Final Projects

PHY 110C
Evan Ott

March 28, 2014

For your final project, you will need to combine the data analysis and typesetting skills you've used all semester. You will select one of the projects below, then send us the completed LaTeX and *Mathematica* files. As always, send it to data.analysis.physics@gmail.com. The write-up should be a (semi-)formal report on how you did what you did. We want to see the math you used, some graphs / tables that help explain what you're analyzing, and a little about the *Mathematica* constructs used. You don't need to tell us about basic things like using a `Table`, but if you had a difficult integral or used a built-in distribution, you can mention it. We're focused on the data analysis, so you don't need to include historical context, much lab setup (in the case of a Modern Lab report), etc. This is our way of determining if you have mastered the coursework over the course of the semester, so don't hold back on showing off.

1 Home Field Advantage

For this project, you will investigate the truth behind the "Home Field (Court, etc.) Advantage:" professional teams tend to win during games played in their local stadium. The data for this project were condensed from play-by-play data from Basketball Geek [1], selecting all games from the 2008-2009 NBA season (excluding those that went into overtime) for a total of 1125 games. The format of the data is in Table 1. The data is on the assignment page of the online textbook.

Complete the sections outlined below, then write up results in a LaTeX document. Be sure to include some figures, tables, equations, or other structures.

1.1 Game Results

To investigate the Home Field Advantage, we first need to see if such an effect is present in the data at the end of the game. Calculate the difference in score at

Home	$Q1_H$	$Q2_H$	$Q3_H$	$Q4_H$	Away	$Q1_A$	$Q2_A$	$Q3_A$	$Q4_A$
BOS	20	14	19	16	CLE	19	12	13	17
CHI	19	14	22	20	MIL	25	18	19	18
				⋮					

Table 1: Representation of data set. Numerical values are number of points scored *in the quarter* (QX_H is points scored by home team, QX_A is for the away team). Each row is a different game. Included are the abbreviations for the teams.

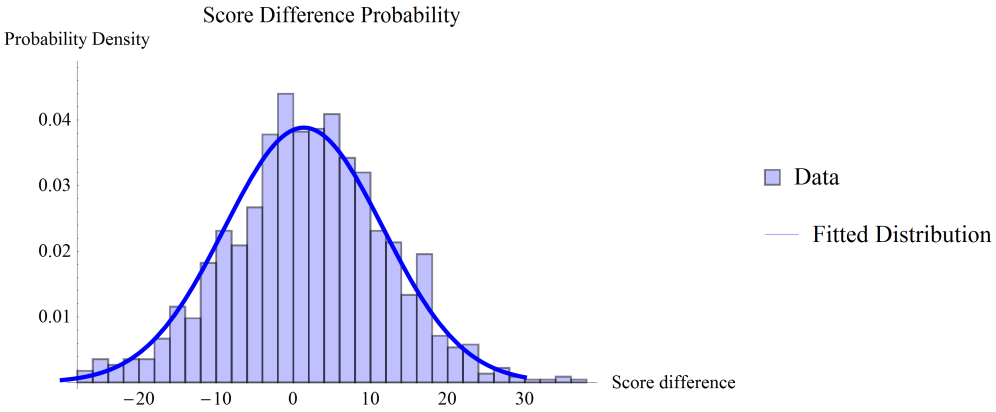

Figure 1: Example graphic indicating data and fitted normal distribution.

the end of each game (scores from each quarter add, in case you aren't the sports type). If a Normal (Gaussian) distribution seems appropriate for the data, fit the mean μ and standard deviation σ of the score difference at the end of the game. Based on the mean, is there an advantage to playing at home? Based on the mean and spread of the data, how often should we expect a home team to win (using no information about player stats, etc.)? Show visually that the data agrees with a Normal distribution by overlaying a probability density function histogram and probability density function of the fitted Normal distribution (something similar to Figure 1).

1.2 In-Game Progression

Now, let's characterize how the point difference changes throughout a game (on average). If we make the assumption that the play in each quarter is independent (first quarter performance has no bearing on fourth quarter, etc.), then we should expect the general properties of adding independent Normally-distributed variables:

- Taking $\mathcal{N}(\mu_1, \sigma_1) \pm \mathcal{N}(\mu_2, \sigma_2)$ gives a combined mean of $\mu = \mu_1 \pm \mu_2$.
- Taking $\mathcal{N}(\mu_1, \sigma_1) \pm \mathcal{N}(\mu_2, \sigma_2)$ gives a combined variance of $\sigma^2 = \sigma_1^2 + \sigma_2^2$ (variances *always* add).

Let's also assume (for the moment – we'll test it momentarily) that each quarter is approximately the same ($\mu_1 \approx \mu_2$, $\sigma_1 \approx \sigma_2$, etc.).

From this, can you come up with a model for the mean difference in score as a function of percentage of time played so far during a game? For the standard deviation of the difference in score as a function of percentage of time played so far during a game? This creates functions $\mu(t)$ and $\sigma(t)$ where $t \in [0, 1]$. We know the final score and spread from subsection 1.1 , and we know the initial score and spread as well (starts at 0–0 with $\sigma = 0$). In your report, please discuss how you arrived at the functions you did.

Hint: $\mu(t)$ and $\sigma(t)$ should have the boundary properties outlined above:

- $\mu(0) = 0$
- $\sigma(0) = 0$
- $\mu(1) = \mu = \mu_1 + \mu_2 + \mu_3 + \mu_4$
- $\sigma^2(1) = \sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2$

then also be able to predict the distribution at half-time ($\mu(.5) = \mu_1 + \mu_2$, and $\sigma^2(.5) = \sigma_1^2 + \sigma_2^2$). If this part stumps you, feel free to ask your student teachers about it.

1.3 Model-Fitting Game Progression

With the model for $\mu(t)$, $\sigma(t)$ from subsection 1.2, use a `LinearModelFit` or `NonlinearModelFit` in *Mathematica* to fit the coefficient based on the data at each quarter (rather than just using the final distribution). For example, you'll have *Mathematica* fit the function $\mu(t)$ to the data:

$$\{\{0,\ 0\},\ \{0.25,\ \mu_1\},\ \{0.5,\ \mu_1+\mu_2\}\ \{0.75,\ \mu_1+\mu_2+\mu_3\},\ \{1,\ \mu_1+\mu_2+\mu_3+\mu_4\}\}$$

These fitting functions provide the estimate of the parameter *and* the standard error in the parameter (use the "`ParameterTable`" field of the fit – see the documentation) along with the p-value.

The p-value reported is the probability of getting an estimated coefficient this far from 0 if there is no dependence on the coefficient. For example, a p-value of 0.9 says that 90% of the time, we would fit a coefficient this different from 0 where there really is no dependence. Based on the reported p-values, was the fit sufficient?

If so, using the estimate and error for the mean vs. time and for the standard deviation vs. time, is the model based only on final scores consistent with the

fitted data? More formally, are the coefficients used in the simple model from subsection 1.2 consistent with those determined by the fits in this subsection? How many "standard errors" away from each other are they? Is the difference statistically significant at the 99% confidence level (i.e., should we be surprised at the result from the simple model if it is indeed drawn from the one we fitted)?

1.4 Results

Using results from subsection 1.3, present a visual indicating how all the games compare with the average and standard deviation of the distribution at the end of each quarter. You may want to use the `ErrorListPlots` module in *Mathematica*. Feel free to explore different representations and include them in your submitted *Mathematica* document.

1.5 Bonus: Single-Game Predictions

Does the Home Field Advantage work for a specific team? Select a team and use a variation of the techniques above to determine if they are more likely to win home games or away games.

1.6 Bonus: Working with Normal Distributions

Using the same "independence" approximations above, find the actual score at half-time for a few games, then predict the probability the team should win assuming the game is representative. In other words, based on $\mu_{3-4} = \mu_3 + \mu_4,\ \sigma^2_{3-4} = \sigma^2_3 + \sigma^2_4$ and $\mu_f = D_2 + \mu_{3-4},\ \sigma_f = \sigma_{3-4}$ (with D_2 being the score difference at half-time), determine what certainty you have that the home team will win the game.

2 Hubble Constant

Details to come.

3 Modern Lab

Details to come.

References

[1] http://www.basketballgeek.com/downloads/2008-2009.regular_season.zip