Lecture notes II

27.8.07

Review of Statistics (SW ch. 3)

- Random sampling:

  random variable of interest: Y

  n units (individuals, firms, etc.) drawn randomly from the population

  Y_i is the observation for unit i (r.v.)

- i.i.d random variables:

  $Y_1, \ldots, Y_n$ are i.i.d draws from the distribution of Y if

  i) All Y_i has identical distribution

  ii) All the Y_i are independent

  $$\Rightarrow P(Y_i = y_i / Y_j = y_j) = P(Y_i = y_i)$$

  if $i \neq j$ (discrete Y)

Distribution of sample average

$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$, Y_i are i.i.d from distribution with mean μ_Y and variance σ_Y^2

What is the distribution of $\bar{Y}$?

$E(\bar{Y}) = \mu_Y$ — $\bar{Y}$ is unbiased estimator of μ_Y

$Var(\bar{Y}) = \frac{1}{n} \sigma_Y^2$

In practice may not know σ_Y^2, but replace with estimator: $\hat{\sigma}_Y^2 = S_Y^2$ where $S_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. Can show that

$E(S_Y^2) = \sigma_Y^2$

The law of large numbers:

$\bar{Y} \xrightarrow{p} \mu_Y \iff P(|\bar{Y} - \mu_Y| > \varepsilon) \to 0$ when $n \to \infty$ for any ε.

That is: $\bar{Y}$ is a consistent estimator of μ_Y

Also: $S_Y^2 \xrightarrow{p} \sigma_Y^2$

The central limit theorem (CLT)

$\bar{Y}$ is approximately normally distributed when n is large : $\bar{Y} \sim N(\mu_Y, \frac{\sigma_Y^2}{n})$

Formally: $\sqrt{n}(\bar{Y} - \mu_Y) \xrightarrow{D} N(0, \sigma_Y^2)$

Estimators

Let $\hat{\mu}$ denote an estimator of μ. Estimators are functions of the data : $\hat{\mu} = h(Y_1, \ldots, Y_n)$ for some function $h()$, i.e. they are statistics.

Therefore they are random variables.

Properties of estimators:

- bias : $E(\hat{\mu}) - \mu$
- consistency $\hat{\mu} \xrightarrow{p} \mu$
- standard error : $\sqrt{E(\hat{\mu} - \mu)^2}$ (S.E.)

$\hat{\mu}$ is unbiased $\Rightarrow$ $SE(\hat{\mu})$ = St. deviation

By CLT, estimators $\hat{\mu}$ ~~they~~ often are approximately normal distributed in large samples: $\hat{\mu} \sim N(\mu, SE(\hat{\mu}))$

Good estimators should be consistent:

$$\hat{\mu} \xrightarrow{p} \mu$$

with ~~small~~ as "small" $SE(\hat{\mu})$ as possible.

Is there a better estimator of $\mu_Y = E(Y)$ than $\bar{Y}$?

Consider all linear estimators

$$\hat{\mu} = \sum_{i=1}^{N} d_i Y_i$$

In order to have $E(\hat{\mu}) = \mu_Y \Rightarrow$

$\sum d_i = 1$

To minimize $Var(\hat{\mu}) = \sum_{i=1}^{n} d_i^2 \sigma_Y^2$

it follows that $d_i = \frac{1}{n}$

$\Rightarrow$ $\bar{Y}$ is Best Linear Unbiased Estimator (BLUE)

Testing Hypotheses

Null hypothesis: H_0
Alternativ hypothesis: H_1

H_0 is believed to be true (maintained) unless clear evidence against H_0 in favour of H_1.

Example

H_0: $E(Y) = \mu^0$
H_1: $E(Y) \neq \mu^0$ (2-sided alternative)

To test H_0 vs H_1 we need a test statistic (test observator) such as "extreme" values of the test statistic can be considered as adverse towards H_0

<u>P-value</u>: The probability that the test statistic is <u>at least</u> as adverse (extreme) as actually observed in the data given that H_0 is true.

Reject H_0 if p-value is sufficiently low, i.e. lower that level of significance (e.g. 5%)

<u>Example</u>:

$H_0: \mu_Y = \mu^0$

$H_1: \mu_Y \neq \mu^0$

Estimator of μ_Y is $\bar{Y}$. If $|\bar{Y} - \mu_Y|$ is large, this is <u>adverse</u> towards H_0.

Natural test-statistic is t-statistic:

$$t = \frac{\bar{Y} - \mu^{\circ}}{\sqrt{Var(\bar{Y})}} \sim N(0,1)$$

if H_0 is true (asymptotically, i.e. for large n)

Large value of $|t^{act}|$ is evidence against H_0.

$Var(\bar{Y}) = \frac{\sigma_Y^2}{n}$ replace by $\frac{S_Y^2}{n}$ since $S_Y^2 \xrightarrow{P} \sigma_Y^2$ $\left(S_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right)$

$$\begin{aligned} \text{p-value} &\equiv P(|t| \geq |t^{act}|) \\ &= P(t \geq |t^{act}|) + P(t \leq -|t^{act}|) \\ &= 2(1 - \Phi(|t^{act}|)) \end{aligned}$$

where $\Phi()$ is c.d.f of standard normal distribution.

One sided test I

$H_0: \mu_Y = \mu^0$

$H_1: \mu_Y < \mu^0$

Test statistic: $t = \dfrac{\bar{Y} - \mu^0}{\sqrt{\dfrac{s_Y^2}{n}}}$

which is approximately $N(0,1)$ when H_0 is true.

If $\mu_Y < \mu^0 \Rightarrow \bar{Y} - \mu^0 < \bar{Y} - \mu_Y$ so that t-statistic negative on average and large negative values of t^{act} is adverse

$$P\text{-value} = P(t \le t^{act}) = \Phi(t^{act})$$

One-sided test II :

$$H_0 : \mu_Y = \mu_0$$
$$H_1 : \mu_Y > \mu_0$$

$$t = \frac{\bar{Y} - \mu^0}{\sqrt{\frac{s_Y^2}{n}}}$$

If $\mu_Y > \mu_0 \Rightarrow \bar{Y} - \mu^0 > \bar{Y} - \mu_Y$

Hence t^{act} tend to be positive. Large positive values of t^{act} is adverse

$$p\text{-value} = P(t \geq t^{act}) = 1 - \Phi(t^{act})$$

<u>Example</u> . . . .

Comparing different population parameters

$Y_i = \beta_0^w + \beta_1^w S_i + u_i$ if person i is woman

$Y_i = \beta_0^m + \beta_1^m S_i + u_i$ if person i is man

$S_i \equiv$ years of schooling for person i

$Y_i \equiv$ log earnings of i

u_i is an error term with

$E(u_i \mid S_i, \text{man}) = 0$

$E(u_i \mid S_i, \text{woman}) = 0$

Two populations: men and women

For men: $E(Y_i \mid S_i) = \beta_0^m + \beta_1^m S_i$

For women: $E(Y_i \mid S_i) = \beta_0^w + \beta_1^w S_i$

Marginal return of schooling is β_1^m for men and β_1^w for women. Are $\beta_1^m = \beta_1^w$?

Hypotesis test:

$H_0 : \beta_1^w = \beta_1^m$

$H_1 : \beta_1^w \neq \beta_1^m$

Test statistic: $T = \dfrac{\hat{\beta}_1^w - \hat{\beta}_1^m}{SE(\hat{\beta}_1^w - \hat{\beta}_1^m)}$

Under H_0: $E(T) = 0$, $Var(T) = 1$

$$SE(\hat{\beta}_1^w - \hat{\beta}_1^m) = \sqrt{Var(\hat{\beta}_1^w - \hat{\beta}_1^m)}$$

$$Var(\hat{\beta}_1^w - \hat{\beta}_1^m) = Var(\hat{\beta}_1^w) + Var(\hat{\beta}_1^m) - 2\underbrace{Cov(\hat{\beta}_1^w, \hat{\beta}_1^m)}_{0} = Var(\hat{\beta}_1^w) + Var(\hat{\beta}_1^m)$$

$= 0$ if different/independent populations

p-value: $P(|T| \geq |T^{act}|)$