* Announcement

① Quiz 10 : Thu (11/19) 9:00 AM

~ Fri (11/20) 9:00 AM

Ch 10.3 ~ Ch 11.1

Exponential Dist'n, $X \sim Exp(\lambda)$

→ $f(x) = \lambda e^{-\lambda x}$ for $x \ge 0$. Density

$E(X) = \frac{1}{\lambda}$, $h = \frac{\log 2}{\lambda}$ (Find h s.t. $F(h) = S(h)$) to find median

half-life = median

such that

Normal Dist'n.

$X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ ⇒ $X+Y \sim$?

If indep, $X+Y \sim N(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$

STAT 88: Lecture 35

95% CI of $\mu_X - \mu_Y$: $\left(\bar{X} - \bar{Y} \pm 2\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right)$

- Bias - Variance Decomposition.

MSE = Bias^2 + Var.

You should be able to tell from figure if Bias/Var small or large

E.g.

Bias ↓ Var ↑   Bias ↑ Var ↓

Contents

Warm up:

German tanks were numbered $1, 2, 3, ..., N$, with N unknown, during World War 2 and the Allies needed to estimate N. They captured 5 tanks numbered $20, 31, 43, 78$ and 92. Can you find an unbiased estimate of N?

① $E(\bar{X}) = \frac{N+1}{2}$ ← population mean Unif$\{1, 2, \cdots, N\}$

⇒ $2E(\bar{X}) - 1 = N$

⇒ $E(\underbrace{2\bar{X} - 1}_{= T_1}) = N$

$T_1 = 2\bar{X} - 1 = 2 \cdot \frac{20+31+43+78+92}{5} - 1 = 109.6$

$Var(T_1) = Var(2\bar{X} - 1)$

$= 4 \cdot Var(\bar{X})$

$= 4 \frac{\sigma^2}{n}$ $\sigma^2 = Var(\text{Unif}\{1, 2, \cdots, N\})$

② $T_2 = \max\{X_1, \cdots, X_n\}$

$\underbrace{\cdots\cdots}_{\text{Gap 1}}\ 20\ \underbrace{\cdots\cdots}_{\text{Gap 2}}\ 31\ \cdots\ 43\ \cdots\ 78\ \cdots\ \underset{T_2}{92}\ \underbrace{\cdots\cdots}_{\text{Gap 6}}\ \underset{N}{|}$

T_2: biased estimator.

Expected length of a gap $= \frac{N-5}{6} \to \frac{N-n}{n+1}$ n sample size.

$E(\text{Gap } 6) = E(N - T_2) = \frac{N-n}{n+1}$

⇒ $N - E(T_2) = \frac{N-n}{n+1}$

⇒ $E(T_2) = N - \frac{N-n}{n+1} = \frac{n}{n+1} \cdot (N+1)$

⇒ $\frac{n+1}{n} \cdot E(T_2) - 1 = N$

⇒ $E(\underbrace{\frac{n+1}{n} \cdot T_2 - 1}_{= T_3}) = N$

$T_3 = \frac{6}{5} \cdot 92 - 1 = 109.4$

$Var(T_3) = Var(\frac{n+1}{n} T_2 - 1)$

$= (\frac{n+1}{n})^2 Var(T_2)$

$= \underbrace{(\frac{6}{5})^2}_{> 1} \cdot Var(T_2)$

Last time

Bias and Variance

We score how good an estimator T of a parameter θ is by

$$\mathrm{MSE}_\theta(T) = E_\theta((T-\theta)^2).$$

$T-\theta$

T E(T) θ

D(T) B(T)

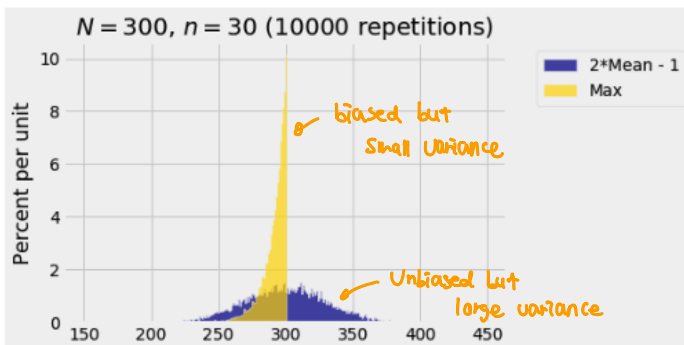
And we showed

$$\mathrm{MSE}_\theta(T) = B_\theta^2(T) + \mathrm{Var}_\theta(T),$$

where

$$\underbrace{B_\theta(T)}_{\text{bias}} = E_\theta(T) - \theta \text{ and } \underbrace{\mathrm{Var}_\theta(T)}_{\text{Variance}} = E_\theta((T - E_\theta(T))^2).$$

The best estimator is *not* always unbiased.



T_1

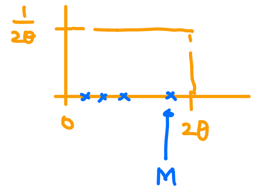
T_2

To find an unbiased estimator, start with a statistic whose expectation is a linear function of the parameter.

11.2. The German Tank Problem

<u>Practice for finding an unbiased estimator</u>

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, 2\theta)$. Let

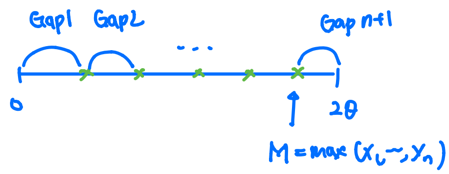


$$M = \max\{X_1, \ldots, X_n\}.$$

Is M a biased estimator?

$E(M) < 2\theta \Rightarrow M$ is biased.

Find $E(M)$.



Expected length of a Gap

$= \frac{2\theta - 0}{n+1}$ ← total length
← # Gaps

$E(\text{Gap } n+1) = \underbrace{E(2\theta - M)}_{= 2\theta - E(M)} = \frac{2\theta}{n+1}$

Discrete:

$\frac{N-n}{n+1}$

$\Rightarrow E(M) = 2\theta - \frac{2\theta}{n+1} = \frac{n}{n+1} \cdot 2\theta$

(Discrete case
$E(T_2) = \frac{n}{n+1}(N+1)$)

(see appendix for another way to calculate $E(M)$ using cdf of M)

Find an unbiased estimator for 2θ.

$E(M) = \frac{n}{n+1} \cdot 2\theta$

$\Rightarrow \frac{n+1}{n} E(M) = 2\theta$

$\Rightarrow \underbrace{E\left(\frac{n+1}{n} M\right)}_{= T} = 2\theta$

$T = \frac{n+1}{n} \cdot M$ unbiased estimator.

Ch11 : Linear regression in population version
Ch12 : " Sample version

11.3. Least Squares Linear Regression

Let (X, Y) be a random pair of father and son heights from the population:

$$X: \text{ father height, and } Y: \text{ son height.}$$

We want to estimate Y, call this $\widehat{Y}$, by the function

$$\widehat{Y} = aX + b,$$

for some slope a and intercept b.

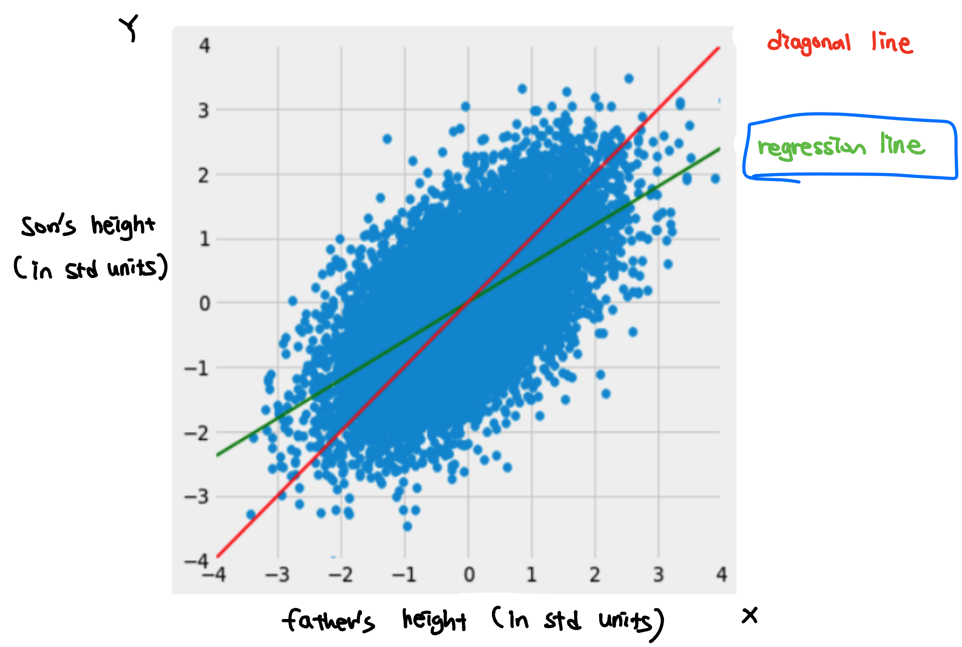
You plug in X into $\widehat{Y} = aX + b$ to predict Y. To find a and b, in Data 8, you collected n pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and made a scatter plot. The regression line is the "best" fitting line $\widehat{Y} = aX + b$ through your scatter plot. The formulas are:

$$\text{slope of the regression line} = r \frac{\text{SD of } Y}{\text{SD of } X},$$

Correlation

and

$$\text{intercept of the regression line} = (\text{average of } Y) - \text{slope} \times (\text{average of } X).$$



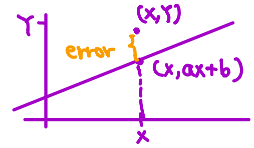
We will now derive the formulas mathematically using calculus and properties of expectation and variance.

$T \rightsquigarrow \theta$

$MSE(T) = E((T-\theta)^2)$

Mean Squared Error For the random point (X, Y), the mean squared error of a linear predictor of Y based on X depends on the slope a and intercept b of the line used. So let us define $\text{MSE}(a, b)$ to be the mean squared error when we use the line $aX + b$ to predict Y. That is,

$$\text{MSE}(a, b) = E((\overbrace{Y - (aX + b)}^{\text{error}})^2).$$



Note that we average over all random (X, Y) pairs in the population. We have to find the values of a and b that minimize this function.

Notation

- $E(X) = \mu_X$, $\text{SD}(X) = \sigma_X$.
- $E(Y) = \mu_Y$, $\text{SD}(Y) = \sigma_Y$.

Best Intercept for a Fixed Slope Fix slope a, and solve $\frac{\partial \text{MSE}(a,b)}{\partial b} = 0$. Since

$$\begin{aligned}
\text{MSE}(a, b) &= E((Y - (aX + b))^2) \\
&= E(((Y - aX) - b)^2) \\
&= E((Y - aX)^2 - 2b(Y - aX) + b^2) \quad \text{additivity} \\
&= E((Y - aX)^2) - 2b \cdot E(Y - aX) + \underset{E(b^2)}{\overset{\shortparallel}{b^2}}.
\end{aligned}$$

Solve $\frac{\partial \text{MSE}(a,b)}{\partial b} = 0$ for b:

$$\begin{aligned}
&\shortparallel \quad -2E(Y-aX) + 2b = 0 \\
&\Rightarrow b = E(Y-aX) \\
&\quad = E(Y) - aE(X) \\
&\quad = \underbrace{\mu_Y - a\mu_X}_{\text{write it as } \hat{b}_a}
\end{aligned}$$

Best Slope For each fixed slope a, we first plug in the best intercept we just found. The the error becomes

$$\text{Error} = Y - (aX + \underbrace{\widehat{b}_a}_{= \mu_Y - a\mu_X}) = Y - (aX + \mu_Y - a\mu_X)$$
$$= Y - aX - \mu_Y + a\mu_X$$
$$= Y - \mu_Y - a(X - \mu_X)$$
$$= \underbrace{D_Y}_{= Y-\mu_Y \text{ (Deviation of } Y)} - a\underbrace{D_X}_{= X-\mu_X}.$$

$E(D_Y^2) = E((Y-\mu_Y)^2) = Var(Y) = \sigma_Y^2$

$E(D_X^2) = \sigma_X^2$

Then

$$\begin{aligned}\text{MSE}(a, \widehat{b}_a) &= E((D_Y - aD_X)^2) \\ &= E(D_Y^2) - 2aE(D_X D_Y) + a^2 E(D_X^2) \\ &= \sigma_Y^2 - 2aE(D_X D_Y) + a^2\sigma_X^2.\end{aligned}$$

Solve $\frac{d\text{MSE}(a,\widehat{b}_a)}{da} = 0$ for a:

$= 0 - 2E(D_X \cdot D_Y) + 2a \cdot \sigma_X^2 = 0$

$\Rightarrow a = \frac{E(D_X \cdot D_Y)}{\sigma_X^2} = \underbrace{\frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X^2}}_{\text{write it as } \widehat{a}}$

So the regression line is

$$\widehat{Y} = \widehat{a}X + \widehat{b},$$

→ Best linear predictor in the sense that it minimizes $MSE = E((Y-(aX+b))^2)$

where

$$\widehat{a} = \frac{E(D_X D_Y)}{\sigma_X^2} \text{ and } \widehat{b} = \mu_Y - \widehat{a} \cdot \mu_X.$$

Correlation $E(D_X D_Y)$ is called the covariance of X and Y. If X is father's height (ft) and Y is son's height (ft), then $E(D_X D_Y)$ has unit ft^2.

If we divide it by $\sigma_X \sigma_Y$,

$$r = \frac{E(D_X D_Y)}{\sigma_X \sigma_Y} \qquad \frac{\cancel{ft^2}}{\cancel{ft^2}}$$

is unitless and called the correlation coefficient of X and Y. This tells you

$$\text{Covariance } E(D_X D_Y) = r\sigma_X \sigma_Y,$$

so

$$\widehat{a} = \frac{E(D_X D_Y)}{\sigma_X^2} = \frac{r\cancel{\sigma_X}\sigma_Y}{\sigma_X^{\cancel{2}}} = \frac{r\sigma_Y}{\sigma_X}.$$

Appendix

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, 2\theta)$. Let $M = \max\{X_1, \ldots, X_n\}$. Calculate the density of M by first calculating the CDF of M.

$$\begin{aligned}
F(m) &= P(M \leq m) \\
&= P(X_1 \leq m, \ldots, X_n \leq m) \\
&= P(X_1 \leq m) \cdots P(X_n \leq m) \\
&= P(X_1 \leq m)^n = \left(\frac{m}{2\theta}\right)^n.
\end{aligned}$$

So,

$$f(m) = \frac{dF(m)}{dm} = nm^{n-1} \cdot \frac{1}{(2\theta)^n}.$$

Now we calculate

$$\begin{aligned}
E(M) &= \int_0^{2\theta} mf(m)dm \\
&= \frac{n}{(2\theta)^n} \int_0^{2\theta} m^n dm \\
&= \frac{n}{(2\theta)^n} \frac{m^{n+1}}{n+1} |_0^{2\theta} \\
&= (2\theta) \frac{n}{n+1}.
\end{aligned}$$