MS&E 226: Fundamentals of Data Science

Lecture 11: Hypothesis testing

Ramesh Johari

Hypotheses

Quantifying uncertainty

Recall the two key goals of inference:

- *Estimation*. What is our best guess for the process that generated the data?
- *Quantifying uncertainty*. What is our uncertainty about our guess?

Hypothesis testing provides another way to quantify our uncertainty.

Example: biased coin flipping

Suppose that we flip a coin 10 times. We observe 9 heads.

We estimate the bias as $\hat{q} = 0.9$. How likely are we to observe an estimate this extreme, *if* the coin really had bias $1/2$?

- In that case, the number of heads in ten flips is $\text{Binomial}(10, 1/2)$.
- The chance of seeing at least 9 heads is ≈ 0.0107.

In other words, it is *very unlikely* that we would have seen so many heads if the true bias were $1/2$.

Hypothesis testing gives a general architecture for making such statements.

The Wald test

Assumption: Asymptotic normality

Suppose that we have data $\mathbf{Y}$ that comes from an unknown distribution determined by the parameter θ.

Suppose we use $\mathbf{Y}$ to compute an estimator $\hat{\theta}$ of θ that is

- consistent and
- asymptotically normal.

(One example: The maximum likelihood estimate.)

I.e., for large n, the sampling distribution of $\hat{\theta}$ is:

$$\hat{\theta} \sim \mathcal{N}(\theta, \widehat{\mathsf{SE}}^2),$$

where θ is the true parameter.

The Wald test

Given a fixed value θ_0, suppose we construct the following *test statistic*:

$$T(\mathbf{Y}) = \frac{\hat{\theta} - \theta_0}{\widehat{\mathsf{SE}}}.$$

Note that:

The Wald test

Given a fixed value θ_0, suppose we construct the following *test statistic*:

$$T(\mathbf{Y}) = \frac{\hat{\theta} - \theta_0}{\widehat{\mathsf{SE}}}.$$

Note that:

- *If* the true parameter was θ_0, *then* the sampling distribution of the Wald test statistic should be approximately $\mathcal{N}(0, 1)$.

The Wald test

Given a fixed value θ_0, suppose we construct the following *test statistic*:

$$T(\mathbf{Y}) = \frac{\hat{\theta} - \theta_0}{\widehat{\mathsf{SE}}}.$$

Note that:

- *If* the true parameter was θ_0, *then* the sampling distribution of the Wald test statistic should be approximately $\mathcal{N}(0, 1)$.
- Look at the observed value of the test statistic; call it T_{obs}.

The Wald test

Given a fixed value θ_0, suppose we construct the following *test statistic*:

$$T(\mathbf{Y}) = \frac{\hat{\theta} - \theta_0}{\widehat{\mathsf{SE}}}.$$

Note that:

- *If* the true parameter was θ_0, *then* the sampling distribution of the Wald test statistic should be approximately $\mathcal{N}(0, 1)$.
- Look at the observed value of the test statistic; call it T_{obs}.
- Intuitively, if a value as extreme as T_{obs} is unlikely under the $\mathcal{N}(0, 1)$ distribution, then we can rule confidently out θ_0 as the true value of the parameter θ.

The Wald test

Given a fixed value θ_0, suppose we construct the following *test statistic*:

$$T(\mathbf{Y}) = \frac{\hat{\theta} - \theta_0}{\widehat{\mathsf{SE}}}.$$

Note that:

- *If* the true parameter was θ_0, *then* the sampling distribution of the Wald test statistic should be approximately $\mathcal{N}(0, 1)$.
- Look at the observed value of the test statistic; call it T_{obs}.
- Intuitively, if a value as extreme as T_{obs} is unlikely under the $\mathcal{N}(0, 1)$ distribution, then we can rule confidently out θ_0 as the true value of the parameter θ.

Here the *null hypothesis* is that $\theta = \theta_0$.
The *alternative hypothesis* is that $\theta \neq \theta_0$.

We are asking if we have enough evidence to *reject the null*.

Note: Rejecting the null does not necessarily mean we accept the alternative!

Hypothesis testing and binary classifcation

There are two kinds of mistakes we can make:

- *False positive*: In fact the null is true, but we mistakenly reject the null.
- *False negative*: In fact the null is false, but we mistakenly fail to reject the null.

The false positive probability $\mathbb{P}(\text{reject}|\theta = \theta_0)$ is called the *size*.

For any specific alternative $\tilde{\theta} \neq \theta_0$, $\mathbb{P}(\text{reject}|\theta = \tilde{\theta})$ is called the *power* at $\tilde{\theta}$. (This is the *true positive probability*.)

Size of the Wald test: A picture

Note that under the null, $|T_{\text{obs}}| \leq 1.96$ with probability 0.95.

So if we reject the null when $|T_{\text{obs}}| > 1.96$, the *size* (i.e., false positive probability) of the test is exactly 0.05. (Smaller size requires a larger threshold.)

Power of the Wald test

Now suppose the true $\theta = \tilde{\theta} \neq \theta_0$. What is the chance we (correctly) reject the null?

Note that in this case, the sampling distribution of the Wald test statistic is still approximately normal with variance 1, but now with mean $(\tilde{\theta} - \theta_0)/\widehat{\mathsf{SE}}$.

Therefore the power at $\tilde{\theta}$ is approximately $\mathbb{P}(|Z| > 1.96)$, where:

$$Z \sim \mathcal{N}\left(\frac{\tilde{\theta} - \theta_0}{\widehat{\mathsf{SE}}}, 1\right).$$

Power of the Wald test: A picture

Size and power: The general case

More generally, let $z_{\alpha/2}$ be the unique point such that $\mathbb{P}(|Z| > z_{\alpha/2}) = \alpha$, for a standard normal r.v. Z. Then the Wald test of size α rejects the null when $|T_{\text{obs}}| > z_{\alpha/2}$.

The power of this test at $\tilde{\theta}$ is approximately $\mathbb{P}(|Z| > z_{\alpha/2})$, where:

$$Z \sim \mathcal{N}\left(\frac{\tilde{\theta} - \theta_0}{\widehat{\text{SE}}}, 1\right).$$

Note that increasing power requires increasing size, as is usually the case in binary classification!

The Wald test and confidence intervals

Recall that a (asymptotic) 95% confidence interval for the true θ is:

$$[\hat{\theta} - 1.96\widehat{\mathsf{SE}}, \hat{\theta} + 1.96\widehat{SE}].$$

Thus the Wald test of size 5% is equivalent to *rejecting the null if* θ_0 *is not in the* 95% *confidence interval*.

The Wald test and confidence intervals

Recall that a (asymptotic) 95% confidence interval for the true θ is:

$$[\hat{\theta} - 1.96\widehat{\mathsf{SE}}, \hat{\theta} + 1.96\widehat{SE}].$$

Thus the Wald test of size 5% is equivalent to *rejecting the null if* θ_0 *is not in the* 95% *confidence interval*.

More generally, recall that a (asymptotic) $1 - \alpha$ confidence interval for the true θ is:

$$[\hat{\theta} - z_{\alpha/2}\widehat{\mathsf{SE}}, \hat{\theta} + z_{\alpha/2}\widehat{\mathsf{SE}}].$$

The Wald test of size α is equivalent to *rejecting the null if* θ_0 *is not in the* $1 - \alpha$ *confidence interval*.

Example: Significance of an OLS coefficient

Suppose given $\mathbf{X}, \mathbf{Y}$, we run a regression and find OLS coefficients $\hat{\boldsymbol{\beta}}$.

We test whether the true β_j is zero or not. *Thus* $\theta_0 = 0$. The Wald test statistic is $\hat{\beta}_j / \widehat{\mathsf{SE}}_j$.

If this statistic has magnitude larger than 1.96, then we say the coefficient is *statistically significant* (at the 95% level).

This is equivalent to the following: if zero is not in the 95% confidence interval for a particular regression coefficient $\hat{\beta}_j$, the coefficient is statistically significant at the 95% level.

More on statistical significance

Lots of warnings:

- Statistical significance of a coefficient suggests it is worth including in your regression model; but don't forget all the other assumptions that have been made along the way!
- Conversely, just because a coefficient is *not* statistically significant, does not mean that it is not important to the model!
- Statistical significance is very different from *practical* significance! Even if zero is not in a confidence interval, the relationship between the corresponding covariate and the outcome may still be quite weak.

p-values

The *p-value* of a test gives the probability of observing a test statistic as extreme as the one observed, *if the null hypothesis were true*.

For the Wald test:

$$p = \mathbb{P}(|Z| > |T_{\text{obs}}|),$$

where $Z \sim \mathcal{N}(0, 1)$ is a standard normal random variable.

Why? Under the null, the sampling distribution of the Wald test statistic is approximately $\mathcal{N}(0, 1)$.

p-values

Note that:

- If the p-value is small, the observed test statistic is very unlikely under the null hypothesis.

p-values

Note that:

- If the p-value is small, the observed test statistic is very unlikely under the null hypothesis.
- In fact, suppose we reject when $p < \alpha$. This is *exactly* the same as rejecting when $|T_{\text{obs}}| > z_{\alpha/2}$.

p-values

Note that:

- If the p-value is small, the observed test statistic is very unlikely under the null hypothesis.
- In fact, suppose we reject when $p < \alpha$. This is *exactly* the same as rejecting when $|T_{\text{obs}}| > z_{\alpha/2}$.
- In other words: *The Wald test of size α is obtained by rejecting when the p-value is below α.*

p-values: A picture

The sampling distribution of p-values: A picture [$*$]

Use and misuse of p-values

Why p-values? They are *transparent*:

- Reporting "statistically significant" (or not) depends on *your* chosen value of α.
- What if *my* desired α is different (more or less conservative)?
- p-values allow different people to interpret the data using their own desired α.

Use and misuse of p-values

Why p-values? They are *transparent*:

- Reporting "statistically significant" (or not) depends on *your* chosen value of α.
- What if *my* desired α is different (more or less conservative)?
- p-values allow different people to interpret the data using their own desired α.

But note: the p-value is *not* the probability the null hypothesis is true!

The t-test

The z-test

We assume that $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are i.i.d. $\mathcal{N}(\theta, \sigma^2)$ random variables.

If we know σ^2:

- The variance of the sampling distribution of $\overline{Y}$ is σ^2/n, so its exact standard error is $\text{SE} = \sigma/\sqrt{n}$.
- Thus *if* $\theta = \theta_0$, then $(\overline{Y} - \theta_0)/\text{SE}$ should be $\mathcal{N}(0, 1)$.
- So we can use $(\overline{Y} - \theta_0)/\text{SE}$ as a test statistic, and proceed as we did for the Wald statistic. This is called a *z-test*.

The only difference from the Wald test is that *if* we know the Y_i's are normally distributed, *then* the test statistic is exactly normal even in finite samples.

The t-statistic

What if we don't know σ^2? Let $\hat{\sigma}^2$ be the unbiased estimator of σ^2. Then with $\widehat{\mathsf{SE}} = \hat{\sigma}/\sqrt{n}$,

$$\frac{\overline{Y} - \theta_0}{\widehat{\mathsf{SE}}}$$

has a *Student's t distribution* under the null hypothesis that $\theta = \theta_0$. This distribution can be used to implement the *t-test*.

For our purposes, just note that again this looks a lot like a Wald test statistic! Indeed, the t distribution is very close to $\mathcal{N}(0, 1)$, even for moderate values of n.

Example: Linear normal model [$*$]

Assume the linear normal model $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$ with i.i.d. $\mathcal{N}(0, \sigma^2)$ errors ε_i.

OLS estimator is:

$$\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

Now note that given $\mathbf{X}$, *the sampling distribution of the coefficients is exactly normal*, because the coefficients are linear combinations of the Y_i's (which are independent normal random variables).

This fact can be used to show the exact sampling distribution of the test statistic $\hat{\beta}_j/\widehat{\mathsf{SE}}_j$ under the null that $\beta_j = 0$ is also a t distribution. (See [SM], Section 5.6.)

Interpreting regression output in R

R output from a linear regression:

```
Call:
lm(formula = Ozone ~ 1 + Solar.R + Wind + Temp, data = airqualit

Residuals:
    Min      1Q  Median      3Q     Max
-40.485 -14.219  -3.551  10.097  95.619

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Solar.R       0.05982    0.02319   2.580  0.01124 *
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
Temp          1.65209    0.25353   6.516 2.42e-09 ***
```

Statistical significance in R

In most statistical software (and in papers), statistical significance is denoted as follows:

- *** means "statistically significant at the 99.9% level".
- ** means "statistically significant at the 99% level".
- * means "statistically significant at the 95% level".

The hypothesis testing recipe

The hypothesis testing recipe

Nearly all hypothesis testing follows a common underlying logic:

- If the true parameter was θ_0...

The hypothesis testing recipe

Nearly all hypothesis testing follows a common underlying logic:

- If the true parameter was θ_0...
- then the *test statistic* $T(\mathbf{Y})$ should look like it came from $f(Y|\theta_0)$.

The hypothesis testing recipe

Nearly all hypothesis testing follows a common underlying logic:

- If the true parameter was θ_0...
- then the *test statistic* $T(\mathbf{Y})$ should look like it came from $f(Y|\theta_0)$.
- We compare the *observed* $T(\mathbf{Y})$ to the *sampling distribution of the test statistic under* θ_0.

The hypothesis testing recipe

Nearly all hypothesis testing follows a common underlying logic:

- If the true parameter was θ_0...
- then the *test statistic* $T(\mathbf{Y})$ should look like it came from $f(Y|\theta_0)$.
- We compare the *observed* $T(\mathbf{Y})$ to the *sampling distribution of the test statistic under* θ_0.
- If the observed $T(\mathbf{Y})$ is unlikely under the sampling distribution of the test statistic given θ_0, we *reject the null hypothesis that* $\theta = \theta_0$.

Decision rules

For many tests (e.g., Wald test, t-test), large test statistic magnitudes are unlikely under the null, so the decision rule takes the form:

"If $|T(\mathbf{Y})| \geq s$, then reject the null; otherwise accept the null."

The choice of s then governs both the size, and the power at specific alternatives.

Caution

A word of warning

Used correctly, hypothesis tests are powerful tools to quantify your uncertainty.

However, they can easily be misused, as we will see later in the course. Some questions for thought:

- Suppose with 1000 covariates, you use the t (or Wald) statistic on each coefficient to determine whether to include or exclude it. What might go wrong?
- Suppose that you test and compare many models by repeatedly using F tests. What might go wrong?