Lecture 24

Last time we discussed sampling distributions. Today we will continue that discussion a little more. Then we will move onto discussing linear combinations of random variables. Next time we will discuss the CLT.

How do we estimate the sampling distribution if we do not know the population distribution?

In the real world we do not know the ~~~~ true population distribution. In practice we either assume it to be some distribution (based on past knowledge or another modelling assumption) or we could use a computer intensive technique to estimate it. In particular this method is called the bootstrap.

Bootstrap

The idea with this method is to use the sample data to approximate the population by resampling from the sample data with replacement.

In particular, suppose that you have a sample $x_1, x_2, \ldots\ldots, x_n$ and ~~population~~ statistic estimate $\hat{\theta}$ based upon the data. To estimate the sampling distribution of θ do the following.

0. set $B=1$
1. Take a sample, with replacement, of size n from $x_1, x_2, \ldots, x_n$. Call this sample $x_1^*, x_2^*, \ldots x_n^*$.
2. Based upon $x_1^*, x_2^*, \ldots, x_n^*$ compute $\hat{\theta}_b$. Store $\hat{\theta}_b$
3. Increment $b = b+1$. Go back to step 1 until $b = B$ (where B is large).

The $\hat{\theta}_b$'s can be used to estimate the sampling distribution of θ.

In a later assignment you will have the opportunity use the bootstrap.

Linear Combination

Given r.v. $X_1, \ldots, X_n$ and constants $a_1, a_2, \ldots a_n$ the rv

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n = \sum_{i=1}^{n} a_i X_i$$

is called a linear combination of the X_i's.

Note if $a_i = \frac{1}{n}$ for all i then

$$Y = \frac{1}{n} X_1 + \frac{1}{n} X_2 + \cdots + \frac{1}{n} X_n = \bar{X}$$

and $a_i = 1$ gives

$$Y = X_1 + X_2 + \cdots + X_n = T_n$$

where T_n is a symbol for sum of n rv.

Note in this formula we don't assume that the X_i are iid.

Expectations/Variances of linear combinations

Let $X_1, X_2, \ldots X_n$ be r.v with means $\mu_1, \ldots \mu_n$ and variances $\sigma_1^2, \ldots \sigma_n^2$. then

1) $E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n)$

2) if $X_1, \ldots X_n$ are all independent

$$Var(a_1X_1 + a_2X_2 + \cdots + a_nX_n)$$
$$= a_1^2 Var(X_1) + a_2^2 Var(X_2) + \cdots + a_n^2 Var(X_n)$$
$$= a_1^2\sigma_1^2 + a_2\sigma_2^2 + \cdots + a_n\sigma_n^2$$

(note this implies $SD(a_1X_1 + \cdots + a_nX_n)$

$$= \sigma_{a_1X_1 + \cdots + a_nX_n} = \sqrt{a_1^2\sigma_1^2 + \cdots + a_n\sigma_n^2}$$

3 For $X_1, \dots, X_n$ not all independent

$$Var(a_1X_1 + \dots + a_nX_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j Cov(X_i, X_j)$$

Proofs (for continuous case)

1) $$E[a_1X_1 + \dots + a_nX_n] = \int \cdots \int (a_1x_1 + \dots + a_nx_n) f(x_1, \dots x_n)\, dx_1 \dots dx_n$$

$$= \int\int\cdots\int a_1x_1 f(x_1, \dots, x_n)\, dx_1 \dots dx_n$$

$$+ \dots + \int\int\cdots\int a_nx_n f(x_1, \dots, x_n)\, dx_1 \dots dx_n$$

$$= a_1 \int x_1 f_{X_1}(x_1)\, dx_1 + \dots + a_n \int x_n f_{X_n}(x_n)\, dx_n$$

$$= a_1 E[X_1] + \dots + a_n E[X_n]$$

2) because $X_1, \dots, X_n$ are all independent

$$f(x_1, \dots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_n}(x_n)$$

then use definition of variance or see proof of 3

3)

$Var(a_1X_1 + a_2X_2 + \dots + a_nX_n)$

$$= E\left(\left(a_1X_1 + a_2X_2 + \dots + a_nX_n - a_1E(X_1) - \dots - a_nE(X_n)\right)^2\right)$$

$$= E\Big(a_1^2(X_1 - E(X_1))^2 + a_2^2(X_2 - E(X_2))^2 + a_1a_2(X_1 - E(X_1))(X_2 - E(X_2))$$
$$+ \dots + a_{n-1}a_n(X_{n-1} - E[X_{n-1}])(X_n - E(X_n))\Big)$$

$$= a_1^2E((X_1 - E(X_1))^2) + a_2^2E((X_2 - E(X_2))^2)$$
$$+ a_1a_2E((X_1 - E(X_1))(X_2 - E(X_2)))$$
$$+ \dots a_{n-1}a_nE((X_{n-1} - E[X_{n-1}])(X_n - E(X_n)))$$

$$= a_1^2 Var(X_1) + a_2^2 Var(X_2) + a_1a_2 cov(X_1, X_2)$$
$$+ \dots + a_{n-1}a_n cov(X_{n-1}, X_n)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j cov(X_i, X_j)$$ since $Var(X_i) = cov(X_i, X_i)$

Note that above imply

$E(X_1 - X_2) = E(X_1) - E(X_2)$

$Var(X_1 - X_2) = Var(X_1) + Var(X_2)$ (X_1, X_2 independent)

$Var(X_1 - X_2) = Var(X_1) + Var(X_2) - 2cov(X_1, X_2)$.

Distribution of a linear combination of normal r.v.

---

when $X_1, \ldots X_n$ are independent normal rV (with possibly different means/sds), then any linear combination of these X_i's is also a normal rV with mean $E[a_1 X_1 + \cdots a_n X_n]$ an variance $Var(a_1 X_1 + \cdots a_n X_n)$.

Some Examples

1) Let $X_1, X_2, \ldots X_5$ be independent normal rV with means $\mu_1 = \mu_2 = 20$ $\mu_3 = \mu_4 = \mu_5 = 21$ and $\sigma_1^2 = \sigma_2^2 = 4$ $\sigma_3^2 = \sigma_4^2 = \sigma_5^2 = 3.5$

Let $y = \frac{X_1 + X_2}{2} - \frac{X_3 + X_4 + X_5}{3}$

what is $P(0 \leq Y)$? $P(-1 \leq Y \leq 1)$?

First what is $E[Y]$?

$$E[Y] = \frac{1}{2}E[X_1] + \frac{1}{2}E[X_2] - \frac{1}{3}E[X_3] - \frac{1}{3}E[X_4] - \frac{1}{3}E[X_5]$$

$$= \frac{40}{2} - \frac{63}{3} = 20 - 21 = -1$$

what is $Var(Y)$?

$$Var(Y) = \left(\frac{1}{2}\right)^2 Var(X_1) + \left(\frac{1}{2}\right)^2 Var(X_2) + \left(\frac{1}{3}\right)^2 Var(X_3)$$

$$+ \left(\frac{1}{3}\right)^2 Var(X_4) + \left(\frac{1}{3}\right)^2 Var(X_5)$$

$$= \left(\frac{1}{2}\right)^2 (8) + \left(\frac{1}{3}\right)^2 (10.5)$$

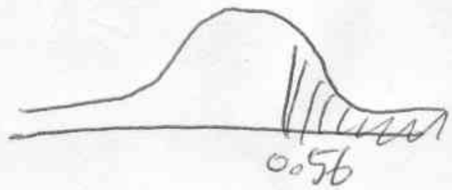
$$= 3.1666 \text{ (4dp)}$$

$$\Rightarrow SD(Y) = \sqrt{3.1666} = 1.7795$$

Since $X_1, \dots X_5$ are all normal r.v so is Y.
so use standard methodology

$$P(Y \geq 0) = P\left(\frac{Y - -1}{\sqrt{3.1666}} \geq \frac{0 - -1}{1.7795}\right)$$

$$= P(Z > 0.56)$$

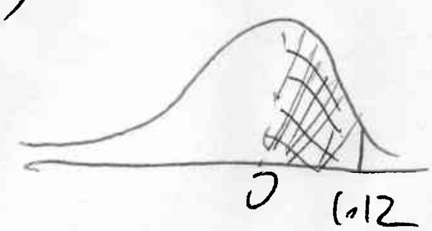


$$= 1 - \Phi(0.56)$$

$$= 1 - .7123 = .2877$$

$$P(-1 \leq Y \leq 1) = P\left(\frac{-1 - -1}{1.7795} \leq Z \leq \frac{1 - -1}{\sqrt{3.1666}}\right)$$

$$= P(0 \leq Z \leq 1.12)$$



$$= \Phi(1.12) - \Phi(0)$$

$$= .8686 - .5$$

$$= .3686$$