# Statistics of the Kolmogorov-Smirnov Type (Conover Chapter Six)

STAT 345-01: Nonparametric Statistics *

Fall Semester 2018

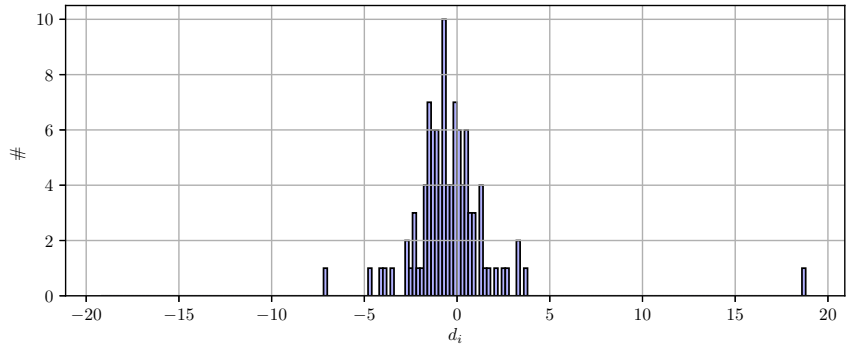
## Contents





## Thursday 25 October 2018 – Read Section 6.1 of Conover

From time to time one considers the question: does a data sample $\{x_i\}$ seem like it could have come from a particular distribution. E.g., on the first prelim exam, you examined the relative power of the $t$-test and sign test on data with the following histogram:




Since the data were not drawn from a normal distribution (as evidenced by the long tails), the $t$-test was not the most powerful option. But how does one quantify how different a data sample

looks from the expectation under some hypothesized distribution? That is the topic of the tests we consider now, starting with variants of the Kolmogorov-Smirnov test.

# 1 The Kolmogorov Test

## 1.1 Test Statistic(s)

In the example above we drew a histogram, and one could imagine comparing the shape of that histogram to the probability density function of a hypothesized continuous distribution. However, that involves a somewhat arbitrary choice in how the bins of the histogram are chosen. Kolmogorov-Smirnov tests instead base their comparisons on the *cumulative* distribution function. The hypotheses concern a comparison between the true cdf $F(x)$ associated with the sampling distribution and some hypothesized cdf $F^*(x)$. The null hypothesis $H_0$ is that $F(x) = F^*(x)$ for all $x$. The alternative hypothesis can be:

- Two-sided: $F(x) \neq F^*(x)$ for some (unspecified) $x$
- One-sided: $F(x) < F^*(x)$ for some (unspecified) $x$
- One-sided: $F(x) > F^*(x)$ for some (unspecified) $x$

Recall that back at the start of the course, we defined the empirical cdf $\hat{F}(x; \{x_i\})$ (which Conover calls $S(x)$) as the fraction of sample values with $x_i \leq x$ for a given $x$, i.e.,

$$\hat{F}(x; \{x_i\}) = \frac{1}{n} \sum_{i=1}^{n} I[x_i \leq x] \tag{1.1}$$

We use as test statistics the maximum separation between the curves $\hat{F}(x; \{x_i\})$ and $F^*(x)$:

$$T^+ = \sup_x [F^*(x) - \hat{F}(x; \{x_i\})] \tag{1.2a}$$

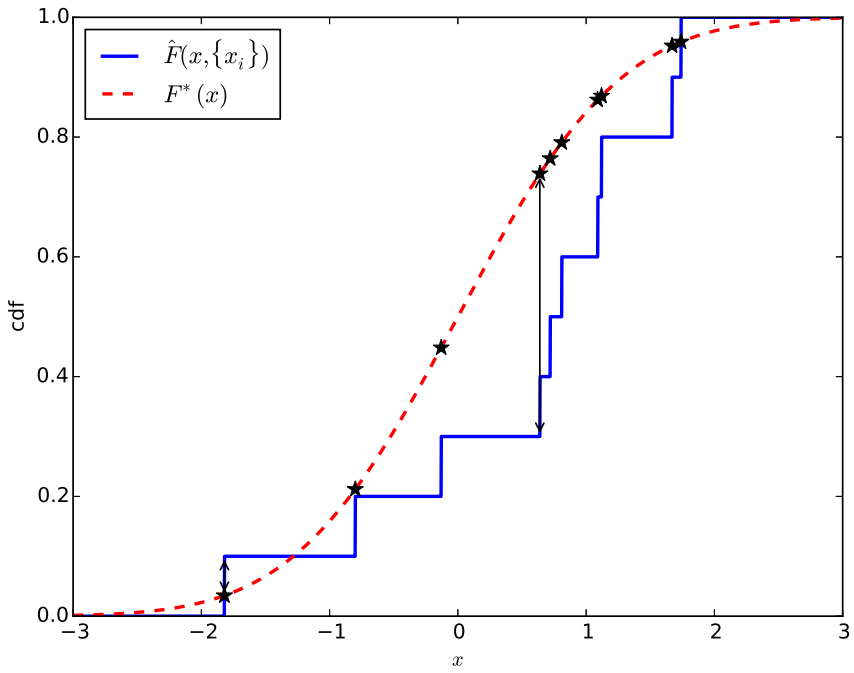
$$T^- = \sup_x [\hat{F}(x; \{x_i\}) - F^*(x)] \tag{1.2b}$$

$$T = \max(T^+, T^-) = \sup_x \left| F^*(x) - \hat{F}(x; \{x_i\}) \right| \tag{1.2c}$$

We can illustrate this for a particular sample of data, under the null hypothesis that they are drawn from a standard normal distribution, plotting the empirical and hypthesized distributions:

```
In [1]: from __future__ import division

In [2]: import numpy as np

In [3]: from scipy import stats

In [4]: xi = np.array([-1.82, 0.72, 1.67, 1.09,
   0.64, 0.81, 1.74, -0.80, -0.13, 1.12])

In [5]: n = len(xi); n
Out[5]: 10

In [6]: x = np.linspace(-3,3,6001)

In [7]: Fhat = np.mean(xi[None,:] <= x[:,None],axis
   =-1)

In [8]: Fstar = stats.norm.cdf(x)

In [9]: figure();
```

```
In [10]: plot(x,Fhat,'b-',lw=2,label=r'$\hat{F}(x,\{
   x_i\})$');

In [11]: plot(x,Fstar,'r--',lw=2,label=r'$F^*(x)$');

In [12]: ip = np.argmax(Fstar-Fhat); x[ip]
Out[12]: 0.63900000000000023

In [13]: im = np.argmax(Fhat-Fstar); x[im]
Out[13]: -1.8200000000000001

In [14]: annotate('',xy=(x[ip],Fstar[ip]),xycoords='
   data',xytext=(x[ip],Fhat[ip]),textcoords='data',
   arrowprops=dict(arrowstyle='<->'));

In [15]: annotate('',xy=(x[im],Fstar[im]),xycoords='
   data',xytext=(x[im],Fhat[im]),textcoords='data',
   arrowprops=dict(arrowstyle='<->'));

In [16]: xlabel(r'$x$');

In [17]: ylabel('cdf');

In [18]: legend(loc='upper left');

In [19]: Tp = max(Fstar-Fhat); Tp
Out[19]: 0.43858853405513848

In [20]: Tm = max(Fhat-Fstar); Tm
Out[20]: 0.065620497554110035
```




We've indicated on the plot the location of $T^+ = \sup_x[F^*(x) - \hat{F}(x;\{x_i\})] \approx 0.439$ and $T^- = \sup_x[\hat{F}(x;\{x_i\}) - F^*(x)] \approx 0.066$ on the plot. We notice that $T^-$ occurs right at the data point $x_i = -1.82$ and $T^+$ occurs right before the data point $x_i = 0.64$.

```
In [21]: x[ip+1]
Out[21]: 0.64000000000000012
```

If the hypothesized distribution is continuous, the maximum and minimum of $[F^*(x) - \hat{F}(x;\{x_i\})$ will always occur at the location of actual data points, where the empirical distribution function $\hat{F}(x;\{x_i\})$ makes discontinous jumps. In particular, $T^-$ always occurs right after a jump and $T^+$ right before a jump. That means we don't actually need to check the cdfs for all $x$ values, only those right before and after actual data points:

```
In [22]: xisort = np.sort(xi)

In [23]: Fstari = stats.norm.cdf(xisort)

In [24]: plot(xisort,Fstari,'k*',ms=10);

In [25]: savefig('notes06_cdfs.eps',bbox_inches='tight');

In [26]: Fhatip = np.arange(n)/n

In [27]: Fhatim = (1+np.arange(n))/n

In [28]: xisort[np.argmax(Fstari-Fhatip)]
Out[28]: 0.64000000000000001

In [29]: max(Fstari-Fhatip)
Out[29]: 0.43891370030713844

In [30]: xisort[np.argmax(Fhatim-Fstari)]
Out[30]: -1.8200000000000001

In [31]: max(Fhatim-Fstari)
Out[31]: 0.065620497554110035
```

## 1.2 $p$-Value for Continuous Distributions

To get the $p$-value assocated with $T^+$ and/or $T^-$, we need to consider their null distribution; for either statistic, the null distribution is the Kolmogorov distribution, with a single parameter $p$. This distribution has a number of forms and somewhat complicated derivations, but a relatively straightforward form appears in Z. W. Birnbaum and Fred H. Tingey, *Annals of Mathematical Statistics* **22**, 592 (1951), available at http://dx.doi.org/10.1214/aoms/1177729550:

$$P(T^{\pm} \geq t^{\pm}) = t^{\pm} \sum_{j=0}^{[n(1-t^{\pm})]} \binom{n}{j} \left(1 - t^{\pm} - \frac{j}{n}\right)^{n-j} \left(t^{\pm} + \frac{j}{n}\right)^{j-1} \tag{1.3}$$

where $[n(1-t^{\pm})]$ is the largest integer less than or equal to $n(1-t^{\pm})$, and $\binom{n}{j} = \frac{n!}{j!(n-j)!}$ is the usual binomial coëfficient. Note that we can split out the $j = 0$ term in the sum and write

$$\begin{aligned} &P(T^{\pm} \geq t^{\pm}) \\ &= (1-t^{\pm})^n + t^{\pm} \sum_{j=1}^{[n(1-t^{\pm})]} \binom{n}{j} \left(1 - t^{\pm} - \frac{j}{n}\right)^{n-j} \left(t^{\pm} + \frac{j}{n}\right)^{j-1}; \end{aligned} \tag{1.4}$$

in this form, it's apparent that $P(T^{\pm} \geq 0) = 1$. We can understand that $T^+$ and $T^-$ cannot be negative, since $\hat{F}(x; \{x_i\})$ will be zero for $x$ below the lowest value in the sample (which means $F^*(x) - \hat{F}(x; \{x_i\})$ cannot be negative everywhere) and $\hat{F}(x; \{x_i\})$ will be one for $x$ above the highest value in the sample (which means $\hat{F}(x; \{x_i\}) - F^*(x)$ cannot be negative everywhere). We can implement this sum, and calculate the one-sided $p$-value associated with $T^+ \approx 0.439$, and likewise with $T^- \approx 0.066$:

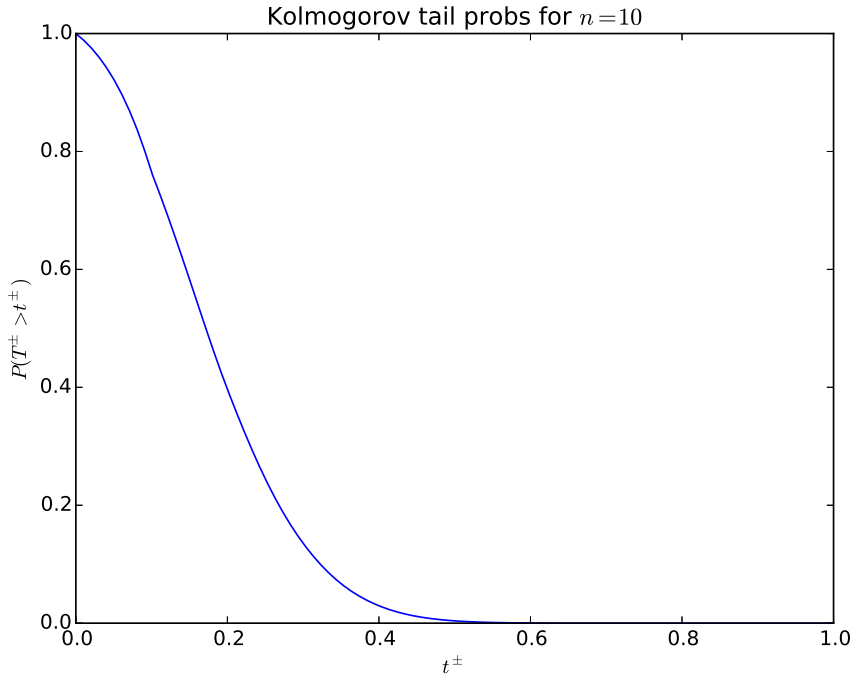
```
In [33]: n_p = int(n*(1-Tp)); n_p
Out[33]: 5

In [34]: Tp * np.sum([binom(n,j)*(1-Tp-j/n)**(n-j)*(
   Tp+j/n)**(j-1) for j in range(n_p+1)])
Out[34]: 0.014367272324540398

In [35]: n_m = int(n*(1-Tm)); n_m
Out[35]: 9
```

```
In [36]: Tm * np.sum([binom(n,j)*(1-Tm-j/n)**(n-j)*(
   Tm+j/n)**(j-1) for j in range(n_m+1)])
Out[36]: 0.88373135523717439
```

And so the two-sided $p$-value is 0.028, which indicates that these data are not consistent, at the 3% level, with being sampled from a standard normal distribution.

Incidentally, we can plot the tail probability for this sample size:

```
In [37]: Tpmvals = np.linspace(1e-4,1,100)

In [38]: tailprobs = np.array([Tpm * np.sum([binom(n
   ,j)*(1-Tpm-j/n)**(n-j)*(Tpm+j/n)**(j-1) for j in
   range(int(n*(1-Tpm))+1)]) for Tpm in Tpmvals])

In [39]: figure();

In [40]: plot(Tpmvals,tailprobs);

In [41]: xlabel(r'$t^{\pm}$');

In [42]: ylabel(r'$P(T^{\pm}>t^{\pm})$');

In [43]: title(r'Kolmogorov tail probs for $n=%d$'%n
   );

In [44]: savefig('notes06_Ktail.eps',bbox_inches='
   tight');
```




For large $n$ (more than about 40), we can use the asymptotic expression

$$P(T^{\pm} \geq t^{\pm}) \approx e^{-2n(t^{\pm})^2} \tag{1.5}$$

Obviously, it won't be good for $n = 10$, but we can check it:

```
In [47]: T95 * np.sum([binom(n,j)*(1-T95-j/n)**(n-j)
   *(T95+j/n)**(j-1) for j in range(int(n*(1-T95))
   +1)])
Out[47]: 0.025111729583049587
```

and the one-sided $p$-value of 2.5% is indeed a bit off from the exact value of 1.4%.

## 1.3 Distribution Confidence Interval

We can use the Kolmogorov test to define a confidence interval associated with the point estimate $\hat{F}(x;\{x_i\})$. E.g., for a 95% confidence interval, we should take the 97.5th percentile of the Kolmogorov distribution, $t^{\pm}_{0.975} = t_{0.95}$, and set the ends of the confidence interval as

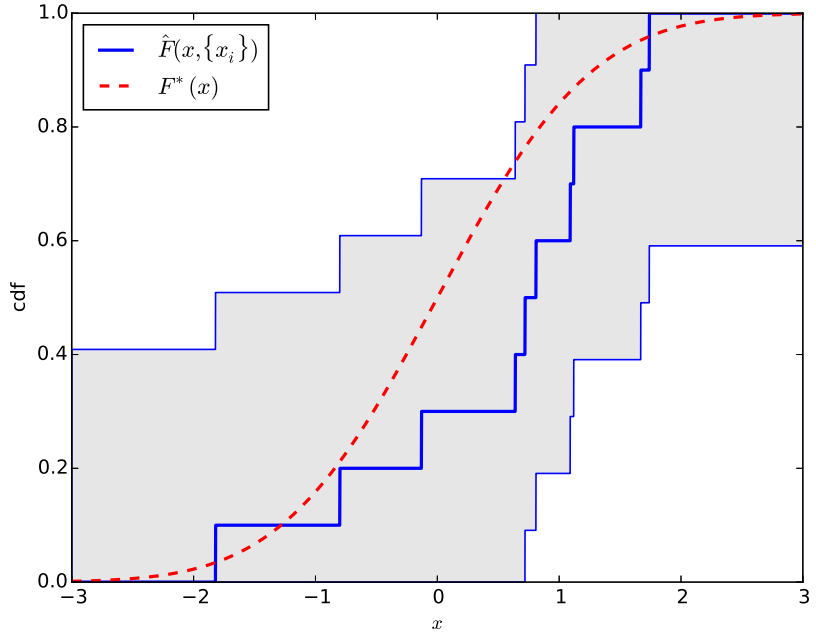
$$\hat{F}(x;\{x_i\}) \pm t_{0.95} \ ; \tag{1.6}$$

if the hypothesized cdf is within those limits, the empirical cdf will be close enough that the $p$-value of the two-sided Kolmogorov test will be above 5%. Of course, we also know that for *any* cdf, $0 \le F(x) \le 1$, so if the upper end of the interval goes above 1, we set it to 1, and likewise if the lower end goes below zero. We can't easily use the expression (1.4) to get the quantiles of the Kolmogorov distribution, but we can look them up, e.g., in Table A13 of Conover, and check that they give the right tail probabilities. That table says that the threshold for a one-tailed test at the 97.5% level, or a two-tailed test at 95%, is 0.409. We check this using our formula:

```
In [45]: T95 = 0.409

In [46]: T95 * np.sum([binom(n,j)*(1-T95-j/n)**(n-j)
    *(T95+j/n)**(j-1) for j in range(int(n*(1-T95))
    +1)])
Out[46]: 0.025111729583049587
```

and it does indeed give a tail probability of about 2.5%. Now we can construct and plot the confidence interval on the cdf:

```
In [47]: F95lower = np.maximum(0.,Fhat-T95)

In [48]: F95upper = np.minimum(1.,Fhat+T95)

In [49]: figure();

In [50]: fill_between(x,F95upper,F95lower,edgecolor='b',
    color=[0.9,0.9,0.9],label='95\% CI');

In [51]: plot(x,Fhat,'b-',lw=2,label=r'$\hat{F}(x,\{x_i
    \})$');

In [52]: plot(x,Fstar,'r--',lw=2,label=r'$F^*(x)$');

In [53]: xlabel(r'$x$');

In [54]: ylabel('cdf');

In [55]: legend(loc='upper left');

In [56]: savefig('notes06_CI.eps',bbox_inches='tight');
```



We've also plotted the standard normal CDF for reference. You can see that it does indeed pass outside the region associated with the 95% confidence interval, which we expect since we found a two-sided $p$-value of $0.028 < 0.05$.

### 1.4 $p$-Value for Discrete Distributions

**Tuesday 30 October 2018**
**– Read Section 6.2 of Conover**

# 2 Tests for Families of Distributions

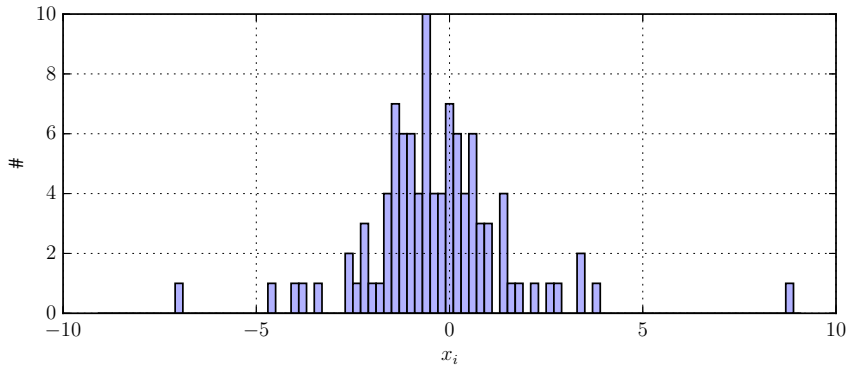
## 2.1 The Lilliefors Test for Normality

The Kolmogorov test assumes we have a single distribution with cdf $F^*(x)$ to which we want to compare the data. But often we want to know if the sampling distribution comes from a family of distributions, e.g., the null hypothesis is that it is a normal distribution with some unspecified parameters $\mu$ and $\sigma$. An obvious generalization is to estimate $\mu$ and $\sigma$ from the data. The statistic is then

$$T = \sup_x \left| \Phi\left(\frac{x-\overline{x}}{s}\right) - \hat{F}(x;\{x_i\}) \right| \tag{2.1}$$

where $\Phi\left(\frac{x-\mu}{\sigma}\right)$ is the cdf of a normal distribution with mean $\mu$ and standard deviation $\sigma$. When the parameters of the distribution are estimated from the data, the test is known as the *Lilliefors Test*.

As an example, consider a very slightly modified version of the data set from the prelim, which has histogram




```
In [1]: from __future__ import division

In [2]: import numpy as np

In [3]: from scipy import stats

In [4]: xi = np.loadtxt('notes06_prelim.dat')

In [5]: n = len(xi); n
Out[5]: 100

In [6]: xbar = np.mean(xi); xbar
Out[6]: -0.34903401460401001

In [7]: s = np.std(xi); s
Out[7]: 1.8482707855708738

In [8]: xi.sort()

In [9]: Fhatip = np.arange(n)/n

In [10]: Fhatim = (1+np.arange(n))/n
```

```
In [11]: stardist = stats.norm(loc=xbar,scale=s)

In [12]: Fstari = stardist.cdf(xi)

In [13]: Tp = max(Fstari-Fhatip); Tp
Out[13]: 0.10591766131171515

In [14]: Tm = max(Fhatim-Fstari); Tm
Out[14]: 0.11233110037923455
```

As before, we don't need to calculate the difference between the empirical cdf $\hat{F}(x;\{x_i\})$ at every possible $x$, just at the actual data values, using

$$\hat{F}(x^i;\{x_i\}) = \frac{i}{n} \tag{2.2a}$$

$$\hat{F}(x^i - \epsilon;\{x_i\}) = \frac{i-1}{n} \tag{2.2b}$$

$$\tag{2.2c}$$

Note that the test is equivalent to first converting the data $\{x_i\}$ using

$$z_i = \frac{x_i - \overline{x}}{s} \tag{2.3}$$

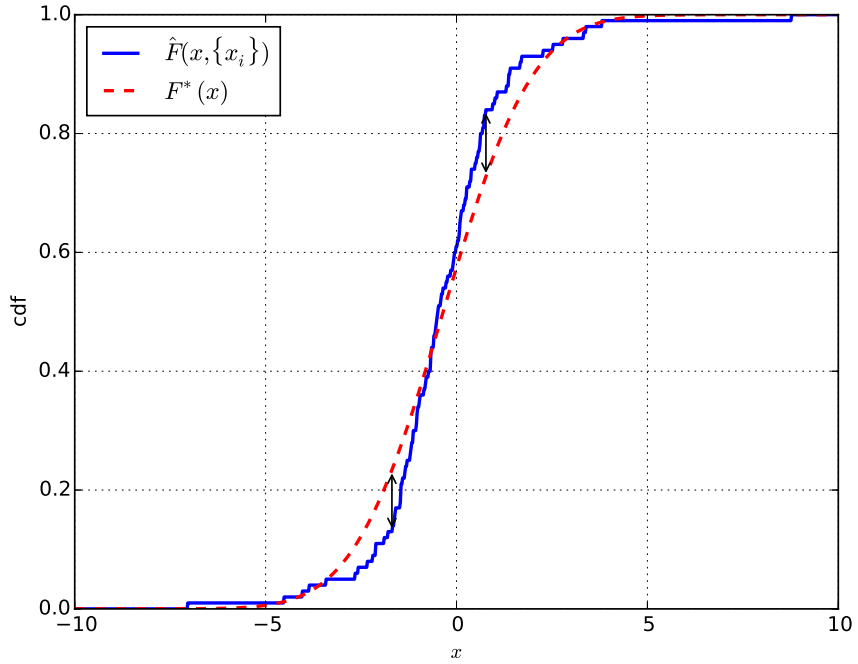
and then constructing the Kolomogorov statistics of the $\{z_i\}$ where the target distribution is standard normal:

```
In [15]: zi = (xi - xbar)/s

In [16]: Fstarzi = stats.norm.cdf(zi)

In [17]: max(Fstari-Fhatip)
Out[17]: 0.10591766131171515

In [18]: max(Fhatim-Fstari)
Out[18]: 0.11233110037923455
```

We can plot the empirical cdf along with the normal one:

```
In [19]: x = np.linspace(-10,10,1000)

In [20]: Fhat = np.mean(xi[None,:] <= x[:,None],axis
   =-1)

In [21]: Fstar = stardist.cdf(x)

In [22]: figure();

In [23]: plot(x,Fhat,'b-',lw=2,label=r'$\hat{F}(x,\{
   x_i\})$');

In [24]: plot(x,Fstar,'r--',lw=2,label=r'$F^*(x)$');

In [25]: ip = np.argmax(Fstar-Fhat)

In [26]: im = np.argmax(Fhat-Fstar)

In [27]: annotate('',xy=(x[ip],Fstar[ip]),xycoords='
   data',xytext=(x[ip],Fhat[ip]),textcoords='data',
   arrowprops=dict(arrowstyle='<->'));

In [28]: annotate('',xy=(x[im],Fstar[im]),xycoords='
   data',xytext=(x[im],Fhat[im]),textcoords='data',
   arrowprops=dict(arrowstyle='<->'));

In [29]: xlabel(r'$x$');

In [30]: ylabel('cdf');

In [31]: legend(loc='upper left');
```

```
In [32]: grid(True);

In [33]: savefig('notes06_lilliefors.eps',
    bbox_inches='tight')
```

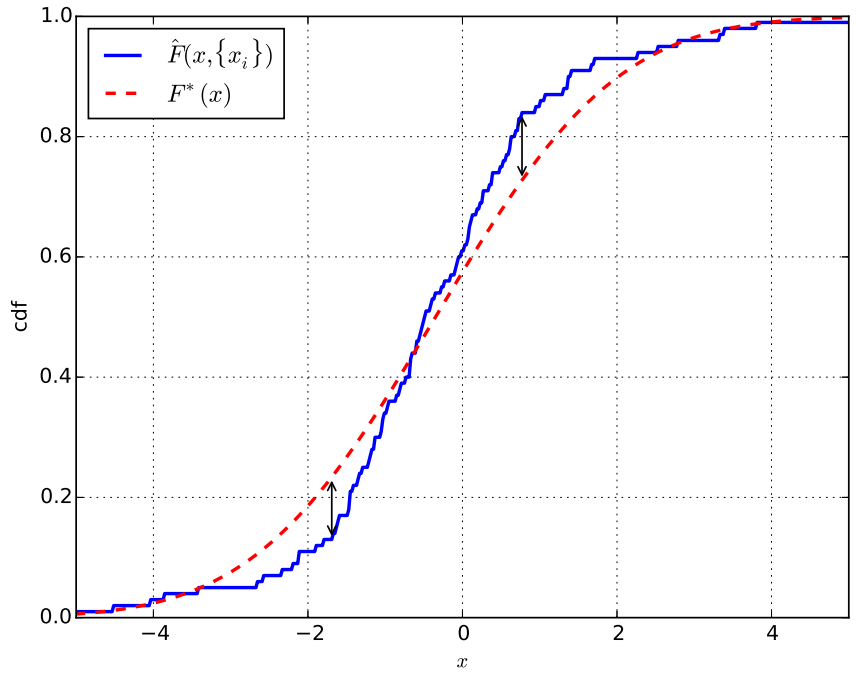



Since we estimated the normal parameters rather than specifying them, the null distribution of the statistic will not be the Kolmogorov distribution. In fact there is no closed-form expression for the "Lilliefors distribution", but we can estimate it with a Monte Carlo, generating a large number of normal samples of size $n$, scaling each by its sample mean and sample variance, and constructing the Kolmogorov statistic with the standard normal as the reference distribution:

```
In [34]: np.random.seed(20181030)

In [35]: Nmonte = 10**5

In [36]: x_Ii = stats.norm.rvs(size=(Nmonte,n))

In [37]: x_Ii.sort(axis=-1)

In [38]: xbar_I = np.mean(x_Ii,axis=-1)

In [39]: s_I = np.std(x_Ii,axis=-1,ddof=1)

In [40]: Fstar_Ii = stats.norm.cdf((x_Ii-xbar_I[:,
    None])/s_I[:,None])

In [41]: Tp_I = np.max(Fstar_Ii-Fhatip,axis=-1)

In [42]: Tm_I = np.max(Fhatim-Fstar_Ii,axis=-1)

In [43]: print(np.mean(np.maximum(Tp_I,Tm_I)>max(Tp,
    Tm)))
0.00342
```

We numerically estimate a $p$-value of 0.00342, since 342 of the 100,000 samples had a higher Kolmogorov statistic. Note that the Monte Carlo uncertainty on that number is about $\sqrt{342} \approx 18.5$, which means the $p$-value is definitely between 0.003 and 0.004. We can get a closer look at what's going on by zooming in a bit:

```
In [44]: xlim(-5,5);

In [45]: savefig('notes06_lilliefors_zoom.eps',
    bbox_inches='tight');
```

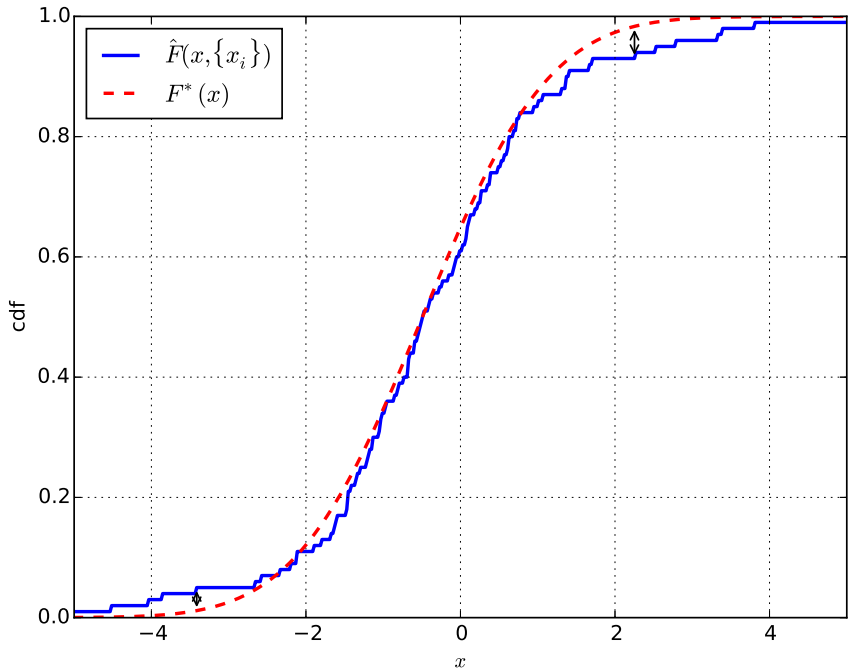



So we see that, although the most egregious feature of the distribution is the large outliers, they are not directly causing the unlikely Lilliefors statistic value. Instead, it's a mismatch in the middle of the distribution. In fact, this is sort of caused by the outliers, since they have a big impact on the estimate of the sample mean and variance used to estimate the parameters of the normal distribution. Note that these parameters do not produce the lowest possible Kolmogorov statistic. We can, for example, get a "better fit" by using the sample median and interquartile spread to estimate the parameters:

```
In [46]: altloc = np.median(xi); altloc
Out[46]: -0.49208918147769198

In [47]: altscale = (np.percentile(xi,75) - np.
   percentile(xi,25))/(stats.norm.ppf(.75) - stats.
   norm.ppf(.25)); altscale
Out[47]: 1.2881030571979752

In [48]: altdist = stats.norm(loc=altloc,scale=
   altscale)

In [49]: Falti = altdist.cdf(xi)

In [50]: altTp = max(Falti-Fhatip); altTp
Out[50]: 0.053847080513066636

In [51]: altTm = max(Fhatim-Falti); altTm
Out[51]: 0.038588594271776813

In [52]: Falt = altdist.cdf(x)

In [53]: figure();

In [54]: plot(x,Fhat,'b-',lw=2,label=r'$\hat{F}(x,\{
   x_i\})$');

In [55]: plot(x,Falt,'r--',lw=2,label=r'$F^*(x)$');

In [56]: altip = np.argmax(Falt-Fhat)

In [57]: altim = np.argmax(Fhat-Falt)

In [58]: annotate('',xy=(x[altip],Falt[altip]),
   xycoords='data',xytext=(x[altip],Fhat[altip]),
   textcoords='data',arrowprops=dict(arrowstyle
   ='<->'));
```

```
In [59]: annotate('',xy=(x[altim],Falt[altim]),
   xycoords='data',xytext=(x[altim],Fhat[altim]),
   textcoords='data',arrowprops=dict(arrowstyle
   ='<->'));

In [60]: xlabel(r'$x$');

In [61]: ylabel('cdf');

In [62]: legend(loc='upper left');

In [63]: grid(True)

In [64]: xlim(-5,5)
Out[64]: (-5, 5)

In [65]: savefig('notes06_altlilliefors_zoom.eps',
   bbox_inches='tight')

In [66]: print(np.mean(np.maximum(Tp_I,Tm_I)>max(
   altTp,altTm)))
0.67988
```

Of course, the fact that the Lilliefors test uses the sample mean and sample median allows it to be influenced by large outliers which would otherwise not have much impact on the Kolmogorov statistic.




## 2.2 Interlude: The Cramér-von Mises and Anderson-Darling Tests

We've seen a drawback to Kolmogorov-Smirnov type tests, which we can consider in two parts:

1. Using the maximum separation between the empirical and hypothesized cdfs bases the results on what's happening on one point rather than overall, and
2. The "worst" mismatch will generally not be in the tails, since the hypothetical cdf is close to 0 or 1 there, and $\left|F^*(x) - \hat{F}(x;\{x_i\})\right|$ is not significantly different if $F^*(x) = 0.01$ or $10^{-6}$.

We can address the first of these by considering some measure like the total area between the cdf curves

$$\int_{-\infty}^{\infty} \left| \hat{F}(x;\{x_i\}) - F^*(x) \right| \, dx \tag{2.4}$$

or a sort of quadratic distance

$$\int_{-\infty}^{\infty} [\hat{F}(x;\{x_i\}) - F^*(x)]^2 \, dx \ . \tag{2.5}$$

One problem with these possible statistics is that they'd change if we made some transformation of the data, like $Y = X^3$, while the cdfs themselves would not, since $P(X^3 \le x^3) = P(X \le x)$. So the natural thing to do is actually to use the hypothesized probability *density* $f^*(x) = F^{*\prime}(x)$ as a "measure" for the integral:

$$\begin{aligned} \frac{T\text{CvM}}{n} &= \int_{-\infty}^{\infty} [\hat{F}(x;\{x_i\}) - F^*(x)]^2 \, f^*(x) \, dx \\ &= \int_0^1 [\hat{F}(F^{*-1}(u);\{x_i\}) - u]^2 \, du \end{aligned} \tag{2.6}$$

where $F^{*-1}(u)$ is the inverse of the hypothesized cdf. This is known as the Cramér-von Mises statistic.

You might think this is a lot harder to evaluate than the Kolmogorov statistic, since you have to integrate over all $x$ (or equivalently all $u$), but you can break up the integral into $n+1$ pieces divided by the actual $x_i$ values. If $x^{(i)}$ denotes the $i$th order statistic of the data (the $i$th value in the list when the data are sorted), then,

- For $x < x^{(1)}$, $\hat{F}(x;\{x_i\}) = 0$ and the integrand is $[F^*(x)]^2$, which makes the integral

$$\int_0^{F^*(x^{(0)})} u^2 \, du = \frac{[F^*(x^{(0)})]^3}{3} \tag{2.7}$$

- For $x^{(i)} < x < x^{(i+1)}$, $i = 1, \ldots n-1$, $\hat{F}(x;\{x_i\}) = \frac{i}{n}$ and the integrand is $[F^*(x) - i/n]^2$, which makes the integral

$$\int_{F^*(x^{(i)})}^{F^*(x^{(i+1)})} (u - i/n)^2 \, du = \frac{[F^*(x^{(i+1)}) - i/n]^3}{3} - \frac{[F^*(x^{(i)}) - i/n]^3}{3} \tag{2.8}$$

- for $x^{(n)} < x$, $\hat{F}(x;\{x_i\}) = 1$ and the integrand is $[F^*(x) - 1]^2$, which makes the integral

$$\int_{F^*(x^{(n)})}^{1} (u-1)^2 \, du = \frac{[F^*(x^{(n)}) - 1]^3}{3} \tag{2.9}$$

Combining these terms with a little bit of algebra gives the expression

$$T_{\text{CvM}} = \frac{1}{12n} + \sum_{i=1}^{n} \left( \frac{2i-1}{2n} - F^*(x^{(i)}) \right)^2 \tag{2.10}$$

This is still not very much impacted by what's happening on the tails, though. To probe that, we modify the statistic further, using the fact that the squared standard error of $\hat{F}(x;\{X_i\})$ as an estimator of $F(x)$ is

$$\text{Var}(\hat{F}(x;\{X_i\})) = \frac{1}{n} F(x)[1 - F(x)] \tag{2.11}$$

We thus define the *Anderson-Darling* statistic as

$$A^2 = n \int_{-\infty}^{\infty} \frac{[\hat{F}(x;\{x_i\}) - F^*(x)]^2}{F^*(x)[1 - F^*(x)]} \, f^*(x) \, dx \tag{2.12}$$

Note that this has a chi-squared-ish construction as a sum of $\frac{(X-\mu)^2}{\sigma^2}$. (Although not exactly, since the different $x$ values don't represent independent points.) The piecewise integration is a little more complicated, but the result is

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left( \ln[F^*(x^{(i)})] + \ln[1 - F^*(x^{(n+1-i)})] \right) \tag{2.13}$$

The exact critical values for $A^2$ can be evaluated with a Monte Carlo, but as an approximation[1], in the case where the target distribution is normal with parameters estimated using the sample mean and variance, if we write

$$(A^*)^2 = A^2\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right) \tag{2.14}$$

some key critical values are

$$P([A^*]^2 \geq 0.631) \approx 0.10 \tag{2.15a}$$
$$P([A^*]^2 \geq 0.752) \approx 0.05 \tag{2.15b}$$
$$P([A^*]^2 \geq 0.873) \approx 0.025 \tag{2.15c}$$
$$P([A^*]^2 \geq 1.035) \approx 0.01 \tag{2.15d}$$
$$P([A^*]^2 \geq 1.159) \approx 0.005 \tag{2.15e}$$

**Thursday 1 November 2018**
**– Read Section 6.3 of Conover**

## 2.3 Consistency Tests for The Exponential Distribution

Another family of distributions for which it's particularly interesting to perform a consistency test with unknown parameters is the exponential distribution. That's because it arises naturally in the description of a Poisson process: if we are noting the arrival times of independent events with an average rate of $1/\tau$, the number of events in some time interval of duration $T$ will be Poisson-distributed with mean $T/\tau$, but the waiting time from one event to the next (or from an arbitrary start time to the first event) will be exponentially distributed, with a cumulative distribution function

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ 1 - e^{-x/\tau} & 0 \leq x < \infty \end{cases} \tag{2.16}$$

and an expectation value $E(X) = \tau$. If we have a set of waiting times $\{x_i\}$ and wish to evaluate the hypothesis that they come from an exponential distribution with unspecified rate, it's reasonable to use the estimated rate $\hat{\tau} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and compare the empirical distribution $\hat{F}(x)$ to the hypothesized distribution

$$F^*(x) = \begin{cases} 0 & -\infty < x < 0 \\ 1 - e^{-x/\overline{x}} & 0 \leq x < \infty \end{cases} \tag{2.17}$$

We define the test statistics as usual as

$$T^+ = \sup_x(F^*(x) - \hat{F}(x)) = \max_i\left(1 - e^{-x^{(i)}/\overline{x}} - \frac{i-1}{n}\right) \tag{2.18a}$$

$$T^- = \sup_x(\hat{F}(x) - F^*(x)) = \max_i\left(\frac{i}{n} - 1 + e^{-x^{(i)}/\overline{x}}\right) \tag{2.18b}$$

$$T = \sup_x\left|\hat{F}(x) - F^*(x)\right| = \max(T^+, T^-) \tag{2.18c}$$

# 3 Two-Sample Consistency Tests

Finally, we consider the case where we have two samples $\{x_i | i = 1, \ldots, n\}$ and $\{y_j | j = 1, \ldots, m\}$, and wish to ask whether they come from the same distribution. We've approached this question as a comparison of location parameters with the Wilcoxon-Mann-Whitney rank sum test and as a comparison of scale parameters with the Conover squared ranks test. But if we wish

[1] citation to be added

to make a general comparison, one option is to compare the empirical distributions

$$\hat{F}_x(x;\{x_i\}) = \frac{1}{n}\sum_{i=1}^{n} I[x_i \le x] = \begin{cases} 0 & x < x^{(1)} \\ \frac{i}{n} & x^{(i)} \le x < x^{(i+1)} \\ 1 & x^{(n)} \le x \end{cases} \tag{3.1a}$$

$$\hat{F}_y(x;\{y_j\}) = \frac{1}{m}\sum_{j=1}^{m} I[y_j \le x] = \begin{cases} 0 & x < y^{(1)} \\ \frac{j}{m} & y^{(j)} \le x < y^{(j+1)} \\ 1 & y^{(m)} \le x \end{cases} \tag{3.1b}$$

where $\{x^{(i)}\}$ and $\{y^{(j)}\}$ are the order statistics of the two samples.

## 3.1 Two-Sample Kolmogorov-Smirnov

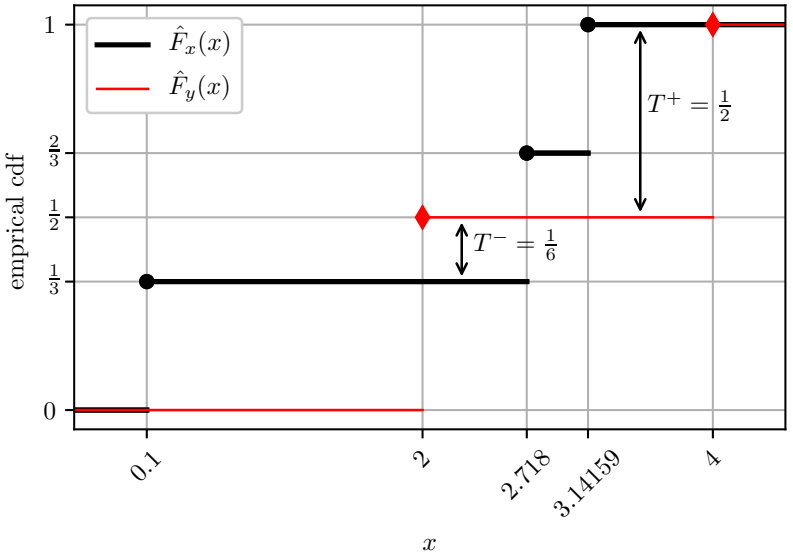
To perform a Kolmogorov-Smirnov test with the two empirical cdfs, we define the statistics

$$T^+ = \sup\left(\hat{F}_x(x) - \hat{F}_y(x)\right) \tag{3.2a}$$

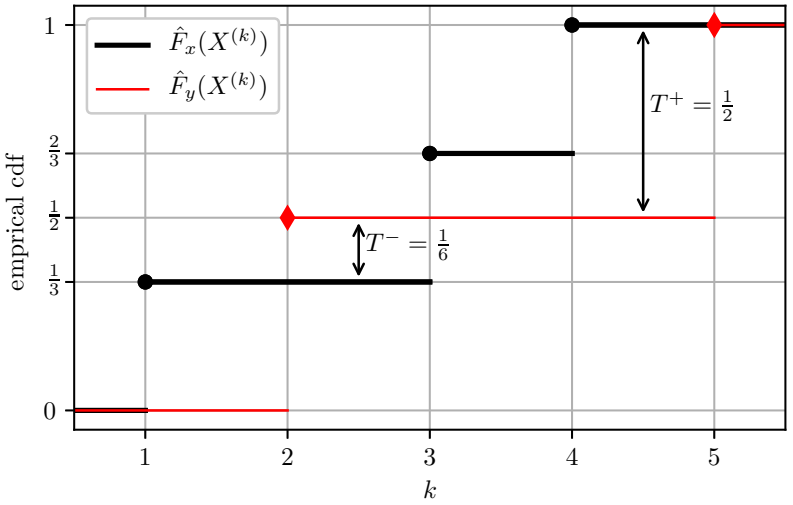
$$T^- = \sup\left(\hat{F}_y(x) - \hat{F}_x(x)\right) \tag{3.2b}$$

$$T = \sup\left|\hat{F}_x(x) - \hat{F}_y(x)\right| = \max(T^+, T^-) \tag{3.2c}$$

To see how this plays out, we consider the very simple case where $\{x_i\} = \{0.1, 2.718, 3.14159\}$ and $\{y_j\} = \{4, 2\}$, and construct the empirical cdfs. We see that $T^+ = 1 - \frac{1}{2} = \frac{1}{2}$, $T^- = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$ and thus $T = \frac{1}{2}$:




Note that nothing about this construction used the numerical values of the $\{x_i\}$ and $\{y_j\}$, just their ordering, so we could also use the ranks of the data rather than the data themselves:



This allows us to construct the null distribution by enumerating the $\binom{n+m}{n}$ possible divisions of the ranks 1 to $n+m$ into $x$-ranks and $y$-ranks, and computing the Kolmogorov-Smirnov statistics for each of them. In this case there are $\binom{5}{3} = 10$ possibilities. Conover goes through them and shows that 1 has $T = \frac{1}{3}$, 3 have $T = \frac{1}{2}$, 4 have $T = \frac{2}{3}$, and 2 have $T = 1$. So the $p$-value for the data we consider is $\frac{4+2}{10} = 0.60$. (We can also look at Table A20 in Conover and see that the threshold to reject a two-tailed hypothesis at the $\alpha = 0.20$ level is $T > \frac{2}{3}$.)

## 3.2 Two-Sample Cramér-von Mises and Anderson-Darling

Recall that the one-sample Cramér-von Mises test compared the empirical distribution $\hat{F}(x)$ to the hypothesized distribution $F^*(x)$ using a statistic constructed from the integral

$$\int_{-\infty}^{\infty} [\hat{F}(x) - F^*(x)]^2 \, f^*(x) \, dx \tag{3.3}$$

We'd like to use the same method in the two-sample case to compare $\hat{F}_x(x)$ and $\hat{F}_y(x)$. There are a couple of challenges related to the "measure" $f^*(x)\,dx = dF^*(x)$. First, if we're treating the two samples the same, which distribution should in place of the hypothesized distribution. One obvious choice is to combine the data sets $\{x_i | i = 1, \ldots, n\}$ and $\{y_j | i = j, \ldots, m\}$ into $\{X_k | k = 1, \ldots, n+m\}$, and use this combined sample to produce an empirical distribution function

$$\begin{aligned} \hat{F}_{x,y}(x; \{X_k\}) &= \frac{1}{n+m} \left( \sum_{i=1}^{n} I[x_i \le x] + \sum_{j=1}^{m} I[y_j \le x] \right) \\ &= \begin{cases} 0 & x < X^{(1)} \\ \frac{k}{n+m} & X^{(k)} \le x < X^{(k+1)} \\ 1 & X^{(n+m)} \le x \end{cases} \end{aligned} \tag{3.4}$$

and then use the definition

$$\int_{-\infty}^{\infty} [\hat{F}_x(x) - \hat{F}_y(x)]^2 \, \hat{f}_{x,y}(x) \, dx \tag{3.5}$$

This brings us to another problem, though. The empirical cdf $\hat{F}_{x,y}(x)$ has discrete jumps, so we can't just take its derivative to calculate a density $\hat{f}_{x,y}(x)$.[2] But if we go back to (3.3) and consider it to be an expectation value constructed using the hypothesized distribution

$$E_{F^*}([\hat{F}(X; \{x_i\}) - F^*(X)]^2) \tag{3.6}$$

we can see a way forward. The empirical distribution $F_{x,y}$ is basically a discrete distribution with a pmf which is non-zero at the values present in the combined sample: $p_{x,y}(X^{(k)}) = \frac{1}{n+m}$ (assuming there are no repeated values in the combined sample). So we can replace the integral $\int_{-\infty}^{\infty}(\cdots)\hat{f}_{x,y}(x)dx$ with the sum $\sum_x(\cdots)p_{x,y}(x)$ and write a statistic based on a discrete expectation value:

$$E_{\hat{F}_{x,y}}([\hat{F}_x(x) - \hat{F}_y(x)]^2) = \sum_{k=1}^{n+m} \frac{[\hat{F}_x(X^{(k)}) - \hat{F}_y(X^{(k)})]^2}{n+m} \tag{3.7}$$

If we construct this sum explicitly for the data set considered above, we get

$$\begin{aligned} &\frac{1}{5}\left[\left(\frac{1}{3} - 0\right)^2 + \left(\frac{1}{3} - \frac{1}{2}\right)^2 + \left(\frac{2}{3} - \frac{1}{2}\right)^2 + \left(1 - \frac{1}{2}\right)^2 + (1-1)^2\right] \\ &\quad = \frac{1}{5}\left[\frac{1}{9} + \frac{1}{36}\frac{1}{36} + \frac{1}{4}\right] = \frac{4+1+1+9}{180} = \frac{15}{180} = \frac{1}{12} \end{aligned} \tag{3.8}$$

[2] Actually, we could carry out a construction based on the Dirac delta function, but that would involve introducing additional mathematical apparatus and possibly induce indigestion in mathematical purists.

More than the numerical value of the sum, it's of interest that the last value is zero. This will always be the case, since it's guaranteed that $F_x(X^{(n+m)}) = 1 = F_x(X^{(n+m)})$ so we can leave that term out of the sum,[3] and write the two-sample Cramér-von Mises statistic (putting an extra $\frac{nm}{n+m}$ out front for convention) as

$$T_2 = \frac{nm}{(n+m)^2} \sum_{k=1}^{n+m-1} [\hat{F}_x(X^{(k)}) - \hat{F}_y(X^{(k)})]^2 \tag{3.9}$$

For the data set in question,

$$T_2 = \frac{3 \times 2}{5} \frac{1}{12} = \frac{1}{10} \tag{3.10}$$

Having laid the groundwork with the Cramér-von Mises statistic, it's straightforward to convert this to an Anderson-Darling statistic, by including $\hat{F}_{xy}(x)[1 - \hat{F}_{xy}(x)]$ in the denominator. Conceptually, this normalizes the squared deviation by the expected variance; practically, it increases the importance of outliers. The statistic is then

$$\begin{aligned} A^2 &= \frac{nm}{(n+m)^2} \sum_{k=1}^{n+m-1} \frac{[\hat{F}_x(X^{(k)}) - \hat{F}_y(X^{(k)})]^2}{\hat{F}_{xy}(X^{(k)})[1 - \hat{F}_{xy}(X^{(k)})]} \\ &= nm \sum_{k=1}^{n+m-1} \frac{[\hat{F}_x(X^{(k)}) - \hat{F}_y(X^{(k)})]^2}{k(n+m-k)} \end{aligned} \tag{3.11}$$

Note that dropping the $k = n+m$ term from the sum has avoided giving us a $0/0$ contribution.

[3] This clears up an apparent asymmetry where the statistic is defined by contributions from points where one of the empirical cdfs is discontinuous, and always uses the value to the right of the jump (since the cdf is by definition right-continuous). The fact that the contribution from the value after the last jump is zero makes up for the lack of a contribution from the value before the first jump.

Obviously, the two-sample Cramér-von Mises and Anderson-Darling statistics can be constructed from the ranks as well, so the associated $p$-values can be calculated as with the Kolmogorov-Smirnov test. Note that they are explicitly one-tailed, though, since the deviation has been squared.