Stat 88: Probability & Math. Statistics in Data Science

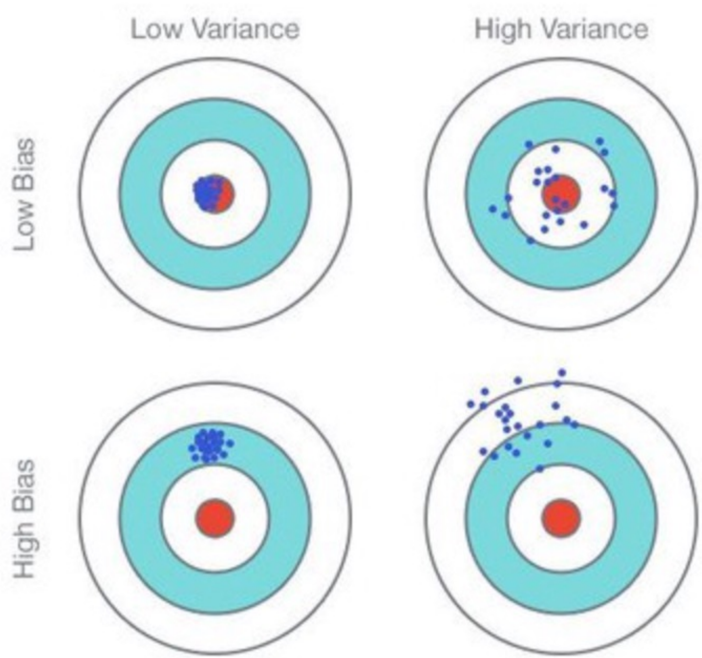


Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off

https://medium.com/@mp32445/understanding-bias-variance-tradeoff-ca59a22e2a83

Chapter 11

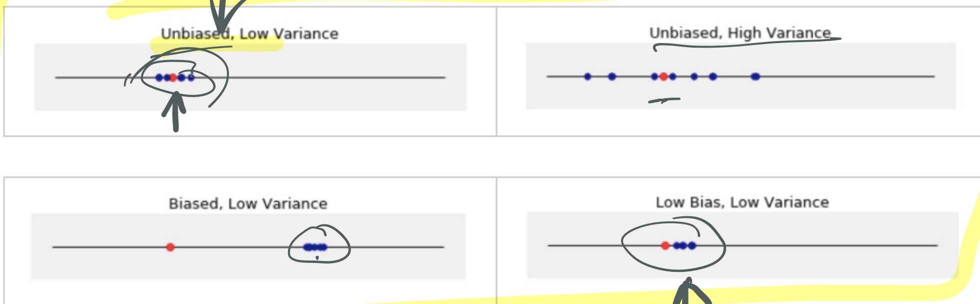
Bias, Variance, and Least Squares

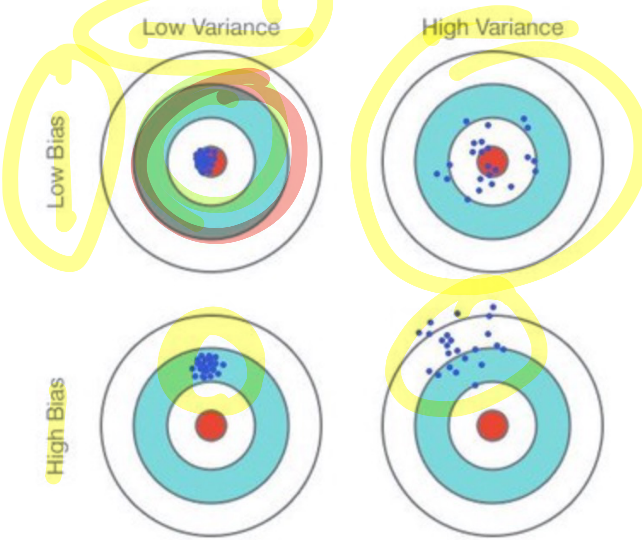
Understanding Bias and Variance

$\bar{X}$ estimates μ

$E(\bar{X}) = \mu$

$\bar{X}$ is an unbiased estimator of μ





T: estimator (rv)

θ: parameter (target, constant)

Say T is *unbiased* if $E(T) = \theta$

$$MSE_\theta(T) = E[(T - \theta)^2]$$

Mean Squared Error

Bias, Variance, and Mean Squared Error

$\mathbb{E}(\bar{X}) - \mu = 0$

estimator of pop. mean ↑ ↑ pop mean

- Bias: $\boldsymbol{B_\theta(T) = E_\theta(T) - \theta}$ (note that $B_\theta(T)$ is a constant)
- Bias is difference between expected value of the estimator and the target.
- Suppose B_θ is positive, what does this mean?

$B(T) > 0 \longrightarrow E(T) - \theta = 0 \Rightarrow E(T) > \theta$ over estimating on average

$B(T) < 0 \Rightarrow E(T) < \theta$ on avg, underestimating

- Deviation (from the mean): $\boldsymbol{D_\theta(T) = T - E_\theta(T)}$ (note that $D_\theta(T)$ is a r.v.)

$\mathbb{E}(D(T))^2 = \mathrm{Var}(T) = \mathbb{E}[(T - E(T))^2]$

$\mathbb{E}(D(T)) = \mathbb{E}(T - \mathbb{E}(T)) = \mathbb{E}(T) - \mathbb{E}(T) = 0$

- Error: $\boldsymbol{T - \theta} = \underbrace{T - E(T)} + \underbrace{E(T) - \theta} = D(T) + B(T)$
- Mean Squared Error: $\boldsymbol{MSE_\theta(T) = E}[(\boldsymbol{T - \theta})^2]$

$\mathbb{E}((T-\theta)^2) = \mathbb{E}[(D(T) + B(T))^2]$

- What is the expected value of $D_\theta(T)$? What about $(D_\theta(T))^2$?

$\mathbb{E}(D(T)) = 0$

$\mathbb{E}((D(T))^2) = \mathrm{Var}(T)$

Make sure you understand

Exercise: Is $\mathbb{E}(X^2) = (\mathbb{E}(X))^2$ for any X? (No.)

$\mathbb{E}(X^2) - (\mathbb{E}(X))^2 = Var(X)$. When is $Var(X) = 0$?

Mean Squared Error & the Bias-Variance Decomposition

$D(T) \equiv D_\theta(T)$

- $MSE_\theta(T) = E[(T-\theta)^2] = \mathbb{E}\left[\left(\underbrace{T - \mathbb{E}(T)} + \mathbb{E}(T) - \theta\right)^2\right]$

$\mathbb{E}(cX) = c\,\mathbb{E}(X)$

$$= \mathbb{E}\left[\left[D(T) + B(T)\right]^2\right]$$

$$= \mathbb{E}\left((D(T))^2 + 2\,D(T)\cdot B(T) + (B(T))^2\right)$$

$$= \underbrace{\mathbb{E}((D(T))^2)} + \underbrace{2B(T)\cdot \mathbb{E}(\cancel{D(T)})}_{0} + (B(T))^2$$

const. so $\mathbb{E}[(B(T))^2] = (B(T))^2$

$$= Var(T) + 0 + (B(T))^2$$

square of bias

$$MSE = \mathbb{E}((T-\theta)^2) = \text{Variance} + \text{Bias}^2$$

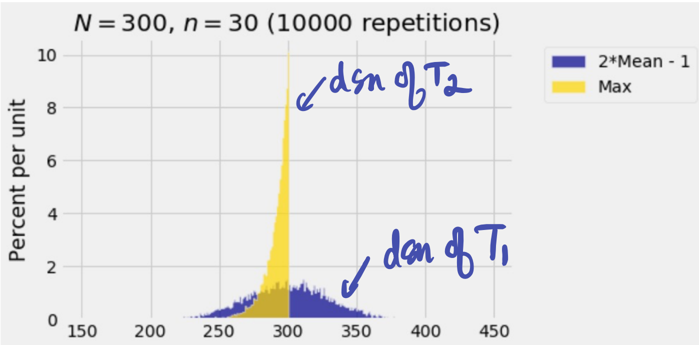
If MSE is small, then both variance & $\underbrace{\text{bias}^2}$ have to be small

if bias^2 is small, then bias is small.

German Tank Problem: $T_1, T_2, \& T_3$

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

$T_2 = \max\{X_1, X_2, \dots X_n\}$



Goal: Estimate # of tanks manufactured N using the tanks seen $X_1, \dots X_n$

Assume $X_1, \dots X_n$ is a SRS from $\{1, \dots N\}$

By symmetry, X_i's have the same dsn

$E(X_i) = \frac{N+1}{2}$, $E(\bar{X}) = \frac{N+1}{2}$

$T_1 = 2\bar{X} - 1 \Rightarrow E(T_1) = N$ ← Target

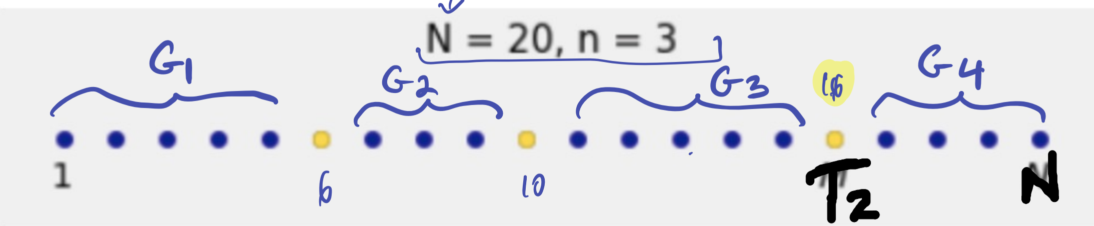
$Var(T_1) > Var(T_2)$

$E(T_1) = N$, $E(T_2) \le N$ (since $X_1, X_2 \dots X_n \le 300$

$E(\max X_i) \le 300$

$Bias(T_1) = 0$

$Bias(T_2) = E(T_2) - N$



gold dots: sampled values

blue dots: pop. values not sampled.

T_2 = max of sampled values

Total length = N

17 blue dots, 4 gaps

avg gap length = $\frac{17}{4}$

We know in this example that $N=20$
so avg gap length is $\frac{17}{4} = \frac{20-3}{4} = \frac{20-3}{3+1}$

In general, avg gap length $= \frac{N-n}{n+1} = \mathbb{E}(G_{n+1})$

$= \frac{\text{total \# of blue dots}}{\text{total \# of gaps}}.$

last gap from T_2 to N

$N = T_2 + G_{n+1}$

$\mathbb{E}(N) = N$ (N constant)

$N = \mathbb{E}(T_2) + \mathbb{E}(G_{n+1})$

$\mathbb{E}(G_{n+1}) = \frac{N-n}{n+1}$

$\boxed{B(T_2) = \mathbb{E}(T_2) - N = -\mathbb{E}(G_{n+1})}$

$\text{Bias}(T_2) = B(T_2) = -\mathbb{E}(G_{n+1}) = -\frac{(N-n)}{n+1}$

up to T_2 : n gold dots & n gaps of blue dots of expected length $\frac{N-n}{n+1}$

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

$$= \frac{n(n+1) + n(N-n)}{n+1} = \frac{\cancel{n^2} + n + nN - \cancel{n^2}}{n+1}$$

$$\mathbb{E}(T_2) = \frac{n(N+1)}{n+1}$$ ← get an unbiased estimator of N from this

Define $T_3 = \left(\frac{n+1}{n}\right) T_2 - 1$

Then $\mathbb{E}(T_3) = N$ (T_3 is an unbiased estimator of N)

$SD(T_3) = SD\left(\left(\frac{n+1}{n}\right) T_2\right) = \left(\frac{n+1}{n}\right) SD(T_2) \approx SD(T_2)$ for large n

Comparing $MSE(T_2)$ & $MSE(T_3)$

$T_2 = \max\{X_1, \ldots X_n\}$

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

$\mathbb{E}(T_3) = N$

$Var(T_3) \approx Var(T_2)$

$Bias(T_3) = 0$

$Bias(T_2) = -\frac{(N-n)}{n+1}$

$$MSE(T_2) = Var(T_2) + (B(T_2))^2$$

$$= Var(T_2) + \left[\frac{(N-n)^2}{(n+1)^2}\right]$$

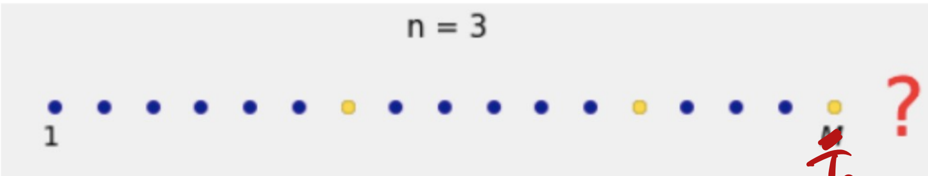
$$MSE(T_3) \approx Var(T_2) + 0$$

T_3 is better. It has same variance & no bias.

T_3 is called AUGMENTED Max.

The Augmented Maximum

See only T_2 = max of observed sample



T_2

Estimate G_{n+1} by Gap length up to T_2

Total # of blue dots upto T_2

$$= \frac{T_2 - n}{n}$$

n ← gold dots

} estimated gap length

↑ # of gaps up to T_2.

estimated

↓

$$N = T_2 + \left(\frac{T_2 - n}{n}\right) = \boxed{\underbrace{T_2\left(\frac{n+1}{n}\right) - 1}_{T_3.}}$$

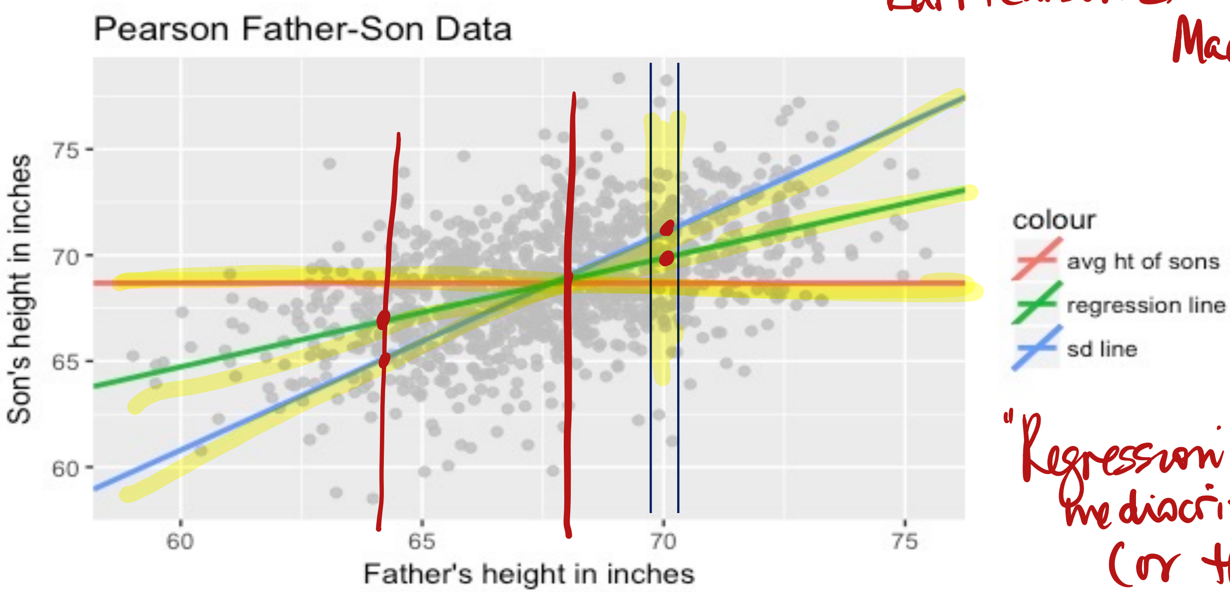
Simple Linear Regression (SLR)

- One of the most used statistical techniques, used for summarizing a scatterplot, and sometimes making inference about the data (understanding the relationships between x and y)
- You have seen SLR before, we will revisit the ideas, using random variables.
- Basically, we want a model that describes the relationship between the predictor (x) and response (y) variables. Question: Can we express the relationship mathematically? Perhaps as

$$Y = f(X) \quad \text{or} \quad Y = f(X) + random\ error$$

- Where we have a random ***pair*** (X, Y)
- Want to use a linear function of X to estimate Y, say $aX + b$
- That is, find the "best" line that fits these data (but have to define "best")

Karl Pearson & Francis Galton
MacTutor
mathematician biographies.



"Regression to mediocrity"
(or the middle)

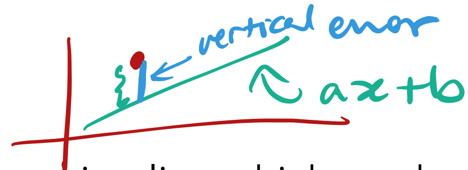
Want to predict y from x. Could use:

- Average of y (so don't use x at all)
- The SD (diagonal) line: better, but not so good (too steep)

Slope: $\frac{SD(y)}{SD(x)}$

- Much better, if the scatter plot shows a linear relationship, to use the **regression method**, which incorporates the correlation r (you have seen it before, but we haven't defined it yet)

The regression method



- The regression method is used to draw the regression line which can be used for prediction.

- It is also called the **least squares line** because it minimizes mean squared error. By *error* we mean the vertical difference between the y-value for some x, and the height of the regression line at that x.

$$e_i = y_i - (ax_i + b), i = 1, 2, \ldots, n$$

$$\sum_{i=1}^{n} e_i^2$$

minimize to find a, b

- From Data 8, do you recall the slope of the regression line? What about the intercept?

$$\text{Slope} = \frac{r \cdot SD(y)}{SD(x)}$$

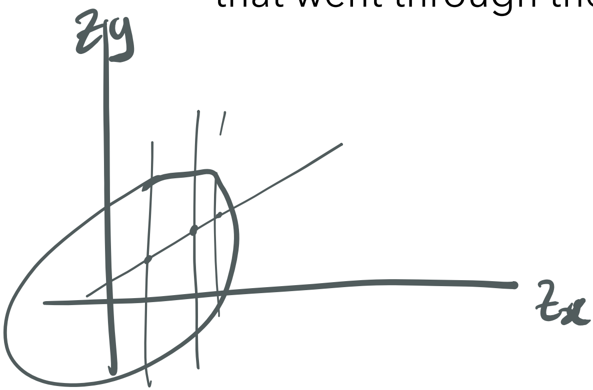
$$b = \bar{y} - a\bar{x}$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

unknown

The regression line aka the least squares line

- In Data 8, you found the slope of the regression line by
  - Using the geometry of the shape of the scatter plot (putting everything in standard units, and looking for the slope of the line that went through the centers of the vertical strips)



slope of reg line on z_x/z_y axes is r

$$z_y = r \cdot z_x$$

$$\frac{y - \bar{y}}{SD(y)} = r \cdot \frac{(x - \bar{x})}{SD(x)}$$

- And also by minimizing (numerically) the ***mean squared error:***
- The regression line is the *unique* straight line that minimizes the mean squared error of estimation among all straight lines, which is why it is called the "Least Squares" line

$$\hat{y} = \frac{r \cdot SD(y)}{SD(x)} \cdot x + \left(\bar{y} - r\frac{SD(y)}{SD(x)}\bar{x}\right)$$

↑ regression estimate