3Y03-3J04 PROBABILITY & STATISTICS FOR (CIVIL) ENGINEERING

C01 - Lecture 17

Last time — LINEAR FUNCTIONS OF RANDOM VARIABLES

If $X_1, \ldots, X_k$ are random variables, then the new random variable $Y = c_1 X_1 + \ldots + c_k X_k$, for c_i real # constants, has **mean** $E(Y) = c_1 E(X_1) + \ldots + c_k E(X_k)$

and **variance** $V(Y) = c_1^2 V(X_1) + \ldots + c_k^2 V(X_k) + 2 \underbrace{\sum_{i<j} c_i c_j \sigma_{X_i X_j}}$

Covariance of X_i, X_j

This term $= 0$ if X_is are independent (as $\sigma_{X_i X_j} = 0$, $i \neq j$)

Special Cases: Sum of random variables with identical means & variances

$$X = X_1 + X_2 + \ldots + X_k$$

has $E(X) = E(X_1) + \ldots + E(X_k) = k \cdot \mu$ — common mean $= \mu$

& if $X_1, \ldots, X_k$ are independent, then

$$V(X) = V(X_1) + \ldots + V(X_k) = k \cdot \sigma^2$$

common variance $= \sigma^2$

Average of random variables with identical means μ and variances σ^2

$$\overline{X} = \frac{1}{k}(X_1 + \ldots + X_k)$$

has $E(\overline{X}) = \frac{1}{k} \cdot k\mu = \mu$

& $V(\bar{X}) = \frac{1}{k^2} \cdot k \cdot \sigma^2 = \frac{\sigma^2}{k}$ ← if X_i independent.

Useful Fact: If $X_1, \dots, X_k$ are independent, each normally distributed with $X_i \sim N(\mu_i, \sigma_i^2)$

Then $Y = c_1 X_1 + \dots + c_k X_k$, c_i real #s,

is also normally distributed,

with mean $= E(Y) = c_1\mu_1 + \dots + c_k\mu_k$

variance $= V(Y) = c_1^2\sigma_1^2 + \dots + c_k^2\sigma_k^2$.

In particular, $X = X_1 + \dots + X_k$, where $X_i \sim N(\mu, \sigma^2)$
is normally distributed $\sim N(k\mu, k\sigma^2)$

and $\bar{X} = \frac{1}{k}(X_1 + \dots + X_k)$, where $X_i \sim N(\mu, \sigma^2)$,
is normally distributed $\sim N(\mu, \sigma^2/k)$.

Example: Drinks bottles filled to a mean volume of 591 ml with standard deviation 5ml

Suppose the volume of each individual bottle is an independent normal random variable.

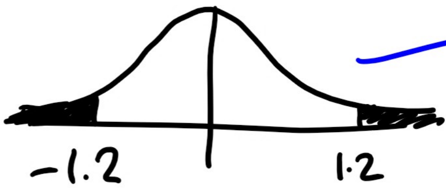
(a) What is the prob. that one bottle has < 585ml?

(b) Now measure 10 bottles. What is the prob. that average volume is < 585ml?

<u>Solution</u> (a) X_1 = volume of bottle $\sim N(591, 25)$

$$P(X_1 < 585) = P\left(Z < \frac{585-591}{5}\right) = P(Z < -1.2)$$

tails are equal in area



$$= P(Z > 1.2) = 1 - P(Z < 1.2)$$

For any r.v. X and $x \in \mathbb{R}$, $P(X > x) = 1 - P(X < x)$

$$= 1 - 0.88493 = \underline{\underline{0.11507}}.$$

(b) X_i = volume of bottle i $(i = 1, \dots, 10)$
$\sim N(591, 25)$

We want $P(\bar{X} < 585)$ where $\bar{X} = \frac{1}{10}(X_1 + \dots + X_{10})$
$\sim N(591, \underset{1.58^2}{\underset{\shortparallel}{2.5}})$

$$P(\bar{X} < 585) = P\left(Z < \frac{585-591}{1.58}\right)$$

$$= P(Z < -3.79)$$

$$= 1 - P(Z < 3.79)$$

area under graph in this region

$$= 1 - 0.999925$$

$$= 0.000075.$$

<u>DATA</u> $\rightarrow$ real #s we observe/measure

e.g. measurements of volumes of bottles

$\rightarrow$ take sample $\{x_1, \dots, x_n\} \Leftarrow$ DATA (measurements) from wider population

(e.g. all bottles' volumes from a certain factory)

→ data are realization of some independent random variables with same distribution. (as one another)

Interested typically in some feature of setup that summarises it e.g. average volume.

Can use data to estimate this feature.

Measures of Location

Sample Mean : $\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$

↓

centre of mass (centre of data set)

- estimates μ, mean of X_i (or of $\overline{X}$)

Sample Median : arrange data from highest to lowest

$$m = \begin{cases} \text{middle value} & \text{if } n \text{ odd} \\ \text{average of 2 middle values} & \text{if } n \text{ even} \end{cases}$$

- doesn't convey as much info. as mean but less affected by extreme observations so better than the mean when random variables skewed (non-symmetric distribution)

Sample Mode : most frequently occurring data value (can have multiple modes)

Measures of Variation

Sample Variance $s^2 = \frac{1}{n-1}\left((x_1-\bar{x})^2 + \dots + (x_n-\bar{x})^2\right)$

NOT $\frac{1}{n}$ (reason has to do with "bias")

We will return to this later, but remember, we want the best estimate of the underlying σ^2, and that turns out to come from $\frac{1}{n-1}$ not $\frac{1}{n}$.

Sample Standard Deviation $s = +\sqrt{s^2}$.

Sample Range $r = \max(x_i) - \min(x_i)$.

Tedious to compute s^2 so we have shortcut:

$$s^2 = \frac{1}{n-1}\left((x_1^2 + \dots + x_n^2) - n\bar{x}^2\right).$$