Statistics in Data Science

overview

1. Statistical Hypothesis Testing
2. Chi-Squared Goodness-Of-Fit Test
3. Linear Regression

Hypothesis Testing

- Greensboro has significantly lower crime rates as compared to the national average.
- Code peer review improves students’ programming abilities.
- Drug A improves students’ ability to not fall asleep during class.
- **Definition:** Hypothesis testing is a statistical method that allows one to make inferences or draw conclusions about a population based on a sample.

Hypothesis Testing

- **Types of Hypotheses:**
  1. **Null Hypothesis ($H0$):**

     Assumes no effect or difference; a statement to be tested.
  2. **Alternative Hypothesis ($H1$):**

     Opposite of the null; what you want to prove.

Example: **- Coin Toss**

Scenario: Testing if a coin is fair.

Null Hypothesis ($H0$): The coin is fair, $P(Heads)=0.5$.

Alternative Hypothesis ($H1$): The coin is not fair, $P(Heads) \neq 0.5$.

Null Hypothesis (*H*0):

- The null hypothesis : value of a parameter is equal to some claimed value.

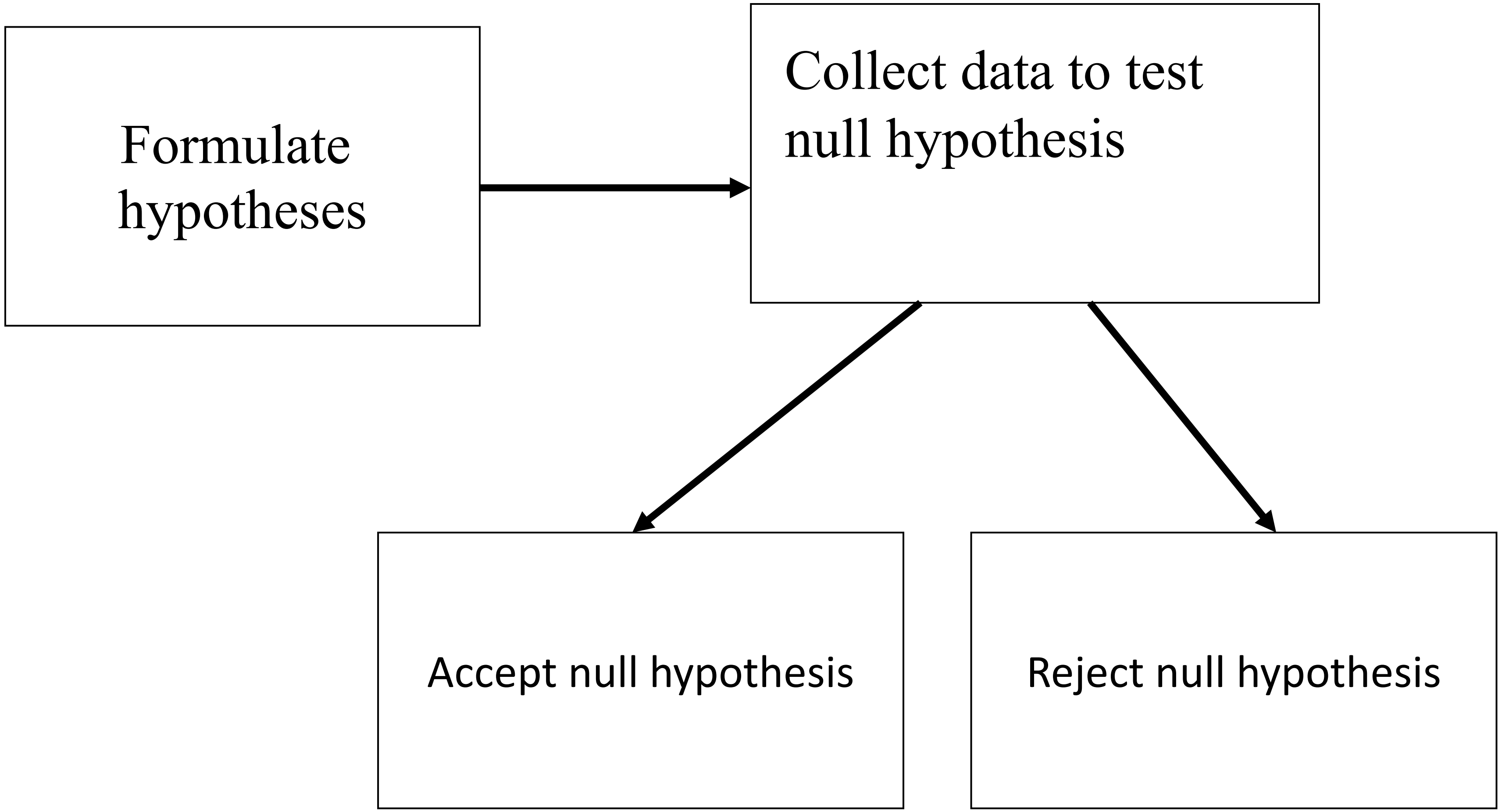
- We test the null hypothesis directly.

- Either reject H_0 or fail to reject H_0.

Alternative Hypothesis ($H1$):

- The alternative hypothesis : parameter that somehow differs from the null hypothesis.
- The symbolic form of the alternative hypothesis must use one of these symbols: $\neq$, $<$, $>$.

Hypothesis Testing



Hypothesis Testing

1. **Null Hypothesis (H_0)**
   - The difference is caused by random chance.

2. **Alternate hypothesis (H_1)**
   - “The difference is real”.
   - (H_1) always contradicts the $H_{0.}$

If you are conducting a study and want to use a hypothesis test to support your claim, the claim must be worded so that it becomes the alternative hypothesis.

Hypothesis Testing

Example - Efficacy Test for New drug

- Drug company has new drug
- FDA tells company that they must demonstrate that new drug is better than current treatment.
- Firm runs clinical trial where some patients receive new drug, and others receive standard treatment
- Numeric response of therapeutic effect is obtained (higher scores are better).

Two-sample tests

- Two-sample tests are statistical procedures used to determine if there are any statistically significant differences between two separate groups or samples based on a particular metric or outcome.

- These tests are used when you have data from two independent samples and you want to make inferences about the population parameters from which these samples were drawn.

One sample vs. two sample tests

Crimes in Greensboro
100
23
34
5
67
89
34
33
20

National average: 78

Old drug- Cholesterol level	New drug- Cholesterol level
54	45
23	33
34	22
32	10
45	8
55	7
34	6
33	5
20	41
33	22

Two-tailed vs. One-tailed Tests

Two-tailed test: "is there a **significant difference**?"

One-tailed tests: "is the sample mean **greater** than P_u?"

"is the sample mean **less** than P_u?"

Paired vs. Unpaired tests

Student	Pre-module score	Post-module score
1	18	22
2	21	25
3	16	17
4	22	24
5	19	16
6	24	29
7	17	20
8	21	23
9	23	19
10	18	20
11	14	15
12	16	15
13	16	18
14	19	26
15	18	18
16	20	24
17	12	18
18	22	25
19	15	19
20	17	16

Conclusions in Hypothesis Testing

We always test the null hypothesis.

1. Reject the null hypothesis.

2. Fail to reject the null hypothesis.

P-Value

- **Definition:** The probability of observing a test statistic as extreme as, or more extreme than, the statistic observed given that the null hypothesis is true.

p-value is the probability of obtaining the observed sample results by chance.

The null hypothesis is rejected if the *p*-value is very small, such as 0.05 or less.

T-test

- One or two samples.

- Test the null hypothesis that the mean of the sample is equal to a given mean.

- Test the null hypothesis that the means of the two samples are equal.

T-test in Python

- scipy.stats.ttest_1samp(a, popmean)
  - Calculates the T-test for the mean of ONE group of scores.
- scipy.stats.ttest_ind(a, b)
  - Calculates the T-test for the means of TWO INDEPENDENT samples of scores.
- scipy.stats.ttest_rel(a, b)
  - Calculates the T-test on TWO RELATED samples of scores, a and b.
- Returns:
  - t-statistic
  - P-value

- **Coding Example - One Sample T-Test**

```python
import numpy as np
from scipy import stats

# Sample data: Ages of 20 individuals
sample_ages = [32, 34, 29, 29, 22, 39, 38, 37, 38, 36, 30, 26, 22, 22, 24,
# Null Hypothesis: Mean age = 30
t_statistic, p_value = stats.ttest_1samp(sample_ages, 30)
print(f"t-statistic: {t_statistic}, p-value: {p_value}")
```

- **Note:** The code tests whether the average age of the sample is 30.

t-statistic: 1.0381786725003321, p-value: 0.3122193752132857

Exercise: Apply T-tests

Crimes in Greensboro
100
23
34
5
67
89
34
33
20

National average: 78

Old drug- Cholesterol level	New drug- Cholesterol level
54	45
23	33
34	22
32	10
45	8
55	7
34	6
33	5
20	41
33	22

Rock-Paper-Scissors

Which did you throw?

a) Rock
b) Paper
c) Scissors

ROCK	PAPER	SCISSORS
34	24	40

How would we test whether all of these categories are equally likely?

Hypotheses

Let p_i denote the proportion in the i^{th} category.

H_0 : All p_i are the same

H_1 : At least one p_i differs from the others

Observed Counts

The ***observed counts*** are the actual counts observed in the study

	ROCK	PAPER	SCISSORS
Observed	34	24	40

Expected Counts

The ***expected counts*** are the expected counts if the null hypothesis were true

	ROCK	PAPER	SCISSORS
Observed	34	24	40
Expected	33	33	33

Chi-Square Statistic

- A ***test statistic*** is one number, computed from the data, which we can use to assess the null hypothesis

- The ***chi-square statistic*** is a test statistic for categorical variables:

$$\boxed{\chi^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}}$$

Rock-Paper-Scissors

	ROCK	PAPER	SCISSORS
Observed	34	24	40
Expected	33	33	33

$$\chi^2 = \frac{\left(34-33\right)^2}{33} + \frac{\left(24-33\right)^2}{33} + \frac{\left(40-33\right)^2}{33} = 3.97$$

Chi-Square Test for Goodness of Fit

Definition: A statistical test used to determine if there's a significant difference between observed frequencies and expected frequencies in one or more categories.

Assumptions:

- Data are randomly sampled from the population.
- The variable is categorical.
- Expected frequencies for each category should be at least 5 for the chi-square approximation to be valid.

Chi-Square Test for Goodness of Fit

Calculate the expected counts for each cell. Make sure they are all greater than 5 to proceed.

1. Calculate the χ^2 statistic

2. Compute the p-value

3. Interpret the p-value in context.

Errors in Hypothesis Testing

1. **Type I Error:** Rejecting $H0$ when it's true.
2. **Type II Error:** Failing to reject $H0$ when it's false.

Type I and Type II errors

Test Result – True State	H_0 True	H_0 False
H_0 True	Correct Decision	Type I Error
H_0 False	Type II Error	Correct Decision

$$\alpha = P(Type\ I\ \ Error) \quad \beta = P(Type\ II\ \ Error)$$

Keep α, β reasonably small

Significance Level (α)

- **Definition:** The significance level, α, is the probability threshold below which the null hypothesis will be rejected. It represents the risk we are willing to take of rejecting a true null hypothesis.
- **Common values:** 0.01, 0.05, 0.10

- **Relationship Between α and Type I Error**

Statement: The significance level, α, is the probability of Type I error.

Explanation: When we set α to 0.05 (or 5%), we're saying we accept a 5% risk of incorrectly rejecting the null hypothesis when it's actually true.

Multiple tests

- As the number of tests increases, the likelihood of observing a rare event (false positive) by chance also increases.
- **Explanation:** If we perform one test at a 5% significance level, our risk of a Type I error (incorrectly rejecting a true null hypothesis) is 5%. But if we perform 20 independent tests at the 5% significance level, the risk increases.

Why multiple testing matters

- In general, if we perform m hypothesis tests, what is the probability of at least 1 false positive?

P(Making an error) = α **= 0.05**

P(Not making an error) = $1 - \alpha$ **= 0.95**

P(Not making an error in m tests) = $(1 - \alpha)^m$

P(Making at least 1 error in m tests) = $1 - (1 - \alpha)^m$ **=0.994, m=100**

Exercise:

- Question: You want to test if the color preference for shirts among males and females is independent of gender. Use the chi-square test to determine this.

- Dataset:

Color	Males	Females
Blue	35	30
Red	15	20
Green	10	15

Solutions:

```
import numpy as np
from scipy.stats import chi2_contingency

data = np.array([[35, 30],
                 [15, 20],
                 [10, 15]])

chi2_stat, p_val, _, _ = chi2_contingency(data)
print("Chi2 Stat:", chi2_stat)
print("P Value:", p_val)
```

```
Chi2 Stat: 1.9019442096365173
P Value: 0.3863652533157555
```