Activities for the statistics classroom

Roger Johnson
roger.johnson@sdsmt.edu
South Dakota School of Mines & Technology

Activity Overview

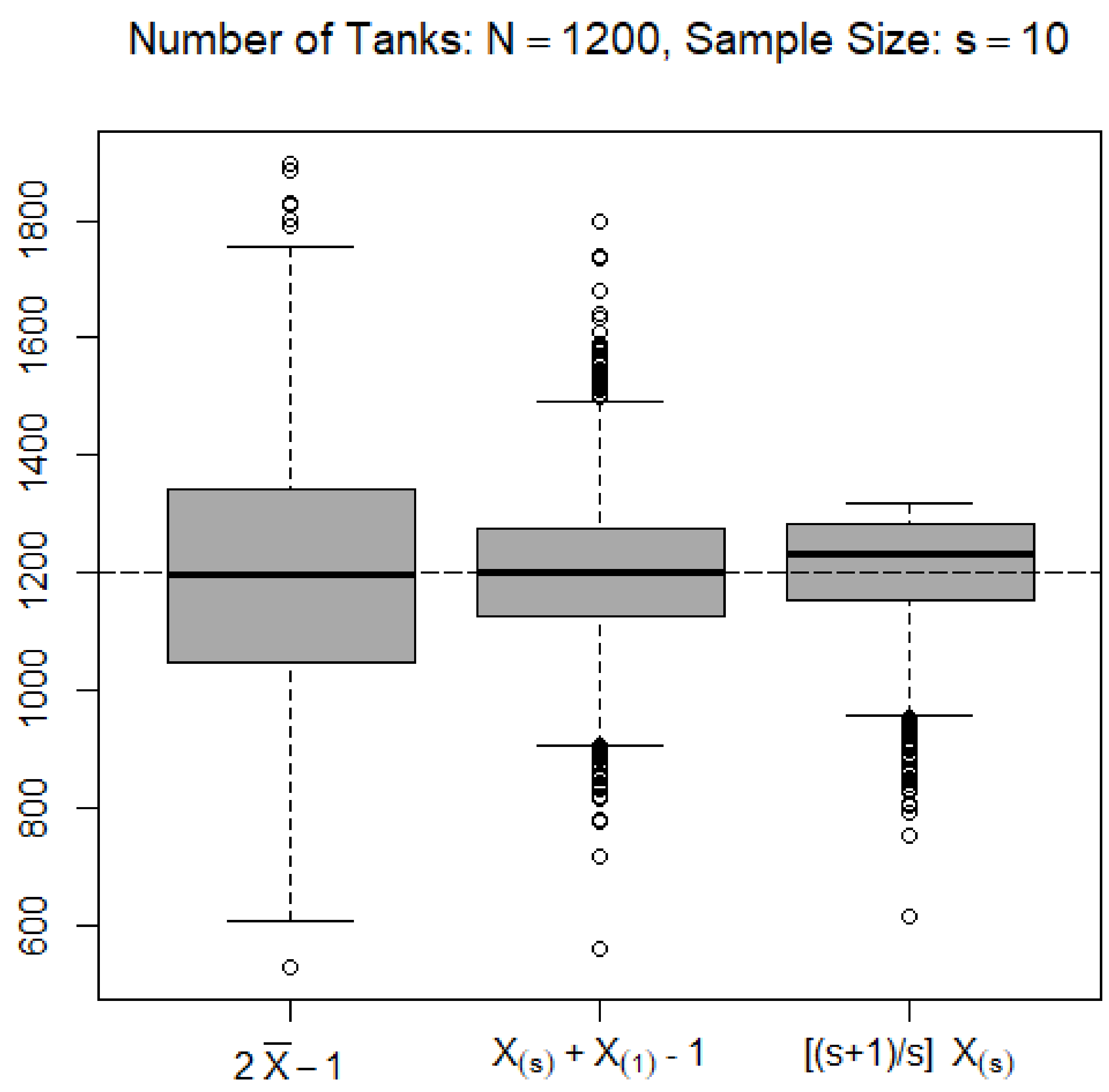
- Impressions:
  - Student learning easier in more active settings
  - Instructors going the ‘extra mile’: Generates student goodwill
- My general activity pattern:
  - Hands-on activity followed by simulation/technology
- This presentation (online at my homepage):
  - Adapt activities as needed for your audience
  - Sometimes short on details
  - *Teaching Statistics* articles (further details/code comments)

Estimation: How many tanks?

- Population: Tanks labeled 1, 2, . . ., ***N***
- Sample: ***s*** captured tanks with labels $X_1, X_2, \ldots, X_s$
- Activity:
  - Use numbered popsicle sticks in place of the tanks, have a few students (10?) sample one stick each
  - Have students, working in groups of 2 or 3, come up with estimates of *N* (instructors can give ‘clues’)
  - Use simulation to demonstrate estimate behavior

Estimation: How many tanks?

- Sample clues for estimates during group work:
  - Sample Average $\approx$ Population Average
  - Use symmetry between $X_{(1)}$ with 1, and $X_{(s)}$ with N
  - Generalize the above clue
- Resulting estimates (in order with clues above):
  - $\bar{X} \approx (1+2+\cdots+N)/N = (N+1)/2$
  - $X_{(1)} - 1 = N - X_{(s)}$
  - $\hat{N} = \frac{(s+1)}{s} X_{(s)}$ (scaled version of the mle)



Number of Tanks: N = 1200, Sample Size: s = 10
1800
1600
1400
1200
1000
800
600
$2\bar{X} - 1$
$X_{(s)} + X_{(1)} - 1$
$[(s+1)/s]\ X_{(s)}$

Estimation: How many fish in the pond?

- Activity illustrates the 'Capture-Recapture' method:
  - Capture and tag some fish (***t*** fish)
  - Return to population (size ***N***) and allow time to mix
  - (Recapture) Sample from tagged/untagged population (***s*** fish) and record the number tagged, say ***r***
- Big Idea:
  - Fraction tagged in population (***t/N***), should be roughly the fraction tagged in the sample of size *s* (***r/s***)

Estimation: How many fish in the pond?

- Can illustrate process with Pepperidge Farm's Goldfish
  - Start with bowl (lake) of one variety
  - Student captured fish replaced by instructor with second variety
- Some details:
  - Compare class estimate with actual population size (don't forget to determine this before class)
  - Easiest to have just a few students involved in the "Capture" and "Recapture" (small handfuls)
  - Thirty roughly 1 ounce bags for less than $10

Estimation/Hypothesis Testing: Earth's surface water percentage?

- Bring a inflatable globe of the Earth to class – Ask how we could estimate the surface water percentage (assuming any needed 'tools')
- Randomly toss globe in class (students call out 'land' or 'water' according to where their right index finger points)
- Can use resulting data to:
  - Give point- and interval-estimates of unknown percentage
  - Test (two-sided alternative) hypothesis of percentage as given at a Website (e.g. 71% according to one site)

Hypothesis Testing: Instructor ESP?

- Instructor draws a sequence of playing cards, face down:
  - Calls out ‘red’ or ‘black’, using his/her psychic ability
  - Then reveals card to classroom – students record ‘correct’ or ‘incorrect’
- I use marked cards attempting to:
  - Suspend belief of ‘cheating’ (I deliberately make some wrong ‘guesses,’ including the first card drawn)
  - Reject the hypothesis of guessing, $p = ½$, in favor of ESP,

    $p > ½$
- Another proportion test: Coke or Pepsi? (Online slides)

Hypothesis Testing (Paired Data): Reduced fat Oreos vs. regular Oreos?

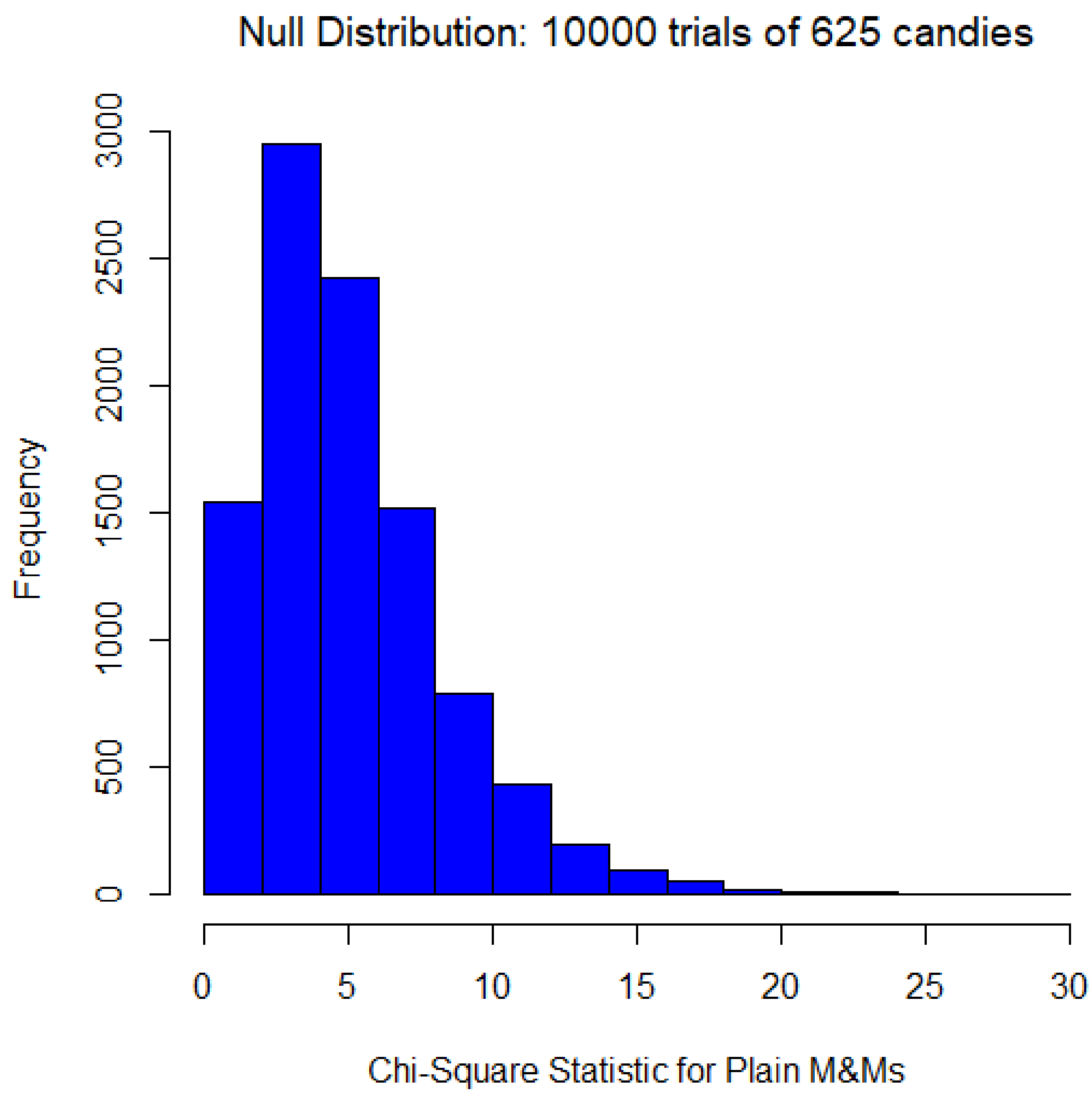
- Population? Random Sample?
- Paired-sample t test and/or nonparametric alternatives
- Discussion of experimental protocol:
  - Have each student try one of each of Cookie A, Cookie B
  - Randomize order of tasting, palate cleansing between
  - Rate each cookie taste on a 1-5 Likert Scale (1='yuck', 5='yum')
- Instructor computes score differences and leads students through a one-sided paired-sample t test or nonparametric alternatives

Hypothesis Testing (Goodness of Fit): Color proportions M&Ms, Skittles, . . .

- Test either an equal color distribution or a claimed color distribution by passing out small ('fun size') bags of candy to students
- Collect data and conduct hypothesis test using Pearson's chi-squared goodness of fit statistic
  - Use technology to compare statistic for actual data with simulated values of the statistic under the null hypothesis
  - Important detail: Random samples are assumed (means that we assume candies are mixed well)

Hypothesis Testing:
Color proportions M&Ms (2017)

Color	Hackettstown, NJ (Code: HKP)	Cleveland, TN (Code: CLV)
Blue	25.0%	20.7%
Orange	25.0%	20.5%
Green	12.5%	19.8%
Yellow	12.5%	13.5%
Red	12.5%	13.1%
Brown	12.5%	12.4%


Null Distribution: 10000 trials of 625 candies
Frequency
0
500
1000
1500
2000
2500
3000
0
5
10
15
20
25
30
Chi-Square Statistic for Plain M&Ms

Data Collected Outside the Classroom

Least Squares:
Fitting calories to food components

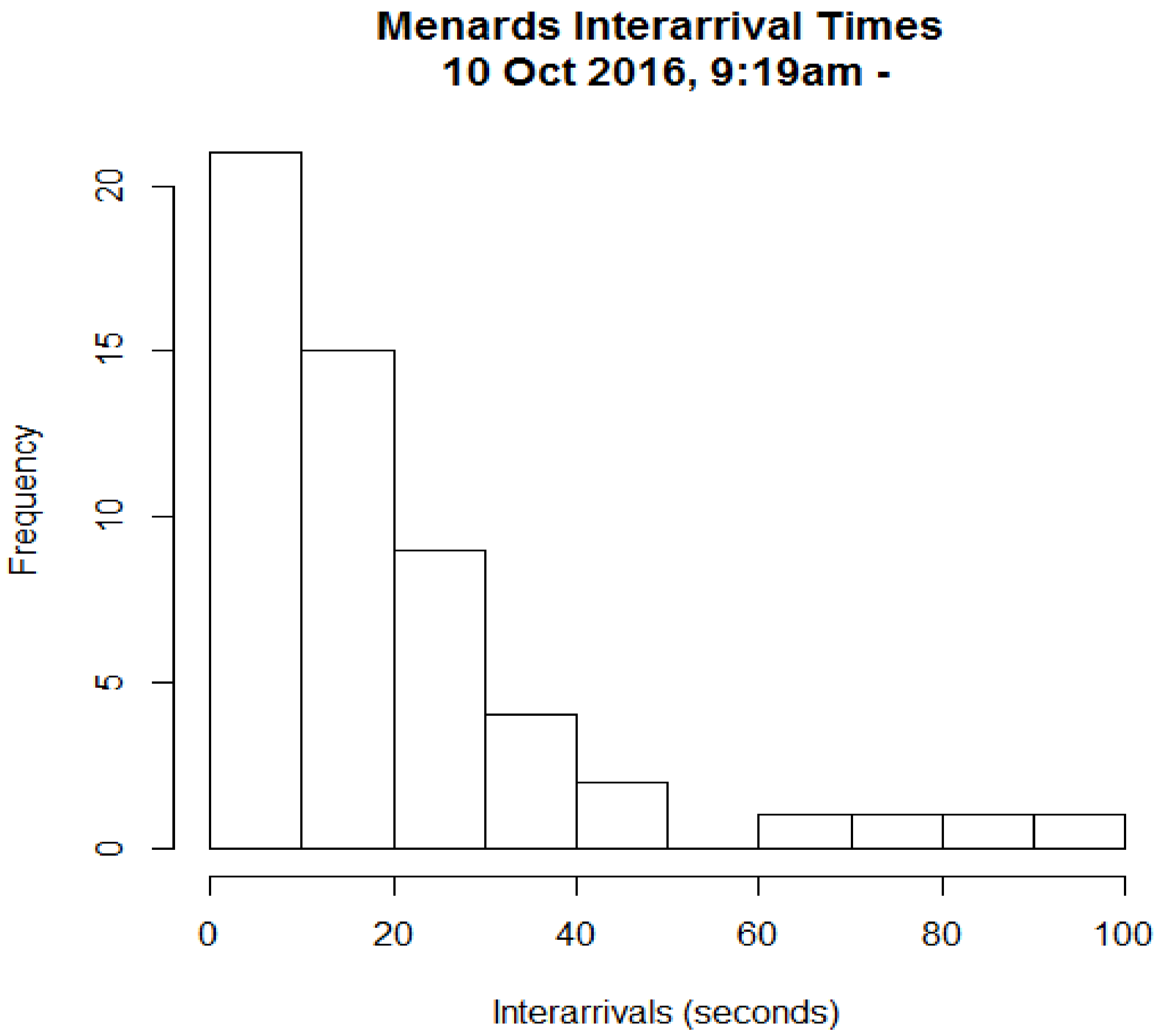
- Have students collect the following on a number of food items (store and/or cupboards):
  - Response: Calories
  - Predictors: Fat, Protein, Carbohydrates – all in grams
- Based on plots of response versus predictors, fit Calories in terms of linear combination of predictors
  - Argue additive constant term is zero (think of water)

Least Squares: Fitting calories to food components

- True model:
  - $\text{Calories} = 4(\text{Carbs}) + 4(\text{Protein}) + 9(\text{Fat})$
- Note:
  - Values of the four values often rounded on food labels (e.g. calories to nearest 5 or 10 calories)

Data Patterns: Interarrival data

- Interarrival: Gap in time between two customers arriving at a service facility (e.g. store/business entrance, fast food drive-up)
- Have students, working alone or in pairs:
  - (**Safely!**) Collect arrival data, then take differences to get interarrivals (disabled students – data from a Webcam)
  - Use technology to compute numerical and graphical summaries
- For "random" arrivals, interarrivals often follow an exponential (population mean = population std deviation)



Menards Interarrival Times
10 Oct 2016, 9:19am -
Frequency
0
5
10
15
20
0
20
40
60
80
100
Interarrivals (seconds)

References

- Gelman, A. and Nolan, D. (2017), *Teaching Statistics: A Bag of Tricks*, 2nd edition, Oxford University Press, New York.
- Johnson, R. (2017), Discovering patterns in interarrival data, *Teaching Statistics*, 39(2), 42-46.
- Johnson, R. (1997), Earth's surface water percentage?, *Teaching Statistics*, 19(3), 66-68.
- Johnson, R. (1996), How many fish are in the pond?, *Teaching Statistics*, 18(1), 2-5.
- Johnson, R. (1995), A multiple regression project, *Teaching Statistics*, 17(2), 64-66. [Fitting calories to food components.]

References

- Johnson, R. (1994), Estimating the size of a population, *Teaching Statistics*, 16(2), 50-52. [Tank problem.]
- Johnson, R. (1993), Testing colour proportions of M&Ms, *Teaching Statistics*, 15(1), 2-4.
- Levine, M. and Rolwing, R. (1993), Coke or Pepsi?, *Teaching Statistics*, 15(1), 4-5.
- Rau, R., Wise, C., and Sommers, P. (1990), Hypothesis Tasting, *Journal of Recreational Mathematics*, 22(2), 98-100. [M&Ms.]

References

- Scheaffer, R., Gnanadesikan, M, Watkins, A, and Witmer, J. (1996), *Activity-Based Statistics: Instructor Resources*, Springer-Verlag, New York.
- Scheaffer, R., Gnanadesikan, M, Watkins, A, and Witmer, J. (1996), *Activity-Based Statistics: Student Guide*, Springer-Verlag, New York.

roger.johnson@sdsmt.edu

This presentation & some R code at:
https://www.mcs.sdsmt.edu/rwjohnso

Teaching Statistics note

Hypothesis Testing: Pepsi or Coke?

- Have a student volunteer – one who claims she can easily distinguish between Pepsi and Coke
  - Lead student discussion on experimental protocol (e.g. palate cleansing between tastes, volunteer outside of classroom as glasses set up, . . .)
  - Decide on specific hypotheses (e.g. $p = 0.70$, say, versus $p > 0.70$)
  - Have student taste several (10?) unknown samples and have her declare her belief for each
- Note that binomial calculations needed to conduct the test

Data Patterns: Interarrival data

- Topics for more advanced classes:
  - (Exponential) Probability plots
  - Fitting an exponential to data (e.g. method of moments)
  - Random number generation (e.g. inverse cdf method)