Solution to Assignment 1

Question 1

(a) T: If $F(x)$ is continuous on $\mathbb{R}$, then by (1.4) of Lecture Notes, for any $x \in \mathbb{R}$,

$$F(2a-x) = \Pr(X \le 2a-x) = \Pr(X \ge x) = \Pr(X > x) = 1 - F(x)$$

If $F(x)$ is not continuous, then there exists $x \in \mathbb{R}$ such that $\Pr(X = x) > 0 \Rightarrow$

$$F(2a-x) = \Pr(X \ge x) = 1 - \Pr(X < x) = 1 - F(x) + \Pr(X = x) > 1 - F(x)$$

(b) T: For any permutation (i, j, k) of (1,2,3),

$$b_i b_j + b_i b_k + b_j b_k - 3 b_i b_j b_k = b_1 b_2 + b_1 b_3 + b_2 b_3 - 3 b_1 b_2 b_3$$

Hence the value of X does not depend on the order of (b_1, b_2, b_3), so that we can take the fixed order $b_1 < b_2 < b_3$ to calculate $\Pr(X = x)$. Under this order, the number of possible outcomes (b_1, b_2, b_3) is $10 \cdot 9 \cdot 8 / 3! = 120$. Therefore the distribution of X can be determined by the formula given in the question.

(c) F: The *p*-value is a probability under H_0, which does not say anything about the probability $\Pr(\text{accept } H_1 \mid H_1)$ to accept a correct H_1.

Question 2

(a) T: Each Y_i has a Bernoulli distribution with parameter $p_i = \Pr(X_i > 0)$, but the sign test for $H_0 : \theta = 0$ based on $X_1, \ldots, X_n$ is nonparametric since it does not involve any parametric distributions of $X_1, \ldots, X_n$.

(b) F: Without the assumption of symmetric distributions for $X_1, \ldots, X_n$, T^+ does not have the same distribution as $S = I_{\{X_1 > 0\}} + 2 I_{\{X_2 > 0\}} + \cdots + n I_{\{X_n > 0\}}$, hence the rejection rule of $H_0 : \theta = 0$ based on the distribution of S is no longer valid.

(c) F: A nonparametric $100(1-\alpha)\%$ confidence interval of θ can be constructed based on ordered values of $X_1, \ldots, X_n$ by $\left(X_{(n+1-b_{\alpha/2})}, X_{(b_{\alpha/2})}\right)$, or the ordered Walsh averages of $X_1, \ldots, X_n$ by $\left(W_{(M+1-t_{\alpha/2})}, W_{(t_{\alpha/2})}\right)$. Neither interval requires a point estimate of θ.

Question 3

(a) T: With $n > 10$, $T^+ = 9$ for 8 outcomes: $(r_1,\ldots,r_B) = (9)$, (1,8), (2,7), (3,6), (4,5), (1,2,6), (1,3,5), (2,3,4). Hence $\Pr(T^+ = 9) = 8/2^n = 2^{3-n}$ under $H_0 : \theta = 0$.

Remark: In this problem, $X_1,\ldots,X_n$ are assumed to be continuous, which implies zero probability of any ties. The distribution of T^+ *conditional on* observed ties in the data is not the same as the (unconditional) distribution of T^+.

(b) F: Under H_0, T^+ is symmetric about its median $a = n(n+1)/4 \geq 11(12)/4 = 33$, hence $\Pr(T^+ \geq 30) > \Pr(T^+ \geq 33) \geq \Pr(T^+ \geq a) \geq 0.5$.

(c) T: $X_{(5)} < 0 \Rightarrow X_{(1)} < \cdots < X_{(5)} < 0$. Hence $X_i + X_j < 0$ for at least $5(6)/2 = 15$ pairs $\{(i,j) : 1 \leq i \leq j \leq n\}$, and each gives a corresponding $W_{ij} < 0$.

Question 4

(a) T: Intuitively, smaller values indicate a smaller median and a wider range points to a greater variance.

(b) T: Smaller values of $(Y_1,\ldots,Y_n)$ lead to lower *Y*-ranks and hence a smaller value of the Wilcoxon rank sum test statistic – which provides evidence for $\theta_X > \theta_Y$.

(c) F: Since $(Y_1,\ldots,Y_n)$ is likely to have a smaller median than that of $(X_1,\ldots,X_m)$, the Ansari-Bradley rank test may be unable to detect the difference between $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$, hence accept $\mathrm{Var}(X) = \mathrm{Var}(Y)$ even if the difference is significant.

Question 5

(a) The values of $Z_i = Y_i - X_i$, $i = 1,\ldots,10$, are calculated as

$$89,\ \ 87,\ \ -60,\ \ 68,\ \ 56,\ \ 114,\ \ -44,\ \ 100,\ \ 91,\ \ 27,\ \ -45$$

The data have 8 positive values in the sample of size $n = 11$ and $B \sim Bin(11, 0.5)$ under $H_0 : \theta = 0$. Hence the exact *p*-value of testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$ by the sign test is $\Pr(B \geq 8) = \Pr(B \leq 3) = (1 + 11 + 55 + 165)(0.5)^{11} = 0.1133$.

(b) The ordered values $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(11)}$ of $Z_1,\ldots,Z_{11}$ are:

$$-60,\ \ -45,\ \ -44,\ \ 27,\ \ 56,\ \ 68,\ \ 87,\ \ 89,\ \ 91,\ \ 100,\ \ 114$$

For $B \sim Bin(11, 0.5)$, $\Pr(B \geq 9) = 0.03271 < 0.05$ and $\Pr(B \geq 8) = 0.1133 > 0.05$. Hence the minimum achievable confidence level above 90% is $1 - \alpha = 1 - 2(0.03271) = 0.9346$ with $\alpha = 0.0654$. It follows that $b_{\alpha/2} = b_{0.03271} = 9$ and $C_\alpha = n + 1 - b_{\alpha/2} = 12 - 9 = 3$. Thus the exact 93.5% confidence interval of θ is given by $(Z_{(3)}, Z_{(9)}) = (-44, 91)$.

(c) The values of Z_i, $|Z_i|$, rank R_i of $|Z_i|$, $\psi_i = I_{\{Z_i > 0\}}$ and $\psi_i R_i$ are listed below:

i	Z_i	$\lvert Z_i \rvert$	R_i	ψ_i	$\psi_i R_i$
1	89	89	8	1	8
2	87	87	7	1	7
3	-60	60	5	0	0
4	68	68	6	1	6
5	56	56	4	1	4
6	114	114	11	1	11
7	-44	44	2	0	0
8	100	100	10	1	10
9	91	91	9	1	9
10	27	27	1	1	1
11	-45	45	3	0	0

The observed Wilcoxon signed rank statistic is $T^+ = 8 + 7 + 6 + 4 + 11 + 10 + 9 + 1 = 56$. Since $M = n(n+1)/2 = 11(12)/2 = 66$, the *p*-value for the Wilcoxon signed test of $H_0 : \theta = 0$ against $H_1 : \theta > 0$ is $\Pr(T^+ \geq 56) = \Pr(T^+ \leq 10)$. By enumeration,

$$p\text{-value} = \Pr(T^+ \leq 10) = \Pr(T^+ = 0) + \Pr(T^+ = 1) + \cdots + \Pr(T^+ = 10)$$

$$= \frac{1 + 1 + 1 + 2 + 2 + 3 + 4 + 5 + 6 + 8 + 10}{2^{11}} = \frac{43}{2048} = 0.0210 < 0.05$$

Thus there is sufficient evidence at the 5% level to reject H_0 in favour of H_1, so that the new technology is effective to increase production.

(d) Let $W_{(1)} \le W_{(2)} \le \cdots \le W_{(66)}$ be the ordered values of $\{(Z_i + Z_j)/2, 1 \le i \le j \le 11\}$ (Walsh averages). The ordered values of Walsh averages are listed below:

$W_{(1)} \le W_{(2)} \le \cdots \le W_{(66)}$											
k	$W_{(k)}$	k	$W_{(k)}$	k	$W_{(k)}$	k	$W_{(k)}$	k	$W_{(k)}$	k	$W_{(k)}$
1	-60	12	5.5	23	22.5	34	56	45	77.5	56	91
2	-52.5	13	6	24	23	35	57	46	78	57	91
3	-52	14	11.5	25	23.5	36	58	47	78.5	58	93.5
4	-45	15	12	26	27	37	59	48	79.5	59	94.5
5	-44.5	16	13.5	27	27	38	62	49	84	60	95.5
6	-44	17	14.5	28	27.5	39	63.5	50	85	61	100
7	-16.5	18	15.5	29	28	40	68	51	87	62	100.5
8	-9	19	20	30	34.5	41	70.5	52	88	63	101.5
9	-8.5	20	21	31	35	42	71.5	53	89	64	102.5
10	-2	21	21.5	32	41.5	43	72.5	54	89	65	107
11	4	22	22	33	47.5	44	73.5	55	90	66	114

The estimate of θ by the signed ranks is

$$\hat{\theta} = \frac{W_{(M/2)} + W_{(M/2+1)}}{2} = \frac{W_{(66/2)} + W_{(66/2+1)}}{2} = \frac{W_{(33)} + W_{(34)}}{2} = \frac{47.5 + 56}{2} = 51.75$$

Next, $\Pr(T^+ \ge 56) = 0.0210 < 0.025$ and

$$\Pr(T^+ \ge 55) = \Pr(T^+ \le 11) = \Pr(T^+ \le 10) + \Pr(T^+ = 11) = \frac{43 + 12}{2048} = 0.0269 > 0.025$$

Hence the exact confidence interval of θ with at least 95% confidence level based on the Wilcoxon signed ranks is $(W_{(66+1-56)}, W_{(56)}) = (W_{(11)}, W_{(56)}) = (4, 91)$ with confidence level $1 - 2(0.0210) = 0.9580 =$ 95.8%.

(e) The sign test has *p*-values greater than 10%, which shows insufficient evidence for $\theta > 0$ at the 10% level. The exact *p*-value of the Wilcoxon signed rank test, on the other hand, is 0.0210, which provides sufficient evidence at the 5% level for $\theta > 0$. This shows that the Wilcoxon signed rank test is more powerful and efficient than the sign test to detect the difference between paired samples based on the same data.

The 95.8% confidence interval $(4, 91)$ of θ based on the Wilcoxon signed ranks is shorter and of higher confidence level, indicating more accurate estimation, than the 93.5% confidence interval $(-44, 91)$ based on the sign statistic.

Question 6

(a) Since $f_i(x)$ is symmetric about 0, $\Pr(X_i > 0) = 0.5 = \Pr(X_i < 0)$, $i = 1,2$. Hence

$$\Pr(S = 2) = \Pr(X_1 < 0, X_2 > 0) = \Pr(X_1 < 0)\Pr(X_2 > 0) = 0.5^2 = 0.25$$

Next, the cdf $F_2(x)$ of X_2 is given by

$$F_2(x) = \begin{cases} \int_{-\infty}^{x} f_2(t)dt = \int_{-\infty}^{x} e^{-2|t|}dt = \int_{-\infty}^{x} e^{2t}dt = 0.5e^{2x} & \text{if } x \le 0 \\ F_2(0) + \int_0^x e^{-2t}dt = 0.5 + 0.5\left(1 - e^{-2x}\right) = 1 - 0.5e^{-2x} & \text{if } x > 0 \end{cases}$$

Since X_1 and X_2 are independent and continuous,

$$\Pr(X_1 > 0, R_1 = 2, X_2 < 0) = \Pr(X_1 > 0, |X_1| > |X_2|, X_2 < 0) = \Pr(0 < -X_2 < X_1)$$

$$= \int_0^\infty \Pr(0 < -X_2 < x \mid X_1 = x) f_1(x)dx = \int_0^\infty \Pr(-x < X_2 < 0) f_1(x)dx$$

$$= \int_0^\infty [F_2(0) - F_2(-x)] f_1(x)dx = \int_0^1 \left[0.5 - 0.5e^{-2x}\right] 0.5dx$$

$$= 0.25\left[1 + 0.5e^{-2x}\Big|_0^1\right] = 0.25\left[1 + 0.5\left(e^{-2} - 1\right)\right] = 0.125 + 0.125e^{-2}$$

Similarly,

$$\Pr(X_1 < 0, R_2 = 2, X_2 > 0) = \Pr(0 < -X_1 < X_2) = \int_{-\infty}^0 \Pr(X_2 > -x) f_1(x)dx$$

$$= \int_{-\infty}^0 \left[1 - F_2(-x)\right] f_1(x)dx = \int_{-1}^0 0.5e^{2x} 0.5dx = 0.25\int_{-1}^0 e^{2x}dx$$

$$= 0.125e^{2x}\Big|_{-1}^0 = 0.125\left(1 - e^{-2}\right) = 0.125 - 0.125e^{-2}$$

Hence

$$\Pr(T^+ = 1) = \Pr(X_1 > 0, R_1 = 2, X_2 < 0) + \Pr(X_1 < 0, R_2 = 2, X_2 > 0) = 0.25$$

(b) The support of $f(x)$ is $[-0.5,1]$ and its cdf $F(x)$ on $[-0.5,1]$ is given by

$$F(x) = \int_{-0.5}^x f(t)dt = \int_{-0.5}^x dt = x + 0.5 \quad \text{if } -0.5 \le x \le 0;$$

$$F(x) = F(0) + \int_0^x f(t)dt = 0.5 + \int_0^x 2(1-t)^3 dt = 0.5 - \frac{2}{4}(1-t)^4\Big|_0^x$$

$$= 0.5 - 0.5\left[(1-x)^4 - 1\right] = 1 - 0.5(1-x)^4 \quad \text{if } 0 \le x \le 1.$$

Therefore, $\Pr(X_i < 0) = F(0) = 0.5 = \Pr(X_i > 0)$, $i = 1,2$. It then follows from the same arguments as in part (a) that $\Pr(S = 2) = 0.25$.

Next, since $f_1(x) = f_2(x) = f(x) = 1$ and $0 \le -x < 0.5$ for $-0.5 \le x < 0$, by the similar arguments as in part (a), we obtain

$$\Pr(X_1 > 0, R_1 = 2, X_2 < 0) = \Pr(X_1 < 0, R_2 = 2, X_2 > 0) = \Pr(0 < -X_1 < X_2)$$

$$= \int_{-\infty}^{0} [1 - F(-x)] f(x) dx = \int_{-0.5}^{0} 0.5(1+x)^4 dx = \frac{0.5}{5}(1+x)^5 \Big|_{-0.5}^{0}$$

$$= \frac{1}{10}[1 - (1-0.5)^5] = \frac{1}{10}\left[1 - \frac{1}{32}\right] = \frac{31}{320}$$

It follows that

$$\Pr(T^+ = 2) = 2\Pr(0 < -X_1 < X_2) = 2 \times \frac{31}{320} = \frac{31}{160} \neq 0.25 = \Pr(S = 2)$$

(c) In Part (a), the densities $f_1(x)$ and $f_2(x)$ are symmetric but not identical in this case, and $\Pr(T^+ = 2) = \Pr(S = 2)$. It is easy to see that $\Pr(T^+ = t) = \Pr(S = t)$ for $t = 0, 3$, hence for $t = 1$ as well. This verifies $T^+ \sim S$ when the distributions of X_1 and X_2 are symmetric but not identical.

Part (b) shows that $\Pr(T^+ = 2) \neq \Pr(S = 2)$ when $f_1(x)$ and $f_2(x)$ are identical but not symmetric. This demonstrates that T^+ does not necessarily have the same distribution as that of S if the distributions of X_1 and X_2 are identical but not symmetric.

Question 7

(a) The ordered values $Z_1 \le \cdots \le Z_{10}$ of combined samples are $(-3,-1,-1,1,1,3,6,8,12,12)$. The ranks of $(Z_1,\ldots,Z_{10})$ would be $(1,2,3,4,5,6,7,8,9,10)$ with no ties. Hence the ranks conditional on observed ties are $(r_1,\ldots,r_{10}) = (1,2.5,2.5,4.5,4.5,6,7,8,9.5,9.5)$. The observed ranks of $(Y_1,\ldots,Y_4) = (Z_3,Z_7,Z_5,Z_{10})$ are $(r_3,r_7,r_5,r_{10}) = (2.5,7,4.5,9.5)$.
It follows that $w = 2.5+7+4.5+9.5 = 23.5$. Each possible value w of W is given by $w = r_i + r_j + r_k + r_l$, where r_i,r_j,r_k,r_l are drown from $(r_1,\ldots,r_{10})$ with $i<j<k<l$.

Under $H_0:\Delta = 0$, each 4-tuple (r_i,r_j,r_k,r_l) with $i<j<k<l$ is equally likely from a total number of $10(9)(8)(7)/4(3)(2) = 10(3)(7) = 210$ possible outcomes. The following outcomes of (r_i,r_j,r_k,r_l) satisfy $w = r_i + r_j + r_k + r_l = 23.5$ with $i<j<k<l$:

Y-ranks	Outcomes	Count
(2.5,4.5,7,9.5)	(r_2,r_4,r_7,r_9), (r_3,r_4,r_7,r_9), (r_2,r_5,r_7,r_9), (r_3,r_5,r_7,r_9) (r_2,r_4,r_7,r_{10}), (r_3,r_4,r_7,r_{10}), (r_2,r_5,r_7,r_{10}), (r_3,r_5,r_7,r_{10})	8
(1,6,7,9.5)	(r_1,r_6,r_7,r_9), (r_1,r_6,r_7,r_{10})	2
(2.5,6,7,8)	(r_2,r_6,r_7,r_8), (r_3,r_6,r_7,r_8)	2

Thus $\Pr(W = w) = \Pr(W = 23.5) = (8+2+2)/210 = 12/210 = 2/35$.

(b) The scores of $Z_1 \le \cdots \le Z_{10}$ for the Ansari-Bradley test statistic C with no ties are $(1,2,3,4,5,5,4,3,2,1)$. Hence the scores of $(Z_1,\ldots,Z_{10})$ conditional on observed ties are $(a_1,\ldots,a_{10}) = (1,2.5,2.5,4.5,4.5,5,4,3,1.5,1.5)$. The observed scores of $(Y_1,\ldots,Y_4)$ for C are $(a_3,a_7,a_5,a_{10}) = (2.5,4,4.5,1.5)$, hence $c = 2.5+4+4.5+1.5 = 12.5$.
Each possible value c of C is given by $c = a_i + a_j + a_k + a_l$, where a_i,a_j,a_k,a_l are drown from $(a_1,\ldots,a_{10})$ with $i<j<k<l$. Under $H_0:\gamma^2 = 1$, each (a_i,a_j,a_k,a_l) is equally likely from a total number of 210 possible outcomes. The following outcomes of (a_i,a_j,a_k,a_l) satisfy $c = a_i + a_j + a_k + a_l = 12.5$, $i<j<k<l$:

Y-scores	Outcomes	Count
(2.5,4.5,4,1.5)	(a_2,a_4,a_7,a_9), (a_3,a_4,a_7,a_9), (a_2,a_5,a_7,a_9), (a_3,a_5,a_7,a_9) (a_2,a_4,a_7,a_{10}), (a_3,a_4,a_7,a_{10}), (a_2,a_5,a_7,a_{10}), (a_3,a_5,a_7,a_{10})	8
(1,2.5,5,4)	(a_1,a_2,a_6,a_7), (a_1,a_3,a_6,a_7)	2
(1,4.5,4,3)	(a_1,a_4,a_7,a_8), (a_1,a_5,a_7,a_8)	2
(4.5,5,1.5,1.5)	(a_4,a_6,a_9,a_{10}), (a_5,a_6,a_9,a_{10})	2
(1,2.5,4.5,4.5)	(a_1,a_2,a_4,a_5), (a_1,a_3,a_4,a_5)	2
(2.5,2.5,4.5,3)	(a_2,a_3,a_4,a_8), (a_2,a_3,a_5,a_8)	2

Thus $\Pr(C = c) = \Pr(C = 12.5) = (8+2\times 5)/210 = 18/210 = 3/35$.

Question 8

First input the data into vectors x and y:

```
x <-c(6.17, 4.78, 3.99, 5.65, 3.87, 4.43, 4.82, 6.68, 4.46, 6.95, 3.02, 4.22, 4.21, 3.97)
y <-c(9.94, 7.08, 7.14, 5.82, 9.60, 10.09, 8.66, 4.74, 4.14, 10.92, 5.61, 6.47, 5.20, 8.21, 3.55, 9.81)
```

(a) The R-command and output of the two-sample Wilcoxon rank sum test of $H_0 : \Delta = 0$ against $H_1 : \Delta > 0$ on the location-shift Δ are shown below:

```
> wilcox.test(y, x, alternative = "greater")

        Wilcoxon rank sum test

data:  y and x
W = 182, p-value = 0.001423
alternative hypothesis: true location shift is greater than 0
```

The p-value 0.001423 shows very strong evidence (even at the 0.2% level) that Y has a greater location parameter than X.

(b) The R-command and output of the Ansari-Bradley rank test of $H_0 : \gamma^2 = 1$ against $H_1 : \gamma^2 \neq 1$ based on X and Y are shown below:

```
> ansari.test(y,x)

        Ansari-Bradley test

data:  y and x
AB = 119, p-value = 0.4846
alternative hypothesis: true ratio of scales is not equal to 1
```

The p-value 0.4846 shows little evidence (even at the 40% level) that the two samples X and Y have different dispersions.

(c) The R-command and output of the Ansari-Bradley rank test of $H_0 : \gamma^2 = 1$ against $H_1 : \gamma^2 \neq 1$ based on $X + 2$ and Y are shown below:

```
> ansari.test(y,x+2)

        Ansari-Bradley test

data:  y and x + 2
AB = 96, p-value = 0.007276
alternative hypothesis: true ratio of scales is not equal to 1
```

The p-value 0.007276 shows very strong evidence (at the 1% level) that the two samples $X + 2$ and Y have different dispersions.

Comment: Part (b) shows that the Ansari-Bradley rank test is unable to detect the difference in dispersions between X and Y due to significantly different locations between X and Y. In Part (c), after the location difference is reduced by adding 2 to sample X, the Ansari-Bradley rank test finds strong evidence for difference in dispersions between $X + 2$ and Y, hence between X and Y.