SAS - Assignment 2

牛圣杰 11910901

# Problem 1

We denote these two test as test 1 and test 2 respectively.

From the problem, we have that for test 1: the rejection region (RR) is $\{\bar{X}-110>c_1\}$ where $c_1$ is a constant to be determined.

and for test 2, $RR^*$ is $\{\bar{X}-110>c_2\}$ where $c_1$ is a constant to be determined.

Note that $\alpha = P_r(\bar{X}-110>c_1) = P_r(\bar{X}-110>c_2) \Rightarrow c_1=c_2. \Rightarrow RR=RR^*$

Thus, for any fixed observations, we have that if we can reject $H_0$ in the test 1, i.e., $\bar{X}-110\leq c_1, \Rightarrow \bar{X}-110\leq c_2$. We can not reject $H_0^*$ in test 2 as well.

In the similar way, if we can not reject $H_0$ in test 1, we can also not reject $H_0^*$ in test 2.

$\Rightarrow$ No matter what the observations are, the test 1 with $\alpha$ and test 2 share the same conclusion, which means a test with level $\alpha$ for testing $H_0: \theta=110$ v.s. $H_1: \theta>110$ is also a test with level $\alpha$ for testing $H_0^*: \theta\leq 110$ v.s. $H_1: \theta>110$.

# Problem 2

1°. P-值可以表示数据与特定统计模型间的不相容程度。我们在一定的假设以零假设条件下计算得到的P值越小,说明我们这组数据就越极端,越不可能得假定的模型中得到,即说明了这组数据和模型同时出现的概率很小,即两者之间的不相容性。

2°. P-值既不能单纯表示零假设的真实性,也不能单纯代表出现观测数据的可能性,它实际代表的是所观察数据与特定的假设解释之间的联系.

3° 不能直接通过P值判断结论和决策,这可能会导致研究和决策的扭曲。应该利用背景因素来得出科学的推断,比如研究设计,测量质量,外部证据以及数据分析基础假设的有效性等

4°. 恰当的统计推断需要完整报告且需保证透明度,比如,不能仅展示那些通过假设检验的显著变量,我们应该完整的报告所有探索的假设,所有数据收集的决策,具体的统计分析以及相对应的P值,实现完整,透明的研究流程.

5° P值的大小并不能代表实际影响的大小,只能用来判断某个效应或影响的存在与否,而不是效应或影响的大小.

6°. 当计算得到的P值与设定的阈值接近时,P值就不再能提供支持或反对原假设的结论,所以在数据分析中,除了P值外,我们也要积极尝试其它合理且有效的方法。

# Problem 3

(1). 1°. We choose the F test.

2°. Test statistic is $F=\frac{S_1^2}{S_2^2} \sim F(n_1-1, n_2-1) = F(16, 18)$ under $H_0$.

3° $F_{obs} = S_1^2/S_2^2 = 5.5^2/4.5^2 = 1.4938$, the rejection region is $\{F_{obs} > F_{\frac{\alpha}{2}}(16,18)\} \cup \{F_{obs} < F_{1-\frac{\alpha}{2}}(16,18)\}$

when $\alpha = 0.05 \Rightarrow$ RR is $\{F_{obs} > 2.6404\} \cup \{F_{obs} < 0.3681\}$. $F_{obs} \notin RR$

4°. P-value $= 2 \times \Pr(F < F_{obs} \mid H_0) = 2 \times 0.7950 = 1.59 > 0.05$

5°. Both from the rejection region and P value, we can not reject the null hypothesis $H_0$.

(2). An estimate interval for $\mu_1-\mu_2$ is $\bar{X}-\bar{Y}$, and $\sigma$ can be estimated by the pooled sample standard deviation.

$$S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{16\cdot 5.5^2+18\cdot 4.5^2}{34}} = 4.9956$$

$\Rightarrow$ test statistic is $T=\frac{\bar{X}_1-\bar{Y}}{\sqrt{S_p(\frac{1}{n_1}+\frac{1}{n_2})}} \sim t(n_1+n_2-2)$ under $H_0: \mu_1=\mu_2$.

$\Rightarrow$ 95% CI for $\mu_1-\mu_2$ is $\bar{x}-\bar{y} \pm t_{\frac{\alpha}{2},34} \times S_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 3 \pm 2.0322 \times 4.9956 \times 0.3338 = 3 \pm 3.389 = [-0.389, 6.389]$

(3) Note that the type II error rate is $P(\underset{\sim}{X} \notin RR \mid H_1 \text{ is true})$

From the problem, $H_0: \mu_1-\mu_2=0$ v.s. $H_1: \mu_1-\mu_2>0 \Rightarrow$ the Z-test statistic is $Z=\frac{(\hat{\mu}_1-\hat{\mu}_2)-(\mu_1-\mu_2)}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2}} \sim N(0,1)$.

Since $\mu_1-\mu_2=2.5$ and $n_1=19$, $n_2=17$, $\sigma=5$, we have that $\frac{\hat{\mu}_1-\hat{\mu}_2}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2}} \overset{d}{=} Z \sim N(1.498, 1)$ under $H_0$.

Thus we have that $1-\beta = \Pr(|\frac{\hat{\mu}_1-\hat{\mu}_2}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2}}| > z_{\frac{\alpha}{2}} \mid H_1) = 0.4404$

$\Rightarrow$ type II error rate $\beta = 0.559$

(4). From (3), we have that $\frac{(\hat{\mu}_1-\hat{\mu}_2)-2.5}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2}} \sim N(0,1)$ under $H_1$. $\Rightarrow \frac{(\hat{\mu}_1-\hat{\mu}_2)}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2}} \overset{d}{=} Z \sim N(\frac{1}{5\sqrt{\frac{n}{n_1 n_2}}}, 1)$

We hope to find the minimum $n=n_1+n_2$ s.t. $\Pr(Z > z_\alpha \mid H_1) \geq 0.8$

$\Rightarrow \Phi(z_\alpha - \frac{1}{5\sqrt{\frac{n}{n_1 n_2}}}) \leq \frac{1}{5} \Rightarrow \sqrt{\frac{n_1 n_2}{n}} \geq 2(z_\alpha - z_{0.8})$

Note that $n_1 n_2 \leq (\frac{n_1+n_2}{2})^2 = \frac{n^2}{4} \Rightarrow n \geq 16(z_\alpha - z_{0.8})^2 \approx 99$

# Problem 4

Consider the difference of two scores $(A-B) \stackrel{d}{=} X$

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 4 | 3 | 4 | 6 | 3 | -1 | -3 | -2 | 0 | -1 | 1 | 2 |
| $\lvert X_i - 0 \rvert$ | 4 | 3 | 4 | 6 | 3 | 1 | 3 | 2 | 0 | 1 | 1 | 2 |
| $R_i$ | 9.5 | 7 | 9.5 | 11 | 7 | 2 | 7 | 4.5 | x | 2 | 2 | 4.5 |
| Signed $R_i$ | 9.5 | 7 | 9.5 | 11 | 7 | -2 | -7 | -4.5 | x | -2 | 2 | 4.5 |

(Since there is 0, we delete this point)

Here. we have $W_{obs} = 9.5+7+9.5+11+7+2+4.5 = 50.5$

Since the alternative hypothesis $H_1: m_0 \neq 0$, and $\alpha = 0.01$, $n = 11$

$\Rightarrow$ from the table. RR is $\{W_{obs} \leq 5\}$ or $\{W_{obs} \geq 61\}$

Since $W_{obs} = 50.5 \notin RR$. $\Rightarrow$ We can't reject the null hypothesis $H_0: m_d = 0$.

# Problem5

## 5(1)

Code:

```
DATA DATA1;
INPUT Convict $ Victim $ Death $ Number;
DATALINES;
White White Yes 19
White White No 132
White Black Yes 0
White Black No 9
Black White Yes 11
Black White No 52
Black Black Yes 6
Black Black No 97
;
RUN;
```

Output:

列

- 全选
- Convict
- Victim
- Death
- Number

总行数: 8 总列数: 4 行 1-8

| | Convict | Victim | Death | Number |
|---|---|---|---|---|
| 1 | White | White | Yes | 19 |
| 2 | White | White | No | 132 |
| 3 | White | Black | Yes | 0 |
| 4 | White | Black | No | 9 |
| 5 | Black | White | Yes | 11 |
| 6 | Black | White | No | 52 |
| 7 | Black | Black | Yes | 6 |
| 8 | Black | Black | No | 97 |

## 5(2)

Code:

```
PROC FREQ DATA=DATA1;
  TABLES Death / Binomial(Level=2 Wald Exact);
  EXACT Binomial;
  WEIGHT Number;
RUN;
```

- Then we have the estimation of the proportion is 0.1104
- 95% Wald CI: $(0.0764, 0.1445)$
- 95% Exact CI: $(0.0786, 0.1496)$

**FREQ 过程**

| Death | 频数 | 百分比 | 累积频数 | 累积百分比 |
|---|---|---|---|---|
| No | 290 | 88.96 | 290 | 88.96 |
| Yes | 36 | 11.04 | 326 | 100.00 |

| 二项式比例 | |
|---|---|
| Death = Yes | |
| 比例 (P) | 0.1104 |
| ASE | 0.0174 |

| 二项式比例的置信限 | | |
|---|---|---|
| 比例 = 0.1104 | | |
| 类型 | 95% 置信限 | |
| Clopper-Pearson（精确） | 0.0786 | 0.1496 |
| Wald | 0.0764 | 0.1445 |

| H0: 比例 = 0.5 的检验 | |
|---|---|
| H0 下的 ASE | 0.0277 |
| Z | -14.0678 |
| 单侧 Pr < Z | <.0001 |
| 双侧 Pr > \|Z\| | <.0001 |
| | |
| 精确检验 | |
| 单侧 Pr <= P | <.0001 |
| 双侧 = 2 * 单侧 | <.0001 |

**样本大小 = 326**

# 5(3)

- Null hypothesis $H_0$: proportion $= 0.09$
- Alternative hypothesis $H_1$: proportion $> 0.09$
- Value of the test statistic
- P-values:

Code:

```
PROC FREQ DATA=DATA1;
  TABLE Death / Binomial(Level=2 P=0.09 Wald Wilson Exact) Alpha=0.1;
  WEIGHT Number;
  EXACT Binomial;
RUN;
```

- Null hypothesis $H_0$: proportion $= 0.09$ v.s. Alternative hypothesis $H_1$: proportion $> 0.09$, $\alpha = 0.1$
- Value of the test statistic: 1.2889
- P-values:

- Z-test: 0.0987 (One Side)
- The exact version: 0.1184 (One Side)

- Conclusion:

  If we consider the Z-test p-value, we can reject the null hypothesis $H_0$, i.e., the proportion of homicide convicts who received death penalty, irrespective of the races of the convict and victim is 0.09 at the significance level $\alpha = 0.1$.

  If we consider the exact version p-value, we can not reject the null hypothesis $H_0$, i.e., the proportion of homicide convicts who received death penalty, irrespective of the races of the convict and victim is 0.09 at the significance level $\alpha = 0.1$.

  Thus, we should not end with the calculation of *p*-value and try other approaches which are appropriate and feasible.

**FREQ 过程**

| Death | 频数 | 百分比 | 累积频数 | 累积百分比 |
|---|---|---|---|---|
| No | 290 | 88.96 | 290 | 88.96 |
| Yes | 36 | 11.04 | 326 | 100.00 |

| 二项式比例 | |
|---|---|
| Death = Yes | |
| 比例 (P) | 0.1104 |
| ASE | 0.0174 |

| 二项式比例的置信限 | | |
|---|---|---|
| 比例 = 0.1104 | | |
| 类型 | 90% 置信限 | |
| Clopper-Pearson（精确） | 0.0831 | 0.1432 |
| Wald | 0.0819 | 0.1390 |
| Wilson | 0.0850 | 0.1423 |

| H0: 比例 = 0.09 的检验 | |
|---|---|
| H0 下的 ASE | 0.0159 |
| Z | 1.2889 |
| 单侧 Pr > Z | 0.0987 |
| 双侧 Pr > \|Z\| | 0.1974 |
| | |
| 精确检验 | |
| 单侧 Pr >= P | 0.1184 |
| 双侧 = 2 * 单侧 | 0.2368 |

**样本大小 = 326**

## 5(4)

```
* Pearson's chi-squared test;
PROC FREQ DATA=DATA1;
  TABLES Convict * Death / RISKDIFF(EQUAL VAR = NULL CL = Wald NORISKS);
  WEIGHT Number;
RUN;
```

- Null hypothesis $H_0$: the proportion of White convicts who received death penalty is equal to that of Black convicts
- Alternative hypothesis $H_1$: the proportion of White convicts who received death penalty is different from that of Black convicts
- We use Pearson's chi-squared test
- Value of the test statistic: 0.4706
- P-values: 0.6379 (two sides)

**FREQ 过程**

| 频数<br>百分比<br>行百分比<br>列百分比 |
|---|

| Convict-Death表 | | | |
|---|---|---|---|
| | Death | | |
| Convict | No | Yes | 合计 |
| Black | 149<br>45.71<br>89.76<br>51.38 | 17<br>5.21<br>10.24<br>47.22 | 166<br>50.92 |
| White | 141<br>43.25<br>88.13<br>48.62 | 19<br>5.83<br>11.88<br>52.78 | 160<br>49.08 |
| 合计 | 290<br>88.96 | 36<br>11.04 | 326<br>100.00 |

**表“Death-Convict”的统计量**

| 风险差值置信限 | | |
|---|---|---|
| 风险差值 = 0.0163 | | |
| 类型 | 95% 置信限 | |
| Wald | -0.0518 | 0.0845 |
| 列 1 (Death = No) | | |

| 风险差值检验 | |
|---|---|
| H0: P1 - P2 = 0 | Wald 方法 |
| 风险差值 | 0.0163 |
| ASE (H0) | 0.0347 |
| Z | 0.4706 |
| 单侧 Pr > Z | 0.3190 |
| 双侧 Pr > lZl | 0.6379 |
| 列 1 (Death = No) | |

```
* Barnard exact test;
PROC FREQ DATA=DATA1;
  TABLES Convict*Death / CHISQ;
  EXACT BARNARD;
  WEIGHT Number;
RUN;
```

- Null hypothesis $H_0$: the proportion of White convicts who received death penalty is equal to that of Black convicts
- Alternative hypothesis $H_1$: the proportion of White convicts who received death penalty is different from that of Black convicts
- We use Barnard exact test
- Value of the test statistic: 0.4706
- P-values: 0.7115 (two sides)

表“Death-Convict”的统计量

| 统计量 | 自由度 | 值 | 概率 |
|---|---|---|---|
| 卡方 | 1 | 0.2214 | 0.6379 |
| 似然比卡方检验 | 1 | 0.2215 | 0.6379 |
| 连续调整卡方 | 1 | 0.0863 | 0.7689 |
| Mantel-Haenszel 卡方 | 1 | 0.2208 | 0.6385 |
| Phi 系数 | | 0.0261 | |
| 列联系数 | | 0.0261 | |
| Cramer V | | 0.0261 | |

| Fisher 精确检验 | |
|---|---|
| 单元格 (1,1) 频数 (F) | 149 |
| 左侧 Pr <= F | 0.7412 |
| 右侧 Pr >= F | 0.3843 |
| | |
| 表概率 (P) | 0.1255 |
| 双侧 Pr <= P | 0.7246 |

| Barnard 精确检验<br>风险差值 | |
|---|---|
| H0: P1 - P2 = 0 | |
| 风险差值 | 0.0163 |
| ASE (H0) | 0.0347 |
| Z | 0.4706 |
| 单侧 Pr >= Z | 0.3631 |
| 双侧 Pr >= \|Z\| | 0.7115 |

样本大小 = 326

**Conclusion**:

From both methods, we can not reject $H_0$, i.e., the proportion of White convicts who received death penalty is equal to that of Black convicts at the significance level $\alpha = 0.05$.

Thus, we can think that the proportion of White convicts who received death penalty is equal to that of Black convicts.

# 5(5)

```
PROC FREQ DATA=DATA1;
  TABLES Victim*Death / CHISQ;
  EXACT BARNARD;
  WEIGHT Number;
RUN;
```

- Null hypothesis $H_0$: Irrespective of the convict's race, does it appear that the proportion of death penalty is
  equal to the situation when the victim's race is White.
- Alternative hypothesis $H_1$: Irrespective of the convict's race, does it appear that the proportion of death penalty is
  higher when the victim's race is White.
- We use Barnard exact test
- Value of the test statistic: 2.3696
- P-values: 0.0147 (One sides)
- Conslusion:

  At the significance level $\alpha = 0.05$, we will reject the null hypothesis $H_0$, i.e., Irrespective of the convict's race, does it appear that the proportion of death penalty is equal to the situation when the victim's race is White.

  Thus, we can think that the proportion of death penalty is higher when the victim's race is White.

表“Death-Victim”的统计量

| 统计量 | 自由度 | 值 | 概率 |
|---|---|---|---|
| 卡方 | 1 | 5.6149 | 0.0178 |
| 似然比卡方检验 | 1 | 6.2497 | 0.0124 |
| 连续调整卡方 | 1 | 4.7678 | 0.0290 |
| Mantel-Haenszel 卡方 | 1 | 5.5977 | 0.0180 |
| Phi 系数 | | 0.1312 | |
| 列联系数 | | 0.1301 | |
| Cramer V | | 0.1312 | |

| Fisher 精确检验 | |
|---|---|
| 单元格 (1,1) 频数 (F) | 106 |
| 左侧 Pr <= F | 0.9964 |
| 右侧 Pr >= F | 0.0116 |
| | |
| 表概率 (P) | 0.0081 |
| 双侧 Pr <= P | 0.0242 |

| Barnard 精确检验 风险差值 | |
|---|---|
| H0: P1 - P2 = 0 | |
| 风险差值 | 0.0866 |
| ASE (H0) | 0.0366 |
| Z | 2.3696 |
| 单侧 Pr >= Z | 0.0147 |
| 双侧 Pr >= \|Z\| | 0.0185 |

样本大小 = 326

# 5(6)

According to the 5(e), we have that proportion of White convicts who received death penalty is equal to that of Black convicts.

But according to the 5(f), we have that the proportion of death penalty is higher when the victim's race is White, which means the punishment will be more serious if the victim's race is White, it maybe invisibly reflects that the value of white lives is more expensive.

As far as I am concerned, people of all races should enjoy the same rights, we can not discriminate against or look down upon someone simply because of their race. We should respect everyone and value their soul rather than color and race.