Mathematical Statistics II

Hypothesis Testing

Jesse Wheeler

Introduction

Introduction

- One way to address parameter uncertainty is considering the variance of our estimator, $\hat{\theta}$.
- In some cases, the exact variance can be calculated, using known distributions. In other cases, it needs to be approximated (e.g., observed-information at the MLE).
- Alternative ways of accounting for uncertainty is in the form of hypothesis tests and confidence intervals.
- This material is largely based on Rice (2007, Chapter 9), Casella and Berger (2024, Chapter 8), as well as some supplemental material from Pawitan (2001).

Definitions

Definition: Hypothesis

A hypothesis is a statement about a population parameter.

- Example: The mean height μ of a population is smaller than 6 ft.
- The definition of a hypothesis is quite general. The important part is that it makes a statement about a population parameter.
- The goal of a hypothesis test is the decide, based on a sample from the population, which hypotheses are true.

Definitions II

- In practice, we usually perform a hypothesis test by considering two complimentary hypotheses. Our goal is to decide which is true, based on data.

Definition: Null and Alternative Hypothesis

The two complementary hypotheses are called the null hypothesis and the alternative hypothesis, denoted H_0 and H_1, respectively.

- For population parameters, we usually consider hypotheses of the form:

$$H_0 : \theta \in \Theta_0 \quad \text{and} \quad H_1 : \theta \in \Theta_0^c,$$

where Θ_0 is some subset of the parameter space.

Definitions III

Blood Pressure medication

Suppose we are interested in testing a new blood pressure medication against existing treatments.

The population parameter of interest is θ, the mean change in blood pressure of this new drug, relative to existing treatments.

We may be interested in testing if $\theta \in \Theta_0 = \{0\}$, against $\theta \in \Theta_0^C = \mathbb{R}/\{0\}$, or:

$$H_0 : \theta = 0, \quad H_1 : \theta \neq 0.$$

Definitions IV

- In a *hypothesis testing problem*, the experimenter must decide to either accept H_0 (or equivalently, reject H_1), or accept H_1 (or equivalently, reject H_0).
- We take the approach of Casella and Berger (2024), and not worry about the particular language used, rather focus on the mathematics and final decision.

Definitions V

Definition: Hypothesis tests

A hypothesis testing procedure, or hypothesis test is a rule that specifies:

(i) For which sample values the decision is made to accept H_0 as true.

(ii) For which sample values H_0 is rejected and H_1 is accepted as true.

The subset of the sample space for which H_0 is rejected is called the rejection region or critical region. The complement of the rejection region is called the acceptance region.

Definitions VI

- Rather than specifying individual possibilities for the sample space, we generalize by considering a test statistic: $W(X) = W(X_1, X_2, \ldots, X_n)$, which is a function of the sample.

Example: Coin Flipping

Suppose we have a new coin, with probability of heads θ. In our experiment, we will flip a coin $n = 4$ times. Consider making a hypothesis test for $H_0 : \theta = 0.5$ against $H_1 : \theta \neq 0.5$.

$\Omega = \{HHHH, HHHT, \ldots, TTTT\}$,

$|\Omega| = 2^4 = 16$.

$R = \{HHHH, TTTT\}$ ←

$A = R^c = \Omega / R.$

$X_i \to 0$ if tails

$X_i \to 1$ if heads

$X_i \sim Bern(\theta)$.

$W(X_1, X_2, X_3, X_4) = \sum_i X_i$,

$R = \{\omega \in \Omega : W(\omega) = 0, 4\}$

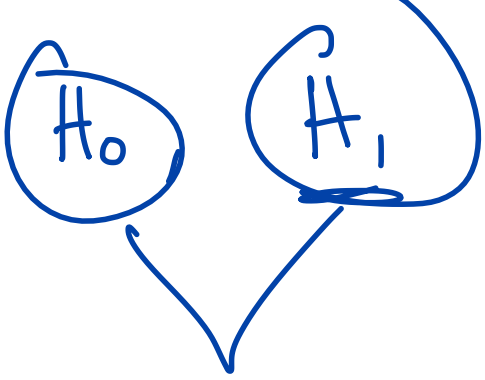
$= \{HHHH, TTTT\}$.

$R = \{HHHH, HHHT, THHH, TTTT\}$ biased towards $\theta > 0.5$

Definitions VII

- In the example above, we could have picked *any* rejection region that we wanted.
- What we would like to do is develop ways to mathematically compare hypothesis testing frameworks.
- We will discuss a few different ways of developing hypothesis tests, and how to compare different testing frameworks.
- Like with point-estimators, this will involve the consideration of some general principles that should be considered.

Error types and testing paradigms



Complimentary, No overlap.

Rejection Region $R = \{\omega : \text{we reject } H_0\}$.

"Acceptance Region" $= R^c = \{\omega : \text{we fail to reject.}\}$

$R: \{HTTTT, THTTT, \ldots, TTTTH\}$

$R: \{\omega: X(\omega) = 1\}$, $X(\omega) = 1$ if ω is sample with 1 head.

$R = \{X = 1\}$

Testing Paradigms

- There are various disagreements about how to perform / what to prioritize when conducting hypotheses tests.
- We will categorize these ideals into three main groups:

1. Bayesian hypothesis testing.
    - The Bayesian treatment of probability as beliefs enables us to consider probabilities that H_0 or H_1 is true, based on the posterior distribution. That is, since $\pi(\theta|x)$ is a density for a random variable $\Theta|X = x$, we can consider
      $P(H_0 \text{ is true}|X) = P(\theta \in \Theta_0|X)$, and
      $P(H_1 \text{ is true } |X) = P(\theta \in \Theta_0^C|X)$.
    - See Chapter 8.2.2 of Casella and Berger (2024) for more details and examples.

Testing Paradigms II

- Frequentist hypothesis testing
  - In the Frequentist setting, we cannot consider $P(H_0 \text{ is true}|X)$, since θ is not random. That is, if $\theta \in H_0$, $P(H_0 \text{ is true } |X) = 1$, and zero otherwise. The thing that is random is the data X, not the parameter θ, and consequently the decision to accept or reject is random. Here, we consider the probability of being wrong under various assumptions.

2. Neyman-Pearson approach. This has largely dominated hypothesis testing in 20th / 21st centuries. If you've seen a hypothesis test, it's likely based on this approach.
3. Fisher approach. Less commonly used, though sometimes mixed with the *Neyman-Pearson* paradigm.

Error Types

- We'll largely focus on the Neyman-Pearson paradigm, since that is the most widely used approach.
- To do so, we need to introduce error types. There are two ways we can make an error:
  1. Conclude that H_0 is false, when it is actually true. This is called a Type I error.
  2. Conclude that H_0 is true, when it is actually false. This is called a Type II error

Table 1: Decision and error types.

	H_0 True	H_0 False
Accept H_0	correct	Type II
Reject H_0	Type I	correct

Error Types II

- The Neyman-Pearson approach focuses on these two types of error.
- Because of randomness, we can't completely eliminate both (or either) type of error.
- There's also a tradeoff: tests that have larger-rejection regions are more likely to reject, meaning more likely to have a Type I error, but a lower Type II error.
- Thus, efforts to decrease one of the error types typically results in an increase of the other type.

Error Types III

- In the frequentist setting, we are interested in long-run frequencies. That is, for any fixed tests, what are the corresponding rates that we make errors.
  - Let α be the Type-I error rate.
  - Let β be the Type-II error rate.

Neyman-Pearson Hypothesis Testing:

- Control the Type I error rate α at some pre-determined level (e.g., $\alpha \leq 0.05$). We call this the significance level of the test.
- Within this constraint, we want to also minimize β, or equivalently, maximize $1 - \beta$, which is known as the power of the test.

Error Types IV

Example: Flipping a coin

Consider the experiment of flipping a coin $n = 10$ times. We would like to test the hypothesis $H_0 : \theta = 0.5$ against $H_1 : \theta \neq 0.5$, where θ is the probability of heads. Find ~~the~~ a rejection region for an $\alpha = 0.05$ level test.

$X = \{0, 1, 2, \ldots, 8, 9, 10\}$

$P_\theta(X = x) = \binom{10}{x} \theta^x (1-\theta)^{n-x}$

$P_{\theta=0.5}(X = 0) = (0.5)^{10} \approx 0.00098$

$P_{\theta=0.5}(X = 1) = 10\,(0.5)^{10} \approx 0.00977$

$P_{\theta=0.5}(X=2) \simeq 0.04395$

$\vdots$

$P_{\theta=0.5}(X=8) = 0.04395$

$P_{\theta=0.5}(X=9) = 0.00977$

$P_{\theta=0.5}(X=10) = 0.00098$.

Rejection Region R, such that

$$P(X \in R \mid H_0 \text{ is true}) \leq \alpha = 0.05$$

$R_1 = \{0\}$, $R_2 = \{10\}$, $R_3 = \{1\}$, $R_4 = \{9\}$,

$R_5 = \{0,1\}$, $R_6 = \{9,10\}$, $R_7 = \{2\}$,

$R_8 = \{8\}$

$R_9 = \{0,1,9,10\}$

$R_{10} = \{2,10\}$

...

$$P(X \in R_i \mid H_0) \leq 0.05.$$

$$\max P(X \in R_i \mid H_0^c)$$

$$P(X \in R_i \mid \theta \neq 0.5)$$

$$R_i \subseteq R_j,$$

$$P(X \in R_i \mid H_0^c) \leq P(X \in R_j \mid H_0^c)$$

plot power, $P_i(\theta)$:

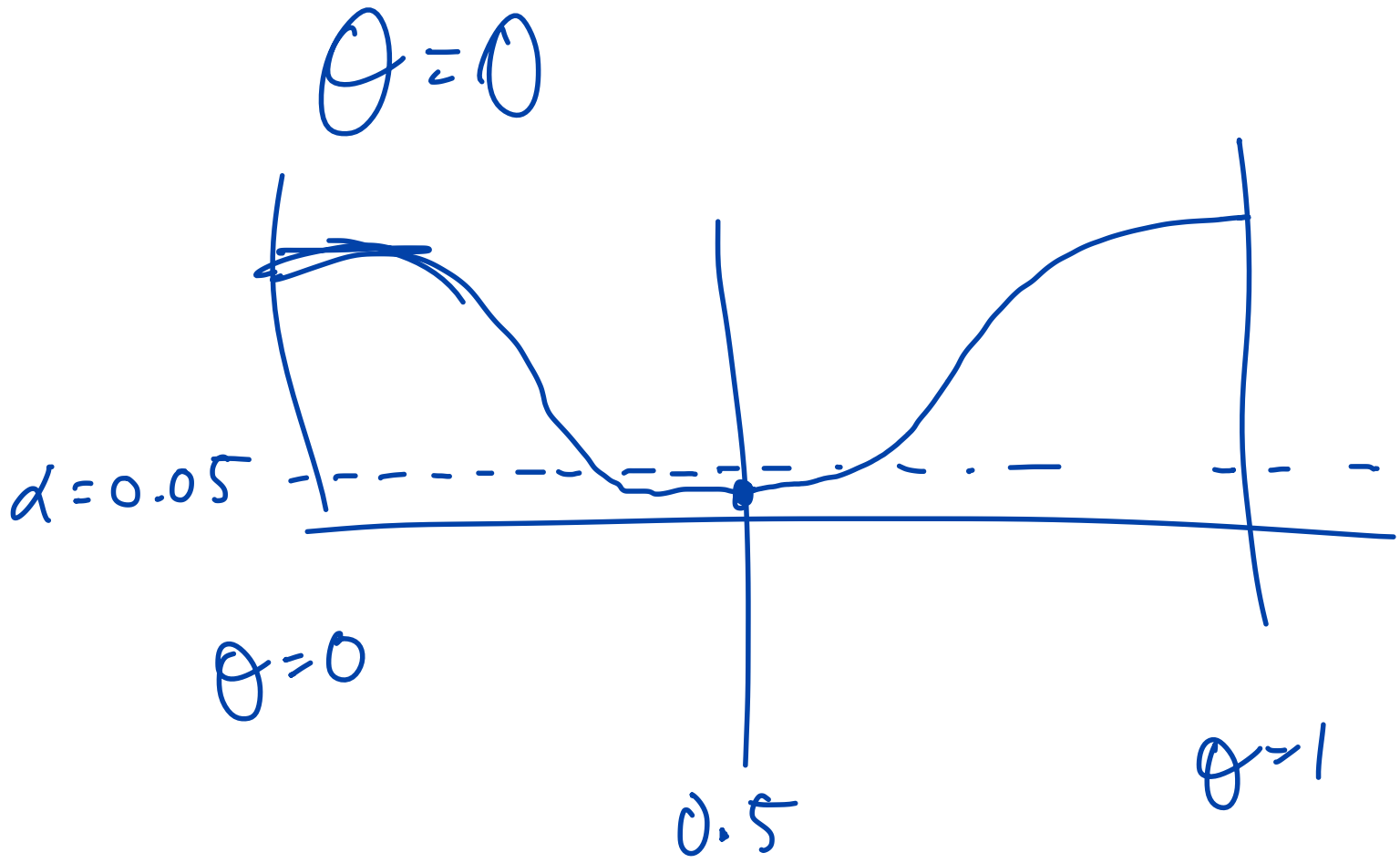
$$P_q(\theta) = P(X \in R_q \mid \theta \neq 0.5)$$

$$P_{10}(\theta) = P(X \in R_{10} \mid \theta \neq 0.5).$$

$$P_{10}(\theta) = P_\theta(X = 2 \text{ or } X = 10)$$

$$= P_\theta(X = 2) + P_\theta(X = 10)$$

$\theta = 0.50001$ close to α



Error Types V

- What we derived on the previous slide is a type of exact hypothesis test. That is, because we knew the exact distribution of both the data $X_1, \ldots, X_n$ and the sample statistic $W(X_1, \ldots, X_n) = \sum_i X_i$, we could pick a rejection region based on the distribution of $W(X_1, \ldots, X_n)$.
- Other type of exact tests are available. Perhaps the most famous is the t-test: If X_i are iid Normal(μ, σ^2), then $T = (\bar{X} - \mu)/(S/\sqrt{n})$ follows exactly a t-distribution with $df = n - 1$ (see Chapter 6 from last semester).
- Most often, however, an exact test is not available, and we will need an alternative way to derive tests and rejection regions.
- Before we present these methods, we will introduce P-values and Fisher's approach to hypothesis testing.

P-values and Fisher's approach

- Informally, the P-value can be interpreted as the probability of observing something as extreme or more extreme, if H_0 is true.
- If the P-value is small, that provides evidence that H_0 is not true.
- A technical definition of a P-value is given in 8.3.26 of Casella and Berger (2024).
- The rejection-region framework presented above is mathematically equivalent to computing a P-value, and comparing to the test size $\alpha = 0.05$.

P-values and Fisher's approach II

- In the binomial example, consider observing $X = 1$. With respect to $H_0 : \theta = 0.5$ and $H_1 : \theta \neq 0.5$, there are 3 other events that are as extreme (or more extreme) than $X = 1$, namely: $X \in \{0, 1, 9, 10\}$.
- Thus, the P-value is: $P_{\theta=0.5}\big(X \in \{0, 1, 9, 10\}\big) \approx 0.022$.
- Since P-value $< \alpha = 0.05$, we reject the null hypothesis and conclude that H_1 is true.

P-values and Fisher's approach III

- The Neyman-Pearson framework used above implies: the long-run error Type I error rate of repeating this test is less than $\alpha = 0.05$, so we have controlled the Type I error rate.
- There are many growing complaints about this framework. Two that have persisted (and are relevant to our discussion) include:
  - The choice $\alpha = 0.05$ (or any other value) is completely arbitrary. What tolerance you have for a Type I error rate should be problem specific.
  - This framework makes the most sense when a study is going to be repeated, since it is formally a statement about the accuracy of many repeated tests, not our single example.

P-values and Fisher's approach IV

- Building off of our example, suppose instead we are testing $H_0 : \theta = 0.5$ against a one-sided alternative: $H_1 : \theta < 0.5$, and then we observe $X = 2$. In this case, the P-value is:

$$P = P_{\theta=0.5}\big(X \in \{0, 1, 2\}\big) \approx 0.055 > \alpha = 0.05.$$

- In this example, the Neyman-Pearson framework would select H_0, and not H_1, without any statement about how close this result is to the arbitrarily chosen $\alpha = 0.05$.
- The Fisherian interpretation of hypothesis tests addresses this issue: rather than picking a pre-determined α value (which is arbitrary), we consider the P-value as a continuous measure of evidence against H_0, rather than a binary decision.

P-value of 0.00001 < α

P-values and Fisher's approach V

- This approach isn't without difficulties, as it is more subjective and does not immediately result in a decision in regards to H_0 and H_1.
- Despite there existing many valid arguments why a Bayesian framework or Fisherian interpretation may be preferred, the Neyman-Pearson paradigm has largely "won" in modern science, to the extent that many publication venues will question why you didn't use Neyman-Pearson if you try another approach.

P-values and Fisher's approach VI

- The American Statistical Association made a formal statement, somewhat discouraging the use of the Neyman-Pearson approach (and P-values in general) in 2016 (Wasserstein and Lazar, 2016). This has been extremely impactful, and has led to debate and some controversy (Ionides *et al.*, 2017).

$\hat{\theta}$, $Var(\hat{\theta})$

→ Bayes

→ Frequentist

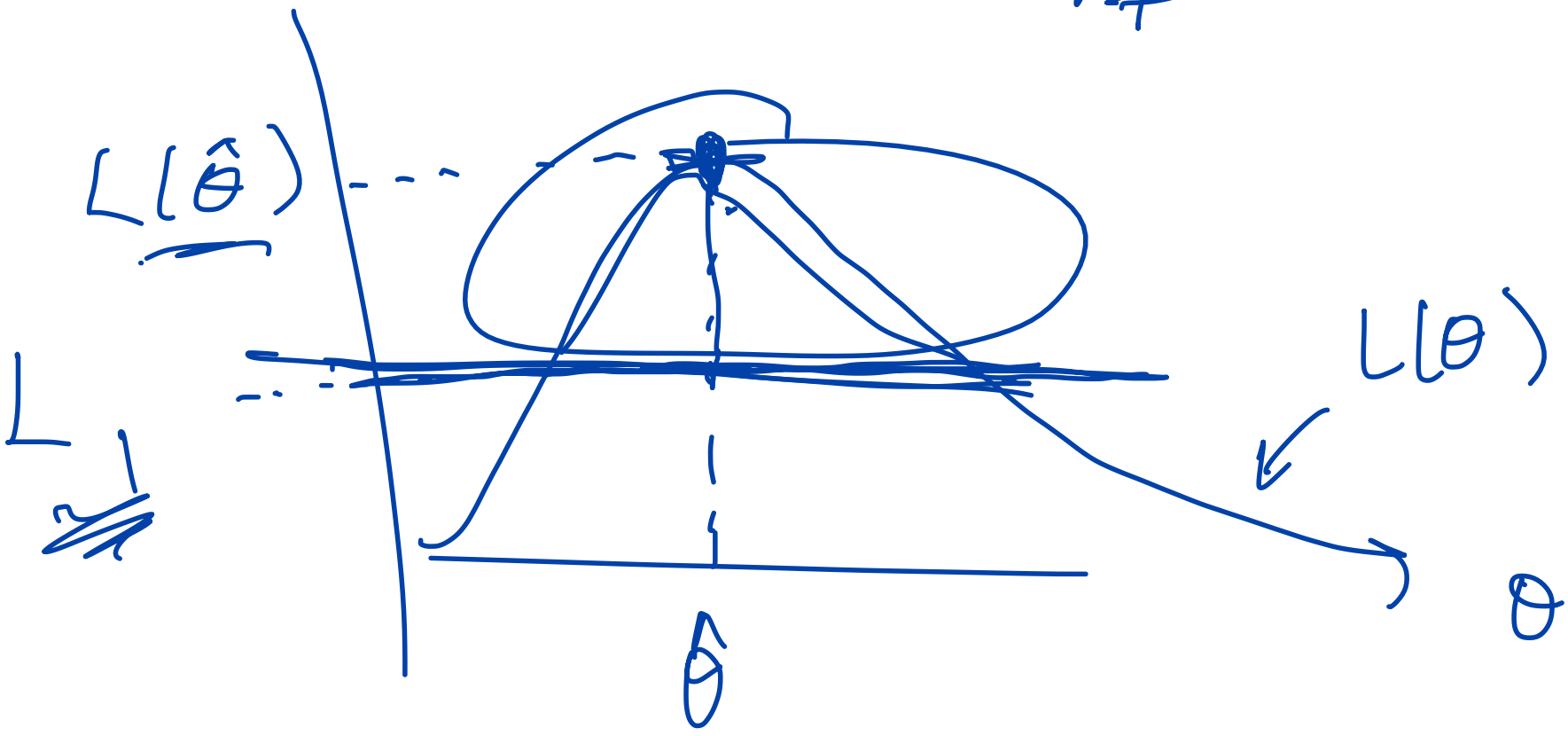
→ Neyman-Pearson

→ Fishers

Likelihood Ratio Tests

$\frac{\alpha}{2} = 0.05$ /

95% of time, we do not have Type I error.



$$1 > \frac{L_1}{L(\hat{\theta})} > c$$

Likelihood Ratio Tests

- We have already considered the likelihood function $L(\theta)$ as a way of measuring uncertainty: the data supports values of θ with higher values of likelihood than lower values.
- What matters is not the value of the likelihood, but how this value compares to other possible likelihood values.
- This intuition gives rise to Likelihood-Ratios, which can be used to build hypothesis tests for general models and hypotheses.
- We'll see that in many situtations, these tests are optimal within the class of tests that are α-level, and unbiased.

Likelihood Ratio Tests II

Likelihood Ratio Tests

Let $L(\theta)$ denote the likelihood function for a fixed dataset and model. If Θ_0 denotes the set of θ values in H_0, and Θ the entire parameter space, then we define

$$\lambda(x) = \frac{\sup_{\Theta_0} L(\theta)}{\sup_{\Theta} L(\theta)}.$$

A likelihood ratio test is any test that has a rejection region of the form:

$$\{x : \lambda(x) \leq c\},$$

where c is any number $0 \leq c \leq 1$.

Likelihood Ratio Tests III

- Idea: If the maximum likelihood corresponding to the null hypothesis is small relative to possible values in the alternative, then we reject H_0.
- It is equivalent to $\lambda^*(x) = \frac{\sup_{\Theta_0} L(\theta)}{\sup_{\Theta_1} L(\theta)}$, but the former is preferred for practical reasons.
- If the null hypothesis is simple (i.e., $H_0 : \theta = d$ for some d), then the LRT is the likelihood of H_0 divided by the likelihood of the MLE.

Likelihood Ratio Tests IV

Likelihood Ratio Test: Normal distribution

Let X_i be iid $N(\theta, 1)$ random variables. Consider testing $H_0 : \theta = \theta_0$, against $H_1 : \theta \neq \theta_0$, where θ_0 is some fixed number. Derive the likelihood ratio test for this hypothesis testing scenario.

$\Theta_0 = \{\theta_0\}$

$\hat{\theta} = \bar{X}$.

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = \frac{L(\theta_0)}{L(\hat{\theta})}$$

$$= \frac{(2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum (x_i - \theta_0)^2\right\}}{(2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum (x_i - \hat{\theta})^2\right\}}$$

$$\lambda(x) = \exp\left\{-\frac{1}{2}\left[\sum (x_i-\theta_0)^2 - \sum (x_i-\bar{x})^2\right]\right\}$$

↓ completing square.

$$= \exp\left\{-\frac{n}{2}(\bar{x}-\theta_0)^2\right\} < C$$

$$-\frac{n}{2}(\bar{x}-\theta_0)^2 < \log C = C'$$

$$(\bar{x}-\theta_0)^2 > C''$$

$H_0: \theta = \theta_0$

$H_1: \theta \neq \theta_0$

type-I error rate to be $< \alpha$

$X_i \overset{iid}{\sim} N(\theta, 1)$

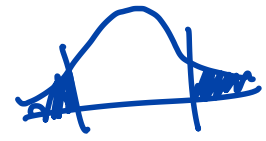
$\bar{X} \sim N\left(\theta, \frac{1}{n}\right)$

If H_0, $\bar{X} \sim N(\theta_0, \frac{1}{n})$.

and

$$\bar{X} - \theta_0 \sim N(0, \frac{1}{n}) \leftarrow$$

$$P(|\bar{X} - \theta_0| > c_n''') \leq \alpha$$



$$2P(\bar{X} - \theta_0 > c''') \leq \alpha$$

$$P(\bar{X} - \theta_0 > c''') \leq \frac{\alpha}{2}$$

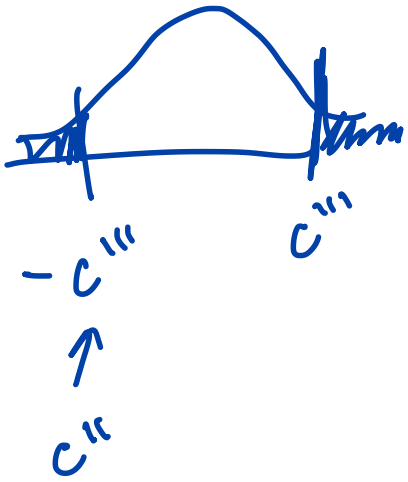
$$P\left(\frac{X-\theta_0}{1/\sqrt{n}} > \frac{c'''}{1/\sqrt{n}}\right) \leq \frac{\alpha}{2}$$

$$P\left(Z > \frac{c'''}{1/\sqrt{n}}\right) \leq \frac{\alpha}{2},$$

$Z \sim N(0,1)$.

$\Phi(\cdot)$ cdf

$P(Z < m)$



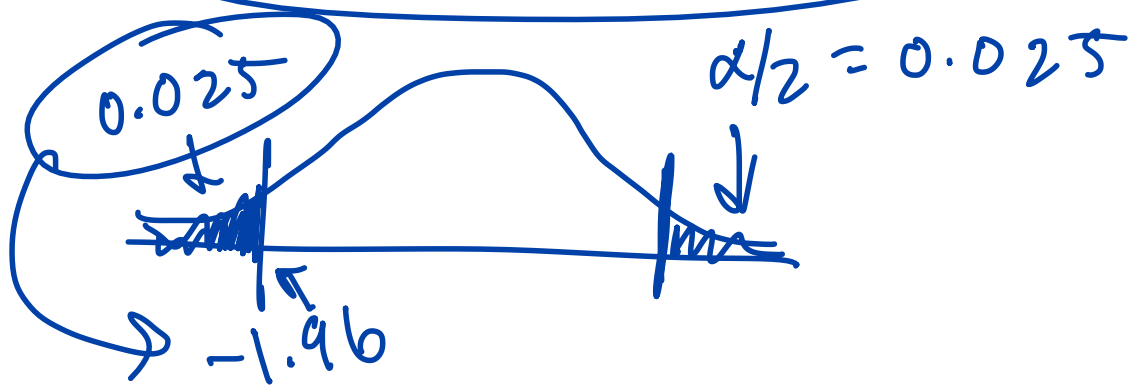
$$P\left(Z < -c'''\sqrt{n}\right) \leq \frac{\alpha}{2}$$

$$P(Z < c_4) = \frac{\alpha}{2}$$

we reject if $\dfrac{\bar{X} - \theta_0}{1/\sqrt{n}} = Z < c_4$

$$Z < \Phi^{-1}\left(\frac{\alpha}{2}\right)$$

$\alpha = 0.05$.



$$R: \left\{x: \frac{\bar{X}_n - \theta_0}{1/\sqrt{n}} < \Phi^{-1}\left(\frac{\alpha}{2}\right)\right\} \cup$$

$$\left\{x: \frac{\bar{X} - \theta_0}{1/\sqrt{n}} > \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right\}$$

if ~~$\alpha = 0.05$~~ ,

$$R = \left\{x: \left|\frac{\bar{X}_n - \theta_0}{1/\sqrt{n}}\right| > 1.96\right\}$$

* Z-test *

t-test, under H_0,

$$\frac{\bar{X} - \theta_0}{S/\sqrt{n}} \sim t_{df = n-1}$$

$$\frac{\bar{X} - \theta_0}{S/\sqrt{n}} \approx t_{df = n-1}$$

Likelihood Ratio Tests V

- Recall that the likelihood function $L(\theta)$ can equivalently be expressed in terms of sufficient statistics.
- Similarly, the likelihood ratio $\lambda(x)$ can be expressed in terms of sufficient statistics, rather than the entire data.
- When possible, it is convenient to consider the likelihood ratio test in terms of minimal sufficient statistics.
- See Theorem 8.2.4 of Casella and Berger (2024) for formal details.

$$\lambda(x) = \frac{g(T(x);\theta_0)\,h(x)}{g(T(x);\hat{\theta})\,h(x)}$$

Likelihood ratio tests in practice

- The LRT is a very common approach to finding a hypothesis test.
- Often, it will result in exact tests, where we pick a rejection region based on the known properties of a sufficient statistic (like the sample mean or variance).
- As previously noted, the likelihood ratio test results in an optimal test in many situations. We will study the theory of when this happens in practice.
- Finally, we will discuss theory that enables likelihood ratio tests in practice, even when the exact distribution of a sufficient statistic is not available.

Likelihood ratio tests in practice II

Definition: Uniformly Most Powerful Test

Let $\beta(\theta)$ be the power function of a hypothesis test testing $H_0 : \theta \in \Theta_0$. We consider the class $\mathcal{C}$ of all level-α tests. A test is said to be a uniformly most powerful (UMP) level α test if:

- The test size (Type-I error rate) is less than or equal to α.
- $\beta(\theta) \geq \beta'(\theta)$ for all $\theta \in \Theta_0^c$ for any other test in class $\mathcal{C}$ that has power function $\beta'(\theta)$.

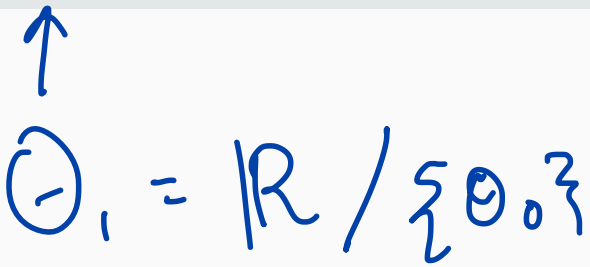
Typically, we assume the class of interest $\mathcal{C}$ is the class of unbiased tests.

Likelihood ratio tests in practice III

Definition: Simple vs Composite Hypotheses

We say that a hypothesis is simple if the corresponding sample space is a single value. E.g., $\Theta_0 = \{\theta_0\}$.

If a sample space is not simple (i.e., it contains more than one parameter), we say that the hypothesis is composite. E.g., $\Theta_0^C = \mathbb{R}/\{\theta_0\}$.



Likelihood ratio tests in practice IV

Neyman-Pearson Lemma

Consider testing two simple hypotheses: $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

Then, a hypothesis test is uniformly most powerful (unbiased) level-α test if and only if it is a likelihood ratio test.

- A formal proof is outside the scope of this class, but see the Neyman-Pearson Lemma proof sketch in Section 9.2 of Rice (2007) for more details.

Likelihood ratio tests in practice V

Definition: Monotone Likelihood Ratios

A family of distributions with pdf (or pmf) given by $f(x;\theta)$ is said to have a monotone likelihood ratio (MLR) if, for every $\theta_2 > \theta_1$, $f(x;\theta_2)/f(x;\theta_1)$ is a monotone (non-decreasing or non-increasing) function of $T(x)$, a univariate statistic (usually $T(x)$ is a sufficient statistic).

Example: Poisson distributions

Let $X_1, \ldots, X_n$ be iid from Poisson(λ). Show that this family of distributions has a monotone likelihood ratio.

$$f(x;\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_{i=1}^{n} x_i!}$$

$$\frac{f(x|\lambda_2)}{f(x|\lambda_1)} = \frac{e^{-n\lambda_2} \cdot \lambda_2^{\sum x_i} \cdot \frac{1}{\prod_{i=1}^{n} x_i!}}{e^{-n\lambda_1} \lambda_1^{\sum x_i} \cdot \frac{1}{\prod_{i=1}^{n} x_i!}}$$

$$= e^{-n(\lambda_2 - \lambda_1)} \left(\frac{\lambda_2}{\lambda_1}\right)^{\sum x_i}$$

$$\star = e^{-n(\lambda_2 - \lambda_1)} \left(\frac{\lambda_2}{\lambda_1}\right)^{T(x)}.$$

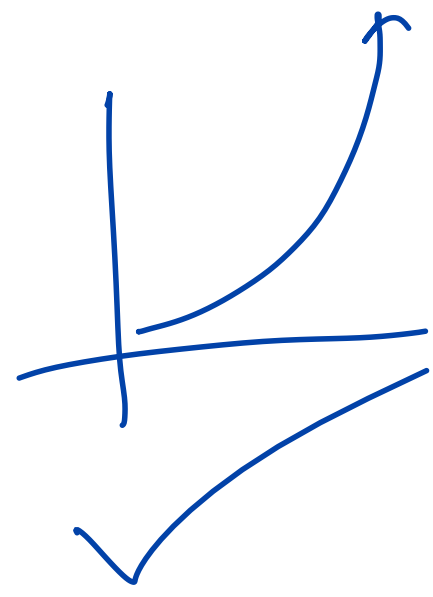
if $\lambda_2 > \lambda_1$,

$\lambda_2 > \lambda_1$ $e^{-n(\lambda_2 - \lambda_1)} > 0$

$\left(\frac{\lambda_2}{\lambda_1}\right) > I$

$\star = C \cdot (\rho)^{t}$

$t_2 > t_1$, $c\rho^{t_2} > c\rho^{t_1}$



Likelihood ratio tests in practice VI

- The MLR property is both convenient, and common.
- In fact, you can show that any (regular) exponential family distribution has the MLR property. This includes the following distributions:
  - Normal, Gamma, Beta, Bernoulli, Poison, Exponential, Chi-squared, Dirichlet, Categorical, Wishart, Inverse Wishart, Geometric
- This list is not all distributions with the MLR property, but going forward you may assume any of these distributions do have this property.
- The MLR property will be used to extend the Neyman-Pearson Lemma to composite hypotheses.

Likelihood ratio tests in practice VII

The Karlin-Rubin Theorem

Consider testing composite hypothesis $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ (or, equivalently, we can change the direction of the inequalities, or test the simple hypothesis $H_0 : \theta = \theta_0$ against the composite $H_1 : \theta > \theta_0$). If there exists a sufficient statistic T for θ and the family of pdfs (or pmfs) has a the MLR property, then an unbiased hypothesis test is uniformly most powerful level-α test if and only if it is a likelihood ratio test.

- If $T(X)$ is the sufficient statistic with MLR property, it can be shown that the Karlin-Rubin Theorem implies that the best possible level-α test is rejecting if and only if $T > c$, with c chosen such that $P_{\theta_0}(T > c) = \alpha$.

Likelihood ratio tests in practice VIII

- See Theorem 8.3.17 of Casella and Berger (2024) for more details.
- This result gives strong support for using a likelihood-ratio test. Common difficulties include:
  - Showing that $f(x;\theta)$ has a MLR.
  - If it does, calibrating c such that $P_{\theta_0}(T > c) = \alpha$. To do this, we either need to know the distribution of T, or approximate it. The later can be done by simulation, or Wilks' Theorem

Likelihood ratio tests in practice IX

Wilks' Theorem

Consider performing the likelihood ratio test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Under H_0 and mild conditions on the likelihood function (See Miscellanea 10.6.2 from Casella and Berger (2024)),

$\lambda(x) = \sup_{\theta_0} L(\theta) / \sup_{\theta} L(\theta)$

$$-2 \log \lambda(x) = -2 \log \frac{L(\theta_0)}{L(\hat{\theta})} \xrightarrow{d} \chi^2_p,$$

as $n \to \infty$, where p is the difference in the number of free-parameters in the null and alternative hypotheses.

design a test s.t. that we reject if $\lambda(x) < c_\alpha$

proof:

$$\ell(\theta) = \ell(\hat{\theta}) + \cancel{\ell'(\hat{\theta})}^{0}(\hat{\theta}-\theta) + \ell''(\hat{\theta})\frac{(\hat{\theta}-\theta)^2}{2!} + \cdots$$

$$-2\log\lambda(x) = -2\log\frac{L(\theta_0)}{L(\hat{\theta})}$$

$$= -2\,\ell(\theta_0) + 2\,\ell(\hat{\theta})$$

$$-2\log\lambda(x) = \chi^2$$

$$(\hat{\theta}-\theta)^2 \cdot \left(-\ell''(\hat{\theta})\right) \to I(\hat{\theta})$$

$$I_n(\hat{\theta})$$

Slutsky's

$$-2 \log \lambda(x) \approx \frac{(\hat{\theta}-\theta)^2}{-l''(\hat{\theta})} \rightarrow \mathcal{N}(0, \mathcal{I}(\hat{\theta}))$$

$$\rightsquigarrow \chi^2_1$$

Likelihood ratio tests in practice X

- Wilks' Theorem gives us a way to calibrate an approximate LRT. Suppose we are testing one-parameter, and want $\alpha = 0.05$. Then, we want to solve for c in the LRT:

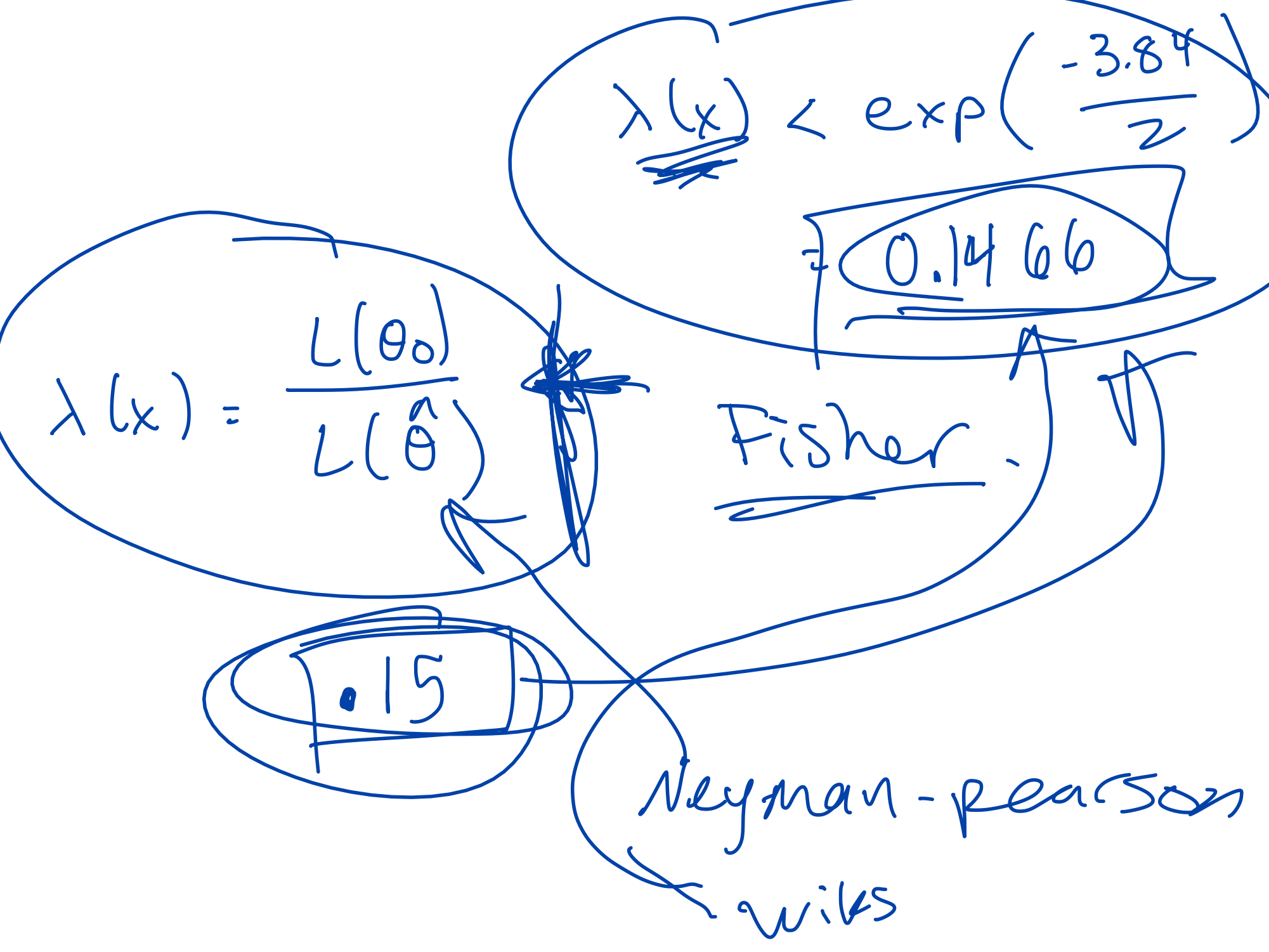
$$P(\lambda(x) < c) = 0.05$$
$$P(\log \lambda(x) < c') = 0.05$$
$$P(-2 \log \lambda(x) > c'') = 0.05$$

  Using the approximation now that $-2 \log \lambda(x) \approx \chi^2_1$, we can find c'' using software $c'' = 3.84$.

- Thus, an approximate test rejects when:

$$-2 \log \lambda(x) > 3.84.$$

$$\log \lambda(x) < \frac{-3.84}{2}$$


$\lambda(x) < \exp\left(\frac{-3.84}{2}\right)$
$= 0.1466$
$\lambda(x) = \frac{L(\theta_0)}{L(\hat{\theta})}$
Fisher.
.15
Neyman-pearson
wilks

4/28 review

Hypothesis: statement about population parameter.

$H_0: \theta \in \Theta_0$

$H_1: \theta \notin \Theta_0$, $\theta \in \Theta_0^c$.

$\theta = \theta_0$

$\theta > \theta_0$

Simple Hypothesis: $\theta = \theta_0$

H_0: $\theta = \theta_0$

H_1: $\theta = \theta_1$.

Neyman - Pearson Lemma

α-level, unbiased test,

$$\lambda(x) = \frac{L(\theta_0)}{L(\theta_1)}$$, reject $\lambda(x) < c$.

if $P(\lambda(x) < c \mid H_0) \leq \alpha$,

then NP lemma states this is UMP.

→ simple Hypotheses.

→ Karlin-Rubin theorem:

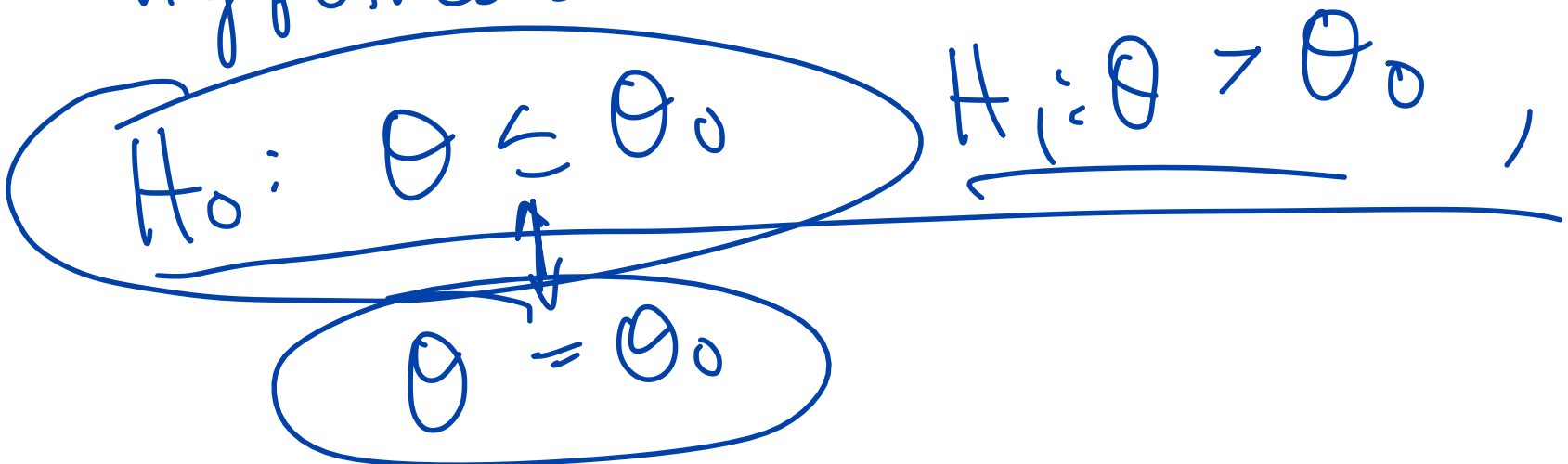
If $\lambda(x)$ is monotone,

MLR property,

then the likelihood ratio test

is UMP for composite

hypotheses.



$T(X)$ is Sufficient, MLR

Reject if $T(x) > c'$.

$\rightarrow P_{H_0}(T(X) > c') \leq \alpha$.

$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t$

$$-2 \log \lambda(x) \xrightarrow{d} \chi^2_p$$

p is the number of "free" parameters.

$x_i \sim \text{Beta}(\alpha, \beta)$. $p = 2$

$p = 1$

$$-2 \log \lambda(x) = -2 \log \frac{L(\theta_0)}{L(\hat{\theta})}$$

$$= 2 \log \frac{L(\hat{\theta})}{L(\theta_0)}$$

$$= 2 \left[\ell(\hat{\theta}) - \ell(\theta_0) \right] \sim \chi^2_p$$

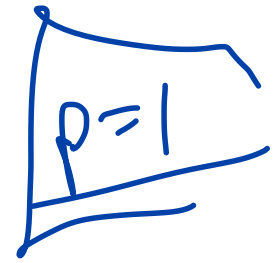
exact LRT:

$$P(\lambda(x) < c) = \alpha = 0.05$$

$$P(\log \lambda(x) < c') = 0.05$$

Wilk's approx.

$$P(-2\log\lambda(x) > c') = 0.05$$



$$P(\chi^2_p > c'') = 0.05.$$

reject if:

$$2\left[\ell(\hat{\theta}) - \ell(\theta_0)\right] > \chi^2_{1,0.95} = 3.84$$

$\alpha = 0.05$ Level.

Reject if

$$\ell(\hat{\theta}) - \ell(\theta_0) > 1.92$$

$$R = \{x : \ell(\hat{\theta}) - \ell(\theta_0) > 1.92\}$$

Likelihood ratio tests in practice XI

- How good is this approximation? very good.
- Not only do we have asymptotic properties, Pawitan (2001, Section 2.9) demonstrates that, due to the invariance property of the MLE and likelihood ratios, the test is exact if there exists some one-to-one transformation $g(\cdot)$ *that we do not need to know* such that $g(X_i)$ are normally distributed.
- Combining these results, even if no such transformation g exists, there likely exists a transformation g where $g(X_i)$ are approximately normal, or at least symmetric, in which case the asymptotic approximation is accurate with even very small n.

Confidence Intervals

Interval Estimation

- We conclude by discussing interval estimators.
- Rather than just getting a point estimate $\hat{\theta}$ for a parameter θ, we estimate an entire interval I.

Definition: Interval estimate

An *interval estimate* of a real-valued parameter θ is any pair of functions $L(X)$ and $U(X)$ of a sample $X = (X_1, \ldots, X_n)$ that satisfy $L(X) \leq U(X)$ for all X.

If $X = x^*$ is observed, then the inference $L(x^*) \leq \theta \leq U(x^*)$ is made.

The random interval $I(X) = \big(L(X), U(X)\big)$ is called an interval estimator.

Interval Estimation II

- There are many different types of interval estimators, both Bayesian and Frequentist.
- Here, we will focus on the Frequentist confidence interval, as these are the most widely used.

Definition: Coverage probability

For an interval estimator $I(X) = \Big(L(X), U(X)\Big)$ of a parameter θ, the coverage probability of $I(X)$ is the probability that the random interval covers the true parameter, θ. That is,

$$P_\theta\big(\theta \in I(X)\big).$$

Above, the interval $I(X)$ is what is random, and θ is fixed.

Interval Estimation III

- For any interval estimator, we're interested in the *worst* performance, or the minimum coverage probability.

Definition: Confidence coefficient.

For an interval estimator $I(X)$ of a parameter θ, the confidence coefficient of $I(X)$ is the infimum of the coverage probabilities:

$$\inf_{\theta} P_\theta\big(\theta \in I(X)\big).$$

- Sometimes we are interested in *confidence sets*, rather than just intervals, though intervals are the most common.
- A given interval estimator $I(X)$, combined with a confidence coefficient, gives rise to a confidence interval.

Interval Estimation IV

Uniform confidence interval

Let $X_1, \ldots, X_n$ be iid from Uniform$(0, \theta)$. Based on previous homework, the MLE for θ is $X_{(n)}$, the maximum of the observations.

Find a $1 - \alpha$ confidence set for θ, of the form:
$I(X) = [aX_{(n)}, bX_{(n)}]$, with $1 \leq a < b$.

$$\left(a \cdot X_{(n)} \;,\; b \cdot X_{(n)} \right)$$

$$\inf_\theta P_\theta\left(\theta \in [a X_{(n)} \;,\; b X_{(n)}] \right) = 1-\alpha = 0.95$$

$I_{[b=\infty]} = \left[(1-\alpha)^{-1/n} X_{(n)}, \infty \right)$

$(1-\alpha)\%$

$I = \left[X_{(n)}, \alpha^{-1/n} X_{(n)} \right]$

Finding Confidence Intervals

- We'll next cover a few ways that confidence intervals can be created.
- If the exact sampling distribution of a point estimator is known, we can often derive exact confidence intervals.
- In practice, however, this often is not possible.

Duality of confidence intervals and hypothesis tests

- The most common way to find confidence intervals is to connect them with a hypothesis test.
- That is, we can find a α-level hypothesis test, and *invert* the test to get a confidence interval.

Inverting a normal test

Let $X_1, \ldots, X_n$ be iid from $N(\mu, \sigma^2)$. Treating σ^2 as fixed, find a $1 - \alpha$ confidence set for μ by inverting a hypothesis test.

$$R = \left\{ x : |\bar{x} - \mu_0| > z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}$$

level α test.

reject if $\dfrac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2}$

$$A = \left\{ x : |\bar{x}-\mu_0| \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\}$$

$$= \left\{ x : \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\}$$

$$\inf_{\mu} P_\mu\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right) \ge 1-\alpha .$$

✓

$$I_{1-\alpha} : \left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} , \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

$\alpha = 0.05$, $z_{\alpha/2} = 1.96$

$$\left[\bar{X} - 1.96 \left(\frac{\sigma}{\sqrt{n}} \right) , \bar{X} + 1.96 \left(\frac{\sigma}{\sqrt{n}} \right) \right]$$

$X_i \sim N(\mu, \sigma^2)$, σ known

Duality of confidence intervals and hypothesis tests II

- This example can be extended to a more general setting. Proof is left as an exercise.

Theorem: Duality of hypothesis tests and confidence intervals

Consider the hypothesis test of $H_0 : \theta = \theta_0$. This involves the specification of a rejection region $R_{\theta_0}(x)$, which depends on the null value θ_0. Similarly, we define the acceptance region as $A_{\theta_0}(x) = R^c_{\theta_0}(x)$.

The set of all parameter values θ such that a fixed x lies in the acceptance region $A_{\theta_0}(x)$ is denoted $C(x) = \{\theta_0 : x \in A_{\theta_0}(x)\}$.

$C(x)$ is a $1 - \alpha$ confidence set for θ if and only if $A_{\theta_0}(x)$ is the acceptance region of a α-level hypothesis test.

Duality of confidence intervals and hypothesis tests III

- The last theorem gives us a way to systematically build confidence intervals: create a hypothesis test, and find the values of θ where the null hypothesis would be accepted.
- Theorem 9.3.5 of Casella and Berger (2024) implies that if a UMP level-α test is inverted, the resulting confidence set will have uniformly better coverage than any other $1 - \alpha$ confidence set.
- This gives strong motivation to invert likelihood ratio tests to obtain confidence sets.

Duality of confidence intervals and hypothesis tests IV

- Like hypothesis testing, we also can consider two primary types of confidence intervals:
  - *Exact* confidence intervals, which require knowing the distribution of the lower and upper bound of $I(X)$. Where possible, these are preferred.
  - *Approximate* confidence intervals, which can be found using various strategies

Duality of confidence intervals and hypothesis tests V

- There are several ways to get approximate confidence intervals. Here are perhaps the three most common:
  - Wald-Intervals. This is likely what you have seen in Math 3350, and was hinted at in Chapter 05. Since $\sqrt{\mathcal{I}_n(\theta_0)}(\hat{\theta} - \theta_0)$ limits to a standard normal, an approximate $1 - \alpha$ confidence interval is given by:

$$I(X) = \left(\hat{\theta} - z_{\alpha/2} I_n^{-1/2}(\hat{\theta}), \hat{\theta} + z_{\alpha/2} I_n^{-1/2}(\hat{\theta})\right)$$

  - Likelihood based sets. This involves inverting a likelihood ratio test, and using the approximation $-2\lambda(x) \sim \chi^2_1$, assuming we are estimating only one parameter. In this case, the approximate confidence set is given by:

$$C(X) = \left\{\theta : 2 \log \frac{L(\hat{\theta})}{L(\theta)} < \chi^2_{df=1,1-\alpha}\right\}.$$

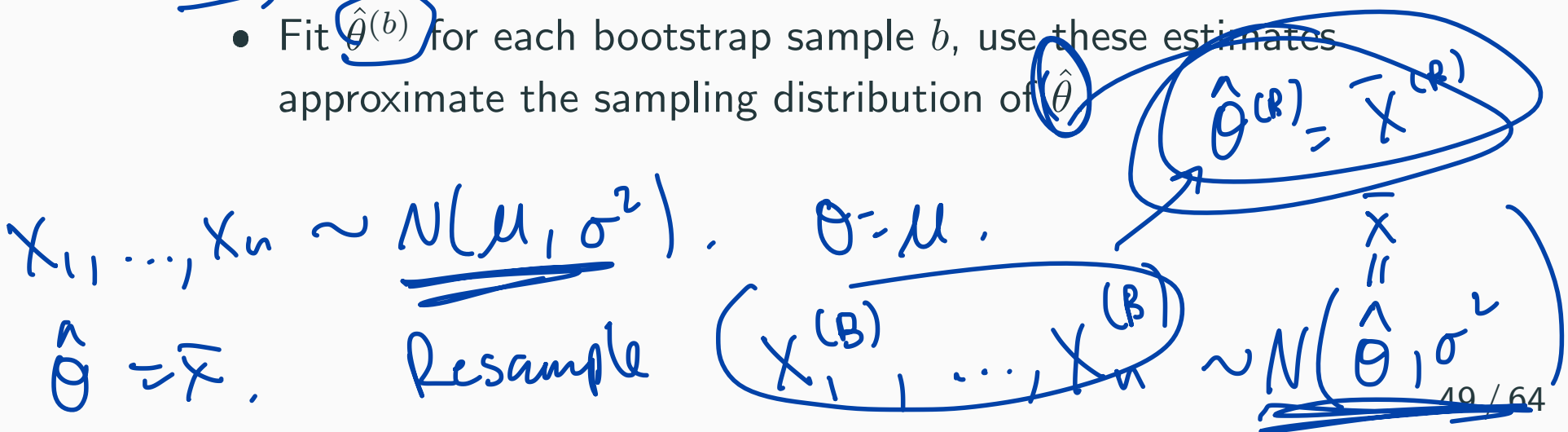
  - Bootstrap based intervals. Both parametric and non-parametric bootstrapping methods.

Introduction

- The Boostrapping technique was first formally described by Brad Efron in the late 1970s.
- The idea is quite simple, but theoretical guarantees are very technical.
- Idea: When the distribution of a sample statistic (or estimator) is unkown or complex, we will approximate the distribution using randomly drawn samples.
- Perhaps surprisingly, the bootstrap can be shown to have faster convergence rates than the standard Wald confidence intervals (normal approximation).
- Two main forms: parameteric and non-parametric.

Parametric Bootstrap

- The *parametric bootstrap* focuses on parameters of fixed families of distributions.
- Idea: The MLE converges very quickly to θ_0, so even in finite samples, $\hat{\theta} \approx \theta_0$. For a fixed model F_θ, we want the sampling distribution of $\hat{\theta}$ under F_{θ_0}, so replace this with the sampling distribution of $F_{\hat{\theta}}$.
  - Fit $\hat{\theta}$ using data $X_1, \ldots, X_n$, given model F_θ.
  - For $b = 1, 2, \ldots, B$, generate a bootstrap sample: $X_1^{(b)}, X_2^{(b)}, \ldots, X_n^{(b)}$.
  - Fit $\hat{\theta}^{(b)}$ for each bootstrap sample b, use these estimates approximate the sampling distribution of $\hat{\theta}$.



Nonparametric Bootstrap

- The other form of Bootstrap has similar convergence guarantees, but works more generally in situations where we cannot simulate from the probability distribution in question.
- As with all bootstrap, we will base our estimates on samples drawn from a new distribution. This time, we will draw from the empirical distribution.
  - Fit $\hat{\theta}$ using data $X_1, \ldots, X_n$, given model F_θ.
  - Create $b = 1, \ldots, B$ bootstrap samples by sampling iid from the empirical distribution $\hat{F}$. (This is done by sampling, with replacement, from the observed data).
  - Recompute estimate $\hat{\theta}^{(b)}$, for each sample $X_1^{(b)}, X_2^{(b)}, \ldots, X_n^{(b)}$, and use these values to approximate the sampling distribution.

Example: Poisson confidence intervals

- We will wrap up this discussion by building a confidence interval for λ using 5 different approachs:
  1. Wald confidence intervals (easy).
  2. Likelihood-Ratio based intervals (moderate).
  3. Exact intervals (difficult).
  4. Parametric Bootstrap (easy).
  5. Non-parametric Bootstrap (easy).

Example: Poisson confidence intervals II

- We have $n = 25$ observations, $X_1, \ldots, X_n$.
- We wish to model the data as Poisson(λ).
- A minimal sufficient statistic for λ is $\sum_i X_i$. We will use:

$$\sum_i X_i = 389$$

- Real data will be given later for the non-parametric bootstrap example.

Profile likelihoods

- The profile likelihood is used when θ is multivariate.
- In this setting, the likelihood function $L(\theta)$ is difficult to visualize.
- Informally, the profile likelihood considers the contribution to the likelihood of a single parameter, after integrating out other parameter values.
- This is particularly useful if there are nuisance parameters, or we are particularly interested in only one of the parameters in the vector θ.

Profile likelihoods II

- We'll split the parameter vector into two parts: $\theta = (\phi, \nu)$, and ϕ and ν can both be multivariate themselves.
- For this class, we'll assume that $\phi \in \mathbb{R}$ is the parameter of interest, and $\nu \in \mathbb{R}^{\nu_d}$ is a nuisance parameter vector.

Profile likelihoods III

Definition: Profile likelihood

If $\theta = (\phi, \nu)$ is a parameter vector with likelihood $L(\theta) = L(\phi, \nu)$, the profile likelihood of ϕ is defined as:

$$L_p(\phi) = \sup_{\nu} L(\phi, \nu),$$

where the maximization of ν is taken with the fixed value of ϕ.

Formally:

$$L_p(\phi) = \sup_{\theta \in \{(\phi_0, \nu) \in \Theta : \phi_0 = \phi\}} L(\theta).$$

Profile likelihoods IV

Gaussian profile likelihood

Let $X_1, \ldots, X_n$ be iid from $N(\mu, \sigma^2)$, where both parameters are unknown. We're often primarily interested in the parameter μ, and not σ^2. Find the profile likelihood function $L_p(\mu)$, treating σ^2 as a nuisance parameter. Compare $L_p(\mu)$ to the function $L^*(\mu) = L(\mu, \hat{\sigma}^2)$.

Profile vs Slice Comparison

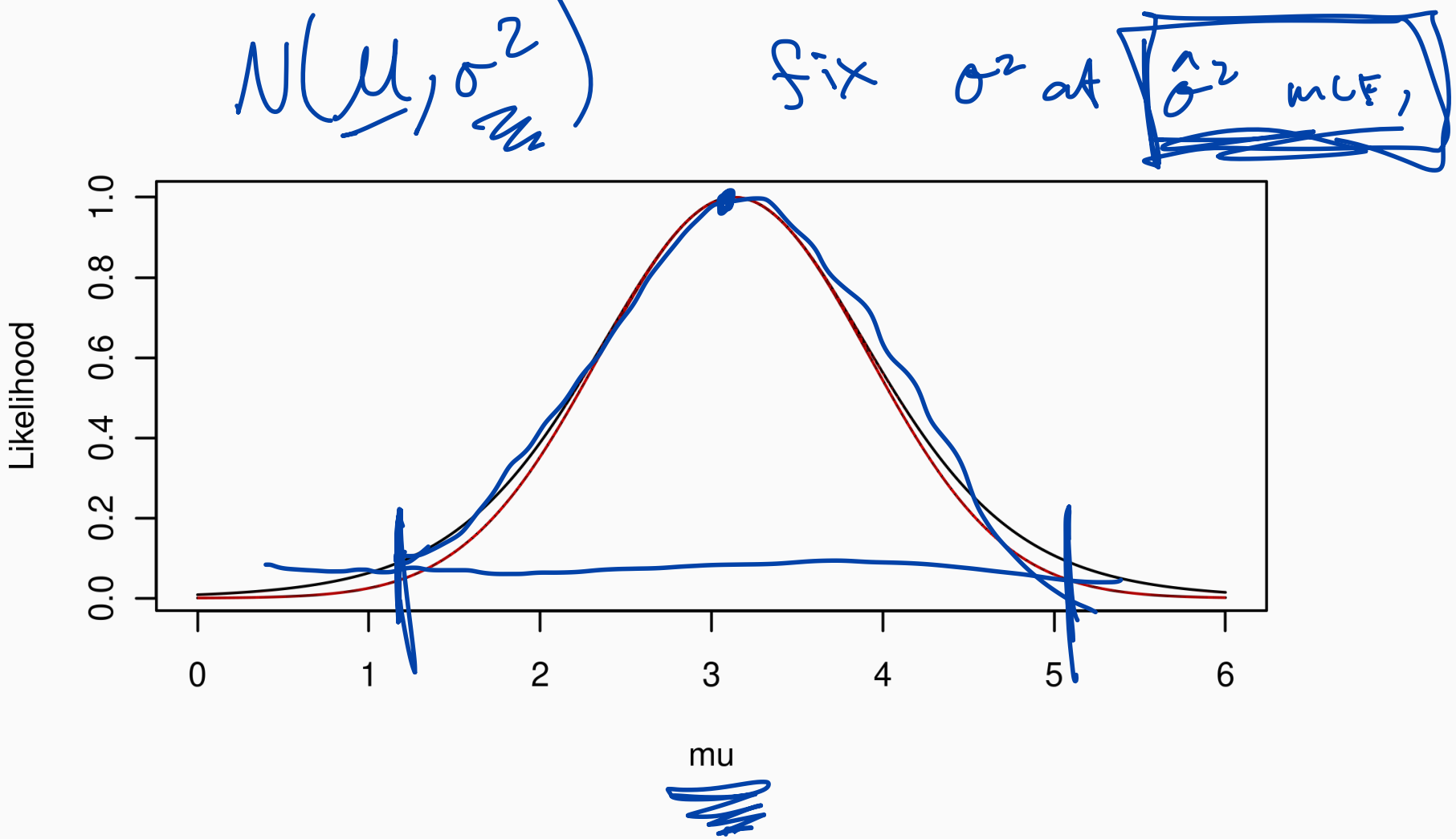


Figure 1: Profile for μ (black curve) vs slice at $\hat{\sigma}^2_{MLE}$ (red line). The profile accounts for uncertainty involved in the estimation of σ^2.

Calibrating Profiles Likelihoods

- Profile likelihoods can readily be used to compute confidence intervals.
- Formally, this is a form of a likelihood-ratio-test based confidence interval, and application of Wilk's Theorem.
- In particular, if we fix nuisance parameters η at the MLE, then Wilk's theorem will also apply to profile likelihoods.
- This works because the maximum of the profile likelihood is the maximum of the complete likelihood.

Calibrating Profiles Likelihoods II

Profile Likelihood Tests and Intervals

Let $\theta = (\phi, \nu)$ be the parameter vector of interest, and ν a vector of nuisance parameters. Then,

$$-2 \log \frac{L_p(\phi_0)}{L_p(\hat{\phi})} \sim \chi^2_p,$$

where p is the dimension of ϕ.

Calibrating Profiles Likelihoods III

- This result implies that, for even high-dimensional models, we can get approximate $1-\alpha$ level confidence intervals for any parameter ϕ by considering values of ϕ with large profile (log)-likelihoods:

$$C(x) = \left\{ \phi : 2\big(\ell_p(\hat{\phi}) - \ell_p(\phi)\big) < \chi^2_{p,(1-\alpha)} \right\}.$$

- Because the maximum of the profile is the same point as the MLE, then we just need to compare profile likelihood values of ϕ to the likelihood at the MLE.

Calibrating Profiles Likelihoods IV

- For instance, if $\alpha = 0.05$ and the dimension of ϕ is 1, then the confidence interval is any value ϕ that is within 1.92 units of the MLE:

$$C(x) = \left\{ \phi : \ell(\hat{\theta}) - \ell_p(\phi) < 1.92 \right\}$$

- In practice, these types of intervals are easy to compute numerically, even when likelihood function is intractable.
- Multiple studies have shown that these intervals are extremely accurate, especially in regards to alternatives like Wald-based intervals (one of my own: Wheeler and Ionides, 2025).

jesse.wheeler@gmail.com.

References and Acknowledgements

Casella G, Berger R (2024). *Statistical inference*. Chapman and Hall/CRC.

Ionides EL, Giessing A, Ritov Y, Page SE (2017). "Response to the ASA's statement on p-values: context, process, and purpose." *The American Statistician*, **71**(1), 88–89.

Pawitan Y (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA.

References and Acknowledgements II

Wasserstein RL, Lazar NA (2016). "The ASA statement on p-values: context, process, and purpose."

Wheeler J, Ionides EL (2025). "Revisiting inference for ARMA models: Improved fits and superior confidence intervals." *PLoS One*, **20**(10), e0333993.

- Compiled on April 16, 2026 using R version 4.5.3.

References and Acknowledgements III

- We acknowledge students and instructors for previous versions of this course / slides.