STAT 200B 2019 Week06

Soyeon Ahn

1 Definitions

- Null vs. alternative
  - H_0: the null hypothesis
  - H_1: the alternative hypothesis (research hypothesis)
- Simple vs. composite
  - a simple hypothesis $\theta = \theta_0$
  - a composite hypothesis. $\theta > \theta_0$ or $\theta < \theta_0$
- One-sided vs. Two-sided
  - a one-sided test. $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$
  - a two-sided test. $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$
- The forms of hypotheses
  - simple null vs. simple alternative
  - simple null vs. composite two-sided alternative
  - simple null vs. composite one-sided alternative
  - composite one-sided null vs. composite one-sided alternative
  - composite two-sided null vs. composite two-sided alternative
- The power function of a hypothesis test with rejection region R is the function of θ: $\beta(\theta) = P_\theta(X \in R)$
- Type I error and Type II error
  - Type I error: rejecting the null hypothesis when it is indeed true
  - Type II error: failing to reject the null hypothesis when it is false
  - For $\theta \in \Theta_0$, the probability of making a type I error = $\alpha(\theta) = P_\theta(X \in R|\theta \in \Theta_0)$
  - For $\theta \in \Theta_1$, Power = 1 - the probability of making a type II error = $\beta(\theta) = P_\theta(X \in R|\theta \in \Theta_1)$

- Size vs. level
  - A size α test if $\sup_{\theta\in\Theta_0}\beta(\theta)=\alpha$.
  - A level α test if $\sup_{\theta\in\Theta_0}\beta(\theta)\leq\alpha$.
  - Simply speaking, the size is the largest the power function under H_0.
  - Note that the set of level α tests contains the set of size α tests. Sometimes it is often computationally impossible to construct a size α test.
- P-value: the lowest value for the significance level that would result in rejection of the null hypothesis.
- Let $\mathbb{C}$ be a class of tests for testing $H_0:\theta\in\Theta_0$ vs. $H_1:\theta\in\Theta_0^c$. A test in class $\mathbb{C}$, with power function $\beta(\theta)$, is a uniformly most powerful (UMP) class $\mathbb{C}$ test if $\beta(\theta)\geq\beta(\theta)'$ for every $\theta\in\Theta_0^c$ that is a power function of a test in class $\mathbb{C}$

2 Neyman-Pearson lemma: the optimality of Likelihood Ratio Test under simple hypotheses

Likelihood Ratio Test (LRT) $H_0:\theta=\theta_0$ vs. $H_1:\theta=\theta_1$

$$T(X)=\frac{\mathcal{L}_n(\theta_1)}{\mathcal{L}_n(\theta_0)}=\frac{\prod f_{\theta_1}(x_i)}{\prod f_{\theta_0}(x_i)}$$

Decision: reject H_0 if $T(X)>k$

Neyman-Pearson lemma: Let δ be a hypothesis test where $H_0:\theta=\theta_0$ and $H_1:\theta=\theta_1$ are simple hypotheses, where the pdf corresponding to θ_i is $f_{\theta_i}(x)$. Suppose our decision rule is the following:

- $x\in R$ [reject H_0] if $f_{\theta_1}(x)>kf_{\theta_0}(x)$.
- $x\notin R$ [retain H_0] if $f_{\theta_1}(x)<kf_{\theta_0}(x)$.

for some $k\geq 0$. If we choose k so that $P_{\theta_0}(X\in R)=\alpha$ then this test is the most powerful, size α test. That is, among all tests with size α, this test maximizes the power $\beta(\theta_1)$.

proof uniformly most powerful (UMP) level α test.

Note that any test satisfying $P_{\theta_0}(X\in R)=\alpha$ is a size α, and a level α test, because $\sup_{\theta\in\Theta_0}P_\theta(X\in R)=P_{\theta_0}(X\in R)=\alpha$, since Θ_0 has only one point.

We define a test function, an indicator function of the rejection region on the sample space that is 1 if $x\in R$ and 0 if $x\notin R$.

Let $\phi(x)$ be the test function of a test satisfying the above decision rule. Let $\phi^{'}(x)$ be the test function of any other level α test, and let $\beta(\theta)$ and $\beta^{'}(\theta)$ be the power functions corresponding to the tests $\phi(x)$, and $\phi^{'}(x)$, respectively. We can show that

$$(\phi(x) - \phi^{'}(x))(f_{\theta_1}(x) - kf_{\theta_0}(x)) \geq 0$$

for every x. [Why? When $\phi = 1$, $(\phi(x) - \phi^{'}(x)) \geq 0$ since $\phi^{'}(x) \leq 1$, test ϕ rejects H_0 if $f_{\theta_1}(x) > kf_{\theta_0}(x)$. When $\phi = 0$, $(\phi(x) - \phi^{'}(x)) \leq 0$ and $f_{\theta_1}(x) < kf_{\theta_0}(x)$).] Thus,

$$0 \leq \int [\phi(x) - \phi^{'}(x)][f_{\theta_1}(x) - kf_{\theta_0}(x)]dx = \beta(\theta_1) - \beta^{'}(\theta_1) - k[\beta(\theta_0) - \beta^{'}(\theta_0)]$$

Since $\phi^{'}$ is a level α test and ϕ is a size α test, $\beta(\theta_0) - \beta^{'}(\theta_0) = \alpha - \beta^{'}(\theta_0) \geq 0$. Since $k \geq 0$,

$$0 \leq \beta(\theta_1) - \beta^{'}(\theta_1) - k[\beta(\theta_0) - \beta^{'}(\theta_0)] \leq \beta(\theta_1) - \beta^{'}(\theta_1)$$

showing that $\beta(\theta_1) \geq \beta^{'}(\theta_1)$. ϕ has greater power than $\phi^{'}$. Since $\phi^{'}$ was an arbitrary level α test and θ_1 is the only point in Θ_1, ϕ is a uniformly most powerful (UMP) level α test.

proof every uniformly most powerful (UMP) level α test is a size α test

Let $\phi^{'}$ now be the test function for any UMP level α test. $\phi(x)$ is also a UMP level α test, therefore, $\beta(\theta_1) = \beta^{'}(\theta_1)$. This fact and $k \geq 0$ imply that

$$\alpha - \beta^{'}(\theta_0) = \beta(\theta_0) - \beta^{'}(\theta_0) \leq 0.$$

Since $\phi^{'}$ is a level α test, $\alpha \leq \beta^{'}(\theta_0)$. Thus, $\beta^{'}(\theta_0) = \alpha$, that is, $\phi^{'}$ is a size α test.

The Neyman-Pearson Lemma shows that α and β cannot both be arbitrarily small. It showed the optimality of the likelihood ratio test; the most powerful test (a test that minimize the probability of type II error) among tests with size smaller than or equal α is the size α likelihood ratio test.

Example (z-test) Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. For each of the following cases, derive the power function of the size 0.05 LRT. $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$

$$T(X) = \frac{\mathcal{L}_n(\theta_1)}{\mathcal{L}_n(\theta_0)} = \exp\left(\frac{(\mu_1 - \mu_0)n\bar{X}}{\sigma^2} + \frac{n(\mu_0^2 - \mu_1^2)}{2\sigma^2}\right)$$

It is an increasing function of $\bar{X}$, thus for any k

$$T(X) > k \text{ is equivalent to } \bar{X} > d.$$

We reject H_0 if $\bar{X} > d$ where d is a value satisfying $P_{\theta_0}(\bar{X} > d) = \alpha$. Under H_0, $\bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$ and therefore the size α test rejects H_0 if

$$z = \frac{n(\bar{x} - \mu_0)}{\sigma} > z_\alpha$$

3 P-values

In the above z test example, LRT rejects if $z > k$ for some constant k. In other words, the size of the test is $\alpha = P(Z > k|H_0) = 1 - \Phi(k)$, and is decreasing as k increases. If our observed values z in the rejection region, then

$$z > k \text{ is equivalent to } \alpha > p_{obs} = P(Z > z|H_0)$$

The quantity p_{obs} is called the p-value of the observed data z. In general, the p-value is sometimes called the observed significance level of x and is the probability under H_0 of seeing data that are more extreme than our observed data x.

In the lecture note, we defined p-value as the smallest level at which we can reject H_0. Suppose that for every $\alpha \in (0, 1)$ we have a size α test with rejection region R_α. Then When $R_\alpha = \{x : T(x) \geq c_\alpha\}$,

$$\text{p-value} = \sup_{\theta \in \Theta_0} P_\theta(T(X) \geq T(x))$$

where x is the observed data.

Therefore, the p-value is the probability under H_0 of observing a value $T(X)$ the same as or more extreme than what was actually observed. Extreme observations are viewed as providing evidence againt H_0.

Theorem If the test statistic has a continuous distribution, then under $H_0 : \theta = \theta_0$, the p-value has a $Unif(0, 1)$ distribution. Therefore, if we reject H_0 when the p-value is less than α, the probability of a Type I error is α.

For a z-test,

$$P(p_{obs} < p|H_0) = P(1 - \Phi(Z) < p|H_0) = P(Z > \Phi^{-1}(1 - p)|H_0)$$

$$= 1 - \Phi(\Phi^{-1}(1 - p)) = 1 - (1 - p) = p$$

For a general case, under H_0, $p(x) = P(T(X) > T(x))$ since H_0 is simple, therfore, $p(x) = 1 - F(t)$.

$$P(p(X) \leq z) = P(1 - F_T(T(X)) \leq z) = P(F_T(T(X)) \geq 1 - z) =$$

$$= P(U \geq 1 - z) = 1 - P(U \leq 1 - z)) = 1 - (1 - z) = z$$

Since $P(p(X) \leq z) = z$, $p(X)$ is uniformly distributed, that is, Uniform(0,1).

4 Composite null and alternative hypotheses

4.1 Composite alternative

For composite hypotheses, the Type I error or Type II error probabilities do not have a single value. Thus the size of the test is defined to take a supreme. The Neyman-

Pearson theory can be extended to one-sided alternatives. For example, the above z-test example, we showed that the most powerful size α test for $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ is

$$z = \frac{n(\bar{x} - \mu_0)}{\sigma} > z_\alpha.$$

This critical value does not depend upon μ_1. Therefore, this test can be uniformly most powerful size α for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$.

In general, it is rare to have an UMP test for the composite alternative (Casella and Berger Example 8.3.19). However, for one-sided alternative, UMP exists under some conditions.

4.2 Composite null: Monotone likelihood ratio (MLR) and Uniformly most powerful test (UMP)

definition A family of pdf (or pmf) $\{g(t;\theta) : \theta \in \Theta\}$ for a univariate random variable with real-valued parameter θ has a monotome likelihood ratio (MLR) in $T(X)$ if for every $\theta_2 > \theta_1$, $g(t;\theta_2)/g(t;\theta_1)$ depends on values of the data X only through the value of statistic $T(X)$ and the ratio is a monotone (nonincreasing or nondecreasing) function of $T(X)$. Note that $c/0$ is defined as ∞ if $0 < c$.

Note that any regular exponental family with $g(t;\theta) = h(t)c(\theta)\exp^{w(\theta)t}$ has an MLR if $w(\theta)$ is a nondecreasing function.

Theorem (Karlin-Rubin) Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose that T is a sufficient statistic for θ and the family of pdfs (or pmfs) $\{g(t;\theta) : \theta \in \Theta\}$ of T has a MLR. Then for any t_0, the test rejects H_0 if and only if $T > t_0$ is a UMP level α test, where $\alpha = P_{\theta_0}(T > t_0)$ (Casella and Berger Theorem 8.3.17)

proof: Let $\beta(\theta) = P_\theta(T > t_0)$ be the power function of the test. Fix $\theta^{'} > \theta_0$ and consider testing $H_0^{'} : \theta = \theta_0$ vs. $H_1^{'} : \theta = \theta_1$.

- $\sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$ and this is a level α test.

  For this, we will show that $\beta(\theta)$ is nondecreasing. Since the family of pdfs or pmfs of T has an MLR, so For $\theta_2 > \theta_1$,

  $$\frac{d}{dt_0}[\beta(\theta_1) - \beta(\theta_2)] = -g(t_0;\theta_1) + g(t_0;\theta_2) = g(t_0;\theta_1)(\frac{g(t_0;\theta_2)}{g(t_0;\theta_1)} - 1)$$

  Because g has an MLR, the ratio on the right-hand side is nondecreasing, so the derivative can only change sign from negative to positive showing that any interior extremum is a minimum. $\beta(\theta_1) - \beta(\theta_2)$ is maximized by its value at ∞ or $-\infty$, i.e., zero.

- If we define

$$k^{'} = \inf_{t \in \mathcal{T}} \frac{g(t;\theta^{'})}{g(t;\theta_0)}$$

where $\mathcal{T} = \{t : t > t_0 \text{ and either } g(t;\theta^{'}) > 0 \text{ or } g(t;\theta_0) > 0\}$, it follows that

$$T > t_0 \iff \frac{g(t;\theta^{'})}{g(t;\theta_0)} > k^{'}$$

It implies $\beta(\theta') \geq \beta^*(\theta')$, where $\beta^*(\theta)$ is the power function for any other level α test of $H_0^{'}$, that is, any test satisfying $\beta(\theta_0) \leq \alpha$ (NP lemma). Howerver, any level α test of H_0 satisfies $\beta^*(\theta_0) \leq \sup_{\theta \in \Theta_0} \beta^*(\theta) \leq \alpha$. Thus, $\beta(\theta') \geq \beta^*(\theta')$ for any level α test of H_0.

Fix $\theta^{'} > \theta_0$ and consider testing $H_0^{'} : \theta = \theta_0$ vs $H_1^{'} : \theta = \theta^{'}$. Since the faimily of pdfs (or pmfs) of $T(X)$ has an MLR, $\beta(\theta)$ is nondecreasing. Thus, $\sup_{\theta \geq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$ and this is a level α test. If we define

$$k^{'} = \inf_{x \in \mathcal{X}} \frac{g(x;\theta^{'})}{g(x;\theta_0)}$$

where $\mathcal{X} = \{x : t(x) > t_0 \text{ and either } g(x;\theta^{'}) > 0 \text{ or } g(x;\theta_0) > 0\}$, it follows that

$$T(x) > t_0 \iff \frac{g(x;\theta^{'})}{f(x;\theta_0)} > k^{'}$$

It implies $\beta(\theta') \geq \beta^*(\theta')$, where $\beta^*(\theta)$ is the power function for any other level α test of $H_0^{'}$, that is, any test satisfying $\beta(\theta_0) \leq \alpha$ (NP lemma). Howerver, any level α test of H_0 satisfies $\beta^*(\theta_0) \leq \sup_{\theta \in \Theta_0} \beta^*(\theta) \leq \alpha$. Thus, $\beta(\theta') \geq \beta^*(\theta')$ for any level α test of H_0.

Example: the power function of the normal distribution

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. For each of the following cases, derive the power function of the size 0.05 LRT.

1. $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$
2. $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$

1. $\Theta = \mathbb{R}$, $\Theta_0 = (-\infty, 0]$. The likelihood ratio statistic is

$$\lambda = 2 \log \left(\frac{\mathcal{L}_n(\hat{\mu})}{\sup_{\mu \in \Theta_0} \mathcal{L}_n(\mu)} \right),$$

where $\hat{\mu}$ is the MLE on Θ.

$\hat{\mu} = \bar{X}_n$.

- If $\bar{X}_n \leq 0$, then $\sup_{\mu\in\Theta_0} \mathcal{L}_n(\mu) = \mathcal{L}_n(\hat{\mu})$.

$$\lambda = 2\log\left(\frac{\mathcal{L}_n(\hat{\mu})}{\mathcal{L}_n(\hat{\mu})}\right) = 0.$$

- If $\bar{X}_n > 0$, then

$$\ell_n(\mu) = constant - n\log\sigma - \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2}.$$

$$\Rightarrow \frac{\partial \ell_n}{\partial \mu}(\mu) = \frac{\sum_{i=1}^n (X_i-\mu)}{\sigma^2} = \frac{n(\bar{X}_n-\mu)}{\sigma^2} > 0, \quad \text{for } \mu \in \Theta_0.$$

So $\mathcal{L}_n(\mu)$ is an increasing function of μ for $\mu \in (-\infty, 0]$, and $\sup_{\mu\in\Theta_0} \mathcal{L}_n(\mu) = \mathcal{L}_n(0)$.

$$\lambda = 2\log\left(\frac{\mathcal{L}_n(\hat{\mu})}{\mathcal{L}_n(0)}\right) = 2(\ell_n(\hat{\mu}) - \ell_n(0)) = -\frac{\sum_{i=1}^n (X_i-\bar{X}_n)^2}{2\sigma^2} + \frac{\sum_{i=1}^n X_i^2}{2\sigma^2} = \frac{n(\bar{X}_n)^2}{2\sigma^2} \geq 0.$$

Since $\lambda \geq 0$, the rejection region should be $\lambda > c$, with some $c > 0$ to be determined by the size $\alpha = 0.05$. We can also see from the above that this rejection region is equivalent to $\bar{X}_n > d$, with $d > 0$ to be determined by the size $\alpha = 0.05$.

Since the test has size 0.05,

$$\begin{aligned}
0.05 &= \sup_{\mu\in\Theta_0} P_\mu(\bar{X}_n > d) \\
&\text{(Since } \bar{X}_n \sim N(\mu, \sigma^2/n),) \\
&= \sup_{\mu\in\Theta_0} P_\mu\left(\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} > \frac{\sqrt{n}(d-\mu)}{\sigma}\right) = \sup_{\mu\in\Theta_0}\left[1 - \Phi\left(\frac{\sqrt{n}(d-\mu)}{\sigma}\right)\right] \\
&= 1 - \Phi(\sqrt{n}d/\sigma).
\end{aligned}$$

$$\Rightarrow d = \frac{\Phi^{-1}(0.95)\sigma}{\sqrt{n}} = \frac{z_{0.05}\sigma}{\sqrt{n}}.$$

The power function of the test is

$$\begin{aligned}
\beta(\mu) = P_\mu(\bar{X}_n > d) &= P_\mu\left(\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} > \frac{\sqrt{n}(d-\mu)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(d-\mu)}{\sigma}\right) \\
&= 1 - \Phi\left(\Phi^{-1}(0.95) - \frac{\sqrt{n}\mu}{\sigma}\right) = 1 - \Phi\left(z_{0.05} - \frac{\sqrt{n}\mu}{\sigma}\right).
\end{aligned}$$

2. $\Theta = \mathbb{R}$, $\Theta_0 = \{0\}$. The likelihood ratio statistic is

$$\lambda = 2\log\left(\frac{\mathcal{L}_n(\hat{\mu})}{\sup_{\mu\in\Theta_0}\mathcal{L}_n(\mu)}\right) = 2\log\left(\frac{\mathcal{L}_n(\hat{\mu})}{\mathcal{L}_n(0)}\right) = 2(\ell_n(\hat{\mu}) - \ell_n(0)),$$

where $\hat{\mu}$ is the MLE on Θ.

From (a),

$$\lambda = 2(\ell_n(\hat{\mu}) - \ell_n(0)) = \frac{n(\bar{X}_n)^2}{2\sigma^2} \geq 0.$$

The rejection region is $\lambda > c$, with $c > 0$ to be determined by the level $\alpha = 0.05$. This is equivalent to $|\bar{X}_n| > d$, with $d > 0$ to be determined by $\alpha = 0.05$.

$$\begin{aligned}
0.05 &= P_{\mu=0}(|\bar{X}_n| > d) \\
&\quad \text{When } \mu = 0, \bar{X}_n \sim N(0, \sigma^2/n) \\
&= P_{\mu=0}\left(\sqrt{n}\bar{X}_n/\sigma > \sqrt{n}d/\sigma\right) + P_{\mu=0}\left(\sqrt{n}\bar{X}_n/\sigma < -\sqrt{n}d/\sigma\right) \\
&= 1 - \Phi(\sqrt{n}d/\sigma) + \Phi(-\sqrt{n}d/\sigma) = 2(1 - \Phi(\sqrt{n}d/\sigma)).
\end{aligned}$$

$$\Rightarrow d = \frac{\Phi^{-1}(0.975)\sigma}{\sqrt{n}} = \frac{z_{0.025}\sigma}{\sqrt{n}}.$$

The power function of the test is

$$\begin{aligned}
\beta(\mu) &= P_\mu\left(|\bar{X}_n| > d\right) \\
&= P_\mu\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} > \frac{\sqrt{n}(d-\mu)}{\sigma}\right) + P_\mu\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < -\frac{\sqrt{n}(d-\mu)}{\sigma}\right) \\
&= 1 - \Phi\left(\frac{\sqrt{n}(d-\mu)}{\sigma}\right) + \Phi\left(-\frac{\sqrt{n}(d-\mu)}{\sigma}\right) \\
&= 1 - \Phi\left(\Phi^{-1}(0.975) - \frac{\sqrt{n}\mu}{\sigma}\right) + \Phi\left(-\Phi^{-1}(0.975) - \frac{\sqrt{n}\mu}{\sigma}\right) \\
&= 1 - \Phi\left(z_{0.025} - \frac{\sqrt{n}\mu}{\sigma}\right) + \Phi\left(-z_{0.025} - \frac{\sqrt{n}\mu}{\sigma}\right).
\end{aligned}$$

5 likelihood ratio test (LRT)

$$T(X) = \frac{\sup_{\theta\in\Theta} \mathcal{L}_n(\theta)}{\sup_{\theta\in\Theta_0} \mathcal{L}_n(\theta)}$$

If $\hat{\theta}_n$ is the MLE and $\hat{\theta}_{n,0}$ is the MLE restricting $\theta \in \Theta_0$, then

$$T(X) = \frac{\mathcal{L}_n(\hat{\theta}_n)}{\mathcal{L}_n(\hat{\theta}_{n,0})}$$

Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

Theorem When the power function can not be calculated exactly, and Θ_0 consists of fixing certain elements of θ (e.g., as in a point-null hypothesis), we can use the limiting distribution

$$\lambda(X) = 2\log T(X) \xrightarrow{D} \chi^2_{r-q}$$

where r is the dimension of θ and q is the number of restricted elements.

We assume that there exist an MLE $\hat{\theta}$ and $\ell(\theta) = \log L(\theta|x)$. There are some θ^* bewteen θ_0 and $\hat{\theta}$ such that

$$\begin{aligned} 2\log T(X) &= -2(\ell(\theta_0) - \ell(\hat{\theta})) \\ &\approx -2\left(\frac{\partial \ell(\hat{\theta})}{\partial \theta}(\theta_0 - \hat{\theta}) + \frac{1}{2}\frac{\partial^2 \ell(\theta^*)}{\partial \theta^2}(\theta_0 - \hat{\theta})^2\right) \\ &= -\frac{\partial^2 \ell(\theta^*)}{\partial \theta^2}(\theta_0 - \hat{\theta})^2 \end{aligned}$$

Since $\hat{\theta} \xrightarrow{P} \theta_0$, thus $\theta^* \xrightarrow{P} \theta_0$ under the null hypothesis. By the uniform law of large numbers,

$$-\frac{1}{n}\frac{\partial^2 \ell(\theta^*)}{\partial \theta^2} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i|\theta)}{\partial \theta^2} \xrightarrow{P} I(\theta_0)$$

where $I(\theta) = -E_\theta\left[\frac{\partial^2}{\partial \theta^2}\log f(X_i;\theta)\right]$

By the Slutsky theorem,

$$-\frac{1}{nI(\theta_0)}\frac{\partial^2 \ell(\theta^*)}{\partial \theta^2} \xrightarrow{P} 1$$

In addition,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0)).$$

By continous mapping,

$$I(\theta_0)n(\hat{\theta} - \theta_0)^2 \xrightarrow{D} \chi_1^2$$

By the Slutsky theorem,

$$2\log T(X) = \frac{1}{nI(\theta_0)}\frac{\partial^2 \ell(\theta^*)}{\partial \theta^2} I(\theta_0)n(\hat{\theta} - \theta_0)^2 \xrightarrow{D} \chi_1^2$$

Theorem Suppose that $\theta = (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots \theta_r)$. Let

$$\Theta_0 = \{\theta : (\theta_{q+1}, \ldots \theta_r) = \theta_{0,q+1}, \ldots \theta_{0,r})\}.$$

Let λ be the likelihood ratio test statistic. Under $H_0 : \theta \in \Theta_0$,

$$\lambda(x^n) \xrightarrow{D} \chi_{r-q}^2$$

where $r - q$ is the dimension of Θ minus the dimension of Θ_0.

The twice of the LRT is asympototically distributed as χ_{r-q}^2 ($r - q$ is the number of restrictions in the null hypothesis). The LRT requires us to maximize the log-likelihood and also to evaluate it at θ.

6 The Wald test

Assume that $\hat{\theta}$ is asymptotically Normal:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0))$$

The size α Wald test is: reject H_0 when $W > z_{\alpha/2}$ where

$$W = \frac{\hat{\theta} - \theta_0}{\hat{se}}$$

The idea behind the Wald test is to replace the log-likielhood by a quadratic approximation at $\hat{\theta}$.

$$\ell(\theta) \approx \ell(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T H(\hat{\theta})(\theta - \hat{\theta})$$

where H stands for the Hessian of the likilihood function. Since $\hat{\theta}$ is the MLE, the first derivative will be zero. We approximate $-H$ by $I(\hat{\theta})$, the Fisher information at the MLE.

7 The score test

The score test is to replace the log-likielhood by a quadratic approximation which agrees with L and its first two derivatives at θ_0.

$$\ell(\theta) \approx \ell(\theta_0) + (\theta - \theta_0)^T l^{'}(\theta_0) + \frac{1}{2}(\theta - \theta_0)^T H(\theta_0)(\theta - \theta_0)$$

Please note that $l^{'}(\theta) = u(\theta)$ is called score and again $l^{''}(\theta) = I(\theta)$ is the expected information.

The score test is based on the following test statistics

$$R = \frac{\ell^{'}(\theta_0)}{\sqrt{I_n(\theta_0)}}$$

8 Graphical representation

the Holy Trinity; d works on the log-likelihood scale, w on the parameter scale, and s on the first derivative scale.

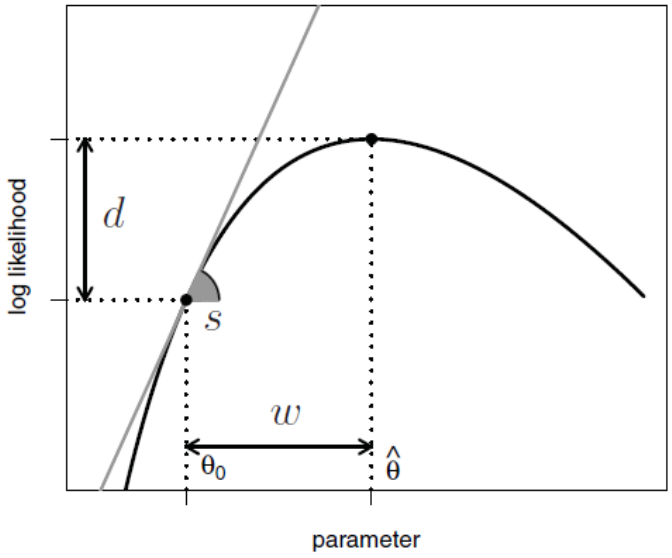

Figure 1. Comparing the three test statistics according to the traditional plot: Likelihood ratio is reported on the y scale, Wald on the x scale, and the score on the first derivative scale. The different scales do not favor understanding of the underlying connections.

(Figure 1. from V. Muggeo, G. Lovison, The three plus one likelihood-based test statistics: unified geometrical and graphical interpretations, Am. Stat. 68 (2014) 302-306.)

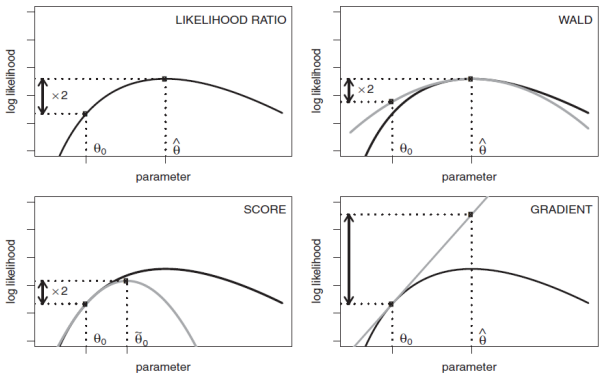

Figure 2. Comparing the four test statistics on the log-likelihood scale. On each plot the log-likelihood is illustrated (black line) along with the relevant approximation underlying the test statistic: in the Wald panel the gray line is $\mathcal{P}_w(\theta)$, in the score panel it is $\mathcal{P}_s(\theta)$, and in the gradient panel it is $\mathcal{P}_g(\theta)$. The arrows on the left side quantify the corresponding observed test statistic; the longer the arrow, the larger the evidence against H_0. Notice that for the likelihood ratio, Wald, and score, the arrow lengths have to be doubled to obtain the actual values comparable to those from the gradient statistic.

(Figure 2. from V. Muggeo, G. Lovison, The three plus one likelihood-based test statistics: unified geometrical and graphical interpretations, Am. Stat. 68 (2014) 302-306.)

When n is large, the LR method will tend to produce intervals very similar to those based on the observed or expected information. Unlike the information-based intervals, however, the LR intervals are scale-invariant. That is, if we find the LR interval for a transformed version of the parameter such as $\phi = \log p/(1-p)$ and then transform the endpoints back to the p-scale, we get exactly the same answer as if we apply the LR method directly on the p-scale. For that reason, the LR method is preferred.

If the loglikelihood function expressed on a particular scale is nearly quadratic, then a information-based interval calculated on that scale will agree closely with the LR interval. Therefore, if the information-based interval agrees with the LR interval, that provides some evidence that the normal approximation is working well on that particular scale. If the information-based interval is quite different from the LR interval, the appropriateness of the normal approximation is doubtful, and the LR approximation is probably better.

9 Example: Binomial distribution

We would like to test whether the response rate is p_0 or not.

$$H_0 : p = p_0 \text{ versus } H_1 : p \neq p_0$$

where $X_i \sim Bernoulli(p)$

The expected Fisher's information $I(p)$,

$$I(p) = -E\left(\frac{\partial^2 \ell(p)}{\partial p^2}\right) = -E\left(-\frac{\sum_i X_i}{p^2} - \frac{n - \sum_i X_i}{(1-p)^2}\right) = \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)}$$

- We know that
$$\hat{p} = \sum_i X_i/n$$
and
$$\hat{p} - p_0 \xrightarrow{D} N\left(0, \frac{p_0(1-p_0)}{n}\right)$$
The Wald statistic is
$$\frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}}$$
- The score statistic is
$$\frac{\hat{p} - p_0}{\sqrt{\hat{p_0}(1-\hat{p_0})/n}}$$
- The LRT is
$$2\left(Y \log \frac{\hat{p}}{p_0} + (n-Y)\log \frac{1-\hat{p}}{1-p_0}\right)$$
where $Y = \sum_i X_i$

We would like to show the asympototic property of the log likehood ratio test. Use $\hat{p} = \frac{Y}{n}$ and $log(1-x) \approx -x$ with small x. Ignore the constant term, then

$$\begin{aligned}
\left(n\hat{p}\log\frac{\hat{p}}{p_0} + n(1-\hat{p})\log\frac{1-\hat{p}}{1-p_0}\right) &= n\left(\hat{p}\log\frac{p_0-(p_0-\hat{p})}{p_0} + (1-\hat{p})\log\frac{1-p_0-(\hat{p}-p_0)}{1-p_0}\right) \\
&\approx 2n\left(\hat{p}\frac{(\hat{p}-p_0)}{p_0} - (1-\hat{p})\frac{(\hat{p}-p_0)}{1-p_0}\right) \\
&= n(\hat{p}-p_0)\left(\frac{\hat{p}}{p_0} - \frac{1-\hat{p}}{1-p_0}\right) \\
&= n(\hat{p}-p_0)\frac{(\hat{p}-p_0)}{p_0(1-p_0)} = \frac{(\hat{p}-p_0)^2}{p_0(1-p_0)/n}
\end{aligned}$$

It is approximately the square of a standard normal.

$$\frac{(\hat{p}-p_0)^2}{p_0(1-p_0)/n}$$

. The Wald statsitic is asymptotically equivalent to the score statistic.

[Agresti, Categorical data analysis, 2013] The Wald statistic uses the behavior of $L(\theta)$ at the MLE $\hat{\theta}$, having chi-squared form $(\hat{\beta}/SE)^2$. The SE of $\hat{\theta}$ depends on the curvature of $L(\theta)$ at $\hat{\theta}$. The score test is based on the slope and curvature of $L(\theta)$ at $\theta = 0$. The likelihood-ratio test combines information about $L(\theta)$ at both $\hat{\theta}$ and θ_0.

The three tests asymptotic equivalence as $n \to \infty$ (Cox and Hinkely, 1974 Sec. 9.3). From small to moderate sample sizes, the likelihood-ratio and score tests are usually more reliable than the Wald test, having actual error closer to the nominal test.