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)
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.

Announcement: Hao Wang, Chenxi Wang, Wai Wang see me after class.

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:

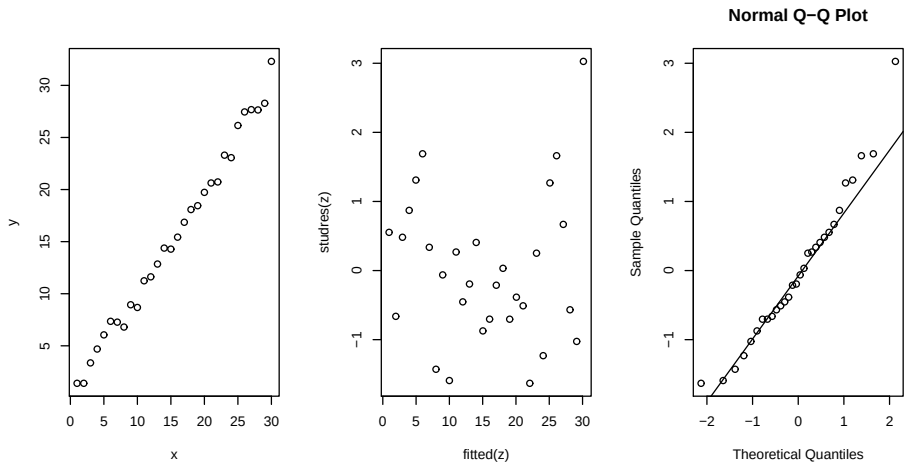
	Estimate	*Std. Error*	*t value*	$Pr(>\|t\|)$	
$(Intercept)$	-0.06299	0.32684	-0.193	0.849	—
x	1.00636	0.01841	54.663	$< 2e-16$	$***$

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 $\hat{\gamma} = 3.4$ with $\hat{\sigma}_{\hat{\gamma}} = 0.13$ and $\hat{\tau} = 2.8$ with $\hat{\sigma}_{\hat{\tau}} = 0.23$.

$\tilde{F}(t) = 1 - \exp(-(t/3.4)^{2.8})$, $t > 0$ 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 $N(\tau, \hat{\sigma}^2_{\hat{\tau}})$.

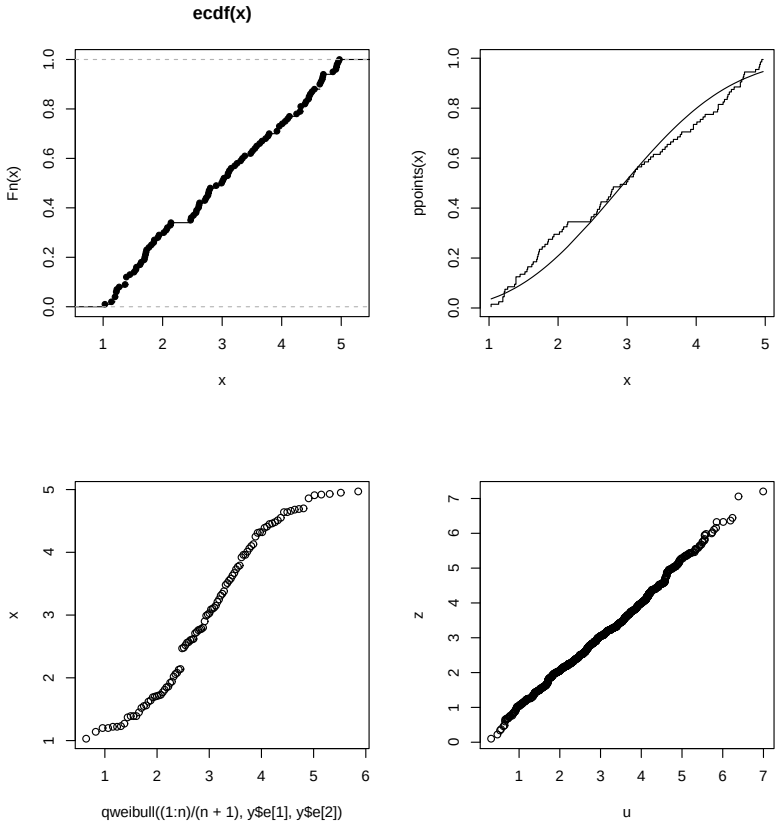
   H_0: $\gamma = 1$ v.s. H_1: $\gamma \neq 1$. Check $|\hat{\gamma} - 1| < 2\hat{\sigma}_{\hat{\gamma}}$?

   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 1000 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(1000,y$e[1],y$e[2])
plot(qweibull(ppoints(x),y$e[1],y$e[2]),x) # or qqplot(u,x)
z=rweibull(1000,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)
```

$$\underbrace{\text{(Power or size of the test } \phi \text{ ?. Or an estimate ?)}}_{E(\phi(\mathbf{U}))\ or\ \ \mathrm{P}(H_1|H_o)}$$

```
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 ?
```

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

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:

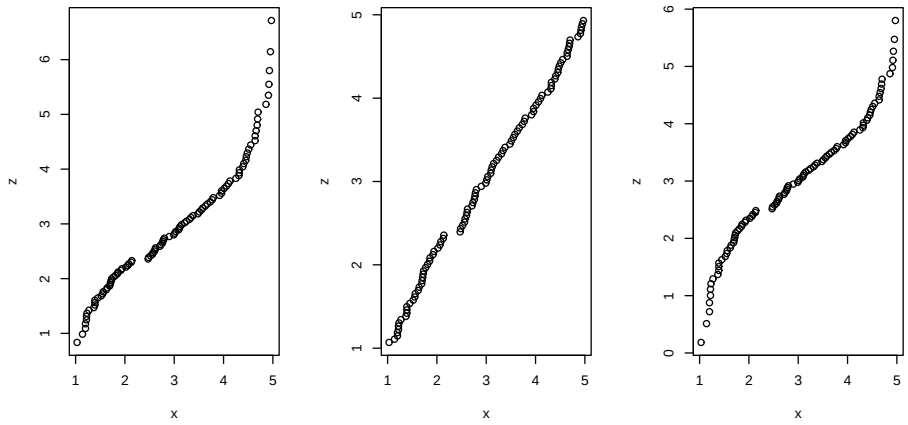
```
    mean(x)
    sum((x>1& x<=2))/length(x)
```

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.

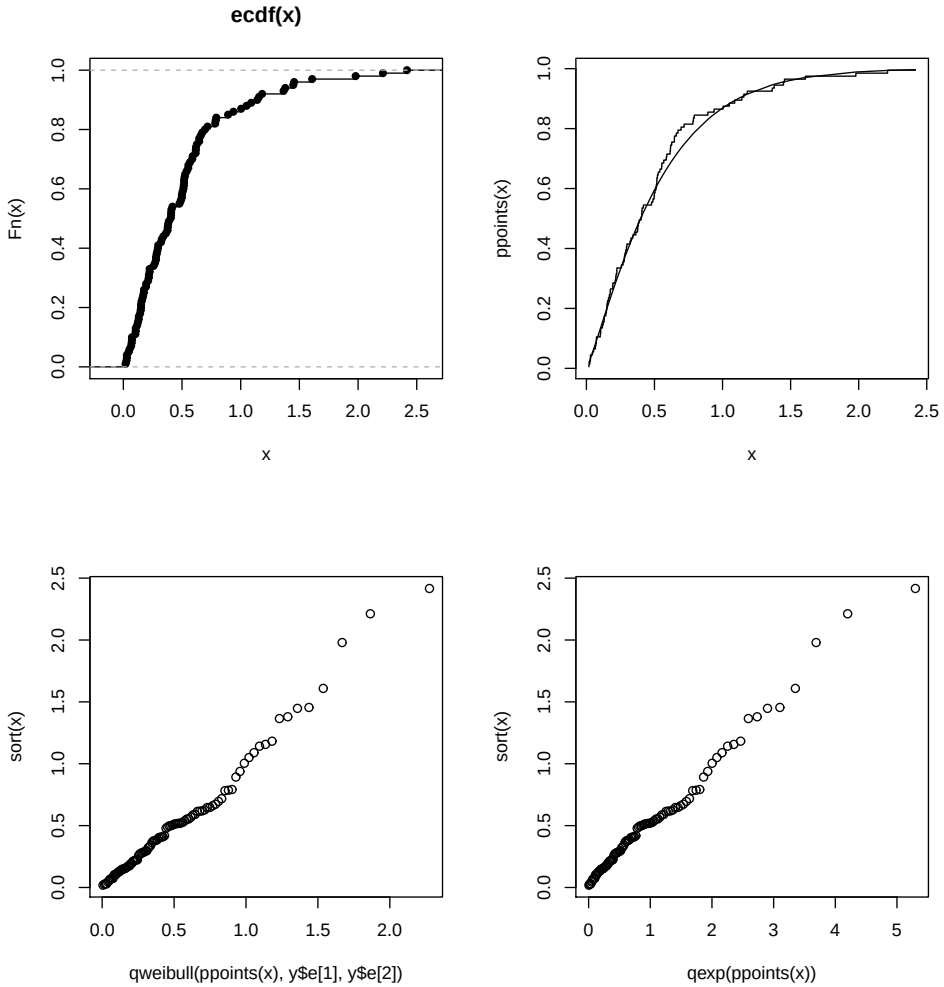
Weibull MLE=> $\check{P}(X \in (1,2]) = 0.18$, differ≈ half due to wrong assumption.

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
```

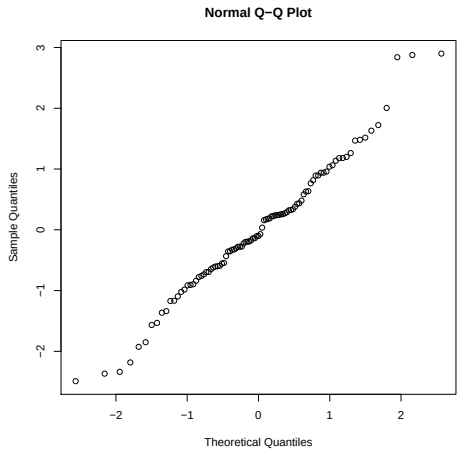
Q: Is it possible that the simulation study suggests that the data do not fit the Weibull model ?

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.

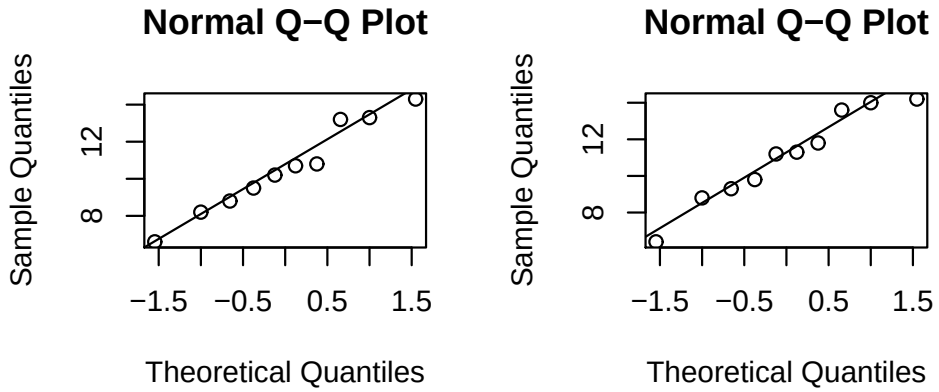
Example. Observations: $1, 3, 7$, H_o: $\mu = 4$. $S_n = ?$ $S_p = ?$ $S = ?$

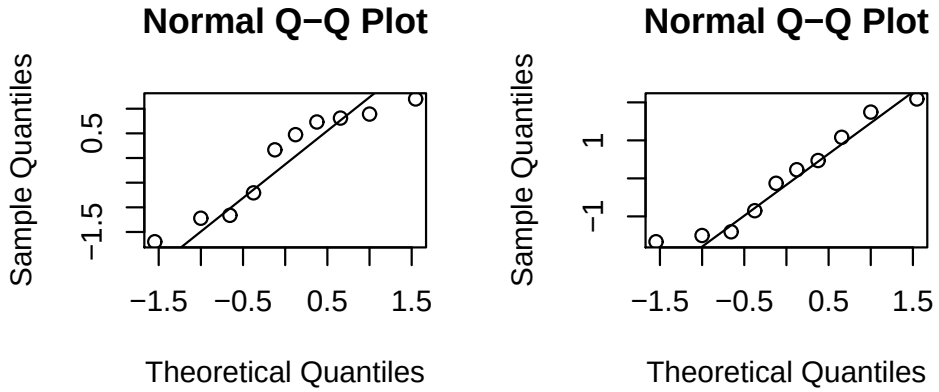
Remark: If n is large, t.test is very close to z.test by CLT on $\overline{X}$ under the assumption: X_1, ..., X_n are i.i.d., provided $\sigma_X < \infty$.

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 ?