Data Analysis (534)

Please turn on your cameras.

Textbook: Modern Applied Statistics with S, 4th ed.
by Venables and Ripley
Office: WH 132
Office hours: M Tu, 3:00pm-4:00pm
Classroom: On-line 1:10-2:10pm
Homework due: Wednesday before class.
Email me at qyu@math.binghamton.edu before 1:10pm on Wednesday.
Grading policy: 40% homework+10% quizzes + 20% midterm+30% final.
B = 75 $\pm$
Midterm: Mar. 29 (M)
Final: May 24-26
You can bring one page with R commands and formulas in exams.
During quizzes or exams, use another camera that can show your table, screen and you.
Quiz: Once a week at a random day,
quiz problems: formulas for Math 447-448 (see my website)
right after quiz, take a picture of your answer, output as a pdf file and email me.
Homework assigned during a week is due next Wednesday.
It is on my website: http://www.math.binghamton.edu/qyu/qyu_personal
Remind me if you do not see it by Saturday morning !
The lecture note is also on my website
http://www.math.binghamton.edu/qyu/qyu_personal
note and note2 are updated one,

Chapter 0. Introduction.

Data analysis is to teach how to analyze data (using R program). 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 (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, *e.g.* $P(Y \in A | X = x) = ?$.

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.

You can find R download site through Google.

Or login to department computer.

ssh qyu@ssh1.math.binghamton.edu (ssh2, ssh3)

sftp qyu@ssh1.math.binghamton.edu

```
> 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:

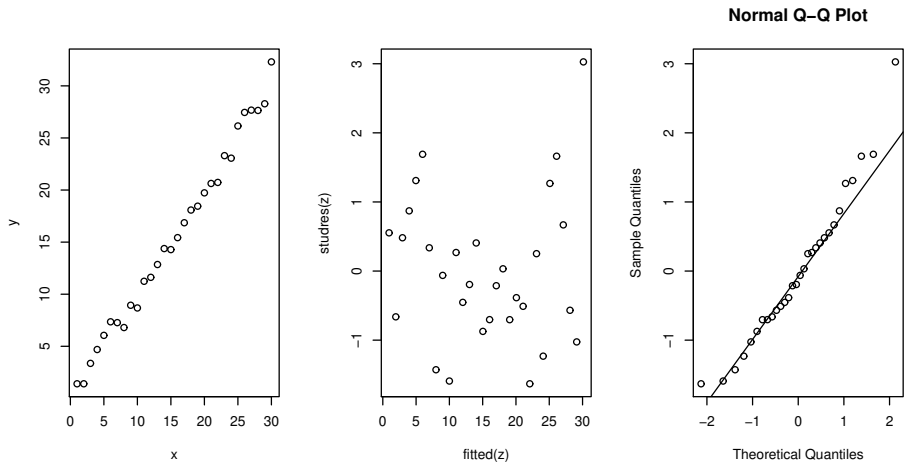
	Estimate	*Std. Error*	*t value*	$Pr(>\|t\|)$	
(*Intercept*)	−0.06299	0.32684	−0.193	0.849	—
x	1.00636	0.01841	54.663	$< 2e-16$	***

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),

attach relevant R outputs into your report (see the sample homework I email you).

Some commands in the linex system:

tex report.tex (create report.dvi file) (or latex ..., or pdflatex ...)

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 (do not compress them):

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 given in Math 531

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

A sample of homework was emailed to you. Mimic it in your homework.

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 ?**

survival function $S(t) = 1 - F(t)$.

Example 1. $X \sim$ Weibull distribution with cdf $F(x|\gamma, \tau) = 1 - \exp(-(x/\tau)^\gamma), \ x > 0$, $S(x|\gamma, \tau) = \exp(-(x/\tau)^\gamma \mathcal{I}(x > 0))$ and $E(X) = \int x f(x) dx = \tau\Gamma(1 + 1/\gamma))$

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*	$f(x;\theta)$
beta	*beta*	$shape1, shape2$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \ x \in (0,1)$
uniform	*unif*	min, max	$\frac{1}{b-a}, \ x \in (a,b)$
gamma	*gamma*	$shape, scale$	$\frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}, \ x, \ \alpha, \ \beta > 0$
exponential	*exp*	*rate*	$\rho e^{-\rho x}, \ x > 0$
chi − square	*chisq*	*df*	
Cauchy	*cauchy*	$location, scale$	$\frac{1}{\pi(1+x^2)} \to \frac{1}{\beta} f(\frac{x-\alpha}{\beta})$
binomial	*binom*	$size, prob$	$\binom{n}{x} p^x (1-p)^{n-x}, \ x \in \{0, 1, ..., n\}$
negative binomial	*nbinom*	$size, prob$	
geometric	*geom*	*prob*	$p(1-p)^x, \ x = 0, 1, ...$
hypergeometric	*hyper*	m, n, k	
normal	*norm*	$mean, sd$	
log − normal	*lnorm*	$meanlog, sdlog$	
F	*f*	$df1, df2$	
T	*t*	*df*	
logistic	*logis*	$location, scale$	
Poisson	*pois*	*lambda*	
Weibull	*weibull*	$shape, scale$	
Wilcox	*wilcox*	m, n	

Example 1 (contitued).

R

> x=rweibull(100,1,5)

> 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 $\tilde{F}$ is a step function, and
   $\tilde{F}(X_{(i)}) = \frac{i}{n}$ (ecdf), or $\frac{i-\frac{1}{2}}{n}$ (ppoints(x)), or $\frac{i}{n+1}$.
   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.
It is desirable to do parameteric analysis, say assume that 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)                    different from summary(lm())
          Length  Class   Mode
estimate  2       -none-  numeric
sd        2       -none-  numeric
vcov      4       -none-  numeric
loglik    1       -none-  numeric   #loglikelihood
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, 2.776,3.3746)-pweibull(1, 2.776,3.3746)) # $P(X \in (1,2])$

[1] 0.1750694

> 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

Ans: The MLEs under the Weibull model are $\hat{\tau} = 3.4$ with $\hat{\sigma}_{\hat{\tau}} = 0.13$ and $\hat{\gamma} = 2.8$ with $\hat{\sigma}_{\hat{\gamma}} = 0.23$.
$\hat{F}(t) = 1 - \exp(-(t/3.4)^{2.8})$, $t > 0$ and $\hat{P}(X \in (1,2]) \approx 0.175$.
$\hat{E}(X) = \hat{\tau}\Gamma(1 + 1/\hat{\gamma}) \approx 3.004$ verses $\overline{X} = 2.992$.

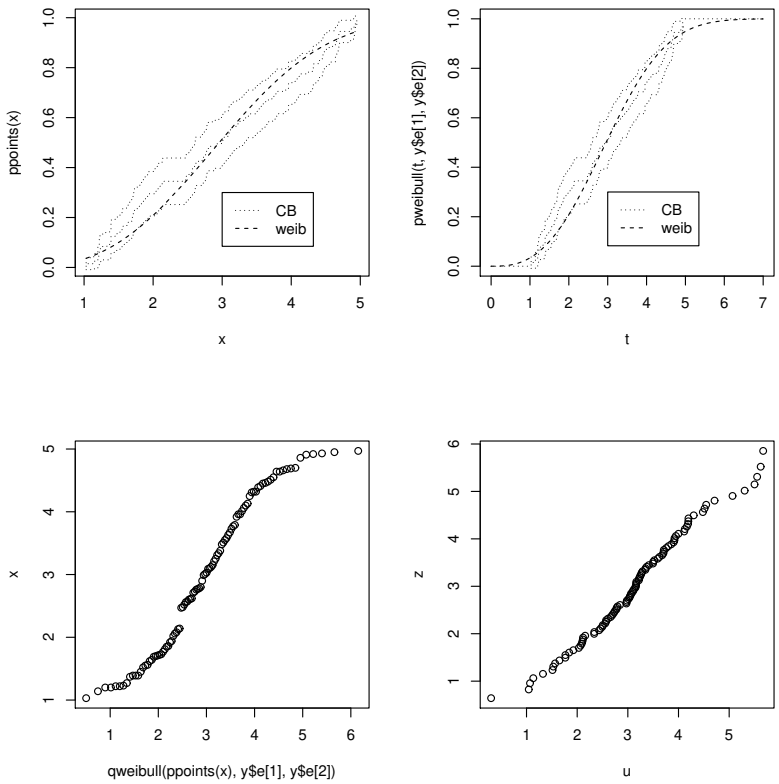
Question:

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. $\hat{E}(X) = \hat{\tau}\Gamma(1 + 1/\hat{\gamma}) \approx 3.004$ is the MLE of $E(X)$ and $\overline{X} = 2.992$ is the non-parametric estimator of $E(X)$ ($\overline{X} = \sum_i x\hat{f}(x)$, where $\hat{f}(x) = \sum_{i=1}^n \mathbf{1}(X_i = x)/n$ is the density of the edf $\hat{F}(x) = \sum_{i=1}^n \mathbf{1}(X_i \leq x)/n$, a non-parametric MLE (NPMLE) of $F_X(t)$. **Which is better ?**
3. Is the model assumption valid ?
   We can use the qqplot, confidence band (CB) of the edf and ks.test to check.
   CB of the edf is the pointwise confidence interval based on the edf.

Example of qqplot and CB codes:

```
makepsfile = function() {
ps.options(horizontal = F)
ps.options(height=8.0, width=7.5)
postscript("ch1.2.ps")
par(mfrow =c(2,2))
x=sort(x)
plot(x,pweibull(x,y$e[1],y$e[2]),type="l",lty=2)
lines(x,ppoints(x),type="S",lty=3)
s=1.96*sqrt(ppoints(x)*(1-ppoints(x))/100)
lines(x,ppoints(x)+s,type="S",lty=3)
lines(x,ppoints(x)-s,type="S",lty=3)
leg.names=c("CB", "weib")
```

```
    legend(3, 0.3, leg.names, lty=c(3,2),cex=1.0)
    t=(0:70)/10
    plot(t,pweibull(t,y$e[1],y$e[2]),type="l",lty=2)
    lines(x,ppoints(x),type="S",lty=3)
    s=1.96*sqrt(ppoints(x)*(1-ppoints(x))/100)
    lines(x,ppoints(x)+s,type="S",lty=3)
    lines(x,ppoints(x)-s,type="S",lty=3)
    lines(c(0,1),c(0,0),type="l",lty=3)
    lines(c(5,7),c(1,1),type="l",lty=3)
    leg.names=c("CB", "weib")
    legend(3, 0.3, leg.names, lty=c(3,2),cex=1.0)
    u=rweibull(100,y$e[1],y$e[2])
    plot(qweibull(ppoints(x),y$e[1],y$e[2]),x) # or qqplot(u,x) compare weibull to data
    z=qweibull((1:100)/101,y$e[1],y$e[2])
    qqplot(u,z) # compare to qqplot weibull v.s. weibull
    dev.off()
    }
    makepsfile()
> pweibull(1,y$e[1],y$e[2]) +1-pweibull(5,y$e[1],y$e[2])
    [1] 0.08444813
```



$Fig(1,1):$ *similar to next* $Fig(1,2):$ *cdf of weibull v.s. edf with its confidence band*
$Fig(2,1):$ *qqplot weibull* $Fig(2,2):$ *qqplot of* 100 *data from Weibull*

Figure 5.1.

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 one sided test $\mathbf{1}(D > c)$, where $D = \sup\{|\tilde{F}(t) - F_o(t)| : t \in R\}$.
(can we use a two-sided test $\mathbf{1}(D \notin [c_1, c_2])$?
Remark. P-value= $P\{D > D_o\}$ (given in R),

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, the P-value is asymptotically true. Thus it is assumed that

(1) θ is the true value and

(2) the sample size n is very large.

However, θ is estimated by its MLE here and n is not large, this changed its true P-value.
One can find the critical value in D by empirical quantiles of 0.05 for a given sample size n.

(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) with ks(). **Why uniform of U(1,5) ?**
Question: What is the difference between the last two tests ?
Summarize the findings.
How to find the MLE of the parameter θ for:

Weibull ? **θ ? true value of θ ?**

Uniform ? **θ ? true value of θ ?**

> (y=fitdistr(x,"unif"))

Error in fitdistr(x, "unif") : unsupported distribution

MLE of U(a,b) ?

Uniform(1,5) ?

```
> 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)) }
```

> fun3(100) # **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)
}
>c(mean(u), mean(v), mean(w))
[1] 1 0.05 0.053   # Are they expected ?
```

Power or size of the test ϕ ? Or an estimate ?)

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

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. OW, $P(H_0|H_1)$ can be 99%, instead of $< 50\%$, 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 an approximation when n is very large. Otherwise, it is arbitrary.

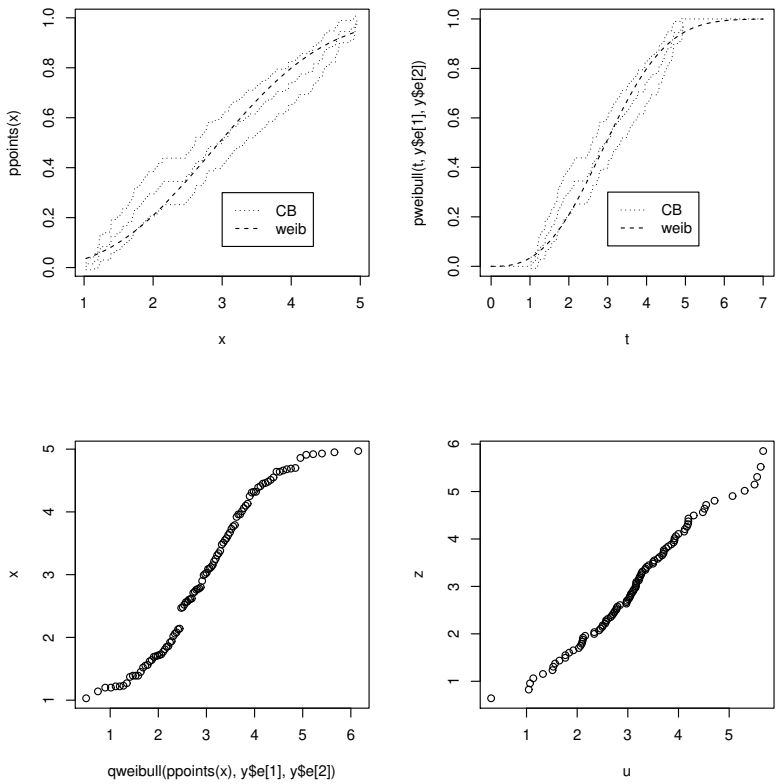
Example 2. Given X_1, ..., X_{100}, 100 observations in the file data_ex2.
It is desirable to do parameteric analysis, say assume that they are from a Weibull distribution. $F(x|\gamma,\tau) = 1 - \exp(-(x/\tau)^\gamma),\ x > 0$
Estimate F and $P\{X \in (1,2]\}$ and $E(X)$.

Ans: The MLEs under the Weibull model are $\hat{\tau} = 3.4$ with $\hat{\sigma}_{\hat{\tau}} = 0.13$ and $\hat{\gamma} = 2.8$ with $\hat{\sigma}_{\hat{\gamma}} = 0.23$.
$\hat{F}(t) = 1 - \exp(-(t/3.4)^{2.8})$, $t > 0$ and $\hat{P}(X \in (1,2]) \approx 0.175$.
$\hat{E}(X) = \hat{\tau}\Gamma(1 + 1/\hat{\gamma}) \approx 3.004$ verses $\overline{X} = 2.992$.

Question:

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. $\hat{E}(X) = \hat{\tau}\Gamma(1 + 1/\hat{\gamma}) \approx 3.004$ is the MLE of $E(X)$ and $\overline{X} = 2.992$ is the non-parametric estimator of $E(X)$ ($\overline{X} = \sum_i x\hat{f}(x)$, where $\hat{f}(x) = \sum_{i=1}^n \mathbf{1}(X_i = x)/n$ is the density of the edf $\hat{F}(x) = \sum_{i=1}^n \mathbf{1}(X_i \leq x)/n$, a non-parametric MLE (NPMLE) of $F_X(t)$. **Which is better ?**
3. Is the model assumption valid ?
   We can use the qqplot, confidence band (CB) of the edf and ks.test to check.

> pweibull(1,y$e[1],y$e[2]) +1-pweibull(5,y$e[1],y$e[2])
[1] 0.08444813



$Fig(1,1)$: *similar to next* $Fig(1,2)$: *cdf of weibull v.s. edf with its confidence band*
$Fig(2,1)$: *qqplot weibull* $Fig(2,2)$: *qqplot of* 100 *data from Weibull*

Figure 5.1.

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).

> 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, the P-value is asymptotically true. Thus it is assumed that

(1) θ is the true value and

(2) the sample size n is very large.

However, θ is estimated by its MLE here and n is not large, this changed its true P-value. One can find the critical value in D by empirical quantiles of 0.05 for a given sample size n.

(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) with ks(). **Why uniform of U(1,5) ?**

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. OW, $P(H_0|H_1)$ can be 99%, instead of $< 50\%$, 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 an approximation when n is very large. Otherwise, it is arbitrary.

Example 4. Generate 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])      true value of (γ, τ) ?
    c=ks.test(x, "pweibull", 1, 0.2)                  (y$e[1], y$e[2]) = (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 ?

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[2]<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[2]<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	$\hat{P}(H_1\|H_0)$	$\hat{P}(H_1\|H_0)$
50	0	0.045
1000	0	0.046
ideal	0.05	0.05

Finding: $\hat{P}(H_1|H_0) = 0$ if Weibull data test Weibull with (τ, γ) replaced by the MLE.
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]), u=rep(0,1000)
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$s
}
> sort(u)[950] # what is this ?
    [1] 0.08622978
> b                    b= ks.test(x, "pweibull", y$e[1], y$e[2])
    [1] 0.09650574 # D_o= 0.09650574
```

Q: Can we have conclusion now ?

```
> mean(u>b)
    # 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 to generate pseudo random numbers.
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 ($\overline{\mathbf{1}(D > b)} \to P(D > b)$ **anything wrong ??**).

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)
```