Introduction & Random Sample

Contents:

- Random Sample
- Distribution of sum of random variables from a random sample
- Random sample from a normal distribution
- Large sample behavior of the important statistics

Definition

If $X_1, \cdots, X_n$ are independent random variables with common marginal distribution with cdf $F(x)$ then we say that they are independent and identically distributed (iid) with common cdf $F(x)$ or $X_1, \cdots, X_n$ are random sample from a infinite population with distribution $F(x)$.

Introduction & Random Sample

$X_1, \cdots, X_n$ is a random sample from $F(x)$

- $$X_1, \cdots, X_n \overset{iid}{\sim} F(x)[\ or \ \ f(x)]$$

- $$\begin{aligned} F_{X_1,\cdots,X_n}(x_1, \cdots, x_n) &= P[X_1 \leq x_1, \cdots, X_n \leq x_n] \\ &= \prod_{i=1}^{n} P[X_i \leq x_i] \\ &= \prod_{i=1}^{n} F(x_i). \end{aligned}$$

- $$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i).$$

Sum of RV from a RS

Definition

Any function of random variables $X_1, \cdots, X_n$ is called *Statistic*.
[function of $X_1, \cdots, X_n$ only not parameters]

▷ Example:

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

$$T(X_1, \cdots, X_n) = \max(X_1, \cdots, X_n)$$

- Statistic is also a random variable
- Interest in distribution of a statistic

Sum of RV from a RS

Definition

The distribution of the statistic is called *sampling distribution* (of the statistic) in contrast to the population distribution

Definition

$$\text{Sample Mean: } \bar{X} = \frac{X_1 + \cdots + X_n}{n},$$

$$\text{Sample Variance: } S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

Sum of RV from a RS

[Theorem 5.2.6]

Theorem

Let $X_1, \cdots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$*. Then*

$$E\left(\bar{X}\right) = \mu, \;\; Var\left(\bar{X}\right) = \sigma^2/n, \;\; E\left(S^2\right) = \sigma^2 \; .$$

Proof: [See textbook for another proof.]

Sampling from the normal distribution

• Find the sampling distributions of a certain statistics that are functions of random sample $X_1, \cdots, X_n$ from a normal distribution.

Lemma

Let $X_1, \cdots, X_n$ be independent random variables. Let $g_i(x_i)$ be a function of x_i. Then the random variables $U_i = g_i(X_i),\ i = 1, \cdots, n$ are mutually independent.

Theorem

1. If $Z \sim N(0,1)$ then $Z^2 \sim \chi^2(1)$

2. $X_i, i = 1 \cdots, n$ are independent random variables, $X_i \sim \chi^2(p_i)$. Then $X_1 + \cdots + X_n \sim \chi^2(p_1 + \cdots, p_n)$.

Sampling from the normal distribution

Theorem

$$X_1, \cdots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$$

1. $\bar{X}$ *and* S^2 *are independent*
2. $\bar{X} \sim N(\mu, \sigma^2/n)$
3. $[(n-1)S^2]/\sigma^2 \sim \chi^2(n-1)$

• Other sampling distributions of sample mean and the ratio of sample variances.

:Student's *t*-distribution and Snedecor's *F*-distribution.

Sampling from the normal distribution

Definition (Students's t distribution)

The t-distribution with d.f. ν is the distribution of

$$T = \frac{Z}{\sqrt{W/\nu}},$$

where Z and W are independent with $Z \sim N(0,1)$, $W \sim \chi^2(\nu)$.
The pdf of the t distribution is

$$f_T(t) = \frac{\Gamma\left[(\nu+1)/2\right]}{\Gamma\left[\nu/2\right]} \frac{1}{\sqrt{\pi\nu}} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \quad -\infty < t < \infty.$$

- Similar shape with $N(0,1)$
- Approaches to $N(0,1)$ as $\nu \to \infty$
- Has a heavier and flatter tail than $N(0,1)$

Sampling from the normal distribution

Definition (Snedecor's F distribution)

Let $W \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$. Assume W and V are independent. Then the distribution of

$$F = \frac{W/\nu_1}{V/\nu_2}$$

has a F distribution with d.f.'s (ν_1, ν_2). The pdf of the F distribution is

$$f_F(x) = \frac{\Gamma\left[(\nu_1+\nu_2)/2\right]}{\Gamma\left[\nu_1/2\right]\Gamma\left[\nu_2/2\right]}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}\frac{x^{(\nu_1/2)-1}}{\left[1+(\nu_1/\nu_2)x\right]^{(\nu_1+\nu_2)/2}}$$

where $0 < x < \infty$.

Sampling from the normal distribution

▷ Example of t-statistics

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

▷ Example of F-statistics

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$, and $Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$. Then

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{(n-1),(m-1)}$$

Sampling from the normal distribution

Theorem

1. *If* $X \sim F(\nu_1, \nu_2)$ *then* $1/X \sim F(\nu_2, \nu_1)$
2. *If* $X \sim t(\nu)$ *then* $X^2 \sim F(1, \nu)$
3. *If* $X \sim F(\nu_1, \nu_2)$ *then*

$$\frac{(\nu_1/\nu_2)X}{1 + (\nu_1/\nu_2)X} \sim Beta(\nu_1/2, \nu_2/2)$$

Proof: See Exercise 5.17 and 5.18.

If $X \sim F(\nu_1, \nu_2)$, then

- $E(X) = \frac{\nu_2}{\nu_2 - 2}$ for $\nu_2 > 2$
- $Var(X) = \frac{2\nu_2^2(\nu_1+\nu_2+2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ for $\nu_2 > 4$

Convergence Concepts

- Investigate the large sample behaviors of the sequence of random variables.
  - ◁ Convergence in probability
  - ◁ Almost sure convergence
  - ◁ Convergence in distribution
  - ◁ Delta method

Convergence Concepts

Convergence in probability

Definition ($X_n \xrightarrow{P} X$)

A sequence of random variables $X_1, X_2, \cdots$ *converges in probability* to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P[|X_n - X| \geq \epsilon] = 0$$

or equivalently

$$\lim_{n\to\infty} P[|X_n - X| < \epsilon] = 1.$$

▷ Example: $X \sim F_X(x)$, $X_n = [(n-1)/n]X$. Then $X_n \xrightarrow{P} X$.

Theorem (WLLN)

If $X_1, \cdots X_n \overset{iid}{\sim} (\mu, \sigma^2)$, $\sigma^2 < \infty$ *Then* $\bar{X}_n \xrightarrow{P} \mu$.

Proof.

Theorem

If $X_n \xrightarrow{P} X$ *and* g *is a function defined on the range of* X *such that* $D_g = \{x | g \text{ is discontinuous at } x\}$ *has* $P[X \in D_g] = 0$, *then* $g(X_n) \xrightarrow{P} g(X)$.

▷ Examples: $S^2 \xrightarrow{P} \sigma^2$? $\quad S \xrightarrow{P} \sigma$?

Definition ($X_n \to X$ *a.s.* (Or, $X_n \overset{a.s.}{\to} X$))

A sequence of random variables $X_1, X_2, \cdots$ *converges almost surely* to a random variable X if, for every $\epsilon > 0$,

$$P\left\{\lim_{n\to\infty} [|X_n - X| < \epsilon]\right\} = 1.$$

Note: X_n converges to X almost surely if the functions $X_n(s)$ converges to $X(s)$ for all $s \in S$ (S: sample space) except for some singletons. (except for $s \in N$, where $N \subset S$ and $P(N) = 0$)

$\triangleright$ Example: $X \sim F_X(x)$, $X_n = [(n-1)/n]X$. Then $X_n \to X$ *a.s.*

Convergence Concepts

- Almost sure convergence $\longrightarrow$ Convergence in probability

$\triangleright$ Example: Almost sure convergence $\longleftarrow$ Convergence in probability ? [Example 5.5.8]

Let $U \sim \text{Uniform}(0,1)$, $X_n = U + I_n$ and $X = U$, where

$$\begin{array}{llcl}
I_1 = I(0 < U < 1) & , & p_1 = P[I_1 = 1] = 1 \\
I_2 = I(0 < U \leq 1/2) & , & p_2 = P[I_2 = 1] = 1/2 \\
I_3 = I(1/2 < U \leq 1) & , & p_3 = P[I_3 = 1] = 1/2 \\
I_4 = I(0 < U \leq 1/3) & , & p_4 = P[I_4 = 1] = 1/3 \\
I_5 = I(1/3 < U \leq 2/3) & , & p_5 = P[I_5 = 1] = 1/3 \\
I_6 = I(2/3 < U \leq 1) & , & p_6 = P[I_6 = 1] = 1/3 \\
\vdots & & \vdots
\end{array}$$

Convergence Concepts

Almost sure convergence

$$P[|X_n - X| \geq \epsilon] = P[I_n \geq \epsilon]$$

Is the sequence I_n for a given value of $U = u$ converge ?
(For example, consider when $u = 1/4$. Then observe the values of X_n.)

Theorem (SLLN)

If $X_1, \cdots X_n \overset{iid}{\sim} (\mu, \sigma^2)$, $\sigma^2 < \infty$. *Then* $\bar{X}_n \overset{a.s.}{\rightarrow} \mu$.

Definition ($X_n \xrightarrow{D} X$)

A sequence of random variables $X_i \sim F_i, i = 1, \cdots$, i.e., $F_i(t) = Pr[X_i \leq t]$. Suppose that X is a random variable with cdf F, i.e., $F(t) = Pr[X \leq t]$. Then the sequence of random variables $X_1, X_2, \cdots$ *converges in distribution* to a random variable X if

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

for all continuity points of F.

▷ Example: Let $X_1, X_2, \cdots$ be iid U(0,1). Let $X_{(n)}$ be the $\max_{1\leq i\leq n} X_i$. Then $n(1 - X_{(n)}) \xrightarrow{D}$ Exp(1). [Example 5.5.11]

Theorem (CLT)

If $\mathbf{X}_1, \mathbf{X}_2, \cdots$ are iid p-dimensional random vectors with finite mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then

$$\sqrt{n}\left(\bar{\mathbf{X}}_n - \boldsymbol{\mu}\right) \xrightarrow{D} N_p\left(\mathbf{0}, \boldsymbol{\Sigma}\right),$$

where $\mathbf{X}_i = (x_{1i}, x_{2i}, \cdots, x_{pi})'$, $\bar{\mathbf{X}}_n = \sum_{i=1}^n \mathbf{X}_i$,
$\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)'$, *and*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

Theorem (CLT with $p = 1$ (Theorem 5.5.14))

If $X_1, X_2, \cdots$ are iid random variables with finite mean μ and variance σ^2 whose mgfs exist in a neighborhood of 0. Then

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \xrightarrow{D} N\left(0, 1\right).$$

Proof: [The proof in the textbook implies the following theorem.]

Theorem (Continuity Theorem)

If sequence of mgf $M_n(t) \to M(t)$ for all t in an open interval containing zero, then the corresponding cdfs $F_n(x) \to F(x)$ at all continuity point of F. That is $X_n \xrightarrow{D} X$.

Theorem (Slutsky's theorem)

Let $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c$, where c is a constant. Then

1. $Y_n X_n \xrightarrow{D} cX$
2. $Y_n + X_n \xrightarrow{D} c + X$
3. $g(Y_n, X_n) \xrightarrow{D} g(c, X)$ *in general when g is continuous.*

▷ Example:

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S} \xrightarrow{D}$$

Theorem (Theorem 5.5.24)

Let $Y_n,\ n = 1, 2, \cdots$ be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \xrightarrow{D} N(0, \sigma^2)$. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then

$$\sqrt{n}\,[g(Y_n) - g(\theta)] \xrightarrow{D} N\left[0, \sigma^2 \left(g'(\theta)\right)^2\right].$$

$\triangleright$ Example: $X_1, \cdots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$. $g(\mu) = e^{\mu}$, $g(\mu) = 1/\mu$.

Find the limiting distributions of $\sqrt{n}(g(\overline{X_n}) - g(\mu))$, where $\overline{X_n} = \sum_{i=1}^{n} X_i/n$

Theorem (Theorem 5.5.26)

Let $Y_n,\ n = 1, 2, \cdots$ be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \xrightarrow{D} N(0, \sigma^2)$. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then

$$n\left[g(Y_n) - g(\theta)\right] \xrightarrow{D} \sigma^2 \frac{g''(\theta)}{2} \chi_1^2.$$

Proof: By Taylor expansion

$$\begin{aligned} g(Y_n) &= g(\theta) + g'(\theta)(Y_n - \theta) + \frac{g''(\theta)}{2}(Y_n - \theta)^2 + \frac{g'''(\xi)}{3!}(Y_n - \theta)^3 \\ &= g(\theta) + \frac{g''(\theta)}{2}(Y_n - \theta)^2 + \text{ Remainder} \end{aligned}$$

Order Statistics

The order statistics of a random sample $X_1, \cdots, X_n$ are the sample values placed in ascending order. They are denoted by

$$X_{(1)}, X_{(2)}, \cdots, X_{(n)},$$

where $X_{(i)}$ is the i^{th} smallest of $X_1, \cdots, X_n$. Then joint pdf is

$$f_{X_{(1)},\cdots,X_{(n)}}(y_1, \cdots, y_n) = n! f_X(y_1) \cdots f_X(y_n),$$

for $-\infty < y_1 < \cdots < y_n < \infty$, and $f_X(\cdot)$ is pdf of X_i's.

- Statistics defined in terms of order statistics.
  1. Sample Range: $= X_{(n)} - X_{(1)}$
  2. Sample Median:

$$\begin{array}{ll} X_{([n+1]/2)}, & \text{if } n \text{ is odd} \\ [X_{(n/2)} + X_{(n/2+1)}]/2, & \text{if } n \text{ is even} \end{array}$$

Order Statistics

- Distribution of order statistics

Let $X_1, \cdots, X_n$ be a random sample with a common pdf $f_X(x)$ and a common cdf $F_X(x)$. Then the marginal and joint distributions of order statistics are as follow:

▷ Marginal distribution

$$f_{X_{(j)}}(y_j) = \frac{n!}{(j-1)!(n-j)!} [F_X(y_j)]^{j-1}$$
$$[1 - F_X(y_j)]^{n-j} f_X(y_j),$$

for $-\infty < y_j < \infty$.

Proof:

Order Statistics

▷ Joint distribution

$$f_{X_{(j)},X_{(k)}}(y_j, y_k) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!} [F_X(y_j)]^{j-1} [F_X(y_k) - F_X(y_j)]^{k-j-1} [1 - F_X(y_k)]^{n-k} f_X(y_j) f_X(y_k),$$

for $-\infty < y_j < y_k < \infty$.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Uniform(0,1). Find the distribution of $R = X_{(n)} - X_{(1)}$.

Simulation

- Approximation of the parameter or distribution of statistic

$\triangleright$ Example: Suppose that a particular electrical component is to be modeled with an exponential(λ) life time.

$$\begin{aligned} p_1 =& P[\text{component lasts at least } h \text{ hours}] \\ =& P[X \geq h; \lambda] = e^{-h/\lambda}. \end{aligned}$$

Assuming the components are independent. Consider the probability that out of c components, at least t will last h hours.

$$\begin{aligned} p_2 =& P[\text{at least } t \text{ components last } h \text{ hours}] \\ =& \sum_{k=t}^{c} \binom{c}{k} p_1^k (1 - p_1)^{c-k}. \end{aligned}$$

Simulation

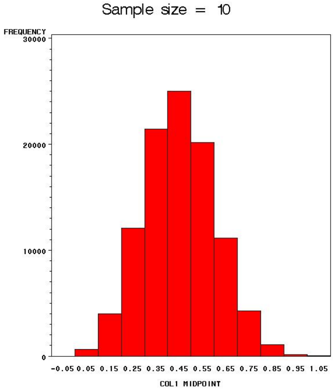
With $c = 20,\ t = 15,\ h = 150,\ \lambda = 300,\ p_1 = 0.60653$ and $p_2 = 0.1382194$.

If the distribution is complicate like Gamma distribution, there is no close form of the probability for p_1. $\rightarrow$ approximation using simulation.

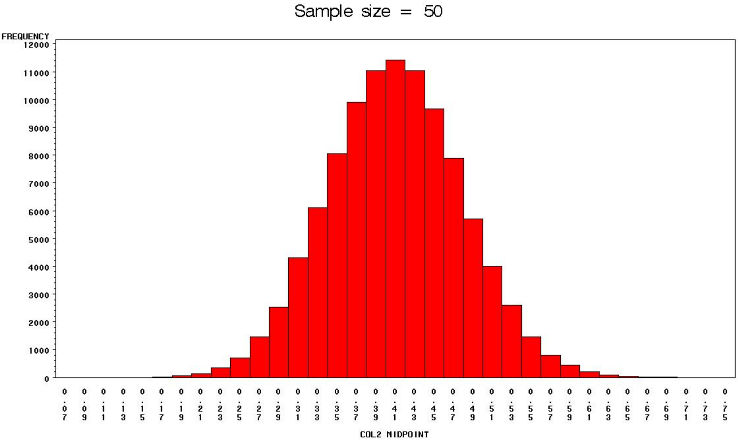
1. Generate $X_1, \cdots, X_{n=20}$ from Exponential$(\lambda = 300)$
2. Define $Y_j = 1$ if at least $t = 15$ X_j's are greater than or equal to $h = 150$, otherwise $Y_j = 0$.

Simulation

▷ Example: Distribution of $\bar{X}_n$ for $n = 10$ and $n = 50$, where $X_1, \cdots, X_n \overset{iid}{\sim}$ Bernoulli$(p = 0.4)$.



Simulation


Sample size = 50
FREQUENCY
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0.07 0.09 0.11 0.13 0.15 0.17 0.19 0.21 0.23 0.25 0.27 0.29 0.31 0.33 0.35 0.37 0.39 0.41 0.43 0.45 0.47 0.49 0.51 0.53 0.55 0.57 0.59 0.61 0.63 0.65 0.67 0.69 0.71 0.73 0.75
COL2 MIDPOINT