A demonstration of critical values for t-distribution
NPFL 054 lab session (Hladká and Holub, 2016)

********************************************************************************
*** Testing a classifier -- a simulation
********************************************************************************

We will simulate a classifier with a given accuracy e.g. 83%.

We will do an experiment that shows test results for a randomly selected test samples. Each test sample will have the same size = 100 examples.

Dataset D will represent the whole population of all possible test examples.  
> D = c(rep(T, 830000), rep(F, 170000))

Simulation of 10 experiments with random test samples  
> set.seed(123)  
> for(i in 1:10){ cat(sum(sample(D, 100, rep=T)), ' ') }; cat('
')
85 82 84 82 85 80 83 86 88 83

-- the number of correctly classified examples obviously has binomial distribution.

When we want to approximate our binomial distribution by a normal one, we will take mean=np and variance=np(1-p) where n and p are the parameters of binomial distribution Bin(n,p), hence

N( mean=83, var=100*0.83*(1-0.83) )

********************************************************************************
*** Demo
********************************************************************************

D = c(rep(T, 830000), rep(F, 170000))

set.seed(123); acc = numeric(10); acc.avg = numeric(10000)
for(j in 1:10000){
   for(i in 1:10){ acc[i] = sum(sample(D, 100, rep=T)) }  
   acc.avg[j] = mean(acc)
}

### Compare distribution of acc and acc.avg
#  
#  -- acc has binominal distribution Bin(100, 0.83), which implies  
#  mean = np = 83 and variance = npq = 14.11  
#  
#  -- while acc.avg empirically has  
#  mean(acc.avg) = 83.009  
#  var(acc.avg) = 1.399
Empirical distribution of observed averages

`hist(acc.avg)`
### Computing t-statistics

```r
set.seed(123); T.values = numeric(10000)
for(j in 1:10000){
    for(i in 1:10){ acc[i] = sum(sample(D, 100, rep=T)) }
    T.values[j] = (mean(acc) - 83)/sd(acc) * sqrt(10)
}
```

### The distribution of t-values is the t distribution with 9 degrees of freedom,
# which is close to the standard normal N(0,1)
#
# mean(T.values) = 0.038
# var(T.values)  = 1.29

# Empirical distribution of observed t-values

plot(cut(T.values, breaks=seq(-5, 5, 0.4)))
Computing critical values

-- look at the number of T.values that fulfill $|T| < \text{critical}_\text{value}$

for alpha = 1%, 2%, 5%, 10%, 20%

The results below illustrate how critical values work.

for(a in c(1, 2, 5, 10, 20)){
    alpha = a/100
    crit.value = qt( (1 - alpha/2), df=9 )
    message("alpha = ", a, "% \t \sim ", sum(abs(T.values) < crit.value) )
}

alpha = 1% ~ 9895
alpha = 2% ~ 9802
alpha = 5% ~ 9499
alpha = 10% ~ 8990
alpha = 20% ~ 7964