Sampling Distributions

Yanrong Yang

Research School of Finance, Actuarial Studies and Statistics
The Australian National University

April 19, 2017

Content

1. Review and Prospect
2. Topic today: statistic and sampling distribution
3. four specific statistics
4. summary

Recall Knowledge

1. **Descriptive Statistics**: summarizing data
2. **Probability**: set theory, random variable, distribution, expectation, moment function, multivariate random variable, limit theory.

Look into Future Work

Let $X_1, X_2, \ldots, X_n$ be i.i.d. sample from a population X. The goal is to utilize the sample $X_1, X_2, \ldots, X_n$ to infer some information for the population X. This is also called **statistical inference**.

1. Statistics and their sampling distributions;
2. Estimation for some parameter about the population, including point estimation and interval estimation;
3. Hypothesis test for some conclusions about the population.

Statistic

In order to study a **random variable** Y, we have available **sample** $Y_1, Y_2, \ldots, Y_n$ with sample size being n.
How to conduct this study or research? The key is **statistic**.

Definition (statistic)

A **statistic** is any function of the observable random variables in a sample and known constants.

In mathematical form, a statistic $S = f(Y_1, Y_2, \ldots, Y_n; c)$ with c representing some known constants.

1. the sample mean: $\bar{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i$;
2. $\bar{Y} - \mu$ with μ being known.

Sampling Distribution

Definition (sampling distribution)

The **sampling distribution** of a statistic is the probability distribution of a statistic.

1. A statistic is processed as random in theory while in practice it is only one value.
2. The sampling distribution of a statistic depends on the probability distribution of the original sample.

Four Specific/Common Statistics (1)

Statistic: **the sample mean**
Sampling distribution:

Theorem (1)

Suppose that $Y_1, Y_2, \ldots, Y_n \sim i.i.dN(\mu, \sigma^2)$. Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then $\bar{Y} \sim N(\mu, \sigma^2/n)$.

1. i.i.d. means that independent and identical distributed;
2. linear combinations of independent/uncorrelated normal rv's are still normal;
3. linear combinations of jointly normally distributed rv's are also normal.

Four Specific/Common Statistics (1)

Statistic: **the standardised sample mean**
Sampling distribution:

Corollary (1)

Under the assumptions of Theorem 1, the statistic $Z = \frac{\bar{Y}-\mu}{\sigma/n} \sim N(0, 1)$.

This result utilizes a property of normal distribution.

Four Specific/Common Statistics (1)

Proof of Theorem 1:

Method 1.

1. $\bar{Y}$ is still normal;
2. $\mathbb{E}(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}Y_i = \frac{1}{n} \times n\mu = \mu$;
3. $Var(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}Var(Y_i) = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$.

□

Method 2.

1. moment function $m_{\bar{Y}}(t)$ determines the distribution of $\bar{Y}$;
2. $$m_{\bar{Y}}(t) = \mathbb{E}e^{\bar{Y}t} = \mathbb{E}e^{t\left(\frac{Y_1+\ldots+Y_n}{n}\right)} = \mathbb{E}\left(e^{t\frac{Y_1}{n}} \times \ldots \times e^{t\frac{Y_n}{n}}\right)$$
$$= \left(\mathbb{E}e^{t\frac{Y_1}{n}}\right)\cdots\left(\mathbb{E}e^{t\frac{Y_n}{n}}\right) = m_{Y_1}\left(\frac{t}{n}\right)\cdots m_{Y_n}\left(\frac{t}{n}\right) = \left[m_{Y_1}\left(\frac{t}{n}\right)\right]^n$$
$$= \left(e^{\mu\left(\frac{t}{n}\right) + \frac{1}{2}\sigma^2\left(\frac{t}{n}\right)^2}\right)^n = e^{\mu t + \frac{1}{2}\frac{\sigma^2}{n}t^2}.$$

□

Four Specific/Common Statistics (1)

Example 1:

A bottling machine discharges volumes of drink that are independent and normally distributed with standard deviation 1 ml.

1. Find the sampling distribution of the mean volume of 9 randomly selected bottles that are filled by the machine.
2. Calculate the probability that this sample mean will be within 0.3 ml of the mean volume of all bottles filled by the machine.

Four Specific/Common Statistics (1)

Analysis of Example 1:

1. Define the rv: let Y_i be the volume of the ith bottle in the sample, $i = 1, \ldots, n$ with $n = 9$;
2. $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ with $\sigma^2 = 1$ and μ unknown;
3. $\bar{Y} \sim N(\mu, 1/9)$;
4. $P\left(|\bar{Y} - \mu| < 0.3\right) = P\left(\left|\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\right| < \frac{0.3}{1/3}\right) = P(|Z| < 0.9) = 1 - 2P(|Z| > 0.9) = 1 - 2 \times 0.1841 = 0.6318.$

Four Specific/Common Statistics (2)

Statistic: **the sample variance**

Sampling distribution:

Theorem (2)

Suppose that $Y_1, Y_2, \ldots, Y_n \sim N(\mu, \sigma^2)$. Let $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Then

1. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$;
2. S^2 is independent of $\bar{Y}$.

Four Specific/Common Statistics (2)

Example 2: Find an interval which we can be 90% sure will contain the sample variance of the 9 sampled volumes.

Analysis:

1. Find (a, b) s.t. $P(a < S^2 < b) = 0.9$;
2. $P(a < S^2 < b) = P\left(\frac{(n-1)a}{\sigma^2} < \frac{(n-1)S^2}{\sigma^2} < \frac{(n-1)b}{\sigma^2}\right) = P(8a < U < 8b) = 0.9$;
3. Due to $U \sim \chi^2(8)$, we derive an interval $(2.73264, 15.5073)$ by χ^2 tables. Then $(a, b) = (0.342, 1.938)$.

The Goal to Study Specific Statistics

Statistics are used to infer the population information.

1. **The sample mean** can be seen as an estimator for **the population mean** μ. If σ^2 is known, Corollary 1 can be used to infer μ.
2. **The sample variance** is an estimator for **the population variance** σ^2. Theorem 2 can be utilized to infer σ^2.

Remark

As σ^2 is unknown, a new statistic named $\mathbf{T}$ **statistic** is proposed.

Four Specific/Common Statistics (3)

Statistic: T statistic
Sampling distribution:

Theorem (3)

Suppose that $Y_1, \ldots, Y_n \sim i.i.d. N(\mu, \sigma^2)$. Let $T = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$. Then $T \sim t(n-1)$.

1. $T = \frac{Z}{\sqrt{U/(n-1)}} \sim t(n-1)$;
2. $Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$; $U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$; $Z \perp U$.

A Specific Distribution: T Distribution

Definition

Suppose that $Z \sim N(0,1)$, $U \sim \chi^2(k)$, $Z \perp U$. The rv $Y = \frac{Z}{\sqrt{U/k}}$ is named the t-distribution with k degrees of freedom. The pdf of Y is

$$f(y) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma(k/2)} \left(1 + \frac{y^2}{k}\right)^{-\frac{1}{2}(k+1)}, \quad -\infty < y < +\infty.$$

1. The pdf for t distribution looks like a standard normal distribution but with fatter tails.
2. The t distribution converges to the standard normal distribution as k tends to infinity.

Four Specific/Common Statistics (3)

Example 3: Find the probability that the mean of the 9 sample volumes will be distant from the population mean by no more than half the sample standard deviation of those 9 volumes.

Analysis:

1. $P\left(\left|\bar{Y}-\mu\right| < 0.5S\right) = P\left(\left|\frac{\bar{Y}-\mu}{S/\sqrt{n}}\right| < \frac{0.5S}{S/3}\right) = P\left(|T| < 1.5\right) = 1 - 2P(T > 1.5)$;
2. By tables $P(T > 1.397) = 0.10$ and $P(T > 1.860) = 0.05$;
3. $1 - 2P(T > 1.5)$ is between $1 - 2 \times 0.10$ and $1 - 2 \times 0.05$.

Why to study F statistic

If we are interested in comparing the variability of two populations, F statistic is useful.

Four Specific/Common Statistics (4)

Theorem (4)

Two samples $X_1, \dots, X_n \sim i.i.d. N(\mu_X, \sigma_X^2)$, $Y_1, \dots, Y_m \sim i.i.d. N(\mu_Y, \sigma_Y^2)$, and $(X_1, \dots, X_n) \perp (Y_1, \dots, Y_m)$. Let $W = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$. Then $W \sim F(n-1, m-1)$.

1. $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^m Y_i$;
2. $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n \sum_{i=1}^n (X_i - \bar{X})^2$, $S_Y^2 = \frac{1}{m-1}\sum_{i=1}^m \sum_{i=1}^m (Y_i - \bar{Y})$.

F Distribution

Definition

Suppose that $U \sim \chi^2(a)$, $V \sim \chi^2(b)$, and $U \perp V$. Let $Y = \frac{U/a}{V/b}$. Then the pdf of Y is $f(y) = \frac{\Gamma\left(\frac{a+b}{2}\right)}{\Gamma(a/2)\Gamma(b/2)} a^{a/2} b^{b/2} y^{\frac{a}{2}-1} (b + ay)^{-\frac{1}{2}(a+b)}$, $y > 0$. Write $Y \sim F(a, b)$.

1. If $X \sim F(a, b)$, then $1/X \sim F(b, a)$;
2. if $X \sim t(a)$, then $X^2 \sim F(1, a)$;
3. if $X \sim F(a, b)$, then $\frac{(a/b)X}{1+(a/b)X} \sim beta(a/2, b/2)$.

Four Specific/Common Statistics (4)

Example 4: Suppose that another sample of 5 bottles is to be taken from the output of the same bottling machine. Find the probability that the sample variance of the volumes in these 5 bottles will be at least 7 times as large as the same variance of the volumes in the 9 bottles that were initially sampled.

Analysis:

$P(S_X^2 > 7S_Y^2) = P\left(\frac{S_X^2}{S_Y^2} > 7\right) = P\left(\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} > 7\right) = P(U > 7) = 0.10$, where we utilize the fact that $\sigma_X^2 = \sigma_Y^2$ and $U \sim F(4, 8)$.

Mean of F Distribution

1. $$\mathbb{E}\left(\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}\right) = \mathbb{E}F(n-1, m-1);$$

2. $$\begin{aligned}\mathbb{E}F(n-1,m-1) &= \mathbb{E}\left(\frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)}\right) \\ &= \mathbb{E}\left(\frac{\chi^2_{n-1}}{n-1}\right)\mathbb{E}\left(\frac{m-1}{\chi^2_{m-1}}\right) \\ &= \left(\frac{n-1}{n-1}\right)\left(\frac{m-1}{m-3}\right) = \frac{m-1}{m-3};\end{aligned}$$

3. $$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \approx \frac{m-1}{m-3} \approx 1, \quad for \;\; large \;\; m.$$

Summary

1. Why we consider **statistic** and its **sampling distribution**?
2. Four specific statistics and their sampling distributions
   the sample mean;
   the sample variance;
   T statistic;
   F statistic
3. the motivations for studying specific statistics.