Hypothesis Testing: Testing an Association

Two-Sample T-Test

We can test an association between a quantitative variable and a binary categorical variable by using a two-sample t-test. The null hypothesis for a two-sample t-test is that the difference in group means is equal to zero. A two-sample t-test can be implemented in Python using the `ttest_ind()` function from `scipy.stats`. The example code shows a two-sample t-test for testing an association between claw length and species of bear (grizzly or black).

```
from scipy.stats import ttest_ind

#separate out claw lengths for two
species
grizzly_bear =
data.claw_length[data.species=='grizzly'
]
black_bear =
data.claw_length[data.species=='black']

#run the t-test here:
tstat, pval = ttest_ind(grizzly_bear,
black_bear)
```

Multiple Two-Sample T-Tests

In order to test an association between a quantitative variable and a non-binary categorical variable, one *could* use multiple two-sample t-tests. However, running multiple tests increases the probability of a false positive (type I error) so that it is greater than the significance threshold for each test. To avoid this issue, a better solution is to run an ANOVA; then, if the p-value for the ANOVA is significant, run Tukey's range test.

codecademy

Analysis of Variance

An Analysis Of Variance (ANOVA) is used to test an association between a non-binary categorical variable and a quantitative variable while limiting the probability of a type I error. The null hypothesis for ANOVA is that the group means are all equal. The alternative hypothesis is that at least one pair of group means are different. An ANOVA can be implemented in Python using the `f_oneway()` function from `scipy.stats`. The example code shows an ANOVA test for an association between tree height and tree species (pine, oak, or spruce).

```
from scipy.stats import f_oneway
fstat, pval = f_oneway(heights_pine,
heights_oak, heights_spruce)
```

Tukey's Range Test

Tukey's range test should be used after ANOVA (if the p-value is significant) to simultaneously compare group means for all possible pairs of groups while maintaining some pre-chosen probability of a type I error. For each pair of groups, Tukey's range test will indicate whether to "reject the null" and conclude that those two groups are significantly different. Tukey's range test can be implemented with the `pairwise_tukeyhsd()` function from `statsmodels.stats.multicomp`. The example code shows how to use this function for examining an association between tree height and tree species using an overall type I error rate of 0.05.

```
# Tukey's Range Test
from statsmodels.stats.multicomp import
pairwise_tukeyhsd
tukey_results =
pairwise_tukeyhsd(tree_data.height,
tree_data.species, 0.05)
```

codecademy

Hypothesis Testing Assumptions

Before using two-sample t-tests, ANOVA, or Tukey's range test, it is important to check whether the assumptions of the tests are true:
(1) All observations should be independently and randomly sampled
(2) The standard deviations of the groups should be equal
(3) The data should be normally distributed or the sample size should be large
(4) The groups created by the categorical variable should be independent

Chi-Square Test

To test for an association between two categorical variables, we can use a Chi-Square test. The null hypothesis for a Chi-Square test is that there is no association between the variables and the alternative hypothesis is that there is an association between the variables. A Chi-Square test can be implemented in Python using the `chi2_contingency()` function from `scipy.stats`. The example code shows how to implement a Chi-Square test for investigating an association between what version of a website someone saw and whether or not they subscribed.

```
import pandas as pd
from scipy.stats import chi2_contingency

# create contingency table
ab_contingency =
pd.crosstab(data.Web_Version,
data.Subscribed)

# run a Chi-Square test
chi2, pval, dof, expected =
chi2_contingency(ab_contingency)
```

codecademy

Chi-Square Assumptions

Proper use of the Chi-Square test requires certain assumptions to be met. The first assumption is that observations are independent and random to ensure the sample properly represents the population. The next assumption is that categories of both variables be mutually exclusive; this is so observations can only fall into one category or the other, but not both. Finally, groups created by the categorical variables should be independent; neither group should have any influence on the other.

Print Share