Stats 200 **Autumn, 2016**

Solutions to Homework 4

Solutions by Pragya Sur

4.1 Problem 1

4.1.1 Part a

Suppose the given values of $|D_1|, \cdots, |D_n|$ are $d_1, \cdots, d_n$. Then the only values T can take are its values corresponding to the arguments $(\pm d_1, \cdots, \pm d_n)$, and due to symmetry under the null each of them is equally likely. That is, T is equal to $T(a_1, \cdots, a_n)$ with probability $1/2^n$ for $(a_1, \cdots, a_n)$ being each of the 2^n tuples $(\pm d_1, \cdots, \pm d_n)$.

4.1.2 Part b

Generate n IID Bernoulli(1/2) random variables and define variables $Z_i = 1$ if the ith Bernoulli variable is 1 and $Z_i = -1$ otherwise. (Then $Z_i's$ are n IID signs.) Then compute $T(Z_1 D_1, \cdots, Z_n D_n)$. Repeat this procedure a large number of times, say $B = 10000$ times, to generate B values for the statistic T. This approximates the conditional distribution of T given $|D_1|, \ldots, |D_n|$ under H_0. To perform a level α test of H_0 based on T, one can reject H_0 if $T(D_1, \ldots, D_n)$ exceeds the $(\alpha B)^{\text{th}}$ largest simulated value.

4.1.3 Part c

If each $D_i = X_i - Y_i$, then assigning a random sign to the ith coordinate is equivalent to permuting X_i and Y_i, so the test in part b may be interpreted as a permutation test.

In the general paired sample case, to determine the rejection threshold of a test of H_0 based on T, one can do the following. For each paired sample, generate a Bernoulli(1/2) variable. If it is 1, swap X_i and Y_i, otherwise do not swap. Call the new values $X_1^*, \cdots, X_n^*, Y_1^*, \cdots, Y_n^*$, and compute $T = T(X_1^*, \cdots, X_n^*, Y_1^*, \cdots, Y_n^*)$. Repeat this procedure a large number of times, say $B = 10000$ times, and compute the value of T each time. The rejection threshold may be taken as the $(\alpha B)^{\text{th}}$ largest simulated value, as in part (a).

4.2 Problem 2

4.2.1 Part a

Given that $X \sim N(\frac{h}{\sqrt{n}}, 1)$, we have

$$\mathbb{P}[X > 0] = \mathbb{P}\left[X - \frac{h}{\sqrt{n}} > -\frac{h}{\sqrt{n}}\right] = 1 - \Phi\left(-\frac{h}{\sqrt{n}}\right) = \Phi\left(\frac{h}{\sqrt{n}}\right).$$

A first order Taylor expansion for a differentiable function f suggests that

$$f(x+h) \approx f(x) + hf'(x)$$

Applying this to the above and noting $\Phi'(x)$ is the normal PDF $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$,

$$\Phi(\frac{h}{\sqrt{n}}) \approx \Phi(0) + \frac{h}{\sqrt{n}}\phi(0) = \frac{1}{2} + \frac{h}{\sqrt{2\pi n}}$$

4.2.2 Part b

The sign statistic S can be written as

$$S = \sum_i Y_i, \quad \text{where} \quad Y_i \sim \text{Bernoulli}(\mathbb{P}[X_i > 0]).$$

By the CLT, $\sqrt{n}(\frac{S}{n} - \mathbb{E}[Y_i])$ is approximately distributed as $\mathcal{N}(0, \text{Var}[Y_i])$. Applying part (a), $\mathbb{E}[Y_i] \approx \frac{1}{2} + \frac{h}{\sqrt{2\pi n}}$, so

$$\sqrt{n}\left(\frac{S}{n} - \mathbb{E}[Y_i]\right) \approx \sqrt{n}\left(\frac{S}{n} - \frac{1}{2} - \frac{h}{\sqrt{2\pi n}}\right) = \frac{1}{\sqrt{n}}\left(S - \frac{n}{2}\right) - \frac{h}{\sqrt{2\pi}}.$$

For large n,

$$\text{Var}[Y_i] \approx \left(\frac{1}{2} + \frac{h}{\sqrt{2\pi n}}\right)\left(1 - \left(\frac{1}{2} + \frac{h}{\sqrt{2\pi n}}\right)\right) \approx \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

So $\frac{1}{\sqrt{n}}(S - \frac{n}{2})$ is approximately distributed as $\mathcal{N}(\frac{h}{\sqrt{2\pi}}, \frac{1}{4})$. Multiplying by 2, $\sqrt{\frac{4}{n}}(S - \frac{n}{2})$ is approximately distributed as $\mathcal{N}(\frac{2h}{\sqrt{2\pi}}, 1)$.

The power of the sign test against the alternative $N(\frac{h}{\sqrt{n}}, 1)$ is given by

$$\mathbb{P}\left[S > \frac{n}{2} + \sqrt{\frac{n}{4}}z(\alpha)\right] = \mathbb{P}\left[\sqrt{\frac{4}{n}}\left(S - \frac{n}{2}\right) - \frac{2h}{\sqrt{2\pi}} > z(\alpha) - \frac{2h}{\sqrt{2\pi}}\right] \approx 1 - \Phi\left(z(\alpha) - \frac{2h}{\sqrt{2\pi}}\right) = \Phi\left(\frac{2h}{\sqrt{2\pi}} - z(\alpha)\right).$$

4.2.3 Part c

Note that $\mu = h/\sqrt{n}$ and $n = 100$, which implies $h = 1, 2, 3, 4$ respectively. Plugging this in the power formula, we get the powers of the sign test are $0.1985, 0.4804, 0.773$ and 0.939 respectively. These are close to the answers from Homework 3.

4.2.4 Part d

For $\mu = 0.2$, $h = 0.2\sqrt{n}$, and we obtain the sample size by solving

$$\Phi\left(\frac{0.4\sqrt{n}}{\sqrt{2\pi}} - z(0.05)\right) = 0.9$$

This gives $n = 336.2917$, rounding up gives 337.

4.3 Problem 3

4.3.1 Part a

Each person in each group is selected independently from either the high risk group or the low risk group. So the cholesterol level for each person in each group is a random variable independent of that for any other person. Also, since for both the treatment and control groups, with probability 1/2 a high risk individual is chosen and with probability 1/2 a low risk individual is chosen, they must have the same distribution.

So, the variables $X_1, \cdots, X_n, Y_1, \cdots, Y_n$ are IID from a common distribution. To compute the mean and variance, we may write X_i as

$$X_i = Z_i H_i + (1 - Z_i) L_i \tag{4.1}$$

where $H_i \sim \mathcal{N}(\mu_H, \sigma^2)$, $L_i \sim \mathcal{N}(\mu_L, \sigma^2)$, $Z_i \sim \text{Bernoulli}(1/2)$, and these are independent. Then

$$\begin{aligned}\mathbb{E}[X_i] &= \mathbb{E}[Z_i]\,\mathbb{E}[H_i] + \mathbb{E}[1 - Z_i]\,\mathbb{E}[L_i] \quad \text{(independence)} \\ &= \frac{1}{2}\mu_H + \frac{1}{2}\mu_L.\end{aligned}$$

To compute the variance, we have

$$\mathbb{E}[X_i^2] = \mathbb{E}[Z_i^2 H_i^2 + 2Z_i(1 - Z_i)H_i L_i + (1 - Z_i)^2 L_i^2].$$

Note that since $Z_i \in \{0, 1\}$, $Z_i(1 - Z_i) = 0$, $Z_i^2 = Z_i$, and $(1 - Z_i)^2 = (1 - Z_i)$. Then

$$\mathbb{E}[X_i^2] = \mathbb{E}[Z_i]\,\mathbb{E}[H_i^2] + \mathbb{E}[1 - Z_i]\,\mathbb{E}[L_i^2] = \frac{1}{2}\,\mathbb{E}[H_i^2] + \frac{1}{2}\,\mathbb{E}[L_i^2].$$

We have $\mathbb{E}[H_i^2] = \text{Var}[H_i] + (\mathbb{E}[H_i])^2 = \mu_H^2 + \sigma^2$, and similarly $\mathbb{E}[L_i^2] = \mu_L^2 + \sigma^2$. So

$$\mathbb{E}[X_i^2] = \frac{1}{2}(\mu_L^2 + \mu_H^2) + \sigma^2,$$

and

$$\text{Var}[X_i] = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = \frac{1}{2}(\mu_L^2 + \mu_H^2) + \sigma^2 - \frac{1}{4}(\mu_L^2 - 2\mu_L\mu_H + \mu_H^2) = \sigma^2 + \frac{1}{4}(\mu_H - \mu_L)^2.$$

4.3.2 Part b

As the X_i's and Y_i's are all IID, by the Central Limit Theorem, $\sqrt{n}(\bar{X} - \mathbb{E}[X_i]) \to \mathcal{N}(0, \text{Var}[X_i])$ and $\sqrt{n}(\bar{Y} - \mathbb{E}[X_i]) \to \mathcal{N}(0, \text{Var}[X_i])$ in distribution, so their difference $\sqrt{n}(\bar{X} - \bar{Y}) \to \mathcal{N}(0, 2\,\text{Var}[X_i])$. The pooled variance is

$$S_p^2 = \frac{1}{2n - 2}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right) = \frac{1}{2}S_X^2 + \frac{1}{2}S_Y^2,$$

where $S_X^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$ and $S_Y^2 = \frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2$ are the individual sample variances. By the result at the end of Lecture 10, $S_X^2 \to \text{Var}[X_i]$ and $S_Y^2 \to \text{Var}[Y_i] = \text{Var}[X_i]$ in probability, so the Continuous Mapping Theorem implies $S_p^2 \to \text{Var}[X_i]$. Then

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{n}}} = \frac{\sqrt{\text{Var}[X_i]}}{S_p}\,\frac{\sqrt{n}(\bar{X} - \bar{Y})}{\sqrt{2\,\text{Var}[X_i]}} \to \mathcal{N}(0, 1)$$

in distribution by Slutsky's lemma. Hence, a test that rejects for $T > z(\alpha)$ is approximately level α for large n.

4.3.3 Part c

The difference in this part from Part a is that here

$$X_i = Z_i H_i + (1 - Z_i) L_i, \tag{4.2}$$

where H_i, L_i are defined as before but $Z_i \sim \text{Bernoulli}(p)$. Then

$$\mathbb{E}[X_i] = p\mu_H + (1 - p)\mu_L.$$

Similarly, $\mathbb{E}[Y_i] = q\mu_H + (1-q)\mu_L$.

For the variances, we compute as in part (a)

$$\mathbb{E}[X_i^2] = \mathbb{E}[Z_i]\,\mathbb{E}[H_i^2] + \mathbb{E}[1-Z_i]\,\mathbb{E}[L_i^2] = p(\mu_H^2+\sigma^2) + (1-p)(\mu_L^2+\sigma^2) = p\mu_H^2 + (1-p)\mu_L^2 + \sigma^2,$$

so

$$\mathrm{Var}[X_i] = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = p\mu_H^2 + (1-p)\mu_L^2 + \sigma^2 - (p\mu_H + (1-p)\mu_L)^2 = \sigma^2 + (\mu_H-\mu_L)^2 p(1-p).$$

Similarly, $\mathrm{Var}[Y_i] = \sigma^2 + (\mu_H-\mu_L)^2 q(1-q)$.

4.3.4 Part d

In this case $S_X^2 \to \mathrm{Var}[X_i]$ and $S_Y^2 \to \mathrm{Var}[Y_i]$ in probability, so

$$S_p^2 \to \frac{1}{2}(\mathrm{Var}[X_i] + \mathrm{Var}[Y_i]) = \sigma^2 + \frac{1}{2}\big(p(1-p) + q(1-q)\big)(\mu_H-\mu_L)^2 =: c.$$

By the CLT, $\sqrt{n}(\bar{X} - \mathbb{E}[X_i]) \to \mathcal{N}(0, \mathrm{Var}[X_i])$ and $\sqrt{n}(\bar{Y} - \mathbb{E}[Y_i]) \to \mathcal{N}(0, \mathrm{Var}[Y_i])$. The X_i's and Y_i's are independent, so the difference $\sqrt{n}(\bar{X} - \bar{Y} - \mathbb{E}[X_i] + \mathbb{E}[Y_i]) \to \mathcal{N}(0, \mathrm{Var}[X_i] + \mathrm{Var}[Y_i])$.

Then

$$T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{2/n}} = \frac{1}{\sqrt{2S_p^2}}(\sqrt{n}(\bar{X}-\bar{Y}))$$

is approximately distributed as

$$\frac{1}{\sqrt{2c}}\mathcal{N}\left(\sqrt{n}(\mathbb{E}[X_i] - \mathbb{E}[Y_i]), \mathrm{Var}[X_i] + \mathrm{Var}[Y_i]\right) = \mathcal{N}\left(\frac{\sqrt{n}(\mathbb{E}[X_i] - \mathbb{E}[Y_i])}{\sqrt{2c}}, 1\right).$$

Let

$$m := \frac{\sqrt{n}(\mathbb{E}[X_i] - \mathbb{E}[Y_i])}{\sqrt{2c}} = \frac{\sqrt{n}(p-q)(\mu_H-\mu_L)}{\sqrt{2c}} = \frac{\sqrt{n}(p-q)(\mu_H-\mu_L)}{\sqrt{2\sigma^2 + \big(p(1-p)+q(1-q)\big)(\mu_H-\mu_L)^2}},$$

so T is approximately $\mathcal{N}(m,1)$. Then the rejection probability is

$$\mathbb{P}[T > z(\alpha)] = \mathbb{P}[T - m > z(\alpha) - m] \approx 1 - \Phi(z(\alpha) - m) = \Phi(m - z(\alpha)).$$

This probability is increasing in m, and only equals α when $m = 0$. If $(p-q)(\mu_H-\mu_L) > 0$, then $m \to \infty$ as $n \to \infty$, and we expect to falsely reject H_0 with probability close to 1 for large n. If $(p-q)(\mu_H-\mu_L) < 0$, then $m \to -\infty$ as $n \to \infty$, and we expect the significance level of the test to in fact be close to 0 for large n.

4.4 Problem 4

4.4.1 Part a

We know that $P_1, \cdots, P_n \sim U(0,1)$ (IID). So for any $t \in (0,1)$,

$$\begin{aligned}
\mathbb{P}[\min_{i=1}^{n} P_i \le t] &= 1 - \mathbb{P}[\min_{i=1}^{n} P_i > t] \\
&= 1 - \mathbb{P}[P_i > t \;\;\forall\;\; i = 1, \cdots, n] \\
&= 1 - \prod_{i=1}^{n} \mathbb{P}[P_i > t] \\
&= 1 - (1-t)^n
\end{aligned}$$

4.4.2 Part b

If all the tests are performed at significance level $1-(1-\alpha)^{1/n}$,

$$\begin{aligned}\mathbb{P}(\text{rejecting any of the } n \text{ null hypotheses}) &= \mathbb{P}(P_i < 1-(1-\alpha)^{1/n} \text{ for any i}) \\ &= \mathbb{P}(\min_{i=1}^{n} P_i < 1-(1-\alpha)^{1/n}) \\ &= 1-(1-1+(1-\alpha)^{1/n})^n = \alpha.\end{aligned}$$

Hence, the probability of (falsely) rejecting any of the n null hypothesis is exactly α.

The Bonferroni procedure rejects when $P_i \le \alpha/n$ and the above procedure rejects when $P_i \le 1-(1-\alpha)^{1/n}$. Note that

$$\left(1-\frac{\alpha}{n}\right)^n > 1-\alpha,$$

so $1-(1-\alpha)^{1/n} > \alpha/n$. Hence whenever the Bonferroni test rejects, this procedure also rejects, so this procedure is more powerful than the Bonferroni test.

4.4.3 Part c

Suppose there are k true null hypotheses and without loss of generality let us assume that these are the first k. If all the tests are performed at significance level $1-(1-\alpha)^{1/n}$, and V is the number of true null hypotheses that are rejected, then the FWER is

$$\begin{aligned}\mathbb{P}(V \ge 1) &= \mathbb{P}(\min_{i=1}^{k} P_i \le 1-(1-\alpha)^{1/n}) \\ &= 1-\mathbb{P}(\min_{i=1}^{k} P_i > 1-(1-\alpha)^{1/n}) \\ &= 1-\mathbb{P}(P_i > 1-(1-\alpha)^{1/n} \ \forall\ i=1,\cdots,k) \\ &= 1-(1-1+(1-\alpha)^{1/n})^k \\ &= 1-(1-\alpha)^{k/n}\end{aligned}$$

Since $k \le n$ and $\alpha < 1$, $(1-\alpha)^{k/n} > (1-\alpha)$ and hence $1-(1-\alpha)^{k/n} < \alpha$, so the FWER is controlled.