STAT 703/J703
April 17th, 2007
-Lecture 25-

Instructor: Brian Habing
Department of Statistics
LeConte 203
Telephone: 803-777-3578
E-mail: habing@stat.sc.edu

Today

Methods Based on the CDF

- The Empirical Distribution Function
- Some Statistical Properties
- Kolmogorov-Smirnov Test
- The Nonparametric Bootstrap

Recall that the definition of the cumulative distribution function (CDF) is:

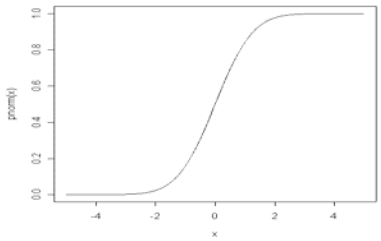
$$F_X(x)=P(X\leq x)$$

Note that:

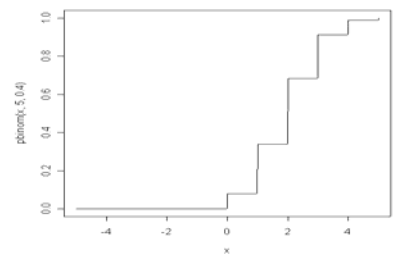
- $F_X(x)$ is non-decreasing
- $F_X(x) \rightarrow 1$ as $x \rightarrow \infty$
- $F_X(x) \rightarrow 0$ as $x \rightarrow -\infty$

The advantage of the CDF is that every random variable has one, and it has the same definition for both discrete and continuous random variables.


```
x<-(-500:500)/100
plot(x,pnorm(x),type="l")
```




```
plot(x,pbinom(x,5,.4),type="l")
```



The empirical distribution function (or empirical cumulative distribution function) is defined as:

$$F_n(x) = \frac{1}{n}\{\#x_i \le x\}$$

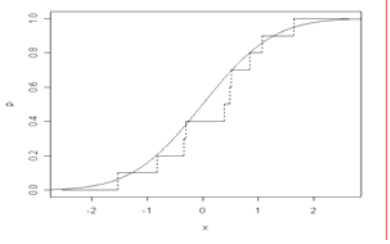
Unlike a histogram, there is only one way to plot an EDF.


```
edf<-function(y){
  x<-sort(y)
  plot(c(min(x)-1,max(x)+1),
    c(0,1),type="n",
    xlab="x",ylab="p")
 lines(c(x[1]-1,x[1]),c(0,0),
    lty=1)
 lines(c(x[1],x[1]),
    c(0,1/length(x)),lty=2)
```



```
for (i in 1:(length(x)-1)){
 lines(c(x[i],x[i+1]),
    c(i/length(x),
    i/length(x)),  lty=1)
 lines(c(x[i+1],x[i+1]),
    c(i/length(x),
    (i+1)/length(x)),lty=2)}
 lines(c(x[length(x)],
    x[length(x)]+1),c(1,1),
    lty=1) }
```

```
edf(rnorm(10))
lines(x,pnorm(x))
```



Statistical properties of the EDF

Note that we could write the EDF as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$$

This leads directly to the fact that

$F_n(x) \to F(x)$ as $n \to \infty$ for each x

With more theory we could prove that

$\sup_x |F_n(x) \to F(x)| \to 0$ as $n \to \infty$

The Kolmogorov-Smirnov test uses this quantity to construct a test of the null hypothesis that the data is drawn from a population with cdf F.

The test statistic is

$$\sup_x |F_n(x) \to F(x)|$$

The command in R is ks.test

```
ks.test(x,"pnorm",0,1)
```


It is interesting that the distribution of the Kolmogorov-Smirnov statistic does not depend on F !!!

Nonparametric Bootstrap

We previously examined the parametric bootstrap for the case when we assumed the data came from some distribution $F(\theta)$ with unknown parameter θ.

Estimating θ, we then generated "bootstrap samples" from the distribution $F(\hat{\theta})$. The statistic $\hat{\theta}^*$ is then calculated for each sample.

We then use the analogy that the sampling distribution of $\hat{\theta}$ is to θ sampling distribution of $\hat{\theta}^*$ is to $\hat{\theta}$.

The nonparametric bootstrap uses the same basic analogy… except that we don't have a specific distribution in mind for *F*.

Because of this we the parameter θ that we are focusing on is usually something like the mean, variance, or median that is "universally defined."

Example: Estimate the variance and bias of the sample standard deviation *s* for the sample:

```
3.95  3.79  3.75  2.71  5.52
6.12  1.74  6.05  3.92  5.69
```

*Generated using x<-10*rbeta(10,3,4) so the population has mean 30/7≈4.29 and variance 150/49 ≈3.06 (sd ≈ 1.75).*


```
sdboot<-function(x,nboots=10000){
sampsize<-length(x)
bootsamps<-
  matrix(sample(x,sampsize*nboots,
  replace=T),ncol=sampsize)
bootstats<-apply(bootsamps,1,sd)
est.bias<-mean(bootstats)-sd(x)
est.se<-sd(bootstats)
c(est.bias,est.se)
}
```


How well does it work?

When can it have trouble?

- Small sample sizes (but doesn't everything?)
- Statistic is not smooth