Goodness of Fit Tests
Bios 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-10-01 11:05

Assessing Fit

- Graphical displays such as qqplot
- Tests
  - χ^2
  - Kolmogorov-Smirnov one-sample (page 279 text)
  - Others

KS GOF Test

- Kolmogorov-Smirnov GOF test (one sample test)
- We want to test that our data come from a known and completely specified distribution: $F_0(y)$

KS GOF Test

- The empirical distribution function (EDF) for a given data set is

$$F_n(y) = \begin{cases} 0 \text{ if } y < y_{(1)} \\ k/n \text{ if } y_{(k)} \le y < y_{(k+1)} \\ 1 \text{ if } y > y_{(n)} \end{cases}$$

  Note: text calls this *empirical cumulative distribution* (ECD); see age 32

KS GOF Test

- H_0: $Y_1, \ldots, Y_n \sim F_0(y)$
- The KS statistic for GOF is

$$D = \max_y |F_0(y) - F_n(y)|$$

- Exact and asymptotic distribution of D have been derived, tabulated
- Critical values on next slide are appropriate for $F_0(y)$ continuous

KS GOF Test

- Critical values for KS one sample test

n	0.05	0.01
10	.409	.489
15	.338	.404
16	.327	.392
17	.318	.381
18	.309	.371
19	.301	.363
20	.294	.352
25	.264	.317
30	.242	.290
35	.224	.269
> 35	$\frac{1.36}{\zeta}$	$\frac{1.63}{\zeta}$

where $\zeta = (n + \sqrt{n/10})^{1/2}$. Source: Conover, *Practical Nonparametric Statistics*, 1980, page 462.

KS GOF Test

- The KS statistic for GOF is

$$D = \max_y |F_0(y) - F_n(y)|$$

- Equivalently

$$D = \max\{D_1, \ldots, D_n\}$$

where

$$D_i \equiv \max\{i/n - z_{(i)}, z_{(i)} - (i-1)/n\}$$

and

$$z_{(i)} = F_0(y_{(i)})$$

KS GOF: Example

- A random sample of size 10

y_1	0.621
y_2	0.503
y_3	0.203
y_4	0.477
y_5	0.710
y_6	0.581
y_7	0.329
y_8	0.480
y_9	0.554
y_{10}	0.382

KS GOF: Example

- It is hypothesized the distribution of these samples is $U(0,1)$

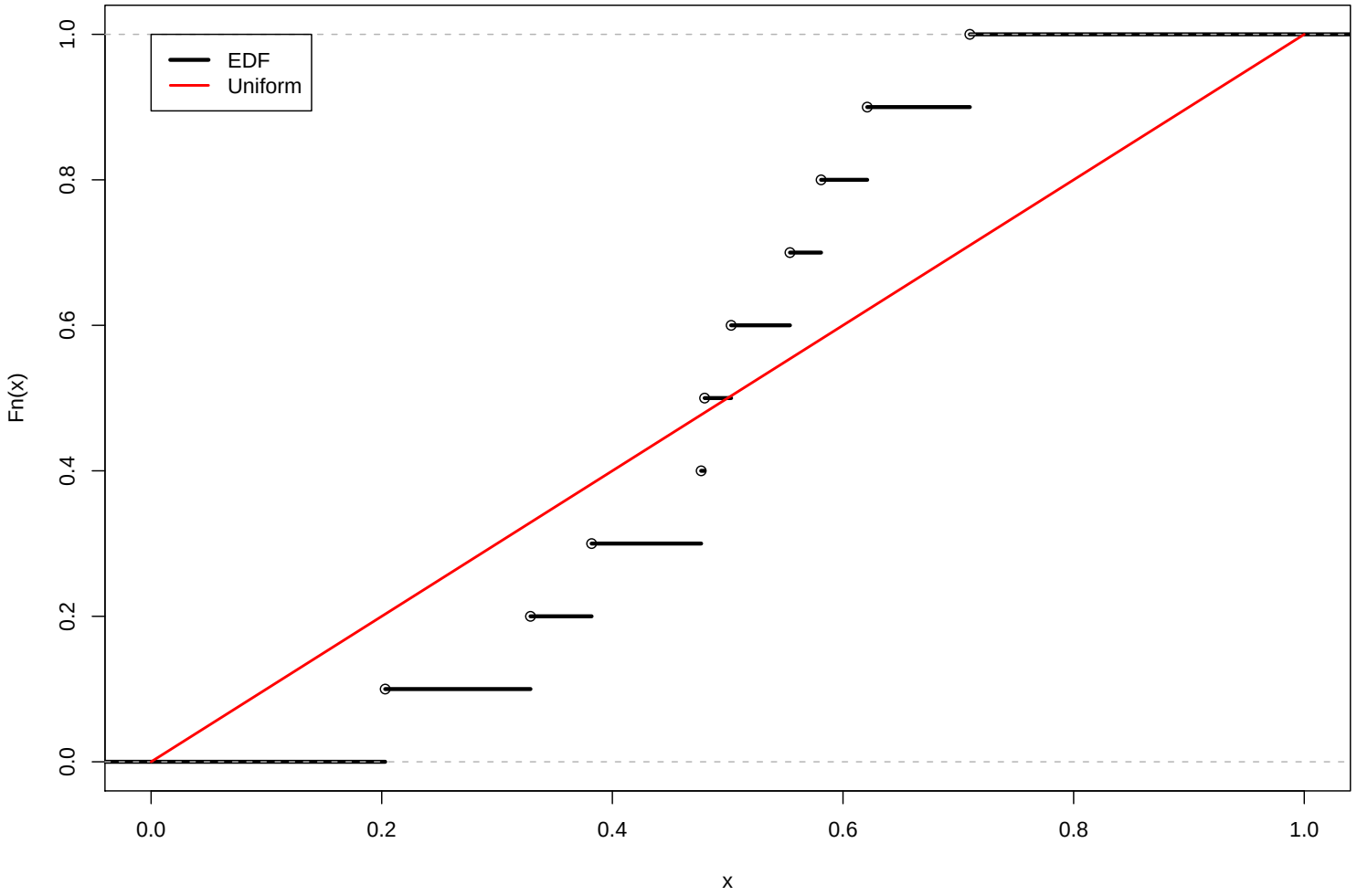
$$F_0(y) = \begin{cases} 0 \text{ if } y < 0 \\ y \text{ if } 0 \leq y \leq 1 \\ 1 \text{ if } 1 < y \end{cases}$$

- $n = 10$
- $C_{.05} = \{D > 0.409\}$
- On next slide we show $D = 0.290$; thus we do not reject H_0

KS GOF: Example

$y_{(i)}$	$F_0(y_{(i)})$	i/n	$(i-1)/n$	D_i
$y_{(1)}$	0.203	.1	0	.203
$y_{(2)}$	0.329	.2	.1	.229
$y_{(3)}$	0.382	.3	.2	.182
$y_{(4)}$	0.477	.4	.3	.177
$y_{(5)}$	0.480	.5	.4	.180
$y_{(6)}$	0.503	.6	.5	.097
$y_{(7)}$	0.554	.7	.6	.146
$y_{(8)}$	0.581	.8	.7	.219
$y_{(9)}$	0.621	.9	.8	.279
$y_{(10)}$	0.710	.10	.9	.290

KS GOF: Example



KS GOF

- The KS test requires that the parameters of $F_0(y)$ are known
- If they are estimated from the data, the distribution of D is not as in the table above.
- Critical values for KS statistic for testing normality when μ and σ^2 are estimated are given by Lilliefors (JASA 1967, p399)

Lilliefors KS GOF Test

- Critical values for KS test of normality

n	0.05	0.01
10	.258	.294
15	.220	.257
16	.213	.250
17	.206	.245
18	.200	.239
19	.195	.235
20	.190	.231
25	.173	.200
30	.161	.187
> 30	$\frac{.886}{\sqrt{n}}$	$\frac{1.031}{\sqrt{n}}$

- Source: Conover, *Practical Nonparametric Statistics*, 1980, page 463.

KS GOF: Example

- A random sample of size 10

y_1	0.621
y_2	0.503
y_3	0.203
y_4	0.477
y_5	1.160
y_6	0.581
y_7	0.329
y_8	0.480
y_9	0.554
y_{10}	0.382

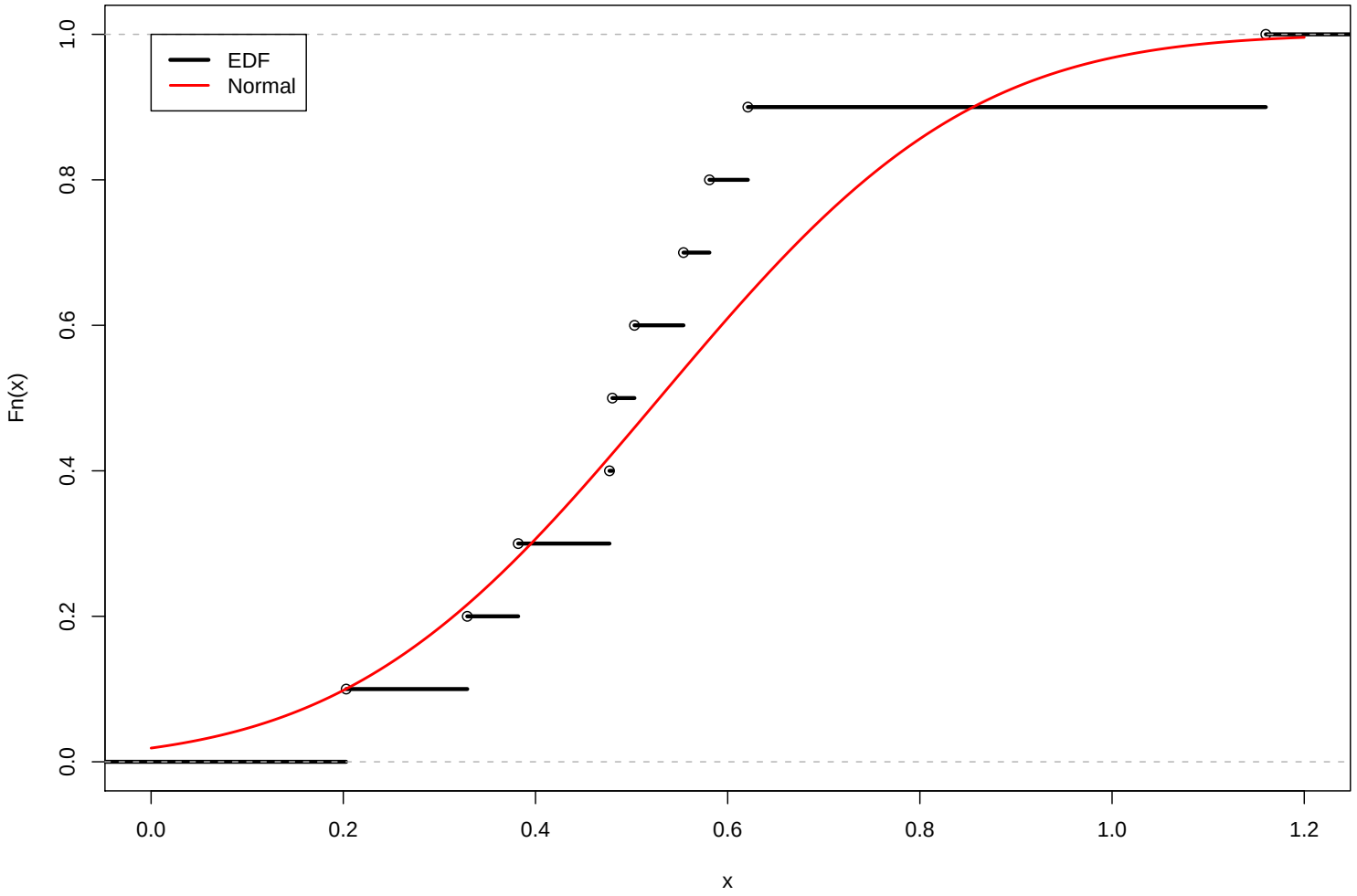
KS GOF: Example

- It is hypothesized the distribution of these samples is normal
- $\hat{\mu} = \bar{y} = 0.529$ and $\hat{\sigma} = s = 0.2546501$
- $C_{.05} = \{D > 0.258\}$
- For these data $D = 0.259$; $p \approx 0.05$

KS GOF: Example

$y_{(i)}$	$F_0(y_{(i)})$	i/n	$(i-1)/n$	D_i
$y_{(1)}$	0.100	.1	0	.100
$y_{(2)}$	0.216	.2	.1	.116
$y_{(3)}$	0.282	.3	.2	.082
$y_{(4)}$	0.419	.4	.3	.119
$y_{(5)}$	0.424	.5	.4	.076
$y_{(6)}$	0.459	.6	.5	.141
$y_{(7)}$	0.539	.7	.6	.161
$y_{(8)}$	0.581	.8	.7	.219
$y_{(9)}$	0.641	.9	.8	.259
$y_{(10)}$	0.993	1	.9	.093

KS GOF: Example



KS GOF: SAS

- SAS: use Proc Univariate w/ NORMAL option or HISTOGRAM statement

```
proc univariate normal; var x;
```

Tests for Normality

Test		--Statistic---	-----p Value------	
Shapiro-Wilk	W	0.835123	Pr < W	0.0386
Kolmogorov-Smirnov	D	0.258945	Pr > D	0.0560
Cramer-von Mises	W-Sq	0.116363	Pr > W-Sq	0.0587
Anderson-Darling	A-Sq	0.710057	Pr > A-Sq	0.0444

KS GOF: R

- R: *ks.test()*; however, beware of ties:

```
>  ks.test(rnorm(100000,0,1),"pnorm",0,1)

        One-sample Kolmogorov-Smirnov test

data:  rnorm(1e+05, 0, 1)
D = 0.0021, p-value = 0.783
alternative hypothesis: two.sided

>  ks.test(rpois(100000,3),"ppois",3)

        One-sample Kolmogorov-Smirnov test

data:  rpois(1e+05, 3)
D = 0.2237, p-value < 2.2e-16
alternative hypothesis: two.sided

Warning message:
cannot compute correct p-values with ties in: ks.test(rpois(1e+05, 3), "ppois", 3)
```

KS GOF: R

- Lilliefors

  SAS: automatic

  R: use "nortest" package

```
> ks.test(x,"pnorm",mean(x),sd(x))

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.2589, p-value = 0.4402
alternative hypothesis: two-sided


> install.packages("nortest")
> library(nortest)
> lillie.test(x)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  x
D = 0.2589, p-value = 0.05602
```

KS vs χ^2 GOF Tests

- If data continuous, KS preferred. Why?
  - If sample size small, KS is exact, while χ^2 relies on large sample approximation
  - KS test more powerful than χ^2 in most situations (Conover, *Practical Nonparametric Statistics*, 1980 p 346)
  - Do not need to bin
- If discrete/categorical, χ^2 preferred

Other GoF Tests

- Wilk-Shapiro: see Conover page 363, Tables A.17, A.18

$$\frac{[\sum_{i=1}^{k} a_i(X_{(n-i+1)} - X_{(i)})]^2}{s^2}$$

  where s^2 is the sample variance and a_i are given

- Under null (i.e., normality), numerator and denominator both estimating (up to a constant) σ^2

- R: shapiro.test()

Other GoF Tests

- Class of GoF test statistics

$$n \int \{F_n(y) - F_0(y)\}^2 \psi(y) dy$$

- Anderson-Darling $\psi(y) = \{F_0(y)(1 - F_0(y))\}^{-1}$
- Cramer-von Mises $\psi(y) = 1$
- R nortest package: ad.test(), cvm.test()