Unit 8: Final Review
1. Final Exam Review

STA 104 - Summer 2017

Duke University, Department of Statistical Science

Prof. van den Boom

Slides posted at
http://www2.stat.duke.edu/courses/Summer17/sta104.001-1/

Activity:

Course evaluations

Project

Some comments on the project:

- Try to tell a story around your one research question
- State hypotheses
- Check conditions in the slides, that includes model diagnostics for the prediction model
- Give rationale for final choice for prediction model, e.g. backwards elimination. Only need to include 1/final model.

Exam setup

Final exam:

- Out of 97 points total
- 37 points “open” questions
- Remaining 60 points multiple choice

Overview of tests: When to use them, and conditions

Reminder: Conditions often relate to the Central Limit Theorem. It states that the sample mean has (nearly) normal sampling distribution when conditions are met.

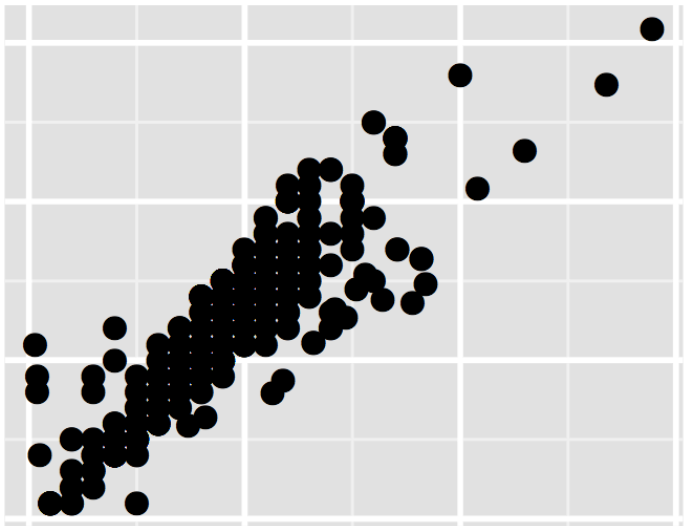
- Z-test for sample means
  - Numerical variable, possibly with categorical variable of 2 levels
  - Independence, sample size/skew
- Z-test for proportions
  - Up to two categorical variable of 2 levels
  - Independence, success-failure
- T-test for sample means
  - Same as Z test for sample means except for the sample size/skew condition
  - Data need to be nearly normal or sample size large enough to correct for skew.

Overview of tests: When to use them, and conditions

- ANOVA
  - Numerical variable with at least 1 categorical variable with more than 2 levels
  - Independence, constant variance, approximate normality
- Chi-square
  - 1 categorical variable with more than 2 levels: Test of goodness of fit
  - 2 categorical variables, with at least one with more than 2 levels: Test of independence
  - Independence, expected counts greater than 5
- Linear regression
  - 1 numerical variable plus other explanatory variables
  - Linearity, constant variance, independence, normality of residuals
- Simulation tests
  - Randomization and bootstrapping
  - No conditions

Leverage and influential points

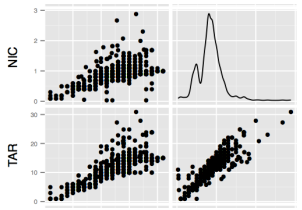
- Which points have the most leverage in the plot below?
- Are they influential?



Confounding variable

What would the slope be for nicotine be if tar was not included in the model?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.5489	0.5395	-1.02	0.3092
NIC	-4.0406	0.4342	-9.31	0.0000
TAR	1.0485	0.0441	23.80	0.0000
LEN	0.0350	0.0055	6.38	0.0000
FLTRNF	-6.4925	0.3577	-18.15	0.0000
PACKSOFT	0.5128	0.1046	4.90	0.0000
STRENGTHLIGHT	1.6804	0.2110	7.96	0.0000
STRENGTHMEDIUM	0.7339	0.4607	1.59	0.1114
STRENGTHREGULAR	0.2801	0.3059	0.92	0.3600
STRENGTHFULL FLAVOR	2.2447	0.3287	6.83	0.0000



Compute a 95% confidence interval for the slope of nicotine.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.5489	0.5395	-1.02	0.3092
NIC	-4.0406	0.4342	-9.31	0.0000
TAR	1.0485	0.0441	23.80	0.0000
LEN	0.0350	0.0055	6.38	0.0000
FLTRNF	-6.4925	0.3577	-18.15	0.0000
PACKSOFT	0.5128	0.1046	4.90	0.0000
STRENGTHLIGHT	1.6804	0.2110	7.96	0.0000
STRENGTHMEDIUM	0.7339	0.4607	1.59	0.1114
STRENGTHREGULAR	0.2801	0.3059	0.92	0.3600
STRENGTHFULL FLAVOR	2.2447	0.3287	6.83	0.0000

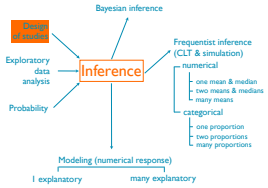
Number of slopes $k = 9$, and sample size $n = 1225$

The average height of Americans is 1.71 meter with a standard deviation of 0.07 meter.

- What is the probability that a randomly selected American is taller than 1.8 meter?
- What is the probability that a random sample of 100 Americans has an average height below 1.7 meter?
- We collect 4 random samples of size 100. What is the probability that at least one results in an average height below 1.7 meter?

Clicker question

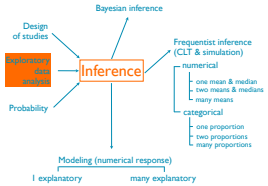
A recent research study randomly divided participants into groups who were told that they were given different levels of Vitamin E to take daily. Actually, one group received only a placebo pill, and the other received Vitamin E. The research study followed the participants for eight years to see how many developed a particular type of cancer during that time period. Which of the following responses gives the best explanation as to the purpose of the random assignment in this study?



(a) To prevent skewness in the results.
(b) To reduce the amount of sampling variability.
(c) To ensure that all potential cancer patients had an equal chance of being selected for the study.
(d) To produce treatment groups with similar characteristics.
(e) To ensure that the sample is representative of all cancer patients.

Clicker question

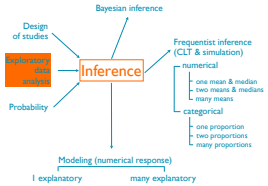
Which of the following is the most appropriate visualization for evaluating the relationship between a numerical and a categorical variable?



(a) a mosaic plot
(b) a segmented frequency bar plot
(c) a frequency histogram
(d) a relative frequency histogram
(e) side-by-side box plots

Clicker question

Which of the following is false?



(a) Box plots are useful for highlighting outliers, but we cannot determine skew based on a box plot.

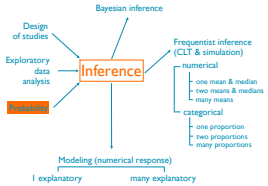
(b) Median and IQR are more robust statistics than mean and SD, respectively, since they are not affected by outliers or extreme skewness.

(c) When the response variable is extremely right skewed, it may be useful to apply a log transformation to obtain a more symmetric distribution, and model the logged data.

(d) Segmented frequency bar plots are "good enough" for evaluating the relationship between two categorical variables if the sample sizes are the same for various levels of the explanatory variable.

Clicker question

Which of the following is false?



(a) If A and B are independent, then having information on A does not tell us anything about B.

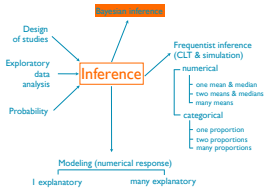
(b) If A and B are disjoint, then knowing that A occurs tells us that B cannot occur.

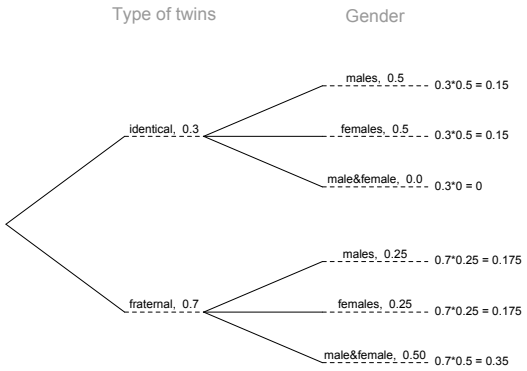
(c) Disjoint (mutually exclusive) events are always dependent since if one event occurs we know the other one cannot.

(d) If A and B are independent, then P(A and B) = P(A) + P(B).

(e) If A and B are not disjoint, then P(A or B) = P(A) + P(B) - P(A and B).

About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the posterior probability that they are identical?

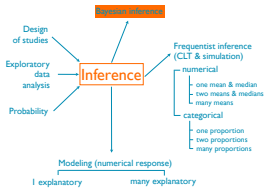




$$
\begin{aligned}
P(\text{iden} \mid f) &= \frac{P(\text{iden } \& \ f)}{P(f)} \\
&= \frac{0.15}{0.15 + 0.175} \\
&= 0.46
\end{aligned}
$$

Clicker question

Which of the following is false?



(a) Suppose you're evaluating 4 claims. If prior to data collection you don't have a preference for one claim over another, you should assign 0.25 as the prior probability to each claim.

(b) Posterior probability and the p-value are the equivalent.

(c) One advantage of Bayesian inference is that data can be integrated to the inferential scheme as they are collected.

(d) Suppose a patient tests positive for a disease that 2% of the population are known to have. A doctor wants to confirm the test result by retesting the patient. In the second test the prior probability for "having the disease" should be more than 2%.

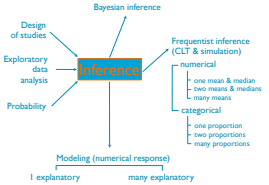
Activity:

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	8:30am section	10:05am section	11:45am section
$p - value$			
$n = 5000$	1:25pm section	3:05pm section	4:40pm section
$p - value$			

Clicker question

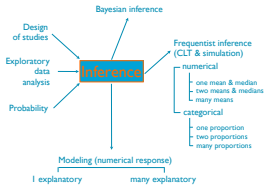
Which of the following is the best method for evaluating the if the distribution of a categorical variable follows a hypothesized distribution?



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

Example - Breast Cancer & Age

It is theorized that an important risk factor for breast cancer is age at first birth. An international study was set up to test this hypothesis. Breast-cancer cases were identified among women in selected hospitals in the United States, Greece, Yugoslavia, Brazil, Taiwan, and Japan. Controls were chosen from women of comparable age who were in the hospital at the same time as the cases but who did not have breast cancer. All women were asked about their age at first birth.

The set of women with at least one birth was arbitrarily divided into two categories: (1) women whose age at first birth was less than or equal to 29 years and (2) women whose age at first birth was greater than of equal to 30 years. The following results were found among women with at least one birth: 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than or equal to 30. How can we assess whether this difference is significant or simply due to chance?

Breast Cancer & Age - set-up

We are comparing two categorical variables (breast cancer status vs. age at first birth), this can be summarized by a contingency table.

We are given 683 of 3220 women with breast cancer (case women) and 1498 of 10,245 women without breast cancer (control women) had an age at first birth greater than 30.

	Breast Cancer (case)	No Breast Cancer (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465

Breast Cancer & Age - set-up

$$n_{case} = 3220,\ n_{ctrl} = 10245$$

- cases: 13465 women (hospital patients) with at least one child
- variable(s): (1) breast cancer status - categorical, (2) age at first birth - categorical
- parameter of interest: $p_{case} - p_{ctrl}$
  - Note: $p_{case} = P(age \geq 30|case)$ and $p_{ctrl} = P(age \geq 30|ctrl)$
- test: compare two population proportion of independent groups
- hypotheses: (two-tailed)

  $H_0 : p_{case} = p_{ctrl}$
  $H_A : p_{case} \neq p_{ctrl}$

Breast Cancer & Age - point estimate

Clicker question

Which of the following is the correct point estimate for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465

(a) $\frac{683}{2181} - \frac{1498}{2181}$
(b) $\frac{683}{13465} - \frac{1498}{13465}$
(c) $\frac{2537}{11284} - \frac{683}{2181}$
(d) $\frac{683}{3220} - \frac{1498}{10245}$
(e) $\frac{683}{2181} - \frac{683}{3220}$

Breast Cancer & Age - standard error

Clicker question

Which of the following is the correct standard error for this HT?

	BC (Case)	No BC (Controls)	Total
≤ 29	2537	8747	11284
≥ 30	683	1498	2181
Total	3220	10245	13465
$\hat{p}$	0.212	0.146	0.162

(a) $\sqrt{\frac{0.212\times(1-0.212)}{3220}} + \sqrt{\frac{0.146\times(1-0.146)}{10245}}$
(b) $\sqrt{\frac{0.212\times(1-0.212)}{3220} + \frac{0.146\times(1-0.146)}{10245}}$
(c) $\sqrt{\frac{0.162\times(1-0.162)}{3220} + \frac{0.162\times(1-0.162)}{10245}}$
(d) $\sqrt{\frac{0.212\times(1-0.212)}{13465} + \frac{0.146\times(1-0.146)}{13465}}$
(e) $\sqrt{\frac{0.162\times(1-0.162)}{13465} + \frac{0.162\times(1-0.162)}{13465}}$

$$Z = \frac{\hat{p}_{case} - \hat{p}_{ctrl} - 0}{SE} = \frac{0.212 - 0.146}{0.0074} = 8.92$$

$$\text{p-value} = P(Z > 8.92) + P(Z < -8.92) \approx 0$$

- Confidence level: 98%
- Theoretical: Using a critical value based on the Z distr. ($z^\star$):

$$\begin{aligned} &\text{point estimate} \pm ME \\ &= \text{point estimate} \pm z^\star \times SE \end{aligned}$$

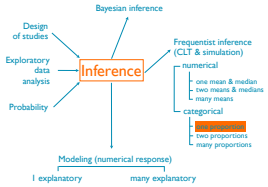
For a confidence interval,

$$\begin{aligned} SE &= \sqrt{\frac{\hat{p}_{case}(1-\hat{p}_{case})}{n_{case}} + \frac{\hat{p}_{ctrl}(1-\hat{p}_{ctrl})}{n_{ctrl}}} \\ &= \sqrt{\frac{0.212(1-0.212)}{3220} + \frac{0.146(1-0.146)}{10245}} = 0.008 \end{aligned}$$

$$\begin{aligned} (0.212 - 0.146) \pm 2.33 \times 0.008 &\approx 0.066 \pm 0.0186 \\ &= (0.0474,\ 0.0846) \end{aligned}$$

Clicker question

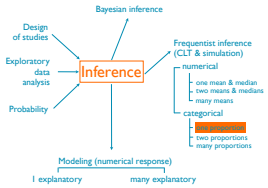
$n = 30$ and $\hat{p} = 0.6$. Hypotheses: $H_0 : p = 0.8; H_A : p < 0.8$. Which of the following is an appropriate method for calculating the p-value for this test?



(a) CLT-based inference using the normal distribution
(b) simulation-based inference
(c) exact calculation using the binomial distribution

Clicker question

$n = 30$ and $\hat{p} = 0.6$. Hypotheses: $H_0 : p = 0.8; H_A : p < 0.8$. Suppose we wanted to use simulation-based methods. 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 60 red and 40 blue chips in a bag. Sample, with replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.8$.
(b) Place 80 red and 20 blue chips in a bag. Sample, without replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.
(c) Place 80 red and 20 blue chips in a bag. Sample, with replacement, 30 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.6$.
(d) Place 80 red and 20 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 $\hat{p}_{sim} \leq 0.6$.