Popular hypothesis tests

A cook book for popular hypothesis tests

Recap

Data science pipeline

1. Question driven
2. Collect data
3. Explorative data analysis
4. Select or design an appropriate model
5. Analysis

Two types of errors

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

The null hypothesis

- Confident when we reject H_0 (small type 1 error rate)

- Less confident when we accept H_0 (unknown type 2 error rate). In this case, we usually say "we don't have enough evidence to reject H_0)

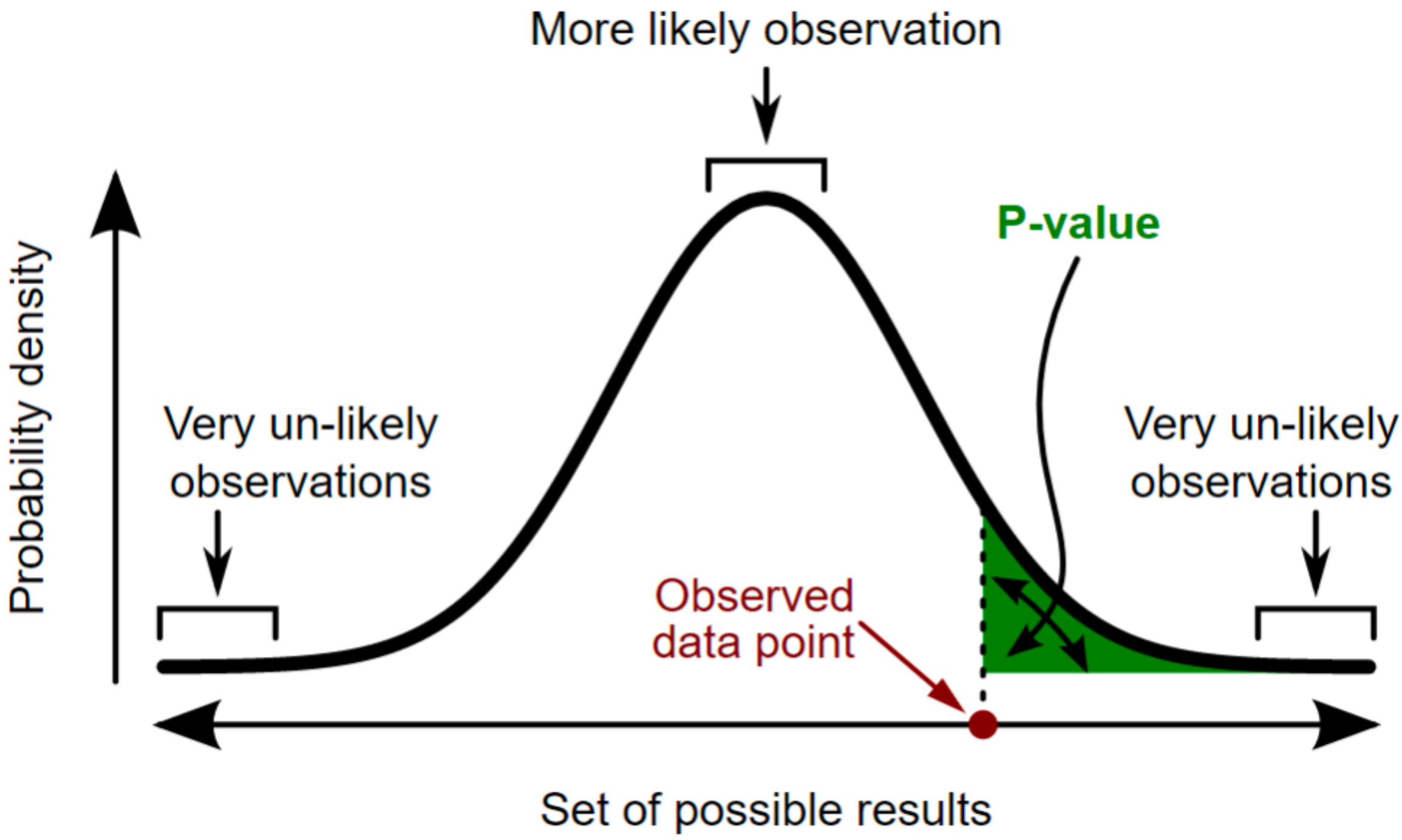
- 決策錯誤代價嚴重的答案放在 H_0

- 想驗證正確的答案放在 H_A

Null distribution

- Test statistics are random
- The sampling distribution of a test statistic when the null hypothesis was true
- Useful for determining the cut-off between the two hypothesis and controls the type 1 error rate

The p–value


More likely observation
P-value
Probability density
Very un-likely observations
Very un-likely observations
Observed data point
Set of possible results

Agenda

- Parametric tests
- Nonparametric tests
- The “rpy2” package

Motivative example: A/B testing

- Example: google ads
- Example: UX/UI design (Reference: Designing with Data)

Parametric tests

Two sample t–test

- Assume $X_i \overset{iid}{\sim} N\left(\mu_1, \sigma_1^2\right)$, $Y_i \overset{iid}{\sim} N\left(\mu_2, \sigma_2^2\right)$, and $X_i \perp Y_j$

- Test on means:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_A : \mu_1 \neq \mu_2$$

or

$$H_0 : \mu_1 \leq \mu_2 \quad \text{vs} \quad H_A : \mu_1 > \mu_2$$

- scipy.stats.ttest_ind for two–tailed tests

One–sample t–test

- Assume $X_1, \ldots, X_n \overset{iid}{\sim} N\left(\mu, \sigma^2\right)$

- Test on mean:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_A : \mu \neq \mu_0$$

or $$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_A : \mu > \mu_0$$

- $T = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$
- scipy.stats.ttest_1samp for two–tailed tests

Paired t–test

- Assume $(X_1, Y_1), \cdots, (X_n, Y_n)$ are paired sample with $X_i \overset{iid}{\sim} N\left(\mu_1, \sigma_1^2\right)$ and $Y_i \overset{iid}{\sim} N\left(\mu_1 2, \sigma_2^2\right)$

- Test on means:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_A : \mu_1 \neq \mu_2$$

or

$$H_0 : \mu_1 \leq \mu_2 \quad \text{vs} \quad H_A : \mu_1 > \mu_2$$

- Let $Z_i = X_i - Y_i$ and use t–test to test on Z_i

- scipy.stats.ttest_rel for two–tailed tests

F–test for variance

- Assume $X_i \overset{iid}{\sim} N\left(\mu_1, \sigma_1^2\right)$, $Y_j \overset{iid}{\sim} N\left(\mu_2, \sigma_2^2\right)$, and $X_i \perp Y_j$

- Test on variances:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad H_A : \sigma_1^2 \neq \sigma_2^2$$

$$\text{or} \qquad H_0 : \sigma_1^2 \leq \sigma_2^2 \quad \text{vs} \quad H_A : \sigma_1^2 > \sigma_2^2$$

- $F = \dfrac{s_X^2}{s_Y^2} \sim F_{n-1,m-1}$

Bartlett's test

- Assume $X_{ij} \overset{iid}{\sim} N\left(\mu_j, \sigma_j^2\right)$ for $i = 1,\ldots,n_j$ and $j = 1,\ldots,J$

- Test on means:

$$\begin{cases} H_0\colon \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_J^2 \\ H_A\colon \sigma_i^2 \neq \sigma_j^2 \text{ for some } i \neq j \end{cases}$$

- scipy.stats.bartlett

Nonparametric tests

Mann–Whitney U test

- Nonparametric version of one-sided two sample t-test without normal assumptions (but less powerful)

- scipy.stats.mannwhitneyu

Wilcoxon signed-rank test

- Nonparametric version of one-sample t-test and paired t-test without normal assumption (but less powerful)
- scipy.stats.wilcoxon for two-tailed tests

Pearson's chi-squared goodness-of-fit test

- Test for fit of a discrete distribution

- scipy.stats.chisquare

Chi-squared independence test

- Test of statistical independence for two discrete random variables

- scipy.stats.chi2_contingency

Kolmogorov–Smirnov test

- Test for fit of a continuous distribution
- scipy.stats.kstest for one sample
- scipy.stats.ks_2samp for two samples

Shapiro–Wilk test

- Test for normality
- scipy.stats.shapiro

Runs test

- Test for randomness: check if a sequence of (binary) random sample is mutually independent

- statsmodels.sandbox.stats.runs.Runs

- The “randtests” package in R

rpy2

Readings

- Chapters 3 and 4 of “Introductory Statistics with Randomization and Simulation“
- Chapter 3 of “Practical Statistics for Data Scientists: 50 Essential Concepts”
- Chapter 6 of “Python for Data Analysis (2nd edition)”

Homework: PM 2.5 Concentrations

Use an appropriate hypothesis test to verify if the PM 2.5 concentrations at 忠明 station in 2017 is higher than those in 2014. Check the model assumptions of a test before applying it.

Data source

Multiple comparison

- Applying multiple tests simultaneously is less preferable
- Bonferroni correction
- Controlling the false discovery rate