Chapter 6: Sampling and Sampling Distributions

Wang Shujia

Contents

1 Basic Concepts of Statistical Inference

Statistical inference

The ***objective*** of statistical analysis is to gain knowledge about certain properties in a population that are of interest to the researcher.

- ***Statistics*** is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data.
- A ***population*** is the collection or set of all objects or measurements that are of interest to the collector.
- The ***sample*** is a subset of data selected from a population. The size of a sample is the number of elements in it.
- A ***statistical inference*** is an estimate, a prediction, a decision, or a generalization about the population based on information contained in a sample.

Descriptive Statistics and Inferential Statistics

- The methods consisting mainly of organizing, summarizing, and presenting data in the form of tables, graphs, and charts are called ***descriptive statistics***.
- The methods of drawing inferences and making decisions about the population using the sample are called ***inferential statistics***.

Observational Studies and Experimental Studies

- Generally, data in ***observational studies*** are collected only by monitoring what occurs, while studies where the researchers assign treatments to cases are called ***experiments***.
- When this assignment includes randomization, e.g. using a coin flip to decide which treatment a patient receives, it is called a ***randomized experiment***.
- Making ***causal*** conclusions based on experiments is often reasonable. However, making the same causal conclusions based on observational data can be treacherous and is not recommended.

Example 1. Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer.

Does this mean sunscreen causes skin cancer?

- Sun exposure is what is called a ***confounding*** variable, which is a variable that is correlated with both the explanatory and response variables.

Simulation and Sampling

- ***Simulation*** studies typically generate numbers according to a researcher-specified model.
  - The effects of natural disasters, such as earthquakes, on buildings and highways are often modeled with simulation.
- ***Sampling*** is the process of performing repetitions of an experiment and gathering data from it.
  - Simple random sampling
  - Stratified sampling
  - Cluster sampling
- A ***random sample*** of size n from a population is a set of n independent and identically distributed (iid) observable random variables $X_1, X_2, \ldots, X_n$.

Parameters and Estimators

- A ***parameter*** is a characteristic of a population.
  - Parameters are treated as constants in *classical statistics* and as random variables in *Bayesian statistics*.
- A ***statistic*** is a characteristic of a sample. More exactly, it is a function T of observable random variables $X_1, X_2, \ldots, X_n$ that does not depend on any unknown parameters.
  - Statistics (data) are treated as random variables in *classical statistics* and as constants in *Bayesian statistics*.
- An ***estimator*** is a function of the sample, while an ***estimate*** (a number) is the realized value of an estimator that is obtained when a sample is actually taken.
- The probability distribution of a sample statistic is called the ***sampling distribution***.

2 Sampling Distributions for Sample Mean and Proportion

2.1 Sample Mean and Proportion: Variance Known

The Distribution of the Sample Mean

Theorem 2. *Let X_1, X_2, ..., X_n be an iid sample from $N(\mu, \sigma^2)$ distribution, then*

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

or

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Sampling distribution of a Proportion

Suppose the r.v. X is the number of successes in a binomial sample of n trials, whose probability of success is p.

The *sample proportion* is

$$\hat{p} = X/n$$

Properties:

- $E(\hat{p}) = p$
- It has standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- If n is large and ($np(1-p) > 9$ or roughly $n \geq 40$), then

$$\hat{p} \sim N(p, \sigma_{\hat{p}}^2) = N(p, \frac{p(1-p)}{n})$$

The Distribution of $\bar{X}$: when sampling is not from Normal

Theorem 3. ***The Central Limit Theorem.*** *Let X_1, X_2, $\ldots$, X_n be an iid sample from a population distribution X with mean μ and finite standard deviation σ. Then the sampling distribution of*

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

approaches the $N(0,1)$ distribution as $n \to \infty$.

2.2 Sampling Distributions of Two-Sample Difference

Difference of Independent Sample Means

Theorem 4. *Let X_1, X_2, $\ldots$, X_{n_1} be an iid sample from a $N(\mu_X, \sigma_X^2)$ distribution and let Y_1, Y_2, $\ldots$, Y_{n_2} be an iid sample from a $N(\mu_Y, \sigma_Y^2)$ distribution. Suppose that X_1, X_2, $\ldots$, X_{n_1} and Y_1, Y_2, $\ldots$, Y_{n_2} are independent, then the quantity*

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/\,n_1 + \sigma_Y^2/\,n_2}} \sim N(0,1)$$

Equivalently,

$$\overline{X} - \overline{Y} \sim N\big(\mu_X - \mu_Y, \sqrt{\sigma_X^2/\,n_1 + \sigma_Y^2/\,n_2}\ \big)$$

Difference of Independent Sample Proportions

Theorem 5. *Let X_1, X_2, $\ldots$, X_{n_1} be an iid sample from a $\mathsf{binom}(1, p_1)$ distribution and let Y_1, Y_2, $\ldots$, Y_{n_2} be an iid sample from a $\mathsf{binom}(1, p_2)$ distribution. Suppose that X_1, X_2, $\ldots$, X_{n_1} and Y_1, Y_2, $\ldots$, Y_{n_2} are independent samples. Define*

$$\hat{p}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \quad \text{and} \quad \hat{p}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j.$$

Then the sampling distribution of

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

approaches a $N(0,1)$ distribution as both n_1, $n_2 \to \infty$.

3 Sampling Distributions Associated with the Normal Distribution

3.1 Chi-Square Distribution (χ^2)

Chi-Squared Distribution

If $X_1, X_2, \cdots, X_n$ are iid $N(0,1)$, then

$$X = \sum_{i=1}^{n} X_i^2$$

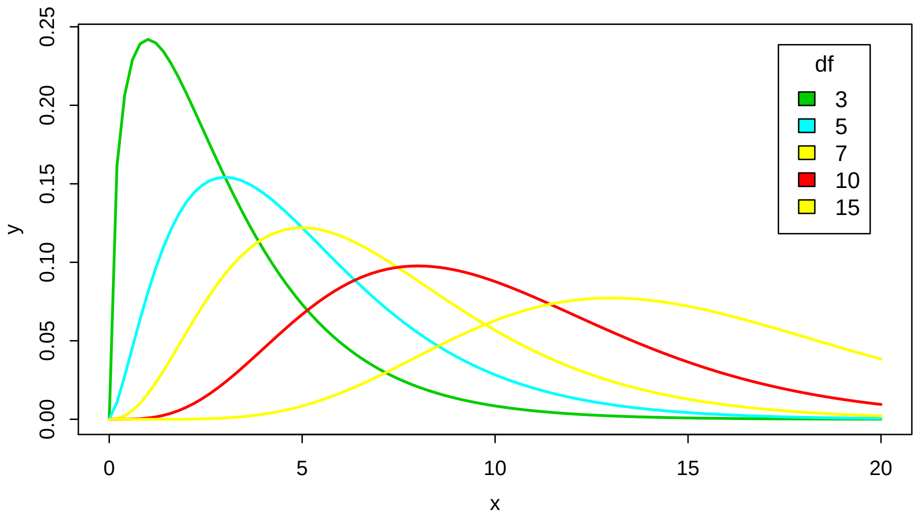
follows a **Chi-Squared distribution** with n degrees of freedom, We write $X \sim \chi^2(n)$ or $X \sim \chi_n^2$

Mean $\mu = n$, variance $\sigma^2 = 2n$ and mgf $M(t) = (1-2t)^{-n/2}$.

R functions: dchisq(x, df),...

Density Curves of $\chi^2(m)$

```
> curve(dchisq(x,df=3),from=0,to=20,ylab="y")
> ind<-c(3,5,7,10,15)
> for (i in ind){
+ curve(dchisq(x,df=i),0,20,col=i,lwd=2,add=TRUE)}
> legend("topright",inset=.05,title="df",as.character(ind),fill=ind)
```



Properties

1. If $Z \sim \text{norm}(0,1)$, then $Z^2 \sim \chi^2(1)$.

2. The chi-square distribution is supported on the positive x-axis, with a *right-skewed* distribution.

3. If $X_1 \sim \chi^2(n_1)$, $X_2 \sim \chi^2(n_2)$ and they are independent, then $X_1 + X_2 \sim \chi^2(n_1 + n_2)$.

4. $\chi^2(n) = \text{gamma}(n/2, 1/2)$

The Distribution of the Sample Variance

Theorem 6. *Let X_1, X_2, ..., X_n be an iid sample from $N(\mu, \sigma^2)$, and let*

$$\overline{X} = \sum_{i=1}^{n} X_i \quad \textit{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Then

1. *$\overline{X}$ and S^2 are independent, and*

2. *The rescaled sample variance*

$$\frac{(n-1)}{\sigma^2}S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

3.2 Student t-Distribution

Student t Distribution

Assume that $Z \sim N(0,1)$, $V \sim \chi^2(m)$, Z and V are independent, then

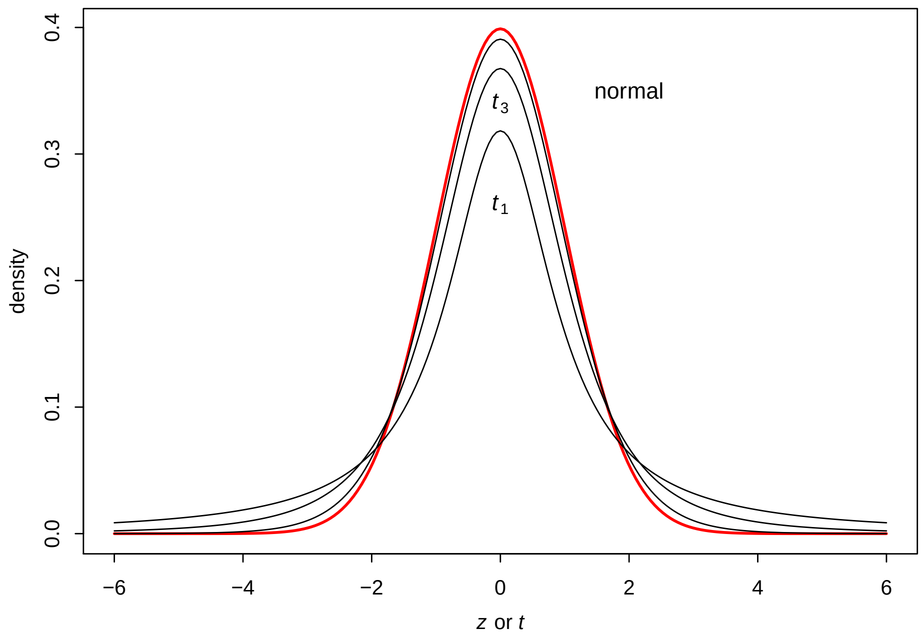
$$t = \frac{Z}{\sqrt{V/m}} \sim t(m)$$

is said to have a **Student's t-distribution** with m degrees of freedom, and we write $X \sim t(m)$ or $X \sim t_m$.

The associated R functions are `dt(x,m)`, `pt()`, `qt()`, and `rt()`.

Mean $\mathrm{E}(X) = 0$ and variance $\mathrm{Var}(X) = n/(n-2)$.

Student t Density Curves



The Distribution of Sample Mean: Variance Unknown

Theorem 7. *Let X_1, X_2, ..., X_n be an iid sample from $N(\mu, \sigma^2)$, then the quantity*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

- For $n > 30$, the t-distribution is close to a $N(0, 1)$.

The Sampling Distribution for $(\bar{X} - \bar{Y})$ ($\sigma_X^2 = \sigma_Y^2 = \sigma^2$,unknown)

Theorem 8. *Given two random samples X_1, X_2, ..., X_{n_1} and Y_1, Y_2, ..., Y_{n_2} that are taken from independent normal populations where $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$, and assume $\sigma_X = \sigma_Y$, then*

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{S_w^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2)$$

where

$$S_w^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}$$

3.3 The F Distribution

F-Distribution

If $U \sim \chi^2(m_1)$, $V \sim \chi^2(m_2)$ and independent , then

$$X = \frac{U/m_1}{V/m_2} \sim F(m_1, m_2)$$

is said to have an F**-distribution** with (m_1, m_2) degrees of freedom.

$$\mathrm{E}(X) = \frac{m_2}{m_2 - 2}, \ \mathrm{Var}(X) = \frac{2m_2^2(m_1 + m_2 - 2)}{m_1(m_2 - 2)^2(m_2 - 4)} \ (m_2 > 4)$$

The associated R functions are `df(x, df1, df2)`, `pf()`, `qf()`, and `rf()`

Properties

1. If $X \sim F(m_1, m_2)$ and$Y = 1/X$, then
$$Y \sim F(m_2, m_1)$$

2. If $X \sim t(n)$, then
$$X^2 \sim F(1, n)$$

Ratio of Independent Sample Variances

Theorem 9. *Let X_1, X_2, ..., X_{n_1} be an iid sample from a $N(\mu_X, \sigma_X^2)$ distribution and let Y_1, Y_2, ..., Y_{n_2} be an iid sample from a $N(\mu_Y, \sigma_Y^2)$ distribution. Suppose that X_1, X_2, ..., X_{n_1} and Y_1, Y_2, ..., Y_{n_2} are independent. Then the ratio*

$$F = \frac{\sigma_Y^2/S_Y^2}{\sigma_X^2/S_X^2} \sim F(n_2 - 1, n_1 - 1)$$