STAT 621 Lecture Notes
CDF, EDF and GOF

As we discussed in the introduction, a traditional approach to nonparametric statistics is primarily concerned with relaxing distributional assumptions on a random sample. We'll start our discussion with one of the most basic concerns in this field – actually estimating the probability distribution that generated any particular random sample. In addition we'll define a formal testing procedure to decide if a random sample follows any specific probability distribution.

CDF: The Cumulative Distribution Function

Recall that the CDF of a random variable X is defined as

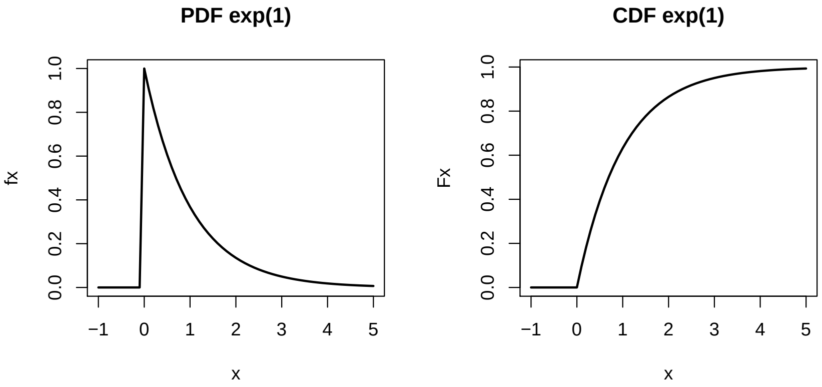
$$F(x) = P(X \leq x) \qquad -\infty < x < \infty$$

All random variables have a CDF, and that CDF is defined over the entire real line. What are some of the properties of the CDF?

Example: The probability density function (PDF) of an exponential random variable with parameter θ is $f(x) = \theta e^{-\theta x}$ for $0 \leq x$. Therefore the CDF is

$$F(x) = P(X \leq x) = \int_0^x \theta e^{-\theta t} dt = 1 - e^{-\theta x} \qquad 0 \leq x,$$

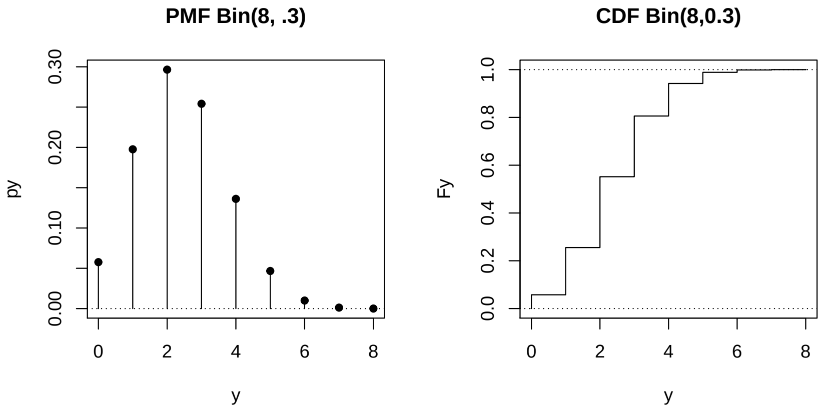
and $F(x) = 0$ if $x < 0$. The PDF and CDF of the exponential distribution with $\theta = 1$ are plotted below.



Example: If Y has a binomial distribution with n trials and success probability p, then its probability mass function (PMF) is $p(y) = \binom{n}{y} p^y (1-p)^{n-y}$. So its CDF is

$$F(y) = P(Y \leq y) = \sum_{k=0}^{y} \binom{n}{k} p^p (1-p)n - k \qquad 0 \leq y,$$

and $F(y) = 0$ if $y < 0$. Here is the PMF and CDF of the Binomial(0, 0.3).



EDF: The Empirical Distribution Function

The empirical distribution function (EDF) is an estimate of a random variable's cumulative distribution function. First of all, why might we want to estimate this function? Why not try the estimate the PDF or PMF instead? (Note: we will discuss estimating PDF's later in the semester).

Consider a random sample $X_1, X_2, \ldots, X_n$. Now for a given constant t, we would like to estimate

$$F(t) = P(X \leq t),$$

the probability that a random value of our variable X will be less than or equal to t. A sensible way to estimate this is with the proportion of values in our sample that are less than or equal to t. This is exactly the EDF,

$$\widehat{F}_n(t) = \frac{1}{n} \sum_i^n I(X_i \leq t).$$

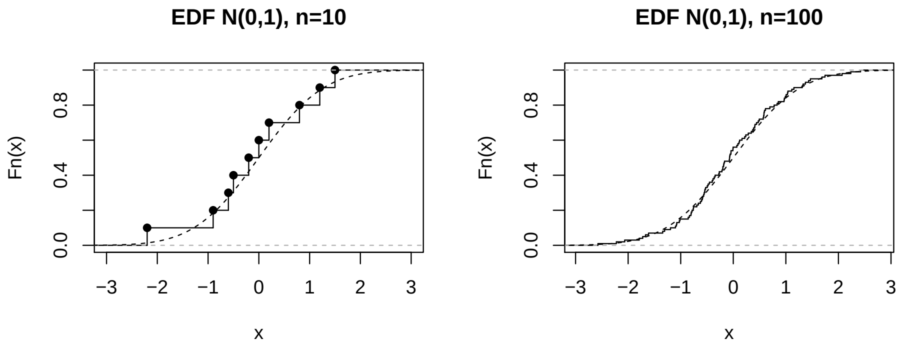
Here $I()$ is the *indicator function* that takes the value 1 if its argument is true, and 0 otherwise.

Example: Here is a random sample of size $n = 10$ from the $N(0,1)$ distribution.

```
> x1=sort(round(rnorm(10), 1))
> x1
      -2.2    -0.9    -0.6    -0.5    -0.2    0.0     0.2     0.8     1.2     1.5
```

Compute the EDF for a couple of values, say $t = -3, -.5, 1, 3$.

Below on the left is the EDF for the above sample, plotted along with the true CDF of the N(0,1) distribution. The plot on the right shows the EDF from a sample of $n = 100$.



```
x1=sort(round(rnorm(10), 1))
x1
Fhat1=ecdf(x1)
plot(Fhat1,verticals=T, xlim=c(-3,3), main="EDF N(0,1), n=10",pch=16)
x=seq(-4,4,.1)
Ftrue=pnorm(x)
lines(x,Ftrue, lty=2)

x2=rnorm(100)
Fhat2=ecdf(x2)
plot(Fhat2,verticals=T,do.points=F, main="EDF N(0,1), n=100")
lines(x,Ftrue,lty=2)
```

The statistical properties of the EDF are fairly straightforward to derive. This is because the EDF is really just a set of averages. Let's let $Y_i = I(X_i \leq t)$ for some constant t. What can we say about Y_i?

Given this, what can we say about the estimator $\widehat{F}_n(t)$ for a fixed t? Think about Bias, Variance, Consistency, Asymptotic Distribution...

GOF: Goodness of Fit Testing

The basic idea behind a Goodness of Fit test is to determine if a random sample follows some specified distribution. For example, one might want to formally test whether a sample is normally distributed or not. A GOF test is one way to do that. In other contexts, one may wish to determine whether two independent samples follow the same probability distribution.

There are many options for GOF tests, but probably the most well known is the **Kolmogorov-Smirnov Test**. This test can be used for either one or two-sample problems. In the case of the two-sample problem, distributions are considered continuous, but otherwise are unrestricted. Here are details and examples.

One-Sample KS Test:

Consider a random sample $X_1, \ldots X_n$ from an unknown probability distribution having CDF $F(t)$. Let $F_0(t)$ be the CDF of a known distribution. Interest is in testing the hypothesis

$$H_0 : F(t) = F_0(t) \text{ for all } t \qquad \text{vs.} \qquad H_A : F(t) \neq F_0(t) \text{ for some } t$$

The KS test statistic is given by

$$D = \sup_t |\widehat{F}_n(t) - F_0(t)|$$

where $F_n(t)$ is the EDF and sup is the *supremum* over all t. So what does this test statistic measure? What should we expect under the null? The alternative?

Kolmogorov derived the approximate distribution of the test statistic under the null, and this distribution is used in most software programs to report p-values.

Example: The following R code generates a random sample from a $N(0, 1)$ distribution. A plot of the EDF is made. The R funciton `ks.test` us used to test whether the data come from a standard normal distribution (which we know they do, since I generated them that way).

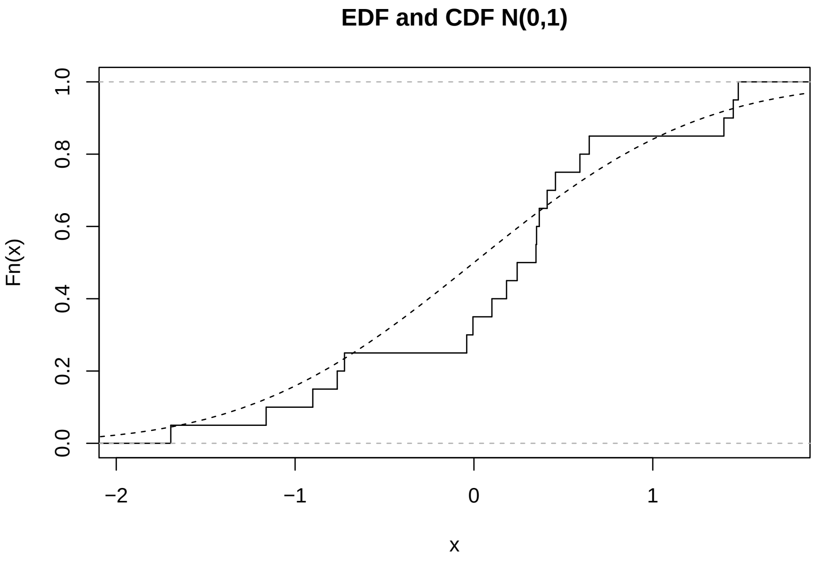
```
# KS test
x=rnorm(20)
Fhat=ecdf(x)
plot(Fhat, verticals=T, do.points=F, main="EDF and CDF N(0,1)")
temp=seq(-4,4,.1)
lines(temp,pnorm(temp),lty=2)
```

```
# The K-S test for one sample

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

One-sample Kolmogorov-Smirnov test

data:  x
D = 0.2339, p-value = 0.1909
alternative hypothesis: two-sided
```



Two-Sample KS Test:

Now consider two independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots Y_n$. Let $\widehat{F}_m(t)$ be the EDF for the X sample and let $\widehat{G}_n(t)$ be the EDF for the Y sample. Here we'd like to test whether these two samples are generated by the same underlying distribution. That is,

$$H_0 : F(t) = G(t) \text{ for all } t \qquad \text{vs.} \qquad H_A : F(t) \neq G(t) \text{ for some } t$$

where $F(t)$ and $G(t)$ are the true CDFs of X and Y. The KS test statistic is

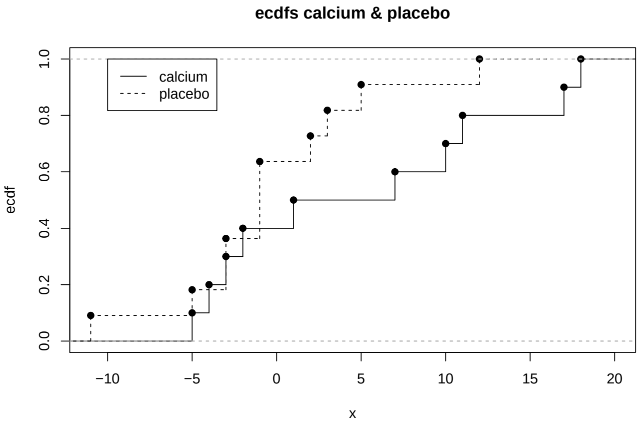
$$D = \sup_t |\widehat{F}_m(t) - \widehat{G}_n(t)|$$

Example: Does calcium in the diet lower blood pressure? A randomized experiment gave one group of 10 African American men a calcium supplement for 12 weeks and a second group of 11 African American men a placebo. The changes in blood pressure are listed below:

Calcium	7	-4	18	17	-3	-5	1	10	11	-2	
Placebo	-1	12	-1	-3	3	-5	5	2	-11	-1	-3

Is the calcium-treated population different from the placebo population, in general? First let's look at the empirical CDFs.

```
calcium=c(7, -4, 18, 17, -3, -5, 1, 10, 11, -2)
placebo=c(-1, 12, -1, -3, 3, -5, 5, 2, -11, -1, -3)
F=ecdf(calcium)
G=ecdf(placebo)
plot(F,verticals=T,ylab="ecdf",main="ecdfs calcium & placebo")
lines(G,verticals=T,lty=2)
legend(-6,1,c("calcium","placebo"),lty=c(1,2))
```



Below I compute the EDFs for both samples at each observed data value, and find the absolute differences. Identify the value of the KS test statistic.

```
m=length(calcium); n=length(placebo)
Z=sort(c(calcium,placebo))
Fm=rep(NA,m+n)
for(j in 1:(n+m)) Fm[j]=sum(calcium<=Z[j])/m
Gn=rep(NA,m+n)
for (i in 1:(m+n)) Gn[i]=sum(placebo<=Z[i])/n
absdiff=abs(Fm-Gn)
cbind(Z,Fm,Gn,absdiff)
```

```
        Z  Fm        Gn    absdiff
 [1,] -11 0.0 0.0909091 0.09090909
 [2,]  -5 0.1 0.1818182 0.08181818
 [3,]  -5 0.1 0.1818182 0.08181818
 [4,]  -4 0.2 0.1818182 0.01818182
 [5,]  -3 0.3 0.3636364 0.06363636
 [6,]  -3 0.3 0.3636364 0.06363636
 [7,]  -3 0.3 0.3636364 0.06363636
 [8,]  -2 0.4 0.3636364 0.03636364
 [9,]  -1 0.4 0.6363636 0.23636364
[10,]  -1 0.4 0.6363636 0.23636364
[11,]  -1 0.4 0.6363636 0.23636364
[12,]   1 0.5 0.6363636 0.13636364
[13,]   2 0.5 0.7272727 0.22727273
[14,]   3 0.5 0.8181818 0.31818182
[15,]   5 0.5 0.9090909 0.40909091
[16,]   7 0.6 0.9090909 0.30909091
[17,]  10 0.7 0.9090909 0.20909091
```

The `ks.test` function also does the two-sample test when given two data vectors as arguments.

```
ks.test(calcium,placebo)

          Two-sample Kolmogorov-Smirnov test

  data:  calcium and placebo
  D = 0.4091, p-value = 0.3446
  alternative hypothesis: two-sided

  Warning message: cannot compute correct p-values with ties in: ks.test(calcium, placebo)
```

Notes:

1. Many software programs report a scaled version of the KS test statistic, for example multiplied by $\sqrt{n}$ or other constant. In R it appears taht the scaling is done to the null distribuiton rather than the test statistic.

2. Ties are often problematic for Distribution-Free procedures. Here the warning message indicates that exact p-values are not available in the presence of ties. The reported p-value uses an approximation to the null distribution.

3. The KS test is very simple to understand and perform, but it's probably not the best option out there. It suffers from low power so it is hard to reject a null for small sample sizes. On the other hand, it's also fairly sensitive with large samples so that even small differences in EDFs can result in a rejection. Despite this it still seems to me widely used.

4. There are a number of other options including the *Cramer-von Mises* test and the *Anderson-Darling* test, as well as tests based on quantile plots.