

LECTURE 15

Bias and Variance

Exploring the different sources of error in the predictions that our models make.

Data 100, Fall 2020 @ UC Berkeley

Anthony D. Joseph, Fernando Pérez
(content by Ani Adhikari & Suraj Rampure)

Random Variables and Expectations:

A quick recap

What is a random variable?

A random variable is a **numerical function of a random sample.**

Another name for a numerical function of a sample is a “statistic.”

We typically denote random variables with uppercase letters late in the alphabet (e.g. X , Y).

Why **random**? Because the sample on which it is a function was drawn at random.

Why **variable**? Because its value depends on how the sample came out.

Definition of expectation

The **expectation** of a random variable X is the weighted average of the values of X , where the weights are the probabilities of the values.

The most common formulation applies the weights one possible value at a time:

$$\mathbb{E}(X) = \sum_{\substack{\text{all possible} \\ x}} x \mathbb{P}(X = x)$$

However, an equivalent formulation applies the weights one sample at a time:

$$\mathbb{E}(X) = \sum_{\substack{\text{all samples} \\ s}} X(s) \mathbb{P}(s)$$

Linearity

Two of the properties we just established were

- Linear transformations apply to expectations.
- Expectation is additive.

Combining these gives us a single property, which is sometimes referred to as the **linearity of expectation**. For any random variables X , Y and constants a , b :

$$E(aX + bY) = aE(X) + bE(Y)$$

This more general form won't appear often in this class, but it is good to be aware of.

Summary

- Random variables are functions of our sample.
- The expectation of a random variable is the weighted average of its possible values, weighted by their probabilities.
 - Expectation behaves nicely with linear transformations of random variables.
 - Expectation is also additive.

Variance and Covariance: Random Variables Revisited

Definition of variance

- Variance is the **expected squared deviation from the expectation** of X .
- It is defined as follows:

$$\mathbb{V}ar(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

- The units of the variance are the square of the units of X .
- To get back to the right scale, we look at the **standard deviation** of X :

$$\mathbb{SD}(X) = \sqrt{\mathbb{V}ar(X)} = \sqrt{\mathbb{E}((X - \mathbb{E}(X))^2)}$$

- Both standard deviation and variance must be non-negative.

Interpretation of variance

- The main use of variance is to **quantify chance error**.
 - How far away from the expectation can X be, just by chance?
- By Chebyshev's inequality from Data 8:
 - No matter what the shape of the distribution of X is,
 - The vast majority of the probability lies in the interval “expectation plus or minus a few SDs”.
 - Specifically, if $\mu = E[X]$ and $\sigma = \text{SD}[X]$, then $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$.
 - We will not be using this formula in this class; it's just here to remind you of it.
 - Here's a [link to a discussion in the Data 8 book](#) about this.

An alternative calculation

There's a more convenient form of variance for use in calculations.

$$\mathbb{V}ar(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

To derive this, we make repeated use of the linearity of expectation. (A more detailed walkthrough is in the lecture video.)

$$\begin{aligned}\mathbb{V}ar(X) &= E((X - E(X))^2) \\ &= E(X^2 - 2XE(X) + (E(X))^2) \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 \\ &= E(X^2) - (E(X))^2\end{aligned}$$

An alternative calculation

$$\mathbb{V}ar(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

For example, to compute the variance of one roll of a die, we can find

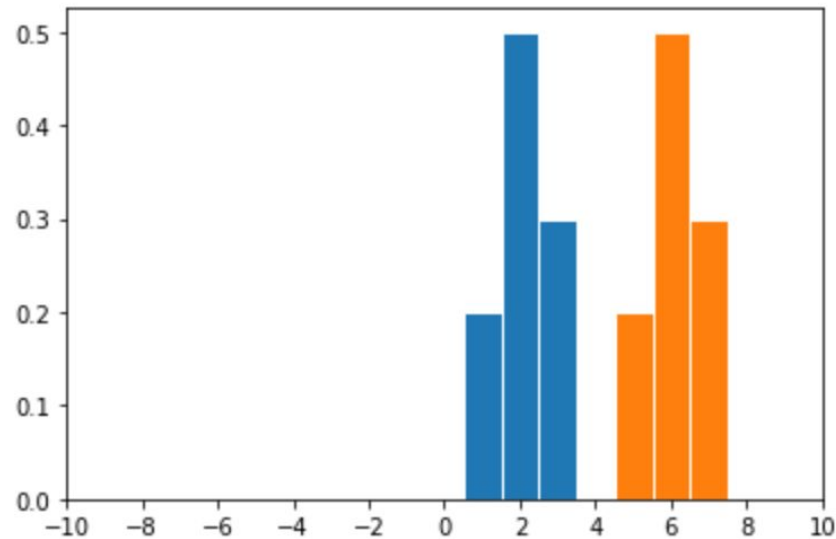
$$\mathbb{V}ar(X) = (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) \cdot \frac{1}{6} - (3.5)^2 = 2.92$$

- This formulation also makes clear that if X is **centered**, i.e. $\mathbb{E}(X) = 0$, then $\mathbb{V}ar(X) = \mathbb{E}(X^2)$.
- Since $\mathbb{V}ar(X)$ is non-negative, this property also shows us that $\mathbb{E}(X^2) \geq (\mathbb{E}(X))^2$.
Equality is if and only if X is a constant.
- If you know the expectation and variance of a random variable, you can easily determine the expectation of its square: $\mathbb{E}(X^2) = \mathbb{V}ar(X) + (\mathbb{E}(X))^2$.

Linear transformations

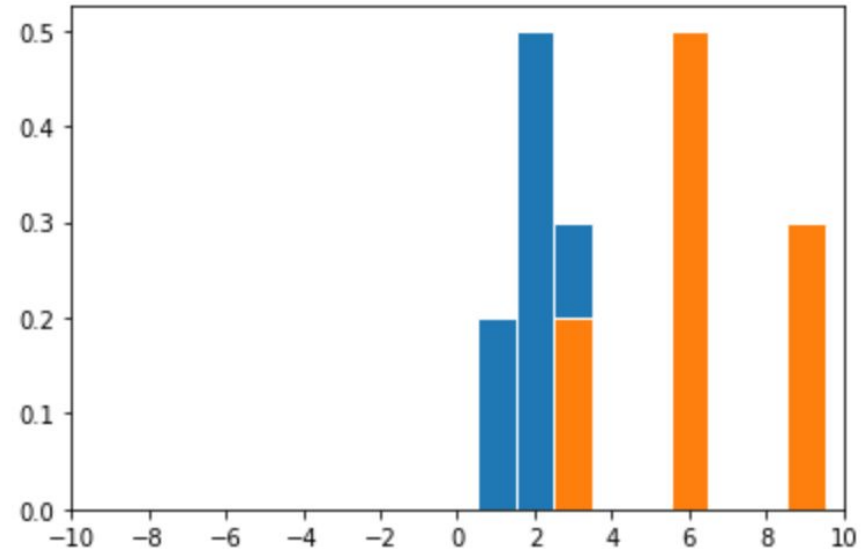
We know that $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$. In order to compute $\text{Var}(aX + b)$, consider:

A shift by **b** units **does not** affect spread:



Here, the distribution of X is in blue, and the distribution of $X+4$ is in orange.

But scaling by **a** units **does** affect spread:



The distribution of X is in blue, and the distribution of $3X$ is in orange.

Linear transformations

We know that $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.

In order to compute $\text{Var}(aX + b)$, consider:

- A shift by **b** units **does not** affect spread. Thus, $\text{Var}(aX + b) = \text{Var}(aX)$.
- The multiplication by **a** **does** affect spread!

Then,

$$\begin{aligned}\text{Var}(aX + b) &= \text{Var}(aX) = E((aX)^2) - (E(aX))^2 \\ &= E(a^2 X^2) - (aE(X))^2 \\ &= a^2 (E(X^2) - (E(X))^2) \\ &= a^2 \text{Var}(X)\end{aligned}$$

In summary:

$$\begin{aligned}\text{Var}(aX + b) &= a^2 \text{Var}(X) \\ \text{SD}(aX + b) &= |a| \text{SD}(X)\end{aligned}$$

Don't forget the absolute values and squares!

Standardization of random variables

X in **standard units** is the random variable $X_{su} = \frac{X - \mathbb{E}(X)}{\text{SD}(X)}$.

- X_{su} measures X on the scale “**number of SDs from expectation.**”
- It is a linear transformation of X . By the linear transformation rules for expectation and variance:

$$\mathbb{E}(X_{su}) = 0, \quad \text{SD}(X_{su}) = 1$$

- Since X_{su} is centered (has expectation 0):

$$\mathbb{E}(X_{su}^2) = \text{Var}(X_{su}) = 1$$

You should prove these facts yourself.

Variance of a sum, Covariance

Recap

Thus far, we've established the following:

- Linear transformations of random variables apply to their expectation.

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

- Expectation is additive.

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

- Linear transformations of random variables transform their variance as follows:

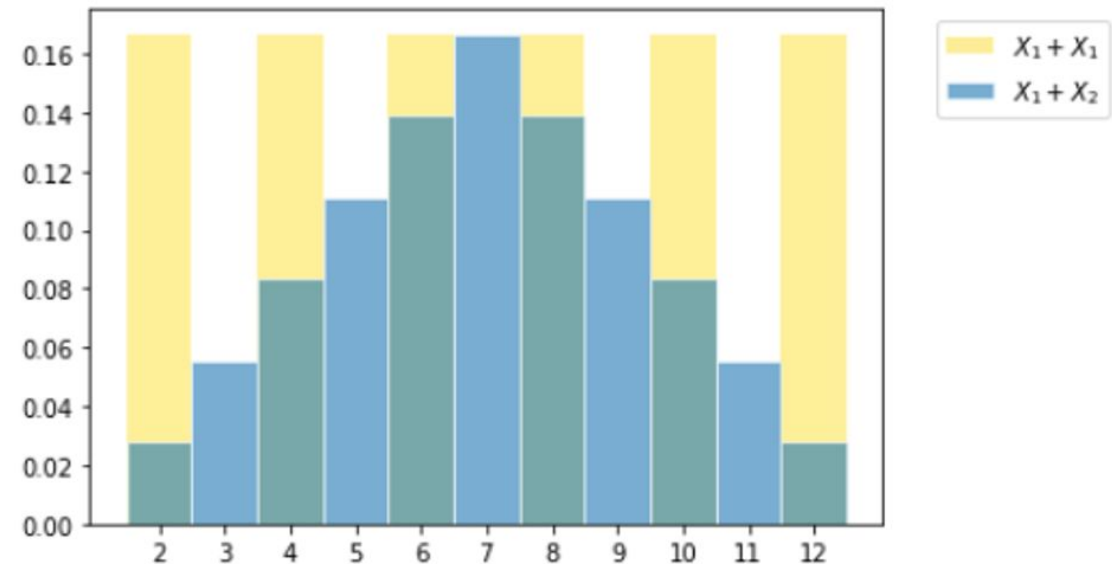
$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

We haven't yet talked about the variance of a sum of random variables. Let's now do that!

Distributions of sums

Suppose X_1 and X_2 are the numbers on two rolls of a die.

- **X_1 and X_2 have the same distribution** (they are both Uniform in $\{1, 2, 3, 4, 5, 6\}$).
- But the **distributions of $X_1 + X_1$ and $X_1 + X_2$ are different.**
- Both $X_1 + X_1$ and $X_1 + X_2$ have the same expectation (7).
- But $X_1 + X_2$ seems to have less spread, indicating that $X_1 + X_1 = 2X_1$ has a larger variance.



Variance of a sum

The variance of a sum is affected by the dependence between the two random variables that are being added. Let's expand out the definition of $\text{Var}(X + Y)$ to see what's going on.

Let $\mu_x = E[X]$, $\mu_y = E[Y]$.

$$\begin{aligned}\text{Var}(X + Y) &= E[(X + Y - E(X + Y))^2] \\ &= E[((X - \mu_x) + (Y - \mu_y))^2] \\ &= E[(X - \mu_x)^2 + 2(X - \mu_x)(Y - \mu_y) + (Y - \mu_y)^2] \\ &= E[(X - \mu_x)^2] + E[(Y - \mu_y)^2] + 2E[(X - \mu_x)(Y - \mu_y)] \\ &= \text{Var}(X) + \text{Var}(Y) + 2\underbrace{E[(X - E(X))(Y - E(Y))]}_{\text{Cov}(X, Y)}\end{aligned}$$

By the linearity of expectation, and the substitution.

We see that the variance of a sum is equal to the sum of variances, PLUS this weird term...

Covariance

The covariance of two random variables is their **expected product of deviations**.

$$\mathbb{C}ov(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

- It is a generalization of variance. Note: $\mathbb{C}ov(X, X) = \mathbb{V}ar(X)$.
- Using the linearity of expectation and some algebra, you can show the following equality, which is a generalization of the alternative calculation for variance:

$$\mathbb{C}ov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

- To see whether variance is ever additive, we need to look at covariance differently.

When is variance additive?

For any two random variables X and Y :

$$\mathbb{V}ar(X + Y) = \mathbb{V}ar(X) + \mathbb{V}ar(Y) + 2\mathbb{C}ov(X, Y)$$

In order for variance to be additive, the covariance between X and Y needs to be 0.

$$\mathbb{V}ar(X + Y) = \mathbb{V}ar(X) + \mathbb{V}ar(Y) \iff \mathbb{C}ov(X, Y) = 0$$

When is the covariance between two random variables 0?

- A sufficient condition is that X and Y are **independent**. If X and Y are independent, knowing the value of X tells you nothing about the value of Y . Independence is a **strong statement**.
- This is not the only case when the covariance is 0, as we will shortly see.

Sometimes called the **addition rule for variance**:

$$\mathbb{V}ar(X + Y) = \mathbb{V}ar(X) + \mathbb{V}ar(Y) \text{ if } X \text{ and } Y \text{ are independent}$$

Correlation

The units of the covariance are hard to interpret (e.g. “inch pounds”). In order to get rid of the units, we can scale it:

$$\begin{aligned}\frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)} &= \frac{\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))}{\text{SD}(X)\text{SD}(Y)} \\ &= \mathbb{E}\left(\frac{X - \mathbb{E}(X)}{\text{SD}(X)} \cdot \frac{Y - \mathbb{E}(Y)}{\text{SD}(Y)}\right) \\ &= \mathbb{E}(X_{su}Y_{su}) \\ &= r(X, Y)\end{aligned}$$

Recall from Data 8:
correlation is the average
product in standard units.
This is the random variable
equivalent of that!

Correlation is covariance scaled by the two SDs.

Uncorrelated random variables

The correlation between X and Y is

$$r(X, Y) = \frac{\mathbb{C}ov(X, Y)}{\mathbb{SD}(X)\mathbb{SD}(Y)}$$

This means that either both correlation and covariance are 0, or neither are.

$$\mathbb{C}ov(X, Y) = 0 \iff r(X, Y) = 0$$

- “Covariance equal to 0” is the same as “**uncorrelated**”.
- **Independent random variables are uncorrelated.**
- But not all uncorrelated random variables are independent!
 - For instance: Let X be uniform on $\{-1, 1\}$. X and X^2 are uncorrelated, but dependent.

More properties of sums

Addition rule for variance

If X and Y are **uncorrelated** (in particular, if they are **independent**), then

$$\mathbb{V}ar(X + Y) = \mathbb{V}ar(X) + \mathbb{V}ar(Y)$$

Therefore, under the same conditions,

$$\mathbb{SD}(X + Y) = \sqrt{\mathbb{V}ar(X) + \mathbb{V}ar(Y)} = \sqrt{(\mathbb{SD}(X))^2 + (\mathbb{SD}(Y))^2}$$

- Think of this as “Pythagorean theorem” for random variables.
- Uncorrelated random variables are like orthogonal vectors.



I.I.D. sample sum

die example

- “i.i.d.” is short for “independent and identically distributed”.
- Draws at random with replacement from a population are i.i.d.
- Let the sample X_1, X_2, \dots, X_n be i.i.d. draws from a numerical population that has mean μ and SD σ .
- Let the sample sum be $S_n = \sum_{i=1}^n X_i$.

Then, $\mathbb{E}(S_n) = n\mu$, $\text{Var}(S_n) = n\sigma^2$, $\text{SD}(S_n) = \sqrt{n}\sigma$

SD is not additive, even when each RV in the sum is independent.

An example revisited

Suppose X_1 and X_2 are the numbers on two rolls of a die. X_1 and X_2 have the same expectation and variance:

$$\mathbb{E}(X_1) = 3.5 = \mathbb{E}(X_2) \quad \text{SD}(X_1) = 1.71 = \text{SD}(X_2)$$

$X_1 + X_1$ and $X_1 + X_2$ have the same expectation:

$$\mathbb{E}(X_1 + X_1) = 2 \times 3.5 = 7 = \mathbb{E}(X_1 + X_2)$$

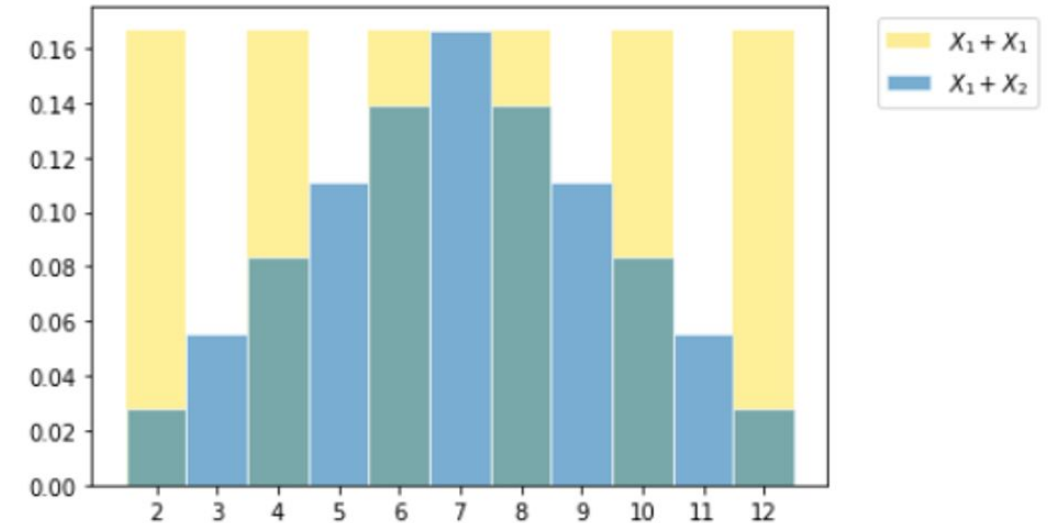
But their variances (and hence SDs) are different:

$$\text{SD}(X_1 + X_1) = \text{SD}(2X_1) = 2 \times 1.71 = 3.42$$

$$\text{SD}(X_1 + X_2) = \sqrt{2} \times 1.71 = 2.42$$

Since X_1 and X_2 are independent, we can use the result from the previous slide.

As we reasoned about earlier, the spread of $X_1 + X_2$ is less than the spread of $X_1 + X_1$.



Variance of the Bernoulli distribution

En-route to computing the variance of the binomial distribution, let's first compute the variance of the Bernoulli (p) distribution. We can do this using the alternate calculation for variance.

$$E(X) = p$$

$$E(X^2) = 1^2 \cdot p + 0^2 \cdot (1 - p) = p$$

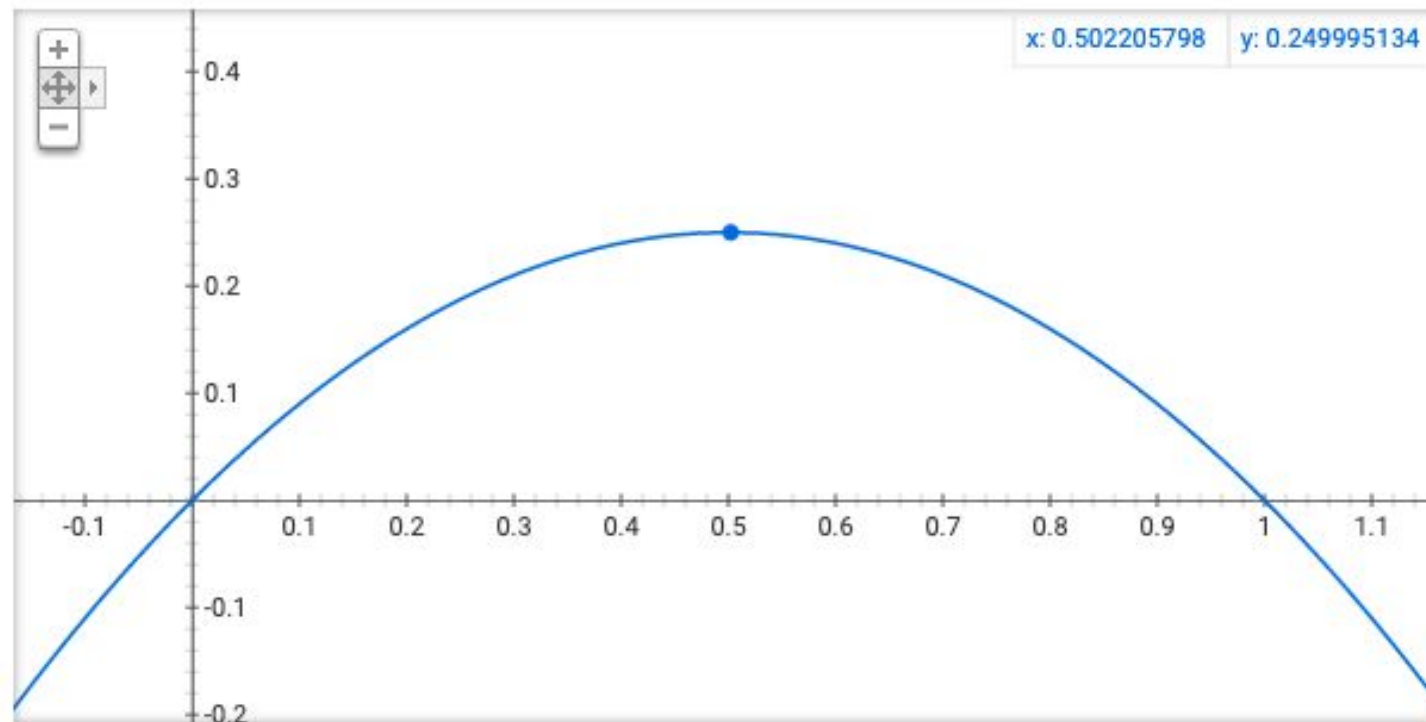
Putting these together:

$$Var(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1 - p)$$

A quick bit of intuition - when is $\text{Var}(\text{Ber}(p))$ maximal?

$$\text{Var}(X) = p - p^2 = p(1 - p)$$

Graph for $x-x^2$



ps - tip, plot above: google(plot x - x^2)

Variance of the binomial distribution

Let X have the **binomial** (n, p) distribution. We know that X is the number of “successes” in n independent trials of some event, each of which occur with probability p .

- Each trial can be thought of as a single Bernoulli (p) trial.
- We can then write:

$$X = I_1 + I_2 + \cdots + I_n$$

where I_j is the **indicator** of success on trial j . $I_j = 1$ if trial j is a success, and 0 else.

- For each j , $\mathbb{E}(I_j) = p$, $\text{Var}(I_j) = p(1 - p)$
- As established before, $\mathbb{E}(X) = np$.
- Using the fact that each indicator is independent:

$$\text{Var}(X) = np(1 - p), \quad \text{SD}(X) = \sqrt{np(1 - p)}$$

Distribution of sample means

Sample mean

Earlier, we looked at the expectation and SD of the sample sum. Let's explore the sample mean.

- Consider an i.i.d. sample X_1, X_2, \dots, X_n .
- For each i , $E(X_i) = \mu$, $SD(X_i) = \sigma$.
- Define:

$$S_n = \sum_{i=1}^n X_i, \quad \bar{X}_n = \frac{1}{n} S_n$$

- The sample mean is a linear transformation of the sample sum (scaled by $1/n$).
 - By linear transformation rules:

$$\mathbb{E}(\bar{X}_n) = \mu, \quad \text{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

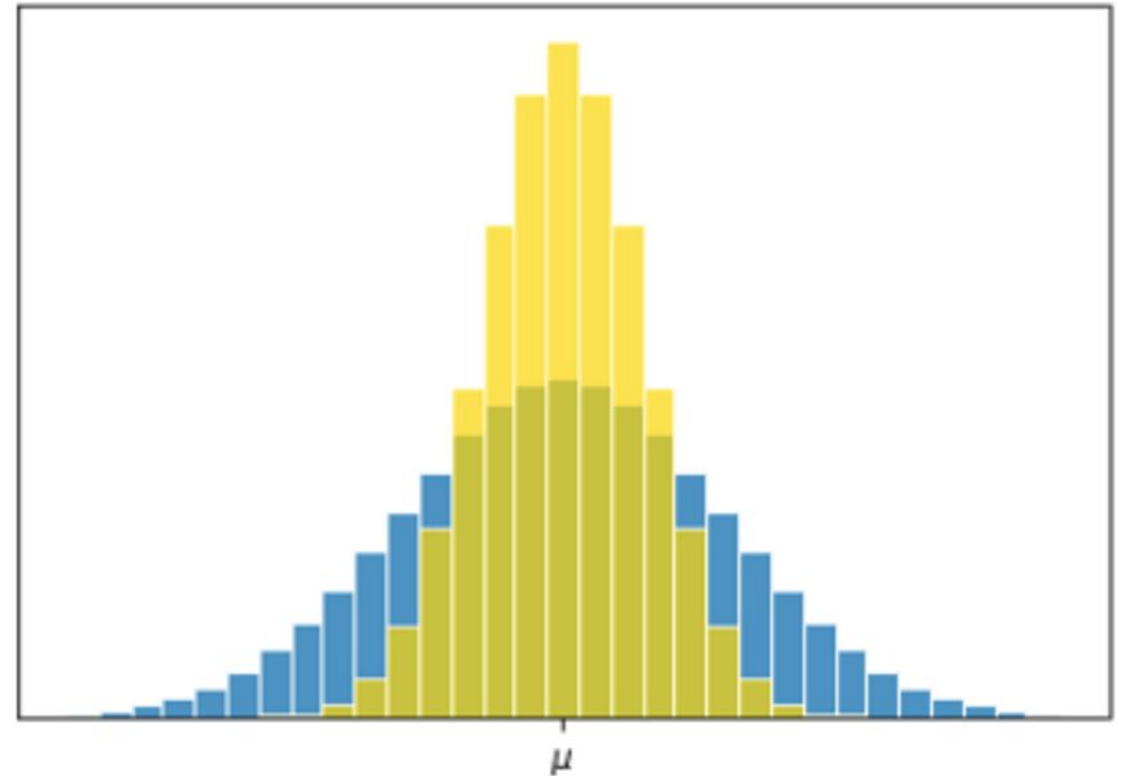
Accuracy of the sample mean

- Our goal will often be to estimate some characteristic of a population.
 - For instance, the average height of undergrad students at UC Berkeley.
 - To do this, we typically go out and collect a single sample. It has just one average.
 - Since that sample was random, it *could have* come out differently. As such, we need to look at the distribution of all possible sample means.
- For any sample size, the expected value of the sample mean is the population mean.
 - $\mathbb{E}(\bar{X}_n) = \mu$.
 - We call the sample mean an unbiased estimator of the population mean.
 - The term “unbiased” will reappear later in the semester.

Shape of the distribution

- As the sample size increases, the SD of the sample mean decreases.
 - The sample mean is more likely to be close to the population mean if we have a larger sample size.
- **Square root law:** If you increase the sample size by a factor, the SD decreases by the square root of the factor.
 - This is from Data 8. But we've just proven it, using the properties of variance!

$$\text{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$



Here are the distributions of two sample means. Both are drawn from the same population, but with different sample sizes.

Central limit theorem

The **central limit theorem (CLT)** states that no matter what population you are drawing from, the probability distribution of **the sum of an i.i.d. sample is roughly normal** if the sample size is large.

- Since the sample mean is a linear transformation of the sample sum, the sample mean is also roughly normal (with scaled parameters).

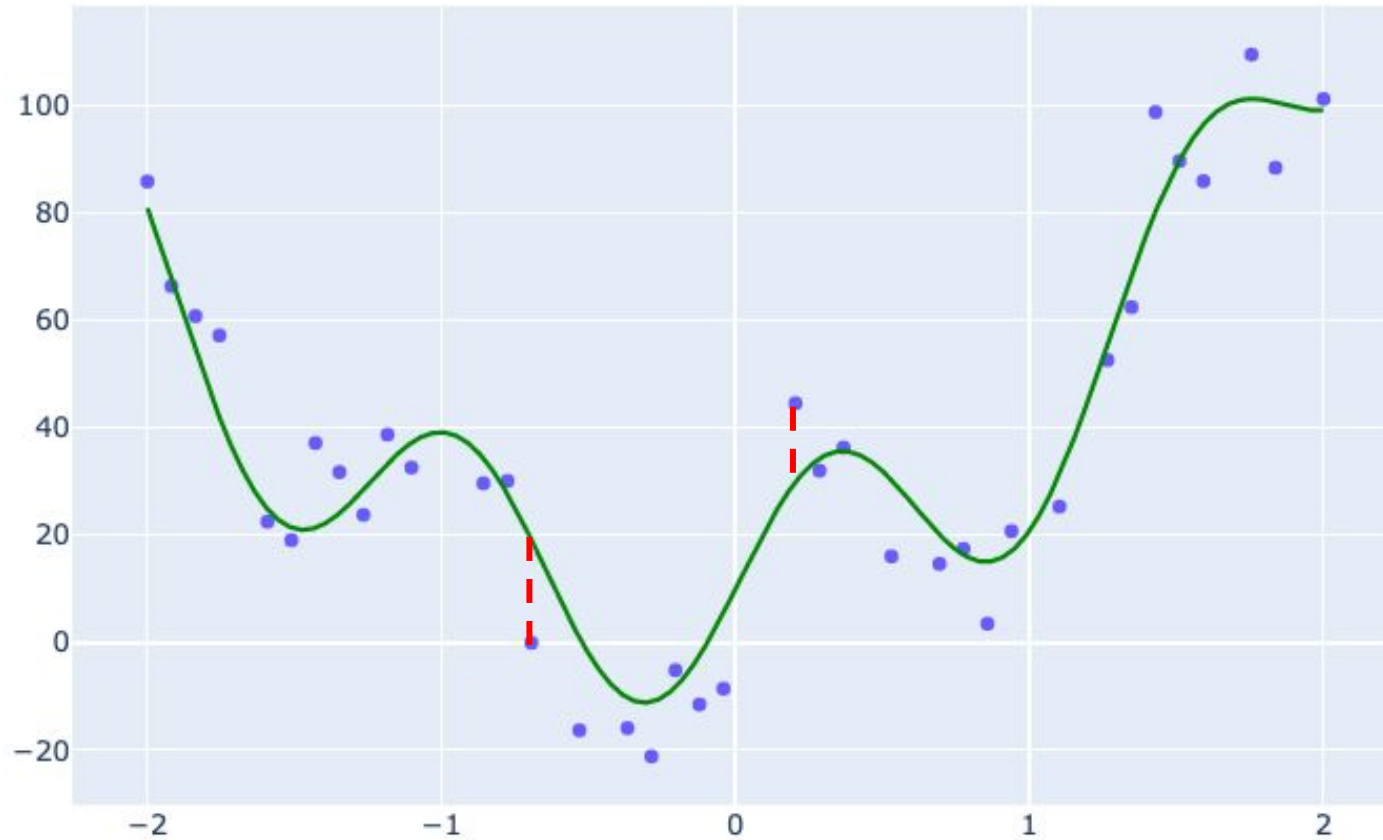
Bias and Variance in Modeling

Questions

- **How** can we describe the randomness in our data generation process?
- **What** does it mean for a new individual to be “similar” to those for whom we already have data?
- **What** are the main sources of error in prediction?
- **Why** do these errors occur, and how can we reduce them?
- **How** do all the different errors affect our overall risk?

Assumptions of Randomness

Data Generation Process



- True relation g
- For each individual,
 - fixed value of x and hence also $g(x)$
 - Random error ϵ
 - Observation is $Y = g(x) + \epsilon$

Errors have expectation 0 and are i.i.d. across individuals

The Data

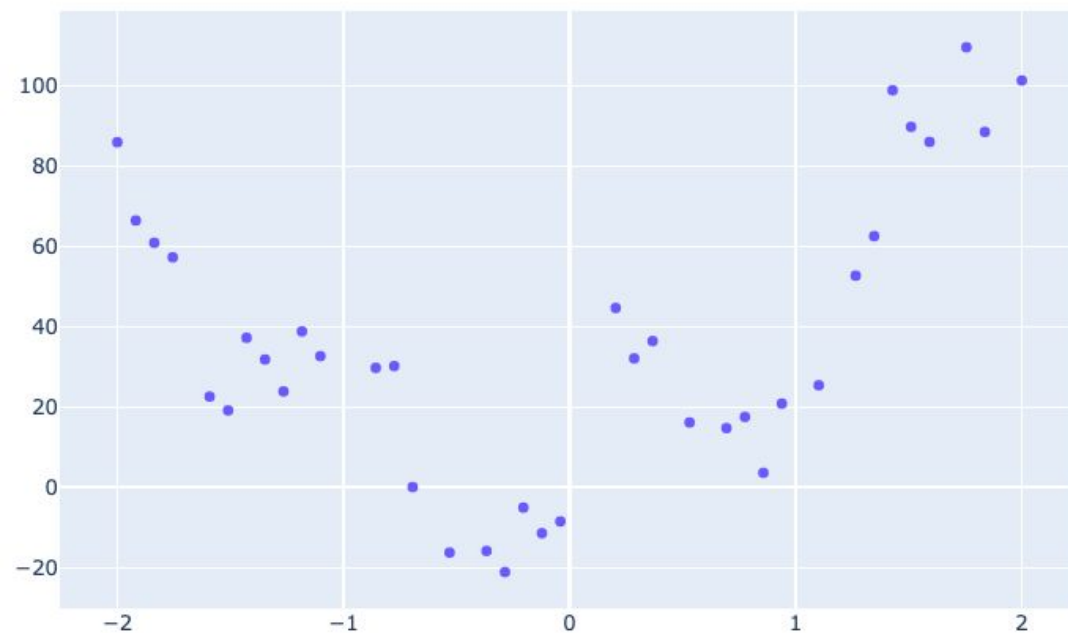
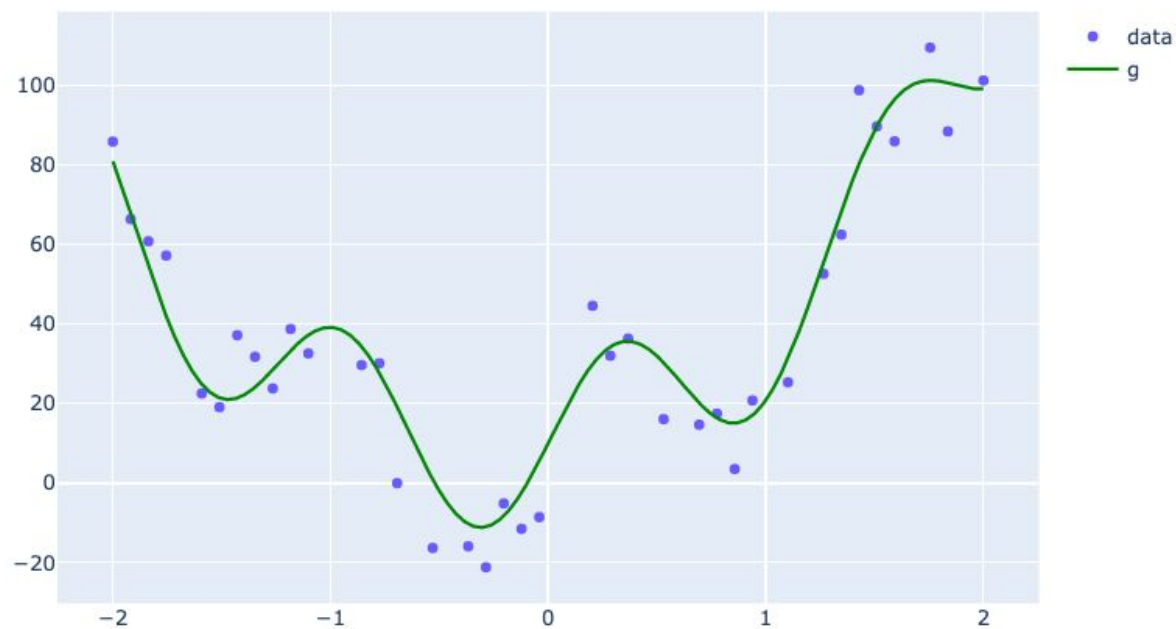
At each x ,

➤ truth = $g(x)$

➤ noise = ϵ

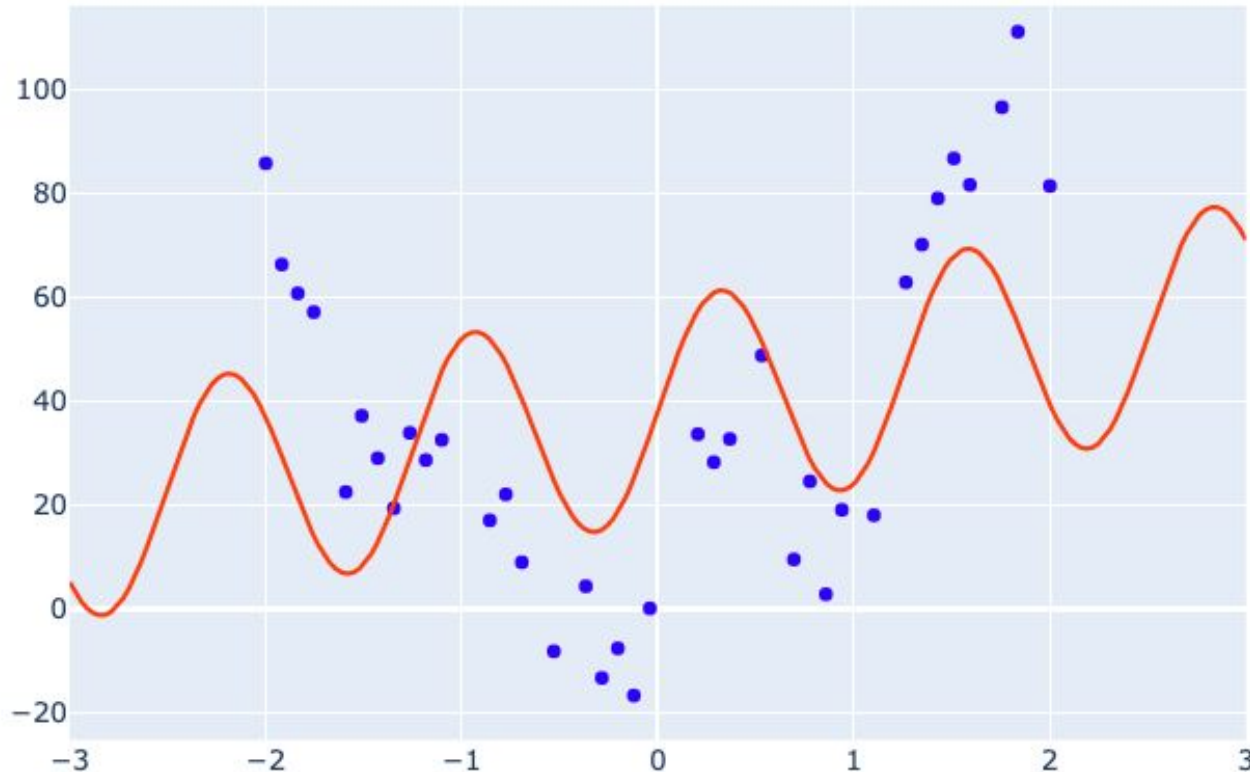
➤ Observation $Y = g(x) + \epsilon$

We only see Y



Our Predictions

We choose a model and fit it to our data.
The red line is our fitted function.



At every x , our prediction for Y is

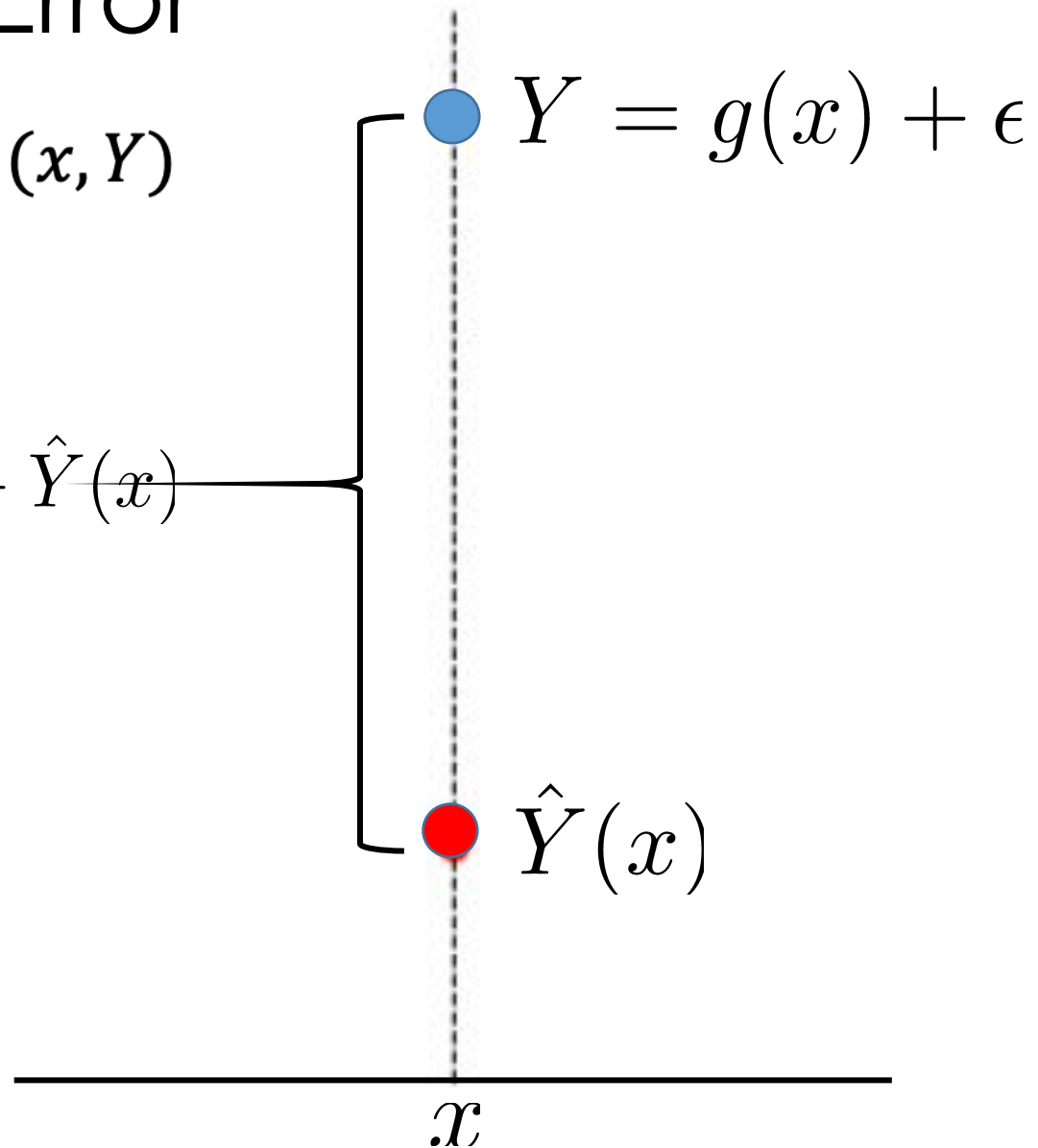
- The height of the red line at x
- Denote this $\hat{Y}(x)$

Measuring Prediction Error

Prediction Error

New individual: (x, Y)

$$\text{error} = Y - \hat{Y}(x)$$



Model Risk

- For a new individual at (x, Y) :
- Mean squared error of prediction

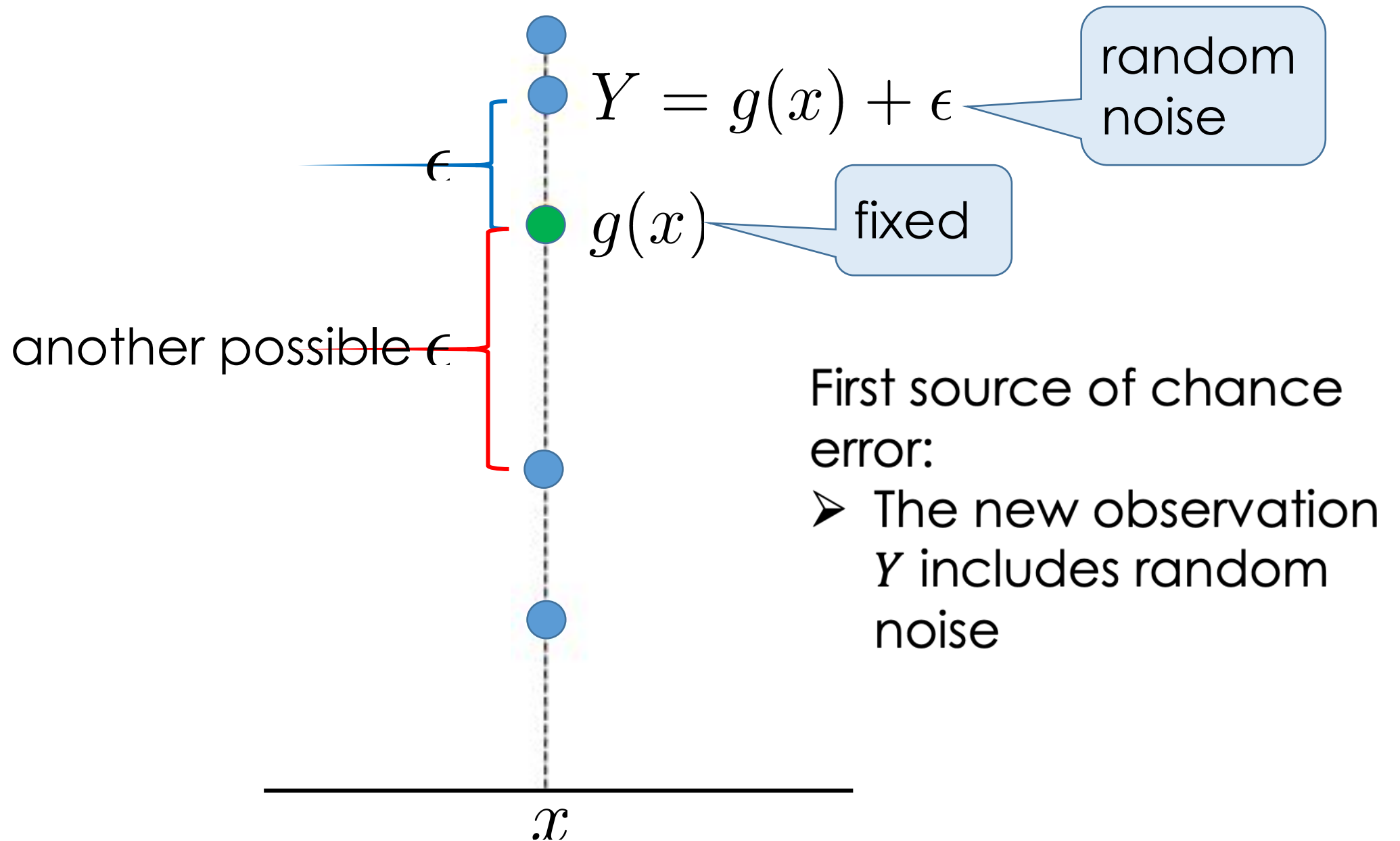
$$\text{model risk} = \mathbb{E}((Y - \hat{Y}(x))^2)$$

- The expectation is an average over all samples:
 - all possible samples we could have got for fitting our model
 - all possible new observations at the fixed x

Two Kinds of Error

- Chance error:
 - Due to randomness alone
 - In the new observations
 - Also in the sample we used for fitting our model
- Bias:
 - Non-random error
 - Due to our model being different from the true underlying function g

Chance Error in the New Observation



Observation Variance

- The new observation is Y
- $Y = g(x) + \epsilon$
- ϵ is random $\mathbb{E}(\epsilon) = 0$ $\mathbb{V}ar(\epsilon) = \sigma^2$
 $\mathbb{V}ar(Y) = \mathbb{V}ar(g(x) + \epsilon) = \mathbb{V}ar(\epsilon) = \sigma^2$

observation variance = σ^2

Reasons and Remedies

Some reasons:

- Measurement error
- Missing information acting like noise

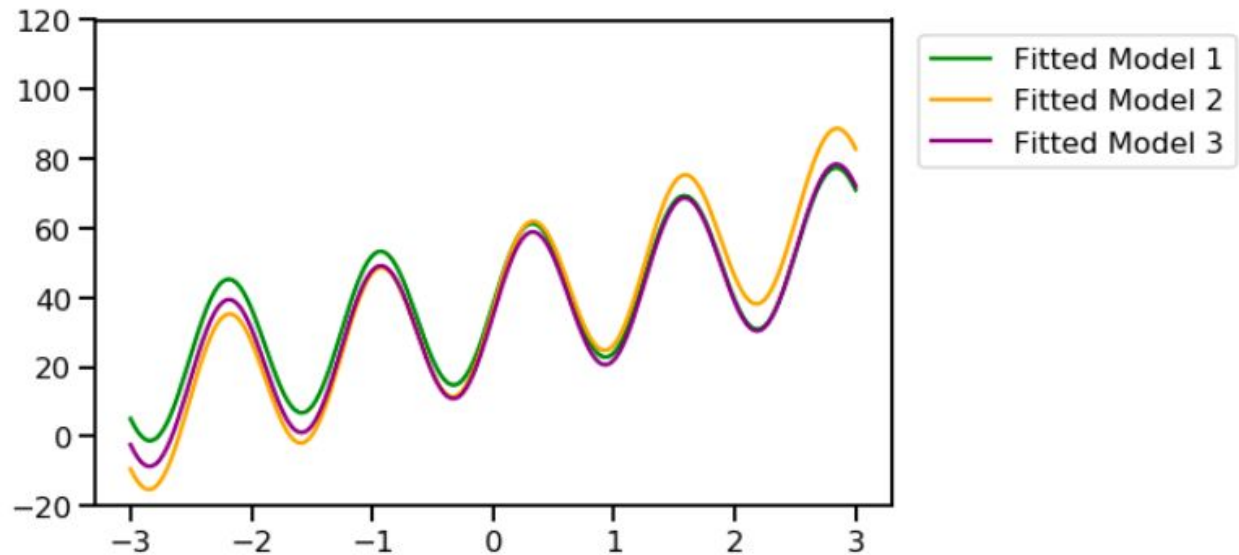
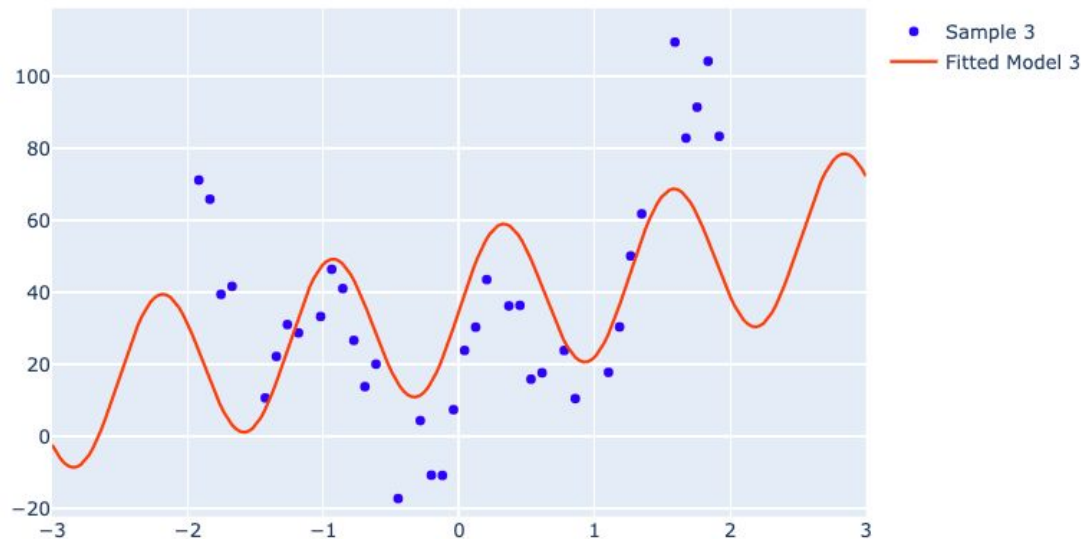
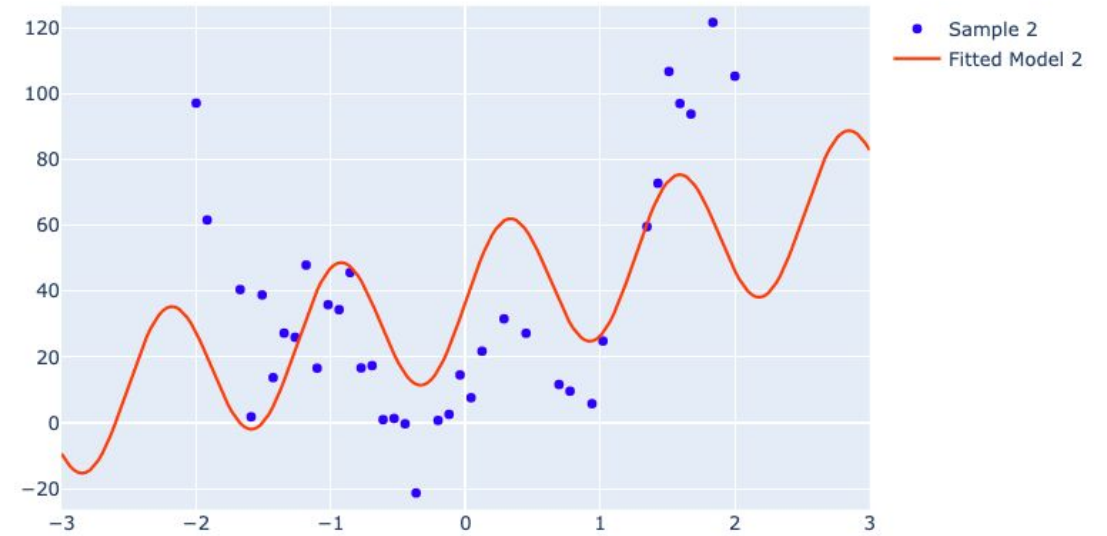
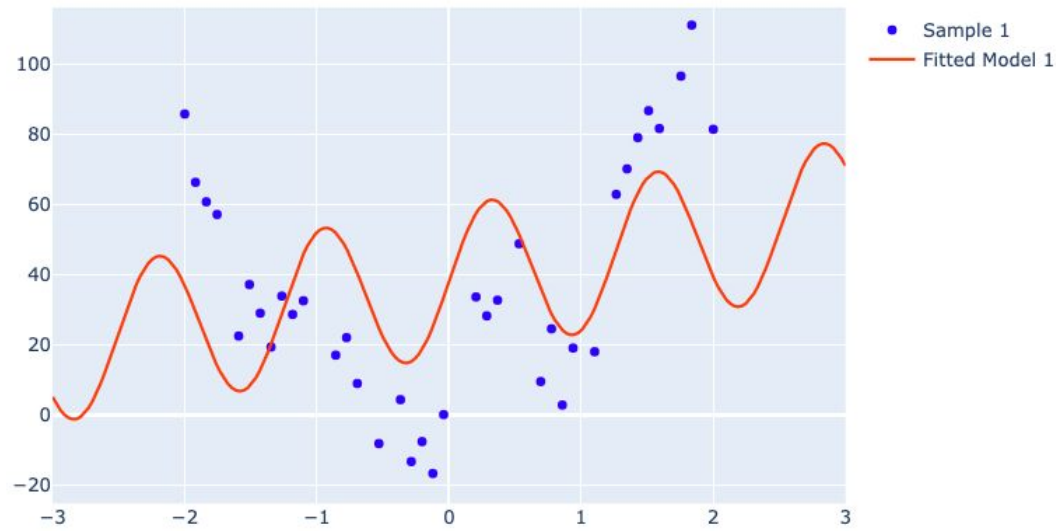
Could try to get more precise measurements

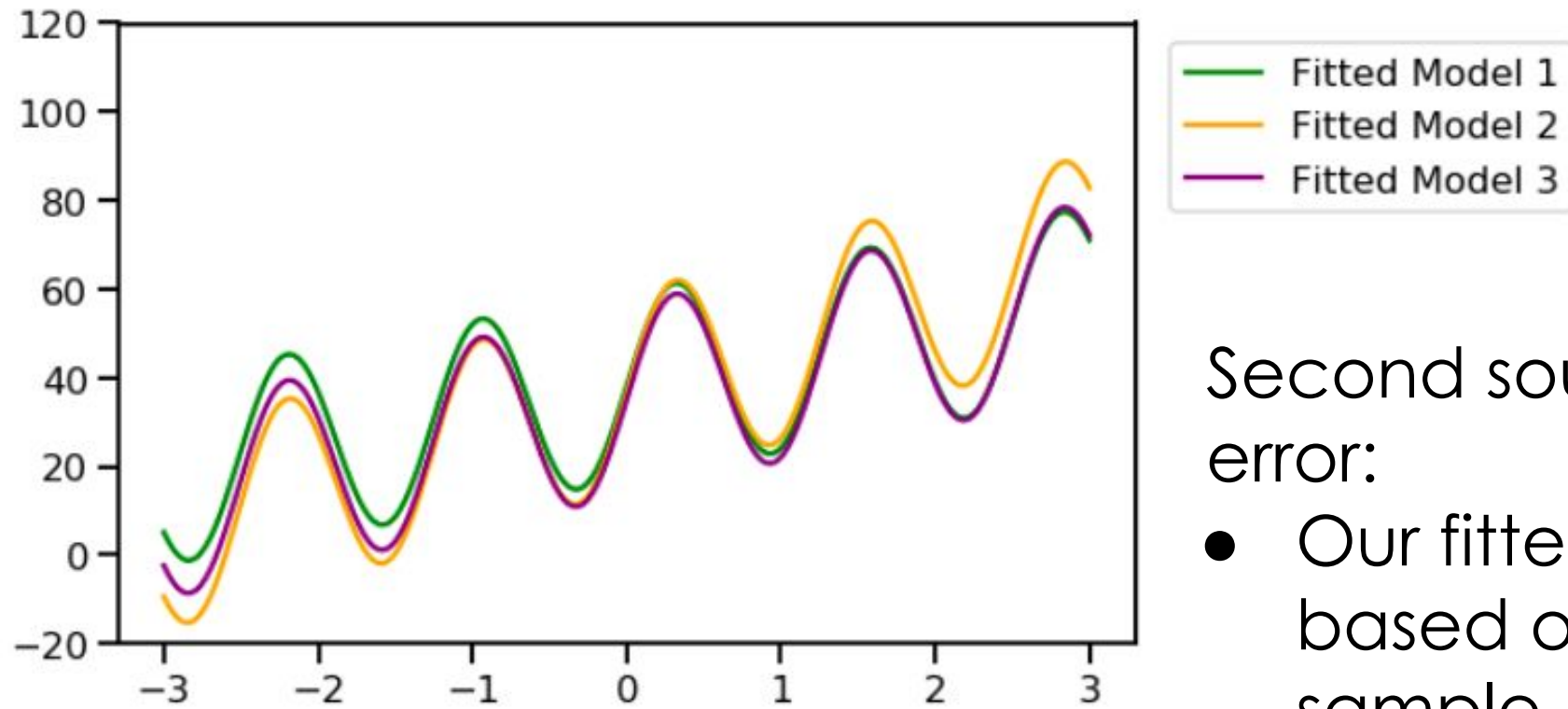
Often this is beyond the control of the data scientist.

Chance Error in Our Fitted Model

Model Variability

\hat{Y} depends on the sample





Second source of chance error:

- Our fitted model is based on a random sample
- The sample could have come out differently
- Then the fitted model would have been different

Model Variance

- Our prediction at x is $\hat{Y}(x)$
- The average of these predictions across all possible samples is $\mathbb{E}(\hat{Y}(x))$
- The variance of our prediction is

$$\text{model variance} = \text{Var}(\hat{Y}(x)) = \mathbb{E}((\hat{Y}(x) - \mathbb{E}(\hat{Y}(x)))^2)$$

Reasons and Remedies

Main reason:

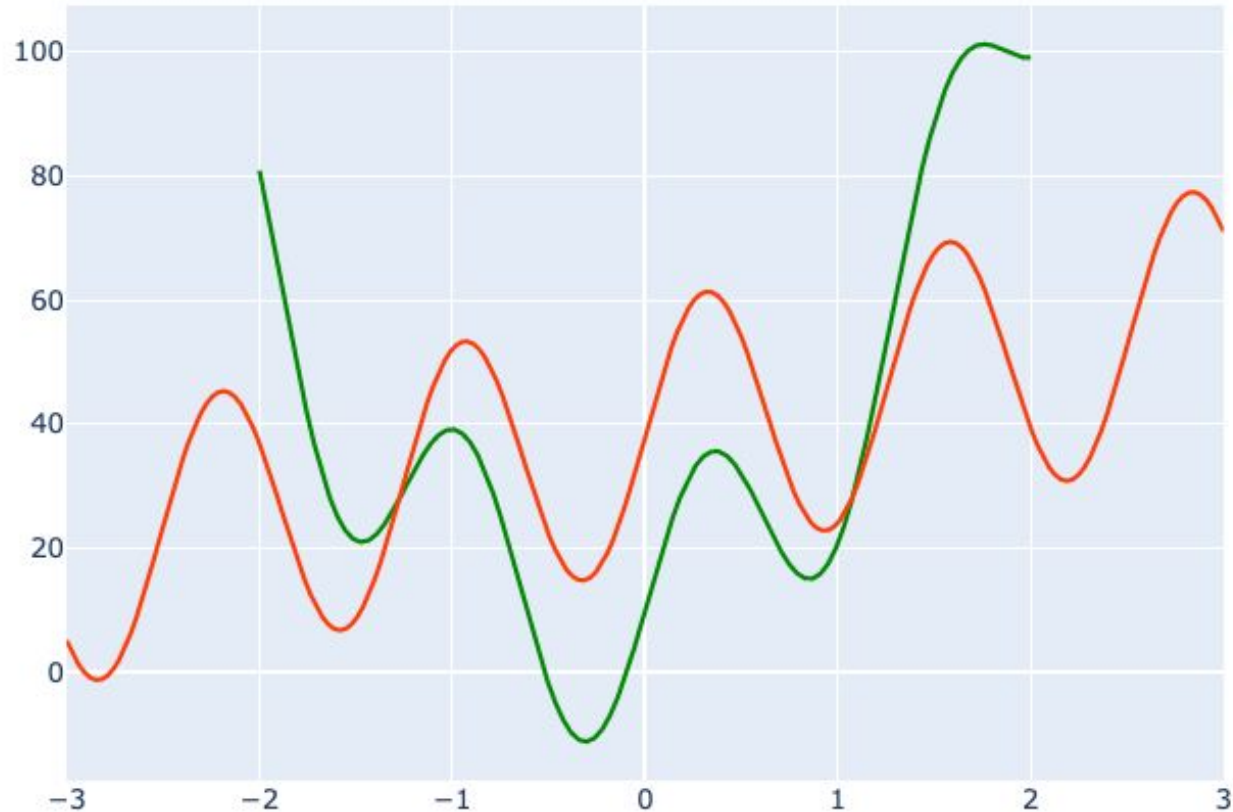
- Overfitting: small differences in random samples lead to large differences in the fitted model

Remedy:

- Reduce model complexity
- Don't fit the noise

Bias

Our Model Versus the Truth



- The green line is the fixed truth g
- The red line is our fitted model
- Bias measures how far off these two are, on average over all possible samples

Model Bias

- The difference between our predicted value and the true $g(x)$
 - averaged over all possible samples

$$\mathbb{E}(\hat{Y}(x) - g(x)) = \mathbb{E}(\hat{Y}(x)) - g(x)$$

$$\text{model bias} = \mathbb{E}(\hat{Y}(x)) - g(x)$$

- Bias depends on x but is **not random**
 - If positive, the model tends to overestimate at x
 - If negative, the model tends to underestimate at x

Reasons and Remedies

Some reasons:

- Underfitting
- Lack of domain knowledge

Remedies:

- Increase model complexity (but don't overfit)
- Consult domain experts to see which models make sense

Components of Prediction Error

Prediction Error

New individual: (x, Y)

$$\text{error} = Y - \hat{Y}(x)$$

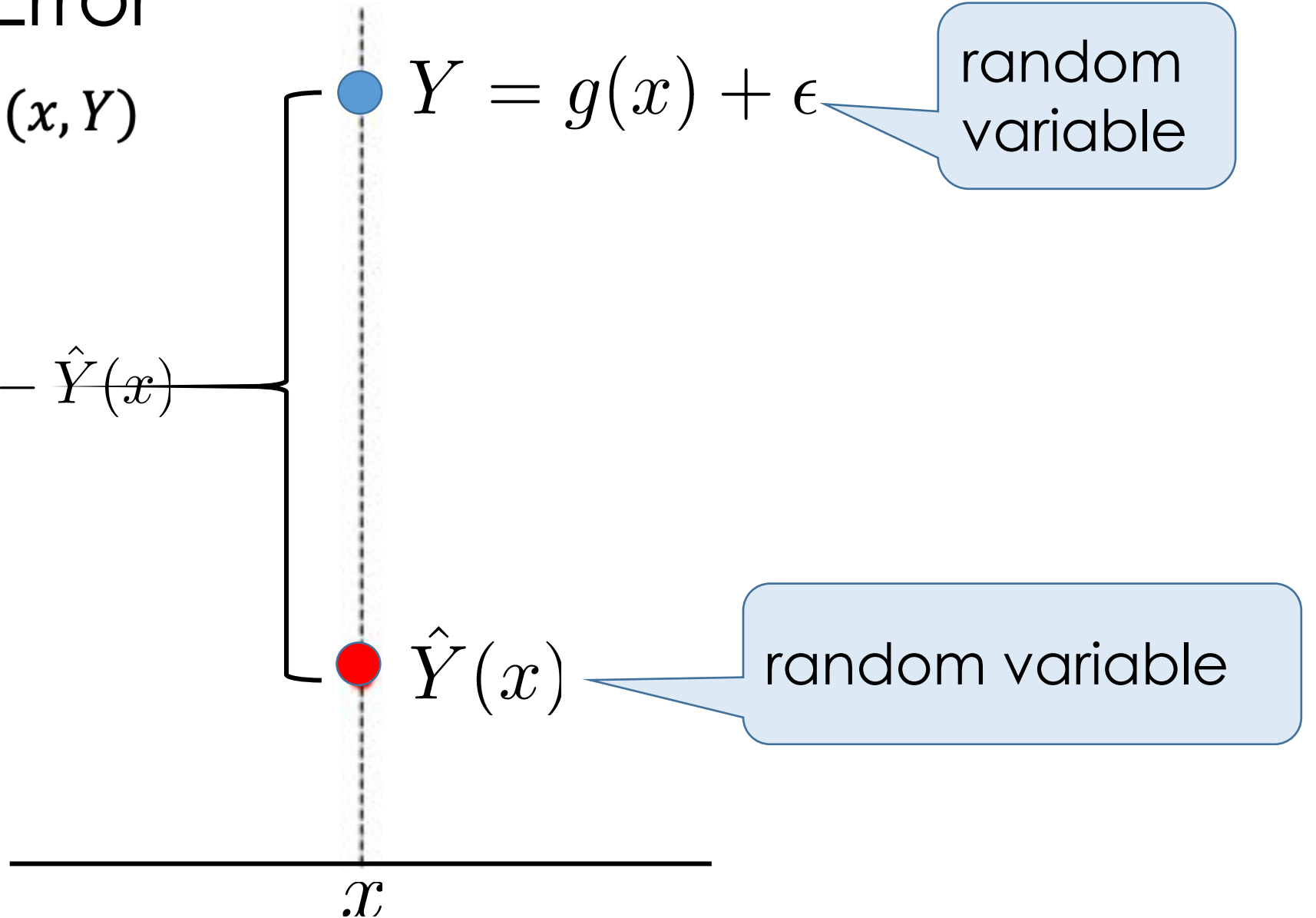
$$Y = g(x) + \epsilon$$

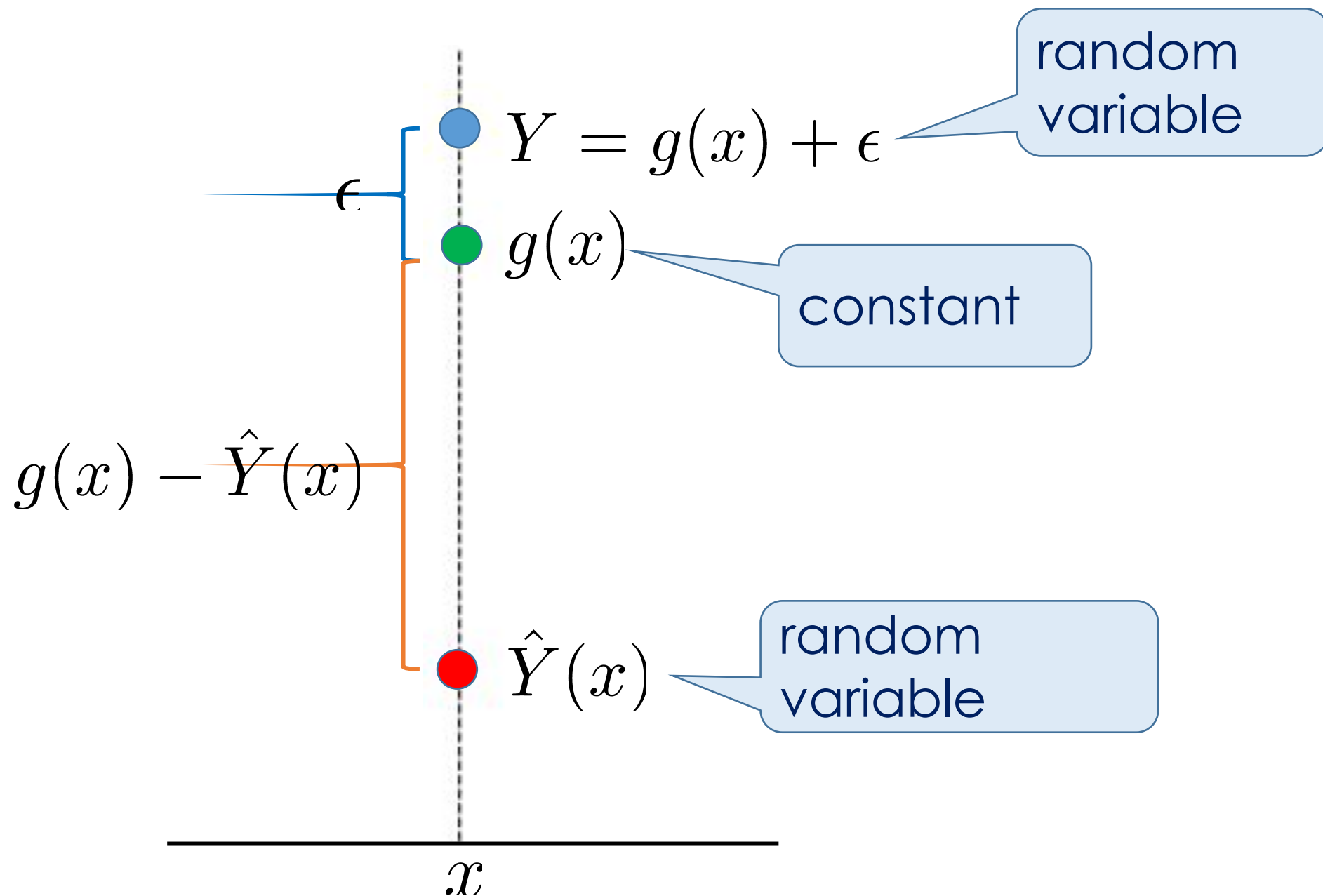
random
variable

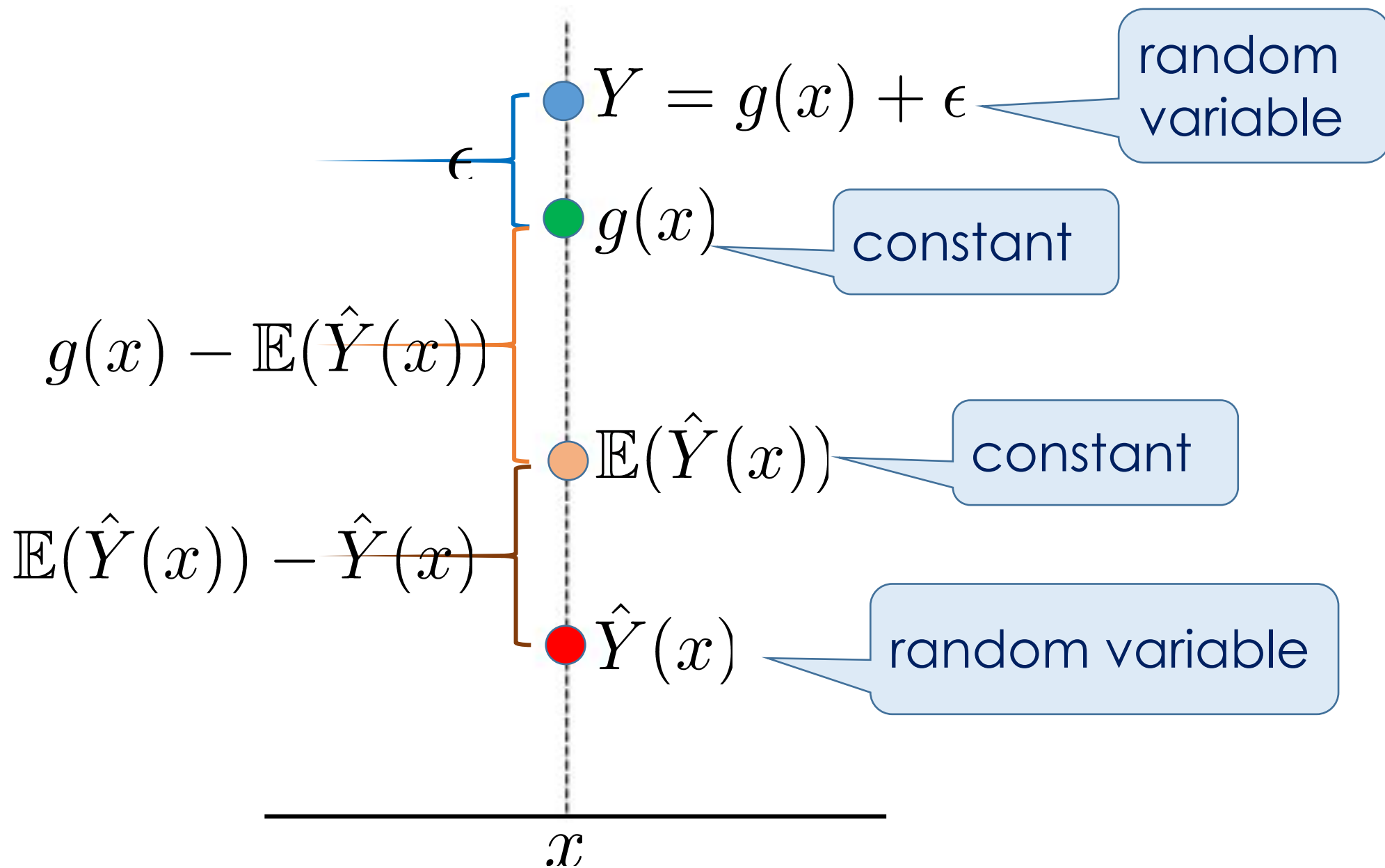
$$\hat{Y}(x)$$

random variable

x







Decomposition of Model Risk

Decomposition of Error and Risk

Decomposition of the prediction error into three pieces:

$$Y - \hat{Y}(x) = \epsilon + (g(x) - \mathbb{E}(\hat{Y}(x))) + (\mathbb{E}(\hat{Y}(x)) - \hat{Y}(x))$$

Decomposition of the model risk into three pieces:

$$\begin{aligned}\mathbb{E}((Y - \hat{Y}(x))^2) &= \mathbb{E}(\epsilon^2) \\ &\quad + (g(x) - \mathbb{E}(\hat{Y}(x)))^2 \\ &\quad + \mathbb{E}((\mathbb{E}(\hat{Y}(x)) - \hat{Y}(x))^2)\end{aligned}$$

The
cross-product
terms are 0

$$\text{model risk} = \sigma^2 + (\text{model bias})^2 + \text{model variance}$$

Bias Variance Decomposition

model risk = observation variance + (model bias)² + model variance

$$\mathbb{E}((Y - \hat{Y}(x))^2) = \sigma^2 + (\mathbb{E}(\hat{Y}(x)) - g(x))^2 + \mathbb{E}((\hat{Y}(x) - \mathbb{E}(\hat{Y}(x)))^2)$$

Note: these three terms are dimensionally consistent.

Predicting by a Function with Parameters

If our model is a non-random function f that has an unknown parameter vector θ

- θ is not random but has to be estimated from the sample
- The estimate $\hat{\theta}$ is random
- So our fitted function $f_{\hat{\theta}}$ is random and is just another name for \hat{Y}

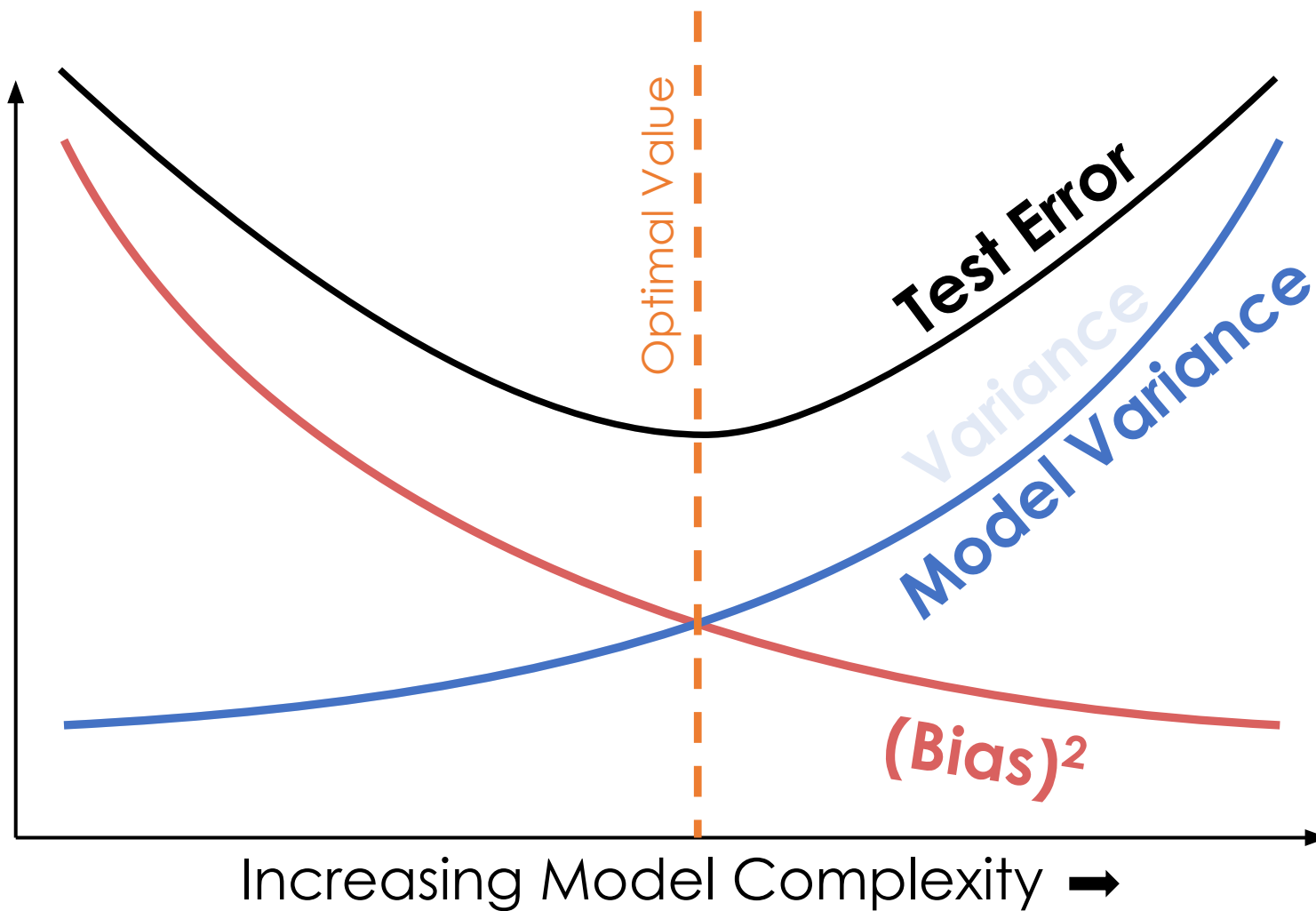
$$\mathbb{E}((Y - f_{\hat{\theta}}(x))^2) = \sigma^2 + (\mathbb{E}(f_{\hat{\theta}}(x)) - g(x))^2 + \mathbb{E}((f_{\hat{\theta}}(x) - \mathbb{E}(f_{\hat{\theta}}(x)))^2)$$

Derivation of Bias-Variance Formula

See the lecture notebooks

Summary for Modeling

Bias Variance Plot



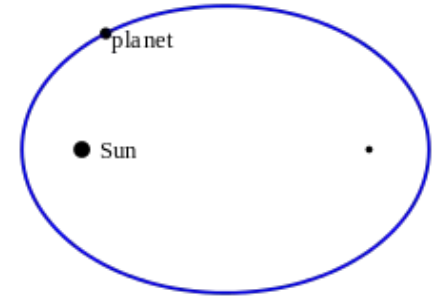
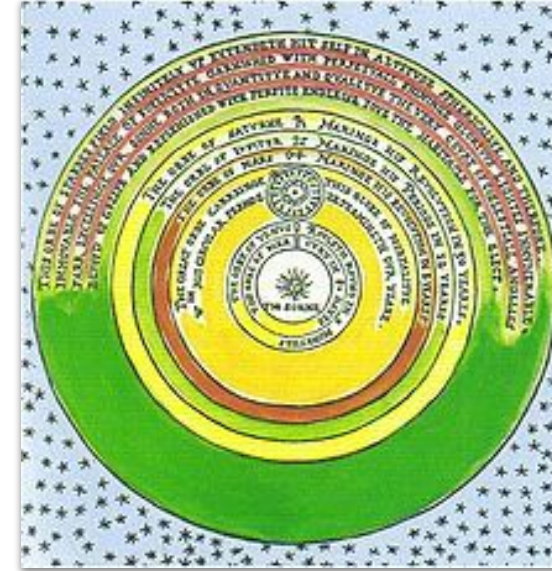
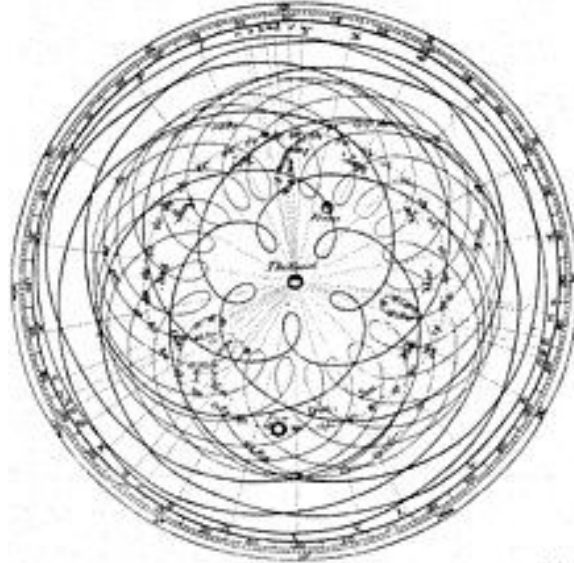
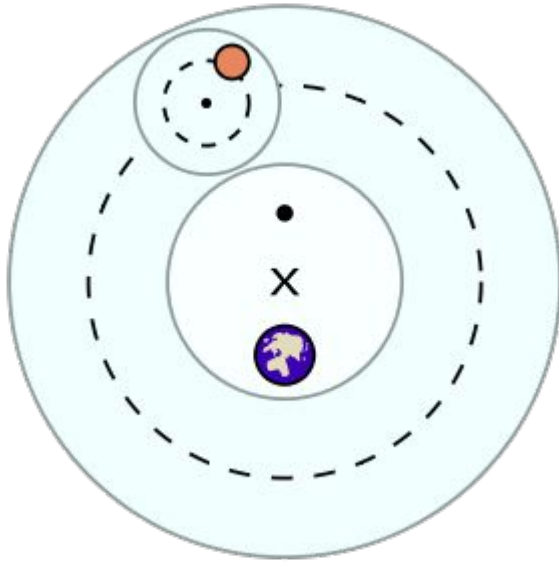
Modeling Goals

- Try to minimize all three of observation variance, model bias, and model variance.

But

- Observation variance is often out of our control
- Reducing complexity to reduce model variance can increase bias
- Increasing model complexity to reduce bias can increase model variance
- Domain knowledge matters: the right model structure!

The right model structure matters!



Ptolemaic Astronomy, a geocentric model based on circular orbits (*epicycles and deferents*).

High accuracy but very high model complexity.

Copernicus and Kepler: a heliocentric model with elliptical orbits.

Small model complexity yet high accuracy.