

Lecture Note 18: Hypothesis Testing and Confidence Intervals, CB 8.1, 8.3.1, 9.1, 9.2.1

Introductory Example

We now turn to a different type of statistical problem. Suppose we have some population of individuals, and want to see whether their average height is equal to some particular value, say 70 inches. Let X_1, \dots, X_n be a random sample from the population of interest. Assume that $X_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, and suppose that σ is known for the time being. We want to use the data to see whether $\mu = \mu_0$, where μ_0 is a fixed value. In this example, $\mu_0 = 70$.

A natural estimator of μ is $\hat{\mu} = \bar{X}_n$, the sample average. This is the MLE and the method of moments estimator. We know that $\hat{\mu} \sim N(\mu, \sigma^2/n)$. So if our estimate $\hat{\mu}$ is “far” from μ_0 , this would seem to provide evidence *against* the hypothesis that $\mu = \mu_0$. But we need to determine how far is far enough to reject the hypothesis. Note that

$$\frac{\sqrt{n}}{\sigma}(\hat{\mu} - \mu) \sim N(0, 1).$$

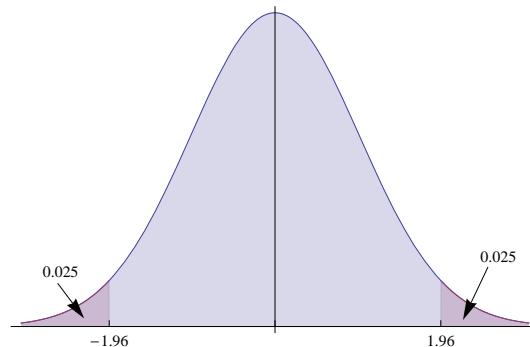
Also, it is known that if Z is a standard normal random variable, then $Pr(|Z| > 1.96) = 0.05$. Putting these two together, we conclude that

$$Pr\left(\left|\frac{\sqrt{n}}{\sigma}(\hat{\mu} - \mu)\right| > 1.96\right) = 0.05.$$

This suggests looking at the statistic

$$T(X_1, \dots, X_n) = \frac{\sqrt{n}}{\sigma}(\hat{\mu} - \mu_0).$$

(Note that this is valid statistic: we can calculate this from the data, because we know μ_0 and σ by assumption.) We can therefore calculate T , and if $|T| > 1.96$, reject the hypothesis that $\mu = \mu_0$.¹ For this procedure, we know that if $\mu = \mu_0$, then T is standard normal and $Pr(|T| > 1.96) = 0.05$. In other words, if the hypothesis $\mu = \mu_0$ is true, we have only a 5% chance of (incorrectly) rejecting it. Graphically, this can be represented as:

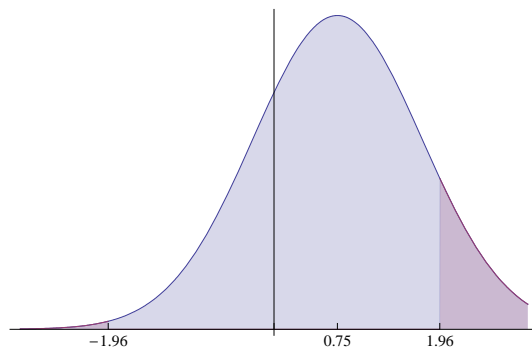


¹Note that since T is standard normal under the hypothesis, T^2 will be $\chi^2(1)$. Since $Pr(\chi^2(1) > 3.84) = .05$, we could have calculated T^2 and rejected the hypothesis if $T^2 > 3.84$. This would give exactly the same conclusion.

On the other hand, if $\mu \neq \mu_0$, the distribution of T is shifted:

$$\begin{aligned} \frac{\sqrt{n}}{\sigma} (\hat{\mu} - \mu_0) &= \frac{\sqrt{n}}{\sigma} (\hat{\mu} - \mu) + \frac{\sqrt{n}}{\sigma} (\mu - \mu_0) \\ &\sim N(0, 1) + \frac{\sqrt{n}}{\sigma} (\mu - \mu_0) \\ &\sim N\left(\frac{\sqrt{n}}{\sigma} (\mu - \mu_0), 1\right). \end{aligned}$$

For example, if $\mu > \mu_0$, then the distribution of T will be shifted to the right:



We see that the probability of $|T| > 1.96$ is now larger. That is desirable: if the hypothesis that $\mu = \mu_0$ is *false*, we want to have a high probability of rejecting it (the dark shaded regions). Of course, it is possible that $\mu \neq \mu_0$ and we fail to reject it (the light shaded region). So there are two types of mistakes we can make: we can mistakenly reject the hypothesis when it is true; and we can fail to reject the hypothesis when it is false.

Hypothesis Testing: Formal Framework

We now present a general setup for problems of testing a hypothesis. We have a random variable X , with a PDF/PMF $f_X(x; \theta)$ for some $\theta \in \Theta$. (Note that X could be a vector, in which case f_X is the joint PDF/PMF.) We assume that there is some true value of θ , in other words that the statistical model is correctly specified.

We have a hypothesis regarding the value of θ . The hypothesis is that $\theta \in \Theta_0$, where Θ_0 is some subset of the overall parameter space Θ . We refer to this as the null hypothesis, or H_0 . If the null hypothesis is not true, then $\theta \in \Theta_0^c$, where Θ_0^c is the complement of Θ_0 in the parameter space Θ . In that case the null hypothesis is false and the alternative hypothesis H_a is true. Formally:

$$\begin{aligned} H_0 : & \quad \theta \in \Theta_0, \\ H_a : & \quad \theta \in \Theta_0^c. \end{aligned}$$

(In some texts, H_1 is used to denote the alternative, rather than H_a .)

Given the null and alternative hypothesis and given a realization of the random variable X we are faced with making a decision, or taking an action. We decide either to reject the

null hypothesis, or not to reject the null hypothesis, that is to accept the null hypothesis. (Some people prefer the more neutral phrase “fail to reject the null hypothesis.”)

We make the decision on the basis of the value of the realization of the random variable X . Typically, we calculate some *test statistic* $T(X)$, and then reject the null hypothesis if $T(X) \in C_T$, where C_T is the *critical region* of the test. In the example at the start of the note, the test statistic was $T = \sqrt{n}(\hat{\mu} - \mu_0)/\sigma$, and the critical region was $C_T = (-\infty, -1.96) \cup (1.96, \infty)$. As a result there are four possible outcomes: the null hypothesis can be true or false, and we can accept or reject the null.

Table 1: HYPOTHESIS TESTING

State of Nature ↓	Decision	
	Accept H_0 (Do Not Reject H_0)	Reject H_0
H_0	Correct Decision	Type I Error
H_a	Type II Error	Correct Decision

Some of these decisions are wrong: if the null hypothesis is true and we reject it, or if the null hypothesis is false and we accept it. These are known as Type I and Type II errors respectively. Table 1 illustrates the four possibilities.

In general, there is a tradeoff between Type I and Type II errors. For example, instead of using 1.96 in our original test, we could have used, say, 2.5. That would reduce the probability of Type I error. (Consider the first graph.) However, it would also have raised the probability of Type II error (second graph).

How to decide how to trade off Type I and II errors? One natural way would be to specify the costs under all the possible outcomes, and then try to minimize the expected costs. This decision-theoretic approach is conceptually appealing, but requires a great deal of work to specify the costs completely.

The classical approach to testing in statistics, initiated in the work of Neyman and Pearson, avoids specifying these costs explicitly. Instead, we start by requiring that the probability of a Type I error is less than some preset number, the significance level of the test.² The level of the test is conventionally chosen to be either 0.05, or 0.01 or 0.10. For example, if the level is $\alpha = 0.05$, we require that for all $\theta \in \Theta_0$,

$$Pr_{\theta}(T(X) \in C_T) \leq \alpha.$$

²Sometimes we also use the term “size of the test” to refer to the significance level.

(Here Pr_θ means we calculate the probability assuming that X is distributed according to $f_X(x; \theta)$.)

Next, define the power function $\beta(\theta)$ as the probability of rejecting the null hypothesis if the true value of the parameter is θ :

$$\beta(\theta) = Pr_\theta(T(X) \in C_T).$$

Note that for $\theta \in \Theta_0^c$, the power function is one minus the probability of a Type II error:

$$\beta(\theta) = 1 - Pr_\theta(\text{Type II error}).$$

Intuitively, we want the power function to be as large as possible for values of $\theta \in \Theta_0^c$, subject to the restriction that the probability of a Type I error is not greater than the level α . In other words, we want to maximize $\beta(\theta)$ for $\theta \in \Theta_0^c$ subject to

$$\beta(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

Example 2

Suppose a college accepts applicants at an overall rate of 25%. There is a particular sub-population from which 20 applicants applied and only 2 were accepted. Is this evidence that the college discriminates against this group? First let us consider the model. It may be reasonable to model the number of accepted applicants from this group as a binomial $Bin(N, p)$ random variable with parameters $N = 20$ and probability p . Recall that the binomial distribution has PMF

$$f_X(x; N, p) = \binom{N}{x} p^x (1-p)^{N-x}.$$

The parameter space is $p \in P = [0, 0.25]$. The null hypothesis is

$$H_0 : p = 0.25,$$

and the alternative hypothesis is

$$H_a : p < 0.25.$$

(We can argue a little about the model or the formulation of the hypotheses. For example, if the college attempts to get a diverse student body, they may not treat applications as independent, and instead try to achieve a particular mix. That may lead the number of acceptances from a particular group to still have mean $N \cdot p$, but variance lower than $N \cdot p(1-p)$, so the Binomial model might not be appropriate. One might also argue that the parameter space is $P = [0, 1]$ and that the null hypothesis should be $p \geq 0.25$ rather than $p = 0.25$. Here we stick for the time being with our initial formulation.)

In this case, X is simply an integer between 0 and 20, so we might start by taking our test statistic $T(X) = X$. Now consider the critical region. Suppose we reject the null hypothesis of $p = 0.25$ for some value of x (so $x \in C_T$). What would we decide if we observed $x - 1$?

This would appear to be even stronger evidence against the null hypothesis than x (given that p cannot be larger than 0.25), and so it would appear reasonable to have $x - 1 \in C_T$ as well. This implies that the critical region should be of the form $C_T = \{0, 1, \dots, k\}$ for some integer k . (It could be that the critical region is empty, but ignore this for the time being.)

Now suppose we use the critical region $C_T = \{0, 1, \dots, k\}$. What is the probability that we make a Type I error?

$$\begin{aligned} Pr(\text{Type I error}) &= Pr(X \in C_T | N = 20, p = .25) \\ &= \sum_{i=0}^k \binom{20}{i} \cdot 0.25^i \cdot 0.75^{20-i}. \end{aligned}$$

What is the probability of a Type II error? This depends on the value of the probability under the alternative.

$$\begin{aligned} Pr(\text{Type II error}) &= Pr(X \notin C_T | N = 20, p) \\ &= \sum_{i=k+1}^{20} \binom{20}{i} \cdot p^i \cdot (1-p)^{20-i}. \end{aligned}$$

Now suppose we expand the critical region from $C_T = \{0, 1, \dots, k+1\}$. The probability of a type I error increases by

$$\binom{20}{k+1} \cdot 0.25^{k+1} \cdot 0.75^{19-k}.$$

The probability of a Type II error decreases by

$$\binom{20}{k+1} \cdot p^{k+1} \cdot (1-p)^{19-k}.$$

This illustrates fundamental tradeoff between Type I and Type II errors. By expanding the critical region we increase the probability of Type I errors and decrease the probability of Type II errors. If we contract the critical region, we do the reverse.

Suppose we decide we want the probability of a Type I error to be less than 0.01. The probability of a Type II error goes down if we increase k , so this means we should set k equal to the largest integer such that

$$Pr(\text{Type I}) = \sum_{i=0}^k \binom{20}{i} \cdot 0.25^i \cdot 0.75^{20-i} \leq 0.01.$$

Suppose we set $k = 0$. Then we have

$$Pr(\text{Type I}) = \sum_{i=0}^0 \binom{20}{i} \cdot 0.25^i \cdot 0.75^{20-i} = \binom{20}{0} \cdot 0.25^0 \cdot 0.75^{20} = 0.0032.$$

Suppose we choose $k = 1$. Then

$$Pr(\text{Type I}) = \sum_{i=0}^1 \binom{20}{i} \cdot 0.25^i \cdot 0.75^{20-i} = 0.0032 + 0.0211 = 0.0243.$$

Hence, $k = 1$ is too large and we conclude that $C_T = \{0\}$ is the optimal critical region given the (self-imposed) restriction that we want the probability that the probability of a Type I error should be less than or equal to 0.01.

What does the power function look like?

$$\begin{aligned} \beta(p) &= Pr(X \in \{0\}; N = 20, p) \\ &= (1 - p)^{20}. \end{aligned}$$

It is clearly decreasing in p . At 0.25 it is the same as the probability of a Type I error, 0.0032. At other values of p consistent with the alternative hypothesis ($p < 0.25$) it is larger than the probability of a Type I error.

□

In general, we have seen that constructing a test requires the following steps:

1. Specify the statistical model and the null and alternative hypotheses.
2. Choose a test statistic T . Intuitively, we want T to (a) have a known distribution under the null hypothesis; and (b) be sensitive to departures from the null hypothesis.
3. Select a significance level α and find a critical region for the test statistic such that the Type I error probability is less than or equal to α .

This still leaves a great deal of leeway in constructing valid hypothesis tests. The next few notes will discuss some strategies for coming up with tests, but for now we turn to a slightly different topic.

Confidence Intervals

There is a closely related problem to testing, which is to construct *interval estimates* for a parameter. A *confidence interval* is a range (rather than a point estimate) which is highly likely to contain the true parameter value. (We often use “statistical inference” to refer to hypothesis testing and confidence interval estimation together.)

Return to the example at the start of the note, where $X = (X_1, \dots, X_n)$ with $X_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, with σ^2 known. We know that

$$\frac{\sqrt{n}}{\sigma}(\hat{\mu} - \mu) \sim N(0, 1),$$

for any value of μ , and since the standard normal density is symmetric about 0,

$$\frac{\sqrt{n}}{\sigma}(\mu - \hat{\mu}) \sim N(0, 1),$$

Then we can say that

$$Pr\left(-1.96 \leq \frac{\sqrt{n}}{\sigma}(\mu - \hat{\mu}) \leq 1.96\right) = 0.95.$$

Rearranging the terms inside the probability statement,

$$Pr\left(\hat{\mu} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

So if we report the interval

$$CI = \left[\hat{\mu} - 1.96 \frac{\sigma}{\sqrt{n}}, \quad \hat{\mu} + 1.96 \frac{\sigma}{\sqrt{n}} \right],$$

this random interval will be guaranteed to contain the true value of μ 95% of the time. We say that CI is a 95% confidence interval for μ .

Note that the confidence interval endpoints cannot depend on any unknown parameters, otherwise we would have no way of actually determining it from the data.

In general, when $X \sim f_X(x; \theta)$, let $L(x)$ and $U(x)$ be functions such that $L(x) \leq U(x)$ for all x . We say that the random interval $[L(X), U(X)]$ is an interval estimator. Its *confidence level* is

$$\min_{\theta \in \Theta} Pr_{\theta}(L(X) \leq \theta \leq U(X)),$$

Typically we try to construct the interval estimator so that its confidence level is some prespecified value, such as .95. Then the interval $[L(X), U(X)]$ is guaranteed to contain the true θ at least 95% of the time.

The following result shows that if we can construct a hypothesis test, we can “invert” it to obtain a confidence interval:

Result: *Let X have PDF/PMF $f_X(x; \theta)$. Suppose that $T(X; \theta_0)$ and $C_T(\theta_0)$ are test statistics and critical regions for testing the null hypothesis that $\theta = \theta_0$ against the alternative that $\theta \neq \theta_0$, with significance level α . Then the region*

$$CI(X) = \{\theta \in \Theta : T(X; \theta) \notin C_T(\theta)\}$$

is a $1 - \alpha$ level confidence interval for θ .

In other words, we look for all potential values of θ that are not rejected by an α -level hypothesis test, and the set of such values is our confidence interval.

Let’s see how this works in our normal model example. Our test accepts the null that $\mu = \mu_0$ if

$$\left| \frac{\sqrt{n}}{\sigma}(\hat{\mu} - \mu_0) \right| \leq 1.96.$$

This is equivalent to

$$-1.96 \leq \frac{\sqrt{n}}{\sigma}(\mu_0 - \hat{\mu}) \leq 1.96$$

or

$$\hat{\mu} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \hat{\mu} + 1.96 \frac{\sigma}{\sqrt{n}}.$$

So the result says to report the interval as

$$\left\{ \mu : \hat{\mu} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + 1.96 \frac{\sigma}{\sqrt{n}} \right\}$$

which is exactly what we derived before.