1 Motivation: the Selection Problem

The way in which statistical inference is applied has changed considerably over the past two decades or so. Indeed, if statistical inference was originally conceived for a setting in which a scientist began by selecting a model, gathered data, and checked the pertinence of that model on her dataset, we have to acknowledge that this order has been recently more or less reversed: scientists are now more generally confronted to situations in which they have access to a vast amount of data, from which they have to select a ”pertinent” model. However, the reversal of these stages is not an insignificant statistical technicality: inference, and in particular, p-values and confidence intervals, are distorted and thus have to be adapted to this new setting where one ”peeks” at the data before modeling.

In the previous two lectures, we focused on a theory developed by Benjamini and Yekutieli (2005), who tackle this adaptation-after-selection problem by constructing adjusted confidence intervals for a number $R$ of selected parameters.

The aim of this lecture is to provide yet another new framework for developing inference after model selection (based on the work of Will Fithian, Dennis Sun and Jonathan Taylor [2]) which, in particular, tackles the problem of adjusting the p-values of our selected model.

To get a clearer idea of what the problem truly is, and what we are looking for, let us consider the following motivating examples:

- **Example 1: the Iowa Republican Poll:**
  The following table lists the results of a May 2015 poll for the Iowa Republican Vote (there are 667 samples in this poll):
<table>
<thead>
<tr>
<th>Candidate</th>
<th>Percentage of votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scott Walker</td>
<td>21 %</td>
</tr>
<tr>
<td>Rand Paul</td>
<td>13 %</td>
</tr>
<tr>
<td>Marco Rubio</td>
<td>13 %</td>
</tr>
<tr>
<td>Ted Cruz</td>
<td>12 %</td>
</tr>
</tbody>
</table>

The question that we want to ask ourselves is this: is Scott Walker truly winning? This is a different question than asking before the poll if Scott Walker is winning, because in the first one, what we are truly asking is this: “is the current leader in the poll truly leading the vote?” whereas the second one focuses on Scott Walker specifically. Is the parameter we are interested in $\pi_{SW}/\pi_{RP}$ or $\pi_{1st\ best}/\pi_{2nd\ best}$, and how should we estimate these? We face the winner’s curse and regression to the mean, and we should expect that our naive estimate of these parameters will be biased upwards. This particular example will be our guideline thread throughout this lecture.

**Example 2:** Suppose we have a set of data $X \in \mathbb{R}^{n \times p}$ with a response variable $Y$, and we want to model $Y$ as:

$$Y = X\beta + \epsilon$$

selecting only a subset of variables $M \subset \{1 \cdots p\}$, and we want to test the regression coefficients on the selected model. But how do we construct p-values for these regression coefficients conditionally on the fact that they have been selected?

## 2 Random Model, random null

Similarly to traditional statistics, we want our level $\alpha$-test for testing a hypothesis $H_0$ under model $M$ to control the error rate at level $\alpha$. However, we now make a distinction between:

- **the Selective Error Rate:** $\mathbb{P}_{M_0,H}(\text{reject } H_0 \mid (M,H_0) \text{ selected})$.
- **the Nominal Error Rate:** $\mathbb{P}_{M_0,H}(\text{reject } H_0)$.

(This Nominal Error Rate is the one that we are used to in traditional statistics)

### 2.1 Data Splitting

A valid adaptive selection procedure would be to split the available data into two independent sets $(Y_1,Y_2)$ and :

(a) Use the first set for model selection ($Y_1$ is our exploratory dataset)
(b) Use the second set for inference ($Y_2$ is the confirmatory dataset)
In this setting, the selective error rate and nominal Error Rate are identical, since:

\[ P_{M_0,H}(\text{reject } H_0 | (M,H_0) \text{ selected}) = P_{M_0,H}(\text{reject } H_0) \]

computed on a new set of data that is independent of the first.

This procedure is known as data splitting, and is attractive in that it mimics the traditional statistical framework: we are selecting a hypothesis based on our observations on \( Y_1 \), and making inference on \( Y_2 \) as though our hypothesis was observed “ahead of time” (i.e., it behaves similarly as if we formulated a hypothesis thanks to \( Y_1 \), forgot all about \( Y_1 \), and gathered data \( Y_2 \) to test our hypothesis). Hence this setting, no inference adjustment is required.

There are however several objections to data splitting:

- it reduces the amount of data available for selection
- it reduces the amount of data available for inference
- and it might not always be possible to achieve (we might have autocorrelated data)

### 2.2 Data Carving

Another way around the problem of adapting inference is “Data carving”, which tries to use some of the information and randomness contained in the observed sample to control directly the selective error rate.

In this setting, the data is again revealed in stages. The main idea is that we look at sufficiently enough data to select the model, and inference is then realized on the entire dataset conditioning on our selection. The two stages (model selection and inference) can be expressed mathematically in terms of filtrations as follows:

\[ F_0 \subseteq F(1_{\{Y \in A\}}) \subseteq F(Y) \]

where \( A \) is the event that hypothesis \( H_0 \) is selected under model \( M \) \( (A = \{(M,H_0) \text{ selected}\} = \{\text{Scott Walker has higher counts than the rest}\} ) \).

In other words, our two-step selection-inference procedure can thus be seen as follows:

- 0. We start off with no knowledge whatsoever on the data (i.e, we are in \( F_0 \))
- 1. Based on our observations on \( Y \), we select \( (M,H_0) \), and we thus move from \( F_0 \) to \( F(1_{\{Y \in A\}}) \)
- 2. In stage 2 (the inference-making stage), we want to find the true value of \( Y \), which mathematically corresponds to moving from \( F(1_{\{Y \in A\}}) \) to \( F(Y) \).

Now, in data splitting, where we condition on \( Y_1 \) rather than \( 1_{\{Y \in A\}} \), the process could be described as follows:

\[ F_0 \subseteq F(1_{\{Y_1 \in A\}}) \subseteq F(Y_1) \subseteq F(Y_1,Y_2) \]
Hence it becomes more obvious that data carving is less stringent than data splitting, because by conditioning on \(1_{Y_1 \in A}\) instead of \(Y_1\), we are leaving a little more room for randomness (and hence, discarding less information): we are no longer automatically rejecting \(Y_1\) if the event \(Y_1 \in A\) is extremely surprising, but rather waiting till stage 2 to reject it if we observe it again.

To get a clearer idea of what data carving achieves, let us go back to example 2 (the regression problem), which we want to solve using the Lasso procedure. Now, depending on the value of \(Y\), the Lasso will select different variables. We can thus partition the space into "Lasso parameter-selection" regions, as shown on the following graph:

![Figure 1: Partition of \(\mathbb{R}^2\) into parameter-selection regions by the Lasso](image)

In this particular example, the Lasso will select variables \(X_1\) and \(X_3\), so we will want to test in stage 2:

\[
H_0^A : \beta_1^A = 0 \quad H_0^A : \beta_3^A = 0
\]

where \(A\) is the event that \(X_1\) and \(X_3\) have been selected (i.e, the yellow region in the plot).

### 3 Selective Hypothesis testing and the Republican Primary

#### 3.1 One goal: achieving Selective Hypothesis Testing

Suppose we observe \(Y\) sampled from an unknown distribution \(F\). Our testing problem (or “question \(q\”) can be modeled by a pair \(q = (M, H_0)\) where \(M(q)\) is a family of distributions, and \(H_0(q) \subseteq M(q)\) is our null hypothesis (without loss of generality, \(H_1 = M \setminus H_0\))

In order to test our hypotheses, we want to use the 2-stage adaptive process of data carving as follows:

- **1. Selection**: defining \(Q\) to be the set of all possible questions, we select a subset \(\hat{Q}(Y) \subseteq Q\) of questions to test on the data
2. **Inference:** for each $q = (M, H_0) \in \hat{Q}$, we test $H_0$ on $M \setminus H_0$

To be more concrete, let us go back to the Republican primaries problem: suppose we want to model the Republican Iowa scores by a sample of data $Y$ drawn from a multinomial model:

$$Y \sim \text{Multinomial} \left( n = 667, \pi \right)_{M(q)}$$

where $\pi = (\pi_1, \ldots, \pi_{667})$ is the associated vector of probabilities corresponding to each of the 667 candidates. The hypothesis that we want to test under that model is $H_0(q) = \left\{ \pi_{SW} \leq \max_{j \neq SW} \pi_j \right\}$

Let us remind ourselves that we want to condition on statistics on the event that the question $q$ was asked, which we denote as $A_q = \{ q \in \hat{Q}(y) \}$. Letting $\phi_q$ be our hypothesis test, we want $\phi_q$ to control **selective type I error rate at level** $\alpha$, i.e. we want to have

$$\mathbb{E}_F [\phi_q(Y) | A_q] \leq \alpha \quad \forall F \in H_0.$$

### 3.2 Application: the Iowa Republican Poll

Now, concretely, back to our Iowan problem, how could we test if Scott Walker is winning? i.e., how can we test the null $H_0 : \pi_{SW} \leq \max_{i \neq SW} \pi_i$?

As just pointed out, we simply want to test (for a fixed $i$): $H_i = \cup_{j \neq i} H_{i \leq j}$ (where $H_{i \leq j}$ denotes the hypothesis $\pi_i \leq \pi_j$), where $i$ realizes the maximum of observed counts. From now on, we are therefore interested in the selection event $A_i = \{ X_i > \max_{j \neq i} X_j \}$ (i.e. it is not really Scott Walker that we are interested in, but whoever comes first in the poll results). Hence, in this example we are not effectively selecting a model, just a simple hypothesis.

Without loss of generality, let $i = 1$. Our selective test is as follows.

1. First we construct a selective $p$-value $p_{12}$ for $H_{1 \leq 2}$ on $A_1$. If we condition on the counts for everyone except candidates 1 and 2 and on $X_1 + X_2$, we have $X_1 | (X_1 + X_2, X_3, \ldots, X_k) \sim \text{Bin}(X_1 + X_2, \frac{\pi_1}{\pi_1 + \pi_2})$. If we further condition on $X_1 > X_2$, we have a truncated binomial, on which we want to test $\pi_1 \leq \pi_2$, or equivalently, the binomial probability $\frac{\pi_1}{\pi_1 + \pi_2} \leq \frac{1}{2}$. To construct the $p$-value $p_{12}$, with $m = X_1 + X_2$ for this test we compute

$$p_{12} = \mathbb{P}(\text{Binom}(m, 1/2) \geq X_1 | \text{Binom}(m, 1/2) > m/2) = \frac{\mathbb{P}(\text{Binom}(m, 1/2) \geq X_1)}{\mathbb{P}(\text{Binom}(m, 1/2) > m/2)}$$

$$= 2\mathbb{P}(\text{Binom}(m, 1/2) \geq X_1) = 2 \sum_{k=X_1}^{X_1+X_2} \binom{X_1+X_2}{k} 2^{-(X_1+X_2)}.$$

Note that a serious downside to the approach is that you have to condition on so much that very little room is left for randomness... Which decreases the power of our test.
2. We iterate this procedure to construct p-values for the different hypotheses \( H_{i \leq j} \) for all \( j \).

To test the union \( H_1 = \bigcup_{j>1} H_{1 \leq j} \), we combine these p-values by defining: \( p_1 = \max_{j \geq 2} p_{1j} \), and reject if \( p_1 \leq \alpha \). This is valid since

\[
\mathbb{P}(p_1 \leq \alpha | A_1) \leq \min_{j>1} \mathbb{P}(p_{1j} \leq \alpha) \leq \alpha
\]

since we constructed level-\( \alpha \) tests for each \( H_{1,j} \).

In particular, if we apply this procedure to the Iowa Republican Poll, we deduce that Scott Walker’s support is at least 20 % higher than the other candidates.

4 Estimation

4.1 Estimation of multivariate normal mean

Consider the problem of estimating vector \( \mu \) in the model

\[ X \sim N(\mu, \sigma^2 I) \]

with \( \sigma \) known, which can be equivalently written as

\[ X_i = \mu_i + \sigma Z_i, \quad Z_i \iid N(0, 1). \]

We use loss function \( \ell(\hat{\mu}, \mu) = \| \hat{\mu} - \mu \|^2 = \sum_{i=1}^p (\hat{\mu}_i - \mu_i)^2 \), which corresponds MSE risk \( R(\hat{\mu}, \mu) = \mathbb{E}_\ell(\hat{\mu}, \mu) \).

The most natural estimator for \( \mu \), which is also the maximum likelihood estimator, is \( X \) itself. This is reasonable because no one has information about \( \mu_i \) except \( X_i \), so intuitively we should not expect our estimate of \( \mu_i \) to depend on \( X_j \) for \( j \neq i \). This estimator has constant risk

\[ R(\hat{\mu}_{\text{MLE}}, \mu) = \mathbb{E}[\|X - \mu\|^2] = \sum_{i=1}^p \mathbb{E}(X_i - \mu_i)^2 = p\sigma^2. \]

The setup here is more general than the case of having only one observation for each \( \mu_i \). If we have \( n \) observations for each \( \mu_i \), we can take the average of the observations for each \( i \) to get \( \overline{X}_i \), which is also normal with some variance \( \sigma^2 \). Any estimator based on \( X_i \) we have in the sample size one case corresponds to the same estimator applied to \( \overline{X}_i \) in the sample size \( n \) case, with the same risk (with \( \sigma^2 \) the variance of the sample mean rather than of a single observation).

For a long time, the folklore was that the maximum likelihood estimator was the ”best” way to estimate the population mean, where, to be precise, what we mean by best is admissible. That is, we cannot find another estimator \( \hat{\mu} \) which achieves a better MSE for all \( \mu \) simultaneously. The ”for all” qualifier is important, because we can easily construct estimators which perform better for some values of the parameter \( \hat{\mu} \). For example, \( \hat{\mu} = 0 \) has 0 risk when \( \mu = 0 \) but has arbitrarily large risk when \( \mu \) is far from 0.
4.2 Stein’s Phenomenon

It turns out that for $p = 1$ and 2 this folklore theorem is true: the MLE is admissible. But surprisingly, for $p \geq 3$, this folklore theorem is false! Stein hinted at this in 1956 and proved it with James in 1961, when they introduce an estimator $\hat{\mu}_{JS}$ which has lower MSE for all values of the parameter $\mu$. Their estimator is

$$\hat{\mu}_{JS} = \left[1 - \frac{p - 2}{\sigma^2 \|X\|^2} = w\right] X.$$

What this estimator does is shrink our observation $X$ towards the origin, where the amount of shrinkage is data dependent, controlled by parameter $w$.

**Theorem 1 (JS ’61).** For all $\mu \in \mathbb{R}^p$, $\mathbb{E}\|\hat{\mu}_{JS} - \mu\|^2 < \mathbb{E}\|\hat{\mu}_{MLE} - \mu\|^2$.

The risk function for the MLE is flat, and the JS estimator has risk function below this. What this theorem seems to say is that you can get an improved estimate by aggregating unrelated information.

**References**
