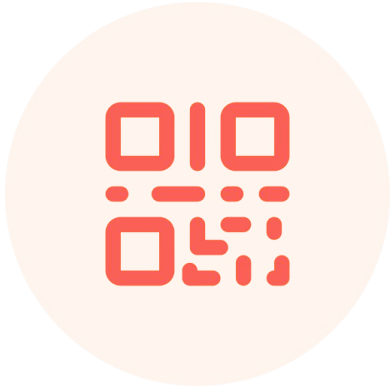
slido



Join at slido.com
#1465597

ⓘ Start presenting to display the joining instructions on this slide.

LECTURE 16

More on Regularization & Random Variables

Numerical functions of random samples and their properties; sampling variability.

Data 100/Data 200, Spring 2023 @ UC Berkeley

Narges Norouzi and Lisa Yan

Content credit: Acknowledgments

Today's Roadmap

Lecture 16, Data 100 Spring 2023

1465597

Continue on Regularization

- **L2 Regularization (Ridge)**
- Scaling Data for Regularization
- L1 Regularization (LASSO)

Random Variables and Distributions

Expectation and Variance

Sums of Random Variables

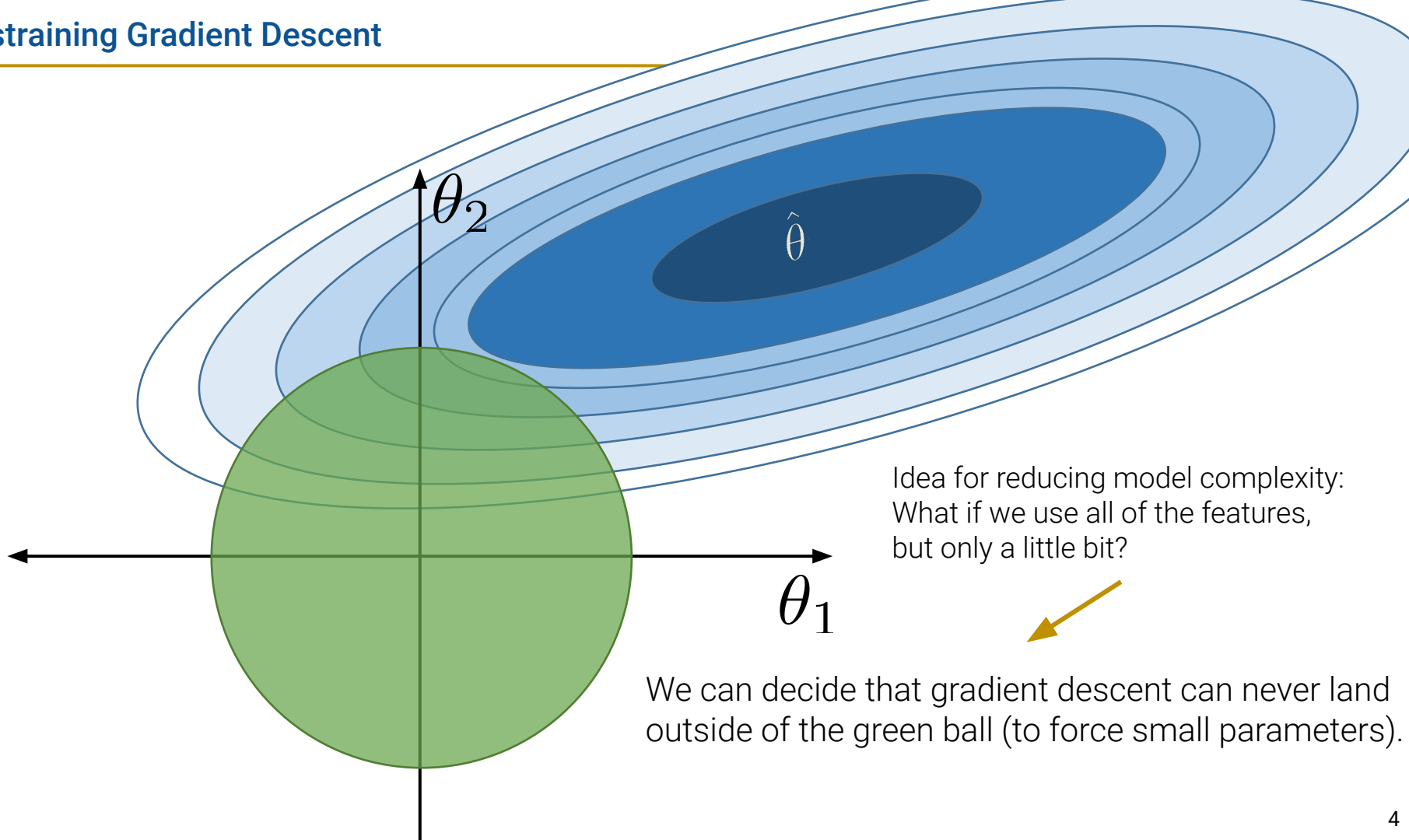
- Equality vs Identically Distributed vs. IID
- Properties of Expectation and Variance
- Covariance, Correlation

Bernoulli and Binomial Random Variables

Sample Statistics

- Sample Mean
- Central Limit Theorem

Constraining Gradient Descent



Idea for reducing model complexity:
What if we use all of the features,
but only a little bit?

We can decide that gradient descent can never land outside of the green ball (to force small parameters).

L2 Regularization

Constraining our model's parameters to a ball around the origin is called **L2 Regularization**.

- The smaller the ball, the simpler the model.

Ordinary least squares. Find thetas that minimize:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1\phi_{i,1} + \ldots + \theta_d\phi_{i,d}))^2$$

Ordinary least squares with **L2 regularization**. Find thetas that minimize:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1\phi_{i,1} + \ldots + \theta_d\phi_{i,d}))^2$$

Such that θ_1 through θ_d live inside a ball of radius Q.

L2 Regularization

1465597

Constraining our model's parameters to a ball around the origin is called **L2 Regularization**.

- The smaller the ball, the simpler the model.

Ordinary least squares. Find thetas that minimize:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1\phi_{i,1} + \ldots + \theta_d\phi_{i,d}))^2$$

Ordinary least squares with **L2 regularization**. Find thetas that minimize:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1\phi_{i,1} + \ldots + \theta_d\phi_{i,d}))^2 \quad \textbf{such that} \quad \sum_{j=1}^{d}\theta_j^2 \leq Q$$

Note, intercept term not included!

Quick Detour into EECS127 (not tested in this class but worth mentioning)

1465597

In 127, you'll learn (through the magic of Lagrangian Duality) that the two problems below are equivalent:

Problem 1: Find thetas that minimize:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1\phi_{i,1} + \cdots + \theta_d\phi_{i,d}))^2 \quad \textbf{such that} \quad \sum_{j=1}^{d}\theta_j^2 \leq Q$$

Problem 2: Find thetas that minimize:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1\phi_{i,1} + \cdots + \theta_d\phi_{i,d}))^2 + \lambda\sum_{j=1}^{d}\theta_j^2$$

The "objective function" that gradient descent is minimizing now has an extra term.

Intuitively, this extra **right term penalizes large thetas**.

L2 Regularized Least Squares in sklearn

1465597

We can run least squares with an **L2 regularization term** by using the "Ridge" class.

```
from sklearn.linear_model import Ridge
ridge_model = Ridge(alpha=10000)
ridge_model.fit(vehicle_data_with_squared_features, vehicle_data["mpg"])
```

Coefficients we get back:

```
ridge_model.coef_
```

```
array([-5.56414449e-02, -7.93804083e-03, -8.22081425e-02, -6.18785466e-04,
       -2.55492157e-05,  9.47353944e-04,  7.58061062e-07,  1.07439477e-05,
       -1.64344898e-04])
```

Note: sklearn's "alpha" parameter is equivalent to λ in the linear regression with L2 regularizer equation

- Alpha is inversely related to the ball radius! Large alpha means small ball.

L2 Regularized Least Squares in sklearn

1465597

We can run least squares with an **L2 regularization term** by using the "Ridge" class.

```
from sklearn.linear_model import Ridge
ridge_model = Ridge(alpha=10**-5)
ridge_model.fit(vehicle_data_with_squared_features, vehicle_data["mpg"])
```

For a tiny alpha, the coefficients are larger:

```
ridge_model.coef_
```

```
array([-1.35872588e-01, -1.46864458e-04, -1.18230336e-01, -4.03590098e-04,
       -1.12862371e-05,  8.25179864e-04, -1.17645881e-06,  2.69757832e-05,
       -1.72888463e-04])
```

Note: sklearn's "alpha" parameter is equivalent to λ in the linear regression with L2 regularizer equation

- Alpha is inversely related to the ball radius! Large alpha means small ball.

L2 Regularized Least Squares in sklearn

We can run least squares with an **L2 regularization term** by using the "Ridge" class. For a tiny alpha, the coefficients are also about the same as a standard OLS model's coefficients!

```
ridge_model.coef_
```

```
array([-1.35872588e-01, -1.46864458e-04, -1.18230336e-01, -4.03590098e-04,
       -1.12862371e-05,  8.25179864e-04, -1.17645881e-06,  2.69757832e-05,
       -1.72888463e-04])
```

```
from sklearn.linear_model import LinearRegression
linear_model = LinearRegression()
linear_model.fit(vehicle_data_with_squared_features, vehicle_data["mpg"])
```

```
linear_model.coef_
```

```
array([-1.35872588e-01, -1.46864447e-04, -1.18230336e-01, -4.03590097e-04,
       -1.12862370e-05,  8.25179863e-04, -1.17645882e-06,  2.69757832e-05,
       -1.72888463e-04])
```

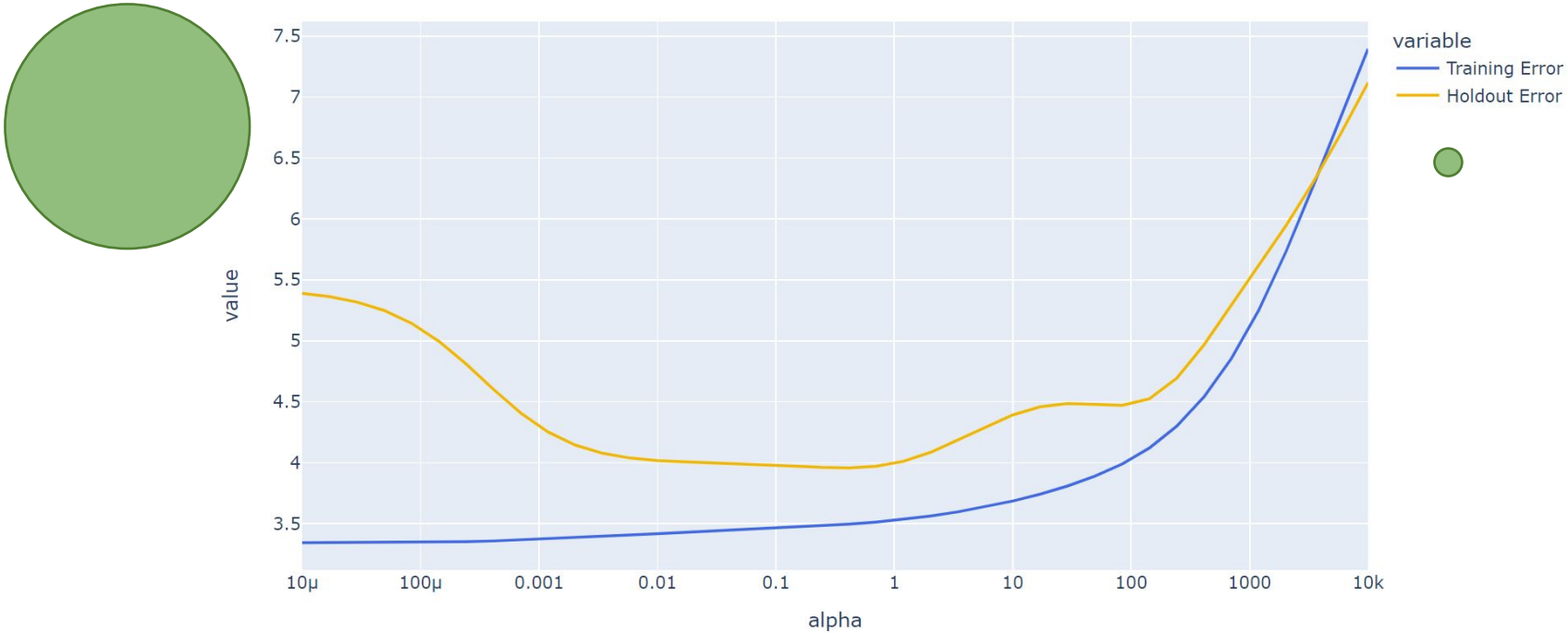
Green ball includes the OLS solution!

Figure (from lab 8)

1465597

In lab8, you'll run an experiment for different values of alpha. The resulting plot is shown below.

- Note: Since alpha is the inverse of the ball radius, the complexity is higher on the left!



Terminology Note

1465597

Why does sklearn use the word “Ridge”?

Because least squares with an **L2 regularization term** is also called “**Ridge Regression**”.

- Term is historical. Doesn’t really matter.

Why does sklearn use a hyperparameter which is the inverse of the ball radius?

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1\phi_{i,1} + \cdots + \theta_d\phi_{i,d}))^2 + \lambda\sum_{j=1}^{d}\theta_j^2$$

Mathematical Note

1465597

Ridge Regression has a closed form solution which we will not derive.

- Note: The solution exists even if the feature matrix has collinearity between its columns.

$$\hat{\theta}_{ridge} = (\mathbb{X}^T\mathbb{X} + n\lambda I)^{-1}\mathbb{X}^T\mathbb{Y}$$

Identity matrix

Scaling Data for Regularization

Lecture 16, Data 100 Spring 2023

1465597

Continue on Regularization

- L2 Regularization (Ridge)
- **Scaling Data for Regularization**
- L1 Regularization (LASSO)

Random Variables and Distributions

Expectation and Variance

Sums of Random Variables

- Equality vs Identically Distributed vs. IID
- Properties of Expectation and Variance
- Covariance, Correlation

Bernoulli and Binomial Random Variables

Sample Statistics

- Sample Mean
- Central Limit Theorem

One Issue With Our Approach

1465597

Our data from before has features of quite different numerical scale!

- Optimal theta for hp will probably be much further from origin than theta for weight^2.

hp	weight	displacement	hp^2	hp weight	hp displacement	weight^2	weight displacement	displacement^2
130.0	3504.0	307.0	16900.0	455520.0	39910.0	12278016.0	1075728.0	94249.0
165.0	3693.0	350.0	27225.0	609345.0	57750.0	13638249.0	1292550.0	122500.0
150.0	3436.0	318.0	22500.0	515400.0	47700.0	11806096.0	1092648.0	101124.0
150.0	3433.0	304.0	22500.0	514950.0	45600.0	11785489.0	1043632.0	92416.0
140.0	3449.0	302.0	19600.0	482860.0	42280.0	11895601.0	1041598.0	91204.0
...	...	...	...	...	...	...	...	...
86.0	2790.0	140.0	7396.0	239940.0	12040.0	7784100.0	390600.0	19600.0
52.0	2130.0	97.0	2704.0	110760.0	5044.0	4536900.0	206610.0	9409.0
84.0	2295.0	135.0	7056.0	192780.0	11340.0	5267025.0	309825.0	18225.0
79.0	2625.0	120.0	6241.0	207375.0	9480.0	6890625.0	315000.0	14400.0
82.0	2720.0	119.0	6724.0	223040.0	9758.0	7398400.0	323680.0	14161.0

Theta will tend to be smaller for weight^2 than other parameters

Coefficients from Earlier

1465597

hp	weight	displacement	hp^2	hp weight	hp displacement	weight^2	weight displacement	displacement^2
130.0	3504.0	307.0	16900.0	455520.0	39910.0	12278016.0	1075728.0	94249.0
165.0	3693.0	350.0	27225.0	609345.0	57750.0	13638249.0	1292550.0	122500.0
150.0	3436.0	318.0	22500.0	515400.0	47700.0	11806096.0	1092648.0	101124.0
150.0	3433.0	304.0	22500.0	514950.0	45600.0	11785489.0	1043632.0	92416.0
140.0	3449.0	302.0	19600.0	482860.0	42280.0	11895601.0	1041598.0	91204.0
...	...	...	...	...	...	...	...	...
86.0	2790.0	140.0	7396.0	239940.0	12040.0	7784100.0	390600.0	19600.0
52.0	2130.0	97.0	2704.0	110760.0	5044.0	4536900.0	206610.0	9409.0
84.0	2295.0	135.0	7056.0	192780.0	11340.0	5267025.0	309825.0	18225.0
79.0	2625.0	120.0	6241.0	207375.0	9480.0	6890625.0	315000.0	14400.0
82.0	2720.0	119.0	6724.0	223040.0	9758.0	7398400.0	323680.0	14161.0

```
ridge_model.coef_
```

```
array([-1.35872588e-01, -1.46864458e-04, -1.18230336e-01, -4.03590098e-04,
       -1.12862371e-05,  8.25179864e-04, -1.17645881e-06,  2.69757832e-05,
       -1.72888463e-04])
```

Making Things Fair

1465597

Ideally, our data should all be on the same scale.

- One approach: Standardize the data, i.e. replace everything with its Z-score.

$$z_k = \frac{x_k - \mu_k}{\sigma_k}$$

- Resulting model coefficients will be all on the same scale.

```
from sklearn.preprocessing import StandardScaler
ss = StandardScaler()
rescaled_df = pd.DataFrame(ss.fit_transform(vehicle_data_with_squared_features),
                           columns = ss.get_feature_names_out())


ridge_model = Ridge(alpha=10000)
ridge_model.fit(rescaled_df, vehicle_data["mpg"])
ridge_model.coef_
```

```
array([-0.1792743 , -0.19610513, -0.18648617, -0.1601219 , -0.18015125,
       -0.16858023, -0.18779478, -0.18176294, -0.17021841])
```

L1 Regularization (LASSO)

Lecture 16, Data 100 Spring 2023

1465597

Continue on Regularization

- L2 Regularization (Ridge)
- Scaling Data for Regularization
- **L1 Regularization (LASSO)**

Random Variables and Distributions

Expectation and Variance

Sums of Random Variables

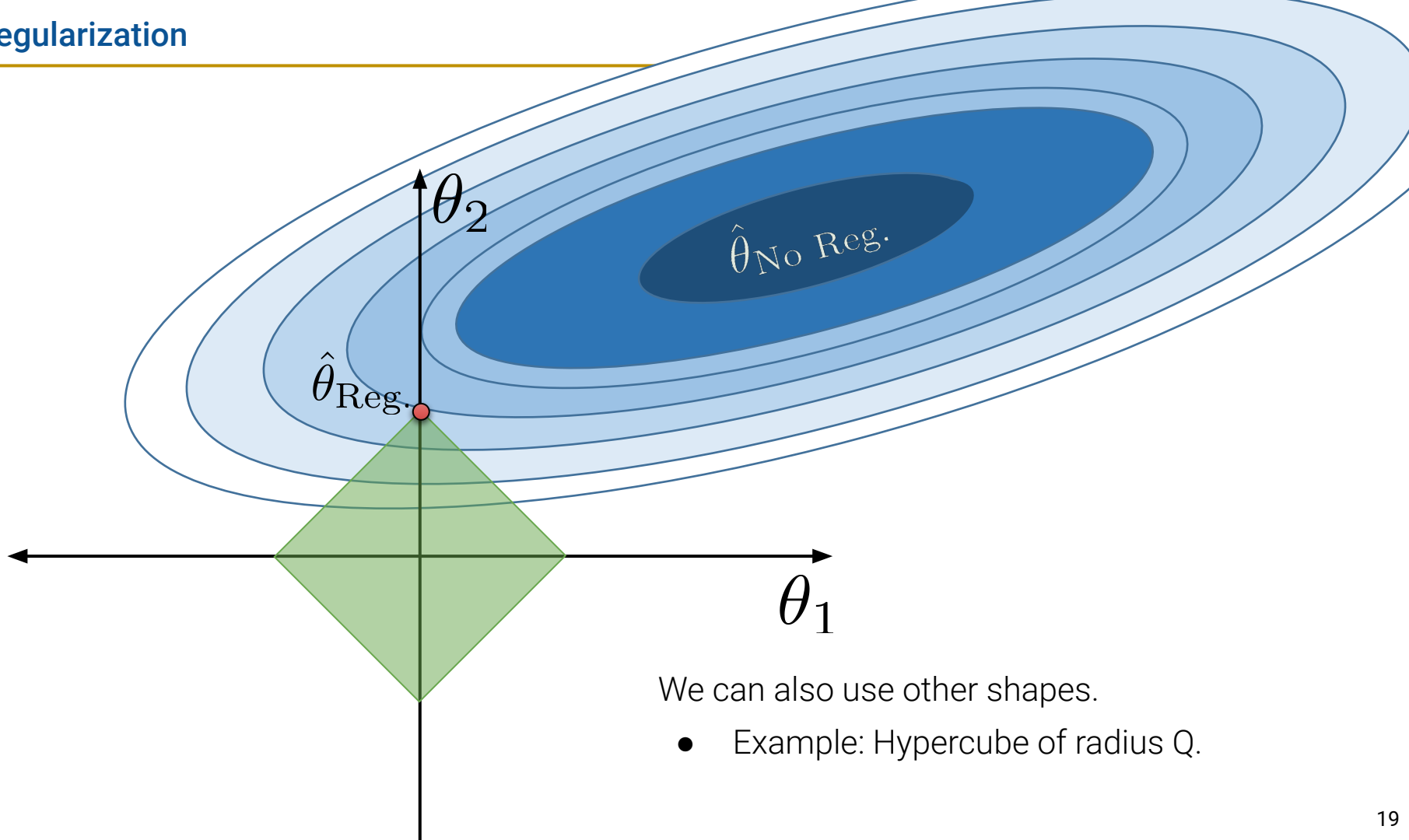
- Equality vs Identically Distributed vs. IID
- Properties of Expectation and Variance
- Covariance, Correlation

Bernoulli and Binomial Random Variables

Sample Statistics

- Sample Mean
- Central Limit Theorem

L1 Regularization



We can also use other shapes.

- Example: Hypercube of radius Q.

L1 Regularization in Equation Form

Using a hypercube is known as **L1 regularization**. Expressed mathematically in the two equivalent forms below:

Problem 1: Find thetas that minimize:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1\phi_{i,1} + \cdots + \theta_d\phi_{i,d}))^2 \quad \textbf{such that} \quad \sum_{j=1}^{d}|\theta_j| \leq Q$$

Problem 2: Find thetas that minimize:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1\phi_{i,1} + \cdots + \theta_d\phi_{i,d}))^2 + \lambda\sum_{j=1}^{d}|\theta_j|$$

L1 Regularized OLS in sklearn

1465597

In sklearn, we use the Lasso module.

- Note: Performing OLS with L1 regularization is also called **LASSO regression**.

```
from sklearn.linear_model import Lasso
lasso_model = Lasso(alpha = 10)
lasso_model.fit(vehicle_data_with_squared_features, vehicle_data["mpg"])
lasso_model.coef_
```

```
lasso_model.coef_
```

```
array([-0.00000000e+00, -1.88104942e-02, -0.00000000e+00, -1.19625308e-03,
        8.84657720e-06,  8.77253835e-04,  3.16759194e-06, -3.21738391e-05,
       -1.29386937e-05])
```

LASSO and “Feature Selection”

1465597

The optimal parameters for a LASSO model tend to include a lot of zeroes! In other words, LASSO effectively selects only a subset of the features.

```
from sklearn.linear_model import Lasso
lasso_model = Lasso(alpha = 10)
lasso_model.fit(vehicle_data_with_squared_features, vehicle_data["mpg"])
lasso_model.coef_
```

```
lasso_model.coef_
```

```
array([-0.00000000e+00, -1.88104942e-02, -0.00000000e+00, -1.19625308e-03,
        8.84657720e-06,  8.77253835e-04,  3.16759194e-06, -3.21738391e-05,
       -1.29386937e-05])
```

Intuitive reason:

- Imagine expanding a 3D cube until it intersects a balloon. More likely to intersect at a corner or edge than a face (especially in high dimensions)

Summary of Regression Methods

1465597

Our regression models are summarized below.

- The “Objective” column gives the function that our gradient descent optimizer minimizes.
- Note that this table uses lambda instead of alpha for regularization strength. Both are common.

Name	Model	Loss	Reg.	Objective	Solution
OLS	$\hat{\mathbb{Y}} = \mathbb{X}\theta$	Squared loss	None	$\frac{1}{n}\|\|\mathbb{Y} - \mathbb{X}\theta\|\|_2^2$	$\hat{\theta}_{\text{OLS}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$
Ridge Regression	$\hat{\mathbb{Y}} = \mathbb{X}\theta$	Squared loss	L2	$\frac{1}{n}\|\|\mathbb{Y} - \mathbb{X}\theta\|\|_2^2 + \lambda\sum_{j=1}^{d}\theta_i^2$	$\hat{\theta}_{\text{ridge}} = (\mathbb{X}^T\mathbb{X} + n\lambda I)^{-1}\mathbb{X}^T\mathbb{Y}$
LASSO	$\hat{\mathbb{Y}} = \mathbb{X}\theta$	Squared loss	L1	$\frac{1}{n}\|\|\mathbb{Y} - \mathbb{X}\theta\|\|_2^2 + \lambda\sum_{j=1}^{d}\|\theta_i\|$	**No closed form**

Expectation and Variance

Lecture 16, Data 100 Spring 2023

1465597

Continue on Regularization

- L2 Regularization (Ridge)
- Scaling Data for Regularization
- L1 Regularization (LASSO)

Random Variables and Distributions

Expectation and Variance

Sums of Random Variables

- Equality vs Identically Distributed vs. IID
- Properties of Expectation and Variance
- Covariance, Correlation

Bernoulli and Binomial Random Variables

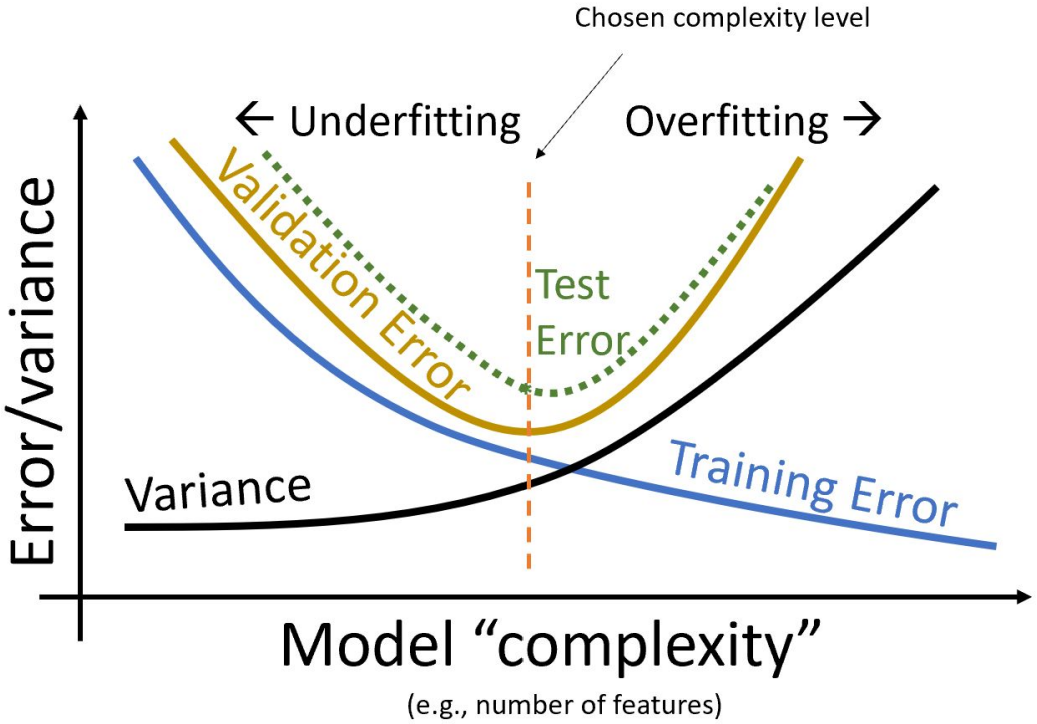
Sample Statistics

- Sample Mean
- Central Limit Theorem

The Bias-Variance Tradeoff

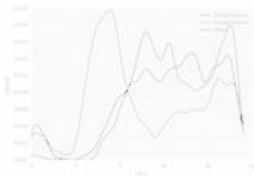
1465597

What is the mathematical underpinning of this plot?



We'll come back to this…

Why Probability?



Question & Problem Formulation

Data Acquisition

Prediction and Inference

Exploratory Data Analysis

Reports, Decisions, and Solutions

(today)

Model Selection Basics: Cross Validation Regularization	**Probability I:** Random Variables Estimators	**Probability II:** Bias and Variance Inference/Multicollinearity

We Will See . . .

1465597

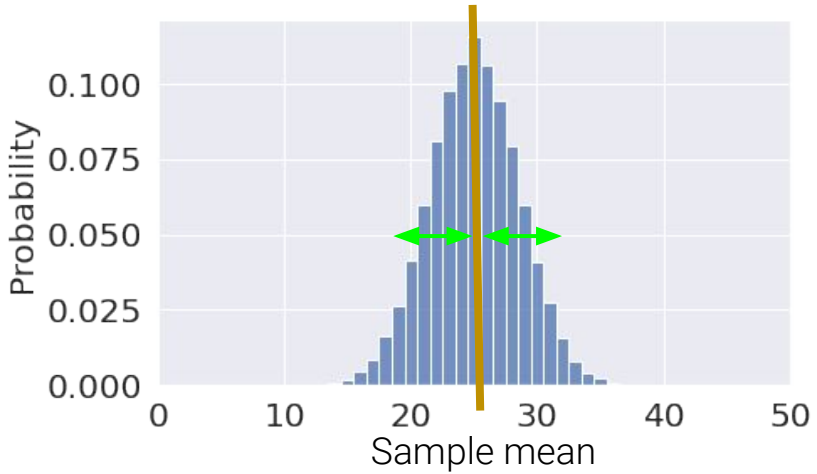
Formalize the notions of **sample statistic, population parameter** from Data 8.

From Data 8:

1. def **sample mean** - the mean of your random sample. `np.mean(data)`

2. The **Central Limit Theorem**: If you draw a large random sample with replacement, then, regardless of the population distribution, the probability distribution of the sample mean:

- Is roughly normal
- Is centered at the **population mean**
- Has an SD = $\frac{\text{population SD}}{\sqrt{\text{sample size}}}$



We Will See → One Important Probability Concept

1465597

Formalize the notions of **sample statistic, population parameter** from Data 8.

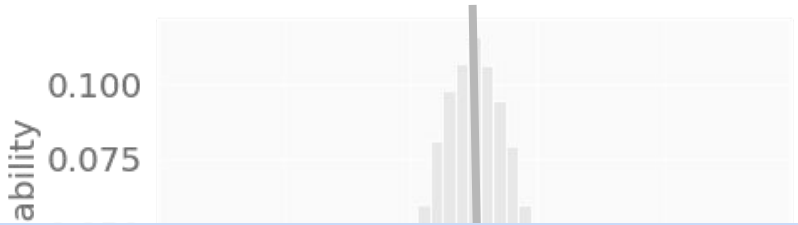
From Data 8:

1. def **sample mea**

Random Variable

2. The Central Limit Theorem: If you draw a **large random sample with replacement**, then, regardless of the population distribution, the **probability distribution** of the **sample mean**:

- Is roughly normal
- Is centered at the **population mean**
- Has an SD = $\frac{\text{population SD}}{\sqrt{\text{sample size}}}$



We will go over **just enough probability** to help you understand its implications for modeling.

For more probability, take STAT 140, EECS 70, and/or EECS 126.

[Terminology] Random Variable

1465597

Suppose we generate random data, like a random sample of size n from some population.

A **random variable** is a numerical function of the randomness

sample was drawn at random — value depends on how the sample came out

- Often denoted with uppercase "variable-like" letters (e.g. X, Y).
- It's value on any given draw is called a **realization.**
- Domain (input): all random samples of size n / all possible outcomes of our random process
- Range (output): number line.

[Terminology] Random Variable

1465597

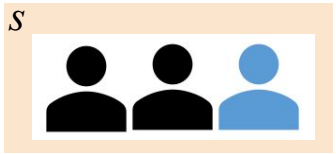
Suppose we generate random data, like a random sample of size n from some population.

A **random** **variable** is a numerical function of the randomness

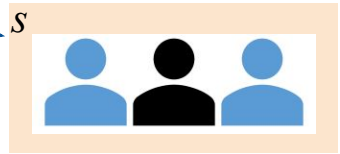
sample was drawn at random — value depends on how the sample came out

- Often denoted with uppercase "variable-like" letters (e.g. X, Y).
- It's value on any given draw is called a **realization.**
- Domain (input): all random samples of size n / all possible outcomes of our random process.
- Range (output): number line.

Suppose you draw a random sample s of size 3 from the following population:

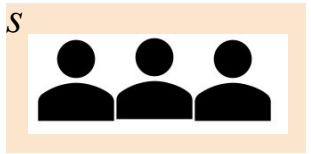


$X(s) = 1$



$X(s) = 2$

$X(s) = 2$



$X(s) = 0$

Define X **= # of blue people**.

X is a random variable!

From Population to Distribution

	X(s)
0	3
1	4
2	4
3	6
4	8
...	...
79995	6
79996	6
79997	4
79998	6
...	...

X(s) from all possible samples

$$P(X = x) = \frac{\#\text{ times where } X = x}{\text{pop. size}}$$

x	P(X = x)
3	0.1
4	0.2
6	0.4
8	0.3

Probability Distribution Table

[Terminology] Distribution

1465597

The **distribution** of a random variable X is a description of how the total probability of 100% is split over all the possible values of X.

A distribution fully defines a random variable.

x	P(X = x)
3	0.1
4	0.2
6	0.4
8	0.3

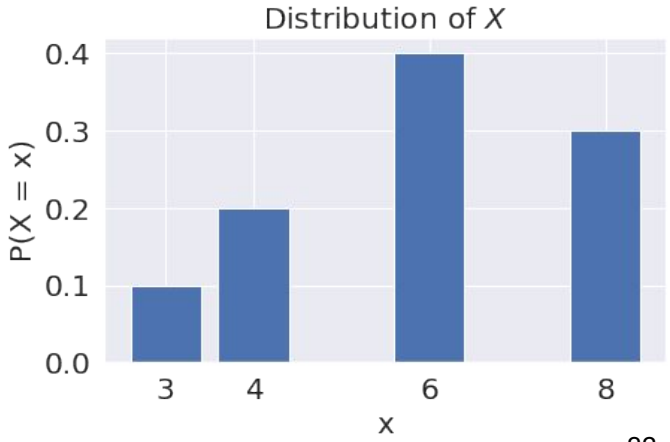
Assuming (for now) that X is discrete, i.e., has a finite number possible values:

$$P(X = x)$$

The probability that random variable X takes on the value x.

$$\sum_{\text{all } x} P(X = x) = 1$$

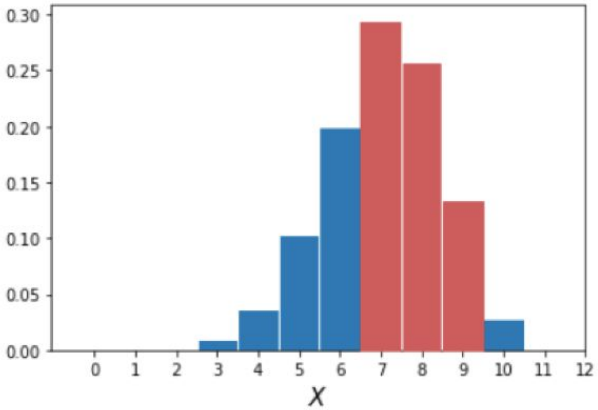
Probabilities must sum to 1.



Distributions Can Be Represented as Histograms or Densities

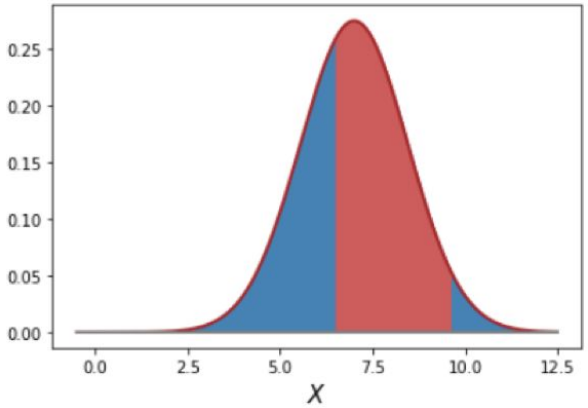
1465597

Distribution of **discrete** random variable X



The area of the red bars is **P(7 <= X <= 9)**.

Distribution of **continuous** random variable Y



The red area under the curve is **P(6.8 <= Y <= 9.5)**.

Take STAT 140 to learn more about discrete vs continuous distributions.

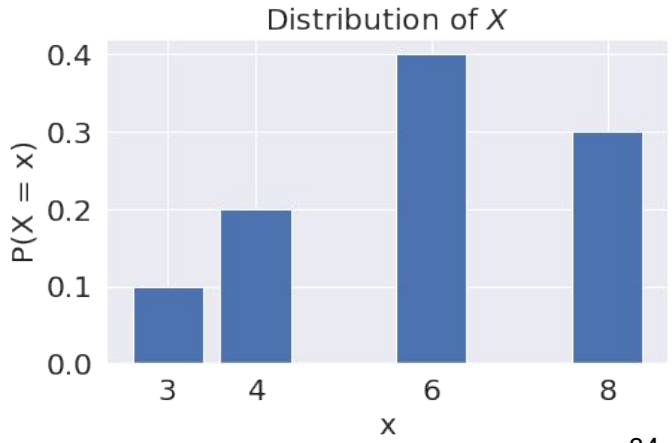
Understanding Random Variables

1465597

Compute the following probabilities for the random variable X.

x	P(X = x)
3	0.1
4	0.2
6	0.4
8	0.3

1. $P(X = 4) =$ 0.2
2. $P(X < 6) =$ $0.1 + 0.2 = 0.3$
3. $P(X \leq 6) =$ $0.1 + 0.2 + 0.4 = 0.7$
4. $P(X = 7) =$ 0
5. $P(X \leq 8) =$ 1



Common Random Variables

1465597

Bernoulli(p)

- Takes on value 1 with probability p, and 0 with probability 1 - p.
- AKA the “indicator” random variable.

Binomial(n, p)

- Number of 1s in 'n' independent Bernoulli(p) trials.

Uniform on a finite set of values

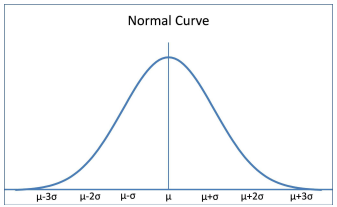
- Probability of each value is 1 / (size of set).
- For example, a standard/fair die.

Uniform on the unit interval(0, 1)

- Density is flat on (0, 1) and 0 elsewhere.

Normal(μ, σ^2)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



The numbers in parentheses are the **parameters** of a random variable, which are constants. Parameters define a random variable’s shape (i.e., distribution) and its values.

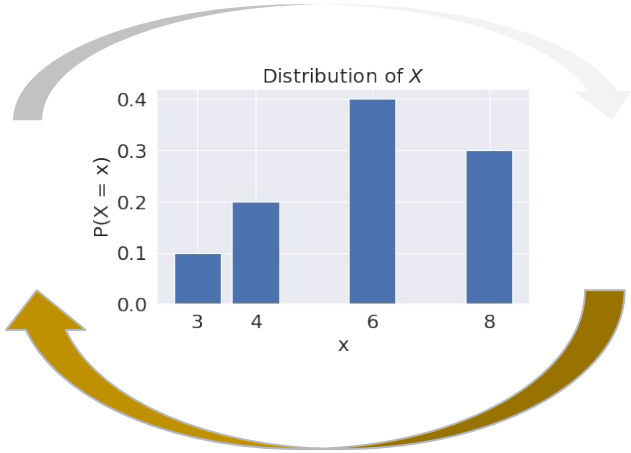
From Distribution to (Simulated) Population

1465597

?

X(s) from all possible samples?

$$P(X = x) = \frac{\#\text{ times where } X = x}{\text{pop. size}}$$



Given a random variable's distribution, how could we **generate/simulate** a population?

x	P(X = x)
3	0.1
4	0.2
6	0.4
8	0.3

Probability Distribution Table

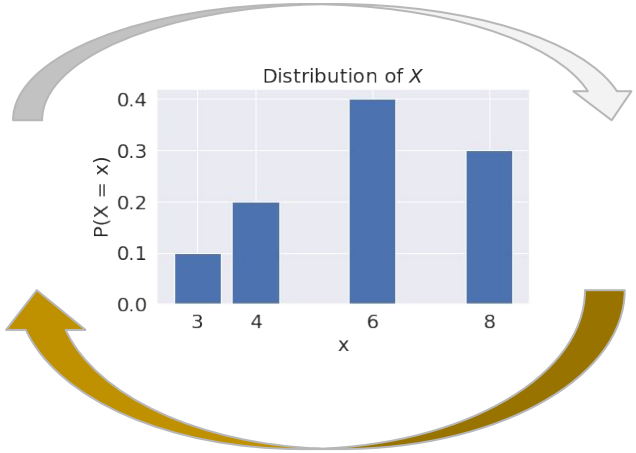
From Distribution to (Simulated) Population

1465597

	X(s)
0	3
1	4
2	4
3	6
4	8
...	...
79995	6
79996	6
79997	4
79998	6
...	...

X(s) from many, many **(simulated)** samples

$$P(X = x) = \frac{\#\text{ times where } X = x}{\text{pop. size}}$$



Simulate: Randomly pick values of X according to its distribution

`np.random.choice` or `df.sample`

x	P(X = x)
3	0.1
4	0.2
6	0.4
8	0.3

Probability Distribution Table

1465597

Expectation and Variance

Lecture 16, Data 100 Spring 2023

Continue on Regularization

- L2 Regularization (Ridge)
- Scaling Data for Regularization
- L1 Regularization (LASSO)

Random Variables and Distributions

Expectation and Variance

Sums of Random Variables

- Equality vs Identically Distributed vs. IID
- Properties of Expectation and Variance
- Covariance, Correlation

Bernoulli and Binomial Random Variables

Sample Statistics

- Sample Mean
- Central Limit Theorem

Descriptive Properties of Random Variables

1465597

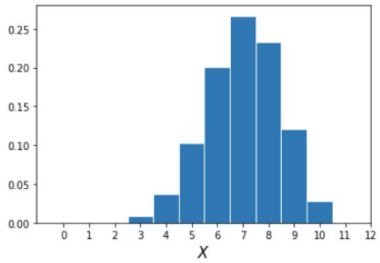
There are several ways to describe a random variable:

Table of all samples s , $X(s)$

Distribution Table

$P(X = x)$

Histogram



Definition (fully describes)

Properties (summarizes)

Random Variable X

Expectation

the "average value" of X

$\mathbb{E}[X]$

Variance

the "spread" of X

$\mathrm{Var}(X)$

The expectation and variance of a random variable are **numerical summaries** of X.

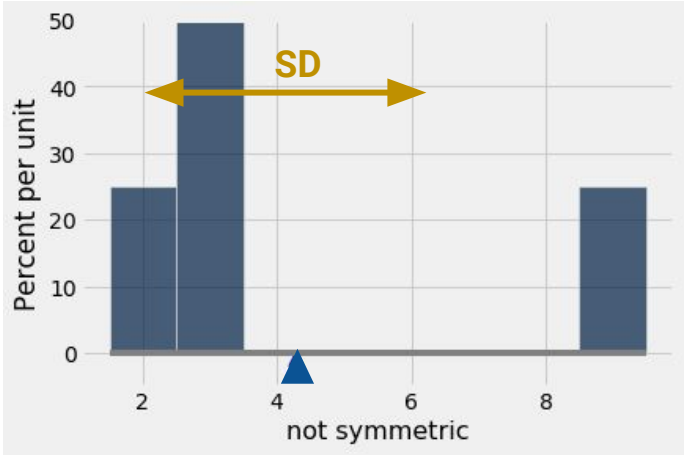
They are numbers and are not random!

You've Seen This Before

1465597

> The **mean** (Data 100: **expectation**) is the center of gravity or **balance point** of the **histogram** (Data 100: of a random variable). [textbook]

> The **variance** is a measure of spread.
> It is the **expected squared deviation from the mean** (Data 100: of a random variable). [textbook]



In Data 8, you computed these from the datapoints themselves (i.e., the sample of data).

In Data 100, we redefine these terms with respect to probability distributions.

Definition of Expectation

1465597

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.

Two equivalent ways to apply the weights:

1. One sample at a time: $$\mathbb{E}[X] = \sum_{\substack{\text{all samples} \\ s}} X(s)P(s)$$

2. One possible value at a time: $$\mathbb{E}[X] = \sum_{\substack{\text{all possible} \\ x}} xP(X = x)$$

More common (we are usually given the distribution, not all possible samples)

Expectation is a number, not a random variable!

- It is analogous to the **average** (same units as the random variable).
- It is the center of gravity of the probability histogram.
- It is the long run average of the random variable, if you simulate the variable many times.

Example

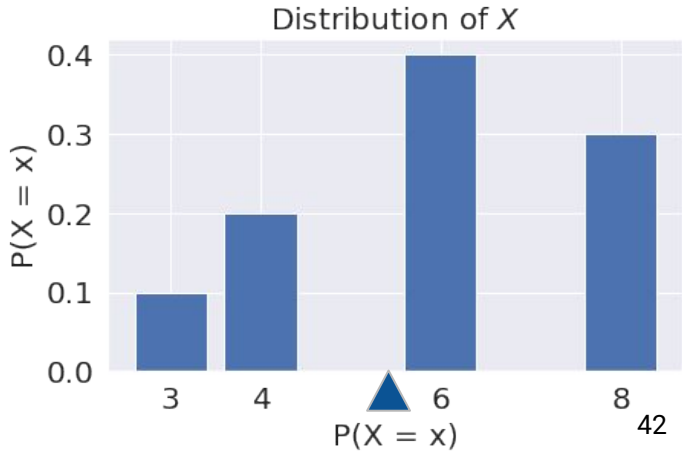
$$\mathbb{E}[X] = \sum_x xP(X = x)$$

1465597

Consider the random variable X we defined earlier.

x	P(X = x)
3	0.1
4	0.2
6	0.4
8	0.3

$$\begin{aligned}\mathbb{E}[X] &= \sum_x x \cdot P(X = x) \\ &= 3 \cdot 0.1 + 4 \cdot 0.2 + 6 \cdot 0.4 + 8 \cdot 0.3 \\ &= 0.3 + 0.8 + 2.4 + 2.4 \\ &= 5.9\end{aligned}$$



Note, $E[X] = 5.9$ is not a possible value of X!
It is an average.

The expectation of X does not need to be a value of X.

Definition of Variance

1465597

Variance is the **expected squared deviation from the expectation** of X.

$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

- The units of the variance are the square of the units of X.
- To get back to the right scale, use the **standard deviation** of X: $\text{SD}(X) = \sqrt{\text{Var}(X)}$

Variance is a number, not a random variable!

- The main use of variance is to **quantify chance error**. How far away from the expectation could X be, just by chance?

By Chebyshev's inequality (which you saw in Data 8, and which we won't prove here either):

- 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."

Variance: Alternate Calculation

There's a more convenient form of variance:

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

- Proof (involves expanding the square and properties of expectation/summations): link
- Useful in Mean Squared Error calculations
  - If X is centered (i.e. E[X] = 0), then $E[X^2] = Var(X)$
- When computing variance by hand, often used instead of definition.
- See STAT 140/EECS 126 for more on how to interpret this expression.

Dice Is the Plural; Die Is the Singular

1465597

Let X be the outcome of a single die roll.

X is a random variable.

$$P(X = x) = \begin{cases} 1/6 & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

1. What is the expectation, $E[X]$?

$$\mathbb{E}[X] = \sum_x xP(X = x)$$

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \end{aligned}$$

(definitions/properties)

2. What is the variance, $\mathrm{Var}(X)$?

slido

Let X be the outcome of a single fair die roll. What is the expectation of X (E[X])?

ⓘ Start presenting to display the poll results on this slide.

Dice Is the Plural; Die Is the Singular

1465597

Let X be the outcome of a single die roll.

X is a random variable.

$$P(X = x) = \begin{cases} 1/6 & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

1. What is the expectation, $E[X]$?

$$\mathbb{E}[X] = 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6)$$
$$= (1/6)(1 + 2 + 3 + 4 + 5 + 6) = \frac{7}{2}$$

2. What is the variance, $\mathrm{Var}(X)$?

$$\mathbb{E}[X] = \sum_x xP(X = x)$$
$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$
$$= \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

slido

Let X be the outcome of a single fair die roll. What is the variance of X (Var[X])?

ⓘ Start presenting to display the poll results on this slide.

Dice Is the Plural; Die Is the Singular

1465597

Let X be the outcome of a single die roll.

X is a random variable.

$$P(X = x) = \begin{cases} 1/6 & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

1. What is the expectation, $E[X]$?

$$\mathbb{E}[X] = 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6)$$
$$= (1/6)(1 + 2 + 3 + 4 + 5 + 6) = \frac{7}{2}$$

$$\mathbb{E}[X] = \sum_x xP(X = x)$$
$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$
$$= \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

2. What is the variance, $\text{Var}(X)$?

Approach 1: Definition

$$\text{Var}(X) = (1/6)\left((1 - 7/2)^2 + (2 - 7/2)^2 + (3 - 7/2)^2 + (4 - 7/2)^2 + (5 - 7/2)^2 + (6 - 7/2)^2\right)$$
$$= 35/12$$

Approach 2: Property

$$\mathbb{E}[X^2] = \sum_x x^2 P(X = x) \quad = 91/6$$

$$\text{Var}(X) = 91/6 - (7/2)^2 = 35/12$$

Sums of Random Variables

Lecture 16, Data 100 Spring 2023

1465597

Continue on Regularization

- L2 Regularization (Ridge)
- Scaling Data for Regularization
- L1 Regularization (LASSO)

Random Variables and Distributions

Expectation and Variance

Sums of Random Variables

- Equality vs Identically Distributed vs. IID
- Properties of Expectation and Variance
- Covariance, Correlation

Bernoulli and Binomial Random Variables

Sample Statistics

- Sample Mean
- Central Limit Theorem

Functions of Multiple Random Variables

1465597

A function of a random variable is also a random variable!

If you create multiple random variables based on your sample…

…then functions of those random variables are also random variables.

For instance, if $X_1, X_2, \ldots, X_n$ are random variables, then so are all of these:

$$X_n^2 \qquad \#\{i : X_i > 10\}$$

$$\max(X_1, X_2, \ldots, X_n) \qquad \frac{1}{n}\sum_{i=1}^{n}(X_i - c)^2$$

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

Many functions of RVs that we care about (**counts**, **means**) involve **sums of RVs**, so we expand on properties of sums of RVs.

Equal vs. Identically Distributed vs. IID

1465597

Suppose that we have two random variables X and Y.

X and Y are **equal** if:

- $X(s) = Y(s)$ for every sample s.
- We write $X = Y$.

Covered up to here on 3/9

X and Y are **identically distributed** if:

- The distribution of X is the same as the distribution of Y
- We say "X and Y are equal in distribution."
- If $X = Y$, then X and Y are identically distributed;
  but the converse is not true (ex: **Y = 7-X**, X is a die)

X and Y are **independent and identically distributed (IID)** if:

- X and Y are identically distributed, *and*
- Knowing the outcome of X does not influence your belief of the outcome of Y, and vice versa ("X and Y are independent.")
- Independence is covered more in STAT 140/EECS 70.
- In Data 100, you will never be expected to prove that RVs are IID.

IID RVs

The remaining slides will be covered in lecture 17

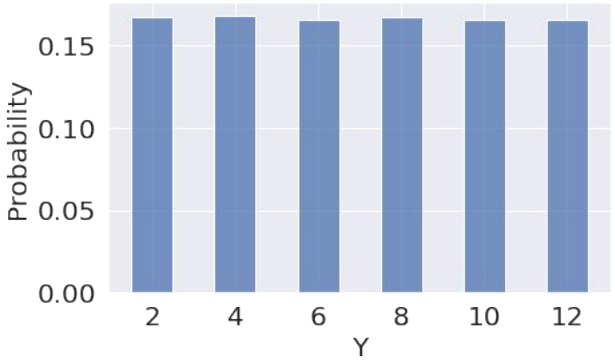
Distributions of Sums

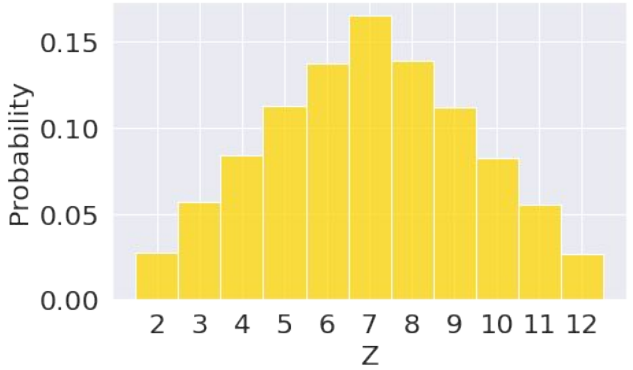
1465597

Let X_1 and X_2 be numbers on two rolls of a die.

- X_1, X_2 are **IID**, so X_1, X_2 have the same distribution.
- But the sums $Y = X_1 + X_1 = 2X_1$ and $Z = X_1 + X_2$ have different distributions!

Let's show this through simulation:





- Same expectation…
- But $Y = 2X_1$ has larger variance!

	Y	Z
E[•]	6.984400	6.984950
Var(•)	11.669203	5.817246
SD(•)	3.416021	2.411897

How can we directly compute $E[Y]$, $\mathrm{Var}(Y)$, **without** simulating distributions?

Demo

Properties of Expectation [1/3]

1465597

Instead of simulating full distributions, we often just compute expectation and variance directly.

Recall definitions of expectation:

$$\mathbb{E}[X] = \sum_{x} xP(X = x)$$

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

Properties:

1. **Expectation is linear**.
Intuition: summations are linear. Proof

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Properties of Expectation [2/3]

1465597

Instead of simulating full distributions, we often just compute expectation and variance directly.

Recall definitions of expectation:

$$\mathbb{E}[X] = \sum_{x} xP(X = x)$$

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

Properties:

1. **Expectation is linear**.
Intuition: summations are linear. Proof

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

2. Expectation is linear in sums of RVs, for any relationship between X and Y. Proof

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Properties of Expectation [3/3]

1465597

Instead of simulating full distributions, we often just compute expectation and variance directly.

Recall definitions of expectation:

$$\mathbb{E}[X] = \sum_x xP(X = x)$$

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

Properties:

1. **Expectation is linear**.
Intuition: summations are linear. Proof

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

2. Expectation is linear in sums of RVs,
for any relationship between X and Y. Proof

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

3. If g is a non-linear function, then in general

$$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$$

.

- Example: if X is *-1* or *1* with equal probability, then $E[X] = 0$ but $E[X^2] = 1 \neq 0$.

Properties of Variance [1/2]

1465597

Recall definition of variance: $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$

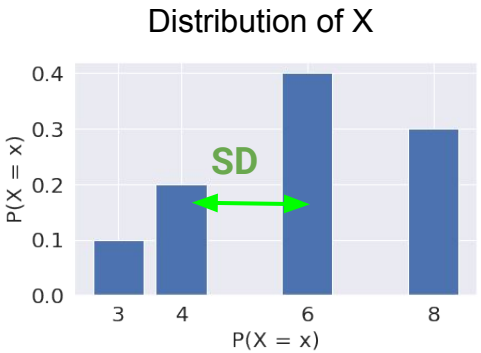
Properties:

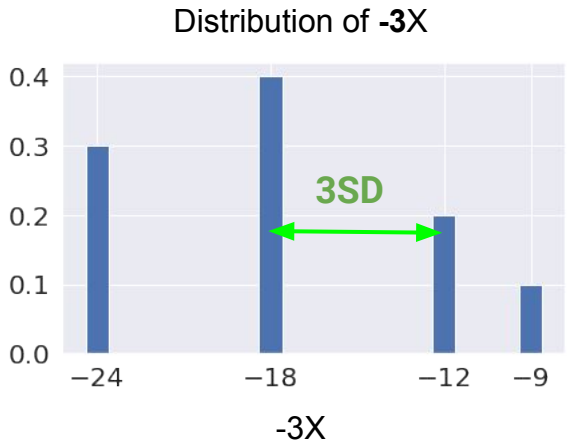
1. **Variance is non-linear**:

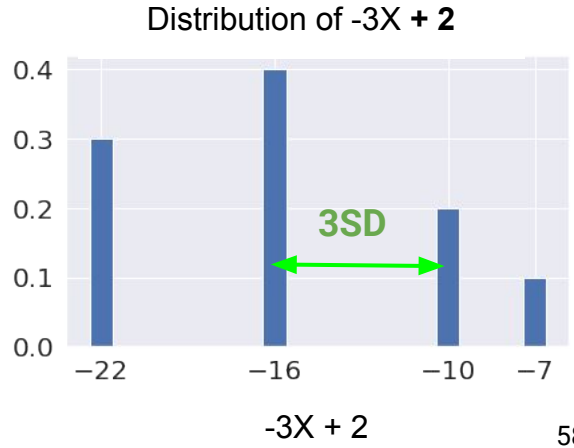
$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

Intuition (full proof): Consider the Standard Deviation for $Y = -3X+2$:

$$\mathrm{SD}(aX + b) = |a|\mathrm{SD}(X)$$







Properties of Variance [2/2]

1465597

Recall definition of variance: $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$

Properties:

1. **Variance is non-linear**:

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

Intuition (full proof): Consider the Standard Deviation for $Y = -3X+2$:

$$\mathrm{SD}(aX + b) = |a|\mathrm{SD}(X)$$

2. Variance of sums of RVs is affected by the (in)dependence of the RVs (derivation):

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X, Y)$$

Covariance of X and Y (next slide).
If X, Y independent,
then Cov(X, Y) = 0.

Covariance and Correlation: The Basics

1465597

Covariance is the expected product of deviations from expectation.

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

- A generalization of variance. Note $\mathrm{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{Var}(X)$.
- Interpret by defining **correlation** (yes, *that* correlation!):

$$r(X, Y) = \mathbb{E}\left[\underbrace{\left(\frac{X - \mathbb{E}[X]}{\mathrm{SD}(X)}\right)}_{\text{standard units of } X \text{ (link)}} \left(\frac{Y - \mathbb{E}[Y]}{\mathrm{SD}(Y)}\right)\right] = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)}$$

Correlation (and therefore covariance) measures a linear relationship between X and Y.

Covariance and Correlation: The Basics

1465597

Covariance is the expected product of deviations from expectation.

$$\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

- A generalization of variance. Note $\text{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \text{Var}(X)$.
- Interpret by defining **correlation** (yes, *that* correlation!):

$$r(X, Y) = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\text{SD}(X)}\right)\left(\frac{Y - \mathbb{E}[Y]}{\text{SD}(Y)}\right)\right] = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)}$$

standard units of X (link)

Correlation (and therefore covariance) measures a linear relationship between X and Y.

- If X and Y are correlated, then knowing X tells you something about Y.
- "X and Y are uncorrelated" is the same as "Correlation and covariance equal to 0"
- **Independent X, Y are uncorrelated**, because knowing X tells you nothing about Y.
- The converse is not necessarily true: **X, Y could be uncorrelated but not independent**.
- For more info, see extra slides + take STAT 140/EECS 70.

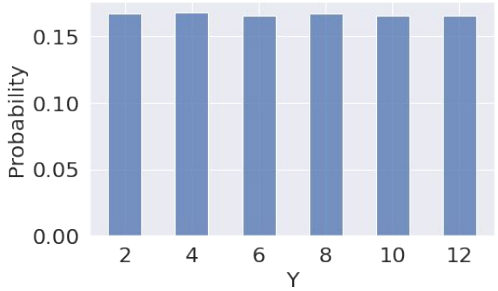
Dice, Our Old Friends: Expectation

1465597

Let X_1 and X_2 be numbers on two rolls of a die.

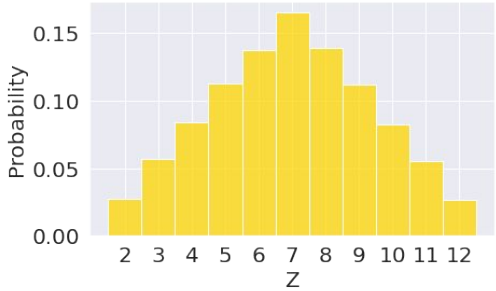
- X_1, X_2 are **IID**, so X_1, X_2 have the same distribution.
- Therefore $E[X_1] = E[X_2] = 7/2 \quad \text{Var}(X_1) = \text{Var}(X_2) = 35/12$

$Y = 2X_1$



$$E[Y] = E[2X_1] = 2E[X_1] = 7$$

$Z = X_1 + X_2$



$$E[Z] = E[X_1] + E[X_2] = (7/2) + (7/2) = 7$$

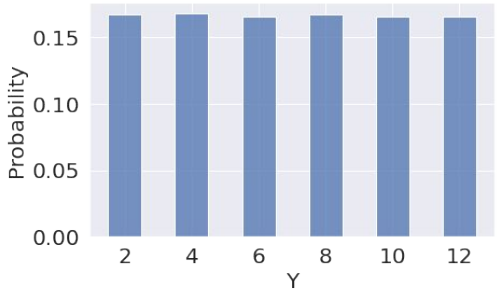
Dice, Our Old Friends: Variance

1465597

Let X_1 and X_2 be numbers on two rolls of a die.

- X_1, X_2 are **IID**, so X_1, X_2 have the same distribution.
- Therefore $E[X_1] = E[X_2] = 7/2$ $\quad \text{Var}(X_1) = \text{Var}(X_2) = 35/12$

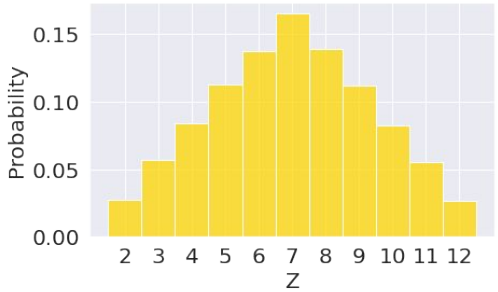
$Y = 2X_1$



$E[Y] = E[2X_1] = 2E[X_1] = 7$

$$\text{Var}(Y) = \text{Var}(2X_1) = 4\text{Var}(X_1)$$
$$= 4(35/12)$$
$$\approx \mathbf{11.67}$$

$Z = X_1 + X_2$



$E[Z] = E[X_1] + E[X_2] = (7/2) + (7/2) = 7$

$$\text{Var}(Z) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X1, X2)$$
$$= (35/12) + (35/12) + 0$$
$$\approx \mathbf{5.83}$$

0

X_1, X_2 independent

[Summary] Expectation and Variance for Linear Functions of Random Variables

1465597

Let X be a random variable with distribution $P(X = x)$.

$$\mathbb{E}[X] = \sum_x xP(X = x)$$

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \quad \text{(definition)}$$

$$= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \quad \text{(easier computation)}$$

Let a and b be scalar values.

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

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

Let Y be another random variable.

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \underbrace{2\mathrm{Cov}(X, Y)}$$

Zero if X, Y independent.

Bernoulli and Binomial Random Variables

Lecture 16, Data 100 Spring 2023

1465597

Continue on Regularization

- L2 Regularization (Ridge)
- Scaling Data for Regularization
- L1 Regularization (LASSO)

Random Variables and Distributions

Expectation and Variance

Sums of Random Variables

- Equality vs Identically Distributed vs. IID
- Properties of Expectation and Variance
- Covariance, Correlation

Bernoulli and Binomial Random Variables

Sample Statistics

- Sample Mean
- Central Limit Theorem

Common Random Variables

1465597

Bernoulli(p)

- Takes on value 1 with probability p, and 0 with probability 1 - p
- AKA the “indicator” random variable.

Binomial(n, p)

- Number of 1s in n independent Bernoulli(p) trials

Uniform on a finite set of values

- Probability of each value is 1 / (size of set)
- For example, a standard die

Uniform on the unit interval(0, 1)

- Density is flat on (0, 1) and 0 elsewhere

Normal(μ, σ^2)

We’ll now revisit these to solidify our understanding of expectation/variance.

Properties of Bernoulli Random Variables

$$\mathbb{E}[X] = \sum_x xP(X = x)$$
$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

Definitions

1465597

Let X be a **Bernoulli**(p) random variable.

- Takes on value 1 with probability p, and 0 with probability 1 - p.
- AKA the "indicator" random variable.

Expectation:

$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$$

We will get an average value of p across many, many samples

Variance:

$$\mathbb{E}[X^2] = 1^2 \cdot p + 0^2 \cdot (1 - p) = p$$
$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= p - p^2 = p(1 - p)\end{aligned}$$

Lower Var: p = 0.1 or 0.9
Higher Var: p close to 0.5

More info: google("plot x(1 - x)")

Properties of Binomial Random Variables

1465597

Let Y be a **Binomial**(n, p) random variable.

- Y is the number (i.e., count) of 1s in n independent Bernoulli(p) trials.
- Distribution of Y given by the binomial formula (Lecture 2).

We can write: $Y = \sum_{i=1}^{n} X_i$

A count is a **sum** of 0's and 1's.

- X_i is the indicator of success on trial *i*. X_i = 1 if trial i is a success, else 0.
- All X_i s are **IID** (independent and identically distributed) and **Bernoulli**(p).

Properties of Binomial Random Variables

1465597

Let Y be a **Binomial**(n, p) random variable.

- Y is the number (i.e., count) of 1s in n independent Bernoulli(p) trials.
- Distribution of Y given by the binomial formula (Lecture 2).

We can write: $Y = \sum_{i=1}^{n} X_i$

A count is a sum of 0's and 1's.

- X_i is the indicator of success on trial *i*. X_i = 1 if trial i is a success, else 0.
- All X_i s are **IID** (independent and identically distributed) and **Bernoulli**(p).

Expectation: $\mathbb{E}[Y] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np$

Variance: Because all X_i s are independent, $\mathrm{Cov}(X_i, X_j) = 0$ for all i, j.

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1-p)$$

1465597

Example

Suppose you win cash based on the number of heads you get in a series of 20 coin flips.

Let $X_i = 1$ if the i-th coin is heads,
0 otherwise

Which payout strategy would you choose?

Hint: Compare expectations and variances.

A. $Y_A = 10 \cdot X_1 + 10 \cdot X_2$

B. $Y_B = \left(\sum_{i=1}^{20} X_i \right)$

C. $Y_C = 20 \cdot X_1$

slido

Suppose you win cash based on the number of heads you get in a series of 20 coin flips.
Let Xi = 1 if the i-th coin is heads, 0 otherwise.

Which payout strategy would you choose?

ⓘ Start presenting to display the poll results on this slide.

Which Would You Pick?

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X, Y)$$
$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

1465597

Suppose you win cash based on the number of heads you get in a series of 20 coin flips.

Let X_1 , X_2 , ..., X_{20} be 20 **IID** Bernoulli(0.5) random variables.

- Since X_i s are independent: $\mathrm{Cov}(X_i, X_j) = 0$ for all i, j.
- Since X_i is Bernoulli(p = 0.5): $E[X_i] = p = 0.5$, $\mathrm{Var}(X_i) = p(1\text{-}p) = 0.25$.

Which payout strategy would you choose?

	A. $Y_A = \$10 \cdot X_1 + \$10 \cdot X_2$	**B.** $Y_B = \$\left(\sum_{i=1}^{20} X_i\right)$	**C.** $Y_C = \$20 \cdot X_1$
Expectation			
Variance			
Std. Deviation			

Which Would You Pick?

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X, Y)$$
$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

1465597

Suppose you win cash based on the number of heads you get in a series of 20 coin flips.

Let X_1 , X_2 , ..., X_{20} be 20 **IID** Bernoulli(0.5) random variables.

- Since X_i s are independent: $\mathrm{Cov}(X_i, X_j) = 0$ for all i, j.
- Since X_i is Bernoulli(p = 0.5): $E[X_i] = p = 0.5$, $\mathrm{Var}(X_i) = p(1\text{-}p) = 0.25$.

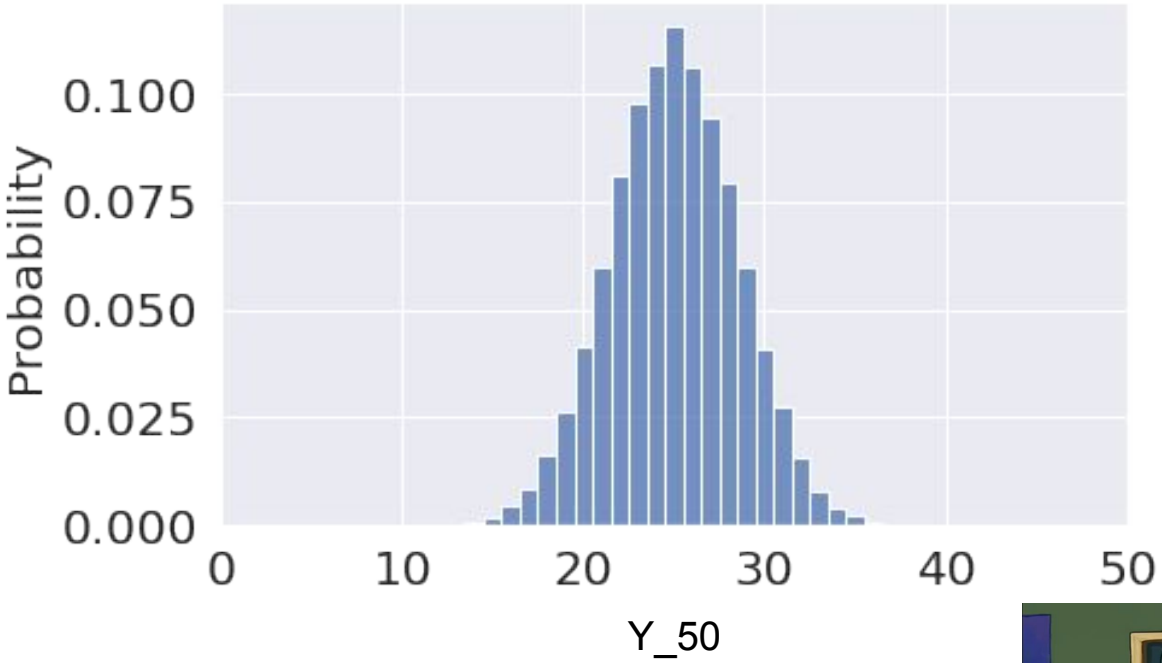
Which payout strategy would you choose?

	A. $Y_A = \$10 \cdot X_1 + \$10 \cdot X_2$	**B.** $Y_B = \$\left(\sum_{i=1}^{20} X_i\right)$	**C.** $Y_C = \$20 \cdot X_1$
Expectation	$E[Y_A] = 10(0.5) + 10(0.5) = 10$	$E[Y_B] = 0.5 + \ldots + 0.5 = 10$	$E[Y_C] = 20(0.5) = 10$
Variance	$\mathrm{Var}(Y_A) = 10^2(0.25) + 10^2(0.25) = 50$	$\mathrm{Var}(Y_B) = 0.25 + \ldots + 0.25 = 20(0.25) = 5$	$\mathrm{Var}(Y_C) = 20^2(0.25) = 100$
Std. Deviation	$\mathrm{SD}(Y_A) \approx \mathbf{7.07}$	$\mathrm{SD}(Y_B) \approx \mathbf{2.24}$	$\mathrm{SD}(Y_C) = \mathbf{10}$

Binomial(n, p) for Large n

1465597

For p = 0.5, n = 50 (i.e. number of heads in 50 fair coin flips):



Sample Statistics

Lecture 16, Data 100 Spring 2023

1465597

Continue on Regularization

- L2 Regularization (Ridge)
- Scaling Data for Regularization
- L1 Regularization (LASSO)

Random Variables and Distributions

Expectation and Variance

Sums of Random Variables

- Equality vs Identically Distributed vs. IID
- Properties of Expectation and Variance
- Covariance, Correlation

Bernoulli and Binomial Random Variables

Sample Statistics

- Sample Mean
- Central Limit Theorem

From Populations to Samples

1465597

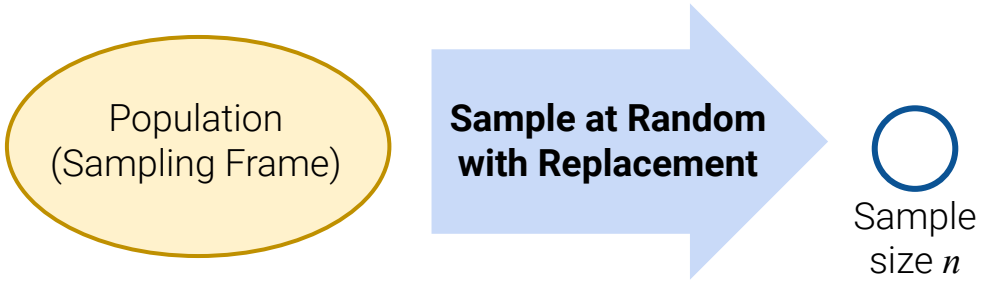
Today, we've talked extensively about **populations**:

- If we know the **distribution of a random variable**, we can reliably compute expectation, variance, functions of the random variable, etc.

However, in Data Science, we often collect **samples**.

- We don't know the distribution of our population.
- We'd like to use the distribution of your sample to estimate/infer properties of the population.

The **big assumption** we make in modeling/inference:



The Sample is a Set of IID Random Variables

1465597

Population (Sampling Frame)

Sample at Random with Replacement

Sample size n

x	P(X = x)
3	0.1
4	0.2
6	0.4
8	0.3

or

	X(s)
0	3
1	4
2	4
3	6
4	8
...	...
79995	6
79996	6
79997	4
79998	6
79999	6
...	...

Population
(really large N)

`df.sample(n, replace=True)`
[documentation]

	X
0	6
1	8
2	6
3	6
4	3
...	...
95	8
96	6
97	6
98	3
99	8

Each observation in our sample is a **Random Variable** drawn **IID** from our population distribution.

Sample
(n << N)

$X_1, X_2, \ldots, X_n$

The Sample is a Set of IID Random Variables

1465597

Population (Sampling Frame)

Sample at Random with Replacement

Sample size n

x	P(X = x)
3	0.1
4	0.2
6	0.4
8	0.3

or

	X(s)
0	3
1	4
2	4
3	6
4	8
...	...
79995	6
79996	6
79997	4
79998	6
79999	6
...	...

$E[X] = 5.9$

Population Mean
A **number**,
i.e., fixed value

`df.sample(n, replace=True)`
[documentation]

	X
0	6
1	8
2	6
3	6
4	3
...	...
95	8
96	6
97	6
98	3
99	8

Sample Mean
A **random variable**!
Depends on our randomly drawn sample!!

`np.mean(...) = 5.71`

Sample $X_1, X_2, \ldots, X_n$

[Terminology] Sample Mean

1465597

Consider an IID sample $X_1, X_2, \ldots, X_n$ drawn from a population with mean μ and SD σ.

Define the **sample mean**:

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

Expectation:

$$\begin{aligned}\mathbb{E}[\bar{X}_n] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \\ &= \frac{1}{n}(n\mu) = \mu\end{aligned}$$

Variance/Standard Deviation:

$$\begin{aligned}\mathrm{Var}(\bar{X}_n) &= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\underbrace{\left(\sum_{i=1}^{n}\mathrm{Var}(X_i)\right)}_{\text{IID} \rightarrow \mathrm{Cov}(X_i, X_j) = 0} \\ &= \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}\end{aligned}$$

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

Distribution?

$\bar{X}_n$ is **normally distributed** by the **Central Limit Theorem**.

Central Limit Theorem

1465597

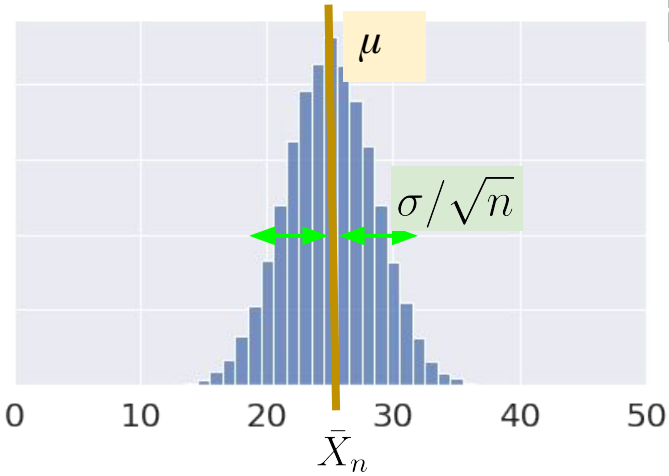
No matter what population you are drawing from:

If an IID sample of size n is large,

the probability distribution of the **sample mean**

is **roughly normal** with mean μ and SD $\sigma/\sqrt{n}$.

(STAT 140/EECS 126) (previous slide)



Any theorem that provides the rough distribution of a statistic
and **doesn't need the distribution of the population** is valuable to data scientists.

- Because we rarely know a lot about the population!

For a more in-depth demo: https://onlinestatbook.com/stat_sim/sampling_dist/

How Large Is “Large”?

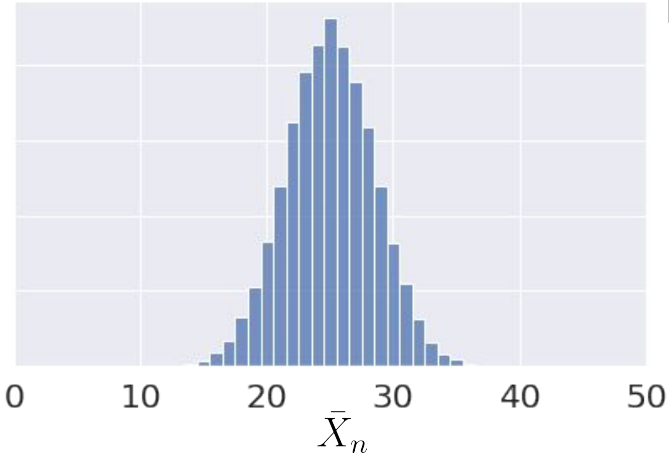
1465597

No matter what population you are drawing from:

If an IID **sample of size n is large**,

the probability distribution of the sample mean

is **roughly normal** with mean μ and SD $\sigma/\sqrt{n}$.



How large does n have to be for the normal approximation to be good?

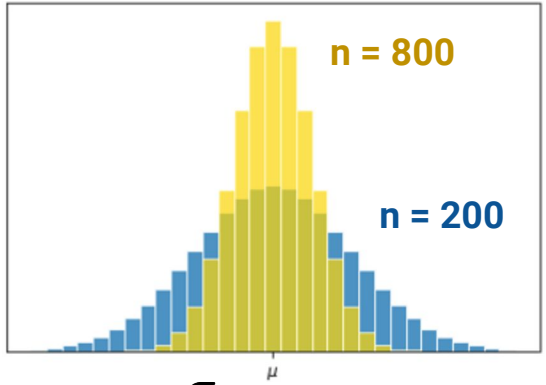
- …It depends on the shape of the distribution of the population…
- If population is **roughly symmetric and unimodal**/uniform, could need as few as **n = 20**. If population is very skewed, you will need bigger n.
- If in doubt, you can bootstrap the sample mean and see if the bootstrapped distribution is bell-shaped.

Accuracy and Spread of the Sample Mean

1465597

Our goal is often to **estimate** some characteristic of a population.

- Example: average height of Cal undergraduates.
- We typically can collect a **single sample**. It has just one average.
- Since that sample was random, it *could have* come out differently.



We should consider the **average value and spread** of all possible sample means, and what this means for how big n should be.

$$\mathbb{E}[\bar{X}_n] = \mu$$

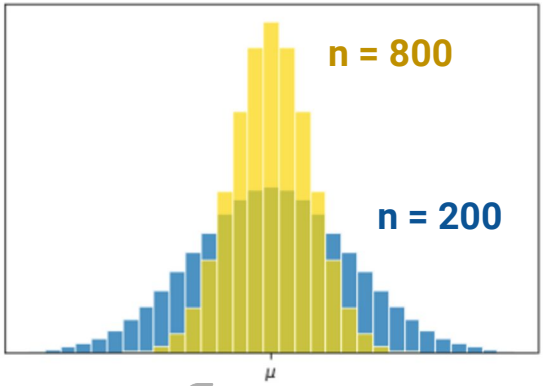
$$\text{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

Accuracy and Spread of the Sample Mean

1465597

Our goal is often to **estimate** some characteristic of a population.

- Example: average height of Cal undergraduates.
- We typically can collect a **single sample**. It has just one average.
- Since that sample was random, it *could have* come out differently.



We should consider the **average value and spread** of all possible sample means, and what this means for how big n should be.

$$\mathbb{E}[\bar{X}_n] = \mu$$

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

For every sample size, the expected value of the sample mean is the population mean.

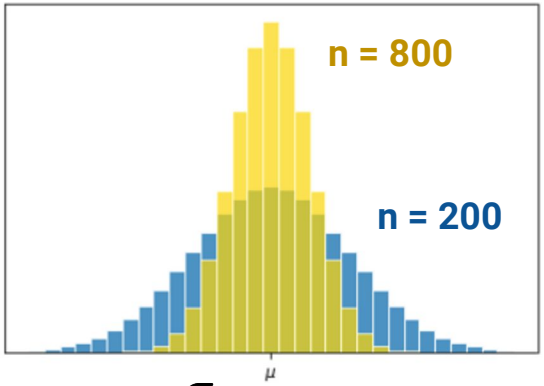
We call the sample mean an **unbiased estimator** of the population mean. (more in next lecture)

Accuracy and Spread of the Sample Mean

1465597

Our goal is often to **estimate** some characteristic of a population.

- Example: average height of Cal undergraduates.
- We typically can collect a **single sample**. It has just one average.
- Since that sample was random, it *could have* come out differently.



We should consider the **average value and spread** of all possible sample means, and what this means for how big n should be.

$$\mathbb{E}[\bar{X}_n] = \mu$$

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

For every sample size, the expected value of the sample mean is the population mean.

We call the sample mean an **unbiased estimator** of the population mean. (more in next lecture)

Square root law (Data 8): If you increase the sample size by a factor, the SD decreases by the square root of the factor.

The sample mean is more likely to be close to the population mean if we have a larger sample size.

Have a Normal Day!

1465597

[Extra Slides] Derivations

Lecture 16, Data 100 Spring 2023

Random Variables and Distributions

- Expectation and Variance
- Equality vs Identically Distributed
- Common RVs: Bernoulli, Binomial

Functions of Random Variables

- Distributions through Simulation, I.I.D.
- Properties of Expectation and Variance
- Covariance, Correlation
- Standard Units

Sample Statistics

- Sample Mean
- Central Limit Theorem

Standardization of random variables

Jump back: link

1465597

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

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:

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

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

$$\mathbb{E}(X_{su}^2) = \mathbb{V}ar(X_{su}) = 1$$

You should prove these facts yourself.

Variance: An Alternate Derivation

Jump back: link

1465597

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.

$$\begin{aligned} Var(X) &= E((X - E(X))^2) \\ &= E\big(X^2 - 2XE(X) + (E(X)^2)\big) \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 \\ &= E(X^2) - (E(X))^2 \end{aligned}$$

Properties of Expectation #1

Jump back: link

Recall definition of expectation: $\mathbb{E}[X] = \sum_x xP(X = x)$

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

1. **Expectation is linear**:

(intuition: summations are linear)

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Proof:

$$\begin{aligned}\mathbb{E}[aX + b] &= \sum_x (ax + b)P(X = x) = \sum_x \big(axP(X = x) + bP(X = x)\big) \\ &= a\sum_x xP(X = x) + b\sum_x P(X = x) \\ &= a\mathbb{E}[X] + b \cdot 1\end{aligned}$$

Properties of Expectation #2

Jump back: link

Recall definitions of expectation: $\mathbb{E}[X] = \sum_x xP(X = x)$

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

3. **Expectation is linear in sums of RVs**:

For any relationship between X and Y.

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Proof:

$$\mathbb{E}[X+Y] = \sum_s (X+Y)(s)P(s) = \sum_s \big(X(s)+Y(s)\big)P(s)$$

$$= \sum_s \big(X(s)P(s) + Y(s)P(s)\big)$$

$$= \sum_s X(s)\,P(s) + \sum_s Y(s)P(s)$$

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

Properties of Variance #1

Jump back: link

1465597

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

In order to compute Var(aX + b), consider:

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

Then,

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

In summary:

$$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$$

$$\mathbb{V}ar(aX+b) = a^2\mathbb{V}ar(X)$$

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

Don't forget the absolute values and squares!

Properties of Variance #2

Jump back: link

1465597

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 Var(X + Y) to see what's going on.

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

$$
\begin{aligned}
Var(X+Y) &= E\big[(X+Y-E(X+Y))^2\big] \\
&= E\big[((X-\mu_x)+(Y-\mu_y))^2\big] \\
&= E\big[(X-\mu_x)^2 + 2(X-\mu_x)(Y-\mu_y) + (Y-\mu_y)^2\big] \\
&= E\big[(X-\mu_x)^2\big] + E\big[(Y-\mu_y)^2\big] + 2E\big[(X-\mu_x)(Y-\mu_y)\big] \\
&= Var(X) + Var(Y) + 2E\big[(X-E(X))(Y-E(Y))\big]
\end{aligned}
$$

By the linearity of expectation, and the substitution.

We see

Addition rule for variance

1465597

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.

LECTURE 16

Random Variables

Narges Norouzi and Lisa Yan

Content credit: Acknowledgments