Data Analysis (534)

Textbook: Modern Applied Statistics with S, 4th ed.
by Venables and Ripley
Office: WH 132
Office hours: M Tu, 3:00pm-4:00pm
Classroom: FA 245 9:40-10:40am
Homework due: Wednesday
Grading policy: 60% homework + 40% midterm+final.
Midterm: Oct. 18 (W)
Final: Dec. 13, W 5:40pm LN 1402
You can bring one page with R commands and formulas.
You can find R download site through Google.
The first week homework due on this Friday.

Chapter 0. Introduction.

Data analysis is to teach how to analyze data. Usual steps in data analysis:

1. For a random sample, *e.g.*, regression data,
   (X_i, Y_i), $i = 1, ..., n$, input them to a computer software, say R or S-plus.
2. Assume a proper probability model, say a parametric model
   $Y_i = \beta' X_i + \epsilon_i$, where $\epsilon \sim N(\alpha, \sigma^2)$;
   or a semiparametric model
   $Y_i = \beta' X_i + \epsilon_i$, where $\epsilon \sim F$, an unknown cumulative distribution function (cdf),
   or a non-parametric model
   $(X_i, Y_i) \sim F(x, y)$, where F is unknown.
3. Compute an estimate of (α, β, σ) if it is parametric,
   or an estimate of (α, β, F) if it is semi-parametric,
   or an estimate of F, if it is non-parametric.
4. Check whether the model assumption is valid.
5. If No, go to Step 2, otherwise, carry out the other statistics inferences, e.g.,
   testing statistical hypotheses,
   or constructing confidence intervals,
   or drawing inferences on some other parameters.

Example 1. An example how to hand in homework.

X_i: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
21 22 23 24 25 26 27 28 29 30
Y_i: 1.40 1.40 3.36 4.69 6.05 7.35 7.27 6.80 8.94
8.68 11.24 11.62 12.85 14.38 14.28 15.43 16.86 18.08 18.45
19.73 20.63 20.73 23.30 23.06 26.15 27.45 27.67 27.64 28.28 32.30

Suppose it is in a file called "data" in a directory /home/qyu/try in a PC.
cd /home/qyu/try
Two ways to work on R:

1. Write a program file, say ch0,
   R - -vanilla < ch0 # figure is in the file Rplots.pdf
   R - -vanilla < ch0 > output # all commands and output in the file called "output".
2. Open R in that directory directly by typing:
   R or click the icon of R on a laptop.

```
> library(MASS)
> sink("ch0.out") # put output in ch0.out file
> x=matrix(scan("data"), ncol=1, byrow=T)
> y=x[31:60]
> x=x[1:30]
> z=lm(y~x)
> summary(z)
> plot(x,y)   # scatter plot
> plot(fitted(z),studres(z))
> qqnorm(studres(z))
> qqline(studres(z))
> makepsfile = function() {
    ps.options(horizontal = F)
    ps.options(height=4.0, width=7.5)
    postscript("ch1.ps")
    par(mfrow =c(1,3))
    plot(x,y)
    plot(fitted(z),studres(z))
    qqnorm(studres(z))
    qqline(studres(z))
    dev.off()
    }
> makepsfile()
> sink() # close sink function
> rm(x,y)
> q()
```

The output is as follows.

Call:

lm(formula = y ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-1.3470	-0.5934	-0.1120	0.4434	2.1720

Coefficients:

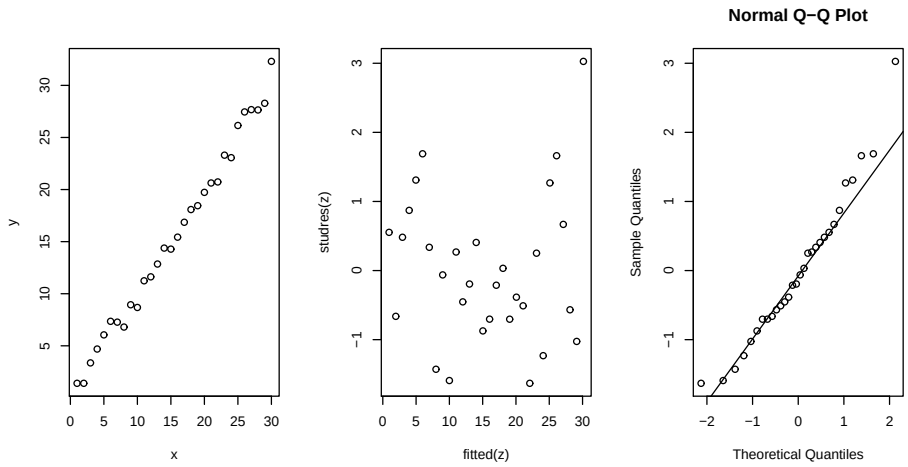
	Estimate	*Std. Error*	*t value*	$Pr(>\|t\|)$	
$(Intercept)$	-0.06299	0.32684	-0.193	0.849	—
x	1.00636	0.01841	54.663	$< 2e-16$	$***$

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.8728 on 28 degrees of freedom

Multiple R-squared: 0.9907, Adjusted R-squared: 0.9904

F-statistic: 2988 on 1 and 28 DF, p-value: < 2.2e-16



Write a report using Tex (or LaTex).

Edit a file called report.tex (see example),

need a postscript file: ch1.ps (which is created by makepsfile in page 2),

Some commands in the linex system:

tex report.tex (create report.dvi file)

xdvi report (view the file)

dvipdf report (create a pdf file)

dvips report -o report.ps (create a postscript file)

dvips -p 2 -l 3 report -o page2.ps

ps2pdf page2.ps (create a two-page pdf file)

pdf2ps report.pdf

For each homework, send me **3 files** by email:

1. junk.pdf — the formal report file (pdf file)
2. junk.tex – the Tex file preparing junk.pdf
3. junk —- a dos file collecting R commands used and output of R.

You need to orgainize them so that they are readable.

A brief manual for Latex is on my website: short-math-guide

A brief introduction of R is in prof. Xu's lecture note.

One can google the pdf file "An introduction to R".

Chapter 5. Univariate Statistics

5.1. Probability Distributions.

Let X be a random variable (rv).

Its cdf $F(t) = P\{X \le t\}$, **domain ?**

density function (df) $f(t) = \begin{cases} F'(t) & \text{if } X \text{ is continuous} \\ F(t) - F(t-) & \text{if } X \text{ is discrete,} \end{cases}$ **domain ?**

quartile $Q(u) = F^{-1}(u) = \min\{t : \ F(t) \ge u\}$ **domain ?**

Example 1. $X \sim$ Weibull distribution with cdf

$$F(x|\gamma,\tau) = 1 - \exp(-(x/\tau)^\gamma),\ x > 0$$

γ– shape, τ – scale,
pweibull(x,shape,scale) — $F(x)$,
qweibull(x,shape,scale) — $Q(x)$,
dweibull(x ,shape,scale) — $f(x)$,
rweibull(10 ,1 ,3) — 10 observations from Exp(3) with $E(X) = 3$.

Remark. The list of all distributions is given in Table 5.1.

Distributions	*R name*	*parameters*
beta	*beta*	$shape1, shape2$
binomial	*binom*	$size, prob$
Cauchy	*cauchy*	$location, scale$
$chi - square$	*chisq*	*df*
exponential	*exp*	*rate*
F	*f*	$df1, df2$
gamma	*gamma*	$shape, rate$
geometric	*geom*	*prob*
hypergeometric	*hyper*	m, n, k
$log - normal$	*lnorm*	$meanlog, sdlog$
logistic	*logis*	$location, scale$
negative binomial	*nbinom*	$size, prob$
normal	*norm*	$mean, sd$
Poisson	*pois*	*lambda*
T	*t*	*df*
uniform	*unif*	min, max
Weibull	*weibull*	$shape, scale$
Wilcox	*wilcox*	m, n

Example 1 (contitued).
R
> x=rweibull(10,1,5)
> round(x,2)
> mean(x)
Q: What will you see ?
QQplot: quantile-quantile plot.

1. Given data X_i, $i = 1$, ..., n.
2. Order them as $X_{(1)} \le \cdots \le X_{(n)}$.
3. Plot $(X_{(i)}, F^{-1}(\tilde{F}(X_{(i)})))$, where $\tilde{F}(X_{(i)}) = \frac{i-\frac{1}{2}}{n}$ (ppoints(x)), or $\frac{i}{n+1}$, or $\frac{i}{n}$ (ecdf).
   Since $\tilde{F}(t) \to F(t)$ w.p.1, we expect the qqplot is roughly a straight line.

Remark. If the assumption $X_i \sim F$ is correct
(and thus $\tilde{F} = F$ in the ideal situation),
then qqplot is plotting (X_i, X_i), $i = 1$, ..., n, as $F^{-1}(F(X_i)) = X_i$.
Thus the qqplot is expected to be a straight line roughly.

Example 2. Given X_1, ..., X_{100}, 100 observations in the file data_ex2.

Suppose they are from a Weibull distribution.

$F(x|\gamma,\tau) = 1 - \exp(-(x/\tau)^\gamma),\ x > 0$

Estimate F and $P\{X \in (1,2]\}$ and $E(X)$.

Solution: We first find the MLE of (γ,τ), that is,

a value of (γ,τ) that maximizes the joint density function

$\mathcal{L}(\gamma,\tau) = \prod_{i=1}^n f(X_i|\gamma,\tau)$, where $f(t) = F'(t)$, $t > 0$.

Carry out data analysis using **R codes:**

```
x=matrix(scan("data_ex2"), ncol=1, byrow=T)
summary(x)
y=fitdistr(x,"weibull") # compute MLE
y
summary(y)
pweibull(2,y$e[1],y$e[2])-pweibull(1,y$e[1],y$e[2]) # P(X ∈ (1,2])
(y$e[2])*gamma(1+1/y$e[1]) # E(X) = ∫ xf(x)dx = τΓ(1 + 1/γ))
```

Output:

```
> summary(x)
       V1
 Min.   :1.030
 1st Qu.:1.840
 Median :3.000
 Mean   :2.992
 3rd Qu.:4.070
 Max.   :4.970
> y
      shape       scale
  2.7761986   3.3746473
 (0.2257762) (0.1280903)
> summary(y)
          Length Class  Mode
estimate  2      -none- numeric
sd        2      -none- numeric
vcov      4      -none- numeric
loglik    1      -none- numeric # log likelihood
n         1      -none- numeric
```

Question: What is the use of summary(y) here ?

```
> y$estimate
   shape    scale
2.776199 3.374647
> y$e
   shape    scale
2.776199 3.374647
> y$v # y$vcov
```

$$\begin{pmatrix} & shape & scale \\ shape & 0.050974887 & 0.009118663 \\ scale & 0.009118663 & 0.016407135 \end{pmatrix}$$

```
> pweibull(2, 2.7761986 ,3.3746473 )-pweibull(1, 2.7761986 ,3.3746473 )) # P(X ∈ (1,2])
> pweibull(2,y$e[1],y$e[2])-pweibull(1,y$e[1],y$e[2])
```

[1] 0.1750694

> ((y$e[2])*gamma(1+1/y$e[1])) # E(X)

3.003995

Our estimates under the Weibull model are $\hat{\tau} = 3.4$ with $\hat{\sigma}_{\hat{\tau}} = 0.13$ and $\hat{\gamma} = 2.8$ with $\hat{\sigma}_{\hat{\gamma}} = 0.23$.

$\tilde{F}(t) = 1 - \exp(-(t/3.4)^{2.8})$, $t > 0$ and $P(X \in (1, 2]) \approx 0.175$.

$E(X) = \tau\Gamma(1 + 1/\gamma) \approx 3.0$.

Qiuestion:

1. Can the model be simplified ?

   e.g., $X \sim Exp(1)$? ($\tau = 1$ or $\gamma = 1$ as $F(x) = 1 - e^{-(\frac{x}{\tau})^{\gamma}}$, $x > 0$).

   If the model is valid, then it can be shown that the MLEs $\hat{\gamma}$ and $\hat{\tau}$ have approximately normal distributions, $N(\gamma, \hat{\sigma}^2_{\hat{\gamma}})$ and $N(\tau, \hat{\sigma}^2_{\hat{\tau}})$.

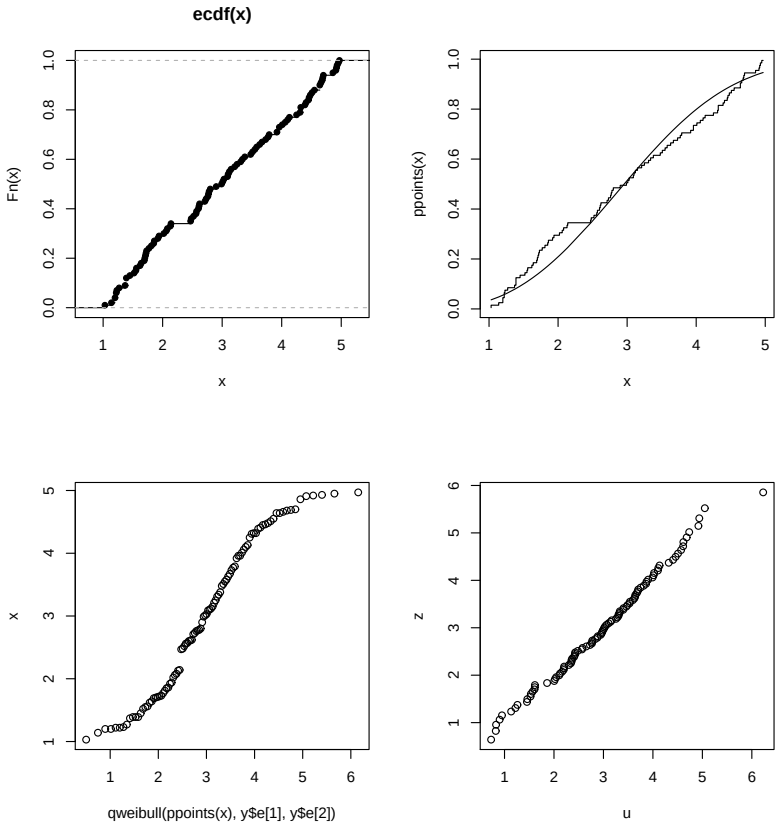
   H_0: $\gamma = 1$ v.s. H_1: $\gamma \neq 1$. Check $|\hat{\gamma} - 1| < 2\hat{\sigma}_{\hat{\gamma}}$?

   H_0: $\tau = 1$ v.s. H_1: $\tau \neq 1$. Check $|\hat{\tau} - 1| < 2\hat{\sigma}_{\hat{\tau}}$?

   Ans: It seems that the model cannot be simplified **Why ?**

2. Is the model assumption valid ?

   We can use the qqplot and ks.test.



Fig(1,1): empirical cdf, Fig(1,2): cdf of weibull v.s. edf

Fig(2,1): qqplot weibull, Fig(2,2): qqplot of 100 data from Weibull

Figure 5.1.

makepsfile = function() {

ps.options(horizontal = F)

ps.options(height=8.0, width=7.5)
postscript("ch1.2.ps")
par(mfrow =c(2,2))
plot(ecdf(x)) # edf
x=sort(x)
plot(x,ppoints(x),type="S") # new curve, edf
lines(x,pweibull(x,y$e[1],y$e[2])) # attach to the previous one
u=rweibull(100,y$e[1],y$e[2])
plot(qweibull(ppoints(x),y$e[1],y$e[2]),x) # or qqplot(u,x)
z=qweibull((1:100)/101,y$e[1],y$e[2])
qqplot(u,z)
dev.off()
}
makepsfile()

It seems that the Weibull assumption is not valid.

qqplot is quite subjective.

ks.test in R is a test.

Kolmogorov-Smirnov Goodness-of-Fit Test

Performs a one or two sample Kolmogorov -Smirnov test, which tests the relationship between two distributions.

One-sample. Suppose that X_1, ..., X_n are a random sample from F.

ks.test(x, "pweibull", shape, scale)

H_0: $F = F_o$ a Weibull distribution(shape,scale), verse

H_1: $F \neq F_o$, where F_o is given (together with the parameter).

The test statistic is $D = \sup\{|\tilde{F}(t) - F_o(t)| : t \in R\}$. P-value is given in R.

Remark. P-value= $P\{D > D_o\}$,

where D_o is the observed value of D for the given X_1, ..., X_n.

We reject H_0: $F = F_o(\cdot|\theta)$ assuming θ is known if P-value is small (< 0.05). θ = **?**

> ks.test(x, "pweibull", y$e[1],y$e[2])

One-sample Kolmogorov-Smirnov test

data: x

D = 0.0965, p-value = 0.3094

alternative hypothesis: two-sided

Question: What is our conclusion about the test ?

Does it agree with qqplot ?

Remark. In ks.test, θ is the true value. However, θ is estimated by its MLE here, this changed its true P-value.

Thus 0.05 needs to be adjusted to a bigger value approxmately 0.43 to be explained next.

One can find the critical value in D by empirical quantiles of 0.05

(See the simulation exercises in Examples 3, 4 and 5.)

Example 3. Generate data from U(1,5) with $n = 100$ or 1000.

Test against Weibull, Uniform and Uniform(1,5).

Question: What is the difference between the last two tests ?

Summarize the findings.

How to find the MLE ?

Weibull ?

Uniform ?

Uniform(1,5) ?

R codes:

```
fun3 = function(n) {
    x=runif(n,1,5)
    y=fitdistr(x,"weibull")
    a=ks.test(x, "pweibull", y$e[1], y$e[2])
    b=ks.test(x, "punif", min(x),max(x))
    c=ks.test(x, "punif", 1, 5)
    return(c(u=a$p.value, v=b$p, w=c$p))
}
n=100
fun3(n) # What is the output ?
What do you expect ?
```

u	v	w	
0.4267747	0.7190210	0.6058055	Are they expected ?
?	?	?	**Is it possible ?**

Repeat 1000 times:

```
m=1000
u=rep(0,m)
v=rep(0,m)
w=rep(0,m)
for(i in 1:m) {
    z=fun3(n)
    u[i]=as.numeric(z[1]<0.05)
    v[i]=as.numeric(z[2]<0.05)
    w[i]=as.numeric(z[3]<0.05)
}
mean(u)
[1] 0.013  # (Power or size of the test ϕ ?. Or an estimate ?)
mean(v)
[1] 0.043  # (Power or size of the test ? Or an estimate ?)
mean(w)
[1] 0.044  # (Power or size of the test ? Or an estimate ?)
n=1000 # Repeat but with larger sample size n size
u=rep(0,m)
v=rep(0,m)
w=rep(0,m)
for(i in 1:m) {
z=fun3(n)
u[i]=as.numeric(z[1]<0.05)
v[i]=as.numeric(z[2]<0.05)
w[i]=as.numeric(z[3]<0.05)
```

$E(\phi(\mathbf{U}))$ *or* $\mathrm{P}(H_1|H_o)$

}
>c(mean(u), mean(v), mean(w))
[1] 1 0.05 0.053 # Are they expected ?

Summary: Uniform(1,5) data test for

	$Weibull$		$Uniform$	$Uniform(1,5)$
n	$\hat{P}(H_0\|H_1)$		$\hat{P}(H_1\|H_0)$	$\hat{P}(H_1\|H_0)$
$ideal$	0		0.05	0.05
1000	0	**why ?**	0.05	0.053
100	0.987	**?**	0.043	0.044

Findings:

1. If n is very large, then it seems that ks.test works.
2. If n is moderate (n=100), $P(H_0|H_1)$ can be 99%, instead of $\leq 5\%$, this explains the discrepancy in Ex. 2.
3. If n is moderate, the level of the ks.test seems fine.

Remark. The P-value given in ks.test is under the assumption that n is very large. Otherwise, it is arbitrary.

Example 4. Generate 100 data from Weibull(1,0.2) with $n = 100$ or 1000.
Test against Weibull and Weibull(1,0.2). Summarize the findings.

R codes:

```
fun3 = function(n) {
x=rexp(n,5) # Why not rexp(n,0.2) ?
y=fitdistr(x,"weibull")
a=ks.test(x, "pweibull", y$e[1], y$e[2])
c=ks.test(x, "pweibull", 1, 0.2)
return(c(u=a$p.value, w=c$p))
}
n=100
fun3(n)
```

output:

u	w	
0.4647952	0.5927737	**Are they what you expect ?**

It seems OK, but we repeat 1000 times again.

```
m=1000
u=rep(0,m)
w=rep(0,m)
for(i in 1:m) {
z=fun3(n)
u[i]=as.numeric(z[1]<0.05)
w[i]=as.numeric(z[3]<0.05)
}
> c(mean(u) , mean(w))
[1] 0 0.045
```

What happens if n is larger ?

```
n=1000
u=rep(0,m)
w=rep(0,m)
for(i in 1:m) {
z=fun3(n)
u[i]=as.numeric(z[1]<0.05)
w[i]=as.numeric(z[3]<0.05)
}
> c(mean(u) , mean(w))
[1] 0 0.046
```

Summary: Weibull(1,0.2) data test for

	$Weibull$	$Weibull(1, 0.2)$
n	$\hat{P}(H_1\|H_0)$	$\hat{P}(H_1\|H_0)$
100	0	0.044
1000	0	0.046
ideal	0.05	0.05

Finding: $P(H_1|H_0) = 0$ if Weibull data test Weibull.

It is too small or the critical value for size 0.05 is too large.

Notice that for a test ϕ if $P(H_1|H_0) = 0$, then it is often $P(H_0|H_1) =$ **??**

Examples 3 and 4 suggest that ks.test is not reliable for $n = 100$.

Example 2 (continued). Examples 3 and 4 suggest that one needs to modify the ks.test for the case θ is replaced by the MLE.

The test statistic is $D = \sup_t |\tilde{F}(t) - F_o(t)|$ if $H_1 : \tilde{F}(t) \neq F_o(t)$.

How to find the empirical critical value of size 0.05 for ks.test:

```
x=matrix(scan("data_ex2"), ncol=1, byrow=T)
y=fitdistr(x,"weibull"))
b=ks.test(x, "pweibull", y$e[1], y$e[2])$s # What is b ?
```

Ans:

```
z=ks.test(x, "pweibull", y$e[1], y$e[2])
summary(z)
```

	$Length$	$Class$	$Mode$
statistic	1	$-none-$	*numeric*
p.value	1	$-none-$	*numeric*
alternative	1	$-none-$	*character*
method	1	$-none-$	*character*
data.name	1	$-none-$	*character*

```
for (i in 1:1000){
x=rweibull(100, y$e[1], y$e[2])
z=fitdistr(x,"weibull")
a=ks.test(x, "pweibull", z$e[1], z$e[2])
u[i]=a$s
}
> sort(u)[950] # what is this ?
[1] 0.08622978
> b
```

[1] 0.09650574 # D_o= 0.09650574

Q: Can we have conclusion now ?

> sum((u>b))/length(u) # length(u[u>b])/length(u)

 # what is this ?

[1] 0.024

What is the reasoning of this approach ?

1. First derive the test statistic value b from the data.
2. Pretend the true θ =MLE.

3 Repeat the ks.test m times with the same n and unknown θ.

4. It results i.i.d. ks.test statistic value D_i, $i = 1, ..., m$
5. SLLN ($\overline{\mathbf{1}(D > b)} \to P(D > b)$??).

What is conclusion for testing H_0: the data are from Weibull distribution in Ex. 2 ?

Question: Ideally, if we reject H_0 when ks.test\$$p < 0.05$, the size of the test is **??**

How to find a "ks.test\$<**??** for a size 0.05 for the data in Ex. 2 ?

Ans:

```
x=matrix(scan("data_ex2"), ncol=1, byrow=T)
y=fitdistr(x,"weibull"))
for (i in 1:10000){
    x=rweibull(100, y$e[1], y$e[2])
    z=fitdistr(x,"weibull")
    a=ks.test(x, "pweibull", z$e[1], z$e[2])
    u[i]=as.numeric(a$p<0.05) # mean(u)=0.00
    u[i]=as.numeric(a$p<0.43) # try to increase from 0.05 to achieve mean(u)≈ 0.05
}
mean(u)
```

[1] 0.0494 (≈ 0.05)

Since the ks.test and qqplots suggest that the data are not from a Weibull distribution. Then there are two choices:

1. empirical distribution function (edf),
2. other parametric distributions.

1. Use the edf to estimate F, $\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le t)$.

R codes:

```
mean(x)
sum((x>1& x<=2))/length(x)
```

Outcomes: $\hat{\mu} = \overline{X} = 2.99$ and $\hat{P}(X \in (1, 2]) = 0.29$

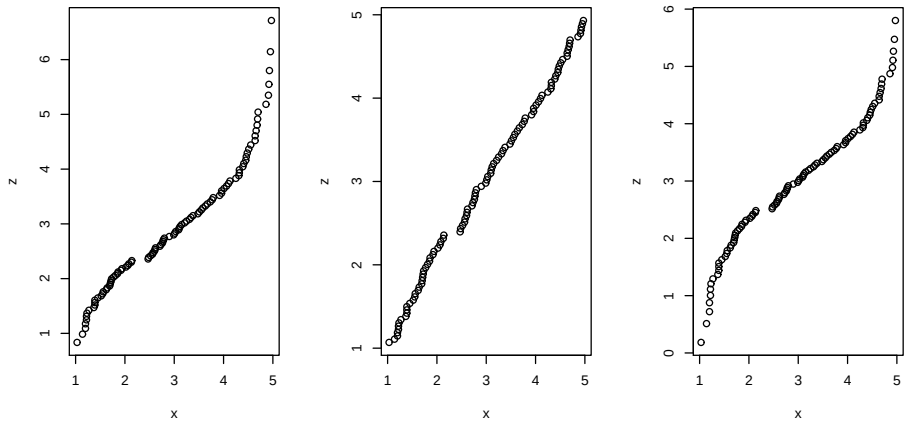
2. Try other parameteric cdf's,

Notice that in the program fitdistr(), distributions "beta", "cauchy", "chi-squared", "exponential", "f", "gamma", "geometric", "log-normal", "lognormal", "logistic", "negative binomial", "normal", "Poisson", "t" and "weibull" are recognised.

Only try gamma, uniform, normal as as follows.

```
par(mfrow =c(1,3))
y=fitdistr(x,"gamma")
n=length(x)
s=(1:n)/(n+1) # or s=(1:n)/n, s=ppoint(sort(x))
z=qgamma(s,y$e[1],y$e[2]) # or z=rgamma(n,y$e[1],y$e[2])
qqplot(x,z)
```

z=qunif(s,min(x),max(x))
qqplot(x,z)
qqnorm(x)



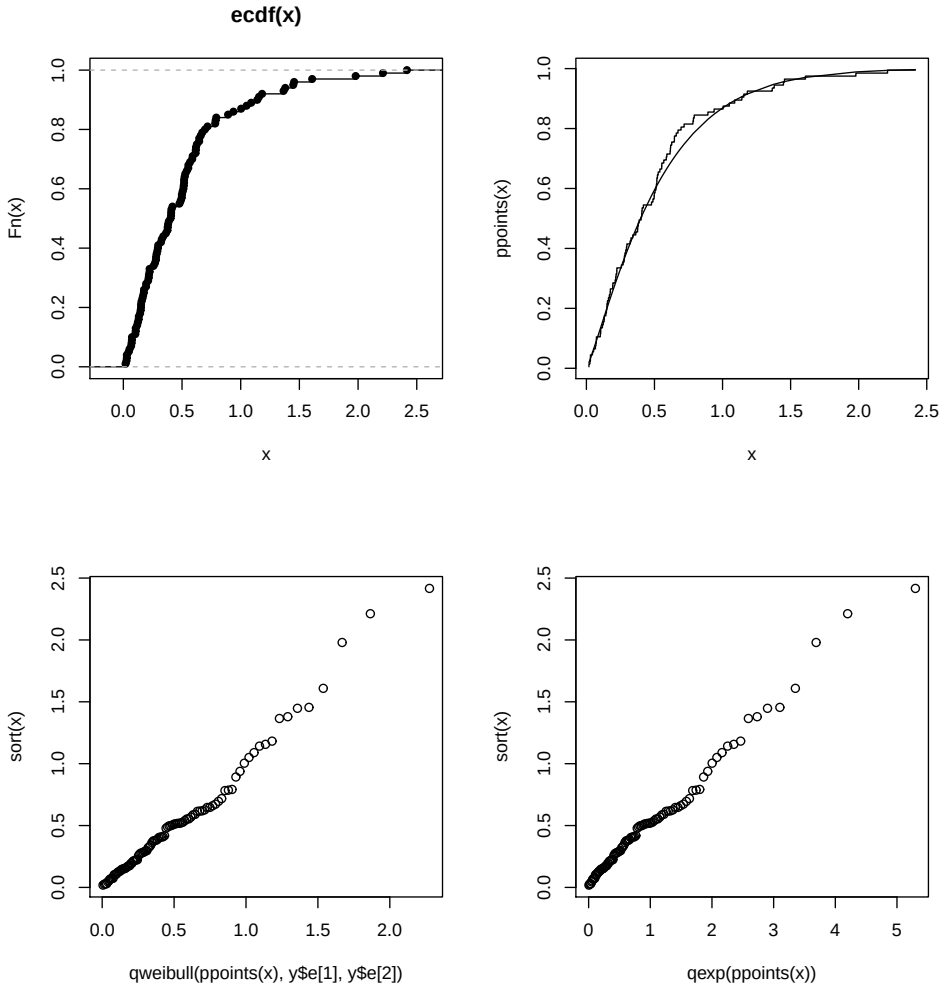
In view of the qqplots, we may test whether the data are from a uniform distribution,
> ks.test(x, distribution = "punif", min(x),max(x))
data: x and min(x)
D = 0.99, p-value = 0.2864
alternative hypothesis: two-sided
> ks.test(x, "punif", 1,5) # **Why (1,5) ?**
data: x
D = 0.055, p-value = 0.9228
alternative hypothesis: two-sided
What is the difference between these two ks.test ?
Which is more appropriate ?
Example 3 suggests that if $X \sim U(a,b)$, both work for $n = 100$.
Example 4 suggests that if $X \sim$ Weibull, MLE does not work for $n = 100$.
Can we assume $X \sim U(a,b)$?
$F(t) = \begin{cases} \frac{t-a}{b-a} & \text{if } t \in (a,b), \\ 1 & \text{if } t \geq b. \end{cases}$
then the MLE is $(\hat{a}, \hat{b}) = (\min_i X_i, \max_i X_i) = (1.03, 4.97)$,
as it maximizes the likelihood function $\mathcal{L}(a,b) = \prod_{i=1}^n \frac{1}{b-a} \mathbf{1}(X_i \in (a,b))$.
R codes:
(max(x)+min(x))/2
punif(2,min(x),max(x))-punif(1,min(x),max(x))
Or assume $X \sim U(1,5)$ based on ks.test(x,"unif", 1,5).
Final solution:
$\hat{F}$ is $U(1,5)$.
$\tilde{\mu} = 3$ and
$\tilde{P}(X \in (1,2]) = 0.25$
Comments:
edf=> $\hat{P}(X \in (1,2]) = 0.29$, with SE $\sqrt{\hat{P}(1-\hat{P})/n} \approx 0.045$ smaller difference.
Weibull MLE=> $\check{P}(X \in (1,2]) = 0.18$, differ$\approx$ half due to wrong assumption.

Homework: Compare the lengths of the CI of $P(X \in (1,2])$ due to the EDF and the CI under the Weibull distribution for the given data in Ex. 2. Check whether $\hat{P}(X \in (1,2])$ falls in the CI of $P(X \in (1,2])$ due to the EDF or under the Weibull assumption ? Then explain what it implicates.

Question: # of parameters using the EDF ? (non-parametric model)

# of parameters using the uniform distribution ? (parametric model)

Both models are correct, but there are more parameters in the edf.



Fig(1,1): empirical cdf, Fig(1,2): cdf of weibull v.s. edf

Fig(2,1): qqplot weibull, Fig(2,2): qqplot exp(1).

Figure 2. QQplots in Example 5

Example 5. Generate 100 data from Exp(1/2) (=Exp(μ)).

Now pretend that we do not know the underlying distribution of the data. Assume Weibull distribution. Estimate F and $P\{X \in (1,2]\}$ and $E(X)$.

Sol. **Simulation data:**

> x=rexp(100,2)

> mean(x)

[1] 0.5153382 # **rate =2 or scale=2 ?**

Now pretend we assume but do not really know the true distribution is

$F(t) = 1 - \exp(-(t/\tau)^{\gamma})$, $t > 0$.

The MLE is computed:
>fitdistr(x,"weibull")

shape	*scale*
1.16690389	0.54517822
(0.08866768)	(0.04934344)

We may test
H_o: $\gamma = 1$ v.s. H_1: $\gamma \neq 1$,
or
H_o: $\tau = 1$ v.s. H_1: $\tau \neq 1$.
That is, we check whether the data is from Exp(μ) or further Exp(1).
If $X \sim Weibull(\gamma, \tau)$, $\hat{\mu} = \hat{\tau}\Gamma(1 + 1/\hat{\gamma})$ with 2 parameters and SE by Delta method;
If $X \sim Exp(\mu)$, $\hat{\mu} = \overline{X}$ with 1 parameter and SE= $\hat{\sigma}_X/n$=?
If $X \sim Exp(1)$, $\hat{\mu} = 1$ with no parameter and SE=?
Conclusion ?

$\hat{\mu}_X = 0.52$,
$\hat{F}(t) = 1 - e^{-t/0.52}$, $t > 0$.
$\hat{P}(X \in (1, 2)) = e^{-1/0.52} - e^{-2/0.52}$.
Done ?

The qqplots (see Fgiure 2) appear linear.
It supports that the data are from the Weibull model or Exponential model.
> ks.test(x, "pexp", 1/mean(x)) # **Do we need to test weibull or others ?**
One-sample Kolmogorov-Smirnov test
data: x
D = 0.079181, p-value = 0.5575
Done ?

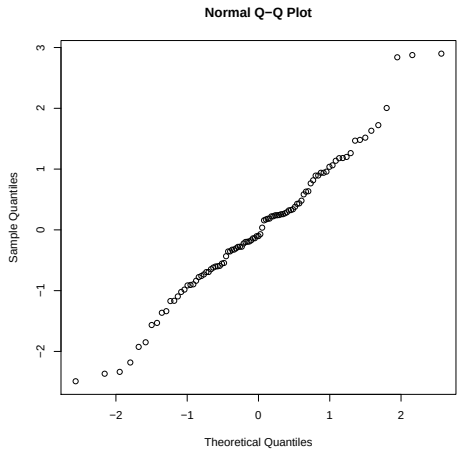
```
> n=100
> b=ks.test(x, "pexp", 1/mean(x))$s
> for (i in 1:m){
    z=rexp(n, 1/mean(x))
    u[i]=ks.test(z, "pexp", 1/mean(z))$s
    }
> sort(u)[950]
    [1] 0.1068376
```
Q: Is it possible that the simulation study suggests that the data do not fit the Weibull model ?

Example 6. (Prostate data).
library(MASS)
>library(faraway)
>prostate[96:98,]

	lcavol	*lweight*	*age*	*lbph*	*svi*	*lcp*	*gleason*	*pgg*45	*lpsa*
96	2.882564	3.7739	68	1.558145	1	1.55814	7	80	5.47751
97	3.471967	3.9750	68	0.438255	1	2.90417	7	20	5.58293
NA	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*

>y=lm(lpsa∼lweight,data=prostate)

We need to check whether $\epsilon \sim N(0, \sigma^2)$ in model $Y = \alpha + \beta x + \epsilon$
> x=y$resid
> qqnorm(x)



> sd(x)
[1] 1.079527

> ks.test(x, "pnorm", 0,1)
Output:
One-sample Kolmogorov-Smirnov test
data: x
D = 0.05809, p-value = 0.8798
alternative hypothesis: two-sided

Section 5.2. Tests on means

t.test, wilcox.test, binom.test.

1. t.test: (based on normal assumption).

Performs a one-sample, two-sample, or paired t-test, or a Welch modified two-sample t-test.

t.test(x, y=NULL, alternative=c("two.sided", "less", "greater"),
mu=0, paired=F, var.equal=T, conf.level=.95)

2. wilcox.test: (nonparametric)

Computes Wilcoxon rank sum test for two sample data (equivalent to the Mann-Whitney test) or the Wilcoxon signed rank test for paired or one sample data.

wilcox.test(x, y=NULL, alternative="two.sided", mu=0, paired=F,
exact=T, correct=T, conf.level=.95)

3. binom.test: (binomial distribution)

Test hypotheses about the parameter p in a binomial(n,p) model given x, the number of successes out of n trials.

binom.test(x, n, p=0.5, alternative="two.sided")

One sample, H_0: $\mu = \mu_0$, v.s. H_1: $\mu \neq \mu_0$ (or $>$, or $<$).
Two-sample, H_0: $\mu_X - \mu_Y = \mu_0$, v.s. H_1: $\mu_X - \mu_Y \neq \mu_0$ (or $>$, or $<$).
Remark. The small-sample t.test is a parameter inference, making use of $N(\mu, \sigma^2)$, whereas

wilcox.test is a non-parametric test, not assuming any parametric distributions, whereas the large sample t-test or binomial test can be either way.

Section 5.2.1. One sample.

t.test.

Test statistic $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$.

Assumption:

The random sample size is large $n > 30$, otherwise, $X_1, ..., X_n$ are i.i.d. from $N(\mu, \sigma^2)$.

wilcox.test:

Rank $X_i - \mu$'s by their absolute values.

Let S_n (S_p) be the sum of negative (positive) ranks.

Let S be the smallest among $|S_n|$ and $|S_p|$.

The Wilcoxon sign rank test statistic is $Z = \frac{S+\frac{1}{2}-\frac{1}{4}n(n+1)}{\sqrt{n(n+1)(2n+1)/24}}$

Assumptions: X_i's are i.i.d. from a symmetric distribution.

Example. Observations: $1, 3, 7$, H_o: $\mu = 4$. $S_n = ?$ $S_p = ?$ $S = ?$

Remark: If n is large, t.test is very close to z.test by CLT on $\overline{X}$ under the assumption: X_1, ..., X_n are i.i.d., provided $\sigma_X < \infty$.

Steps in one-sample test on mean μ:

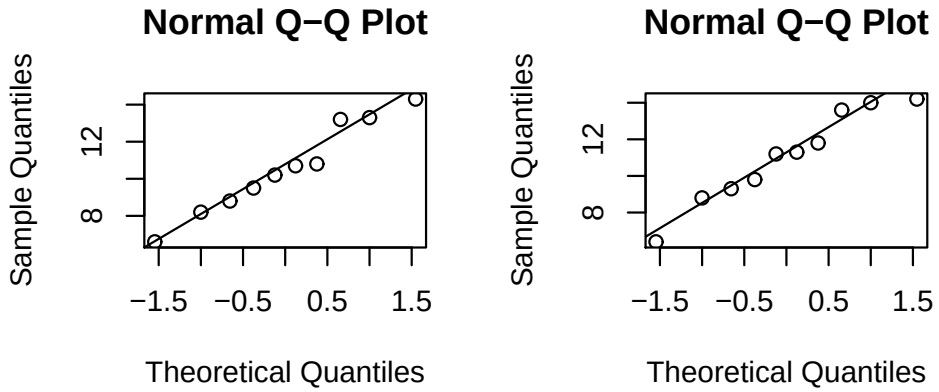
1. Input data;
2. qqnorm or ks.test to check normality;
3. If $X \sim N(\mu, \sigma^2)$ then t.test;
4. O.W. use hist() or stem() to check symmetry;
5. If it is symmetric, use wilcox.test.
6. O.W. let $Y = \sum_{i=1}^{n} \mathbf{1}(X_i > \mu)$, binom.test(Y,n,0.5)

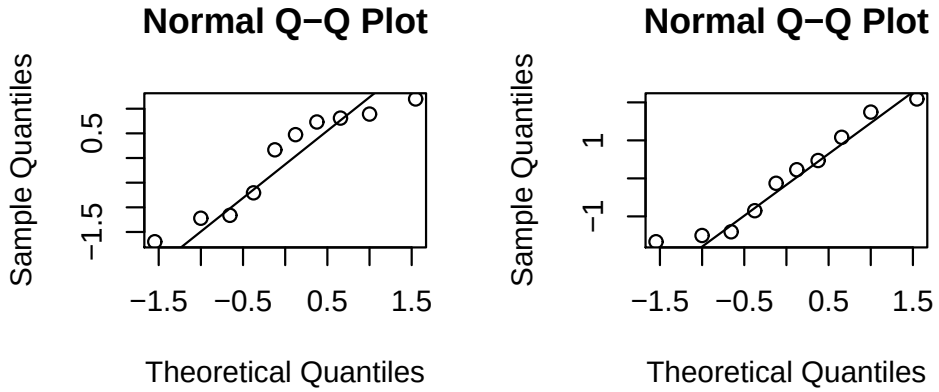
Example 1. Data on shoe wear (10 pairs).

shoe=list(A=c(13.2, 8.2, 10.2, 14.3, 10.7, 6.6, 9.5, 10.8, 8.8, 13.3),

B=c(14.0, 8.8, 11.2, 14.2, 11.8, 6.4, 9.8, 11.3, 9.3, 13.6))

Mean = 10 ?





$qqplot(A)$ $qqplot(B)$
$qqplot(rnorm(10))$ $qqplot(rnorm(10))$ **(why 10 ?)**

It seems from qqplot that the normal assumption is OK. No need of ks.test.

> (z=t.test(A,mu=10))

One Sample t-test

data: A

t = 0.722, df = 9, p-value = 0.4886

alternative hypothesis: true mean is not equal to 10

95 percent confidence interval:

8.805406 12.314594

sample estimates:

mean of x

10.56

> z$c

[1] 8.805406 12.314594

attr(,"conf.level")

[1] 0.95

Conclusion:

For testing H_0: $\mu = 10$. P-value > 0.4. Do not reject H_0.

Mean = 10

> stem(A) # **Do we need to do this ?**

06 | 6
08 | 285
10 | 278
12 | 23
14 | 3
The decimal point is at the "|".

What can we conclude ?

> wilcox.test(A,mu=10)

Wilcoxon signed rank test
data: A
V = 33, p-value = 0.625
alternative hypothesis: true location is not equal to 10

Comments: For this data set, both tests are valid, and they do not reject H_o.
But it is more appropriate to use the t.test. **Why ?**

5.2.2. Two-sample.

Data: X_1, ..., X_n, Y_1, ..., Y_m.

H_0: $\mu_X - \mu_Y = \mu_0$, v.s. H_1: $\mu_X - \mu_Y \neq \mu_0$

If both sample-sizes are very large a Z-test

$$\phi = \begin{cases} \mathbf{1}(\frac{|\overline{X}-\overline{Y}|}{\sqrt{S_X^2/n+S_Y^2/m}} > z_{\alpha/2}) & \text{if two samples are independent,} \\ t.test(x-y) & \text{if two samples are paired.} \end{cases}$$

Steps if n and m are small or moderate :

1. Check normal assumptions by qqnorm or ks.test.
   Use t.test if normal, o.w. use wilcox.test.
2. Determine independence by data feature (*e.g.* $n \neq m$?) or use cor.test.
   If dependent, use one-sample test with $Z_i = X_i - Y_i$. Otherwise, go on.
3. If normal, check whether $\sigma_X = \sigma_Y$ by var.test.

Questions:

X and Y are uncorrelated => $X \perp Y$?
X and Y are uncorrelated <= $X \perp Y$?
X and Y are correlated => $X \not\perp Y$?
X and Y are correlated <= $X \not\perp Y$?

t.test.

Test statistic $T = (\overline{X} - \overline{Y} - \mu_0)/\hat{\sigma}$, where
$\hat{\sigma}$ is an estimate of σ, depending on the assumption.

Assumptions:

1. $X_i \sim N(\mu_X, \sigma_X^2)$ and $Y_i \sim N(\mu_Y, \sigma_Y^2)$,
2. $\sigma_X = \sigma_Y$?
3. Are two samples dependent ?

cor.test.

cor.test(x,y,method="pearson", "kendall","spearman")
Given (X_i, Y_i), $i = 1$, ..., n, test for correlation ρ (**= ?**).
"pearson" test statistics:

$$T = \sqrt{n-2} * R/\sqrt{1-R^2}$$

where $R = S_{xy}/\sqrt{S_{xx}S_{yy}}$. $T \sim t_{n-2}$ if $(X, Y) \sim N(\mu, \Sigma)$.

"kendall" test statistics:

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

where $n_c = \sum_{i<j} \mathbf{1}((Y_i - Y_j)(X_i - X_j) > 0)$, the number of concordant

(i.e., numbers of $b = \frac{Y_i - Y_j}{X_i - X_j} > 0$ or $\begin{matrix} * & (X_i, Y_i) \\ * & (X_j, Y_j) \end{matrix}$), and

$n_d = \sum_{i<j} \mathbf{1}((Y_i - Y_j)(X_i - X_j) < 0)$, the number of discordant

(i.e., numbers of $b = \frac{Y_i - Y_j}{X_i - X_j} < 0$ or **??**).

Critical values for testing Kendall's tau is tabulated.

"spearman" test statistics:

$$\hat{\rho} = \frac{S_{rs}}{S_r S_s} = \frac{\sum_i r_i s_i - C}{\sqrt{\sum_i r_i^2 - C}\sqrt{\sum_i s_i^2 - C}}$$

where $C = n(n+1)^2/4$,

$r_i = rank$ of x_i among x_j's and

$s_i = rank$ of y_i among y_j's.

Critical values for testing Spearman's rho is tabulated.

Steps:

1. Input data,
2. qqnorm and qqline on X_is and Y_is separatel,
3. If normal assumption is valid use pearson,

otherwise, use kendal or spearman. (**Does it has anything to do with t.test ?**)

var.test

Performs an F test to compare variances of two independent samples from $N(\mu_i, \sigma_i^2)$'s.

var.test(x, y, alternative="two.sided", conf.level=.95)

H_0: $\sigma_X = \sigma_Y$.

Test statistics $F = \sqrt{S_X^2/S_Y^2}$

wilcox.test. Wilcoxon Rank Sum Tests for testing two means.

Data: X_1, ..., X_n, Y_1, ..., Y_m.

Assumptions: The X_i's and Y_j's are independent samples

H_o: $F_X(t) = F_Y(t - \mu)$

The test statistic is $W = \sum_{j=1}^{m} R_{n+j}$,

where R_{n+j} =rank(Y_j) among $X_i - \mu$'s and Y_j's.

Example 1 (continued). Data on shoe wear (10 pairs).

Which of them are appropriate ?

cor.test(x,y,alternative="two.sided",method="pearson")

var.test(x,y)

t.test(x,y,pair=T)

t.test(x,y)

t.test(x, y, alternative="two.sided", paired=F, var.equal=T)

wilcox.test(x,y)

wilcox.test(x-y)

Applying tests to this data set yields output as follows.

> cor.test(A,B,alternative="two.sided",method="pearson")

Pearson's product-moment correlation
data: A and B
t = 16.50071, df = 8, p-value = 1.831e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.9383049 0.9967172
sample estimates:
cor
0.9856358

Are A and B correlated ?

> var.test(A,B)

F test to compare two variances
data: A and B
F = 0.9485, num df = 9, denom df = 9, p-value = 0.9385
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.2355932 3.8186432
sample estimates:
ratio of variances
0.948497

Q: $\sigma_A^2 = \sigma_B^2$ **?** Yes, No, DNK.

> t.test(A,B)

Welch Two Sample t-test
data: A and B
t = -0.4318, df = 17.987, p-value = 0.671
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.815702 1.855702
sample estimates:
mean of x mean of y
10.56 11.04

Do we reject H_o **?** Yes, No, DNK.

> t.test(A,B,var.equal=T)

Two Sample t-test
data: A and B
t = -0.4318, df = 18, p-value = 0.671
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.815585 1.855585
sample estimates:
mean of x mean of y
10.56 11.04

Do we reject H_o **?** Yes, No, DNK.

> t.test(A,B,pair=T)

Paired t-test
data: A and B
t = -3.5602, df = 9, p-value = 0.006118

alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.7849953 -0.1750047
sample estimates:
mean of the differences
-0.48

Do we reject H_o ? Yes, No, DNK.

> wilcox.test(A,B)

Wilcoxon rank sum test with continuity correction
data: A and B
W = 42.5, p-value = 0.5966
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(A,B,pair=T)

Wilcoxon signed rank test with continuity correction
data: A and B
V = 3, p-value = 0.01437
alternative hypothesis: true location shift is not equal to 0

Conclusion:

It seems from qqplot that the normal assumption is OK.
cor.test gives $\rho = 0.98$ and P-value 0.00. X and Y are paired.
Thus the var.test is not valid, even though
it seems from var.test that the variances are equal (P-value= 0.94).
If we use correct test (paired t.test), P-value is 0.006 and
we reject H_0. That is, there is a difference in mean.
If we use the incorrect test (two sample test), P-value is 0.67
and we do not reject H_0.
The paired Wicoxon test gives P-value 0.014, which is not as siginificant as the paired t.test.

Example 2 (a simulation study).

Generate two independent samples from N(0,1) and N(0,25).
Test for equal means.

```
x=rnorm(10)
y=rnorm(10,0,5)
qqnorm(x)
qqline(x)
qqnorm(y)
qqline(y)                    # expect to reject H_o ? Yes, No, DNK
cor.test(x,y,method="pearson")          # expect to reject H_o ? Yes, No, DNK
var.test(x,y)                  # expect to reject H_o ? Yes, No, DNK
t.test(x,y,pair=T)                  # expect to reject H_o ? Yes, No, DNK
t.test(x,y)                  # expect to reject H_o ? Yes, No, DNK
t.test(x, y, alternative="two.sided", paired=F, var.equal=T) # What do you expect ?
```

Output

Pearson's product-moment correlation

data: x and y
t = 1.8239, df = 8, p-value = 0.1056
alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:
-0.1331101 0.8735067
sample estimates:
cor
0.5419356

F test to compare two variances

data: x and y
F = 0.0227, num df = 9, denom df = 9, p-value = 4.522e-06
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.005646828 0.091527345
sample estimates:
ratio of variances
0.0227341

Paired t-test

data: x and y
t = 0.2158, df = 9, p-value = 0.834

Welch Two Sample t-test

data: x and y
t = 0.1978, df = 9.409, p-value = 0.8474

Two Sample t-test

data: x and y
t = 0.1978, df = 18, p-value = 0.8454

What is your conclusion ?

Example 3 (a simulation study).
Generate two independent samples from N(0,1) and N(2.0,9)
Test for equal means.

```
x=rnorm(10)
y=rnorm(10,2.0,3)
qqnorm(x)
qqline(x)
qqnorm(y)
qqline(y)
cor.test(x,y,alternative="two.sided",method="pearson")
var.test(x,y)
t.test(x,y,pair=T)          # expect to reject H_o ? Yes, No, DNK
t.test(x,y)          # expect to reject H_o ? Yes, No, DNK
t.test(x, y, alternative="two.sided", paired=F, var.equal=T)
          # expect to reject H_o ? Yes, No, DNK
wilcox.test(x,y)          # expect to reject H_o ? Yes, No, DNK
wilcox.test(x-y)          # expect to reject H_o ? Yes, No, DNK
```

Q: Which of the 7 tests is valid ? (that is, the model assumptions is valid).
Q: Which of the last 5 tests is more appropriate ?
Output

Pearson's product-moment correlation

data: x and y
t = -1.3413, df = 8, p-value = 0.2167
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.8332983 0.2754580
sample estimates:
cor
-0.4284824

F test to compare two variances

data: x and y
F = 0.2261, num df = 9, denom df = 9, p-value = 0.03726
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.05615047 0.91012210
sample estimates:
ratio of variances
0.2260615

Paired t-test

data: x and y
t = -2.0923, df = 9, p-value = 0.06594
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-3.1563001 0.1231082
sample estimates:
mean of the differences
-1.516596

Welch Two Sample t-test

data: x and y
t = -2.4151, df = 12.871, p-value = 0.03136
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.874619 -0.158573
sample estimates:
mean of x mean of y
0.3581944 1.8747904

Two Sample t-test

data: x and y
t = -2.4151, df = 18, p-value = 0.02659
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.8359078 -0.1972842
sample estimates:
mean of x mean of y
0.3581944 1.8747904

Wilcoxon rank sum test

data: x and y
W = 27, p-value = 0.08921

alternative hypothesis: true location shift is not equal to 0

Wilcoxon signed rank test

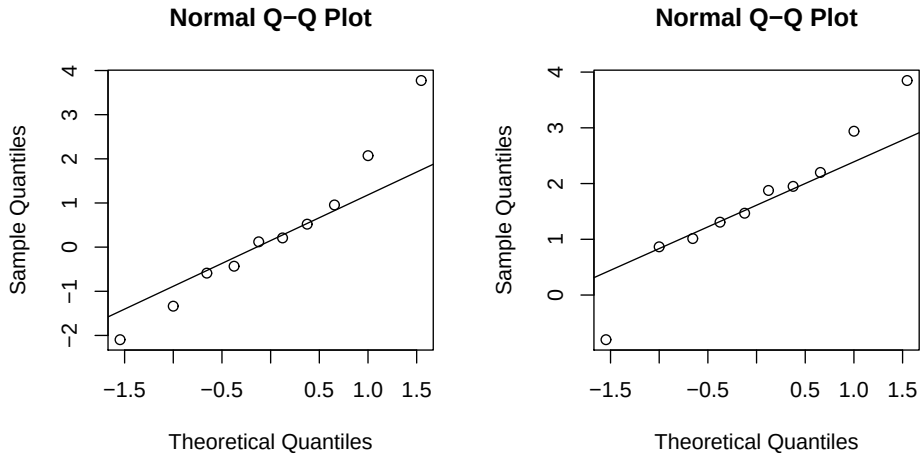
data: x - y

V = 10, p-value = 0.08398

alternative hypothesis: true location is not equal to 0

What is the conclusion ?

Example 4 (a simulation study). Generate two independent samples. Test for equal means.



Does it seem straight lines ?

What will you do if you are not sure ?

Can we use ks.test ?

It seems from qqplot that the normal assumption is not likely.

> cor.test(x,y,alternative="two.sided",method="kendall")

Kendall's rank correlation tau

Kendall's rank correlation tau

data: x and y

T = 15, p-value = 0.2164

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

-0.3333333

Try:

t.test(x,y,pair=T)

t.test(x,y)

t.test(x, y, alternative="two.sided", paired=F, var.equal=T)

wilcox.test(x,y)

wilcox.test(x-y)

Which of the previous tests is valid ?

Two possible answers:

a. DNK.

b. Based on QQ-plot, the last two.

Which of the previous tests is more appropriate ?

Output:

Paired t-test

data: x and y

t = -1.7041, df = 9, p-value = 0.1226

Welch Two Sample t-test

data: x and y

t = -2.0313, df = 16.632, p-value = 0.05852

Two Sample t-test

data: x and y

t = -2.0313, df = 18, p-value = 0.05725

Wilcoxon rank sum test

data: x and y

W = 23, p-value = 0.04326

Wilcoxon signed rank test

data: x - y

V = 13, p-value = 0.1602

Conclusion ?

a. We correctly reject H_0: equal mean if use wilcox(x,y).

b. We incorrectly do not reject H_0: equal mean if use wilcox(x-y).

Remark. The two samples are from double exponential+0, +2.

```
x=rexp(10)
z=c(-1,1)
u=sample(z,10,replace=T)
x=u*x
y=rexp(10)
z=c(-1,1)
u=sample(z,10,replace=T)
y=u*y+2
```

5.2.3 Tests on mean with multiple samples

1. **kruskal.test**. (Kruskal-Wallis (K-W) Rank Sum Test).

Performs a Kruskal-Wallis rank sum test on data following a one-way layout.

```
kruskal.test(y, groups)
```

Assumption: There are t (independent) samples, the ith sample has size n_i (> 1) and

$$(X_{i1}, ..., X_{in_i}),\ X_{ij} = \mu_i + \epsilon_{ij} \text{ and } F_{\epsilon_{ij}} = F_o.$$

H_o: all samples are from the same distribution, thus $\mu_1 = \cdots = \mu_t$,

H_1: $\mu_i \neq \mu_j$ for at least one pair.

This is a nonparametric alternative to one-way anova, which needs $N(\mu_i, \sigma^2)$

```
z=aov(y~ groups)
summary(z)
```

# same as
z= lm(y∼ groups)
anova(z)

The K-W test statistic is

$$T = \frac{(N-1)(S_t^2 - C)}{S_r^2 - C}.$$

Here $N = \sum_i n_i$.

Rank all N observations from 1 to N.

Let $r_{ij} = \text{rank}(X_{ij})$ and

s_i be the sum of the ranks in the ith sample, $i = 1, ..., t$.

Let $S_r^2 = \sum_{i,j} r_{ij}^2$, $S_t^2 = \sum_i (s_i/n_i)^2$ and $C = N(N+1)^2/4$.

T has approximately $\chi^2(t-1)$ distribution for moderate N.

Critical values for T is tabulated for small N.

Example 1. Total of 14 data from 3 groups.

>holl.y = c(2.9,3.0,2.5,2.6,3.2,3.8,2.7,4.0,2.4,2.8,3.4,3.7,2.2,2.0)

>holl.grps = factor(c(1,1,1,1,1,2,2,2,2,3,3,3,3,3),

> labels=c("Normal Subjects","Obstr. Airway Disease","Asbestosis"))

$t = 3$, $n_1 = n_3 = 5$, $n_2 = 4$. Test for equal means.

>kruskal.test(holl.y, holl.grps)

Kruskal-Wallis rank sum test

data: holl.y and holl.grps

Kruskal-Wallis chi-squared = 0.7714, df = 2, p-value = 0.68

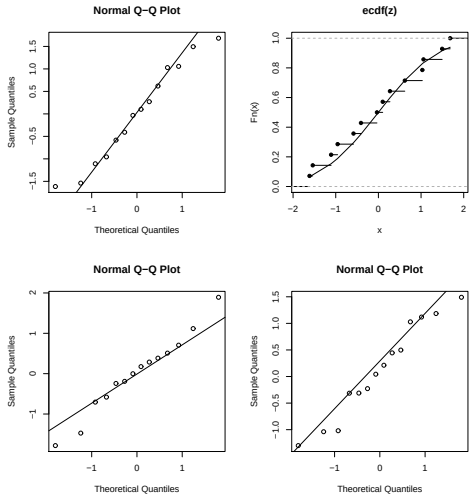
>z=lm(holl.y∼ holl.grps)

>anova(z) Analysis of Variance Table

Response: holl.y

	Df	$Sum\ Sq$	$Mean\ Sq$	$Fvalue$	$Pr(>F)$
holl.grps	2	0.4468	0.22339	0.5601	0.5866
Residuals	11	4.3875	0.39886		

>z=studres(z)



$qqnorm(z)$ $cedf(z)$
$qqnorm(x)$ $qqnorm(y)$, where x, y =rnorm(14)

Q: Does the normal assumption hold ?

Conlcusion of the test ?

Example 2 (Simulation study). Generate 4 random samples.
Test H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.
The output:
> kruskal.test(x,y)

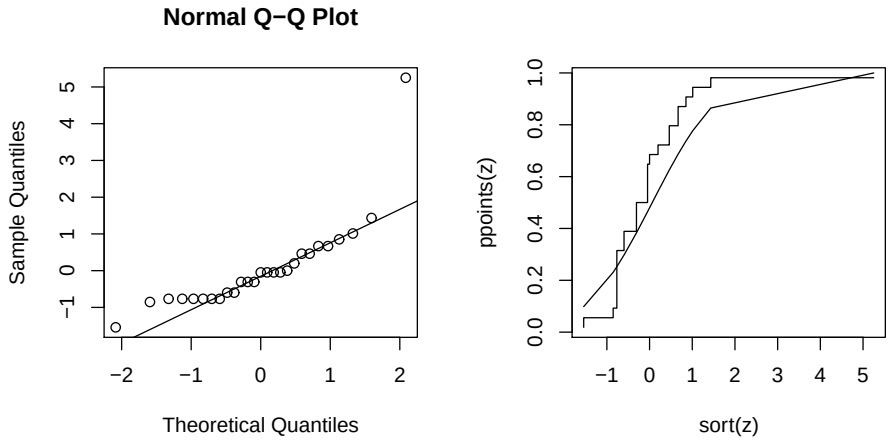
Kruskal-Wallis rank sum test

data: x and y

Kruskal-Wallis chi-squared = 8.0894, df = 3, p-value = 0.0442

> summary(aov(x∼y))

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$
y	3	13.78	4.594	2.254	0.109
$Residuals$	23	46.88	2.038		



Remark. In order to use aov, one needs to check the normal assumption.
Notice that there are obvious steps in qqplot.
It implies that there are many ties in residuals.
Q: Which test is more appropriate ?
Conclusion of the test ?

The 4 random samples are from $\mathcal{P}(\lambda_i)$,
with 4 different λ_i and 4 different sample sizes n_i.

```
n=c(3,4,5,15)
p=c(0.8,1,0.9,2)
x1=rpois(n[1],p[1])
x2=rpois(n[2],p[2])
x3=rpois(n[3],p[3])
x4=rpois(n[4],p[4])
x=c(x1,x2,x3,x4)
rm(x1,x2,x3,x4)
y=c(rep(1,n[1]),rep(2,n[2]),rep(3,n[3]),rep(4,n[4]))
y=as.factor(y)
z=lm(x∼ y)
z=studres(z)
qqnorm(z)
```

qqline(z)
plot(sort(z),ppoints(z),type="S") Or plot(ecdf(z)) #if normal assumption is not likely
kruskal.test(x,y) #if normal assumption seems likely
anova(z)
z= aov(x∼y)
summary(z)

2. **friedman.test** (Rank sum test).
friedman.test(B) # matrix $B_{b\times t}$ v.s. column factor (called treatment)

Remark. The test is a non-parametric alternative of two-way anova (parametric one).

Review of two-way anova: Suppose we have t treatments each applied to one of the t treatments in each of b blocks in a randomized block design. We denote by X_{ji} the response (observation) from treatment i in block j.

$$\begin{array}{c} \\ block\ 1 \\ \cdot \\ block\ b \end{array} \begin{array}{c} \begin{array}{ccc} treatment\ 1 & \cdots & treatment\ t \end{array} \\ \begin{pmatrix} X_{11} & \cdots & X_{1t} \\ \cdot & \cdots & \cdot \\ X_{b1} & \cdots & X_{bt} \end{pmatrix} \end{array} \stackrel{def}{=} B \quad (friedman.test(B)) \tag{1}$$

Assumption for two-way anova:
X_{ij}'s are i.i.d. from $N(\mu+\alpha_i+\beta_j,\sigma^2)$, $i=1$, ..., b, $j=1$, ..., t.
It is to test H_0: $\beta_1=\cdots=\beta_t$ or
H_0^*: $\alpha_1=\cdots=\alpha_b$.

R commands:
z=aov(y∼column+row)
summary(z)
anova(lm(y∼column+row) # present the same output
It gives two p-values.
The command lm(y∼column+row) means

$$\begin{aligned} y_{ij} =& \mu+\alpha_2\mathbf{1}_{(i=2)}+\cdots+\alpha_b\mathbf{1}_{(i=b)}+\beta_2\mathbf{1}_{(j=2)}+\cdots+\beta_t\mathbf{1}_{(j=t)}+\epsilon_{ij} \\ & (=\mu+\alpha_2\mathbf{1}_{(i=2)}+\beta_2\mathbf{1}_{(j=2)}+\epsilon_{ij} \text{ if } t=b=2). \end{aligned} \tag{2}$$

Recall the LSE is $\hat{\theta}=(X'X)^{-1}X'Y$, where $\theta=$?

If $b=t=2$, then

$$\begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix} \text{ is changed to } \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}}_{rank=3} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix} => Y=X\theta \text{ ?}$$

$\alpha_1=\beta_1=0$ in Eq. (2) is the default identifiability condition.
Other identifiability conditions:
$\mu=\alpha_1=0$, or

$\mu = \beta_1 = 0$, or
$\sum_i \alpha_i = 0 = \sum_j \beta_j$.

In order to use aov, one needs to check:

(1) the normal assumption and

(2) independent samples.

In friedman.test, H_o: no treatment effect difference, i.e., $\beta_1 = \cdots = \beta_t = 0$,

where $F_{X_{ij}}(x) = F(x - \alpha_i - \beta_j)\ \forall\ x$ and $\begin{pmatrix} X_{11} \\ \vdots \\ X_{b1} \end{pmatrix}, ..., \begin{pmatrix} X_{1t} \\ \vdots \\ X_{bt} \end{pmatrix}$ are independent.

Data in friedman.test are input as the matrix $B = \begin{pmatrix} X_{ij} \end{pmatrix}_{b \times t}$.

Thus "treatment" is the colume factor.

We replace the observations in each block by ranks 1 to t.

This ranking is carried out separately for each block.

The sum of the ranks is then obtained for each treatment, denoted by

$s_j = \sum_{i=1}^{b} r_{ij}$, $j = 1, ..., t$,

where r_{ij} denotes the rank (or mid-rank if there are ties) of X_{ij} within block i,

let $S_r^2 = \sum_{i,j} r_{ij}^2$ $(= bt(t+1)(2t+1)/6$ if there is no tie),

let $S_t^2 = \sum_j s_j^2/b$ and $C = bt(t+1)^2/4$,

$T = b(t-1)(S_t^2 - C)/(S_r^2 - C)$

has approximately $\chi^2(t-1)$ distribution, if b, t are not too small.

Q: What is the difference between the two assumptions ?

(1) $F_{ij}(x) = F(x - \alpha_i - \beta_j)\ \forall\ x$;

(2) $X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$, where ϵ_{ij} are i.i.d., with $E(\epsilon_{ij}) = 0$.

If X_{ij} has Cauchy distribution, which model is applicable ?

Remark. aov() can test both equal row effects and equal column effects in the same time, but friedman.test can only do column effects.

Example 3. Test for equal treatment effects. 12 data from 3 treatments and 4 subjects.

	$Trt1$	$Trt2$	$Trt3$
$Subject1$	0.73	0.48	0.51
$Subject2$	0.76	0.78	0.03
$Subject3$	0.46	0.87	0.39
$Subject4$	0.85	0.22	0.44

Input data:

```
treatment = factor(rep(c("Trt1", "Trt2", "Trt3"), each=4))
sub = factor(rep(c("Subject1", "Subject2", "Subject3", "Subject4"), 3))
y = c(0.73,0.76,0.46,0.85,0.48,0.78,0.87,0.22,0.51,0.03,0.39,0.44)
z=lm(y~ treatment+ sub)
z
summary(aov(y~ treatment+ sub))   # usual approach
qqnorm(studres(z))   # check assumptions
```

qqline(studres(z))
dim(y)=c(4,3)
v=sample(1:3,2)
cor.test(y[,v[1]],y[,v[2]], method="kendall") # **For aov() or friedman.test() ?**
friedman.test(y) # if aov() is not applicable).
kruskal.test(as.vector(y),treatment) # **Is y a vector or matrix ?**
Output:
> z

Call:

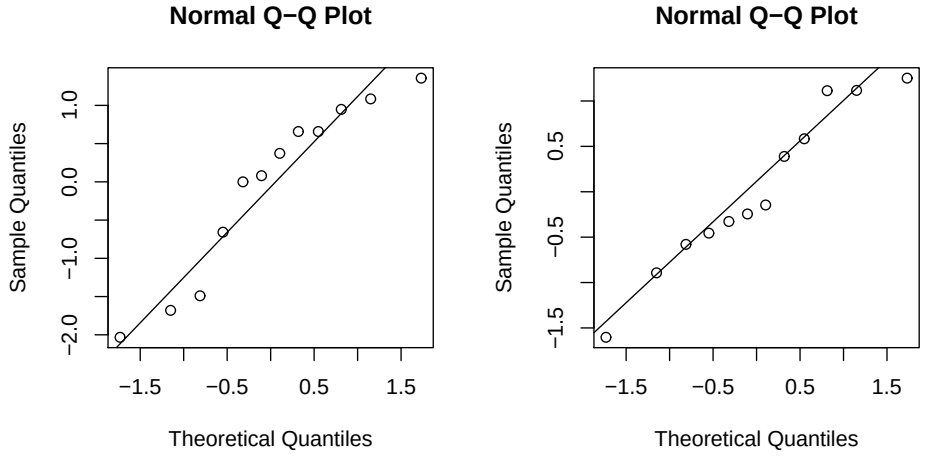
lm(formula = y ∼ treatment + sub)

Coefficients:

$(Intercept)$	$treatmentTrt2$	$treatmentTrt3$	$subSubject2$	$subSubject3$	$subSubject4$
$7.300e-01$	$-1.125e-01$	$-3.575e-01$	$-5.000e-02$	$1.408e-16$	$-7.000e-02$

> summary(aov(y∼ treatment+ sub))

	Df	$Sum\ Sq$	$MeanSq$	$F\ value$	$Pr(>F)$
$treatment$	2	0.2673	0.13366	1.691	0.262
sub	3	0.0114	0.00380	0.048	0.985
$Residuals$	6	0.4741	0.07902		



$qqnormdata \quad qqnorm(rnorm)$

Q: Do we need to continue ?
> cor.test(y[v[1],],y[v[2],], method="kendall") # **Is it a good choice here ?**

T = 9, p-value = 0.7194

> friedman.test(y)

Friedman rank sum test

data: y

Friedman chi-squared = 2, df = 2, p-value = 0.3679

> kruskal.test(as.vector(y),treatment)

Kruskal-Wallis rank sum test

data: y and treatment

Kruskal-Wallis chi-squared = 3.5769, df = 2, p-value = 0.1672

Conclusion: H_o ? H_1 ? α ? statistic ? ?

One-way anova: $Y_{ij} = \mu_j + \epsilon_{ij}$, $i = 1, ..., n_j$, $j = 1, ...,t$, where ϵ_{ij}'s are i.i.d..

$$e.g., \begin{matrix} Y_{11} & Y_{12} & Y_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ Y_{31} & Y_{32} \end{matrix}$$

Two-way anova: $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$, $i = 1, ..., b$, $j = 1, ...,t$, where ϵ_{ij}'s are i.i.d..

$$\begin{matrix} Y_{11} & Y_{12} & \cdots & Y_{1t} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{b1} & Y_{b2} & \cdots & Y_{bt} \end{matrix}$$

Example 4 (Simulation exercise). Generate 4 random samples from Exp(θ_i), with different θ_i and same sample sizes, say n. Form a $4 \times n$ matrix. Test for equal column effects, and then for equal row effects.

```
n=20
p=matrix(rep(c(1,3,2,5),n),4)
x=matrix(rexp(4*n),ncol=n)
x=x+p
gr= factor(as.vector(row(x)))
bl = factor(as.vector(col(x)))
# skip checking independence
friedman.test(x) # what to expect for P-value
friedman.test(t(x)) # what to expect for P-value
friedman.test(as.vector(x),bl,gr)
friedman.test(as.vector(x),gr,bl)
kruskal.test(as.vector(x),gr)
kruskal.test(as.vector(x),bl)
```

Output:

```
> friedman.test(x)
Friedman chi-squared = 18.257, df = 19, p-value = 0.5053
> friedman.test(t(x))
Friedman chi-squared = 50.22, df = 3, p-value = 7.172e-11
> friedman.test(as.vector(x),bl,gr)
Friedman chi-squared = 18.257, df = 19, p-value = 0.5053
> friedman.test(as.vector(x),gr,bl)
Friedman chi-squared = 50.22, df = 3, p-value = 7.172e-11
> kruskal.test(x,gr)
Kruskal-Wallis chi-squared = 59.436, df = 3, p-value = 7.757e-13
> kruskal.test(as.vector(x),bl)
Kruskal-Wallis chi-squared = 5.0241, df = 19, p-value = 0.9994
```

3. **prop.test** (Proportions Tests).

Compare proportions against hypothesized values p (a vector).
Alternately, tests whether underlying proportions are equal.
Suppose that X_i, $i = 1, ..., k$, are independent and $X_i \sim bin(n_i, p_i)$
prop.test(x, n, p, alternative="two.sided", conf.level=.95, correct=T)

x, n, p are $k \times 1$ vectors.

H_o: p_i's are as given.

The test statistic is $T = U'U$, approximately $\chi^2(k)$,
where $U' = (U_1, ..., U_k)$ and $U_i = \frac{X_i - n_i p_{io}}{\sqrt{n_i p_{io}(1-p_{io})}}$.
prop.test(x, n) for H_0: $p_1 = \cdots = p_k$.

```
> n=c( 18,22,19,23,15)
> x=c( 16,10, 9, 9 ,9)
> p=c(0.9,0.5,0.5,0.5,0.5)
> prop.test(x,n,p)
    5-sample test for given proportions without continuity correction
    X-squared = 1.9461, df = 5, p-value = 0.8566
    alternative hypothesis: two.sided
    null values:
    prop1 prop2 prop3 prop4 prop5
      0.9   0.5   0.5   0.5   0.5
    sample estimates:
    prop1     prop2     prop3     prop4     prop5
0.8888889 0.4545455 0.4736842 0.3913043 0.6000000

> prop.test(x,n)
    5-sample test for equality of proportions without continuity correction
    X-squared = 12.079, df = 4, p-value = 0.01677
```

Q: Consider testing problem of a mean or median μ, H_o: $\mu = 0$ with $n = 6$ in three cases: (1) $N(\mu, 1)$, (2) $U(a, b)$, (3) Cauchy Distribution.

1. If the sample is from $N(0, 1)$ what is the size of the test if we reject with p.value=0.05 using t.test? 0.05 ?
2. If the sample is from $N(0, 1)$ what is the size of the test if we reject with p.value=0.05 using wilcox.test? 0.05 ?
3. If the sample is from $N(1, 1)$ what is the size of the test if we reject with p.value=0.05 using t.test? 0.05 ?
4. If the sample is from $N(1, 1)$ what is the size of the test if we reject with p.value=0.05 using wilcox.test? 0.05 ?
5. If the sample is from $U(-1, 1)$ what is the size of the test if we reject with p.value=0.05 using t.test? 0.05 ?
6. If the sample is from $U(-1, 1)$ what is the size of the test if we reject with p.value=0.05 using wilcox.test? 0.05 ?
7. If the sample is from $U(0, 1)$ what is the size of the test if we reject with p.value=0.05 using t.test? 0.05 ?
8. If the sample is from $U(0, 1)$ what is the size of the test if we reject with p.value=0.05 using wilcox.test? 0.05 ?
9. How about Cauchy distribution ? Difference between it and U(a,b) ?

If one is not sure the distribution, one can use wilcox.test.
Otherwise, the size of the test is not what you selected.
If one is sure of normal distribution, both tests can be used, as they have the same level.
However, t.test is more powerful.
How to find the size of the test $P(H_1|H_0)$?
$P(H_1|H_0) = \int \cdots \int_{RR} \prod_{i=1}^{n} (f(x_i)dx_i)$,
where $RR = \{|T| > c\}$ and $T = \frac{\overline{X} - \mu_o}{S/\sqrt{n}}$.
n=6

```
m=1000
fun=function(n){
x=runif(n,-1,1)
a=t.test(x)$p.value
return(a)
}
u=matrix(rep(0,m*n),m)
s=apply(u,1,fun)
mean((s<0.05)) # 0.06
```

5.3 Some classical tests in R

5.3.1. Tests for contingency tables:

$r \times c$ contingency table $\left(N_{ij}\right)_{r\times c}$ for testing H_o: row factor $\perp$ column factor,

where N_{ij} are counts.

$r \times c \times l$ contingency table: $\left(N_{ijk}\right)_{r\times c\times l}$

Example 1. Consider a special case of 2×2 contingency table.

Let A – a randomly selected person is a male,

B – a randomly selected person is a democrat.

H_o: $A \perp B$.

<=> $P(AB) = P(A)P(B)$

<=> $P(AB^c) = P(A)P(B^c)$ <=> $P(A^cB^c) = P(A^c)P(B^c)$ <=> $P(A^cB) = P(A^c)P(B)$.

$n = 9$ people are sampled. Data are $\begin{pmatrix} x \\ y \end{pmatrix}$:

```
> x = factor(c(1,1,2,1,2,1,1,2,2), labels=c("male", "female")) # old days
> y = factor(c(1,1,1,2,1,2,2,1,1), labels=c("democrat", "none-democrat"))
> table(x,y)
```

	y	
x	$democrat$	$none-democrat$
$male$	2	3
$female$	4	0

is called a 2×2 contingency table.

One tests H_o base on the data in the form of contingency table.

(1) $r \times c$ **tables**

Original data: Data: X_1, ..., X_n, where

$X_i \in \{(a, b) : \ a \in \{a_1, ..., a_r\}, b \in \{b_1, ..., b_c\}\}$.

$r \times c$ contingency table: $\begin{array}{c|ccc} & b_1 & \cdots & b_c \\ a_1 & N_{11} & \cdots & N_{1c} \\ \vdots & \vdots & \cdots & \vdots \\ a_r & N_{r1} & \cdots & N_{rc} \end{array}$, leads to probability table $\begin{array}{ccc} p_{11} & \cdots & p_{1c} \\ \vdots & \cdots & \vdots \\ p_{r1} & \cdots & p_{rc} \end{array}$,

where $N_{ij} = \sum_{k=1}^{n} \mathbf{1}_{(X_k=(a_i,b_j))}$, $\sum_{ij} p_{ij} = 1$ and $p_{ij} \geq 0$.

Test H_o: The column and row fatcotrs are independent,

that is, $p_{ij} = p_{i\cdot}p_{\cdot j}$ $\forall$ (i, j), where $p_{i\cdot} = \sum_j p_{ij}$ and $p_{\cdot j} = \sum_i p_{ij}$.

Three tests will be introduced:

fisher.test. chisq.test. mcnemar.test.

(2) $r \times c \times l$ tables

Original data: X_1, ..., X_n.

$X_i \in \{(a, b, w) : \ a \in \{a_1, ..., a_r\}, b \in \{b_1, ..., b_c\}, w \in \{w_1, ..., w_l\}\}$.

Let $N_{ijk} = \sum_{h=1}^{n} \mathbf{1}_{(X_h=(a_i,b_j,w_k))}$,

Array: $(N_{ijk})_{r\times c\times l}$.

mantelhaen.test will be introduced.

1. **fisher.test.** Performs a Fisher's exact test on a two-dimensional contingency table.

e.g., 2×2 table.

	B	B^c	sum
A	N_{11}	N_{12}	r_1
A^c	N_{21}	N_{22}	r_2
sum	c_1	c_2	n

Data: X_1, ..., X_n.

$X_i \in \{(A, B), (A^c, B), (A, B^c), (A^c, B^c)\}$.

N_{11}, N_{21}, N_{12}, N_{22} are numbers of the 4 types of X_i's.

H_0: the row and column factors are independent ($i.e., P(AB) = P(A)P(B)$).

The test statistic is

$$\phi = \mathbf{1}(N_{11} \geq q_1, \ or \ N_{12} \geq q_2)$$

where q_1 and q_2 are chosen from the hypergeometric tables to make

$$\sum_{s\leq q_1} \frac{\binom{c_1}{s}\binom{c_2}{r_1-s}}{\binom{n}{r_1}} \text{ and } \sum_{s\geq q_2} \frac{\binom{c_1}{s}\binom{c_2}{r_1-s}}{\binom{n}{r_1}}$$

each as close to $\alpha/2$ (level of the test) as possible, but not larger.

Remark. The P-value is exact, not an approximation.

Example 1. Is it true that gender $\perp$ policital affiliation ?

```
x = factor(c(1,1,2,1,2,1,1,2,2), labels=c("male", "female"))
y = factor(c(1,1,1,2,1,2,2,1,1), labels=c("democrat", "republican"))
fisher.test(x,y) # x and y are factors
x=table(x,y) # A second way
fisher.test(x) # x is a matrix of counts
```

```
> fisher.test(x,y)
    Fisher's Exact Test for Count Data
    data: x and y
    p-value = 0.1667
    alternative hypothesis: true odds ratio is not equal to 1
    95 percent confidence interval:
    0.00000 2.64606
```

Ans: ??

2. chisq.test. (Pearson's Chi-square Test for Count Data).

Performs a Pearson's chi-square test on a two-dimensional contingency table.

(A large sample test for independence of $r \times c$ contingency table).

H_o, the row and column effects are independent.

That is $\mathrm{P}\{C = c_j, R = r_i\} = P\{C = c_j\}P\{R = r_i\} \ \forall \ (i, j)$.

C =Column factor and R = row factor.

Test statistic is

$$T = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

where n_{ij} is the count in cell (i,j),

e_{ij} is expected n_{ij}

(in defalt, $e_{ij} = n_{i+}n_{+,j}/n$) and $n = \sum_{i,j} n_{ij}$.

T is approximately $\chi^2_{(c-1)(r-1)}$ under H_o.

df = df in Θ - df in Θ_o (under H_o)

$(rc - 1 - ((r-1) + (c-1)) = (r-1)(c-1))$

There are various functions of chisq.test:

x=c(762, 327, 468)

y= c(484, 239, 477)

Case A. H_o : independent of 3×2 contingency table

	A	B
a	762	484
b	327	239
c	468	477

$n = 762 + \cdots + 477$

(z=matrix(c(x,y),3)) # matrix(c(x,y),nrow=3)

chisq.test(z)

Case B. 762 is treated as a factor rather than # in cell (1,1) as in case A.

H_o : independent of 3×3 contingency table

$n = 3$,

(z=table(x,y))

chisq.test(z)

chisq.test(x,y)

	$A =$ "762"	$B =$ "327"	$C =$ "468"
$a =$ "484"	1	0	0
$b =$ "239"	0	1	0
$c =$ "477"	0	0	1

Case C. test equal probabilities of the 6 elements in $c(x, y)$. H_0: $p_1 = \cdots = p_6$.

chisq.test(c(x,y))

Case D. test for (#x[1]:#x[2]:#x[3]) =(#y[1]:#y[2]:#y[3])

chisq.test(x, p = y, rescale.p = TRUE) # if y is not a probability vector

chisq.test(x, p = y/sum(y)) # same as above

Example 2. A 3×3 table corresponds to X_1, ..., $X_{12} \in \mathcal{R}^2$.

y=factor(c(2,2,2,3,3,3,2,1,1,2,1,2),label=c("A","B","C"))

z=factor(c("a","b","a","b","c","c","c","a","a","a","a","b"))

table(z,y) #

	A	B	C
a	3	3	0
b	0	2	1
c	0	1	2

chisq.test(z,y) P-value 0.13,

fisher.test(z,y) P-value 0.24

$1 \times m$ contingency table application.

Suppose that n_i is the count of observations fall in cell i, with expected frequency np_i, $i = 1, ..., m$ and $n = \sum_i n_i$. Pearson's χ^2 Goodness-of-fit statistics

$$T = \sum_{i=1}^{m} (n_i - n\hat{p}_i)^2/(n\hat{p}_i)$$

Two applications
First application:

H_o: $p_i = p_i(\theta)$ where $\theta \in \Omega_o$,

where $\hat{p}_i$ is the MLE of p_i $(= p_i(\theta))$. $T \sim \chi^2(m-k)$ asymptotically, where k is the degrees of freedom on θ or the dimension of Θ_o.

df = df in Θ - df in Θ_o (under H_o see Case C).

If p_i does not depend on some θ, $k = 0$ (see Case D).

An alternative application.

Data: independent X_1, ..., X_n with distribution F.

H_0: $F = F_0$, where F_0 is known, except for some parameters.

Divide the range into a grid of m cells.

Let n_i be the count of observations fall in cell i,

Proceed as before.

Remark. This is an alternative to ks.test().

Example 3.

(x = runif(100,0,4))
breaks = quantile(x)
y=fitdistr(x,"weibull")
z=pweibull(breaks, y$e[1], y$e[2])
(u=z[2:5]-z[1:4])
u=c(z[1],u,1-z[5])
(x=c(0,25,25,25,25,0))
chisq.test(x,p=u)
P-value < 0.01.

3. **mantelhaen.test.** Performs a Mantel-Haenszel chi-square test on a three-dimensional contingency table.

Data: independent X_1, ..., $X_n \in \left\{(x_1, x_2, x_3), x_h \in A_h, h \in \{1, 2, 3\}\right\}$,

where A_h are sets of sizes r, c and i (Row, Column and Item). *e.g*

X_1, ,..., X_4 are input by

$\vec{x}_1$=factor(c(1,2,1,2), labels=c("NoResponse","Response")),
$\vec{x}_2$=factor(c(1,2,2,1), labels=c("Male", "Female")),
$\vec{x}_3$=factor(c(1,2,1,1), labels=c("Nodular", "Diffuse")).

where $r = c = i = 2$.

There are 3 factors, taking r, c and i values respectively.

Factor 1 takes values a_{11}, ..., a_{1r},
Factor 2 takes values a_{21}, ..., a_{2c},
Factor 3 takes values a_{31}, ..., a_{3i},

Contingency table:

$$\begin{array}{ccccc} & a_{21} & \cdots & a_{2c} \\ a_{11} & w_{111} & \cdots & w_{1c1} \\ \vdots & \cdots & \cdots & \vdots \\ \underbrace{a_{1r}}_{a_{31}} & \underbrace{w_{r11}} & \cdots & \underbrace{w_{rc1}} \end{array} \quad \cdots \quad \begin{array}{ccccc} & a_{21} & \cdots & a_{2c} \\ a_{11} & w_{11k} & \cdots & w_{1ci} \\ \vdots & \cdots & \cdots & \vdots \\ \underbrace{a_{1r}}_{a_{3i}} & \underbrace{w_{r1k}} & \cdots & \underbrace{w_{rci}} \end{array}$$

$n = \sum_{k,j,h} w_{kjh}$.

H_0: Conditional on $I = h$, Column factor $\perp$ row factor.

$\mathrm{P}\{R = a_i, C = b_j | I = h\} = P\{R = a_i | I = h\} P\{C = b_j | I = h\}$ for each h.

For example, suppose that we have a sequence of 2×2 tables from different age groups, obtained from independent observations $X_{ijh} = (x, y)$, $i = 1, 2$, $j = 1, 2$, $h = 1, ..., k$, where x and y are the indicator functions that the (i, j) person belongs to group R and C, respectively. Here $R \cap C$ may not be empty (e.g, democratic and artist), called cross-classified.

$$\begin{array}{ccccc} item\ 1 & C & C^c & & \\ R & w_{111} & w_{121} & n_{11} & \\ R^c & w_{211} & w_{221} & n_{12} & \\ & m_{11} & m_{12} & n_1 & \end{array} , \cdots , \begin{array}{ccccc} item\ k & C & C^c & & \\ R & w_{11k} & w_{12k} & n_{k1} & \\ R^c & w_{21k} & w_{22k} & n_{k2} & \\ & m_{k1} & m_{k2} & n_k & \end{array}$$

H_o: $p_{11} = p_{12}$, ..., $p_{k1} = p_{k2}$, where $p_{i1} = P(C|R, I = i)$ and $p_{i2} = P(C|R^c, I = i)$.
Is it the same as $P(RC|I = i) = P(R|I = i)P(C|I = i)$, $i = 1, ..., k$?

Test statistic is

$$MH = \frac{\sum_{j=1}^{k} (w_{11j} - E_0(w_{11j}))}{\sqrt{\sum_{j=1}^{k} Var_0(w_{11j})}}, \text{ where } MH^2 \sim \chi^2(1).$$

```
x=factor(rep(c(1,2,1,2),c(3,10,15,2)),labels=c("NoResponse","Response"))
y=factor(rep(c(1,2,1,2,1,2,1,2), c(1,2,4,6,12,3,1,1)), labels=c("Male", "Female"))
z=factor(rep(c(1,2), c(13,17)), labels=c("Nodular", "Diffuse"))
mantelhaen.test(x,y,z)
x=table(x,y,z)
mantelhaen.test(x) # same answer
```

Output:

```
Mantel-Haenszel chi-squared test without continuity correction
data: x and y and z
Mantel-Haenszel X-squared = 0.15182, df = 1, p-value = 0.6968
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
0.2064438 10.5283756
sample estimates:
common odds ratio
1.474286
```

How to generate simulation data for contingency table ?

```
x=mvrnorm(90,c(0,0),matrix(c(4,0.4,0.4,3),2,2))          # dimension of x ?
```

```
x=round(x/4)                    # What does x represent ? Factors or counts ?
fisher.test(x[,1],x[,2])
fisher.test(x)                  # Do we have the same answer ? Does it work ?
chisq.test(x[,1],x[,2])

n=30
x=rbinom(2*n,1,0.5)
dim(x)=c(2,n)
y=matrix(c(1,1,0,1),ncol=2)
for(i in 1:n)
x[,i]=x[,i]%*%y
fisher.test(x[1,],x[2,])
chisq.test(x[1,],x[2,])
```

4 **mcnemar.test.** (McNemar's Chi-Square Test for Count Data).

Performs a McNemar's chi-square test on a 2-dimensional $R \times R$ contingency table.
Data X_i, $i = 1$, ..., n (may be dependent).
$X_i = (x, y)$, $x \in A$ and $y \in B$, $||A|| = ||B|| = R$.
H_0: $P\{X_1 = (x, y)\} = P\{X_1 = (y, x)\}$ $\forall (x, y)$ or
$p(i, j) = p(j, i)$ $\forall$ (i, j).

Remark. Usual $R \times C$ contingency table consists of counts from independent observations
$X_i = (x, y)$, $i = 1$, ..., n, where
x and y indicate the ith person belongs to which row and column classifications, respectively, and the row and column classifications may not have the same numbers. And test
H_o^*: $p(i, j) = p(i, \cdot)p(\cdot, j)$ $\forall$ (i, j).

Assumption: The observations maybe dependent (*e.g.*, some objects may be measure twice), sometime apart.

If the contingency table consists of N observations cross-classified on the row and column variables, which would typically have the same levels, then McNemar's statistic could be used to test the null hypothesis of symmetry.

Under H_0, McNemar's statistic approximately $\sim \chi^2_{R(R-1)/2}$ (similar to the LRT).

df of Θ ? $\left(p_{ij}\right)_{R\times R}$

df of Θ_0 ? H_0: $p_{ij} = p_{ji}$ $\forall$ (i, j)

For $R = 2$: Let n_{ij} be the count in cell [i,j].
The test statistic is $T = Z^2$ with

$$Z = \frac{n_{12} - (n_{12} + n_{21})/2}{\sqrt{(n_{12} + n_{21})(\frac{1}{2})^2}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}.$$

	A_1	A_2
B_1	n_{11}	n_{12} .
B_2	n_{21}	n_{22}

```
> x = factor(c(1,1,2,1,2,1,1,2,2), labels=c("male", "female"))
> y = factor(c(1,1,1,2,1,2,2,1,1), labels=c("democrat", "republican"))
> mcnemar.test(x,y)
    McNemar's Chi-squared test with continuity correction
    McNemar's chi-squared = 0, df = 1, p-value = 1
```

Conclusion ?

> (x=table(x,y))

	y	
x	$democrat$	$republican$
$male$	2	3
$female$	4	0

Ans. Proporpotion of the female democrats and male republicans are the same.

Any problem with the analysis ?

5. **ks.test.** Kolmogorov-Smirnov Goodness-of-Fit Test. Performs a one or two sample Kolmogorov -Smirnov test, which tests the relationship between two distributions.

5.1. One-sample. Suppose that X_1, ..., X_m are a random sample from F. To test against H_1: $F \neq F_o$, where F_o is given (upto a parameter). The test statistic is

$J = \sup\{|F_m(t) - F_o(t)| : t \in R\}$. P-value is given in R.

Remark. Most of the time, we do not know the parameters in F_o and has to estimate the parameters. The statistic is changed this way. For instance, under normal assumption, for n large, the critical values for the ks.test with parameters known and for ks.test with estimated parameters (called Lilliefors' test) are

	0.90	0.95	0.99
with estimators	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.52/\sqrt{n}$
with known parameters	$0.82/\sqrt{n}$	$0.89/\sqrt{n}$	$1.04/\sqrt{n}$

5.2. Two-sample. Suppose that X_1, ..., X_m and Y_1, ..., Y_n are two independent samples with continuous cdfs F and G, respectively. To test against H_1: $F \neq G$, let F_m and G_n be the edf's of F and G, respectively, let d = greatest common divisor of m and n. The test statistic is

$$J = \frac{mn}{d} \sup\{|F_m(t) - G_n(t)| : t \in R\}$$

P-value is given in R.

A simulation example. Test whether two random sample have the same distribution.

```
m=1000
n=90
fun1=function(){
    x = rnorm(n)
    y = rnorm(n, mean = 0.5, sd = 1)
    a=t.test(x,y)$p.value
    return(a)
}
fun2=function(){
    x = rnorm(n)
    y = rnorm(n, mean = 0.5, sd = 1)
    a=ks.test(x,y)$p.value
    return(a)
}
u=matrix(rep(0,m*n),m)
s=apply(u,1,fun1)
mean((s<0.05))
```

[1] 0.909

mean((apply(u,1,fun2) <0.05)) # > 0.909 or < 0.909 ? **Why ?**

[1] 0.775

If the parametric assumption is correct, mean test is better.

Otherwise, ks.test is more powerful. $F_Y = F_X \begin{cases} => \\ <\neq \end{cases} \quad \mu_Y = \mu_X$.

§5.6. Density Estimation

Given a random sample, X_1, ..., X_n from X with a cdf F, where $F(t) = P(X \le t)$.

Q: F =?

Two approaches:

Parametric. $F(t) = F_o(t;\theta)$, $\theta \in \Theta \subset \mathcal{R}^p$. $\hat{F}(t) = F_o(t;\hat{\theta})$.

Non-parametric. $\hat{F}$ is the edf.

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le t).$$

The d.f. f is $f = \begin{cases} F' & \text{if } X \text{ is continuous} \\ F(t) - F(t-) & \text{if } X \text{ is discrete} \end{cases}$.

Q: Why do we need to know F ?

(1) Estimate $P(X \in A)$ by $\int \mathbf{1}(x \in A) d\hat{F}(x)$.

(2) Estimate $E(g(X))$ by $\int g(x) d\hat{F}(x)$.

(3) Compare two distributions H_o: $F(t) \le G(t)\ \forall\ t$.

Note that $E(g(X)) = \int g(x) f(x) dx$ is X is continuous.

$\int g(x) d\hat{F}(x) = \sum_{i=1}^{n} g(x_i)\frac{1}{n}$ ($= \overline{X}$ if $g(x) = x$ and $\hat{F}$ is the edf.)

Q: Why do we need to know f ?

One Example: If X_1 is continuous, the sample median $med(X)$ satisfies

$$\sqrt{n}(med(X) - m) \xrightarrow{D} N(0, \sigma^2), \text{ with } \sigma^2 = \frac{1}{4(f(m))^2}.$$

Q: f =?

Two approaches:

A. Parametric: $\hat{f}(x) = f_o(x;\hat{\theta})$, where $\hat{\theta}$ is an estimate.

B. Non-parametric:

B.1. If X is discrete, $\hat{f}(t) = \hat{F}(t) - \hat{F}(t-) = \sum_{i=1}^{n} \mathbf{1}(X_i = t)/n$.

B.2. If X is continuous, $\hat{f}$ above is not proper.

Possible estimators in case B.2:

(1) Histograms. (hist() (not really an estimator of f) or truehist()).

(2) Kernel estimators.

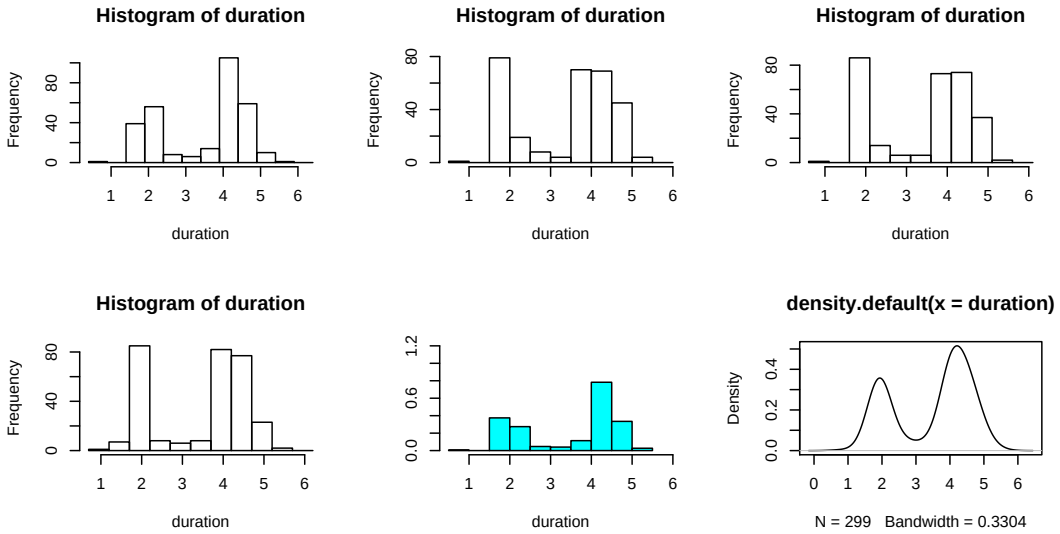
Drawbacks of histograms: It depends on the initial point and nclass. See display, two graphs below and their R programs

attach(geyser)

geyser[1:3,]

	waiting	*duration*
1	80	4.016667
2	71	2.150000
3	57	4.000000

length(geyser[,2])
[1] 299
hist(duration,breaks=seq(0.4,6.4,0.5))
hist(duration,breaks=seq(0.5,6,0.5))
hist(duration,breaks=seq(0.6,6.1,0.5))
hist(duration,breaks=seq(0.7,6.2,0.5))
truehist(duration,nbin=15,xlim=c(0.5,6),ymax=1.2)
# shaped area =1, # of block ≤ 15
plot(density(duration),lty=1,type="l")



(2) Kernel estimators.
density(x, adjust = 1, window = kernel, width, n = 512, from, to)
kernel = c("gaussian", "epanechnikov", "rectangular",
"triangular", "biweight", "cosine", "optcosine"),

$$\hat{f}(t) = \frac{1}{b}\int K(\frac{x-t}{b})d\hat{F}(x) = \sum_{i=1}^{n} K(\frac{x_i - t}{b})\frac{1}{nb},$$

where $b = width$ (bandwidth), $K(\cdot)$ is a kernel, satisfying $\int K(x)dx = 1$ (and $K(x) \geq 0$).
Examples of kernels:

$$g\ (gaussian): \quad K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$
$$r\ \ (rectangular): \quad K(x) = \frac{1}{2}\mathbf{1}(|x| \leq 1)$$
$$t\ \ (triangular): \quad K(x) = (1-|x|)\mathbf{1}(|x| \leq 1)$$
$$c\ \ (cosine): \quad K(x) = \frac{1}{2}(1+\cos(\pi x))\mathbf{1}(|x| \leq 1)$$
$$e\ \ (epanechnikov): \quad K(x) = \frac{3}{4}(1-x^2)\mathbf{1}(|x| \leq 1)$$

Bandwith selection:

Minimize the mean integrated squared error (MISE)

$$\begin{aligned} MISE =& E(\int (\hat{f}(x;b) - f(x))^2 dx) \\ =& E(\int \hat{f}^2(x;b)dx - 2E\hat{f}(X;b) + \int f^2(x)dx) \\ =& \frac{1}{nb}\int K^2 + \frac{b^4}{4}\int (f'')^2\{\int x^2 K\}^2 + O(\frac{1}{nb} + b^4) + \int f^2 \\ \to& \infty \text{ if } b \to 0+ \text{ or } b \to \infty. \end{aligned}$$

The optimal bandwidth would be

$$b = (\frac{\int K^2}{n\int (f'')^2\{\int x^2 K\}^2})^{1/5}$$

with f'' given. Since f'' needs to be estimated, a compromise is

$$b = nrd = 1.06 min(\hat{\sigma}, IQR/1.34)n^{-1/5}, \text{ where IQR=3rd quantile -1st quantile}$$

Another choice is width="SJ" (Sheather and Jones (1991)).

Q: Why MISE, not MSE $E((\hat{f}(t) - F(t)^2)$?

Example. Compute the SD of the sample median, using
galaxy data (velocities in km/sec of 82 galaxies), where $\sigma^2 = \frac{1}{4(f(m))^2}$.

```
> min(galaxies)
[1] 9172
> gal=galaxies/1000
> median(gal)
[1] 20.8335
> (u=density(gal, from=20.8335, to=20.8335))
```

how many output ?

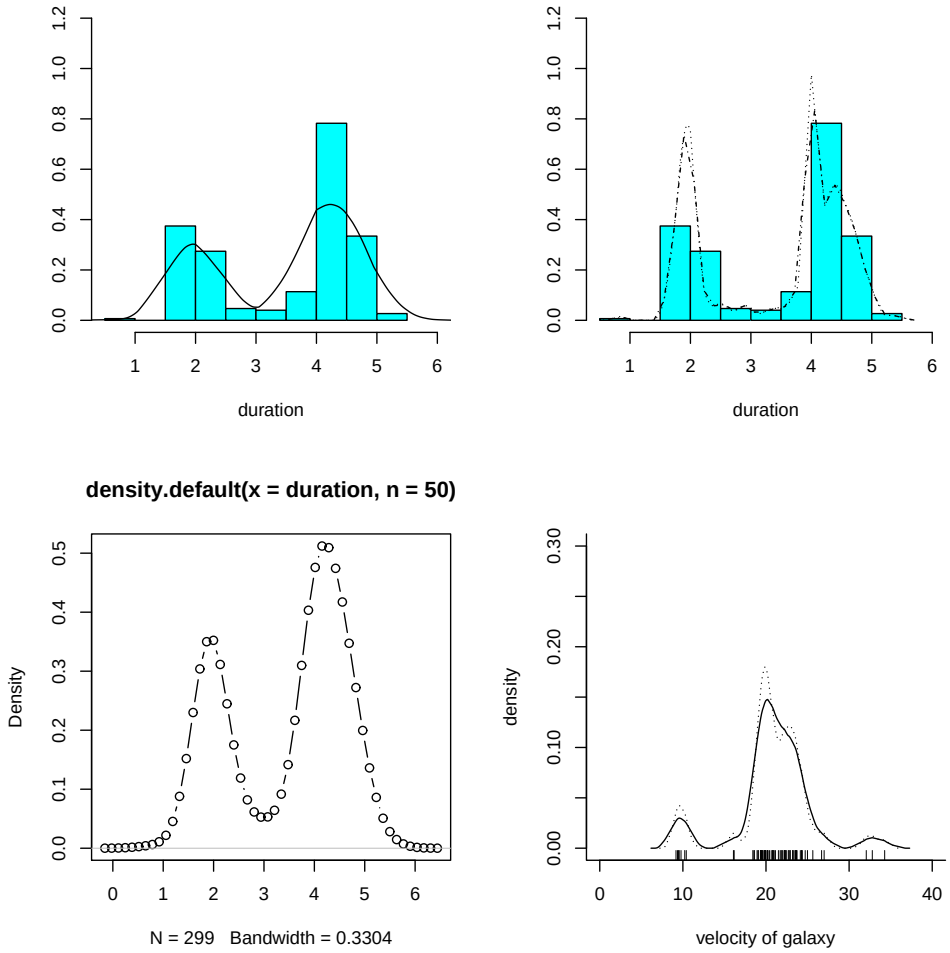
Output: .....

x	y
$Min.$: 20.83	$Min.$: 0.1353
⋮	⋮
$Max.$: 20.83	$Max.$: 0.1353

$2^9 = 512$

```
> (u=density(gal,window="triangular",width="SJ", n=1, from=20.8335, to=20.8335))
> u$x    #≈median(gal)
[1] 20.83
> (f=u$y) [1] 0.1353
> 1/(2*sqrt(length(gal))*f) # formula for SD of sample median
[1] 0.4079768
```

Density estimation using geyser data and galaxy data




```
par(mfrow =c(2,2))
truehist(duration,nbin=15,xlim=c(0.5,6),ymax=1.2)
lines(density(duration,window="triangular",width="nrd"))
truehist(duration,nbin=15,xlim=c(0.5,6),ymax=1.2)
lines(density(duration,window="triangular",width="SJ",n=2**8),lty=3)      #256
lines(density(duration,window="triangular",width="SJ",n=2**5),lty=4)      #32
plot(density(duration,n=50),lty=1,type="b")
plot(x=c(0,40),y=c(0,0.3),type="n", bty="l",xlab="velocity of galaxy", ylab="density")
rug(gal)
lines(density(gal,window="triangular",width="SJ",n=256),lty=3)
lines(density(gal,window="triangular", n=256),lty=1)
```

For a given data set, the density estimators vary.

Do we know the true density ?

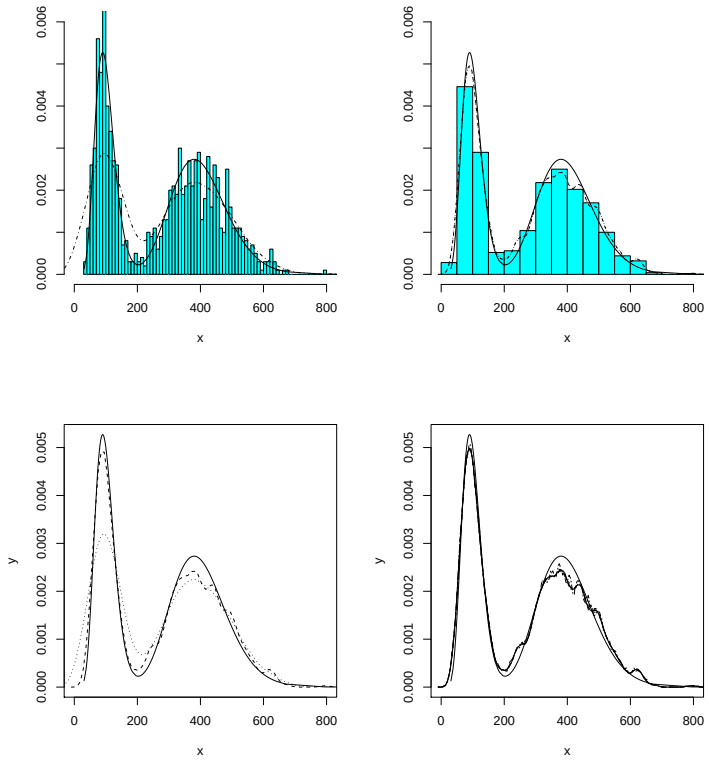
A simulation study can let us see the difference between the real one and guesses.

A simulation study. Generate data from the mixture of two Gamma distributions $Gamma(shape, scale)$. The density is then $f(x) = f_W(x) = 0.4 * f_X(x) + 0.6 * f_Y(x)$, where f_X and f_Y are the densities of $Gamma(10, 10)$ and $Gamma(20, 20)$, respectively, or $W = \begin{cases} X & \text{if } Z = 0 \\ Y & \text{if } Z = 1 \end{cases}$, where $Z \sim bin(1, 0.6)$. Plot and compare the density, the histogram,

various estimates of density in the graphs.

```
p=(rbinom(1000,1,0.6)+1)*10
x=rgamma(1000,shape=p, scale=p)                    # x=rgamma(1000,p, 1/p)
x=sort(x)
y=0.4*dgamma(x,10,0.1)+0.6*dgamma(x,20,0.05)
truehist(x,nbin=80,xlim=c(0,800),ymax=0.006)
lines(density(x,window="triangular",width="nrd",n=500),lty=4)
lines(x,y,lty=1)
truehist(x,nbin=15,xlim=c(0,800),ymax=0.006)
lines(density(x,window="triangular",width="SJ",n=100),lty=3)
lines(density(x,window="triangular",width="SJ",n=500),lty=2)
lines(x,y,lty=1)
plot(x,y,type="l",lty=1,xlim=c(0,800))
lines(density(x,window="triangular",width="SJ",n=100),lty=2)
lines(density(x,window="triangular", n=100),lty=3)
plot(x,y,type="l",lty=1,xlim=c(0,800))
lines(density(x,window="g",width="SJ"),lty=4)
lines(density(x,window="c",width="SJ"),lty=5)
lines(density(x,window="r",width="SJ"),lty=6)
lines(density(x,window="t",width="SJ"),lty=7)
lines(density(x,window="e",width="SJ"),lty=2)
lines(x,y,lty=1)
```

Density estimation using simulation data



§5.7 Bootstraping

Q: Why bootstrapping?
Ans: To solve (1) the variance of an estimator = ? and (2) confidence interval = ?

Under parametric approach, say $X \sim F_o(\cdot;\theta)$, the MLE of θ, say
$\hat{\theta}$ is often asymptotically distributed as $N(\theta, \sigma_n^2)$,
where σ_n^2 can be estimated by

$$\hat{\sigma}_n^2 = -\left(\frac{\partial^2 \log \prod_{i=1}^n f_o(X_i;\theta)}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} \right)^{-1} \tag{1}$$

and a 95% CI of θ can be approximated by

$$\hat{\theta} \pm 1.96\sqrt{\hat{\sigma}_n^2}. \tag{2}$$

However, Eq. (1) and Eq. (2) may not hold if

under non-parametric approach

or an estimator is not asymptotically normally distributed.

Bootstrap method may provide a solution in such cases.

Suppose we want to estimate θ, by a statistic $\hat{\theta}(\underline{X})$ based on observations $\underline{X} = (X_1, ..., X_n)$.

Method: A random sample with replacement of size n is taken from the set $\{X_1, ..., X_n\}$, denoted by $\underline{X}^{(i)} = (X_1^{(i)}, ..., X_n^{(i)})$, where $i = 1$.

$\underline{X}^{(i)}$ is called resampling sample.

One can compute $\hat{\theta}_1 = \hat{\theta}(\underline{X}^{(i)})$.

Repeat to get m resampling samples and m $\hat{\theta}(\underline{X}^{(i)})$s, say $\hat{\theta}^{(i)}$s.

Estimator of $\sigma_{\hat{\theta}}^2$= the sample variance of $\hat{\theta}^{(i)}$'s

Example 1 (simulation study). Obtain $\hat{\sigma}_{sample\ median}$.

```
> n=100
> m=1000
> da=rcauchy(n*n)
> da=abs(da)
> y=rep(0,n)
> t=rep(0,n)
> res=numeric(m)
> dim(da)=c(n,n)
> for (j in 1:n) {
    y[j]=median(da[j,])
    for(i in 1:m) res[i] = median(sample(da[j,],replace=T))
    t[j] = sd(res)
    }
> sd(y)
    [1] 0.1605435
> mean(t)
    [1] 0.1705241
```

What do sd(y) and mean(t) mean ?

$sd(y) \approx \sigma_{sample\ mean}$,

mean(t) $\approx$ the average of the bootstraping estimates of $\sigma_{sample-median}$.

What does the simulation result suggest ?

Another R function for bootstraping: boot() and boot.ci().
Example 2. Using Galaxies data.
First way:
> for(i in 1:m) res[i] = median(sample(gal,replace=T))
> sd(res)
[1] 0.5254444 **What is the answer next time ?**
The second way:
> temp=boot(gal,function(x,i) median(x[i]),R=1000)
> temp

Bootstrap Statistics :

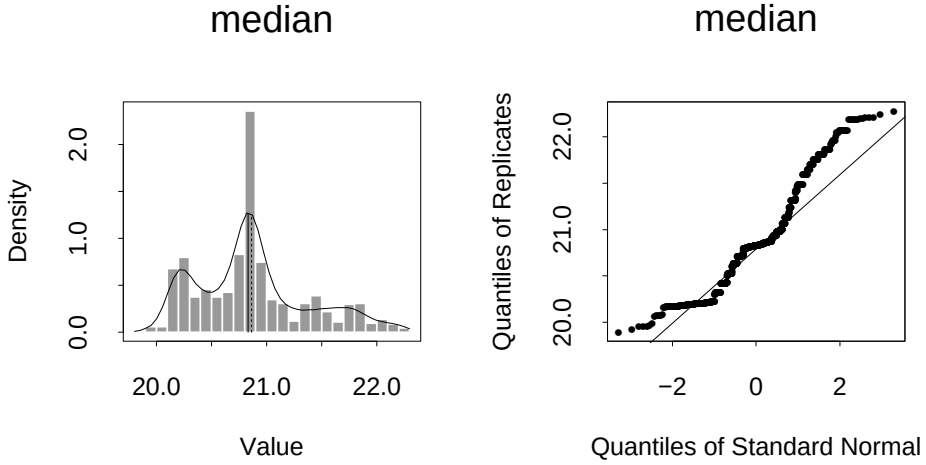
	original	*bias*	*std.error*
$t1*$	20.8335	0.0808045	0.5317111

> summary(temp)

	Length	*Class*	*Mode*	
$t0$	1	$-none-$	*numeric*	
t	1000	$-none-$	*numeric*	$= res[1:1000]$
R	1	$-none-$	*numeric*	
data	82	$-none-$	*numeric*	
seed	626	$-none-$	*numeric*	
...				

> sd(temp$t) **=sd(res) ?**

[1] 0.5317111 **Why sd(res)=0.525444 ?**
Recall $\hat{\sigma}_{median} = \frac{1}{2\hat{f}(median)} = 0.4079768.$ **=sd(res) ?**
> plot(temp) # yields the next figure



edf of bootstraping med(galaxies/1000) using existing program

The curves of truehist() and qqplot based on the bootstraping sample of med($\underline{X}$) suggest that the cdf of the med($\underline{X}$) does not have a normal distribution.

Confidence interval of a parameter (L, R):

e.g. 95% approximate CI of θ satisfies $\mathrm{P}\{L < \theta < R\} \approx 0.95$.

If $\hat{\theta}$ is approximately $N(\theta, \sigma^2_{\hat{\theta}})$, then

$(L, R) = (\hat{\theta} - 1.96\hat{\sigma}_{\hat{\theta}},\ \hat{\theta} + 1.96\hat{\sigma}_{\hat{\theta}})$,
as $P\{\hat{\theta} - \theta < 1.96\hat{\sigma}_{\hat{\theta}}\} \approx 0.025$.

$\hat{\theta}$ may not be approximately normal, then there are several approaches.

percentile CI: One expects that $\hat{\theta} \approx \theta$, the edf $\hat{F}_{\hat{\theta}^*}$ of $\hat{\theta}^* \approx$ the cdf of $F_{\hat{\theta}}$,
the empirical quantile $\hat{Q}_{\hat{\theta}^*}(0.025)$ and $\hat{Q}_{\hat{\theta}^*}(0.975)$ satisfies
$P\{\hat{Q}_{\hat{\theta}^*}(0.025) \le \theta \le \hat{Q}_{\hat{\theta}^*}(0.975)\} \approx 0.95$,
thus

$$(L, R) = (\hat{Q}(0.025), \hat{Q}(0.975))$$

is called the percentile CI.
L (R) is the 2.5th (97.5th) quantile of the edf of the m $\hat{\theta}_1^*$, ..., $\hat{\theta}_m^*$s.

Basic CI: One expects that
$0.95 = P(a \le \hat{\theta} - \theta \le b) \approx P(a \le \hat{\theta}^* - \hat{\theta} \le b)$.
Thus a and b are the 2.5th and 97.5th quantiles of the edf based on $\hat{\theta}_i^* - \hat{\theta}$, $i = 1$, ..., m.
A 95% CI is $(L, R) \approx (\hat{\theta} - b, \hat{\theta} - a)$. It can be shown that
$(L, R) \approx (2\hat{\theta} - \hat{Q}_{\hat{\theta}^*}(0.975), 2\hat{\theta} - \hat{Q}_{\hat{\theta}^*}(0.025))$.
R also gives another CI denoted by **bca** or **BCa** (see BC_a on page 136 of V & R).

Three program for bootstraping galaxies data:

```
gal=galaxies/1000
(1) m=1000; res=numeric(m)            # res=rep(0,m)
    for(i in 1:m) res[i] = median(sample(gal,replace=T))
    s=sd(res) # sample SD of sample median
    x=median(gal)
    c(x-1.96*s,x+1.96*s) # normal CI
    (y=quantile(res,p=c(0.025,0.975))) # percentile CI   y=sort(res)[c(25,975)]
    2*x-y[2:1]                        # ??
(2) temp=boot(gal,function(x,i) median(x[i]),R=1000)
    boot.ci(temp, type = c("norm", "basic", "perc", "stud"))
(3) fun = function(d, i) {
        m = median(d[i])
        n = length(i)
        v = (n-1)*var(d[i])/n**2                    # var(x) = 1/(n-1) sum_{i=1}^n (x_i - xbar)^2
        c(m, v)
    }
    temp=boot(gal,fun, R=1000)
    boot.ci(temp, type = c("norm", "basic", "perc", "stud"))
    boot.ci(temp)
```

Output:

(1)

```
> c(x-1.96*s,x+1.96*s)
[1] 19.79584 21.87116
> (y=quantile(res,p=c(0.025,0.975)))
     2.5%      97.5%
 20.17245  22.05300
```

```
> 2*x-y[2:1]
      97.5%       2.5%
   19.61400   21.49455
```

(2)

```
> temp=boot(gal,function(x,i) median(x[i]),R=1000)
> boot.ci(temp, type = c("norm", "basic", "perc", "stud"))
    BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
    Based on 1000 bootstrap replicates
    CALL :
    boot.ci(boot.out = temp, type = c("norm", "basic", "perc", "stud"))
    Intervals :
      Level      Normal            Basic          Percentile
      95%    (19.78, 21.81)   (19.62, 21.50)   (20.17, 22.05)
```

(3)

```
> boot.ci(temp, type = c("norm", "basic", "perc", "stud"))
    Intervals :
      Level      Normal            Basic
      95%    (19.78, 21.81)   (19.62, 21.50)
      Level    Studentized      Percentile
      95%    (19.57, 21.59)   (20.17, 22.05)
> boot.ci(temp)
    Intervals :
    Level      Normal            Basic           Studentized
    95%    (19.78, 21.81 )  (19.62, 21.50 )  (19.57, 21.59 )
    Level      Percentile          BCa
    95%    (20.17, 22.05 )  (20.08, 21.92 )
```

§5.5. Robust estimators.

Suppose that X_1, ..., X_n are i.i.d. from a df $f = f(\cdot;\theta)$, $\theta \in \Theta \subset \mathcal{R}^p$.

θ =? (estimation of θ).

Several methods:

1. MLE. $\hat{\theta} = argmax_\theta \prod_{i=1}^n f(X_i;\theta)$.
2. MME. Solution to $\overline{X^k} = E_\theta(X^k),\ k = 1, ..., p$.
3. MDE. $\hat{\theta} = argmin_\theta \sum_{i=1}^n |X_i - \theta|$,
   or $\hat{\theta} = argmin_\theta \sum_{i=1}^n |X_i - \theta|^2 = argmin_\theta \sqrt{\sum_{i=1}^n |X_i - \theta|^2}$
4. Bayes estimator. $\hat{\theta} = argmin_{\tilde{\theta}} r(\pi, \tilde{\theta})$ (r is the Bayes risk of $\tilde{\theta}$),
   π is a density function of θ, $r(\pi, \hat{\theta}) = E(E((\hat{\theta}(X) - \theta)^2|\theta))$
5. Robust estimator ?

Location or scale parameter example. Suppose X_i are from a cdf F with median m or mean μ, and scale τ or standard deviation σ.

Example 1. Exp(1) distribution. μ_X =1, median m =ln2 and σ_X =1.

The mean and median are both called location parameters (centers).

The SD is called a scale parameter.

Example 2. Cauchy Distribution $f(x;\theta,\tau) = f_o(\frac{x-\theta}{\tau})$, where $f_o(x) = \frac{1}{\pi(1+x^2)}$.

θ is the median, a location parameter, the mean = ?

τ is a scale parameter, the standard deviation = ?

If the distribution is symmetric about the center (e.g., $N(\mu, \sigma)$),

mean and median are the same (**Is it correct ??**)

$\overline{X}$ and med(X) are two location estimators.

$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ is a scale estimator.

Note that μ and *median* are often quite close.

$\overline{X} \neq \mu$! (estimate $\hat{\theta} \neq \theta$!)

Example 3. (A simulation study).

> x=rnorm(10)

> x=c(x,100)

The data x can be roughly viewed as from

$$F_X(t) = \frac{1}{11}[\mathbf{1}(t \geq 100) + 10\Phi(t)], \text{ where } \Phi(t) \text{ is the cdf of } N(0,1).$$

Q: $E(X) = ?$ The median of $X = ?$

> mean(x)

[1] 9.138084 # compare to the mean of N(0,1)

> median(x)

[1] 0.1937377 # compare to the median of N(0,1)

> x=rexp(40)

> x=c(x,100)

> mean(x)

[1] 3.254283 # compare to the mean of Exp(1)

> median(x)

[1] 0.6262377 # compare to the median of Exp(1) *i.e.*, $\log(2) \approx 0.693$

Q: Why med(X) is more stable than $\overline{X}$?

Ans: Outliers in data, med(X) is less sensitive to outliers than $\overline{X}$.

Outliers distort some estimators greatly.

In fact,

$$\lim_{X_1 \to \pm\infty} \overline{X} = \pm\infty.$$

$\lim_{X_1 \to \pm\infty}$ med(X) is finite.

Q: How to quantify outliers ?

Use boxplot for detecting outliers.

Example 4. Boxplots of data chem and abbey.

> summary(chem) # (24 determinations of copper in wholemeal flour)

Min.	1*st Qu.*	*Median*	*Mean*	3*rd Qu.*	*Max.*
2.200	2.775	3.385	4.280	3.700	28.950

> summary(abbey) # (31 Daily Price Returns Of Abbey National Shares)

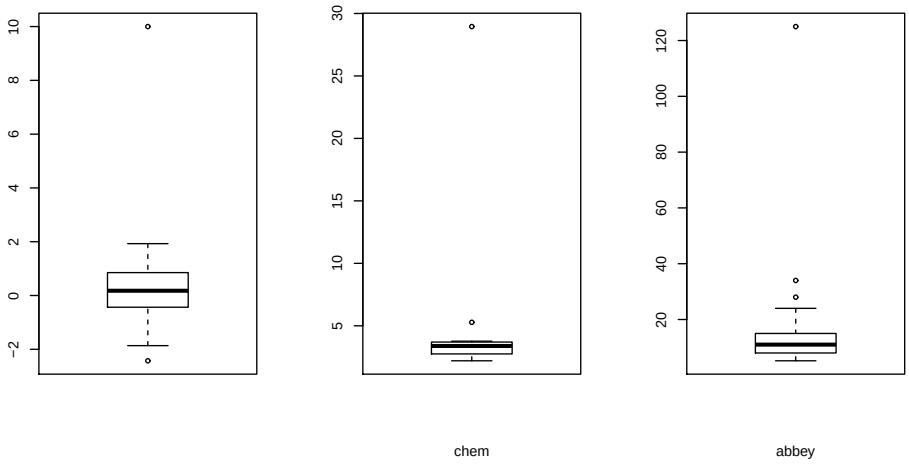
Min.	1*st Qu.*	*Median*	*Mean*	3*rd Qu.*	*Max.*
5.20	8.00	11.00	16.01	15.00	125.00

> x=c(rnorm(40),10)

> boxplot(x)

> boxplot(chem, sub="chem")

> boxplot(abbey, sub="abbey")



Def. If an observation is more than 4 IQR (inter-quartile-range or $Q_3 - Q_1$) away from the center of the data, it is called an outlier

center —- median

lower and upper hinges —– 1st and 3rd quartiles

whiskers ———- 1.5 IQR from hinges (unless max or min $<$ 1.5 IQR).

Observations outside whiskers are suspected outliers.

It is definitely an ourlier if it is $\begin{cases} 3 \text{ IQR away from hinges, or} \\ 3 \text{ SD away from the center.} \end{cases}$

Outliers may due to typos, errors or maybe true observations.

It can be viewed as $X = \begin{cases} U & \text{if } Y = 0 \\ V & \text{if } Y = 1, \end{cases}$ or

$F_X(x;\theta) = (1-\alpha)F_U(x;\theta) + \alpha F_V(x)$, $\alpha \in [0, \epsilon)$, and $Y \sim bin(1, \alpha)$.

Q: How to quantify that an estimator is less sensitive to outliers ?

Ans: Stability and resistancy.

Def. An estimator is **stable** if it does not change by a large amount when an outlier is added in. In particular, the estimator is bounded no matter what the outlier is equal to.

The **breakdown value** (or point) of an estimator is the supremum value of the proportion p of the sample that can be moved to ∞ without the statistic moving to ∞ (in the case that the sample size is as large as necessary).

An estimator with a large breakdown point is said to be "**resistant** (to gross errors)".

Remark. Definition from "Robust Nonparametric Statistical Methods"

by T.P. Hettmansperger and J.W. Mckean (1998) is as follows:

Asymptotic breakdown point.

Let $\mathbf{x} = (x_1, ..., x_n)$ represent a realization of a sample and

let $\mathbf{x}^{(m)} \in \mathcal{R}^n$ represent the corruption of any m of the n observations, that is,

x_{i_1}, ..., $x_{i_{n-m}}$ among $\{x_1, ..., x_n\}$ are fixed at their original values,

but the rest m observations are changing (possibly to $\pm\infty$).

Let $\mathcal{X}$ be the collection of all combination of choosing m x_i's among $\{x_1, ..., x_n\}$ to be corrupted with the rest fixed. Of course, $\mathcal{X}$ depends on the sample and it contains sequences of elements that $||\mathbf{x}^{(m)}||$ tend to ∞. We define the bias of an estimator $\hat{\theta}$ to be

$$bias(m; \hat{\theta}, \mathbf{x}) = \sup\{|\hat{\theta}(\mathbf{x}^{(m)}) - \hat{\theta}(\mathbf{x})| : \ \mathbf{x}^{(m)} \in \mathcal{X}\}$$

If the bias is infinite, we say the estimator has broken down and the

$$\textbf{sample breakdown value} = \min\{m/n : \ bias(m; \hat{\theta}, \mathbf{x}) = \infty\}$$

Its limit as $n \to \infty$, if it exists, is called the (asymptotic) breakdown value,

$$p = \textbf{breakdown value} = \lim_{n\to\infty} \min\{m/n : \ bias(m; \hat{\theta}, \mathbf{x}) = \infty\}$$

For med(X), p=50%.
e.g., if n is 5 and less than half of the sample are moved to $+\infty$,
med(X) is bounded by original values of $X_{(1)}$ and $X_{(n)}$.
However, if 3 observations are moved to $+\infty$, $med(X) \to \infty$.
Thus the sample breakdown value of med(X) is 3/5 if $n = 5$,
and is $\begin{cases} \frac{1+n/2}{n} & \text{if } n \text{ is even} \\ \frac{n+1}{2n} & \text{if } n \text{ is odd.} \end{cases}$
The limit or the (population) breakdown value is thus $p = 1/2$.
For $\overline{X}$, p=0 (asymptotically), as the sample breakdown value is $1/n$.
Q: Robustness ?
Def. Relative efficiency (RE) of $\tilde{\theta}$ to $\hat{\theta}$ is
$RE(\tilde{\theta}, \hat{\theta}) = \lim_{n\to\infty}(\sigma_{\hat{\theta}}/\sigma_{\tilde{\theta}}) = \lim_{n\to\infty}(\hat{\sigma}_{\hat{\theta}}/\hat{\sigma}_{\tilde{\theta}})$ a.e.,
where $\hat{\sigma}^2_{\hat{\theta}}$ – asymptotic variance of $\hat{\theta}$ ($\lim_{n\to\infty} \hat{\sigma}_{\hat{\theta}}/\sqrt{Var(\hat{\theta})} = 1$).
Under the exponential type family, $n\hat{\Sigma}^2_{\hat{\theta}} \approx (E(\frac{d\ln f(X;\theta)}{d\theta} \frac{d\ln f(X;\theta)}{d\theta'}))^{-1}$, $\theta \in \mathcal{R}^p$.
$\hat{\Sigma}^2_{\hat{\theta}} \approx (\sum_{i=1}^n \frac{d\ln f(X_i;\theta)}{d\theta} \frac{d\ln f(X_i;\theta)}{d\theta'})^{-1}|_{\theta=\hat{\theta}}$.
$\hat{\sigma}^2_{\hat{\theta}} \approx (\sum_{i=1}^n (\frac{d\ln f(X_i;\theta)}{d\theta})^2)^{-1}|_{\theta=\hat{\theta}}$.
$\hat{\sigma}^2_{\hat{\theta}} \approx (\sum_{i=1}^n \frac{-d^2\ln f(X_i;\theta)}{d\theta^2})^{-1}|_{\theta=\hat{\theta}}$?
$ARE(\tilde{\theta}, \hat{\theta})$ — asymptotic RE of $\tilde{\theta}$ to $\hat{\theta}$.
Robust method studies how to find a stable or resistant estimator $\tilde{\theta}$
with a **large** ARE$(\tilde{\theta}, \hat{\theta})$ to a (possibly efficient or standard) estimator $\hat{\theta}$
under $F(\cdot;\theta) = (1-\alpha)F_o(\cdot;\theta) + \alpha F_1$.
The resulting estimator is called a **robust** estimator.
It is often that the standard situation is under the normal assumption.

Example 1. If X has the density $f(x) = f_o(x - \mu)$,
ARE(med(X),$\overline{X}$) $= \begin{cases} 64\% & \text{if } f_o \text{ is } N(0, \sigma^2) \\ 96\% & \text{if } f_o \text{ is } t_5 \\ > 1 & \text{if } f_o \text{ is the standard double exponential.} \end{cases}$
They are all symmetric distributions and thus $m = \mu$.
Q: What are the candidates of robust estimators ?
A class of location estimators:
M-estimators (MLE-like estimators).
Consider a location parameter related to $f(x - \mu)$, where
$f(x)$ is a density with $\int f(x)dx = 1$.
MLE satifies:
$\hat{\mu} = argmin_\mu(-\ln \prod_{i=1}^n f(X_i - \mu))$.

$\hat{\mu} = zero.point_\mu \sum_{i=1}^n (\ln f(X_i - \mu))'$ if the zero point is unique.
Rewrite

$$\hat{\mu} = argmin_\mu \sum_{i=1}^n \rho(X_i - \mu) \qquad \text{with } \rho = -\ln f$$

$$\hat{\mu} = zero.point_\mu \sum_{i=1}^n \psi(X_i - \mu) \qquad \text{with } \psi = \tfrac{f'(X_i-\mu)}{f(X_i-\mu)} = \rho'.$$

$\begin{cases} \text{The MLE } \overline{X} \text{ under } N(\mu, \sigma^2): & \rho = x^2/2 \ \& \ \psi(x) = x. \\ \text{The MLE med(X) under the DE (double exponential)} & \rho = |x| \ \& \ \psi(x) = sign(x) \end{cases}$
$f(x) = \frac{1}{2\tau} e^{\frac{-|x-\mu|}{\tau}}$.
Remark: The MLE $\overline{X}$ under $N(\mu, \sigma^2)$ is not robust if $X \not\sim N(\mu, \sigma^2)$.
The MLE med(X) under the DE is robust even if X is no longer the DE.
But they motivate the MLE-like function ρ and the score function ψ $(= \rho')$.
Now the ρ does not have to related to $-\ln f$, and ψ does not need to related to $-(\ln f)'$.

Example 1. med(X): $\rho = |x|$ and $\psi(x) = sign(x)$. That is,

$$med(X) = zero.point_\mu \sum_{i=1}^n \psi(X_i - \mu) = zero.point_m \sum_{i=1}^n \text{sign}(X_i - m)$$

If $n = 4$, $X_i = i$, Med(X)= **??**
$zero.point_m \sum_{i=1}^n \text{sign}(X_i - m) =$ **?? How to derive it ?**

If $n = 3$, $X_i = i$, Med(X)= **??**
$zero.point_m \sum_{i=1}^n \text{sign}(X_i - m) =$ **?? How to derive it ?**

If $n = 5$, X_i's are 1, 2, 2, 3, 4, Med(X)= **??**
$zero.point_m \sum_{i=1}^n \text{sign}(X_i - m) =$ **?? How to derive it ?**

$$\sum_{i=1}^5 \text{sign}(X_i - m) = \begin{cases} \geq 3 & \text{if } m < 2 \\ 1 & \text{if } m = 2 \\ -1 & \text{if } m \in (2, 3) \\ < -1 & \text{if } m \geq 3 \end{cases}$$

There is no solution in this case to $zero.point_m \sum_{i=1}^n \text{sign}(X_i - m)$
How to handle it ??

$zero.crossing.point \sum_{i=1}^n \text{sign}(X_i - m)$
Other M-estimators
Metric trimming M-est: (by Huber) (robust) (bisquare)

$$\psi(x) = \begin{cases} x & |x|/c < 1 \\ 0 & \text{otherwise} \end{cases} \text{ (delete large outliers).}$$

Metric Winsorizing M-est: (by Huber) (robust) (attributed to C.P. Winsor)

$$\psi(x) = \begin{cases} -c & x < -c \\ x & |x|/c < 1 \\ c & x > c \end{cases}.$$

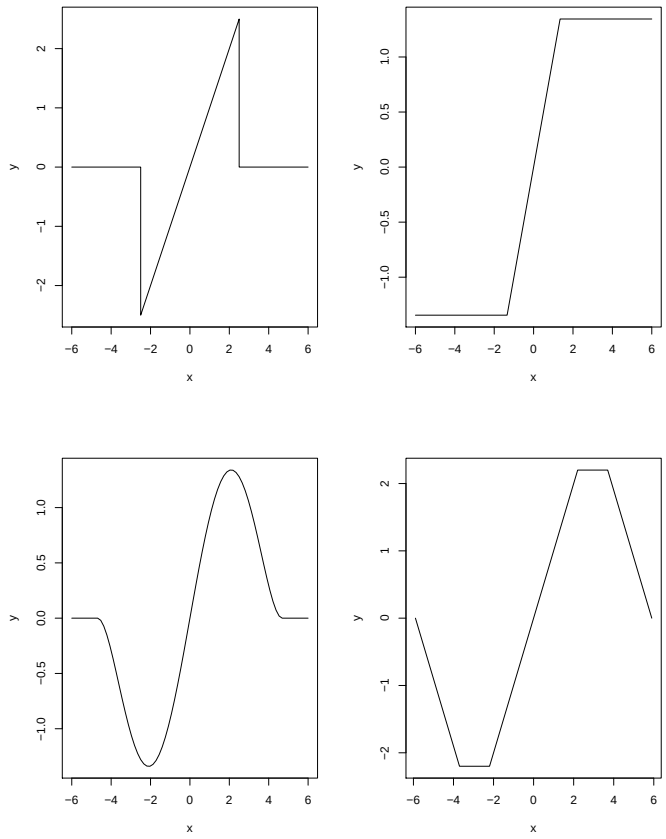
(bring large outliers to $\mu \pm c$). ARE to $\overline{X}$ is 95% under $N(\mu, \sigma^2)$ if $c = 1.345$.
Tukey's biweight M-est.:

$$\psi(x) = x\big[1 - (x/R)^2\big]_+^2$$

where $[x]_+ = x\mathbf{1}(x > 0)$. The value $R = 4.685$ gives 95% ARE at the normal.
Hampel's M-est.:

$$\psi(x) = \begin{cases} x & 0 < |x| < a \\ \text{sign}(x) \cdot a & a < |x| < b \\ \text{sign}(x)\frac{a(c-|x|)}{c-b} & b < |x| < c \\ 0 & \frac{|x|}{c} > 1. \end{cases}$$

e.g., $a = 2.2s$, $b = 3.7s$, $c = 5.9s$.



Graph of the last 4 score functions.

Remark. There is scaling problem above (c, R and s are unknown).

It can be replaced by an estimate of a scale parameter.
Possible estimators of scale parameter:

$$\textbf{non-robust} \begin{cases} S = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2} & \text{(breakdown point } p = 0). \\ \hat{\sigma}_m = \frac{1}{n}\sum_{i=1}^n |X_i - \overline{X}|\sqrt{\pi/2} & \text{(breakdown point } p = 0) \end{cases}$$

$$\textbf{robust} \begin{cases} \text{mad(X)}{=}med(|X - med(X)|)/0.6745 \\ \hat{\sigma}_q = IQR/1.35 \end{cases}$$

The coefficients are made so that they equal σ under the normal distribution.
huber(y, k = 1.5, tol = 1e-06)

Finds the Huber M-estimator of location with MAD scale.
hubers(y, k = 1.5, mu, s, initmu = median(y), tol = 1e-06)
Finds the Huber M-estimator for location with scale specified,
scale with location specified, or both if neither is specified.
mad(x, center = median(x), constant = 1.4826)

```
> length(chem)
[1] 24
> x=sort(chem)
> mean(x)
[1] 4.280417
> mean(x[2:23])
[1] 3.253636
> median(x)
[1] 3.385
> median(x[2:23])
[1] 3.385
> mean(chem,trim=0.05)
[1] 3.253636
> mean(chem,trim=0.1)
[1] 3.205
> sd(chem)
[1] 5.297396
> mad(chem)
[1] 0.526323
> huber(chem)
    $mu
[1] 3.206724
    $s
[1] 0.526323
> unlist(huber(chem))
       mu         s
 3.206724  0.526323
> unlist(hubers(chem))
       mu         s
 3.205498  0.673652
> fitdistr(chem,"t",list(m=3,s=0.5),df=5)
         m             s
 3.1853947     0.6422023
(0.1474804) (0.1279530)
> fitdistr(chem,"t",df=5)                    # same results
```

§**6.9. A Comment on the MLE with regression data.**

Let (X_1, Y_1), ..., (X_n, Y_n) be i.i.d. observations from $F_{X,Y}$,
where $X \in \mathcal{R}^p$, a column vector.

The LSE

$$\hat{\beta} = (\overline{XX'} - \overline{X}(\overline{X'}))^{-1}(\overline{X'Y} - \overline{X'}(\overline{Y})). \tag{1}$$

$$\hat{\beta} \overset{a.s.}{\to} \beta^* = \Sigma^{-1}(E(X'Y) - E(X')E(Y)) \tag{2}$$

if the expectations exist. This is due to **????**

Remark 1. Equations (1) and (2) do not rely on the assumption that (X, Y) satisfies the linear regression model:

$$Y_i = X_i'\beta + \alpha + \epsilon_i$$

where ϵ_is are i.i.d. with $E(\epsilon_i|X_i) = 0$ and $\sigma^2 = Var(\epsilon_i|X_i)$.
It is often to further assume $\epsilon_i \perp X_i$ and/or $\epsilon_i \sim N(0, \sigma^2)$.

Remark 2. $Y = \beta X + \alpha + \epsilon$

$<=>$ $Y - \beta X - \alpha = \epsilon$ and $f_{Y|X}(t|x) = f_\epsilon(\underbrace{t - \alpha - \beta x}_{u})$.

$<=>$ $Y - \beta X = W$ and $f_{Y|X}(t|x) = f_W(t - \beta x)$. $W = ??$

Example 1. If $Y|(X = x) \sim N(\beta x + \alpha, \sigma^2)$, then $W \sim N(\alpha, \sigma^2)$. $Y = \beta X + \alpha + \epsilon$.

Example 2. If $Y = \beta X + W$, $W \sim Exp(1)$, then

$f_{Y|X}(t|x) = f_W(t - \beta x) = e^{-(t-\beta x)}$, $t > ??$

Does $Y|X$ have an Exponential distribution ?

Remark 3. $\ln Y = \beta X + \alpha + \epsilon$,

$<=>$ $Y = e^{\beta X} W$

$<=>$ $f_{Y|X}(t|x) = f_W(\frac{t}{e^{\beta x}})$ $(= f_W(w))$, and $\ln W = \alpha + \epsilon$.

Example 3. If $S_{Y|X}(t|x) = \exp(-e^{-\beta x}t) = S_W(t/e^{\beta x})$, $t > 0$, then $W \sim ?$

$Y|(X = x) \sim ??$

$Y = W/e^{bX}$ **? or** $Y = W/e^{-bX}$ **?**

$\ln Y = \beta X + \alpha + \epsilon$? or $\ln Y = -\beta X + \alpha + \epsilon$?

Remark 4. If $Y = \beta X + \alpha + \epsilon$, then the LSE

$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$, where $\hat{\beta}$ is as in Eq. (1). Or the simpler formula:

$(\hat{\alpha}, \hat{\beta}')' = (B'B)^{-1}B'\mathbf{Y}$, where $B = \begin{pmatrix} 1 & X_i' \\ \vdots & \vdots \\ 1 & X_n' \end{pmatrix} = (\mathbf{1}, \mathbf{X})_{n\times(1+p)}$ and $\mathbf{Y} = (Y_1, ..., Y_n)'$.

The LSE satisfies $E(\hat{\beta}) = \beta$ and $V(\hat{\alpha}, \hat{\beta}'|\mathbf{X}) = \sigma^2(B'B)^{-1}$.

If $\epsilon \not\sim N(0, \sigma^2)$, then the anova table is not valid as it is based on F distribution; and the LSE is not an efficient estimator.

One can also consider regression models under the parametric assumption.

Assume that $Y_i|(X_i = x) \sim F$, where $F = F_o(y|x, \beta)$ has a parametric form,

and F_o is known except β.

Then in order to find $\hat{F}$, it suffices to find $\hat{\beta}$.

A standard estimator is the MLE, that maximizes

Ł$(b) = \prod_{i=1}^n f_o(Y_i|\mathbf{X}_i, b)$,

where f_o is the density of F_o and $S_o = 1 - F_o$.

The MLE is efficient if the parametric assumption is valid and certain regularity conditions are satisfied.

1. Gaussian distribution

Common form $f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(t-\mu)^2}{2\sigma^2})$, $t > 0$

With covariate in Splus or R, reparametrization:

$f_Y(y|\mathbf{x}, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y-\beta'\mathbf{x})^2}{2\sigma^2})$

or $Y = \beta'\mathbf{x} + \sigma Z$, $Z \sim N(0,1)$. $E(Z) = 0$ **???**

2. Exponential distribution

Common form $S(t) = exp(-t/\theta)$, $t > 0$.

With covariate in Splus or R, reparametrization:

$S_Y(y|\mathbf{x}, \beta) = exp(-e^{-\beta'\mathbf{x}}y)$, $y > 0$. $E(Y|\mathbf{x}, \beta) = e^{\beta x}$.

$\ln Y = \beta'\mathbf{x} + \ln Z$, $Z \sim Exp(1)$. $E(\ln Z) = 0$ **???**

$E(\ln(Z)) \approx -0.585$.

3. Weibull distribution

Common form $S(t) = exp(-t^\gamma/\theta)$, $t > 0$.

With covariate in Splus or R, reparametrization:

$S_Y(y|\mathbf{x}, \beta) = exp(-e^{-\beta'\mathbf{x}/\tau}y^{1/\tau})$, $y > 0$.

$\ln Y = \beta'\mathbf{x} + \tau \ln Z$, $Z \sim Exp(1)$. $E(\tau \ln Z) = 0$ **???**

$Z = (e^{-\beta'\mathbf{x}}Y)^{1/\tau}$.

That is, $S(t) = exp(-(t/\mu)^{1/\tau})$, $t > 0$.

4. Logistic distribution

Common form $S(t) = \frac{1}{1+exp(t)}$,

With covariate in Splus or R, reparametrization:

$S_Y(y|\mathbf{x}, \beta) = \frac{1}{1+exp(\frac{y-\beta'\mathbf{x}}{\tau})}$,

$Y = \beta'\mathbf{x} + \tau Z$, $Z \sim logistic(0,1)$. $E(Z) = 0$, $\sigma_Z = \pi/\sqrt{3}$.

5. Lognormal distribution

Assume $\ln Y = \beta'\mathbf{x} + \sigma Z$, where $Z \sim N(0,1)$. $E(Z) = 0$ **???**

6. Loglogistic distribution

$\ln Y = \beta'\mathbf{x} + \tau Z$, $Z \sim logistic(0,1)$. $E(Z) = 0$ **???**

Remark. About $\ln Y = \beta X + Z$. If $f_{Y|X}(t|x) = e^{-\beta x} f_o(te^{-\beta x})$, where f_o is a df.

Then $e^Z = Y/e^{\beta x}$ has d.f. f_o.

This is due to u-substitution.

$$\int_{-\infty}^{y} e^{-\beta x} f_o(t/e^{\beta x}) dt = \int_{-\infty}^{y/e^{\beta x}} f_o(u) du, \qquad \text{where } u = t/e^{\beta x}.$$

R command:

The parametric MLE is efficient under certain regularity assumptions. In particular,

if the residual plot suggests that certain parametric family is plausible,

one can apply the R functions as follows.

zz=survreg(Surv(y)∼x, dist="exponential")

dist: (default: weibull), gaussian, logistic, lognormal and loglogistic

glm()

The generalized linear model includes a subset of the exponential family, which is also parametric distributions. The conditional distribution $f_{Y|X}$ may not satisfy the linear regression model or log linear regression model. We can compute the MLE of the parameters based on regression data. The GLM includes

$N(\mu, \sigma^2)$,

$G(\alpha, \beta)$,
$bin(m, p)$
Poisson(μ)
inverse-Gaussian
We only review $G(\alpha, \beta)$ here.

The gamma distribution family belongs to the generalized linear model (glm) but does not satisfy the ordinary linear regression model or the ordinary log linear regression model.

7. Gamma Distribution. $f_Y(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}$, $y, \alpha, \beta > 0$.

If $\alpha \neq 1$, then treat α as known. Thus $\mu = \alpha\beta$, it can be shown that

$$\ln f_Y(y) = \underbrace{\alpha}_{\frac{A_i}{\phi}}[y \underbrace{(-1/\mu)}_{\theta_i} - \underbrace{\ln\mu}_{\gamma(\theta_i)}] + \underbrace{[\alpha\ln\alpha - \ln\Gamma(\alpha) + (\alpha-1)\ln y]}_{\tau(y, \frac{A_i}{\phi})}$$

(the form of the generalized linear model).
Then $\theta_i = \frac{-1}{\mu_i}$,
$A_i/\phi = \alpha$,
$\gamma(\theta_i) = -\ln(-\theta_i)$.
The link $l(\mu_i) = \beta' X_i$ and $\mu_i = l^{-1}(\beta' X_i)$.
The default link is $l(\mu_i) = -1/\mu_i = \beta' X_i$. $\mu_i = -1/(\beta' X_i)$
The other links are $l(\mu) = \mu$ and $\ln(\mu)$,
The identity leads to $\mu_i = \beta' X_i$.
=> $Y = \beta' X + W$, $E(Y|X) = \beta' X$?
$V(Y|X) = \frac{(\beta' X)^2}{\alpha}$?
The log leads to $\mu_i = \exp(\beta' X_i)$
=> $\ln Y = \beta' X + W$, $E(Y|X) = e^{\beta' X}$?
$E(\ln Y|X) = \beta' X$??
inverse leads to $\mu_i = -1/(\beta' X_i)$.

$$\ln f_{Y|X}(y|x_i) = \begin{cases} \underbrace{\alpha}_{\frac{A_i}{\phi}}[y \underbrace{(-1/\mu)}_{\beta x_i} - \underbrace{\ln\mu}_{\ln(\frac{-1}{\beta x_i})}] + \underbrace{[\alpha\ln\alpha - \ln\Gamma(\alpha) + (\alpha-1)\ln y]}_{\tau(y, \frac{A_i}{\phi})} & \text{inverse link} \\ \underbrace{\alpha}_{\frac{A_i}{\phi}}[y \underbrace{(-1/\mu)}_{-1/\beta x_i} - \underbrace{\ln\mu}_{\ln(\beta x_i)}] + \underbrace{[\alpha\ln\alpha - \ln\Gamma(\alpha) + (\alpha-1)\ln y]}_{\tau(y, \frac{A_i}{\phi})} & \text{identity link} \\ \underbrace{\alpha}_{\frac{A_i}{\phi}}[y \underbrace{(-1/\mu)}_{-1/e^{\beta x_i}} - \underbrace{\ln\mu}_{\ln(e^{\beta x_i})}] + \underbrace{[\alpha\ln\alpha - \ln\Gamma(\alpha) + (\alpha-1)\ln y]}_{\tau(y, \frac{A_i}{\phi})} & \text{log link} \end{cases}$$

Example 1. Carry out a simulation study on the LSE and MLE of β ($\beta = 2$), under the assumption that
$\ln Y = 2x + lnZ$ or
$Y = e^{\beta x} Z$, where $Z \sim Exp(1)$.

```
library(MASS); library(survival);
n=500
x=sample(1:4,n,replace=T)
b=2
y=rgamma(n,1,exp(-b*x))
# y=rexp(n,scale=exp(b*x))
# y=rweibull(n,1,exp(-b*x))
z=lm(log(y)~x)
z=survreg(Surv(y)~x) #weibull
z=glm(y~x,family=Gamma(link=log),maxit=50)
z=survreg(Surv(y)~x, dist="exponential")
z=survreg(Surv(y)~x-1, dist="exponential")
z=lm(log(y)~x-1)
summary(z)                    # for each of the 6 z
predict(z,data.frame(x=4),se=T)          # estimate E(Y|X = 4) with SE
```

	Value	*Std. Error*	*z*	*p*
$\exp(\alpha+\beta x) \sim x$				
LSE				
(*Intercept*)	−0.59293	0.15615	−3.797	0.000164
x	1.96718	0.05792	33.964	$< 2e-16$
survreg(*weib*)				
(*Intercept*)	0.0591	0.1219	0.485	0.628
x	1.9563	0.0449	43.530	0.000
Log(*scale*)	0.0721	0.0351	2.053	0.040
glm(*gamma*)				
(*Intercept*)	0.09658	0.11475	0.842	0.4
x	1.95446	0.04256	45.920	$< 2e-16$
survreg(*exp*)				
(*Intercept*)	0.0966	0.1123	0.86	0.39
x	1.9545	0.0419	46.66	0.00
$\exp(\beta x) \sim x-1$				
survreg(*exp*)				
x	1.99	0.0168	119	0
LSE				
x	1.76649	0.02401	73.57	$< 2e-16$

Remark. One can find some interesting facts from the table.

1. Notice that $\hat{\alpha}$ $\begin{cases} \approx 0 & \text{in the MLE of } \exp(\alpha+\beta x) \text{ or Weibull,} \\ \approx -0.59 & \text{for the LSE (**Anything wrong ?**)} \end{cases}$
2. **Relation between the** $\hat{\sigma}$ of
   $\ln(y \sim x)$, Gamma $y \sim x$, Weibull $y \sim x$, Exp $y \sim x$ and Exp $y \sim x-1$.
   Why such a relation ?
   Which estimator of β **is better ?**

Answer to question in (1):

$\alpha = 0$ in the MLE approach under Exp($e^{\beta x}$), but
$\alpha = E(\ln(Z)) \approx -0.585$ in the LSE approach $\ln Y = \beta X + \ln Z = \beta X + \alpha + \epsilon$.

Answer to question in (2):

(a) Semi-parametric v.s. parametric approach;

(b) The # of parameters go from 4 to 1.

Simulation Example 2. About ~ 1 v.s. $\sim x$ or $\sim x-1$.

Let $Y \sim Exp(\mu)$.

$S(t) = \exp(-\lambda t)\mathbf{1}(t>0)$, $\lambda = 1/\mu$.

$= \exp(-\frac{t}{\mu})\mathbf{1}(t>0$,

$Z = \frac{Y}{\mu}$, where $Z \sim Exp(1)$.

$\ln Y = \ln\mu + \ln Z$.

y=rexp(500)

fitdistr(y,"exponential")

survreg(Surv(y)$\sim$ 1, dist="exponential")

glm(y$\sim$1,family=Gamma(link=log))

lm(log(y)$\sim$1)

Question: What is the true value that is estimated ?

How many distinct values of the estimates ?

$method$	$estimate$	
$fitdistr$	1.047001	$rate$
$1/\overline{Y}$	1.047001	**Are these 2 always the same ?**
$survreg$	-0.04593031	
$log\overline{Y}$	-0.04593031	
$glm()$	-0.04593	**Are these 3 always the same ?**
$lm(\ln Y \sim 1)$	-0.5651	
$\overline{\ln Y}$	-0.5650986	**Are these two the same ?**

Suppose $Y \sim Exp(\mu)$ where $E(Y) = \mu$.

The MLE of $E(Y)$ is $\overline{Y}$.

$method$	$estimate$		
$\overline{Y}$	0.9551085	$\hat{\mu}$	
$fitdistr$	1.047001	$\hat{\lambda}$	$S(t) = \exp(-\lambda t) = \exp(-t/\mu)$
$1/\overline{Y}$	1.047001	$1/\hat{\mu}$	
$survreg$	-0.04593031	$-\ln(\hat{\lambda})$	$\ln Y = \log\mu + \ln Z$
$log\overline{Y}$	-0.04593031	$\log\hat{\mu}$	
$glm()$	-0.04593	$\tilde{\ln\mu}$	
$lm(\ln Y \sim 1)$	-0.5651	$\hat{\alpha}$ for $\ln Y = \alpha + \epsilon$	$\alpha = E(\ln Z) \approx -0.58$
$\overline{\ln Y}$	-0.5650986	$\hat{\alpha}$	$\approx -0.58 + 0.01490145$

Remark. In both the MLE approach and the LSE approach,
the SE is derived through the approximation (delta method or Slutsky's theorem).

If $g(\cdot)$ has continuous gradient and X_i's are i.i.d. with mean μ and covariance matrix Σ, then

$$nCov(g(\overline{X})) \approx \frac{\partial g}{\partial \mu'} Cov(X) \frac{\partial g}{\partial \mu}$$

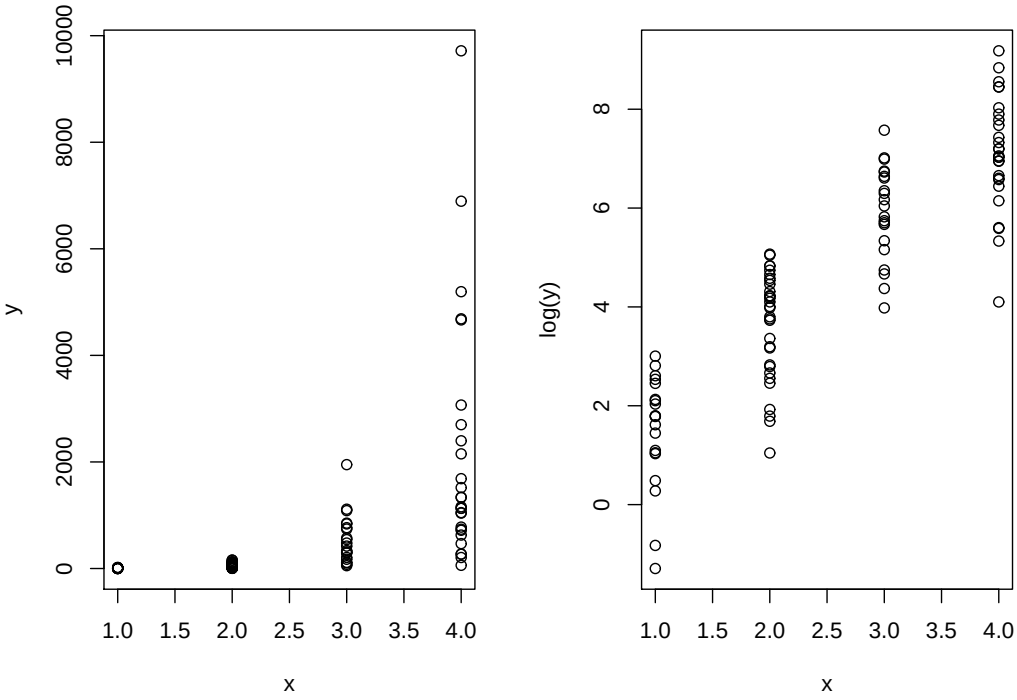
which can be further estimated by the consistent estimates of the unknown parameters.

From simulation example 1, we can see that the MLE of β is better than the LSE.
How to check whether a parametric family fit the regression data ?
Ans: Use qqplot of residuals against the quantiles of the targetting distribution.
Example 3. Estimate $E(Y|X=4)$ based on the regression data of 100 pairs of (x,y).
> n=length(y)
> ss=lm(y∼ x)
> summary(ss)

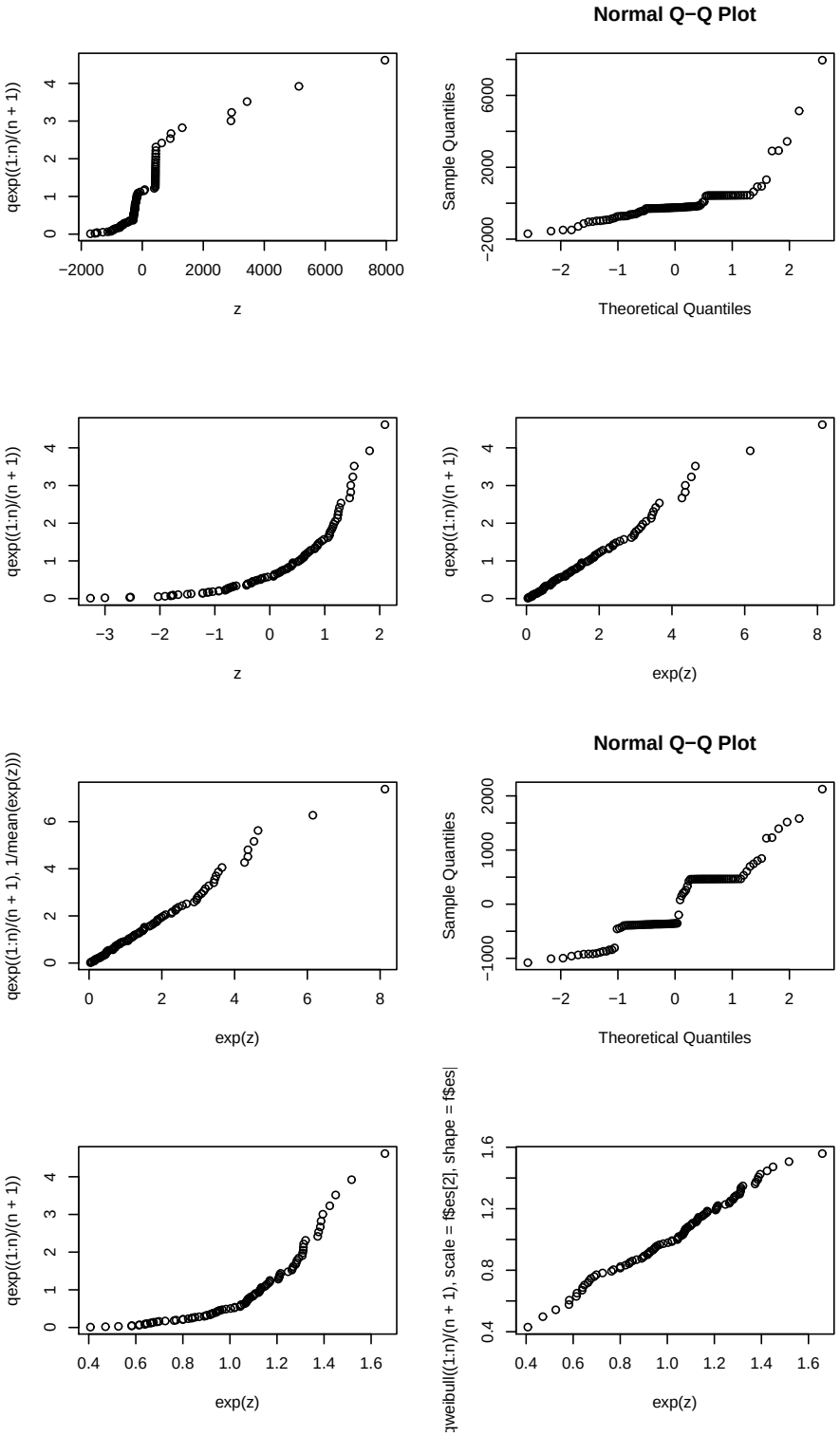
Coefficients:

	Estimate	*Std.Error*	*tvalue*	$Pr(>\|t\|)$	
$(Intercept)$	-1154.0	320.0	-3.606	0.000492	$***$
x	728.4	115.4	6.312	$8.05e-09$	$***$

Multiple R-squared: 0.289, Adjusted R-squared: 0.2818
Diagnostic Plots.
> plot(x,y)
> plot(x,log(y))



> z=resid(ss)
> qqplot(z,qexp((1:n)/(n+1))) # qqplot(1,1)
> qqnorm(z) # qqplot(1,2)
> tt=lm(log(y)∼ x)
> z=resid(tt)
> zz=survreg(Surv(y)∼ x)
> qqplot(z,qexp((1:n)/(n+1))) # qqplot(2,1)
qqplot(z,log(qexp((1:n)/(n+1))))
> qqplot(exp(z),qexp((1:n)/(n+1))) # qqplot(2,2)
> qqplot(exp(z),qexp((1:n)/(n+1),1/mean(exp(z)))) # qqplot(3,1)
Can we test it ?



Left qqplots	*Right qqplots*	*model*
$\hat{\epsilon} \sim Exp(1)$	$\hat{\epsilon} \sim N(0, \sigma^2)$	$Y = \alpha + \beta X + \epsilon$
$\ln Z \sim Exp(1)$	$Z = \exp(\log Z) \sim Exp(1)$	$\log Y = \alpha + \beta X + \log Z$
$Z \sim Exp(\hat{\mu})$		$Y = e^{\alpha+\beta X} Z$
	$\hat{\epsilon} \sim N(0, \sigma^2)$	$Y = \alpha + \beta X + \epsilon$
$Z \sim Exp(1)$	$Z = \exp(\log Z) \sim Weibull$	$\log Y = \alpha + \beta X + \tau \log Z$

Figure 1. QQ-plots in Examples 3 & 4

> ks.test(exp(z), "pexp") # see qqplot(2.2)

D = 0.21618, p-value = 0.0001745

> ks.test(exp(z), "pexp",1/mean(exp(z))) # see qqplot(3.1)

D = 0.052345, p-value = 0.9469

> ks.test(exp(z), rexp(n))

D = 0.24, p-value = 0.006302

> ks.test(exp(z), rexp(n,1/mean(exp(z))))

D = 0.08, p-value = 0.9062

Notice that the qqplots (2,2) and (3,1) in Figure 1 appear linear,

but their ks.tests are different. **Why ???**

If the quantiles of the two distribution satisfy $Q_2 = a + bQ_2$, then the qqplot appears linear.

If $Q_2 = bQ_1$, it passes the origin, as in qqplots (2,2) and (3.1).

If $Q_2 = Q_1$, **the slope is ??** #compare the slope of qqplot(3,1).

> mean(exp(z))

[1] 1.599217 #compare $1/slope$ of qqplot(2,2).

Summary:

H_0: $\ln Y = \beta X + \ln Z$, $Z \sim Exp(\hat{\mu})$.

H_0: $\ln Y = \beta X + \ln Z$, $Z \sim Exp(\mu)$.

H_0: $\ln Y = \beta X + \ln Z$, $Z \sim Exp(1.6)$.

Which of them is more appropriate ?

Conclusion: It is appropriate to fit the data to Weibull or Exponential distribution.

> predict(ss,data.frame(x=4)) # lm(y$\sim$ x)

[1] 1759.536

> predict(tt,data.frame(x=4)) # lm(log(y)$\sim$ x)

[1] 7.364874

> exp(predict(tt,data.frame(x=4)))

[1] 1962.809

> predict(zz,data.frame(x=4)) # MLE

[1] 2819.518

One may further make use the existing results as follows.

> summary(tt)

lm(formula = log(y) $\sim$ x)

Coefficients:

	Estimate	*Std.Error*	*tvalue*	$Pr(>\|t\|)$	
(*Intercept*)	-0.1765	0.2816	-0.627	0.532	
x	1.8854	0.1015	18.566	$< 2e-16$	$***$

Multiple R-squared: 0.7786, Adjusted R-squared: 0.7764

> summary(zz)

survreg(formula = Surv(y) $\sim$ x)

	Value	*Std.Error*	z	p
(*Intercept*)	0.214	0.2458	0.872	$3.83e-01$
x	1.932	0.0898	21.528	$8.49e-103$
Log(*scale*)	-0.115	0.0784	-1.461	$1.44e-01$

Scale= 0.892

Weibull distribution

Question: What can be concluded from the last 2 summaries ?
Questions: $\alpha = 0$ or $\beta = 2$ or scale=1 ??
> exp(4*2) # (= $E(Y|X = 4)$, as $Y = e^{\beta X} Z$, $Z \sim Exp(1)$).
[1] 2980.958 # **Final answer to $E(Y|X = 4)$.**
One may also construct a CI for $E(Y|X = 4)$ through information in summary(zz).
exp(4*(2+1.96*tt$coef[2,2]*c(-1,1)))
> u=predict(zz,data.frame(x=4),se=T)
> c(u$fit-2*u$se.fit ,u$fit+2*u$se.fit) # based on MLE
(1906.055, 3732.980) # it contains 2981.
> u=predict(tt,data.frame(x=4),se=T)
> c(u$fit-2*u$se.fit ,u$fit+2*u$se.fit) # Which of the two is appropriate here
> exp(c(u$fit-2*u$se.fit ,u$fit+2*u$se.fit)) # for prediction of Y ?
based on LSE of $\ln Y = \beta X + \alpha + \epsilon$.
(1092.792, 2283.027) # **Why doesn't it contain 2981 ?**
> u=predict(ss,data.frame(x=4),se=T)
> u$fit+(2*u$se.fit*c(-1,1))
based on LSE of $Y + \beta X + \alpha + \epsilon$.
(1340.891 2178.180) # **Why doesn't it contain 2981 ?**

Example 4.
> n=100
> x=sample(1:4,n,replace=T)
> b=2
> y=rweibull(n,scale=exp(b*x),shape=5)
> zz=lm(y~ x)
> summary(zz)
lm(formula = y ~ x)
Coefficients:

	Estimate	*Std.Error*	*tvalue*	$Pr(>\|t\|)$	
(*Intercept*)	-1335.26	152.38	-8.763	$5.84e-14$	$***$
x	877.70	57.35	15.304	$< 2e-16$	$***$

> z=resid(zz)
> qqnorm(z)

> zz=lm(log(y)∼ x)
> summary(zz)
lm(formula = log(y) ∼ x)
Coefficients:

	Estimate	*Std.Error*	*tvalue*	$Pr(>\|t\|)$	
(*Intercept*)	−0.10830	0.06395	−1.693	0.0935	.
x	1.99218	0.02407	82.769	$< 2e-16$	$***$

> z=resid(zz)
> zz=survreg(Surv(y)∼ x)
> summary(zz)
survreg(formula = Surv(y) ∼ x)

	Value	*Std.Error*	z	p
(*Intercept*)	0.0139	0.0469	0.297	$7.66e-01$
x	1.9925	0.0174	114.360	$0.00e+00$
Log(*scale*)	−1.5616	0.0790	−19.774	$5.03e-87$

Scale= 0.21
Weibull distribution
> qqplot(exp(z),qexp((1:n)/(n+1)))
> qqplot(exp(z),qweibull((1:n)/(n+1),scale=1,shape=1/zz$sc))
(as $\ln Y = \beta X + \tau \ln Z$ and $S_{Y/e^{\beta X}|X}(z|x) = \exp(-z^{\frac{1}{\tau}})\mathbf{1}(z > 0)$)
(f=fitdistr(exp(z),"weibull"))
qqplot(exp(z),qweibull((1:n)/(n+1),scale=f$es[2],shape=f$es[1]))

For testing Weibull:
> ks.test(exp(z), "pweibull", scale=1,shape=5)
D = 0.2271, p-value = 6.624e-05

> (f=fitdistr(exp(z),"weibull"))

shape	*scale*
4.79440995	1.12009538
(0.35680848)	(0.02467239)

> ks.test(exp(z), "pweibull", scale=f$es[2],shape=f$es[1])
D = 0.067978, p-value = 0.7446
Why are the results of the 2 tests different ?
Reason: Recall $Y/e^{\beta X} = Z$ or $\ln Y = \beta X + \ln Z$.
The 1st command uses parameters for fitting $S_{Y|X} = \exp(-(\frac{y}{e^{\alpha+\beta x}})^{\frac{1}{\tau}})$ to (X_i, Y_i)'s.
The 2nd command uses parameters for fitting $f_Z = \exp(-(\frac{z}{\theta})^{\frac{1}{\tau}})$ to exp(residuals).
The ks.test tests whether residuals fit f_Z.

> ks.test(exp(z), "pweibull", scale=1,shape=1/zz$sc)
D = 0.22605, p-value = 7.292e-05
> ks.test(exp(z), rweibull(n, scale=f$es[2],shape=f$es[1]))
D = 0.12, p-value = 0.4676

Chapter 13. Survival Analysis

§1. Introduction

Typical data in survival analysis:

Mortality data (population census).

To compute the life expectancy of the population,

n people are sampled.

Let X_i be the age at which the i-th person died,

Then we record $\begin{cases} X_i & \text{if the person died} \\ C_i+ & \text{if he/she was alive,} \end{cases}$

where C_i was his/her current age then.

Let F be the cdf of X_i, $F(t) = P(X_i \le t)$.

We shall find out $F =?$, d.f. $f =?$ and $E(X_i) =?$, etc.

Two main characters in survival analysis:

(1) $X_i \ge 0$,

(2) the exact value of X_i may not be observed.

X_i is called the failure time or survival time.

Definition: An observation on X_i is called

$\begin{cases} \text{exact} & \text{if the exact value of } X_i \text{ is observed;} \\ \textit{right censored } (RC) & \text{if } X_i \text{ is only known to be larger than some } C_i. \end{cases}$

C_i is called the censoring time.

A data set is called

$\begin{cases} \textit{complete} & \text{if all observations are exact;} \\ \textit{right censored} & \text{if each observation is either exact or RC .} \end{cases}$

Representation of RC data:

Simple way: X_i or C_i+, $i = 1, ..., n$.

Standard way: (M_i, δ_i), $i = 1, ..., n$, —- random vectors,

where $M_i = \min\{X_i, C_i\}$, $\delta_i = \mathbf{1}(X_i \le C_i)$, and

$\mathbf{1}(A) = \begin{cases} 1 & \text{if } A \text{ happens} \\ 0 & \text{o.w.} \end{cases}$ is the indicator function of the event A.

Example of RC data:

1. Mortality data (population census).
2. Type I censoring.

   Each individual was followed by a fixed time c.

   Each X was recorded unless $X > c$.

   $X_{(1)},\ ...,\ X_{(i)}, \underbrace{c+,\ ...,\ c+}_{n-i\ terms}$

   where $X_{(1)} \le \cdots \le X_{(i)}$ are order statistics of $X_1, ..., X_n$.

   Drawback ?
3. Type II censoring.

   Observation ceases after a predetermined number d of failures have been observed.

$$X_1, \cdots, \underbrace{X_d,\ c,\ ...,\ c}_{k\ terms}, c+,\ ...,\ c+$$

where $X_d = c$ and $k \ge 1$. **Advantage over Type I censoring ?**

4. Random censoring.

In a medical follow-up study of 5 years, n cancer patients are enrolled (not necessary from the beginning). X is the time to death of a patient since a certain treatment (after the enrollment). We either observe X or observe $X > 5 - B$, where B is the beginning time of the treatment for the individual since the start of the study.

Leukaemia data

Gehan, 1965 recorded times of remission of leukaemia patients.
Some were treated with drug 6-mercaptopurine (6-MP),
the others were serving as a control (placebo).

Time of remission (weeks).

Group 0 (6-MP): (m=21),

6+, 6, 6, 6, 7, 9+, 10+, 10, ..., 34+, 35+,

Group 1 (control): (n=21).

1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 17, 22, 23

The right censorship model (RC model) (Kaplan and Meier, 1958, JASA). Assume:

X — random survival time,

C — random censoring variable,

X and C are independent,

Observable random vector (M, δ) $(= (\min(X, C), \mathbf{1}_{(X \le C)}))$.

Given (M_i, δ_i), $i = 1$, ..., n, the likelihood is

$$Ł = \prod_{i=1}^{n} (f(M_i))^{\delta_i} (1 - F(M_i))^{1-\delta_i}$$

Simple explaination on Ł: In discrete case,

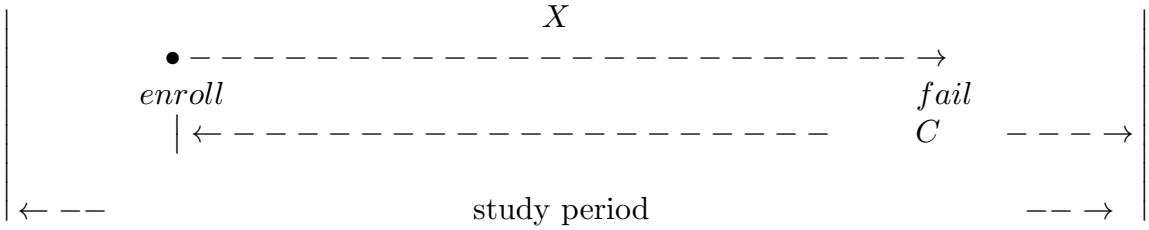
$$P((M, \delta) = (m, t)) = \begin{cases} f(m)(1 - F_C(m-)) & \text{if } t = 1 \\ (1 - F(m)) f_C(m) & \text{if } t = 0 \end{cases}$$

We assume that F_C does not contain information of F. *e.g.*, $X \sim Exp(\mu)$, $C \sim U(0, 5)$, μ is unknown. Then F_c contains no information about μ.

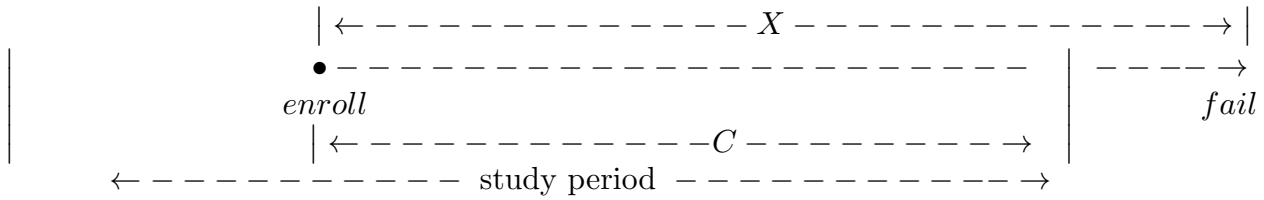
Thus F_c does not need to be in the likelihood.

Graphical illustration for C and X

Exact observation



Right-censored observation



Remark.

1. (M_i, δ_i)'s are observations on F_X.
2. $(M_i, 1-\delta_i)$'s are observations on F_C.
3. M_i's are observations on F_M.

§2. **Some functions.**

Definition.

$S(t) = P(X > t)$ is often called the *survival function* of X.

$S(t-) = P(X \geq t)$ is <u>*sometimes*</u> called the *survival function* of X, *e.g.* in Cox and Oakes book.

If X is continuous, $S(t-) = S(t)$.

In general, $S(t-) = \lim_{u \uparrow t} S(u)$.

The p.d.f. of X,

$$f(t) = \begin{cases} -S'(t) & \text{if } X \text{ is continuous} \\ S(t-) - S(t) & \text{if } X \text{ is discrete.} \end{cases}$$

$$S(t) = \begin{cases} \int_t^\infty f(x)dx & \text{if } X \text{ is continuous} \\ \sum_{x>t} f(x) & \text{if } X \text{ is discrete.} \end{cases}$$

Definition.

$h(t) = f(t)/S(t-)$ is called the *hazard function* of X.

$H(t) = -\log S(t-)$ is called the *integrated hazard* of X.

Interpretation of h and H:

$$h(x) = \begin{cases} \lim_{\triangle \to 0} \frac{P(X \in [x, x+\triangle) | X \geq x)}{\triangle} & \text{if } X \text{ is continuous} \\ P(X = x | X \geq x) & \text{if } X \text{ is discrete.} \end{cases}$$

If X is continuous,

$$\begin{aligned} H'(t) =& (-\log S(t))' \\ =& f(t)/S(t) \\ =& h(t) \end{aligned}$$

That is

$$H(t) = \int_{-\infty}^t h(x)dx,$$

integrated hazard.

If X is discrete,

$$\begin{aligned} S(t-) &= \prod_{x_i < t, x_i \in D_f} \frac{S(x_i)}{S(x_i-)} \\ &= \prod_{x_i < t} (1 - h(x_i)) \end{aligned}$$

where $D_f = \{x : \ f(x) > 0\}$. Thus

$$H(t) = -\log \prod_{x_i < t} (1 - h(x_i)) \approx \sum_{x_i < t} h(x_i),$$

cumulative hazard.

Example 1. (Exponential distribution)

$f(x) = \rho e^{-\rho x}$, $x > 0$;

$S(x) = e^{-\rho x}$, $x > 0$;

$h(x) = \frac{f(x)}{S(x)} = \rho$, $x > 0$ (constant hazard);

$H(x) = \int_0^x h(t)dt = \rho x$, $x > 0$;

$E(X) = 1/\rho$.

§**3. Parametric approach**.

Consider RC regression data: (M_i, δ_i, X_i)s, with

$M_i = \min\{Y_i, C_i\}$, $\delta_i = \mathbf{1}(Y_i \le C_i)$, and

X_is are $p \times 1$ dimensional covariate vector.

Assume that $Y_i|X_i \sim F$, where

$F = F_o(y|X_i, \beta)$ has a parametric form, and F_o is known except β.

Then in order to find $\hat{F}$, it suffices to find $\hat{\beta}$.

A standard estimator is the MLE, that maximizes

$\text{Ł}(b) = \prod_{i=1}^n (f_o(M_i|b, \mathbf{X}_i))^{\delta_i} (S_o(M_i|b, \mathbf{X}_i))^{1-\delta_i}$,

where f_o is the density of F_o and $S_o = 1 - F_o$.

1. weibull distribution

Standard form can be written as $S(t) = exp(-(t/\tau)^\gamma)$, $t > 0$.

With covariate reparametrization:

$S_Y(y|\mathbf{x}, \beta) = exp(-(\frac{y}{e^{\beta'\mathbf{X}}})^\gamma)$, $y > 0$, or

$\ln Y = \beta'\mathbf{x} + \frac{1}{\gamma}\ln Z$, or $Z = (\frac{Y}{e^{\beta'\mathbf{X}}})^\gamma$, $Z \sim Exp(1)$.

$E(\ln(Z)) \approx -0.585$ and $E(Z) = \tau\Gamma(1 + \frac{1}{\gamma})$.

2. exponential distribution

Standard form $S(t) = exp(-t/\theta)$, $t > 0$.

With covariate reparametrization:

$S_Y(y|\mathbf{x}, \beta) = exp(-\frac{y}{e^{\beta'\mathbf{X}}})$, $y > 0$.

$E(Y|\mathbf{x}, \beta) = e^{\beta x}$.

$\ln Y = \beta'\mathbf{x} + \ln Z$, or $Z = \frac{Y}{e^{\beta'\mathbf{X}}}$, $Z \sim Exp(1)$.

3. gaussian distribution

Standard form $f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(t-\mu)^2}{2\sigma^2})$, $t > 0$

With covariate reparametrization:

$f_Y(y|\mathbf{x}, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y-\beta'\mathbf{x})^2}{2\sigma^2})$

$Y = \beta'\mathbf{x} + \sigma Z$, or $Z = \frac{Y-\beta'\mathbf{x}}{\sigma}$, $Z \sim N(0, 1)$.

4. logistic distribution

Standard form $S(t) = \frac{1}{1+exp(t)}$,

With covariate reparametrization:

$S_Y(y|\mathbf{x}, \beta) = \frac{1}{1+e^{\frac{y-\beta'\mathbf{X}}{\sigma}}}$,

$Y = \beta'\mathbf{x} + \sigma Z$, or $Z = \frac{Y-\beta'\mathbf{x}}{\sigma}$, $Z \sim logistic(0, 1)$.

$E(Z) = 0$, $\sigma_Z = \pi/\sqrt{3}$.

5. lognormal distribution

Assume $\ln Y = \beta'\mathbf{x} + \sigma Z$, where $Z \sim N(0, 1)$.

6. loglogistic distribution

$\ln Y = \beta'\mathbf{x} + \sigma Z$, $Z \sim logistic(0, 1)$.

Remark. $\frac{1}{\gamma}$ in the cases 1 and 2 or σ in the latter cases is called the *scale* in R output. If $X_i = 1 \ \forall \ i$, it reduces to the ordinary parameteric MLE, with i.i.d. (M_i, δ_i).

Splus or R command:

survreg(Surv(M,δ)∼x)

predict()

survreg(Surv(time)∼ag*log(wbc),data=leuk, dist="weibull")

dist: (default: weibull), gaussian, logistic, lognormal and loglogistic

survreg(Surv(time,cens)∼factor(pair)+treat,data=gehan, dist="exponential"))

Example 1. Gehan Data Set. A trial of 42 leukaemia patients. Some were treated with the drug 6-mercaptopurine and the rest are controls. The trial was designed as matched pairs, both withdrawn from the trial when either came out of remission.

> gehan[1:3,]

	pair	*time*	*cens*	*treat*
1	1	1	1	*control*
2	1	10	1	$6-MP$
3	2	22	1	*control*

21 pairs indexed by pair 1, ..., pair 21.

>(x=survreg(Surv(time,cens)∼factor(pair)+treat,data=gehan, dist="exponential"))

	$Value$	$Std.\ Error$	z	p
$(Intercept)$	2.070	0.725	2.854	$4.31e-03$
$factor(pair)2$	2.148	1.032	2.082	$3.73e-02$
$factor(pair)3$	1.833	1.225	1.497	$1.35e-01$
$factor(pair)4$	1.772	1.014	1.747	$8.06e-02$
$factor(pair)5$	1.468	1.009	1.455	$1.46e-01$
$factor(pair)6$	1.895	1.031	1.838	$6.60e-02$
$factor(pair)7$	0.558	1.000	0.558	$5.77e-01$
$factor(pair)8$	2.519	1.231	2.046	$4.08e-02$
$factor(pair)9$	2.297	1.229	1.870	$6.16e-02$
$factor(pair)10$	2.486	1.235	2.013	$4.41e-02$
$factor(pair)11$	1.052	1.226	0.858	$3.91e-01$
$factor(pair)12$	1.827	1.229	1.487	$1.37e-01$
$factor(pair)13$	1.677	1.227	1.367	$1.72e-01$
$factor(pair)14$	1.778	1.030	1.726	$8.44e-02$
$factor(pair)15$	2.086	1.235	1.689	$9.12e-02$
$factor(pair)16$	3.063	1.239	2.473	$1.34e-02$
$factor(pair)17$	0.800	1.020	0.784	$4.33e-01$
$factor(pair)18$	1.586	1.020	1.554	$1.20e-01$
$factor(pair)19$	1.408	1.234	1.141	$2.54e-01$
$factor(pair)20$	0.402	1.226	0.328	$7.43e-01$
$factor(pair)21$	1.970	1.241	1.588	$1.12e-01$
$treatcontrol$	-1.767	0.437	-4.041	$5.32e-05$

Scale fixed at 1

Exponential distribution

Loglik(model)= -101.6 Loglik(intercept only)= -116.8

Chisq= 30.27 on 21 degrees of freedom, p= 0.087

$$\log Y_i = \beta_1 + \sum_{j=1}^{20} \beta_{j+1} \mathbf{1}(Y_i \text{ is in pair } (j+1)) + \beta_{22} \mathbf{1}(Y_i \text{is in treatment group}) + \epsilon_i ?$$

$$\log Y_i = \beta_1 + \sum_{j=1}^{20} \beta_{j+1} \mathbf{1}(Y_i \text{ is in pair } (j+1)) + \beta_{22} \mathbf{1}(Y_i \text{is in control group}) + \epsilon_i ?$$

How can we tell ?

```
> predict(x)
 [1] 1.354090 7.927091 11.597863 67.895994 8.466179 49.562544
 [7] 7.964408 46.625088 5.878999 34.416725 9.012454 52.760541
[13] 2.366545 13.854181 16.807815 98.395994 13.466179 78.833450
[19] 16.270452 95.250175 3.878999 22.708363 8.416362 49.270906
[25] 7.245544 42.416725 8.012454 46.906359 10.903907 63.833450
[31] 28.978633 169.646169 3.012454 17.635453 6.610318 38.697997
[37] 5.537363 32.416725 2.024908 11.854181 9.708181 56.833450
> exp(sum(x$coef[c(1,22)]))
[1] 1.35409
> exp(sum(x$coef[c(1)]))
[1] 7.927091
> exp(sum(x$coef[c(1,2,22)]))
[1] 11.59786
> exp(sum(x$coef[c(1,2)]))
[1] 67.89599
> exp(sum(x$coef[c(1,3,22)]))
[1] 8.466179   # E(Y|control group in the 3rd pair) = exp(β̂1 + β̂3 + β̂treatcontrol)
>survreg(Surv(time,cens)~treat,data=gehan,dist="exponential"))
```

	$Value$	$Std.\ Error$	z	p
$(Intercept)$	3.69	0.333	11.06	$2.00e-28$
$treatcontrol$	−1.53	0.398	−3.83	$1.27e-04$

Scale fixed at 1

Loglik(model)= -108.5 Loglik(intercept only)= -116.8

Chisq= 16.49 on 1 degrees of freedom, p= 4.9e-05

From the two outputs of survreg(), We can conlude:

"treat" but not factor(pair) is significant.

Reason:

Chisq= 30.27 on 21 degrees of freedom, p= 0.087,

but Chisq= 16.49 on 1 degrees of freedom, p= 4.9e-05

The aforementioned Chisqs are due to the likelihood ratio test (LRT).

$LRT = \frac{L(\hat{\theta}_o)}{L(\hat{\theta}_\Omega)}$

where $\hat{\theta}_o$ is the MLE under the restricted parameter space Ω_o, and
$\hat{\theta}_\Omega$ is the MLE under the unrestricted parameter space Ω.

In case 1 of this example,

$\Omega = \{(\beta \in \mathcal{R}^{22}\}$,

$\Omega_o = \{\beta \in \Omega : \ \beta_2 = \cdots = \beta_{21} = \beta_{treatcontrol} = 0\}$

$-2log(LRstatistic)$ $(=-2(-101.6-(-101.6))) \sim \chi^2(22-1)$

How about case 2 ?

Example 2 (Ordinary MLE) Simulation.

```
> n=500
> z=rexp(n)
> y=exp(-2+3*log(z))                    lnY + βX + (1/γ)lnZ or Z = (Y/e^{β'x})^γ.
# y=rweibull(n, scale=exp(-2),shape=1/3)
> c=runif(n,0,90)
> d=as.numeric((y<=c))
> m=d*y+(1-d)*c
> d
> zz=survreg(Surv(m,d)~1)
> summary(zz)
```

	$Value$	$Std.Error$	z	p
$(Intercept)$	-1.94	0.1536	-12.7	$1.1e-36$
$Log(scale)$	1.18	0.0354	33.2	$2.2e-242$

Scale= 3.25

Remark. $\log(3.25) \approx 1.18$.

Example 3 (regression data). Simulation

```
> n=100
> z=rexp(n)
> x=1:n
> y=exp(x+3*log(z))            #log y=x+3* ln z
# y=rweibull(n, scale=exp(x),shape=1/3)
> c=runif(n,0,90)
> d=as.numeric((y<=c))
> m=d*y+(1-d)*c
> tau=exp(3)
> g=3
> tau*gamma(1+1/g)          μ = τΓ(1 + 1/γ)
    [1] 17.93597          # =E(Y|X=3)
> n=500
> t=rweibull(n,scale=exp(3),shape=3)
> mean(t)
    [1] 18.24878
> n=50000
> t=rweibull(n,scale=exp(3),shape=3)
> mean(t)
    [1] 17.99746
> (zz=survreg(Surv(y)~x, dist="weibull"))
    (Intercept) x
    0.3934833 0.9936918
    Scale= 2.466862
```

Loglik(model)= -5162.7 Loglik(intercept only)= -5394.2
> predict(zz,data.frame(x=3,se=T))
$fit 29.21139
$se.fit 13.34247
> (zz=survreg(Surv(y)∼x, dist="exponential"))
(Intercept) x
1.004578 1.004807
Scale fixed at 1
Loglik(model)= -5274.7 Loglik(intercept only)= -15340.8
> predict(zz,data.frame(x=3,se=T))
$fit 55.64534
$se.fit 9.957346
> (zz=survreg(Surv(y)∼x, dist="gaussian"))
(Intercept) x
-1.863609e+41 5.681060e+39
Scale= 6.433197e+41
Loglik(model)= -9768.6 Loglik(intercept only)= -9771.8
> predict(zz,data.frame(x=3,se=T))
$fit -1.693177e+41
$se.fit 0
> (zz=survreg(Surv(y)∼x, dist="lognormal"))
> predict(zz,data.frame(x=3,se=T))
$fit 17.25435
$se.fit 10.94379
> (zz=survreg(Surv(y)∼x, dist="logis"))
> predict(zz,data.frame(x=3,se=T))
$fit -1.693177e+41
$se.fit 0
> (zz=survreg(Surv(y)∼x, dist="loglogis"))
> predict(zz,data.frame(x=3,se=T))
$fit 17.14267
$se.fit 9.342127

True value ≈ 17.93. Summary of predictions:

weibull	*Exp*	*normal*	*lognorm*	*logis*	*loglogis*
29.21(13.34)	55.65(9.96)	$-1.69e+41(0)$	17.25(10.94)	$-1.69e+41(0)$	17.14(9.34)

Which is better ?

Consider another case of predict $E(Y|X = 10)$ with the same data above:
19669.18 True value
9312.827 (4709.439) log-logistic prediction
7355.4 (3819.678) log-normal prediction
142181.5 (26684.16) Exp() prediction
31438.62 (14672.82) Weibull prediction

Which prediction should we trust if we are dealing with real data ?
We need to check validity of models with analogs of edf and LSE.

Remark. What does predict() estimate ?
zz=survreg(Surv(y)∼x, dist="")
predict(zz,data.frame(x=3),se=T))

$$= \begin{cases} \hat{\alpha}+3\hat{\beta} & \text{if } Y=\alpha+\beta X+\sigma Z \text{ (logis or guassian)} \\ e^{\hat{\alpha}+3\hat{\beta}} & \text{if } \ln Y=\alpha+\beta X+\tau Z \text{ (log-normal or loglogis, or Exp)} \\ & \text{or weibull,} \end{cases}$$

$$=\text{estimate of } \begin{cases} E(Y|X=3) & \text{if logis or guassian or Exp()} \\ exp(E(\log Y|X=3)) & \text{if log-normal or loglogis or weibull or Exp(),} \end{cases}$$

$\neq$ estimate of $E(Y|X=3)$.

§4. Nonparametric estimation.

Suppose X_1, ..., X_n are i.i.d. from a cdf F_o and survival function S_o.
We call M_i a death if $\delta_i = 1$.
Given RC data (M_i, δ_i), $i = 1$, ..., n, the likelihood is

$$Ł(S) = \prod_{i=1}^{n} (f(M_i))^{\delta_i} (S(M_i))^{1-\delta_i}$$

if F is given. In nonparametric set-up, F is unknown, the generalized MLE (GMLE), or the non-parametric MLE (NPMLE) of S_o is a survival function $\hat{S}$ that maximizes

$$Ł(S) \text{ with } f(t) = S(t-) - S(t),$$

over all possible survival function S.
With complete data, the GMLE is $\hat{S}$ $(= 1 - \hat{F})$, where $\hat{F}$ is the edf.

$$\hat{S}(t) = \sum_{i=1}^{n} \frac{1}{n} \mathbf{1}(M_i > t) = \prod_{t \geq M_{(i)}} (1 - \frac{1}{n-i+1}) = \prod_{t \geq t_j} (1 - \frac{d_j}{r(t_j)})$$

where $M_{(1)} \leq \cdots \leq M_{(n)}$ are order statistics of M_is.
$t_1 < \cdots < t_m$ are distinct values of deaths.
d_j is the number of deaths at t_j,
$r(t_j) = \sum_{i=1}^{n} \mathbf{1}(M_i \geq t_j)$.
$\sigma^2_{\hat{S}(t)} = S(t)F(t)/n$, estimated by $\hat{\sigma}^2_{\hat{S}(t)} = \hat{S}(t)\hat{F}(t)/n$.
$\frac{n-1}{n}$
$\frac{n-2}{n} = \frac{n-1}{n}\frac{n-2}{n-1}$
$\cdots$

$$\frac{n-i}{n} = \underbrace{\frac{n-1}{n}\frac{n-2}{n-1}\cdots\frac{n-i+1}{n-i+2}\frac{n-i}{n-i+1}}_{i \ factors}$$

With RC data, $\hat{S}$ is called the Kaplan & Meier estimator (KME) or product-limit-estimator (PLE).

$$\hat{S}(t) = \prod_{t \geq M_{(i)}} (1 - \frac{\delta_{(i)}}{n-i+1}) = \prod_{t \geq M_{(i)}} (1 - \frac{1}{n-i+1})^{\delta_{(i)}} = \prod_{t \geq t_j} (1 - \frac{d_j}{r(t_j)}).$$

where $\delta_{(i)}$ is the δ_j associated with $M_{(i)}$.

$$\text{Its variance } \approx \frac{1}{n}(S_o(t))^2 \int_0^t \frac{1}{S_o(x-)S_C(x-)S_o(x)} dF_o(x)$$

and can be estimated by

$$\hat{\sigma}^2_{\hat{S}(t)} = \frac{1}{n}(\hat{S}(t))^2 \int_0^t \frac{1}{\hat{S}(x-)\hat{S}_C(x-)\hat{S}(x)} d\hat{F}(x)$$

where $\hat{S}_C$ is the PLE based on $(M_i, 1-\delta_i)$'s.

Another estimator (under continuity assumption) is

$$\hat{\sigma}^2_{\hat{S}(t)} = (\hat{S}(t))^2 \sum_{t \geq t_j} \frac{d_j}{r(t_j)[r(t_j) - d_j]}$$

Question: Why does $\frac{1}{n}$ disappear in $\hat{\sigma}^2_{\hat{s}(t)}$?

GMLE can be computed by Splus or R as follows.

leuk[15:17,]

	wbc	*ag*	*time*
15	100000	*present*	1
16	52000	*present*	5
17	100000	*present*	65

u=survfit(Surv(time) ∼ ag,data=leuk) # complete data with grouping

u$time[1:n]

u$surv[1:n]

> summary(u)

ag=absent

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower*95%*CI*	*upper*95%*CI*
2	16	1	0.9375	0.0605	0.82609	1.000
3	15	3	0.7500	0.1083	0.56520	0.995
4	12	3	0.5625	0.1240	0.36513	0.867
7	9	1	0.5000	0.1250	0.30632	0.816
8	8	1	0.4375	0.1240	0.25101	0.763
⋮						
56	2	1	0.0625	0.0605	0.00937	0.417
65	1	1	0.0000	*NaN*	*NA*	*NA*

ag=present

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower*95%*CI*	*upper*95%*CI*
16	13	1	0.7059	0.1105	0.51936	0.959
22	12	1	0.6471	0.1159	0.45548	0.919
26	11	1	0.5882	0.1194	0.39521	0.876
39	10	1	0.5294	0.1211	0.33818	0.829
56	9	1	0.4706	0.1211	0.28423	0.779
65	8	2	0.3529	0.1159	0.18543	0.672
100	6	1	0.2941	0.1105	0.14083	0.614
⋮						
143	2	1	0.0588	0.0571	0.00879	0.394
156	1	1	0.0000	*NaN*	*NA*	*NA*

> attach(gehan)
> Surv(time,cens)[1:7]
[1] 1 10 22 7 3 32+ 12
> Surv(time)[1:7]
[1] 1 10 22 7 3 32 12
x=survfit(Surv(time,cens) ~ treat,data=gehan) # RC data with grouping
x=survfit(Surv(time,cens) ~treat,data=gehan,conf.type="log-log") # CI uses log-log form
> summary(x)
treat=6-MP

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower*95%*CI*	*upper*95%*CI*
6	21	3	0.857	0.0764	0.620	0.952
7	17	1	0.807	0.0869	0.563	0.923
⋮						
23	6	1	0.448	0.1346	0.188	0.680

treat=control

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower*95%*CI*	*upper*95%*CI*
1	21	2	0.9048	0.0641	0.67005	0.975
2	19	2	0.8095	0.0857	0.56891	0.924
3	17	1	0.7619	0.0929	0.51939	0.893
4	16	2	0.6667	0.1029	0.42535	0.825
5	14	2	0.5714	0.1080	0.33798	0.749
8	12	4	0.3810	0.1060	0.18307	0.578
⋮						
22	2	1	0.0476	0.0465	0.00332	0.197
23	1	1	0.0000	NaN	NA	NA

> c(x$time[4],x$surv[4]) # gives $(M_{(4)}, S(M_{(4)}))$

> x=survfit(Surv(time,cens)~1,data=gehan,conf.type="log-log") # without grouping
> c(x$l[1],x$u[1])
[1] 0.8227431 0.9878735
> x=survfit(Surv(time,cens)~1,data=gehan)
> c(x$l[1],x$u[1])
[1] 0.8901054 1.0000000

Estimation of $(1-\alpha)100\%$ confidence intervals:

1. $\hat{S}(t) \pm z_{\alpha/2}\hat{\sigma}_{\hat{S}(t)}$.
2. $exp(-\hat{H}(t) \pm z_{\alpha/2}\hat{\sigma}_{\hat{H}(t)})$, where $\hat{H}(t) = -\ln(\hat{S}(t))$, as $S(t) = exp(-H(t))$ and $e^x \geq 0$,
3. $exp(-exp(\hat{G}(t) \pm z_{\alpha/2}\hat{\sigma}_{\hat{G}(t)}))$ $(e^{-e^x} \downarrow \& \in (0,1))$
   where $G(t) = \ln(-\ln S(t)) = \ln H(t)$,
   $\hat{G}(t) = \ln(-\ln\hat{S}(t))$, and

$$\hat{\sigma}^2_{g(\hat{s}(t))} = (\frac{\partial g(x)}{\partial x}\big|_{x=\hat{S}(t)})^2 \hat{\sigma}^2_{\hat{S}(t)}.$$

Comments: Behaviors of CI (L, U) in the three cases:
In case 1, it is possible that $L < 0$ and $U > 1$.
In case 2, $L \geq 0$ but it is possible that $U > 1$.
In case 3, $0 \leq L \leq U \leq 1$.

Comparing two treatments.

In medical research, two treatments are compared.

Figure 13.1 presents the PLE curves, the pair of data, and the CI of the PLE curves for Gehan data in R.

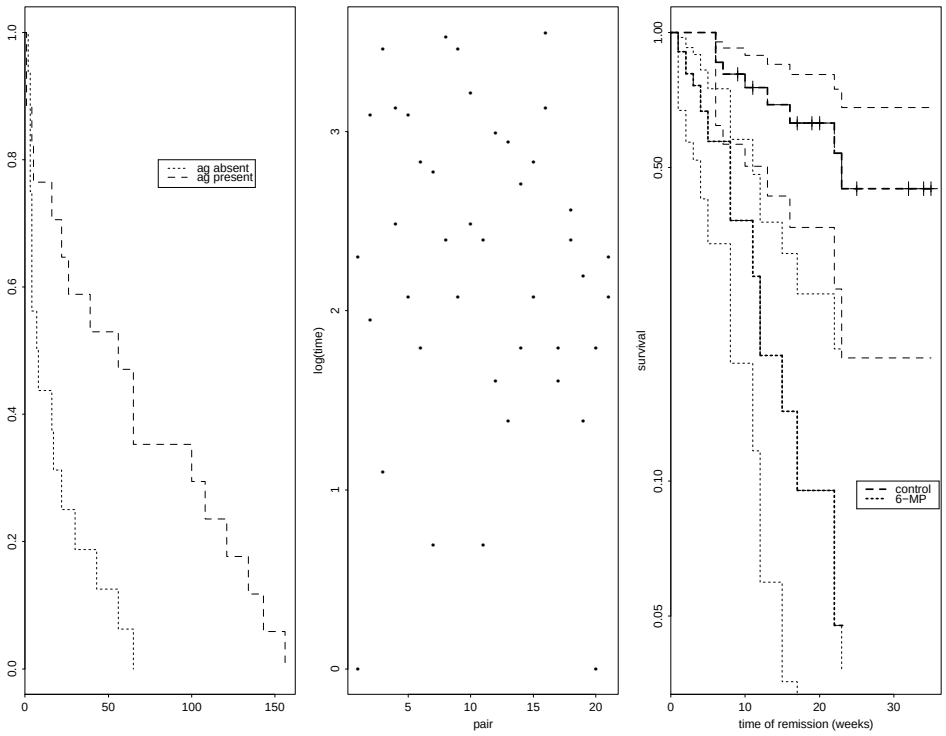


Figure 13.1. Figures for Gehan data

```
> plot(u, lty=2:3)                          # Fig.13.1 (1,1)
> legend(80,0.8,c("ag absent","ag present"),lty=2:3)
> plot(log(time)~pair,data=gehan) # plot(pair,log(time))          # Fig.13.1 (1,2)
> plot(x,conf.int=T,lty=3:2,log=T, xlab="time of remission (weeks)", ylab="survival")
> lines(x,lty=3:2,cex=2,lwd=2)
> legend(25,0.1,c("control","6-MP"),lty=3:2,lwd=2)          # Fig.13.1 (1,3)
```

Question: Is treatment 1 better than treatment 2 ?

It corresponds to a hypothesis:

$$H_0\text{: } S_1 = S_2 \text{ v.s. } H_1\text{: } \begin{cases} S_1(t) \geq S_2(t) & \forall\ t \\ S_1(t) > S_2(t) & \text{for some } t. \end{cases}$$

Splus or R implements several tests using

survdiff(Surv(time,cens)$\sim$ treat, data=gehan,rho=0)

rho=0: log-rank test, a common test.

rho=1 Peto-Peto modification of the Wilcoxon test,

which is more sensitive to early differences in S_is.

```
> survdiff(Surv(time,cens)~treat)
```

	N	$Observed$	$Expected$	$(O-E)^2/E$	$(O-E)^2/V$
$treat = 6-MP$	21	9	19.3	5.46	16.8
$treat = control$	21	21	10.7	9.77	16.8

Chisq= 16.8 on 1 degrees of freedom, p= 4.17e-05

Conclusion ?

§5. Semi-parametric approach.

5.1. Accelerate Lifetime model.

Given $X = x$, if $\ln Y = \beta x + E$, where E is a random variable, then we say Y has accelerate lifetime (AL) model or log linear model. The name AL comes from $Y = e^{\beta x} e^E = e^{\beta x} Z$, where $Z = e^E$.

Weibull, log-normal and log-logistic are all special cases of parametric AL models, with E having an exponential, normal and logistic distribution, respectively.

Note that AL models are special case of linear models.

Let S_o be the survival function of E. If we do not assume that we know S_o, then the set-up becomes a semiparametric one.

With complete data, the LSE is a semiparametric estimator, as it is a solution of

$$\hat{\beta} = argmin_{b \in \mathcal{R}^p} \sum_{i=1}^{n} ||Y_i - bX_i||^2,$$

which does not depend on any parametric distribution.

$$S_{Y|X}(t|x) = P(Y > t|X = x) = P(Ye^{-\beta x} > te^{-\beta x}) = S_Z(t/e^{\beta x})$$

The standard extension of the LSE with RC data is called the Buckley-James estimator (BJE) (1979).

The BJE is often derived by the numerical method proposed by Buckley and James. Yu and Wong (2002) found the method to derive all possible exact solutions to the BJE in finite steps.

We shall not go into the details of the computation of the BJE, but use it as an example of illustration in model justification. The R program for BJE can be searched from the internet.

Example 1.

```
library(MASS), library(survival), library(rms)
> n=200
> b=2
> y=rnorm(n)
> fun1=function(){
    z=sample(1:5,n,replace=T)
    c=runif(n,0,8)
    y=b*z+y
    d=ifelse(y>c, 0,1)                    # d=as.numeric(y<=c)
    m=y*d+c*(1-d)
    f=bj(Surv(m, d) ~ z, link="identity")                    # y = α + βz + ε
    # f=bj(Surv(m, d) ~ z)                  # log y = α + βz + ε
    # f=bj(Surv(m, d) ~ z, link="log")
    return(f)
    }
> fun1()
    Buckley-James Censored Data Regression
    bj(formula = Surv(m, d) ~ z, link = "identity")
```

				Discrimination Indexes	
Obs	200 (=?)	*Regressiond.f.*	1	*g* 3.345	
Events	64 (=?)	*sigma*	0.8657		
		d.f.	62		
	Coef		*S.E.*	*Wald Z*	$Pr(>\|Z\|)$
Intercept	−0.1333	(= 0 ? *why*?)	0.2427	−0.55	0.5827
z	2.0227		0.1264	16.01	< 0.0001

```
> y=rnorm(n)
> fun1()
    No convergence in 50 steps
    Failure in bj.fit
    $fail
    [1] TRUE
> y=rexp(n)
> fun1()
```

				Discrimination Indexes	
Obs	200	*Regressiond.f.*	1	*g* 3.268	
Events	83	*sigma*	0.8054		
		d.f.	81		
	Coef		*S.E.*	*Wald Z*	$Pr(>\|Z\|)$
Intercept	1.0230	(= 0 ? *why*?)	0.1818	5.63	< 0.0001
z	1.9661		0.0741	26.55	< 0.0001

```
> y=runif(n,-1,1)
> fun1()
```

	Coef		*S.E.*	*Wald Z*	$Pr(>\|Z\|)$
Intercept	0.0975	(= 0 ? *why*?)	0.1167	0.84	0.4037
z	1.9524		0.0435	44.88	< 0.0001

How to make use of the BJE to check parametric assumptions ?

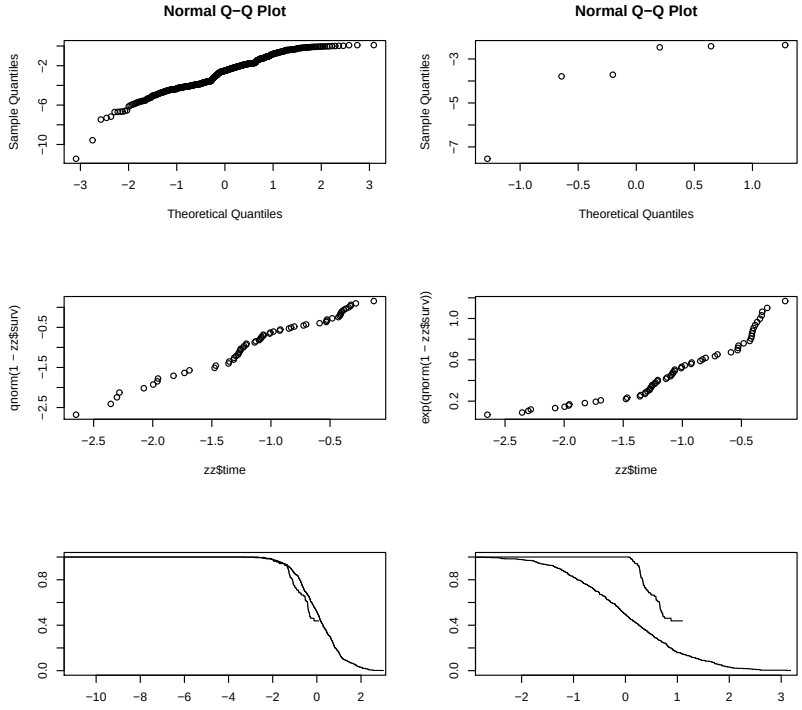
1. Compute the BJE $\hat{\beta}$ based on the RC data (M_i, X_i, δ_i)'s.
2. Derive the "censored residuals": $(M_i - \hat{\beta}' X_i, \delta_i)$'s.
3. Derive the MLE based on the censored residuals and targetting parametric distribution.
4. Compare the KME of the censored residual to the MLE of the targetting parametric cdf with the parameter being the MLE.
5. Test using survdiff() for RC data or ks.test for complete data.

```
> n=500
> b=2
> y=rnorm(n)
> x=sample(1:3,n,replace=T)
> c=runif(n,0, 10)
> y=exp(b*x+y)
> d=ifelse(y>c, 0,1)
> m=y*d+c*(1-d)
    # f= bj(Surv(m, d) ~ x, link="identity")
```

```
> f= bj(Surv(m, d) ~ x)
> f
```

	$Coef$	$S.E.$	$Wald\ Z$	$Pr(>\|Z\|)$
$Intercept$	0.3554	0.1789	1.99	0.0470
x	1.8296	0.1187	15.41	< 0.0001

```
> m=log(m)-f$coef[1]-f$coef[2]*x
> zz=summary(survfit(Surv(m,d)~1))
> par(mfrow=c(3,2))
> qqnorm(m)                    # Anything wrong ?
> qqnorm(log(m))
> plot(zz$time,qnorm(1-zz$surv))                    Why is it correct ?
> plot(zz$time,exp(qnorm(1-zz$surv)))
> M=c(m,rnorm(n))                    # combine two samples as regression data
# zz=survreg(Surv(m,d)~1, dist="gaussian") # if fit other distribution
# M=c(m,rnorm(n,zz$coef[1],zz$scale))
> z=c(rep(1,n),rep(0,n))                    # covariates
> d=c(d,rep(1,n))
> plot(survfit(Surv(M,d) ~z))
> survdiff(Surv(M,d)~z)
    Chisq= 0.2 on 1 degrees of freedom, p= 0.683
> M=c(exp(m),rnorm(n))
> plot(survfit(Surv(M,d) ~z))
> survdiff(Surv(M,d)~z)
    Chisq= 248 on 1 degrees of freedom, p= 0
```



Example 2. (Simulation example). Let $\epsilon \sim Exp(1)$, $X \sim U(0,1,2)$, $\beta = 1$ and $Y = \beta * X + \epsilon$. Then compute the BJE $(\hat{\alpha}, \hat{\beta}) \approx$?

n=10000
c=1.08
e=rexp(n)
x=sample(0:2,n,replace=T)
y=x+e
d=ifelse(y>c, 0,1)
m=y*d+c*(1-d)
bj(Surv(m, d) ∼ x, link="identity")

				Discrimination	*Indexes*
Obs	10000	*Regressiond.f.*	1	*g*	0.467
Events	2470	*sigma*	0.2872		
		d.f.	2468		
	Coef	*S.E.WaldZ*	*Pr*(> |*Z*|)		
Intercept	0.7222	0.0061	117.93	< 0.0001	
z	0.5243	0.0185	28.36	< 0.0001	

(α, β)= ?

Does the results suggest that the BJE is consistent ?

Remark. With RC data, the PLE of $S_X(t)$ is consistent if t is observable, but not $E(X)$.

What value of t is observable in Example 2 ?

Example 3. Based on the output with RC data, derive $\hat{S}(t)$, $\hat{f}(t)$ and $E(X)$.

> summary(survfit(Surv(m,d)∼1))

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower*95%*CI*	*upper*95%*CI*
0	1000	313	0.687	0.0147	0.659	0.716
1	687	371	0.316	0.0147	0.288	0.346

Possible observations: (M_i, δ_i) = ?

$(0,1)$, $(1,1)$, ... ?

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < 0 \\ 0.687 & \text{if } t \in [0,1) \\ 0.316 & \text{if } t \geq 1 \end{cases}$$

$\hat{f}(t) = 0.313\mathbf{1}(t=0) + 0.371\mathbf{1}(t=1)$ **Why ?**

Is it correct ?

$\hat{f}(t) = 0.313\mathbf{1}(t=0) + 0.371\mathbf{1}(t=1) + 0.316\mathbf{1}(t=\infty)$ **Why ?**

a degenerated distribution, that is, $P(X = \pm\infty) > 0$.

$\hat{\mu}_X$ = ?

Ancient way is to treat RC data as complete data.

Then $\hat{S}_M(t) = \begin{cases} 1 & \text{if } t < 0 \\ \frac{1}{n}\sum_{i=1}^n \mathbf{1}(M_i > 0) & \text{if } t \in [0,1) \\ 0 & \text{if } t \in [1,\infty) \end{cases}$

$\hat{f}_M(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(M_i = t) = a\mathbf{1}(t=0) + (1-a)\mathbf{1}(t=1)$, where a = ?

$\tilde{\mu}_X$ = ?

The data are generated by simulation.

> n=1000

> b=1

> c=1

> x=sample(c(0,1,2),n,replace=T)

> d=ifelse(x>c, 0,1)

> m=x*d+c*(1-d)

The model is $X \sim U(0,1,2)$, $C = 1$, $X \perp C$.

In general, under this model, the PLE $\hat{S}(t)$

$= \prod_{j=0}^{1} \mathbf{1}(t > j)(1 - \frac{d_j}{r_j})$

$$= \begin{cases} 1 & \text{if } t < 0 \\ \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i > 0) & \text{if } t \in [0,1) \\ \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i > 1) & \text{if } t \in [1,\infty) \end{cases}$$

$\overset{a.s}{\to} S_X(t)\mathbf{1}(t \le 1) + S_X(1)\mathbf{1}(t > 1)$

$$= \begin{cases} 1 & \text{if } t < 0 \\ 2/3 & \text{if } t \in [0,1) \\ 1/3 & \text{if } t \ge 1 \end{cases}.$$

$\hat{S}(2) \overset{a.s}{\to} S_X(2)$?

Can we get a consistent estimator of $E(X)$?

$E(X) = ?$ 1

$E(M) = ?$ 1-a

Does this example suggest that the estimates are consistent ?

Remark. Under right censoring $M = Y \wedge C$ and $\delta = \mathbf{1}(Y \le C)$, where $Y \perp C$, let $\mathcal{R}_Y$ and $\mathcal{R}_M$ be the ranges of Y and M, respectively,

If $\mathcal{R}_M \neq \mathcal{R}_Y$ and $\mathcal{R}_M \cap \mathcal{R}_Y$ contains a non-empty open set, then

$\not\exists$ consistent PLE of $S_Y(t)$ $(\forall\ t)$ and $E(Y)$.

$\not\exists$ consistent BJE of $S_{Y|X}(t|\cdot)$ $(\forall\ t)$ and $E(Y|X)$.

$\exists$ consistent MLE of $S_Y(t)$ and $E(Y)$.

Example 4. Let $X \sim bin(3,p)$ and $C = 0$. MLE of $p = ?$

$\mathcal{L} = \prod_{i=1}^{n} (f(M_i))^{\delta_i} (S(M_i))^{1-\delta_i}$

$= (\binom{3}{0} p^0 q^{3-0})^m (1 - \binom{3}{0} p^0 q^{3-0})^{n-m}$

$= (q^3)^m (1 - q^3)^{n-m}$

where $m = \sum_{i=1}^{n} \mathbf{1}(M_i = 0, \delta_i = 1)$

$= \sum_{i=1}^{n} \mathbf{1}(\delta_i = 1)$.

$\hat{q^3} = \frac{m}{n}$

$\overset{a.s.}{\to} P(\delta = 1)$

$= P(X = 0 \le C)$

$= P(X = 0) = q^3$,

$\hat{q} = (m/n)^{1/3}$, and $\hat{p} = 1 - \hat{q}$.

Thus $\hat{\mu} = 3\hat{p} \to 3p$? **Is it correct ?**

Q: Why is the parametric approach is more preferable than the non-parametric approach ?

Example 5. Suppose that

$X \sim bin(3, p)$,
$C \sim U(\{0.5, 1.5\})$,
$X \perp C$,
$M = X \wedge C$ and
$\delta = \mathbf{1}(X \le C)$.

Given n RC data (M_i, δ_i)'s, find

the NPMLE's of S_X and f_X,
the MLE's of p, S_X and f_X.

Sol. $(M_i, \delta_i) = ?$

How many distinct types of (M_i, δ_i) ?

$$\hat{S}(t) = \prod_{i:M_{(i)} \le t} (1 - \frac{\delta_{(i)}}{n-i+1}) = \prod_{t_j \le t} (1 - \frac{d_j}{r_j})$$

$X:$	0		1		2	3
$C:$		0.5		1.5		
	M_1	M_2	M_3	M_4		
$\#\ of\ M_i:$	m_0	$m_{0.5}$	m_1	$m_{1.5}$		
$d_j:$	?	?	?	?		
$r_j:$	n		$m_1 + m_{1.5}$			

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < 0 \\ 1 - m_0/n & \text{if } t \in [0, 1) \\ (1 - m_0/n)(1 - \frac{m_1}{m_1 + m_{1.5}}) & \text{if } t \ge 1. \end{cases}$$

$$\hat{f}(t) = \begin{cases} m_0/n & \text{if } t = 0 \\ (1 - m_0/n)\frac{m_1}{m_1 + m_{1.5}} & \text{if } t = 1 \\ ? & \text{if } t = ? \end{cases} = \begin{cases} m_0/n & \text{if } t = 0 \\ m_{0.5}/n & \text{if } t = 0.5 \\ m_1/n & \text{if } t = 1 \\ \cdots \end{cases} \quad \textbf{???}$$

$$\begin{aligned} &\mathcal{L} \\ =&\prod_{i=1}^{n} (f(M_i))^{\delta_i} (S(M_i))^{1-\delta_i} \\ =&\left(\binom{3}{0} p^0 q^{3-0}\right)^{m_0} \left(1 - \binom{3}{0} p^0 q^3\right)^{m_{0.5}} \left(\binom{3}{1} p^1 q^{3-1}\right)^{m_1} \left(1 - \binom{3}{0} p^0 q^3 - \binom{3}{1} p^1 q^2\right)^{m_{1.5}} \\ =&\left((1-p)^3\right)^{m_0} \left(1 - (1-p)^3\right)^{m_{0.5}} \left(3p^1(1-p)^2\right)^{m_1} \left(1 - (1-p)^3 - 3p^1(1-p)^2\right)^{m_{1.5}} \end{aligned}$$

where $m_j = \sum_{i=1}^{n} \mathbf{1}(M_i = j)$, $j \in \{0, 0.5, 1, 1.5\}$

How to find the MLE of p ?

Generate a data set from the model with $n = 40$, find numerically.

```
n=1000
p=(0:n)/n
f=L(p)                    Is it an R code ?
p[f==max(f)]
```

§5.2. Proportional Hazards Model.

Def. Define $\tau = \tau_T = \sup\{t : F_T(t) < 1\}$ for a random variable T.
Let (X, Y) be a random vector.
The baseline hazards function is $h_o(t) = \frac{f_o(t)}{S_o(t-)}$.
If for each given $X = x$, the conditional hazard of Y is

$$h(y|x) = h_o(y)c(x), \text{ for } y < \tau_C, \text{ where } h_o \text{ is an hazard and } c(x) \geq 0, \tag{1}$$

where C is the censoring variable, then (X, Y) is said to be from a proportional hazards (PH) model or Cox's regression model.
If for each given $X = x$, the conditional survival function of Y satisfies

$$S_{Y|X}(t|x) = (S_o(t))^{c(x)} \text{ where } S_o \text{ is a survival function and } c(x) \geq 0,$$

then we say (X, Y) is from a Lehmann family or Lehmann model.
Q: Why do we need Cox' model, the Lehmann model and the linear regression model ?
Example 1. If $S_o(t)$ is a survival function of continuous random variable, then

$$S(t|x) = (S_o(t))^{c(x)} \text{ satisfies the PH model.} \tag{2}$$

Reason: $f(t|x) = -c(x)(S_o(t))^{c(x)-1}S_o'(t) = c(x)S(t|x)\frac{f_o(t)}{S_o(t)}$,
$h(t|x) = \frac{f(t|x)}{S(t|x)} = c(x)h_o(t)$.
Special cases:
a. Weibull: $S_o(t) = e^{-t^\gamma}$, $t > 0$.
$S(t|x) = exp(-e^{\beta x}t^\gamma)$.
$h(t|x) = e^{\beta x}\gamma t^{\gamma-1}$.
b. logistic: $S_o(t) = \frac{1}{1+e^t}$.
$S(t|x) = (\frac{1}{1+e^t})^{e^{\beta x}}$.
Remark 1. It is common to set $c(x) = e^{\beta x}$ so that $c(x) \geq 0$.
If one assume h_o is known, then the PH model is just a parametric problem.
β can be estimated by survreg() or other numerical methods.
If one assume h_o is unknown, then the PH model becomes a semiparametric model:
$h(t|x) = h_o(t)e^{\beta x}$ with β and h_o unknown.
Note that $h_o(\cdot)$ is a function on $\mathcal{R}^1$, thus it is non-parametric. On the other hand β can be viewed as a parameter. Thus it is a semi-parametric approach.

The semi-parametric MLE of (β, h_o) or (β, S_o) is to find a (b, S) such that it maximizes

$$L(b, S) = \prod_{i=1}^{n}(f(M_i|X_i))^{\delta_i}(S(M_i|X_i))^{1-\delta_i}, \ b \in \mathcal{R}^p \text{ and } S(\cdot) \text{ is a survival function.}$$

This approach involved both b and S.
Cox proposes a smart partial likelihood MLE approach:
(a) first estimate β by maximizing the partial likelihood $\mathcal{L}(b)$ (in (3)) over all $b \in \mathcal{R}^p$,
(b) then estimate h_o or S_o nonparametrically.

$$\mathcal{L}(b) = \prod_{i=1}^{n}(\frac{e^{\beta X_i}}{\sum_{j:Y_j \geq Y_i} e^{\beta X_j}})^{\delta_i}. \tag{3}$$

Notice that $L \not\propto \mathcal{L}$ and it does not involves S and f.
It is implemented in Splus or R by

coxph(Surv(time,cens)∼treat,data=gehan)

coxph(Surv(time,cens)∼treat,data=gehan,method="exact") # for ties.

which presents $\hat{\beta}$, default method is "efron" for the case of no ties.
Cox proposes an estimator for S_o: $\hat{S}_o(t) = exp(-\sum_{j:\ M_j \le t} \frac{\delta_j}{\sum_{M_h \ge M_j)} e^{\hat{\beta} x_h}})$.

It is implemented in Splus or R by

survfit()

Example 2. Motors data.

> motors[1:3,]

	temp	*time*	*cens*
1	150	8064	0
11	170	1764	1
26	190	1680	0

Consider Cox regression model $S(t|temp) = (S_o(t))^{e^{\beta temp}}$.

1. Find the estimate of β.
2. Draw the survival function at temp=200 with CI.
3. $S(3000|200)$ =?
4. $median(time|temp = 200)$ = ? **Why not** $E(time|temp = 200)$ **?**
5. Test whether there is an effect of temperature on the survival

   That is, H_0: $\beta = 0$ v.s. H_1: $\beta \neq 0$.

Sol: Splus or R commands:

> x=coxph(Surv(time,cens)∼temp,motors,method="efron")

	coef	*exp(coef)*	*se(coef)*	*z*	*p*
temp	0.0919	1.1	0.0274	3.36	0.00079

Likelihood ratio test=25.6 on 1 df, p=4.3e-07 n= 40, number of events= 17

> (y=survfit(x,newdata=data.frame(temp=200), conf.type="log-log"))

records	*n.max*	*n.start*	*events*	*median*	0.95*LCL*	0.95*UCL*
40	40	40	17	1440	1344	3444

> summary(y)

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower*95%*CI*	*upper*95%*CI*
408	40	4	$9.38e-1$	$4.19e-2$	$7.78e-01$	0.984
...						
1344	28	2	$5.14e-1$	$2.12e-1$	$1.06e-01$	0.821
1440	26	1	$3.74e-1$	$2.15e-1$	$4.51e-02$	0.731
1764	20	1	$9.62e-2$	$2.01e-1$	$1.50e-06$	0.665
2772	19	1	$2.18e-2$	$7.67e-2$	$8.40e-11$	0.532
3444	18	1	$4.23e-3$	$2.15e-2$	$2.13e-15$	0.413
3542	17	1	$6.78e-4$	$4.63e-3$	$1.43e-20$	0.312
...						

1. The estimate of β: $\hat{\beta} = 0.09$.
2. plotting:
   Splus or R commands:
   plot(survfit(x,newdata=data.frame(temp=200), conf.type="log-log"))
3. $S(3000|200) \approx 0.0218$ **Why ?**
4. $median(time|temp = 200) \approx 1440$.

5. Reject H_0 as P-value = 0 (see LRT test).
   For testing, we use likelihood ratio test $\phi = \mathbf{1}(-2\ln\lambda > \chi^2_{1,\alpha})$. where

$$\lambda = \frac{\prod_{i=1}^n (\hat{f}(M_i))^{\delta_i} (\hat{S}(M_i))^{1-\delta_i}}{\prod_{i=1}^n (\tilde{f}(M_i))^{\delta_i} (\tilde{S}(M_i))^{1-\delta_i}}$$

   $\hat{S}$ is PLE and $\tilde{S}$ is estimate under the PH model.

Question: Suppose that we make parametric assumption, say S_o is a weibull distribution.
   Then how to use Splus or R to settle the previous 5 questions?

Sol.:

> x=survreg(Surv(time,cens)∼temp,data=motors)
> summary(x)

	$Value$	$Std.$ $Error$	z	p
$(Intercept)$	16.3185	0.62296	26.2	$3.03e-151$
$temp$	−0.0453	0.00319	−14.2	$6.74e-46$
$Log(scale)$	−1.0956	0.21480	−5.1	$3.38e-07$

Scale= 0.334
Weibull distribution
Loglik(model)= -147.4 Loglik(intercept only)= -169.5
Chisq= 44.32 on 1 degrees of freedom, p= 2.8e-11
n= 40

Under the Weibull distribution, Splus or R assumes the model: Given $X = x$,

$$\ln Y = \beta_0 + \beta_1 x + \sigma \ln E, \text{ where } X \perp E \text{ and } E \sim exp(1).$$

That is $Y = e^{\beta_0+\beta_1 x} E^\sigma$ or $E = e^{\frac{\beta_0+\beta_1 x}{-\sigma}} Y^{1/\sigma}$. It follows that

$$\begin{aligned} S_{Y|X}(y|x) = P\{Y > y|X = x\} &= P\{Y^{1/\sigma} e^{\frac{-\beta_0-\beta_1 x}{\sigma}} > y^{1/\sigma} e^{\frac{\beta_0+\beta_1 x}{-\sigma}} | X = x\} \\ &= P\{E > y^{1/\sigma} e^{\frac{-\beta_0-\beta_1 x}{\sigma}} | X = x\} \\ &= e^{-y^{1/\sigma} e^{\frac{-\beta_0-\beta_1 x}{\sigma}}} \\ &= e^{-y^{1/\sigma} / e^{\frac{\beta_0+\beta_1 x}{\sigma}}} \end{aligned} \quad (1)$$

1. From survreg(), we have $\hat{\beta}_0 = 16.3$, $\hat{\beta}_1 = -0.0453$ and $\hat{\sigma} = 0.334$.
2. plot$(y, \hat{S}(y|200))$ with $y = 1 : 9000$ (see Eq. (1)).
3. Thus $\hat{S}(3000|200) = exp(-3000^{1/0.334} / e^{\frac{16.3-0.0453\times 200}{0.334}})$
4. We can estimate both $E(Y|X = x)$ and median$(Y|X = x)$.
   $E(Y|X = x) = \theta^{\frac{1}{\gamma}} \Gamma(1 + \frac{1}{\gamma})$ by the textbook, where $S(t) = exp(-t^\gamma/\theta) = \exp(-(\frac{t}{\eta})^\gamma)$.
   By comparison, $\gamma = 1/\sigma$, $\theta = e^{\frac{\beta_0+\beta_1 x}{\sigma}}$.
   It leads to $\hat{E}(Y|X = x)$.
   The median can be solved by the equation

$$e^{-y^{1/\hat{\sigma}} / e^{\frac{\hat{\beta}_0+\hat{\beta}_1 x}{\hat{\sigma}}}} \Big|_{x=200} = 0.5$$

5. P-values =0 from survreg(), reject H_0.

Correction 1 in a Remark. What does predict() estimate ?
zz=survreg(Surv(y)∼x, dist="")
predict(zz,data.frame(x=3,se=T))

$$= \begin{cases} \hat{\alpha}+3\hat{\beta} & \text{if } Y=\alpha+\beta X+\sigma Z \text{ (logis or guassian)} \\ e^{\hat{\alpha}+3\hat{\beta}} & \text{if } \ln Y=\alpha+\beta X+\tau Z \text{ (log-normal or loglogis, or Exp)} \\ \underbrace{\hat{\tau}\Gamma(1+exp(-(\hat{\alpha}+3\hat{\beta}))}_{\text{a mistake, delete it}} & \underbrace{if}_{delete\ it} \text{ weibull} \end{cases}$$

$$=\text{estimate of } \begin{cases} E(Y|X=3) & \text{if logis or guassian or Exp() or weibull} \\ exp(E(\log Y|X=3)) & \text{if (log-normal or loglogis)} \end{cases}$$

$\neq$ estimate of $E(Y|X=3)$.

Remark. What does predict() estimate ?
zz=survreg(Surv(y)∼x, dist="")
predict(zz,data.frame(x=3),se=T))

$$= \begin{cases} \hat{\alpha}+3\hat{\beta} & \text{if } Y=\alpha+\beta X+\sigma Z \text{ (logis or guassian)} \\ e^{\hat{\alpha}+3\hat{\beta}} & \text{if } \ln Y=\alpha+\beta X+\tau Z \text{ log-normal or loglogis, or Exp, or weibull} \end{cases}$$

$$=\text{estimate of } \begin{cases} E(Y|X=3) & \text{if logis or guassian or Exp()} \\ exp(E(\log Y|X=3)) & \text{if log-normal or loglogis or weibull or Exp(),} \end{cases}$$

$\neq$ estimate of $E(Y|X=3)$.

Correction 2 in Homework Solution 14.

4. Carry out the following simulation study:
Generate 200 RC observations from Weibull with $S_{Y|X}(t|x) = exp(-exp(4x)t^2)$, $t>0$, and with censoring rate around 50% and x from bin(1,0.3).
n=200
x=rbinom(n,1,0.3)
y=rexp(n)
y=exp(4*x+0.5*log(y))

It is observations from Weibull with $S_{Y|X}(t|x) = exp(-(t/exp(4x))^2)$, $t>0$, not

$$S_{Y|X}(t|x) = exp(-exp(4x)t^2) = \exp(-(\frac{t}{e^{-2x}})^2),\ t>0.$$

y=exp(-2*x+0.5*log(y))

Remark 2. Cox's model is often specified by

$$h_{Y|X}(t|x) = e^{\beta' x} h_o(t),\ t < \tau = \tau_C.$$

If the random variable is discrete, then the choice of $c(x) = e^{\beta x}$ may cause problem, *e.g.*,
Example 3. Let $f_o(y) = 0.5\mathbf{1}(y \in \{0, 1\})$, and $c(x) = 3$. Then
$h_o(0) = \frac{f_o(0)}{S_o(0-)} = 0.5$,
$P(Y = y|Y \geq y, X = x)$
$= h(y|x) = h_o(y)c(x)$
> 1 **Anything wrong ?**
If we choose $c(x) = exp(-e^{\beta x})$, it ensures that $c(x)$ is between 0 and 1.
For continuous random variable, we only need $c(x) \geq 0$,
as the PH model $S_{Y|X}(y|x) = (S_o(y))^{c(x)} \in [0, 1]$ for all $c(x) \geq 0$.
Remark 3. In the original definition of PH model, $y < \tau$ is omitted.
We show in Example 4 that if h_o corresponds to a discrete random variable, the statement
$h_{Y|X}(t|x) = c(x)h_o(t)$, $t < \tau$ $(= \tau_C)$,
without the restriction does not define a hazard function.
Example 4. If T is discrete at τ, then

$$h_T(\tau) = f_T(\tau)/S_T(\tau-) = f_T(\tau)/f_T(\tau) = 1$$

which is always true. It follows that
$h(\tau|x) = h_o(\tau)c(x)$ does not hold at τ, as

$$h(\tau|x) = 1 \neq 1 \times c(x) = h_o(\tau)c(x) \text{ (how about } c(x) = 1) \text{ ?}$$

It does not matter for continuous random variables, as one can define
$f(t) = 0$ for $t \geq \tau$.
Example 5. For continuous random variables, $S_{Y|X}(\cdot|x) = (S_o(\cdot))^{e^{\beta x}}$ <=> Cox's model.
The statement does not hold for discrete random variable. When Cox proposes the PH model, he distinguishes it from the Lehmann model. However, later in the literature, the Lehmann model is identified as the PH model. To see that they are different, consider the following example.
Suppose $Y_0 \sim bin(2, p)$. Then its hazard function is

$$h_o(t) = \begin{cases} (1-p)^2 & \text{if } t = 0, \\ \frac{2(1-p)p}{1-(1-p)^2} & \text{if } t = 1, \\ p^2/p^2 & \text{if } t = 2, \end{cases} = \begin{cases} (1-p)^2 & \text{if } t = 0, \\ \frac{2(1-p)}{2-p} & \text{if } t = 1, \\ 1 & \text{if } t = 2, \end{cases}$$

Suppose $h(y|x) = h_{Y|X}(y|x) = h_o(y)exp(-e^{\beta x})$ for $y = 0$ or 1.
Then
$h(0|x) = (1-p)^2 e^{-e^{\beta x}}$ yields $f(0|x) = (1-p)^2 e^{-e^{\beta x}}$ as $S(0-|x) = 1$, where $S(y|x) = S_{Y|X}(y|x)$.
$h(1|x) = \frac{2(1-p)}{2-p} e^{-e^{\beta x}}$ yields $f(1|x) = \frac{2(1-p)}{2-p} e^{-e^{\beta x}}(1 - (1-p)^2 e^{-e^{\beta x}})$
$f(2|x) = 1 - f(0|x) - f(1|x)$.
It follows that

$$S(1|x) = 1 - (1-p)^2 e^{-e^{\beta x}}.$$

However,

$$(S_o(1))^{e^{-e^{\beta x}}} \approx 0.808 \neq 0.7360219 \approx \left(1-(1-p)^2\right)^{e^{-e^{\beta x}}},$$

if $p = 0.2$ and $e^{-e^{\beta x}} = 0.3$,

This example indicates that if h_o is a hazard function of a discrete random variable, and $h(t|x) = h_o(t)e^{-e^{\beta x}}$, its cdf is not of the form

$S(t|x) = (S_o(t))^{e^{-e^{\beta x}}}$.

$S(t|x) = (S_o(t))^{e^{-e^{\beta x}}}$ defines a class of survival functions, even in the case that S_o is discrete. In particular,

$$H(t|x) = -\ln S(t|x) = e^{-e^{\beta x}}(-\ln S_o(t)) = e^{-e^{\beta x}} H_o(t).$$

Thus one can say that it follows the proportional cumulative hazards (PIH) model.

Remark. In linear regressions, we often have $\beta'\mathbf{X} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, that is, with the intercept term. Then the error term ϵ has mean zero. However, in PH models, we only have $\beta'\mathbf{X} = \beta_1 X_1 + \cdots + \beta_p X_p$, that is, without the intercept term. In fact

$(S_1(t))^{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} = (S_1(t)^{e^{\beta_0}})^{e^{\beta_1 X_1+\cdots+\beta_p X_p}} = (S_o(t))^{e^{\beta_1 X_1+\cdots+\beta_p X_p}}$,

with $(S_1(t)^{e^{\beta_0}} = S_o(t)$.