STATISTICS AND EXPLORATORY DATA ANALYSIS

LIGHTHOUSE LABS

AGENDA

- Hypothesis testing
- p-value
- Scaling data
- Normalization

LIGHTHOUSE LABS

HYPOTHESIS TESTING

LIGHTHOUSE LABS

HYPOTHESIS TESTING

- Suppose you are about to study for an exam, and you are interested in if the number of study hours is correlated with the test score. Perhaps you hypothesize that they are correlated.
- You collect a bunch of data from your friends, family, classmates, etc. and calculate the correlation coefficient between study hours and test scores to be 0.6
- What do you conclude? Is 0.6 high enough for you to say there is a correlation? Is there enough evidence to support your hypothesis? If it’s not high enough, what about 0.7? 0.8? It would be nice if there was a formal way to test if the correlation is significant.

HYPOTHESIS TESTING

- A statistical test used to determine if there is enough evidence to support a hypothesis
- A formal (statistical) way to test for significance.
  - E.g. Is the correlation significant or not?
  - E.g. Is our data normally distributed or not?
  - E.g. Are two variables independent or not?
  - E.g. Is there a difference in the average between two groups of data or not?

NULL AND ALTERNATIVE HYPOTHESIS

- Hypothesis tests consist of a null hypothesis (H_0) and an alternative hypothesis (H_a or H_1).
- H_0 is the default (assumed) belief, and we are interested in if there is enough evidence overturn H_0 and instead conclude that H_a is true.
    - H_0 = not-guilty, H_a=guilty
    - H_0 = correlation is zero, H_a=correlation is not zero
    - H_0 = data is normally distributed, H_a=data is not normally distributed
    - H_0 = variables are independent, H_a=variables are not independent
    - H_0 = two groups have the same average, H_a=two groups do not have the same average

NULL AND ALTERNATIVE HYPOTHESIS

- You need enough evidence to overturn H_0, since it is the default belief.
  - If your evidence tells you that you are “unsure”, hypothesis testing will have you stay with H_0.
- So how do we determine what is “enough evidence”?
  - P-values.

NULL AND ALTERNATIVE HYPOTHESIS

- You need enough evidence to overturn H_0, since it is the default belief.
  - If your evidence tells you that you are "unsure", hypothesis testing will have you stay with H_0.
- So how do we determine what is "enough evidence"?
  - P-values.

P-VALUES

LIGHTHOUSE LABS

P-VALUES

- **P-value**: the probability of observing the data you did (or something more extreme), assuming the null hypothesis was true.
  - How likely was my data, assuming the null hypothesis was true.
  - How likely is the null hypothesis true.
- The lower the p-value, the more evidence you have to reject the null hypothesis
  - By convention, people reject the null hypothesis when $p < 0.05$.
  - If my null hypothesis was true, there is a less than 5% chance I would have observed my data.
- If $p < 0.05$, reject H_0
- If $p \geq 0.05$, do not reject H_0

COMMON HYPOTHESIS TESTS

LIGHTHOUSE LABS

SHAPIRO-WILK TEST

- Normality test: Shapiro-Wilk Test
  - H_0: data is normally distributed
  - H_a: data is not normally distributed

```python
# Example of the Shapiro-Wilk Normality Test
from scipy.stats import shapiro
data = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
stat, p = shapiro(data)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably Gaussian')
else:
    print('Probably not Gaussian')
```

PEARSON'S CORRELATION COEFFICIENT

- Correlation test: Pearson's Correlation Coefficient (numerical values)
  - H_0: no correlation between the two variables
  - H_a: correlation between the two variables

```
# Example of the Pearson's Correlation test
from scipy.stats import pearsonr
data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
data2 = [0.353, 3.517, 0.125, -7.545, -0.555, -1.536, 3.350, -1.578, -3.537, -1.579]
stat, p = pearsonr(data1, data2)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably independent')
else:
    print('Probably dependent')
```

CHI-SQUARED TEST

- Independence test: Chi-Squared Test (categorical values)
  - H_0: variables are independent
  - H_a: variables are not independent

```python
# Example of the Chi-Squared Test
from scipy.stats import chi2_contingency
table = [[10, 20, 30],[6,  9,  17]]
stat, p, dof, expected = chi2_contingency(table)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably independent')
else:
    print('Probably dependent')
```

T-TEST

- Two equal averages test: T-Test
  - H_0: averages are equal
  - H_a: averages are not equal

```python
# Example of the Student's t-test
from scipy.stats import ttest_ind
data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
data2 = [1.142, -0.432, -0.938, -0.729, -0.846, -0.157, 0.500, 1.183, -1.075, -0.169]
stat, p = ttest_ind(data1, data2)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution')
else:
    print('Probably different distributions')
```

ONE-WAY ANOVA TEST

- Multiple equal averages test: One-way ANOVA Test
  - H_0: all averages are equal
  - H_a: one or more average are not equal

```python
# Example of the Analysis of Variance Test
from scipy.stats import f_oneway
data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
data2 = [1.142, -0.432, -0.938, -0.729, -0.846, -0.157, 0.500, 1.183, -1.075, -0.169]
data3 = [-0.208, 0.696, 0.928, -1.148, -0.213, 0.229, 0.137, 0.269, -0.870, -1.204]
stat, p = f_oneway(data1, data2, data3)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution')
else:
    print('Probably different distributions')
```

SELECTING THE RIGHT TEST

- As you explore your data during EDA, you may want to perform some of these hypothesis tests to check your assumptions.
- For each of these hypotheses, determine what is the appropriate test and what are the null and alternative hypotheses?

1. "The scatterplot between these two variables suggest they may be linearly correlated"
2. "It seems like a lot of smokers develop lung cancer"
3. "The average blood pressures for diabetic people and non-diabetic people seem to be different"
4. "The amount of sleep I get is not the same for every day of the week"

SELECTING THE RIGHT TEST

- As you explore your data during EDA, you may want to perform some of these hypothesis tests to check your assumptions.
- For each of these hypotheses, determine what is the appropriate test and what are the null and alternative hypotheses?

1. "The scatterplot between these two variables suggest they may be linearly correlated"
   Pearson correlation coefficient test
2. "It seems like a lot of smokers develop lung cancer"
3. "The average blood pressures for diabetic people and non-diabetic people seem to be different"
4. "The amount of sleep I get is not the same for every day of the week"

SELECTING THE RIGHT TEST

- As you explore your data during EDA, you may want to perform some of these hypothesis tests to check your assumptions.
- For each of these hypotheses, determine what is the appropriate test and what are the null and alternative hypotheses?

1. "The scatterplot between these two variables suggest they may be linearly correlated"
   Pearson correlation coefficient test
2. "It seems like a lot of smokers develop lung cancer"
   Chi-squared test for independence
3. "The average blood pressures for diabetic people and non-diabetic people seem to be different"

4. "The amount of sleep I get is not the same for every day of the week"

SELECTING THE RIGHT TEST

- As you explore your data during EDA, you may want to perform some of these hypothesis tests to check your assumptions.
- For each of these hypotheses, determine what is the appropriate test and what are the null and alternative hypotheses?

1. "The scatterplot between these two variables suggest they may be linearly correlated"
   Pearson correlation coefficient test
2. "It seems like a lot of smokers develop lung cancer"
   Chi-squared test for independence
3. "The average blood pressures for diabetic people and non-diabetic people seem to be different"
   T-test comparing two averages
4. "The amount of sleep I get is not the same for every day of the week"

SELECTING THE RIGHT TEST

- As you explore your data during EDA, you may want to perform some of these hypothesis tests to check your assumptions.
- For each of these hypotheses, determine what is the appropriate test and what are the null and alternative hypotheses?

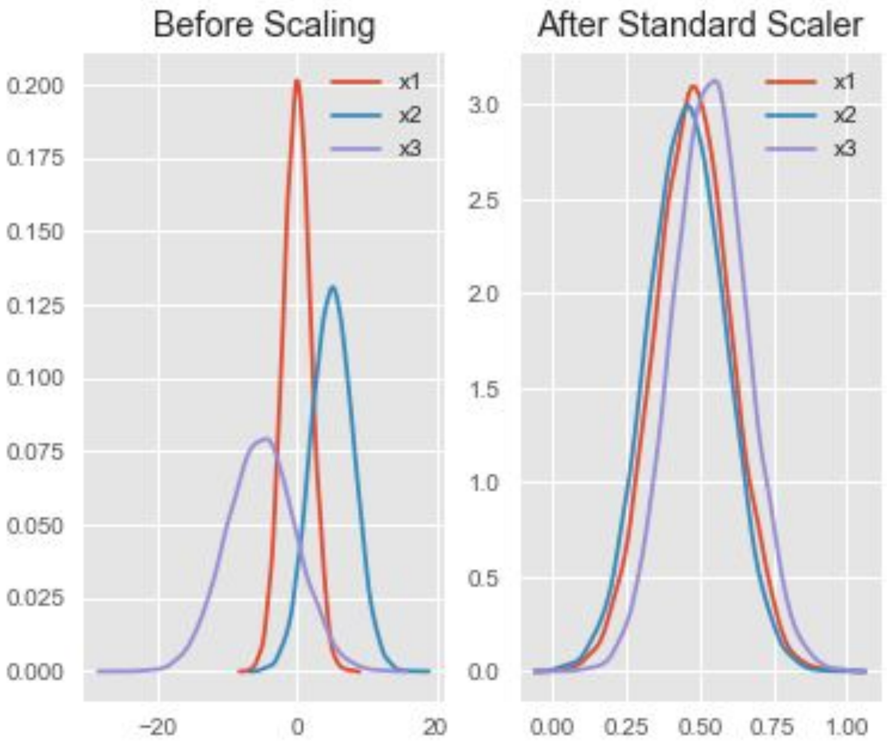
1. "The scatterplot between these two variables suggest they may be linearly correlated"
   Pearson correlation coefficient test
2. "It seems like a lot of smokers develop lung cancer"
   Chi-squared test for independence
3. "The average blood pressures for diabetic people and non-diabetic people seem to be different"
   T-test comparing two averages
4. "The amount of sleep I get is not the same for every day of the week"
   ANOVA test comparing multiple averages

SCALING DATA

SCALING DATA

- Later in the course, we will see it can be advantageous to have “unitless” data.
  - Based on the idea that our data shouldn’t depend on what unit it was measured in.
- We can scale our data to instead be measured as “standard deviations away from the mean” (called standard units or z-scores).
  - Take each data value, subtract the mean, and then divide by the standard deviation.
- Works best on normally distributed data.
- “Standardization” or “StandardScaler”

SCALING DATA

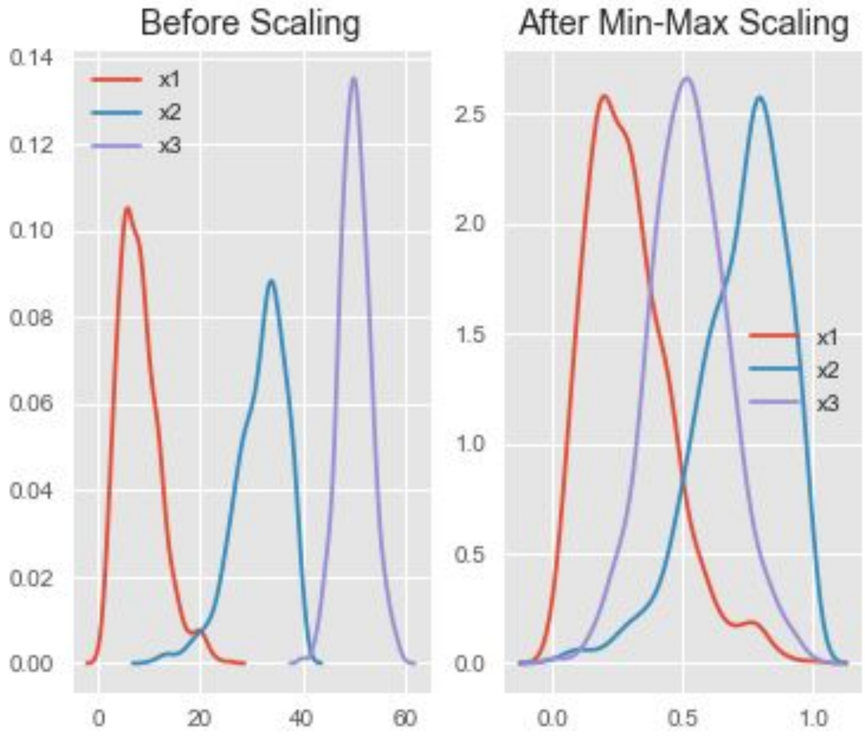


$$z = \frac{x - \mu}{\sigma}$$

SCALING DATA

- Another type of scaling “squishes” the data into a desired range (0 to 1, by default).
    - Based on the idea that variables with larger values shouldn’t automatically have a bigger impact.
- We can do this by taking each data value, subtracting the minimum, and then dividing by the range.
- Distribution shape is maintained.
- “Normalization” or “MinMaxScaler”

SCALING DATA



$$u = \frac{x - \min(x)}{\max(x) - \min(x)}$$

TRANSFORMING DATA

- Lastly, some statistical models have the assumption that the data is normally distributed.
- We should run a normality test (Shapiro-Wilk) prior to applying these models.
- But what if a normality test fails?
- We can try to use a mathematical transformation!
- Transformations can also be used to change the scale of the data
    - E.g. log or square-root transformation