# Semana 8

Tuesday, September 21, 2021 5:21 PM

## Comparing distributions

### Total variation distance (TVD)

Every distance has a computational recipe

- For each category, compute the difference in proportion between two distributions.
- Take the absolute value of each difference.
- Sum, and then divide the sum by 2

### Summary of the method

To assess whether a sample was drawn randomly from a known categorical distribution:

- Use TVD as the statistic because it measures the distance between categorical distributions.
- Sample at random from the population and compute the TVD from the random sample; repeat numerous times
- Compare:
    - Empirical distribution of simulated TVDs
    - Actual TVD from the sample in the study

## Decisions and uncertainty

### Incomplete information

- We are trying to choose between two views of the world, based on data in a sample.
- It is not always clear whether the data are consistent with one view or the other.
- Random sample can turn out quite extreme. It is unlikely, but possible.

### Testing Hypotheses

- A test chooses between two vies of how data were generated.
- The views are called hypotheses
- The test picks the hypothesis that is better supported by the observed data.
- Method:
    - Simulate data under one of the hypotheses.
    - Compare the simulation results and the observed data.
    - Pick one of the hypotheses, based on whether or not the simulation results and observed data are consistent.

### Null and alternative

The method only works if we simulate data under one of the hypotheses.

- Null hypothesis.
    - A well defined chance model about how the data were generated.
    - We can simulate data under the assumptions of this model - "under the null hypothesis"
- Alternative hypothesis
    - A different view about the origin of the data.

### Test statistic

- The statistic that we choose to simulate, to decide between the two hypotheses

Questions before choosing the statistic:

- What values of the statistic will make us lean towards the null hypothesis?
- What values will make us lean towards the alternative?
    - Preferably, the answer should be just "high" or just "low". Try to avoid "both high and low".

### Prediction under the null hypothesis

- Simulate the test statistic under the null hypothesis; draw the histogram of the simulated values.
- This displays the empirical distribution of the statistic under the null hypothesis.
- It is a prediction about the statistic, made by the null hypothesis.
    - It shows all the likely values of the statistic
    - Also how likely they are **(If the null hypothesis is true)**
- The probabilities are approximate, because we cant generate all the possible random samples

### Conclusion of the test

Resolve choice between null and alternative hypotheses

- Compare the observed test statistic and its empirical distribution under the null hypothesis
- If the observed value is not consistent with the distribution, then the test favors the alternative - "rejects the null hypothesis"

Whether a value is consistent with a distribution:

- A visualization may be sufficient.
- If not, there are conventions about "consistency".

## Statistical significance

### Conventions about inconsistency

- Inconsistent: the test statistic is in the tail of the empirical distribution under the null hypothesis.
- "In the tail," first convention:
    - The area in the tail is less than 5%.
    - The result is "statistically significant"
- "In the tail," second convention:
    - The area in the tail is less than 1%
    - The result is "highly statistically significant"

### Definition of the P-value

Formal name: observed significance level.

The *P*-value is the chance,

- Under the null hypothesis,
- That the test statistic
    - Is equal to the value that was observed in the data
    - Or is even further in the direction of the alternative

## An error probability

- The cut off for the *P*-value is an error probability.
- If:
    - Your cutoff is 5%
    - And the null hypothesis happens to be true.
- Then there is about a 5% chance that your test will reject the null hypothesis

## The basic structure of testing a hypothesis is the following:

1. Setting up your question:
    - Null Hypothesis:  For a fully specified chance model under which you can simulate the data, it hypothesizes that the data were generated as if by random selection, according to the particulars of the chance model.
    - Alternative hypothesis: States that there is some process other than the chance model that generated the data.

2. Choose a Test Statistic:
    - A value that can be computed for the observed data and also for random samples from the chance model.
    - The magnitude of the Test Statistic should help you determine whether the distribution of data are consistent with the Null Hypothesis.
    - When this value is computed for the observed data, it is called the Observed value of the test statistic.
3. Simulate under the Null Hypothesis:
    - Use random sampling to create a(n) Empirical distribution of the test statistic in order to approximate the Probability distribution of this statistic, which is called its Sampling distribution, assuming that the chance model is true.
4. Conclusion:
    - P-value: The probability under the chance model that the Test Statistic is equal to the Observed value of the test statistic or ever further in the direction of the Alternative hypothesis.
    - P-value Cutoff: The maximum P-value for which you choose to reject the null hypothesis in favor of the Alternative hypothesis. A typical value is 5%.