Unit 5: Inference for categorical data

4. MT2 Review

Sta 101 - Fall 2015

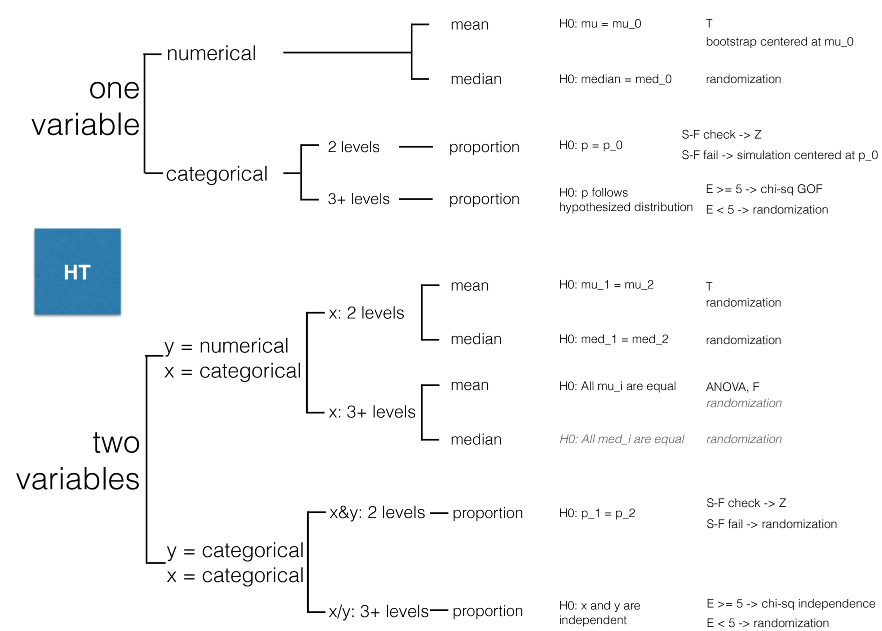
Duke University, Department of Statistical Science

Dr. Çetinkaya-Rundel

Slides posted at *http://bit.ly/sta101_f15*

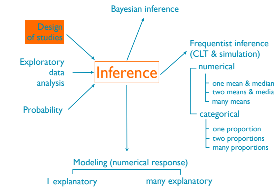
- MT 2 next week
  - Bring a calculator + cheat sheet + writing utensil
  - Tables will be provided
- MT 2 review session: Sat, Nov 7, 4-5pm, Old Chem 116
  - office hours as usual: *https://stat.duke.edu/courses/Fall15/sta101.002/info/#oh*
  - extra office hours from Dr. Monod: Friday, 1:30-3pm (Old Chem 122A)
- MT 2 review materials posted on the course website
- Project 1 due Friday evening (+ work on it in lab on Thursday)
- PS 5 due Friday evening, PA 5 due Saturday evening (note day change to allow for review before midterm)

inference	HT	CI
theoretical	Z, T, F, chi-sq	Z, T
simulation	bootstrap centered at null randomization	bootstrap



CI

one variable	numerical		mean	mu	T, bootstrap
			median	median	bootstrap
	categorical	2 levels	proportion	p	S-F check -> Z, S-F fail -> bootstrap
		3+ levels	proportion	N/A	
two variables	y = numerical, x = categorical	x: 2 levels	mean	mu_1 - mu_2	T, bootstrap
			median	med_1 - med_2	bootstrap
		x: 3+ levels	mean	N/A	
			median	N/A	
	y = categorical, x = categorical	x&y: 2 levels	proportion	p_1 - p_2	S-F check -> Z, S-F fail -> bootstrap
		x/y: 3+ levels	proportion	N/A	



Clicker question

Which of the following is true?

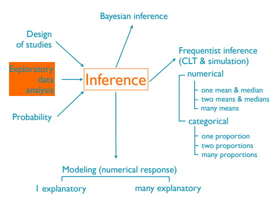
(a) If the sample size is large enough, conclusions can be generalized to the population.

(b) If subjects are randomly assigned to treatments, conclusions can be generalized to the population.

(c) *Blocking in experiments serves a similar purpose as stratifying in observational studies.*

(d) Representative samples allow us to make causal conclusions.

(e) Statistical inference requires normal distribution of the response variable.



Clicker question

Which of the following is the best visualization for evaluating the relationship between two categorical variables?

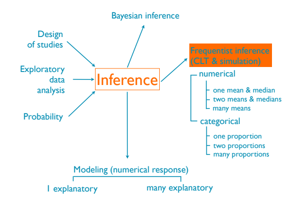
(a) side-by-side box plots

(b) *mosaic plot*

(c) pie chart

(d) segmented frequency bar plot

(e) relative frequency histogram



Clicker question

Two students in an introductory statistics class choose to conduct similar studies estimating the proportion of smokers at their school. Student A collects data from 100 students, and student B collects data from 50 students. How will the standard errors used by the two students compare? Assume both are simple random samples.

(a) *SE used by Student A $<$ SE used as Student B.*

(b) SE used by Student A $>$ SE used as Student B.

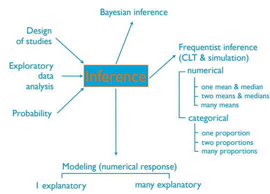
(c) SE used by Student A $=$ SE used as Student B.

(d) SE used by Student A $\approx$ SE used as Student B.

(e) Cannot tell without knowing the true proportion of smokers at this school.

Clicker question

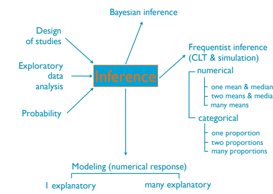
Which of the following is the best method for evaluating the relationship between two categorical variables?



(a) *chi-square test of independence*
(b) chi-square test of goodness of fit
(c) anova
(d) t-test

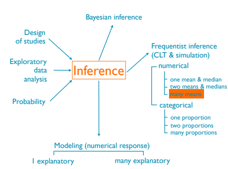
Clicker question

Which of the following is the best method for evaluating the relationship between a numerical and a categorical variable with many levels?

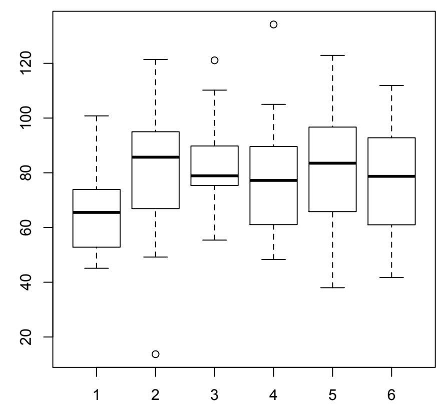


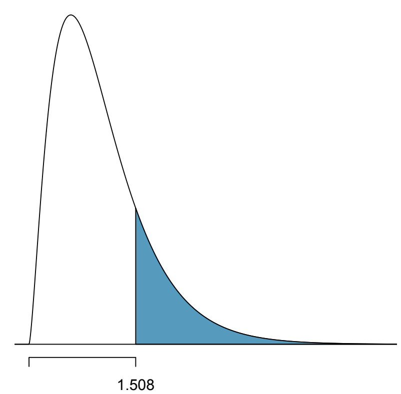
(a) z-test
(b) chi-square test of goodness of fit
(c) *anova*
(d) t-test

Data are collected at a bank on 6 tellers' randomly sampled transactions. Do average transaction times vary by teller?



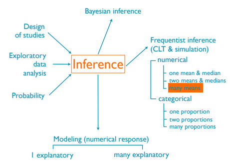

```
Response variable: numerical, Explanatory variable: categorical
ANOVA

Summary statistics:
n_1 = 14, mean_1 = 65.7857, sd_1 = 15.2249
n_2 = 23, mean_2 = 79.9174, sd_2 = 23.284
n_3 = 15, mean_3 = 82.66, sd_3 = 18.1842
n_4 = 15, mean_4 = 77.9933, sd_4 = 23.2754
n_5 = 44, mean_5 = 81.7295, sd_5 = 21.5768
n_6 = 29, mean_6 = 75.3069, sd_6 = 20.4814

H_0: All means are equal.
H_A: At least one mean is different.
Analysis of Variance Table

Response: data
           Df Sum Sq Mean Sq F value Pr(>F)
group       5   3315  663.06   1.508 0.1914
Residuals 134  58919  439.69
```



Activity:

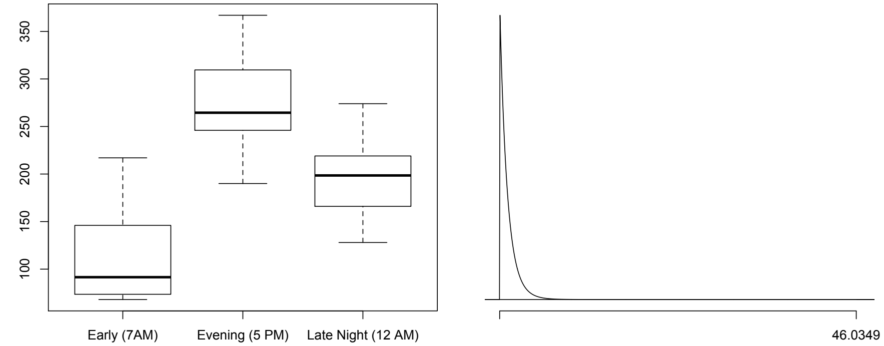
Data are collected on download times at three different times during the day. We want to evaluate whether average download times vary by time of day. Fill in the ??s in the ANOVA output below.




```
Response variable: numerical, Explanatory variable: categorical
Summary statistics:
n_Early (7AM) = 16, mean_Early (7AM) = 113.375, sd_Early (7AM) = 47.6541
n_Eve (5 PM) = 16, mean_Eve (5 PM) = 273.3125, sd_Eve (5 PM) = 52.1929
n_Late (12 AM) = 16, mean_Late (12 AM) = 193.0625, sd_Late (12 AM) = 40.9023
```

```
Analysis of Variance Table
Response: data
          Df Sum Sq Mean Sq F value    Pr(>F)
group     ??     ??      ??      ?? 1.306e-11
Residuals ?? 100020      ??
Total     ?? 304661
```


What is the result of the ANOVA?



Since 1.306e-11 < 0.05, we reject the null hypothesis. The data provide convincing evidence that the average download time is different for at least one pair of times of day.

Activity:

The next step is to evaluate the pairwise tests. There are 3 pairs of times of day

1. Early vs. Evening: left side of class (facing the board)
2. Evening vs. Late Night: center of class
3. Early vs. Late Night: right side of class

Determine the appropriate significance level for these tests, and then complete the test assigned to your team.

$$\alpha^\star = 0.05/3 = 0.0167$$

(1) Early vs. Evening

$$T_{45} = \frac{113.375 - 273.3125}{\sqrt{\frac{2223}{16} + \frac{2223}{16}}} = \frac{-159.9375}{16.67} = -9.59$$

$p - val < 0.01$

(2) Evening vs. Late Night

$$T_{45} = \frac{113.375 - 193.0625}{\sqrt{\frac{2223}{16} + \frac{2223}{16}}} = \frac{-79.6875}{16.67} = -4.78$$

$p - val < 0.01$

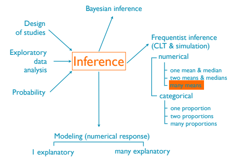
(3) Early vs. Late Night

$$T_{45} = \frac{273.3125 - 193.0625}{\sqrt{\frac{2223}{16} + \frac{2223}{16}}} = \frac{80.25}{16.67} = 4.81$$

$p - val < 0.01$

Clicker question

What percent of variability in download times is explained by time of day?




```
Response: data
          Df Sum Sq Mean Sq F value    Pr(>F)
group      2 204641  102320  46.035 1.306e-11
Residuals 45 100020    2223
```

(a) $\frac{204641}{204641+100020}$ *= 0.67*

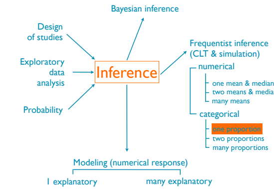
(b) $\frac{204641}{100020}$

(c) $\frac{100020}{204641}$

(d) $\frac{102320}{102320+2223}$

Clicker question

$n = 50$ and $\hat{p} = 0.80$. Hypotheses: $H_0 : p = 0.82; H_A : p \neq 0.82$. We use a randomization test because the sample size isn't large enough for $\hat{p}$ to be distributed nearly normally $(50 \times 0.82 = 41 < 10; 50 \times 0.18 = 9 < 10)$. Which of the following is the correct set up for this hypothesis test? Red: success, blue: failure, $\hat{p}_{sim}$ = proportion of reds in simulated samples.



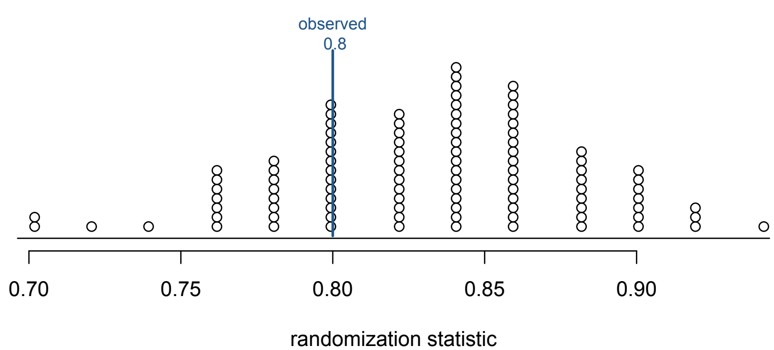
(a) Place 80 red and 20 blue chips in a bag. Sample, with replacement, 50 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \neq 0.82$.

(b) Place 82 red and 18 blue chips in a bag. Sample, without replacement, 50 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \neq 0.80$.

(c) *Place 82 red and 18 blue chips in a bag. Sample, with replacement, 50 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.80$ or $\hat{p}_{sim} \geq 0.84$.*

(d) Place 82 red and 18 blue chips in a bag. Sample, with replacement, 100 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.80$ or $\hat{p}_{sim} \geq 0.84$.

What is / should be the center of the randomization distribution?
What is the result of the hypothesis test?



Inference for numerical data:

- One numerical:
  - Parameter of interest: μ
  - *T*
  - HT and CI
- One numerical vs. one categorical (with 2 levels):
  - Parameter of interest: $\mu_1 - \mu_2$
  - *T*
  - HT and CI
  - If samples are dependent (paired), first find differences between paired observations
- One numerical vs. one categorical (with $3+$ levels) - mean:
  - Parameter of interest: N/A
  - ANOVA
  - HT only
- For all other parameters of interest: simulation

Binary outcome:

- One categorical:
  - Parameter of interest: p
  - S/F condition met $\rightarrow$ Z, if not simulation
  - HT and CI
- One categorical vs. one categorical, each with only 2 outcomes:
  - Parameter of interest: $p_1 - p_2$
  - S/F condition met $\rightarrow$ Z, if not simulation
  - HT and CI
- S/F: use observed S and F for CIs and expexted for HT

$3+$ outcomes:

- One categorical, compared to hypothetical distribution:
  - Parameter of interest: N/A
  - At least 5 expected successes in each cell $\rightarrow \chi^2$ GOF, if not simulation
  - HT only
- One categorical vs. one categorical, either with $3+$ outcomes:
  - Parameter of interest: N/A
  - At least 5 expected successes in each cell $\rightarrow \chi^2$ Independence, if not simulation
  - HT only