

## Turning data into knowledge to solve real world problems



Christopher R. Bilder, Ph.D.  
Department of Statistics  
University of Nebraska-Lincoln  
[www.chrisbilder.com](http://www.chrisbilder.com)



## 12 years ago...

- The year is 1993
- Pearl Jam records second CD, Vs.,
  - Daughter, Go, Elderly Woman Behind the Counter in a Small Town
- Bill Clinton was inaugurated as the 42<sup>nd</sup> president
- Movies - Jurassic Park, Sleepless in Seattle
- Husker football - 1993 regular season went undefeated
- University of Nebraska-Omaha undergraduate student
  - Math major
  - Planned to be an actuary
    - Two summer internships
    - Passed 4 exams under old system
  - Wanted to go on to graduate school
    - Math?
    - Actuarial Science?

## 12 years ago...

- University of Nebraska-Omaha student (continued)
  - Taking Introduction to Probability and Statistics II
    - Like UWL's Mathematical Statistics II (MTH 442)
    - Hypothesis testing
      - Use for decision making!
      - Scientifically prove a hypothesis or statement
    - Go to graduate school for statistics!
- Continuing the timeline
  - BS 1994 in Mathematics from University of Nebraska-Omaha
  - MS 1996 and PhD 2000 in Statistics from Kansas State University
    - Internships – Pharmaceutical company, Dept. of Energy government lab
    - Consult with students and professors in sociology, agriculture,...
    - Taught courses like UWL's Elementary Statistics (MTH 145)
  - Assistant Professor at University of Nebraska-Lincoln in its new Department of Statistics

## Purpose

- Tell you a little about statistics
- 3 actual examples that come from my teaching and research
  - Turning data into knowledge to solve real world problems
- About statistics at University of Nebraska-Lincoln
- See [www.chrisbilder.com/statistics](http://www.chrisbilder.com/statistics) for more information

## Grocery store prices

- Undergraduate teaching example for an introductory course like UWL's MTH 145
- How could you determine which grocery store, Super Wal-Mart or Festival Foods, has lower average prices?



- Paired or dependent two sample hypothesis test for  $\mu_{\text{Wal-Mart}} - \mu_{\text{Festiv.Foods}}$
- Sample the same items at each store

## Grocery store prices

- Undergraduate teaching example for an introductory course like UWL's MTH 145
- How could you determine which grocery store, Dillon's or Food-4-Less in Manhattan, KS, has lower average prices?



- Paired or dependent two sample hypothesis test for  $\mu_{\text{Dillon's}} - \mu_{\text{Food-4-Less}}$
- Sample the same items at each store
- Only cereals from Fall 1998

## Grocery store prices

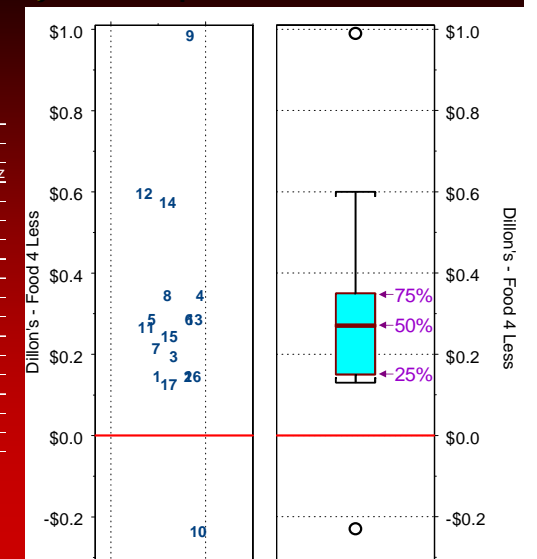
- Sample:

	Item	Dillon's	Food-4-Less	Difference
2	Malt-o-meal - Golden Puffs, 18oz	\$1.99	\$1.84	\$0.15
3	Quaker Oats - Life Cereal: Original, 21oz	\$3.69	\$3.49	\$0.20
4	Cheerios, 20oz	\$4.59	\$4.24	\$0.35
5	Cheerios, 15oz	\$3.79	\$3.50	\$0.29
6	Wheaties, 18oz	\$3.89	\$3.60	\$0.29
7	Kellogg's Funpack, 8 9/16oz	\$2.89	\$2.67	\$0.22
8	Kellogg's Variety Pack 9 5/8oz.	\$3.49	\$3.14	\$0.35
9	Kellogg's Frosted Mini-Wheats Bite Size	\$3.49	\$2.50	\$0.99
10	Kellogg's Frosted Mini-Wheats, 16oz	\$2.50	\$2.73	-\$0.23
11	Kellogg's Frosted Flakes, 15oz	\$3.19	\$2.92	\$0.27
12	Our Family Frosted Flakes, 20oz.	\$2.50	\$1.90	\$0.60
13	Kellogg's Crispix, 12oz.	\$3.49	\$3.20	\$0.29
14	Our Family - Raisin Bran, 20oz	\$2.50	\$1.92	\$0.58
15	Kellogg's Smart Start, 13.3oz	\$3.49	\$3.24	\$0.25
16	Grape Nuts, 24oz	\$3.00	\$2.85	\$0.15
17	Frosted Alpha Bits, 15oz	\$3.00	\$2.87	\$0.13

## Grocery store prices

- Do you think there are mean differences?

	Item
1	Malt-o-meal - Tootie Fruities, 15oz
2	Malt-o-meal - Golden Puffs, 18oz
3	Quaker Oats - Life Cereal: Original, 21oz
4	Cheerios, 20oz
5	Cheerios, 15oz
6	Wheaties, 18oz
7	Kellogg's Funpack, 8 9/16oz
8	Kellogg's Variety Pack 9 5/8oz.
9	Kellogg's Frosted Mini-Wheats Bite Size
10	Kellogg's Frosted Mini-Wheats, 16oz
11	Kellogg's Frosted Flakes, 15oz
12	Our Family Frosted Flakes, 20oz.
13	Kellogg's Crispix, 12oz.
14	Our Family - Raisin Bran, 20oz
15	Kellogg's Smart Start, 13.3oz
16	Grape Nuts, 24oz
17	Frosted Alpha Bits, 15oz



## Grocery store prices

- Paired two sample hypothesis test
  - $H_0: \mu_{\text{Dillon's}} - \mu_{\text{Food-4-Less}} = 0$
  - $H_a: \mu_{\text{Dillon's}} - \mu_{\text{Food-4-Less}} \neq 0$
  - $t = 4.77$ ,  $p\text{-value} = 0.0002$ ,  
95% confidence interval:  $0.1644 < \mu_{\text{Dillon's}} - \mu_{\text{Food-4-Less}} < 0.4274$
  - Reject equal mean prices
- If price was the only consideration, what store should one shop at?
- Assumptions
  - Are prices and selection at these two stores indicative of all stores?
  - Normal populations
  - The sample was taken in 1998; what about now?
  - Finite populations

## Placekicking

- MS report – applying statistics to new problems or investigating new methodology
  - 120 page book!
  - Reduced version published in *Chance* in 1998
- Find a model to estimate the probability of success for placekicks in the NFL
- Video
  - January 7, 1996
  - Playoff game
  - Indianapolis Colts 10  
Kansas City Chiefs 7
  - Lin Elliott of KC will attempt a 42-yard field goal to tie the game and send it into overtime
  - [Field goal video](#)

## Placekicking

- What factors affect the probability of success for NFL placekicks?
  - Distance
  - Pressure – How do you quantitatively measure?
  - Wind
  - Grass vs. artificial turf
  - Dome vs. outdoor stadium
- Collected data >1,700 placekicks during the 1995 NFL season
- Find the best logistic regression model of the form

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}$$

where  $p$  is the probability of success

$x_i$  for  $i=1, \dots, k$  are independent variables

$\beta_i$  measures the effect of  $x_i$  on  $p$  for  $i=1, \dots, k$

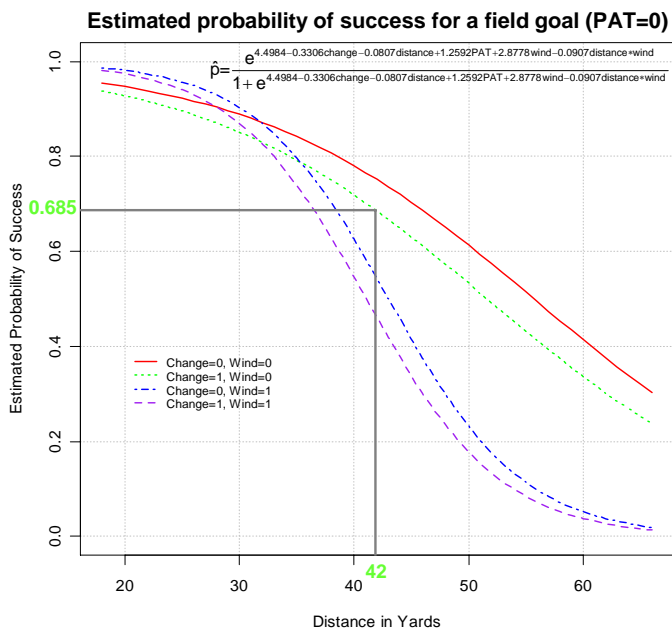
$e \approx 2.718$ ;  $\ln(e) = 1$

## Placekicking

- The  $\beta_i$ 's are parameters which are estimated through maximum likelihood estimation
- Estimated model
 
$$\hat{p} = \frac{e^{4.4984 - 0.3306 \text{change} - 0.0807 \text{distance} + 1.2592 \text{PAT} + 2.8778 \text{wind} - 0.0907 \text{distance} * \text{wind}}}{1 + e^{4.4984 - 0.3306 \text{change} - 0.0807 \text{distance} + 1.2592 \text{PAT} + 2.8778 \text{wind} - 0.0907 \text{distance} * \text{wind}}}$$
  - Change: lead change = 1, non-lead change = 0
  - Distance: distance in yards
  - PAT: point after touchdown = 1, field goal = 0
  - Wind: windy (speed > 15 MPH) = 1, non-windy = 0
- What is the estimated probability of success for Elliott's field goal?

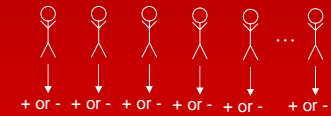
Change	Distance	PAT	Wind
1	42	0	0

- Conditions:
- Estimated probability of success:  $\hat{p} = 0.6850$
- 90% confidence interval for probability of success:  
 $0.6298 < p < 0.7402$



## Hepatitis C prevalence

- MS/PhD research – forwarding statistical science
- Hepatitis C
  - Viral infection that causes cirrhosis and cancer of the liver
  - Screening blood donors is important to prevent transmission
- Questions:
  - How can people be tested in a cost effective and timely manner?
    - Blood bank setting
  - What is the prevalence of Hepatitis C in a population?
    - Public health study
- Individual testing
  - Each blood sample is tested individually
  - Problems:
    - Costly
    - Time



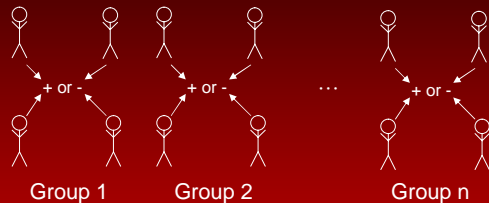
www.chrisbilder.com

14 of 26

## Hepatitis C prevalence

### Group testing

- Pool the blood samples together to form  $n$  groups of size  $s$



- If the GROUP sample is negative, then all  $s$  people do not have the disease
- If the GROUP sample is positive, then at least ONE of the  $s$  people have the disease
  - Follow-up procedures can be done to determine which people are positive
- Strategy works well when prevalence of a disease is small

www.chrisbilder.com

15 of 26

## Hepatitis C prevalence

### Notation

- $p$  = probability an **INDIVIDUAL** is positive
- $\theta$  = probability a **GROUP** is positive
- $s$  = group size
- $n$  = number of groups
- $T$  be a random variable denoting the number of positive **GROUPS**
  - $T$  has a binomial probability distribution with “ $n$  trials” and “ $\theta$  as the probability of success”

$$P(T = t) = \binom{n}{t} \theta^t (1 - \theta)^{n-t} \text{ for } t=0, 1, 2, \dots, n$$

www.chrisbilder.com

16 of 26

## Hepatitis C prevalence

- How can we estimate  $p$ ?
  - We observe information about the groups, not individuals!
  - Estimate  $\theta$  with  $\hat{\theta} = T/n = \# \text{ positive} / \# \text{ of groups}$ 
    - Maximum likelihood estimate of  $\theta$
  - $\theta = P(\text{group is positive})$ 
    - $= P(\text{at least one individual is positive})$
    - $= 1 - P(\text{no individuals are positive})$  **using complement rule**
    - $= 1 - P(\text{all individuals are negative})$
    - $= 1 - (1 - p)^s$  **since  $p = P(\text{individual is positive})$  and  $s$  individuals per group**
  - $p = 1 - (1 - \theta)^{1/s}$
  - Then  $\hat{p} = 1 - (1 - \hat{\theta})^{1/s} = 1 - (1 - T/n)^{1/s}$ 
    - This is the maximum likelihood estimate of  $p$  by the invariance property

## Hepatitis C prevalence

- Estimation of Hepatitis C prevalence in Xuzhou City, China
  - Data from Liu et al. (*Transfusion*, 1997)
    - 1,875 blood donors screened for Hepatitis C
      - There were 42 positive individuals
  - In order to test the usefulness of group testing, blood samples were also pooled
    - $n = 375$  groups
    - $s = 5$  individuals per group
    - $t = 37$  positive groups
  - Estimates of  $p$ , probability individual is positive
    - Using individual data:  $42/1875 = 0.0224$
    - Using group data:  $\hat{p} = 1 - (1 - 37/375)^{1/5} = 0.0206$
  - Which is easier and more cost effective?
    - 1875 tests using individual testing
    - 375 tests using group testing

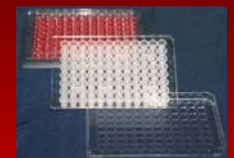
## Hepatitis C prevalence

- New research – MS/PhD research
  - What factors could affect  $p$ ?
  - Include independent variables to help model  $p$ 

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}$$
  - Problem: Do not have the individual outcomes
  - After a group is tested positive, how can you find what individuals have the disease?
    - Use model to help decide who to retest if get a positive group
  - Multiple diseases
    - Hepatitis C
    - HIV
    - Other disease
    - Simultaneously model

## Hepatitis C prevalence

- Other examples of where group testing is used
  - Multiple vector transfer design experiments
    - Estimate probability an insect vector transfers a pathogen to a plant
  - Drug discovery experiments
    - Screen hundreds of thousands of chemical compounds to look for potentially good ones
  - Veterinary
    - Bovine viral diarrhea virus (BVDV) infection in cattle



## Why statistics?

### Statistics is used in many diverse areas!

- Statistics is the “science of science”
- Florence Nightingale quote:

the most important science in the whole world: for upon it depends the practical application of every other science and of every art: the one science essential to all political and social administration, all education, all organization based on experience, for it only gives results of our experience.

### Where do statistics graduates work?

- Pharmaceutical – Pfizer, Merck
- Marketing – Target, Hallmark
- Government research labs – INEEL, Los Alamos, Sandia, Argonne
- Agriculture – Pioneer Hi-Bred
- Contract research organizations – Quintiles
- Every MS and PhD statistics student that I have known has had a job offer before they graduated!

## Why statistics?

### Salaries

- Non-academic starting (2003 American Statistical Association survey)

Degree	Sample size	Percentile		
		25 <sup>th</sup>	50 <sup>th</sup>	75 <sup>th</sup>
MS	102	45.5K	50K	59K
PhD	99	60K	65K	75K

- Survey response rate was 23.5% by organizations surveyed
- See salary surveys at the American Statistical Association’s website
- I hope you have an interest to take more statistics courses
  - University of Wisconsin-La Crosse
  - Graduate school in statistics or non-statistics programs
- Of course, I want you to consider coming to the University of Nebraska-Lincoln!

## Statistics at Nebraska

### Facts

- July 1, 2003 formed
- 11 faculty + hiring for 2 more in spring 2006
- No undergraduate major
- 40+ MS and PhD graduate students (mostly MS)
- Newly renovated building (March, 2005)

### Background of new graduate students

- A few statistics courses – like UWL MTH 441 and 442
- Majority have mathematics degrees

### Recommendations for courses

- MTH 441 and 442 (Mathematical Statistics I & II)
- Applied statistics – like MTH 445, Correlation & Regression Analysis
- **MTH 413** (Topics in Linear Algebra), MTH 371 (Intro. to Numerical Methods); MTH 408 (Real Analysis II) is helpful if go on for a PhD
- Some background in computer programming

## Statistics at Nebraska

### Assistantships

- Work 20 hours a week
- 25 students have an assistantship (others supported outside of dept.)
- Teaching – \$13K per school year + tuition (MS students)
- Research – variable depending on grants
  - Statistics and non-statistics faculty grants
- Average GPA for incoming teaching assistants in 2005: 3.69

### What makes Nebraska unique?

- Consulting course and help desk
- STAT 971 – Statistical Modeling
- STAT 832 – Statistics in Sports; work with Nebraska athletic department, NASCAR, Denver Nuggets, ...
- Survey Research and Methodology program and Gallup Organization
- Bioinformatics

## Statistics at Nebraska

- Applying for graduate school in statistics (starting fall 2006)
  - Send out applications before end of fall semester
  - Apply to more than one school
  - Visit schools in fall or early spring
    - Helps to show you are interested!
    - WE DO HAVE FUNDING AVAILABLE TO HELP PAY FOR YOUR VISIT!
  - Initial assistantship offers usually go out in early March
    - Sometimes make offers as well during the summer

## Statistics at Nebraska

- For more information...
  - E-mail me at [chris@chrisbilder.com](mailto:chris@chrisbilder.com)
    - Advice
    - Visit and sit in on a class
  - Website: [www.chrisbilder.com/statistics](http://www.chrisbilder.com/statistics)
    - This PowerPoint presentation
    - Links to
      - Introductory information about being a statistician
      - Jobs (including internships)
      - Salary information
      - List of all Departments of Statistics
      - Professional societies
      - MS and PhD course websites that myself and others teach
      - Newspaper and magazine articles about statistical applications