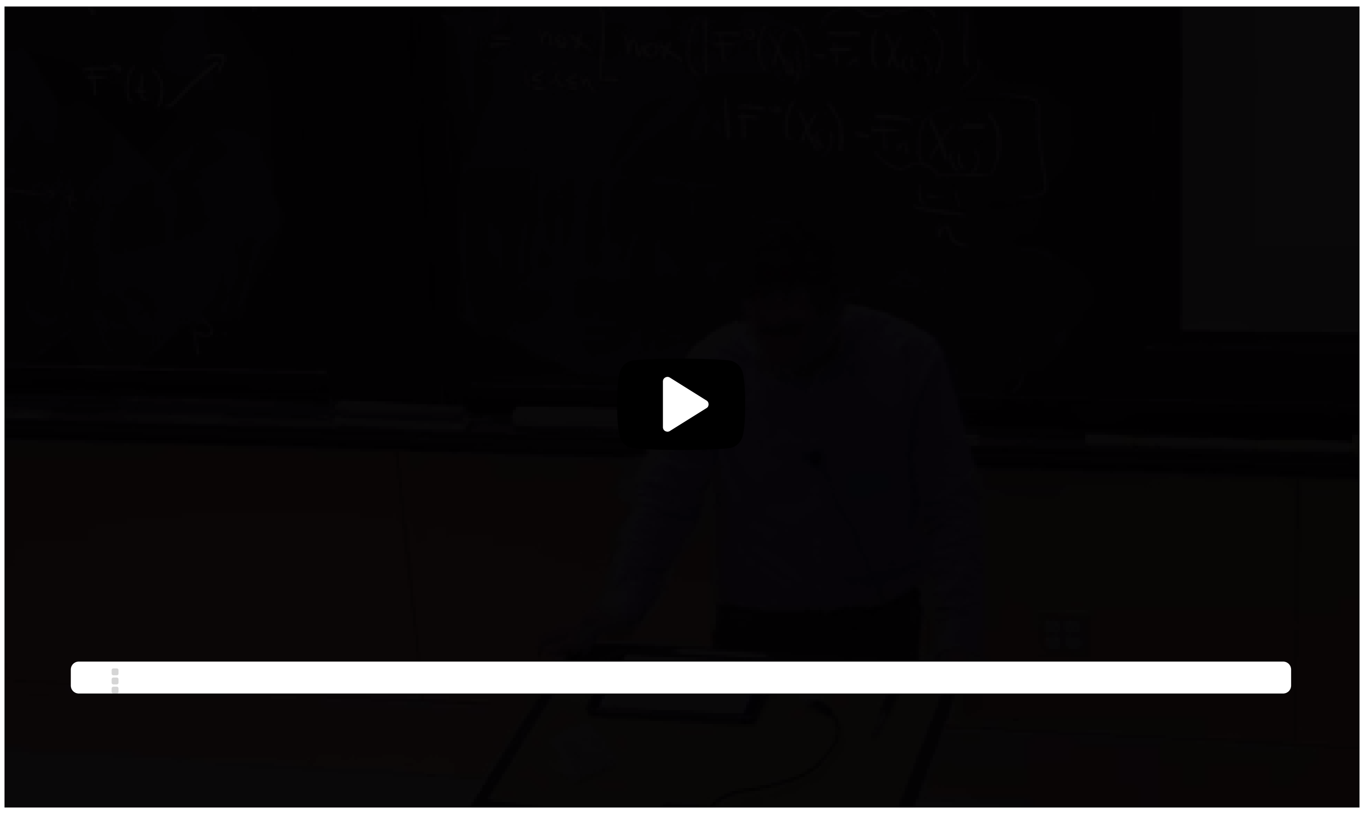
7. Kolmogorov-Smirnov Test: Computational Issues

Kolmogorov-Smirnov Test: Computational Issues



8:50 / 8:50 1.0x

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.

End of transcript. Skip to the start.

Video
Download video file

Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

xuetangX.com 学堂在线

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

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

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

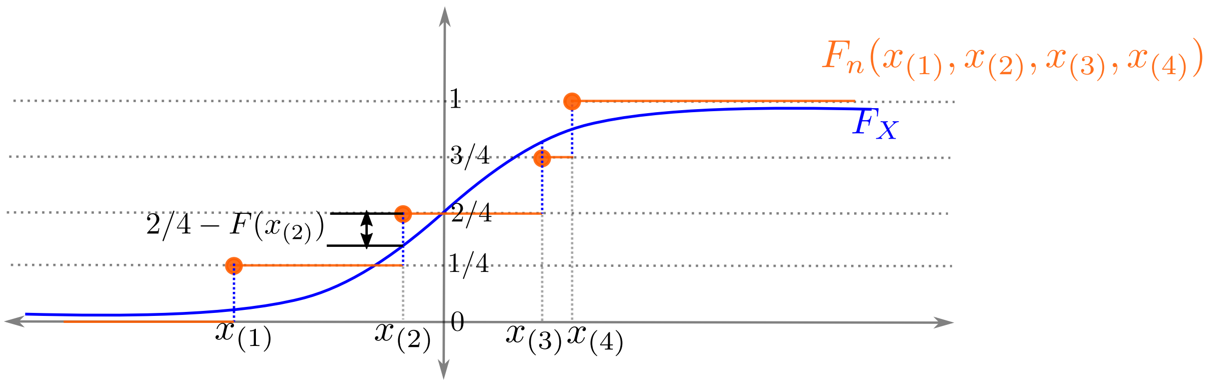
$$\mathbf{1}(T_n > q_\alpha) \qquad \text{where } q_\alpha = q_\alpha \left(\sup_{t \in [0,1]} |\mathbb{B}(t)| \right).$$

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:

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

where $X_{(i)}$ is the **ordered 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 increaseing, 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

0/1 point (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.)

- [x] T_n converges in distribution to a Brownian motion.
- [x] T_n converges to a pivotal distribution under H_0. ✔
- [x] 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.

Solution:

We examine the choices in order.

- The first choice is incorrect. If H_0 holds, then T_n converges to the **supremum** of a Brownian bridge. A Brownian motion is a **random curve** while its supremum over the interval $[0,1]$ is a **random variable**. Since T_n is also a random variable, it cannot converge to a random curve.

- The second choice is correct. Independent of the distribution of $X_1,\ldots,X_n$, we have that

$$\sqrt{n}\sup_{t\in\mathbb{R}}\left|F_n(t)-F(t)\right| \xrightarrow[n\to\infty]{(d)} \sup_{x\in[0,1]}\left|\mathbb{B}(x)\right|.$$

  That is, the limiting distribution is **independent** of the distribution of the $X_1,\ldots,X_n$ (as long as F is continuous). By definition, T_n is a pivotal statistic under H_0.

- The third choice is correct. In general, pivotal distributions can be understood by consulting a table of quantiles. Using computational tools, Brownian motions (and their suprema) can be simulated, so this is another approach to computing the quantiles.

- The fourth choice is incorrect. If the sample size is n, then the formula on the slide "Kolmogorov-Smirnov test (3)" provides a formula that involves computing $2n$ maxima. This is certainly doable if $n = 1000$.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Practice: Compute the Kolmogorov-Smirnov Test Statistic

1/1 point (graded)

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

0.6708203932499368 ✔ **Answer:** 0.6708

Solution:

First we reorder the given data set to get

$$0.2, 0.4, 0.7, 0.7, 0.8.$$

Now $X_{(i)}$ is defined to be the i-th element in the list above. Our goal is to compute

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

for $n = 5$ and plugging in the above reordered data set. We first need to compute the maximum of the following list of numbers:

$$\max(|0-0.2|, |0.2-0.2|) = 0.2$$
$$\max(|0.2-0.4|, |0.4-0.4|) = 0.2$$
$$\max(|0.4-0.7|, |0.8-0.7|) = 0.3$$
$$\max(|0.4-0.7|, |0.8-0.7|) = 0.3$$
$$\max(|0.8-0.8|, |1-0.8|)) = 0.2$$

which comes out to be 0.3. Therefore, $T_5 = \sqrt{5} \cdot 0.3 \approx 0.6708$.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Discussion

Show 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

© All Rights Reserved