

Introduction to Statistical Modeling

Theory Construction and Statistical Modeling



**Utrecht
University**

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Statistical Reasoning

Statistical Testing

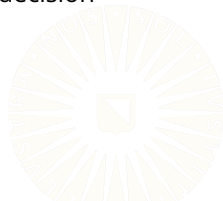
Statistical Modeling



Motivating Example

Imagine you are working for an F1 team. Your job is to use data from past seasons to optimize the baseline setup of your team's car.

- Suppose you have two candidate setups that you want to compare.
- For each setup, you have 100 past lap times.
- How do you distill those 200 lap times into a succinct decision between the two setups?



Motivating Example

Suppose I tell you that the mean lap time for Setup A is 118 seconds and the mean lap time for Setup B is 110 seconds.

- Can you confidently recommend Setup B?
- What caveats might you consider?



Motivating Example

Suppose I tell you that the standard deviation for the times under Setup A is 7 seconds and the standard deviation for the times under Setup B is 5 seconds.

- How would you incorporate this new information into your decision?



Motivating Example

Suppose I tell you that the standard deviation for the times under Setup A is 7 seconds and the standard deviation for the times under Setup B is 5 seconds.

- How would you incorporate this new information into your decision?

Suppose, instead, that the standard deviation of times under Setup A is 35 seconds and the standard deviation under setup B is 25 seconds.

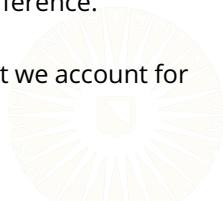
- How should you adjust your appraisal of the setups' relative benefits?



Statistical Reasoning

The preceding example calls for *statistical reasoning*.

- The foundation of all good statistical analyses is a deliberate, careful, and thorough consideration of uncertainty.
- In the previous example, the mean lap time for Setup A is clearly longer than the mean lap time for Setup B.
- If the times are highly variable, with respect to the size of the mean difference, we may not care much about the mean difference.
- The purpose of statistics is to systematize the way that we account for uncertainty when making data-based decisions.



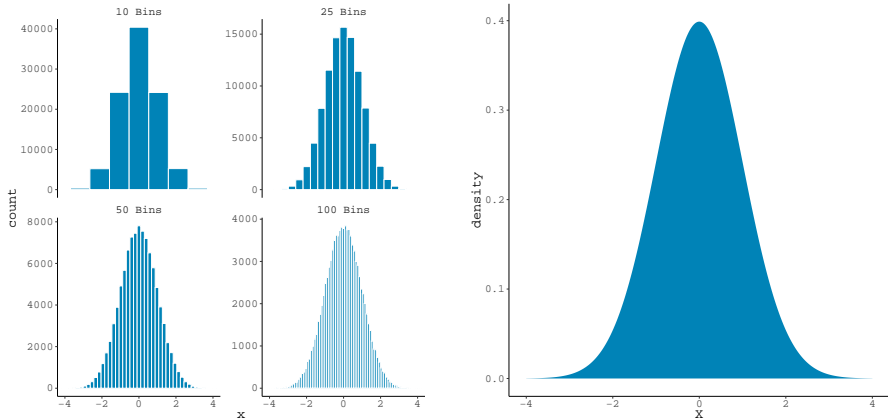
Probability Distributions

Statisticians (and anyone who uses statistics) quantify uncertainty using probability distributions.

- Probability distributions quantify how likely it is to observe each possible value of some probabilistic entity.
- Probability distributions are re-scaled frequency distributions.
- We can build up the intuition of a probability density by beginning with a histogram.

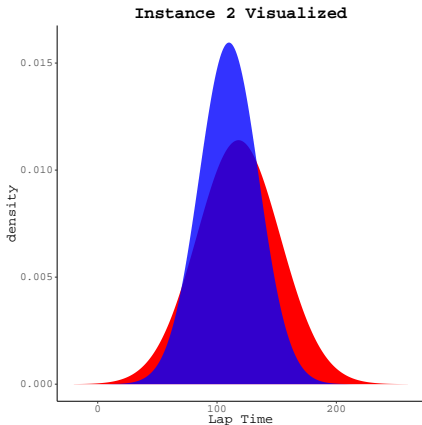
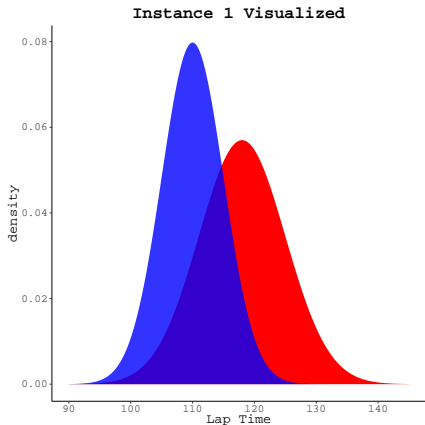


Probability Distributions



Reasoning with Distributions

We will gain insight by conceptualizing our example problem in terms of the underlying distributions of lap times.



Statistical Testing

In practice, we may want to distill the information in the preceding plots into a simple statistic so we can make a judgment.

- One way to distill this information and control for uncertainty when generating knowledge is through statistical testing.
 - When we conduct statistical tests, we weight the estimated effect by the precision of the estimate.
- A common type of statistical test, *Student's t-test*, follows this pattern:

$$t = \frac{\text{Estimate} - \text{Null-Hypothesized Value}}{\text{Variability}}$$



Statistical Testing

If we want to test the nil-null hypothesis of a zero mean difference, we get the following values for our example:

$$t = \frac{(\bar{X}_A - \bar{X}_B) - 0}{\sqrt{S_{A-B}^2 (n_A^{-1} + n_B^{-1})}}$$

where

$$\text{Estimate} = \bar{X}_A - \bar{X}_B$$

and

$$\begin{aligned} \text{Variability} &= \sqrt{S_{A-B}^2 (n_A^{-1} + n_B^{-1})} \\ &= \sqrt{\frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2} \left(\frac{1}{n_A} + \frac{1}{n_B} \right)} \end{aligned}$$

Statistical Testing

Applying the preceding formula to the first instantiation of our example problem produces:

$$\begin{aligned} t &= \frac{118 - 110 - 0}{\sqrt{\frac{(100-1)7^2 + (100-1)5^2}{100+100-2} \left(\frac{1}{100} + \frac{1}{100}\right)}} \\ &\approx \frac{8}{0.86} \\ &\approx 9.30 \end{aligned}$$



Statistical Testing

If we consider the second instantiation of our example problem, the effect does not change, but our measure of variability does:

$$V = \sqrt{\frac{(100 - 1)35^2 + (100 - 1)25^2}{100 + 100 - 2} \left(\frac{1}{100} + \frac{1}{100} \right)}$$
$$\approx 4.30$$

As a results, our test statistic changes to reflect our decreased certainty:

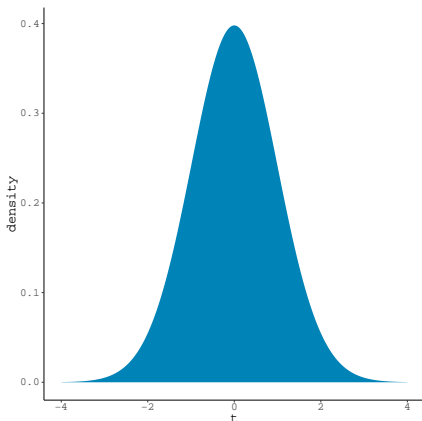
$$t \approx \frac{8}{4.30} \approx 1.86$$



Sampling Distributions

To evaluate the extremity of our test statistics, we compare our estimated statistic to its null-hypothesized sampling distribution.

- The *population* is defined by an infinite sequence of repeated tests.
 - The sampling distribution quantifies the relative frequencies of possible values the test statistic can take.
- The area of a region under the curve represents the probability of observing a *test statistic* within the corresponding interval.



Nice demo: http://onlinestatbook.com/stat_sim/sampling_dist/

Statistical Modeling

Statistical testing is a very useful tool, but it quickly reaches a limit.

- In experimental contexts, real-world “messiness” is controlled through random assignment, and statistical testing is a sufficient method of knowledge generation.
- Data scientists rarely have the luxury of being able to conduct experiments.
- Data scientists work with messy observational data and usually don't have questions that lend themselves to rigorous testing.

Data scientists need *statistical modeling*.



Statistical Modeling

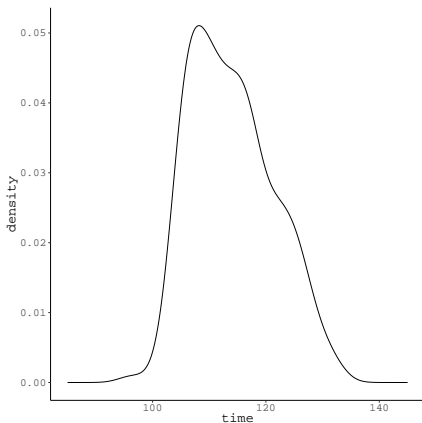
- Modelers attempt to build a mathematical representation of the (interesting aspects) of a data distribution.
- The model succinctly describes whatever system is being analyzed.
- Beginning with a model ensures that we are learning the important features of a distribution.
- The modeling approach is especially important in messy data science applications where clear a priori hypotheses are rare.



Statistical Modeling

To apply a modeling approach to our example problem we consider the combined distribution of lap times.

- The model we construct will explain variation in lap times based on interesting features.
- In this simple case, the only feature we consider is the type of setup.



Modeling our Example

Let's say we're willing to assume that the (conditional) distribution of lap times is normal.

$$Y_{time} \sim N(\mu, \sigma^2)$$

To get the same answer as our statistical test, we model the mean of the distribution of lap times, μ , using a single grouping factor.

$$\mu = \beta_0 + \beta_1 X_{setup}$$

$$Y_{time} \sim N(\beta_0 + \beta_1 X_{setup}, \sigma^2)$$



Modeling our Example

Since we're mostly interested in describing the mean lap time, we can express the above differently:

$$Y_{time} = \beta_0 + \beta_1 X_{setup} + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

After we fit this model to a sample, the parameters β_0 and β_1 are replaced by estimated statistics.

$$\begin{aligned}\hat{Y}_{time} &= \hat{\beta}_0 + \hat{\beta}_1 X_{setup} \\ &= 110 + 8X_{setup}\end{aligned}$$



Modeling our Example

We can easily fit this model in R:

```
lmOut <- lm(time ~ setup, data = exData)

partSummary(lmOut, -c(1, 2))

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 110.0000     0.6083   180.8   <2e-16
## setupA       8.0000     0.8602     9.3   <2e-16
##
## Residual standard error: 6.083 on 198 degrees of freedom
## Multiple R-squared:  0.304, Adjusted R-squared:  0.3005
## F-statistic: 86.49 on 1 and 198 DF,  p-value: < 2.2e-16
```

Two Modeling Traditions

Breiman (2001) defines two cultures of statistical modeling:

- Data models
- Algorithmic models

Two Modeling Traditions

Breiman (2001) defines two cultures of statistical modeling:

- Data models
- Algorithmic models

Data scientists use both types of models.

- Both types of model have strengths and weaknesses.
 - Data models tend to support a priori hypothesis testing more easily.
 - Data models also tend to provide more interpretable results.
 - Algorithmic models can't be beat for pure power.

Two Modeling Traditions

Breiman (2001) defines two cultures of statistical modeling:

- Data models
- Algorithmic models

Data scientists use both types of models.

- Both types of model have strengths and weaknesses.
 - Data models tend to support a priori hypothesis testing more easily.
 - Data models also tend to provide more interpretable results.
 - Algorithmic models can't be beat for pure power.
- Algorithmic models are currently preferred in cutting edge prediction/classification applications.

Two Modeling Traditions

Breiman (2001) defines two cultures of statistical modeling:

- Data models
- Algorithmic models

Data scientists use both types of models.

- Both types of model have strengths and weaknesses.
 - Data models tend to support a priori hypothesis testing more easily.
 - Data models also tend to provide more interpretable results.
 - Algorithmic models can't be beat for pure power.
- Algorithmic models are currently preferred in cutting edge prediction/classification applications.
- Many models can be viewed as data models or algorithmic models, depending on how they're used.

Characteristics of Models

Data models share several core features:

- Data models are built from probability distributions.
 - Data models are modular.
- Data models encode our hypothesized understanding of the system we're exploring.
 - Data models are constructed in a "top-down", theory-driven way.

Characteristics of Models

Data models share several core features:

- Data models are built from probability distributions.
 - Data models are modular.
- Data models encode our hypothesized understanding of the system we're exploring.
 - Data models are constructed in a "top-down", theory-driven way.

Algorithmic models are distinct from data models in several ways:

- Algorithmic models do not have to be built from probability distributions.
 - Often, they are based on a set of decision rules (i.e., an algorithm).
- Algorithmic models begin with an objective (i.e., a problem to solve) and seek the optimal solution, given the data.
 - They are built in a "bottom-up", data-driven way.

Data Modeling Example

Suppose we believe the following:

1. BMI is positively associated with disease progression in diabetic patients after controlling for age and average blood pressure.
2. After controlling for age and average blood pressure, the effect of BMI on disease progression is different for men and women.

We can represent these beliefs with a moderated regression model:

$$Y_{prog} = \beta_0 + \beta_1 X_{BMI} + \beta_2 X_{sex} + \beta_3 X_{age} + \beta_4 X_{BP} + \beta_5 X_{BMI} X_{sex} + \varepsilon$$

Data Modeling Example

We can use R to fit our model to some patient data:

```
library(rockchalk)

## Load the data:
dataDir <- "../..data/"
dDat    <- readRDS(paste0(dataDir, "diabetes.rds"))

## Fit the regression model:
fit <- lm(progress ~ bmi * sex + age + bp, data = dDat)
```

Data Modeling Example

```
partSummary(fit, -c(1, 2))

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -174.7986    27.0004  -6.474 2.58e-10
## bmi          7.2106     0.8922   8.082 6.34e-15
## sexmale     -90.1718    35.1134  -2.568 0.0106
## age          0.1691     0.2322   0.728 0.4670
## bp           1.4032     0.2385   5.884 7.97e-09
## bmi:sexmale   3.0257     1.3090   2.311 0.0213
##
## Residual standard error: 59.68 on 436 degrees of freedom
## Multiple R-squared:  0.4075, Adjusted R-squared:  0.4007
## F-statistic: 59.98 on 5 and 436 DF,  p-value: < 2.2e-16
```

Data Modeling Example

We can do a simple slopes analysis to test the group-specific effects of BMI on disease progression:

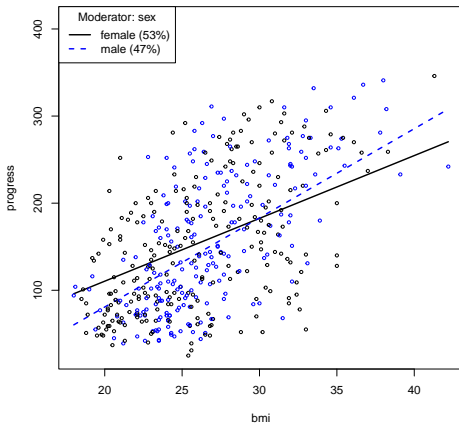
```
psOut <- plotSlopes(fit, plotx = "bmi", modx = "sex")
tsOut <- testSlopes(psOut)
```

```
tsOut$hypotests[ , -1]
```

##		slope	Std. Error	t value	Pr(> t)
## female	7.210575	0.8921929	8.081856	6.335264e-15	
## male	10.236323	1.0328739	9.910525	5.137409e-21	

Data Modeling Example

We can also visualize the simple slopes:



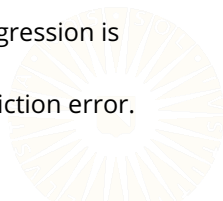
Algorithmic Modeling Example

Suppose we want to find the best predictors of disease progression among the variables contained in our dataset:

- Age
- BMI
- Blood Pressure
- Blood Glucose
- Sex
- Total Cholesterol
- LDL Cholesterol
- HDL Cholesterol
- Triglycerides
- Lamorigine

We could try *best-subset selection*.

- Fit a series of regression models wherein disease progression is predicted by all possible subsets of X variables.
- Choose the set of X variables that minimizes the prediction error.



Algorithmic Modeling Example

```
library(leaps)

## Save the predictor variables' names:
xNames <- grep(pattern = "progress",
                x       = colnames(dDat),
                invert  = TRUE,
                value   = TRUE)

## Train the models:
fit <- regsubsets(x      = progress ~ .,
                  data   = dDat,
                  nvmax  = ncol(dDat) - 1)

## Summarize the results:
sum <- summary(fit)
```

Algorithmic Modeling Example

```
sum$outmat
```

```
##          age bmi bp  tc  ldl hdl tch ltg glu sexmale
## 1  ( 1 )  " " "*" " " " " " " " " " " " " " " " "
## 2  ( 1 )  " " "*" " " " " " " " " " " "*" " " " "
## 3  ( 1 )  " " "*" "*" " " " " " " " " "*" " " " "
## 4  ( 1 )  " " "*" "*" "*" " " " " " " " "*" " " " "
## 5  ( 1 )  " " "*" "*" " " " " "*" " " " "*" " " "*"
## 6  ( 1 )  " " "*" "*" "*" "*" " " " " "*" " " "*"
## 7  ( 1 )  " " "*" "*" "*" "*" " " "*" "*" " " "*"
## 8  ( 1 )  " " "*" "*" "*" "*" " " "*" "*" "*" "*"
## 9  ( 1 )  " " "*" "*" "*" "*" "*" "*" "*" "*" "*"
## 10 ( 1 )  "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"

```

Algorithmic Modeling Example

```
## Variables selected by BIC:
xNames[with(sum, which[which.min(bic), -1])]

## [1] "bmi" "bp" "hdl" "ltg" "sex"

## Variables selected by Adjusted R^2:
xNames[with(sum, which[which.max(adjr2), -1])]

## [1] "bmi" "bp" "tc" "ldl" "tch" "ltg" "glu" "sex"

## Variables selected by Mallows's Cp:
xNames[with(sum, which[which.min(cp), -1])]

## [1] "bmi" "bp" "tc" "ldl" "ltg" "sex"
```

Prediction & Estimation

There are two other common objectives of statistical analyses.

1. Prediction/Classification
2. Estimation

Prediction/Classification involves building a statistical model to “guess” future values of some outcome.

- Weather forecasting
- Predicting the winner of an election
- Financial projections

Estimation focusses on getting the most accurate possible estimate of some real-world quantity.

- The number of refugees in a country
- The rates of obesity in a certain population
- The number of traffic accidents in a given area



Formal Modeling

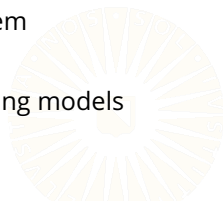
Smaldino (2017) distinguishes between two ways in which we can translate theories into models.

VERBAL MODELS

- Vague descriptions of the theory or phenomenon
- The description does not describe the theory explicitly enough to define a single, unambiguous representation
- The model could describe multiple phenomena

FORMAL MODELS

- Rigorously defines all the important aspects of a system
- Implies only one representation of the phenomenon
- Can be used to rule-out potential theories by comparing models



Example



References

- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231.
- Smaldino, P. E. (2017). Models are stupid, and we need more of them. *Computational Social Psychology*, 311–331.

