Statistical Modeling for Poets

Vittorio Addona

Mathematics, Statistics, and Computer Science
Macalester College
addona@macalester.edu

MOSAIC Workshop
Institute for Mathematics and its Applications
July 1, 2010.

Introduction

- What do we mean by *Statistics for Poets*?
  - A book by Bernard Berkowitz (1991): *Statistics for Poets: A manual For those So Inclined*
  - Statistics as a liberal art.
  - The lowest level statistics class.
- At Macalester, the course that most closely alligns with this description is *Math 153: Data Analysis and Statistics*.
- What do we mean by *Statistical Modeling for Poets*?
  - Actually, I'm not really sure. It's a catchy title! But it does conjure up the right mental image ...

Motivation

- There is an on-going revolution at Macalester in the way introductory statistics is taught. Math 155 incorporates modeling, statistics, calculus, and computation.
- Many reasons why Math 153 should evolve with Math 155:
  1. Several justifications used to argue for Math 155-style course can be applied to Math 153 (e.g. students entering with more, "cookbook" procedures are poor, formulas are of little value).
  2. Not constructive to have a big gap between Math 153 and Math 155 (e.g. we would be handicapping students who take Math 153, like it, and want to do more statistics!).
  3. Not everyone takes Math 155 (and "poets" use statistics too!).

The Old Math 153

1. Numerical and Graphical descriptive statistics: mean, median, sd, histograms, boxplots, etc.
2. Probability: Unions, Intersections, Complements, Conditional.
3. Simpson's Paradox.
4. Discrete and Continuous Probability Distributions, in particular, the Binomial and the Normal.
5. Sampling Distributions, in particular, the CLT.
6. One sample confidence intervals.
7. One sample testing, including power, multiple testing, etc.
8. Two sample inference.
9. Correlation and Simple linear regression.

The New Math 153

1. Numerical and Graphical descriptive statistics (**Intro to R**)
2. **Correlation and Univariate models (w/out inference)**
3. Probability: Unions, Intersections, Complements, Conditional.
4. Simpson's Paradox, **Multivariate models (w/out inference)**
5. Binomial and Normal Distributions, and, briefly, the CLT.
6. Confidence intervals (**on general model coefficients**).
7. Hypothesis testing (**on general model coefficients**).
8. Power, multiple testing, etc.
9. **Interaction models**.
10. **Multi-collinearity**, $\mathbf{R^2}$, **Adjusted-**$\mathbf{R^2}$, **F-tests**.

Example # 1: Body fat data

- **BodyFat.csv** contains body circumference measurements for 252 men, along with estimates of the percentage of body fat determined by underwater weighing.

	BodyFat	Age	Weight	Height	Neck	Chest	Abdomen
1	12.3	23	154.25	67.75	36.2	93.1	85.2
2	6.1	22	173.25	72.25	38.5	93.6	83.0
3	25.3	22	154.00	66.25	34.0	95.8	87.9
4	10.4	26	184.75	72.25	37.4	101.8	86.4
5	28.7	24	184.25	71.25	34.4	97.3	100.0
6	20.9	24	210.25	74.75	39.0	104.5	94.4

- For more details, read the handout on this data set.
- We will fit models for BodyFat variable.

Example # 1: Body fat data

- What do you think the relationship is between BodyFat and Height: positive, negative, or none?

Example # 1: Body fat data

- What do you think the relationship is between BodyFat and Height: positive, negative, or none?

```
> lm(BodyFat~Height,data=bf)$coefficients
(Intercept)      Height
 33.4944938  -0.2044753
```

Example # 1: Body fat data

- What do you think the relationship is between BodyFat and Height: positive, negative, or none?

```
> lm(BodyFat~Height,data=bf)$coefficients
(Intercept)      Height
 33.4944938  -0.2044753
```

```
> summary(lm(BodyFat~Height,data=bf))$coefficients
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 33.4944938 10.1095831  3.313143 0.001059050
Height      -0.2044753  0.1439210 -1.420747 0.156636257
```

Example # 1: Body fat data

- Nearly all students say that there is a *negative* relationship. Why?

Example # 1: Body fat data

- Nearly all students say that there is a *negative* relationship. Why?
- Because they are (tacitly) holding Weight fixed.

Example # 1: Body fat data

- Nearly all students say that there is a *negative* relationship. Why?
- Because they are (tacitly) holding Weight fixed.

```
> summary(lm(BodyFat~Height+Weight,data=bf))$coefficients
                Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 32.4046404 7.46991673  4.338019 2.090472e-05
Height      -0.7025959 0.11178118 -6.285457 1.452798e-09
Weight       0.2013835 0.01393161 14.455145 8.444594e-35
```

Example # 1: Body fat data

- Consider the relationship between BodyFat and Weight.

Example # 1: Body fat data

- Consider the relationship between BodyFat and Weight.

```
> summary(lm(BodyFat~Weight,data=bf))$coefficients
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-12.0515789	2.58138682	-4.668645	4.950144e-06
Weight	0.1743886	0.01423722	12.248779	2.473116e-27

Example # 1: Body fat data

- Consider the relationship between BodyFat and Weight.

```
> summary(lm(BodyFat~Weight,data=bf))$coefficients
                Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -12.0515789 2.58138682 -4.668645 4.950144e-06
Weight        0.1743886 0.01423722 12.248779 2.473116e-27
```

- What happens if we include Abdomen circumference?

Example # 1: Body fat data

- Consider the relationship between BodyFat and Weight.

```
> summary(lm(BodyFat~Weight,data=bf))$coefficients
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-12.0515789	2.58138682	-4.668645	4.950144e-06
Weight	0.1743886	0.01423722	12.248779	2.473116e-27

- What happens if we include Abdomen circumference?

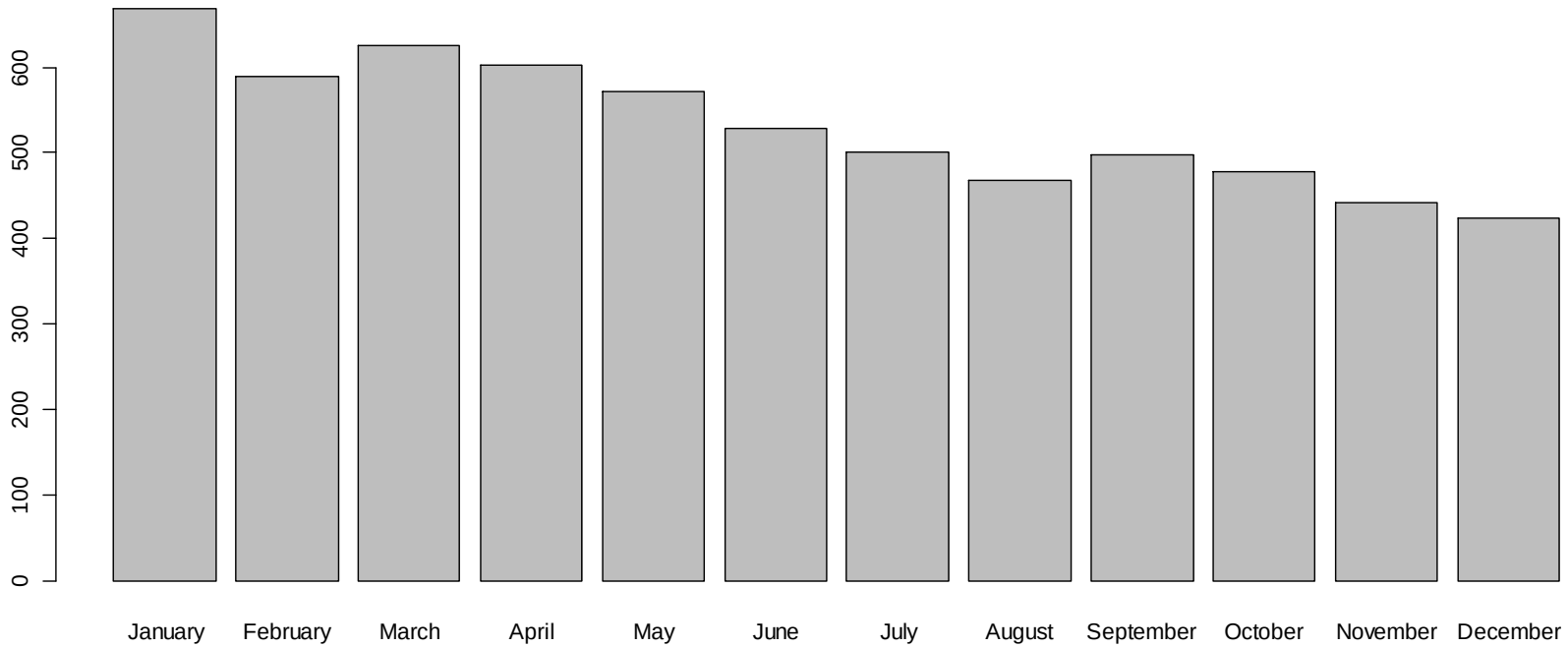
```
> summary(lm(BodyFat~Weight+Abdomen,data=bf))$coefficients
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-45.9523732	2.60501273	-17.639980	9.790287e-46
Weight	-0.1480031	0.02080957	-7.112259	1.207060e-11
Abdomen	0.9895044	0.05671626	17.446573	4.486591e-45

Example # 2: Birthdays of hockey players ... a possible M-CAST?

- This example was motivated by Gladwell, in *Outliers: The Story of Success*.
- Data was gathered on every player who played in the National Hockey League's (NHL's) regular season through the 2008-09 season. We are interested in the birthdays of 6,391 players.
- We covered this after introducing discrete distributions, including the Binomial. The students are shown a plot of the birth month frequencies, and asked to comment.

Birth month frequencies of NHL players through 2008-09



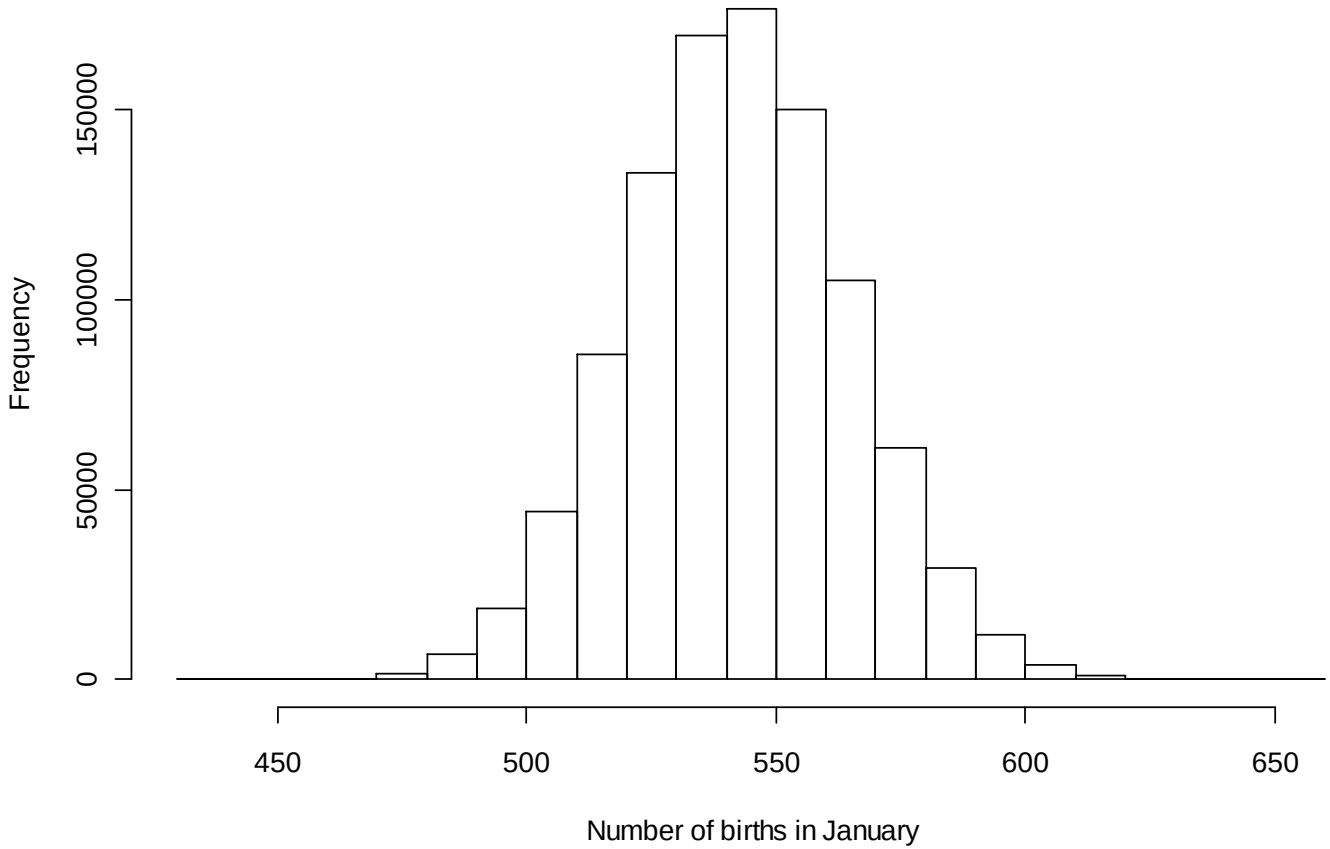
Example # 2: Birthdays of hockey players ... a possible M-CAST?

- Observations:
  - "There is a downward trend", or
  - "January has many more observations than December"
- The students explain why the graph is surprising: they had expected to see something "flat".
- If appropriate, the professor can introduce a goodness-of-fit test. I use this example to allude to hypothesis testing, by performing a simple simulation.

Example # 2: Birthdays of hockey players ... a possible M-CAST?

- At this point, the students have seen *rbinom*, the function in *R* which simulates the flip of a coin. Here, we simply have 6,391 coin flips.
- How do we know the chance of "heads" (born in January)?
- We don't ... but we believe that it should be, roughly, 1/12, or 31/365. This is purely an assumption (our H_0).
- Under this H_0, we can replay history 1,000,000 times, say: *JanDist = rbinom(1000000, size=6391, prob=31/365)*
- A histogram of JanDist represents the sampling distribution for the number of births in January.

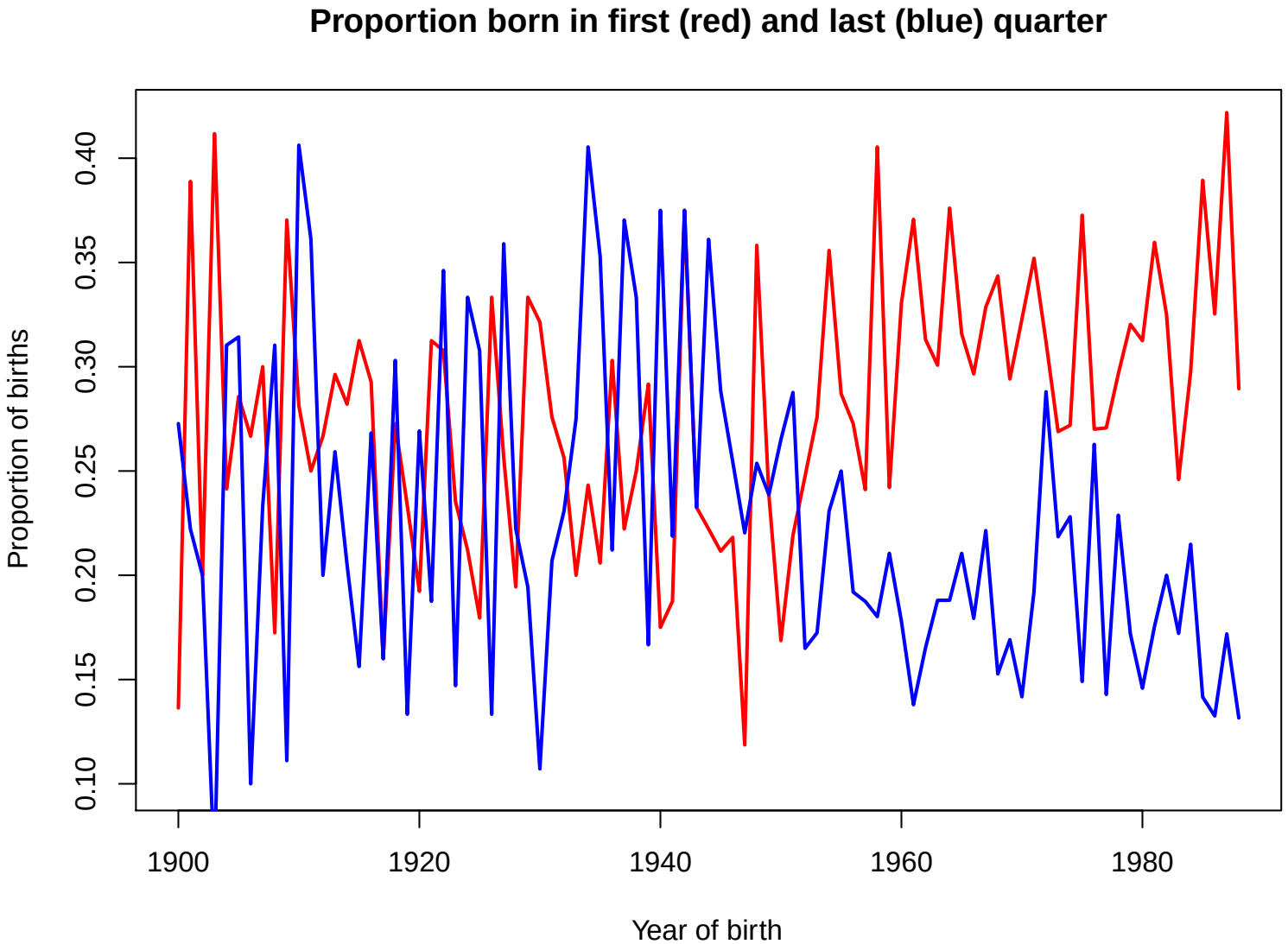
Sampling distribution, under H_0, for January births



Example # 2: Birthdays of hockey players ... a possible M-CAST?

- How compatible is the observation of 668 births in January with H_0?
- How likely is it that we observe (at least) 668 births, under H_0? (easily computable with 1 command line).
- Students may question the assumption of uniform births, and this can lead to an excellent discussion (e.g. what data could we obtain to form a more reasonable H_0?).
- A more intriguing topic for debate: Why is this happening?
- Phenomenon is known as the *relative age effect*: in a group of kids, there are performance advantages of being the eldest.

When did the RAE begin to manifest itself?



Example # 3: Judging bias in diving scores

- **Diving2000.csv** contains information on all 10,787 dives at the 2000 Olympics in Sydney. Relevant variables are: Diver, Country, JScore, Judge, JCountry, and Same. JScore is the judge's score, and Same (Yes or No) indicates whether the judge is from the same country as the diver.

	Diver	Country	JScore	Same
1	ABALLI Jesus-Iory	CUB	7.0	No
2	ABALLI Jesus-Iory	CUB	7.5	No
3	ABALLI Jesus-Iory	CUB	7.5	No
4	ABALLI Jesus-Iory	CUB	8.0	No

- Is there a bias in favor of divers when they are from the same country as a judge?

Example # 3: Judging bias in diving scores

- There are many possible ways to answer this question, but let's start with something simple:

```
> summary(lm(JScore~Same,data=dive))$coefficients
             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 6.813711 0.01436646 474.279200 0.000000e+00
SameYes     0.648072 0.08420448   7.696408 1.521706e-14
```

Example # 3: Judging bias in diving scores

- There are many possible ways to answer this question, but let's start with something simple:

```
> summary(lm(JScore~Same,data=dive))$coefficients
             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 6.813711 0.01436646 474.279200 0.000000e+00
SameYes     0.648072 0.08420448   7.696408 1.521706e-14
```

- Possible conclusions?

Example # 3: Judging bias in diving scores

- There are many possible ways to answer this question, but let's start with something simple:

```
> summary(lm(JScore~Same,data=dive))$coefficients
             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 6.813711 0.01436646 474.279200 0.000000e+00
SameYes     0.648072 0.08420448   7.696408 1.521706e-14
```

- Possible conclusions?
- What does this ignore? Many modeling paths can be taken, and this always leads to a good class discussion.

Example # 3: Judging bias in diving scores

- Previous analysis treats all judges the same ...

Example # 3: Judging bias in diving scores

- Previous analysis treats all judges the same ...
- Can control for individual judges to account for leniency/severity of each.

Example # 3: Judging bias in diving scores

- Previous analysis treats all judges the same ...
- Can control for individual judges to account for leniency/severity of each.
- Can fit an interaction model between *Same* and *Judge*.

Example # 3: Judging bias in diving scores

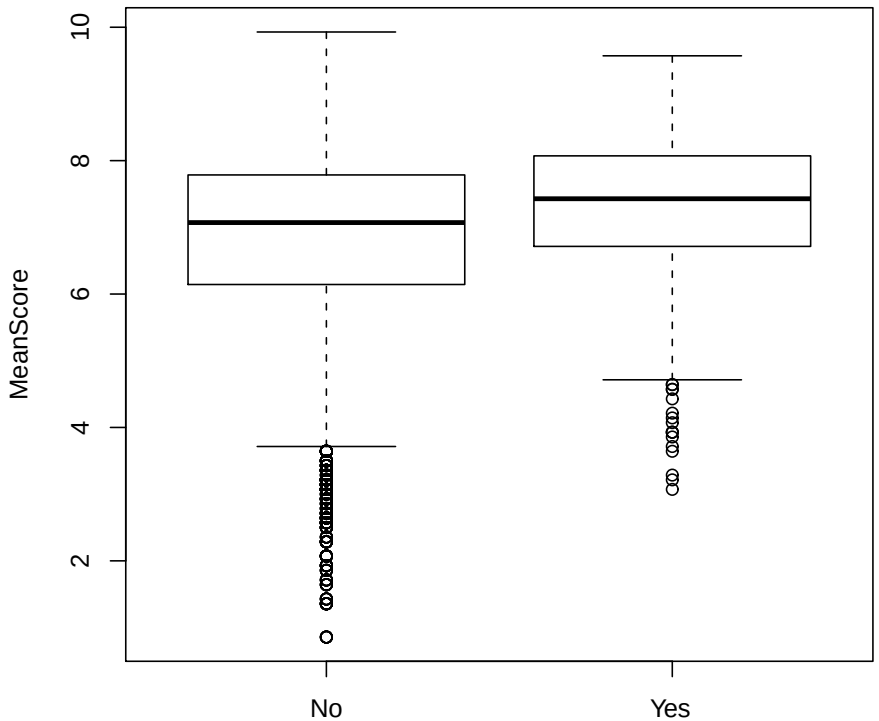
- Previous analysis treats all judges the same ...
- Can control for individual judges to account for leniency/severity of each.
- Can fit an interaction model between *Same* and *Judge*.
- Perhaps a bigger issue is that dives associated with "Same=Yes" are better. Why might this be?

Example # 3: Judging bias in diving scores

- Previous analysis treats all judges the same ...
- Can control for individual judges to account for leniency/severity of each.
- Can fit an interaction model between *Same* and *Judge*.
- Perhaps a bigger issue is that dives associated with "Same=Yes" are better. Why might this be?
- Can try to control for the quality of the dive ...

Example # 3: Judging bias in diving scores

- My naive way of doing this is by finding the mean score given to each dive using all other judges other than the one being considered. Call this variable *MeanScore*.



Example # 3: Judging bias in diving scores

- Now model the deviances from MeanScore by Same:

```
> Deviances = dive$JScore-MeanScore
> summary(lm(Deviances~Same,data=dive))$coefficients
                Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.007256755 0.00398040 -1.823122 6.831257e-02
SameYes      0.249294972 0.02332987 10.685658 1.612278e-26
```

Example # 3: Judging bias in diving scores

- Now model the deviances from MeanScore by Same:

```
> Deviances = dive$JScore-MeanScore
> summary(lm(Deviances~Same,data=dive))$coefficients
                Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.007256755 0.00398040 -1.823122 6.831257e-02
SameYes      0.249294972 0.02332987 10.685658 1.612278e-26
```

- Magnitude of bias has decreased, but it is still positive, and significant ...

Concluding Remarks

- Math 153 has no pre-requisite.
- Math 153 sometimes includes a few virulent "anti-statistics" types, but the material is accessible to them, and the response has been very favorable!
- With regards to material in the introductory statistics class, we have focused on changing *what* we cover, in addition to *how* we cover it.
- The "it depends" answer does not fit well with the traditional introductory statistics course.
- Gathering relevant, interesting, data is a constant challenge ...

References

- Addona, V. and P.A. Yates (2009). A Closer Look At the Relative Age Effect in the National Hockey League. Submitted to *Journal of Quantitative Analysis in Sports*.
- Emerson, J.W., M. Seltzer, and D. Lin (2009). Assessing Judging Bias: an Example from the 2000 Olympic Games. *The American Statistician* 63(2): 124-131.
- Johnson, R.W. (1996). Fitting Percentage of Body Fat to Simple Body Measurements. *Journal of Statistics Education*, 4(1).