10/31: Hypothesis Testing

Assignment 5

Is available on `http:/www.cs.cmu.edu/~10-751`.

Due on Tuesday, November 12.

Using Hypothesis Testing

- There is a "duality" between confidence intervals and hypothesis testing.
- Hypothesis testing can be overused, prone to misinterpretation.
- Only use where there is a very specific test that is to be evaluated, such as a claim that two predictors have the same accuracy.
- Confidence intervals are generally more useful.

Recall: t distribution

If $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_n$, then

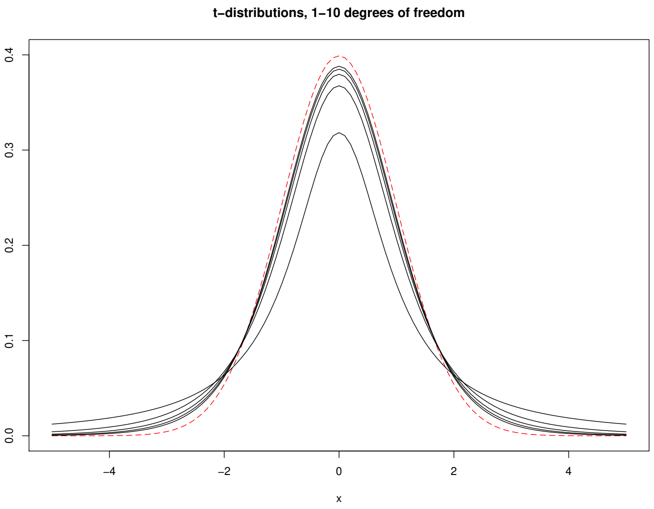
$$\frac{Z}{\sqrt{\frac{1}{n}Y}} \sim t_n$$

Appropriate when we don't know σ^2, since

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

has a χ^2_{n-1} distribution (and is independent of $\overline{X}_n$)

t Distributions



Simple t test

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where μ and σ unknown.

Want to test $\mu = \mu_0$.

Let $T = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S_n}$, where S_n^2 is the sample variance.

Suppose $\Pr(|T| > c_{\alpha/2}) = \alpha$, where $c_{\alpha/2} = F^{-1}_{t_{n-1}}(1 - \alpha/2)$ is the quantile for the t-distribution.

For a "size α" test, we reject the hypothesis that $\mu = \mu_0$ if $|T| > c_{\alpha/2}$.

Simple t test (cont.)

This illustrates the basic duality between confidence intervals and hypothesis testing.

We'll return to this notion later.

Basic Idea

Suppose that one group of rats is given a drug, the other a placebo. Some laboratory experiments are carried out, and the occurrence of a certain type of cancer is recorded for each of the two groups.

We want to test hypothesis that the cancer rates in the two groups are the same.

Null Hypothesis: The cancer rate is the same.

Alternative Hypothesis: The cancer rate is different.

Hypotheses and Acceptance Regions

Null Hypothesis H_0: $\theta \in \Theta_0$

Alternative Hypothesis H_1: $\theta \in \Theta_1$

Typically $\Theta_1 = \Theta - \Theta_0$.

Given a statistic $T = T(X_1, X_2, \ldots, X_n)$, a *rejectance region* is defined as $R = \{(x_1, \ldots, x_n) : T(x_1, \ldots, x_n) > c\}$.

If $\boldsymbol{x} \in R$, reject H_0, otherwise do not reject H_0.

Types of Errors

	Reject Null	Retain Null
Null is true	type I error	ok
Null is false	ok	type II error

Power and Size

The *power function* of a test is $\pi(\theta) = \mathrm{Pr}_\theta(T \in R)$.

Want $\pi(\theta)$ low for $\theta \in \Theta_0$, and $\pi(\theta)$ high for $\theta \in \Theta_1$

The *size* of a test is $\alpha = \sup_{\theta \in \Theta_0} \mathrm{Pr}_\theta(T \in R)$, which is the largest probability of making a type I error.

For $H = \theta = \theta_0$, this is just $\mathrm{Pr}_{\theta_0}(T \in R_{\theta_0})$

A test has *level* α in case $\pi(\theta) \leq \alpha$ for $\theta \in \Theta_0$.

Example – One-Sided Test

$X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, σ known.

$H_0 : \mu \in \Theta_0 = (-\infty, 0]$

$H_1 : \mu \in \Theta_1 = (0, \infty)$.

Simple test: reject H_0 if $T = \overline{X}_n > c$.

Thus $R = \{\boldsymbol{x} : T(\boldsymbol{x}) > c\}$.

Example (cont.)

$$\begin{aligned} \text{Power} \quad \pi(\mu) &= \Pr_\mu(\overline{X}_n > c) \\ &= 1 - \Phi\left(\frac{\sqrt{n}(c-\mu)}{\sigma}\right) \end{aligned}$$

To get a size α test,

$$c = \mu + \frac{\sigma\Phi^{-1}(1-\alpha)}{\sqrt{n}}$$

(we take $\mu = 0$ in the example.)

Note the relation with confidence intervals

Wald Test

The Wald test is a natural extension of asymptotic normality and approximate confidence intervals for the normal.

Can you guess how it's going to work?

Wald Test

Consider $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$.

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

$$\frac{\hat{\theta} - \theta_0}{\widehat{se}} \rightsquigarrow \mathcal{N}(0,1)$$

The size α Wald test is: reject H_0 when

$$\left|\frac{\hat{\theta} - \theta_0}{\widehat{se}}\right| > z_{\alpha/2}$$

Asymptotically, this test has size α.

Comparing Accuracies

Let $X \sim \text{Binomial}(m, p_1)$, and $Y \sim \text{Binomial}(n, p_2)$.

Want to test the hypothesis $p_1 = p_2$. Let $\delta = p_1 - p_2$.

$H_0 : \delta = 0$.

The maximum likelihood estimators are $\hat{p}_1$ and $\hat{p}_2$, the fraction of errors.

The estimated standard errors are

$$\begin{aligned} \widehat{\mathsf{se}}_1 &= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m}} \\ \widehat{\mathsf{se}}_2 &= \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n}} \end{aligned}$$

Comparing Accuracies (cont.)

Wald test: reject H_0 when $|W| \geq z_{\alpha/2}$, where

$$W = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}}$$

When is the power of this test large?

Wald and Confidence Intervals

Here we see the duality plainly:

The size α Wald test rejects $H_0 : \theta = \theta_0$ if and only if $\theta \notin C$, where C is the confidence interval

$$C = (\hat{\theta} - \hat{\mathsf{se}}\, z_{\alpha/2}, \hat{\theta} + \hat{\mathsf{se}}\, z_{\alpha/2})$$

p-values

For a given test and data set, the smallest value of α at which the null hypothesis is rejected is called the *p-value* of the test (for the data).

Think of the confidence interval around $\hat{\theta}$ expanding until θ_0 falls within it.

p-values (cont.)

The p-value is a measure of the evidence against H_0. Customarily:

p-value		implication
< 0.01	:	very strong evidence against H_0
0.01–0.05	:	strong evidence against H_0
0.05–0.10	:	weak evidence against H_0
> 0.1	:	little evidence against H_0

p-values (cont.)

The p-value may be high because the test has low power (or of course because H_0 is true)

p-values (cont.)

Theorem. If the test is of the form

$$\text{reject } H_0 \text{ if } T(X) > c_\alpha$$

then

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

Thus, the p-value is the probability of observing a statistic that was at least as large as what was actually observed.

p-values for the Wald Test

For the Wald test, under the null hypothesis, $\frac{\hat{\theta}(X)-\theta_0}{\hat{\text{se}}}$ is approximately standard normal.

Thus, the p-value is approximately given by

$$p\text{-value} \approx \Pr(|Z| > |w|) = 2\Phi(-|w|)$$

where

$$w = \frac{\hat{\theta}(x) - \theta_0}{\hat{\text{se}}(x)}$$

p-values for the Wald Test: Example

Suppose that we observe binomial data with $\overline{X}_n = 216.2$ and $\overline{Y}_n = 195$, and $\hat{\sigma}_X^2 = 5^2$ and $\hat{\sigma}_Y^2 = 2.4^2$.

$$W = \frac{2.16.2 - 195.3}{\sqrt{5^2 + 2.4^2}} = 3.78$$

$$p\text{-value} \Pr(|Z| > 3.78) = 2\Phi(-3.78) = 0.002$$

considered very strong evidence against the null hypothesis that the error rates are the same.

Confidence Regions from Tests

Suppose that for every θ_0 there is a level α test for $H_0 : \theta = \theta_0$.

Let R_{θ_0} denote the rejectance region of the test.

Let $C(X) = \{\theta : X \notin R_\theta\}$.

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

(Of course, this won't give an *interval* in general)

Confidence Regions from Tests

Proof:
$$\begin{aligned} \text{Pr}_{\theta_0}(\theta_0 \in C(X)) &= \text{Pr}_{\theta_0}(X \notin R_{\theta_0}) \\ &\geq 1-\alpha \end{aligned}$$

Flip Side: Tests from Confidence Regions

Suppose that $C(X)$ is a $100(1-\alpha)$ confidence region for θ. Thus,

$$\text{Pr}_\theta(\theta \in C(X)) = 1-\alpha$$

Then a rejectance region for a level α test of the null hypothesis $H_0 : \theta = \theta_0$ is

$$R_{\theta_0} = \{X \mid \theta_0 \notin C(X)\}$$