MSiA 400 Lab Basic Statistics with R

Oct 6, 2014

Young Woong Park

Data Set

- Average heights of men and women of 71 countries (collected from Wikipedia)

	A	B	C	D
1		Male	Female	
2	Argentina	1.735	1.608	
3	Australia	1.784	1.645	
4	Austria	1.792	1.676	
5	Azerbaijan	1.718	1.654	
6	Bahrain	1.651	1.542	
7	Belgium	1.786	1.681	
8	Bolivia	1.6	1.422	
9	Brazil	1.707	1.588	
10	Bulgaria	1.752	1.632	
11	Cameroon	1.706	1.613	
12	Canada	1.76	1.633	
13	Chile	1.71	1.591	
14	China, People's Republic of	1.663	1.57	
15	Colombia	1.706	1.587	
16	Côte d'Ivoire	1.701	1.591	
17	Croatia	1.805	1.663	
18	Cuba	1.68	1.56	
19	Czech Republic	1.803	1.672	
20	Denmark	1.826	1.687	
21	Dinaric Alps	1.856	1.711	
22	Egypt	1.703	1.589	

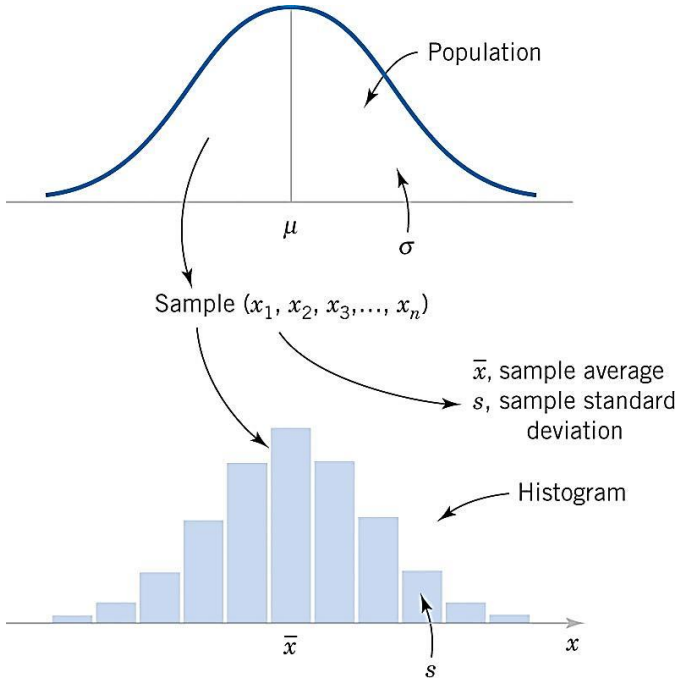
- Load height.txt from the last lab

```
> height = read.table("z:\\ msia400lab1 \\height.txt", header=T)
```

Numerical Summary of Data

- Definitions
  - Data: the numeric observations of a phenomenon of interest
  - Population: the complete set of objects of interest
  - Sample: A subset of the population, or a portion used for analysis

- Description of population or sampled data
  - center: measured by the mean
  - spread: measured by the variance



* Source: Applied Statistics and Probability for Engineers, Montgomery and Runger

Mean

- Sample mean: n observations $x_1, x_2, \cdots, x_n$ in a random sample

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Population mean: N observations $x_1, x_2, \cdots, x_N$ in a population

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- R: Calculate the sample mean of male and female

```
> M = height$Male;
> F = height$Female;
> n = length(F);
> mean(M);          # equivalent to mean(height[,1]) or mean(height[, "Male"])
> mean(F);
```

Variance

- Sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Population variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- R: Calculate the variance of male and female

```
> var(M);                  # sample variance
> var(F);                  # sample variance
> (n-1)/n*var(M);          # population variance if we assume data is from population
> (n-1)/n*var(F);          # population variance if we assume data is from population
> sd(F);                   # standard deviation
```

Covariance

- Covariance of two variables x and y
  - measures how the two are linearly related

- Sample covariance of x and y

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Population covariance of x and y

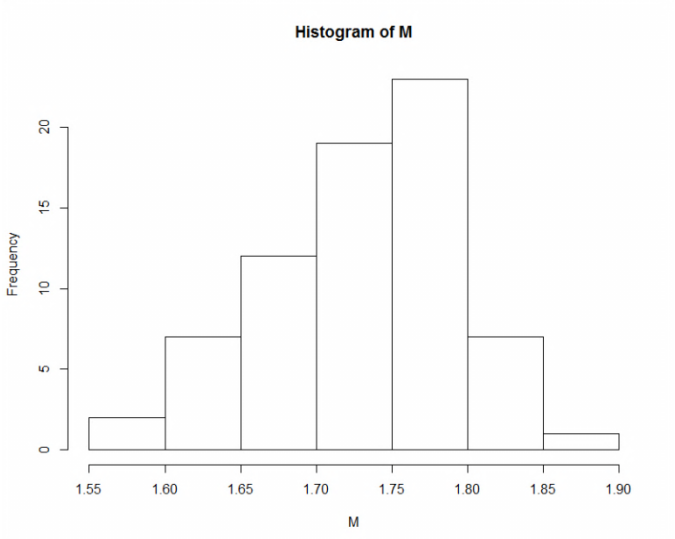
$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

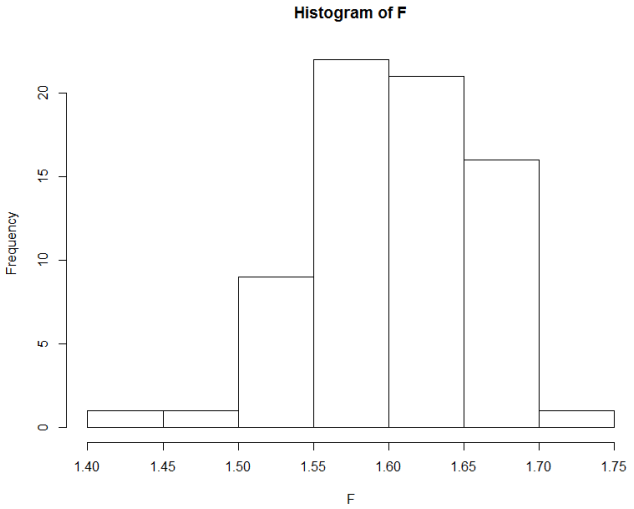
- R: Calculate covariance of male and female heights

```
> cov(M,F);
```

Histogram

- Histogram (Frequency Distribution)
  - presents the counts of observations grouped within pre-specified classes or groups





- R: Generate histograms

```
> hist(M);
> hist(F);
> hist(F, breaks=4);                          # Try
> hist(F, breaks=seq(1.40,1.75,by=0.03);      # Try
```

Point estimation of population mean

- Sample mean: n observations $x_1, x_2, \cdots, x_n$ in a random sample

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- R: Find a point estimation of women's height

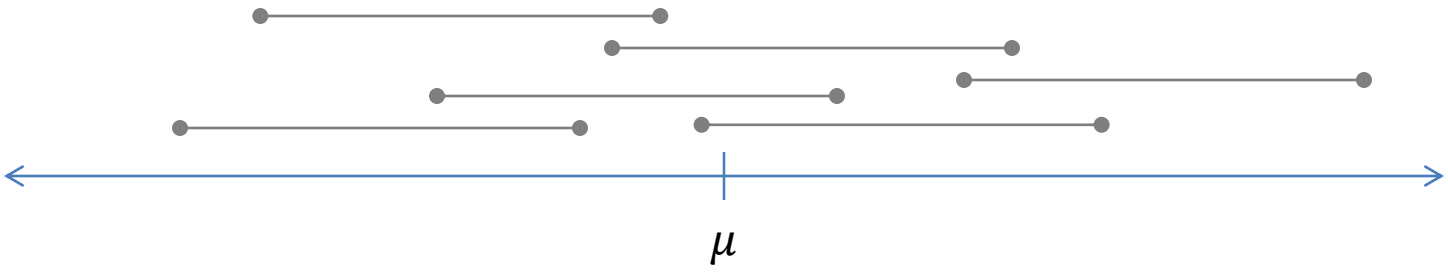
```
> mean(F);
```

Confidence Interval

- Confidence interval (CI) on μ with known population variance (σ^2)

$$(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ , \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

- Interpretation of CI
  - Note: The confidence interval is a random interval!!
  - Yes: The observed interval includes the true value of μ with confidence $100(1-\alpha)$
  - No: The true value of μ lies in the observed interval with confidence $100(1-\alpha)$



Confidence Interval of population mean

- Known population variance (σ^2)

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

  - R: Confidence interval of men's height with known σ =0.0036 and $\alpha = 0.05$

```
> sigma = 0.0036; Mbar = mean(M);
> E = qnorm(0.975)*sigma/sqrt(n);      # margin of error
> CI_M = Mbar + c(-E,E);
```

- Unknown population variance (s^2)

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

  - R: interval estimation of men's height with unknown σ and $\alpha = 0.05$

```
> S = sd(M);
> E = qt(0.975, df=n-1)*S/sqrt(n);      # margin of error
> CI_M = Mbar + c(-E,E);
```

Hypothesis testing

- Statistical hypotheses
  - Two-tailed alternative hypothesis
    $H_0: \mu_m = 1.728$
    $H_1: \mu_m \neq 1.728$
  - Lower-tailed alternative hypothesis
    $H_0: \mu_m = 1.728$
    $H_1: \mu_m < 1.728$
- Hypothesis testing
  - Obtains information in a random sample from the population
  - If the information is consistent with the hypothesis, we conclude the hypothesis is true, otherwise, false.

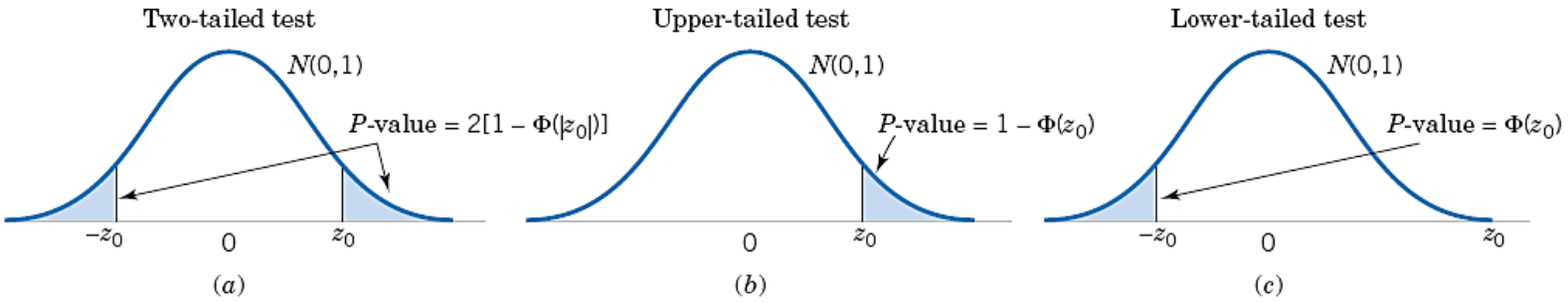


Figure 9-7 The P-value for a z-test. (a) The two-sided alternative $H_1: \mu \neq \mu_0$. (b) The one-sided alternative $H_1: \mu > \mu_0$. (c) The one-sided alternative $H_1: \mu < \mu_0$.

* Source: Applied Statistics and Probability for Engineers, Montgomery and Runger

Two tailed test with known variance

- Two-tailed test with known variance

$$\mathrm{H}_0\colon \mu_m = 1.729$$
$$\mathrm{H}_1\colon \mu_m \neq 1.729$$

- Test statistics

$$z_0 = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

* Fail to reject H_0 if $-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2}$

- R: Test the hypothesis with $\alpha = 0.05$

```
> z = (Mbar – 1.729)/(sigma/sqrt(n));
> z.half.alpha = qnorm(1-0.05/2);
> c(-z.half.alpha, z.half.alpha);          # check if z is inside the interval
```

Lower-tailed test with known variance

- Two-tailed test with known variance

$$\mathrm{H}_0\colon \mu_m = 1.729$$
$$\mathrm{H}_1\colon \mu_m < 1.729$$

- Test statistics

$$z_0 = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

* Fail to reject H_0 if $z_0 \leq -z_\alpha$

- R: Test the hypothesis with $\alpha = 0.05$

```
> z = (Mbar – 1.729)/(sigma/sqrt(n));
> z.alpha = –qnorm(1-0.05);
```

Two tailed test with unknown variance

- Two-tailed test with unknown variance

$$\mathrm{H}_0\colon \mu_m = 1.729$$
$$\mathrm{H}_1\colon \mu_m \neq 1.729$$

- Test statistics

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

* Fail to reject H_0 if $-t_{\alpha/2} \leq t_0 \leq t_{\alpha/2}$

- R: Test the hypothesis with $\alpha = 0.05$

```
> t = (Mbar – 1.729)/(sd(M)/sqrt(n));
> t.half.alpha = qt(1-0.05/2, df=n-1);
> c(-t.half.alpha, t.half.alpha);          # check if t is inside the interval
```

Example: Tensile Strength

- Tensile Strength

A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. Product engineering thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5 and 20%. A team of engineers decided to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%. They decided to make up six test specimens at each concentration level, using a pilot plant. All 24 specimens are tested on a laboratory tensile tester, in random order.

* Source: Applied Statistics and Probability for Engineers, Montgomery and Runger

- Tensile Strength of Paper (psi)

Hardwood concentration (%)	Observations						Totals	Averages
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17

ANOVA (1)

- Goal
  Check if the tensile strength for 4 different concentrations are equal

- Hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$$
$$H_1: \mu_i \neq 0 \text{ for at least one } i$$

- ANOVA table

Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	$\boldsymbol{F_0}$
Treatments	$SS_{Treatments}$	$k-1$	$MS_{Treatments}$	$\frac{MS_{Treatments}}{MS_E}$
Error	SS_E	$k(n-1)$	MS_E	
Total	SS_T	$kn-1$		

ANOVA (2)

- Import data

  > tensile = read.table("z:\\ msia400lab1 \\tensile.txt", header=T)

- Constructing ANOVA table

Step 1. Turn the data into a single vector

> resp = c(t(as.matrix(Tensile)));

Step 2. Setup values

> treats = c("HC5","HC10","HC15","HC20");

> k = 4;

> n = 6;

Step 3. Create a vector of treatment factors that corresponds to each element of resp

> tm = gl(k,1,n*k,factor(treats));

Step 4. Apply the function **aov** and print summary

> myANOVA = aov(resp ~ tm);

> summary(myANOVA);

Note

gl(): generate levels

(# levels, # replications, length, labels)

Note

as.matrix() : convert table to matrix

t() : transpose

ANOVA (3)

- ANOVA table

Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F_0	P-value
Treatments	382.8	3	127.60	19.61	3.59 E-6
Error	130.2	20	6.51		
Total	513.0	23			

- Two approaches
  - $F_{0.01,3,20} = 4.94 < 19.61 = F_0$, reject H_0
  - P-value =3.59 E-6 < $\alpha = 0.01$, reject H_0

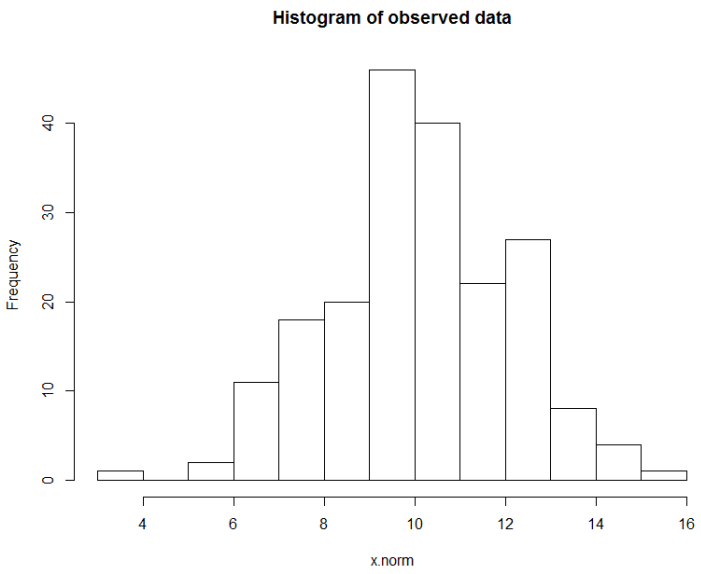
* Note
> qf(1-0.01,df1=3,df2=20)

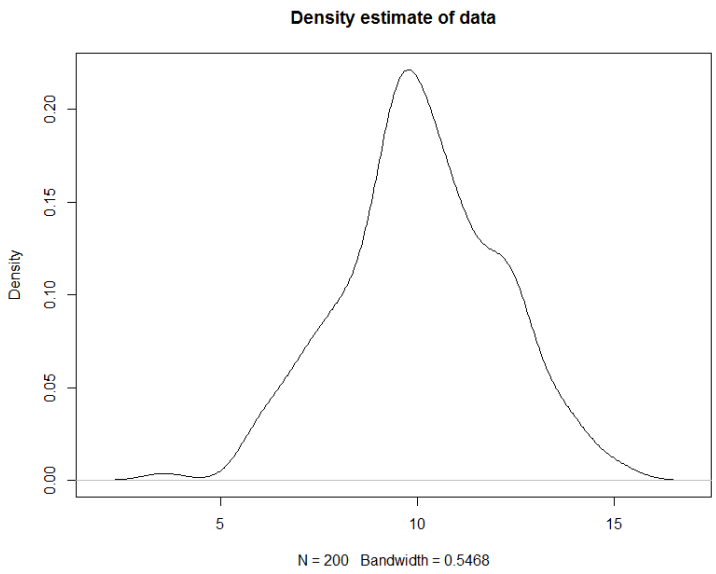
- Conclusion: Hardwood concentration affects the strength of the paper

Distribution Fitting 1

- Samples from Normal distribution

```
> x.norm = rnorm(n=200,m=10,sd=2);  # Generate 200 random samples from N(10,2)
> hist(x.norm,breaks = 10,main="Histogram of observed data");
> plot(density(x.norm),main="Density estimate of data");
```

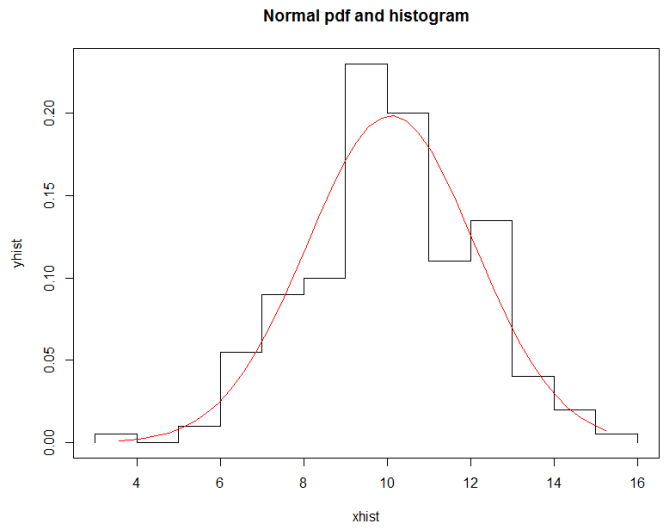


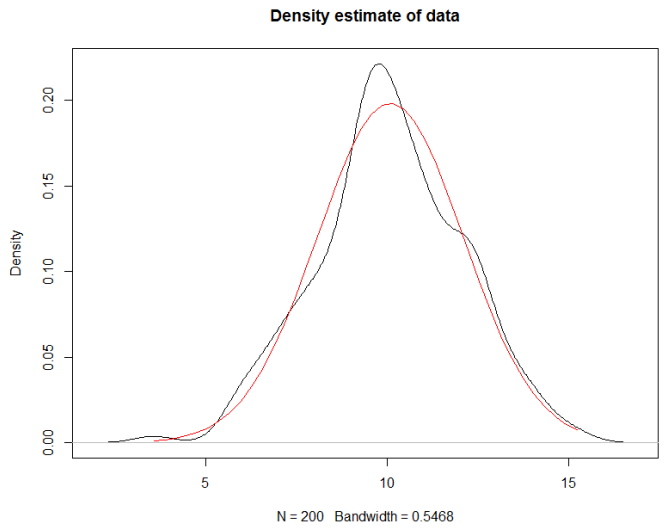


Distribution Fitting 2

- Comparisons

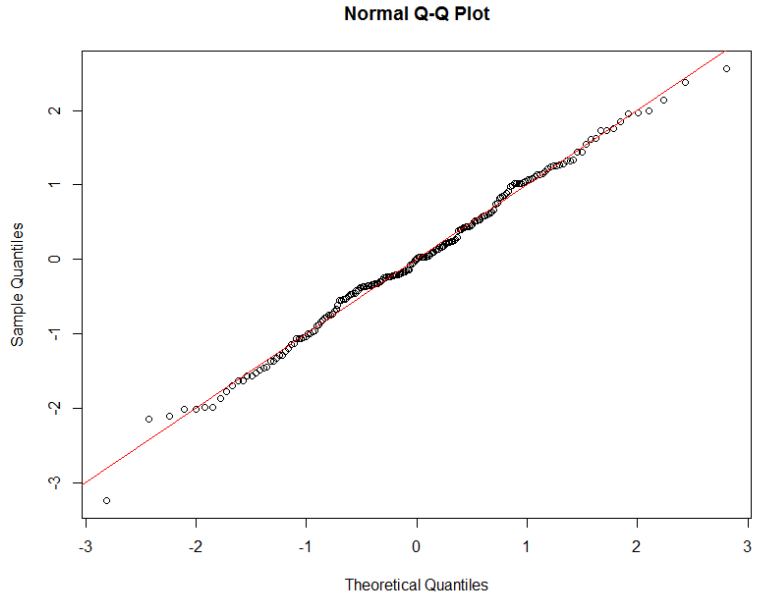
```
> h = hist(x.norm,breaks=10);
> x.hist = c(min(h$breaks),h$breaks); y.hist = c(0,h$density,0);
> x.fit = seq(min(x.norm),max(x.norm),length=40);
> y.fit = dnorm(x.fit,mean=mean(x.norm),sd=sd(x.norm));
> plot(x.hist,y.hist,type="s",ylim=c(0,max(y.hist,y.fit)), main="Normal pdf and histogram")
> lines(x.fit,y.fit, col="red");

> plot(density(x.norm),main="Density estimate of data")
> lines(x.fit,y.fit, col="red")
```





Distribution Fitting 3

- QQ plot
  A graphical method for comparing two probability distributions by plotting their quantiles against each other



- R: QQ plot generation

```
> z.norm = (x.norm-mean(x.norm))/sd(x.norm);
> qqnorm(z.norm);
> abline(0,1, col ="red");
```