DS-UA 112
Introduction to Data Science

Week 10: Lecture 1
Correlation – Relating Attributes of Data

How can we use a known quantity to predict an unknown quantity?

DS-UA 112
Introduction to Data Science

Week 10: Lecture 1

Correlation – Predicting Attributes of Data

Adapted from Nolan, Speed, Gonzalez, Lau

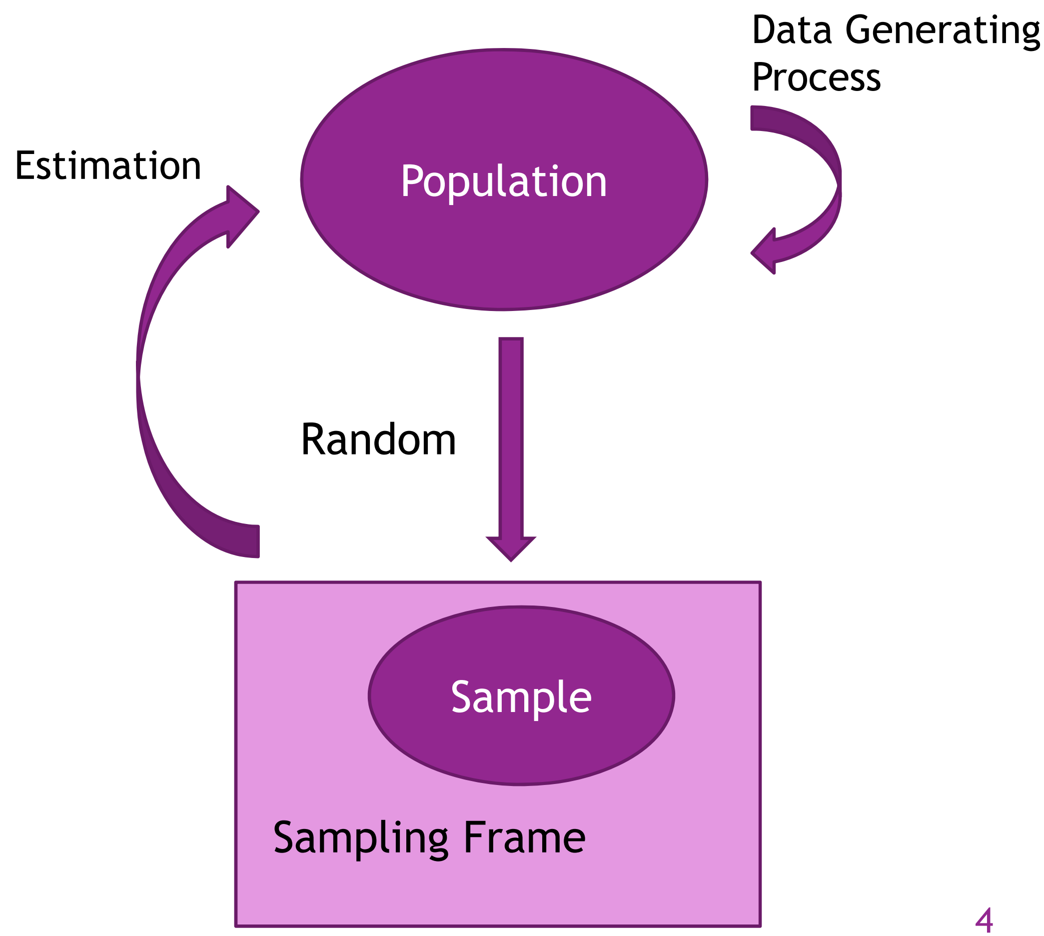
Announcements

- Please check Week 10 agenda on NYU Classes
  - Lab 9
    - Due on Friday April 3 at 12PM
  - Project 1
    - Due on Monday April 6 at 12PM
  - Survey

https://nyu.qualtrics.com/jfe/form/SV_3DCWUa4yc08L0wt

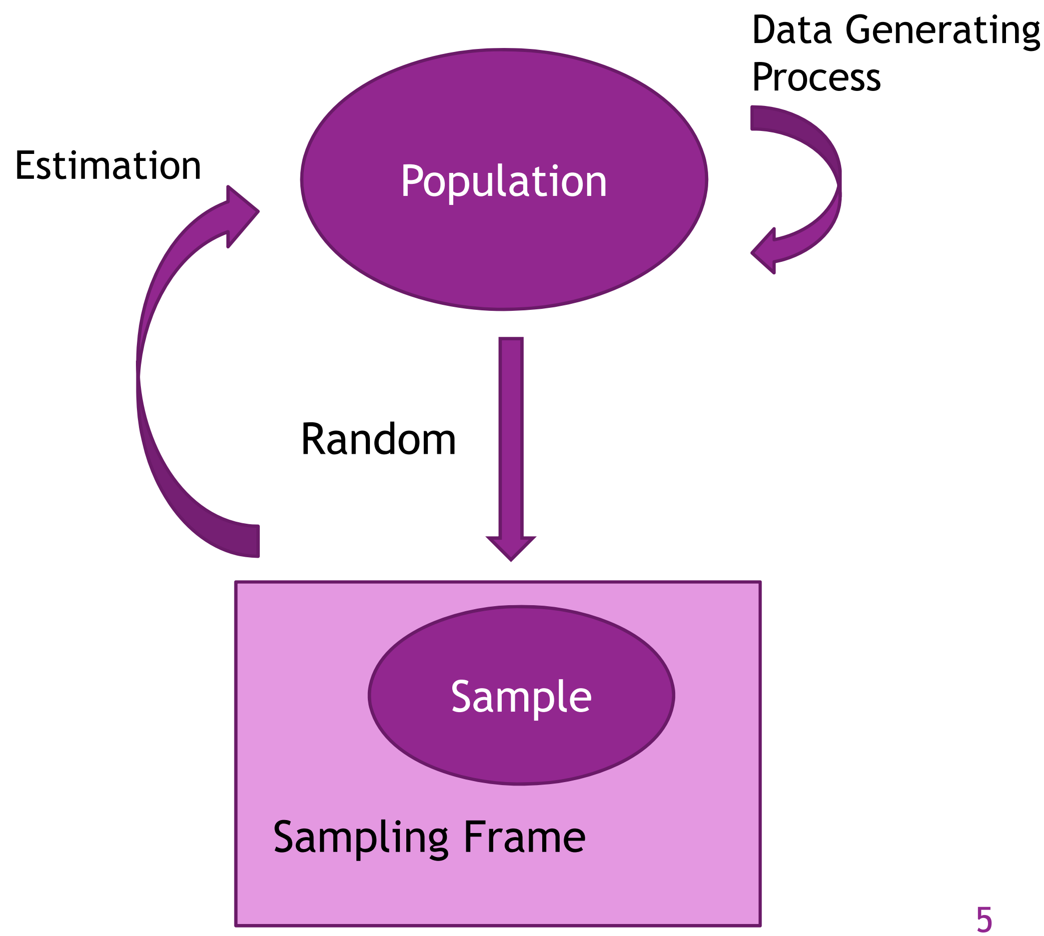
Review

- Modelling
  - We want to have models that are simple but not too simple.
  - For example, if I can spot ten or more clouds in the sky, then I should bring an umbrella because I might get caught in the rain
  - Our models should build on our experiences. We can use observations to inform our experiences. Sometimes the data helps us to change our minds



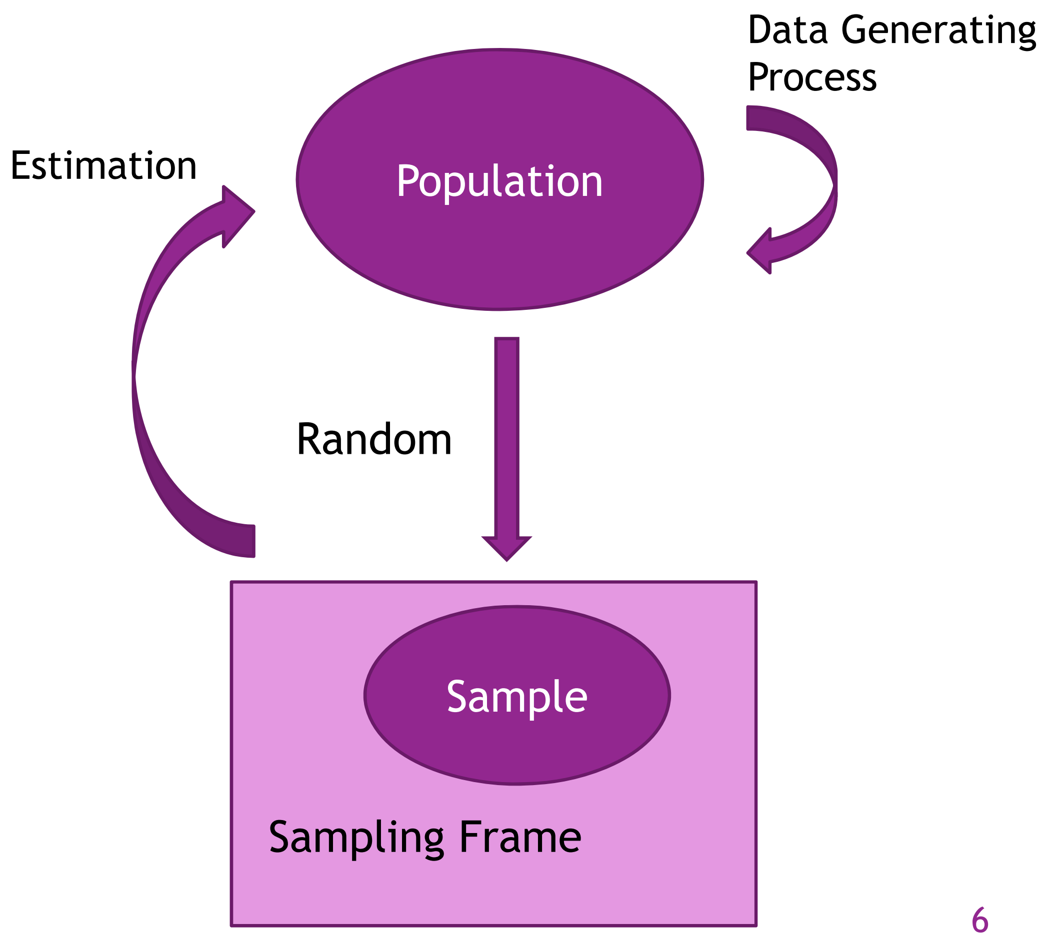
Review

- Modelling
  - Experience with a population leads to assumptions about trends generating the data in samples
  - Models represent these assumptions by probability distributions for relevant random variables
  - When the probability distribution depends on parameters, these unknown quantities are the missing pieces of the model



Review

- Modelling
  - We want to generalize findings beyond a sample to a population.
  - For example, does a sample of polling preferences suggest a winner of the election?
  - Estimation involves generalizing from the sample to the population. We use random variables to compute the chance that observations appears in our random samples



Exercise

See Week 10 Lecture 1 Notes

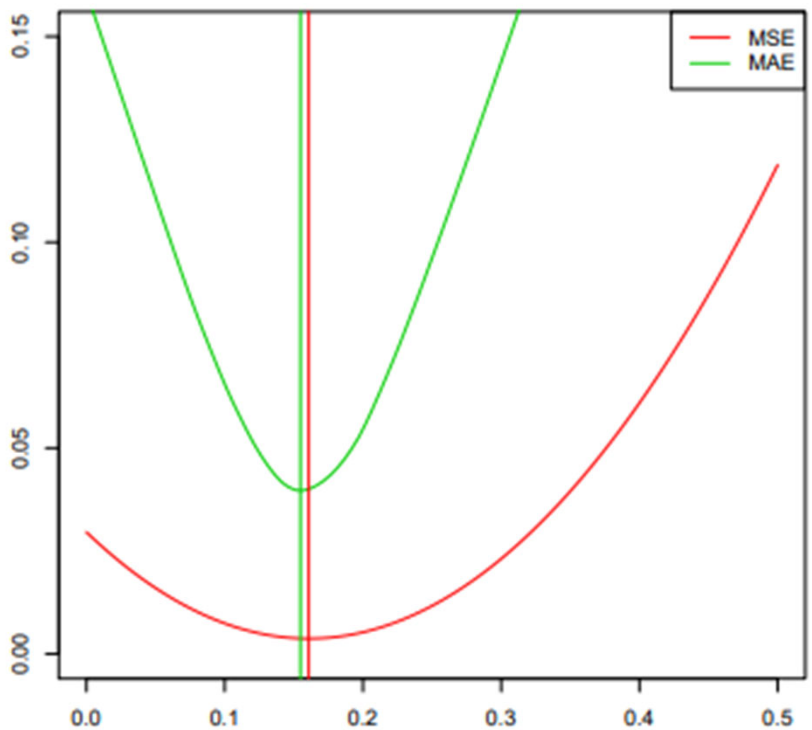
- 0-1 Random Variables
  - Suppose that a person chosen at random from a population has chance p of possessing a characteristic.
  - For example, the characteristic could be voting Democrat.
  - The chance of not possessing the characteristic is 1-p by the complement rule
  - We call random variables taking the value 0 or 1 Bernoulli random variables
  - Denote the characteristics as 1 and 0. Use $X_{1,\dots,}X_n$ to denote the corresponding random variables for *n* observations.
  - How can we make estimates about the number of 1's and 0's in the population based on the number of 1's and 0's in the sample of size *n*

Review

- Learning from data involves
  - Selecting an appropriate model
    - Does the model reflect our understanding?
  - Determining an estimator to fill in the missing pieces of the model
    - Can we fit the model to the data?
  - Assessing the validity of the model
    - Has the model provide accurate and robust insights?

- We can use loss functions to fit the model to data. Remember a loss function inputs
  - Data corresponding to the observations in a random sample
  - Unknown quantity that should estimate the parameters
- The output measures the accuracy and consistency of the model for a choice of parameters.

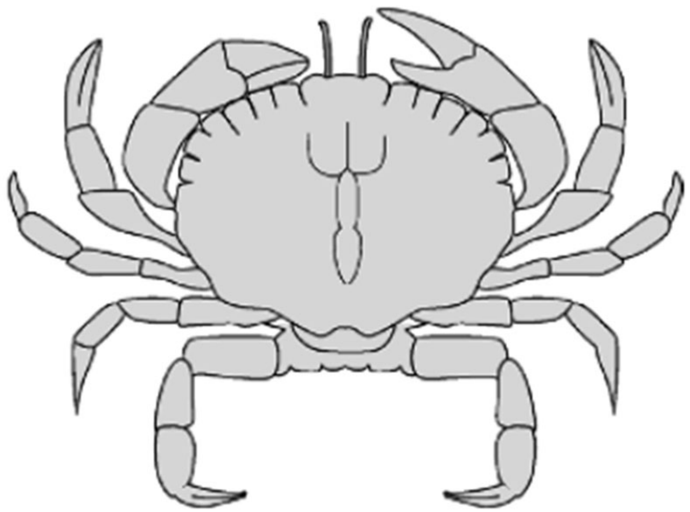
Review



- Remember that we have different choices for loss functions because we have different ways to measure accuracy and consistency of a model
- Compared to the mean absolute error, the mean square error has large output for large input.
- So mean absolute error is more robust to outliers

Agenda

- Risk
  - Expectation of Loss Functions
  - Bias and Variance
- Prediction
  - Joint Distribution
  - Conditional Distribution
- Correlation



Dungeness Crab (*Cancer magister*).

Estimators

Estimator: $\hat{\theta}(X_1, \ldots, X_n)$

Sample data: $x_1, \ldots, x_n$

Estimate: $\hat{\theta}(x_1, \ldots, x_n)$

Parameter: θ^*

- Note that the estimator is a function of random variables and the estimate is a function of numbers.

- Remember that we use capital letters to denote random variables and lowercase letters to denotes values obtained from the random variables
- The subscripts labels each random variable or value across the different samples
- Often the Greek letter theta denotes the unknown quantity in a loss function. The asterisk means the parameter. The hat means the estimate we choose from minimizing the loss function.

Risk

- Suppose we have a loss function L depending on dataset $x_{1,\dots,}x_n$ and unknown quantity θ
- For fixed value of θ, we can compare the value of $L(\theta, x_1,\dots,x_n)$ across different datasets
- We can take the datasets to correspond to different values of a random variable. If we repeatedly observe *n* values of a random variable X, then we can compute $L(\theta, x_1,\dots,x_n)$ for each dataset

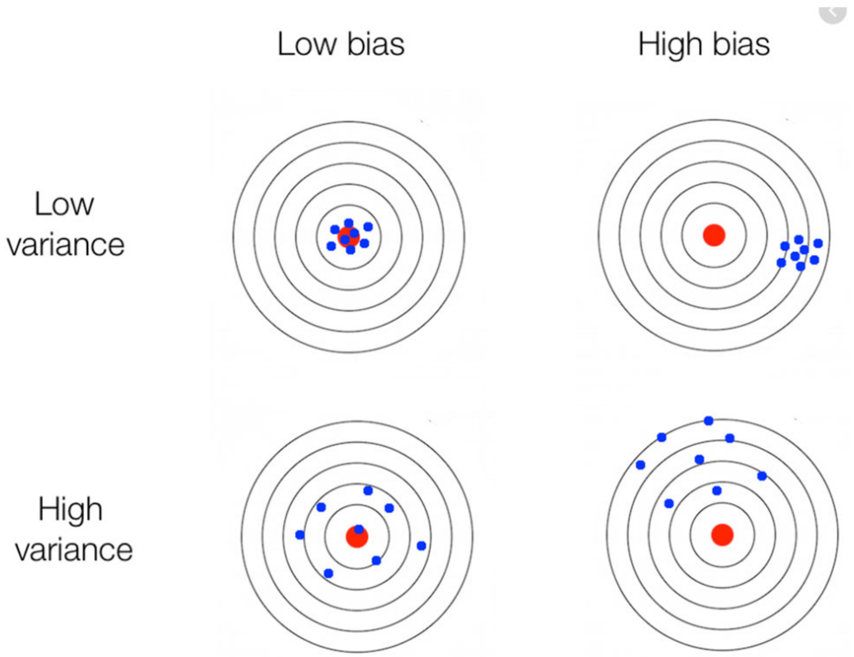
For example the square loss is $L(\theta, X) = (\theta - X)^2$

- Risk is the expectation of a loss function for random variable

$$E[L(\theta, X)]$$

- We need to know the probability distribution of X to compute the expectation.
- If we can compute the expectation, then we better understand the value of the loss function across different random samples

Bias and Variance



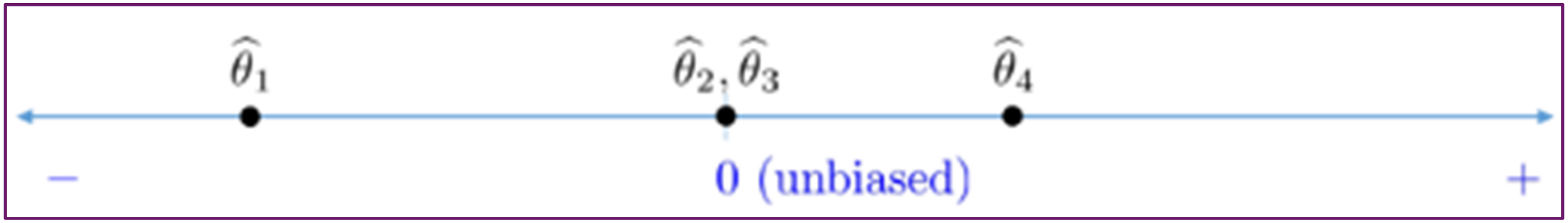
- We want to choose θ to make E[L(θ, X)] small. The choice of θ that makes the expectation smallest is $\hat{\theta}$. We can write $\hat{\theta}_n$ to remind us that the estimate from *n* samples.
- For the square loss, we can break the risk into two components
  - Bias measuring the accuracy of the estimator
  - Variance measuring the consistency of the estimator
- Here bias does not refer to a property of data but to a tendency of estimators

Bias and Variance

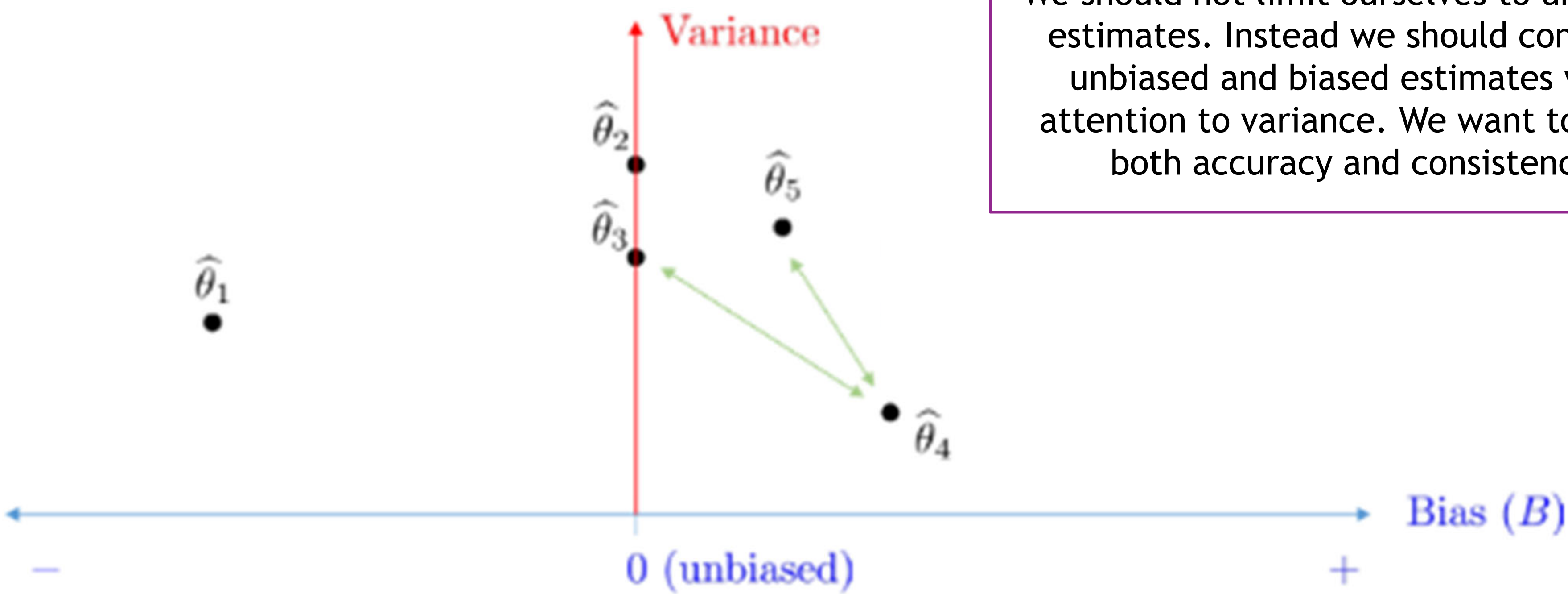
$$\begin{aligned}
& \mathrm{E}[(\hat{\theta}_n - \theta)^2] \\
= \;& \mathrm{E}[(\hat{\theta}_n - \mathrm{E}[\hat{\theta}_n] + \mathrm{E}[\hat{\theta}_n] - \theta)^2] \\
= \;& \mathrm{E}[(\hat{\theta}_n - \mathrm{E}[\hat{\theta}_n])^2 + 2(\hat{\theta}_n - \mathrm{E}[\hat{\theta}_n])(\mathrm{E}[\hat{\theta}_n] - \theta) \\
& +(\mathrm{E}[\hat{\theta}_n] - \theta)^2] \\
= \;& \mathrm{E}[(\hat{\theta}_n - \mathrm{E}[\hat{\theta}_n])^2] + 2(\mathrm{E}[\hat{\theta}_n] - \theta)\,\mathrm{E}[(\hat{\theta}_n - \mathrm{E}[\hat{\theta}_n])] \\
& + \mathrm{E}[(\mathrm{E}[\hat{\theta}_n] - \theta)^2] \\
= \;& \mathrm{Var}[\hat{\theta}_n] + 2(\mathrm{E}[\hat{\theta}_n] - \theta) \times 0 + (\mathrm{E}[\hat{\theta}_n] - \theta)^2 \\
= \;& \mathrm{Var}[\hat{\theta}_n] + (\mathrm{Bias}[\hat{\theta}_n])^2
\end{aligned}$$

Bias and Variance

- Remember that an estimator is a function of a random variables. So an estimator is a random variable.
- Since an estimator is a random variable, we can compute its variance
- If variance is a small number, then the estimator is close to its expectation. So for an unbiased estimator, small variance tells us we have a good approximation of the population parameter.



Bias and Variance



We should not limit ourselves to unbiased estimates. Instead we should compare unbiased and biased estimates with attention to variance. We want to have both accuracy and consistency

Joint Distribution

- We can define distributions for multiple random variables
- The joint distribution of multiple random variables specifies the chances that simultaneously each random variable takes a value
- For example, suppose that we roll two dice. One die X is fair and one die Y is not fair.

$$
\begin{aligned}
\Pr(Y = 1) = \Pr(Y = 2) &= \frac{1}{16} \\
\Pr(Y = 3) = \Pr(Y = 4) &= \frac{3}{16} \\
\Pr(Y = 5) = \Pr(Y = 6) &= \frac{4}{16}
\end{aligned}
$$

Joint Distribution

- The values in the rows for a fixed column are the probabilities of X conditional on Y.
- The values in the columns for a fixed row are the probabilities of Y conditional on X.
- The sum of a row across the columns is the marginal value of X. The sum of a column down the rows is the marginal value of Y

Loaded die, Y

		1	2	3	4	5	6	
	1	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
	2	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
Fair die, X	3	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
	4	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
	5	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
	6	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$	$\frac{1}{6}$
		$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{4}{16}$	

Joint Distribution

What is the probability of Y – X >2?

- The values in the rows for a fixed column are the probabilities of X conditional on Y.
- The values in the columns for a fixed row are the probabilities of Y conditional on X.
- The sum of a row across the columns is the marginal value of X. The sum of a column down the rows is the marginal value of Y

Loaded die, Y

Fair die, X	1	2	3	4	5	6
1	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
2	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
3	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
4	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
5	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$
6	$\frac{1}{96}$	$\frac{1}{96}$	$\frac{3}{96}$	$\frac{3}{96}$	$\frac{4}{96}$	$\frac{4}{96}$

Prediction

- Estimation means determining the parameters of probability distributions in a model for a population. Here we study $P(X=x, Y=y)$ for all values
- Remember that the multiplication rule tells us that

$$P(X=x, Y=y) = P(Y=y \mid X=x)\, P(X=x)$$

- If we want to use known information to guess unknown information then we should focus on conditional probability

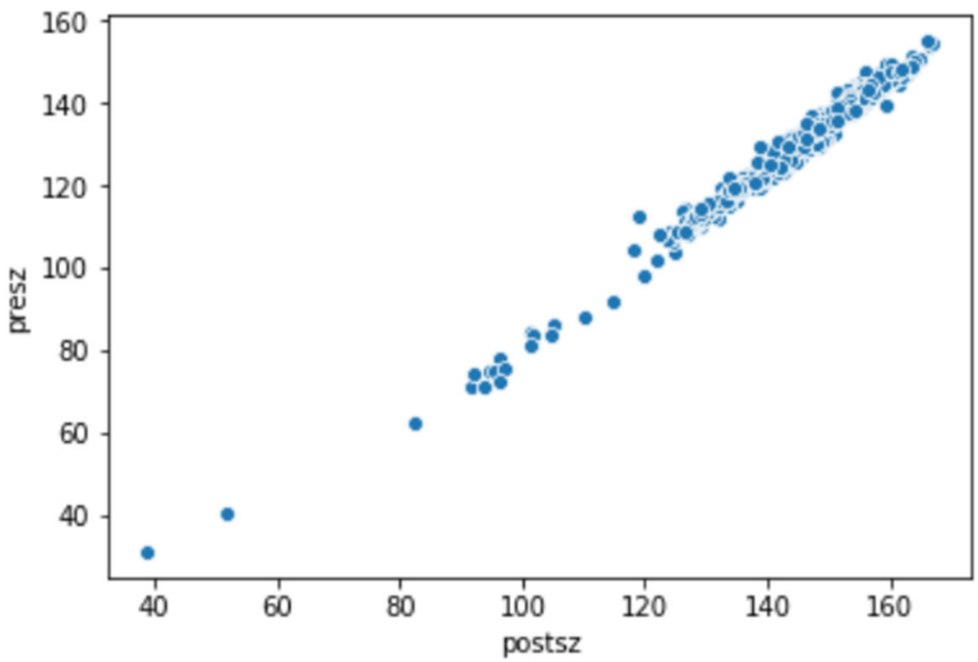
- Prediction means studying

$$P(Y=y \mid X=x)$$

Here $X=x$ is known and Y is unknown

- Sometimes we can make accurate and consistent predictions without knowing the probability distribution. In particular, we can try to make predictions with estimation of parameters.

Correlation



- We want X to relate to Y for prediction. Given information about X, we need to generate information about Y.
- If it appears the X and Y differ by a transformation that involves scaling

$$X \longrightarrow c\,X$$

and shifting

$$c\,X \longrightarrow c\,X + d$$

then maybe we can find a linear relationship between the random variables

Correlation

Covariance

Standard Deviation

Correlation measures the cosine of the angle between the dataset thought of as vectors

Standard Deviation

$$\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}} \frac{1}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

$$= \boxed{\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}} \frac{\sqrt{\sum_{i=1}^{n} y_i^2}}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Summary

- Risk
  - Expectation of Loss Functions
  - Bias and Variance
- Prediction
  - Joint Distribution
  - Conditional Distribution
- Correlation

Goals

- How does risk combine loss functions and random variables?
- What is the connection between correlation and causation?

Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

Can you find two dependent random variables with correlation equal to 0?