SAMPLING DISTRIBUTIONS AND THE CENTRAL LIMIT THEOREM

Normal and chi-squared distributions

Moving towards statistical inference

- Statistical inference refers to making conclusions about a *population* based on *sample data*.
- Definition 1: Random variables $Y_1, Y_2, \dots, Y_n$ constitute a RANDOM SAMPLE if they are independent and identically distributed (iid).
- A very common statistical inference problem involves estimating a *parameter* with a (sample) *statistic*.
- Definition 2: A STATISTIC is a function of observed random variables $Y_1, Y_2, \dots, Y_n$.

Sampling distributions

- The value of a statistic will depend on the specific sample observed.
- In order to perform inference, we need to know the distribution of a statistic.

Random variable

- Definition 3: The SAMPLING DISTRIBUTION of a statistic is the distribution of possible values of the statistic across many repeated samples.
- The sampling distribution naturally depends on the distribution of the sample data $Y_1, Y_2, \ldots, Y_n$.

Sampling distributions for normally distributed data

- We will start by considering the situation in which $Y_1, Y_2, \dots, Y_n \sim iid\ N(\mu, \sigma^2)$.
- Theorem 1: If $Y_1, Y_2, \dots, Y_n \sim iid\ N(\mu, \sigma^2)$, then $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$.

$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$

Proof Before (method of mgfs) showed: MGFs

$Y_1 + Y_2 + \dots + Y_n = \sum_{i=1}^{n} Y_i \sim N(n\mu, n\sigma^2)$

$\bar{Y} = \frac{1}{n} \sum Y_i$

$Y \sim N(\mu, \sigma^2)$ What is the dist. of $\frac{Y}{a}$?

$U = \frac{Y}{a}$ dist. of U?

mgf of U $\quad E[e^{tU}] = E[e^{t\frac{Y}{a}}] = E[e^{\frac{t}{a}Y}]$

$Y \sim N(\mu, \sigma^2)$

$= e^{\mu \frac{t}{a}} e^{\frac{1}{2}\sigma^2 \frac{t^2}{a^2}} = e^{\frac{\mu}{a}t} e^{\frac{1}{2}\frac{\sigma^2}{a^2}t^2}$

$$\boxed{\begin{array}{c}\text{mgf of } N(\mu, \sigma^2) \\ e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2}\end{array}}$$

$$\text{"} \; E[e^{tY}]$$

mgf of $N\left(\frac{\mu}{a}, \frac{\sigma^2}{a^2}\right)$

$$Y \sim N(\mu, \sigma^2) \Rightarrow \frac{Y}{a} \sim N\left(\frac{\mu}{a}, \frac{\sigma^2}{a^2}\right)$$

$$\sum_{i=1}^{n} Y_i \sim N(n\mu, n\sigma^2) \Rightarrow \bar{Y} = \frac{\sum Y_i}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$\square$

Normally distributed data: Example

- Example 1: Australian men aged 50-60 currently using antihypertensive drugs have mean diastolic blood pressure 94.9 mm Hg and standard deviation 11.5 Hg. Assuming that these measures are normally distributed, what is the probability that the average DBP of a sample of 8 men from this population will be at least 100 mm Hg?

μ

σ

$n = 8$

$P(\bar{Y} > 100)$

Normally distributed data: Another example

- Example 2: In the previous example how large must the sample be for the mean to be within 2 mm Hg from the mean with 95% probability?

sample

$\bar{Y}$

μ

$$\frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

- Example 3: If $Z \sim N(0,1)$ find the distribution of $\underbrace{Z^2}_{Y}$.

Method of mgf

mgf of $Y = Z^2$: $E[e^{tY}] = E[e^{tZ^2}] = \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}(1-2t)} dz$$

Restrict to $t < \frac{1}{2}$
so $1 - 2t > 0$

$$K = (1-2t)^{-\frac{1}{2}}$$

$$= K \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}K} e^{-\frac{z^2}{2K^2}} dz}_{=1}$$

pdf of $N(0, K^2)$

$$= K = (1-2t)^{-\frac{1}{2}}$$

mgf of Z^2

check book

mgf of Gamma $(\alpha = \frac{1}{2}, \beta = 2)$

$Z^2 \sim \text{Gamma}(\frac{1}{2}, 2)$

$$\boxed{Z^2 \sim \chi^2_1}$$

The χ^2 distribution

1. - We saw before that the χ^2 distribution with ν degrees of freedom is simply the $gamma\left(\frac{\nu}{2}, 2\right)$ distribution.
2. - In Example 3 we saw that if $Z \sim N(0,1)$ then $Z^2 \sim \chi_1^2$.
- Theorem 2: If $Y_1, Y_2, \dots, Y_n$ are iid $N(\mu, \sigma^2)$ random variables, then

$$\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 \sim \chi_n^2$$

0. $\frac{Y_i - \mu}{\sigma} \sim N(0,1)$

Let $Z_i = \frac{Y_i - \mu}{\sigma}$

$$\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2$$

By ② $Z_i^2 \sim \chi^2_1$ by ① $\sim \text{gamma}(\frac{1}{2}, 2)$

we did this for exp but same method show ~~β~~ (mgf):

③ $W_1, W_2, \ldots, W_n \sim \text{Gamma}(\alpha_i, \beta)$ all indep.

then $\sum_{i=1}^{n} W_i \sim \text{Gamma}\left(\sum_{i=1}^{n} \alpha_i, \beta\right)$

(check)

$$\sum_{i=1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2 \overset{③}{\sim} \text{Gamma}\left(\sum_{i=1}^{n} \frac{1}{2}, 2\right)$$

$$\sim \text{Gamma}\left(\frac{n}{2}, 2\right)$$

$$\sim \chi^2_n$$

①

$\square$

The distribution of the sample variance

iid

- Definition 4: Given data $Y_1, Y_2, \dots, Y_n$, the SAMPLE VARIANCE is defined to be

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the sample mean.

- Theorem 3: If $Y_1, Y_2, \dots, Y_n$ are iid $N(\mu, \sigma^2)$ random variables, then

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Furthermore, $\bar{Y}$ and s^2 are independent.

Look at $n=2$ case Y_1, Y_2 iid $N(\mu, \sigma^2)$

$$s^2 = \frac{1}{2-1}\left[(Y_i - \bar{Y})^2 + (Y_2 - \bar{Y})^2\right]$$

$$\bar{Y} = \frac{1}{2}(Y_1 + Y_2) = \underline{\frac{1}{2}Y_1 + \frac{1}{2}Y_2}$$

$$s^2 = \left(Y_1 - \frac{1}{2}Y_1 - \frac{1}{2}Y_2\right)^2 + \left(Y_2 - \frac{1}{2}Y_1 - \frac{1}{2}Y_2\right)^2$$

$$= \left(\frac{1}{2}Y_1 - \frac{1}{2}Y_2\right)^2 + \left(\frac{1}{2}Y_2 - \frac{1}{2}Y_1\right)^2$$

$$= \frac{1}{4}(Y_1 - Y_2)^2 + \frac{1}{4}(Y_1 - Y_2)^2 = \frac{1}{2}(Y_1 - Y_2)^2$$

What is the dist of $Y_1 - Y_2$?

$a_1 Y_1 + a_2 Y_2$

$a_1 = 1$

$a_2 = 1$

$$Y_1 - Y_2 \sim N(\mu - \mu, \sigma^2 + \sigma^2)$$
$$\sim N(0, 2\sigma^2)$$

$$\left(\frac{Y_1 - Y_2}{\sqrt{2}\,\sigma}\right)^2 = \frac{(Y_1 - Y_2)^2}{2\sigma^2} \sim \chi^2_1 \qquad \frac{(2-1)s^2}{\sigma^2} \sim \chi^2_1$$

Z