2. Introduction to Goodness of Fit Tests

Recap of Parametric Hypothesis Testing: The Uniform Statistical Model

0/1 point (graded)

Let X be a uniform random variable with the distribution $\mathrm{Unif}\,[0, \theta^*]$.

We would like to test whether $H_0 : \theta^* = 2$ or $H_1 : \theta^* \neq 2$, with $\Theta = (0, \infty)$.

Let $X_1, \ldots, X_n$ be iid samples of X.

- Let $\overline{X_n}$ denote sample mean.
- Let $\widetilde{S}_n$ denote the unbiased sample variance of $X_1, \ldots, X_n$.
- Let $\widehat{\theta_n}^{\mathrm{MLE}}$ denote the maximum likelihood estimator of θ.
- Let $\ell_n\left(\widehat{\theta_n}^{\mathrm{MLE}}\right)$ denote the log-likelihood of n samples evaluated at the maximum likelihood estimator and $\ell_n\,(2)$ denote the log-likelihood of n samples under H_0.

Select from the following the tests that are **technically correct (that is, can be applied under this scenario) and that have the required level** $\alpha \in [0, 1]$.

Note: By asymptotic level of α, we require the probability of type-1 error under H_0 be at most α as $n \to \infty$. By non-asymptotic level of α, we require the probability of type-1 error under H_0 be at most α for every n.

- [x] $\mathbf{1}\left\{\frac{|\overline{X}_n-1|}{\sqrt{\widetilde{S}_n/n}} > q_{\alpha/2}\right\}$ for non-asymptotic level α, where $q_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the Student's T distribution with $n-1$ degrees of freedom.

- [x] $\mathbf{1}\left\{\sqrt{n}\frac{|2\overline{X}_n-2|}{\sqrt{4/3}} > q_{\alpha/2}\right\}$ for asymptotic level α, where q_α is the $(1-\alpha)$-quantile of the standard normal random variable. ✔

- [ ] $\mathbf{1}\left\{\widehat{\theta_n}^{\text{MLE}} > 2 \text{ or } \widehat{\theta_n}^{\text{MLE}} \leq 1\right\}$ for asymptotic level α. ✔

- [x] $\mathbf{1}\left\{2\left(\ell_n\left(\widehat{\theta_n}^{\text{MLE}}\right) - \ell_n(2)\right) > q_\alpha\right\}$ for asymptotic level α, where q_α is the $(1-\alpha)$-quantile of χ^2_1.

✖

Solution:

Recall from Lecture 10 that the maximum likelihood estimator for the uniform statistical model is $\max_{i=1,\ldots,n} X_i$.

Technically incorrect: First, let us elaborate on the choices that are technically incorrect under this scenario: **Choices 1 and 4**. The first choice is attempting a Student's T test for non-asymptotic level α and this is technically incorrect because X is not a Gaussian random variable. Only if X is a Gaussian random variable will the test statistic (at least to our knowledge in this course) follow a Student's T distribution for a finite number of samples n.

Choice 4 is attempting a likelihood ratio test using the log-likelihoods evaluated at the maximum likelihood estimator and under H_0 and this choice is also technically incorrect because the MLE technical conditions are not satisfied for the uniform statistical model (recall the technical conditions for asymptotic normality of the MLE). Only if the MLE conditions are satsfied can this test be applied according to Wilk's theorem.

Technically correct and have an asymptotic level α: **Choices 2 and 3**. The third choice has an asymptotic level 0 because of the following:

$$\begin{aligned} P_{H_0}\left[\widehat{\theta_n}^{\text{MLE}} > 2 \text{ or } \widehat{\theta_n}^{\text{MLE}} \leq 1\right] &= P_{H_0}\left[\widehat{\theta_n}^{\text{MLE}} > 2\right] + P_{H_0}\left[\widehat{\theta_n}^{\text{MLE}} \leq 1\right] \\ &= P_{H_0}\left[\widehat{\theta_n}^{\text{MLE}} \leq 1\right] \end{aligned}$$

$$= P_{H_0}\left[\max_{i=1,\ldots,n} X_i \leq 1\right]$$
$$= \left(\frac{1}{2}\right)^n \to 0.$$

The second choice is a test that is both technically correct and has an asymptotic level α. This can be seen from the following:

- $2\overline{X_n}$ has an expectation equal to θ^*.
- The variance of $2\overline{X_n}$ under H_0 is

$$\begin{aligned}\mathsf{Var}_{H_0}\left(2\overline{X_n}\right) &= \frac{4}{n}\mathsf{Var}_{H_0}(X)\\ &= \frac{4}{n}\frac{4}{12}\\ &= \frac{4}{3n}.\end{aligned}$$

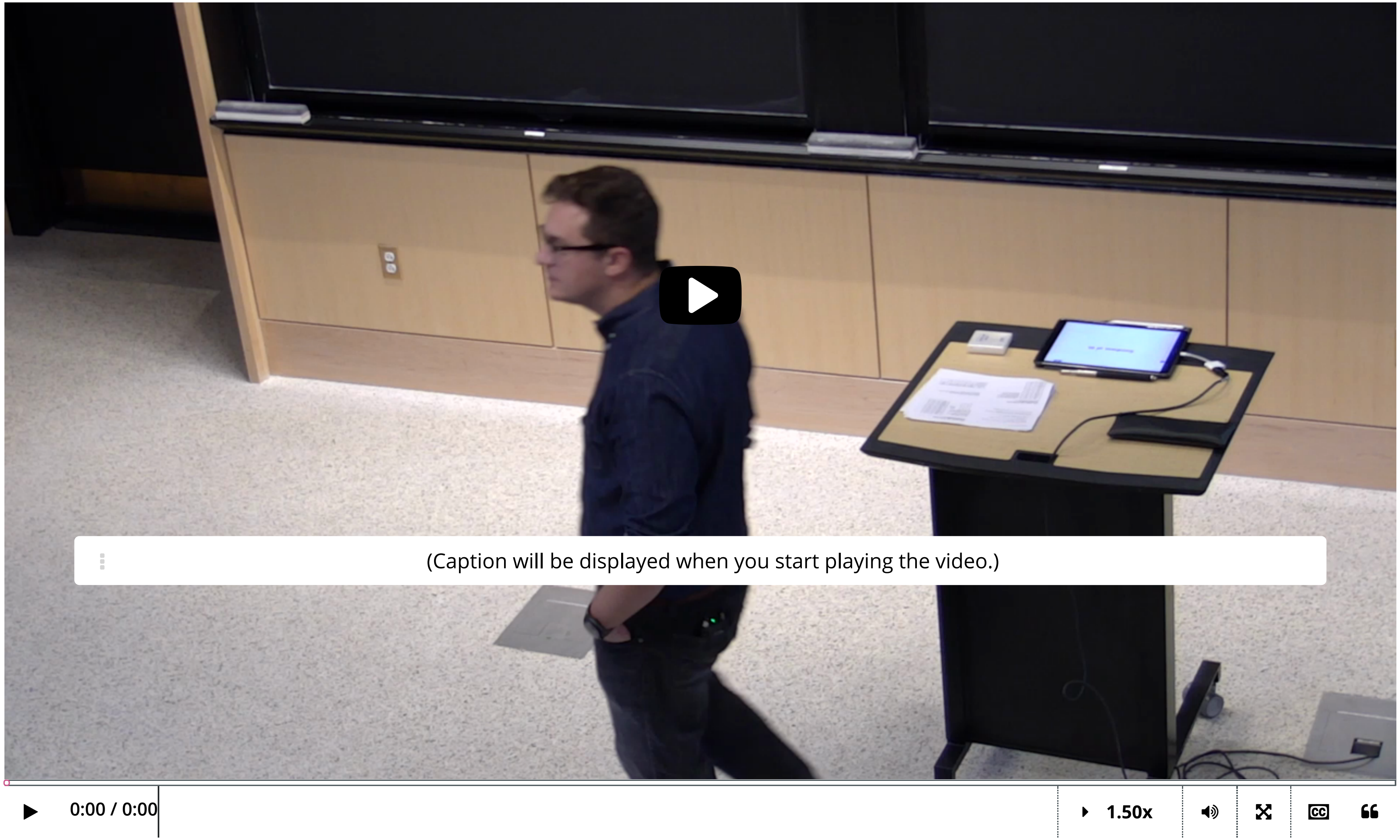
- An application of central limit theorem.

Remark: Wald's test cannot be written out because Fisher information does not exist for the uniform random variable.

Submit You have used 3 of 3 attempts

Answers are displayed within the problem

Goodness of Fit Tests: Motivation



Video
Download video file

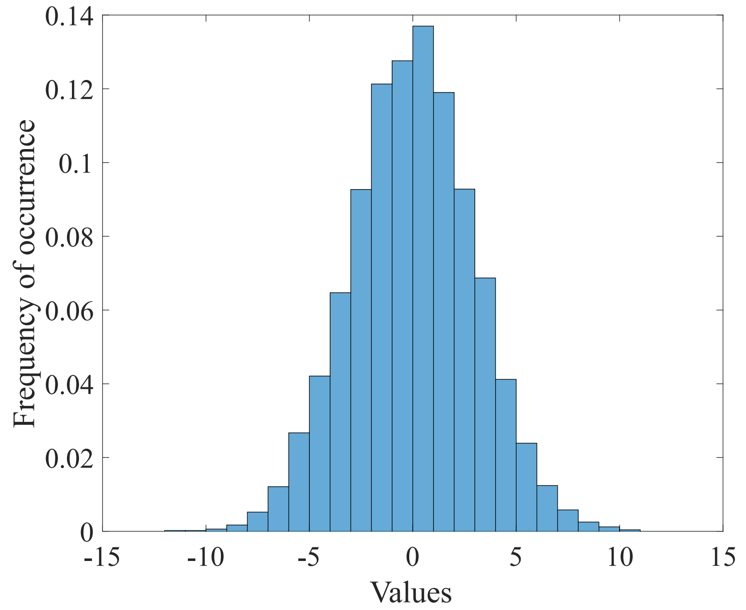
Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

Intuition for Goodness of Fit Tests

1/1 point (graded)

In the topic goodness of fit testing, we want to decide whether our data can be modeled by a specific type of distribution (**e.g.**, uniform, Gaussian, Poisson). In practice, a useful tool for making such a decision is to use a **histogram** of the data set.

A histogram for a sample data set is shown below. The x-axis, which represents the sample space, is divided into the intervals $[i, i+1]$ for all $i \in \mathbb{Z}$. The bar over the interval $[i, i+1]$ represents **how many** data points took values in that interval.



Based on the histogram above, which of the following is a reasonable choice of distribution for our data?

- Bernoulli
- Uniform on the interval $[0, 1]$
- Gaussian

χ^2 with some number of degrees of freedom

Solution:

First we examine the incorrect choices.

- The choices "Bernoulli" and "Uniform on the interval $[0,1]$" are incorrect. We observe data points outside of the interval $[0,1]$, so there is no way that our data could be distributed as either of these distributions.
- The choice "χ^2 with some number of degrees of freedom" is also incorrect. A χ^2 distributed random variable will not take negative values. Since our data set includes negative observations, it is impossible for the underlying distribution to be χ^2.

By process of elimination, the correct response is "Gaussian". The histogram seems to interpolate a bell-curved profile, so it is intuitive to conclude that our data is distributed as a Gaussian.

Remark: The histogram is a useful tool for gaining intuition as to what distribution the data follows. In the remainder of this unit, we will develop more rigorous methods to examine whether a particular data set is drawn from a given distribution.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Concept Check: Terminology

3/3 points (graded)

Suppose you observe iid samples $X_1, \ldots, X_n \sim P$ from some **unknown** distribution $\mathbf{P}$. Let $\mathcal{F}$ denote a family of known distributions (for example, $\mathcal{F}$ could be the family of normal distributions $\{\mathcal{N}(\mu, \sigma^2)\}_{\mu \in \mathbb{R}, \sigma^2 > 0}$).

In the topic of **goodness of fit testing**, our goal is to answer the question "**Does $\mathbf{P}$ belong to the family $\mathcal{F}$?**".

In particular, parametric hypothesis testing is a particular case of goodness of fit testing (why?). However, we emphasize that in the context of parametric hypothesis testing, you must have a **parametric** statistical model for the data.

Categorize the following problems as examples of parameter estimation (as studied in Unit 2), parametric hypothesis testing (as studied in the previous two lectures), or goodness of fit testing (as introduced in the above video). (Choose all categories that apply.)

Problem 1: Estimate the bias of an unfair coin.

- [x] Parameter estimation
- [ ] Parametric hypothesis testing
- [ ] Goodness of fit testing

✔

Problem 2: Decide if a 6-sided die is fair or not.

- [ ] Parameter estimation
- [x] Parametric hypothesis testing
- [x] Goodness of fit testing

✔

Problem 3: Decide if the heights of pine trees in Canada have a Gaussian distribution. Assume that the statistical model for this data is $\{\mathbb{R}, \Pi\}$ where Π denotes the set of **all** probability distributions with sample space $\mathbb{R}$. (In particular, this model is **non-parametric**.)

- [ ] Parameter estimation
- [ ] Parametric hypothesis testing
- [x] Goodness of fit testing

✔

Solution:

We examine the problems in order.

- The first choice is an example of **parameter estimation**. The key-word **estimate** confirms that this is the correct response. We are not performing a hypothesis test on where the true parameter (the bias) lies in the parameter space. Rather, we want to come up with a good approximiation for the bias.

- The second choice is an example of **parametric hypothesis testing** and **goodness of fit testing**. A six-sided die has a statistical model $(\{1,2,3,4,5,6\}, \{\mathbf{P}_{\mathbf{p}}\}_{\mathbf{p}})$, where $\mathbf{p}$ could be any valid pmf. Testing if the die is fair is the same as testing if the true parameter is $(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$. Moreover, this is a goodness of fit test, because we want to figure out whether or not the **uniform distribution** on $\{1,2,3,4,5,6\}$ is a good fit for our data.

- The third choice is an example of **goodness of fit testing** only. For this example, the family of distributions is $\mathcal{F} = \{\mathcal{N}(\mu, \sigma^2)\}_{\mu \in \mathbb{R}, \sigma^2 > 0}$, and we want to figure out if the tree-height data can be modeled by some distribution in this family. Because our statistical model is non-parametric, this is not an example of parametric hypothesis testing or parameter estimation.

Remark: In general, goodness of fit testing is considered a topic in **non-parametric statistics**, in contrast to the material we have covered so far. You should keep in mind though that the topic of parametric hypothesis testing is a special case of goodness of fit testing. However, to handle non-parametric models we will need to develop new techniques.

Submit You have used 3 of 3 attempts

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 4 Hypothesis testing:Lecture 15: Goodness of Fit Test for Discrete Distributions / 2. Introduction to Goodness of Fit Tests

Add a Post

All Posts

MLE estimator technical conditions for uniform distribution

question posted 7 days ago by riskia

In the solution of the review problem, it was stated that MLE estimator for uniform distribution does not satisfy some technical conditions required for asymptotic normality. I went through the course notes and can't find the exact details of this explanation. Could someone point me in the right direction? Thanks.

This post is visible to everyone.

Add a Response

6 responses

yangyf
7 days ago

+1

Support depends on the parameters is usually one problem.

posted 6 days ago by **markweitzman** (Community TA)

θ^* being on the boundary being another problem.

posted 6 days ago by **dna47a**

@markweitzman, thanks on the comment! In particular, how parameter dependence of distribution support render invalidate MLE asymptotic normality? Is there any place where one can read a bit more?

P.S.: Ideally, something more specific than "just read this elementary 1753 pages book on Statistics"? =)

posted 6 days ago by **es87f6es87f6**

Add a comment

Analesdey
5 days ago

I also didnt understand what the conditions are. The professor only mentioned MLE conditions have to be satisfied. Could be because the mle is not found from the derivative per se. as mentioned its θ^{MLE} being on the boundary?

Add a comment

KhaledRA
3 days ago

+1

Add a comment

arsenyspb
3 days ago

This course... We are constantly pushed but not taught... knowledge isn't shared... never again in my life.

Add a comment

bradchattergoon
2 days ago

Here is a link to a description of the Regularity Conditions for the MLE: https://courses.cit.cornell.edu/econ620/reviewm5.pdf

You can see on page 2: 7. The log-likelihood function must be twice continuously differentiable in a neighborhood of the true value. The reason this is not satisfied in the Uniform Distribution case is that the the true value is at the boundary of the distribution so it is not continuous. This point is for the asymptotic convergence of the distribution and since Wilk's Theorem is about asymptotic convergence option 4 wouldn't work.

I also got this wrong and I agree with the others that it is somewhat unfair to casually mention that there are Regularity Conditions but never actually spell them out before testing on it. That said, I appreciate the question so that the mathematical nuances of what we are doing are highlighted.

Add a comment

marijas_data_kitchen
about 9 hours ago

+
...

Thank you for asking this question @riskia. I also got it wrong and was at first a bit maddened by it. Slide 38, Unit3, does list out the conditions for asymptotic normality of the MLE. Now I think I learned more from this question by getting it wrong, than if I had gotten it right :-)

Add a comment

Showing all responses

Add a response:

Preview

Submit