Sample Measures and Their Distribution

Jiří Neubauer

Department of Econometrics FVL UO Brno
office 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Survey Sampling

Survey Sampling

- entire, total, complete – census
- incomplete – sample survey

We would like to get sample which represents the characteristics of the population as closely as possible – representative sample.

Elementary Statistical Terms

A sample can be

- **random** – A sample is drawn in such a way that each element of the population has a chance of being selected. If all samples of the same size selected from a population have the same chance of being selected, we call it **simple random sampling**. Such a sample is called a **simple random sample**.
- **non-random** – The elements of the sample are not selected randomly but with a view of obtaining a representative sample.

Independent Random Variables

Random variables $X_1, X_2, \ldots, X_n$ are independent if and only if for any $x_1, x_2, \ldots, x_n \in \mathbb{R}$ is

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \cdot P(X_2 \leq x_2) \cdots P(X_n \leq x_n).$$

Independent Random Variables

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random vector which components $X_1, X_2, \ldots, X_n$ are random variables. Let

$$F(\mathbf{x}) = F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

be a **joint distribution function** and $F(x_1), F(x_2), \ldots, F(x_n)$ be distribution functions of the random variables $X_1, X_2, \ldots, X_n$. The random variables $X_1, X_2, \ldots, X_n$ are independent if and only if

$$F(x_1, x_2, \ldots, x_n) = F(x_1) \cdot F(x_2) \cdots F(x_n).$$

Independent Random Variables

If $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a random vector which components $X_1, X_2, \ldots, X_n$ are discrete random variables,

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

is a **joint probability** and $p(x_1), p(x_2), \ldots, p(x_n)$ are probability functions of random variables $X_1, X_2, \ldots, X_n$, then:
The random variables $X_1, X_2, \ldots, X_n$ are independent if and only if

$$p(x_1, x_2, \ldots, x_n) = p(x_1) \cdot p(x_2) \cdots p(x_n).$$

Independent Random Variables

If $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a random vector which components $X_1, X_2, \ldots, X_n$ are continuous random variables,

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$$

is **joint probability density function** and $f(x_1), f(x_2), \ldots, f(x_n)$ are probability density functions of random variables $X_1, X_2, \ldots, X_n$, then: The random variables $X_1, X_2, \ldots, X_n$ are independent if and only if

$$f(x_1, x_2, \ldots, x_n) = f(x_1) \cdot f(x_2) \cdots f(x_n).$$

Random Sample

- We measure some characteristic (variable) x_i $(i = 1, 2, \ldots, n)$ in given random sample – we obtain data.

Random Sample

- We measure some characteristic (variable) x_i $(i = 1, 2, \ldots, n)$ in given random sample – we obtain data.
- We can consider each value of characteristic as a possible value of a random variable X_i. Every random variable $X_i, (i = 1, \ldots, n)$ has the same distribution.

Random Sample

Definition

The random sample of size n is a sequence of independent random variables $X_1, X_2, \ldots, X_n$ with the same distribution.

The random sample can be considered as a vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Measured data we denote $x_1, x_2, \ldots, x_n$ and they are called measurements or (empirical) data.

Random Sample

If $X_1, X_2, \ldots, X_n$ is the random sample (i.i.d. – independent identically distributed random variables) then a distribution function $F(\mathbf{x})$ of the random sample is

$$F(\mathbf{x}) = F(x_1)F(x_2)\cdots F(x_n), \quad x_i \in \mathbb{R}.$$

Example

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from a uniform distribution on an interval $(0, 1)$. Find a distribution function $F(\mathbf{x})$ of the random sample.

Example

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from a uniform distribution on an interval $(0, 1)$. Find a distribution function $F(\mathbf{x})$ of the random sample.

Solution:
$X_i \sim R(0, 1)$ thus $F(x_i) = x_i$ for $0 < x_i < 1$,

$$F(\mathbf{x}) = F(x_1)F(x_2)\cdots F(x_n) = x_1 \cdot x_2 \cdots x_n.$$

Random Sample

If $X_1, X_2, \ldots, X_n$ is the random sample (i.i.d. random variables) then a probability function $p(\mathbf{x})$ of the random sample is

$$p(\mathbf{x}) = p(x_1)p(x_2)\cdots p(x_n), \quad x_i \in \mathbb{R}.$$

Example

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from a Poisson distribution with a parameter λ. Find a probability function $p(\mathbf{x})$ of the random sample.

Example

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from a Poisson distribution with a parameter λ. Find a probability function $p(\mathbf{x})$ of the random sample.

Solution:

$X_i \sim Po(\lambda)$ thus $p(x_i) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$ for $x_i = 0, 1, 2, \ldots,\ i = 1, 2, \ldots, n$

$$p(\mathbf{x}) = \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \cdots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} = \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda} \frac{1}{x_1! \cdot x_2! \cdots x_n!}.$$

Random Sample

If $X_1, X_2, \ldots, X_n$ is the random sample (i.i.d. random variables) then a probability density function $f(\mathbf{x})$ of the random sample from a distribution with the probability density function $f(x)$ is

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

Example

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from normal distribution $N(\mu, \sigma^2)$. Find the probability density function $f(\mathbf{x})$ of the random sample.

Example

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from normal distribution $N(\mu, \sigma^2)$. Find the probability density function $f(\mathbf{x})$ of the random sample.

Solution:

$X_i \sim N(\mu, \sigma^2)$ thus $f(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$ for $x_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$

$$f(\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Sample Measures

Definition

A function of random variables $X_1, X_2, \ldots, X_n$ is called **statistics**

$$T = T(X_1, X_2, \ldots, X_n) = T(\mathbf{X}).$$

Sample Measures

- Sample sum

$$M = \sum_{i=1}^{n} X_i$$

Sample Measures

- Sample sum

$$M = \sum_{i=1}^{n} X_i$$

- Sample mean

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

Sample Measures

- Sample variance

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

Sample Measures

- Sample variance

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

- Sample standard deviation

$$S = \sqrt{S^2}$$

Sample Measures

- Sample variance

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

- Sample standard deviation

$$S = \sqrt{S^2}$$

- Sample (moment) variance

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

Sample Measures

- Sample r^{th} moment

$$M'_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

Sample Measures

- Sample r^{th} moment

$$M_r' = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

- Sample r^{th} central moment

$$M_r = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^r$$

Sample Measures

- Sample skewness

$$A_3 = \frac{M_3}{M_2^{3/2}}$$

Sample Measures

- Sample skewness

$$A_3 = \frac{M_3}{M_2^{3/2}}$$

- Sample kurtosis

$$A_4 = \frac{M_4}{M_2^2} - 3$$

Distribution of the Sample Sum

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with the expected value (the mean) μ and the variance σ^2 ($E(X_i) = \mu$, $D(X_i) = \sigma^2$, for $i = 1, 2 \ldots, n$). The expected value and the variance of the sample sum are

$$E(M) = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E(X_i) = n\mu$$

$$D(M) = D\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} D(X_i) = n\sigma^2$$

Distribution of the Sample Sum

Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution $N(\mu, \sigma^2)$, then the sample sum also has a normal distribution

$$M \sim N(n\mu, n\sigma^2).$$

Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with the expected value (the mean) μ and the variance σ^2. The expected value and the variance of the sample mean are

$$E(\overline{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

$$D(\overline{X}) = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Distribution of the Sample Mean

Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution $N(\mu, \sigma^2)$, then the sample mean also has a normal distribution

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

A standardized random variable

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

has standard normal distribution $N(0, 1)$.

Distribution of the Sample Mean

If $X_1, X_2, \ldots, X_n$is a random sample from a distribution with the mean μ and variance σ^2, then a random variable

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

has for $n \geq 30$ approximately a standard normal distribution $N(0, 1)$ – see the central limit theorem.

Distribution of the Sample Variance

To derive the expected value of the sample variance we need following formulas:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

Distribution of the Sample Variance

To derive the expected value of the sample variance we need following formulas:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$D(X_i) = E(X_i^2) - E(X_i)^2 \rightarrow E(X_i^2) = D(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

Distribution of the Sample Variance

To derive the expected value of the sample variance we need following formulas:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$D(X_i) = E(X_i^2) - E(X_i)^2 \rightarrow E(X_i^2) = D(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

$$D(\overline{X}) = E(\overline{X}^2) - E(\overline{X})^2 \rightarrow E(\overline{X}^2) = D(\overline{X}) + E(\overline{X})^2 = \frac{\sigma^2}{n} + \mu^2$$

Distribution of the Sample Variance

$$
\begin{aligned}
E(S_n^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - E(\overline{X}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2) = \frac{1}{n} n(\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \\
&= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2
\end{aligned}
$$

Distribution of the Sample Variance

$$
\begin{aligned}
E(S_n^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - E(\overline{X}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2) = \frac{1}{n} n(\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \\
&= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2
\end{aligned}
$$

$$
E(S^2) = E\left(\frac{n}{n-1} S_n^2\right) = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2
$$

Distribution of the Sample Variance

Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with the mean μ and the variance σ^2. A random variable

$$\chi^2 = \frac{n-1}{\sigma^2} S^2$$

has χ^2-distribution with $n-1$ degrees of freedom.

Sample Distribution

Let us assume a random sample from a normal distribution with the mean μ and variance σ^2. We know that $Z = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$ and $\chi^2 = \frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$. A random variable

$$T = \frac{Z}{\sqrt{\frac{\chi^2}{n-1}}} = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \cdot \frac{\sqrt{n-1}}{\sqrt{\frac{n-1}{\sigma^2}S^2}} = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \cdot \frac{\sigma}{S} = \frac{\overline{X}-\mu}{S}\sqrt{n}$$

has a Student t-distribution with $n-1$ degrees of freedom.

Sample Distribution

Theorem

Let us have a random sample from a normal distribution with the mean μ and variance σ^2. A random variable

$$T = \frac{\overline{X} - \mu}{S}\sqrt{n}$$

has a Student t-distribution with $n - 1$ degrees of freedom.

Distribution of the Sample Proportion

Let us assume that distribution in a population can be described as a distribution of a Bernoulli random variable. A random sample can contain either ones or zeros. A random variable $X = X_1 + X_2 + \cdots + X_n$ denotes the number of ones (co called a **sample frequency**). A ratio

$$P = \frac{X}{n}$$

is called a **sample relative frequency** or a **sample proportion**

Distribution of the Sample Proportion

Let us assume that n is big enough. The random variable $P = \frac{X}{n}$ has approximately normal distribution with the mean π and the standard deviation $\sqrt{\pi(1-\pi)/n}$ – see the central limit theorem.
A standardized random variable

$$Z = \frac{P - \pi}{\sqrt{\pi(1-\pi)/n}}$$

has for large n approximately normal distribution $N(0,1)$. Approximation can be used if $n\pi \geq 5$ and $n(1-\pi) \geq 5$.