STAT 345/445 Lecture 17

Properties of a random sample

Random samples – Section 5.1
Sum of a random sample – Section 5.2

Definition of a random sample

Random sample

Random variables $X_1, \dots, X_n$ are called a

random sample of size n from the population $f(x)$

if $X_1, \dots, X_n$ are

- mutually independent, and
- marginal pmf/pdf of each X_i is $f(x)$

- Alternative name for a random sample:
  independent and identically distributed (iid) random variables with pdf or pmf $f(x)$

$$\text{i.i.d. } f(x) \quad = \quad \text{random sample from } f(x)$$

Random samples and statistical inference

- We view data as observations of random variables
- Usually have more than one observation

$$X_1, X_2, \ldots, X_n$$

  - Can often assume that $X_1, X_2, \ldots, X_n$ is a *random sample*
- We model the data by specifying a joint distribution

$$f(x_1, x_2, \ldots, x_n \mid \theta)$$

unknown

  with the goal of
  - learning about (estimating) θ and/or
  - predicting observations of $X_{n+1}, X_{n+2}, \ldots$
- Use some *summary* of the data to do this
  - Need to find the distribution of that summary $\rightarrow$ *sampling distribution*

Example: Say we collected data on temps at CLG airport. Say data points are

$x_1 = 11^\circ C$, $x_2 = 12^\circ C$, $x_3 = 12.2^\circ C$, ... $x_n = 17^\circ C$.

We will assume that $x_1, x_2, \dots x_n$ are realizations of random variables.

$X_1, X_2, \dots, X_n$

if $X_1, X_2, \dots, X_n$ are a random sample, from $f(x)$

then $f(x)$ can be seen as the (population) distribution of temps at all times at CLB.

Calculate e.g. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

dist of $\bar{x}$?

sampling distribution

About random samples

- Recall: Random variables $X_1, \ldots, X_n$ are (mutually) *independent* iif
$$f(x_1, \ldots, x_n) = f_1(x_1) \times \cdots \times f_n(x_n)$$
where $f_i(x_i)$ is the marginal pdf/pmf of X_i
- So, what is the joint pdf/pmf of a *random sample* from $f(x)$?
  - $f_i(x_i) = f(x_i)$ for all i, so

$$\begin{aligned} f(x_1, \ldots, x_n) &= f_1(x_1) \times \cdots \times f_n(x_n) \\ &= f(x_1) \times \cdots \times f(x_n) \\ &= \prod_{i=1}^{n} f(x_i) \end{aligned}$$

all marginals are the same.

Mutually independent

- Recall again: Random variables $X_1, \ldots, X_n$ are (mutually) *independent* iif

$$f(x_1, \ldots, x_n) = f_1(x_1) \times \cdots \times f_n(x_n)$$

  where $f_i(x_i)$ is the marginal pdf/pmf of X_i

- $\Rightarrow$ any subcollection of $X_1, \ldots, X_n$ are also (mutually) independent.

- For example:

$$\begin{aligned} f(x_1, x_2) &= \int \cdots \int f(x_1, \ldots, x_n)\, dx_3 \cdots dx_4 \\ &= \int \cdots \int f_1(x_1) \times \cdots \times f_n(x_n)\, dx_3 \cdots dx_n \\ &= f_1(x_1) f_2(x_2) \int f_3(x_3) dx_3 \times \cdots \times \int f_n(x_n) dx_n \\ &= f_1(x_1) f_2(x_2) \end{aligned}$$

More about random samples

- Not all *collections* of random variables are random *samples*
  - Need both independence and same (marginal) distributions
- If population is finite and we sample without replacement we don't get a random sample.

Example:

- Draw cards from a standard deck or 52 cards.
- Let $X_i =$ the card we get in draw i, $i = 1, \ldots, 10$.
- All X_i have the same (marginal) distribution, but they are not independent since e.g.:

$$P(X_1 = 3\spadesuit) = \frac{1}{52} \qquad \text{but} \qquad P(X_1 = 3\spadesuit \mid X_4 = 2\diamondsuit) = \frac{1}{51}$$

A simple random sample

- Sampling without replacement from a finite population is a very common
- A **simple random sample** of size n, $X_1, \ldots, X_n$ from a finite population of size N comes from a selection procedure were:
  - Any subset of n elements have the same probability of being selected.
- simple random sample $\neq$ random sample
- If N is huge we have simple random sample $\approx$ random sample

$\hookrightarrow \frac{1}{N} \approx \frac{1}{N-1}$ "almost independence"

More definitions

You just have to be able to evaluate $T(x_1, x_2, \ldots x_n)$ for any possible value of $x_1, \ldots, x_n$.

A statistic

Let

- $X_1, \ldots, X_n$ be a random sample of size n
- $T(x_1, \ldots, x_n)$ be a real-valued (or vector-valued) function with domain that includes the sample space of $(X_1, \ldots, X_n)$

then

- The random variable (or random vector) $Y = T(X_1, \ldots, X_n)$ is called a **statistic**.
- The probability distribution of Y is called the **sampling distribution of Y**

- In short: A statistic is a function of a random sample.
- Note: Cannot be a function of a parameter.

In general, a statistic is a function of a collection of random variabes (does not have to be a random sample).

Commonly seen statistics

- **sample mean:**

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **sample variance:**

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

- **sample standard deviation:**

$$S = \sqrt{S^2}$$

The random variables $\overline{X}$, S^2 and S all have a (sampling) distribution

Sampling distributions

- A lot of Statistical inference is based on sampling distributions
- Hence our focus on distributions of functions of random variables!
  - A bit easier if we have a random sample
- Example: If the mgf for the population exists:

Let $X_1, \ldots, X_n$ be a *random sample* of size n from a population with mgf $M_X(t)$. The the mgf of the sample mean is

$$M_{\overline{X}}(t) = (M_X(t/n))^n$$

- Only useful if we recognize the mgf on the right side

$\overline{X} = \frac{1}{n} X_1 + \ldots + \frac{1}{n} X_n$ i.e. $b = 0$ and $a_i = \frac{1}{n} \; \forall i$

$\boxed{Y = b + a_1 X_1 + \ldots + a_n X_n}$

Examples of sampling distributions

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$. What is the distribution of $\overline{X}$? ... ①

  $\bar{X} \sim N(\mu, \sigma^2)$

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathrm{Gamma}(\alpha, \beta)$. What is the distribution of $\overline{X}$? ... ②

① $X_1, X_2, \cdots X_n$, iid $N(\mu, \sigma^2)$

From last example: $\bar{x}$ is normal with mean

$$\sum_{i=1}^{n} \frac{1}{n}\mu = \frac{1}{n} n \mu = \mu$$

and variance $\sum_{i=1}^{n} \frac{1}{n^2}\sigma^2 = n\frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$

② $X_1, X_2, \dots X_n \sim \text{Gamma}(\alpha, \beta)$

$$M_X(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}$$

$$M_{\bar{X}}(t) = \left(M_X(t/n)\right)^n = \left(\left(\frac{1}{1-\beta t/n}\right)^{\alpha}\right)^n$$

$$= \left(\frac{1}{1-\frac{\beta}{n}t}\right)^{\alpha n} = \text{mgf of Gamma}\left(\alpha n, \frac{\beta}{n}\right)$$

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

Sampling distributions: convolution formula

E.g. if mgfs are not available,

Convolution formula

Let X and Y be independent random variables with pdfs $f_X(x)$ and $f_Y(y)$. Then the pdf of $Z = X + Y$ is

X, Y indep, Z=X+Y distri; distribu..

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

Set W=X

proof...

inverse function

$x = w = h_1(z, w)$,

$y = z - x = z - w = h_2(z, w)$

$(X, Y) \to (Z, W)$

$$\frac{\partial x}{\partial z} = 0, \quad \frac{\partial x}{\partial w} = 1 \qquad \frac{\partial y}{\partial z} = 1, \quad \frac{\partial y}{\partial w} = 1$$

$$\Rightarrow J = \left| \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} \right| = 0 - 1 = -1$$

Joint Pdf of X, Y:

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

Therefore: $f_{ZW}(z, w) = f_X(w) f_Y(z-w) \cdot |-1|$

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

$n > 2$, Just iterate:

$$Z_1 = X_1 + X_2$$

$$Z_2 = X_1 + X_2 + X_3 = Z_1 + X_3$$

$$\vdots$$

etc.

Moments of sampling distributions

Lemma

Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $g(x)$ be a function such that $\mathrm{E}\,(g(X_1))$ and $\mathrm{Var}\,(g(X_1))$ exist. Then

$$\mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathrm{E}\,(g(X_1))$$

$$\text{and } \mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathrm{Var}\,(g(X_1))$$

finding the first 2 moments of distribution

of $\sum X_i$ or $\frac{1}{n}\sum X_i$

$$E\left(\sum_{i=1}^{n} g(x_i)\right) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(g(x_1) + g(x_2) + \cdots + g(x_n)\right) \cdot f(x_1, x_2, \ldots x_n)\, dx_1, \cdots dx_n$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1) \cdot f(x_1, x_2, \ldots x_n)\, dx_1, \ldots x_n + \cdots + \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_n) f(x_1, x_2, \ldots x_n)\, dx_1, \cdots dx_n$$

$$\int_{-\infty}^{\infty} g(x_n) \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-1 \text{ terms}} f(x_1, \ldots x_n)\, dx_1, \ldots dx_n$$

$$= f_n(x_n)$$

$$= \int_{-\infty}^{\infty} g(x_1) f_1(x_1)\, dx_1 + \cdots + \int_{-\infty}^{\infty} g(x_n) f_n(x_n)\, dx_n$$

$$= E(g(x_1)) + \cdots + E(g(x_n))$$

$$= \mu + \cdots + \mu = n\mu$$

$\therefore$ (1)

For simplicity, show: $Var(\sum_{i=1}^{n} X_i) = n\sigma^2$,

$\sigma^2 = V(X_1)$,

$$Var(\sum_{i=1}^{n} X_i) = E\left(\left[\sum_{i=1}^{n} X_i - \underbrace{E(\sum_{i=1}^{n} X_i)}_{= n\mu \text{ from } (1)}\right]^2\right)$$

$$= E((X_1 - \mu + X_2 - \mu \cdots X_n - \mu)^2)$$

$$= E((X_1-\mu)^2 + \cdots + (X_n-\mu)^2 + 2(X_1-\mu)(X_2-\mu) + \cdots + 2(X_{n-1}-\mu)(X_n-\mu))$$

$$= E((X_1-\mu)^2) + \cdots + E((X_n-\mu)^2) + \underbrace{E(2(X_1-\mu)(X_2-\mu))}_{=2Cov(X_1,X_2)} + \cdots + 2E((X_{n-1}-\mu)(X_n-\mu))$$

$$= \underbrace{\sigma^2 + \cdots + \sigma^2}_{n\sigma^2} + \underbrace{2\sum_{i<j} \mathrm{Cov}(X_i, X_j)}_{=0 \text{ if } X_1, \ldots X_n \text{ are indep}}$$

$$= n\sigma^2$$

if indep

Moments of sampling distributions

Theorem

Let $X_1, \ldots, X_n$ be a random sample of size n from a population with mean μ and variance $\sigma^2 < \infty$. Then

1. $\mathrm{E}\left(\overline{X}\right) = \mu$ $\quad E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \cdot n\mu = \mu$
2. $\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$ $\quad Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$
3. $\mathrm{E}\left(S^2\right) = \sigma^2$ $\quad \ldots$ ③

Useful fact: For any numbers $x_1, \ldots, x_n$ we have

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

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

③: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$$E(S^2) = \frac{1}{n-1} E\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} E(x_i^2) - nE(\bar{x}^2)\right) \cdots (*)$$

Recall: $Var(X) = E(X^2) - E(X)^2 \Rightarrow E(X^2) = \sigma^2 + \mu^2$

$E(\bar{x}^2) = Var(\bar{x}) + E(\bar{x})^2$

$(*)$: $= \frac{1}{n-1}\left(\sum_{i=1}^{n}(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right)$

$$= \frac{1}{n-1}\left(n\Delta^2 + n\mu^2 - \Delta^2 - n\mu^2\right) = \frac{1}{n-1}\left((n-1)\Delta^2\right) = \Delta^2$$