Chapter 1

Creating vectors: c(1,2,3) = 1 2 3, c("a", "b") = "a", "b"
rep(2,5) = 2 2 2 2 2
seq(from = 2, to=10, by=3)= 2,5,8
seq(1,10,length=3) = 1.0, 5.5, 10.0
class(a) = "numeric" / "character" (cannot add both tgt)

Creating matrix: matrix(1:12, byrow=T, nrow=3)
Matrix multiplication: %*%, Element multiplication: *
Dim(), +, -, transpose:t(), inverse: solve(), determinant: det
diag(c(1:2)) diagonal 1,2 others 0
[-r,-c]: remove row and col
r/cbind(1:3, c(2,3,5)) : add row/col

Creating dataframes:

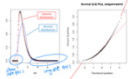
```
Customer = c('Ian', 'John', 'Keegan')
Married = c(TRUE, FALSE, TRUE)
Age = c(35, 26, 48)
Opinion = c('Excellent', 'Not bad', 'Good')

survey = data.frame(Customer, Married, Age, Opinion)
```

```
survey
Customer Married Age   Opinion
     Ian    TRUE  35 Excellent
    John   FALSE  26   Not bad
  Keegan    TRUE  48      Good
```

```
Subject=rep(1001:1027, rep(3,27))
Treatment=c("Placebo", "Low Dose", "High Dose")
Period=c(1,2,3,3,1,2,2,3,1)
print(cbind(Subject, Treatment, Period))
```

survey[2,]; ➔ 2nd row. survey[,3]; ➔ 3rd column
subset(survey, Age>30)
survey[survey$Age>30 &survey$a>2,]
data[order(data$var, decreasing=F),]

```
painscore=c(0:10, 0:8)
painvector=NULL
for(i in c(1:length(painscore)))
        {painvector[i]=painlevel(painscore[i])}
pain_level = cbind(painscore,painvector)
print(pain_level)
```



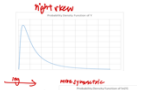
ifelse(a %% 2 == 0,"even","odd")
C= function(n) { value = 0
i = 2
while(i <= n){value =value+ i^3; i=i +1}
return(value) }

```
best3_mean <- function(score) {y=score[order(score, decreasing =
TRUE)]
return(mean(y[1:3])) }

best3_ave = apply(CA_result[,2:6],1,best3_mean)
```

average of best 3 test scores

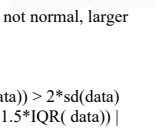
Chapter 2

sum(data)/length(data) ==mean(data) mean(data,trim=0.1)
# -10% each end



hist(r) (lg shift mean right, exp left)
xpt<- seq(-001,100,by=0.1)
n_den<-dnorm(xpt,mean(r),sd(r))
ypt<-n_den*length(r)*10
lines(xpt,ypt,col="blue")

See if data is normally distributed
1. q=seq(from=0.02,to=0.98,by=0.02)
x=2*3qnorm(q) ⇒N(2,32)
OR x=qt(q,df=1) ⇒ t distribution v=1
qqnorm(x)
qqline(x, col="blue")
2.Shapiro.test(return) ⇒ small values not normal, larger sample more stat significant



boxplot(data) ⇒ Identify outliers
Classic technique: abs(data - mean(data)) > 2*sd(data)
Boxplot: (data < quantile(data, 0.25)-1.5*IQR(data)) | (data > quantile(data,0.75)+1.5*IQR(data))
median and mean ➔ boxplot, bigger spread (var) ➔ hist

Chapter 3

CLT: for n>=30

- When σ known, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ as $n \to \infty$

$$\Pr\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha \quad \bar{X} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$$

- When σ unknown, use sample sd $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})$

$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ ~ t_{n-1}, then

CI: $\bar{X} - t_{\alpha/2,n-1}\cdot\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2,n-1}\cdot\frac{s}{\sqrt{n}}$

Construct Confidence Interval 100(1-α)%
1. norm: z <- qnorm(1-alpha/2), z1<- qnorm(1-alpha) //one sided
zCI90p<- c(xbar-z*s/sqrt(n), xbar+z*s/sqrt(n)) //two-sided
zCI90p <- c("-infty", xbar+ z1*s/sqrt(n)) //one-sided (-infty, u)
2. t-statistic: CI90p = c(xbar=t*s/sqrt(n), xbar+t*s/sqrt(n)
t <- qt(1-alpha/2, df=n-1) if 90% ➔ alpha<- 0.10
- T distribution approaches normal when degree of freedom ↑
3. t.test(x, conf.level=0.9, alt="less") "greater/two.sided"
t-tests are for means. If no value for mu is given, 0 is set default.
4.Estimation of proportion: for large n, CI for p will be
•Bernoulli: mean = p, var = p(1-p)
•Binomial: mean = np, var = np(1-p)

$\hat{p} - z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

•F approximation: $v_1 = 2(n-np+1)$, $v_2 = 2np$, $v_1' = v_2+2$, $v_2' = v_1-2$.

```
CI90 <- c(phat-z*sqrt(phat*(1-phat)/n), phat+z*sqrt(phat*(1-phat)/n))
```

$$\text{Lower Limit} = \frac{n\hat{p}}{n\hat{p} + (n-n\hat{p}+1)F_{\alpha/2,v_1,v_2}}$$

$$\text{Upper Limit} = \frac{(n\hat{p}+1)F_{\alpha/2,v_1',v_2'}}{n-n\hat{p}+(n\hat{p}+1)F_{\alpha/2,v_1',v_2'}}$$

approximation.

v1 = 2*(n-n*phat+1)
v2 = 2*n*phat
v1p= v2+2
v2p= v1-2

```
Lower = n*phat/(n*phat+(n-n*phat+1)*qf(1-alpha/2, v1p, v2p))
Upper = (n*phat+1)*qf(1-alpha/2, v1p, v2p)/(n-n*phat+(n*phat+1)*qf(1-alpha/2, v1p, v2p))
```

•prop.test(53,100,conf.level=0.9, correct=FALSE) ⇒ w/o correction
-discrete is approximated by continuous, np and np(1-p) >= 5

Hypothesis testing
H_1: what we want to show
•P(Type I error) = α ⇒ reject H0 when H_0 is TRUE
•P(Type II error) = β ⇒ did not reject H_0 when H_0 is FALSE

$\alpha = Pr(X \geq r | H_0 \text{ true}) = Pr(Z \geq r)$ ← pnorm(r)
$\beta = Pr(X < r | H_1 \text{ true}) = Pr(Z < (r-1))$ ← pnorm(r-1)

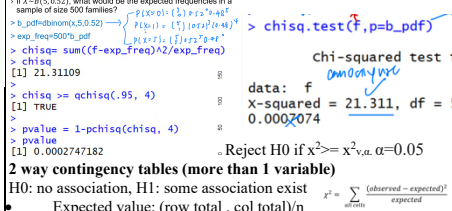
Power of a test ⇒ 1 – β
probability of rejecting H_0 when H_1 is true. When n ↑, α↓, β ↑ ⇒ power ↓

P value (reject H_0 when p <= α, do not reject H_0 when p>α)
- Normal: z = (xbar - mu0)/ sd * sqrt(n)
z_pvalue = 2*pnorm(z) //for 2 sided
- T-approximations: t=(xbar-mu0)/sd * sqrt(n)
t_pvalue= 2*pt(z,df=n-1) // 2 sided
- t.test(wind, conf.level = 0.95, mu=2, alt="greater") H_0: 2,H_1:>2
- pvalue = 1-pnorm((phat-p0)/sqrt(p0*(1-p0)/n)) //for large n
p0: H_0= 0.1
- prop.test(x=5850, n=50000, p=0.113 (H0 value), alt="greater")

Chapter 4

Chi-square goodness of fit

- Degree freedom v= (#categories) – 1 – #parameters
- H0: no significant diff bet observed and expected
  H1: have significant diff bet observed and expected



If $X \sim B(5,0.52)$, what would be the expected frequencies in a sample of size 500 families?

```
> b_pdf=dbinom(x,5,0.52)
> exp_freq=500*b_pdf
> chisq= sum((f-exp_freq)^2/exp_freq)
> chisq
[1] 21.31109
>
> chisq >= qchisq(.95, 4)
[1] TRUE
>
> pvalue = 1-pchisq(chisq, 4)
> pvalue
[1] 0.0002747182
```

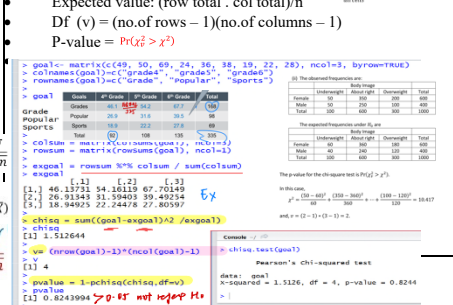
```
> chisq.test(f,p=b_pdf)

        Chi-squared test f

data:  f
X-squared = 21.311, df = 5
0.0002074
```

Reject H0 if x^2>= $x^2_{v,\alpha}$ α=0.05

2 way contingency tables (more than 1 variable)
H0: no association, H1: some association exist

$$\chi^2 = \sum_{\text{all cells}} \frac{(observed - expected)^2}{expected}$$

- Expected value: (row total . col total)/n
- Df (v) = (no.of rows – 1)(no.of columns – 1)
- P-value = $Pr(\chi^2_v > \chi^2)$

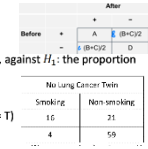


```
> goal<- matrix(c(49, 50, 69, 24, 36, 38, 19, 22, 28), ncol=3, byrow=TRUE)
> colnames(goal)=c("grade4", "grade5", "grade6")
> rownames(goal)=c("Grade", "Popular", "Sports")
>
> goal
        grade4 grade5 grade6
Grade       49     50     69
Popular     24     36     38
Sports      19     22     28
> colsum = matrix(c(sum(goal[,1]), sum(goal[,2]), sum(goal[,3])), ncol=3)
> rowsum = matrix(rowsums(goal), ncol=1)
>
> exgoal = rowsum %*% colsum / sum(colsum)
> exgoal
          [,1]     [,2]     [,3]
[1,] 46.13731 54.16119 67.70149
[2,] 26.91343 31.59403 39.49254
[3,] 18.94925 22.24478 27.80597
>
> chisq = sum((goal-exgoal)^2 /exgoal)
> chisq
[1] 1.512644
>
> vw (nrow(goal)-1)*(ncol(goal)-1)
> v
[1] 4
>
> pvalue = 1-pchisq(chisq,df=v)
> pvalue
[1] 0.8243994
```

```
> chisq.test(goal)

        Pearson's Chi-squared test

data:  goal
X-squared = 1.5126, df = 4, p-value = 0.8244
```

Paired 2 way contingency tables (2 conditions)
H0: no. of changes in each direction is equally likely

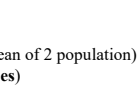
- tables(patients$Before, patients$After)
- **Chi-square statistic**
  Without correction: (B-C)²/(B+C);
  With correction: (|B-C| -1)² / B+C with df=1
- Use mcnemar.test()



We shall use a McNemar Change test.
We test H_0: the r lung cancer is not related to smoking, against H_1: the proportion of people having lung cancer is related to smoking.

R code:

```
smoke = matrix(c(16, 21, 4, 59), nrow=2, byrow = T)
mcnemar.test(smoke)
```



Chapter 5

Two independent sample (comparing mean of 2 population)

- 2 normal samples (**known variances**)

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

- **Unknown variances**

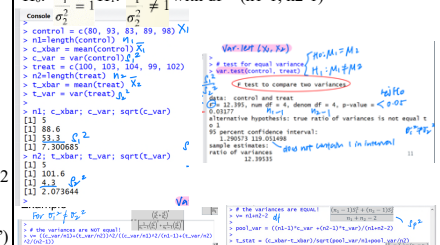
$$S_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(X_{1j} - \bar{X}_1)^2$$

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_v - \text{distribution}$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2$$

- [Case 1: $\sigma_1^2 = \sigma_2^2$] ⇒ v = n_1+n_2-2
- [Case 2: $\sigma_1^2 \neq \sigma_2^2$] ⇒ $v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$

F-test➔ var.test() (determine if var1=var2) ⇒ $F = S_1^2/S_2^2$.
- H_0: $\sigma_1^2 = \sigma_2^2$ H_1: $\sigma_1^2 \neq \sigma_2^2$ with df = (n1-1, n2-1)



To test whether equal var: var.test(X1, X2)
After conducting equal or not,t.test(X1, X2, var.equal=T/F)

- p-value : 2*(1-pf(Fstar, n1-1, n2-1)
- p-value: 2*pt(t_stat, df=v)

t.test(runtime$Mile ~ runtime$Athlete, var.equal=T, alternative = "two.sided")

Two Normal Dependent Samples ➔ paired t.test()
Eg. whether there is drug effect (control-treatment)

$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \qquad S_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$

$\frac{\bar{D} - \delta}{\sqrt{S_d^2/n}} \sim t_v$ – distribution V = n-1 & n is small

- H_0: $\mu_1 - \mu_2 = 0$ H_1: $\mu_1 - \mu_2 \neq 0$
- t.test(drug$control, drug$treat, mu=0, paired=T)
OR t.test(drug$control - drug$treat)

Multiple (>2) samples ⇒ $\sigma_1^2 = \sigma_2^2$ (ANOVA)
Eg. Non-smoker, ex-smoker, smoker

- H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$, H_1 = $\mu_i \neq \mu_j$ /not all μ_i are equal
- **if** $\mu_1 = \mu_2 = \mu_3$ ⇒ spread out ⇒ easy see diff in μ
- if σ^2 is large, difficult to see diff in μ
- From boxplots, ANOVA model looks appropriate. This is because distribution of samples are roughly symmetric (median is roughly in the middle) and have approximately equal spread. ➔ normal distribution

Within-group Mean Square:

$$S_w^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)S_i^2$$

Between-group Mean Square:

$$S_b^2 = \frac{1}{k-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2.$$

Between, Within group sum of squares:

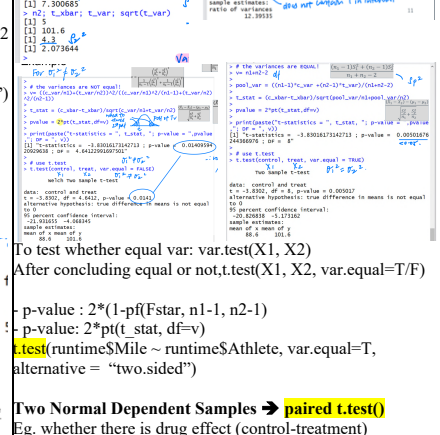
$$SSB = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2 \quad SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2,$$

Total sum of squares (SSB+SSW)

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2 + (X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2$$

F-statistic: $F = \frac{S_b^2}{S_w^2} \sim F(k-1, n-k)$ K-1: numerator df
N-k: denominator df

ANOVA test (1) (diet example)
H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_1: not all μ_i are equal



```
> for (i in c(1:3)) {gpx[i]=length(diet$wloss[diet$Diet==i])}
> for (i in c(1:3)) {gpxbar[i]=mean(diet$wloss[diet$Diet==i])}
> for (i in c(1:3)) {gpvar[i] =var(diet$wloss[diet$Diet==i])}
>
> gpx
[1] 24 27 27
> gpxbar
[1] 3.300000 3.025926 5.148148
> gpvar
[1] 5.018261 6.367379 5.738746
>
> xbar=mean(diet$wloss)
>
> w_var=sum((gpx-1)*gpvar)/(sum(gpx)-length(gpx))
> b_var=sum(gpx*(gpxbar-xbar)^2)/(length(gpx)-1)
>
> fstat = b_var/w_var
> pvalue = 1- pf(fstat, df1=length(gpx)-1, df2 = sum(gpx)-length(gpx
```

```
> aov(diet$wloss~factor(diet$Diet))
Call:
   aov(formula = diet$wloss ~ factor(diet$Diet))

Terms:
                factor(diet$Diet) Residuals
Sum of Squares             71.0937 430.1793
Deg. of Freedom                 2        75

Residual standard error: 2.394937
Estimated effects may be unbalanced
> summary(aov(diet$wloss~factor(diet$Diet)))
                  Df Sum Sq Mean Sq F value
factor(diet$Diet)  2  71.09  35.55   6.197
Residuals         75 430.18   5.74
                   Pr(>F)
factor(diet$Diet) 0.00323 **
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

when H_0 rejected, determine which grp diff

```
> pairwise.t.test(diet$wloss, diet$Diet, p.adjust.method = "none")

        Pairwise comparisons using t tests with pooled SD

data:  diet$wloss and diet$Diet

  1      2
2 0.6845 -
3 0.0075 0.0017

P value adjustment method: none
```

```
> anova(aov(data = ratdata, Wt_gain~as.factor(Diet)))
Analysis of Variance Table
Response: Wt_gain
                  Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(Diet)    3 5480.5 1826.83 7.943 0.0002973 ***
Residuals         39 8969.7  229.99
---
```

```
> pairwise.t.test(ratdata$Wt_gain, ratdata$Diet, p.adjust.method = "none")

        Pairwise comparisons using t tests with pooled SD

data:  ratdata$Wt_gain and ratdata$Diet

  A      B      C
B 0.3245 -      -
C 0.0048 0.0519 -
D 7e-05  0.0017 0.1960

P value adjustment method: none
```

Group (A&B) diff from (C&D)

Non parametric test: use for not normal

For one sample: Quantile test (binomial test)

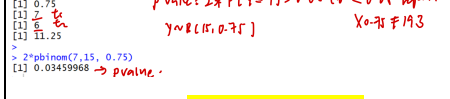
- P(X⩽50) = 0.1 ⇒ $X_{0.1}$=50 (10% students get ⩽50pts)
- H_0: p^{th} quantile = X_p, H_1: p^{th} quantile != X_p
- t_1 : no. of observations ⩽ X_p, t_2: # observations < X_p

If t_1 is too small (<np) ⇒ p-value= 2*Pr(Y⩽t_1) (2side)
If t_2 is too large (>np) ⇒ p-value = 2*Pr(Y⩾ t_2)
= 2* [1-Pr(Y⩽t_2 -1)]

```
> x=c(189, 233, 195, 160, 212,176, 231, 185, 199, 213, 202, 193, 174, 166, 248)
> n=length(x)
> p=0.75
>
> t1=sum(x<=193)
> t2=sum(x<193)
>
> n; p; t1; t2; n*p
[1] 15
[1] 0.75
[1] 7
[1] 6
[1] 11.25
>
> 2*pbinom(7,15, 0.75)
[1] 0.03459968
```

H0: 75th quantile of population is 193
H1: 75th quantile of population is NOT 193
p-value: 2* P[Y⩽7] = 0.0746 < 0.05 reject H0.
Y ~ B(15, 0.75)
X0.75 ≠ 193

2 independent samples: Wilcoxon rank sum test
H_0 : Pr(X<Y) =½ (identically distributed)
H_1: P(X<Y) != ½ (not identically distributed)
Find out whether 2 populations are identically distributed
Reject H0 if T is too small near n(n+1)/2 or too large [n(n+1)/2 + nm]



$T = \sum_{i=3}^{n} R(X_i) = (6+8+9+\cdots+18) = 111$

$W = T - \frac{n(n+1)}{2} = 111 - \frac{9\times 10}{2} = 66$

```
> # Data in two numeric vectors
> men_weight <- c(67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4)
> women_weight <- c(38.9, 61.2, 73.3, 21.8, 63.4, 64.6, 48.4, 48.8, 48.5)
> # Create a data frame
> my_data <- data.frame(
+   group = rep(c("Man", "Woman"), each = 9),
+   weight = c(men_weight,  women_weight)
+ )
>
> my_data$rank <- rank(my_data$weight,ties.method = "average")
```

```
sum(my_data[my_data$group=="Man", "rank"])
wilcox.test(men_weight, women_weight)
```

$T = \sum R(x_i)$

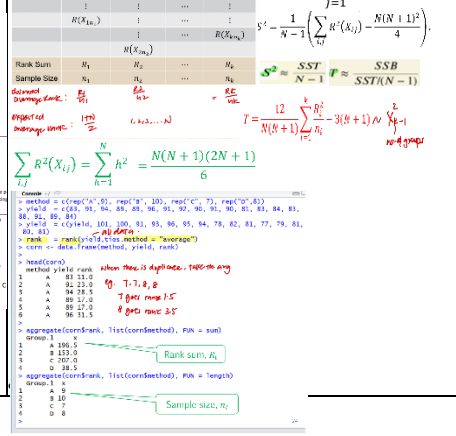
```
        Wilcoxon rank sum test with continuity correction

data:  men_weight and women_weight
W = 66, p-value = 0.02712
alternative hypothesis: true location shift is not equal to 0
```

Value	12	23	30	34	35	36	37	39	41
Rank	1	2	3	4	5	6	7	8	9
Diet	X	X				X	X	X	X

Value	42	43	44	52	57	59	60
Rank	10	11	12	13	14	15	16
Diet			X				

Rank sum of X, T = (1+2+6+7+8+9+12) = 45
Wilcoxon test statistic, W = T – n(n+1)/2 = 45 – 7*8/2 = 17, where n is the sample size of X.
From the output, the p-value is 0.069.
X: {23, 12, 37, 36, 39, 41, 44}, and *Y*: {30, 35, 43, 60, 59, 52, 57, 42, 34}.

Multiple Ind. Samples: Kruskal-Wallies test



	Sample 1	Sample 2	…	Sample k
	$R(X_{11})$	$R(X_{21})$	…	$R(X_{k1})$
	$R(X_{12})$	$R(X_{22})$	…	$R(X_{k2})$
	⋮	⋮	…	⋮
	$R(X_{1n_1})$	⋮	…	⋮
		⋮	…	$R(X_{kn_k})$
		$R(X_{2n_2})$		
Rank Sum	R_1	R_2	…	R_k
Sample Size	n_1	n_2	…	n_k

$R_i = \sum_{j=1}^{n_i} R(X_{ij})$

$S^2 = \frac{1}{N-1}\left(\sum_{ij} R^2(X_{ij}) - \frac{N(N+1)^2}{4}\right)$.

$S^2 \approx \frac{SST}{N-1}$ $F \approx \frac{SSB}{SST(N-1)}$

$T = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1) \sim \chi^2_{k-1}$

$\sum_{ij} R^2(X_{ij}) = \sum_{i=1}^{N} h^2 = \frac{N(N+1)(2N+1)}{6}$

```
> method = c(rep("A",9), rep("B", 10), rep("C", 7), rep("D",8))
> yield = c(81, 91, 94, 89, 89, 96, 91, 92, 90, 91, 90, 81, 83, 84, 83, 88, 91, 89, 84)
> rank = c(yield, 101, 100, 91, 93, 96, 95, 94, 78, 82, 81, 77, 79, 81, 80, 81)
> rank = rank(yield,ties.method = "average")
> corn <- data.frame(method, yield, rank)
>
> head(corn)
  method yield rank
1      A    83 11.0
2      A    91 23.0
3      A    94 28.5
4      A    89 17.0
5      A    89 17.0
6      A    96 31.5
>
> aggregate(corn$rank, list(corn$method), FUN = sum)
  Group.1     x
1       A 196.5
2       B 153.0
3       C 207.0
4       D  98.5
> aggregate(corn$rank, list(corn$method), FUN = length)
  Group.1  x
1       A  9
2       B 10
3       C  7
4       D  8
>
```

Rank sum, R_i

Sample size, n_i


```
> ranksum1 = aggregate(corn$rank, list(corn$method), FUN = sum)
> n1 = aggregate(corn$rank, list(corn$method), FUN = length)
> N= length(method)
>
> s.sq = (sum(rank^2) -N*((N+1)^2/4))/(N-1)
> T.stat1 = (sum(rank1^2)/n1[,2] - (N*(N+1)^2/4)) /(s.sq)
> pvalue = 1-pchisq(T.stat, 3)
>
> s.sq; T.stat; pvalue
[1] 98.3303
[1] 25.62864
[1] 1.140573e-05
```

- Use Kruskal.test()

```
>
> kruskal.test(corn$yield, corn$method)

        Kruskal-Wallis rank sum test

data:  corn$yield and corn$method
Kruskal-Wallis chi-squared = 25.629, df = 3, p-value =
1.141e-05
```

Dependent samples

1) Sign Test: (form of quantile test with p=0.5)

Tests whether one RV in a pair (Xi, Yi) tends to be larger than the other variable.

- "+" if $X_i < Y_i$, "-" if $X_i > Y_i$, "0" if $X_i = Y_i$ (discard tied pairs)
- H_0: P(+) = P(-) (p=0.5), H_1: P(+) != P(-)
- For H_0 # of '+' should be ½ sample size T = # of '+'
- T~B(n,½), E(T) = n/2, Var(T) = n/4 ⇒ T~N(n/2, n/4)

Suppose at the end of the period, 8 preferred B, 1 preferred A, and 1 reported no preference.

So, $n = 9$, $T = 8$. The p-value is $\Pr(T \geq 8 | T \sim B(9,0.5)) = 0.0195$

The binomial probability can be calculated using a probability table, or R code: 1-pbinom(7, 9, 0.5), or binom.test(8, 9, 0.5, alternative="greater").

- Can use prop.test(80,90,p=0.5,alternative="greater")
- Can also use normal approximations:T~Bin(90,0.5) ⇒ N(45,90/4), P(Z>=(80-45)/root(90/4)) <0.05

2) Wilcoxon signed rank: wilcox.test(y,x,paired=T)

(paired = T: dependent, if don't specify is independent)

H_0: distribution of X = distribution of Y

H_1: distribution of X not equal to distribution of Y

Steps:

1) find difference between Xi and Yi
2) absolute the difference
3) rank the difference eg. 1, 1, 2, 2 ⇒ rank of 1=1.5
4) Add back the signs to the rank
5) T = sum(r)/sqrt(sum(r^2))

x	86	71	77.0	68	91.0	77	91.0	70.0	71	88	87
y	88	77	76.0	64	96.0	65	90.0	65.0	80	81	72
D	2	6	-1.0	-4	5.0	-12	-1.0	-5.0	9	-7	-15
rank	3	7	-1.5	-4	5.5	-10	-1.5	-5.5	9	-8	-11

$$T = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}}$$

Multiple Dependent Sample: Friedman.test rank sum

H0: all populations identical dist

H1: not all population have identical dist

Eg. Customer be given several items and asked to rank the items

$$A = \sum_{i,j} R^2(X_{ij}),\ B = \frac{1}{n}\sum_i R_i^2 \qquad T = \frac{(k-1)n[B - nk(k+1)^2/4]}{A - nk(k+1)^2/4}$$

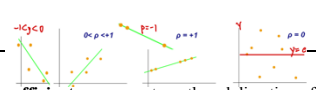
1:

```
> dataframe2 = dataframe[order(dataframe$store),]
>
> rownank=NULL
> for (i in 1:7){ rownank= c(rownank,
+ rank(dataframe2[dataframe2$store==i, 'sale']))}
>
> dataframe2= data.frame(dataframe2,rownank)
> dataframe2
   store brand sale rownank
1      1     A    5     7.0
8      1     B    4     1.0
15     1     C    7     3.0
22     1     D   10     4.0
29     1     E   12     5.0
2      2     A    1     2.5
9      2     B    3     4.0
16     2     C    1     2.5
23     2     D    0     1.0
30     2     E    2     4.0
3      3     A   16     2.0
10     3     B   12     1.0
17     3     C   22     3.5
24     3     D   22     3.5
31     3     E   35     5.0
4      4     A    4     4.5
11     4     B    4     2.5
18     4     C    1     1.0
```

```
> ranksum =
+ aggregate(dataframe2$rownank, list(dataframe2$brand), FUN
= sum) K_i
>
> n=7
> k=5
> A = sum(dataframe2$rownank^2) Σ
> B = sum(ranksum[,2]^2)/n B
>
> T = (k-1)*n*(B -n*k*(k+1)^2/4)/(A -n*k*(k+1)^2/4) (k-1)T
> T
[1] 8.328358
> 1-pchisq(T,df=4) k-1=4
[1] 0.08026409 2005
>
```

```
> friedman.test(sale~brand|store, data = dataframe)

        Friedman rank sum test

data:  sale and brand and store
Friedman chi-squared = 8.3284, df = 4, p-value
= 0.08026

>
> friedman.test(dataframe$sale,dataframe$brand,dataframe$store
e)

        Friedman rank sum test
```

3:



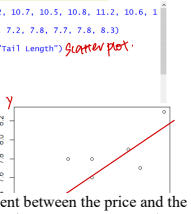
Chapter 7

Correlation coefficient: measure strength and direction of linear relationship (strong/weak/no) between 2 variables

0: no linear relation, strong: 0.8-1, moderate: 0.4-0.8, weak:0-0.4

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[X-\mu_X]^2}\sqrt{E[Y-\mu_Y]^2}}$$ If ρ=-1, inverse X=Y, ρ =1.

```
> x= c(10.4, 10.8, 11.1, 10.2, 10.3, 10.2, 10.7, 10.5, 10.8, 11.2, 10.6, 1
1.4)
> y= c(7.4, 7.6, 7.9, 7.2, 7.4, 7.1, 7.4, 7.2, 7.8, 7.7, 7.8, 8.3)
>
> plot(x,y, xlab = "Wing Length", ylab= "Tail Length")
>
> cov(x,y)
[1] 0.120303
> var(x)
[1] 0.1560606
> var(y)
[1] 0.1224242
> r= cov(x,y)/ sqrt(var(x)*var(y))
> r
[1] 0.8703546
> cor(x,y)
[1] 0.8703546
> se = sqrt((1-r^2)/(length(x)-2))
> se
[1] 0.1557186
```



Let ρ be the correlation coefficient between the price and the size. We test $H0$: $\rho = 0$ (no correlation) against $H1$: $\rho \neq 0$. From the output, we see that $\hat{\rho}$ = 0.927, with a p-value = 1.3e-12 (from the regression model). Hence, we reject $H0$, and conclude that there is a correlation between the selling price of an apartment and its size.

Sample correlation coefficient

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{X})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})}}$$

Standard error

$$S_R = \sqrt{\frac{1-R^2}{n-2}}$$

```
> 1/se
> 2*pt(-t, df=10)
[1] 0.0002311059
>
> f=(1+abs(r))/(1-abs(r))
> 2*(1-pf(f,10,10))
[1] 0.0002311059
>
> cor.test(x,y)
```

Hypothesis Testing

Case1: (H_0: ρ =0, H_1: ρ != 0)

1: T-test: $T = \frac{R}{S_R}$ df= n-2, P(|T|>= .87/.156) = p-value `2*(1-pt(r/sr, 10))`

2: F-test: $F = \frac{1+|R|}{1-|R|}$ df = (n-2,n-2)

```
(1+abs(r))/(1-abs(r))
2*(1-pf((1+abs(r))/(1-abs(r)),10, 10))
```

- cor.test(w,t)

can only use t and F distribution for testing H_0: ρ=0, <= 0 , >=0

Case2:H_0: ρ = any value not 0 eg, (H_0: ρ =0.75, H1:ρ!= 0.75(p_0))

- Fisher transformation $z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right),\ z_0 = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)$

$$Z = \frac{z - z_0}{\sqrt{1/(n-3)}} \sim \mathcal{N}(0,1).$$

```
z <- 0.5*log((1+r)/(1-r)); z
z0 <- 0.5*log((1+.75)/(1-.75)); z0
Z <- (z-z0)/sqrt(1/(12-3)); Z
```

`2*(1-pnorm(Z))`

Confidence Interval for correlation

- 95% CI for z:

```
zu <- z + qnorm(1-0.025)*(1/sqrt(12-3));
zl <- z - qnorm(1-0.025)*(1/sqrt(12-3));
```

- 95% CI for p:

```
(exp(2*zl)-1)/(exp(2*zl)+1)
(exp(2*zu)-1)/(exp(2*zu)+1)
```

Confidence Interval for Correlation Coefficient

- For example, in previous example, we have $r = 0.870$, $z = 1.33$ and $\sigma_z = \sqrt{1/(n-3)} = 1/\sqrt{12-3} = 0.333$.
- A 95% CI for z_p will be

$$z \pm Z_{0.025} \times \sigma_z = 1.33 \pm 1.96 \times 0.333 = (0.677, 1.983)$$

- Hence,

$$r_l = \frac{\exp(2\times 0.677) - 1}{\exp(2\times 0.677) + 1} = 0.5896 \text{ , and } r_u = \frac{\exp(2\times 1.983) - 1}{\exp(2\times 1.983) + 1} = 0.9628$$

```
zeelo = zee-qnorm(0.975)*szee
zeeup = zee+qnorm(0.975)*szee
rlo = (exp(2*zeelo)-1)/(exp(2*zeelo)+1)
rup = (exp(2*zeeup)-1)/(exp(2*zeeup)+1)
```

Simple Linear regression (X⇒ind.) (Y⇒ dependent)

Eg. Higher midterm ➔ higher end of term scores

Simple linear regression model based on plots

The plots suggest that the distribution of the residuals has slightly longer tails at both ends. This may be due to the two possible outliers at both ends. Other than that, the distribution is fairly symmetric, hence a simple linear regression model is considered reasonable

- Model: $Y_i = \beta_0 + \beta_1 X + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- Transform non-linear to linear: $y = \beta_0 e^{\beta_1 X} \Rightarrow \ln y = \ln(\beta_0) + \beta_1 X$
- Fitted regression line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left[n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2\right]$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\left[n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2\right]$$

Estimated regression line and residuals $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot X$

- SSE: $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\left[n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right]$$

- Estimators minimise SSE:

intercept $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ slope $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$

```
attach(waste)
n = length(x)
xbar = mean(x)
ybar = mean(y)
sxx = 1/n*(n*sum(x^2)-sum(x)^2)
syy = 1/n*(n*sum(y^2)-sum(y)^2)
sxy = 1/n*(n*sum(x*y)-sum(x)*sum(y))
b1hat = sxy/sxx
b0hat = ybar-b1hat*xbar
sse = syy-b1hat*sxy
```

The predicted model is: $\hat{Y} = 3.8296 + 0.9036X$

With $SSE = 323.3273$

- Use lm()

```
lm(waste$y ~ waste$x, data = waste)

Call:
lm(formula = waste$y ~ waste$x, data = waste)

Coefficients:
(Intercept)      waste$x
     3.8296       0.9036
```

Inferences:

can be

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

When given a dataset, an estimate would be

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}$$

s^2 is also called the **MSE**, **mean squared error**

Coefficient of Determination

- R^2 = 1- (SSE/S_{yy}) Best model: highest R^2 smaller SSE
- If R^2 = 0.8 80% variation of Y can be explained by this

Mean square error: $S^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{SSE}{n-2}$

β1 hat (slope parameter) has normal distribution

$$\hat{\beta}_1 = \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i - \bar{X})Y_i$$, E[β1hat] = β1, Var = $\frac{\sigma^2}{S_{XX}}$

T Statistic: β1hat has a normal distribution

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{XX}}}$$ df = n-2

If σ known, replace S with σ

- 95% Confidence interval

```
s2 <- sse/(n-2); s2
b1hat + qt(1-0.025, n-2)*sqrt(s2)/sqrt(sxx)
b1hat - qt(1-0.025, n-2)*sqrt(s2)/sqrt(sxx)
```

- P value for H0: β1 = 1, H1: β1 < 1

```
T <- (b1hat-1)/(sqrt(s2)/sqrt(sxx)); T
pt(T, df=31)
```

Example

Previously, we had $\hat{\beta}_1 = 0.9036$, $S_{xx} = 4152.182$ and $SSE = 323.3273$, so

$$s^2 = \frac{SSE}{n-2} = \frac{323.3273}{33-2} = 10.4299$$

A 95% CI for β_1 can be computed as:

$$\hat{\beta}_1 \pm t_{0.025}\frac{s}{\sqrt{S_{xx}}} = 0.9036 \pm 2.0395 \times \frac{\sqrt{10.4299}}{\sqrt{4152.182}} = (0.801, 1.006)$$

If we want to check whether β_1 is less than 1.0, then we may test, H_0: $\beta_1 = 1.0$ against H_1: $\beta_1 < 1.0$,

Under H_0, the test statistics

$$T = \frac{\hat{\beta}_1 - 1}{s/\sqrt{S_{xx}}}$$ has a t-distribution with DF = n − 2 = 31

The observed t-statistic is:

$$t = \frac{0.9036 - 1}{\sqrt{10.4299}\sqrt{4152.182}} = -1.9234$$

The p-value:

$$\Pr(T \leq t) = \Pr(T < -1.9234) = 0.0318$$

β0 (intercept parameter) hat

$$\hat{\beta}_0 = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{X}(X_i - \bar{X})}{S_{XX}}\right)Y_i$$ E[β0hat] = β0,

$$Var[\hat{\beta}_0] = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)$$

$$T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}}$$ df = n-2

Previously, we had $\hat{\beta}_0 = 3.8296$, $S_{xx} = 4152.182$, $\bar{x} = 33.45455$ and $s^2 = 10.4299$

A 95% CI for β_0 can be computed as:

$$\hat{\beta}_0 \pm t_{0.025} \times S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

$$= 3.8296 \pm 2.0395 \times \sqrt{10.4299} \times \sqrt{\frac{1}{33} + \frac{33.45455^2}{4152.182}}$$

$$= 3.8296 \pm 3.6067 = (0.223, 7.436)$$

If we want to check whether β_0 is zero, then we may test, H_0: β_0 = 0 against H_1: $\beta_0 \neq 0$,

Under H_0, the test statistics

$$T = \frac{\hat{\beta}_0 - 0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$ has a t-distribution with DF = n − 2 = 31

The observed t-statistic is:

$$t = \frac{3.8296 - 0}{\sqrt{10.4299} \times \sqrt{\frac{1}{33} + \frac{33.45455^2}{4152.182}}} = 2.1655$$

The p-value:

$$2 \times \Pr(T \geq t) = 2 \times \Pr(T \geq 2.1655) = 0.0382$$

- Use summary() to get estimates and p values

```
> model1 = lm(y~x)
> summary(model1)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-5.939 -1.783 -0.228  1.506  8.157

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.82963    1.76845   2.166   0.0382 *
x           0.90364    0.05012  18.030  <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.23 on 31 degrees of freedom
Multiple R-squared:  0.9129,    Adjusted R-squared:  0.9101
F-statistic: 325.1 on 1 and 31 DF,  p-value: < 2.2e-16
```

- Use confint() to get CI

```
confint(model1, level=0.95)
```

	2.5%	97.5%
(Intercept)	0.2228611	7.436405
waste$x	0.8014249	1.005862

Point estimate

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \qquad E[\hat{Y}_0] = \beta_0 + \beta_1 x_0$$

$$Var[\hat{Y}_0] = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}}\right)$$

- Confidence interval

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{XX}}}$$

For the paired t-test, we test H0: ur=ub against H1: ur>ub

p<0.05, Therefore, there is sufficient evidence to conclude that the use of radial tires instead of regular belted tires improves fuel economy. Assume that the paired differences approx normal

For the Wilcoxon signed-rank test, we test, H0: dist radial = dist belted, H1: dist radial > dist belt

p<0.05 Therefore, there is sufficient evidence to conclude that the use of radial tires instead of regular belted tires improves fuel economy.

Assume that if there is a difference between population distributions, the difference is only in the location parameter, and there is a pairing of the data.

c) For the Wilcoxon rank-sum test, we test,

H0: the distribution of radial tire = the distribution of belted tire against H1: the distribution of radial tire > the distribution of belted tire .

p>0.05, Therefore, there is no sufficient evidence to conclude H1

Assume that if there is a difference between population distributions, the difference is only in the location parameter, and the 2 samples are independent (not paired).

d) We may do a boxplot or a qqplot on the paired differences to check whether a normal distribution can be assumed. The Shapiro-Wilk normality test may also be used.

CI : $(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) \pm t_{\alpha/2} \times S\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$

Given new x value, replace x0 with x new

On the other hand, suppose we want to predict what a new observation would look like for the percent reduction in chemical oxygen demand (y), when the percentage reduction in total solids (x) is 20%.

Then we may construct a 95% confidence interval for Y_{new} when x = 20,

We know $\hat{Y} = 3.8296 + 0.9036X$ with $SSE = 323.3273$

So, $\hat{Y}_{new} = 3.8296 + 0.9036 \times 20 = 21.902$

We also know $S_{xx} = 4152.182$, $\bar{x} = 33.45455$ and $s^2 = 10.4299$, $t_{0.025,31} = 2.0395$

A 95% CI for Y_{new} when $x = 20$ is

$$(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) \pm t_{\alpha/2} \times S\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$

$$= \hat{Y}_{new} \pm t_{0.025,31} \times \sqrt{10.4299} \times \sqrt{1 + \frac{1}{33} + \frac{(20 - 33.45455)^2}{4152.182}}$$

$$= 21.902 \pm 2.0395 \times 3.3467 = (15.08, 28.73)$$

- Use predict()

```
newx <- data.frame(x=c(20,25,30))
# change names so new data and old have the same names
x <- waste$x
y <- waste$y
predm <- predict(lm(y~x), newx, level =0.95, interval = 'confidence')
cbind(newx,predm)
```

x	fit	lwr	upr
20	21.90250	20.11193	23.69307
25	26.42071	24.98491	27.85652
30	30.93893	29.73919	32.13867

Interval param = confidence (w.r.t average) / prediction (w.r.t a single value)

```
> new = data.frame(costcube=c(100^3, 150^3))
> predm = predict(lm(sales~costcube), new, level = 0.90, interval =
"confidence")
> cbind(new,predm)
  costcube      fit      lwr      upr
1  1000000 642.2084 443.3315 841.0852
2  3375000 1552.7211 1386.7324 1718.7098
>
> predi = predict(lm(sales~costcube), new, level = 0.90, interval =
"prediction")
> cbind(new,predi)
  costcube      fit      lwr      upr
1  1000000 642.2084 16.64022 1267.777
2  3375000 1552.7211 936.81873 2168.624
```

d) From the output, the average sales amount of all companies that spend 100 (x $1,000) on advertising cost has an average of 642.2 (x $1,000), with a 90% confidence interval of (443.3, 841.1) thousand dollars.

e) To spend 150 (x $1,000) on advertising this year, the manager could expect the sales to be between 936.8 and 2168.6 thousand dollars, with 90% confidence.

a)

(i) What do you think the researcher was trying to determine?

-whether there was difference in spending between the two type of household

(i) State the null and alternative hypotheses. State any assumptions that you would need to perform the t-test, with clear definition of any notations

that you would use? Let u1 and u2 be mean of public and private respectively.

H0: u1=u2. H1: u1 not equal u2. Assume the amount spent is approximately normal and independent to perform t-test.

(ii) Would you assume equal or unequal variances for the two datasets? Why?

Based on var.test(), we would assume 2 var to be equal as the F test shows that pvalue is 0.627 and we do not reject equal variance claim at 0.05 level sig

(iv) At 0.05 level of significance, would you reject the null hypothesis? Why?

No as pvalue =0.05341 is greater than 0.05

(v) What would be your conclusion after the t-test? Explain your answer.

There is not enough evidence to show a difference in the means

b) Under the tests based on In_Public and In_Private,

(i) What do you think the researcher was trying to determine?

Whether there is difference bet log-transformed amt spend between 2 neighb.

(ii) State the null and alternative hypotheses. State any assumptions that you would need to perform the t-test, with clear definition of any notations

- Ua and Ub be mean of log-transformed amount spent H0: Ua=Ub, H1: Ua! Ub

Assume that log-transformed amount spent are approximately normally dist.

c) In The analyses above, which do you think is better in determining the difference in amount spent on a child's tuition? Justify your answer.

-In the analyses above, I think the t-test with log-transformed data is better in determining the difference in amount spent on a child's tuition. This is because the boxplots of the data show a right-skewed distribution, and a log-transformation would bring the data closer to normal distribution, which is required for t-test.

a) Yes, from the scatter plots, there is a strong linear correlation between body weight and heart weight. Yes, it is appropriate to use body weight as independent variable and heart weight as dependent variable in a linear regression analysis because it has quite a higher correlation coefficient 0.7186, and the scatter plot seems to show a linear relationship.

b) The fitted linear regression model is: heart_wt = 8.7753 + 0.9457 * body_wt.

Yes, the body weight contributes significantly to the heart weight. This is because from the regression model, we see that the coefficient of body_wt is significantly different from zero, with p-value = 0.000527.

c) From the output, the average heart weight of all the animals that have 2.5 kg body weight is estimated to have an average of 11.14 grams heart weight, with a 90% confidence interval of (10.69, 11.59) grams.

d) An Animal that has 3.5 kg as body weight, is estimated to have 12.09 grams of heart weight, with a 90% confidence interval of (10.35, 13.82) grams.

e) The plots suggest that the distribution of the residuals has a symmetric distribution, and possibly shorter tails than a normal distribution. Other than that, a simple linear regression model is considered reasonable.

f) Itis not advisable to use this model to predict what would be the heart weight when the body weight is 10 kg, because 10 kg is way too far away from the observed range of the body weight (roughly 2 to 5.5 kg) in building this model.