IE605: Engineering Statistics

Lecture 07

Manjesh K. Hanawal

Previous Lecture:

- Exponential Family of Distributions
- Population and Random Sampling
- Sample mean, variance and standard deviation
- Sampling from Normal distribution
- Student's t-distribution

This Lecture:

- F-distributions
- Convergence of RVs
- Consistency
- Order Statistics
- Generating Random Samples

F-distributions

We would be interested in variability of populations:

$$(X_1, X_2, \ldots, X_n) \text{ are iid and } \quad X_i \sim \mathcal{N}(\mu_X, \sigma_X^2) \quad \forall\, i$$
$$(Y_1, Y_2, \ldots, Y_m) \text{ are iid and } \quad Y_j \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \quad \forall j$$

We would estimate $\frac{S_X^2}{S_Y^2}$. What is its distribution?

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \frac{(n-1)S_X^2/(n-1)\sigma_X^2}{(m-1)S_Y^2/(m-1)\sigma_Y^2} = \frac{\chi_{n-1}^2/(n-1)}{\chi_{m-1}^2/(m-1)}$$

$\dfrac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2}$has F-distribution with (n-1) and (m-1) degree of freedom

F-dsitribution is named in the honor of Sir Ronald Fisher!

F-distributions contd...

PDF of F distribution with p and q degrees of freedom ($F_{p,q}$)

$$f_F(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^2 \frac{x^{p/2-1}}{[1+(p/q)x]^{(p+q)/2}} \quad x > 0$$

$$F_{p,q} = \frac{U/p}{V/q} \text{ where } U \sim \chi^2_p, V \sim \chi^2_q, \text{ and independent}$$

Derivation of pdf of $F_{p,q}$

- $X = \frac{U/p}{V/q} = \frac{q}{p}\frac{U}{V}$ and $Y = V$
- As U, V are independent $f(U, V) = f(U)f(V)$
- Find joint distribution of (X, Y) be applying transformations
- Find marginal distribution of X

Properties of F-distribution

- Claim 1: If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$
  $X = \frac{U/p}{V/p}$ where $U \sim \chi^2_p, V \sim \chi^2_q$ and are independent
  $1/X = \frac{V/q}{U/p}$, hence $1/X \sim F_{q,p}$

- Claim 2:if $X \sim t_p$, then $X^2 \sim F_{1,p}$
  $X = \frac{U}{\sqrt{V/p}}$, where $U \sim \mathcal{N}(0,1), V \sim \chi^2_p$ and are independent
  $X^2 = U^2/(V/p) = \chi^2_1/(V/p) = (\chi^2_1/1)/(\chi^2_p/p) \sim F_{1,p}$

- Claim 3: if $X \sim F_{p,q}$, then $\frac{(p/q)X}{1+(p/q)X} \sim beta(p/2, q/2)$
  (Exercise!)

Convergence of Sequence of RVs

What happens when the number of samples goes to infinity (theoretical artifact)

Convergence in Probability: A sequence of RVs $X_1, X_2, \ldots,$ converge in probability to a random variable X if, for an $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0 \text{ or } \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$$

- In the definition $X_1, X_2, \cdots$ need not be i.i.d or independent
- Compactly written as $X_n \xrightarrow{p} X$ in probability.

- Suppose $X_1, X_2, \ldots$ are i.i.d. with common mean μ and variance $\sigma^2 > \infty$. From LLN, we know

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$$

- For an $\epsilon > 0$

$$\begin{aligned} P(|\bar{X}_n - \mu| \geq \epsilon) &\leq \frac{\mathbb{E}\left(|\bar{X}_n - \mu|^2\right)}{\epsilon^2} \\ &= \frac{Var(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon} \to 0 \end{aligned}$$

- Sample mean converges to population mean!
- $\mathbb{E}(\bar{X}_n) = \mu$ (unbiased).

Consistency of Sample mean an Sample Variance

Consistency: A sample quantity is consistent if its sequence converges to a constant

- Sample mean is consistent: $\bar{X}_n \xrightarrow{p} \mu$ (by LLN)
- Is sample variance consistent?
  $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_n - \bar{X})^2$. We know $\mathbb{E}(S_n^2) = \sigma^2$

$$P(|S_n^2 - \sigma^2| \geq \epsilon) \leq \frac{\mathbb{E}\left(\left(S_n^2 - \sigma^2\right)^2\right)}{\epsilon^2} = \frac{Var(S_n^2)}{\epsilon^2}$$

  if $Var(S_n^2) \to 0$, then $S_n^2 \xrightarrow{p} \sigma^2$ (hence consistent)
- Is sample standard deviation consistent? (Exercise!)

Other Convergence types

Almost sure convergence: A sequence of RVs $X_1, X_2, \cdots$ convergence to X almost surely if $P\left(\lim_{n\to\infty} X_n = X\right) = 1$. Denoted as $X_n \xrightarrow{a.s} X$.

Convergence in distribution: A sequence of RVs $X_1, X_2, \cdots$ convergence to X in distribution if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all continuity points of F_X. Denoted as $X_n \xrightarrow{d} X$

$$X_n \xrightarrow{a.s} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X.$$

Order Statistics

Smallest, largest, middle observation of a random sample are useful

- Highest temperature in the last 50 years
- Lowest rainfall in the last 50 years
- median value of stock index in the last month

The order static of a random sample $X_1, X_2, \ldots, X_n$ are the sample value placed in the ascending order, denotes by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ where $X_{(1)} \leq X_{(2)} \leq \ldots, \leq X_{(n)}$

$$
\begin{aligned}
X_{(1)} &= \min_{1 \leq i \leq n} X_i \\
X_{(2)} &= \text{second smallest } X_i \\
&\vdots \\
X_{(n)} &= \max_{1 \leq i \leq} X_i
\end{aligned}
$$

Sample mean vs Sample Median

- Sample range: $X_{(n)} - X_{(1)}$
- Sample median:

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

- Example:
  Random sample: $24, 89, 59, 34, 55, 81, 45, 93, 85, 50$
  Order statistic: $24, 34, 45, 50, 55, 59, 81, 85, 89, 93$
  Sample range: $93 - 24 = 69$
  Sample mean: 61.5
  Median: 57
- Median gives better indication of "typical" values than means!

Sample Percentile

For any $p \in [0, 1]$, the $(100p)$the percentile is the observation such that approximately np of the observations are less than this observation and $n(1-p)$ of the observation are greater.

- For $p = 0.5$, 50the percentile gives median
- For any $b \in \mathbb{R}_+$, define

$$\{b\} = \begin{cases} \lceil b \rceil & if \quad \lceil b \rceil \leq b + 0.5 \\ \lfloor b \rfloor & if \quad b - 0.5 < \lfloor b \rfloor \end{cases}$$

- $\frac{1}{2} < np < n - \frac{1}{2} \implies \frac{1}{2n} < p < 1 - \frac{1}{2n}$

Lower and Upper Quartile

$$(100p)\text{th sample percentile is} = \begin{cases} X_{(\{np\})} & \text{if } p < 0.5 \\ X_{(n+1-\{n(1-p)\})} & \text{if } p > 0.5 \end{cases}$$

Example 1: $n = 50, p = .35, np = 17.5, \{np\} = 18$. 35th sample percentile is $X_{(18)}$
Example 2: $n = 50, p = .65, n(1-p) = 17.5, \{n(1-p)\} = 18$
$n + 1 - \{n(1-p)\} = 50 + 1 - 18 = 33$. 65th sample percentile is $X_{(33)}$

- For $p < 0.5$ and $p > 0.5$ sample percentiles exhibit symmetry
- if $(100p)$th sample percentile is $i$the smallest observation, then $100(1-p)$the sample percentile is the ith largest observation
- 25th sample percentile is called lower quartile
- 75the sample percentile is called upper quartile

Distribution of Order Statistics

<u>Discrete Case:</u>
Random sample $X_1, X_2, \ldots, X_n$ come from a discrete distributions with pmf $P_X(x_i) = p_i$, where $x_1 < x_2 < \cdots$ are the possible realizations in ascending order. For any x_i, what is $P(X_{(j)} \le x_i)$?

$$\begin{aligned} P_0 =& 0 \\ P(X \le x_1) = P_1 =& p_1 \\ P(X \le x_2) = P_2 =& p_1 + p_2 \\ &\vdots \\ P(X \le x_i) = P_i =& p_1 + p_2 + \ldots, + p_i \\ &\vdots \end{aligned}$$

Discrete Case contd..

- Fix some x_i. Define $Y_j = \mathbb{1}_{\{X_j \leq x_i\}}$ for all $j = 1, 2, \ldots, n$
- $P(Y_j = 1) = P_i$ for all $j = 1, 2, \ldots, n$
- $Y = \sum_{j=1}^{n} Y_j$, $Y \in \{0, 1, 2, \ldots, n\}$
- As X_is are i.i.d, Y_js are i.i.d. $Y_j \sim Ber(P_i)$.
- $Y \sim Bin(n, P_i)$. Y is sum of n $Ber(P_i)$ RVs
- $\{X_{(j)} \leq x_i\} = \{Y \geq j\}$. Hence $P(X_{(j)} \leq x_i) = P(Y \geq i)$
- $P(Y \geq j) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$.

$$P(X_{(j)} \leq x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

$$\begin{aligned} P(X_{(j)} = x_i) &= P(X_{(j)} \leq x_i) - P(X_{(j)} \leq x_{i-1}) \\ &= \sum_{k=j}^{n} \binom{n}{k} \left(P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k} \right) \end{aligned}$$

Continuous case

Random sample $X_1, X_2, \ldots, X_n$ come from a population with pdf $f_X(x)$, and CDF $F_X(x)$. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the order statistics. Then, pdf of $X_{(j)}$ is

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

Joint pdf of $X_{(i)}$ and $X_{(j)}$ for $1 \leq i < j \leq n$ is

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