统计学方法及其应用

Statistical Methods with Applications

Rui Jiang, PhD

Associate Professor

Ministry of Education Key Laboratory of Bioinformatics
Bioinformatics Division, TNLIST/Department of Automation
Tsinghua University, Beijing 100084, China

Random Sampling

统计学方法及其应用 > 统计学基础 > 随机变量的函数

"A random variable is a quantity whose values are random and to which a probability distribution is assigned."

通过对部分的观测推断整体的性质

- 研究全中国儿童（6~7岁）的身高
  - $\sim 10^7$ 人
  - 测量全国儿童的身高
- 选择一些具有代表性的儿童进行测量
  - 全体儿童的集合

    $\mathbf{K} = \{k_1, k_2, \ldots, k_N\}$, N: 儿童的数量
  - 儿童身高的集合

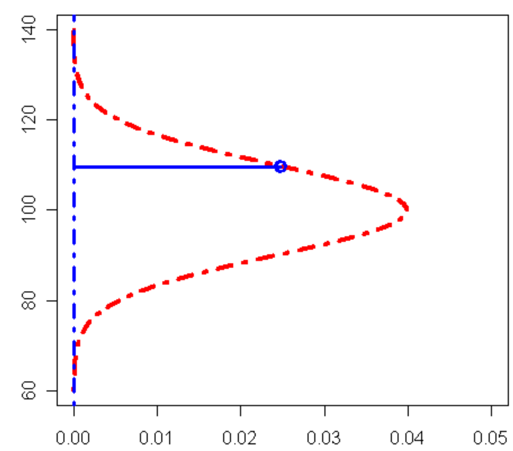
    $\mathbf{H} = \{h_1, h_2, \ldots, h_M\}$, M: 儿童身高的可能取值的数量
  - 每次选取一名儿童进行观测得到一个随机变量

    $X: \mathbf{K} \mapsto \mathbf{H}$

    pmf (pdf): $f_X(x) = f(x)$
  - 进行 n 次观测得到 n 个随机变量

    $X_i: \mathbf{K} \mapsto \mathbf{H}$

    pmf (pdf): $f_{X_i}(x) = f(x)$



Population

- 概念上，总体指研究对象的全体
  - **K**
- 统计上，总体指与全体相联系的某一数值特征的概率分布
  - $f(x)$
- 例如
  - 研究全中国儿童的身高

    总体为全体中国儿童

    因为关心的是身高这一数值特征，总体又指儿童身高的分布
  - 研究降压药物的降压作用

    总体为全体高血压病人

    因为关心的是血压这一数值特征，总体又指病人血压的分布

Random sampling with replacement

- 全体儿童的集合
  - $\mathbf{K} = \{k_1, k_2, \ldots, k_N\}$, N: 儿童的数量
- 儿童身高的集合
  - $\mathbf{H} = \{h_1, h_2, \ldots, h_M\}$, M: 儿童身高的可能取值的数量, $M = N$
- 每次选取一名儿童测量身高，不排除已经观察过的儿童
  - $P(K_i = k_j) = 1/N,\ i = 1,2,\ldots,n,\ j = 1,2,\ldots,N$
- Random sampling with replacement
  - **Mutually independent**
  - **Identically distributed**

Random sampling without replacement

- 每次选取一名儿童进行观测，排除已经观察过的儿童
  - $P(K_2 = k_i \mid K_1 = k_j) = 0, \qquad i = j$
  - $P(K_2 = k_i \mid K_1 = k_j) = 1/(N-1), \qquad i \neq j$
- Random sampling without replacement
  - **Not independent**
  - **Identically distributed**

$$
\begin{aligned}
P(K_2 = k_i) &= \sum_{j=1}^{N} P(K_2 = k_i \mid K_1 = k_j) P(K_1 = k_j) \\
&= \sum_{j \neq i} P(K_2 = k_i \mid K_1 = k_j) P(K_1 = k_j) + \underbrace{P(K_2 = k_i \mid K_1 = k_i) P(K_1 = k_i)}_{0} \\
&= (N-1)\left(\frac{1}{N-1}\frac{1}{N}\right) \\
&= \frac{1}{N}
\end{aligned}
$$

Random sampling from infinite population

- When $N \gg n$
  - $P(K_i = k_k \mid K_1, \ldots, K_{i\text{-}1}) = 1/(N-i+1) \approx 1/N = P(K_i = k_k)$
  - "Nearly independent"
- When $N \to \infty$
  - Mutually independent
  - Identically distributed
  - **Random sampling from infinite population**

Random sample

Random sample

The random variables $X_1,\ldots,X_n$ are called a **random sample of size n from the population $f(x)$** if $X_1,\ldots,X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1,\ldots,X_n$ are called **independent and identically distributed (iid) random variables with pdf or pmf $f(x)$**. A realization of these random variables, $x_1,\ldots,x_n$, is called **an observation** of the sample $X_1,\ldots,X_n$.

Descriptive statistics

What does descriptive statistics really do?

Suppose that the random variables $X_1, \ldots, X_n$ are a sample of size n from a certain population. Let $x_1, \ldots, x_n$ be an observation of $X_1, \ldots, X_n$. Descriptive statistics aims at **representing this observation by means of figures and tables, making the information contained in the sample obvious.**

Inferential statistics

What does inferential statistics really do?

Since the random variables $X_1, \ldots, X_n$ are independent and identically distributed, the joint pdf or pmf of $X_1, \ldots, X_n$ is given by

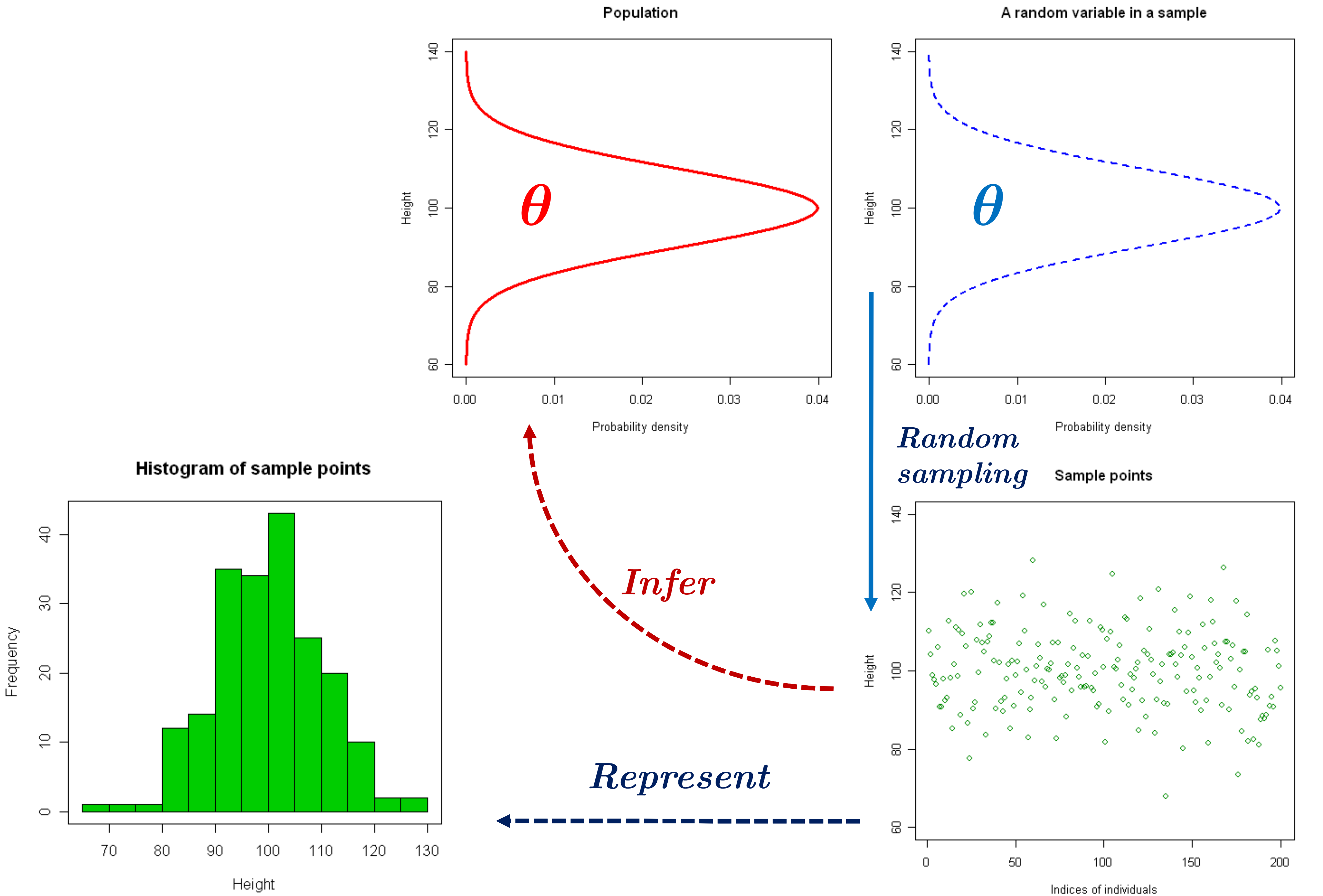
$$f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n) = \prod_{i=1}^{n} f(x_i).$$

In particular, if $f(x)$ is a member of a paramtric family, say, $f(x \mid \theta)$, then

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

Inferential statistics then aims at **using observations of the sample to infer properties associated with θ.**

Population, sample, descriptive statistics, inferential statistics



Statistic

Let $X_1,\ldots,X_n$ be a random sample of size n from a population and let $T(x_1,\ldots,x_n)$ be a **real-valued** or **vector-valued function** whose domain includes the sample space of $(X_1,\ldots,X_n)$. Then the random variable or random vector $T=(X_1,\ldots,X_n)$ is called a **statistic**. The probability distribution of a statistic Y is called the **sampling distribution of $\boldsymbol{Y}$**.

统计量就是样本的函数，唯一的要求是不依赖于决定总体的参数。统计量的分布称为抽样分布。

Sample mean

The **sample mean** is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \bar{X}(X_1, \ldots, X_n) = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sample variance and standard deviation

The **sample variance** is the statistic defined by

$$S^2 = S^2(X_1, \dots, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Since

$$\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \sum_{i=1}^{n} X_i^2 - 2\bar{X}\underbrace{\sum_{i=1}^{n} X_i}_{n\bar{X}} + \underbrace{\sum_{i=1}^{n}\bar{X}^2}_{n\bar{X}^2} = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

We have

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right), \text{ or say } (n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

The **sample standard derivation** is the statistic defined by

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

Computational issue

For a series of iid random variables $X_1, \ldots,$

let $\bar{X}_k = \frac{1}{k}\sum_{i=1}^{k} X_i$ and

$$S_k^2 = \frac{1}{k-1}\sum_{i=1}^{k}(X_i - \bar{X}_k)^2 = \frac{1}{k-1}\left(\sum_{i=1}^{k} X_i^2 - k\bar{X}_k^2\right).$$

Can we express $\bar{X}_{k+1}$ using $\bar{X}_k$ and S_{k+1}^2 using S_k^2 ?

Recurrence

$$\left.\begin{aligned} k\bar{X}_k &= \sum_{i=1}^{k} X_i \\ (k+1)\bar{X}_{k+1} &= \sum_{i=1}^{k+1} X_i \end{aligned}\right\} \Rightarrow (k+1)\bar{X}_{k+1} = k\bar{X}_k + X_{k+1}.$$

$$\left.\begin{aligned} (k-1)S_k^2 &= \sum_{i=1}^{k} X_i^2 - k\bar{X}_k{}^2 \\ kS_{k+1}^2 &= \sum_{i=1}^{k+1} X_i^2 - (k+1)\bar{X}_{k+1}{}^2 \end{aligned}\right\} \Rightarrow kS_{k+1}^2 - (k-1)S_k^2 = \frac{k}{k+1}(X_{k+1} - \bar{X}_k)^2.$$

Expectation of a random sample

Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $\mathrm{E}g(X_1)$ and $\mathrm{Var}\,g(X_1)$ exist. Then

$$\mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathrm{E}g(X_1),$$

and

$$\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathrm{Var}\,g(X_1).$$

Proof

$$\mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right) = \sum_{i=1}^{n} \mathrm{E}g(X_i) = n\mathrm{E}g(X_1)$$

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) &= \mathrm{E}\left[\sum_{i=1}^{n} g(X_i) - \mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right)\right]^2 \\
&= \mathrm{E}\left[\sum_{i=1}^{n} g(X_i) - \sum_{i=1}^{n} \mathrm{E}g(X_i)\right]^2 \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \left[g(X_i) - \mathrm{E}g(X_i)\right]\right]^2 \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \left[g(X_i) - \mathrm{E}g(X_i)\right]^2\right] + \mathrm{E}\left[\sum_{i \neq k} \left[g(X_i) - \mathrm{E}g(X_i)\right]\left[g(X_k) - \mathrm{E}g(X_k)\right]\right] \\
&= \sum_{i=1}^{n} \mathrm{E}\left[g(X_i) - \mathrm{E}g(X_i)\right]^2 + \sum_{i \neq k} \mathrm{E}\left[(g(X_i) - \mathrm{E}g(X_i))(g(X_k) - \mathrm{E}g(X_k))\right] \\
&= n\mathrm{Var}\, g(X_1) + \sum_{i \neq k} \mathrm{Cov}\left[g(X_i), g(X_k)\right] \\
&= n\mathrm{Var}\, g(X_1)
\end{aligned}$$

Sample mean and sample variance

Let $X_1, \ldots, X_n$ be a random sample from a population with μ and $\sigma^2 < \infty$, then

1. $\mathrm{E}\bar{X} = \mu$,
2. $\mathrm{Var}\bar{X} = \dfrac{\sigma^2}{n}$,
3. $\mathrm{E}S^2 = \sigma^2$.

Proof

$$\mathrm{E}\bar{X} = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} n\mathrm{E}X_1 = \mathrm{E}X_1 = \mu$$

$$\mathrm{Var}\bar{X} = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} n\mathrm{Var}X_1 = \frac{\sigma^2}{n}$$

$$\begin{aligned}
\mathrm{E}S^2 &= \mathrm{E}\left[\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\right] \\
&= \frac{1}{n-1}\left(n\mathrm{E}X_1^2 - n\mathrm{E}\bar{X}^2\right) \\
&= \frac{1}{n-1}\left[n(\mathrm{Var}X_1 + (\mathrm{E}X_1)^2) - n(\mathrm{Var}\bar{X} + (\mathrm{E}\bar{X})^2)\right] \\
&= \frac{1}{n-1}\left[n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] \\
&= \sigma^2
\end{aligned}$$

Sampling from Normal Populations

统计学方法及其应用 | 统计学基础 | 随机变量的函数

“A random variable is a quantity whose values are random and to which a probability distribution is assigned.”

Sampling from normal populations

- One-sample — Sampling from a univariate normal population
- Paired-sample — Sampling from a bivariate normal population
- Two-sample — Sampling from two univariate normal populations

One-sample mean and variance

Sample mean and sample variance

Let $X_1, \ldots, X_n$ be a random sample from a normal population $N(\mu, \sigma^2)$, then $\bar{X}$ and S^2 are independent random variables.

Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$. Then X and Y are independent random variables **if and only if** there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \Re$ and $y \in \Re$,

$$f(x, y) = g(x)h(y).$$

One-sample mean

Distributions of sample mean

Let $X_1, \ldots, X_n$ be a random sample from a normal population $N(\mu, \sigma^2)$, then

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

has a standard normal distribution.

Summation of two Normals

Summation of two independent normal random variables

Let $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\upsilon, \tau^2)$ be two independent normal random variables, then

$$Z = X + Y \sim N(\mu + \upsilon, \sigma^2 + \tau^2),$$
$$Z = X - Y \sim N(\mu - \upsilon, \sigma^2 + \tau^2)$$

$$\begin{aligned}
\int_{-\infty}^{\infty} f_X(w) f_Y(z-w)dw &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(z-w-\upsilon)^2}{2\tau^2}\right) dw \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\tau^2}} \exp\left(-\frac{(z-\mu-\upsilon)^2}{2(\sigma^2+\tau^2)}\right) \times \\
&\quad \underbrace{\frac{1}{\sqrt{2\pi}} \sqrt{\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2} \frac{\sigma^2+\tau^2}{\sigma^2\tau^2} \left(w - \frac{\mu\tau^2 + (z-\upsilon)\sigma^2}{\sigma^2+\tau^2}\right)^2\right] dw}_{1}
\end{aligned}$$

Summation of multiple iid normals

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

$$X_1 \sim N(\mu, \sigma^2)$$

$$X_1 + X_2 \sim N(2\mu, 2\sigma^2)$$

$$\cdots$$

$$X_1 + X_2 + \cdots + X_{n-1} \sim N((n-1)\mu, (n-1)\sigma^2)$$

$$(X_1 + X_2 + \cdots + X_{n-1}) + X_n \sim N((n-1)\mu, (n-1)\sigma^2) + N(\mu, \sigma^2) \sim N(n\mu, n\sigma^2)$$

One-sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Y = \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \Rightarrow f_Y(y) = \frac{1}{\sqrt{2\pi(n\sigma^2)}} \exp\left(-\frac{(y-n\mu)^2}{2(n\sigma^2)}\right)$$

$$Z = \frac{1}{n}Y \Rightarrow Y = nZ, \frac{dy}{dz} = n$$

$$f_Z(z) = f_Y(nz)\left|\frac{dy}{dz}\right| = \frac{n}{\sqrt{2\pi n\sigma^2}} \exp\left(-\frac{(nz-n\mu)^2}{2n\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi(\sigma^2/n)}} \exp\left(-\frac{(z-\mu)^2}{2(\sigma^2/n)}\right)$$

mgf of the sample mean

Let $X_1,\ldots,X_n$ be a random sample from a population with mgf $M_X(t)$, then the mgf of sample mean $\bar{X} = (X_1 + \cdots + X_n) / n$ is

$$M_{\bar{X}}(t) = [M_X(t / n)]^n.$$

Let $Y = X_1 + \cdots + X_n$, then

$$M_Y(t) = [M_X(t)]^n$$

Let $\bar{X} = Y / n$, then the mgf of $\bar{X}$ is

$$M_{\bar{X}}(t) = M_Y(t / n) = [M_X(t / n)]^n.$$

mgf of the sample mean

Let $X_1,\ldots,X_n$ be iid random variables with mgf $M_X(t)$, then the mgf of $Z = X_1 + \cdots + X_n$ is

$$M_Z(t) = \mathrm{E}[e^{tZ}] = \mathrm{E}[e^{t\Sigma_{i=1}^n X_i}] = \mathrm{E}\left[\prod_{i=1}^n e^{tX_i}\right] = \prod_{i=1}^n \mathrm{E}[e^{tX_i}] = [M_X(t)]^n$$

Let Y be a random variable with pdf $f_Y(y)$. Let $Z = aY + b$, then

$$Y = \frac{Z-b}{a}, \quad \frac{dy}{dz} = \frac{1}{a}$$

$$f_Z(z) = \frac{1}{a} f_Y\left(\frac{z-b}{a}\right)$$

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} e^{tz} f_Z(z)dz = \int_{-\infty}^{\infty} e^{t(ay+b)} \frac{1}{a} f_Y\left(\frac{z-b}{a}\right) d(ay+b) \\ &= e^{tb} \int_{-\infty}^{\infty} e^{(at)y} f_Y(y)dy \\ &= e^{tb} M_Y(at) \end{aligned}$$

Particularly, if $Z = Y/n$, then $M_Z(t) = M_Y(t/n)$

Normal sample mean

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population. Then, the mgf of the sample mean is

$$M_{\bar{X}}(t) = \left[M_X\left(\frac{t}{n}\right) \right]^n = \left[\exp\left(\mu \frac{t}{n} + \frac{\sigma^2}{2} \left(\frac{t}{n}\right)^2 \right) \right]^n = \exp\left(\mu t + \frac{\sigma^2 / n}{2} t^2 \right).$$

Therefore,

$$\bar{X} \text{ has a } N\left(\mu, \frac{\sigma^2}{n}\right) \text{ distribution.}$$

Equivalently,

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \text{ has a standard normal distribution.}$$

Summation of sample squares

Distribution of the sum of sample squares

Let $X_1, \ldots, X_n$ be a random sample from a normal population $N(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$$

has a χ^2 distribution with n degrees of freedom

Each $\frac{X_i - \mu}{\sigma}$ is a standard normal.

Chi-squared distribution

- pdf

$$\chi_p^2 = \frac{1}{\Gamma(p\,/\,2)2^{p/2}} x^{\frac{p}{2}-1} e^{-\frac{x}{2}}, 0 < x < \infty, p > 0$$

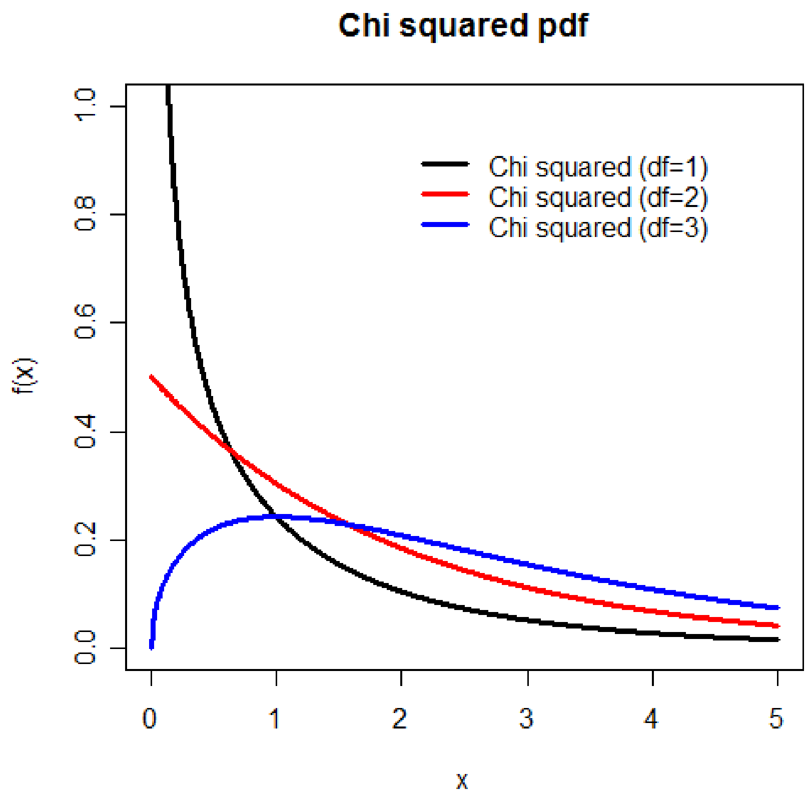
- Mean

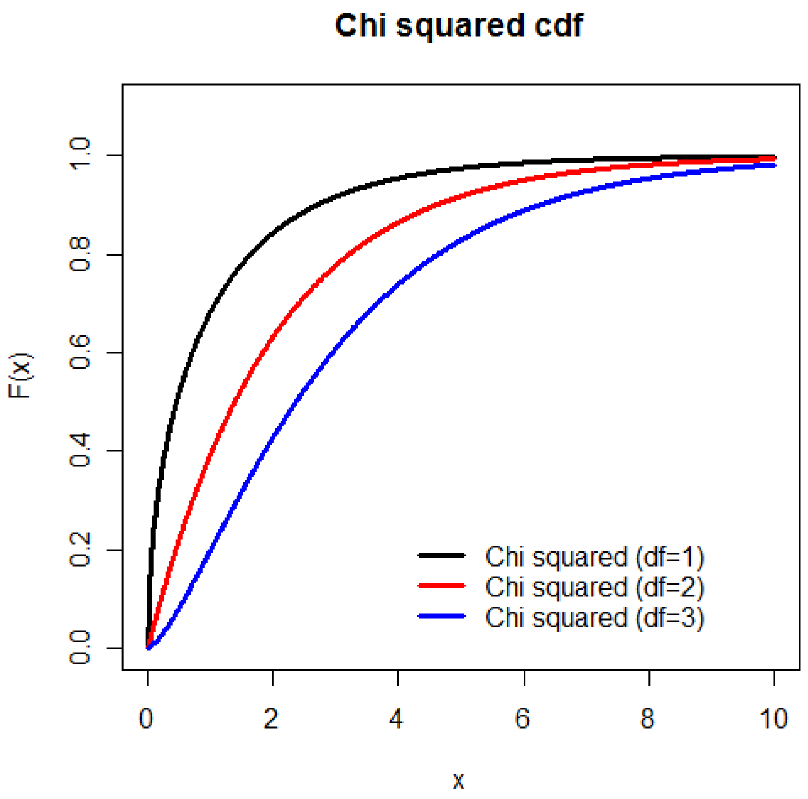
$$\mathrm{E}X = p$$

- Variance

$$\mathrm{Var}X \quad = 2p$$

Illustration of χ^2 pdf and cdf





χ^2 distribution in R

- pmf

  `dchisq(x, df)`

- cdf

  `pchisq(q, df)`

- Quantile function

  `qchisq(p, df)`

- Random numbers

  `rchisq(n, df)`

Gamma(***p*/2, scale=2**) $\rightarrow \chi_p^{\,2}$

Gamma pdf

$$f(x \mid \alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} e^{-x/\theta}, 0 < x < \infty, \alpha > 0, \theta > 0$$

$\chi_p^{\,2}$ pdf

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{\frac{p}{2}-1} e^{-\frac{x}{2}}, 0 < x < \infty, p > 0$$

A chi-squared random variable with *p* degrees of freedom is a gamma random variable with shape *p*/2 and scale 2

Gamma(**1/2, scale=2**) $\rightarrow \chi_1{}^2$

Gamma pdf

$$f(x \mid \alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} e^{-x/\theta}, 0 < x < \infty, \alpha > 0, \theta > 0$$

$\chi_1{}^2$ pdf

$$f(x) = \frac{1}{\Gamma(1/2)2^{1/2}} x^{\frac{1}{2}-1} e^{-\frac{x}{2}} = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{x}} e^{-\frac{x}{2}}, 0 < x < \infty$$

A chi-squared random variable with 1 degrees of freedom is a gamma random variable with shape 1/2 and scale 2

Chi-squared mgf

Gamma mgf can be calculated as

$$
\begin{aligned}
M(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} e^{-x/\theta} dx \\
&= \frac{1}{\Gamma(\alpha)\theta^{\alpha}} \int_{-\infty}^{\infty} x^{\alpha-1} e^{-x(1/\theta - t)} dx \\
&= \left[\frac{1}{\Gamma(\alpha)\theta^{\alpha}}\right]\left[\frac{1}{\Gamma(\alpha)[\theta\,/\,(1-\theta t)]^{\alpha}}\right]^{-1} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\Gamma(\alpha)[\theta\,/\,(1-\theta t)]^{\alpha}} x^{\alpha-1} e^{-x/[\theta/(1-\theta t)]} dx}_{=1} \\
&= \left(\frac{1}{1-\theta t}\right)^{\alpha}
\end{aligned}
$$

χ_p^2, as $Gamma(p\,/\,2, 2)$, then has mgf

$$
M_{\chi_p^2}(t) = \left(\frac{1}{1-2t}\right)^{p/2}
$$

Summation of multiple independent χ^2

$$\chi^2_{p_1} + \cdots + \chi^2_{p_n} \sim \chi^2_{\sum_{i=1}^n p_i}$$

$$M_{\chi^2_p}(t) = \left(\frac{1}{1-2t}\right)^{\frac{p}{2}} \Rightarrow$$

$$M_{\chi^2_{p_1}+\cdots+\chi^2_{p_n}}(t) = \left(\frac{1}{1-2t}\right)^{\frac{p_1}{2}} \cdots \left(\frac{1}{1-2t}\right)^{\frac{p_n}{2}} = \left(\frac{1}{1-2t}\right)^{\frac{1}{2}\sum_{i=1}^n p_i} \Rightarrow$$

$$\chi^2_{p_1} + \cdots + \chi^2_{p_n} \sim \chi^2_{\sum_{i=1}^n p_i}$$

Independent chi-squared random variables add to a chi-squared random variable, and the degrees of freedom also add

(Standard normal)$^2 \rightarrow {\chi_1}^2$

For a random variable $X \sim N(0,1)$ and the transformation $Y = g(X) = X^2$

$x \in (-\infty, 0), y = g_1(x) = x^2,\ h_1(y) = -\sqrt{y}$, decreasing;

$x \in (0, +\infty), y = g_2(x) = x^2,\ h_1(y) = \sqrt{y}$, increasing;

$x = 0$ (with probability 0).

Define $\quad A_0 = \{0\}; A_1 = (-\infty, 0); A_2 = (0, \infty)$.

Then $\quad A_0 \cap A_1 \cap A_2 = \emptyset$ and $A_0 \cup A_1 \cup A_2 = (-\infty, \infty)$.

In A_1, $\quad f_1(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(-\sqrt{y})^2}{2}} \left| -\frac{1}{2}\frac{1}{\sqrt{y}} \right| = \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}} e^{-\frac{y}{2}}$

In A_2, $\quad f_2(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\sqrt{y})^2}{2}} \left| \frac{1}{2}\frac{1}{\sqrt{y}} \right| = \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}} e^{-\frac{y}{2}}$

Then, $\quad f(y) = f_1(y) + f_2(y) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}} e^{-\frac{y}{2}} \sim \chi_1^2$

The square of a standard normal random variable is a chi-squared random variable with 1 degree of freedom

One-sample variance

Distributions of sample variance

Let $X_1, \ldots, X_n$ be a random sample from a normal population $N(\mu, \sigma^2)$, then

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

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

The variance of the sample variance

Because

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

we have

$$\mathrm{E}\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1,$$

$$\mathrm{Var}\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1).$$

This is to say that

$$\mathrm{E}\left[S^2\right] = \left(\frac{\sigma^2}{n-1}\right)(n-1) = \sigma^2,$$

$$\mathrm{Var}\left[S^2\right] = \left(\frac{\sigma^2}{n-1}\right)^2 2(n-1) = \frac{2\sigma^4}{n-1}.$$

One-sample mean

Distributions of sample mean

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population, then

$$\frac{\bar{X} - \mu}{S / \sqrt{n}}$$

has a **Student's t distribution** with $n-1$ degrees of freedom.

When variance is unknown

Let $X_1,\ldots,X_n$ be a random sample from a $N(\mu,\sigma^2)$ population, then

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

However, in most cases, the true variance σ^2 is unknown. Since

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

We have

$$\frac{\bar{X}-\mu}{S / \sqrt{n}} = \frac{\bar{X}-\mu}{S / \sqrt{n}}\frac{\sigma}{\sigma} = \frac{\dfrac{\bar{X}-\mu}{\sigma / \sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}\Big/(n-1)}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-1} / (n-1)}}$$

This distribution is desirable because the unknown σ^2 is not involved.

But what is this distribution?

Transformation

Let $U \sim N(0,1)$ and $V \sim \chi_p^2$ be two independent random variables. Then

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

$$f(v) = \frac{1}{\Gamma(p/2)2^{p/2}} v^{p/2-1} e^{-v/2}$$

Consider the transformation

$$T = \frac{U}{\sqrt{V/p}}, W = V$$

Since

$$U = T\sqrt{W/p}, V = W$$

the Jacobian is

$$J = \begin{vmatrix} (w/p)^{1/2} & [1/(2p)](w/p)^{-1/2} t \\ 0 & 1 \end{vmatrix} = \left(\frac{w}{p}\right)^{1/2}$$

Student's t distribution

The joint pdf of (U,V) is

$$f(u,v) = \frac{1}{\sqrt{2\pi}}\frac{1}{\Gamma(p/2)2^{p/2}}e^{-u^2/2}v^{p/2-1}e^{-v/2}$$

The joint pdf of (T,W) is therefore

$$\begin{aligned} f(t,w) &= \frac{1}{\sqrt{2\pi}}\frac{1}{\Gamma(p/2)2^{p/2}}\left(\frac{w}{p}\right)^{1/2} e^{-t^2w/(2p)}w^{p/2-1}e^{-w/2} \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{\Gamma(p/2)2^{p/2}p^{1/2}}w^{(p/2+1/2)-1}e^{-(1/2)(t^2/p+1)w} \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{\Gamma(p/2)2^{p/2}p^{1/2}}w^{\tilde{\alpha}-1}e^{-w/\tilde{\theta}}, \end{aligned}$$

where $\tilde{\alpha} = (p+1)/2, \tilde{\theta} = 2/(1+t^2/p)$. Because $w^{\tilde{\alpha}-1}e^{-w/\tilde{\theta}}$ is the kernel of a $Gamma(\tilde{\alpha}, \text{scale} = \tilde{\theta})$ pdf,

$$\int w^{\tilde{\alpha}-1}e^{-w/\tilde{\theta}}dw = \Gamma(\tilde{\alpha})\tilde{\theta}^{\tilde{\alpha}} = \Gamma\left(\tfrac{p+1}{2}\right)\left(\tfrac{2}{1+t^2/p}\right)^{\frac{p+1}{2}}$$

Therefore,

$$f(t) = \int f(t,w)dw = \frac{1}{\sqrt{2\pi}}\frac{1}{\Gamma(p/2)2^{p/2}p^{1/2}}\Gamma\left(\tfrac{p+1}{2}\right)\left(\tfrac{2}{1+t^2/p}\right)^{\frac{p+1}{2}} = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)}\frac{1}{\sqrt{p\pi}}\frac{1}{(1+t^2/p)^{(p+1)/2}}.$$

Student's t distribution

- pdf

$$f(x \mid p) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{\sqrt{p\pi}} \frac{1}{(1+x^2\,/\,p)^{(p+1)/2}}, -\infty < x < \infty, p = 1, \ldots$$

- Mean

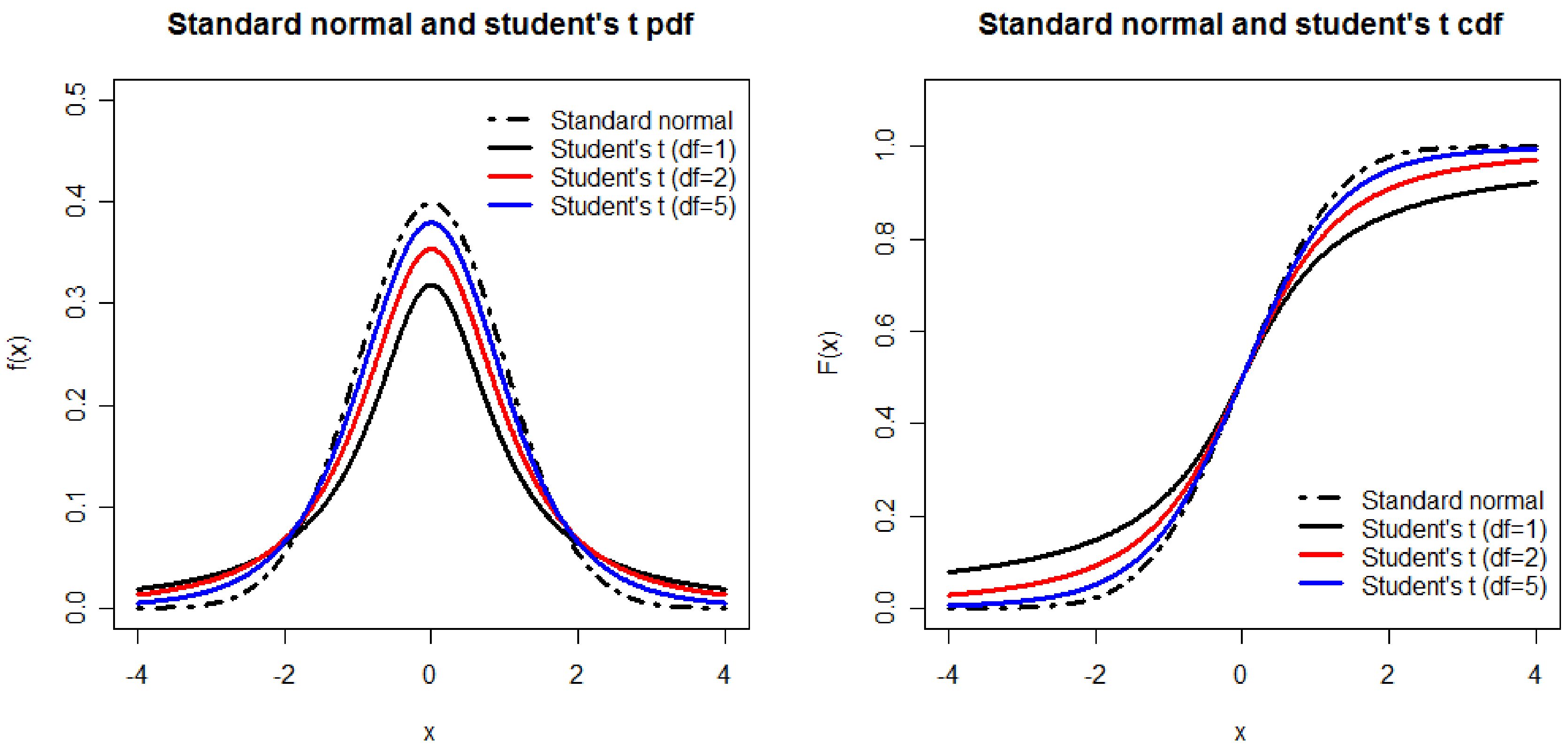
$$\mathrm{E}X = 0, p > 1$$

- Variance

$$\mathrm{Var}X = \frac{p}{p-2}, p > 2$$

- t_1 has no mean, t_2 has no variance.
- A t distribution becomes a Cauchy distribution when $p = 1$ (sample size 2, illness appears in ordinary situation)
- A t distribution becomes a standard normal distribution when the degree of freedom tends to infinity

Student's *t* distribution



Student's t distribution in R

- pmf

  `dt(x, df)`

- cdf

  `pt(q, df)`

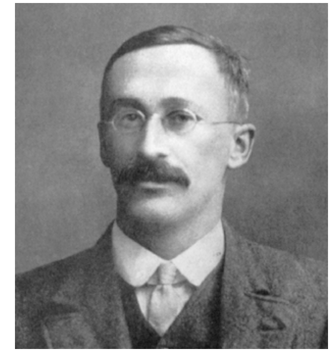
- Quantile function

  `qt(p, df)`

- Random numbers

  `rt(n, df)`

VOLUME VI MARCH, 1908 No. 1



William Sealy Gosset

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series

Bivariate normal distribution

Bivariate normal distribution

A random vector (X, Y) is said to has a bivariate normal distribution if their joint pdf is

$$f_{XY}(x, y \mid \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times$$

$$\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}.$$

Marginal distributions

Marginal distributions

If $(X,Y) \sim$ Bivariate normal$(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, the marginal distribution of X is $N(\mu_X, \sigma_X^2)$, the marginal distribution of Y is $N(\mu_Y, \sigma_Y^2)$.

$Z = aX + bY$ has a $N(\mu_Z, \sigma_Z^2)$ distribution, where

$$\mu_Z = a\mu_X + b\mu_Y,$$
$$\sigma_Z^2 = a^2\sigma_X^2 + 2ab\rho\sigma_X\sigma_Y + b^2\sigma_Y^2.$$

Particularly, $W = X - Y$ has a $N(\mu_W, \sigma_W^2)$ distribution, where

$$\mu_W = \mu_X - \mu_Y,$$
$$\sigma_W^2 = \sigma_X^2 - 2\rho\sigma_X\sigma_Y + \sigma_Y^2.$$

Paired-sample mean

Paired-sample mean

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal population with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$, then

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma_{X-Y} / \sqrt{n}}$$

has a standard normal distribution, and

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_{X-Y} / \sqrt{n}}$$

has a student's t distribution with $n-1$ degrees of freedom,

where $\sigma_{X-Y}^2 = \sigma_X^2 - 2\rho\sigma_X\sigma_Y + \sigma_Y^2$, and

$$S_{X-Y}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left[(X_i - Y_i) - (\bar{X} - \bar{Y}) \right]^2.$$

Two-sample means

Two sample means

Let $X_1,\ldots,X_m$ be a random sample from a $N(\mu_X,\sigma_X^2)$ population; let $Y_1,\ldots,Y_n$ be a random sample from an independent $N(\mu_Y,\sigma_Y^2)$ population. Assume $\sigma_X^2=\sigma_Y^2=\sigma^2$, then

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

has a standard normal distribution.

Two-sample means

Two sample means

Let $X_1,\ldots,X_m$ be a random sample from a $N(\mu_X,\sigma_X^2)$ population; let $Y_1,\ldots,Y_n$ be a random sample from an independent $N(\mu_Y,\sigma_Y^2)$ population. Assume $\sigma_X^2=\sigma_Y^2=\sigma^2$, then

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

has a student's t distribution with $m+n-2$ degrees of freedom. Here,

$$S_p^2=\frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}$$

is called the **pooled variance estimate**.

Two-sample variances

Distributions of sample variances

Let $X_1, \ldots, X_m$ be a random sample with size n from a normal population $N(\mu_X, \sigma_X^2)$ and let $Y_1, \ldots, Y_n$ be a random sample with size m from an independent normal population $N(\mu_Y, \sigma_Y^2)$. Then, the random variable

$$\frac{S_X^2 / S_Y^2}{\sigma_X^2 / \sigma_Y^2} = \frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2}$$

has a Snedecor's ***F* distribution** with $m-1$ and $n-1$ degrees of freedom.

$$\frac{S_X^2 / S_Y^2}{\sigma_X^2 / \sigma_Y^2} = \frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2} = \frac{\dfrac{(m-1)S_X^2}{\sigma_X^2}\dfrac{1}{m-1}}{\dfrac{(n-1)S_Y^2}{\sigma_Y^2}\dfrac{1}{n-1}} = \frac{\dfrac{\chi_{m-1}^2}{m-1}}{\dfrac{\chi_{n-1}^2}{n-1}} = \frac{U/p}{V/q}, U \sim \chi_p^2, V \sim \chi_q^2$$

Consider transformation $F = (U/p)/(V/q), W = U + V$

F distribution

George W. Snedecor

- pdf

$$f(x \mid p,q) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2}\frac{x^{p/2-1}}{(1+(p/q)x)^{(p+q)/2}}, 0 \le x < \infty, p,q = 1,\ldots$$

- Mean

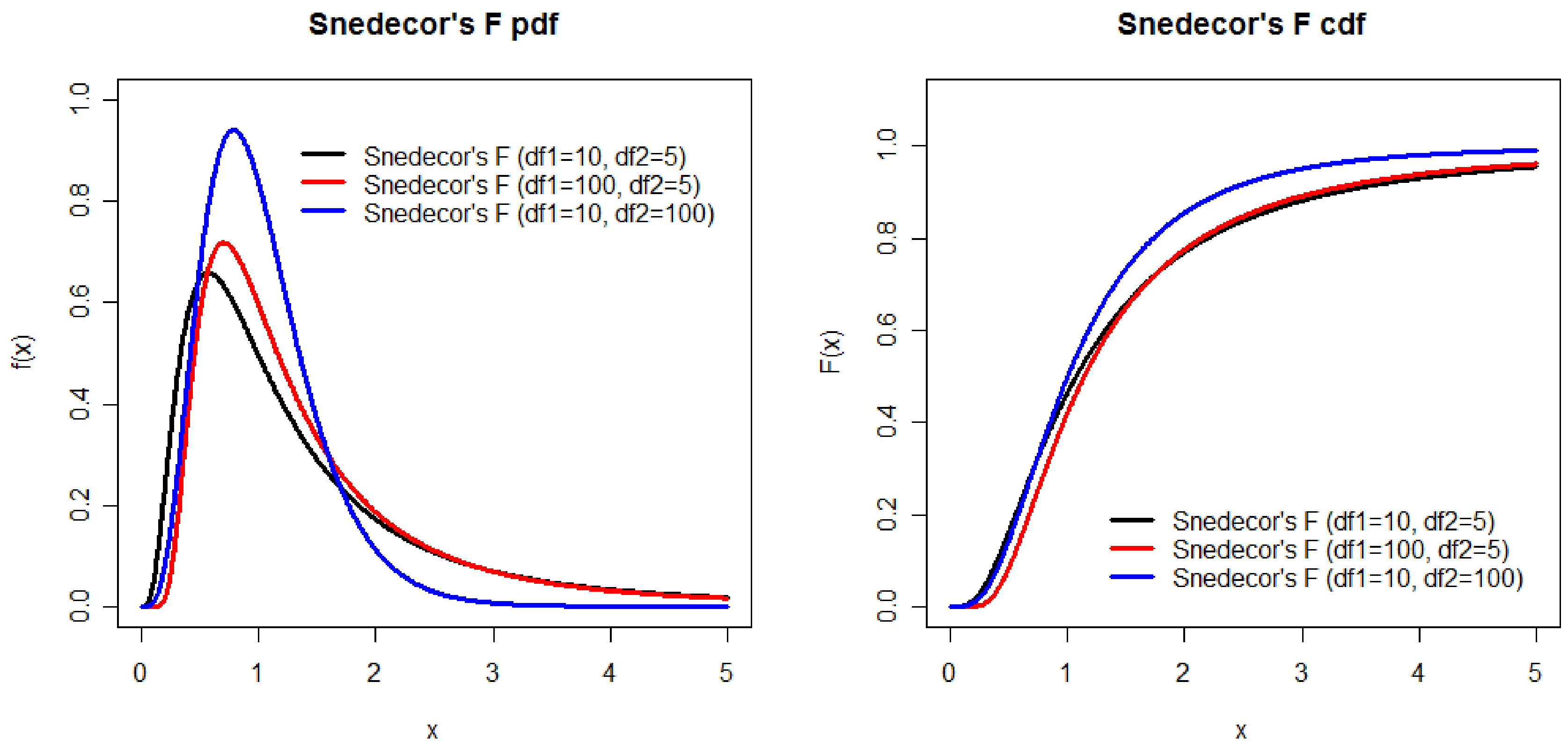
$$\mathrm{E}X = \frac{q}{q-2}, q > 2$$

- Variance

$$\mathrm{Var}X = 2\left(\frac{q}{q-2}\right)\frac{p+q-2}{p(q-4)}, q > 4$$

- If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$; that is, the reciprocal of an F random variable is again an F random variable
- If $X \sim t_q$, then $X^2 \sim F_{1,q}$; that is, the square of a t random variable is an F random variable
- If $X \sim F_{p,q}$, then $(p/q)X/(1+(p/q)X) \sim \mathrm{Beta}(p/2,\ q/2)$

pdf and cdf


Snedecor's F pdf
Snedecor's F (df1=10, df2=5)
Snedecor's F (df1=100, df2=5)
Snedecor's F (df1=10, df2=100)
f(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
x
Snedecor's F cdf
Snedecor's F (df1=10, df2=5)
Snedecor's F (df1=100, df2=5)
Snedecor's F (df1=10, df2=100)
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
x

F distribution in R

- pmf
  ```
  df(x, df1, df2)
  ```
- cdf
  ```
  pf(q, df1, df2)
  ```
- Quantile function
  ```
  qf(p, df1, df2)
  ```
- Random numbers
  ```
  rf(n, df1, df2)
  ```

Summary

- One-sample

1. $\bar{X}$ and S^2 are independent random variables;
2. $\dfrac{\bar{X}-\mu}{\sigma / \sqrt{n}} \sim N(0,1)$;
3. $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$;
4. $\dfrac{\bar{X}-\mu}{S / \sqrt{n}} \sim t_{n-1}$

- Two-sample

1. $\dfrac{(\bar{X}-\bar{Y})-\left(\mu_X-\mu_Y\right)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t_{m+n-2}$
2. $\dfrac{S_X^2 / S_Y^2}{\sigma_X^2 / \sigma_Y^2} = \dfrac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2} \sim F_{m-1,n-1}$

Sampling from Other Populations

统计学方法及其应用 统计学基础 随机变量的函数

"A random variable is a quantity whose values are random and to which a probability distribution is assigned."

Order statistics

Order statistics

The **order statistics** of a random sample, $X_1,\ldots,X_n$, are the sample values placed in ascending order. They are denoted by $X_{(1)},\ldots,X_{(n)}$.

$X_{(1)} =$ the smallest X_i; $\quad X_{(1)} = \min_{1 \le i \le n}\{X_i\}$

$X_{(2)} =$ the second smallest X_i;

$\ldots$

$X_{(n)} =$ the largest X_i; $\quad X_{(n)} = \max_{1 \le i \le n}\{X_i\}$

Sample range

Sample range

The **sample range** is the difference between the maximum and the minimum values in a random sample, $X_1, \ldots, X_n$, denoted by

$$R = \max_{1 \le i \le n} X_i - \min_{1 \le i \le n} X_i = X_{(n)} - X_{(1)}.$$

Sample median

Sample median

The **sample median**, usually denoted by M, is a number such that approximately one-half of the observations are less than M and one-half are greater,

$$M = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(X_{(n/2)} + X_{(n/2+1)}\right) & \text{if } n \text{ is even} \end{cases}.$$

A single order statistic

Define

$$I(X_i \leq x) = \begin{cases} 1 & X_i \leq x \\ 0 & X_i > x \end{cases}.$$

Then

$$P(I(X_i \leq x) = 1) = P(X_i \leq x) = F_X(x). \Rightarrow I(X_i \leq x) \sim \text{Bernoulli}(F_X(x)).$$

Define

$$Y = \sum_{i=1}^{n} I(X_i \leq x). \Rightarrow Y \sim \text{Binomial}(n, F_X(x)).$$

Consider

$$F_{X_{(k)}}(x) = P(X_{(k)} \leq x) = P(Y \geq k) = \sum_{i=k}^{n} \binom{n}{i} F_X(x)^i (1 - F_X(x))^{n-i}.$$

Thus

$$f_{X_{(k)}}(x) = \frac{d}{dx} F_{X_{(k)}}(x) = \frac{d}{dx} \sum_{i=k}^{n} \binom{n}{i} F_X(x)^i (1 - F_X(x))^{n-i}.$$

Distribution of a single order statistic

pdf of the k-th order statistic

Let $X_{(1)},\ldots,X_{(n)}$ denote the order statistics of a random sample, $X_1,\ldots,X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the pdf of $X_{(k)}$ is

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} f_X(x)\left[F_X(x)\right]^{k-1}\left[1-F_X(x)\right]^{n-k}.$$

One uniform order statistic

When $X_i \sim \text{uniform}(0,1)$, that is, $f_X(x) = 1, F_X(x) = x$, we have

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}$$
$$= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} x^{k-1}(1-x)^{(n-k+1)-1}.$$

So $f_{X_{(k)}}(x) \sim \text{Beta}(k, n-k+1)$.

$$\mathrm{E}X_{(k)} = \frac{k}{n+1},$$
$$\mathrm{Var}X_{(k)} = \frac{k(n-k+1)}{(n+1)^2(n+2)}.$$

Particularly,

$$\mathrm{E}\min_{1\le i\le n} X_i = \frac{1}{n+1}, \mathrm{E}\max_{1\le i\le n} X_i = \frac{n}{n+1},$$
$$\mathrm{Var}\min_{1\le i\le n} X_i = \mathrm{Var}\max_{1\le i\le n} X_i = \frac{n}{(n+1)^2(n+2)}.$$

http://home.jesus.ox.ac.uk/~clifford/a5/chap2/node8.html

Joint distribution of two order statistics

Joint pdf of two order statistics

Let $X_{(1)},\ldots,X_{(n)}$ denote the order statistics of a random sample, $X_1,\ldots,X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the joint pdf of $X_{(i)}$ and $X_{(j)}$, $1 \leq i < j \leq n$, is

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} f_X(u) f_X(v) \left[F_X(u)\right]^{i-1} \left[F_X(v) - F_X(u)\right]^{j-i-1} \left[1 - F_X(v)\right]^{n-j}.$$

Two uniform order statistics

When $X_i \sim \text{uniform}(0,1)$, that is, $f_X(x) = 1, F_X(x) = x$, we have

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} u^{i-1}(v-u)^{j-1-i}(1-v)^{n-j}$$

$$= \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(j-i)\Gamma(n+1-j)} u^{i-1}(v-u)^{(j-i)-1}(1-v)^{(n+1-j)-1}.$$

So $f_{X_{(i)},X_{(j)}}(u,v) \sim \text{Dir}(i, j-i, n+1-j)$.

Let $R = X_{(n)} - X_{(1)}$ and $V = (X_{(n)} + X_{(1)})/2$, then

$$f_R(r) = n(n-1)r^{n-2}(1-r) = \text{Beta}(n-1,2),$$

and

$$f_V(v) = n\left[2(1-v)\right]^{n-1}.$$

Joint distribution of all order statistics

pdf and cdf of all order statistics

Let $X_{(1)},\ldots,X_{(n)}$ denote the order statistics of a random sample, $X_1,\ldots,X_n$, from a continuous population with pdf $f_X(x)$. Then the joint pdf of $X_{(1)},\ldots,X_{(n)}$ is

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots,x_n) = n!\,f_X(x_1)f_X(x_2)\cdots f_X(x_n)$$

if $-\infty < x_1 < \cdots < x_n < \infty$ and 0 otherwise.

From the above pdf, the joint cdf of $X_{(1)},\ldots,X_{(n)}$ is

$$F_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots,x_n) =$$

$$n!\int_{-\infty}^{x_1}\left[\int_{t_1}^{x_2}\left[\int_{t_2}^{x_3}\cdots\int_{t_{n-1}}^{x_n} f_X(t_n)dt_n\right]f_X(t_2)dt_2\right]f_X(t_1)dt_1.$$

Joint distribution of uniform order statistics

pdf and cdf of uniform order statistics

Let $X_{(1)},\ldots,X_{(n)}$ be the order statistics of a random sample, $X_1,\ldots,X_n$, from a uniform(0, 1) population. Then $f_X(x) = 1$ for $x \in [0,1]$ and 0 otherwise. Therefore, the joint pdf of $X_{(1)},\ldots,X_{(n)}$ is

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots,x_n) = n!$$

if $0 \le x_1 < \cdots < x_n \le 1$ and 0 otherwise.

From the above pdf, the joint cdf of $X_{(1)},\ldots,X_{(n)}$ is

$$F_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots,x_n) = n!\int_0^{x_1}\int_{t_1}^{x_2}\cdots\int_{t_{n-1}}^{x_n} dt_n \cdots dt_2 dt_1.$$

Convergence in probability

Convergence in probability

A sequence of random variables $X_1, \ldots, X_n$, **converges in probability** to a random variable X if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0,$$

or equivalently,

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1.$$

Suppose that $X_1, X_2, \ldots$ converges in probability to a random variable X and that h is a continuous function. Then $h(X_1), h(X_2), \ldots$ converges in probability to $h(X)$.

Example I

Let the sample space $\mathcal{S}$ be the closed interval [0,1] with sample points uniformly distributed.

Define random variable

$$X(s) = s,\ \ s \in S.$$

Define random variables

$$X_n(s) = s + s^n,\ \ s \in S.$$

Then

$$X_n(s) - X(s) = s^n, s \in S.$$

Now, for every $\varepsilon > 0$,

$$\begin{aligned}
&\lim_{n\to\infty} P(|\, X_n(s) - X(s)\,| < \varepsilon) \\
&= \lim_{n\to\infty} P(s^n < \varepsilon) \\
&= P(s \in [0,1)) \\
&= 1.
\end{aligned}$$

Therefore, X_n converges in probability to X.

Example II

Let the sample space $\mathcal{S}$ be the closed interval [0,1] with sample points uniformly distributed.

Define random variable

$$X(s) = s, \ \ s \in S.$$

Define random variables X_n as follows

$$X_1(s) = s + I_{[0,1]}(s),$$
$$X_2(s) = s + I_{[0,1/2]}(s), X_3(s) = s + I_{[1/2,1]}(s),$$
$$X_4(s) = s + I_{[0,1/3]}(s), X_5(s) = s + I_{[1/3,2/3]}(s), X_6(s) = s + I_{[2/3,1]}(s),$$
$$\ldots$$

Then, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\, X_n(s) - X(s)\,| \ \ < \varepsilon)$$
$$= \lim_{k\to\infty} P(\text{length}(I[0,1/k]) < \varepsilon)$$
$$= 1.$$

Therefore, X_n converges in probability to X.

Almost sure convergence (Convergence with probability 1)

Almost sure convergence

A sequence of random variables $X_1, \ldots, X_n$, **converges almost surely** to a random variable X if, for every $\varepsilon > 0$,

$$P\left(\lim_{n\to\infty} | X_n - X | < \varepsilon\right) = 1.$$

Almost sure converge is much **stronger** than converges in probability.
Almost sure converge implies converges in probability.

Example I

Let the sample space $\mathcal{S}$ be the closed interval $[0,1]$ with sample points uniformly distributed.

Define random variable

$$X(s) = s, \;\; s \in S.$$

Define random variables

$$X_n(s) = s + s^n, \;\; s \in S.$$

For every $s \in [0,1)$, as $n \to \infty$,

$$X_n(s) \to s = X(s).$$

However, for $s = 1$,

$$X_n(s) = 1 + 1^n = 2 \neq X(s).$$

Since

$$P(s=1) = 0 \text{ and } P(s \in [0,1)) = 1,$$

X_n converges almost surely to X.

Example II

Let the sample space $\mathcal{S}$ be the closed interval [0,1] with sample points uniformly distributed.

Define random variable

$$X(s) = s, \;\; s \in S.$$

Define random variables X_n as follows

$$X_1(s) = s + I_{[0,1]}(s),$$

$$X_2(s) = s + I_{[0,1/2]}(s), X_3(s) = s + I_{[1/2,1]}(s),$$

$$X_4(s) = s + I_{[0,1/3]}(s), X_5(s) = s + I_{[1/3,2/3]}(s), X_6(s) = s + I_{[2/3,1]}(s),$$

$$\ldots$$

Then, for every $s \in S$, the value $X_n(s)$ alternates between the value of s and $s+1$ infinitely often. Therefore, there is no value of $s \in S$ for which $X_n(s) \to s = X(s)$. In other words,

altuhough X_n converges in probability to X,

X_n does **NOT** converge almost surely to X.

Convergence in distribution

Convergence in distribution

A sequence of random variables $X_1, \ldots, X_n$, **converges in distribution** to a random variable X if,

for every $\varepsilon > 0$,

$$\lim_{n\to\infty} F_{X_n} = F_X(x)$$

at all points x where $F_X(x)$ is continuous.

If the sequence of random variables, $X_1, X_2, \ldots$ converges in probability to a random variable X, the sequence also converges in distribution to X.

Example

Let $X_1, X_2, \cdots$ be iid $uniform(0,1)$ random variables. Let $X_{(n)} = \max_{1 \le i \le n} X_i$.
As $n \to \infty$, $X_{(n)}$ gets close to 1, but must necessarily be less than 1. Therefore

$$
\begin{aligned}
P(|X_{(n)} - 1| \ge \varepsilon) &= \underbrace{P(X_{(n)} \ge 1 + \varepsilon)}_{=0} + P(X_{(n)} \le 1 - \varepsilon) \\
&= P(X_{(n)} \le 1 - \varepsilon).
\end{aligned}
$$

However,

$$
\begin{aligned}
P(X_{(n)} \le 1 - \varepsilon) &= P(\max_{1 \le i \le n} X_i \le 1 - \varepsilon) \\
&= P(X_i \le 1 - \varepsilon, i = 1, \dots, n) \\
&= (1 - \varepsilon)^n \\
&\to 0, \text{ as } n \to \infty.
\end{aligned}
$$

Therefore, $X_{(n)}$ converges to 1 in probability.

Example (continued)

Furthermore, let $\varepsilon = t / n,$ then

$$P(X_{(n)} \le 1 - t / n) = (1 - t / n)^n \to e^{-t},$$

that is

$$P(n(1 - X_{(n)}) \le t) \to 1 - e^{-t}.$$

Recall the *exponential*(1) distribution.

$$f(x) = e^{-x},$$

$$F(x) = \int_0^x e^{-t} dt = -e^{-t} \Big|_0^x = 1 - e^{-x}.$$

Hence,

$n(1 - X_{(n)})$ converges in distribution to an exponential random variable.

Chebychev's inequality

Chebychev's inequality

Let X be a random variable and let $g(x)$ be a nonnegative function. Then for any $r > 0$

$$P(g(X) \geq r) \leq \frac{\mathrm{E}g(X)}{r}.$$

$$\begin{aligned}
\mathrm{E}g(X) &= \int_{-\infty}^{\infty} g(x)f(x)dx \\
&\geq \int_{\{x:g(x)\geq r\}} g(x)f(x)dx \\
&\geq r\int_{\{x:g(x)\geq r\}} f(x)dx \\
&= rP(g(X)\geq r)
\end{aligned}$$

Weak law of large numbers (WLLN)

Let $X_1, \ldots, X_n$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$.
Define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Then,

$$P(|\bar{X}_n - \mu| \geq \varepsilon) = \underbrace{P(|\bar{X}_n - \mu|^2 \geq \varepsilon^2) \leq \frac{\mathrm{E}(\bar{X}_n - \mu)^2}{\varepsilon^2}}_{\text{Chebychev's inequality}} = \frac{\mathrm{Var}\bar{X}_n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Hence,

$$P(|\bar{X}_n - \mu| < \varepsilon) = 1 - P(|\bar{X}_n - \mu| \geq \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2} \rightarrow 1, \text{ as } n \rightarrow \infty.$$

In other words

$$\lim_{x \rightarrow \infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1.$$

Weak law of large numbers

Weak law of large numbers

Let $X_1, \ldots, X_n$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Then, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1;$$

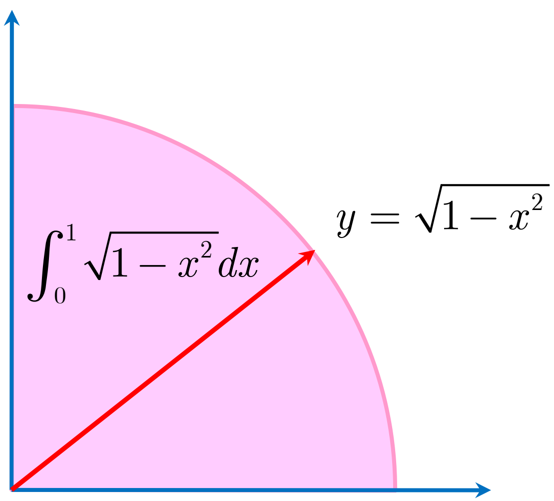
that is, $\bar{X}_n$ converges in probability to μ.

Sample mean becomes population mean when the sample size tends to infinity.

Monte Carlo integration

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathrm{E}_{p(x)} f(X)\right| < \varepsilon\right) = 1 \Rightarrow \mathrm{E}_{p(x)} f(X) \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i), \text{ as } n \to \infty$$

$$\mathrm{E}_{p(x)}\left[f(X)\right] = \int_{-\infty}^{\infty} f(x)p(x)dx \Rightarrow \underbrace{\int_{-\infty}^{\infty} h(x)dx = \int_{-\infty}^{\infty} f(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)}_{\text{Monte Carlo integration}}$$

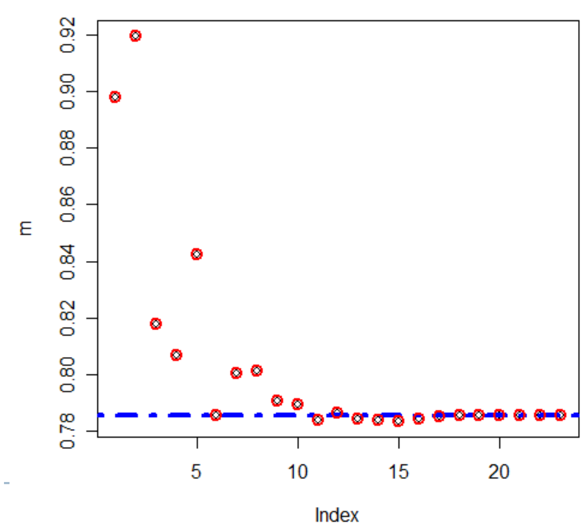


$$\int_0^1 \sqrt{1-x^2}\,dx = \frac{1}{2}\left(x\sqrt{1-x^2} + \arcsin x\right)\Big|_0^1 = \frac{\pi}{4}$$

$$\int_0^1 \sqrt{1-x^2}\,dx = \int_0^1 \underbrace{\sqrt{1-x^2}}_{f(x)} \cdot \underbrace{1}_{p(x)}\, dx$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} \sqrt{1-x_i^2},$$

where every x_i is sampled from a uniform(0,1) distribuiton.



Convergence of sample variance

Weak law of large numbers for sample variance

Let $X_1, \ldots, X_n$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$. Define

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

Then, for every $\varepsilon > 0$,

$$P(|\bar{S}_n^2 - \sigma^2| \geq \varepsilon) \leq \frac{\mathrm{E}(\bar{S}_n^2 - \sigma^2)^2}{\varepsilon^2} = \frac{\mathrm{Var}\bar{S}_n^2}{\varepsilon^2}.$$

So, if $\mathrm{Var}\bar{S}_n^2 \to 0$, then $\bar{S}_n^2$ converges to σ^2 in probability.

Strong law of large numbers (SLLN)

Strong law of large numbers

Let $X_1,\dots,X_n$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$.

Then, for every $\varepsilon > 0$,

$$P\left(\lim_{n\to\infty} |\bar{X}_n - \mu| < \varepsilon\right) = 1;$$

that is, $\bar{X}_n$ converges almost surely to μ.

The central limit theorem (CLT)

The central limit theorem

Let $X_1,\ldots,X_n$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Then, for any x, $-\infty < x < \infty$,

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \le x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

$\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting standard normal distribution.

The distribution of normalized sample mean becomes standard normal distribution when sample size tends to infinity.

Normal approximation of binomial

- Bernoulli trial

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

- Multiple Bernoulli trials
  - From concept

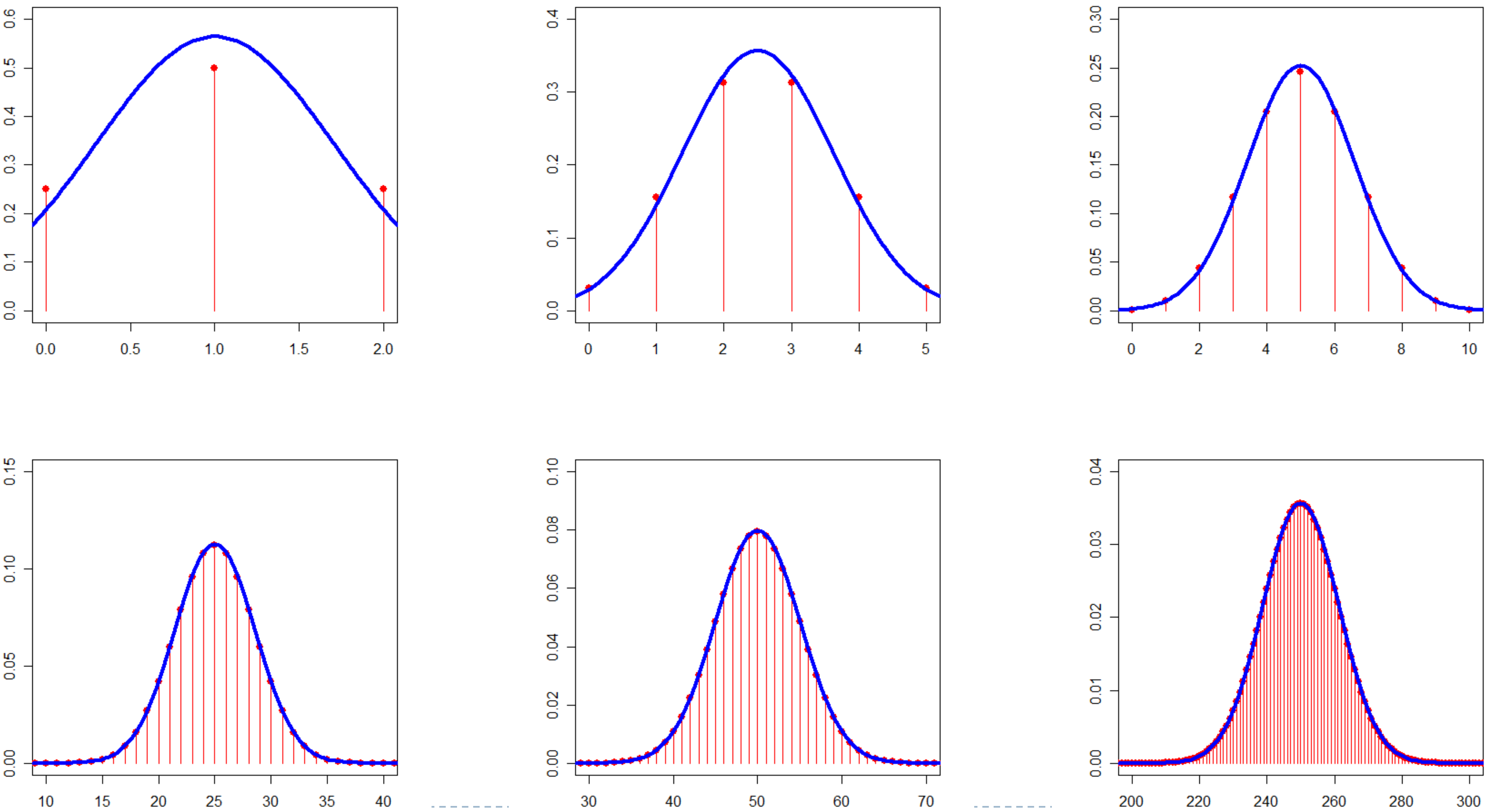
$$Y_i = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$$

  - From the central limit theorem

$$Z_n = \frac{Y_i - np}{\sqrt{np(1-p)}} = \frac{Y_i / n - p}{\sqrt{p(1-p) / n}} \sim N(0,1), \text{ as } n \to \infty$$

$$Y_i \sim N\left(np, np(1-p)\right), \text{ as } n \to \infty$$

Normal approximation of binomial



Thank you very much