1] 2.5 2.5 1.0 7.0 9.0 5.5 8.0 4.0 5.5
```

```
> sign.r = r * sign(d)
> sign.r
```

```
[1] 2.5 -2.5 1.0 7.0 9.0 -5.5 8.0 4.0 5.5
```

```
> w.plus = sum(r[d > 0])
> w.plus
```

```
[1] 37
```

```
> w.minus = sum(r[d < 0])
> w.minus
```

```
[1] 8
```

```
> w = min(w.plus, w.minus)
> w
```

```
[1] 8
```

Since `sign.r` contains non-integral values, there are ties. Normal approximation should be used in calculating p-value.

### Question 2 (b)

```
> wilcox.test(pulse.sf, alternative = "greater", mu = 70, exact = F,
+             correct = F)
```

Wilcoxon signed rank test

```
data: pulse.sf
```

```
V = 37, p-value = 0.04264
```

```
alternative hypothesis: true location is greater than 70
```

The test statistic is  $W^+ = 37$  and the  $p$ -value is 0.04264. Since the  $p$ -value is less than 0.05, we reject  $H_0$  and conclude that there is evidence in the data against  $H_0$ . The pulse for female smoker is greater than 70.

**Question 2 (c)** Under the same symmetric data distribution assumption as the sign test, the Wilcoxin signed rank (WSR) test uses the information of both sign and rank (magnitude) of the data. Hence, it is more powerful than the sign test and is preferred to sign-test. The result agrees with that from a t-test showing that the wilcoxon signed rank test and t-test have similar power of rejecting  $H_0$  for this data set. On the other hand, t-test is most powerful because it uses all information of the data. If normality assumption is approximately satisfied, t-test should be used.

### Question 2 (d)

```
> ew.plus = sum(r[d != 0])/2
> ew.plus
```

```
[1] 22.5
```

```
> varw.plus = sum((r[d != 0])^2)/4
> varw.plus

[1] 71

> z0 = (w.plus - ew.plus)/sqrt(varw.plus)
> z0

[1] 1.720833

> p.value = 1 - pnorm(z0)
> p.value

[1] 0.04264053
```

The  $p$ -value agrees with that from 2(b).