Sampling Distributions

September 17, 2013

Sampling Distributions

R supports a large number of distributions. Usually, four types of functions are provided for each distribution:

- d: density function
- p: cumulative distribution function, $P(X \leq x)$
- q: quantile function
- r: draw random numbers from the distribution

Central Limit Theorem

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a distribution with mean μ and variance σ^2. Then for large n, $\bar{X}$ is approximately normal with mean μ and variance σ^2/n. This means

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

as $n \to \infty$

CLT- Example

Example: If a sample of size 16 is drawn from a normal population that has a mean 27 and standard deviation of 2, what is the probability that the mean of the sample will be less than 26?
We have $n = 16, \mu = 27, \sigma = 2$. We want to find $P(\bar{X} < 26)$.
We know that

$$\begin{aligned} P(\bar{X} \leq 26) &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{26 - \mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z \leq \frac{26 - 27}{2/\sqrt{16}}\right) \\ &= P(Z \leq -2) \\ &= 0.0228. \end{aligned}$$

Hence the probability that the mean of the sample will be less than 26 is 0.0228.

```
> pnorm(26,27,2/sqrt(16))
[1] 0.02275013
```

Simulating CLT

```
n = 30            # sample size
k = 1000         # number of samples
mu = 5; sigma = 2; SEM = sigma/sqrt(n)
x = matrix(rnorm(n*k,mu,sigma),n,k) # creates a matrix
x.mean = apply(x,2,mean)
x.down = mu - 4*SEM; x.up = mu + 4*SEM; y.up = 1.5
hist(x.mean,prob= T,xlim= c(x.down,x.up),ylim= c(0,y.up),
main= 'Sampling distribution of the sample mean, Normal case')
par(new= T)
x = seq(x.down,x.up,0.01)
y = dnorm(x,mu,SEM)
plot(x,y,type= 'l',xlim= c(x.down,x.up),ylim= c(0,y.up))
```

Student's -t distribution

If $X_1, X_2, \cdots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2 then

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

This is an important result, but the major difficulty arise on application in which cases σ is unknown. In this case we replace σ with its estimate s and we study the distribution of $\frac{\overline{X}-\mu}{s/\sqrt{n}}$. The distribution of this expression will have the student's t-distribution.
A random variable X is said to have t-distribution with n degrees of freedoms if its pdf is given by

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left\{1 + \frac{x^2}{n}\right\}^{-\frac{n+1}{2}} \text{ for } -\infty < x < \infty.$$

R codes for t- distribution

- dt: density function of t-distribution
- pt: cumulative distribution function
- qt: quantile function of t- distribution
- rt: draw random numbers from the t-distribution

Need to choose the parameter n.
$t(df, ncp)$? noncentral t distribution with noncentrality parameter ncp

```
> pt(-1,df=10)
[1] 0.1704466
> pt(0, df=10)
[1] 0.5
> pt(1, df=10)
[1] 0.8295534

# Calculating percentiles
> # Find the 25th percentile with a degree of freedom=4
> qt(.25, df=4)
[1] -0.7406971
```

Continuous Distributions

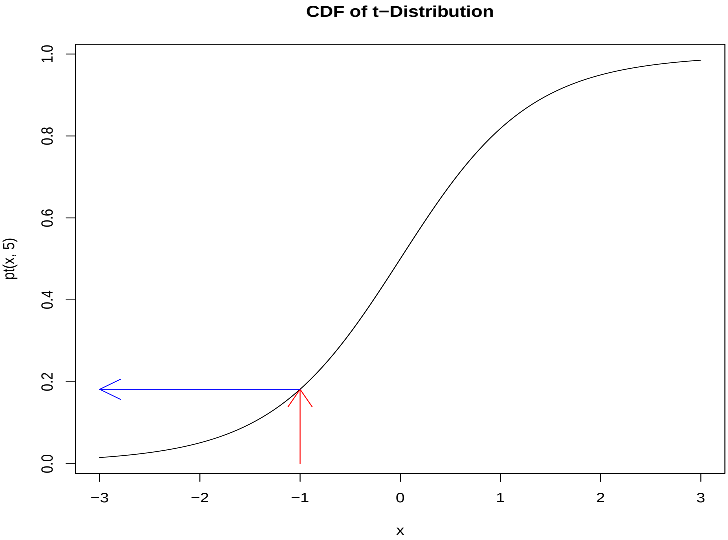
The cumulative probability function is a straightforward notion: it is an S-shaped curve showing, for any value of x, the probability of obtaining a sample value that is less than or equal to x. Here is what it looks like for the normal distribution:

```
>curve(pt(x,5),-3,3, main="CDF of t-Distribution")
>arrows(-1,0,-1,pt(-1,5),col="red")
>arrows(-1,pt(-1,5),-3,pt(-1,5),col="blue")
```

The value of $x(-1)$ leads up to the cumulative probability (red arrow) and the probability associated with obtaining a value of this size (-1) or smaller is on the y axis (blue arrow). The value on the y axis is 0.1816087:

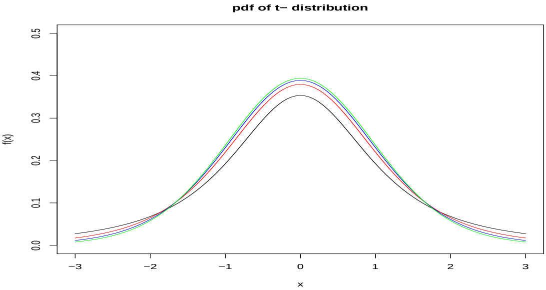
```
> pt(-1,5)
[1] 0.1816087
```

CDF- Student's t- Distribution



PDF: t- Distribution

Superimpose many PDFs:




```
curve(dt(x,2),from=-3,to=3,col="black", ylim=c(0,0.5),
xlim=c(-3,3),ylab="f(x)",xlab="x",main="pdf of t-distribution")
curve(dt(x,5),from=-3, to=3, col="red", add=T)
curve(dt(x,10),from=-3,to=3,col="blue", add=T)
curve(dt(x,20),from=-3, to=3, col="green", add=T)
```

The "from" and "to" can be omitted.

Shading

```
cord.x <- c(-3,seq(-3,-1,0.01),-1)
cord.y <- c(0,dt(seq(-3,-1,0.01),5),0)
curve(dt(x,5),xlim=c(-3,3),main='Student t- distribution')
polygon(cord.x,cord.y,col='blue')
```

Few Examples- Marking and shading

1) Plot the t distribution with 5 degrees of freedom and mark the 90th percentile:

```
curve(dt(x,5),-3,3)
lines(qt(0.9,5),dt(qt(0.9,5),5),type="h", col="red")
```

2) Shade the area under the pdf of t distribution with 5 degrees of freedoms to the right of the 90th percentile:

```
x1=seq(qt(0.9,5),3,0.01);
y1=dt(x1,5)
curve(dt(x,5),-3,3); lines(x1,y1,type="h",col="red")
```

Example

1) Generate 10 random numbers from t- distribution with 20 degrees of freedom.
2) For a Student's t-distribution with 12 degrees of freedom what is the probability that $P(X \leq 2)$?
3) What is the "x" value from a Student's t-distribution with 12 degrees of freedom so that there is a 99% probability that a random value is below x?
4) Obtain 95% quantile for student's t -distribution with 15 degrees of freedom.

Chi-Square Distribution

A random variable is X is said to have χ^2-distribution with n-degrees of freedom if its pdf is given by

$$f(x) = \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-x/2} \qquad x > 0$$

Here n is called the degrees of freedom.
If $X \sim N(0,1)$ then $X^2 \sim \chi^2(1)$. Therefore, if $X \sim N(\mu, \sigma^2)$ then the random variable $Z^2 = (X-\mu)^2/\sigma^2$ is $\chi^2(1)$
chisq(df)– central χ^2 with df degrees of freedom (default)
chisq(df,ncp) – noncentral χ^2 with noncentrality parameter ncp

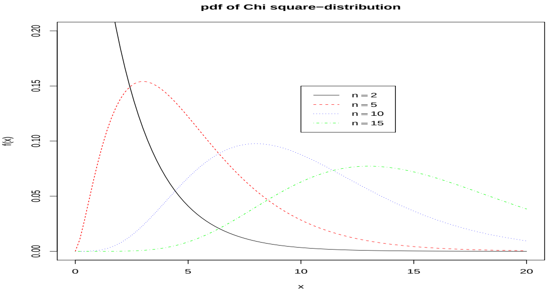
Chi-square Example

- dchisq(x, df, ncp = 0, log = FALSE)
- pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
- qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
- rchisq(n, df, ncp = 0)

Note that *ncp* is the non-centrality parameter. If omitted the central chi-square is assumed.

```
> x <- seq(0,20,by=.5)
> y <- dchisq(x,df=10)
> plot(x,y)
      Or
>curve(dchisq(x, 10),0,20)
```

Chi-square distribution




```
curve(dchisq(x,2),from=0,to=20,col="black", ylim=c(0,0.2),xlim=c(0,20),
ylab="f(x)",xlab="x",main="pdf of Chi square-distribution", lty=1)
curve(dchisq(x,5),from=0, to=20, col="red", add=T,lty=2)
curve(dchisq(x,10),from=0,to=20,col="blue", add=T,lty=3)
curve(dchisq(x,15),from=0, to=20, col="green", add=T,lty=4)
legend(10,0.15,legend=c(expression(n==2),expression(n==5),
expression(n==10), expression(n==15)),lty=1:4,
col=c("black","red","blue", "green"))
```

Chi-Square Distribution

CDF of chi-square distribution: **pchisq(q,df)**

```
> pchisq(2,5)
[1] 0.150855
> pchisq(10,5)
[1] 0.9247648
> pchisq(10,20)
[1] 0.03182806
```

Quantiles of chi-square distribution: **qchisq(p,df)**

```
> qchisq(0.2,5)
[1] 2.342534
> qchisq(0.5,5)
[1] 4.35146
> qchisq(0.95,5)
[1] 11.0705
```

Generating random numbers from chi-square distribution: **rchisq(n,df)**

F- distribution

A random variable X is said to have a F-distribution with n_1 numerator degrees of freedom and n_2 denominator degrees of freedom, denoted as $X \sim F(n_1, n_2)$, if its pdf is given by

$$f(x) = \begin{cases} \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}(\frac{n_1}{n_2})^{\frac{n_1}{2}} x^{\frac{n_1-2}{2}} (1 + \frac{n_1}{n_2}x)^{-(n_1+n_2)/2} & \text{if} \quad x > 0 \\ \\ 0 & \text{otherwise} \end{cases}$$

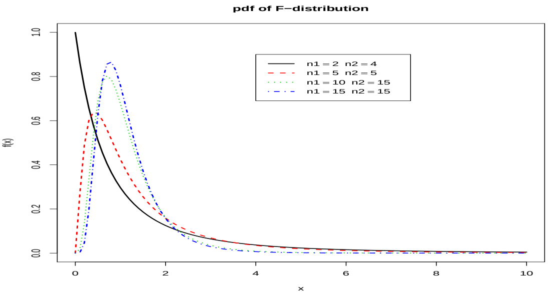
F -Distribution

- df(x, df1, df2, ncp, log = FALSE)
- pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
- qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
- rf(n, df1, df2, ncp)

Note that *ncp* is the non-centrality parameter. If omitted the central F is assumed.

```
> x <- seq(0,20,by=.5)
> y <- df(x,df1=10, df2=5)
> plot(x,y)
      Or
>curve(df(x, df1=10,df2=5),0,20)
```

F-distribution




```
curve(df(x,2,4),from=0,to=10,col=1, ylim=c(0,1),xlim=c(0,10), lwd=2,
ylab="f(x)",xlab="x",main="pdf of F-distribution", lty=1)
curve(df(x,5,5),from=0, to=10, col=2, add=T,lty=2,lwd=2)
curve(df(x,10,15),from=0,to=10,col=3, add=T,lty=3,lwd=2)
curve(df(x,15,15),from=0, to=10, col=4, add=T,lty=4,lwd=2)
legend(4,0.9,legend=c(expression(n1==2~~n2==4),expression(n1==5~~n2==5),
expression(n1==10~~n2==15), expression(n1==15~~n2==15)),lty=1:4,lwd=2,
col=c(1,2,3,4))
```

F- Distribution

CDF of F- distribution: **pf(q,df1,df2)**

```
> pf(2,df1=5,df2=10)
[1] 0.835805
> pf(10,df1=5,df2=10)
[1] 0.9987942
> pf(10,df1=20,df2=10)
[1] 0.9996589
```

Quantiles of F-distribution: **qf(p,df1,df2)**

```
> qf(0.2,10,5)
[1] 0.5547161
> qf(0.5,10,5)
[1] 1.073038
> qf(0.95,10,5)
[1] 4.735063
```

Generating random numbers from chi-square distribution: **rf(n,df1,df2)**

Quantile-Quantile Plots for Normal Distributions

One of the most useful graphical procedure for assessing distributions is the quantile-quantile plot. A quqntile-quantile (Q-Q) plot plots the quantiles of one distribution against the quantiles of another distribution as (x,y) points. When two distributions have similar shapes, the points will fall along a straight line. The R function to draw a quantile-quantile plot is qqplot(x,y). Histograms can be used to compare two distributions. However, it is rather challenging to put both histograms on the same graph. R offers to statements: qqnorm(), to test the goodness of fit of a gaussian distribution, or qqplot() for any kind of distribution.

Example

```
x.norm<-rnorm(n=200,m=10,sd=2)
hist(x.norm,main="Histogram of observed data")
plot(density(x.norm),main="Density estimate of data")
plot(ecdf(x.norm),main="Empirical CDF")
z.norm<-(x.norm-mean(x.norm))/sd(x.norm) ## standardized data
qqnorm(z.norm) ## drawing the QQplot
abline(0,1) ## drawing a 45-degree reference line
```