MATH 5010 –Foundations of Statistical Theory and Probability

❖ Sampling, Order Statistics

Instructor: He Wang
Department of Mathematics
Northeastern University

1. Random Samples
2. Statistic of Random variables
3. Sum of random samples
4. Sampling from Normal Distribution
5. Order Statistics

Random Samples

Suppose random variables $X_1, \dots, X_n$ are mutually independent and the marginal pdf or pmf of each X_i is the same function $f(x)$.

Then, $X_1, \dots, X_n$ are called **independent and identically distributed (IID)** random variables with pdf or pmf $f(x)$.

$\{X_1, \dots, X_n\}$ is also called **random sample of size n from the population** $\boldsymbol{f(x)}$.

The joint pdf of $X_1, \dots, X_n$ is given by

$$f(x_1, \dots, x_n) = f(x_1) f(x_2) \cdot \cdots \cdot f(x_n) = \prod_{i=1}^{n} f(x_i).$$

Example.

Let $\{X_1, \ldots, X_n\}$ be a random sample from exponential population

$$f(x;\beta) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$$

As a real world question, $X_1, \ldots, X_n$ might correspond to the times until failure (in years) for n identical lightbulbs that are put on test and used until they fail

The joint pdf of the sample is

$$f(x_1, \ldots, x_n|\beta) = \prod_{i=1}^{n} f(x_i|\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-(x_1+\cdots+x_n)/\beta}.$$

Random sampling a sequence $X_1, \ldots, X_n$ can be obtained if

1. There is an **infinite** population.

2. **Sampling with Replacement** (also called **Bootstrap**) from finite population.

Sampling without replacement will not give us the independent Ramom Samples.

Suppose x, y are two distinct elements of population $\{x_1, \dots, x_N\}$

$$P(X_2 = y | X_1 = y) = 0$$

$$P(X_2 = y | X_1 = x) = \frac{1}{N-1}$$

X_1 and X_2 are not independent, but they are close to independent when N is large.

$$P(X_1 = x) = \frac{1}{N}$$

$$P(X_2 = x) = \sum_{i=1}^{N} P(X_2 = x | X_1 = x_i) P(X_1 = x_i).$$

$$= (N-1)\left(\frac{1}{N-1}\frac{1}{N}\right) = \frac{1}{N}.$$

They are identical distribution.

❖ Statistic of Random variables

Let $\{X_1, \dots, X_n\}$ be a random sample from a population.

Let $Y = T(X_1, \dots, X_n)$ be a transformation of the random sample, called a **statistic**.

The probability distribution of a statistic Y is called the **sampling distribution** of Y.

Notation: We will use lower case letters $x_1, \dots, x_n$ denote the observed data values.

Example: $\boldsymbol{T} = \mathbf{max}\{X_1, \dots, X_n\}$

Example: $\boldsymbol{T} = \mathbf{3}$ (Strange, ignore random sample, but it is a statistics)

Example: Sample Mean $$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

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

The sample standard deviation S is also a statistic.

Observed values of the above statistics are denoted by $\bar{x}, s^2$ *and* s.

Two formulas:

1. $$\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

2. $$(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Theorem: Let $\{X_1, \ldots, X_n\}$ be a random sample from a population with mean μ and variance σ^2.

1.) $E[\bar{X}] = \mu$

2.) $Var(\bar{X}) = \frac{\sigma^2}{n}$

3.) $E[S^2] = \sigma^2$

So, $\bar{X}$ is an unbiased estimator of μ.

S^2 is an unbiased estimator of σ^2.

❖ Sum of random samples

The Sample Mean $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ relates to the sum

$$Y = X_1 + \cdots + X_n$$

- The pdf relation $f_{\bar{X}}(x) = nf_Y(nx)$

- The MGF relation:

$$M_{\bar{X}}(t) = \mathrm{E}e^{t\bar{X}} = \mathrm{E}e^{t(X_1+\cdots+X_n)/n} = \mathrm{E}e^{(t/n)Y} = M_Y(t/n).$$

Theorem: Let $\{X_1, \ldots, X_n\}$ be a random sample from a population with MGF $M_X(t)$.

$$M_{\bar{X}}(t) = \left[M_X\left(\frac{t}{n}\right)\right]^n$$

Example: Suppose $X_i \sim Normal(\mu, \sigma^2)$ for $i = 1, \ldots, n$

Then

$$\bar{X} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$$

Example: Suppose $X_i \sim Gamma(\alpha, \beta)$ for $i = 1, \ldots, n$

Then

$$\bar{X} \sim Gamma\left(n\alpha, \frac{\beta}{n}\right)$$

If the above theorem can not solve the problem, we will consider the next method:

Theorem: If $X \ and \ Y$ are independent continuous random variables with PDFs $f_X(x) \ and \ f_Y(y)$. Then the PDF of $Z = X + Y$ is given by *convolution product:*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z - w) dw$$

Proof: Let $W = X$. The Jacobian of the transformation from $(X, Y) \ to \ (Z, W)$ is 1.

By the transformation theorem, the joint pdf of (Z, W) is

$$f_{Z,W}(z, w) = f_{X,Y}(w, z - w) = f_X(w) f_Y(z - w).$$

The marginal pdf of Z if given by integral over w:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z - w) dw$$

Example(Sum of Cauchy random variables)

Suppose $U\ and\ V$ are independent Cauchy random variables:

$$U \sim Cauchy(0, \sigma)\ and\ V \sim Cauchy(0, \tau)$$

$$f_U(u) = \frac{1}{\pi\sigma}\frac{1}{1+(u/\sigma)^2}, \qquad f_V(v) = \frac{1}{\pi\tau}\frac{1}{1+(v/\tau)^2},$$

So, the PDF for $Z = U + V$ is

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\pi\sigma}\frac{1}{1+(w/\sigma)^2}\frac{1}{\pi\tau}\frac{1}{1+((z-w)/\tau)^2}\,dw,$$

$$= \frac{1}{\pi(\sigma+\tau)}\frac{1}{1+(z/(\sigma+\tau))^2}$$

So $Z = U + V \sim Cauchy(0, \sigma + \tau)$

❖ Sampling from Normal Distribution

Theorem: Let $\{X_1, \ldots, X_n\}$ be a random sample from normal distribution with mean μ and variance σ^2. Then

- $\bar{X}$ and S^2 are independent.
- $\bar{X} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ chi squared distribution with $n-1$ degree of freedom.

Chi Square Random Variable

Chi-squared distribution χ_k^2 with degree freedom k is a special case of $Gamma$ distribution

$$\chi_k^2 \sim Gamma\left(\alpha = \frac{k}{2}, \theta = 2\right)$$

Theorem:

1.) If $Z \sim Normal(0,1)$, then, $Z^2 \sim \chi_1^2$

2.) If X_i $are\ independent$ $\chi_{p_i}^2$ then, $X_1 + \cdots + X_n \sim \chi_{p_1+\cdots+p_n}^2$

Student's t distribution

Let $\{X_1, \dots, X_n\}$ be a random sample from a population with mean μ and variance σ^2.

If we know σ^2, we can measure $\bar{X}$ by

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

However, if σ^2 is unknown, we may consider the distribution of

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

This topic was first addressed by W. S. Gosset (who published under the pseudonym of Student) in the early 1900s.

Notice that

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}}$$

- Numerator $U = (\bar{X}-\mu)/(\sigma/\sqrt{n}) \sim Normal\ (0,1)$
- Denominator $\sqrt{S^2/\sigma^2} \sim \sqrt{\chi^2_{n-1}/(n-1)} =: \sqrt{V/p}$ is independent of the numerator.

$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has **Student's t distribution with $\boldsymbol{p = n-1}$ degree of freedom** with pdf given by

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

Sketch of Proof:

First consider the joint pdf of $U \; and \; V$

$$f_{U,V}(u,v) = \frac{1}{(2\pi)^{1/2}} e^{-u^2/2} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}} v^{(p/2)-1} e^{-v/2}$$

Then consider the the transformation

$$t = \frac{u}{\sqrt{v/p}}$$

$$w = v$$

The pdf of T is calculated by the marginal pdf from $f_{T,W}(t,w)$

If $n = 2$, it becomes the Cauchy distribution.

Student's t has no MGF because it does not have moments of all orders. In fact, if there are p degrees of freedom, then there are only $p - 1$ moments.

$$E[T_p] = 0 \text{ for } p > 1$$

$$Var[T_p] = \frac{p}{p-1} \; for \; p > 2$$

Variance Ratio Distribution (F-distribution)

Let $\{X_1, \ldots, X_n\}$ be a random sample from a $Normal(\mu_X, \sigma_X^2)$

Let $\{Y_1, \ldots, Y_m\}$ be a random sample from a $Normal(\mu_Y, \sigma_Y^2)$

We are interested in the ratio:

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

We know:

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2(n-1)$$

The random variable

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

has **Snedecor's F distribution** with $(n-1)\ and\ (m-1)\ degree\ of\ freedom$.

The PDF of F is given by

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

for $0 < x < \infty$

Theorem:

1.) If $X \sim F_{p,q}$, then $\frac{1}{X} \sim F_{q,p}$

2.) If $X \sim t_q$, then $X^2 \sim F_{1,q}$

3.) If $X \sim F_{p,q}$, then $\frac{\left(\frac{p}{q}\right)X}{\left(1+\left(\frac{p}{q}\right)\right)} \sim Beta\left(\frac{p}{2}, \frac{q}{2}\right)$

$$
\begin{aligned}
\mathbf{E}\left(\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}\right) = \quad \mathbf{E}F_{n-1,m-1} \quad &= \mathbf{E}\left(\frac{\chi_{n-1}^2/(n-1)}{\chi_{m-1}^2/(m-1)}\right) \\
&= \mathbf{E}\left(\frac{\chi_{n-1}^2}{n-1}\right)\mathbf{E}\left(\frac{m-1}{\chi_{m-1}^2}\right) \\
&= \left(\frac{n-1}{n-1}\right)\left(\frac{m-1}{m-3}\right) \\
&= \frac{m-1}{m-3}.
\end{aligned}
$$

For large m, we have the estimatation:

$$
\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \approx \frac{m-1}{m-3} \approx 1
$$

❖ Order Statistics

Let $\{X_1, \ldots, X_n\}$ be a random sample.

The order statistics $Y_1 < Y_2 < \cdots < Y_n$ are the ordered version of these n random variables such that Y_j is the j-th smallest values among $\{X_1, \ldots, X_n\}$.

$$Y_1 = \min\{X_1, \ldots, X_n\}.$$

$$Y_n = \max\{X_1, \ldots, X_n\}.$$

The notation $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ is often used

Theorem: Let $\{X_1, \dots, X_n\}$ be a random sample from **discrete** distribution with pmf $f_X(x_i) = p_i$, where $x_1 < x_2 < \cdots$ are possible values of X.

Define

$$
\begin{aligned}
P_0 &= 0 \\
P_1 &= p_1 \\
P_2 &= p_1 + p_2 \\
&\vdots \\
P_i &= p_1 + p_2 + \cdots + p_i \\
&\vdots
\end{aligned}
$$

Then,

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

$$P\big(X_{(j)} = x_i\big) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}\right].$$

Theorem: Let $\{X_1, \dots, X_n\}$ be a random sample from a **continuous** population with cdf $F_X(x)$ and pdf $f_X(x)$.

- The pdf of $X_{(j)}$ is

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

- The joint pdf of $X_{(i)}$ and $X_{(j)}$ for $1 \le i < j \le n$ is

$$\begin{aligned} f_{X_{(i)},X_{(j)}}(u,v) &= \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v) [F_X(u)]^{i-1} \\ &\quad \times [F_X(v)-F_X(u)]^{j-1-i}[1-F_X(v)]^{n-j} \end{aligned}$$

for $-\infty < u < v < \infty$

- The joint pdf of $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ is

$$f_{X_{(1)},\dots,X_{(n)}}(x_1,\dots,x_n) = \begin{cases} n! f_X(x_1)\cdots f_X(x_n) & -\infty < x_1 < \cdots < x_n < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Example

Let $\{X_1, \ldots, X_n\}$ be a random sample from uniform distribution $U[0,1]$

Then, $p_X(x) = 1$ and $F_X(x) = x$ for $x \in [0,1]$

Thus the pdf

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

$$= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1}(1-x)^{n-j}$$

So, $X_{(j)} \sim Beta(j, n-j+1)$

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

Sample Range and Median

- The **Sample Range $\boldsymbol{R}$** is the distance between the smallest and largest observations.

$$R := X_{(n)} - X_{(1)}$$

It is a measure of the dispersion in the sample and should reflect the dispersion in the population.

- The **sample median $\boldsymbol{M}$** is a number such that approximately one-half of the observations are less than M and one-half are greater.

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

- The lower quartile (25th percentile) and upper quartile (75th percentile) are also commonly used.

Example

Let $\{X_1, \ldots, X_n\}$ be a random sample from uniform distribution $U[0, a]$

$$f_{X_{(1)},X_{(n)}}(x_1, x_n) = \frac{n(n-1)}{a^2}\left(\frac{x_n}{a} - \frac{x_1}{a}\right)^{n-2}$$

$$= \frac{n(n-1)}{a^n}(x_n - x_1)^{n-2}$$

Denote $V := \frac{\left(X_{(n)} + X_{(1)}\right)}{2}$

$$R := X_{(n)} - X_{(1)}$$

The Jacobian for this transformation is -1

Solve the joint pdf of (R, V)

$$f_{R,V}(r,v) = \frac{n(n-1)r^{n-2}}{a^n}, \quad 0 < r < a, \quad r/2 < v < a - r/2.$$

The marginal pdf of R is

$$\begin{aligned} f_R(r) &= \int_{r/2}^{a-r/2} \frac{n(n-1)r^{n-2}}{a^n}\, dv \\ &= \frac{n(n-1)r^{n-2}(a-r)}{a^n}, \quad 0 < r < a. \end{aligned}$$

If $a = 1$, we see that $R \sim Beta\ (n-1, 2)$.

For arbitrary a, it R/a has a beta distribution.

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition (**5.1, 5.2, 5.3, 5.4**)
- **Book 2. [W]: All of Statistics: Larry Wasserman**
- **Book 3. Introduction to Probability**. C.M. Grinstead and J.L. Snell. American Mathematical Society, 2012
- **Book 4. Introduction to Probability Models**, S. Ross, 12th edition (published by Academic Press).

Online books:

https://www.probabilitycourse.com/