Introduction to Statistics

Statistics: The Art and Science of Learning from Data

1. What is Statistics?

- Statistics, as a subject matter, is a set of scientific principles and techniques that are useful in reaching conclusions about populations and processes when the available information is both limited and variable, that is, statistics is the science of learning from data.

According to Google's Chief Economist Hal Varian: "I keep saying the sexy job in the next ten years will be statisticians. ... The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades ...." (McKinsey Quarterly, January 2009)

Questions Statistics Can Answer

- Do people who eat a high fiber cereal for breakfast, on average, weigh less than people who skip breakfast?

  Statistics can NOT be used to determine whether you personally would weigh less if you ate a high fiber cereal, but it can tell us if there is evidence the average weight of high fiber cereal eaters is less than the average weight of all people who skip breakfast.

- Do females live longer than males?

  http://www.dailymail.co.uk/sciencetech/article-1323571/Why-women-live-longer-men-Male-bodies-genetically-disposable.html

  Statistics only can tell on average whether female live longer than male.

An interesting case study example (Outliers (2008) by Malcolm Gladwell)

"One warm, spring day in May of 2007, the Medicine Hat Tigers and the Vancouver Giants met for the Memorial Cup Hockey Championships in Vancouver, British Columbia. The Tigers and the Giants were the two finest teams in the Canadian Hockey League, which in turn is the finest junior hockey league in the world. These were the future stars of the sport – seventeen-, eighteen- and nineteen- year olds who had been skating and shooting pucks since they were barely more than toddlers." (*Outliers page 15*)

"You can't buy your way into Major Junior A hockey. ... Nor does it matter if you live in the most remote corner of the most northerly province in Canada. If you have ability, the vast network of hockey scouts and talent spotters will find you, and if you are willing to work to develop that ability, the system will reward you. The system is based on individual merit – and both of these words are important. Players are judged on their own performance, not anyone else's, and on the basis of their ability, not on some other arbitrary fact.
Or are they?" (*Outliers page 17*)

Here is the player roster of the 2007 Medicine Hat Tigers. Take a close look and see if you can spot anything strange about it. (*Outliers pages 20-21*)

Number	Name	Position	L/R	Height	Weight	Birth Date	Hometown
9	Brennan Bosch	C	R	5' 8"	173	February 14, 1988	Martensville, SK
11	Scott Wasden	C	R	6' 1"	188	January 4, 1988	Westbank, BC
12	Colton Grant	LW	L	5' 9"	177	March 20, 1989	Standard, AB
14	Darren Helm	LW	L	6'	182	January 21, 1987	St. Andrews, MB
15	Derek Dorsett	RW	L	5' 11"	178	December 20, 1986	Kindersley, SK
16	Daine Todd	C	R	5' 10"	173	January 10, 1987	Red Deer, AB
17	Tyler Swynstun	RW	R	5' 11"	185	January 15, 1988	Cochrane, AB
19	Matt Lowry	C	R	6'	186	March 2, 1988	Neepawa, MB
20	Kevin Undershute	LW	L	6'	178	April 12, 1987	Medicine Hat, AB
21	Jerrid Sauer	RW	R	5' 10"	196	September 12, 1987	Medicine Hat, AB
22	Tyler Ennis	C	L	5' 9"	160	October 6, 1989	Edmonton, AB
23	Jordan Hickmott	C	R	6'	183	April 11, 1990	Mission, BC
25	Jakub Rumpel	RW	R	5' 8"	166	January 27, 1987	Hrnciarovce, SLO
28	Bretton Cameron	C	R	5' 11"	168	January 26, 1989	Didsbury, AB
36	Chris Stevens	LW	L	5' 10"	197	August 20, 1986	Dawson Creek, BC
3	Gord Baldwin	D	L	6' 5"	205	March 1, 1987	Winnipeg, MB
4	David Schlemko	D	L	6' 1"	195	May 7, 1987	Edmonton, AB
5	Trever Glass	D	L	6'	190	January 22, 1988	Cochrane, AB
10	Kris Russell	D	L	5' 10"	177	May 2, 1987	Caroline, AB
18	Michael Sauer	D	R	6' 3"	205	August 7, 1987	Sartell, MN
24	Mark Isherwood	D	R	6'	183	January 31, 1989	Abbotsford, BC
27	Shayne Brown	D	L	6' 1"	198	February 20, 1989	Stony Plain, AB
29	Jordan Benfield	D	R	6' 3"	230	February 9, 1988	Leduc, AB
31	Ryan Holfeld	G	L	5' 11"	166	June 29, 1989	LeRoy, SK
33	Matt Keetley	G	R	6' 2"	189	April 27, 1986	Medicine Hat, AB

Quarter	Months	O (Observed Births)
1	January, February, March	14
2	April, May, June	6
3	July, August, September	3
4	October, November, December	2
	Total	25

There is very strong evidence that births are not evenly distributed across the 4 quarters, with births in the 1st quarter more likely. Why is there an uneven birth distribution?

"... an iron law of Canadian hockey: in any elite group of hockey players – the very best of the best – 40 percent of the players will have been born between January and March, 30 percent between April and June, 20 percent between July and September, and 10 percent between October and December...."

The explanation is quite simple. It has nothing to do with astrology, nor is there anything magical about the first three months of the year. It's simply that in Canada the eligibility cutoff for age class hockey is January 1. A boy who turns ten on January 2, then could be playing alongside someone who doesn't turn ten until the end of the year – and at that age, in pre-adolescence, a twelve-month gap in age represents an enormous difference in physical maturity.

This being Canada, the most hockey-crazed country on earth, coaches start to select players for the traveling rep squad – the all-star teams – at the age of nine or ten, and of course they are more likely to view as talented the bigger and more coordinated players, who have the benefit of critical extra months of maturity.

And what happens when a player gets chosen for a rep squad? He gets better coaching, and his teammates are better, and he plays fifty or seventy-five games a season instead of twenty games a season like those left behind in the "house" league, and he practices twice as much as, or even three times more than, he would have otherwise. In the beginning, his advantage isn't so much that he is inherently better but only that he is a little older. But, by the age of thirteen or fourteen, with the benefit of better coaching and all that extra practice under his belt, he really is better, so he's the one more likely to make it to the Major Junior A league and from there into the big leagues.

Barnsley argues that these kinds of skewed age distributions exist whenever three things happen: selection, streaming, and differentiated experience." (*Outliers* pages 22-25)

"In Outliers, I want to convince you that these kinds of personal explanations of success don't work. People do not rise from nothing. ... The people who stand before kings may look like they did it all by themselves. But in fact they are invariably the beneficiaries of hidden advantages and extraordinary opportunities and cultural legacies that allow them to learn and work hard and make sense of the world in ways others cannot. ... The culture we belong to and the legacies passed down by our forebears shape the patterns of our achievements in ways we cannot begin to imagine.

2. Four Main Phases of Statistical Analysis:

Example 2.1: A study conducted by researchers at Pennsylvania State University investigated whether time perception, a simple indication of a person's ability to concentrate, is impaired during nicotine withdrawal. The study results were presented in the paper "Smoking Abstinence Impairs Time Estimation Accuracy in Cigarette Smokers" (*Psychopharmacology Bulletin [2003]: 90-95*).

To test this hypothesis, 22 non-smokers (12 males, 10 females) and 20 daily cigarette smokers (12 males, 8 females) were asked to estimate the duration of a 45-second period of time in a laboratory setting. Smokers participated in two sessions: once after smoking ad-lib (Abbreviation for the Latin "ad libitum" meaning "at pleasure" and "at one's pleasure, as much as one desires, to the full extent of one's wishes.") and once after objectively confirmed 24-hour smoking abstinence.
The researchers wanted to determine whether smoking abstinence had a negative impact on time perception, causing elapsed time to be overestimated.