Lecture 4: Inferences

IST5573

統計方法 Statistical methods

2016/10/05

Scientific experimental process

Define the question

- Do mice fed with chow (control) and high fat (treatment) have different weights?

Identify the population

- In the Jackson Lab,
- weights of female mice fed with chow $x_1, x_2, \cdots, x_m$, and
- weights of female mice fed with high fat $y_1, y_2, \cdots, y_n$

Population parameters

- $\mu_X = \frac{1}{m}\sum_{i=1}^{m} x_i \, , {\sigma_X}^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_X)^2$
- $\mu_Y = \frac{1}{n}\sum_{i=1}^{n} y_i \, , {\sigma_Y}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_Y)^2$

Distribution of the population

- ecdf: $F_x(a) = \Pr(x \leq a)$ (proportion of $x_1, x_2, \cdots, x_m$ that are smaller than a), histogram of $x_1, x_2, \cdots, x_m$
- ecdf: $F_y(a) = \Pr(y \leq a)$ (proportion of $y_1, y_2, \cdots, y_n$ that are smaller than a), histogram of $y_1, y_2, \cdots, y_n$

Design the (random) experiment

- Buy M $(< m)$ female mice fed with chow from the Jackson Lab
- Buy N $(< n)$ female mice fed with high fat from the Jackson Lab

Various random variables from the experiment

- $X_1, X_2, \cdots, X_M$, $\bar{X} = \frac{1}{M}\sum_{i=1}^{M} X_i$, ${s_X}^2 = \frac{1}{M-1}\sum_{i=1}^{M}(X_i - \bar{X})^2$
- $Y_1, Y_2, \cdots, Y_N$, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$, ${s_Y}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$

Distributions of random variables

- If the population data are available, Monte Carlo simulations can be used to generate the distributions of all possible values of the random variables.
- If we do not have the access to the population, some statistical theories (e.g., Central Limit theory) can help us approximate these distributions with some known distributions (e.g., normal, t).

Statistical inferences

- **After performing the experiment**, use the observed sample data to predict the population parameters.
  - Point estimation, confidence interval
  - Hypothesis testing

Statistical inferences

- **After performing the experiment**, we obtain one sample of the random variables:
$$X_1, X_2, \cdots, X_M \rightarrow x_1, x_2, \cdots, x_M$$
$$Y_1, Y_2, \cdots, Y_N \rightarrow y_1, y_2, \cdots, y_N$$
- Statistical inference is the mathematical theory that permits you to approximate the **population parameters** with only the **observed values from your sample**: $x_1, x_2, \cdots, x_M$ and $y_1, y_2, \cdots, y_N$.

Two methods of inferences

- **Estimation**
  - point estimation
  - interval estimation: **confidence interval**
- **Hypothesis testing**

Point estimation

Population parameters	Random variables	Point estimates
μ_X, μ_Y	$\bar{X}, \bar{Y}$	$\bar{x}, \bar{y}$
${\sigma_X}^2, {\sigma_Y}^2$	${s_X}^2, {s_Y}^2$	${s_x}^2, {s_y}^2$
$\mu_Y - \mu_X$	$\bar{Y} - \bar{X}$	$\bar{x} - \bar{y}$

- Point estimates = random variables plugged in the observed sample values
- Use the point estimates as our guess of the population parameters

Confidence interval (CI)

- Point estimation provides us the **effect size** (i.e., the observed difference).
- A confidence interval includes information about your **estimated effect size** and the **uncertainty associated with this estimate**.

CI for population mean

- A 95% confidence interval (we can use percentages other than 95%) is a **random interval** with a 95% probability of falling on the parameter we are estimating.
- Keep in mind that saying 95% of random intervals will fall on the true value (our definition above) is **not the same** as saying there is a 95% chance that the true value falls in our interval.

- To construct it, we note that the CLT tells us that $\sqrt{N}(\bar{X} - \mu_X)/s_X$ follows a normal distribution with mean 0 and SD 1 (i.e., the standard normal distribution $N(0,1)$). This implies that:

$$\Pr\left(-z_{0.025} \leq \frac{\sqrt{N}(\bar{X} - \mu_X)}{s_X} \leq z_{0.025}\right) = 0.95$$

  where $z_{0.025}$ is the upper 2.5 percentage point of the standard normal distribution (i.e., $\Pr(Z > z_{0.025}) = 0.025$ with $Z \sim N(0,1)$).
- Note In R, one can get the value of $z_{0.025}$ by qnorm(1- 0.025)

- Now do some basic algebra to clear out everything and leave μ_X alone in the middle and you get that the following event:

$$\bar{X} - z_{0.025}(\frac{s_X}{\sqrt{N}}) \leq \mu_X \leq \bar{X} + z_{0.025}(\frac{s_X}{\sqrt{N}})$$

  has a probability of 95%.

- Be aware that it is the edges of the interval $\bar{X} \pm z_{0.025}(\frac{s_X}{\sqrt{N}})$, not μ_X, that are random.

- The definition of the confidence interval is that 95% of **random intervals** will contain the true, fixed value μ_X.
  - For a specific interval that has been calculated, (e.g., **the interval calculated by plugging in observed sample values**), the probability is either 0 or 1 that it contains the fixed population mean μ_X.
- Now, we will show how to construct a confidence interval for **the population mean of control female mice**.
  - RMD_example 4.1

CI when small sample size

- We use the CLT to create our confidence intervals, and with $N = 5$ it may not be as useful an approximation.
- This mistake affects us in the calculation of the upper 2.5 percentage point $z_{0.025}$, which assumes a normal distribution.

- Statistical theory offers another useful result. If **the distribution of the population is normal**, then we can work out the exact distribution of $\sqrt{N}(\bar{X} - \mu_X)/s_X$ as a **t-distribution**.
- The t-distribution is a much more complicated distribution than the normal. The t-distribution has a parameter called **degrees of freedom**.

- Then the 95% CI for μ_X is

$$\bar{X} - t_{0.025,N-1}(\frac{s_X}{\sqrt{N}}) \le \mu_X \le \bar{X} + t_{0.025,N-1}(\frac{s_X}{\sqrt{N}})$$

where $t_{0.025,N-1}$ is the upper 2.5 percentage point of the t-distribution with degree of freedom = $N-1$ (i.e., $\Pr\left(t > t_{0.025,N-1}\right) = 0.025$ with $t \sim t_{N-1}$).

  - We can confirm these with a simulation: RMD_example 4.2

Hypothesis testing

- **Statistical hypothesis**: A statement about the **parameters** of one or more **populations**.
- **Test of a hypothesis**:
  - A procedure leading to a decision about a particular hypothesis
  - Hypothesis testing procedures rely on using the information in a **random sample from the population of interest** to judge that the hypothesis is true or false.

Setup the hypotheses

- We consider two hypotheses:
  - The **null hypothesis** H_0
    - Our original knowledge
    - In our mouse diet experiment, H_0: $\mu_X = \mu_Y$
  - The **alternative hypothesis** H_a
    - The hypothesis we seek to prove
    - In our mouse diet experiment, H_a: $\mu_X \neq \mu_Y$

Decision in hypothesis testing

	Truth (you never know)	
Decision	H_0 is true	H_a is true
Not reject H_0 (negative)	Right decision	**Type II error (false negative)**
Reject H_0 (positive)	**Type I error (false positive)**	Right decision

Hypothesis testing procedure

1. Decide the **significance level**:

$$\alpha = \Pr(\text{type I error})$$
$$= \Pr(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

   Usually set $\alpha = 0.05 \text{ or } 0.01$

2. Decide the **test statistic**:

$$t = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{{s_Y}^2}{N} + \frac{{s_X}^2}{M}}} \text{ (the t} - \text{statistic)}$$

   The characteristic used for making the decision

3. Set the **decision rule**:

   Reject H_0 if $|t| > z^*$ (i.e., $|t|$ is big)

Type I, II errors

- In the hypothesis testing procedure, we set the significant level (i.e., the probability of making type I error) as 0.05 or 0.01. **Note that the 0.05 and 0.01 cut-offs are arbitrary!**
- The reason we don't use infinitesimal cut-offs to avoid type I errors at all cost is that there is another error we can commit: to not reject the null when we should (the type II error).

- Thus, in **1.**, we fix the type I error rate at a level that we are comfortable with (e.g., 0.05/0.01).
- Then, via some statistical theories, we decide **2.** and **3.** such that we will commit type II errors as unlikely as possible.

Hypothesis testing procedure (cont'd)

4. Decide z^* in the decision rule:

Select z^* that satisfies
$\Pr(|t| > z^* | H_0 \text{ is true}) = \alpha$

We call $(-\infty, -z^*) \cup (z^*, \infty)$ **the rejection (critical) region**.

Null distribution of t-statistics

- To obtain z^*, we need to know the distribution of the t-statistic when H_0 is true (when there is no difference between μ_X and μ_Y).
- Because we **have access to the control population**, we can actually observe as many values as we want of the t-statistics when the diet has no effect.

- In our mouse diet experiment, we can do this by randomly sampling 24 control mice, giving them the same diet, and then recording the t-statistic between two randomly split groups of 12 and 12. Here is this process written in R code:
  - RMD_example 4.3
- These values are what we call the **null distribution** of t-statistics.
- With the null distribution of t-statistics, we can then calculate z^*.
  - RMD_example 4.3

Normal approximation for the null distribution of t-statistics

- In practice, we **do not have access to the population**.
- Fortunately, we can use CLT approximation for the null distribution of t-statistics.
- When the null is true (i.e., $\mu_Y - \mu_X = 0$) and N, M are large, by CTL

$$Z = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{{\sigma_Y}^2}{N} + \frac{{\sigma_X}^2}{M}}} \sim N(0, 1)$$

- Typically, we don't know the population standard deviations: σ_X and σ_Y. We can use the sample standard deviations s_X and s_Y to **estimate** them.
- We can redefine

$$t = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{{s_Y}^2}{N} + \frac{{s_X}^2}{M}}} \sim N(0, 1)$$

- We call this a **t-statistic**.
- We can then set $z^* = z_{\alpha/2}$ (the upper $100(\alpha/2)$ percentage point of $N(0,1)$)
  - RMD_example 4.4

The t-distribution

- The CLT relies on large samples, what we refer to as **asymptotic results**.
- When the CLT does not apply, there is another option that does not rely on asymptotic results.

- Statistical theory offers another useful result. If **the distribution of the population is normal**, then we can work out the exact distribution of the t-statistic as a **t-distribution**.
- R has a nice function t.test that actually computes everything.

t-distributions in practice

- In our mouse diet experiment, there is a problem. CLT works for large samples, but is 12 large enough?
- The z^* we computed is only a valid approximation if the assumptions hold, which do not seem to be the case here.
- We will now demonstrate how to obtain a valid z^* in a t-test using the t-distribution.
  - RMD_example 4.5

Hypothesis testing procedure (cont'd)

5. Make the decision based on the observed sample:

Calculate the value of the test statistic based on the observed sample t_0, then

If $|t_0| > z^*$, reject H_0 (which implies H_a is true)**-- statistically significant**

If $|t_0| \leq z^*$, not reject H_0 (there is not enough evidence to reject H_0)

6. Calculate the **p-value** of the test, then

If p-value $< \alpha$, reject H_0**-- statistically significant**

If p-value $\geq \alpha$, not reject H_0

Another way of making the decision

p-value

- When the null hypothesis is true (there is no diet effect), the probability that we see a test statistic t as **extreme** as the one we observed t_0: either larger than $|t_0|$ or smaller (more negative) than $-|t_0|$ (i.e.,

$$\text{p} - \text{value}$$
$$= \Pr(t > |t_0| | H_0 \text{ is true}) + \Pr(t < -|t_0| | H_0 \text{ is true})$$

  - **This is what is known as the p-value**.

- If we have access to the control population, the p-value can be calculate as the following: RMD_example 4.3
- We can also use the normal or t approximation for the p-value: RMD_examples 4.4, 4.5
- Notice that the decision results from **5.** and **6.** are **the same**! RMD_examples 4.3, 4.4

Connection between CI and p-value

- We can form a 95% CI for $\mu_Y - \mu_X$ with the observed difference $\bar{Y} - \bar{X}$:

$$(\bar{Y} - \bar{X}) - z_{0.025}\left(\sqrt{\frac{{s_Y}^2}{N} + \frac{{s_X}^2}{M}}\right) \le \mu_Y - \mu_X \le (\bar{Y} - \bar{X}) + z_{0.025}\left(\sqrt{\frac{{s_Y}^2}{N} + \frac{{s_X}^2}{M}}\right)$$

- **If interval does not include 0** (when $H_0\!:\ \mu_Y - \mu_X = 0$), this implies

$$(\bar{Y} - \bar{X}) - z_{0.025}\left(\sqrt{\frac{{s_Y}^2}{N} + \frac{{s_X}^2}{M}}\right) > 0 \text{ or } (\bar{Y} - \bar{X}) + z_{0.025}\left(\sqrt{\frac{{s_Y}^2}{N} + \frac{{s_X}^2}{M}}\right) < 0$$

$$\frac{\bar{Y}-\bar{X}}{\sqrt{\frac{{s_Y}^2}{N}+\frac{{s_X}^2}{M}}} > z_{0.025} \text{ or } \frac{\bar{Y}-\bar{X}}{\sqrt{\frac{{s_Y}^2}{N}+\frac{{s_X}^2}{M}}} < -z_{0.025}$$

which suggests **rejecting H_0 (p-value < 0.05).**

- Example in t-tests
  - RMD_example 4.6

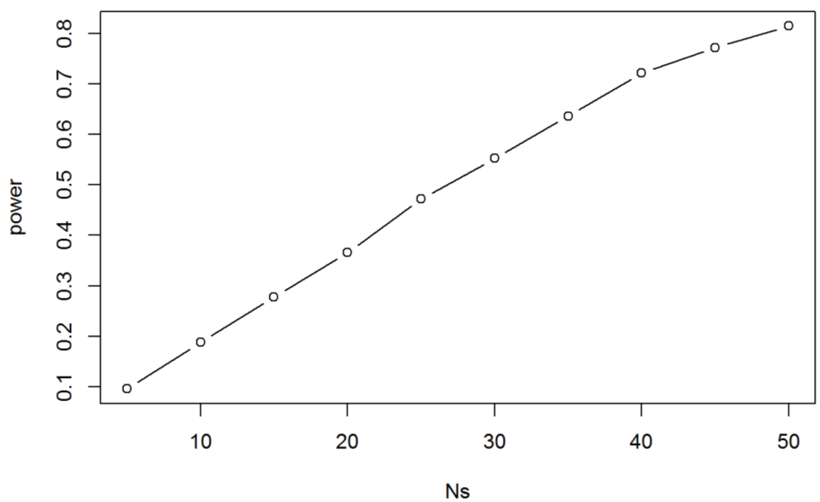
Power calculations

- **Power** is the probability of rejecting the null when the null is false.
- $\text{Power} = 1 - \Pr(\text{type II error})$
- The hypothesis testing procedure fixes the type I error rate at a level that we are comfortable with (e.g., 0.05)., and then adopts some statistical theories to seek the test statistic and decision rule that will **maximize the power** of the test.

- In calculating the power, "when the null is false" is a complicated statement because it can be false in many ways.
  - $\Delta = \mu_Y - \mu_X$ could be anything and the power actually depends on this parameter.
  - It also depends on the standard error of your estimates which in turn depends on **the sample size** and the **population standard deviations**.
- In practice, we don't know these so we usually report power for several plausible values of $\Delta, \sigma_X, \sigma_Y$ and various sample sizes.
- Statistical theory gives us formulas to calculate power. The pwr package performs these calculations for you.

- If we have the access to the population, then we can calculate the powers via the **Monte Carlo simulation**.
    - RMD_example 4.7

Sample size and power



- As we can see that **the power improves with the sample size.**
- **In the planning stage of the study**, one can use this relationship to determine the appropriate sample size that can reach the power set by your study.