MS&E 226: "Small" Data

Lecture 15: Examples of hypothesis tests (v3)

Ramesh Johari
ramesh.johari@stanford.edu

The recipe

The hypothesis testing recipe

In this lecture we repeatedly apply the following approach.

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

The theory of hypothesis testing relies on finding *test statistics* $T(\mathbf{Y})$ for which this procedure yields as high a power as possible, given a particular size.

The Wald test

Assumption: Asymptotic normality

We assume that the statistic we are looking at is in fact an estimator $\hat{\theta}$ of a parameter θ, that is:

- unbiased and
- asymptotically normal.

(Example: MLE.)

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

The *Wald test* uses test statistic:

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

The recipe:

The Wald test

The *Wald test* uses test statistic:

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

The recipe:

- *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

The *Wald test* uses test statistic:

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

The recipe:

- *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

The *Wald test* uses test statistic:

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

The recipe:

- *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}.
- Under the null, $|T_{\mathsf{obs}}| \leq 1.96$ with probability 0.95.

The Wald test

The *Wald test* uses test statistic:

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

The recipe:

- *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}.
- Under the null, $|T_{\mathsf{obs}}| \leq 1.96$ with probability 0.95.
- So if we reject the null when $|T_{\mathsf{obs}}| > 1.96$, the size of the test is 0.05.

The Wald test

The *Wald test* uses test statistic:

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

The recipe:

- *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}.
- Under the null, $|T_{\mathsf{obs}}| \leq 1.96$ with probability 0.95.
- So if we reject the null when $|T_{\mathsf{obs}}| > 1.96$, the size of the test is 0.05.

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 $|T_{\mathsf{obs}}| > z_{\alpha/2}$.

The Wald test: A picture

The Wald test and confidence intervals

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 is equivalent to *rejecting the null if* θ_0 *is not in the* $1-\alpha$ *confidence interval*.

Power of the Wald test

Now suppose the true $\theta \neq \theta_0$. 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 $(\theta - \theta_0)/\widehat{\mathsf{SE}}$.

Therefore the power at θ is approximately $\mathbb{P}(|Z| > z_{\alpha/2})$, where:

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

Power of the Wald test: A picture

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. 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).

(Though remember that "statistical significance" can be far from practical significance...)

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 F test in linear regression

Multiple regression coefficients

We again assume we are in 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.

The Wald (or t) test lets us test whether *one* regression coefficient is zero.

Multiple regression coefficients

We again assume we are in 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.

The Wald (or t) test lets us test whether *one* regression coefficient is zero.

What if we want to know if *multiple* regression coefficients are zero? This is equivalent to asking: *would a simpler model suffice?* For this purpose we use the F test.

The F statistic

Suppose we have fit a linear regression model using p covariates. We want to test the null hypothesis that *all* of the coefficients $\beta_j, j \in S$ (for some subset S) are zero.

Notation:

- $\hat{\mathbf{r}}$: the residuals from the full regression ("unrestricted")
- $\hat{\mathbf{r}}^{(S)}$: the residuals from the regression *excluding* variables in S ("restricted")

The F statistic is:

$$F = \frac{(\|\hat{\mathbf{Y}}\|^2 - \|\hat{\mathbf{Y}}^{(S)}\|^2)/(p - |S|)}{\|\hat{\mathbf{r}}\|^2/(n-p)}.$$

The F test

The recipe for the F test:

- The null hypothesis is that $\beta_j = 0$ for all $j \in S$.
- If this is true, then the test statistic has an F distribution, with $p - |S|$ degrees of freedom in the numerator, and $n - p$ degrees of freedom in the denominator.
- We can use the F distribution to determine how unlikely our observed value of the test statistic is, if the null hypothesis were true.

Under the null we expect $F \approx 1$. Large values of F suggest we can reject the null.

The F test

The F test is a good example of why hypothesis testing is useful:

- We could implement the Wald test by just looking at the confidence interval for β_j.
- The same is not true for the F test: we can't determine whether we should reject the null by just looking at individual confidence intervals for each β_j.
- The F test is a succinct way to summarize our level of uncertainty *about multiple coefficients at once*.

"The" F test

Statistical software such as R does all the work for you.

First, note that regression output always includes information on "the" F statistic, e.g.:

```
Call:
lm(formula = Ozone ~ 1 + Solar.R + Wind + Temp, data = airqualit
...
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

This is always the F test against *the null that all coefficients (except the intercept) are zero*. What does rejecting this null mean? What is the alternative?

F tests of one model against another

More generally, you can use R to run an F test of one model against another:

```
> anova(fm_small, fm_big)
Analysis of Variance Table

Model 1: Ozone ~ 1 + Temp
Model 2: Ozone ~ 1 + Temp + Solar.R + Wind
  Res.Df   RSS Df Sum of Sq     F   Pr(>F)
1    109 62367
2    107 48003  2     14365 16.01 8.27e-07 ***
...
```

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?