MS&E 226: Fundamentals of Data Science

Lecture 11: Hypothesis testing

Ramesh Johari

Introduction to hypothesis testing

The two goals of parametric inference

Recall the following two goals

- *Estimation*. What is our best guess for the true parameters of the population model (e.g., the population mean)?
- *Quantifying uncertainty*. How uncertain are we in our guess?

So far we've talked about quantifiying uncertainty via *standard errors* and *confidence intervals*.

Today we'll talk about a different way to quantify uncertainty: *hypothesis testing*.

Motivating example: Flight delays

Suppose we draw a sample of $n = 500$ flights and obtain:

- Sample mean delay: $\overline{Y} = 11.68$ minutes
- Sample standard deviation: $\hat{\sigma} = 74.45$ minutes
- Estimated SE: $\widehat{\text{SE}} = \hat{\sigma}/\sqrt{n} = 3.33$ minutes

Question: Is our evidence consistent with the *hypothesis* that the true population mean μ is zero? I.e., is the average flight is exactly on time?

The hypothesis testing "recipe"

We want to test whether a specific claim about a parameter is plausible, given our data.

The hypothesis testing "recipe":

- Suppose the claim were true. This is the *null hypothesis*, denoted H_0.

The hypothesis testing "recipe"

We want to test whether a specific claim about a parameter is plausible, given our data.

The hypothesis testing "recipe":

- Suppose the claim were true. This is the *null hypothesis*, denoted H_0.
- Across many "parallel universes" (the sampling distribution), how likely would we be to observe data as extreme as what we actually saw?

The hypothesis testing "recipe"

We want to test whether a specific claim about a parameter is plausible, given our data.

The hypothesis testing "recipe":

- Suppose the claim were true. This is the *null hypothesis*, denoted H_0.
- Across many "parallel universes" (the sampling distribution), how likely would we be to observe data as extreme as what we actually saw?
- If very unlikely, we *reject* the null hypothesis.

Virtually all hypothesis tests work this way!

Example: Testing if mean delay is zero

Null hypothesis H_0: $\mu = 0$ (true population mean is zero)

Alternative hypothesis H_1: $\mu \neq 0$ (true population mean is not zero)

From the CLT, we know that for large n, the sampling distribution is *approximately normal*:

$$\overline{Y} \approx \mathcal{N}(\mu, \widehat{\mathsf{SE}}^2).$$

Question: If H_0 were true ($\mu = 0$), how likely are we to see $|\overline{Y}| \geq 11.68$?

The t statistic

Standardizing the sample mean

To test $H_0 : \mu = \mu_0$, we construct the following test statistic, also called a *t statistic*:

$$\hat{t} = \frac{\overline{Y} - \mu_0}{\widehat{\text{SE}}}.$$

From the CLT, we know that $\overline{Y}$ has a sampling distribution that is approximately $\mathcal{N}(\mu, \widehat{\text{SE}}^2)$ for large n.

Therefore: If H_0 is true ($\mu = \mu_0$), then for large n, the sampling distribution of $\hat{t}$ is approximately $\mathcal{N}(0, 1)$.

The t statistic

We write $\hat{t}$ for our *observed* value of our t statistic.

We can use $\hat{t}$ to "test" whether we believe the null hypothesis $H_0 : \mu = \mu_0$ is true:

- If H_0 is true, then $\hat{t}$ should be "typical" for a $\mathcal{N}(0, 1)$ random variable.
- If H_0 is false, then $\hat{t}$ will tend to be "large" in absolute value.

Example: Flight delays

For our flight delays example:

- $H_0 : \mu = 0$
- $\overline{Y} = 11.68$
- $\widehat{\text{SE}} = 3.33$

Test statistic:

$$\hat{t} = \frac{\overline{Y} - 0}{\widehat{\text{SE}}} = \frac{11.68 - 0}{3.33} = 3.51.$$

Question: Is this plausible if $H_0 : \mu = 0$ is true, i.e., if the sampling distribution of $\hat{t}$ is $\mathcal{N}(0, 1)$?

p-values

The p-value

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

$$\text{p-value} = \mathbb{P}(|Z| \geq |\hat{t}|),$$

where $Z \sim \mathcal{N}(0, 1)$.

The p-value answers the question: "If H_0 is true, is your observation $\hat{t}$ plausible?"

For our example: p-value $= \mathbb{P}(|Z| \geq 3.51) \approx 0.0004$.

Interpreting the p-value

p-value ≈ 0.0004:

- *Interpretation*: If the true mean delay were zero, there's only a 0.04% chance we would observe a sample mean as extreme as 11.68 minutes.
- This is *very unlikely*, i.e., our evidence is inconsistent with the truth of H_0.

How NOT to interpret the p-value

Important: The p-value IS NOT the probability that H_0 is true!

We *cannot* make statements about the "chance" of H_0 being true, because the true μ is *not random*.

The p-value:

- **IS**: Probability of observing a test statistic as extreme as $\hat{t}$, *given* H_0 is true.
- **IS NOT**: Probability H_0 is true, *given* that you observed $\hat{t}$.

The first is a *frequentist* statement; the second is a *Bayesian* statement, which we will see later in the course.

Rejecting the null:
Hypothesis testing as binary classification

Can we reject the null?

In hypothesis testing, we determine whether the evidence allows us to *reject the null* H_0.

Formally we choose a *significance level* α (e.g., $\alpha = 0.05$).

Decision rule: Reject H_0 at significance level α if p-value $\leq \alpha$.

A smaller α means we need *stronger evidence* (more extreme $\hat{t}$) to reject the null.

For our example, p-value $= 0.0004 < 0.05 \implies$
we reject $H_0 : \mu = 0$ at significance level $\alpha = 0.05$.

A note on terminology [$*$]

The statistic we are using is called a *t statistic*; it is also sometimes called a *studentized statistic*. ("Studentizing" refers to normalizing by the estimated standard error $\widehat{\text{SE}}$.)

The hypothesis test defined by the decision rule on the preceding slide is often referred to as a "t-test", though the formal definition of a t-test requires assuming the data generating process is *exactly* normal (not asymptotically normal). (See appendix for more on t-tests.)

Another name for the rule on the previous slide is the *Wald test*.

Hypothesis testing as classification

This decision rule makes hypothesis testing into *binary classification*!

The "truth" (unknown):

- H_0 is true
- H_0 is false

Our decision (based on data):

- Reject H_0
- Don't reject H_0

Just like a classifier, we can make mistakes...

Two types of errors

In hypothesis testing, we can make two types of mistakes:

	Reject H_0	*Don't Reject* H_0
H_0 True	**False Positive**	True Negative
H_0 False	True Positive	**False Negative**

- *Type I error* (False Positive): Reject H_0 when it's actually true
- *Type II error* (False Negative): Fail to reject H_0 when it's actually false

The meaning of "significance"

Recall we called α the "significance level".

Key result: The significance level α *is exactly the false positive (Type I error) probability!*

$$\alpha = \mathbb{P}(\text{reject } H_0 | H_0 \text{ true}).$$

In other words, α is a "tuning knob" that controls how often we make false positive errors.

Why is α = false positive probability?

If we reject H_0 when the p-value $\leq \alpha$ ...

Why is α = false positive probability?

If we reject H_0 when the p-value $\leq \alpha$ ...

then we reject H_0 if the chance of seeing a test statistic as extreme as our observation is $\leq \alpha$...

Why is α = false positive probability?

If we reject H_0 when the p-value $\leq \alpha$ ...

then we reject H_0 if the chance of seeing a test statistic as extreme as our observation is $\leq \alpha$...

which has chance *exactly* α if H_0 is true!

So if we reject when p-value ≤ 0.05, we reject with probability 0.05 when H_0 is true $\implies$ 5% false positive probability.

Power

Power is the probability of *correctly* rejecting H_0 when it's false – informally:

$$\text{Power} = \mathbb{P}(\text{reject } H_0 | H_0 \text{ false}) = 1 - \mathbb{P}(\text{Type II error}).$$

Problem: H_0 can be false in many ways! Power depends on what the true μ actually is.

To formally compute power, we need a *specific alternative* $\mu = \mu_a \neq \mu_0$; see appendix.

Tradeoff between Type I and Type II errors

Reducing α (being more conservative):

- *Decreases* false positive probability
- *Increases* false negative probability, i.e., decreases power

Increasing α (being less conservative):

- *Increases* false positive probability
- *Decreases* false negative probability, i.e., increases power

This is the fundamental tradeoff in hypothesis testing!

Increasing power: The sample size

When the truth is $\mu_a \neq \mu_0$, then for large n the t statistic for testing $H_0 : \mu = \mu_0$ has sampling distribution that is approximately:

$$\mathcal{N}\left(\frac{\mu_a - \mu_0}{\hat{\sigma}/\sqrt{n}}, 1\right).$$

With $\mu_a \neq \mu_0$, as $n \to \infty$, the magnitude of this statistic $\to \infty$.
So at any α, we become increasingly likely to (correctly) reject the null!

In other words: *power increases as the sample size grows*.

Connection to confidence intervals

Equivalent decision rule

The t statistic has approximately a $\mathcal{N}(0,1)$ distribution under H_0.

Therefore the decision rule "reject if the p-value is $\leq \alpha$" is equivalent to rejecting H_0 when:

$$|\hat{t}| > z_{\alpha/2},$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the $\mathcal{N}(0,1)$ distribution.

For $\alpha = 0.05$: $z_{0.025} \approx 1.96$.

For our example: $|3.51| > 1.96 \implies$ Reject H_0 at significance level $\alpha = 0.05$.

Duality between tests and confidence intervals

There is an important connection between hypothesis tests and confidence intervals:

Note that we reject the null with $\alpha = 0.05$ exactly when $|\hat{t}| > 1.96$.

Since $\hat{t} = (\overline{Y} - \mu_0)/\widehat{\text{SE}}$, this is equivalent to rejecting the null exactly when:

$$\mu_0 \notin [\overline{Y} - 1.96\widehat{\text{SE}}, \overline{Y} + 1.96\widehat{\text{SE}}]$$

I.e., we reject $H_0 : \mu = \mu_0$ if μ_0 is not in the 95% confidence interval.

General α [$*$]

Recall the $(1-\alpha)$ CI is: $[\overline{Y} - z_{\alpha/2}\widehat{\text{SE}}, \overline{Y} + z_{\alpha/2}\widehat{\text{SE}}]$.

We reject $H_0 : \mu = \mu_0$ when $|\hat{t}| > z_{\alpha/2}$, which means:

$$\left| \frac{\overline{Y} - \mu_0}{\widehat{\text{SE}}} \right| > z_{\alpha/2}.$$

This is equivalent to: $|\overline{Y} - \mu_0| > z_{\alpha/2}\widehat{\text{SE}}$.

As a result, a significance level α test rejects $H_0 : \mu = \mu_0$ *if and only if* μ_0 is *not* in the $(1-\alpha)$ confidence interval.

Applications to other estimators

Other asymptotically normal estimators

The same approach works for *any asymptotically normal estimator*. Examples:

- Sample mean (CLT)
- Ordinary least squares (OLS) linear regression (an M-estimator under Assumptions (A1)-(A3))
- Logistic regression (an MLE under Assumptions (B1)-(B2))
- Other M-estimators (see Lecture 9)

Generalizing the approach

Suppose $\hat{\theta}$ is an estimator for θ with estimated standard error $\widehat{\mathsf{SE}}$, such that for large n the sampling distribuiton is approximately normal:

$$\hat{\theta} \approx \mathcal{N}\left(\theta, \widehat{\mathsf{SE}}^2\right).$$

To test the null hypothesis $H_0 : \theta = \theta_0$, use the t statistic:

$$\hat{t} = \frac{\hat{\theta} - \theta_0}{\widehat{\mathsf{SE}}}.$$

Now testing of the null hypothesis H_0 is *identical* to the preceding discussion.

Example 1: OLS linear regression (standard output)

OLS produces the following regression table output:

```
lm(formula = price ~ 1 + livingArea + bedrooms, data = sh)
...
              Estimate Std. Error t value Pr(>|t|)
...
livingArea     125.405      3.527  35.555  < 2e-16 ***
...
```

The `t value` is the t statistic value; and `Pr(>|t|)` is the p-value.

But exactly which null hypothesis is being tested here?

Example 1: OLS linear regression (standard output)

A regression table's output always reports t statistics and p-values for the null hypothesis H_0 that *the corresponding coefficient is zero*.

In this case, the p-value on `livingArea` is extremely small, which means that the observed t statistic is extremely unlikely if the true coefficient on `livingArea` was zero.

Important note: This calculation *assumes* Assumptions (A1)-(A3) hold!

Example 2: OLS linear regression (nonzero null)

In fact, we can use the same table to test *other* null hypotheses.

E.g., can test $H_0 : \beta_{\text{livingArea}} = 120$. Form the t-statistic:

$$\hat{t} = \frac{\hat{\beta}_{\text{livingArea}} - 120}{\widehat{\text{SE}}_{\text{livingArea}}} = \frac{125.405 - 120}{3.527} = 1.532.$$

Corresponding p-value (from normal distribution) = $0.125 > 0.05$, so we do not reject the null if $\alpha = 0.05$.

Alternatively, note that 120 is *inside* the 95% confidence interval $[118.49, 132.32]$, so we don't reject the null if $\alpha = 0.05 \implies$ same answer by duality.

A note on OLS linear regression with normal errors [$*$]

The previous discussion on OLS relied on the fact that OLS is an M-estimator under Assumptions (A1)-(A3), so that the estimated coefficient is *asymptotically normal* when n is large, with mean that is the true coefficient.

When, in addition, Assumption (A4) holds – i.e., the errors in the population model are *normally distributed* – then for *any* sample size n, the t statistic has a sampling distribution that is *Student's t distribution*.

As previously noted, the test we have been doing in this case is called a *t test*; see appendix).

Note that Student's t distribution is very close to the $\mathcal{N}(0, 1)$ distribution even for small n (e.g., $n > 50$), so for practical purposes the distinction usually doesn't matter.

Example 3: Logistic regression

Logistic regression on CORIS dataset:

```
glm(formula = chd ~ ., family = "binomial", data = coris)
...
             Estimate Std. Error z value Pr(>|z|)
...
sbp          0.133308   0.117452   1.135 0.256374
...
ldl          0.360181   0.123554   2.915 0.003555 **
...
```

The `z value` is the t statistic value; and `Pr(>|z|)` is the p-value – again for the null hypothesis that *the corresponding coefficient is zero*.

Important note: Again, Assumptions (B1)-(B2) have to hold to have asymptotic normality!

Statistical significance notation

Common notation:

- *** means p-value < 0.001 ("significant at 99.9% level")
- ** means p-value < 0.01 ("significant at 99% level")
- * means p-value < 0.05 ("significant at 95% level")

Common language: "The coefficient on `livingArea` is statistically significant at the 99.9% level."

Interpreting statistical significance

Important caveats:

1. Statistically significant $\neq$ *practically* significant
   - Even tiny effects can be "statistically significant" with large n
2. Not statistically significant $\neq$ unimportant
   - Small n or large $\widehat{\text{SE}}$ can hide important effects
3. Require assumptions that ensure asymptotic normality of coefficients
   - If (A1)-(A3) or (B1)-(B2) violated, tests may be misleading

Appendix: z-tests and t-tests

Asymptotic normality and hypothesis testing [$*$]

So far, we've relied on *asymptotic normality* (large n approximation):

- $\hat{\theta} \approx \mathcal{N}(\theta, \widehat{\text{SE}}^2)$ for large n
- Test statistic $\hat{t} \approx \mathcal{N}(0, 1)$ under H_0

What if we know the *exact* distribution for finite n?

The z-test [$*$]

Setup:

- Data $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. (exactly normal)
- We *know* σ^2 (rare in practice!)
- Want to test $H_0 : \mu = \mu_0$

Test statistic (using true SE $= \sigma/\sqrt{n}$):

$$z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}.$$

Under H_0: $z \sim \mathcal{N}(0, 1)$ *exactly* for any n (not just asymptotically).

This is called a *z-test*. Rarely applicable because we almost never know σ.

The t-test [$*$]

Setup:

- Data $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. (exactly normal)
- We *don't know* σ^2 (the usual case)
- Want to test $H_0 : \mu = \mu_0$

Test statistic (using estimated $\widehat{\mathrm{SE}} = \hat{\sigma}/\sqrt{n}$):

$$t = \frac{\bar{Y} - \mu_0}{\hat{\sigma}/\sqrt{n}}.$$

Under H_0: $t \sim$ *Student's t-distribution* with $(n-1)$ degrees of freedom.

Student's t-distribution [$*$]

The t-distribution:

- Is symmetric around 0 (like $\mathcal{N}(0,1)$)
- Has heavier tails than $\mathcal{N}(0,1)$ (accounts for uncertainty in $\hat{\sigma}$)
- Converges to $\mathcal{N}(0,1)$ as $n \to \infty$
- Is very close to $\mathcal{N}(0,1)$ even for $n \geq 30$

For large n, t-tests and asymptotic tests give nearly identical results.

Appendix: Power computation [∗]

Computing power [$*$]

To compute power, we need a *specific alternative* $\theta = \theta_a \neq \theta_0$.

If $\theta = \theta_a$, then:

$$\hat{t} = \frac{\hat{\theta} - \theta_0}{\widehat{\text{SE}}} \approx \mathcal{N}\left(\frac{\theta_a - \theta_0}{\widehat{\text{SE}}}, 1\right).$$

Power at θ_a:

$$\text{Power}(\theta_a) = \mathbb{P}\left(|Z| > z_{\alpha/2}\right) \text{ where } Z \sim \mathcal{N}\left(\frac{\theta_a - \theta_0}{\widehat{\text{SE}}}, 1\right).$$

Power increases with effect size [$*$]

Power depends on:

1. *Effect size*: $|\theta_a - \theta_0|$ (how far is truth from null?)
2. *Standard error*: $\widehat{\text{SE}}$ (how much uncertainty?)
3. *Significance level*: α

Larger $|\theta_a - \theta_0|/\widehat{\text{SE}}$ → Higher power.

Appendix: One-sided vs two-sided tests [$*$]

Two-sided tests (what we've done so far) [$*$]

Two-sided test:

- $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$
- p-value $= \mathbb{P}(|Z| \geq |\hat{t}|)$ where Z is $\mathcal{N}(0, 1)$
- Tests whether θ differs from θ_0 in *either direction*

This is the most common type of test in practice.

One-sided tests [$*$]

One-sided test (upper tail):

- $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$
- p-value $= \mathbb{P}(Z \geq \hat{t})$ where Z is $\mathcal{N}(0, 1)$

One-sided test (lower tail):

- $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$
- p-value $= \mathbb{P}(Z \leq \hat{t})$ where Z is $\mathcal{N}(0, 1)$

Again, reject H_0 if the p-value is smaller than α.

Two-sided vs. one-sided tests [$*$]

A one-sided test only tests deviations in *one direction*.

They are less commonly used, since if $H_0 : \theta = \theta_0$ is not true, we typically don't have any reason to know in advance whether in fact $\theta > \theta_0$ or $\theta < \theta_0$.

Two-sided vs. one-sided tests [$*$]

Note that $P(|Z| \geq |\hat{t}|) = \mathbb{P}(Z \geq |\hat{t}|) + \mathbb{P}(Z \leq -|\hat{t}|)$.

Therefore, at a fixed significance level α, it is *easier* to reject the null using a one-sided test.

This practice is sometimes viewed as "inflating" significant results, which is one of the reasons that two-sided testing is standard practice.