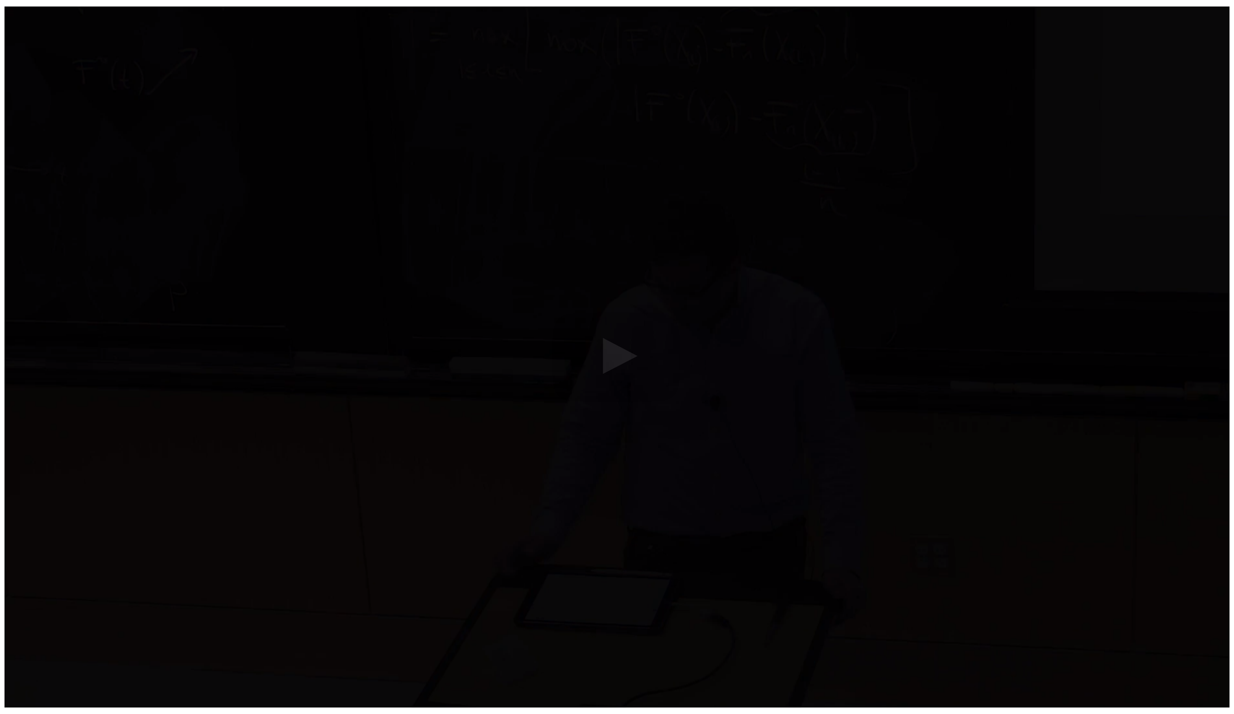
7. Kolmogorov-Smirnov Test: Computational Issues

Exercises due Aug 3, 2021 19:59 EDT

Kolmogorov-Smirnov Test: Computational Issues



8:50 / 8:50 1.50x

bridge has a PDF,

and this PDF I can simulate because I

can do it once and for all.

I can just turn on my computer, and prepare the critical values

for this supremum of the absolute value of a Brownian

bridge, once and for all, and just

disseminate it to the world because it's always

going to be the same PDF, and the same critical values.

So this is something I don't need to redo

every time I have new data.

I can print it to the back of the book.

It will always be the same two Q alphas.

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Let $X_1, \ldots, X_n$ be i.i.d. random variables with unknown cdf F. Our goal is to test the hypotheses:

$$\begin{aligned} H_0 &: F = F^0 \\ H_1 &: F \neq F^0. \end{aligned}$$

The **Kolmogorov-Smirnov test statistic** is defined as

$$T_n = \sup_{t \in \mathbb{R}} \sqrt{n} \left| F_n(t) - F^0(t) \right|$$

and the **Kolmogorov-Smirnov test** is

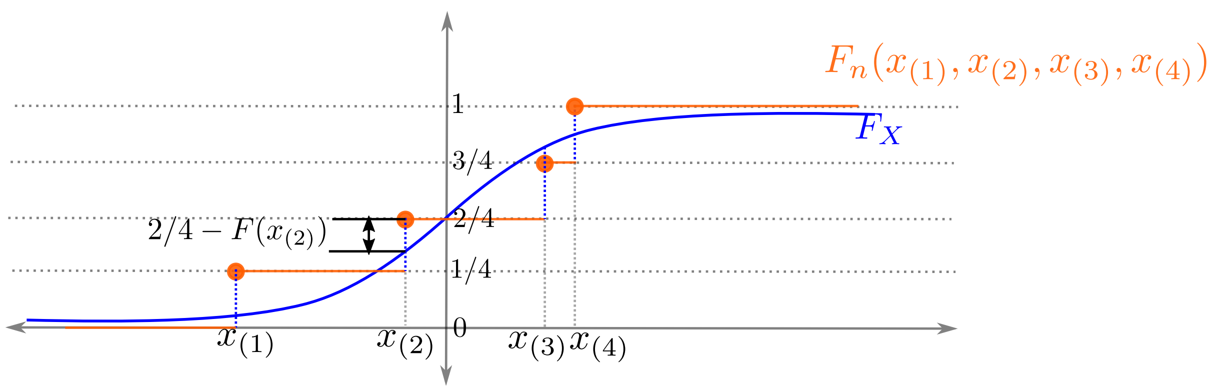
$$\mathbf{1}(T_n > q_\alpha) \qquad \text{where } q_\alpha = q_\alpha\left(\sup_{t \in [0,1]} |\mathbb{B}(t)| \right).$$

Here, $q_\alpha = q_\alpha\left(\sup_{t\in[0,1]} |\mathbb{B}(t)|\right)$ is the $(1-\alpha)$-quantile of the supremum $\sup_{t\in[0,1]} |\mathbb{B}(t)|$ of the Brownian bridge as in Donsker's Theorem.

Even though the K-S test statistics T_n is defined as a supremum over the entire real line, it can be computed explicitly as follows:

$$\begin{aligned} T_n &= \sqrt{n} \sup_{t \in \mathbb{R}} \left| F_n(t) - F^0(t) \right| \\ &= \sqrt{n} \max_{i=1,\ldots,n} \left\{ \max\left(\left| \frac{i-1}{n} - F^0\left(X_{(i)}\right) \right|, \left| \frac{i}{n} - F^0\left(X_{(i)}\right) \right| \right) \right\} \end{aligned}$$

where $X_{(i)}$ is the **order statistic** , and represents the $i^{(th)}$ smallest value of the sample. For example, $X_{(1)}$ is the smallest and $X_{(n)}$ is the greatest of a sample of size n.



An example of the empirical cdf $F_n\left(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}\right)$ for a specific data set $x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}$ of sample size 4, and the cdf $F_X(x)$ under the null hypothesis.

We see that because $F^0(t)$ is increasing, and $F_n(t)$ is piecewise constant, $\left|F_n(t) - F^0(t)\right|$ can only possibly achieve its maximum at $t = x_{(i)}$.

Concept Check: Kolmogorov-Smirnov Test Statistic

1 point possible (graded)

As above, let $X_1, \ldots, X_n$ be iid random variables with unknown cdf F. To decide between the null hypothesis, $H_0 : F = \Phi$, and the alternative hypothesis, $H_1 : F \neq \Phi$, stated in the previous problem, we consider the Kolmogorov-Smirnov test statistic for this hypothesis

$$T_n = \sup_{t \in \mathbb{R}} \sqrt{n} \left|F_n(t) - \Phi(t)\right|.$$

Which of the following are true statements regarding the test statistic T_n? (Choose all that apply.)

- [ ] Under H_0, T_n converges in distribution to a Brownian motion.
- [ ] T_n converges to a pivotal distribution under H_0.
- [ ] If H_0 holds, then T_n converges to a distribution whose quantiles we can either look up in tables or estimate very well using simulations.
- [ ] Given a sample of size $n = 1000$, the value of the test-statistic T_n cannot be computed efficiently.

Submit

You have used 0 of 2 attempts

Practice: Compute the Kolmogorov-Smirnov Test Statistic

1 point possible (graded)

Let $X_1, \ldots, X_n$ be iid samples with cdf F, and let F^0 denote the cdf of $\mathsf{Unif}(0,1)$. Recall that

$$F^0(t) = t \cdot \mathbf{1}(t \in [0,1]) + 1 \cdot \mathbf{1}(t > 1).$$

We want to use goodness of fit testing to determine whether or not $X_1, \ldots, X_n \overset{iid}{\sim} \mathsf{Unif}(0,1)$. To do so, we will test between the hypotheses

$$\begin{aligned} H_0 &: F(t) = F^0 \\ H_1 &: F(t) \neq F^0. \end{aligned}$$

To make computation of the test statistic easier, let us first reorder the samples from smallest to largest, so that

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

is the reordered sample. In this set-up, the Kolmogorov-Smirnov test statistic is given by the formula

$$T_n = \sqrt{n} \max_{i=1,\ldots,n} \left\{ \max \left(\left| \frac{i-1}{n} - X_{(i)} \mathbf{1}\left(X_{(i)} \in [0,1]\right) \right|, \left| \frac{i}{n} - X_{(i)} \mathbf{1}\left(X_{(i)} \in [0,1]\right) \right| \right) \right\}.$$

You observe the data set $\mathbf{x}$ consisting of 5 samples:

$$\mathbf{x} = 0.8, 0.7, 0.4, 0.7, 0.2$$

Using the formula above, what is the value of T_5 for this data set? (You are encouraged to use computational tools.)

Submit You have used 0 of 3 attempts

Discussion

Hide Discussion

Topic: Unit 4 Hypothesis testing:Lecture 16: Goodness of Fit Tests Continued: Kolmogorov-Smirnov test, Kolmogorov-Lilliefors test, Quantile-Quantile Plots / 7. Kolmogorov-Smirnov Test: Computational Issues

Add a Post

Show all posts | by recent activity

- "pivotal statistic" vs "pivotal distribution" 1
- Understanding the formula for the test statistic 5
- ? L16 Q7(a) - Concept Check: Kolmogorov-Smirnov Test Statistic 5
  Community TA
- ? Concept Check: Kolmogorov-Smirnov Test Statistic 7
  In the 3rd bullet of the answer provided for this question, I think there might be an incorrect word, given that the first bullet explained that the p...
- Compute the Kolmogorov-Smirnov Test Statistic. R ks.test() warning 6
  After computing the value of statistic "manually" (not exactly, delegating calculations to R of course) I decided to use available ks.test() to comp...
- ? Regarding null 4
  Why is it that we pose the null hypothesis in a way that it is to be rejected?

edX for Business
Open edX
Careers
News

Legal

Terms of Service & Honor Code
Privacy Policy
Accessibility Policy
Trademark Policy
Sitemap

Connect

Blog
Contact Us
Help Center
Media Kit
Donate