Statistics - Hypothesis Testing and Confidence Intervals

Andrew Andreas

June 14, 2024

1 Introduction

This is a set of brief notes on Hypothesis Testing and Confidence Intervals in statistics and are part of a series of notes on select statistical ideas. These may evolve over time as I look to expand their depth.

Please report any errors to andrew.andreas0@gmail.com

2 Motivation

Recall that $\hat{\theta}$ is the estimator which is a function of the observed data sample.

$$\hat{\theta} = t(x_1, x_2, ..., x_n) \tag{1}$$

Having got an estimate for a particular parameter θ in the form of $\hat{\theta}$, we want to understand how good our estimator is relative to θ.

3 Hypotheses

Hypothesis are statements of beliefs about the particular value a parameter θ can take from the set of possible values Ω. Formally, Ω is the parameter space and a hypothesis H regarding a parameter θ states that $\theta \in \omega$, where $\omega \subset \Omega$.

Definition. Simple Hypothesis A simple hypothesis is one which asserts that the parameter θ takes on a single, specific value

Definition. Composite Hypothesis A composite hypothesis is one which asserts that the parameter θ can take on a range of values ω, i.e., it is not simple.

Definition. Null Hypothesis, H_0 The null hypothesis H_0 is the hypothesis that the claim being investigated does not exist. It is denoted $H_0 : \theta \in \omega$

Definition. Alternate Hypothesis, H_0 The alternate hypothesis H_1 is the **complement** to the null hypothesis $H_1 : \theta \in \Omega - \omega = \theta \notin \omega$. It is the hypothesis that the **Null Hypothesis** is **False**

While we have two competing hypothesis, it is necessary that we have a method for selecting between them according to the data observed. This is achieved through the use of a **rejection region** $R \subseteq \mathcal{R}^n$

Definition. Rejection Region, R The rejection region is defined as a set of values $R = \{x : t(\mathbf{x}) > c\}$, where $t(\mathbf{x})$ is a test statistic and c some defined critical value

To define an test, we need to consider the type of error that can occur so that we may objectively assess its performance.

- **Type-I Error:** A type-I error occurs when we reject H_0 when H_0 is true. The probability of a type-I error is called the **size** of the test and denoted α

- **Type-II Error:** A type-II error occurs when we fail to reject H_0 when H_0 is false. The probability of a type-II error is denoted β and $1-\beta$ is called the **power** of the test

Example 1. Consider a random variable $X \sim N(\theta, 1)$, then a possible test statistic is the sample mean $\bar{X}$. Suppose we want to test the null hypothesis $H_0 : \theta = 0$, such that $\bar{X} \sim N(0, 1/n)$ and $H_1 : \theta = 1$, such that $\bar{X} \sim N(1, 1/n)$.

Then under H_0, α is defined as

$$\begin{aligned}
\alpha &= P\big(t(\mathbf{x}) > c \mid H_0\big) && (2)\\
&= P\big(\bar{X} > c \mid H_0\big) && (3)\\
&= P\big(\sqrt{n}(\bar{X} - 0) > \sqrt{n}(c - 0)\big) && (4)\\
&= P\big(Z > \sqrt{n}(c)\big) && (5)\\
&= 1 - P\big(Z \leq \sqrt{n}(c)\big) && (6)\\
&= 1 - \Phi\big(\sqrt{n}(c)\big) && (7)\\
& && (8)
\end{aligned}$$

Where $Z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2}}$ and Φ is the standard normal cdf. Similarly, we can consider the power $1-\beta$ such that under H_1. First, consider β

$$\begin{aligned}
\beta &= P\big(t(\mathbf{x}) \leq c \mid H_1\big) && (9)\\
&= P\big(\bar{X} \leq c \mid H_1\big) && (10)\\
&= P\big(\sqrt{n}(\bar{X} - 1) \leq \sqrt{n}(c - 1)\big) && (11)\\
&= P\big(Z \leq \sqrt{n}(c - 1)\big) && (12)\\
&= \phi\big(\sqrt{n}(c - 1)\big) && (13)
\end{aligned}$$

Then $1 - \beta = 1 - \phi\big(\sqrt{n}(c-1)\big)$.

The above example illustrates the relationship between the **size** of a test and the **power**. While it is desirable to make α as small as possible and $1-\beta$ as large as possible, we see that for a fixed sample with n observations, if we want to make α small then we must make the critical value c larger but a larger c causes $1-\beta$ to decrease.

In practice, we usually compare tests of fixed **size** and comapre the **power** of the competing tests. A test whose **power** is at least as good as any other test for fixed α is called the **most powerful** test.

4 Testing Hypothesis

4.1 The Likelihood Ratio Test

Definition. The Likelihood Ratio Test utilizes the likelihood ratio as a test statistic, that is the ratio of the likelihoods under H_0 and under the entire parameter set Ω. Suppose $H_0 : \theta \in \omega$ and $H_1 : \theta in \Omega - \omega$ are the null and alternative hypothesis, then the likelihood ratio is

$$\text{LR} = \frac{\max_{\theta \in \Omega} L(\theta)}{\max_{\theta \in \omega} L(\theta)} \tag{14}$$

with a rejection region of the form $R = \{\mathbf{x} : LR > c\}, c \geq 1$.

The numerator of the LR is the MLE, $L(\hat{\theta})$, as this it the value of θ that maximises the likelihood over the entire parameter space, Ω. The denominator is the maximum value of the likelihood over the H_0.

The LR takes values ≥ 1 and its lower bound at 1 is a result of the case where $th\hat{e}ta$ that maximises $L(\theta)$ is contained within the null hypothesis.

The logarithm of the LR is given by

$$\log(\text{LR}) = \log\left(\frac{\max_{\theta\in\Omega} L(\theta)}{\max_{\theta\in\omega} L(\theta)}\right) \tag{15}$$

$$= \log\big(\max_{\theta\in\Omega} L(\theta)\big) - \log\big(\max_{\theta\in\omega} L(\theta)\big) \tag{16}$$

$$= l(\hat{\theta}) - \log\big(\max_{\theta\in\omega} L(\theta)\big) \tag{17}$$

In the case of simple hypothesis, the LR becomes

$$\max\left[\frac{\max_{\theta\in\Omega} L(\theta)}{\max_{\theta\in\omega} L(\theta)}, 1\right] \tag{18}$$

We have seen that the likelihood ratio provides us with a test statistic and in most cases it is equivalent to the results we see from **standard tests** such as the **one-sample t-test**. Please see the proof of this in the appendix.

4.2 Critical Values

A hypothesis test requires a test statistic and a critical value to create a valid rejection region. This far we have focused on the test statistic, specifically the LR ratio.

4.2.1 Fixed Level Tests

It is now time to consider how we select the value of the critical value d. The critical value is the value such that

$$\mathrm{P}\big(\log(\text{LR}) > d \mid H_0\big) = \alpha \tag{19}$$

$$\tag{20}$$

Then for a fixed size α, then d is the $(1-\alpha)$ quantile of the distribution of log(LR) given that the null hypothesis is true. This is called the **null distribution** of the log-likelihood ratio statistic.

4.2.2 P-Values

As an alternative to fixed-level tests, we can consider the idea of a **p-value** which based on the observed sample data, considers the probability of observing data at least as extreme given the null hypothesis is true.

In the general case for a test statistic $t(X)$ (note here X represents for a a random variable and denotes the general case) and its observed value $t(\mathbf{x})$ then the **p-value** is defined as

$$p = \mathrm{P}\big(t(X) \geq t(\mathbf{x}) \mid H_0\big) \tag{21}$$

$$\tag{22}$$

4.2.3 Asymptotic Null Distribution

The **null distribution** is the distribution of the test statistic, in this case the log(LR), under the null hypothesis. In most cases there will not be a closed form solution that can be easily obtained. One method to find an approximation is to simulate it, though that will not be explored for the moment.

What we can do is exploit the asymptotic behaviour of the log(LR) test statistic, that is as the number of random variables grows larger, then the null distribution of $2\log(\text{LR})$ is increasingly well approximated by the $\chi^2(d)$ where $d = d_1 - d_0$, the delta between the free parameters under the alternate and null hypothesis.

$$2\log(\text{LR}) \approx \chi^2(d) \tag{23}$$

Then to obtain a test of size $\approx \alpha$

$$\begin{aligned}
\alpha &= \mathrm{P}\big(\mathrm{LR}) > c \mid H_0\big) && (24)\\
&= \mathrm{P}\big(2\log(\mathrm{LR}) > 2\log(c) \mid H_0\big) && (25)\\
&\simeq \mathrm{P}\big(Y > 2\log(c)\big) && (26)\\
& && (27)
\end{aligned}$$

Where $Y \sim \chi^2(d)$. In order to then determine the value of $k = 2\log(c)$ for a given α, then $k = 2\log(c) = \chi^2_{1-\alpha}(d)$, where $\chi^2_{1-\alpha}(d)$ denotes the $(1-\alpha)$ quantile of the $\chi^2(d)$ distribution.

5 Other Statistical Tests

The likelihood ratio test is a general general statistic which benefits from its dependence on the MLE and thus exhibits useful properties. There are useful approximations of the likelihood ratio test which are equivalent as the sample size grows large.

5.1 Wald Tests

There are two versions of the Wald test but both arise from the expansion of the log-likelihood function at its null distribution around the MLE. Since the MLE is a consistent estimator of $\theta = \theta_0$ under H_0, then $\hat{\theta}$ and θ are close in large samples.

Therefore the Taylor series expansion is appropriate. Expanding $l(\theta_0)$

$$\begin{aligned}
l(\theta_0) =& l(\hat{\theta}) + (\theta_0 - \hat{\theta})\frac{dl(\theta)}{d\theta}\Big|_{\hat{\theta}} + \frac{1}{2}(\theta_0 - \hat{\theta})^2\frac{d^2l(\theta)}{d\theta^2}\Big|_{\hat{\theta}} + ... && (28)\\
\approx& l(\hat{\theta}) + (\theta_0 - \hat{\theta})\frac{dl(\theta)}{d\theta}\Big|_{\hat{\theta}} + \frac{1}{2}(\theta_0 - \hat{\theta})^2\frac{d^2l(\theta)}{d\theta^2}\Big|_{\hat{\theta}} && (29)
\end{aligned}$$

after ignoring terms of order greater than two.

The second term of the expansion is the derivative of the log-likelihood, evaluated at the MLE, $\hat{\theta}$, which by definition is zero. So we can omit this from the expression

$$l(\theta_0) \approx l(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^2\frac{d^2l(\theta)}{d\theta^2}\Big|_{\hat{\theta}} \tag{30}$$

Then by the law of large numbers, the second derivative is well approximated by its expected value

$$l(\theta_0) \approx l(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^2\mathbb{E}\left[\frac{d^2l(\theta)}{d\theta^2}\right]\Big|_{\hat{\theta}} \tag{31}$$

Then given this expression for $l(\theta_0)$, then we can approximate 2log(LR)

$$\begin{aligned}
2log(LR) =& 2(l(\hat{\theta}) - l(\theta_0)) && (32)\\
\approx& 2(l(\hat{\theta}) - \big(l(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^2\mathbb{E}\left[\frac{d^2l(\theta)}{d\theta^2}\right]\Big|_{\hat{\theta}}\big) && (33)\\
=& 2 \times -\frac{1}{2}(\theta_0 - \hat{\theta})^2\mathbb{E}\left[\frac{d^2l(\theta)}{d\theta^2}\right]\Big|_{\hat{\theta}} && (34)\\
=& (\theta_0 - \hat{\theta})^2\mathbb{E}\left[\frac{-d^2l(\theta)}{d\theta^2}\right]\Big|_{\hat{\theta}} && (35)
\end{aligned}$$

This is the first of the two **Wald test statistics**, denoted W_1.

$$W_1 = (\theta_0 - \hat{\theta})^2\mathbb{E}\left[\frac{-d^2l(\theta)}{d\theta^2}\right]\Big|_{\hat{\theta}} \tag{36}$$

The second is just a difference at where the expected value of the negative second derivative is evaluated. It arises from a point we mentioned earlier and is a result of the consistency property

of the MLE. Since the MLE is a consistent estimator of θ then under $H_0, \theta = \theta_0$ which means the MLE lies close to θ_0 for sufficiently large sample sizes. Hence the second test statistic, W_2 is

$$2log(LR) \approx W_2 = (\theta_0 - \hat{\theta})^2 \mathbb{E}\Big[\frac{-d^2 l(\theta)}{d\theta^2}\Big]\Big|_{\theta_0} \tag{37}$$

In both instances, large values of these test statistics provide evidence against the null hypothesis - to see this, remember that the log(LR) is the difference between the log-likelihood evaluated at the MLE and under the null distribution. In addition, while log-likelihood's aren't probabilities, areas of high density correspond to areas of high probability. So a large difference between the two values means that θ_0 is not near the MLE.

The asymptotic null distribution is still the $\chi^2(d)$ distribution

5.2 Score Test

The score test is another approximation of 2log(LR) with a key difference that $\hat{\theta}$ does not need to be known. The creation of the score test also utilizes the Taylor expansion but of the likelihood's first derivative, $l'(\theta)$, rather than the likelihood itself, $l(\theta)$, and it is expanded around θ_0

$$\begin{aligned} l'(\theta) =& l(\theta_0) + (\theta - \theta_0)\frac{dl'(\theta)}{d\theta}\Big|_{\theta_0} + \frac{1}{2}(\theta - \theta_0)^2 \frac{d^2 l'(\theta)}{d\theta^2}\Big|_{\theta_0} + ... && (38)\\ \approx& l(\theta_0) + (\theta - \theta_0)\frac{dl'(\theta)}{d\theta}\Big|_{\theta_0} && (39)\\ =& l(\theta_0) + (\theta - \theta_0)\frac{d^2 l(\theta)}{d\theta^2}\Big|_{\theta_0} && (40) \end{aligned}$$

Again we use the law of large numbers so that

$$\begin{aligned} l'(\theta) \approx& l(\theta_0) + (\theta - \theta_0)\mathbb{E}\Big[\frac{d^2 l(\theta)}{d\theta^2}\Big]\Big|_{\theta_0} && (41)\\ =& l(\theta_0) + (\theta - \theta_0)\mathbb{E}\big[l''(\theta)\big]\big|_{\theta_0} && (42) \end{aligned}$$

Then if we substitute $\hat{\theta}$ for θ we get

$$\begin{aligned} 0 \approx& \, l(\theta_0) + (\hat{\theta} - \theta_0)\mathbb{E}\big[l''(\hat{\theta})\big]\big|_{\theta_0} && (43)\\ (\hat{\theta} - \theta_0) \approx& -\frac{l(\theta_0)}{\mathbb{E}\big[l''(\hat{\theta})\big]\big|_{\theta_0}} && (44)\\ (\hat{\theta} - \theta_0) \approx& \frac{l(\theta_0)}{\mathbb{E}\big[-l''(\hat{\theta})\big]\big|_{\theta_0}} && (45)\\ & && (46) \end{aligned}$$

If we substitute this expression for $(\theta_0 - \hat{\theta})$ into our formula for the second **Wald test statistic**

$$\begin{aligned} 2log(LR) \approx& (\theta_0 - \hat{\theta})^2 \mathbb{E}\Big[\frac{-d^2 l(\theta)}{d\theta^2}\Big]\Big|_{\theta_0} && (47)\\ \approx& \left(\frac{l(\theta_0)}{\mathbb{E}\big[-l''(\hat{\theta})\big]\big|_{\theta_0}}\right)^2 \mathbb{E}\Big[\frac{-d^2 l(\theta)}{d\theta^2}\Big]\Big|_{\theta_0} && (48)\\ =& \frac{\big(l(\theta_0)\big)^2}{\mathbb{E}\big[-l''(\hat{\theta})\big]\big|_{\theta_0}} && (49) \end{aligned}$$

The asymptotic null distribution is again the $\chi^2(d)$.

The benefits of the score test is that there is no need to calculate the MLE, which can otherwise be complicated or difficult to maximise.

6 Confidence Intervals

Confidence intervals are a form of uncertainty quantification about a parameter. We know two associated quantities with an estimator are its variance and its bias.

An estimator is a random variable as it is a function of $\mathbf{X}$ and so its variance can be used to understand the uncertainty in the estimator. One way which we can do this through confidence intervals.

Definition. Confidence Interval Suppose that a single parameter θ takes values in Ω, and let $\alpha \in (0,1)$. A $100(1-\alpha)\%$ confidence interval for θ is a random interval $\big(a(\mathbf{X}), b(\mathbf{X})\big) \subset \Omega$ where the endpoints depend only on the data $\mathbf{X}$ and known constants, such that

$$P\Big[a(\mathbf{X}), b(\mathbf{X}) \text{ contains } \theta^*\Big] = 1-\alpha \tag{50}$$

When we say a random interval, we mean that the endpoints are random variables and in the case of confidence intervals are functions of **only the data**.

Given a sample, $\mathbf{X} = \mathbf{x}$, we can compute a realisation of the the confidence interval $\big(a(\mathbf{x}), b(\mathbf{x})\big)$. The realisation of the confidence interval does not mean that the true value of the parameter θ^* is contained within at a probability of $100(1-\alpha)\%$ since it is not a random interval.

A useful interpretation of confidence intervals is that if there is available a method to construct random intervals independently then if this were to be repeated a large number of times then the proportion of actual confidence intervals $\big(a(\mathbf{x}), b(\mathbf{x})\big)$ which contain θ^* would approach $1-\alpha$

6.1 Constructing Confidence Intervals

In order to construct confidence intervals, it is necessary to rely on **Pivotal Quantities**. A pivotal quantity is defined as a statistic $t(\mathbf{X}, \theta)$ that itself is a function of the data and the parameter. The quantity is either increasing or decreasing with respect to the parameter and its distribution does not depend on the parameter.

Example Suppose data, $X_1, X_2, ..., X_n$, are generated from the normal distribution $N(\mu, \sigma^2)$ where both μ and σ^2 are not known. We can compute sample estimators of the sample mean, $\bar{X}$, and the sample variance, S^2, to estimate them. Then if we define

$$t(\mathbf{X}, \mu) = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

If we want to construct a $100(1-\alpha)\%$ confidence interval we can construct an interval

$$P\big[t_{\alpha/2}(n-1) < t(\mathbf{X}, \mu) < t_{1-\alpha/2}(n-1)\big] = 1-\alpha \tag{51}$$

If we consider the interval and note the symmetry of the t-distribution such that $t_{\alpha/2}(n-1) = -t_{a-\alpha/2}(n-1)$

$$-t_{1-\alpha/2}(n-1) < \frac{\sqrt{n}(\bar{X}-\mu)}{S} < t_{1-\alpha/2}(n-1) \tag{52}$$

$$\bar{X} - \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1) < \mu \ < \bar{X} + \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1) \tag{53}$$

which if we substitute back into equation (50), gives us a $100(1-\alpha)\%$ confidence interval

$$P\big[\bar{X} - \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1) < \mu \ < \bar{X} + \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1)\big] = 1-\alpha \tag{54}$$

6.2 Confidence Intervals from Hypothesis Tests

It is necessary to define confidence regions, which are a generalization of confidence intervals to higher dimensions.

Confidence Regions A $100(1-\alpha)\%$ confidence region, $C(\mathbf{X})$, for a parameter $\theta \in \Omega$ is a random subset of Ω depending only on the data $\mathbf{X}$ such that

$$P\big[C(\mathbf{X}) \text{ contains } \theta^*\big] = 1 - \alpha$$

where θ^* is the true value of the parameter and $\alpha \in (0, 1)$

From our hypothesis tests, we found test statistics and then constructed rejection regions for a fixed size α of the form

$$P\big[t(\mathbf{X}) > c \mid H_0)\big] = \alpha$$

If we invert this test, i.e, reversing the sign of the inequality then we can create a $100(1-\alpha)\%$ confidence region. Formally, for a test statistic $t(\mathbf{x})$, a parameter θ and a fixed level test of size α of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ with rejection region $(t(\mathbf{x}) > c(\theta_0)) = \alpha$. Define

$$c(\mathbf{X}) = \big[t(\mathbf{X}) \leq c(\theta_0)))\big] \tag{55}$$

then $C(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence region, which has probability

$$P\big[t(\mathbf{X}) \leq c(\theta_0) \mid H_0)\big] = 1 - \alpha \tag{56}$$

and tells us that under the null hypothesis, the probability that $C(\mathbf{X})$ contains θ_0 is $1 - \alpha$

6.2.1 Test Inversion Lemma

When we inverted our test, we utilized the Test Inversion Lemma which asserts that plausible values of θ, given the data, form a $100(1-\alpha)\%$ confidence region. Formally, suppose θ is a single parameter and $t(\mathbf{x})$ is the test statistic of size α for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, with rejection region $\{t(\mathbf{x}) > c(\theta_0)\}$. Define

$$c(\mathbf{X}) = \big[t(\mathbf{X}) \leq c(\theta_0)))\big] \tag{57}$$

Then $C(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence region for θ.

6.2.2 The Likelihood Ratio

Using the Test Inversion Lemma we can construct approximate confidence regions using the $2log(LR)$, since this is only approximately distributed as $\chi^2(1)$.

Under the null hypothesis, the true value of the parameter $\theta = \theta_0$ and the rejection region is of the form $\{\mathbf{x} : 2log(LR) > c(\theta_0)\}$, where $c(\theta_0) = \chi^2_{1-\alpha}(1)$ for a test of size α. Formally, a $100(1-\alpha)\%$ confidence region for θ is given by

$$C(\mathbf{X}) = \{\theta : 2log(LR) \leq \chi^2_{1-\alpha}(1)\} \tag{58}$$

and contains all the values of θ such that

$$2\big(l(\hat{\theta}) - l(\theta_0)\big) \leq \chi^2_{1-\alpha}(1)$$

6.2.3 Wald Confidence Intervals

Sometimes we cannot attain explicit expressions for for confidence intervals using the likelihood ratio test, however we can generally get expressions for tests that approximate the likelihood ratio, with one example being the Wald tests. Using the first Wald test statistic, W_1

$$W_1 = (\theta_0 - \hat{\theta})^2 \mathbb{E}\left[\frac{-d^2 l(\theta)}{d\theta^2}\right]\Bigg|_{\hat{\theta}} \tag{59}$$

$$= (\theta_0 - \hat{\theta})^2 I(\hat{\theta}) \tag{60}$$

We know the rejection region for the Wald test is $\{t(\mathbf{x}) : W_1 > \chi^2_{1-\alpha}(1)\}$, therefore by the Test Inversion Lemma we have a confidence region

$$C(\mathbf{X}) = \{(\theta_0 - \hat{\theta})^2 I(\hat{\theta}) \leq \chi^2_{1-\alpha}(1)\} \tag{61}$$

If we using a test with one degree of freedom, $d = 1$, as above, then we can use the fact that the chi-squared variable is just the square of the standard normal variable, then

$$\chi^2_{1-\alpha}(1)\} = (z_{1-\alpha/2})^2 \tag{62}$$

so we can manipulate the inequality in χ^2 into a standard normal confidence interval of the form

$$\hat{\theta} - z_{1-\alpha/2}\frac{1}{\sqrt{I(\hat{\theta})}} \leq \theta \leq \hat{\theta} + z_{1-\alpha/2}\frac{1}{\sqrt{I(\hat{\theta})}} \tag{63}$$

which yields an equal-tailed $100(1-\alpha)\%$ confidence interval that approximates confidence interval for 2log(LR).

7 Appendix

7.1 Likelihood Ratio Test and the One-Sample T-Test