ID: 0856733
Name: Ming-Han Huang

109-1 Statistics
National Chiao Tung university

Homework #3

Due by Sunday 11/22 23:55

Question #1.A

Sudden infant death syndrome (SIDS) is the sudden and unexplained death of a baby younger than 1 year old. A diagnosis of SIDS is made if the baby's death remains unexplained even after a death scene investigation, an autopsy, and a review of the clinical history. The low birth weight children are at risk, and these question compare the distribution between boys and girls.

Question #1.B

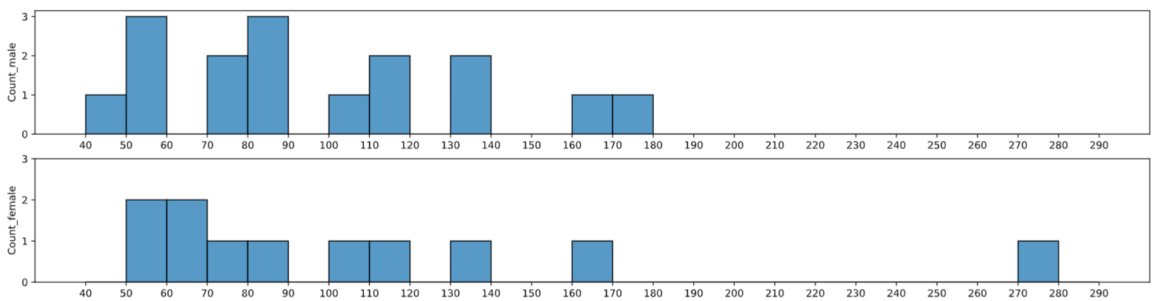


Figure 1: Histogram

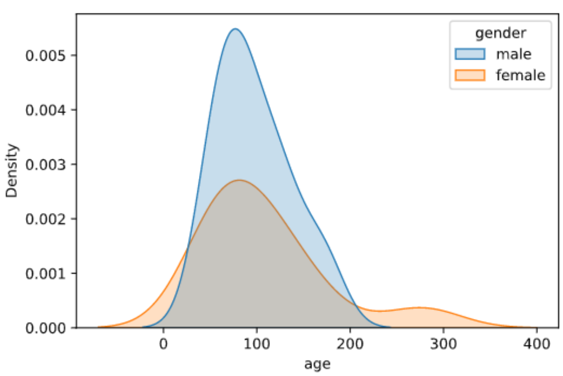


Figure 2: KDE Plot

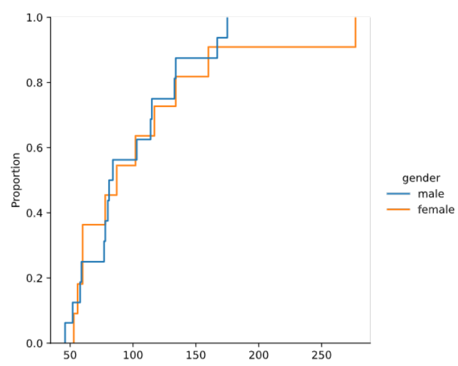


Figure 3: Empirical distribution

成對樣本T檢定的case通常有兩種，一種是「重複量數」，也就是前後測的問題，像是減肥前後、實驗前後的成績，因此每一筆配對的資料都是來自同一位受試者，這種是最常見的配對樣本。但這筆資料不是同一個人 ，意思是不是從男生變女生。
另一種為「配對組法」，雖然每一筆配對的資料是來自兩位的受試者，但是我們會認定他們的某一特質（研究者所關心的）是相同的。意思是，想擁某一些手法，讓那配對其他條件都相同下，只有男女這一個差別，那這樣才適合做T 檢定。
也就是說，不能直接拿男女的存活時間直接做T檢定，即便他們是抽樣出來的。

Question #1.C

The result from scipy.stats.ttest_ind is:

statistic = 0.5118640907091607
pvalue = 0.6132382438368118

Since we have pvalue= $0.61324 > 0.05$, that is, we cannot reject the null hypothesis at confidence interval with level 95%. Therefore, the two distribution are the same.

Question #1.D

The result from scipy.stats.ranksums is:

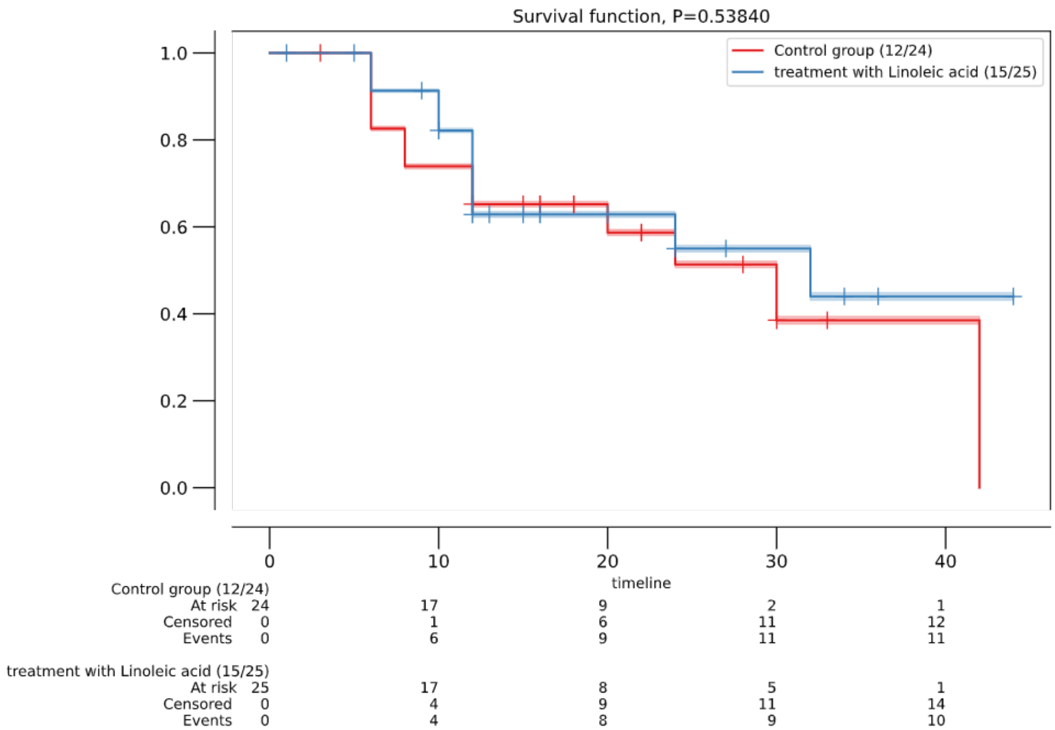
statistic = 0.148039131365948
pvalue = 0.8823118861000214

Since we have pvalue= $0.8823 > 0.05$, that is, we cannot reject the null hypothesis at confidence interval with level 95%. Therefore, the two distribution are the same.
Also, we get a higher p-value than t-test make sense, since Wilcoxon rank-sum test is more appropriate than t-test for the question.

Question #1.E

To answer the question, we pool the two samples, order the observations from the smallest to the largest, retaining their group identity, and rank them from 1 to 27.
Since the sample of female (m=11) is smaller than that of mal (n=16), the test static W_m is the sum of the ranks associated with the female.
Therefore, $W_m = 3 + 4 + 7.5 + ... + 27 = 157$
And, $E(W) = \frac{11(11+16+1)}{2} = 154$,
$Var(W) = \sqrt{\frac{11*16(11+16+1)}{12}} \rightarrow SD = 20.26$
$Z = \frac{157-154}{20.26} = 0.148$
The p-value with z-score=0.148 is .882343 under two-tail hypothesis. The result is very closed to (d), and also we cannot reject $H_0 : F_1 = F_2$.

Question #2.A



The horizontal axis (x-axis) represents time in days, and the vertical axis (y-axis) shows the probability of surviving or the proportion of people surviving. The lines represent survival curves of the two groups. A vertical drop in the curves indicates an event. The vertical tick mark on the curves means that a patient was censored at this time.

- At time zero, the survival probability is 1.0 (or 100% of the participants are alive).
- The median survival is approximately 30 days for Control group and 32 days for treatment with Linoleic acid, suggesting a good survival for treatment with Linoleic acid compared to Control group

Question #2.B

The result from lifelines.statistics.logrank_test is:

statistic = 0.378521
pvalue = 0.538396

The result from gehan.test is:

statistic = 0.3345791
pvalue = 0.5629751

The log rank test for difference in survival gives a p-value of p = 0.538396 and the Gehan test gives p = 0.5629751, indicating that there is no evidence that treatment with Linoleic acid have better effect in survival at confidence interval with level 95%. Also, the fact that we get a larger p-value from Gehan test compared to the log-rank test makes sense

Question #3.A

$$l(\lambda) = logL(\lambda) = \sum_{i=1}^{n} \delta_i(log\lambda - \lambda x_i) + (1-\delta_i)(-\lambda x_i) \tag{1}$$

To find MLE of λ, let$\frac{\partial l(\lambda)}{\partial \lambda} = 0$, that is:

$$\frac{\partial l(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} \frac{\delta_i}{\lambda} - x_i = 0 \tag{2}$$

$$\lambda_{MLE} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i} \tag{3}$$

Question #3.C

$$\frac{\partial^2 l(\lambda)}{\partial \lambda^2} = \frac{\partial}{\partial \lambda} \sum_{i=1}^{n} \frac{\delta_i}{\lambda} - x_i = -\frac{\sum_{i=1}^{n} \delta_i}{\lambda^2} \tag{4}$$

$$Var(\hat{\lambda}) = -(\frac{\sum_{i=1}^{n} \delta_i}{\lambda^2})^{-1} = \frac{\sum_{i=1}^{n} \delta_i}{(\sum_{i=1}^{n} x_i)^2} \tag{5}$$

Question #4.A

$$\begin{aligned} P(Y=y) &= \tbinom{y+r-1}{y} p^y (1-p)^r \\ &= exp(yln(p) + rln(1-p) + ln(\tbinom{y+r-1}{y})) \end{aligned} \tag{6}$$

Thus the canonical parameter $\theta = ln(p)$, and the canonical link function $g(\mu) = ln(\frac{\mu/r}{1+\mu/r})$

Question #4.B

Continue from A, we have canonical parameter $\theta = ln(p)$, that is, $p = e^\theta$

$$P(Y=y) = exp(\theta y + rln(1-e^\theta) + ln(\tbinom{y+r-1}{y})) \tag{7}$$

That is, we have $b(\theta) = -rln(1-e^\theta)$ and $\phi = 1$.

$$E[y] = \frac{\partial b(\theta)}{\partial \theta} = r\frac{e^\theta}{1-e^\theta} = r\frac{p}{1-p} = \frac{rp}{1-p} \tag{8}$$

$$Var[y] = \phi\frac{\partial^2 b(\theta)}{\partial \theta^2} = \frac{\partial}{\partial \theta} r\frac{e^\theta}{1-e^\theta} = \frac{re^\theta}{(1-e^\theta)^2} = \frac{rp}{(1-p)^2} \tag{9}$$