Foundations of Small-Sample-Size Statistical Inference and Decision Making

Vasileios Maroulas

Department of Mathematics
Department of Business Analytics and Statistics

University of Tennessee

November 3, 2016

Outline

Introduction

- Significance test is a formal procedure for comparing observed data with a *hypothesis* whose truth we want to assess.
- The hypothesis is a statement about the parameters in a population or model.
- The results of a test are expressed in terms of a probability that measures how well the data and the hypothesis agree.

Terminology

- **Null Hypothesis** denoted by H_0. The test of significance is designed to assess the strength of the evidence against the null hypothesis. The null hypothesis is usually a statement of “no effect" or “no difference" (the default assumption that nothing happened or changed).

Terminology

- **Null Hypothesis** denoted by H_0. The test of significance is designed to assess the strength of the evidence against the null hypothesis. The null hypothesis is usually a statement of “no effect" or “no difference" (the default assumption that nothing happened or changed).
- **Alternative Hypothesis** denoted by either H_1 or H_a. It is the competing argument with respect to H_0, however it needs to be decided if it is *one-sided* or *two-sided*.

Terminology

- **Null Hypothesis** denoted by H_0. The test of significance is designed to assess the strength of the evidence against the null hypothesis. The null hypothesis is usually a statement of “no effect" or “no difference" (the default assumption that nothing happened or changed).
- **Alternative Hypothesis** denoted by either H_1 or H_a. It is the competing argument with respect to H_0, however it needs to be decided if it is *one-sided* or *two-sided*.
- **Test Statistic** measures compatibility between the null hypothesis and the data. It is employed for calculating the probability needed for our test of significance.

Terminology

- **Null Hypothesis** denoted by H_0. The test of significance is designed to assess the strength of the evidence against the null hypothesis. The null hypothesis is usually a statement of “no effect" or “no difference" (the default assumption that nothing happened or changed).
- **Alternative Hypothesis** denoted by either H_1 or H_a. It is the competing argument with respect to H_0, however it needs to be decided if it is *one-sided* or *two-sided*.
- **Test Statistic** measures compatibility between the null hypothesis and the data. It is employed for calculating the probability needed for our test of significance.
- $p-$**value** is the probability, computed assuming that H_0 is *true*, that the test statistic would a take a value as extreme or more extreme than what was actually observed. The *smaller* the $p-$value, the stronger the evidence against H_0

Terminology

- **Null Hypothesis** denoted by H_0. The test of significance is designed to assess the strength of the evidence against the null hypothesis. The null hypothesis is usually a statement of “no effect" or “no difference" (the default assumption that nothing happened or changed).
- **Alternative Hypothesis** denoted by either H_1 or H_a. It is the competing argument with respect to H_0, however it needs to be decided if it is *one-sided* or *two-sided*.
- **Test Statistic** measures compatibility between the null hypothesis and the data. It is employed for calculating the probability needed for our test of significance.
- $p-$**value** is the probability, computed assuming that H_0 is *true*, that the test statistic would a take a value as extreme or more extreme than what was actually observed. The *smaller* the $p-$value, the stronger the evidence against H_0
- $\alpha-$ **level of significance** is the decisive value of p. If $p \leq \alpha$ then we say that the data is statistically significant at level α.

Example 1

In agricultural modeling earth's temperature plays an important role. We want to compare ground vs air-based temperature sensors. Ground-based sensors are expensive, and air-based (from satellites or airplanes) of infrared wavelengths may be biased. Temperature data were collected by ground and air-based sensors at 10 locations, and we want to test if they are different.

Location	Ground (^{o}C)	Air (^{o}C)	Difference (d_i)
1	46.9	47.3	-0.4
2	45.4	48.1	-2.7
3	36.3	37.9	-1.6
4	31.0	32.7	-1.7
5	24.7	26.2	-1.5
6	22.3	23.3	-1.0
7	49.8	50.2	-0.4
8	40.5	42.6	-2.1
9	37.7	39.4	-1.7
10	35.5	37.9	-2.4

Null vs Alternative hypothesis

- Hypotheses always refer to some population or model, not to a particular outcome. For this, we state $H_0,\ H_1$ in terms of population parameters.
- μ is the population's difference between ground and air temperatures.

$$\boxed{H_0 : \mu = 0 \text{ vs } H_1 : \mu \neq 0}$$

- If there is a reason to believe before any data collection that the parameter being tested is necessarily restricted to one particular "side" of H_0 then H_1 is one-sided.

Left-tailed test $H_0 : \mu = 0$ vs $H_1 : \mu < 0$

or

Right-tailed test $H_0 : \mu = 0$ vs $H_1 : \mu > 0$

Test statistic

- The test is based on a statistic that estimates the parameter that appears in the hypotheses.
- If H_0 is true then we expect the estimate to take a value “close" to the parameter value specified by H_0.
- Values of the estimate far from the parameter value in H_0 yield evidence against H_0.

$$\text{test-statistic} = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of estimate}}$$

- The test statistic is a *random variable* with a distribution that we know.

Test statistic for Example 1

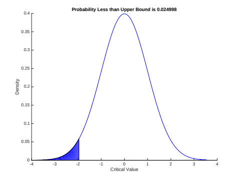
- Recall: test-statistic $= \frac{\text{estimate}-\text{hypothesized value}}{\text{standard deviation of estimate}}$
- The hypothesized value is $\mu = 0$.
- The estimate of the the mean is the average of differences provided by the data., i.e. for this data $\bar{d} = -1.55$.
- Let's **assume** that we know (typically not true) that the standard deviation of population is $\sigma = 2$

$$z^* = \frac{\bar{d} - 0}{\sigma/\sqrt{n}} = \frac{-1.55 - 0}{2/\sqrt{10}} = -2.4508$$

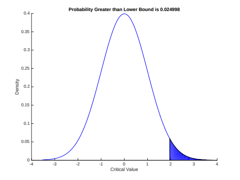
$p-$value

- The key to calculating the $p-$value is the sampling distribution of the test statistic.

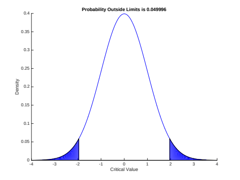
- **Assuming that the data is normal** (needs to be checked), z^* is a realization of Z from the standard normal distribution $N(0,1)$.



$H_1 : \mu < \mu_0, \; p = P(Z \leq z^*)$

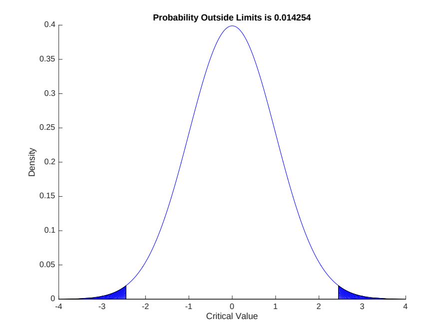


$H_1 : \mu > \mu_0, \; p = P(Z \geq z^*)$



$H_1 : \mu \neq \mu_0, \; p = 2P(Z \geq |z^*|)$

Back to example



Example: $p = 2P(Z \geq |-2.4508|) = 0.0143.$

- A mean difference as large as that observed would occur fewer than 14 times in 1000 samples (of size 10) if the population mean difference were 0.
- This is convincing evidence that the mean difference between ground and air-based measured temperatures is not zero.

$\alpha-$level of significance

- A $p-$value is more informative than a “reject-or-not" the H_0 .
- However, a quick way of assessment is needed.
- $\alpha-$level of significance shows how much evidence against H_0 you need as decisive.

$\alpha-$level of significance

- A $p-$value is more informative than a “reject-or-not" the H_0 .
- However, a quick way of assessment is needed.
- $\alpha-$level of significance shows how much evidence against H_0 you need as decisive.

- If $p-$**value** $\leq \alpha$, reject H_0 (accept H_1).
- If $p-$**value** $> \alpha$, then the data do not provide sufficient evidence to reject H_0.

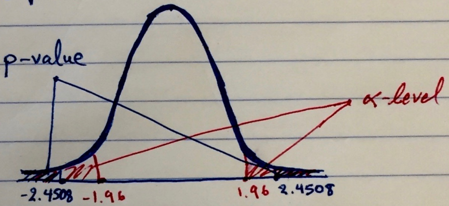
Example 1

H_0: mean difference $\mu = 0$ vs $H_1: \mu \neq 0$

Test-statistic: $z = \dfrac{-1.55 - 0}{2/\sqrt{10}} = -2.4508$ (under H_0 true)

p-value : $p = 2\,\mathbb{P}(Z \geq |z|) = 2\,\mathbb{P}(Z \geq 2.4508) = 0.0143$

α-level of significance : $\alpha = 0.05$ (probability)



p-value < α-level $\Rightarrow$ $\boxed{\text{Reject } H_0}$

Assumption: known variance

- $H_0 : \mu = c$ vs $H_1 : \mu \neq c$
- Recall $z - statistic = \frac{\bar{x}-c}{\sigma/\sqrt{n}}$
- Typically variance is unknown and needs to be estimated
- We do by the sample variance, s
- Test-statistic (mean of population):

$$t - statistic = \frac{\bar{x} - c}{s/\sqrt{n}}$$

- Test follows the same strategy (compute $p-$value and compare it with α)

Example 1

In agricultural modeling earth's temperature plays an important role. We want to compare ground vs air-based temperature sensors. Ground-based sensors are expensive, and air-based (from satellites or airplanes) of infrared wavelengths may be biased. Temperature data were collected by ground and air-based sensors at 10 locations, and we want to test if they are different.

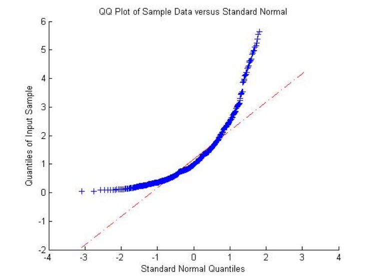
Location	Ground (°C)	Air (°C)	Difference (d_i)
1	46.9	47.3	-0.4
2	45.4	48.1	-2.7
3	36.3	37.9	-1.6
4	31.0	32.7	-1.7
5	24.7	26.2	-1.5
6	22.3	23.3	-1.0
7	49.8	50.2	-0.4
8	40.5	42.6	-2.1
9	37.7	39.4	-1.7
10	35.5	37.9	-2.4

Example 1

- $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$
- $t^* = \frac{-1.55-0}{0.7706/\sqrt{10}} = -6.458$
- $p-$value $= 2P(T_9 \geq 6.458) \approx 0.0002$
- A mean difference as large as that observed would occur fewer than 2 times in 10,000 samples (of size 10) if the population mean difference were 0.
- $p-$value $< \alpha$ so reject H_0.

Robustness of t tests

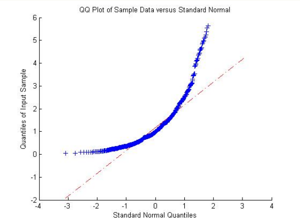
- $t-$tests are not robust against outliers ($\bar{x}, s$ not resistant to outliers).
  - Average height of soybean plants at $R1$ stage of their growth is 16". Imagine 3 plants with height 16" and 3 with 20", their average now is 18".
- $t-$tests robust against deviations from normality but not to outliers and presence of strong skewness



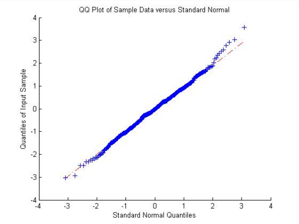
Right-skewed data

Some advice

- *Small sample size:* use $t-$test if the data are close to normal. If outliers are present do not use t.
- *Moderate sample size:* use $t-$test except in the presence of strong skewness or outliers.
- *Large sample size:* use $t-$test even for clearly skewed distributions (transform the data first, e.g. use logarithm)



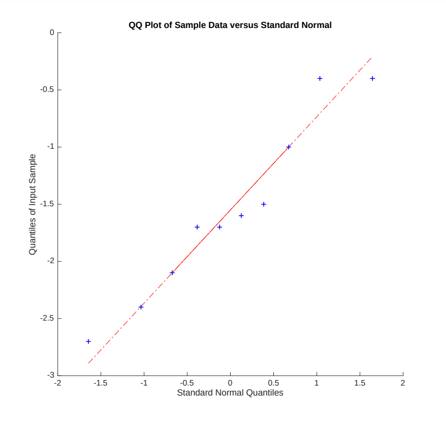
Right-skewed data



Log-transformed data

Checking for outliers and skewness

- Normal quantile plot
- Stemplot
- Boxplot



Example 1

- Inference for standard deviations, or proportions or parameters related to regression.
- Different hypotheses but same strategy.
- What only changes if the test-statistic and its associated distribution.
  - if small sample size: proportions use the binomial distribution
  - if large sample size: proportions use normal distribution

Summary

- The point of a test of significance is to provide a clear statement of the degree of evidence provided by the sample against H_0.
- We wrote $p-$value$\leq \alpha$, however there is no sharp border between significant and not significant.
- There is an increasingly strong evidence to reject H_0 as the $p-$value decreases.
- When H_0 (no effect or no difference) can be rejected at the usual level $\alpha = 0.05$, there is good evidence that an effect is present (could be small).
- Design carefully your study and plot your data.

To p or not to p?

- A Bayesian approach to hypothesis testing
- Attempt a statistical learning approach.
  - classification
  - clustering

Statistical Learning Example: Classification

- Consider a set of data obtained from soybean plants.

Statistical Learning Example: Classification

- Consider a set of data obtained from soybean plants.
- Each soybean has exactly one disease.

Statistical Learning Example: Classification

- Consider a set of data obtained from soybean plants.
- Each soybean has exactly one disease.
- Goal is to “understand" the characteristics of (4) different types of soybean diseases given features extracted from the plant so that when we are given a new soybean crop to be able to predict accurately what kind of disease it may have.

Statistical Learning Example: Classification

- Consider a set of data obtained from soybean plants.
- Each soybean has exactly one disease.
- Goal is to “understand" the characteristics of (4) different types of soybean diseases given features extracted from the plant so that when we are given a new soybean crop to be able to predict accurately what kind of disease it may have.
  - $p = 35$ predictors.
  - Based on condition and attributes of leaves, fruitpods, seeds, etc.

Statistical Learning Example: Classification

- Consider a set of data obtained from soybean plants.
- Each soybean has exactly one disease.
- Goal is to “understand" the characteristics of (4) different types of soybean diseases given features extracted from the plant so that when we are given a new soybean crop to be able to predict accurately what kind of disease it may have.
  - $p = 35$ predictors.
  - Based on condition and attributes of leaves, fruitpods, seeds, etc.
  - Only $n = 12$ **examples**, 3 for each disease class!

Dataset sampled from UC Irvine data Repository: https://archive.ics.uci.edu/ml/datasets/Soybean+(Small)

A Small Dataset of Soybeans

- Want to *maximize* the amount of data we can use to build the model on due to small sample size.

A Small Dataset of Soybeans

- Want to *maximize* the amount of data we can use to build the model on due to small sample size.

- Can we use all of the data to build the model?

A Small Dataset of Soybeans

- Want to *maximize* the amount of data we can use to build the model on due to small sample size.
- Can we use all of the data to build the model?
  - No! Need to validate the model to ensure our accuracy results are not biased!

A Small Dataset of Soybeans

- Want to *maximize* the amount of data we can use to build the model on due to small sample size.
- Can we use all of the data to build the model?
  - No! Need to validate the model to ensure our accuracy results are not biased!
- One option: leave one out cross validation.
  - Train the model on all but one data point, and see how the model performs on the held out instance.
  - Average out the error over all the instances.

Logistic Regression: A Statistics Approach

- We first model using Logistic Regression.

Logistic Regression: A Statistics Approach

- We first model using Logistic Regression.
- Logistic Regression attempts to model the log probability ratio

$$\log \frac{\text{probability of disease 1}}{\text{probability of disease 2}}$$

linearly in the predictors

Logistic Regression: A Statistics Approach

- We first model using Logistic Regression.
- Logistic Regression attempts to model the log probability ratio

$$\log \frac{\text{probability of disease 1}}{\text{probability of disease 2}}$$

  linearly in the predictors
- Parameters are estimated by some optimization method (maximum likelihood approach) and significance of predictors can be tested using significance tests (similar to what we discussed earlier).

Logistic regression for the soybeans dataset

- Employ logistic regression on 11 points
- Predict using the 12th point
- Measure the error (or accuracy) by answering the question “did I get it right?"
- Repeat 12 times so all points get held out once

Logistic regression for the soybeans dataset

- Employ logistic regression on 11 points
- Predict using the 12th point
- Measure the error (or accuracy) by answering the question “did I get it right?"
- Repeat 12 times so all points get held out once

Model	Accuracy
Logistic Regression	**91.67%**

Logistic regression for the soybeans dataset

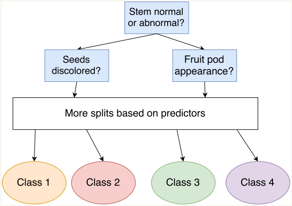
- Employ logistic regression on 11 points
- Predict using the 12th point
- Measure the error (or accuracy) by answering the question “did I get it right?"
- Repeat 12 times so all points get held out once

Model	Accuracy
Logistic Regression	**91.67%**

- 91.67% means that 11 out of 12 times I got it right.

Something different: Decision Tree

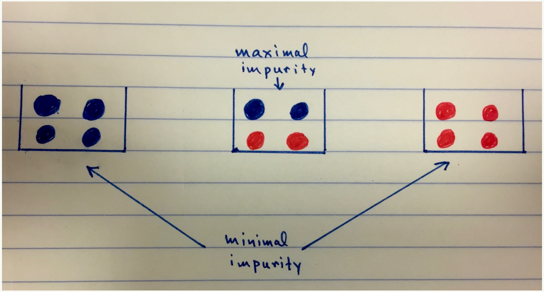
- Decision trees are recursive partitioning algorithms that come-up with a tree-like structure.
- These structures represent patterns in an underlying data set.



- The top node is the *root* node specifying a testing condition of which the outcome corresponds to a branch leading up to an internal node.
- The terminal nodes (*leaf* nodes) of the tree assign the classifications.

Decision tree

- Splitting decision
  - Strategy is to minimize the impurity at the leaves level



- Stopping decision
  - Avoid *overfitting:* if you split too much, one gets many pure classes but with very few members in it.
- Assignment decision: what class to assign to a leaf node?
  - Look at the majority class within the leaf node.

Back to soybean problem

- Now attempt to model using a decision-tree.

Back to soybean problem

- Now attempt to model using a decision-tree.
- Model attempts to build a tree (using 11 data) to create the most "pure" nodes at each step, and leaf nodes are labeled according to the majority class.

Back to soybean problem

- Now attempt to model using a decision-tree.
- Model attempts to build a tree (using 11 data) to create the most “pure” nodes at each step, and leaf nodes are labeled according to the majority class.
- New examples (the 12th) are then sent down the tree and classified according to the label of the leaf they end up in.

Back to soybean problem

- Now attempt to model using a decision-tree.
- Model attempts to build a tree (using 11 data) to create the most "pure" nodes at each step, and leaf nodes are labeled according to the majority class.
- New examples (the 12th) are then sent down the tree and classified according to the label of the leaf they end up in.

Model	Accuracy
Logistic Regression	91.67%
Decision Tree	75%

- This means 9 out of 12 were classified correctly
- Can we do better?

Turning Decision Trees into Random Forests

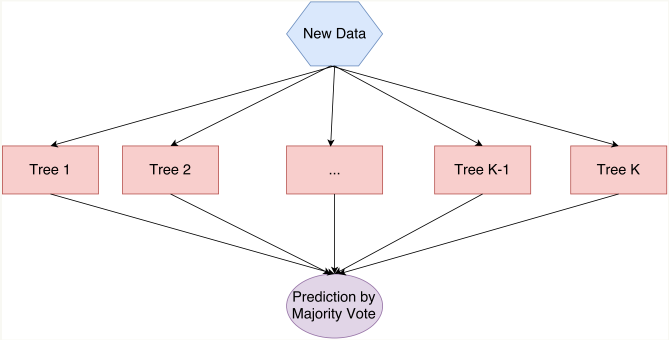
- Stochastically generate a large number of decision trees.

Turning Decision Trees into Random Forests

- Stochastically generate a large number of decision trees.
- At each split within each tree use a random subset of predictors instead of all of them.
- Predict on a new example (soybean) by taking the majority class prediction out of the K trees.

Turning Decision Trees into Random Forests

- Stochastically generate a large number of decision trees.
- At each split within each tree use a random subset of predictors instead of all of them.
- Predict on a new example (soybean) by taking the majority class prediction out of the K trees.



Take home message

Model	Accuracy
Logistic Regression	91.67%
Decision Tree	75%
Random Forest	**100%**

- Statistical Learning methods sometimes may be more appropriate than more “traditional” methods.

Take home message

Model	Accuracy
Logistic Regression	91.67%
Decision Tree	75%
Random Forest	**100%**

- Statistical Learning methods sometimes may be more appropriate than more “traditional” methods.
- When dealing with a small dataset, statistical learning techniques such as leave one out cross validation allow training on a large portion of the dataset while giving a good estimate for the true error.

Conclusion

- Dived into hypothesis testing bolts and nuts
- Use with caution hypothesis testing especially when small sample size data (e.g., look for outliers and skewness)
- Nothing is wrong with $p-$value however need to take it for what it is (a probability such that the smaller it is the stronger the evidence against the H_0).
- There are alternatives, e.g. statistical learning