Sta 111 - Summer II 2017
Probability and Statistical Inference

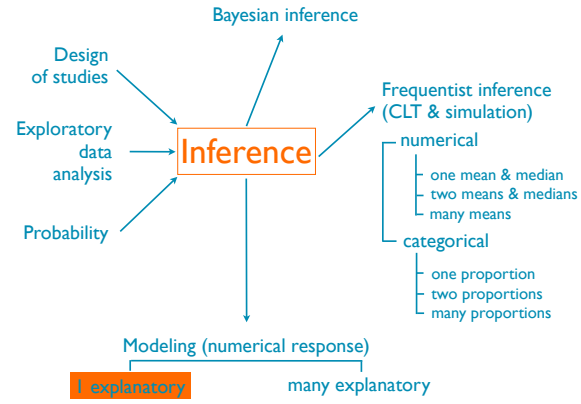
25. Final Exam Review

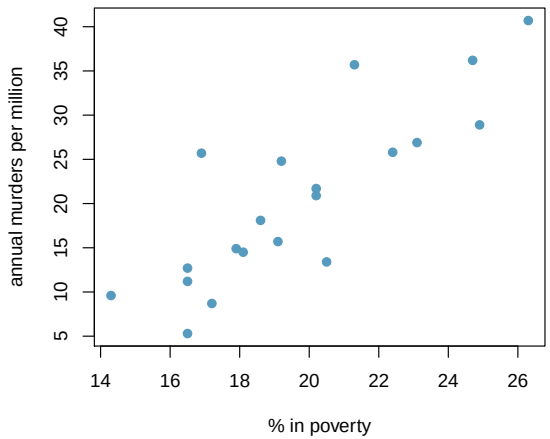
Lu Wang

Duke University, Department of Statistical Science

August 9, 2017

We want to build a model for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

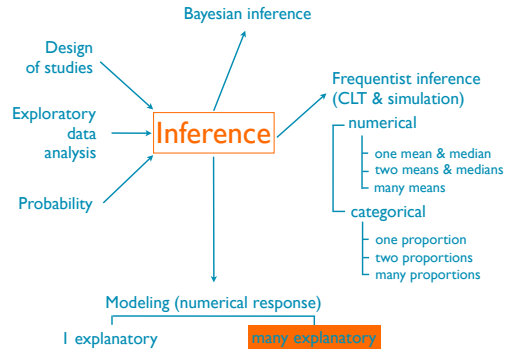




	annual murders / million (y)	% in poverty (x)
mean	$\bar{y} = 20.57$	$\bar{x} = 19.72$
sd	$s_y = 9.88$	$s_x = 3.24$
	correlation	$R = 0.84$

1. Calculate the slope.
2. Calculate the intercept.
3. Write out the linear model.

The Federal Trade Commission annually rates varieties of domestic cigarettes according to their tar and nicotine content. Past studies have shown that increases in the tar and nicotine content of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke.



1. The response variable is carbon monoxide emitted (`CO`). Suppose the full model uses the following explanatory variables: `nicotine`, `tar`, `length`, `filter`, `pack`, `strength`, and `type`. Describe, briefly, in your own words, how you would carry out model selection using the backward elimination method based on adjusted R^2.

2. The pairwise scatterplots show that `NIC` and `TAR` are both positively associated with `CO`. The output of the model resulting from backward elimination with adjusted R^2 is shown below. Evaluate the slopes of `NIC` and `TAR` variables. Are these results surprising if cor(`NIC`, `TAR`)=0.895? Why, or why not? Make sure to use appropriate terminology in your answer.

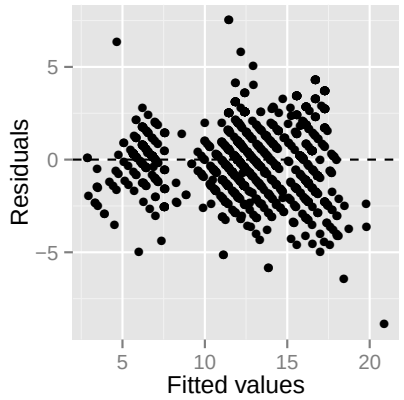
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.5489	0.5395	-1.02	0.3092
NIC	-4.0406	0.4342	-9.31	0.0000
TAR	1.0485	0.0441	23.80	0.0000
LEN	0.0350	0.0055	6.38	0.0000
FLTRNF	-6.4925	0.3577	-18.15	0.0000
PACKSOFT	0.5128	0.1046	4.90	0.0000
STRENGTHLIGHT	1.6804	0.2110	7.96	0.0000
STRENGTHMEDIUM	0.7339	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

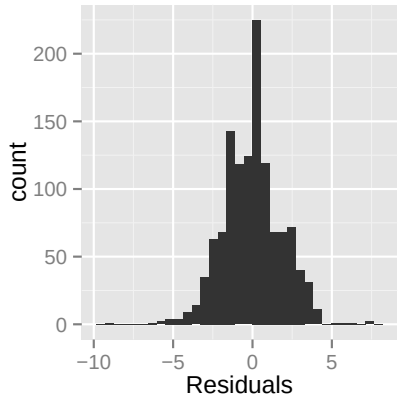
3. Next, we try the following two models, and obtain the following adjusted R^2 values:

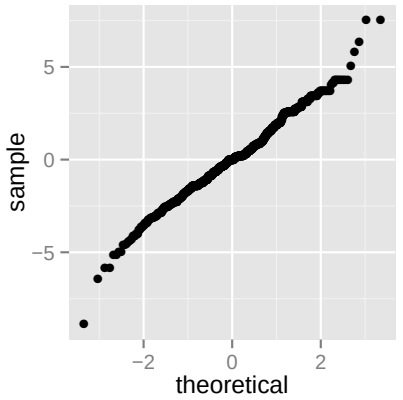
- Option 1, remove TAR: `lm(CO ~ NIC + LEN + FLTR + PACK + STRENGTH, data = cig07)`,
  adjusted $R^2 = 0.7066$
- Option 2, remove NIC: `lm(CO ~ TAR + LEN + FLTR + PACK + STRENGTH, data = cig07)`,
  adjusted $R^2 = 0.7857$

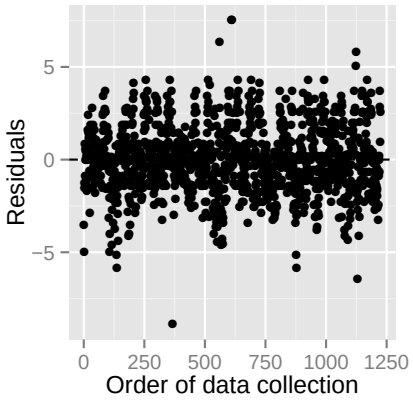
Based on these results which variable should we keep in our full model, nicotine or tar? Why?

4. We will complete some inferential tasks based on the final model. Use the following plots to check conditions before to determine whether we can proceed with these tasks.









5. Provided below is the final model output. Construct a 95% confidence interval for the slope of the `pack` variable (PACKSOFT), and interpret it in context. (Pack type: hard or soft)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0586	0.5555	-0.11	0.9160
TAR	0.7344	0.0293	25.07	0.0000
LEN	0.0267	0.0056	4.76	0.0000
FLTRNF	-6.1949	0.3686	-16.81	0.0000
PACKSOFT	0.5597	0.1081	5.18	0.0000
STRENGTHLIGHT	1.9077	0.2168	8.80	0.0000
STRENGTHMEDIUM	0.7900	0.4766	1.66	0.0976
STRENGTHREGULAR	0.5664	0.3149	1.80	0.0723
STRENGTHFULL FLAVOR	3.0920	0.3268	9.46	0.0000

Residual standard error: 1.836 on 1216 degrees of freedom

Multiple R-squared: 0.7871, Adjusted R-squared: 0.7857

F-statistic: 561.8 on 8 and 1216 DF, p-value: < 2.2e-16

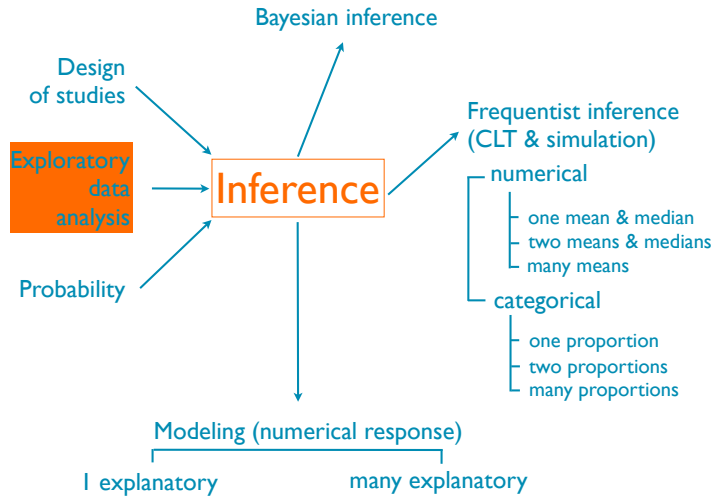
Logistic Regression

Consider the output from fitting a logistic regression to the credit card default data, where `default` is a binary variable (1=default, 0=no default), `student` is an indicator of being a student, `balance` is the average credit card balance after making monthly payments and `income` is the income of the customer.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.87	0.49	-22.08	0.0000 ***
studentYes	-0.65	0.24	-2.74	0.0062 **
balance	5.74×10^{-3}	2.32×10^{-4}	24.74	0.0000 ***
income	3.03×10^{-6}	8.20×10^{-6}	0.37	0.7115

1. Provide an interpretation for the slope of `studentYes` in terms of odds of defaulting.
2. Calculate a 95% confidence interval for $\exp(\beta_s)$.
3. The credit card manager is surprised that the slope of `income` is not significant. Can we conclude that income is unrelated to credit card default? Why or why not?

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



(a) a mosaic plot

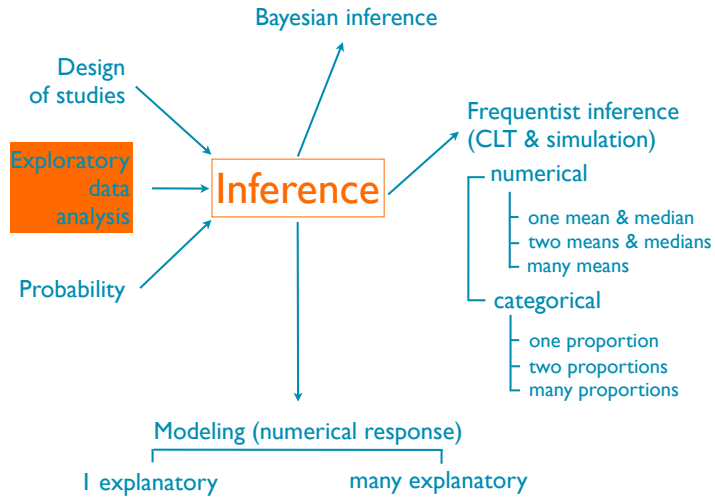
(b) a segmented frequency bar plot

(c) a frequency histogram

(d) a relative frequency histogram

(e) side-by-side box plots

Which of the following is false?



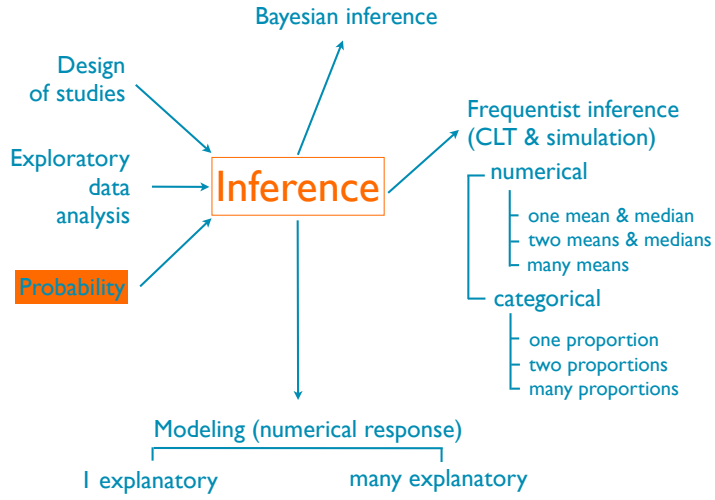
(a) Box plots are useful for highlighting outliers, but we cannot determine skew based on a box plot.

(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.

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.

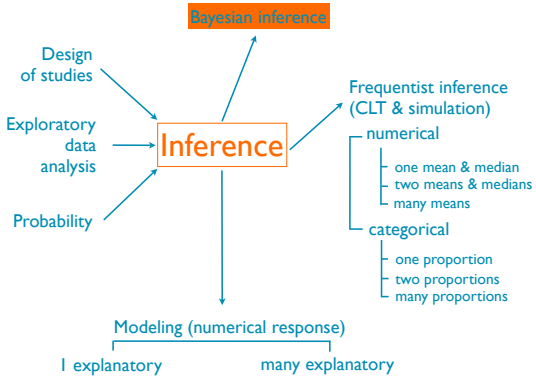
(b) If A and B are disjoint, then knowing that A occurs tells us that B cannot occur.

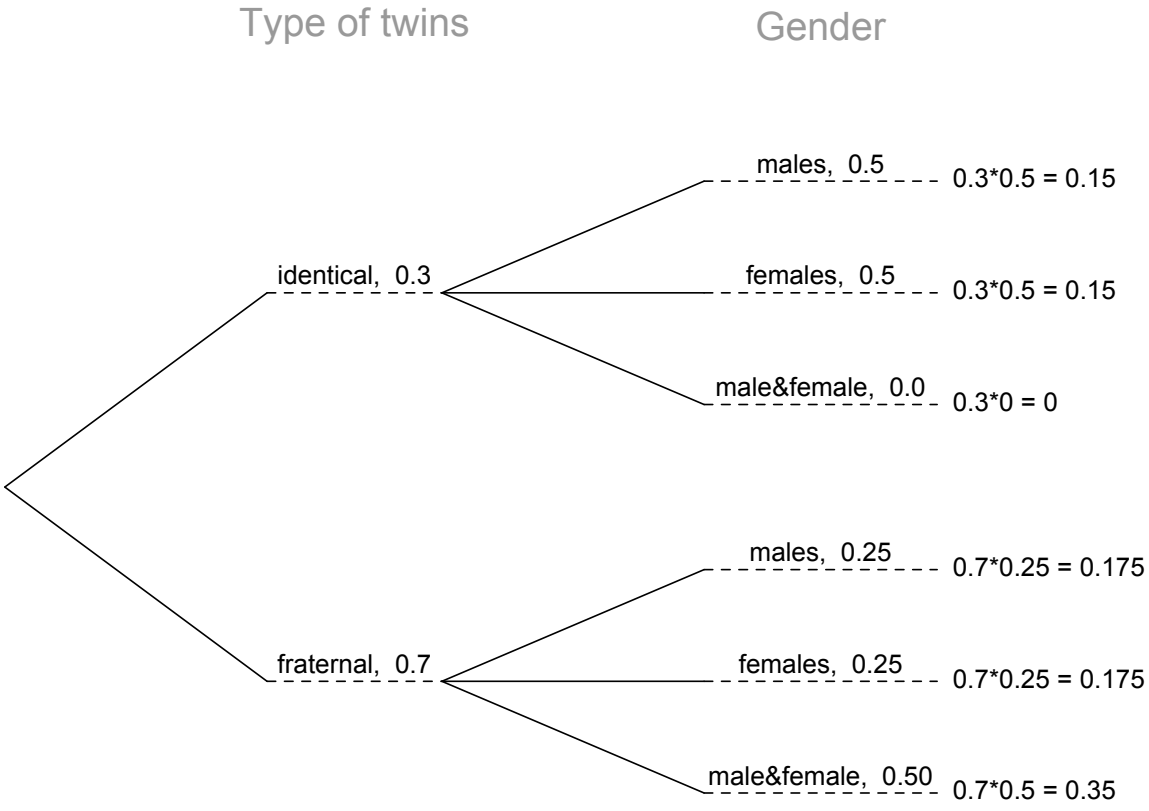
(c) Disjoint (mutually exclusive) events are always dependent since if one event occurs we know the other one cannot.

(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?





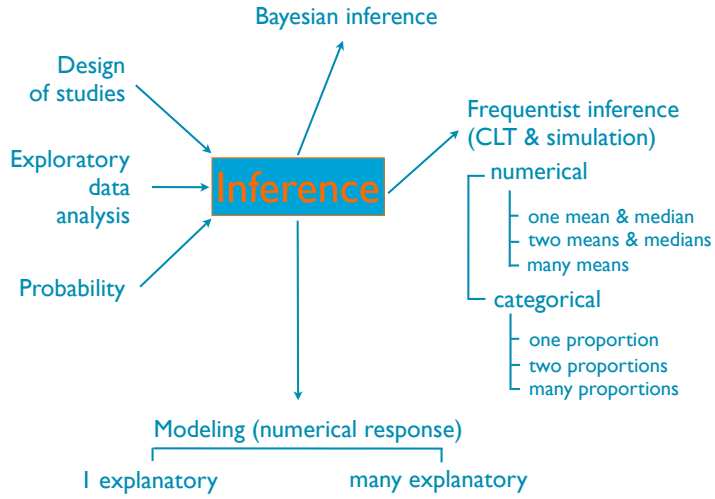
$$P(iden \mid f) = \frac{P(iden \ \& \ f)}{P(f)} = \frac{0.15}{0.15 + 0.175} = 0.46$$

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
$n = 30$, z statistic			
$n = 5000$, z statistic			

Note: The 95% quantile of N(0,1) is 1.64; the 97.5% quantile of N(0,1) is 1.96.

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

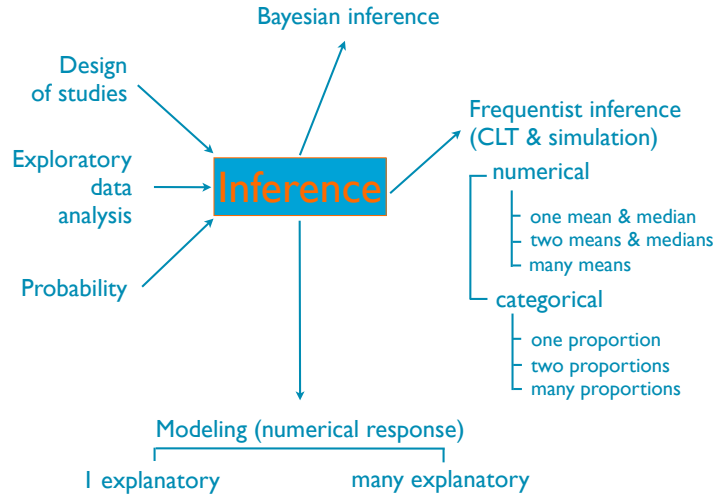
(b) chi-square test of goodness of fit

(c) anova

(d) linear regression

(e) t-test

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.
- Controls were chosen from women of comparable age who were in the hospital at the same time as the Breast-cancer cases but who did not have breast cancer.
- All women were asked about their age at first birth.

Breast Cancer & Age - set-up

The set of women with at least one birth was divided into two categories:

1. women whose age at first birth was less than or equal to 29 years
2. women whose age at first birth was greater than or equal to 30 years

The following results were found among women with at least one birth:

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

Breast Cancer & Age - set-up

- 683 of 3220 women with breast cancer (case women) had an age at first birth ≥ 30.
- 1498 of 10,245 women without breast cancer (control women) had an age at first birth ≥ 30.

How can we assess whether this difference is significant or simply due to chance?

- 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

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

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) $\sqrt{\frac{0.162\times(1-0.162)}{13465} + \frac{0.162\times(1-0.162)}{13465}}$

Breast Cancer & Age - test statistic & p-value

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

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

Breast Cancer & Age - confidence interval

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

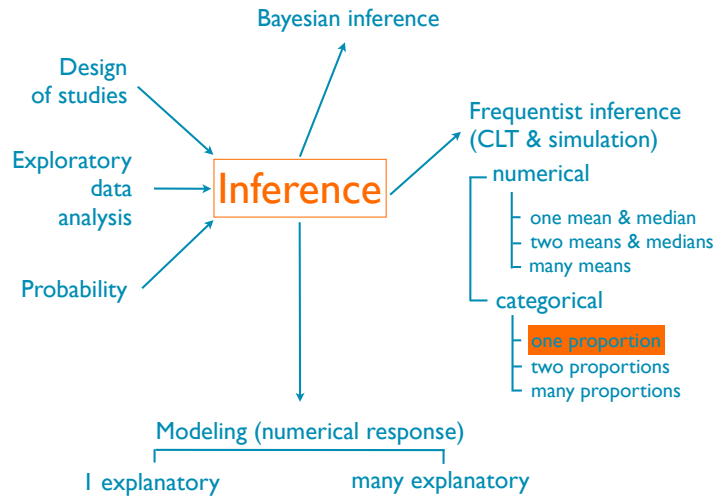
$$\begin{aligned} &\text{point estimate} \pm ME \\ =\ &\text{point estimate} \pm z^\star \times SE \end{aligned}$$

For a confidence interval,

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

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

$n = 30$ 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?

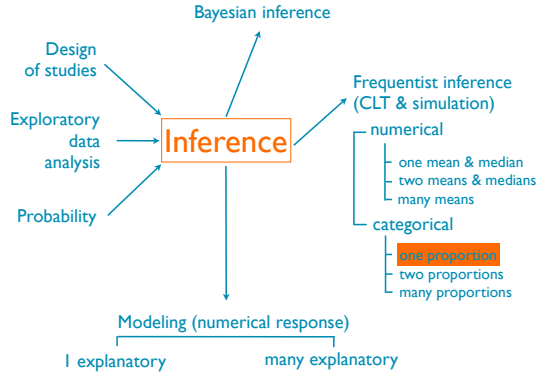


(a) CLT-based inference using the normal distribution

(b) simulation-based inference

(c) exact calculation using the binomial distribution

$n = 30$ and $\hat{p} = 0.7$. Hypotheses: $H_0 : p = 0.8; H_A : p < 0.8$. Suppose we want 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 $\hat{p}_{sim} \leq 0.7$.

(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 $\hat{p}_{sim} \leq 0.7$.

(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 $\hat{p}_{sim} \leq 0.7$.