Lecture 3: Monte Carlo simulations

BTBI30081

統計應用方法Applied Methods in Statistics

2025/3/5

Discrete probability

- For **categorical data**, the subset of probability is referred to as discrete probability.
- Example (RMD_example 3.1)

  If I have 2 red beads and 3 blue beads in an urn and I pick one at random what is the probability of picking a red one?

  - The answer is 2/5 or 40%.
- Definition: The probability of an event happening is the **proportion** of times it happens if we **repeat** the choice **over and over independently** and under the same condition.

Monte Carlo simulations

- Computers provide a way to actually perform the experiment (example) described above
- **Random number generators** permit us to mimic the process of picking at random
- Use R function sample to randomly pick up elements
    - Sampling **with replacement** (RMD_example 3.2)
    - Sampling **without replacement** (RMD_example 3.3)

Continuous probability

- Example: The mouse diets (RMD_example 3.4)

  This data was produced by ordering 24 female mice from the Jackson Lab, where 12 mice were fed with chow diet and 12 high fat (hf) diet. The scientists weighed each mice and obtained this data.

Continuous probability

- For a list of numeric (continuous) values, such as mice weights, it is not useful to construct a distribution that defines a proportion to each possible outcome.
  - If we measure every single mouse in a very large population of size $10{,}000$ with extremely high precision, because no two mice are exactly the same weight, we need to assign the proportion 0.0001 to each observed value and attain no useful summary at all.
  - When defining probability distributions, it is not useful to assign a very small probability to every single weight.

- For numeric (continuous) data, it is much more practical to define a function that operates on intervals rather than single values.
    - The standard way of doing this is using the cumulative distribution functions (CDF)

$$F(a) = \Pr(X \leq a)$$

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

The probability density

- For categorical distributions we can define the probability of a category.
  - For example a roll of a die, let's call it X, can be 1, 2, 3, 4, 5 or 6. The probability of 4 is defined as

    $$\Pr(X = 4) = 1/6$$

    The CDF can then easily be defend:

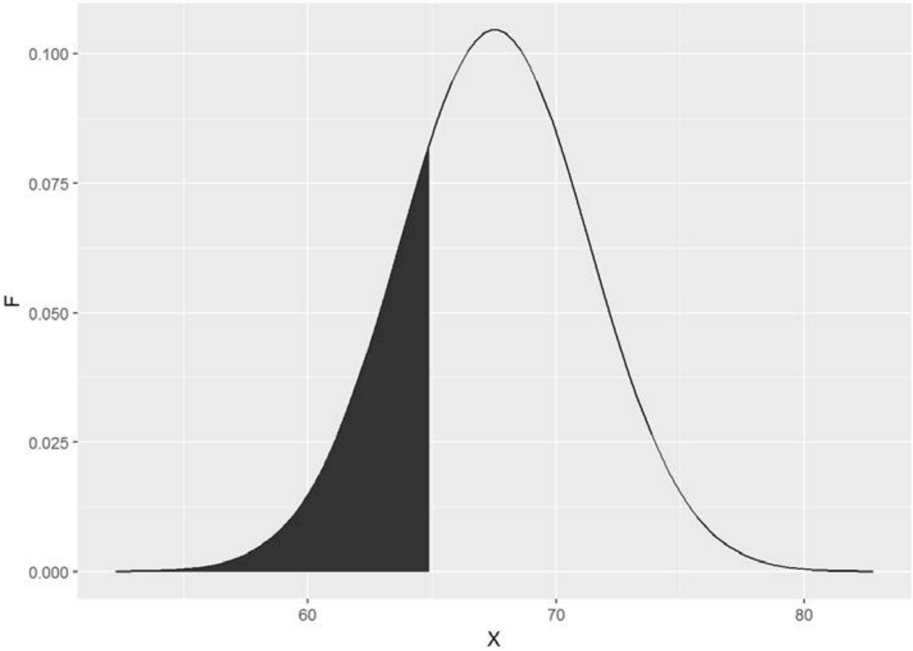
$$F(4) = \Pr(X \leq 4) =$$
$$\Pr(X = 4) + \Pr(X = 3) + \Pr(X = 2) + \Pr(X = 1) = 2/3$$

- With continuous distributions the probability of a singular value $\Pr(X = x)$ is not even defined
- Although for continuous distributions the probability of a single value is not defined, there is a theoretical definition that has a similar interpretation.

- The **probability density** at x is defined as the function $f(x)$ such that

$$F(a) = \Pr(X \leq a) = \int_{-\infty}^{a} f(x)\,dx$$

- You can think of $f(x)$ as a curve for which the area under that curve up the value a gives you the probability of $X \leq a$



Normal approximation

- The **normal distribution** is a useful approximation to many naturally occurring distributions of **continuous data**, including that of weight.

- The cumulative distribution for the normal distribution is defined by a mathematical formula:

$$F_{\mathrm{norm}}(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx$$

    - In R can be obtained with the function pnorm
- We say that a random quantity is normally distributed with mean μ and standard deviation σ if it's cumulative distribution is defined by $F_{\mathrm{norm}}(a)$.
- RMD_example 3.5

Monte Carlo simulations

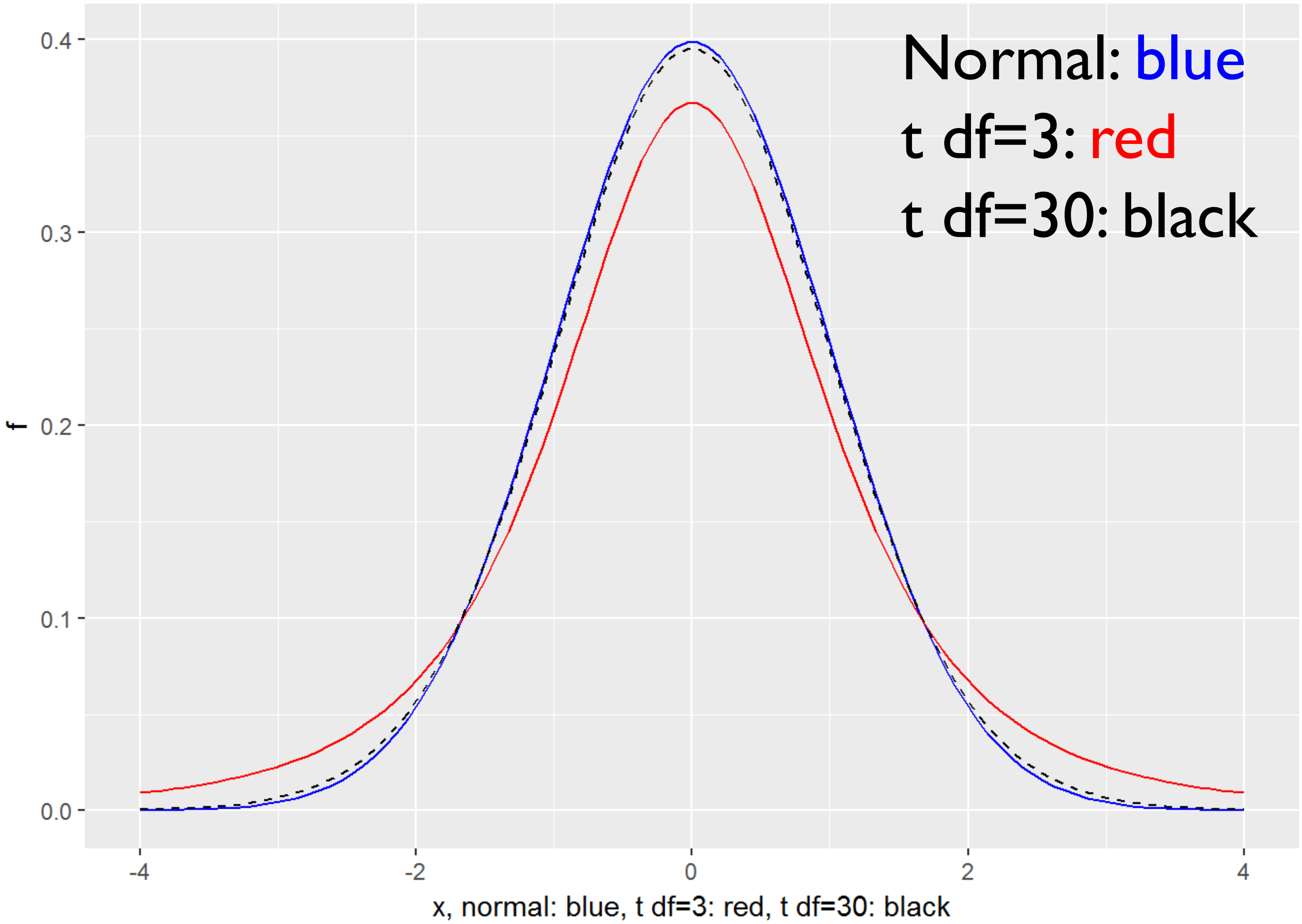
- R provides functions to generate normally distributed outcomes. Specifically, the rnorm function takes three arguments: size, mean (defaults to 0), and standard deviation (defaults to 1) and produced these random numbers.
    - An example of how we could generate data that looks like our reported mouse weights from the normal distribution: RMD_example 3.6

Other continuous distributions

- The normal distribution is not the only useful theoretical distribution.
- Other continuous distribution that we may encounter are the **student-t**, **chi-squared**, **exponential**, **gamma**, **beta**, and **beta-binomial**.

- R provides functions to compute the density, the quantiles, the cumulative distribution functions and to generate Monte Carlo simulations.
- R uses a convention that let’s remember the names. Namely using the letters d, q, p and r in front of a shorthand for the distribution.
- RMD_example 3.7

Density of normal and t



Normal: blue

t df=3: red

t df=30: black

Parametric simulations

- Simulations can also be used to check theoretical or analytical results.
- Many of the theoretical results we use in statistics are based on **asymptotic**s: they hold when **the sample size goes to infinity**.
- In practice, we never have an infinite number of samples so we may want to know how well the theory works with our actual sample size.
- Sometimes we can answer this question analytically, but not always. Simulations are extremely useful in these cases.

- As an example, let's use a Monte Carlo simulation to compare the CLT to the t-distribution approximation for different sample sizes.
    - RMD_example 3.8

Parametric simulations for the observations

- What we did in RMD_example 3.8 was a type of Monte Carlo simulation **when we had access to population data** and generated samples at random.
- In practice, we do not have access to the entire population.
  - The reason for using the approach here was for educational purposes.

- When we want to use Monte Carlo simulations in practice, it is much more typical to **assume a parametric distribution** and generate a population from this, which is called a **parametric simulation**.
  - This means that we take parameters estimated from the real data (here the mean and the standard deviation), and plug these into a model (here the normal distribution).
  - RMD_example 3.9

Parametric vs. nonparametric simulations

- **Parametric simulations** assumed that the random sample was from some known distribution, and we then generated random numbers from this distribution to simulate characteristics of estimators of interest.
- In reality, the distribution that the random sample belongs to is either unknown or too complicated to generate data from.
- We can consider methods that allow us to generate the underlying distribution using the **observed sample** (i.e., **nonparametric (Monte Carlo) simulations**).

What is the bootstrap?

- Bootstrap methods are a class of nonparametric simulations that estimate the distribution of a population by **resampling**.
- Resampling methods treat an observed sample as a finite population, and random samples are generated (resampled) from it to estimate population characteristics and make inferences about the sampled population.
- Bootstrap methods are often used when the distribution of the target population is not specified; the sample is the only information available.

(Nonparametric) bootstrap

- Suppose the observed sample is $x_1, x_2, \cdots, x_M$, θ is the population characteristics we want to estimate (e.g., the population standard deviation σ_X), and we use $\hat{\theta}$ (a random variable) to estimate it (e.g., the sample standard deviation s_X)

1. For each bootstrap replicate, indexed by $b = 1, \cdots, B$:
   a) Draw M values with **replacement** from the set $\{x_1, x_2, \cdots, x_M\}$
   b) Compute $\hat{\theta}$ with values sampled in a), denoted as $\hat{\theta}^{(b)}$
2. The bootstrap estimate can be obtained via $\hat{\theta}^{(1)}, \cdots, \hat{\theta}^{(B)}$

Bootstrap estimate of variance

- The variance of your estimator (i.e., $\mathrm{var}(\hat{\theta})$) (e.g., the variance of the sample standard deviation $\mathrm{var}(s_X)$) can be used to evaluate how accurate your estimator.
- However, this variance is typically difficulty to obtain theoretically. The bootstrap can help here.
- The bootstrap estimate of **variance** of $\hat{\theta}$

$$\widehat{\mathrm{var}}\left(\hat{\theta}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{(b)} - \bar{\hat{\theta}}\right)^2$$

where $\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$

- The bootstrap estimate of **standard error** of $\hat{\theta}$ (i.e., $\sqrt{\mathrm{var}(\hat{\theta})}$) is the square root of the variance estimate.

Bootstrap estimate of confidence interval

- For the bootstrap estimate of 95% confidence interval for θ (we can use percentages other than 95%), one can use

the upper 2.5 percentage point and the upper 97.5 percentage point of $\hat{\theta}^{(1)}, \cdots, \hat{\theta}^{(B)}$

- RMD_example 3.10: Bootstrap for using the sample standard deviation $\hat{\theta} = s_X$ to estimate the population standard deviation $\theta = \sigma_X$

Permutation tests

- Wan to perform a hypothesis testing
- Suppose we have a situation in which none of the standard mathematical statistical approximations apply.
- In practice, we do not have access to all values in the population so we can't perform a simulation as done above.
- Permutation tests can be useful in these scenarios.

- For example, we have computed a summary statistic, the difference in mean, to determine if the observed difference was significant.
- In previous courses, we showed parametric approaches (CLT) that can help.
- **Permutation tests** take advantage of the fact that **if we randomly shuffle the cases and control labels, then the null is true.**
    - Here is how we generate a null distribution by shuffling the data 1,000 times to 1,000 null mean differences: RMD_example 3.11
    - How many of the null means are bigger than the observed value? That proportion would be the p-value for the null.

Notes

- Keep in mind that there is no theoretical guarantee that the null distribution estimated from permutations approximates the actual null distribution.
  - For example, if there is a real difference between the populations, some of the permutations will be unbalanced and will contain some samples that explain this difference.
- This is why permutations result in conservative p-values. For this reason, when we have **few samples**, we **cannot do permutations**.