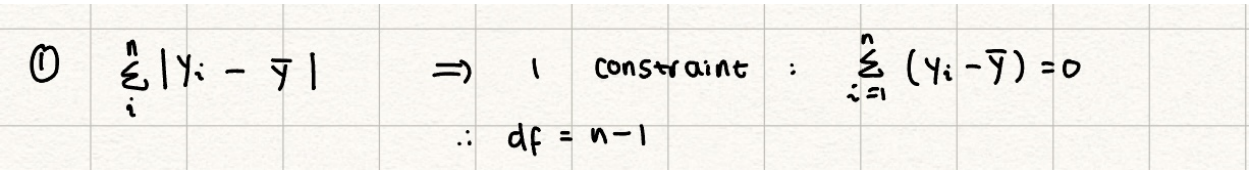
stat421_hw3

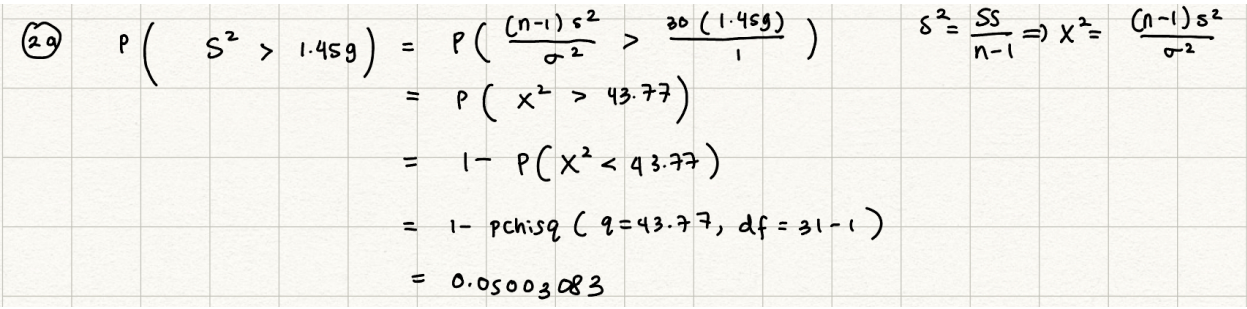
2022-10-14

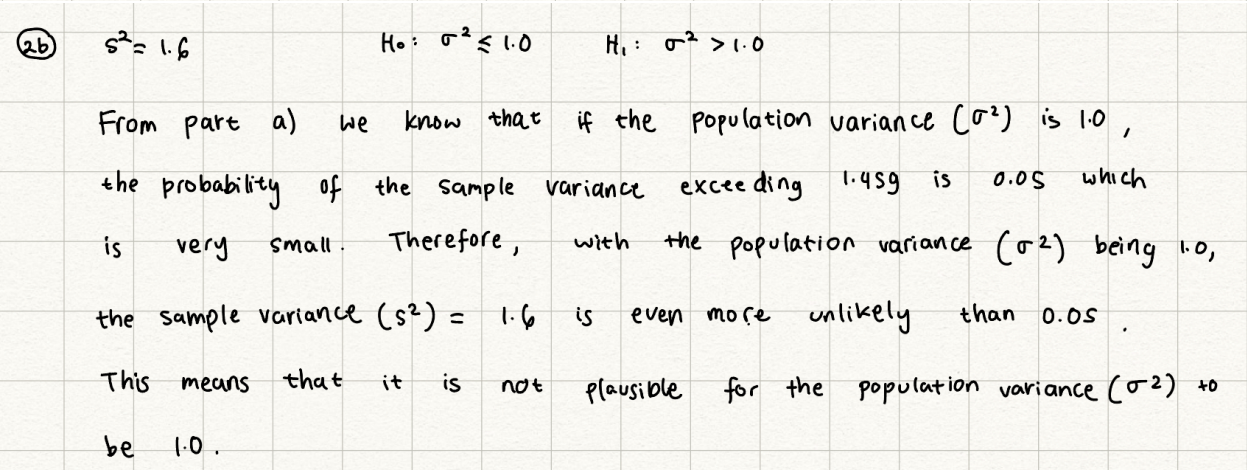
Lecture 6




```
1-pchisq(q=43.77,df=30)
```

```
## [1] 0.05003083
```






```
1-pf(q=2.33,20,24)
```

```
## [1] 0.02484508
```

③ $V[Y|X=1] = V[Y|X=2]$. $H_0: \sigma_1^2 = \sigma_2^2$ $n_1 = 21$ $n_2 = 25$

$$P(S_1^2 > 2.33 S_2^2) = P\left(\frac{S_1^2}{S_2^2} > 2.33\right)$$

$$= P\left(\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} > 2.33 \frac{\sigma_1^2}{\sigma_2^2}\right)$$

$$= P(F > 2.33)$$

$$= 1 - pf(q = 2.33, 20, 24)$$

$$= 0.024508$$

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,\, n_2-1}$$

```
(qt(0.05,df=20)*(sqrt(2/21)))+3
```

```
## [1] 2.46774
```

4. $H_0: \mu = 3$ $H_1: \mu < 3$ $\alpha = 0.05$ Population mean = μ

$s^2 = 2 \Rightarrow n = 21 \Rightarrow df = 21 - 1 = 20$

(a) We assume that $H_0: \mu = 3$ is true.

$$\frac{\bar{y}_{obs} - \mu_0}{s_{obs}/\sqrt{n}} < t_\alpha$$

$$\bar{y}_{obs} < t_\alpha \cdot \frac{s_{obs}}{\sqrt{n}} + \mu_0$$

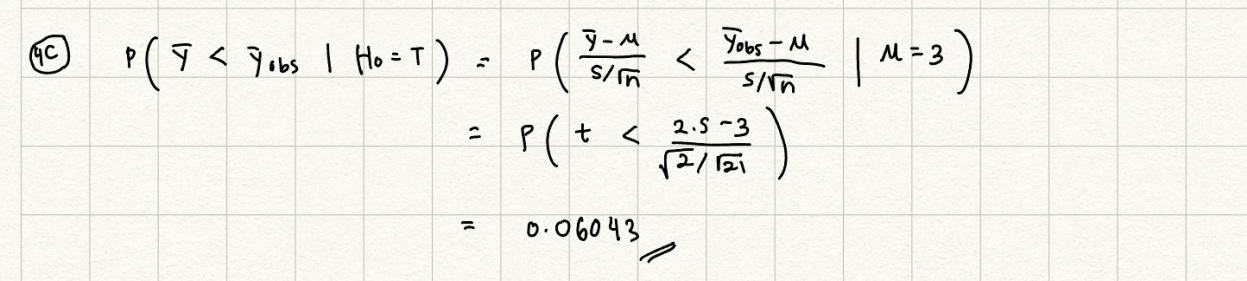
$$\bar{y}_{obs} < t_{0.05} \cdot \frac{\sqrt{2}}{\sqrt{21}} + 3$$

$$\bar{y}_{obs} < -1.724718 \cdot \sqrt{\frac{2}{21}} + 3$$

$$\bar{y}_{obs} < 2.46774 \Rightarrow RR.$$

$t_{0.05} = qt(0.05, df = 20) = -1.724718$

(b) Since the rejection region is $\bar{y}_{obs} < 2.46774$ and the observed sample mean is 2.5, it is not in the rejection region. Therefore, we cannot reject H_0.



```
pt(q=(2.5-3)/(sqrt(2)/sqrt(21)),df=20)
```

```
## [1] 0.06042675
```

4d) We know that the p-value is 0.06042675 from part c, and that $\alpha = 0.05$. Since p-value > 0.05, we cannot reject the H0.

Lecture 7

Lecture 7

① $t = \dfrac{(\bar{y}-\mu)/(\sigma/\sqrt{n})}{\sqrt{x^2/(n-1)}} \sim t_{n-1}$ $\quad X^2 \equiv \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ $\quad T \equiv \dfrac{\frac{\bar{y}-\mu}{\sigma/\sqrt{n}} + \Delta}{\sqrt{x^2/(n-1)}}$

$$-T_{\frac{\alpha}{2},n-1,\Delta} < \frac{\frac{\bar{y}-\mu}{\sigma/\sqrt{n}} + \Delta}{\sqrt{x^2/(n-1)}} < T_{\frac{\alpha}{2},n-1,\Delta}$$

$\frac{(n-1)S^2}{\sigma^2} \cdot \frac{1}{(n-1)} = \frac{S^2}{\sigma^2}$

$$-T_{\frac{\alpha}{2},n-1,\Delta} < \frac{\frac{\bar{y}-\mu}{\sigma/\sqrt{n}} + \Delta}{S/\sigma} < T_{\frac{\alpha}{2},n-1,\Delta}$$

$\dfrac{\bar{y}-\mu+\Delta\left(\frac{\sigma}{\sqrt{n}}\right)}{\frac{\sigma}{\sqrt{n}}\cdot\frac{S}{\sigma}}$

let $\Delta = -z_q\sqrt{n}$ $\qquad z_q\sqrt{n}\left(\frac{\sigma}{\sqrt{n}}\right)$

$$-T_{\frac{\alpha}{2},n-1,-z_q\sqrt{n}} < \frac{\frac{\bar{y}-\mu}{\sigma/\sqrt{n}} - z_q\sqrt{n}}{S/\sigma} < T_{\frac{\alpha}{2},n-1,-z_q\sqrt{n}}$$

$$-T_{\frac{\alpha}{2},n-1,-z_q\sqrt{n}} < \frac{\frac{\bar{y}-\mu-z_q\sigma}{\sigma/\sqrt{n}}}{S/\sigma} < T_{\frac{\alpha}{2},n-1,-z_q\sqrt{n}}$$

$$-T_{\frac{\alpha}{2},n-1,-z_q\sqrt{n}} < \frac{\bar{y}-\mu-z_q\sigma}{S/\sqrt{n}} < T_{\frac{\alpha}{2},n-1,-z_q\sqrt{n}}$$

$$-T_{\frac{\alpha}{2},n-1,-z_q\sqrt{n}} \cdot \frac{S}{\sqrt{n}} < \bar{y} - (\mu + z_q\sigma) < T_{\frac{\alpha}{2},n-1,-z_q\sqrt{n}} \cdot \frac{S}{\sqrt{n}}$$

$$\bar{y} - T_{\frac{\alpha}{2},n-1,-z_q\sqrt{n}} \cdot \frac{S}{\sqrt{n}} < \mu + z_q\sigma < \bar{y} + T_{1-\frac{\alpha}{2},n-1,-z_q\sqrt{n}} \cdot \frac{S}{\sqrt{n}}$$

② $P\left(-t_{\frac{\alpha}{2},n-1} < t < t_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$

$P\left(-t_{\frac{\alpha}{2},n-1} < \frac{\bar{y}-\mu}{S/\sqrt{n}} < t_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$

$\bar{y}_{obs} \pm t_{\frac{\alpha}{2},n-1}\cdot\frac{S_{obs}}{\sqrt{n}}$ observed

$\bar{y} \pm t_{\frac{\alpha}{2},n-1}\cdot\frac{S}{\sqrt{n}}$ random

$$P\left(\bar{y}_{obs} - t_{\frac{\alpha}{2},n-1}\cdot\frac{S_{obs}}{\sqrt{n}} < \bar{y} < \bar{y}_{obs} + t_{\frac{\alpha}{2},n-1}\cdot\frac{S_{obs}}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\frac{\bar{y}_{obs} - t_{\frac{\alpha}{2},n-1}\cdot\frac{S_{obs}}{\sqrt{n}} - \mu}{S/\sqrt{n}} < \frac{\bar{y}-\mu}{S/\sqrt{n}} < \frac{\bar{y}_{obs} + t_{\frac{\alpha}{2},n-1}\cdot\frac{S_{obs}}{\sqrt{n}} - \mu}{S/\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\frac{\bar{y}_{obs}-\mu}{S/\sqrt{n}} - t_{\frac{\alpha}{2},n-1}\cdot\frac{S_{obs}}{S} < \frac{\bar{y}-\mu}{S/\sqrt{n}} < \frac{\bar{y}_{obs}-\mu}{S/\sqrt{n}} + t_{\frac{\alpha}{2},n-1}\cdot\frac{S_{obs}}{\sqrt{n}}\right) = 1-\alpha$$

since we know that $P\left(-t_{\frac{\alpha}{2},n-1} < \frac{\bar{y}-\mu}{S/\sqrt{n}} < t_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$,

Then, $\dfrac{\bar{y}_{obs}-\mu}{S/\sqrt{n}} = 0$

```r
# Part 3a
# H0: mu1 - mu2 = 0
# H1: mu1 - mu2 != 0
y1 = c( 65, 82, 81, 67, 57, 59, 66, 75, 82, 70 )
y2 = c( 64, 56, 71, 69, 83, 74, 59, 82, 65, 79 )
n1 = 10
n2 = 10
mean1 = mean(y1)
mean2 = mean(y2)

# REJECTION REGION
(t_alpha = qt(0.05/2, n1+n2-2))
```

```
## [1] -2.100922
```

```r
(t_alpha1 = qt(1-(0.05/2), n1+n2-2))
```

```
## [1] 2.100922
```

```r
(sp = sqrt(((n1-1)*var(y1) + (n2-1)*var(y2))/(n1+n2-2)))
```

```
## [1] 9.315459
```

```r
(t_obs = (mean(y1) - mean(y2))/(sp*sqrt(1/n1 + 1/n2)))
```

```
## [1] 0.04800768
```

```r
# P-VALUE
(p_val = 2*pt(-t_obs,n1+n2-2))
```

```
## [1] 0.9622388
```

```r
# CONFIDENCE INTERVAL
(lower_bound = (mean1-mean2)+t_alpha*sp*(sqrt((1/n1)+(1/n2))))
```

```
## [1] -8.552441
```

```r
(upper_bound = (mean1-mean2)-t_alpha*sp*(sqrt((1/n1)+(1/n2))))
```

```
## [1] 8.952441
```

The rejection region are <-2.1009 and >2.1009. we know that the t observation is not inside the rejection region, which is why we cannot reject the H0 that (mu1-mu2) = 0. Since alpha=0.05 and p-value=0.96, we cannot reject the null hypothesis because p-value>alpha. The confidence interval is [-8.55244,8.95244]. Since 0 is inside the CI, we can choose to not reject the null hypothesis.

```r
# Part 3b
# H0: sigma1 = sigma2
# H1: sigma1 != sigma2

# REJECTION REGION
(f_alpha = qf(0.05/2, n1-1, n2-1))
```

```
## [1] 0.2483859
```

```r
(f_alpha1 = qf(1-(0.05/2), n1-1, n2-1))
```

```
## [1] 4.025994
```

```r
(f_obs = var(y1)/var(y2))
```

```
## [1] 0.9782168
```

```r
# P-VALUE
(p_val_f = 2*pf(f_obs,n1-1,n2-1)) # 2-sided
```

```
## [1] 0.9743665
```

```r
# CONFIDENCE INTERVAL (why?)
(lower = (var(y1)/var(y2)) * qf(0.05/2, n2-1, n1-1))
```

```
## [1] 0.2429752
```

```r
(upper = (var(y1)/var(y2)) * qf(1-0.05/2, n2-1, n1-1))
```

```
## [1] 3.938295
```

The rejection region are <0.248386 and >4.0259942. We know that the f observation is not inside the rejection region, which is why we cannot reject the H0. Since alpha=0.05 and p-value=0.97, we cannot reject the null hypothesis because p-value>alpha. The confidence interval is [0.24298,3.9383]. Since the CI includes 1, we cannot reject H0.

Lecture 8

Lecture 8

① $SS_{Treatment} = n\sum_i^a (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \frac{1}{n}\sum_i^a y_{i\cdot}^2 - \frac{1}{N} y_{\cdot\cdot}^2$

$$n\sum_i^a (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = n\sum_i^a \left((\bar{y}_{i\cdot})^2 - 2\bar{y}_{i\cdot}\bar{y}_{\cdot\cdot} + (\bar{y}_{\cdot\cdot})^2\right)$$

$$= n\sum_i^a \left(\left(\tfrac{1}{n} y_{i\cdot}\right)^2 - \tfrac{2}{n} y_{i\cdot}\tfrac{1}{an} y_{\cdot\cdot} + \left(\tfrac{1}{an} y_{\cdot\cdot}\right)^2\right)$$

$$= \frac{1}{n}\sum_i^a (y_{i\cdot})^2 - \frac{2}{an} y_{\cdot\cdot}\sum_i^a y_{i\cdot} + \frac{1}{a^2 n}(y_{\cdot\cdot})^2 \sum_i^a 1$$

$$= \frac{1}{n}\sum_i^a (y_{i\cdot})^2 - \frac{2}{an} y_{\cdot\cdot}\sum_i^a y_{i\cdot} + \frac{1}{an}(y_{\cdot\cdot})^2$$

$$= \frac{1}{n}\sum_i^a (y_{i\cdot})^2 - \frac{2}{N} y_{\cdot\cdot}\sum_i^a y_{i\cdot} + \frac{1}{N}(y_{\cdot\cdot})^2$$

$$= \frac{1}{n}\sum_i^a (y_{i\cdot})^2 - \frac{2}{N}(y_{\cdot\cdot})^2 + \frac{1}{N}(y_{\cdot\cdot})^2$$

$$= \frac{1}{n}\sum_i^a (y_{i\cdot})^2 - \frac{1}{N}(y_{\cdot\cdot})^2$$

(2a) $\sum_i^a \sum_j^n (y_{ij} - \mu)^2 \Rightarrow$ 0 constraints

Therefore, df = an

(2b) $\sum_i^a \sum_j^n (y_{ij} - \bar{y}_{\cdot\cdot})^2 \Rightarrow$ 1 constraint: $\sum_i^a \sum_j^n (y_{ij} - \bar{y}_{\cdot\cdot}) = 0$

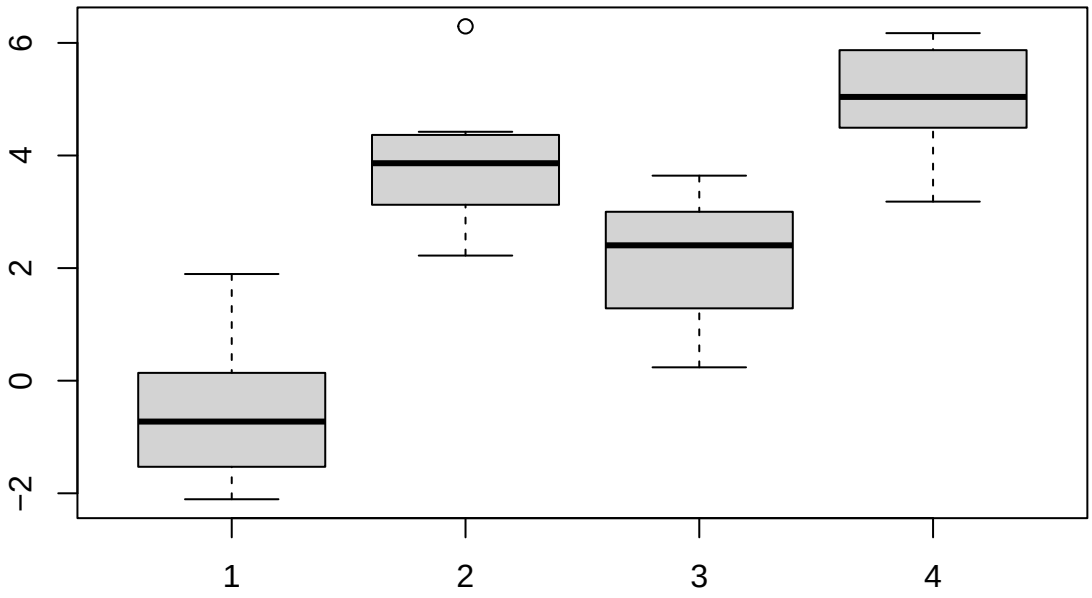
Therefore, df = an − 1

(2c) $\sum_i^a (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \Rightarrow$ 1 constraint: $\sum_i^a (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) = 0$

Therefore, df = a − 1

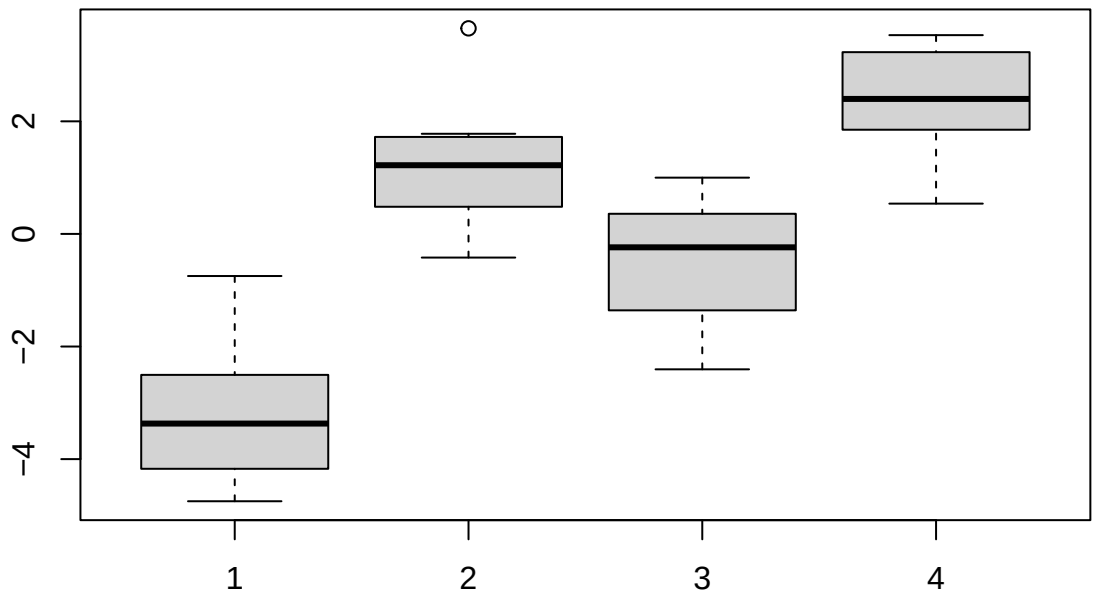
```
y = matrix(nrow=4,ncol=10)
y[1,] = c( -2.10552316, 1.89491371, -1.52919682, -0.99265143, -0.45911960, 1.09271028, -1.54680778, 0.13
y[2,] = c( 4.25667943, 4.36518096, 4.42108835, 3.77229146, 2.22264903, 3.95354759, 6.29377745, 3.585010
y[3,] = c( 3.64209745, 2.76932242, 1.46001019, 0.23739519, 0.27629510, 2.83897173, 2.99999590, 3.546578
y[4,] = c( 3.18088593, 5.44976665, 5.87116946, 4.01275036, 6.00826692, 5.19220036, 6.17338313, 4.888460

# part a
boxplot(y[1,],y[2,],y[3,],y[4,])
```



We can see from the boxplots that x has some effect on y. We can see that y4 has a higher distribution than the other y's, while y1 has the smallest values. We can also see that there is a difference between the y2 and y3.

```
# part b
grand_mean = mean(y)
y_measure = y-grand_mean
boxplot(y_measure[1,],y_measure[2,],y_measure[3,],y_measure[4,])
```



Subtracting the mean does not change anything, but the overall boxplot are just shifted downwards. It still looks like that x has an effect on y.

```
# part c
y_bar = c(mean(y[1,]),mean(y[2,]),mean(y[3,]),mean(y[4,]))
(effects = c(y_bar[1]-grand_mean, y_bar[2]-grand_mean, y_bar[3]-grand_mean, y_bar[4]-grand_mean))
```

```
## [1] -3.0961875  1.2613624 -0.5329403  2.3677655
```

```
# part d
(se = c(sd(y[1,])/sqrt(10),sd(y[2,])/sqrt(10),sd(y[3,])/sqrt(10),sd(y[4,])/sqrt(10)))
```

```
## [1] 0.3986627 0.3433781 0.3964595 0.2976147
```

8-3e) -3.096 +- 0.399 1.261 +- 0.343 -0.533 +- 0.396 2.368 +- 0.298 it looks like the third effect might be zero because zero is inside the interval of mean+-se

```
# part f
n = 10
a = 4
var = c(var(y[1,]),var(y[2,]),var(y[3,]),var(y[4,]))
SS_treatment = n*sum(effects^2)
SS_error = (n-1)*(sum(var))
MS_treatment = SS_treatment/(a-1)
MS_error = SS_error/((n*a)-a)
(F_ratio = MS_treatment / MS_error)
```

```
## [1] 43.54614
```

```
(F_alpha = qf(0.05,a-1,a*n-a,lower.tail=FALSE))
```

```
## [1] 2.866266
```

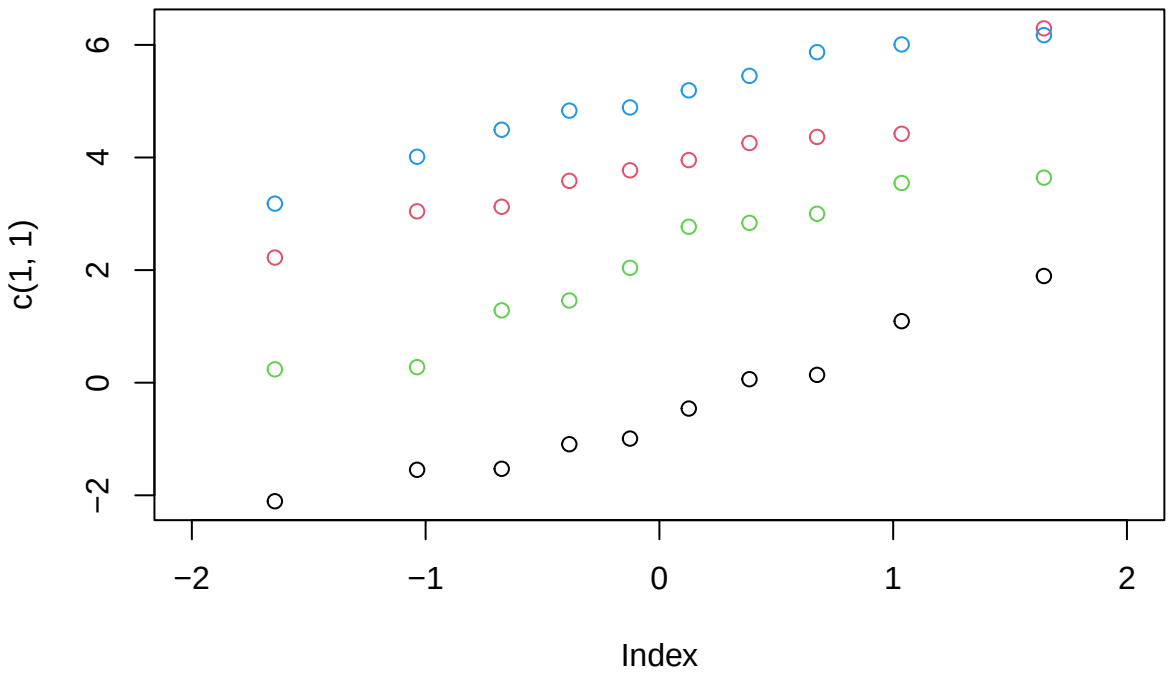
The rejection region is >2.866. Since, the F_ratio is inside the rejection region we choose to reject H0 in favor of H1. This means that x has an effect on y

```
# part g
x_vec=rep(1:4,10)
y_vec = as.numeric(y) # visually confirm y.vector is the correct y in vector form.
summary.aov(lm(y_vec ~ as.factor(x_vec)))
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## as.factor(x_vec)  3 170.68   56.89   43.55 4.58e-12 ***
## Residuals        36  47.03    1.31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 4.58e-12. Since 4.58e-12 is smaller than alpha, then we choose to reject the H0 in favor of H1

```
# part h
plot(c(1,1),cex=0, xlim=c(-2,2), ylim=range(y))
for (i in 1:a) {
  x = y[i,]
  n = length(x)
  probs = seq(0.5/n, 1-0.5/n, length = n)
  q = qnorm(probs,0,1)
  points(q,sort(x),col=i)
}
```



We can see that the qq plots are pretty much straight, and this means that they most likely follow the normal distribution. Their slopes are also pretty similar which means that their variances are also similar.