BIOS 662 Fall 2018

Goodness-of-Fit Tests

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Assessing Fit

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

Kolmogorov-Smirnov Goodness-of-Fit Test

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

Kolmogorov-Smirnov Goodness-of-Fit 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: The text calls this the *empirical cumulative distribution* (ECD) – Definition 3.9 on page 32

Kolmogorov-Smirnov Goodness-of-Fit Test

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

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

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

Kolmogorov-Smirnov Goodness-of-Fit Test

- Critical values for the KS one sample test

n	$\alpha = 0.05$	$\alpha = 0.01$
10	0.409	0.489
15	0.338	0.404
16	0.327	0.392
17	0.318	0.381
18	0.309	0.371
19	0.301	0.363
20	0.294	0.352
25	0.264	0.317
30	0.242	0.290
35	0.224	0.269
>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.

Kolmogorov-Smirnov Goodness-of-Fit Test

- The KS statistic for goodness-of-fit is

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

- Equivalently

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

where

$$D_i \equiv \max\left(\frac{i}{n} - x_{(i)},\, x_{(i)} - \frac{(i-1)}{n}\right)$$

and

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

KS GOF Test: Example

- Consider this 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 Test: Example cont.

- It is hypothesized that this sample is from the $U(0,1)$ distribution

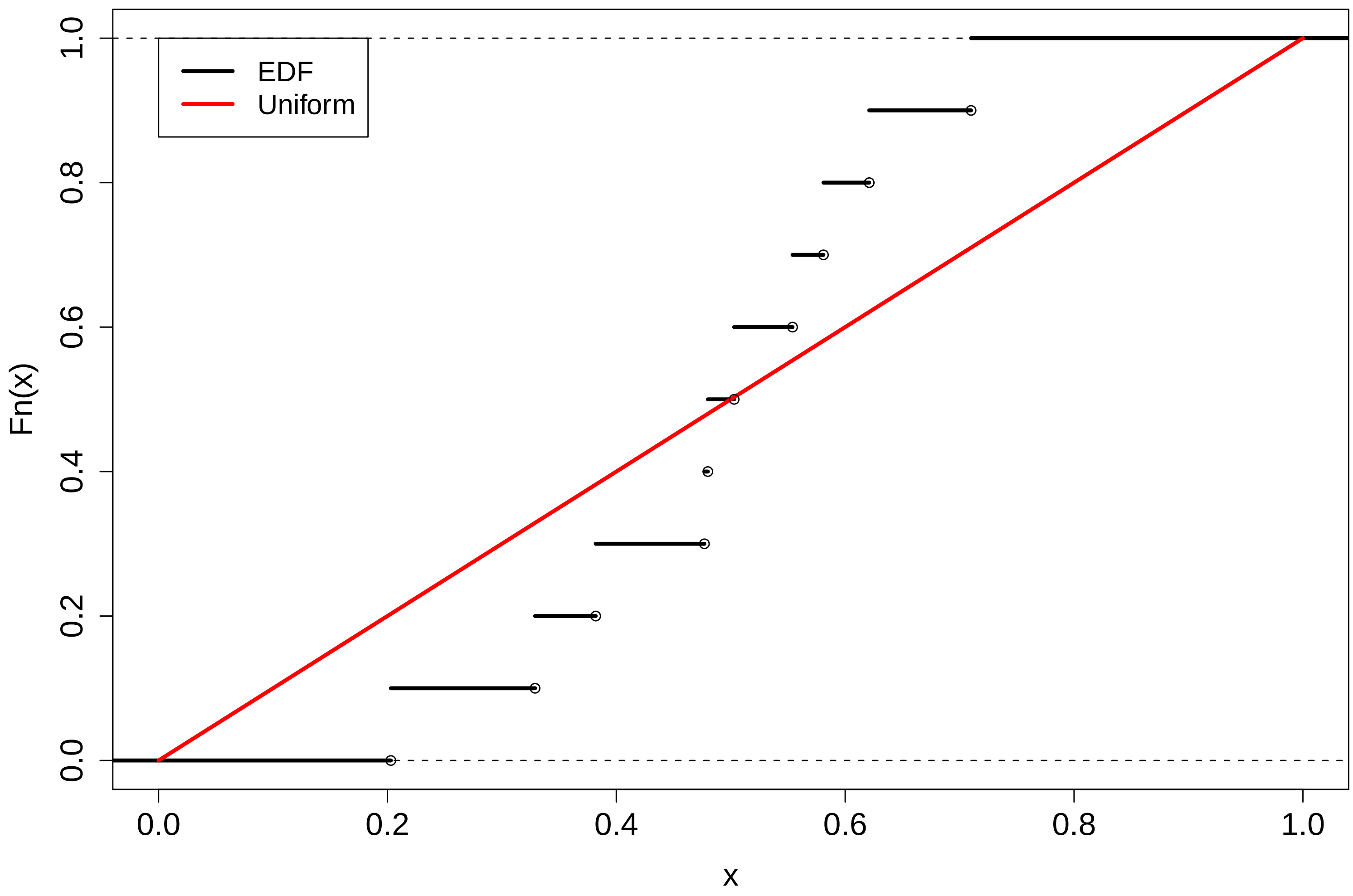
$$F_0(y) = \begin{cases} 0 & \text{if} \quad y < 0 \\ y & \text{if} \quad 0 \le y \le 1 \\ 1 & \text{if} \quad 1 < y \end{cases}$$

- $n = 10$
- $C_{0.05} = \{D : D > 0.409\}$
- On the next page we show that $D = 0.290$; thus we do not reject H_0

KS GOF Test: Example cont.

$y_{(i)}$	$F_0(y_{(i)})$	i/n	$(i-1)/n$	D_i
$y_{(1)}$	0.203	0.1	0.0	0.203
$y_{(2)}$	0.329	0.2	0.1	0.229
$y_{(3)}$	0.382	0.3	0.2	0.182
$y_{(4)}$	0.477	0.4	0.3	0.177
$y_{(5)}$	0.480	0.5	0.4	0.080
$y_{(6)}$	0.503	0.6	0.5	0.097
$y_{(7)}$	0.554	0.7	0.6	0.146
$y_{(8)}$	0.581	0.8	0.7	0.219
$y_{(9)}$	0.621	0.9	0.8	0.279
$y_{(10)}$	0.710	1.0	0.9	0.290

KS GOF Test: Example cont.



Kolmogorov-Smirnov Goodness-of-Fit Test

- 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 several pages back
- Critical values for KS statistic for testing normality when μ and σ^2 are estimated are given by Lilliefors (JASA 1967, p. 399)

Lilliefors KS GOF Test

- Critical values for KS test of normality

n	$\alpha = 0.05$	$\alpha = 0.01$
10	0.258	0.294
15	0.220	0.257
16	0.213	0.250
17	0.206	0.245
18	0.200	0.239
19	0.195	0.235
20	0.190	0.231
25	0.173	0.200
30	0.161	0.187
>30	$\frac{0.886}{\sqrt{n}}$	$\frac{1.031}{\sqrt{n}}$

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

KS GOF: Example

- Consider this 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

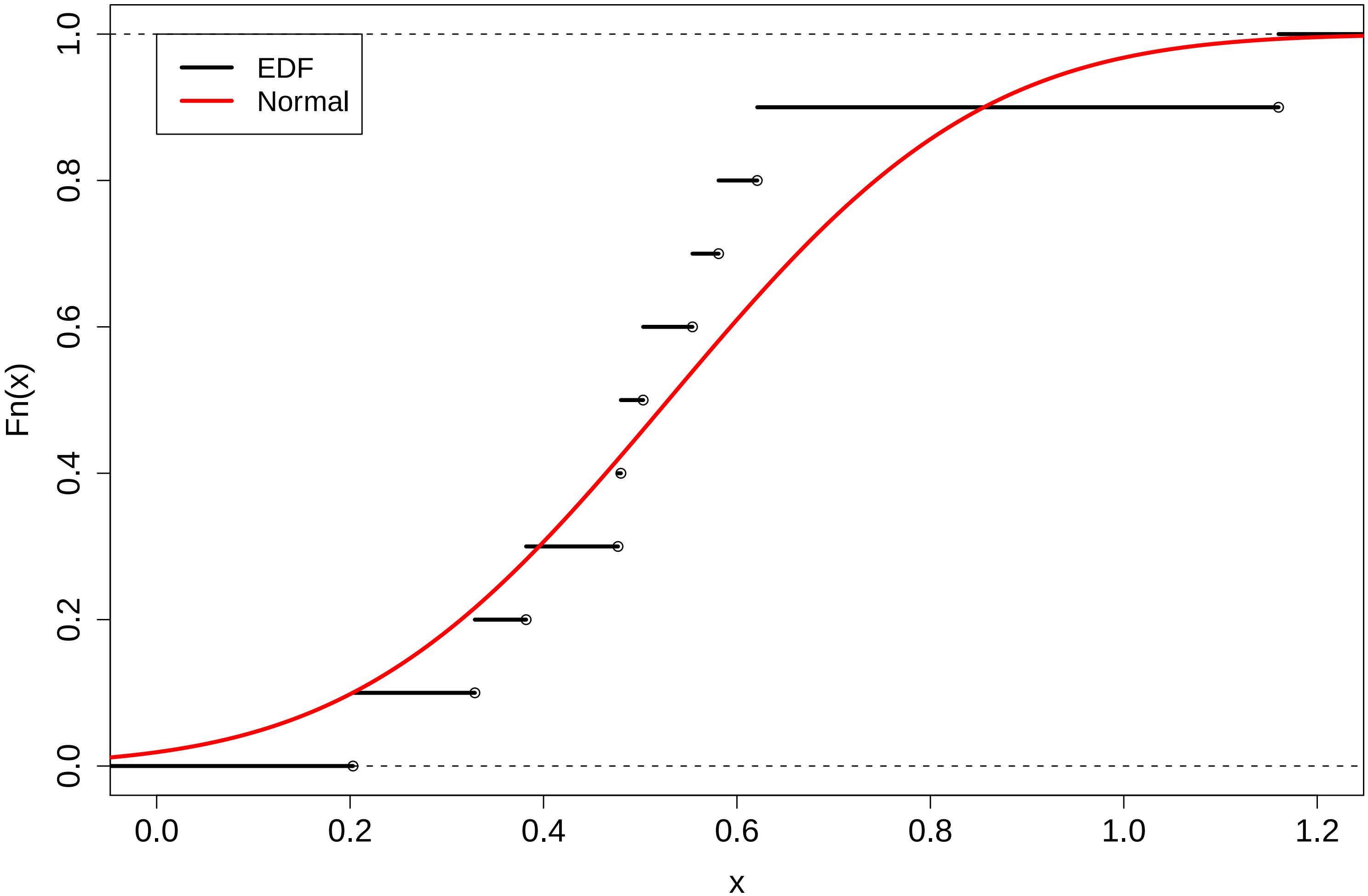
KS GOF: Example cont.

- It is hypothesized that this sample is from a normal distribution
- $\hat{\mu} = \bar{y} = 0.529$ and $\hat{\sigma} = s = 0.2546501$
- $C_{0.05} = \{D : D > 0.258\}$
- For these data $D = 0.259$; $p \approx 0.05$

KS GOF: Example cont.

	$y_{(i)}$	$F_0(y_{(i)})$	i/n	$(i-1)/n$	D_i
$y_{(1)}$	0.203	0.100	0.1	0.0	0.100
$y_{(2)}$	0.329	0.216	0.2	0.1	0.116
$y_{(3)}$	0.382	0.282	0.3	0.2	0.082
$y_{(4)}$	0.477	0.419	0.4	0.3	0.119
$y_{(5)}$	0.480	0.424	0.5	0.4	0.076
$y_{(6)}$	0.503	0.459	0.6	0.5	0.141
$y_{(7)}$	0.554	0.539	0.7	0.6	0.161
$y_{(8)}$	0.581	0.581	0.8	0.7	0.219
$y_{(9)}$	0.621	0.641	0.9	0.8	0.259
$y_{(10)}$	1.160	0.993	1.0	0.9	0.093

KS GOF: Example cont.



KS GOF: SAS

- SAS: use proc univariate with the option "normal" or the "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: SAS

```
proc univariate;
   histogram x / normal;
** The plot isn't meaningful with so few observations;
** The table has a different heading;
```

```
The UNIVARIATE Procedure
Fitted Normal Distribution for x

        Goodness-of-Fit Tests for Normal Distribution

Test                  ----Statistic-----   ------p Value------

Kolmogorov-Smirnov    D       0.25894505   Pr > D        0.056
Cramer-von Mises      W-Sq    0.11636316   Pr > W-Sq     0.059
Anderson-Darling      A-Sq    0.71005670   Pr > A-Sq     0.044
```

KS GOF: R

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

```
> set.seed(34621)
> ks.test(rnorm(100000,0,1),"pnorm",0,1)

        One-sample Kolmogorov-Smirnov test

data:  rnorm(1e+05, 0, 1)
D = 0.0032, p-value = 0.2591
alternative hypothesis: two-sided

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

        One-sample Kolmogorov-Smirnov test

data:  rpois(1e+05, 3)
D = 0.2243, p-value < 2.2e-16
alternative hypothesis: two-sided

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

Lilliefors KS GOF: SAS / R

- SAS: automatic
- R: use "nortest" package

```
> x<-c(0.621,0.503,0.203,0.477,1.16,0.581,0.329,0.480,0.554,0.382)

> 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")
> lillie.test(x)

        Lilliefors (Kolmogorov-Smirnov) normality test

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

KS vs χ^2 Goodness-of-Fit Tests

- If data are continuous, KS preferred. Why?
  - If sample size small, KS is exact, whereas χ^2 relies on large sample approximation
  - KS test is 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 Goodness-of-Fit Tests

- Shapiro-Wilk test for normality: see Conover p. 363, Tables A.17, A.18

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

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

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

- R: shapiro.test()

Other Goodness-of-Fit Tests

- Class of goodness-of-fit 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()
- SAS: Automatic with "proc univariate normal;"