STA130 Review Notes

Jim Gao (*jim.gao@mail.utoronto.ca*)

December 16, 2018

1 Hypothesis Testing

1.1 Terminology

Term	**Definition**
Statistical Inference	An uncertain conclusion made using methods of statistics
Parameter	"True" value of what is interested, holds for the population
Statistic	Value describing the sample(as a subset of the population). May change for different samples.
Test Statistic	Value measuring the compatibility between the null and alternative hypothesis.

1.2 Goal

We wish to verify if a given description of some population is statistically significant (i.e. not occurring due to chance).

Kissing example:

- The statistic: "64.5% of the couples turned their head to the right" describes the behaviour of some subset of the population.
- We wish to know whether if there is some factor or reason driving this behaviour. There also might be nothing at all.
- If this statistic is too extreme for it to be random chance (another example might be getting 100 consecutive heads when flipping a coin), we say that it is statistically significant.

1.3 Procedure

1.3.1 The Null/Alternative Hypothesis

We transform the question of statistical significance into a problem of finding evidence against "the lack of statistical significance".

To do that, we establish two hypotheses.

1. The **null** hypothesis (H_0): We assume that the event occurred by random chance. There would be an equal chance to each outcome (i.e. if there are n total outcomes, the probability for each would be $p = 1/n$).
2. The **alternative** hypothesis (H_a): We assume that the null hypothesis is false. This should be the negation of H_0. If H_0 assumes that $p = 1/n$, H_a would assume that $p \neq 1/n$.

1.3.2 Establishing Test Statistic

We narrow down the outcome of the event down to a single variable. This variable is called the **test statistic**.

In some cases, the test statistic can be simply measured, and in other cases it needs to be computed. For example:

- Kiss example: The proportion of couples is the test statistic. This can be directly measured from the sample taken.
- Tylenol vs Aspirin: As we are measuing the **differences** between two sets of data, this comparison needs to be reflected in the test statistic.

  Let $\hat{p}_T$ and $\hat{p}_A$ measure the effectiveness (propotion of patients helped) for the two types of medication, we wish to measure their difference. Therefore, the test statistic would be:

  $$\hat{p} = \hat{p}_T - \hat{p}_A$$

  In this case, H_0 would be $\hat{p} = 0$, while H_a is $\hat{p} \neq 0$.
- House Prices vs Size: When we are measuring the **relationship** between two variables, the test statistic would be the **slope of the regression line**.

  H_0 would be "no relationship", which is when $m = 0$. Therefore, H_a would be defined as $m \neq 0$ (i.e. there is a relationship).

1.3.3 Simulation

In this step, we assume that H_0 is true.

In the kissing example, we would run simulations assuming that the probability of turning the head to the right is $p = 0.5$.

In this case, the **test statistic** would be the proportion of couples kissing while turning their heads to the right.

Obviously, not every simulation will have exactly half the couples kissing to the right. The distribution from such simulation would produce a **normal distribution centered at the median**.

We call such distribution the **empirical distribution** of the test statistic under the null hypothesis.

1.3.4 The p-value

As our goal is to look for the incompatibility of the null hypothesis (as an evidence to statistical significance), we do so with the p-value.

We define the p-value to be the probability of observing data that is **as least as unusual as the sample**.

That is, for some median M and the test statistic $\hat{x}$, we are looking for the probability of getting:

$$x \in (-\infty, M - \hat{x}) \cup (M + \hat{x}, \infty)$$

for some x generated from the empirical distribution.

In this case, a small p-value would indicate that there is little chance of the statistic observed from the sample to be happening purely by chance (under the assumption of H_0).

1.3.5 Conclusion

We find **evidence against the null hypothesis** when $p < 0.05$.

In this case, we say that the observation is **statistically significant**.

In the case when p is large (i.e. $p > 0.1$), we would **fail to reject the null hypothesis**. In any case, **you cannot support the null hypothesis**.

1.4 Errors in Hypothesis Testing

Recalling the definition of a statistical inference. We are making an **uncertain** conclusion about samples. Therefore, we are bound to make mistakes sometimes.

The mistakes, or errors, can be classified into 2 categories - Type 1 and 2.

	H_0 True	H_0 False
No Reject H_0	No Error	Type 2
Reject H_0	Type 1	No Error

In other words, we say that:

- Type 1 occurs when the data seems to be inconsistent with H_0 when H_0 is true.
- Type 2 occurs when the data seems to be consistent with H_0 when it is not actually true.

2 Bootstrap Sampling

2.1 Terminology

Term	Definition
Population	A group/set of similar items of interest in a statistical study
Sample	A subset representing the population in a statistical study
Bias	A tendency during the process of data collection that may result in misleading conclusions
Variability	Measure of how spread-out a group of data is

2.2 Motivation

Ideally, we take samples from the population, and compute statistics from the sample to gain knowledge of the real world (parameters).

Therefore, we need to ensure that **sample represents the real world**. This ensures that **there is no bias**, and would require multiple samples for the statistic to begin approximating the parameter.

However, sometimes **we only have one sample**.

2.3 Assumption and Consequences

Since our main goal to approximate the parameters of the real world, but we only have one sample, we would want that sample to be **representative of the whole population**. This is a key assumption to bootstrapping.

Of course, the results of sampling from the (one) sample would be different than sampling from the population. However, this result can also vary based on the **size of the sample**.

When comparing the sampling distribution of a larger sample compared to a smaller sample:

- Mode: Neither the location nor the number of modes is unaltered.
- Variability: A larger sample has lower variability (i.e. lower standard deviation)
- Symmetry: Distribution is more symmetric with a larger sample

In general, a larger sample produces better results.

2.4 Procedure

The bootstrapping methods relies on **sampling with replacement**. The goal is to create **"bootstrap samples"** of the same size of the original sample.

2.5 Confidence Intervals

The goal of bootstrapping is to make conclusions about the population from only 1 sample. Therefore, we can produce an interval from bootstrap sampling that describes where the true parameter situates.

In addition, based on how accurate (confident) we need to be about such interval, higher confidence levels would result in larger intervals (more accurate, less precise).

2.5.1 Computing Confidence Intervals

To compute the $p\%$ confidence interval, we repeat the following steps many times:

1. Compute the bootstrap distribution.
2. In the distribution, find the interval that covers $p\%$ of the distribution.

Statistically, $p\%$ of the intervals would capture the true parameter inside it. In other words, we can define the confidence interval as: **the interval where we are $p\%$ certain that it captures the true parameter**.

3 Supervised Learning

3.1 Definition

The algorithm has the goal of discovering patterns in existing data, and to make decisions (classifications or regression) for some new data following such patterns.

In supervised learning, the algorithm is "supervised" by the existing data, which contains the correct answer.

3.2 Terminology

Term	**Definition**
Response Variables	The output/dependent variable.
Predictor Variables	The input variables. We base the prediction on these variables.
Catagorical Variables	Discrete, countable variables (e.g. sex, nationality).
Continuous Variables	Variables described with continuous numbers (e.g. age, height)
Cost/Loss	The total error of a prediction model

3.3 Classification Trees

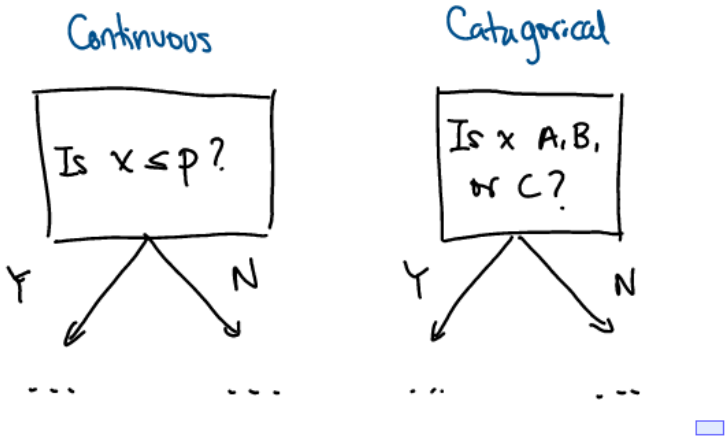
Classificiation trees are used for **catagorical supervised learning**.

3.3.1 Main Features

The classification tree consists of vertices and edges, just like a tree. The vertices are called **"decision nodes"**, in which they contain a predictor variable as well as a rule.

It is important to note that each node makes a **binary decision**, meaning that, for each node, there may only be two possible outcomes.

The predictor variable may be a catagorical one as well as a continuous one. There are different ways to treat each one:



- Continuous: The rule describes a threshold p. Based the value of the input x, the node makes a binary decision. (e.g. if $x < p$ and $x \geq p$)
- Catagorical: The rule describes a subset A. It makes a decision based on if the input x satisfies $x \in A$.

On the leaf nodes (nodes with no children) are 3 values:

1. The decision/output
2. The number of training data points satisfying the conditions
3. The error rate

3.3.2 Measure of Purity/Cost

Purity or cost is a measure of the accuracy of the classification tree. A **higher accuracy** implies a **higher purity/lower cost**. Therefore, we define a better tree to be the one with a higher purity.

Purity is usually measured with the **Gini index** or **Entropy**.

3.3.3 Branching Threshold

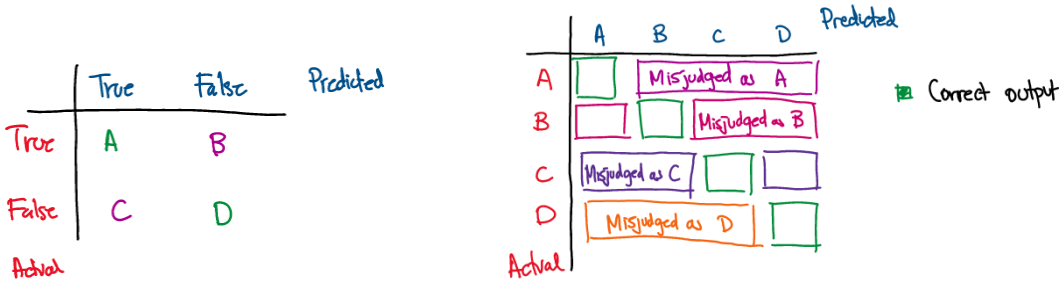
While trying to increase the accuracy of the tree, we also want to keep the tree as simple as possible. This simplifies the decision process, and also **prevents overfitting**.

To do so, we design a rule for the decision tree to stop splitting. Let β be some threshold, and ΔI be the improvement in purity as a result of branching. We keep branching when $\beta > \Delta I$, and stop branching otherwise.

This prevents the tree from unnecessarily branching, when branching creates more complexity while not significantly improving the accuracy of the tree.

3.3.4 Confusion Matrix

The confusion matrix describes the accuracy as well as proportion of different types of errors from a classification tree model.



From the confusion matrix, we can compute the following:

- True-Positive Rate: A/T
- True-Negative Rate: D/T
- False-Positive Rate *(False Alarm)*: C/T
- False-Negative Rate *(Miss)*: B/T

Where T is the total, where $T = A + B + C + D$.

3.3.5 ROC Curve

When the threshold β is used during the generation of the classification tree, we are able to get trees with different levels of complexity and accuracy.

By adjusting β, we can adjust the proportion of **True-Positive** rate and **False-Positive** rate.

Ideally, we would have $TP = 1, FP = 0$. However, we can only approach this point with a statistical model.

For the trade off between TP and FP, we need to find the threshold β such that it fits the scenario. Depending on the penalty for mistakes (e.g. high penalties in medical fields, lower penalties in advertising), we would define β accordingly.

3.4 Linear Regression

Linear regression is used for **continuous supervised learning**.

Given some dataset, $(\vec{x_i}, y_i)$, we wish to find some function $h(\vec{x})$ that fits the dataset.

The linear model h can be written as:

$$\hat{y} = h(\vec{x}) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

where $x_1, x_2, \ldots, x_n$ are predictor variables, and $\hat{y}$ is the output (response). The goal is find the set of scalars $(a_0, a_1, \ldots, a_n)$ which defines the model.

We define the error term ϵ_i as:

$$\epsilon_i = h(\vec{x_i}) - y_i = \hat{y}_i - y_i$$

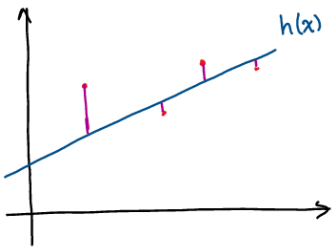
which defines the error between the prediction and the actual value.

3.4.1 Computing Cost/Error

With the dataset $(\vec{x_i}, y_i)$ and function $h(\vec{x})$, we wish to measure the accuracy of the function against the dataset.

We define the cost function as follows:

$$\text{Cost} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(h(\vec{x_i}) - y_i\right)^2$$



For each data point, the **farther** the prediction is from the actual value, the higher the cost would be (since $x^2 = |x|^2$). **A higher cost implies a worse model**.

3.4.2 R-squared Value

The R^2 value measures **how well a model fits the data**.

R^2 produces a number in the range $[0, 1]$, and is defined as follows:

$$R^2 = 1 - \frac{\sum_i \left(h(\vec{x_i}) - y_i\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2}$$

We can observe that the numerator and denominator are both squared-errors. The numerator describes the error for h, and the denominator computes the error of the **null model**, which predicts $\bar{y}$ all the time.

The null model is as bad as the model can be. When h is really bad, we would have a smaller R^2 value. When the error of the function h is insignificant compared to the null model (i.e. h is a good model), the value of R^2 is high.

3.4.3 Root Mean-Squared Error

The RMSE, similar to the cost function, is used to **measure prediction error**. The advantage of RMSE is that **the RMSE value has the same unit as in the dataset** as an **absolute measure of error/loss**.

It is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_i \left(h(\vec{x_i}) - y_i\right)^2}$$

3.4.4 Finding Best-Fit Model

A better model would have a lower cost. Therefore, we want to find the parameters $(a_0, a_1, \ldots, a_n)$ such that the cost is minimal.

This can be done with the `lm` function in R.

3.5 Factors to Consider

3.5.1 Overfitting

Given some dataset with n points, with methods like Lagrange Interpolation, we can always find some polynomial that fits the points perfectly with 0 loss. However, such models are almost always useless.

We cannot only judge the performance of some model based on its performance on the training data it is given, but also other data that **it has not seen before**.

We say that a regression function h is overfitting when there is a significant difference on its performance on the training dataset and the testing dataset.

3.5.2 Confounding Variables

Confounding variables are **predictor variables that change the nature of the relationship**.

Examples of such variables are age in terms of cancer rate.