Lecture 2: Random variables, distributions, and Central Limit Theorem

IST5573

統計方法 Statistical methods

2016/9/21

Random Variables and Distributions

Random variables (隨機變數)

- **Background**: This data was produced by ordering 24 female mice from The Jackson Lab and randomly assigning either chow or high fat (hf) diet. After several weeks, the scientists weighed each mice and obtained this data: femaleMiceWeights.csv
- **Question**: We are interested in determining if following a high fat makes mice heavier after several weeks.
- Examining femaleMiceWeights.csv
  - RMD_example 2.1

- So the hf diet mice are about 10% heavier in average.
- Are we done? Why do we need p-values and confidence intervals? **The reason is that these averages are random variables.**
  - They can take many values.
  - If we repeat the experiment, we obtain 24 new mice from The Jackson Laboratory and, after randomly assigning them to each diet, we get a different mean. Every time we repeat this experiment, we get a different value. **We call this type of quantity a random variable.**

Population (母體)

- Imagine that we actually have the weight of **all control (fed with chow) female mice** in the data femaleControlsPopulation.csv
- In Statistics, we refer to this as the **population** for the control mice.
- These are all the control mice available from which we sampled.
  - Note that **in practice we do not have access to the population**.

- We can continue to sample 12 mice repeatedly from the population, and see how the average changes.
  - RMD_example 2.2

Distributions (分佈)

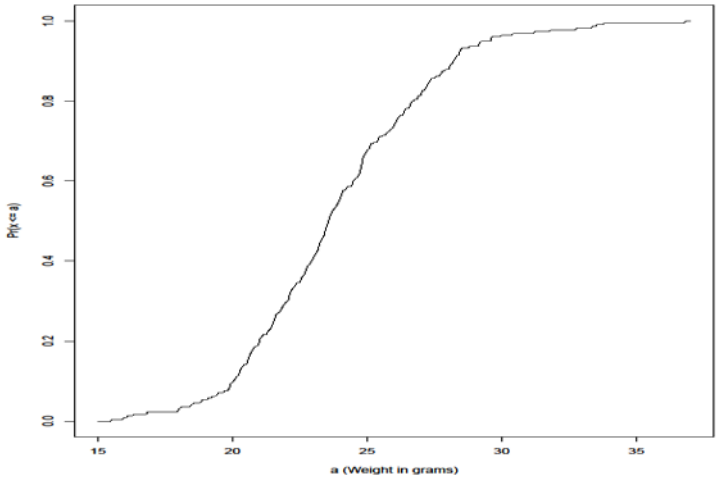
- The simplest way to think of a distribution is as a compact description of many numbers.
  - The weights of female mice in the control population: RMD_example 2.3
- Scanning through these numbers, we start to get a rough idea of what the entire list looks like, but it is certainly inefficient.
- We can quickly improve on this approach by defining and visualizing a distribution.

Cumulative distribution function (累積分佈函數)

- For all possible values of a, we can calculate the proportion of numbers in our list that are below a. We use the following notation:

$$F(a) \equiv \Pr(x \leq a)$$

- This is called the **cumulative distribution function** (CDF). When the CDF is derived from data, as opposed to theoretically, we also call it the empirical CDF (ECDF). We can plot empirical $F(a)$ versus a like this:
  - RMD_example 2.4
  - Note we defined Pr(a) by counting cases



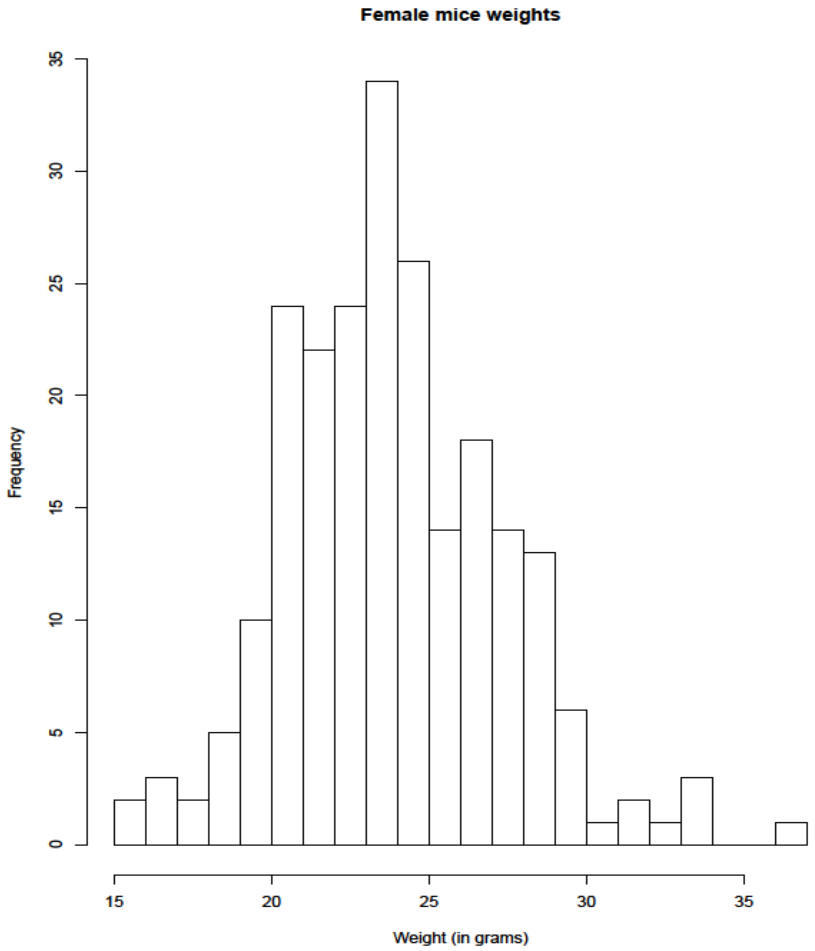
Histograms (直方圖)

- The proportion of values in intervals:

$$\Pr(a < x \leq b) = F(b) - F(a)$$

- Plotting these intervals versus their proportions as bars is what we call a **histogram**.
  - It is a more useful plot because we are usually more interested in intervals, such and such percent are between 22 grams and 25 grams, etc., rather than the percent less than a particular weight.

- A histogram of female mouse weights
  - RMD_example 2.5



Probability distribution

- Summarizing lists of numbers is one powerful use of distribution.
- An even more important use is describing **the possible outcomes of a random variable**.
- Unlike a fixed list of numbers, we don't actually observe all possible outcomes of random variables, so instead of describing proportions, we describe **probabilities**.

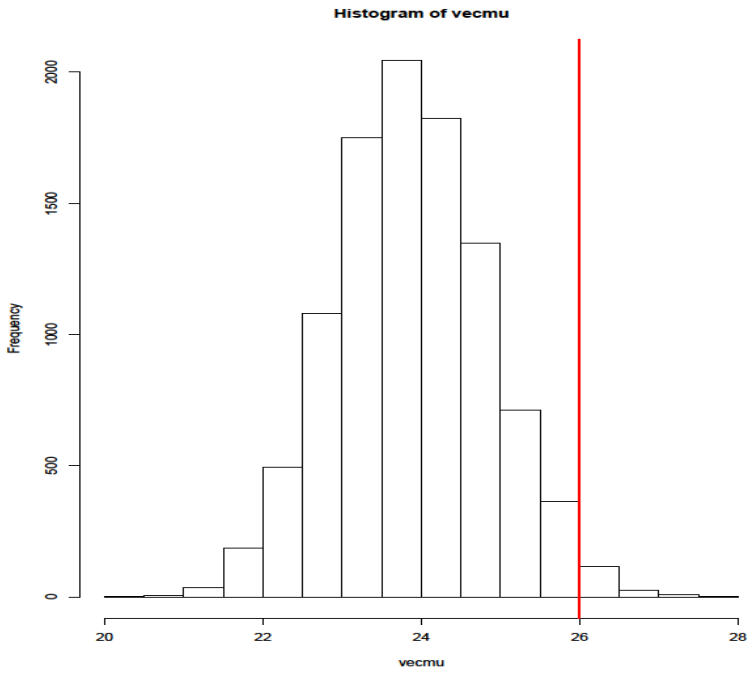
- For instance, if we pick a random weight from our list, then the probability of it falling between a and b is denoted with:

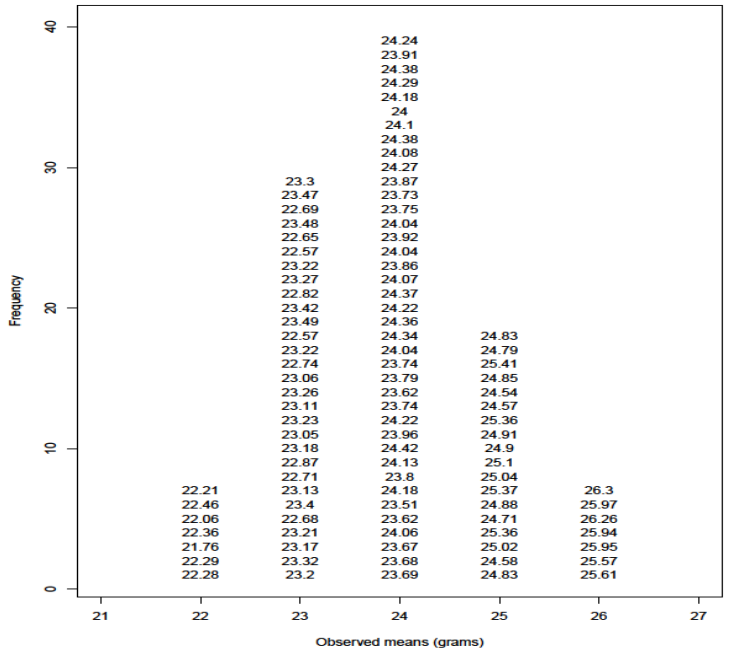
$$\Pr(a < X \leq b) = F(b) - F(a)$$

  - Note that the X is now capitalized to distinguish it as a random variable.
- The equation above defines the probability distribution of the random variable.

- Knowing this distribution is incredibly useful in science.
  - For example, in the mice example above, we can run what is called a **Monte Carlo simulation** and obtain 10,000 means under the control population. This will form the distribution of the means of mouse weights from the control population.
  - With this distribution, we can compute the probability of observing a value as large as, for example, 26 grams.

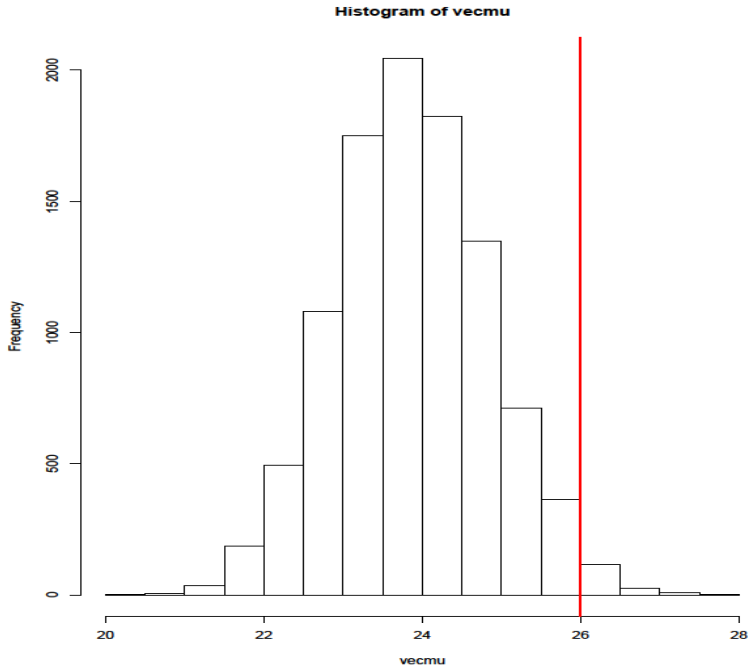
- Example: the probability distribution of the mean of mouse weights
  - RMD_example 2.6





Normal distribution

- While we defined $\Pr(a)$ by counting cases, we will learn that, in some circumstances, mathematics gives us formulas for $\Pr(a)$ that save us the trouble of computing them as we did here.
- One example of this powerful approach uses the normal distribution approximation.



- The probability distribution we see approximates one that is very common in nature: the **bell curve**, also known as the **normal distribution** or **Gaussian distribution**.

- When the histogram of a list of numbers approximates the normal distribution, we can use a convenient mathematical formula to approximate the proportion of values or outcomes in any given interval:

$$\Pr(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx$$

Here μ and σ are referred to as the mean and the standard deviation of the population.

- The formula is stored as pnorm in R which sets a to $-\infty$, and takes b as an argument.
- If this normal approximation holds for our list, then the population mean and standard deviation of our list can be used in the formula above.
- An example of this would be when we noted above that only 1.4% of values on the control distribution were above 26 grams.
  - RMD_example 2.7

- A very useful characteristic of this approximation is that one only needs to know μ and σ to describe the entire distribution.
- From this, we can compute the proportion of values in any interval.

Summary

- To make the above calculation, we needed to buy all the mice available from The Jackson Laboratory and performing our experiment repeatedly to define the distribution. Yet this is not something we can do in practice.
- Statistical inference is the mathematical theory that permits you to approximate this with only the data from your sample, i.e. the original 24 mice.

Populations, Samples, Estimates and Central Limit theorem

Population parameters

- A first step in statistical inference is to understand what **population** you are interested in.
- In the mouse weight example, we have two populations: female mice on control diets and female mice on high fat diets, with weight being the outcome of interest.
- We consider this population to be fixed, and the randomness comes from the sampling.

- The mouse weight population:
  - We happen to have the weights of all the mice of this type.
  - RMD_example 2.8

- We denote the weights of the control population as $x_1, x_2, \cdots, x_m$. Here, $m = 225$ is the size of the control population.
- The weights of the high fat diet population are $y_1, y_2, \cdots, y_n$, where $n = 200$.
- We can define summaries of interest for these populations, such as the mean

$$\mu_X = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \mu_Y = \frac{1}{n}\sum_{i=1}^{n} y_i$$

the variance

$$\sigma^2{}_X = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_X)^2 \qquad \sigma^2{}_Y = \frac{1}{n}\sum_{i=1}^{n} (y_i - \mu_Y)^2$$

- We refer to such quantities that can be obtained from the population as **population parameters**.
- The question we started out asking can now be written mathematically is

  H_0: $\mu_Y - \mu_X = 0$ (**the null hypothesis**)
- Although in our illustration we have all the values and can check if this is true, in practice we do not.
  - For example, in practice it would be prohibitively expensive to buy all the mice in a population.

Sample estimates

- Here we learn how taking a **sample** permits us to answer our questions.
- In the previous chapter, we obtained samples of 12 mice from each population.
- We represent data from samples with **capital letters to indicate that they are random**. So the samples are

$$X_1, X_2, \cdots, X_M \text{ and } Y_1, Y_2, \cdots, Y_N$$

where $M = N = 12$.

- Since we want to know if $\mu_Y - \mu_X = 0$, we consider the sample version: $\bar{Y} - \bar{X} = 0$ with:

$$\bar{X} = \frac{1}{M}\sum_{i=1}^{M} X_i \qquad \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

- Note that this difference of averages is also a random variable.
- Previously, we learned about the behavior of random variables with an exercise that involved repeatedly sampling from the original distribution.
  - In this particular case it would involve buying 24 mice over and over again. This is not an exercise that we can execute in practice.

Central Limit Theorem (CLT)

- Here we described the mathematical theory that mathematically relates $\bar{X}$ to μ_X and $\bar{Y}$ to μ_Y.
- Specifically, we will describe how the **central limit theorem** permits us to use an approximation to answer this question.

CTL

- **When the sample size (M) is large**, the average $\bar{X}$ of a random sample follows a normal distribution centered at the population mean μ_X and with standard deviation equal to the population standard deviation σ_X, divided by the square root of the sample size M

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{M}}\right) \text{ or } \frac{\bar{X} - \mu_X}{\left(\sigma_X / \sqrt{M}\right)} \sim N(0,1)$$

- We don't know the population standard deviation σ_X. We can use the sample standard deviation s_X to estimate it

$$s^2{}_X = \frac{1}{M-1}\sum_{i=1}^{M}(X_i - \bar{X})^2$$

- We can redefine the **t-ratio**

$$t = \frac{\bar{X} - \mu_X}{\left({}^{s_X}/_{\sqrt{M}}\right)} \sim N(0,1)$$

CLT in practice

- Let's use our data to see how well the central limit theorem approximates sample averages from our data.
- We will leverage our entire population dataset to compare the results we obtain by actually sampling from the distribution to what the CLT predicts.
  - RMD_example 2.9

t-test

- Now go back to our question of $\mu_Y - \mu_X = 0$.
- **When the null is true** (i.e., $\mu_Y - \mu_X = 0$) and N, M are large, by CTL

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{{\sigma^2}_Y}{N} + \frac{{\sigma^2}_X}{M}}} \sim N(0,1)$$

- 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

$${s^2}_X = \frac{1}{M-1}\sum_{i=1}^{M}(X_i - \bar{X})^2 \quad {s^2}_Y = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

- We can redefine

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

- We call this a **t-statistic**.

Null distribution of t-statistics

- With our experimental data femaleMiceWeights.csv, we can calculate **the observed t-statistic**: obststat
- How do we know that this obststat is due to the diet?
- What happens if we give all 24 mice the same diet? Will we see a difference this big?
  - Statisticians refer to this scenario as the **null hypothesis** (H_0: $\mu_Y - \mu_X = 0$).

- 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.
- 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 2.10
- These values are what we call the **null distribution** of t-statistics.

p-value

- The values in null form the null distribution.
- Only a small percent of the 10,000 simulations are bigger than obststat.
- When there is no diet effect (i.e., the null hypothesis is true), we see a t-statistic as big as the one we observed only 2.4% of the time.
  - **This is what is known as the p-value**.

Normal approximation for the p-value

- In practice, we **do not have access to the population**.
- Fortunately, we can use CLT approximation for the null distribution of t-statistics and then compute the p-value.
  - The CLT tells us that under the null hypothesis for large sample sizes, the t-statistic is approximately normal with mean 0 (the null hypothesis) and SD 1 (we divided by its SE).
  - So now to calculate a p-value all we need to do is ask: how often does a N(0,1) (standard normal) random variable exceed obststat?

t-tests in practice

- We will now demonstrate how to obtain a p-value in a t-test in practice with R.
  - RMD_example 2.11

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**.
- The t-distribution is a much more complicated distribution than the normal. The t-distribution has a parameter called **degrees of freedom**.
- R has a nice function t.test that actually computes everything.

t-distributions in practice

- In our mouse weight example, there is a problem. CLT works for large samples, but is 12 large enough?
- The p-value 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 p-value in a t-test using the t-distribution.
  - RMD_example 2.12