|  |  |  |
| --- | --- | --- |
|  |  |  |
| 2 | -3 | 9 |
| 4 | -1 | 1 |
| 6 | 1 | 1 |
| 8 | 3 | 9 |

Standard Deviation / Variance / Mean Calculation Example 1:

Consider the data below. Assume this data represents the entire population rather than being a subset of the population. Calculate the variance and standard deviation for the feature “x”.

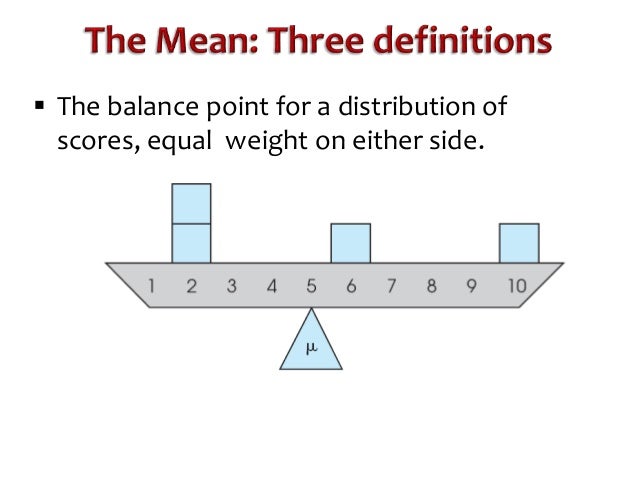
Mean:

Variance

Standard Deviation

**Interpretation:**

**Mean:** A mean can be interpreted as a center of mass. Think of a balancing point. Imagine all the numbers in the dataset are sitting on a number line and the weight of each is it’s value. For example, when then the weight of that datapoint is equal to 2. The mean would be the point on the number line where the weight of each side is equal.



**Variance:** The average distance from the mean is the variance. This measures how spread out the data is. We can calculate a distance by looking at the difference and then adding all of these distances up. But….there is a problem in doing so, which is that positive differences (values above the mean) would cancel with negative differences (values below the mean). We fix this by squaring the difference, which is always positive. . Now, let’s average them by adding them up and dividing by N. , which is the variance.

**Standard Deviation**: Let’s just unsquared the difference by taking the square root. The interpretation is similar to the variance.

**Exercises:**

Calculate the variance, standard deviation and mean for the data below

**Problem 1:**

X = [5, 9, 4, 8, 6, 10]

**Problem 2:**

X = [2, 10, 50, 25, 21, 14]

**Problem 3:**

X = [5, 9, 4, 8, 6, 1000]

**Problem 4:** Compare your results from Problem 1 against your results from Problem 4. Do the calculated statistics represent the data well? Notice that the data supporting Problem 1 and 3 are identical except for one value. How do extreme values impact the mean, variance and standard deviation?

Check your answers in Python:

import numpy as np

X = np.array([2, 10, 50, 25, 21, 14])

print( 'mean =' + str(np.mean(X)) + '  sd = ' + str(np.std(X)) + '   var=' + str(np.var(X)))

SPOILER: ANSWERS on NEXT PAGE. IF YOU LOOK AT THE ANSWERS WITHOUT FIRST ANSWERING THESE YOURSELF THEN YOU ARE HURTING YOUR CHANCES AT GETTING A GOOD QUIZ SCORE.

**Answers:**

Problem 1:

Problem 2:

Problem 3:

Problem 4: The mean in Problem 3 is rather far away from most of the data. It does not seem to represent any of the data well, which is all below the value of 10, except a single number. Whereas the Mean in Problem 1 is a much closer representation of the data set. The lesson: extreme values in our data can make the mean far less “mean”ingful. 😊

# Predictive Analytics:

The goal is to make accurate predictions of a target variable . In our class examples, we tried to predict house prices. There are many different types of predictive models. I introduced you to Linear Regression Models, which is a class of models that takes the form:

In the model above there are predictor variables. Model complexity can be increased by adding more predictors to the model.

There are many packages in Python that can do linear regression. We used SkLearn. The first step is to import the package

**from** sklearn.linear\_model **import** LinearRegression

Next, we initialize a linear regression object and assign it whatever alias we want. In this case I call it “lm”

lm = LinearRegression()

We then fit the model, which means that we estimate the coefficient values for all ’s, which gives us .

lm.fit(X,Y)

Then we combine these estimated coefficients with our data for to give us the predicted value of which we denote . A plot of against gives us our regression line.

Yhat=lm.predict(X)

Now we can measure the fitness of our model by measuring how close the errors are to the actual values of . The most common metric is Mean Squared Error, defined as:

In Python: Import mse from sklearn and then calculate mse.

from sklearn.metrics import mean\_squared\_error

mse = mean\_squared\_error(Y, Yhat)

Models with a lower MSE make more accurate predictions.

**For example:** Does model A or model B make better predictions? MSE(model A) = 5679 and MSE(model B)= 9875? Clearly, model A out performs model B in this case.

We can also plot the distribution for against the distribution of . This visually shows us how accurate our model is at different values of .

**For example:** The figure below shows that this model overestimates values of homes in the 200K range and underestimates homes in the upper range above 700K. We can see this by noting that the blue line (fitted values) is skewed to the right of the red line (actual values) in the 200K range and the blue line ends around the 700K mark even though the red line extends to 800K.

Chart, line chart

Description automatically generated

**Example Exam Question:**

Open the following file on my Github 🡪 mknomics 🡪 teaching repo 🡪 “Final Exam Practice.ipynb”

Question: Which model provides the best in-sample fit and why? Model 1, Model 2, or Model 3. ***Answers that do not provide support (comparing MSE’s) receive no credit.***

# Dummy Variable Creation/Interpretation:

Open the following file on my Github 🡪 mknomics 🡪 teaching repo 🡪 “Final Exam Practice.ipynb”

Go to the section: “Interpret Scatter Plot”

Consider the regression:

Where is a dummy variable equal to 1 if FullBath=3 and equal to zero otherwise.

Example Question: Does the scatter plot hint that this specification might produce a statistically significant value for ? Support your conclusion. Answers not providing support receive no credit.

Write your own Python code that estimates the equation above, report your findings on . Are they statistically significant and if so then give an interpretation.

**Reference Dummy\_Variables.ipynb in the teaching repo for a parallel example.**

**Consider Model 2 in “Final Exam Practice.ipynb”**

Interpret the coefficient on Abnorml.

Interpret the coefficient on the interaction Abnorml\_SF

Interpret the coefficient on WoodDeckSF. Make a decision based on your interpretation. You are considering selling a house that currently does not have a deck. Installing a new deck has the potential to boost the Sale Price that you get for the house. Let’s say installing a new deck costs $65 per square foot. Should you install it or not? Answers without proper statistical support will not receive credit.

**You build a model:**

Consider the month the house was sold. The value range is from 1 to 12. Why might you consider including this variable in your regression? How would you put it into your regression model? Define your model and be very explicit in defining your variables so that I clearly understand how they enter into your equation.