Statistics

2023 Lectures
Part 7 - Random Samples

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Definition of random sample and statistics

Definition 31: An n-tuple of independent identically distributed random variables $X_1, \ldots, X_n$ is called random sample of size n.

Definition 32: A function of random sample is called statistic if it depends on random sample but not on parameters of distribution.

- sample sum: $T_n = X_1 + \cdots + X_n$
- sample mean: $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$
- sample variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- sample deviation: $S = \sqrt{S^2}$
- smallest values in the data: $\min(X_1, \ldots, X_n)$
- largest values in the data: $\max(X_1, \ldots, X_n)$

Random sample is an important link between the observed data and the distribution in the population from which it has been selected.

Properties of sample mean and variance

Example 73: Let $X_1, \ldots, X_n$ be random sample from $EXP(\lambda)$. What is the distribution of $\bar{X}$?

$$m_{\bar{X}}(t) = m_{T_n}\left(\frac{t}{n}\right) = \left(m_X\left(\frac{t}{n}\right)\right)^n = \left(1 - \frac{t}{n\lambda}\right)^{-n}, t \in \mathbb{R}$$

and thus $\bar{X} \sim GAM(n, n\lambda)$.

Theorem 47: Let $X_1, \ldots, X_n$ be a random sample from a distribution with $EX = \mu$ and $VarX = \sigma^2$. Then

$$E\bar{X} = \mu, Var\bar{X} = \frac{\sigma^2}{n},$$
$$ET_n = n\mu, VarT_n = n\sigma^2,$$
$$ES^2 = \sigma^2.$$

Normally distributed random sample and t distribution

Theorem 48: (without proof) Let $X_1, \ldots, X_n$ be a random sample from the normal distribution $N(\mu, \sigma^2)$. Then $\bar{X}$ and

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

are independent random variables with distributions $N(\mu, \frac{\sigma^2}{n})$ and χ^2_{n-1}. On the other hand, if $\bar{X}$ and $\frac{n-1}{\sigma^2} S^2$ are independent, then $X_i \sim N(\mu, \sigma^2)$.

Definition 33: Let $Z \sim N(0,1)$ and $U \sim \chi^2_\nu$ be independent rv's. Then

$$X = \frac{Z}{\sqrt{\frac{U}{\nu}}} \sim t_\nu$$

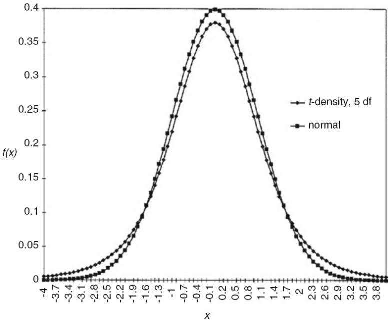
is said to have Student's t distribution with ν degrees of freedom.

Properties of t distribution

If $X \sim t_\nu$ then the density function is

$$f_X(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

for $x \in \mathbb{R}$.



Theorem 49: (without proof) Student's t_n distribution approaches $N(0,1)$ as $n \to \infty$.

- for normally distributed random sample, by definition

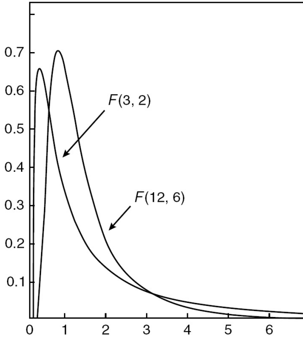
$$\frac{\bar{X} - \mu}{S}\sqrt{n} \sim t_{n-1}.$$

F distribution

Definition 34: Let U and V are independent rv's with $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$, respectively. Then the random variable

$$X = \frac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}} \sim F_{\nu_1,\nu_2}$$

is said to have Fisher Snedecor's F distribution with ν_1 and ν_2 degrees of freedom.



- both t and F distributions are of high importance in statistics, especially for testing statistical hypotheses
- there are usually tables of quantiles for t and F distribution for the most frequently used values of degrees of freedom along with quantiles of standardized normal distribution at the end of books on statistics

Order statistics

Definition 35: Let $X_1, \ldots, X_n$ be a random sample from continuous distribution with a cdf F and density f. Rv's

$$X_{1:n} \leq X_{2:n} \leq \ldots X_{n:n},$$

where $X_{i:n}$ is the ith in magnitude among $X_1, \ldots, X_n$, are called order statistics.

Let G_k be a cdf of $X_{k:n}$, then $X_{k:n} \leq t$ if at least k variables in $X_1, \ldots, X_n$ satisfy $X_i \leq t$, which fits binomial distribution with n and $p = P(X_i \leq t) = F(t)$. Thus

$$G_k(t) = \sum_{r=k}^{n} \binom{n}{r} (F(t))^r (1 - F(t))^{n-r}$$

$k = 1 : G_1(t) = 1 - (1 - F(t))^n$
$k = n : G_n(t) = (F(t))^n$

- Unlike variables $X_1, \ldots, X_n$ in a random sample, order statistics $X_{1:n}, \ldots, X_{n:n}$ are dependent variables.

Convergence in probability

Definition 36: The sequence $\{\xi_n\}$ converges in probability to

- a constant c if for every $\varepsilon > 0$

$$\lim_{n\to\infty} P(|\xi_n - c| \geq \varepsilon) = 0.$$

- a random variable ξ if for every $\varepsilon > 0$

$$\lim_{n\to\infty} P(|\xi_n - \xi| \geq \varepsilon) = 0.$$

We write $\xi_n \xrightarrow{P} c$ or $\xi_n \xrightarrow{P} \xi$.

- meaning: as n increases, it becomes less and less likely that ξ_n will deviate from c (or ξ) by more than ε
- Explicitly:

$$\lim_{n\to\infty} P(s \in S : |\xi_n(s) - \xi(s)| \geq \varepsilon) = 0.$$

Convergence almost surely

Definition 37: Let $\xi_1, \xi_2, \ldots$ be a sequence of random variables defined on $(S, \mathcal{A}, P)$. If $\lim \xi_n(s) = \xi(s)$ for all points $s \in U$, where $U \subset S$ with $P(U) = 1$, then we say that ξ_n converges to ξ almost everywhere (almost surely). We write $\xi_n \xrightarrow{a.s.} \xi$.

- meaning: as n increases, for almost every sample point $s \in S$ the sequence of values $\xi_n(s)$ converges to $\xi(s)$

Theorem 50: (without proof) If $\xi_n \xrightarrow{a.s.} \xi$ then $\xi_n \xrightarrow{P} \xi$.

Convergence a.s. implies convergence in probability

Example 74: Let $X \sim U[0,1]$. Let

$$\begin{aligned}
I_1 &= [0,1]\,, \\
I_2 &= \left[0, \frac{1}{2}\right]\,, \\
I_3 &= \left[\frac{1}{2}, 1\right]\,, \\
&\vdots \\
I_{2^m+i} &= \left[\frac{i}{2^m}, \frac{i+1}{2^m}\right],\ i = 0, 1, \dots, 2^m - 1,\ m = 0, 1, 2, \dots
\end{aligned}$$

Let

$$\xi_n = \begin{cases} 1, & \text{if } X \in I_n; \\ 0, & \text{if } X \notin I_n. \end{cases}$$

Then $\xi_n \xrightarrow{P} 0$ but not a.s. since infinitely many ξ_n are equal 1.

Weakness of convergence in probability

- if sequence of ξ_n converges in probability to a constant, the sequence of expected values $E\xi_n$ can converge to another value or even diverge

Example 75: Let $\xi_1, \xi_2, \ldots$ are independent rv's with $P(\xi_n = 1) = 1 - \frac{1}{n}, P(\xi_n = n) = \frac{1}{n}$.

Then $\xi_n \xrightarrow{P} 1$, yet $E\xi_n \to 2$.

Example 76: Let $\xi_1, \xi_2, \ldots$ are independent rv's with $P(\xi_n = 1) = 1 - \frac{1}{n}, P(\xi_n = n^2) = \frac{1}{n}$.

Then $\xi_n \xrightarrow{P} 1$, yet $E\xi_n \to \infty$.

Convergence in distribution

Definition 38: Let $\xi_0, \xi_1, \xi_2, \ldots$ be a sequence of random variables and let $F_n(t) = P(\xi_n \leq t), n = 0, 1, 2, \ldots$ be their cdf's. The sequence $\{\xi_n\}_1^\infty$ converges in distribution to ξ_0 if

$$\lim_{n\to\infty} F_n(t) = F_0(t)$$

for every t at which $F_0(t)$ is continuous. We write $\xi_n \xrightarrow{d} \xi$.

Example 77: Let $\xi_n = \frac{1}{n}$ with probability 1, $\xi_0 = 0$. Then $\xi_n \xrightarrow{d} \xi_0$.

$$F_n(t) = P(\xi_n \leq t) = \begin{cases} 0, & \text{if } t < \frac{1}{n}; \\ 1, & \text{if } t \geq \frac{1}{n}. \end{cases}$$

at $t \neq 0$: $\lim F_n(t) = F_0(t)$
at $t = 0$: $F_0(0) = 1$ while $F_n(0) = 0$ for every n.

Chain of implications

Theorem 51: (without proof) Let ξ_n and η_n be sequences of random variables, ξ be a random variable and c be a constant. Further,

a) (Slutsky) let $\xi_n - \eta_n \xrightarrow{P} 0$ and $\eta_n \xrightarrow{d} \xi$. Then $\xi_n \xrightarrow{d} \xi$.

b) let $\xi_n \xrightarrow{P} \xi$. Then $\xi_n \xrightarrow{d} \xi$.

c) let $\xi_n \xrightarrow{a.s.} \xi$. Then $\xi_n \xrightarrow{d} \xi$.

d) let $\xi_n \xrightarrow{d} c$. Then $\xi_n \xrightarrow{P} c$.

e) let $\xi_n \xrightarrow{d} \xi$ and a, b, a_n, b_n be constants such that $a_n \to a$, $b_n \to b$. Then $a_n\xi_n + b_n \xrightarrow{d} a\xi + b$.

f) let $m_n(t)$ and $m(t)$, the mgf's of ξ_n and ξ, respectively, exist. Then $\xi_n \xrightarrow{d} \xi$ if and only if $m_n(t) \to m(t)$ for all $t \in \mathcal{O}(0)$ for some neighborhood $\mathcal{O}$ of 0.

Weak and Strong Laws of Large Numbers

Theorem 52: (Weak Law of Large Numbers)
Let $X_1, X_2, \ldots$ be a sequence of iid rv's. Assume that $EX_i = \mu$ and $VarX_i = \sigma^2 > 0$ for all i. Then for every $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) = 0.$$

Theorem 53: (Generalized Weak Law of Large Numbers)
Let $X_1, X_2, \ldots$ be a sequence of independent rv's. Assume that $EX_i = \mu_i$ and $VarX_i = \sigma_i^2 > 0$ such that

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 = 0.$$

Then for every $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \bar{\mu}\right| \geq \varepsilon\right) = 0.$$

Special case and Strong Law of Large Numbers

Example 78: (Weak LLN for binomial distribution)
If S_n has binomial distribution $BIN(n,p)$ then

$$\frac{S_n}{n} \xrightarrow{P} p.$$

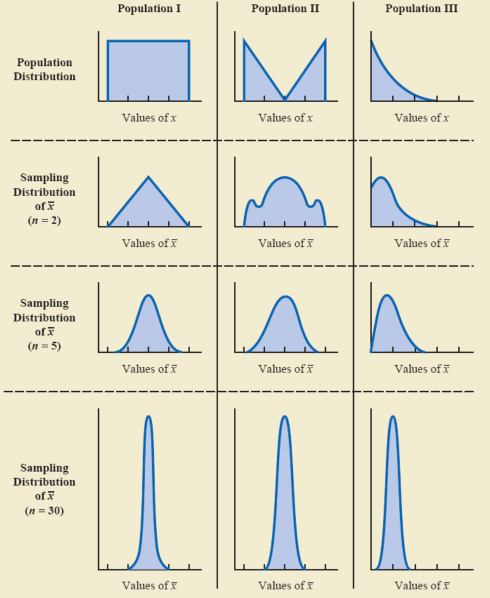
This explains why expected value of $ALT(p)$ is p.

- Idea: if we take the averages of larger and larger numbers of observations then it becomes less and less likely that the average deviates from the “true average” EX

Theorem 54: (without proof) (Strong Law of Large Numbers)
Let $X_1, X_2, \dots$ be independent, with $EX_i = \mu_i, VarX_i = \sigma_i^2$, $i = 1,2,3,\dots$. If $\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} < \infty$ then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{a.s.} 0.$$

Central limit theorem



Central limit theorem

Central limit theorem is a term to designate a theorem that asserts that the sum of large numbers of random variables after standardization have approximately a standard normal distribution.

Theorem 55: (without proof) (Levy - Lindenberg CLT)
Let $X_1, X_2, \ldots$ be a sequence of iid rv's with $EX_i = \mu$ and $VarX_i = \sigma^2, 0 < \sigma^2 < \infty$. Then letting $S_n = X_1 + \ldots X_n$, for every x

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{d}t.$$

We can rephrase: Let $Z \sim N(0,1)$ and $S_n^* = \frac{S_n - ES_n}{\sqrt{VarS_n}}$ then $S_n^* \xrightarrow{d} Z$.

- Special case for binomial distribution, Laplace CLT is historically the oldest version of CLT.

Example

Example 79: Let us sum up 300 numbers rounded up to one decimal. The error of the sum cannot exceed $300 \cdot 0.05 = 15$. Assume the errors $X_k, k = 1, \ldots, 300$ are independent variables with uniform distribution on $(-0.05, 0.05)$ with $EX = 0$ and $VarX = \frac{1}{1200}$.

Standardized error

$$Z_{300} = \frac{\sum_1^{300} X_k}{\sqrt{\frac{300}{1200}}} = 2\sum_1^{300} X_k \approx Z \sim N(0,1).$$

Then $P(|\sum_1^{300} X_k| \leq \varepsilon) = P(|Z_{300}| \leq 2\varepsilon) \doteq 2\Phi(2\varepsilon) - 1$. E.g., for $\varepsilon = 1$ this probability equals roughly 0.9545.

Recommendation:

- For $X_1, \ldots, X_n$ iid, $n \geq 30$
- For Laplace CLT, $\min\{np, nq\} \geq 5$

Another example

Example 80: A fair coin is tossed $n = 15$ times. Find the approximate probability that the number of heads S_{15} will satisfy $8 \le S_{15} < 10$.

$S_{15} = 8$ or 9; $np = 15 \cdot 0.5 = 7.5$; $\sqrt{npq} = \sqrt{15 \cdot 0.5^2} \doteq 1.94$

$$P(8 \le S_{15} < 10) = P(8 \le S_{15} \le 9) \overset{\text{CLT}}{\approx} \Phi\left(\frac{9-7.5}{\sqrt{15/4}}\right) - \Phi\left(\frac{7-7.5}{\sqrt{15/4}}\right) \approx 0.3826.$$

$$P(8 \le S_{15} < 10) \overset{\text{with CC}}{=} P(8 - 0.5 \le S_{15} \le 9 + 0.5)$$

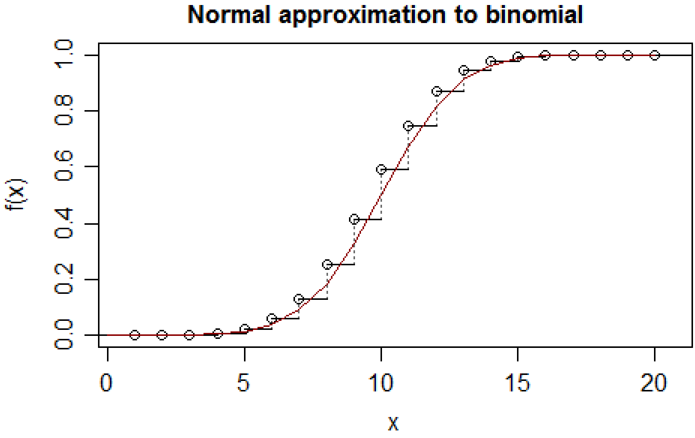
$$\overset{\text{CLT}}{\approx} \Phi\left(\frac{9.5-7.5}{\sqrt{15/4}}\right) - \Phi\left(\frac{7.5-7.5}{\sqrt{15/4}}\right) \approx 0.3492.$$

$$P(8 \le S_{15} < 10) \overset{\text{exact}}{=} \binom{15}{8}\left(\frac{1}{2}\right)^{15} + \binom{15}{9}\left(\frac{1}{2}\right)^{15} = 0.3491.$$

- addition and subtraction of 0.5 is called continuity correction (it may not always help)

Continuity correction visualized

Example 81: Consider $X \sim BIN(20, 0.5)$.
The graph of the cdf of X and the cdf of $N(10, 5)$.



More CLTs

Theorem 56: (without proof) (Liapunov)
Let $X_1, X_2, \ldots$ be a sequence of independent rv's such that $EX_i = \mu_i$, $VarX_i = \sigma_i^2$, $\gamma_i = E|X_i - \mu_i|^3 < \infty$. Put $m_n = \sum_{j=1}^n \mu_j$, $s_n^2 = \sum_{j=1}^n \sigma_j^2$, $\Gamma_n = \sum_{j=1}^n \gamma_j$. If $\lim_{n\to\infty} \frac{\Gamma_n}{s_n^3} = 0$ then

$$\frac{S_n - m_n}{s_n} \xrightarrow{d} Z \sim N(0,1).$$

- Feller-Lindenberg CLT
- many many more CLTs