Handout 3. Probability and distributions

1. Random sampling

The basic notion of a random sample is to deal from a well-shuffled pack of cards or picking numbered balls from a well-stirred urn.

In R, you can simulate these situations with the sample function. If you want to pick five numbers at random from the set 1:40, then you can write

```
> sample(1:40,5)
> sample(c("H","T"), 10, replace=T)
> sample(c("succ", "fail"), 10, replace=T, prob=c(0.9, 0.1))
```

2. Probability calculations and combinatorics

In R, the choose function can be used to calculate the number of ways to choose 5 numbers out of 40.

```
> choose(40,5)

 1/prod(40:36)
 prod(40:36)/prod(1:5)
 ??product
 ?factorial
 factorial(4)
```

3. Discrete and continuous distributions

Discrete: Binomial distribution, geometric distribution, Poisson distribution
Continuous: Normal, Beta, Gamma, log-normal, ….

Distribution	**R name**	**additional arguments**
beta	`beta`	shape1, shape2, ncp
binomial	`binom`	size, prob
Cauchy	`cauchy`	location, scale
chi-squared	`chisq`	df, ncp
exponential	`exp`	rate
F	`f`	df1, df2, ncp
gamma	`gamma`	shape, scale
geometric	`geom`	prob
log-normal	`lnorm`	meanlog, sdlog
logistic	`logis`	location, scale
negative binomial	`nbinom`	size, prob
normal	`norm`	mean, sd
Poisson	`pois`	lambda
Student‟s t	`t`	df, ncp
uniform	`unif`	min, max
Weibull	`weibull`	shape, scale

Prefix the name given here by „d“ for the density, „p“ for the CDF, „q“ for the quantile function and „r“ for simulation (random deviates). The first argument is x for dxxx, q for pxxx, p for qxxx and n for rxxx. We next discuss and give some examples on these functions.

(1) Densities

```
> x <- seq(-4,4,0.1)
> plot(x,dnorm(x),type="l")
> curve(dnorm(x), from=-4, to=4)
```

```
x <- seq(0,1,0.01)
plot(x,dbeta(x,4,5),type="l")
curve(dbeta(x,4,5), from=0, to=1)

x <- seq(0,100,1)
plot(x,dchisq(x,40),type="l")
curve(dchisq(x,40), from=0, to=100)

x <- seq(0,100,1)
plot(x,dgamma(x,30),type="l")
curve(dgamma(x,30), from=0, to=100)
```

For discrete distributions, where variables can take on only distinct values, it is preferable to draw a pin diagram, here for the binomial distribution with n = 50 and p = 0.33:

```
> x <- 0:50
>   plot(x,dbinom(x,size=50,prob=.33),type="h")
#curve(dbinom(x,size=50,prob=.33), from=0, to=50)

x <- 0:50
plot(x,dpois(x,10),type="h")
```

(2) Cumulative distribution functions

```
> pnorm(160,mean=132,sd=13)
> pbinom(16,size=20,prob=.5)

pbeta(0.8,4,5)
pchisq(100,40)
pchisq(10,40)
pgamma(100,30)
pgamma(10,30)

ppois(2,10)

x <- seq(-4,4,0.1)
plot(x,pnorm(x),type="l")
curve(pnorm(x), from=-4, to=4)
```

```
      x <- 0:50
      plot(x,pbinom(x,size=50,prob=.33))
#curve(pbinom(x,size=50,prob=.33), from=0, to=50)
```

(3) Quantiles

If we have n normally distributed observations with the same mean μ and standard deviation sigma, then it is known that the average „xbar" is normally distributed around μ with standard deviation sigma/sqrt(n). A 95% confidence interval for μ can be obtained as

xbar + sigma/sqrt(n) × N0.025 <= μ<= xbar + s/pn × N0.975

where N0.025 is the 2.5% quantile in the normal distribution.

```
qnorm(0.5)
qnorm(0.5,1,2)
qnorm(0.025)
xbar=83
sigma=12
n<-5
sem<-sigma/sqrt(n)
sem
xbar+sem*qnorm(0.025)
xbar+sem*qnorm(0.975)
xbar-sem*qnorm(0.025)

qbinom(0.5,size=20,prob=.5)
qbeta(0.5,4,5)
qbeta(1,4,5)

qchisq(0.5,40)
qchisq(1,40)

qgamma(0.5,30)
qgamma(1,30)

qpois(0.5,10)
qpois(1,10)
```

(4) Random numbers

Computer generates sequences of "pseudo-random" numbers, which for practical purposes behave as if they were drawn randomly

```
      > rnorm(10,mean=7,sd=5)
      > rbinom(10,size=20,prob=.5)
rbeta(10,4,5)
rchisq(30,40)
rgamma(50,30)
rpois(20,10)
```