NYU Center for Data Science

DS-UA 111
Data Science for Everyone

Week 10: Lecture 2
Error Probabilities

NYU Center for Data Science

Could we determine an approach to hypothesis testing with numbers instead of visualizations?

DS-UA 111
Data Science for Everyone

Week 10: Lecture 2
Error Probabilities

Adapted from Adhikari, DeNero, Wagner, Milner

Announcements

- Please check Week 10 agenda on NYU Classes
  - Homework 3/4
  - Lab 6
- Refer to the Calendar linked to NYU Classes
  - Instructor Office Hours set for Friday 8-10AM EST

Review

- Functions
  - Keywords
    - def
    - return
  - Input
    - argument
    - parameter
  - Output
    - multiple outputs

References

- Functions:
  - Chapter 3.3
  - Chapter 8

Review

- Call Expression
  - We perform many operations in Python. Sometimes we want to give these operations a name
  - We can give a name to a block of code with a function
  - Python provides many functions for us

Python has a function for us to compute absolute value

```
abs(-10)
```

Can you guess the result?

Review

- Call Expression
  - We perform many operations in Python. Sometimes we want to give these operations a name
  - We can give a name to a block of code with a function
  - Python provides many functions for us

We can write our own function for absolute value

```
def absolute_value(x):
    if x >= 0:
        return x
    else:
        return -x

absolute_value(-10)
```

Do we have the same result?

Review

- Call Expression
  - We perform many operations in Python. Sometimes we want to give these operations a name
  - We can give a name to a block of code with a function
  - Python provides many functions for us

We can write our own function for absolute value

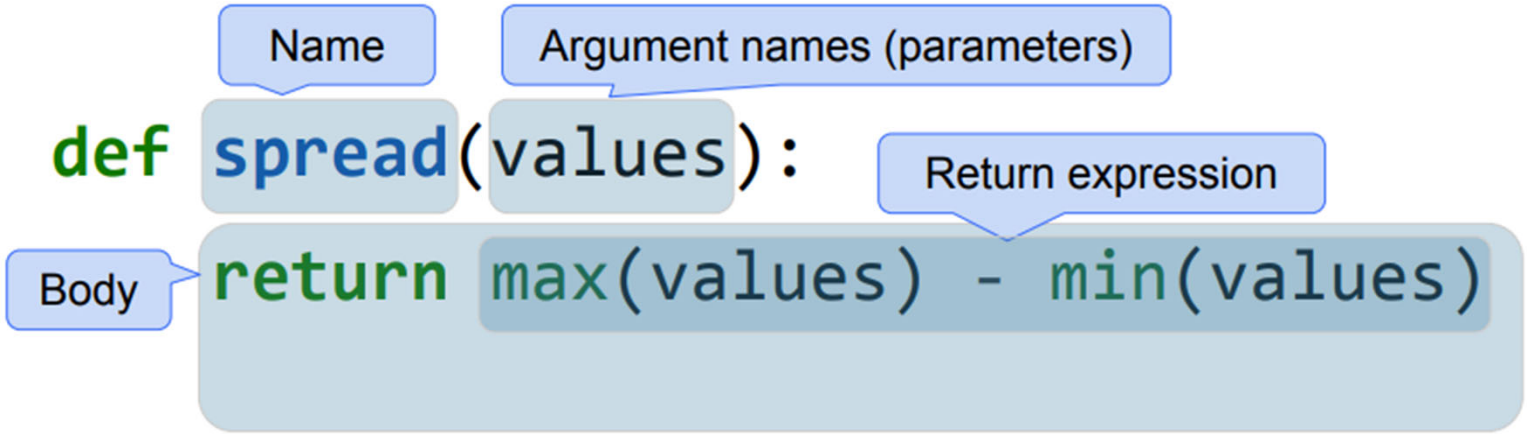
```
def absolute_value(x):
    if x >= 0:
        return x
    else:
        return -x

absolute_value(-10)
```

Do we have the same result?

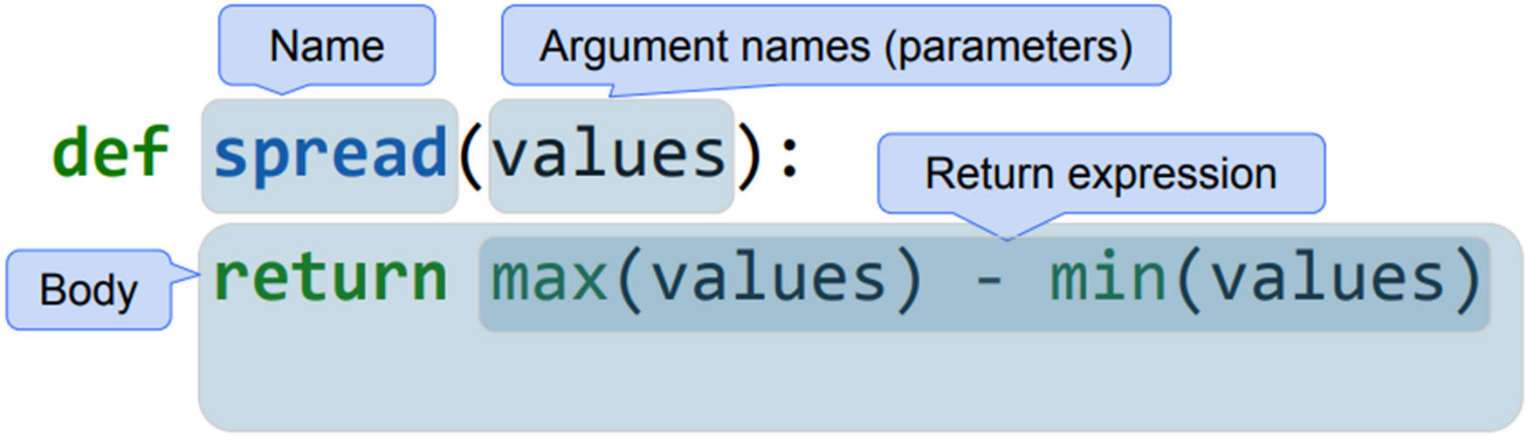
Review

- Keywords
  - We can spot functions by the keyword `def`
  - Think of an abbreviation for definition.
  - We need the keyword to indicate that the subsequent block of code defines the function



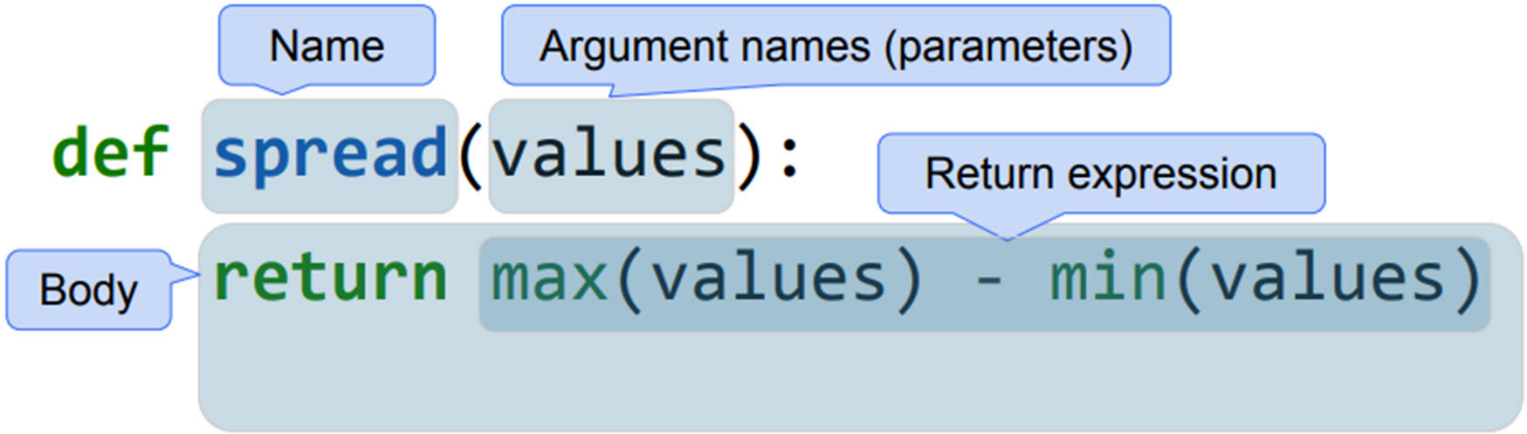
Review

- Format
  - The name of the function appear after the keyword def
  - The line ends with a colon
- The subsequent block of code needs to have indents at each line. Hit the TAB key.



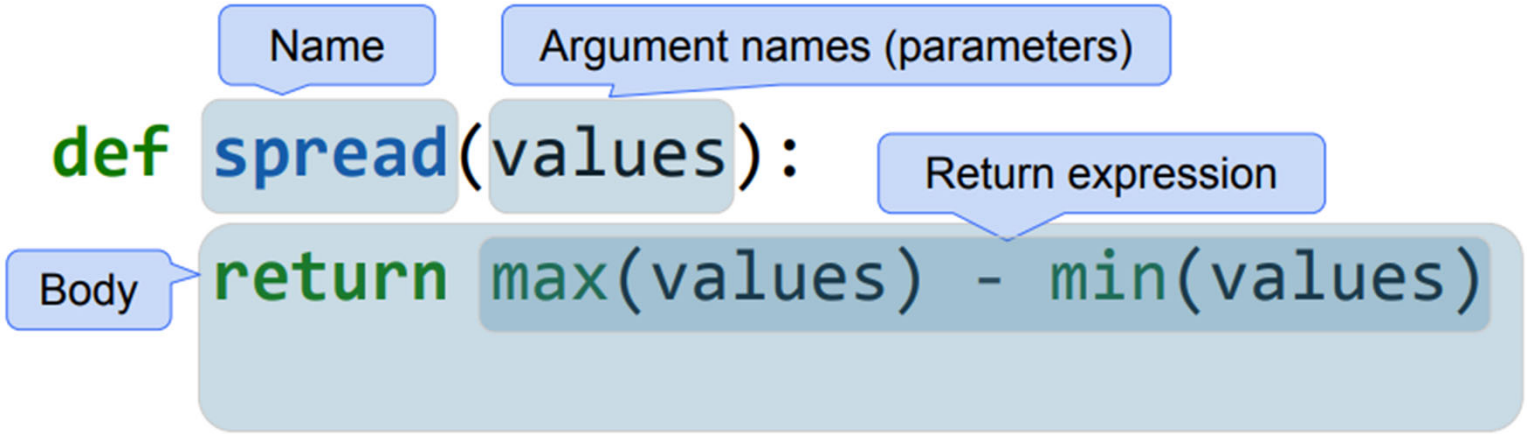
Review

- Input / Output
  - The input of the function appears within parentheses
  - Functions may not need an output
- If we want the function to output something, then we use the keyword **return**



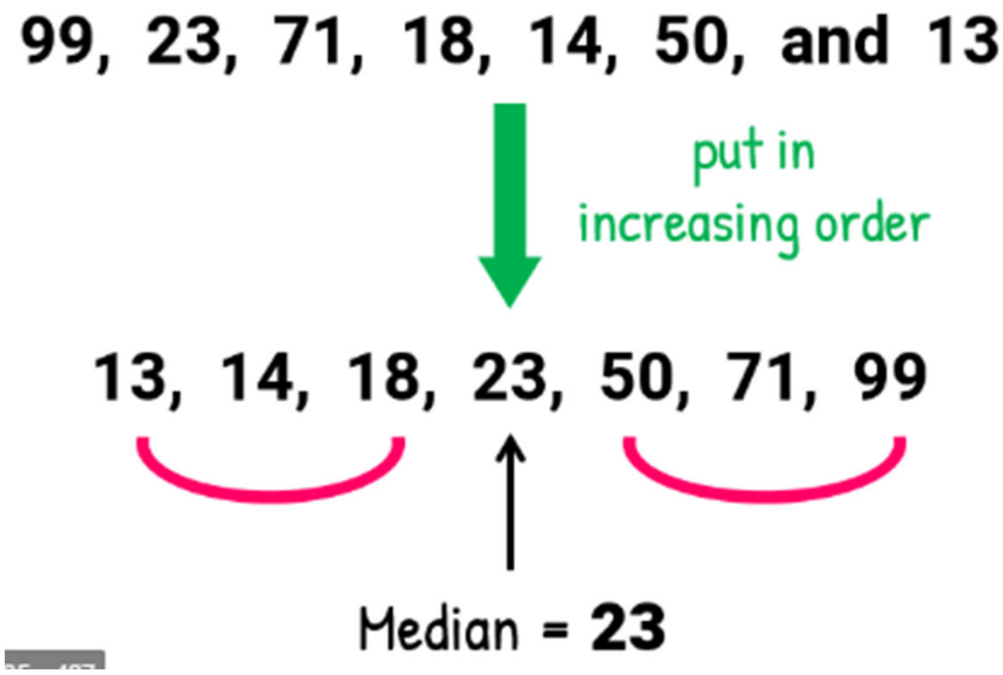
Review

- Input
  - The variables used in the block of code for the function are parameters
  - Think of place-holders
- We assign values to the parameters by calling the function. These values are arguments. Think lists, numbers, strings,…



Exercise

- Suppose we want to summarize the grades of students on an assignment.
- How can we find the middle number? Could we write a function to determine it?



Agenda

- Comparing Distributions
  - Statistics for goodness of fit
- Error Probabilities
  - Observed Significance Levels

References

- Hypothesis Testing
  - Chapters 11.3,11.4

Hypothesis Testing

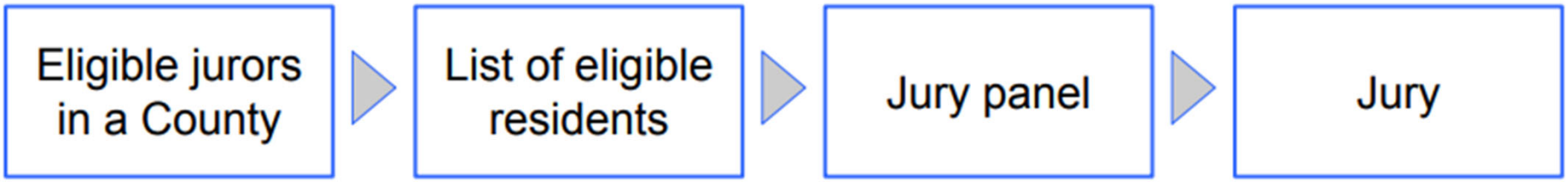
- Suppose we have a sample that might come from randomly sampling a population.
- Assuming we have a guess about the probability distribution for the population, then we can simulate random sampling

Hypothesis Testing

- Step 1: Hypotheses
- Step 2: Statistic
- Step 3: Probability Distribution

Hypothesis Testing

- If we can associate a statistic to the distributions, then we can compare the statistic of the sample to the empirical distribution of the statistics simulated from the population



Statistics for Multiple Categories

- Step 1
  - Take the difference between the proportions corresponding to each category
- Step 2
  - Apply absolute value transformation to obtain positive numbers
- Step 3
  - Add the transformed numbers. Divide the summation by 2.

Total Variation Distance

- Step 1: Differences
- Step 2: Absolute Value
- Step 3: Summation

Validating Hypotheses

- We need to choose between
  - null hypothesis
  - alternative hypothesis
- We compare observed statistic to its empirical distribution under the null hypothesis
- We needed the assumptions behind the null hypothesis to simulate the random samples from the population
- If the observed statistic seems reasonable then we accept the null hypothesis
- How can we check the consistency of a value with a distribution?

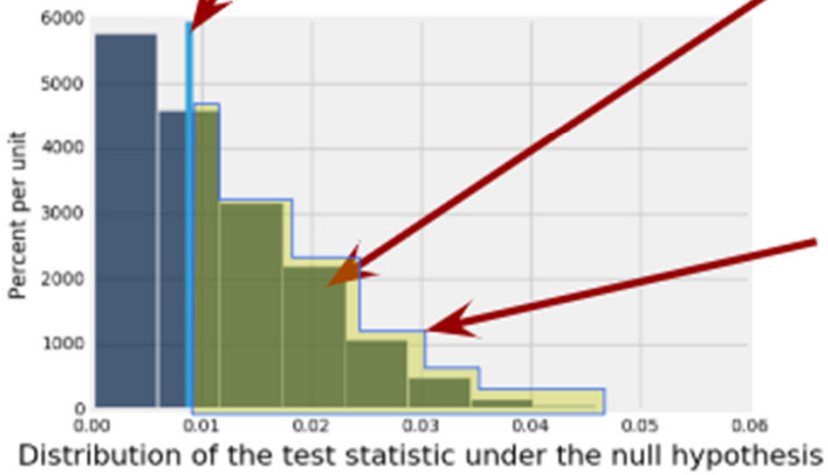
Validating Hypotheses

- While a histogram can be helpful to judge the consistency of the value, we can use numbers to provide a rigorous and reproducible approach to accepting or rejecting the null hypothesis
- The number should represent the chance of obtaining the observed statistic

We define a p-value to be the probability under the null hypothesis that the random statistics are greater than or equal to the observed statistic

Validating Hypotheses

P(the test statistic would be equal to or more extreme than the observed test statistic under the null hypothesis)

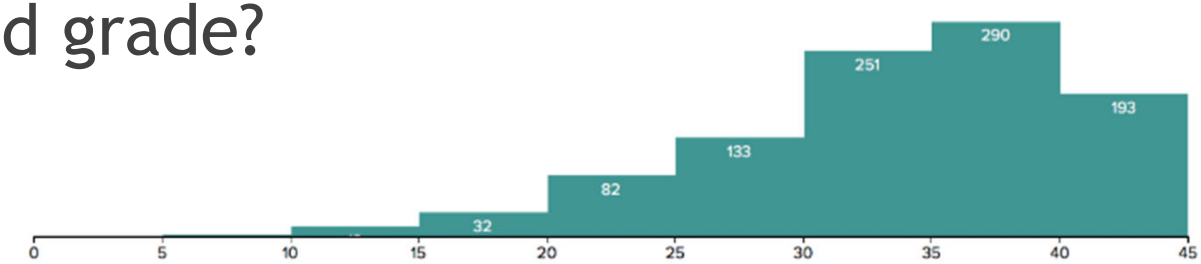


Evaluating Mendel's pea flower hypothesis

This area is the P-value (approximately)

Example

- How can we use hypothesis testing to validate the assumption that each section has the same chance of getting a good grade?
- Suppose we want to assess the grades on assignment across the sections of class



Summary

- Comparing Distributions
  - Statistics for goodness of fit
- Error Probabilities
  - Observed Significance Levels

Goals

- Use functions to give names to blocks of code. Call functions by passing arguments.
- Apply p-values to accept or reject the null hypothesis