# Statistical inferencing

# Agenda

- Reviews
- Decisions under uncertainties
- Hypothesis testing
- Recap

# Reviews

# Disciplines

- Statistics
  - scientific insights
- Machine learning
  - prediction
- Big data
- Data assimilation

# ML v.s. Stat

| Machine learning | Statistics |
| --- | --- |
| network, graphs | model |
| weights | parameters |
| learning | fitting |
| generalization | test set performance |
| supervised learning | regression/classification |
| unsupervised learning | density estimation, clustering |
| large grant = $1,000,000 | large grant = $50,000 |
| nice place to have a meeting:<br>Snowbird, Utah, French Alps | nice place to have a meeting:<br>Las Vegas in August |

# Data science pipeline

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

# Ask questions data science can solve

- Interact with your customer
- Take the following courses:
  - 資料分析
  - 統計諮詢
  - …

# Is your data ready to answer your question?

1. Big data (population)
2. Experimental design
3. Random sampling

# Exploratory Data Analysis

- Making sense of the data
- Discover patterns of data (e.g. distribution, relationships)
- Check assumptions (e.g. normality)
- Very helpful for statistical modeling

# Statistical models

Transform your question to a simplified mathematical formulation

- stochastic
- contains unknown parameters
- answer the question by model parameters

# Analysis

- Estimate the model parameters and
  - machine learning: predict future events by the fitted model
  - statistical analysis: answer the scientific problem by model parameters

# Example: breast cancer

1. Questions

   - machine learning: diagnosis of a future patient
   - statistical analysis: is the tumor size an important factor for diagnosis?

2. Data: breast cancer dataset

3. Explorative data analysis

4. Select or design an appropriate model: use logistic regression to model how likely is the tumor a cancer given the features derived from a medical image

5. Analysis: estimate the model parameters and

   - machine learning: use the fitted logistic regression model to diagnose a future patient

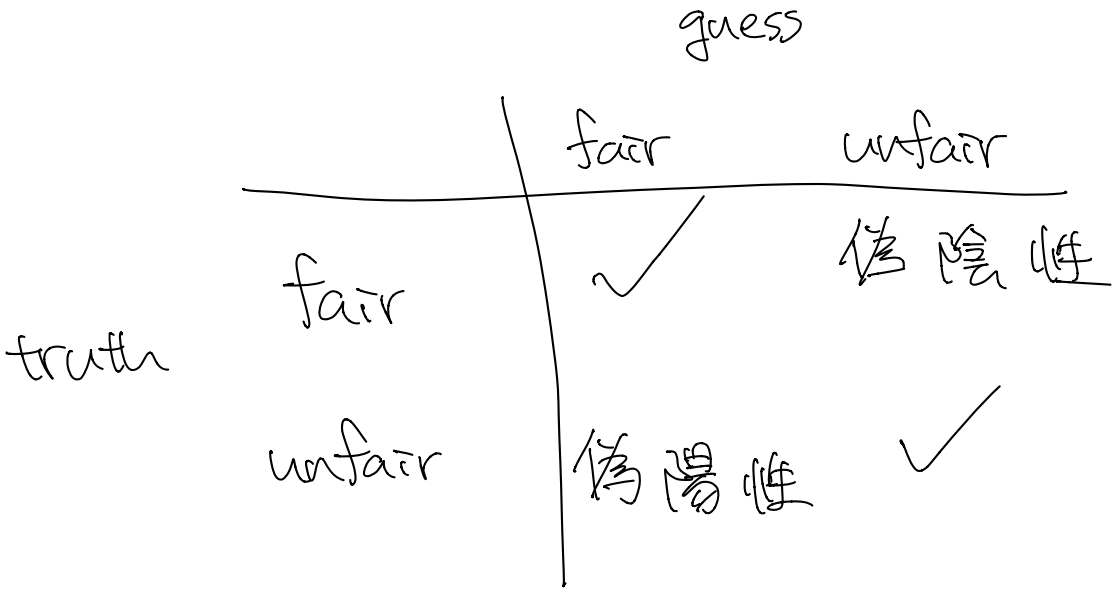
   - statistical analysis: is the regression coefficient of tumor size positive? (hypothesis testing)

# Decision under uncertainties

# Example: is the coin fair?

- Flipping a fair coin for 10 times, you'll expect to observe 5 heads.
- In fact, you may observe 1, 2, ... , 10 heads due to randomness.
- How to make a guess (or decision) under randomness?

# Decision errors



# Hypothesis testing

# Parameters of interest

Transform your question to a proper statistical model with certain parameters of interest

- example: coin tossing

$$X_i = \begin{cases} 1, & \text{with prob. } p \\ 0, & \text{otherwise} \end{cases}, \quad p = \frac{1}{2} ?$$

- example: breast cancer dataset

$\beta_{\text{tumor size}} > 0$ ?

# Hypothesis testing

Only capable for binary decisions

- null hypothesis $H_0$
- alternative hypothesis $H_1$ or $H_A$

# Test statistic

- A statistic specially designed for hypothesis testing
- Derived from point estimations of parameters of interest

# 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* | ✓ |

# Two types of errors

- Good decision: small decision error rate
- Unfortunately, minimizing both types of errors simultaneously is impossible.
  - Example: coin tossing

# Hypothesis testing framework

- Alternatives:
  - control the upper bound of type 1 error (by significance level $\alpha$) and minimize type 2 error (frequentist approach)
  - minimize the total cost of decision errors (bayesian approach)

# Two types of errors

| | | Decision | |
|---|---|---|---|
| | | fail to reject $H_0$ | reject $H_0$ |
| **Truth** | $H_0$ true | ✓ | $P(\text{Type 1 Error}) \le \alpha$ |
| | $H_A$ true | $P(\text{Type 2 Error}) = \beta$ | ✓ $1-\beta$ |

power
檢定力

good decision:

① $P(\text{type I error}) \le \alpha$

② $\max(1-\beta)$

$P(\text{type I error})$

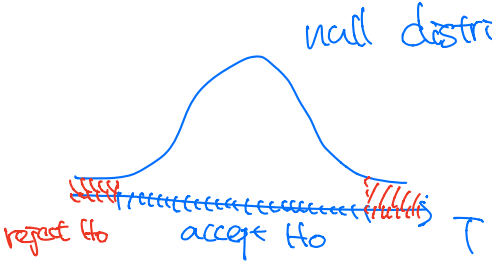
$= P(\text{reject } H_0 \mid H_0 \text{ is true}) \le \alpha$

# 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



# Example: coin tossing

$X_1, X_2, \cdots, X_{10} \overset{iid}{\sim} \text{Bernoulli}(p)$

$$\hat{p} = \frac{\#\{X_i = 1\}}{10} = \frac{\sum_{i=1}^{10} X_i}{10}$$

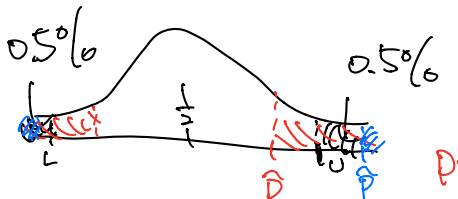
$$\hat{p} \xrightarrow{d} N\left(p, \frac{p(1-p)}{10}\right)$$

when $H_0$ is true,

$$\hat{p} \xrightarrow{d} N\left(\frac{1}{2}, \frac{1}{40}\right)$$ null distribution of $\hat{p}$

$$\begin{cases} H_0: p = \frac{1}{2} \\ H_A: p \neq \frac{1}{2} \end{cases}$$

$\alpha = 1\%$




reject $H_0$ if $\hat{p} < L$ or $\hat{p} > U$

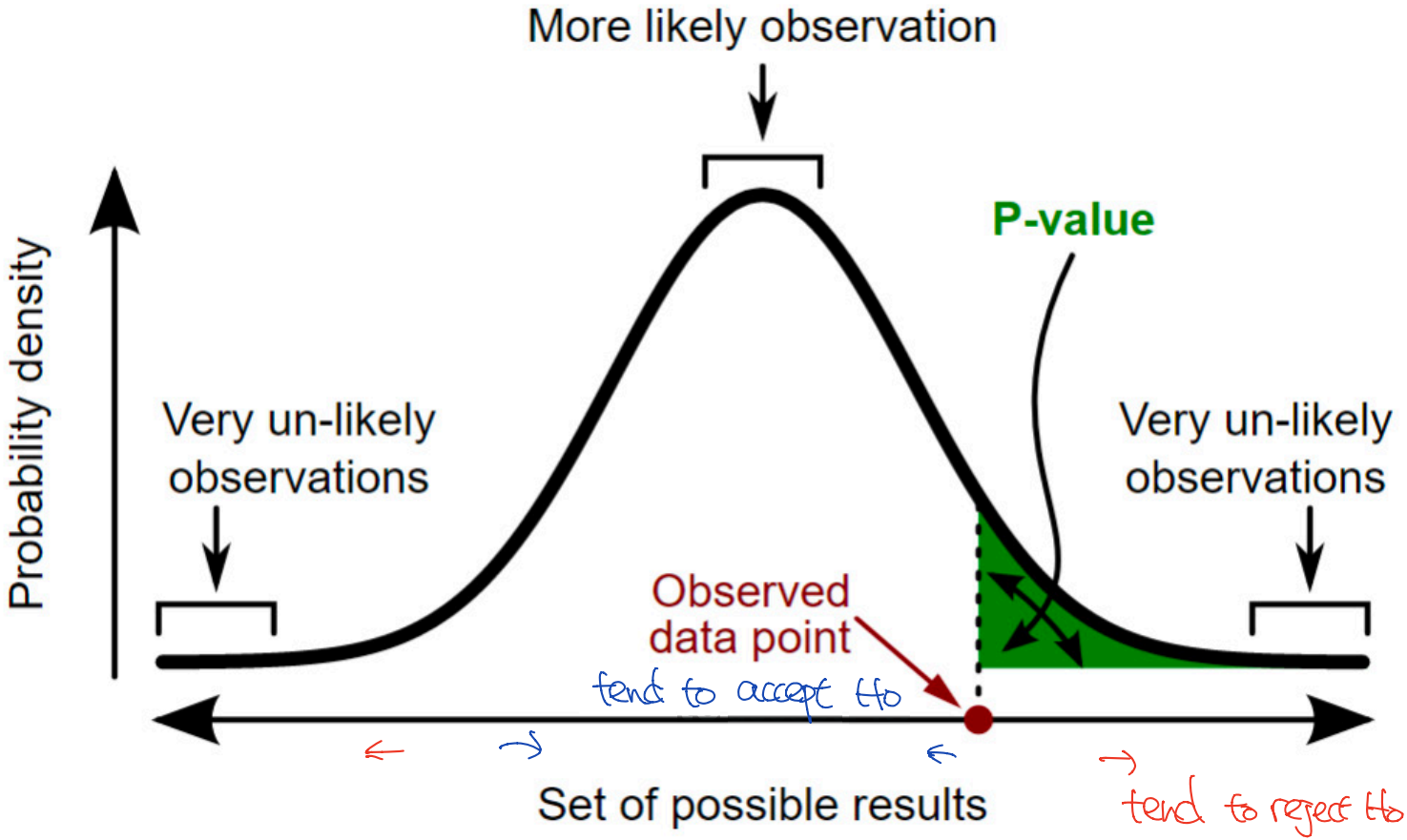
p-value $> \alpha$ p-value $< \alpha$ reject $H_0$ p-value $< \alpha$

# The p–value

- How unlikely to observe the sample and thus the value of testing statistic when the null hypothesis is true

$$p\text{-value} = P(T \mid H_0)$$

# The p–value


More likely observation
P-value
Very un-likely observations
Very un-likely observations
Probability density
Observed data point
tend to accept Ho
Set of possible results
tend to reject Ho

# Recap

- The data science pipeline
- Hypothesis testing framework
  - two types of errors
  - control type 1 error rate by significance level $\alpha$
  - null distribution
  - p–value