Parametric Statistical Inference

Charlie Gibbons
University of California, Berkeley

ARE 210

Fall 2015

Outline

1. Hypothesis testing
   - Framework
   - p-values
   - Significance levels
   - Power
   - Wald tests
   - Likelihood ratio tests
   - Score tests
   - Finite sample tests
2. Confidence intervals
   - Framework
   - Confidence v. prediction intervals
3. Relating these concepts

Null hypothesis

We have seen how to derive an estimator along with its distribution.

Now, we want to ask whether our data are consistent with some belief that we have about the true value of our parameter(s) known as a *null hypothesis*. Examples include:

$$\theta \leq t \quad \text{and}$$
$$\theta = t$$

The first example is a *one-sided* hypothesis and the latter is a *two-sided hypothesis*.

Hypothesis testing

To perform a hypothesis test, we ask, "what's the probability of getting a value of our estimator further away from our null hypothesis (in absolute value) than our particular estimate given that the null hypothesis is true."

A null hypothesis is tested using a *rejection region*, formulated as

$$R = \{x : T(x) > c\};$$

we reject for realizations of our data if some *test statistic* $T(x)$ is above some *critical value* c.

This formulation underscores that we *reject our data as inconsistent with our hypothesis.*

Probability of rejecting the null

Let $T(X)$ be our estimator as a function of a random variable X (*e.g.*, a vector of N observations). Suppose that we calculate the probability of observing our estimator above a threshold c given that the true θ equals our null hypothesis t; this is the probability of rejecting the null hypothesis *given that the null hypothesis is true.*

$$\Pr(T(X) > c \mid \theta \leq t)$$

We use this probability to determine an appropriate threshold c.

Fisherian hypothesis testing

We begin by using the asymptotic distribution of our estimator:

$$T(X) \sim \mathrm{N}\left(\theta, \mathrm{Var}\left(T(X)\right)\right)$$

and a null hypothesis that $\theta \leq t$.

In this case, let $T(X) = \hat{\theta}$.

We have

$$\begin{aligned}
\Pr(T(X) > c \mid \theta \leq t) &= \Pr(T(X) - t > c - t \mid \theta \leq t) \\
&= \Pr\left(\frac{T(X) - t}{\sqrt{\mathrm{Var}(T(X))}} > \frac{c - t}{\sqrt{\mathrm{Var}(T(X))}} \middle| \theta \leq t \right)
\end{aligned}$$

Pivots

For any θ in this context, we have

$$Z = \frac{T(X) - t}{\sqrt{\text{Var}(T(X))}} \sim \text{N}(0, 1);$$

the *distribution* of Z is not a function of θ (though Z itself will be).

Let $X \sim F(x \mid \theta)$ and suppose that some function of X and θ, $Q(X, \theta)$, has the same distribution for all values of θ. Then $Q(X, \theta)$ is a *pivot* or *pivotal quantity* for X.

If the PDF of X can be written as

$$g(Q(x, \theta)) \left| \frac{\partial Q(x, \theta)}{\partial \theta} \right|$$

and that $Q(x, \cdot)$ is monotonic in x, then $Q(X, \theta)$ is a pivot.

Pivots and hypothesis testing

When our test statistic has a pivot, then we can derive its distribution without knowing anything about the true value of θ.

Using our pivot, we have

$$\Pr\left(Z > \frac{c - t}{\sqrt{\mathrm{Var}(T(X))}}\right).$$

p-values

Suppose that we let $c = T(x)$, the value of our estimator for our particular sample (*i.e.*, our estimate). Then

$$\Pr\left(Z > \frac{T(x) - t}{\sqrt{\text{Var}(T(X))}}\right)$$

is called the *p-value* of our estimate.

p value

The probability of observing a $\hat{\theta}$ at least as far from your null hypothesis as your actual estimate given that the null hypothesis is true.

Interpretation

Note that the p value is just a restatement, a one-to-one transformation, of our *test statistic* $\hat{z}$ and is just a means of describing our result relative to the null hypothesis; it is sample-dependent and so too is its interpretation (cf. a frequency interpretation, as in the next case).

The p value is calculated **assuming that the null hypothesis is true**. We calculate the probability of observing our data given this assumption.

Note that this tells us the probability of our data, **not the probability that the null hypothesis is true.**

Fundamental problem of statistics

We learn Pr(data | null hypothesis), not Pr(null hypothesis | data).

How can we go from the former to the latter, the actual quantity of interest?

Frequentists can't; this is called the *fundamental problem of statistics.*

Bayesian inference

Bayes' rule states that

$$\begin{aligned}&\Pr(\text{null hypothesis} \mid \text{data}) \\ &= \frac{\Pr(\text{data} \mid \text{null hypothesis}) \Pr(\text{null hypothesis})}{\Pr(\text{data})}.\end{aligned}$$

What's the problem?

- Pr(data) isn't known, but we actually don't need it and
- The *prior* probability of the null Pr(null hypothesis) is unknown.

This is an illustration of *Bayesian inference*.

Neyman-Pearson

Imagine observing many data sets and calculating many p values. You *reject the null hypothesis* if the p value is less than some level α.

Let $Z = \frac{T(X)-t}{\sqrt{\text{Var}(T(X))}}$. Then

$$\begin{aligned}
\Pr\left(p(Z) < \alpha \mid H_0\right) &= \Pr\left(p(Z) < \alpha\right) \\
&= \Pr\left(\Phi(Z) < \alpha\right) \\
&= \Pr\left(Z < \Phi^{-1}\left(\alpha\right)\right) \\
&= 1 - \Pr\left(Z < \Phi^{-1}\left(1-\alpha\right)\right) \\
&= 1 - \left(1-\alpha\right) \\
&= \alpha.
\end{aligned}$$

Significance

α is called the *significance level* of a test. We reject the null hypothesis if $p(z) < \alpha$.

If you reject the null hypothesis using a level α test, then, if you performed many α-level tests, you would falsely reject the null hypothesis $100 \times \alpha\%$ of the time.

Note that this is a frequency-based interpretation of a hypothesis test (cf. the data-specific p value).

Note that this procedure, too, does not tell us whether *our specific* null hypothesis is true or false; instead, it tells what proportion of the time we make the correct decision of rejecting the null using this procedure.

Alternative hypothesis

We have only mentioned the null hypothesis; we haven't mentioned what happens if the null is in fact false.

We haven't specified an *alternative hypothesis* H_1.

Some test statistics require specifying an alternative in order to derive them.

Type I and type II errors

Four things could happen:

	H_0 is true	H_0 is false
Do not reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

$\beta(\theta) \equiv 1 - \Pr(\text{Type II error})$ is called the *power*.

Neyman and Pearson advocated finding a test that falsely rejects the null hypothesis some specified α proportion of the time and that maximizes the probability of rejecting the null hypothesis when it is false.

We want to minimize the Type II error (or maximize power) subject to some specified level of Type I error.

A courtroom example

Consider being on a jury, with the previous table relabeled under the null hypothesis of not guilty:

	Not guilty	Guilty
Do not convict	Correct	Type II error
Convict	Type I error	Correct

We can minimize the Type I error by never convicting anyone, but that would mean that we let a lot of guilty people go free; in other words, we have a high Type II error.

We could make sure that every guilty person goes to jail by convicting everyone, but that would require convicting a lot of innocent people; minimizing the Type II error leads to a high Type I error.

There is a trade-off between Type I and Type II errors.

Most powerful tests

Actually, we (may) have taken an alternative into account before we even started.

The alternative hypothesis helps us choose the "best" (highest power) tests, but we might not use it in calculating test statistics. These are called *most powerful tests*.

A test may be powerful for only a range of alternatives, while another test is more powerful for alternatives in another range. It is hard to find a test that is the most powerful for all alternatives, a *uniformly most powerful test*.

Power calculations

Let's consider the power of the test that we derived. Suppose, without loss of generality, that our null hypothesis is that $\theta = 0$. We reject the null hypothesis if $z > c$, where c is chosen to give us the appropriate level of our test (*e.g.*, $c = \Phi^{-1}(1 - \alpha)$).

To calculate power, we calculate the probability of rejecting the null hypothesis across all possible values of the true θ. Stated differently, *power is a function of the true parameter value*:

$$\beta(\theta) = \Pr(X \in R \mid \theta).$$

Comprehension check: What is the power of this test at $\theta = 0$?

$$\sup_{\theta \in \theta_0} \beta(\theta) = \alpha.$$

Calculating power

Let z be the test statistic and Z represent a standard normal random variable. Then,

$$\begin{aligned}
\Pr(\text{reject null}) &= \Pr(z > c) \\
&= 1 - \Pr(z < c) \\
&= 1 - \Pr\left(\frac{\hat{\theta} - t}{\sqrt{\text{Var}\left(T(X)\right)}} < c \right) \\
&= 1 - \Pr\left(\frac{\hat{\theta} - \theta}{\sqrt{\text{Var}\left(T(X)\right)}} < c + \frac{t - \theta}{\sqrt{\text{Var}\left(T(X)\right)}} \right) \\
&= 1 - \Pr\left(Z < c + \frac{t - \theta}{\sqrt{\text{Var}\left(T(X)\right)}} \right) \\
&= 1 - \Phi\left(c + \frac{t - \theta}{\sqrt{\text{Var}\left(T(X)\right)}} \right)
\end{aligned}$$

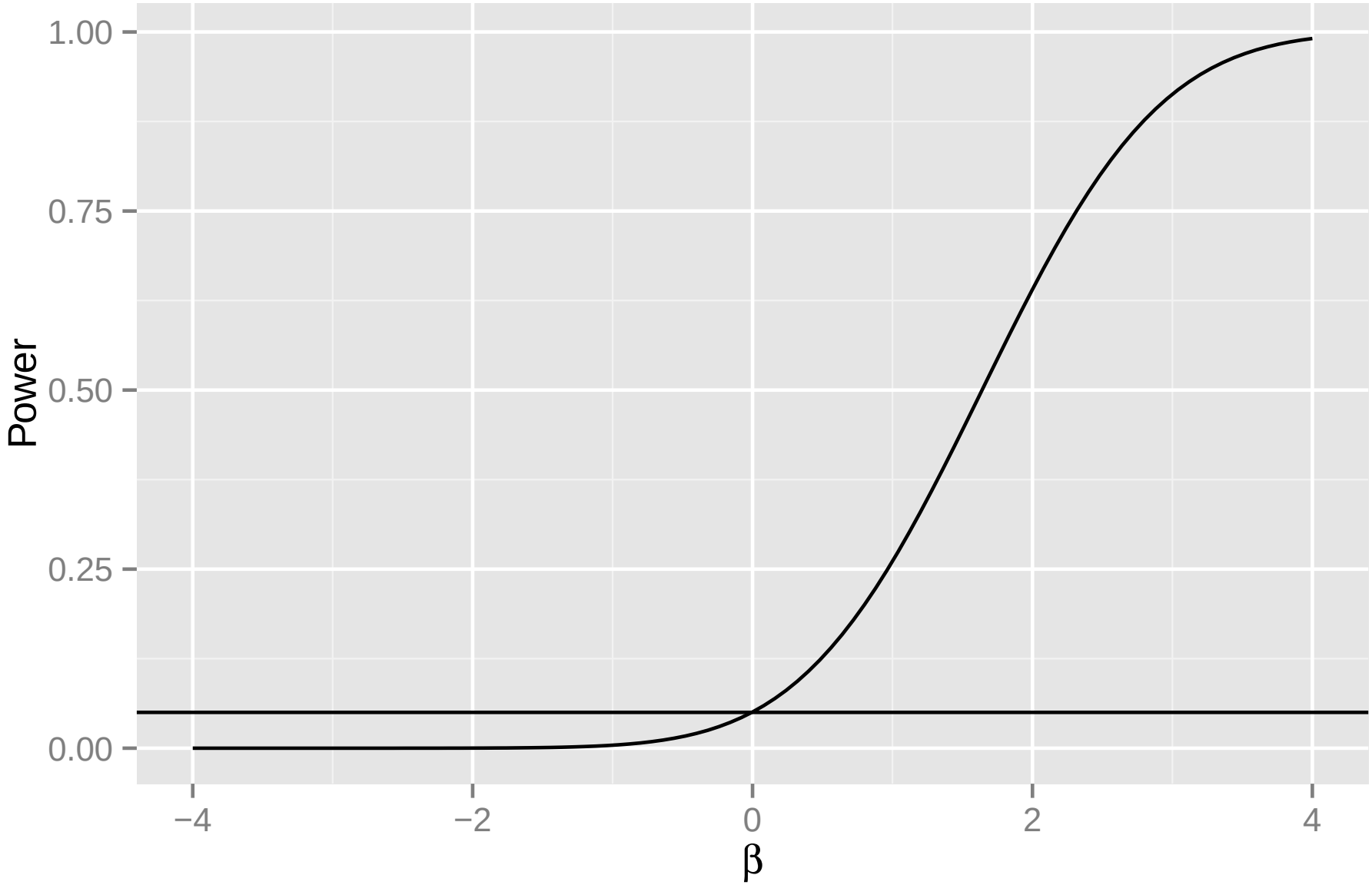

Figure : Power for z test with standard error of 1 and null hypothesis of $\theta < 0$ with $\alpha = 0.05$

Summary so far

We've used the asymptotic distribution of our estimator $\hat{\theta}$ to derive a test of the null hypothesis that $\theta < t$.

We were able to do this because $\hat{\theta}$ has an asymptotically normal distribution, which we can turn into a pivot.

This is a simple version of a *Wald test*, which uses the test statistic

$$\frac{\hat{\theta} - t}{\sqrt{\text{Var}\left(\hat{\theta}\right)}}.$$

Wald test for vector-valued θ

Now let $\hat{\theta}$ be a vector with an asymptotic $\mathrm{N}(\theta, \hat{V})$ distribution. Then, for a matrix R of a set of restrictions with rank r with the null hypothesis that $R\theta = t$. Calculate a test statistic

$$T(X) = \left(R\hat{\theta} - t\right)' \left(R\,\hat{V}\,R'\right)^{-1} \left(R\hat{\theta} - t\right) \overset{H_0}{\sim} \chi^2_r.$$

Reject if $T(x) > \chi^2_{r,\frac{\alpha}{2}}$.

Constrained MLE

Consider a null hypothesis that $\theta \in \Omega_0$ and an alternative $\theta \in \Omega_1$.

Suppose that we find the *constrained* MLEs on the subsets defined by the competing hypotheses:

$$\hat{\theta}_0 = \underset{\theta \in \Omega_0}{\operatorname{argmax}}\, \ell(\theta; x)$$

$$\hat{\theta}_1 = \underset{\theta \in \Omega_1}{\operatorname{argmax}}\, \ell(\theta; x)$$

Practically, when Ω_0 is a single point, we consider Ω_1 to be the full space including that point.

Likelihood ratio test

We can ask how likely one hypothesis is relative to the other.

Define the test statistic

$$\begin{aligned} T(X) &= -2 \log \left(\frac{\mathcal{L}\left(\hat{\theta}_0\right)}{\mathcal{L}\left(\hat{\theta}_1\right)} \right) \\ &= -2 \left[\ell\left(\hat{\theta}_0\right) - \ell\left(\hat{\theta}_1\right) \right]; \end{aligned}$$

this is the *likelihood ratio test* test statistic.

We reject H_0 if this statistic is too large.

Asymptotic LRT

Suppose that *theta* is (potentially) a vector and that $\Omega_0 \in \Omega_1$. Also, assume that the log likelihood conforms to the standard MLE regularity conditions.

Then, the asymptotic distribution of the LRT $T(X)$ is χ^2_{k-m}, where k is the number of free parameters in the alternative space and m is the number of free parameters in the null space (*i.e.*, $k - m$ is the number of restrictions or constraints imposed by the null hypothesis).

Distribution of the score

We don't just know the distribution of $\hat{\theta}$, we also know the distribution of average score:

$$S\left(\hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} s_i\left(\hat{\theta}\right) \xrightarrow{d} \mathrm{N}\left(0, \Sigma\right),$$

where

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} s_i\left(\hat{\theta}\right) s_j\left(\hat{\theta}\right)'$$

We derive the score function under the alternative model, but use the values from the restricted model to calculate the average score and the cross product matrices.

Score test

To conduct the score test, we calculate the scores and cross products under the restricted model. Then, our test statistic is

$$T\left(\hat{\theta}\right) = S'\left(\hat{\theta}\right)\Sigma^{-1}S\left(\hat{\theta}\right),$$

which has a $\chi^2_{k_1-k_0}$ distribution under the null.

Testing with normal random variables

Let $X_i \sim \text{Normal}(\mu, \sigma^2)$.

Facts:

- $Y = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n-1}}$ has a Normal$(0, 1)$ distribution.
- $Z = \frac{\sum_i \left(X_i - \bar{X}\right)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$ has a χ^2_{n-1} distribution.
- Y and Z are independent. *This is only true if X_i is normally distributed.*

t distribution

Consider the test statistic

$$\frac{\bar{X} - \mu}{\sqrt{S^2}}$$

t distribution

If Y is a standard normal random variable and Z is a χ^2_{n-1} random variable and they are independent, then

$$X = \frac{Y}{\sqrt{Z/(n-1)}}$$

has a t_{n-1} distribution with $n-1$ degrees of freedom.

Finite sample test

Hence, for a set of normal random variables, we can find a *finite sample* distribution of its test statistic and that this random variable does not depend upon σ. Hypothesis testing proceeds analogously to the asymptotic case.

Multivariate normal testing

Suppose that X_i is a vector of jointly normal random variables; it is a *multivariate* normal random variable with vector mean μ and variance-covariance matrix Σ: Normal(μ, Σ).

Consider a null hypothesis that, for a matrix R of restrictions with rank r, $R\mu = t$.

The univariate Wald test became the t-test for normal X_i; the multivariate Wald test becomes the F-test for multivariate normal X_i.

F test

F distribution

Suppose that $Y \sim \chi^2_m$, $Z \sim \chi^2_p$, and that these two variables are independent. Then

$$W = \frac{Y/m}{Z/p} \sim F_{m,p};$$

W has an F distribution with m *numerator degrees of freedom* and p *denominator degrees of freedom.*

Returning to our hypothesis that $R\mu = t$, the test statistic

$$T(X) = \frac{(R\hat{\mu} - t)' \left(R\hat{\Sigma}R'\right)^{-1} (R\hat{\mu} - t)}{r} \overset{H_0}{\sim} F_{r,N-K},$$

where K is the dimension of the vector μ.

Confidence interval definition

Suppose that we are interested in the population mean μ. We'd like to find a range of values m^L to m^U such that the expected probability of μ lying in that range is equal to $1 - \alpha$. This is known as a $100(1 - \alpha)\%$ *confidence interval*.

Formally, we want

$$\Pr\left(m^L < \mu < m^U\right) = 1 - \alpha.$$

Solving

Let's subtract our value of $\hat{\mu}$ from all sides:

$$\Pr\left(m^L - \hat{\mu} < \mu - \hat{\mu} < m^U - \hat{\mu}\right) = 1 - \alpha.$$

Now, divide all sides by the standard error of $\hat{\mu}$

$$\Pr\left(\frac{m^L - \hat{\mu}}{\sqrt{\text{Var}\left(\hat{\mu}\right)}} < \frac{\mu - \hat{\mu}}{\sqrt{\text{Var}\left(\hat{\mu}\right)}} < \frac{m^U - \hat{\mu}}{\sqrt{\text{Var}\left(\hat{\mu}\right)}}\right) = 1 - \alpha.$$

Solving, con't

Lastly, multiply everything by -1—this changes the direction of the inequalities!

$$\Pr\left(\frac{\hat{\mu} - m^L}{\sqrt{\text{Var}\left(\hat{\mu}\right)}} > \frac{\hat{\mu} - \mu}{\sqrt{\text{Var}\left(\hat{\mu}\right)}} > \frac{\hat{\mu} - m^U}{\sqrt{\text{Var}\left(\hat{\mu}\right)}}\right) = 1 - \alpha.$$

We see that this gives us a pivot, making it easy to solve for the values of m^L and m^U using the asymptotic distribution of our estimator:

$$\Pr\left(\frac{\hat{\mu} - m^L}{\sqrt{\text{Var}\left(\hat{\mu}\right)}} > Z > \frac{\hat{\mu} - m^U}{\sqrt{\text{Var}\left(\hat{\mu}\right)}}\right) = 1 - \alpha.$$

Two tails

We can write this as

$$1 - \Pr\left(\frac{\hat{\mu} - m^L}{\sqrt{\text{Var}\left(\hat{\mu}\right)}} < Z\right) - \Pr\left(Z < \frac{\hat{\mu} - m^U}{\sqrt{\text{Var}\left(\hat{\mu}\right)}}\right) = 1 - \alpha.$$

For a unimodal distribution, the shortest confidence interval is the one in which

$$f\left(\frac{\hat{\mu} - m^L}{\sqrt{\text{Var}\left(\hat{\mu}\right)}}\right) = f\left(\frac{\hat{\mu} - m^U}{\sqrt{\text{Var}\left(\hat{\mu}\right)}}\right),$$

so long as the mode is in the range of these two endpoints. For a normal distribution, this requirement is equivalent to the probability of being more extreme than these cutpoints being the same.

Equal tail probabilities

Let the probability of being in each tail be the same. Then we have:

$$1 - 2 \times \Pr\left(Z < \frac{\hat{\mu} - m^U}{\sqrt{\text{Var}\left(\hat{\mu}\right)}} \right) = 1 - \alpha$$

$$\Pr\left(Z < \frac{\hat{\mu} - m^U}{\sqrt{\text{Var}\left(\hat{\mu}\right)}} \right) = \frac{\alpha}{2}.$$

Critical values

The *critical value* is defined as the z^c such that, for a normal distribution,

$$\Pr\left(Z > z^c\right) = \Pr\left(Z < -z^c\right) = \frac{\alpha}{2}$$

These probabilities are equal because the Normal distribution is symmetric.

Here we care about our estimate being "too high" (the first probability—the upper tail) and being "too low" (the second probability—the lower tail).

Summary

We can write the general expression for confidence intervals as:

$$\left[\hat{\mu} - z^c \sqrt{\text{Var}\left(\hat{\mu}\right)},\ \hat{\mu} + z^c \sqrt{\text{Var}\left(\hat{\mu}\right)}\right].$$

This means that, if we repeated our estimation on many samples, then the true parameters would lie in these regions 95% of the time.

It does **not** state that there is a 95% chance that μ is in this range—μ is either in this range or not. The probability statement is about the *interval*, which is random because our data are random, not μ.

We say that this range has a 95% probability of *covering* the true value.

Confidence v. prediction intervals

We can create a confidence interval for the expected value of y:

$$\Pr(m^L < \mathbb{E}[y] < m^U) = 1 - \alpha.$$

A *prediction interval* predicts y itself, not its expected value:

$$\Pr(\hat{y}^L < y < \hat{y}^U) = \Pr(\hat{y}^L < \mu + \epsilon < \hat{y}^U) = 1 - \alpha.$$

Though both intervals have the same midpoint, the prediction interval has a higher variance because it takes into account the variability of our estimates as well as the variability of y itself.

Both are constructed as described for confidence intervals; just the estimate of the standard error is different.

Intuitive notions

Confidence intervals and hypothesis tests are very similar.

A confidence interval asks, given a tail probability α and the assumption that $\mu = \hat{\mu}$ (*i.e.*, that our unbiased/consistent estimator gives us an estimate that is the true mean), what data-based boundaries produce this tail probability?

A hypothesis test asks, given a data-based test statistic and the assumption that $\mu = b$ (*i.e.*, that our null hypothesis is true), what is the probability of being in the tails?

Relating these concepts

A confidence interval contains all the values for null hypotheses that cannot be rejected at the α level:

$$C_\alpha(X) = \{\theta_0 : X \in A_\alpha(\theta_0)\}.$$

A hypothesis that is rejected at the α level is outside of the $100(1-\alpha)\%$ confidence interval and a hypothesis that cannot be rejected at that level is contained in that confidence interval:

$$A_\alpha(\theta_0) = \{X : \theta_0 \in C_\alpha(X)\}.$$

Thus, it is said that a hypothesis test is an *inverted* confidence interval and vice-versa.