Stat 88: Probability & Math. Statistics in Data Science

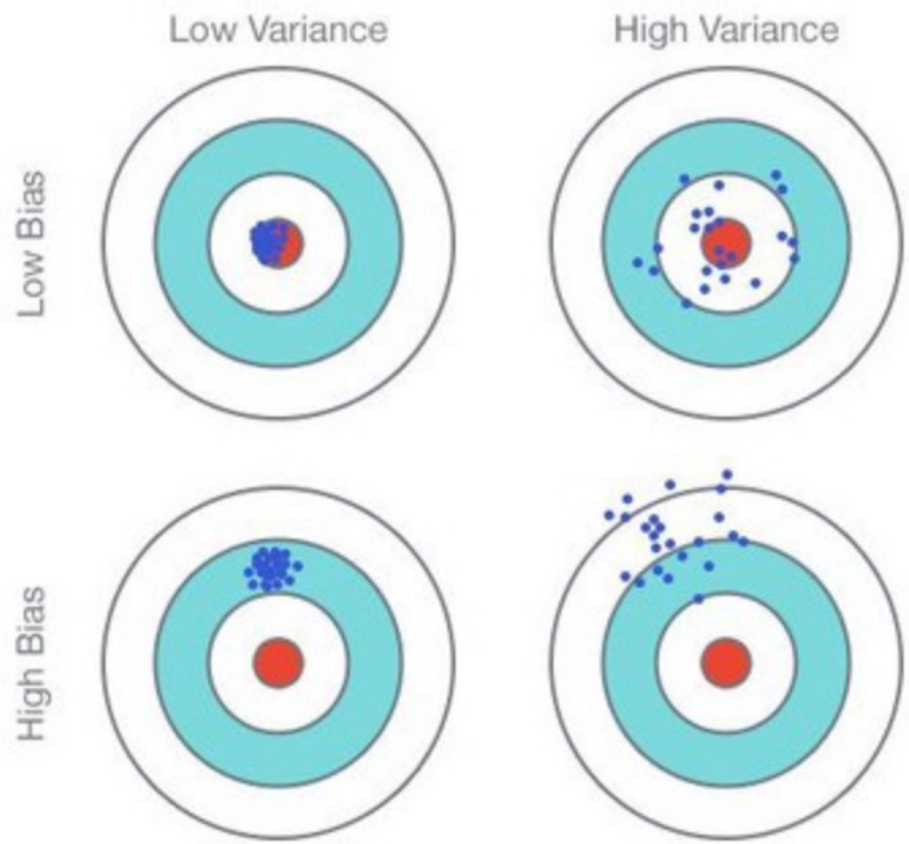


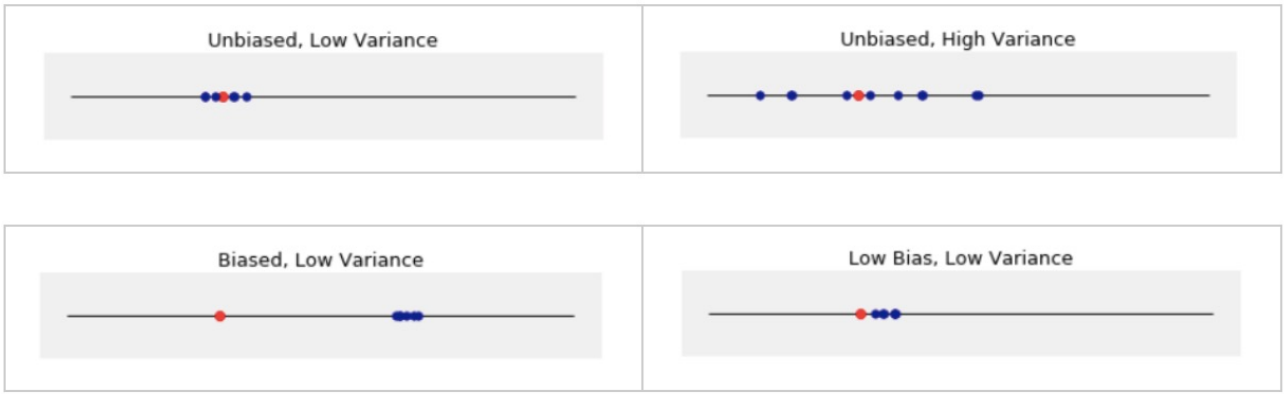
Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off

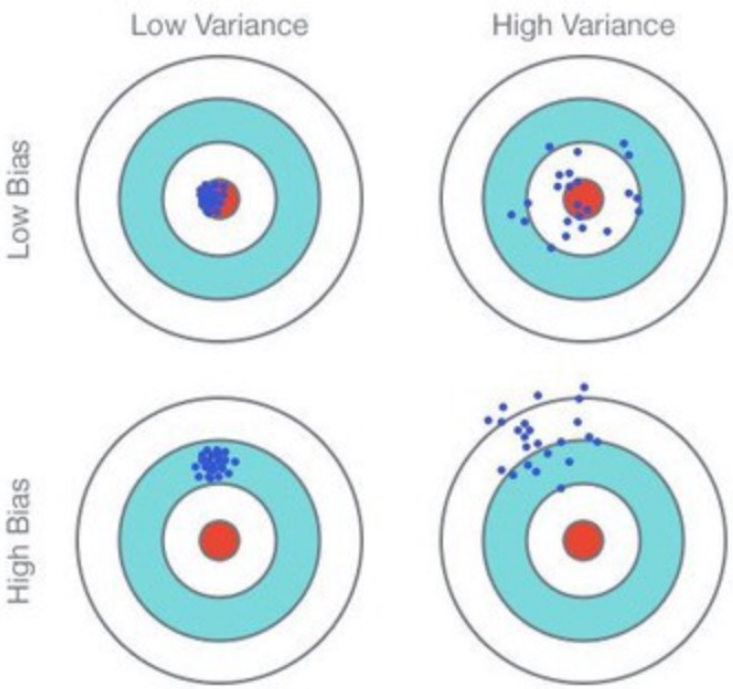
https://medium.com/@mp32445/understanding-bias-variance-tradeoff-ca59a22e2a83

Chapter 11

Bias, Variance, and Least Squares

Understanding Bias and Variance





T: estimator (rv)

θ: parameter (target, constant)

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

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

Bias, Variance, and Mean Squared Error

- 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?

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

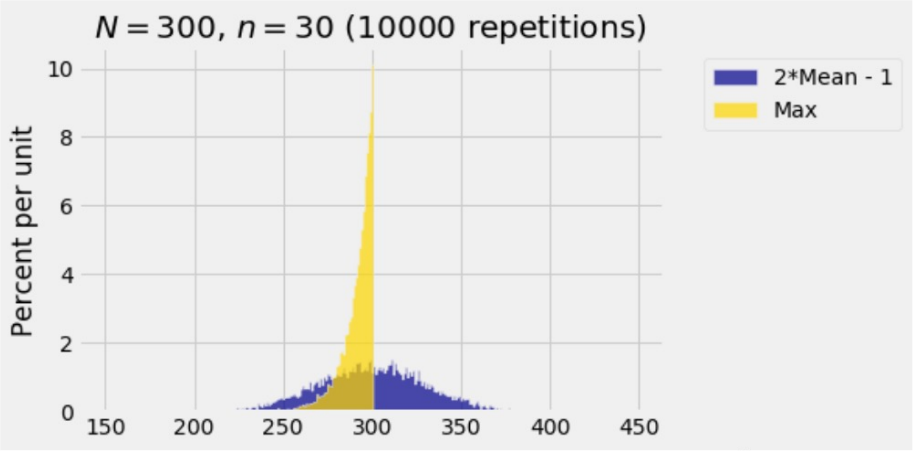
- Error: $\boldsymbol{T - \theta =}$
- Mean Squared Error: $\boldsymbol{MSE_\theta(T) = E[(T - \theta)^2]}$

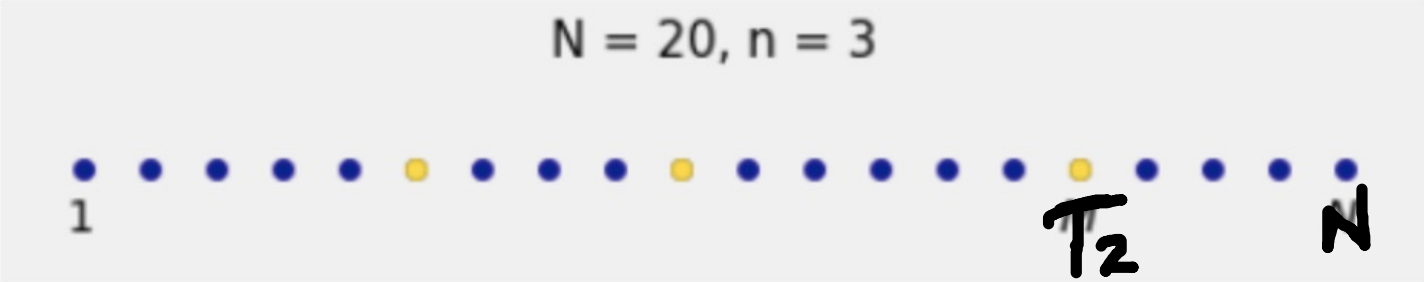
- What is the expected value of $D_\theta(T)$? What about $\big(D_\theta(T)\big)^2$?

Mean Squared Error & the Bias-Variance Decomposition

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

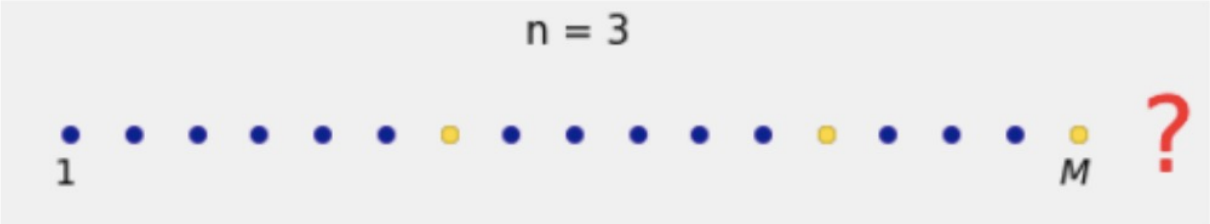
German Tank Problem: $T_1, T_2, \& T_3$





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

The Augmented Maximum

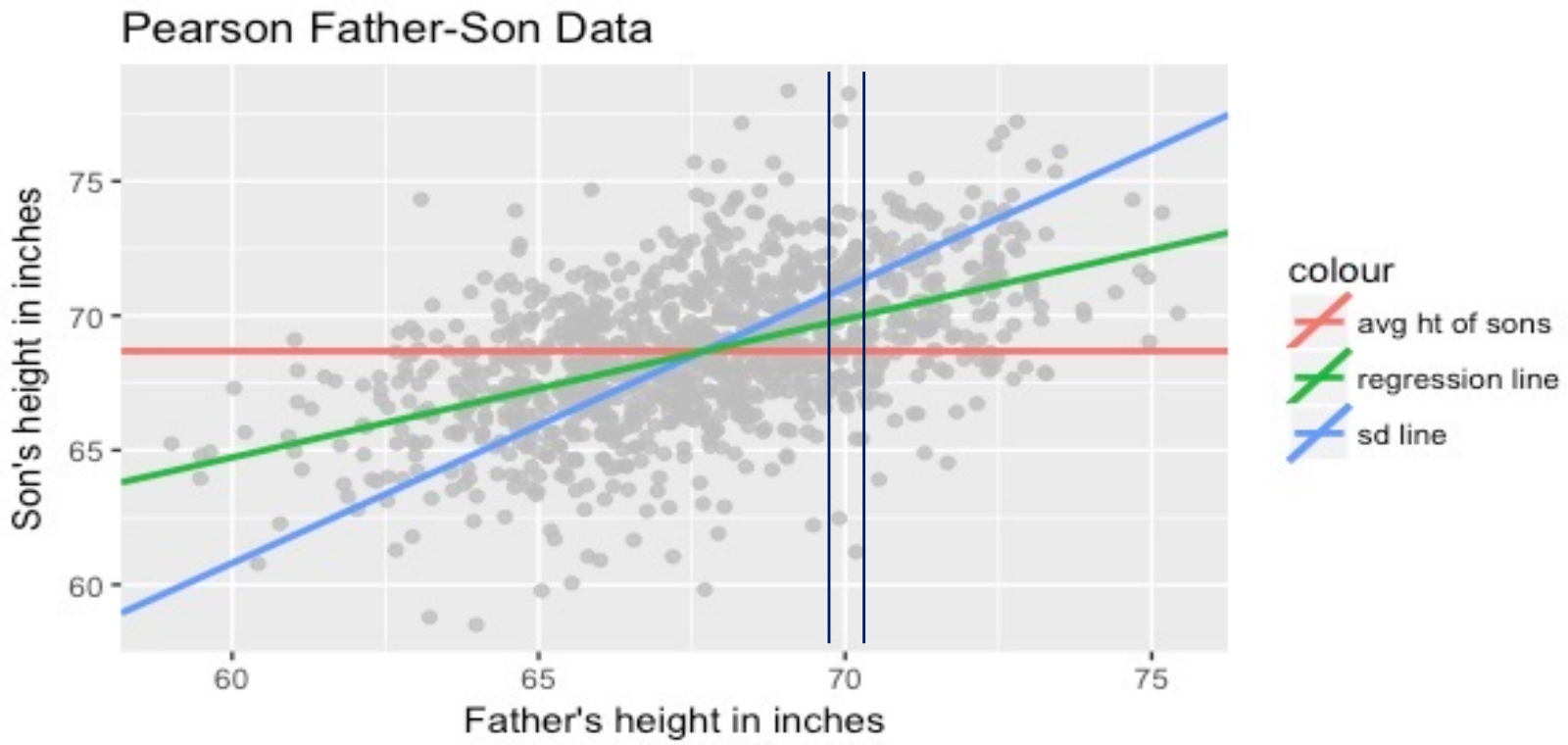


Simple Linear Regression (SLR)

- 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

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

- 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")



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)
- Much better, if the scatter plot shows a linear relationship, to use the **regression method**, which incorporates the correlation $\boldsymbol{r}$ (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.

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

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

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)

- 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

Mathematical derivation of the formulas for a and b

- As usual, $E(X) = \mu_X, SD(X) = \sigma_X; E(Y) = \mu_Y, SD(Y) = \sigma_Y$
- (X, Y) are our random variables, that we *think* are related by a linear function, perhaps with some error: $Y = aX + b + error$
- We want to estimate the equation of the line, that is, find $\hat{Y}$ such that $\hat{Y} = aX + b$

- Find the a and b by minimizing the mean square error, where error is the difference between our estimate $\hat{Y}$ and the original random variable Y.

- Notice that the mean squared error will be a function of a and b:

$$MSE(a, b) = E\left((Y - \hat{Y})^2\right) = E\left(\left(Y - (aX + b)\right)^2\right)$$

- First, we can look for the best intercept for some fixed slope:

Mathematical derivation of the formulas for a and b

- Looking for the best intercept for some fixed slope, that is, fix a, and then see, for this *given* value of a, what would be the b that minimizes the MSE?

- We can write out the MSE as a function of b, take the derivative, and set it equal to 0, and look for the best b.

Equation of the regression line

- $\hat{Y} = \hat{a}X + \hat{b}$
- $\hat{Y}$ is called the fitted value of Y, $\hat{a}$ is the slope, $\hat{b}$ is the intercept where:
- $\hat{a} = \frac{r\sigma_Y}{\sigma_X}, r = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = E(Z_X \times Z_Y)$
- $\hat{b} = \mu_Y - \hat{a}\,\mu_X$

Correlation

- The expected product of the deviations of X and Y, $E(D_X D_Y)$ is called the ***covariance*** of X and Y.
- The problem with using covariance is that the units are multiplied *and* the value depends on the units
- Can get rid of this problem by dividing each deviation by the SD of the corresponding SD, that is, put it in standard units. The resulting quantity is called the ***correlation coefficient*** of X and Y:
- $r(X, Y) =$

- Note that it is a pure number with no units, and now we will prove that it is always between -1 and 1.

Bounds on correlation

- $r = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = E(Z_X Z_Y)$
- (Note that this implies that $E(D_X D_Y) = r\sigma_X\sigma_Y$. We will use this later.)

Correlation as a measure of linear association

- $D = Y - \hat{Y},\ E(D) = 0,\ Var(D) = (1 - r^2)\sigma_Y^2$
- What if the correlation is very close to 1 or -1? What does this tell you about X & Y?

- What about if the correlation is close to 0? What does this tell you about X & Y?

Residual is uncorrelated with X

- What about $r(D, X), D = Y - \hat{Y}$?
- Intuitively, what should this be? Why?
- What should your residual (diagnostic) plot look like?