

## Discretionary Note

Anish Krishna Lakkapragada

**IF YOU USE THIS FILE TO CHEAT, YOU ARE NOT ONLY STUPID BUT YOU ARE CHEATING YOURSELF OUT OF THE ABILITY TO FALL IN LOVE WITH MATH.** Furthermore, I am not smarter than you and my solutions did not always get a perfect score.

**CONTENT STARTS ON NEXT PAGE.**

To access the general instructions for this repository head [here](#).

# STATS 242 HW 5

February 18, 2025

---

**Number of late days: 0; Collaborators: Derek Gao**

1.

- (a) Under  $H_0$ , the median of  $f$  is zero  $\implies$  each sample  $X_i$  has a 50% chance of being above zero. Let us define r.v.  $I_i \sim \text{Bern}(0.5)$  to represent if  $X_i \geq 0$ . Because  $S = \sum_{i=1}^n I_i$ ,  $S \sim \text{Bin}(n, 0.5)$ . Furthermore, with a large  $n$ , the CLT enables us to approximate Binomial distributions with normal distributions<sup>1</sup> and thus  $S \sim \mathcal{N}(n\mathbb{E}[I_i], n\text{Var}[I_i])$  or  $S \sim \mathcal{N}(0.5n, 0.25n)$ . This means that the distribution of  $S - \frac{n}{2} \sim \mathcal{N}(0, 0.25n)$  and the distribution of  $T = \sqrt{\frac{4}{n}}(S - \frac{n}{2}) \sim \mathcal{N}(0, 1)$ . Thus  $T \sim \mathcal{N}(0, 1)$ .

To test  $H_0$  vs.  $H_1$  at the significance level  $\alpha$ , we would compute  $T$  for a sample of data and if it is above the upper- $\alpha$  point of  $\mathcal{N}(0, 1)$ , we will reject  $H_0$ .

- (b) Note that because under  $H'_1$ ,  $X_i \sim \mathcal{N}(\frac{h}{\sqrt{n}}, 1)$  this means that  $X_i - \frac{h}{\sqrt{n}} \sim \mathcal{N}(0, 1)$ .

$$\mathbb{P}_{H'_1}[X_i > 0] = \mathbb{P}_{H'_1}[X_i - \frac{h}{\sqrt{n}} > -\frac{h}{\sqrt{n}}] = 1 - \Phi(-\frac{h}{\sqrt{n}}) = \Phi(\frac{h}{\sqrt{n}})$$

Assuming that  $h$  is a small fixed value and  $n$  is large, then  $\frac{h}{\sqrt{n}}$  is close to zero. This means we can approximate  $\mathbb{P}_{H'_1}[X_i > 0] = \Phi(\frac{h}{\sqrt{n}})$  at zero with the following first-degree Taylor Series approximation:

$$\begin{aligned}\mathbb{P}_{H'_1}[X_i > 0] &= \Phi(\frac{h}{\sqrt{n}}) \\ &\approx \Phi(0) + \Phi'(\frac{h}{\sqrt{n}}) \cdot \frac{h}{\sqrt{n}} = \frac{1}{2} + \phi(\frac{h}{\sqrt{n}}) \cdot \frac{h}{\sqrt{n}} = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} e^{-\frac{h^2}{2n}} \cdot \frac{h}{\sqrt{n}}\end{aligned}$$

where  $\phi$  is the standard normal PDF. Note that because  $h \ll n$ ,  $\frac{h^2}{2n} \approx 0$  and so:

$$\begin{aligned}\mathbb{P}_{H'_1}[X_i > 0] &= \Phi(\frac{h}{\sqrt{n}}) \approx \frac{1}{2} + \frac{1}{\sqrt{2\pi}} e^0 \cdot \frac{h}{\sqrt{n}} \\ \mathbb{P}_{H'_1}[X_i > 0] &\approx \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \cdot \frac{h}{\sqrt{n}}\end{aligned}$$

---

<sup>1</sup>This is because the Binomial Distribution is given by  $n(\bar{I})$  where  $\bar{I}$  is the mean of  $n$  i.i.d Bernoulli Variables.

- (c) Under  $H'_1$ , the previously defined r.v.  $I_i \sim \text{Bern}(\mathbb{P}_{H'_1}[X_i > 0])$  or  $I_i \sim \text{Bern}(\Phi(\frac{h}{\sqrt{n}}))$ . As stated in part (a), r.v.  $S$  is given by a Binomial distribution but for large  $n$  can be approximated by  $\mathcal{N}(n\mathbb{E}[I_i], n\text{Var}[I_i])$ . As such, we first compute  $n\mathbb{E}[I_i]$ :

$$\begin{aligned} n\mathbb{E}[I_i] &= n\mathbb{P}_{H'_1}[X_i > 0] = n\Phi(\frac{h}{\sqrt{n}}) \approx n(\frac{1}{2} + \frac{h}{\sqrt{2\pi n}}) = \frac{n}{2} + \frac{h\sqrt{n}}{\sqrt{2\pi}} \\ n\mathbb{E}[I_i] &\approx \frac{n}{2} + \frac{h\sqrt{n}}{\sqrt{2\pi}} \end{aligned}$$

and then  $n\text{Var}(I_i)$ :

$$\begin{aligned} n\text{Var}(I_i) &= n(\mathbb{P}_{H'_1}[X_i > 0])(1 - \mathbb{P}_{H'_1}[X_i > 0]) = n(\Phi(\frac{h}{\sqrt{n}}))(1 - \Phi(\frac{h}{\sqrt{n}})) \\ &\approx n(\frac{1}{2} + \frac{h}{\sqrt{2\pi n}})(1 - (\frac{1}{2} + \frac{h}{\sqrt{2\pi n}})) = n(\frac{1}{2} + \frac{h}{\sqrt{2\pi n}})(\frac{1}{2} - \frac{h}{\sqrt{2\pi n}}) = n(\frac{1}{4} - \frac{h^2}{2\pi n}) = \frac{n}{4} - \frac{h^2}{2\pi} \\ n\text{Var}(I_i) &\approx \frac{n}{4} - \frac{h^2}{2\pi} \end{aligned}$$

So, for large  $n$  we have that  $S$  can be approximated by  $\mathcal{N}(\frac{n}{2} + \frac{h\sqrt{n}}{\sqrt{2\pi}}, \frac{n}{4} - \frac{h^2}{2\pi})$ . This means that the distribution  $S - \frac{n}{2}$  is given by approximately  $\mathcal{N}(\frac{h\sqrt{n}}{\sqrt{2\pi}}, \frac{n}{4} - \frac{h^2}{2\pi})$  and so the distribution of  $T = \sqrt{\frac{4}{n}}(S - \frac{n}{2})$  is approximately  $\mathcal{N}(\sqrt{\frac{4}{n}} \cdot \frac{h\sqrt{n}}{\sqrt{2\pi}}, \frac{4}{n}(\frac{n}{4} - \frac{h^2}{2\pi}))$  or  $\mathcal{N}(\sqrt{\frac{2}{\pi}}h, 1 - \frac{2h^2}{\pi n})$ . Note that under our assumption  $h \ll n$ , we can drop the  $\frac{2h^2}{\pi n}$  term in the variance. Thus we can give the following normal approximation for  $T$  that only relies on  $h$  and not  $n$ :  $\mathcal{N}(\sqrt{\frac{2}{\pi}}h, 1)$ .

- (d) As stated in part (a), we would reject  $H_0$  if the computed test statistic  $T$  is above the upper- $\alpha$  point of  $\mathcal{N}(0, 1)$  (given by  $z^{(a)}$ ). The power of a test is given by  $\mathbb{P}_{H'_1}[\text{reject } H_0] = \mathbb{P}_{H'_1}[T > z^{(a)}]$ . As shown in part (c), under  $H'_1$ ,  $T$  can be approximated by  $\mathcal{N}(\sqrt{\frac{2}{\pi}}h, 1)$ . Thus, we can compute the power:

$$\begin{aligned} \mathbb{P}_{H'_1}[\text{reject } H_0] &= \mathbb{P}_{H'_1}[T > z^{(a)}] \approx \mathbb{P}[\mathcal{N}(\sqrt{\frac{2}{\pi}}h, 1) > z^{(a)}] \\ &= \mathbb{P}[\mathcal{N}(\sqrt{\frac{2}{\pi}}h, 1) - \sqrt{\frac{2}{\pi}}h > z^{(a)} - \sqrt{\frac{2}{\pi}}h] = \mathbb{P}[\mathcal{N}(0, 1) > z^{(a)} - \sqrt{\frac{2}{\pi}}h] \\ &= 1 - \mathbb{P}[\mathcal{N}(0, 1) \leq z^{(a)} - \sqrt{\frac{2}{\pi}}h] = 1 - \Phi(z^{(a)} - \sqrt{\frac{2}{\pi}}h) = \Phi(\sqrt{\frac{2}{\pi}}h - z^{(a)}) \end{aligned}$$

and so the power of this test against alternative  $H'_1$  is approximately  $\Phi(\sqrt{\frac{2}{\pi}}h - z^{(a)})$ .

- (a) The simulated probability of a Type I Error for the Sign Test was 0.04. The simulated probability of a Type I Error for the t-test was 0.0522. We report the simulated power against each alternative for both tests in the table below:

|           | $\mu = 0.1$ | $\mu = 0.2$ | $\mu = 0.3$ | $\mu = 0.4$ |
|-----------|-------------|-------------|-------------|-------------|
| Sign test | 0.1754      | 0.4509      | 0.7529      | 0.9287      |
| t-test    | 0.1629      | 0.5011      | 0.8408      | 0.9771      |

- (b) The power of the one-sample z-test is given by  $\Phi(\sqrt{n}\mu - z^{(\alpha)})$ . Comparing that to the simulated power of the one-sample t-test, we find that the z-test, across all alternatives, has a consistently greater power.

In problem 1(d), we found that the power of the sign test was approximately  $\Phi(\sqrt{\frac{2}{\pi}} \cdot \sqrt{n}\mu - z^{(\alpha)})$ . Comparing that to the simulated power of the sign test, we find that this approximation, across alternatives, yields a consistently greater power. Furthermore, our simulated power of the sign test is generally lower than our simulated power of the t-test and considerably lower than our power derivation for a one-sample z-test. We report the analytically computed powers for the z-test and sign test below:

|  | $\mu = 0.1$ | $\mu = 0.2$ | $\mu = 0.3$ | $\mu = 0.4$ |
|--|-------------|-------------|-------------|-------------|
| Sign test: $\approx \Phi(\sqrt{\frac{2}{\pi}} \cdot \sqrt{n}\mu - z^{(\alpha)})$ | 0.1985      | 0.4804      | 0.7730      | 0.9390      |
| z-test: $\Phi(\sqrt{n}\mu - z^{(\alpha)})$                                       | 0.2595      | 0.6388      | 0.9123      | 0.9907      |

```

1  # %%
2  import numpy as np
3  import math
4  from scipy import stats
5
6  NUM_SAMPLES = 10000
7  SIG_LEVEL = 0.05
8
9  upper_alpha = stats.norm.ppf(1 - SIG_LEVEL, loc=0, scale=1)
10
11 def get_N_normal_observations(mu, N=100):
12     return np.random.normal(mu, 1, N)
13
14 def reject_ttest(samples):
15     t_stat, p_value = stats.ttest_1samp(samples, popmean=0)
16     return 1 if p_value <= SIG_LEVEL else 0
17
18 def compute_sign_statistic(samples):
19     n = len(samples)
20     return math.sqrt(4 / n) * (np.sum(samples > 0) - 0.5 * n)
21
22 def reject_sign_test(samples):

```

```

23     sign_statistic = compute_sign_statistic(samples)
24     return 1 if sign_statistic >= upper_alpha else 0
25 # %%
26 MEANS = [0, 0.1, 0.2, 0.3, 0.4]
27 REJ_TTEST_COUNTS = [0] * len(MEANS)
28 REJ_SGN_COUNTS = [0] * len(MEANS)
29
30 for m_i, mean in enumerate(MEANS):
31     for _ in range(NUM_SAMPLES):
32         samples = get_N_normal_observations(mean)
33         rej_ttest = reject_ttest(samples)
34         rej_sign_test = reject_sign_test(samples)
35         REJ_TTEST_COUNTS[m_i] += rej_ttest
36         REJ_SGN_COUNTS[m_i] += rej_sign_test
37
38 # %%
39 """Get the simulated Type I Error."""
40 print(f"Type I Error for Sign Statistic: {REJ_SGN_COUNTS[0] /
    ↪ NUM_SAMPLES}")
41 print(f"Type I Error for T-Test Statistic: {REJ_TTEST_COUNTS[0] /
    ↪ NUM_SAMPLES}")# %%
42 """Get the Power Against Each Alternative"""
43 for i, mean in enumerate(MEANS):
44     if mean == 0: continue
45     print(f"Power for sign test @ {mean} mean: {REJ_SGN_COUNTS[i] /
    ↪ NUM_SAMPLES}")
46
47 for i, mean in enumerate(MEANS):
48     if mean == 0: continue
49     print(f"Power for t-test @ {mean} mean: {REJ_TTEST_COUNTS[i] /
    ↪ NUM_SAMPLES}")
50
51 """Compare to z-test"""
52 for i, mean in enumerate(MEANS):
53     if mean == 0: continue
54     print(f"Estimated power for z-test @ {mean} mean:
    ↪ {stats.norm.cdf(math.sqrt(100) * mean - upper_alpha)}")
55 # %%
56 """Compare to sign test"""
57 for i, mean in enumerate(MEANS):
58     if mean == 0: continue
59     print(f"Estimated power for sign-test @ {mean} mean:
    ↪ {stats.norm.cdf(math.sqrt(2 / math.pi) * math.sqrt(100) * mean
    ↪ - upper_alpha)}")

```

3.

- (a) We are given that the FWER is controlled at level  $\alpha \implies \mathbb{P}[\text{reject any true } H_0] \leq \alpha$ . Let us define  $V$  and  $R$  as the number of true null hypotheses rejected and the number of total null hypotheses rejected, respectively. The FDR is controlled at level  $\alpha$  if  $\mathbb{E}[\frac{V}{R}] \leq \alpha$ . Using LOTE, we can write the FDR as the following:

$$\begin{aligned}\mathbb{E}[\frac{V}{R}] &= \mathbb{E}[\frac{V}{R} | \frac{V}{R} = 0]P(\frac{V}{R} = 0) + \mathbb{E}[\frac{V}{R} | \frac{V}{R} \neq 0]P(\frac{V}{R} \neq 0) \\ \mathbb{E}[\frac{V}{R}] &= (0)P(\frac{V}{R} = 0) + \mathbb{E}[\frac{V}{R} | \frac{V}{R} \neq 0]P(\frac{V}{R} \neq 0) = \mathbb{E}[\frac{V}{R} | \frac{V}{R} \neq 0]P(\frac{V}{R} \neq 0)\end{aligned}$$

We first compute  $P(\frac{V}{R} \neq 0) = P(V \neq 0) = \mathbb{P}[\text{reject any true } H_0]$ . Note that this last probability is guaranteed to be  $\leq \alpha$  by the FWER and so  $P(\frac{V}{R} \neq 0) \leq \alpha$ . Given this, we can create the following inequality for the FDR:

$$\mathbb{E}[\frac{V}{R}] \leq \mathbb{E}[\frac{V}{R} | \frac{V}{R} \neq 0]\alpha$$

This question asks us to consider if the FDR is necessarily controlled at level  $\alpha$ , given the FWER is. Note that  $V$  is strictly less than  $R$ , and so  $\frac{V}{R} \leq 1 \implies \mathbb{E}[\frac{V}{R} | \frac{V}{R} \neq 0] \leq 1 \implies \text{FDR} = \mathbb{E}[\frac{V}{R} \neq 0]\alpha \leq \alpha \implies$  the FDR is controlled at level  $\alpha$ .

- (b) The Bonferroni method applied to control  $\text{FWER} \leq \alpha$  will reject any null hypotheses that has a p-value  $\leq \frac{\alpha}{n}$ . The BH procedure will reject hypotheses where their p-value is less than a multiple (i.e. their rank  $r \in \mathbb{N}$ ) of  $\frac{\alpha}{n}$ : the BH procedure rejects hypotheses with p-values  $\leq \frac{\alpha r}{n}$ . Let us say that hypothesis  $H_k$  with p-value  $P_k$ . Let us also suppose that this hypothesis has a rank  $r_k$  when compared to all other hypotheses' p-values in this multiple hypotheses testing experiment. If  $H_k$  was rejected by the Bonferroni method  $\implies P_k \leq \frac{\alpha}{n} \leq \frac{\alpha r_k}{n} \implies P_k \leq \frac{\alpha r_k}{n} \implies P_k$  will be rejected by the BH procedure. Thus, all hypotheses rejected by the Bonferroni method will be rejected by the BH procedure.

4.

- (a) We compute this below:

$$\begin{aligned}\mathbb{P}[\text{reject any true null hypothesis}] &= 1 - \mathbb{P}[\text{reject no true null hypotheses}] \\ &= 1 - \prod_{i=1}^{n_0} \mathbb{P}[\text{accept this null hypothesis}]\end{aligned}$$

The probability of rejecting any true null hypothesis is given by the probability  $P_i \leq t$ . Because this null hypothesis is true,  $P_i \sim \text{Unif}(0, 1) \implies \mathbb{P}[P_i \leq t] = t$ . The probability of accepting any true null hypothesis is the complement of this probability,  $1 - t$ . Thus we have:

$$\mathbb{P}[\text{reject any true null hypothesis}] = 1 - \prod_{i=1}^{n_0} (1 - t) = 1 - (1 - t)^{n_0}$$

- (b) The FWER is given by  $\mathbb{P}[\text{reject any true hypothesis}]$ . As computed in (a), for a given cutoff  $t$ , this probability is given by  $1 - (1 - t)^{n_0}$ . Setting  $t = 1 - (1 - \alpha)^{\frac{1}{n}}$ , we have that:

$$\mathbb{P}[\text{reject any true null hypothesis}] = 1 - (1 - t)^{n_0} = 1 - (1 - (1 - (1 - \alpha)^{\frac{1}{n}}))^{n_0} = 1 - (1 - \alpha)^{\frac{n_0}{n}}$$

Because  $n_0 \leq n$  and  $1 - \alpha \leq 1$ ,  $(1 - \alpha)^{\frac{n_0}{n}} \geq (1 - \alpha) \implies -(1 - \alpha)^{\frac{n_0}{n}} \leq \alpha - 1$  and so:

$$\begin{aligned} \mathbb{P}[\text{reject any true null hypothesis}] &= 1 - (1 - \alpha)^{\frac{n_0}{n}} \leq 1 + (\alpha - 1) \\ \mathbb{P}[\text{reject any true null hypothesis}] &\leq \alpha \end{aligned}$$

and so we have shown for  $t = 1 - (1 - \alpha)^{\frac{1}{n}}$ , the FWER  $\leq \alpha$ , meaning that the FWER is controlled at level  $\alpha$ .

We now compare if this choice of  $t = 1 - (1 - \alpha)^{\frac{1}{n}}$  or  $t = \frac{\alpha}{n}$  will reject more hypotheses. Because whichever choice of  $t$  is greater will reject *more* hypotheses<sup>2</sup>, we aim to find which choice of  $t$  is greater. We first assume that  $n$ , the number of hypotheses tests, is  $\geq 1$  and that  $0 \leq \alpha \leq 1$ . Note by Bernoulli's inequality that for  $0 \leq r \leq 1$  and  $x \geq -1$ ,  $(1 + x)^r \leq 1 + rx$ . Because  $0 \leq \frac{1}{n} \leq 1$  and  $-\alpha \geq -1$ , we can apply Bernoulli's inequality and so  $(1 - \alpha)^{\frac{1}{n}} \leq 1 - \frac{\alpha}{n} \implies -(1 - \alpha)^{\frac{1}{n}} \geq \frac{\alpha}{n} - 1$ . So the choice of  $t = 1 - (1 - \alpha)^{\frac{1}{n}} \geq 1 + \frac{\alpha}{n} - 1 \implies t = 1 - (1 - \alpha)^{\frac{1}{n}} \geq \frac{\alpha}{n}$ . Thus choosing  $t = 1 - (1 - \alpha)^{\frac{1}{n}}$  will reject more hypotheses as it is a greater threshold.

w The procedure of choosing  $t = 1 - (1 - \alpha)^{\frac{1}{n}}$  differs from the Bonferroni correction because it assumes independence between all  $n$  hypothesis tests, which the Bonferroni correction does not assume. This makes sense as this stronger assumption allows this choice of  $t$  to reject more hypotheses when compared to the Bonferroni correction.

---

<sup>2</sup>this means that more computed p-values will meet this threshold  $\implies$  more hypotheses will be rejected