Module 8

Bootstrap / Hypothesis Testing

Kentaro Nakamura

GOV 2003

April 2nd, 2026

Today's Agenda

- **Midterm**: Awesome job for everyone!
  - I will review the question 1 next week
  - If you want to submit the regrading request, please do so asap
- Case-Control design (problem set 3)
- Bootstrap
- Hypothesis test
  - Likelihood ratio test
  - Score test
  - Optimality of Likelihood ratio test

Case-Control Design (1)

- **Setup**: Binary outcome $Y \in \{0, 1\}$ and regressor X
  - Goal is to estimate $\Pr(Y_i = 1 \mid X_i)$

- Suppose that we sample the observation based on outcome Y_i
  - Let $S_i = 1$ if the observation i is in the sample and 0 otherwise
  - The sampling probability is

$$s_y = \Pr(S_i = 1 \mid Y_i = y) = \Pr(S_i = 1 \mid Y_i = y, X_i)$$

    which is chosen by design (i.e., known to researchers)
  - The second equality is because the sampling does not depend on X_i, which means $S \perp X \mid Y$.

- **Important**: Sample conditional probability $\Pr(Y_i = 1 \mid X_i, S_i = 1)$ is NOT equal to $\Pr(Y_i = 1 \mid X_i)$ (selection on dependent variable!)

Case-Control Design (2)

- If $\Pr(Y_i = 1)$, then $\Pr(Y_i = 1 \mid X_i = x)$ is identifiable
- This is because

$$\Pr(Y_i = 1 \mid X_i = x)$$
$$= \frac{\Pr(Y_i = 1)\Pr(X_i = x \mid Y_i = 1)}{\Pr(Y_i = 1)\Pr(X_i = x \mid Y_i = 1) + \Pr(Y_i = 0)\Pr(X_i = x \mid Y_i = 0)}$$

- Also, because $S \perp X \mid Y$,

$$\Pr(X_i = x \mid Y_i = 1) = \Pr(X_i = x \mid Y_i = 1, S_i = 1)$$

which means that if we know $\Pr(Y_i = 1)$, we can identify $\Pr(Y_i = 1 \mid X_i = x)$

- However, $\Pr(Y_i = 1) \neq \Pr(Y_i = 1 \mid S_i = 1)$, so unless you have $\Pr(Y_i = 1)$ from external data, you cannot use it.

Case-Control Design (3)

- Now, the sample odds ratio is

$$\begin{aligned}
&\frac{\Pr(Y_i = 1 \mid X_i = x, S_i = 1)}{\Pr(Y_i = 0 \mid X_i = x, S_i = 1)} \\
&= \frac{\Pr(S_i = 1 \mid X_i = x, Y_i = 1)\Pr(Y_i = 1 \mid X_i = x)}{\Pr(S_i = 1 \mid X_i = x, Y_i = 0)\Pr(Y_i = 0 \mid X_i = x)} \\
&= \underbrace{\frac{\Pr(S_i = 1 \mid Y_i = 1)}{\Pr(S_i = 1 \mid Y_i = 0)}}_{\text{Constant (not depend on } x)} \times \underbrace{\frac{\Pr(Y_i = 1 \mid X_i = x)}{\Pr(Y_i = 0 \mid X_i = x)}}_{\text{Population Odds}}
\end{aligned}$$

- Thus, we can still identify the odds ratio

$$\text{Odds Ratio}(x, x') = \frac{\frac{\Pr(Y_i=1|X_i=x)}{\Pr(Y_i=0|X_i=x)}}{\frac{\Pr(Y_i=1|X_i=x')}{\Pr(Y_i=0|X_i=x')}} = \frac{\frac{\Pr(Y_i=1|X_i=x,S_i=1)}{\Pr(Y_i=0|X_i=x,S_i=1)}}{\frac{\Pr(Y_i=1|X_i=x',S_i=1)}{\Pr(Y_i=0|X_i=x',S_i=1)}}$$

Bootstrap

- **Idea**: Mimic sampling from population by **resampling** from the sample with replacement
  - Suppose that each sample is drawn i.i.d. from population distribution

- Different bootstrap methods
  - Percentile bootstrap
  - Parametric bootstrap
  - (Extra) Wild bootstrap

Percentile Bootstrap

- **Example**: Suppose that $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n \sim \mathcal{P}$ with unknown data distribution $\mathcal{P}$. Then, we fit the regression model

$$Y_i = \boldsymbol{X}_i^\top \beta + \epsilon_i, \quad \mathbb{E}[\epsilon_i \mid \boldsymbol{X}_i] = 0$$

- **Procedure**: for each $b \in \{1, \cdots, B\}$,
  - we first draw the bootstrap sample with replacement. We then obtain the **bootstrap sample**

$$\{(X_1^{(b)}, Y_1^{(b)}\}, \cdots, (X_n^{(b)}, Y_n^{(b)})\}$$

  - Then, we compute the estimator from the bootstrap sample

$$\hat{\beta}^{(b)} = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i^{(b)} (\boldsymbol{X}_i^{(b)})^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i^{(b)} \boldsymbol{Y}_i^{(b)}\right)$$

- Once we repeat the procedure above B times, we obtain $\hat{\beta}^{(b)}$ for $b = 1, \cdots, B$
  - By taking the quantile, we can obtain the distribution and thus can obtain confidence interval.

Parametric Bootstrap

- **Example**: Suppose that $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n \sim \mathcal{P}$ with unknown data distribution $\mathcal{P}$. Then, we fit the regression model

$$Y_i = \boldsymbol{X}_i^\top \beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

  - Notice that we assume parametric assumption on error term!

- **Procedure of Parametric Bootstrap**:
  - We fit the OLS and obtain $\hat{\beta}$ and $\hat{\sigma}^2$
  - For each $b \in \{1, \cdots, B\}$, we then generate the **bootstrap sample** from the fitted model:

$$Y_i^{(b)} = \boldsymbol{X}_i^\top \hat{\beta} + \epsilon_i^{(b)}, \quad \epsilon_i^{(b)} \sim \mathcal{N}(0, \hat{\sigma}^2)$$

  - Then, we compute the estimator from the bootstrap sample

$$\hat{\beta}^{(b)} = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i Y_i^{(b)}\right)$$

- Advantage: Better performance when model is correct
- Disadvantage: Assume parametric structure (e.g., assuming homoskedasticity)

Extra: Wild Bootstrap / Multiplier Bootstrap

- Consider the example in percentile bootstrap
  - We do not want parametric assumption on error term, but we want to use linear model.

- **Procedure of Wild Bootstrap**:
  - We fit the OLS and obtain $\hat{\beta}$ and $\hat{\epsilon}_i$
  - For each $b \in \{1, \cdots, B\}$, we generate the **bootstrap sample**:

$$Y_i^{(b)} = \boldsymbol{X}_i^\top \hat{\beta} + \hat{\epsilon}_i \cdot v_i^{(b)}$$

    where $v_i^{(b)}$ are **Rademacher variable** such that

$$v_i = \begin{cases} 1 & \text{with probablity } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}$$

  - Then, we compute the estimator from the bootstrap sample

$$\hat{\beta}^{(b)} = \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i Y_i^{(b)} \right)$$

- It works since $\mathbb{E}[\epsilon_i v_i] = 0$ and $\mathbb{V}[\epsilon_i v_i] = \mathbb{E}[\epsilon_i^2]$.
  - Allows heteroskedasticity as $\mathbb{E}[\epsilon_i^2]$ can vary by i

Estimation of Bias by Bootstrap

- **Setup**: Data is distributed according to distribution F: $\mathcal{D}_i \sim F$.
  - Empirical distribution is defined as $\hat{F}_n = \frac{1}{n}\sum_i^n \delta_{\mathcal{D}_i}$.
  - The estimator can be written as $\hat{\theta}_n = \theta(\hat{F}_n)$ and the bais is written as

    $$\text{Bias}(F) = \mathbb{E}_F[\theta(\hat{F}_n)] - \theta(F)$$

    but we never be able to obtain it as we cannot observe F

- In bootstrap world, we obtain the bootstrap from $\hat{F}_n$, denoted as $\{D_i^*\}_{i=1}^n$ and obtain the estimator $\theta_{n,b}^*$ for $b = 1, \cdots, B$.
  - Thus, we can estimate the bias by

    $$\widehat{\text{Bias}}(\hat{F}_n) = \mathbb{E}_{\hat{F}_n}[\theta_{n,b}^*] - \theta(\hat{F}_n)$$

  - NOTE: This cannot estimate the bias by data (e.g., selection) (to see, think of bias of estimated bias)

- Using the estimated bias, we can estimate the bias-corrected estimator

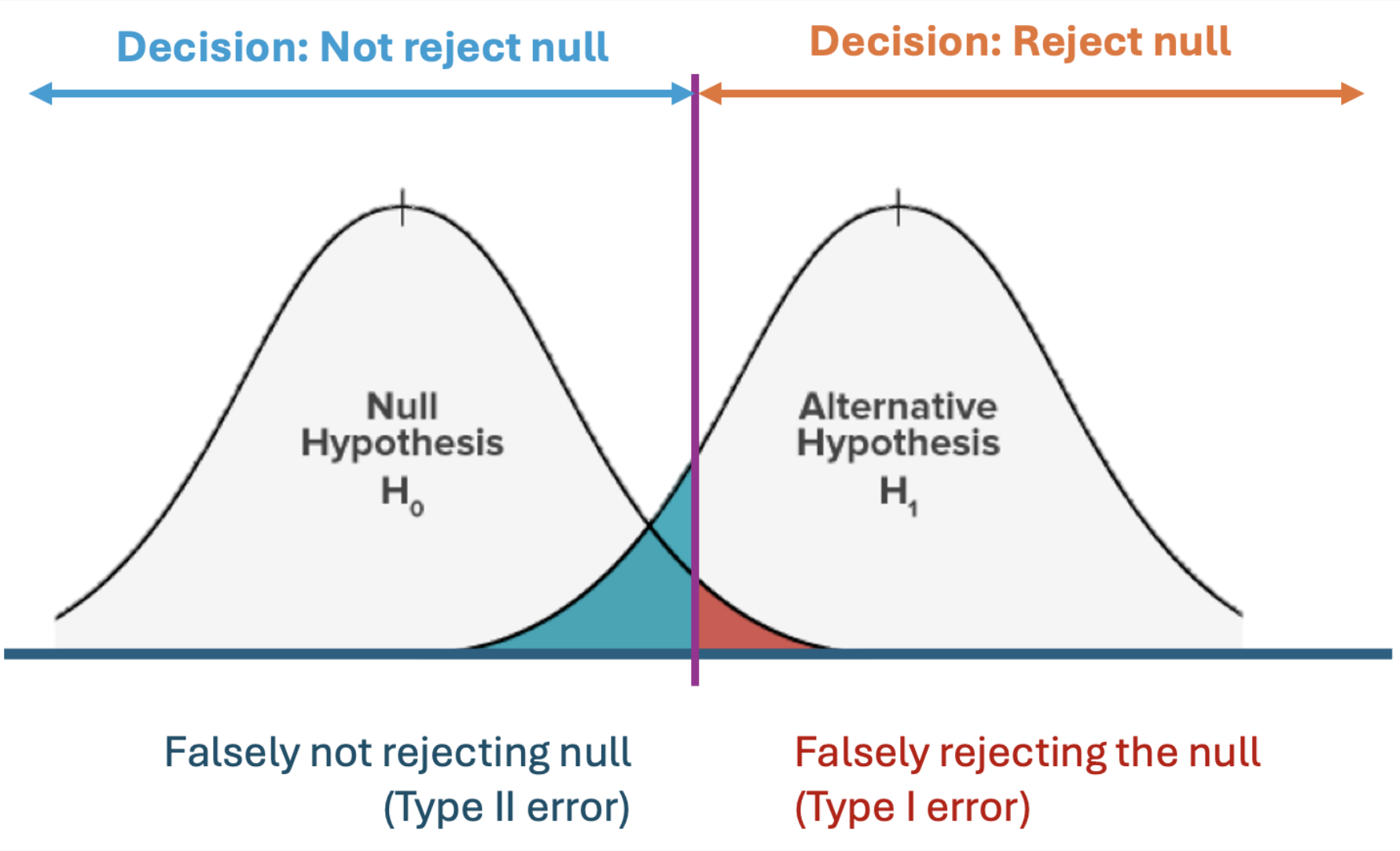
  $$\hat{\theta}_n - \widehat{\text{Bias}}(\hat{F}_n) = 2\hat{\theta}_n - \frac{1}{B}\sum_{b=1}^{B}\theta_{n,b}^*$$

Review of Hypothesis Testing (1)

	Accept H_0	**Reject** H_0
H_0 **true**	Correct Decision	Type I Error
H_1 **true**	Type II Error	Correct Decision

- The probability of Type I error is called the **size** of the test
- The rejection probability under the alternative hypothesis is called the **power** of the test
  - I.e., power equals 1 minus the probability of a Type II error
- There is a trade-off between size and power
  - Understand this trade-off from the next page's illustration

Review of Hypothesis Testing (2)



Likelihood Ratio Test

- **Setup**: Test $H_0 : \theta_0 \in \Theta_0$ versus $H_1 : \theta_0 \in \Theta \setminus \Theta_0$
  - Data: $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^n$

Theorem (Likelihood Ratio Test)

The likelihood ratio test statistic, defined as

$$\hat{\lambda}_n(\mathcal{D}) = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}_n(\theta \mid \mathcal{D})}{\sup_{\theta \in \Theta} \mathcal{L}_n(\theta \mid \mathcal{D})} = \frac{\textit{Resteicted MLE under the null}}{\textit{Unrestricted MLE}}.$$

Asymptotoically, this test statistics converges to chi-square distribution

$$-2 \log \hat{\lambda}_n(\mathcal{D}) \xrightarrow{d} \chi^2_K$$

where $K = \dim(\Theta) - \dim(\Theta_0)$

- **Decision rule**: Reject the null if $-2 \log \hat{\lambda}_n(\mathcal{D})$ is large
  - If Θ_0 is different from Θ (i.e., null is false), then $\hat{\lambda}_n(\mathcal{D})$ should be small and thus $-2 \log \hat{\lambda}_n(\mathcal{D})$ should be huge

Proof of Likelihood Ratio Test

- By Taylor expansion of likelihood function, we get

$$\ell_n(\hat{\theta}_n) = \ell_n(\theta_0) + \ell_n'(\theta_0)^\top(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top H_n(\theta_n^*)(\hat{\theta}_n - \theta_0)$$

  where θ_n^* is between $\hat{\theta}_n$ and θ_0 and H_n is the hessian (second derivative)

- Now, notice that $\ell_n'(\hat{\theta}_n) = 0$ (i.e., score) by the definition of MLE

- Thus,

$$\begin{aligned}
-2\log\hat{\lambda}_n(\mathcal{D}) &= -\{\ell_n(\hat{\theta}_n) - \ell_n(\theta_0)\} \\
&= (\hat{\theta}_n - \theta_0)^\top\{-H_n(\theta_n^*)\}(\hat{\theta}_n - \theta_0) \\
&= \underbrace{\sqrt{n}(\hat{\theta}_n - \theta_0)^\top}_{\to\mathcal{N}(0,\Omega(\theta_0)^{-1})} \underbrace{\left(-\frac{1}{n}H_n(\theta_n^*)\right)}_{\to\Omega(\theta_0)} \underbrace{\sqrt{n}(\hat{\theta}_n - \theta_0)}_{\to\mathcal{N}(0,\Omega(\theta_0)^{-1})} \\
&\xrightarrow{d} \chi^2_K
\end{aligned}$$

Extra: Optimiality of Likelihood Ratio Test

Theorem (Neyman–Pearson Lemma)

Consider testing $H_0 : f = f_0 \quad$ *versus* $\quad H_1 : f = f_1$, *where both hypotheses are simple.*
Then, among all tests ϕ *satisfying*

$$\mathbb{E}_{f_0}[\phi(X)] \leq \alpha,$$

the most powerful level-α *test rejects for large values of*

$$\frac{f_1(X)}{f_0(X)}.$$

- I.e., for simple hypothesis, likelihood ratio is most powerful (**uniformaly most powerful test**)
- You can extend this to one-sided composite null (known as Karlin-Rubin Theorem)
- However, no such test exists for two-sided composite null (see Appendix)

Extra: Proof of Neyman-Pearson Lemma (1)

- Let ϕ^* be the likelihood ratio test defined by

$$\phi^*(X) = \begin{cases} 1 & \text{if } f_1(X)/f_0(X) > c, \\ 0 & \text{if } f_1(X)/f_0(X) < c, \end{cases}$$

  with possible randomization on the boundary so that

$$\mathbb{E}_{f_0}[\phi^*(X)] = \alpha.$$

- Now let ϕ be any other test such that

$$\mathbb{E}_{f_0}[\phi(X)] \leq \alpha.$$

- We want to show that

$$\mathbb{E}_{f_1}[\phi(X)] \leq \mathbb{E}_{f_1}[\phi^*(X)].$$

Extra: Proof of Neyman-Pearson Lemma (2)

- Consider
$$\int (\phi^*(x) - \phi(x))(f_1(x) - cf_0(x))\, dx.$$

- By construction of ϕ^* and $0 \leq \phi \leq 1$, we have:
  - if $f_1(x) - cf_0(x) > 0$, then $\phi^*(x) = 1$, so $\phi^*(x) - \phi(x) \geq 0$
  - if $f_1(x) - cf_0(x) < 0$, then $\phi^*(x) = 0$, so $\phi^*(x) - \phi(x) \leq 0$

- Hence, pointwise,
$$(\phi^*(x) - \phi(x))(f_1(x) - cf_0(x)) \geq 0,$$
and therefore
$$\int (\phi^*(x) - \phi(x))(f_1(x) - cf_0(x))\, dx \geq 0.$$

Extra: Proof of Neyman-Pearson Lemma (3)

- Expanding this gives

$$\int (\phi^*(x) - \phi(x)) f_1(x)\, dx \;\geq\; c \int (\phi^*(x) - \phi(x)) f_0(x)\, dx.$$

- That is,

$$\mathbb{E}_{f_1}[\phi^*(X)] - \mathbb{E}_{f_1}[\phi(X)] \;\geq\; c\Big(\mathbb{E}_{f_0}[\phi^*(X)] - \mathbb{E}_{f_0}[\phi(X)]\Big).$$

- Since $\mathbb{E}_{f_0}[\phi^*(X)] = \alpha$ and $\mathbb{E}_{f_0}[\phi(X)] \leq \alpha$, the right-hand side is nonnegative. Thus,

$$\mathbb{E}_{f_1}[\phi^*(X)] - \mathbb{E}_{f_1}[\phi(X)] \geq 0.$$

- So

$$\mathbb{E}_{f_1}[\phi^*(X)] \geq \mathbb{E}_{f_1}[\phi(X)].$$

- Therefore, ϕ^* is the most powerful level-α test.

Asymptotic equivalence of test statistics (1)

- For intuition, consider the simple null $H_0 : \theta = \theta_0$.
- Let

$$s_n(\theta) = \ell_n'(\theta), \qquad H_n(\theta) = -\frac{1}{n}\ell_n''(\theta).$$

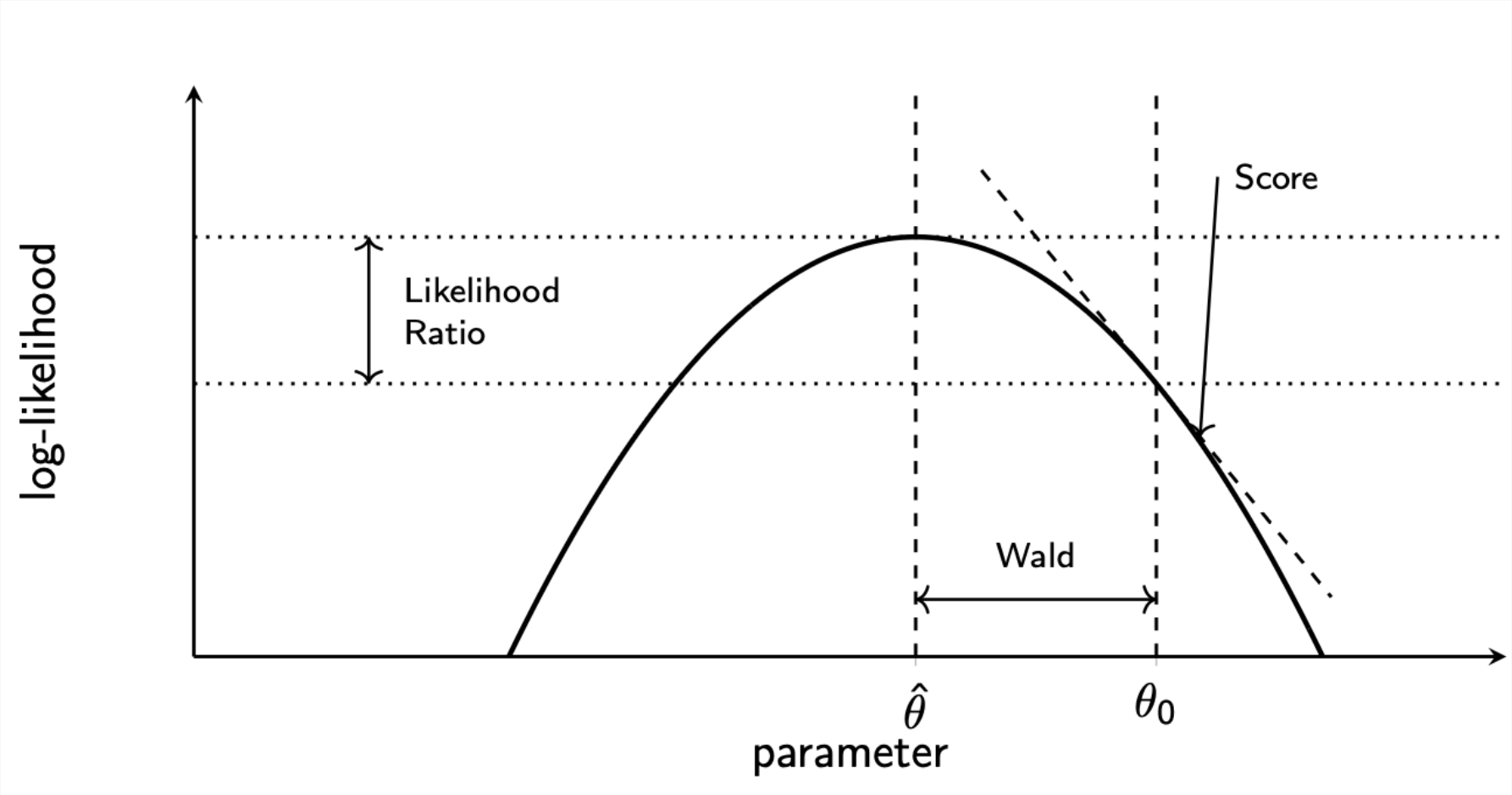
$$\text{Wald:} \quad W_n = n(\hat{\theta}_n - \theta_0)^\top H_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)$$

$$\text{Score:} \quad S_n = \frac{1}{n} s_n(\theta_0)^\top H_n(\theta_0)^{-1} s_n(\theta_0)$$

$$\text{Likelihood ratio:} \quad LR_n = 2\{\ell_n(\hat{\theta}_n) - \ell_n(\theta_0)\}$$

- Same null, three viewpoints:
  - **Wald**: how far $\hat{\theta}_n$ is from the null
  - **Score**: how steep the log-likelihood is at the null
  - **LR**: how much better the unrestricted fit is than the null

Asymptotic equivalence of Test Statistics (2)



Asymptotic equivalence of test statistics (3)

- **Likelihood Ratio Test**: As we derived,

$$LR_n \approx n(\hat{\theta}_n - \theta_0)^\top H(\theta_0)(\hat{\theta}_n - \theta_0).$$

- **Score Test**: By first-order Taylor expansion,

$$s_n(\theta_0) = \ell_n'(\theta_0) = \ell_n'(\hat{\theta}_n) + \ell_n''(\tilde{\theta}_n)(\theta_0 - \hat{\theta}_n) = \ell_n''(\tilde{\theta}_n)(\theta_0 - \hat{\theta}_n)$$

  so

$$S_n \approx n(\hat{\theta}_n - \theta_0)^\top H(\theta_0)(\hat{\theta}_n - \theta_0).$$

- **Wald Test**:

$$W_n \approx n(\hat{\theta}_n - \theta_0)^\top H(\theta_0)(\hat{\theta}_n - \theta_0).$$

- Thus, under H_0, all three have the same asymptotic χ^2 distribution.

Appendix: Optimality for Composite Null (1)

Definition (Monotone Likelihood Ratio)

The family of distribution $\{f_\theta(y) : \theta \in \Theta\}$ has **monotone likelihood ratio** (MLR) in a test statistic $T(Y)$ if the ratio $f_{\theta_2}(Y)/f_{\theta_1}(Y)$ can be expressed as a function of $(T(Y), \theta_1, \theta_2)$ and for each $\theta_1 < \theta_2$, the ratio is non-decreasing in $T(Y)$ when at least one of the numerator and denominator is positive

Example (Exponential Family)

Recall that PDF of exponential family is

$$f_\theta(Y) = \exp\{T(Y)\eta(\theta) - \psi(\eta(\theta))\}h(Y).$$

For $\theta_1 < \theta_2$, the log of the likelihood ratio is

$$\lambda = T(Y)\{\eta(\theta_2) - \eta(\theta_1)\} - \{\psi(\eta(\theta_2)) - \psi(\eta(\theta_1))\}$$

and if η is non-decreasing function, then λ is non-decreasing in $T(Y)$

Appendix: Optimality for Composite Null (2)

Definition (Uniformly Most Powerful (UMP) Test)

A test statistic $\phi(Y_i)$ is UMP at level α if (i) it has size alpha (i.e., $\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi(Y_i)] \leq \alpha$) and for every parameter value in the alternative $\theta \in \Theta_1$,

$$\mathbb{E}_\theta[\phi(Y_i)] \geq \mathbb{E}_\theta[\phi^*(Y_i)] \quad \text{for all other test } \phi^* \text{ with size } \alpha$$

- A UMP test is the test that has the highest power for every possible alternative

Theorem (Karlin-Rubin (Casella and Berger 2002, Chapter 8))

Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose the family $\{f_\theta(y) : \theta \in \Theta\}$ has monotone likelihood ratio (MLR) in some statistic $T(Y)$. Then, for any t_0, the test with rejection region $T > t_0$ is uniformly most powerful at level α

- Proof is at the last page of Appendix

Appendix: Optimality for Composite Null (3)

- UMP tests generally do not exist for two-sided alternatives (e.g., $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$)

- **Intuition**: The "best direction to look for evidence" depends on the alternative, and for two-sided alternatives, these directions conflict.
  - When $\theta > \theta_0$, we should reject for large values of the test statistic
  - When $\theta < \theta_0$, we should reject for small values of the test statistic
  - Thus, the most powerful rejection region depends on the true value of θ
  - No single test can be optimal for all $\theta \neq \theta_0$

Appendix: Proof of Karlin-Rubin (1)

- Fix any alternative value $\theta_1 > \theta_0$. By the monotone likelihood ratio assumption in $T(Y)$, the likelihood ratio

$$\frac{f_{\theta_1}(y)}{f_{\theta_0}(y)}$$

  is a nondecreasing function of $T(y)$.

- Therefore, by the Neyman–Pearson lemma, the most powerful level-α test of the simple null

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1$$

  rejects for large values of $T(Y)$; that is, its rejection region is of the form

$$\{T(Y) > c\}$$

  for some cutoff c chosen to attain size α.

Appendix: Proof of Karlin-Rubin (2)

- Importantly, the argument in previous page holds for *every* $\theta_1 > \theta_0$.
- Hence the same class of tests, namely upper-tail tests based on $T(Y)$, is most powerful against every simple alternative in the composite alternative $\Theta_1 = \{\theta : \theta > \theta_0\}$.
- It remains to verify level-α control under the composite null $\Theta_0 = \{\theta : \theta \leq \theta_0\}$. Since the family has MLR in $T(Y)$, the distribution of $T(Y)$ shifts to the right as θ increases.
- Therefore, for any fixed cutoff c,

$$P_\theta(T(Y) > c)$$

  is nondecreasing in θ.

Appendix: Proof of Karlin-Rubin (3)

- Consequently,

$$\sup_{\theta \le \theta_0} P_\theta(T(Y) > c) = P_{\theta_0}(T(Y) > c).$$

- Thus, if we choose $c = t_0$ so that

$$P_{\theta_0}(T(Y) > t_0) = \alpha,$$

  then the test with rejection region $\{T(Y) > t_0\}$ has size α.

- Finally, let ϕ^* be any other test of size at most α. Since ϕ^* also has size at most α when testing the simple null $\theta = \theta_0$, the Neyman–Pearson lemma implies that for every $\theta_1 > \theta_0$,

$$E_{\theta_1}[\phi(Y)] \ge E_{\theta_1}[\phi^*(Y)],$$

  where $\phi(Y) = \mathbf{1}\{T(Y) > t_0\}$.

- Hence ϕ is uniformly most powerful for testing

$$H_0 : \theta \le \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$