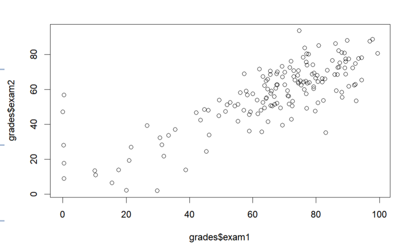
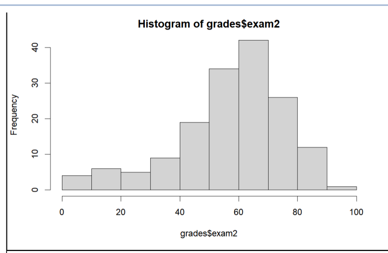
Yuzhou Peng 121090446

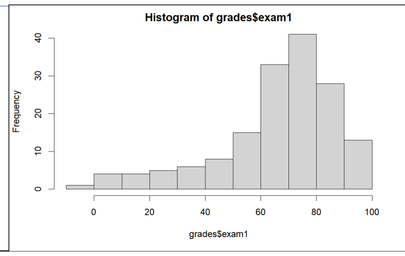
Problem 1

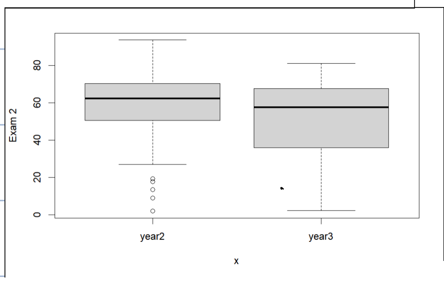
The data **grades.csv** has 158 rows and three columns: **Year**, the year of the student, **exam1**, the score of the first exam jittered with some noise, and **exam2**, the score of the second exam jittered with some noise. With the data, answer the following questions. You can use R/Python/other software to help you answer the questions.

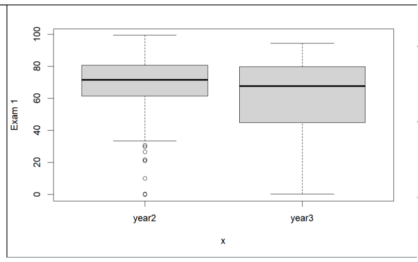
(1) (5 points) Exploratory data analysis: make histograms, scatter plots and other plots you find helpful in exploring the dataset.









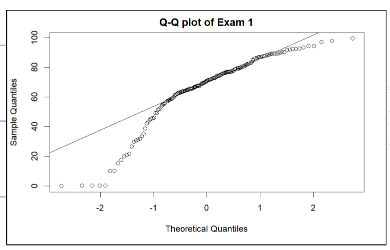


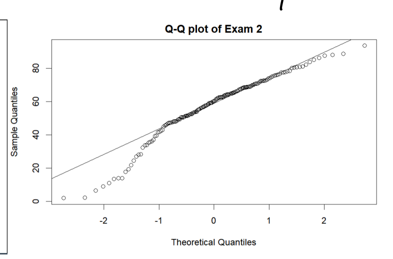
* From scatter plot. there is potential pattern in exam 1 and 2 students with higher exam 1 score tend to have higher exam 2 score
* From histograms, the exam scores skews to the right
* From box plots. Year 3 students have higher variance in both exam 1 and 2.

(5 points) Summarize what you observe and comment on the assumption that the data is iid Gaussian.

① It's reasonable to assume student's scores are independent with each other

② But from Q-Q plot the distribution is not Gaussian





(2) (10 points) Formulate a hypothesis testing problem to evaluate the statement that `exam1` and `exam2` have the same median. You can use the signed Wilcoxon rank sum test. What is the p-value associated with your null hypothesis?

H_0 : exam 1 and exam2 have the same median

$\Leftrightarrow$ location shift is 0 $\Rightarrow$ exam1 $\sim F(t)$, exam2 $\sim G(t)$

```
## (2)
```{r}
wilcox.test(grades$exam1, grades$exam2)
```

	Wilcoxon rank sum test with continuity correction

data: grades$exam1 and grades$exam2
W = 16559, p-value = 5.171e-07
alternative hypothesis: true location shift is not equal to 0
```

$F(t)$ and $G(t)$ have the same median.

$\hat{p} = 5.17 \times 10^{-7}$.

reject the null hypothesis

(3) (15 points) For `exam2`, examine the difference between the group `year2` and the group `year3`. First perform the Kolmogorov-Smirnov test to see if there is any difference. If differences are detected, explore the differences in location, dispersion, and both.

```
## (3)
```{r}
exam2_grades_y3 <- grades$exam2[seq(1,39)]
exam2_grades_y2 <- grades$exam2[seq(40,158)]

#K-S test
ks.test ( exam2_grades_y2 , exam2_grades_y3)

```

	Exact two-sample Kolmogorov-Smirnov test

data: exam2_grades_y2 and exam2_grades_y3
D = 0.22366, p-value = 0.08792
alternative hypothesis: two-sided
```

H_0: exam1 $\sim F(t)$, exam2 $\sim G(t)$, $F = G$.

$\hat{p} = 0.088 > 0.05$.

fail to reject the null.

Reject that the distributions are different.

(4) (10 points) Test the independence between `exam1` and `exam2` with Kendall's τ and Spearman's ρ, respectively.

```
## (4)
```{r}
cor.test(grades$exam1, grades$exam2, method = "kendall")

cor.test(grades$exam1, grades$exam2, method = "spearman")
```

	Kendall's rank correlation tau

data: grades$exam1 and grades$exam2
z = 10.373, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.5562364

	Spearman's rank correlation rho

data: grades$exam1 and grades$exam2
S = 171106, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7397069
```

Both tests have $\hat{p} < 2.2 \times 10^{-16} \ll 0.05$

Reject null for both Kendall's and Spearman's test.

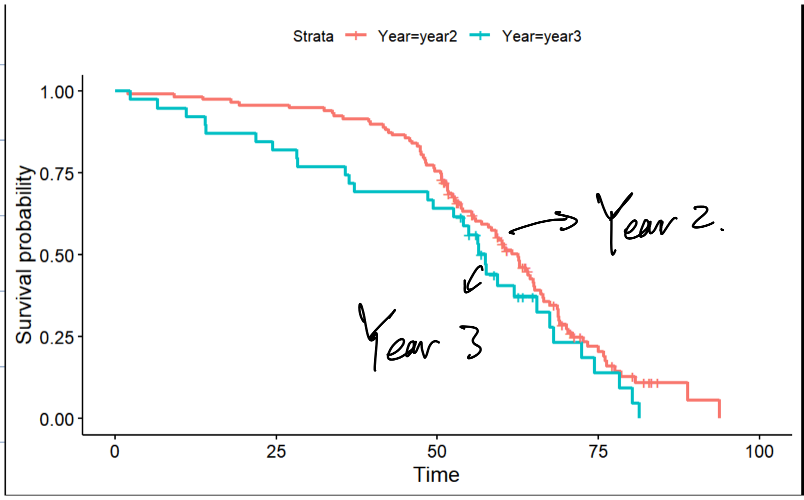
exam 1 and 2 are not independent

Problem 2

The data `grades-censor.csv` has 158 rows and five columns: `Year`, the year of the student, `exam1obs`, the observed score of the `exam1`, `exam2obs`, the observed score of the `exam2`, `delta1`, the indicator if `exam1obs` is uncensored (1 if uncensored, 0 if right-censored), and `delta2`, the indicator if `exam2obs` is uncensored (1 if uncensored, 0 if right-censored). The right censoring is defined as $\min(X_i, C_i)$, where X_i is the score and C_i is the random censoring variable independent of X_i.

With the data, answer the following questions. You can use R/Python/other software to help you answer the questions.

(1) (10 points) For `exam2`, plot the survival functions of `year2` and `year3` in the same figure.



(2) (15 points) Test the hypothesis that `year3` is the same as `year2` in `exam2` with (1) the Kolmogorov-Smirnov test, ignoring the censoring, and (2) the logrank test, taking the censoring into account. Compare the results and comment on/explain the differences.

````
## (2)
```{r}
exam2_censor_y3 <- grades_censor$exam2obs[seq(1,39)]
exam2_censor_y2 <- grades_censor$exam2obs[seq(40, 158)]
#K-S test
ks.test ( exam2_censor_y3 , exam2_censor_y2)

#log-rank test
surv_obj <- Surv(time = grades_censor$exam2obs, event = grades_censor$delta2)
survdiff(surv_obj ~ grades_censor$Year)
```


        Exact two-sample Kolmogorov-Smirnov test

 data:  exam2_censor_y3 and exam2_censor_y2
 D = 0.22366, p-value = 0.08792
 alternative hypothesis: two-sided

 Call:
 survdiff(formula = surv_obj ~ grades_censor$Year)

                             N Observed Expected (O-E)^2/E (O-E)^2/V
 grades_censor$Year=year2 119       89     95.6     0.451      2.24
 grades_censor$Year=year3  39       31     24.4     1.765      2.24

  Chisq= 2.2  on 1 degrees of freedom, p= 0.1
````

(1) K-S test

$\hat{p} = 0.088$. fail to reject

(2) log-rank test.

$\hat{p} = 0.1$. fail to reject

Both tests fail to reject the null. but log-rank test have larger $\hat{p}$-value

K-S test tends to reject the null compared with log-rank test.

(3) (10 points) With the Cox proportional hazards model, study the score difference between **year2** and **year3**, controlling for different exams. To earn full credits, you need to provide both the point estimate and the confidence interval.

```
cox_model <- coxph(Surv(grades, delta) ~ Year + exam, data = grades_cox_combined)
summary(cox_model)
```

```
Call:
coxph(formula = Surv(grades, delta) ~ Year + exam, data = grades_cox_combined)

  n= 316, number of events= 220

       coef exp(coef) se(coef)     z Pr(>|z|)
Year 0.2779    1.3204   0.1521 1.827 0.067660 .
exam 0.5233    1.6875   0.1377 3.801 0.000144 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
Year     1.320     0.7573     0.980     1.779
exam     1.688     0.5926     1.288     2.210
```

Assume the model:

$$r(t) = r_0(t) \exp\{\beta_1 \mathbb{1}\{\text{exam } 2\} + \beta_2 \mathbb{1}\{\text{Year } 3\}\}$$

$\hat{\beta}_2 = 0.278$ with a confidence interval $[\log(0.980), \log(1.779)]$

Interpretation: Controlling for exam, Year 3 students score hazard rate is on averge $e^{0.278} \simeq 1.320$ higher than that of Year 2 students score

(4) (10 points, extra credits) Provide point estimates of the censored data points. Your score of this problem is based on the mean squared error (MSE).

Given data $(X_i^{(1)}\ X_i^{(2)})$ be the exam 1, 2 for i-th student in Year 2;
observed

$(Y_j^{(1)}.\ Y_j^{(2)})$ be the exam 1, 2 for j-th student in Year 3
observed.

The estimation follows 2 cases.

① One of $(X_i^{(1)}\ X_i^{(2)})$ is uncensored, say, $X_i^{(2)}$ is not censored. obtain the rank of $X_i^{(2)}$ in $(X_1^{(2)} \cdots X_n^{(2)})$ that are not censored. denoted as $R_i^{(2)}$. Search in $(X_1^{(1)} \cdots X_m^{(1)})$ that are uncensored, find $X_{i_0}^{(1)}$ with rank $= R_i^{(2)}$, let $X_{i_0}^{(1)}$ be the estimate for $X_i^{(1)}$

② Both $(X_i^{(1)}.\ X_i^{(2)})$ are censored.
obtain ranks of observations in the uncensored data, respectively $(R_i^{(1)}.\ R_i^{(2)})$, Let $R = \min(R_i^{(1)}.\ R_i^{(2)})$
for simplicity consider $R = R_i^{(2)}$ then an estimator for $X_i^{(2)}$ is $\widehat{X_i^{(2)}} = \text{averge}(X_1^{(2)} \cdots X_l^{(2)})$ in which each element has rank higher than R, then egard $\hat{X}_i^{(2)}$ as the uncensored data with and repeat procedure in ① to obtain estimator for $X_i^{(1)}$

Do the same for $(Y_j^{(1)}.\ Y_j^{(2)})$.

Problem 3

Suppose $X_1, \dots, X_n \overset{iid}{\sim} F(\cdot)$. The empirical CDF is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\{X_i \le x\})$$

where $\mathbf{1}(\{X_i \le x\})$ is the indicator function.

(1) (10 points) For some target location x, derive the mean squared error (MSE) of $F_n(x)$ as an estimator for $F(x)$.

(1) $MSE = \mathbb{E}(\hat{F}_n(x) - F(x))^2 = \mathbb{E}(\hat{F}_n^2(x)) + \mathbb{E}(F^2(x))$

$- 2\mathbb{E}(F(x)\hat{F}_n(x))$

$= \mathbb{E}(\hat{F}_n^2(x)) + F^2(x) - 2F(x) \cdot \mathbb{E}(\hat{F}_n(x))$

$\mathbb{E}\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\{X_i \le x\} = \frac{1}{n} \cdot \sum_{i=1}^{n} F(x) = F(x)$

$\mathbb{E}\hat{F}_n^2(x) = \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n} \mathbb{1}\{X_i \le x\}\right)^2$

$(n^2-n)F^2(x)$

$= \frac{1}{n^2}\mathbb{E}\left(\underline{\sum_{i=1}^{n} \mathbb{1}\{X_i \le x\}}\right) + 2\mathbb{E}\left(\sum_{i<j} \underline{\mathbb{1}\{X_i \le x, X_j \le x\}}\right)$

$= \frac{1}{n^2}(nF(x) + (n^2-n)F^2(x))$

$= \frac{1}{n}F(x) + (1-\frac{1}{n})F^2(x)$

$\Rightarrow MSE = \frac{1}{n}F(x) + (1-\frac{1}{n})F^2(x) + F^2(x) - 2F^2(x)$

$= \frac{1}{n}(F(x) - F^2(x))$

(2) (10 points) Suppose $x \neq y$ are two distinct points, find $\mathrm{Cov}(F_n(x), F_n(y))$.

(2) $\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y))$

$= \mathbb{E}(\hat{F}_n(x)\hat{F}_n(y)) - \mathbb{E}(\hat{F}_n(x))\,\mathbb{E}(\hat{F}_n(y))$.

$\mathbb{E}(\hat{F}_n(x)\cdot\hat{F}_n(y)) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\le x\}}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\le y\}}\right)$.

$= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{\{X_i\le x\}}\cdot\sum_{i=1}^{n}\mathbb{1}_{\{X_i\le y\}}\right)$.

$= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{\{X_i\le x,\, X_i\le y\}} + 2\sum_{i<j}\mathbb{1}_{\{X_i\le x,\, X_j\le y\}}\right)$

$= \frac{1}{n^2}\left(n\cdot F(\min\{x,y\}) + (n^2-n)\cdot F(x)F(y)\right)$

$= \frac{1}{n}F(\min\{x,y\}) + \left(1-\frac{1}{n}\right)F(x)F(y)$

(3) (10 points, extra credits) Find a 95% confidence interval of $F(x)$ for a given location x. To earn full credits, you need to justify your answer.

$\mathbb{E}\,\mathbb{1}_{\{X_i\le x\}} = F(x)$

$\mathrm{Var}(\mathbb{1}_{\{X_i\le x\}}) = F(x) - F^2(x)$

By Central Limit Theorem

$$\frac{\left(\sum_{i=1}^{n}\mathbb{1}_{\{X_i\le x\}} - nF(x)\right)}{\sqrt{n}\cdot\sqrt{F(x)-F^2(x)}} \xrightarrow{d} N(0,1).$$

Since $\sup\|\hat{F}_n(x) - F(x)\| \xrightarrow{a.s.} 0$.

$\hat{F}_n(x) \xrightarrow{P} F(x)$

By continuous mapping theorem

$$\sqrt{\hat{F}_n(x) - \hat{F}_n^2(x)} \xrightarrow{P} \sqrt{F(x) - F^2(x)}$$

By Slutsky's Theorem.

$$\frac{\sqrt{n}\,(\hat{F}_n(x) - F(x))}{\sqrt{\hat{F}_n(x) - \hat{F}_n^2(x)}} \xrightarrow{d} N(0,1).$$

given 95% confidence level.

a $CI = \left[\hat{F}_n(x) - \sqrt{\frac{\hat{F}_n(x) - \hat{F}_n^2(x)}{n}}\,\Phi(0.975),\ \hat{F}_n(x) + \sqrt{\frac{\hat{F}_n(x) - \hat{F}_n^2(x)}{n}}\,\Phi(0.975)\right].$