Last time

Sec 10.4 Normal (2nd example on CI and test)

Sec. 5.4 Unbiased estimator

Sec 11.1 Bias / Variance (of estimators)

11.2 German Tank

Today.

Sample variance.

Recall Defn of $\mathrm{var}(X) = \mathbb{E}(X - \underset{\mathbb{E}X}{\mu})^2 = \mathbb{E}D_X^2$

"average of squared deviation"

Suppose $X_1, X_2, \dots, X_n$ i.i.d. from population with mean μ & sd σ ($\mathrm{var}\ \sigma^2$)

$D_X := X - \mu$?

$\bar{X}$, the sample mean, is an unbiased estimator of μ

$\hat{D}_{X_i} := X_i - \bar{X}$ is the estimated deviation

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} \hat{D}_{X_i}^2$$

Note: Use $(n-1)$ not n to get an unbiased estimator

The one with $\frac{1}{n-1}$ is called "sample variance"

$$\sum_{i=1}^{n} \hat{D}_{X_i}^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$$
$$= \sum_{i=1}^{n} (X_i^2 - 2\bar{X}X_i + \bar{X}^2)$$
$$= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2.$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \boxed{\sum_{i=1}^{n} X_i^2 - \cancel{2}n\bar{X}^2} \cancel{+n\bar{X}^2}$$

$$\mathbb{E}\sum_{i=1}^{n} \hat{D}_{X_i}^2 = \mathbb{E}\sum_{i=1}^{n} X_i^2 - n\,\mathbb{E}\bar{X}^2$$
$$= \sum_{i=1}^{n} \mathbb{E}X_i^2 - n\,\mathbb{E}\bar{X}^2$$
$$= n(\cancel{\mu^2} + \sigma^2) - n(\cancel{\mu^2} + \frac{\sigma^2}{n})$$
$$= (n-1)\sigma^2$$

$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$, with $\mathrm{Var}(X) = \sigma^2$, $\mathbb{E}X = \mu$

$\Downarrow$

$\mathbb{E}X^2 = \mu^2 + \sigma^2$

$\bar{X}$ has mean μ & var $\frac{\sigma^2}{n}$

Note: $\mathbb{E}\hat{\sigma}^2 = \sigma^2$ $\Rightarrow$ $\hat{\sigma}^2$ is unbiased w.r.t. σ^2

$\not\Rightarrow$ $\hat{\sigma}$ is Not unbiased w.r.t σ

$X_1, X_2, \dots, X_n$ with mean μ & sd σ

95% CI of μ is given by

$$\bar{X} \pm 2\,\sigma/\sqrt{n}$$

thanks to CLT: for large n, $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P(-2 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 2) = 95\%$$

What if we do not know σ?

We use $\hat{\sigma}$ instead. $\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim N(0,1)$?

Suppose that we know $\frac{\sigma}{\hat{\sigma}} \xrightarrow{d} 1$ (or equivalently, $\hat{\sigma}^2 \xrightarrow{d} \sigma^2$)

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

$\times$

$$\frac{\sigma}{\hat{\sigma}} \times \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} 1 \times N(0,1)$$

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \xrightarrow{d} N(0,1)$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$
$$= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right)$$

by the weak law of large numbers

$\frac{n}{n-1} \to 1$, $\quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{d} \mathbb{E}X_i^2 = \mu^2 + \sigma^2$, $\quad \bar{X}^2 \xrightarrow{d} (\mathbb{E}X_i)^2 = \mu^2$

$(n \to \infty)$ $\xrightarrow{d} \sigma^2$ $\Rightarrow$ $\frac{\sigma}{\hat{\sigma}} \xrightarrow{d} 1$ $\Rightarrow$ $\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \xrightarrow{d} N(0,1)$

Conclusion: for n large, we may use

$\hat{\sigma} := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ to replace the population variance

when using CLT (CI / test) in case we do not know the population variance

Sec 11.3 Least square linear regression.

Setup of regression:

want to estimate Y

we know X

construct an estimator / predictor of Y based on X

(a function $g(X)$ of X)

In particular, linear regression means we use a linear function $g(x) = aX + b$

i.e. we wish to find the optimal pair (a, b) such that

$\hat{Y} = aX + b$ is closest to Y → in terms of minimizing error

$\mathrm{Error} = Y - \hat{Y}$

$\mathrm{MSE} = \mathbb{E}(\mathrm{error}^2) = \mathbb{E}(Y - \hat{Y})^2$

e.g. $\hat{Y} = b$. find b that minimize $\mathrm{MSE} = \mathbb{E}(Y - b)^2$

$\mathbb{E}(Y-b)^2 = \mathbb{E}Y^2 - 2b\,\mathbb{E}Y + b^2$ – a quadratic function of b

method 1: $\mathbb{E}Y^2 + \underbrace{(b - \mathbb{E}Y)^2}_{\geq 0} - (\mathbb{E}Y)^2$ – reaches 0 only when $b = \mathbb{E}Y$

method 2: Set derivative to 0:

$$2b - 2\mathbb{E}Y = 0 \Rightarrow b = \mathbb{E}Y$$

Conclusion: select $b = \mathbb{E}Y$ as the least square constant estimator.

Note. In the regression problem, we treat X and Y as known R.V.'s so that we may focus on the selection of regression parameter

senerio:

Use some data where we know (X_i, Y_i) to construct the linear regression model

$\Rightarrow$ Use the model and some more input X_j to give prediction on Y_j

Find (a, b) to minimize $\mathrm{MSE} = \mathbb{E}(Y - (aX + b))^2$

Step 1: for some fixed a, find $b \overset{= b(a)}{}$ to minimize MSE

(minimizing) $\mathbb{E}((Y - aX) - b)^2 \Rightarrow b = b(a) = \mathbb{E}(Y - aX)$
$$= \mu_Y - a\mu_X$$ → best intercept given slope

Step 2: find a such that $(a, b(a))$ - - - -

$$\mathrm{MSE} = \mathbb{E}\left(Y - (aX + (\mu_Y - a\mu_X))\right)^2$$
$$= \mathbb{E}\left((Y - \mu_Y) - a(X - \mu_X)\right)^2$$
$$= \mathbb{E}(D_Y - aD_X)^2$$
$$= \mathbb{E}D_Y^2 - 2a\,\mathbb{E}D_X D_Y + a^2\,\mathbb{E}D_X^2$$
$$= \sigma_Y^2 - 2a\,\mathbb{E}D_X D_Y + a^2\sigma_X^2.$$

$$\begin{cases} \hat{b} = \mu_Y - \hat{a}\mu_X \\ \hat{a} = \frac{\mathbb{E}D_X D_Y}{\sigma_X^2} \end{cases}$$

Set $\frac{d}{da}(\mathrm{MSE}) = -2\mathbb{E}D_X D_Y + 2a\sigma_X^2 = 0$

$$\frac{\mathbb{E}D_X D_Y}{\sigma_X^2} = a$$ → best slope (with best intercept)

"$\mathbb{E}D_X D_Y$" – Covariance of X & Y

$=: \mathrm{Cov}(X, Y)$. $\quad \mathrm{Cov}(X, X) = \mathrm{Var}(X)$

e.g. X educational level (year)

Y - annual income (dollar)

$\mathrm{Cov}(X, Y)$ – (year · dollar)

Soln: use standardized deviation instead (Recall Chebyshev's)

$$\mathrm{Cov}(X, Y) = \mathbb{E}D_X D_Y$$

$r = r_{XY} = \mathrm{Corr}(X, Y) = \mathbb{E}\frac{D_X}{\sigma_X}\frac{D_Y}{\sigma_Y}$ – a number (with no unit)

"Correlation coefficient" "correlation"

This agrees with the Data 8 definition of correlation:

" $r_{X,Y}$ is the average of the products of X, Y measured in standard units"

$$\mathrm{Cov}(X, Y) = \sigma_X\sigma_Y\,\mathrm{Corr}(X, Y)$$

$$\hat{a} = \frac{\mathbb{E}D_X D_Y}{\sigma_X^2} = \frac{\cancel{\sigma_X}\sigma_Y r}{\sigma_X^{\cancel{2}}} = r \cdot \frac{\sigma_Y}{\sigma_X}.$$ (checked the result from Data 8)

Conclusion.

Least square linear regression line:

$$\hat{Y} = \hat{a}X + \hat{b}$$

where $\hat{Y}$ is the linear regression with the least MSE,

$$\begin{cases} \hat{a} = r \cdot \frac{\sigma_Y}{\sigma_X}, \\ \hat{b} = \mu_Y - \hat{a}\mu_X. \end{cases}$$