Real-World Statistics

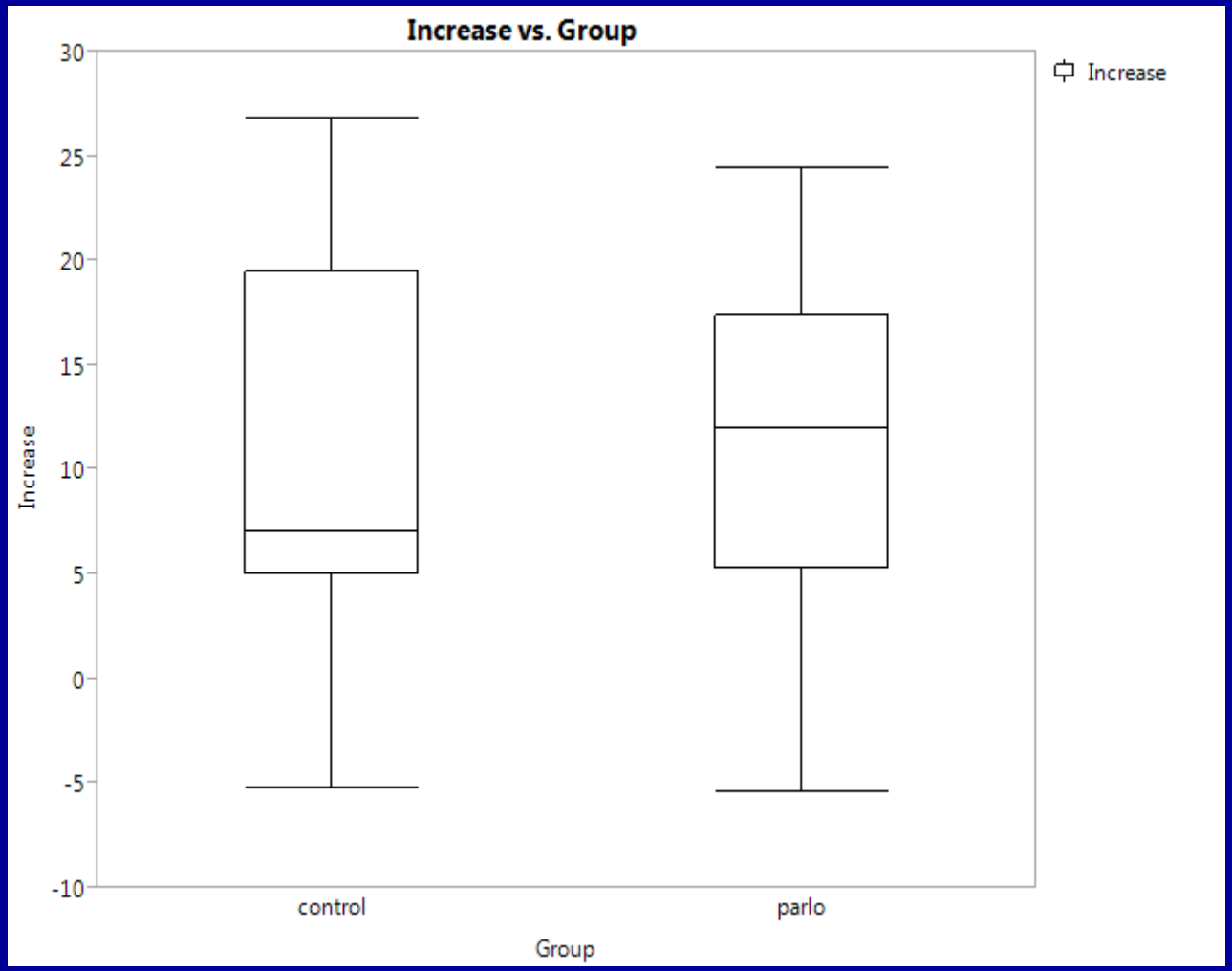
TANGO Stat Ed
USCOTS Conference
May 27, 2015
(www.tinyurl.com/TANGOStatEd)

Michael A. Posner, Ph.D., PStat®
Associate Professor of Statistics
Department of Mathematics and Statistics
Director, Center for Statistics Education
Villanova University

Classroom Data

- Does PARLO impact student learning?

preMean	postMean	group
24.9441	36.9211	parlo
23.6559	38.172	parlo
27.621	33.2661	control
34.1364	42.4663	control
30.1708	24.9209	control
22.4935	49.2589	control
26.4716	40.705	parlo
28.5758	33.5666	control
33.7996	39.4234	parlo
22.0223	41.8114	control
33.5914	38.8817	parlo
32.2581	37.6344	parlo
22.5806	17.1477	parlo
35.2941	49.7154	control
32.7189	38.2488	control
22.212	27.2811	control
10.3226	29.6774	control
20.9217	45.3456	parlo
27.7419	26.7742	parlo
30.4715	48.1886	parlo
36.1851	53.5764	parlo



Real Data - PARLO

- 19 page overview document
  - 18 Barcodes duplicated with different results
  - What happens if you eliminate missing?
- Analysis – Hierarchical Linear Model

Content
Attitudes
Learning Outcomes

Teaching Surveys

External PSSAs

Mar 1, 2010 Funding Begins	53 teachers 18 schools	3,168 Students 61 teachers 29 schools	2,311 Students 44 teachers 22 schools
2009-10	2010-11	2011-12	2012-13
Cohort 1: Recruitment and PD	**Cohort 1: Year 1** 53 teachers 18 schools	**Cohort 1: Year 2** 1,563 students 37 teachers 16 schools	**Cohort 1: Year 3** 1,122 students 23 teachers 11 schools
	Cohort 2: Recruitment and PD	**Cohort 2: Year 1** 1,602 students 24 teachers 13 schools	**Cohort 2: Year 2** 1,189 students 21 teachers 11 schools

Classroom Data vs. Real Data

- Missing data – why missing?
- Dealing with outliers
- Observational studies
- Unstructured problem
- Unstructured data (“Big Data”)
  - 80-90% of the time spent preparing the data
- A single analysis is rarely appropriate

The Bridge Between Classroom and Real Data

- Naked vs. Realistic vs. Real data
- Intrinsic motivation
  - Relevant topics
  - Invest students in the outcome
- Discovery (constructivist approach)
- Discuss design prior to sharing data
- Discuss biases after determining results
  - …in a way that isn't too pessimistic
- Complex problems

Lottery - Expected Value

- Consider playing a lottery where you pick three digits. If you match all three, you win $400. If you match only the first two, you win $10. The lottery costs $1 to play.

Amount you win	Prob
-$1	990/1000
$10	9/1000
$400	1/1000

- How much would you have to make the grand prize ($400) in order to make this lottery fair?

Making Pizza – Expected Value

The operations manager for a pizza store needs to decide how many pizzas to make. The overhead costs for running the business are $1000, no matter how many pizzas they sell (this covers employees, delivery, advertising, etc.). Consider only a single type of pizza – cheese. Cheese pizzas cost $4.50 to make and sell for $9 each. An unsold pizza is considered to have no value. The following table represents the demand for cheese pizza. Create a graph showing expected profit by production level (in units of 100 from 200 to 900). Based on this graph, what is the optimal production level of cheese pizza?

Demand	Probability
200	0.1
300	0.15
400	0.15
500	0.2
600	0.2
700	0.10
800	0.05
900	0.05

Making Pizza - Scaffolding

1. Calculate the profit or loss if they supply 200 pizzas and all 200 pizzas are sold.
2. Calculate the profit or loss if they produce 200 pizzas and the demand is 300, 400, …., 900. Put these results into a probability distribution table with probabilities listed as above.
3. Calculate the expected profit or loss for producing 200 pizzas. This is your first point on the final graph.
4. Repeat this process for producing 300 to 900 pizzas, in increments of 100.
5. Plot these points on the final graph and determine the optimal number of cheese pizzas to produce.

Some Useful Datasets

- JSE Data Archive
  - 4cdata.txt (the basic data file)
    4c1data.txt (includes indicator or "dummy" variables)
    4c.txt (the documentation file)
    NAME: Pricing the C's of Diamond Stones
    TYPE: Observational Regression Analysis Data
    SIZE: 308 observations, 5 variables The article appears in the *JSE*, Volume 9, Number 2 (July 2001).
    SUBMITTED BY: [name w/email address]
- STEW – including lesson plans (K-12)
- Tuva Labs, including email list
- Qandl
- Census at Schools

Some More Useful Datasets

- The Open Data Philly project
- ICPSR (Consortium for Political & Social Research)
- The General Social Survey - monitoring social change & the growing complexity of American society, 30+ years
- Publicly available Health-related datasets
  - National Center for Health Statistics
  - NHANES (National Health and Nutrition Evaluation Survey)
  - HHS (Dept of Health and Human Services)
- NAEP (National Assessment of Educational Progress)
- Free Data from the National Climatic Data Center
- Sports Data (google around - there are many sites)
- DASL - Data and Storage Library
- FedStats.gov - A link to numerous federal databases
- Search using .xls or .xlsx

Some Useful Studies

- Futurity.org
  - Research news from top universities
  - Sign up for daily email
- Various newspapers, blogs, etc.

A Few Videos

- Steve Wang – The Obvious Answer…
- Esther Duflo – Social Justice