Random Samples

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §5.1-§5.3)

Introduction

- Statistical inferences are concerned with two entities: *population* and *sample*.
- A sample drawn from a population is used to make *inferences* about the population.
- In this section, we will be concerned with properties of random samples.

Random Sample

- **Example** Suppose a new drug has been developed for the treatment of hypertension. A sample of 50 hypertensive patients from the UNC Hospital is selected and treated by the new treatment.
- The primary outcome is the reduction in DBP after the treatment, which gives 50 numbers $x_1, x_2, \cdots, x_{50}$.
- *A sample of size* $n = 50$.
- Statistician would say: x_i is the *observed value*, or *realized value*, of a random variable X_i.
- If $X_1, X_2, \cdots, X_n$ are mutually independent with the same marginal pdf or pmf $f(x)$ (i.i.d.); $X_1, X_2, \cdots, X_n$ is called a random sample from the population $f(x)$.

Sampling from a Finite Population

- Sampling from a finite populations *with replacement* allows a unit to appear more than once in the sample.
- Sampling from a finite population *without replacement* allows a unit to appear at most once in the sample.
- Assume there are 25 balls in the urn, with 3 blacks and 22 reds.
- $X_1, \cdots, X_5$ ($n = 5$) is a random sample if drawn from a finite population of $N = 25$ *with replacement*.
- $X_1, \cdots, X_5$ is NOT a random sample if drawn *without replacement* because $P(X_2 = 1|X_1 = 1) = \frac{2}{24} \neq P(X_2 = 1) = \frac{3}{25}$, which implies

$$P(X_1 = 1, X_2 = 1) \neq P(X_1 = 1)P(X_2 = 1).$$

- What happen if N is very large?.

Statistics

- Let $X_1, \cdots, X_n$ be a random sample with $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$.
- A statistic is denoted by $T(x_1, \cdots, x_n)$, which can be real-valued or vector-valued.
- **Example:** *Sample mean* $\bar{X}$ and *sample variance* S^2, which are defined by, respectively,

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

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

- The probability distribution of T is called the sampling distribution of T.

Computational Formula

$$
\begin{aligned}
(n-1)S^2 &= \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n.
\end{aligned}
$$

- The last expression is sometimes described as a “computational formula" for S^2.

Sums of $X_1, \cdots, X_n$

- Sums are attractive mathematically because their means and variances can be calculated using simple rules, like

$$E\bar{X} = n^{-1}E(X_1 + \cdots + X_n) = n^{-1}nEX_1 = \mu,$$
$$Var\bar{X} = n^{-2}Var(X_1 + \cdots + X_n) = n^{-2}nVarX_1 = n^{-1}\sigma^2.$$

- For S^2, $E[(n-1)S^2] = E(\sum_{i=1}^n X_i^2 - n\bar{X}^2) = \sum_{i=1}^n EX_i^2 - nE\bar{X}^2$.
- We have

$$EX_i^2 = VarX_i + (EX_i)^2 = \sigma^2 + \mu^2,$$

and

$$E\bar{X}^2 = Var\bar{X} + (E\bar{X})^2 = n^{-1}\sigma^2 + \mu^2$$

- We get $E[(n-1)S^2] = (n-1)\sigma^2$ and $ES^2 = \sigma^2$.

Unbiased Estimator

- If $ET(X_1, \cdots, X_n) = \theta$, we say that T is an unbiased estimator of θ.
- **Example** If $EX_1 = \mu$ and $VarX_1 = \sigma^2$, then $\bar{X}$ is an unbiased estimator of μ, and S^2 is an unbiased estimator of σ^2.
- If one defines

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

  is T an unbiased estimator of σ^2?
- What happen if $n \to \infty$?

Samples from Normal Distribution

- If X has mgf $M_X(t)$, then $M_{\bar{X}}(t) = \{M_X(t/n)\}^n$.
- Suppose that $X_1, X_2, \cdots, X_n$ is a random sample from $N(\mu, \sigma^2)$.
- Then,

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

- Thus, $\bar{X} \sim N(\mu, \sigma^2/n)$.
- In some cases, the mgf of $\bar{X}$ may not correspond to any distribution we know, or the mgf of X may not exist (e.g. Cauchy).

Samples from Normal Distribution (cont'd)

- Suppose that $X_1, X_2, \cdots, X_n$ is a random sample from $N(\mu, \sigma^2)$.
- We know that $\bar{X} \sim N(\mu, \sigma^2/n)$, and that $ES^2 = \sigma^2$.
- If $\mu = 0$ and $\sigma = 1$, the density of $X_1, X_2, \cdots, X_n$ is

$$\prod_{i=1}^{n} \phi(x_i) = \prod_{i=1}^{n} \frac{e^{-x_i^2/2}}{\sqrt{2\pi}} = (2\pi)^{-n/2} e^{-\sum_{i=1}^{n} x_i^2/2}.$$

- Consider a transformation from $X_1, \cdots, X_n$ to $Y_1, \cdots, Y_n$, where $Y_1 = \bar{X}$ and $Y_i = X_i - \bar{X}$, $2 \leq i \leq n$, with inverse transformations $X_1 = Y_1 - \sum_{i=2}^{n} Y_i$ and $X_i = Y_i + Y_1$, $2 \leq i \leq n$.
- The joint density of $Y_1, \cdots, Y_n$ is

$$f_{Y_1,\cdots,Y_n}(y_1, \cdots, y_n) = n\phi(y_1 - \sum_{i=2}^{n} y_i) \prod_{i=2}^{n} \phi(y_i + y_1),$$

Samples from Normal Distribution (cont'd)

which can be expressed as

$$\left\{ \frac{1}{\sqrt{2\pi(1/n)}} e^{-y_1^2/(2/n)} \right\} \left\{ \frac{n^{1/2}}{(2\pi)^{(n-1)/2}} e^{-c/2} \right\},$$

where $c = \sum_{i=2}^n y_i^2 + (\sum_{i=2}^n y_i)^2$.

- This implies: $Y_1 = \bar{X}$ is independent of $Y_2, \cdots, Y_n$.
- Since

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

one can claim S^2 is a function of $Y_2, \cdots, Y_n$.

Distribution of S^2

- This tells you $\bar{X} \perp S^2$.
- What is the distribution of S^2? Consider $n = 2$. In this case,

$$S^2 = \left(X_1 - \frac{X1 + X2}{2}\right)^2 + \left(X_2 - \frac{X1 + X2}{2}\right)^2 = \left(\frac{X_1}{\sqrt{2}} - \frac{X_2}{\sqrt{2}}\right)^2$$

- Since $X_1 \perp X_2$, $\frac{X_1}{\sqrt{2}} - \frac{X_2}{\sqrt{2}} \sim N(0,1)$ and $S^2 \sim \chi_1^2$.
- How about $n = k$?
- Let $\bar{X}_k$ and S_k^2 denote the sample mean and sample variance, respectively.
- A method of **induction** will be shown to prove that S_{k+1}^2 follows χ_k^2, assuming S_k^2 follows χ_{k-1}^2.

Distribution of S^2 (cont'd)

- One can show

$$\bar{X}_{k+1} = \frac{k\bar{X}_k + X_{k+1}}{k+1}$$

$$kS_{k+1}^2 = (k-1)S_k^2 + \frac{k}{k+1}(X_{k+1} - \bar{X}_k)^2,$$

- The second equation is proved in the following page.

Distribution of S^2 (cont'd)

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

Distribution of S^2 (cont'd)

- The distribution of $kS_{k+1}^2 = (k-1)S_k^2 + \frac{k}{k+1}(X_{k+1} - \bar{X}_k)^2$ is derived as follows:

(1) First, we have known that S_k^2, X_{k+1}, $\bar{X}_k$ are independent.

(2) Since $X_{k+1} - \bar{X}_k \sim N(0, 1 + 1/k)$, $\frac{k}{k+1}(X_{k+1} - \bar{X}_k)^2 \sim \chi_1^2$.

(3) Assuming the statement $(k-1)S_k^2 \sim \chi_{k-1}^2$ is true, one shall show that kS_{k+1}^2 follows χ_k^2.

(4) By induction, we need to show when $k = 2$, S_2^2 follows χ_1^2, which was proved in the previous page.

Extension to $X_i \sim N(\mu, \sigma^2)$

- What if $X_1, \cdots, X_n$ from $N(\mu, \sigma^2)$?
- Define $Z_i = (X_i - \mu)/\sigma$ and let S_X^2 and S_Z^2 denote sample variance of X and Z, respectively.
- We know that $Z_i \sim N(0, 1)$, $i = 1, \cdots, n$.
- Also, $\bar{Z} = (\bar{X} - \mu)/\sigma$, and $S_Z^2 = S_X^2/\sigma^2$.
- Therefore, $((\bar{X} - \mu)/\sigma, S_X^2/\sigma^2)$ has the same distribution as $(\bar{Z}, S_Z^2)$.
- One has

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

and $\bar{X}$ and S_X^2 are independent.

More Transformations

- χ^2_{n-1} distribution has mean $n-1$ and variance $2(n-1)$.
- We have $E(S_X^2) = \sigma^2$ and $Var(S_X^2) = 2\sigma^4/(n-1)$ since

$$Var\{(n-1)S_X^2/\sigma^2\} = 2(n-1).$$

- A test statistic $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ can't be computed if σ unknown.
- If σ unknown, look for the distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$.
- What is the distribution of

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

Student's t Distribution

- If $U \sim N(0,1)$, $V \sim \chi^2_p$ and U and V are independent, then the distribution of $T = U/\sqrt{V/p}$ known as *Student's t distribution with p degrees of freedom*, abbreviated as t_p, with density

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

- Since U and V are independent, the joint density of (U, V) is

$$f_{U,V}(u,v) = f_U(u)f_V(v) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}\frac{1}{\Gamma(p/2)2^{p/2}}v^{p/2-1}e^{-v/2}.$$

- The transformation from (u, v) to $t = \frac{u}{\sqrt{v/p}}$ and $w = v$.

Student's t Distribution (cont'd)

- The inverse is $u = t\sqrt{w/p}$ and $v = w$ with Jacobian $\sqrt{w/p}$.
- The joint density of (T, W) is then

$$f_{T,W} = f_{U,V}(t\sqrt{w/p}, w)\sqrt{w/p}.$$

- The marginal density of T is obtained by integrating out w.
- The t_p density is symmetric about 0.
- It does not have an mgf. In fact, only the first $p-1$ moments exist.
- The mean is 0 if $p > 1$, and the variance is $p/(p-2)$ if $p > 2$.
- The case $p = 1$ is Cauchy distribution ($\Gamma(\frac{1}{2}) = \sqrt{\pi}$).

Student's t Distribution (cont'd)

- If $X_1, \cdots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ density, and we define $U = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ and $V = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.
- $U \sim N(0,1)$, $V \sim \chi^2_{n-1}$, and $U \perp V$. That shows

$$T = \frac{U}{\sqrt{V/(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

- 95% CI: $\bar{x} \pm t_{n-1,1-\alpha/2} s/\sqrt{n}$.
- How about 95% CI for σ^2? We will talk about pivotal quantity in the future.

F Distribution

- If $U \sim \chi^2_p$, $V \sim \chi^2_q$, and $U \perp V$. The distribution of $X = (U/p)/(V/q)$, which is known as *Snedecor's F distribution with p and q degrees of freedom*, abbreviated as $F_{p,q}$.
- This distribution arises in the study of ratios of sample variances. Such ratios arise in the analysis of variance (ANOVA) and in regression analysis.
- What is the distribution of $1/X$?

Other Properties of Normal Variates

- If X has a normal distribution and Y has a normal distribution, then X and Y are independent if and only if $Cov(X, Y) = 0$.
- $\bar{X}$ is normal, $X_i - \bar{X}$ is normal, and

$$Cov(\bar{X}, X_i - \bar{X}) = Cov(\bar{X}, X_i) - Cov(\bar{X}, \bar{X}) = \sigma^2/n - \sigma^2/n = 0.$$

- We can conclude $\bar{X}$ and $X_i - \bar{X}$ are independent, for $1 \leq i \leq n$.
- That can help show $\bar{X}$ is independent of $X_i - \bar{X}$ (check the notes).
- The "zero covariance implies independence" property generally does not apply to other distributions.
- For example, if $X \sim N(0, 1)$ and $Y = X^2 \sim \chi_1^2$, then clearly X and Y are not independent. However,

$$Cov(X, Y) = Cov(X, X^2) = EX^3 - (EX)(EX^2) = 0 - (0)(1) = 0.$$