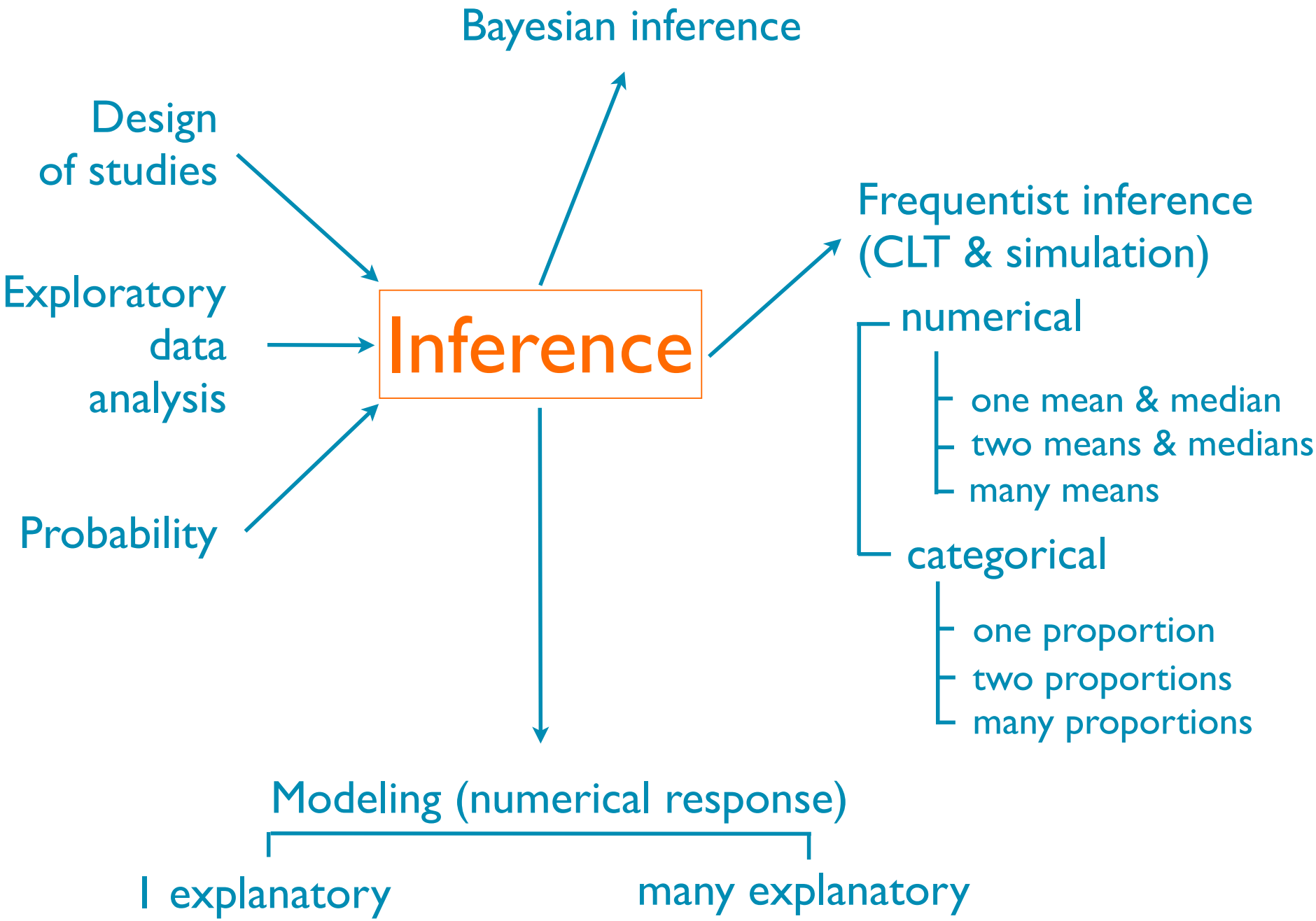
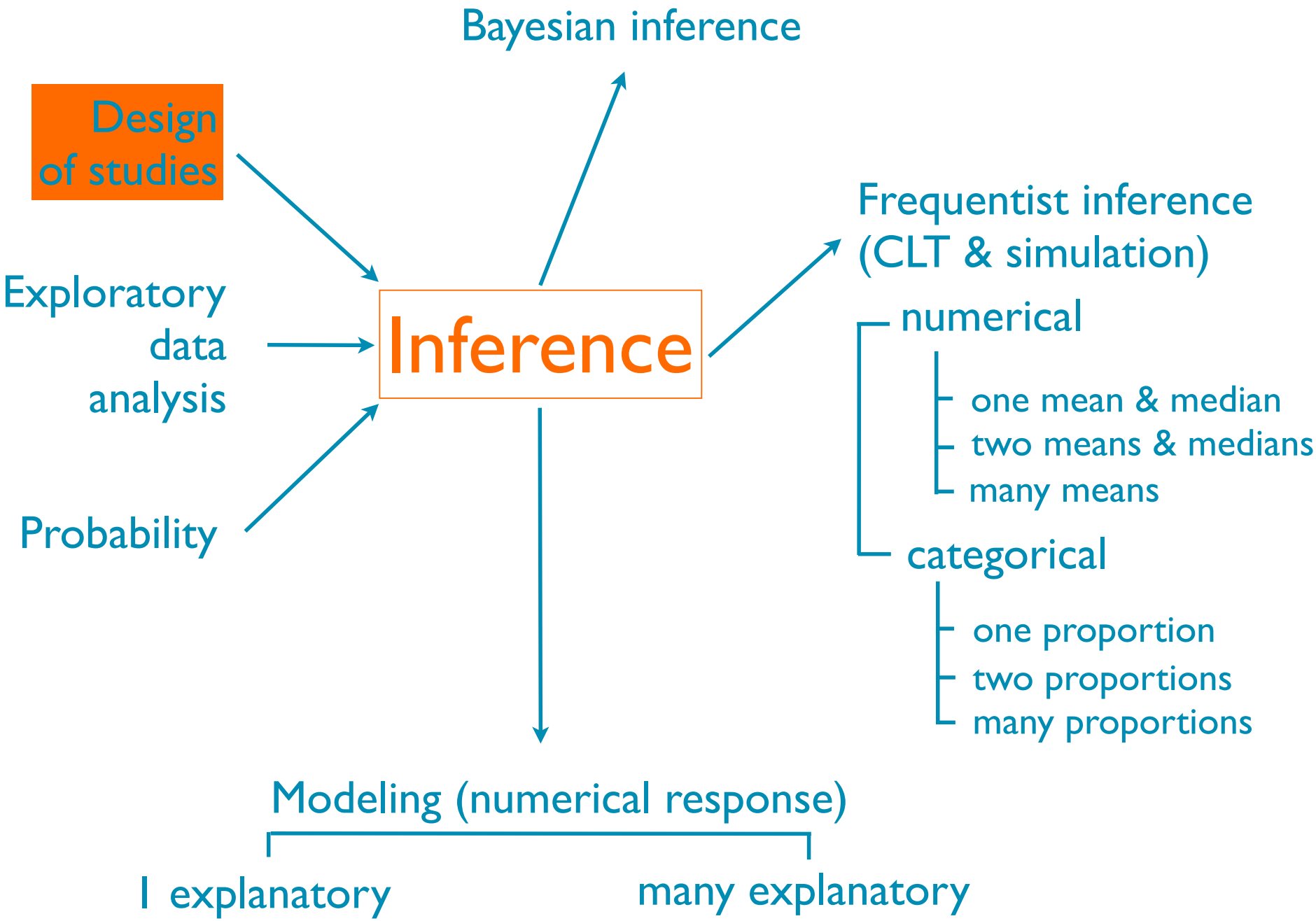
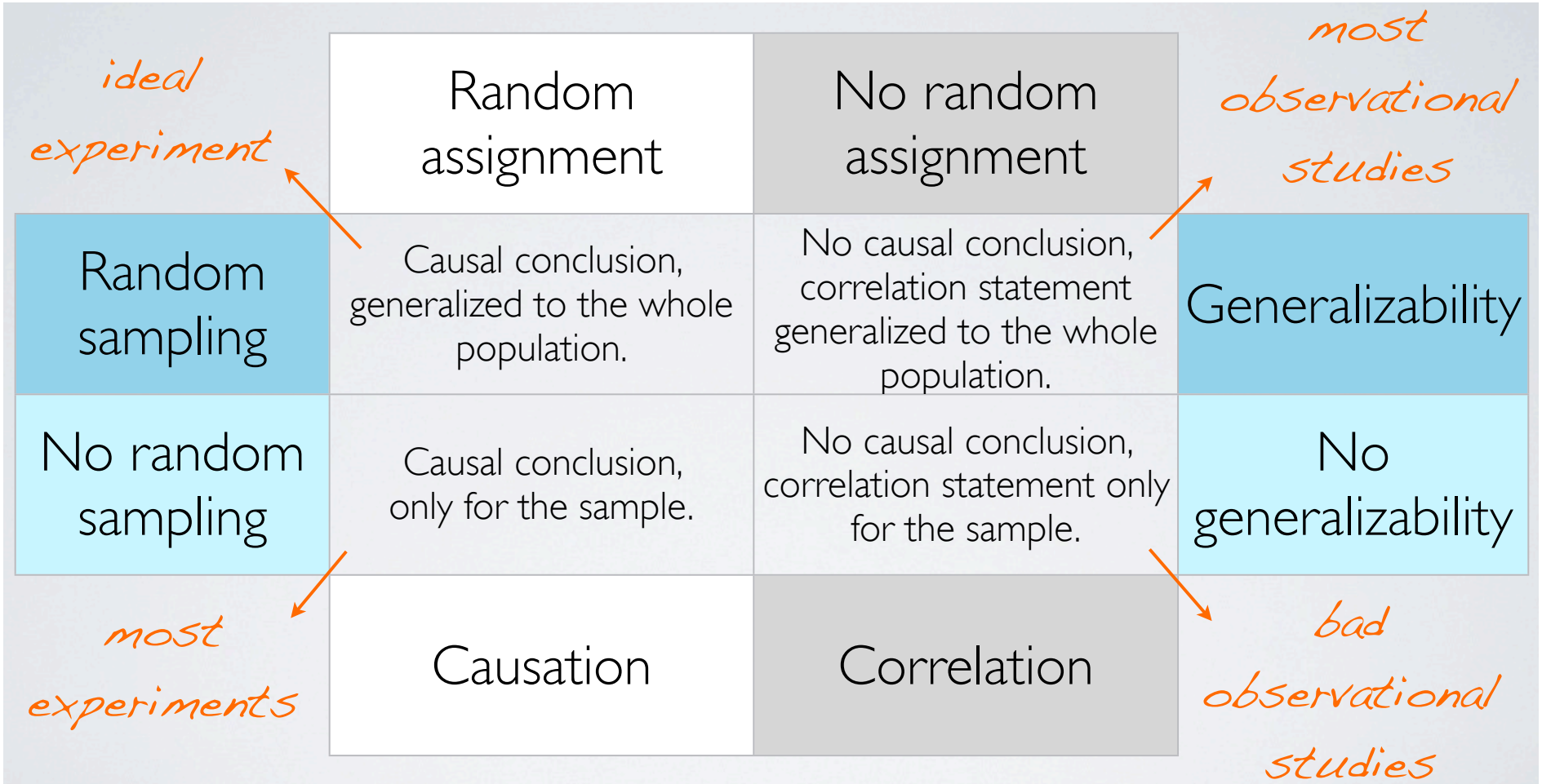
Lecture 27: Final Review

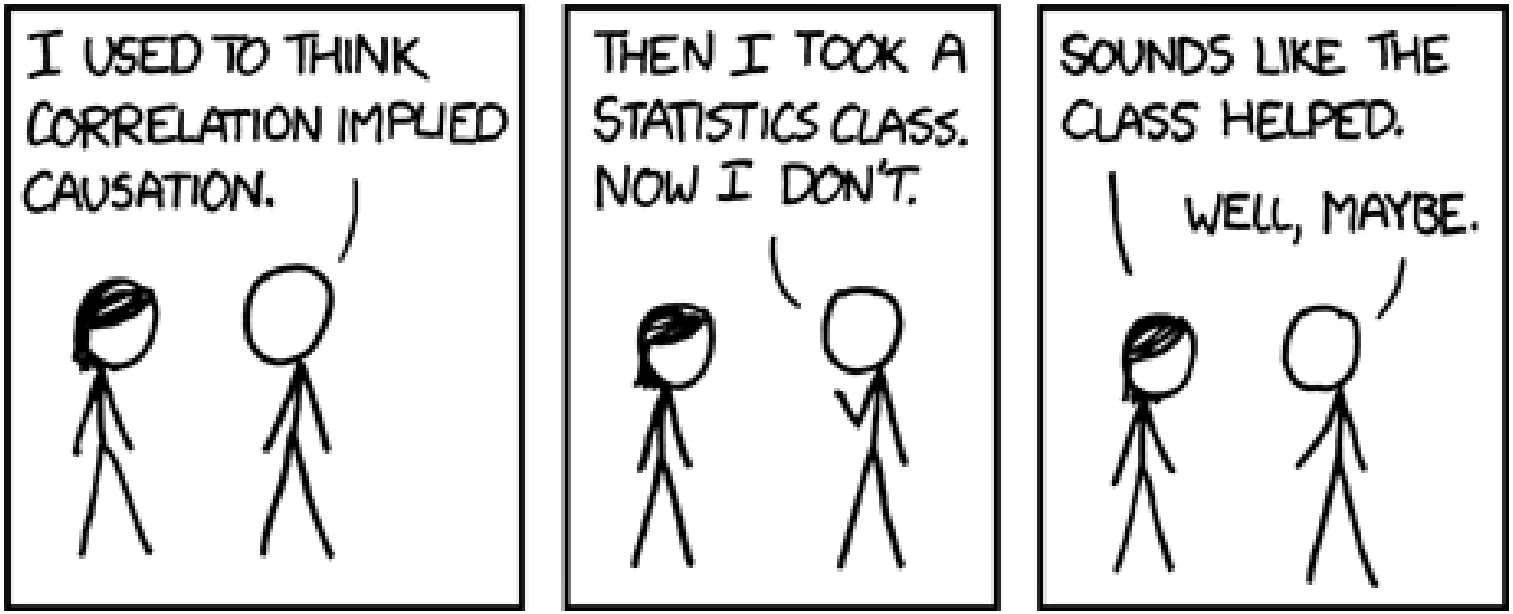

Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory


Bayesian inference
Design of studies
Frequentist inference (CLT & simulation)
Exploratory data analysis
Inference
numerical
one mean & median
two means & medians
many means
Probability
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Random assignment vs. random sampling

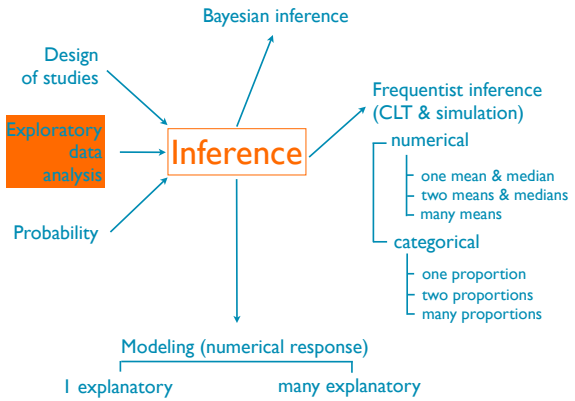


Correlation does not imply causation



http://xkcd.com/552/

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



(a) side-by-side box plots

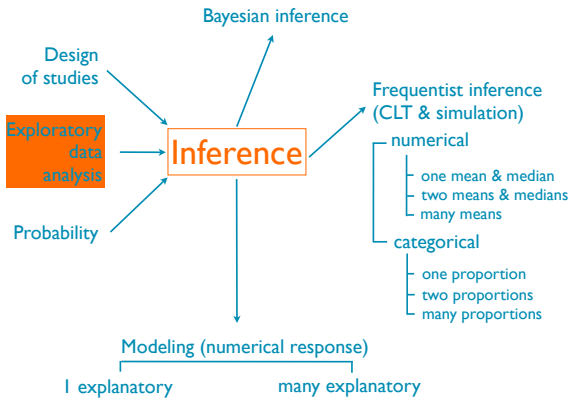
(b) mosaic plot

(c) pie chart

(d) segmented frequency bar plot

(e) relative frequency histogram

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



(a) side-by-side box plots

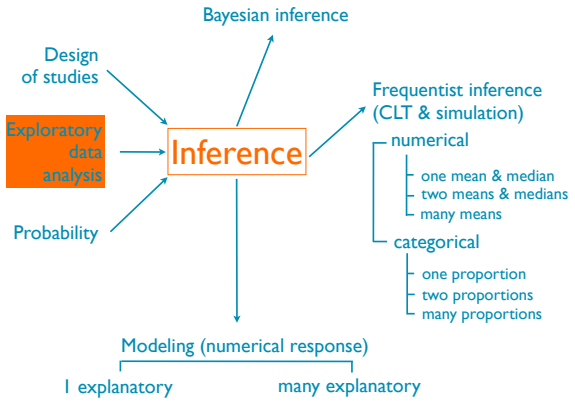
(b) *mosaic plot*

(c) pie chart

(d) segmented frequency bar plot

(e) relative frequency histogram

Which of the following is false?



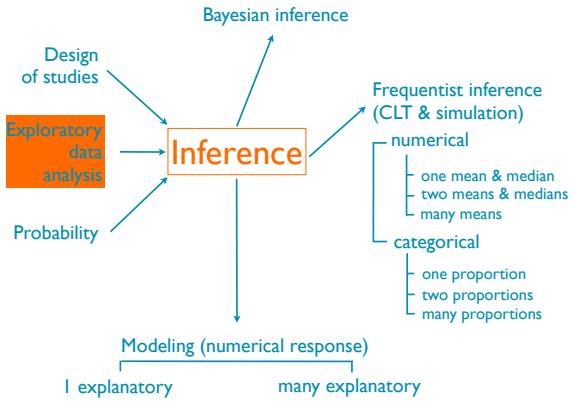
(a) Box plots are useful for highlighting outliers, but we cannot determine skew based on a box plot.

(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?



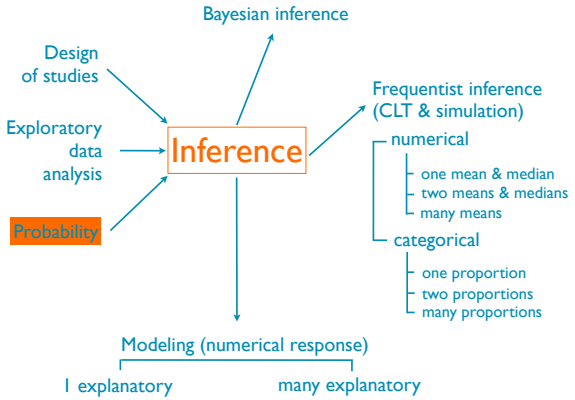
(a) *Box plots are useful for highlighting outliers, but we cannot determine skew based on a box plot.*

(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.

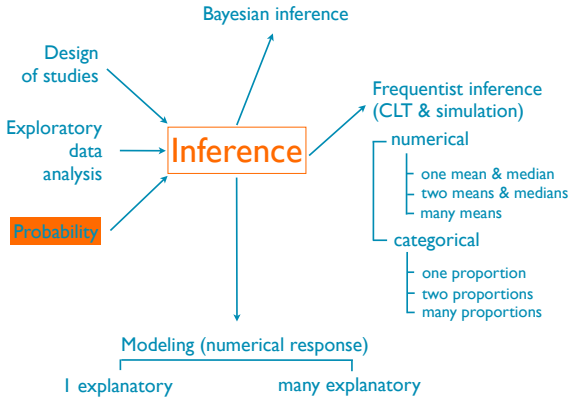
(b) If A and B are disjoint, then knowing that A occurs tells us that B cannot occur.

(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).

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.

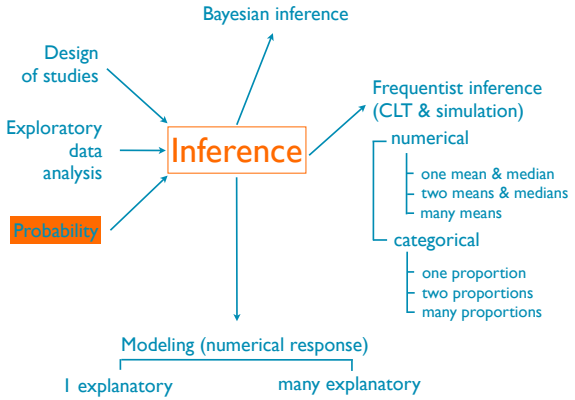
(b) If A and B are disjoint, then knowing that A occurs tells us that B cannot occur.

(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).

Which of the following is the least useful method for assessing if the data follow a normal distribution?



(a) Check if 68% of the data are within 1 SD of the mean, 95% of data are within 2 SDs of the mean, and 99.7% of data are within 3 SDs of the mean.

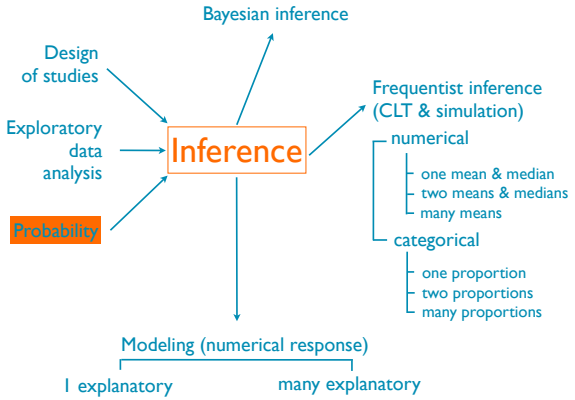
(b) Check if the points are on a straight line on a normal probability plot.

(c) Check if the mean and median are equal.

(d) Check if the distribution is unimodal and symmetric.

(e) Generate normally distributed random data with same mean and standard deviation as the observed data, overlay the plots of the generated and observed data, and check if they line up.

Which of the following is the least useful method for assessing if the data follow a normal distribution?



(a) Check if 68% of the data are within 1 SD of the mean, 95% of data are within 2 SDs of the mean, and 99.7% of data are within 3 SDs of the mean.

(b) Check if the points are on a straight line on a normal probability plot.

(c) *Check if the mean and median are equal.*

(d) Check if the distribution is unimodal and symmetric.

(e) Generate normally distributed random data with same mean and standard deviation as the observed data, overlay the plots of the generated and observed data, and check if they line up.

Estimation

- Our starting point will be a **random sample** of n individuals $X_1, X_2, \ldots, X_n$, which we will assume to be **i.i.d.** (independent and identically distributed) from some **model**
- We want to estimate an unknown characteristic θ of the model from which we assume the data come from. The usual name that we will given to θ is **parameter**. Some examples of parameter could be the mean, median, or the variance of the distribution
- We will find an **estimator** to estimate the parameter, which we will typically denote $\hat{\theta}$. The estimator has to be something we can compute using the data, i.e., a function $g(X_1, X_2, \ldots, X_n)$.

Point Estimates

Statisticians often provide two things:

- a point estimate of some quantity of interest, and
- a statement of the uncertainty in that estimate.

A **parameter** is some property of a distribution (or density function), such as the mean, median, standard deviation, and so forth.

A **point estimate** for a parameter is some statistic $h(X_1, \ldots, X_n)$ which, when evaluated for a random sample, gives a sensible approximation to the parameter.
Notice that a **point estimator** has to be random since it is a function of a random sample from some distribution, but the true parameter itself is constant.

Common Examples

– **Sample mean**: As we have seen, if $X_1, X_2, \ldots, X_n$ are a random sample such that $E(X_i) = \mu$ and $V(X_i) = \sigma^2$, $\overline{X}_n$ is an unbiased estimator of μ, with variance $V(X_i) = \sigma^2/n$ and MSE equal to σ^2/n

Common Examples

– **Sample mean**: As we have seen, if $X_1, X_2, \ldots, X_n$ are a random sample such that $E(X_i) = \mu$ and $V(X_i) = \sigma^2$, $\overline{X}_n$ is an unbiased estimator of μ, with variance $V(X_i) = \sigma^2/n$ and MSE equal to σ^2/n

– **Sample variance**: Suppose that $X_1, X_2, \ldots, X_n$ are random sample such that $E(X_i) = \mu$ and $V(X_i) = \sigma^2$, and now we want to estimate σ^2 from the data. The sample variance is defined as

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

which is an unbiased estimator of σ^2:

$$\begin{aligned} E(s_n^2) &= \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n} X_i^2\right) - nE\left(\overline{X}_n^2\right)\right] \\ &= \frac{1}{n-1}\left[n(\mu^2+\sigma^2) - n(\mu^2+\sigma^2/n)\right] = \sigma^2 \end{aligned}$$

– **Estimating proportions**: Let Y be Binomial(n, p), where n is known and p is unknown. Our intuition tells us that the sample proportion $\widehat{p} = Y/n$ should be a reasonable estimator of p. The sample proportion is unbiased $E(\widehat{p}) = E(Y/n) = p$ and its variance is $V(\widehat{p}) = p(1-p)/n$. Note that $\widehat{p}$ is just a sample average (a sum of Bernoulli(p) over n), so we could have used what we know about sample averages directly.

– **Clinical trial, sample size determination**: We are in charge of designing a clinical trail whose goal is to estimate the proportion of patients that will recover from a disease after taking a new treatment. We know we will model our data as $Y \sim$ Binomial(n, p), but now we want to determine a sample size that would allow us to estimate p sufficiently well. The variance of the sample proportion is $p(1-p)/n$ which depends on p. However, $p(1-p)$ is maximized at $p = 0.5$, so we know that the variance of our estimator will be smaller than (or equal to) $0.25/n$. Therefore, if we want to design an experiment such that the variance of the estimator is less than or equal to 0.005 (standard deviation of approximately 0.071), we should select a sample size of at least 50 individuals.

Maximum Likelihood Estimation

Suppose we have a random sample of i.i.d. random variables $X_1, X_2, \ldots, X_n$ with a PMF or PDF $f_\theta(x)$ which depends on a parameter θ. The joint PMF/PDF is

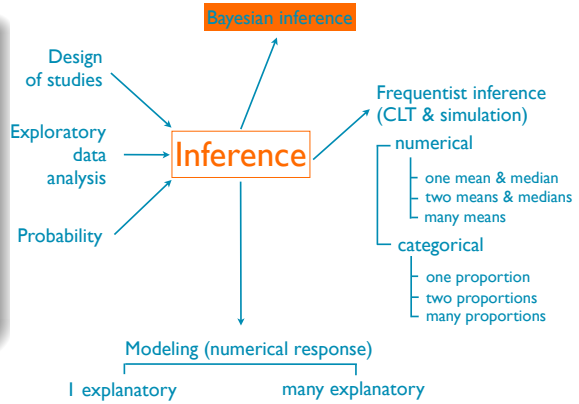
$$f_\theta(x_1, x_2, \ldots, x_n) = f_\theta(x_1) f_\theta(x_2) \ldots f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

Upon observing the data, we can substitute $x_1, x_2, \ldots, x_n$ by the actual values in the sample, so $f_\theta(x_1, x_2, \ldots, x_n)$ becomes a function of θ alone. Seeing $f_\theta(x_1, x_2, \ldots, x_n)$ as a function of θ, we write

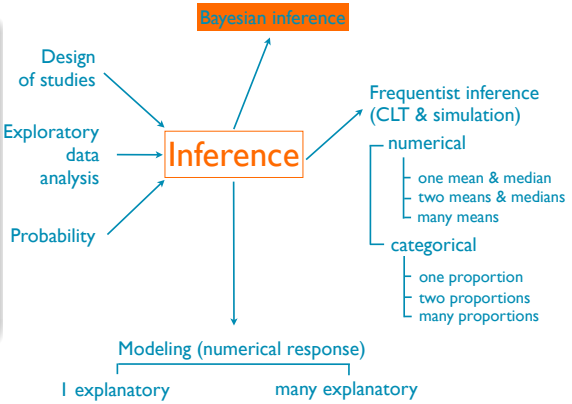
$$\mathcal{L}(\theta) = f_\theta(x_1, x_2, \ldots, x_n),$$

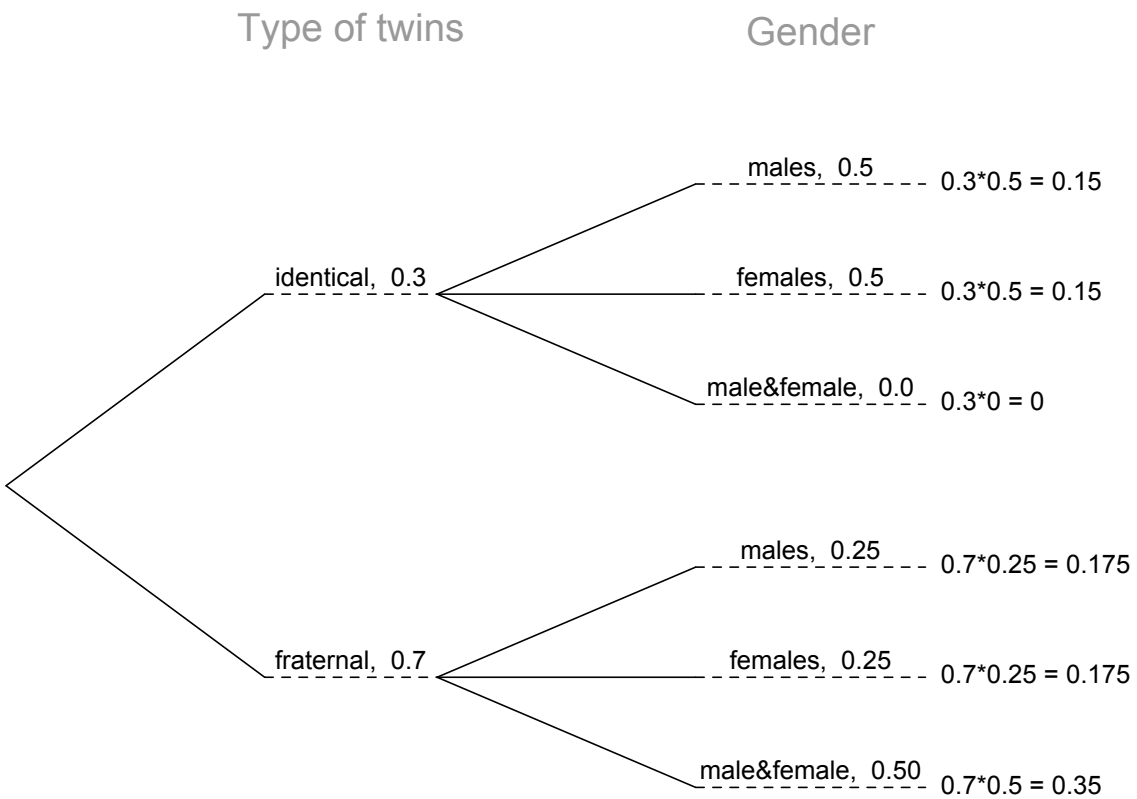
and call $\mathcal{L}(\theta)$ the **likelihood** of the data. The **Maximum Likelihood Estimator of θ (MLE)** is the value $\widehat{\theta}$ that maximizes the likelihood. Products are typically hard to maximize, so we usually take logarithms and maximize the **log-likelihood** $\ell(\theta) = \log \mathcal{L}(\theta)$ instead.

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?

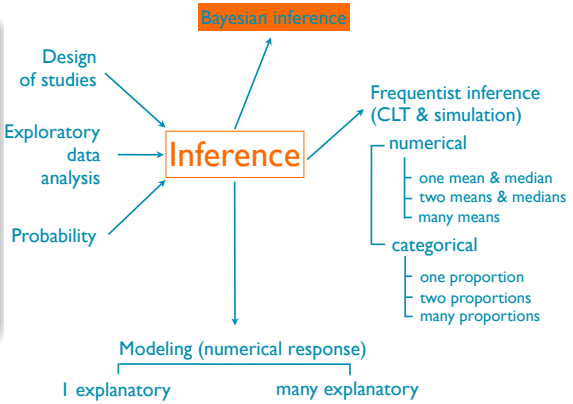


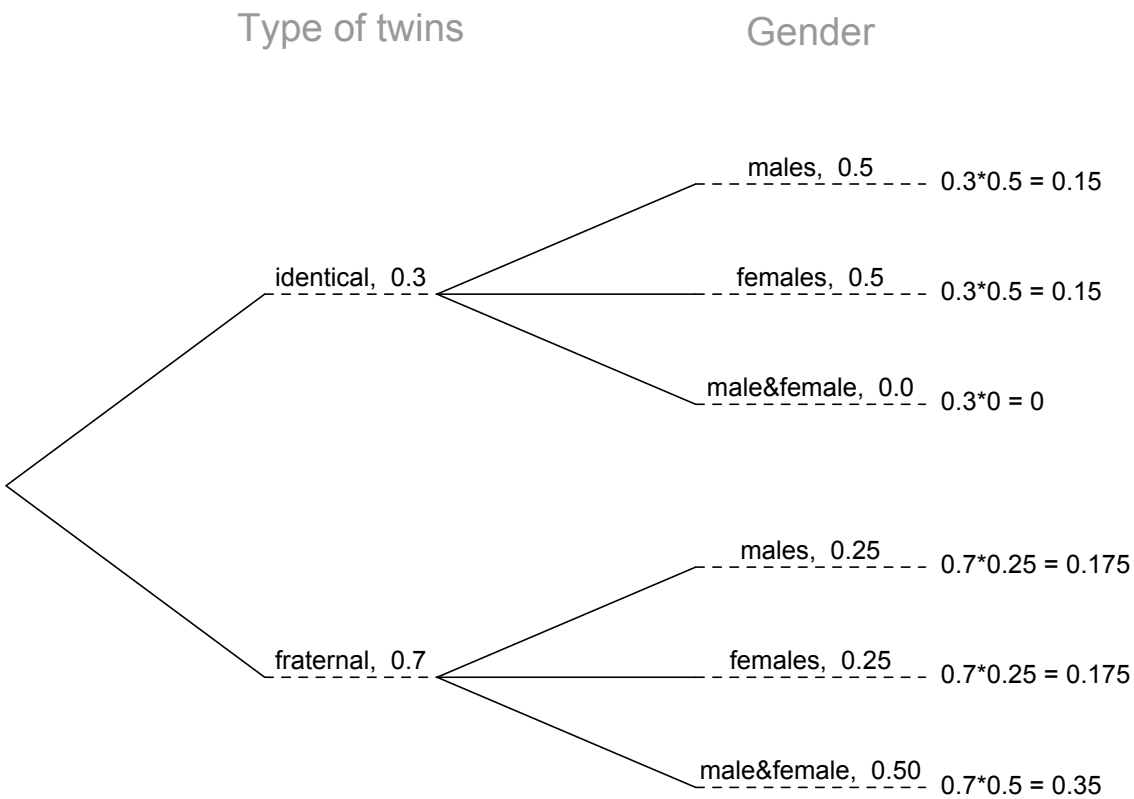
About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a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



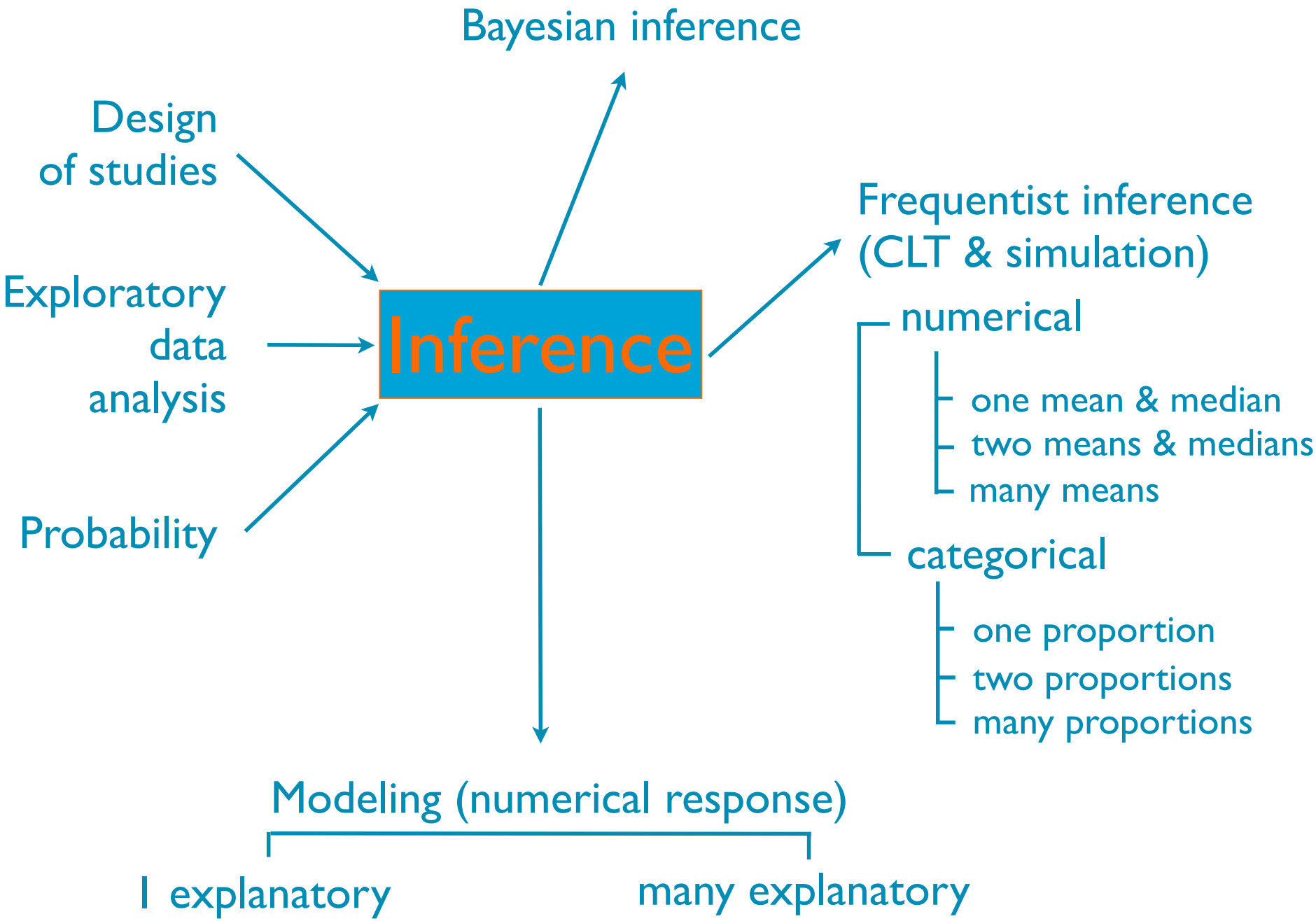


About 30% of human twins are identical and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a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





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


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Confidence Interval

GENERAL FORM

two-sided interval	**upper interval**	**lower interval**
U,L = $pe \pm (se)(cv_C)$	**U =** $pe + (se)(cv_C)$	**L =** $pe + (se)(cv_{1-C})$

Here

- pe is the point estimate of the parameter of interest,
- se is the standard error of our estimate, or an estimate of that standard error,
- cv_C is the value from a table that has area C under the curve in the appropriate place (i.e., middle, left tail, or right tail, respectively).

"In 95% of similarly constructed intervals, the true mean will lie within the interval." or "95% of the time, we will draw a sample that generates a confidence interval that contains the true value."

Normal, Known variance

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from a Normal(μ, σ^2) where μ is unknown but σ^2 is known. As usual, let $\overline{X}_n$ be the sample mean,

$$Z = \sqrt{n}(\overline{X}_n - \mu)/\sigma \sim \text{Normal}(0,1),$$

and let $z_{\alpha/2}$ be the value such that

$$P(Z \geq z_{\alpha/2}) = P(Z \leq -z_{\alpha/2}) = \alpha/2$$

(The value can be found on normal probability table; for example, if $\alpha = 0.05$, $z_{\alpha/2}$ is approximately 1.96), where $Z \sim \text{Normal}(0,1)$.

Normal, Known variance

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from a Normal(μ, σ^2) where μ is unknown but σ^2 is known. As usual, let $\overline{X}_n$ be the sample mean,

$$Z = \sqrt{n}(\overline{X}_n - \mu)/\sigma \sim \text{Normal}(0, 1),$$

and let $z_{\alpha/2}$ be the value such that

$$P(Z \geq z_{\alpha/2}) = P(Z \leq -z_{\alpha/2}) = \alpha/2$$

(The value can be found on normal probability table; for example, if $\alpha = 0.05$, $z_{\alpha/2}$ is approximately 1.96), where $Z \sim \text{Normal}(0, 1)$. Then

$$\overline{X}_n \pm z_{\alpha/2}\sigma/\sqrt{n}$$

is a $(1-\alpha)\%$ confidence interval for μ.

Normal, unknown variance

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from a Normal(μ, σ^2) with both μ and σ^2 unknown. Let $\overline{X}_n$ be the **sample mean** and s_n^2 be the **sample variance** $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2/(n-1)$. A 95% confidence interval for μ is

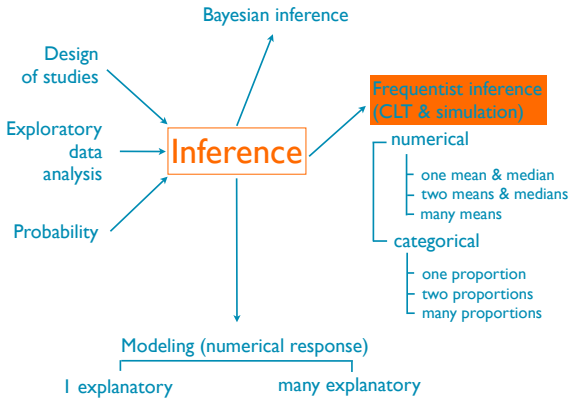
$$\overline{X}_n \pm t_{n-1,\alpha/2} s_n / \sqrt{n}$$

where $t_{n-1,\alpha/2}$ is the value such that

$$P(T \geq t_{n-1,\alpha/2}) = P(T \leq -t_{n-1,\alpha/2}) = \alpha/2$$

for $T \sim$ Student-$t(n-1)$, which you can find on the Student-t table.

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.

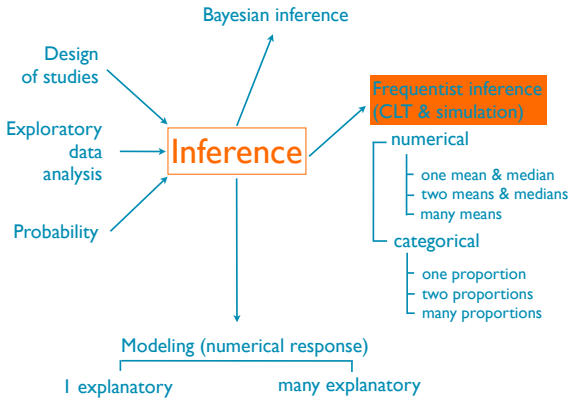
(b) SE used by Student A > SE used as Student B.

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

(d) SE used by Student A ≈ SE used as Student B.

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

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.



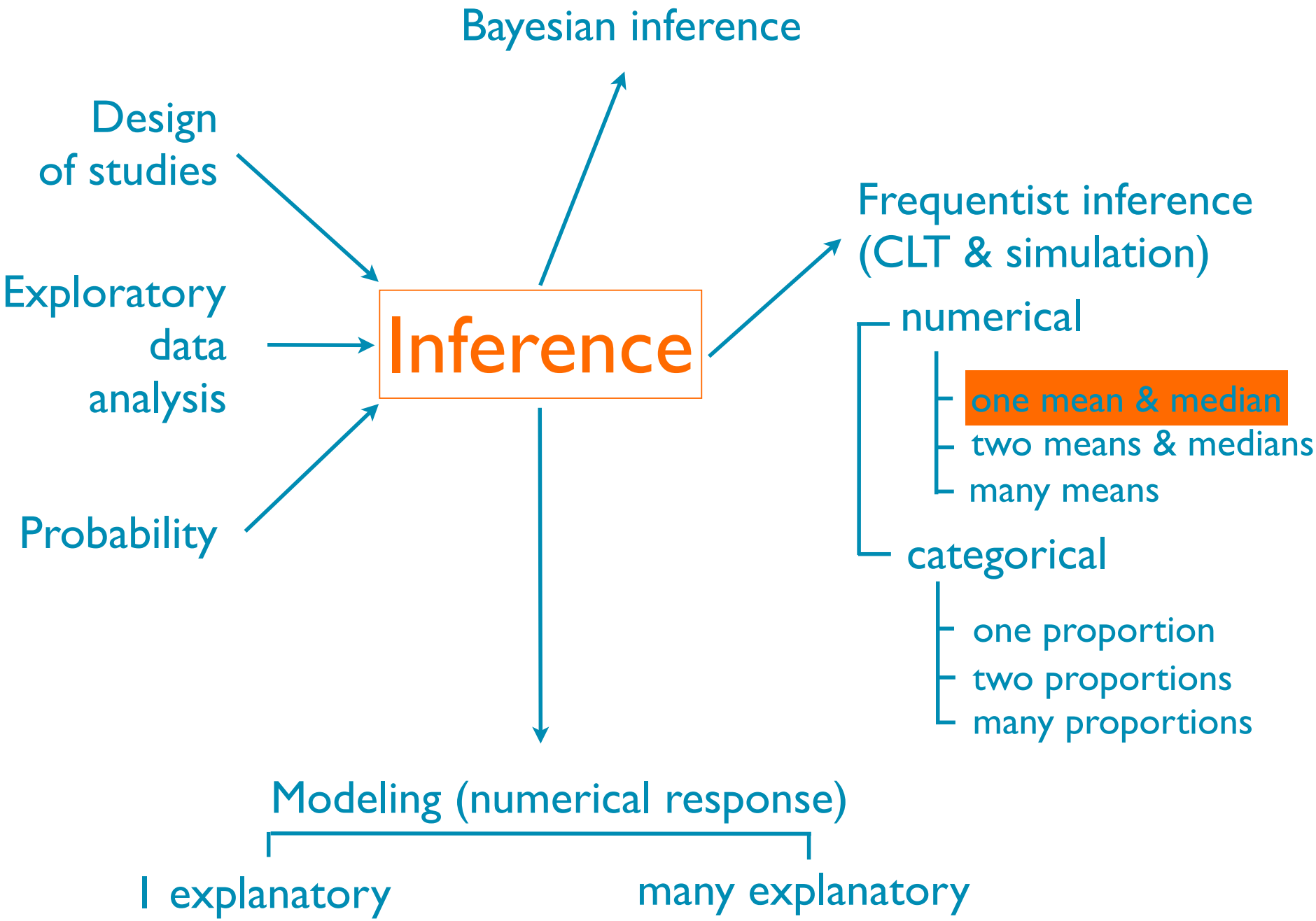
(a) *SE used by Student A < SE used as Student B.*

(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 ≈ SE used as Student B.

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

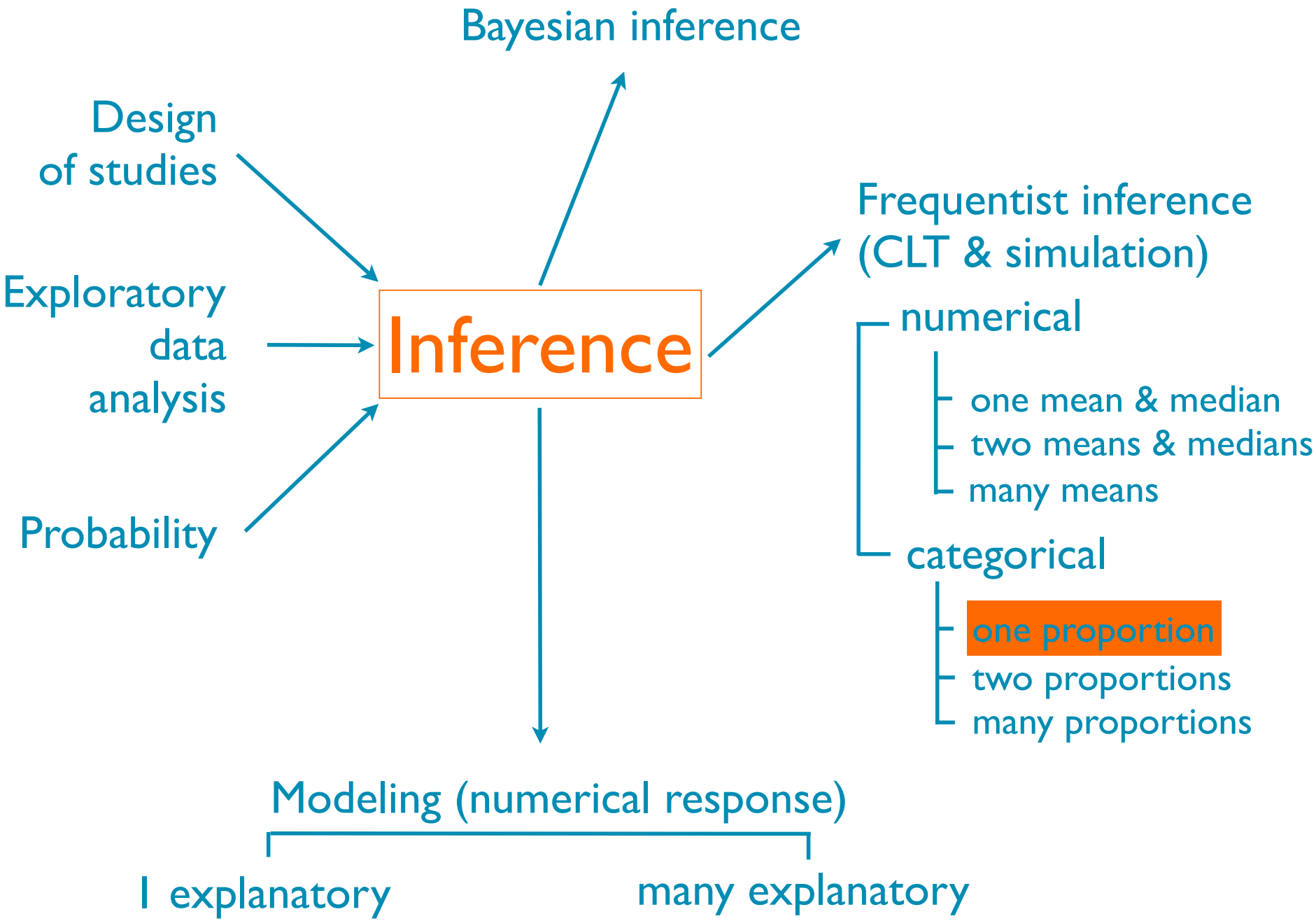

Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Approximate Interval

1. For a CI on the population mean μ, when either
   - the population standard deviation σ is known, or
   - $n > 31$, so the population σ is accurately estimated by the sample standard deviation

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

   then the *pe* is $\overline{X}$ and the *se* is $\sigma/\sqrt{n}$ or $\hat{\sigma}/\sqrt{n}$, as appropriate. The *cv* comes from the z table.
2. For a CI on the population mean μ when $n \leq 31$ and one estimates the population σ by the sample $\hat{\sigma}$, then the *pe* is $\overline{X}$ and the *se* is $\hat{\sigma}/\sqrt{n}$. The *cv* comes from the t_{n-1} table.


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Proportions

Let $X_1, X_2, \ldots X_n$ be i.i.d. Bernoulli(p). Let $\widehat{p} = \sum_{i=1}^{n} X_i/n$. By CLT

$$\widehat{p} \approx \text{Normal}(p, p(1-p)/n)$$

Therefore, we could try to reuse our work for the Normal and find the approximate interval $(1-\alpha)$ confidence interval

$$\widehat{p} \pm z_{\alpha/2}\sqrt{p(1-p)/n}$$

Proportions

Let $X_1, X_2, \ldots X_n$ be i.i.d. Bernoulli(p). Let $\widehat{p} = \sum_{i=1}^{n} X_i/n$. By CLT

$$\widehat{p} \approx \text{Normal}(p, p(1-p)/n)$$

Therefore, we could try to reuse our work for the Normal and find the approximate interval $(1-\alpha)$ confidence interval

$$\widehat{p} \pm z_{\alpha/2}\sqrt{p(1-p)/n}$$

But note that $\sqrt{p(1-p)/n}$ depends on p. We can approximate $\sqrt{p(1-p)/n}$ by $\sqrt{\widehat{p}(1-\widehat{p})/n}$ and still have that

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}$$

is an approximate $(1-\alpha)$ confidence interval.

Finite Population Correction Factor

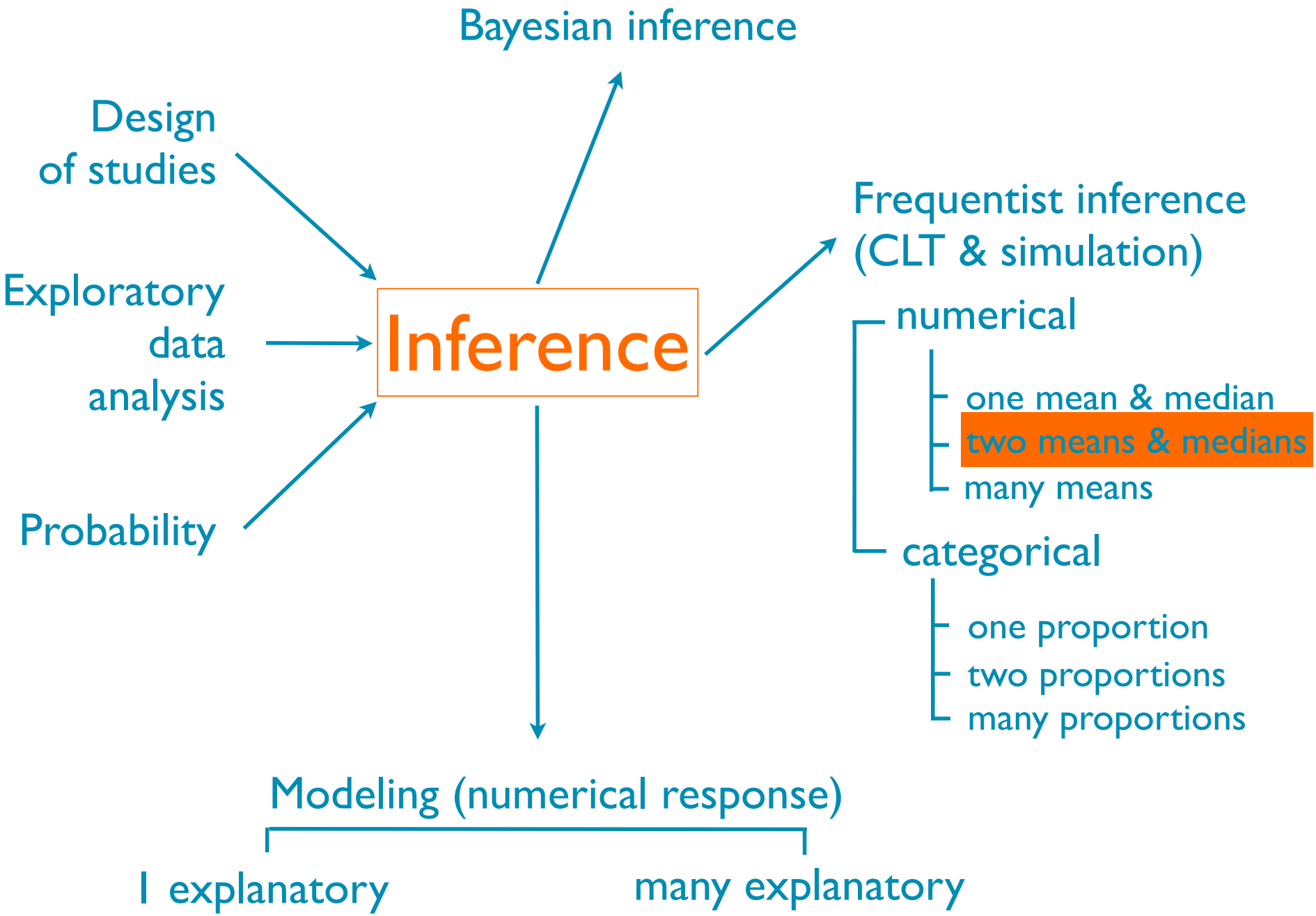
The CLT and the standard errors (deviations) of sample averages or mean are based on samples selected **with replacement**. However, in virtually all survey research, you sample **without replacement** from populations that are of a finite size, M.

If the sample size n is small compared to the population size M, one can ignore the distinction between sampling with replacement and without replacement. Then the standard deviation of an average is $\sigma/\sqrt{n}$.

If n is large relative to the population size N (say n, is more than 5% of the population size, N), then use the **Finite Population Correction Factor (FPCF)** multiply estimates of the standard error by the Finite Population Correction Factor (FPCF):

$$FPCF = \sqrt{\frac{N-n}{N-1}}.$$

In finite populations, if the sampling is without replacement, FPCF shrinks the sample standard deviation $\hat{\sigma}$.


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Normal, known σ^2

Suppose we have observations from two groups

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Normal}(\mu_1, \sigma_1^2)$$
$$Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} \text{Normal}(\mu_2, \sigma_2^2)$$

and assume that $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_m)$ are independent. Suppose further that σ_1^2 and σ_2^2 are **known** and we want to find a confidence interval for the difference in means $\mu_1 - \mu_2$. By properties of Normals, we have

$$\overline{X}_n - \overline{Y}_m \sim \text{Normal}(\mu_1 - \mu_2, \sigma_1^2/n + \sigma_2^2/m)$$

The setup looks awkward, but it is actually the same as in the interval for μ with one group and a known variance σ^2 (**why?**), so a $(1-\alpha)$ confidence interval is

$$(\overline{X}_n - \overline{Y}_m) \pm z_{\alpha/2} \sqrt{\sigma_1^2/n + \sigma_2^2/m}$$

Normal, unknown σ^2

Suppose we have observations from two groups

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Normal}(\mu_1, \sigma^2)$$

$$Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} \text{Normal}(\mu_2, \sigma^2)$$

with $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_m)$ independent. The two groups have the same variance. The population variance σ^2 is **unknown** and estimated from the data using a weighted average of sample variances:

$$s^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

where s_1^2 is the sample variance in group 1 (the group with the Xs) and s_2^2 is the sample variance in group 2 (the group with the Ys). Some textbook called s^2 the pooled sample variance. A $(1-\alpha)$ confidence interval is

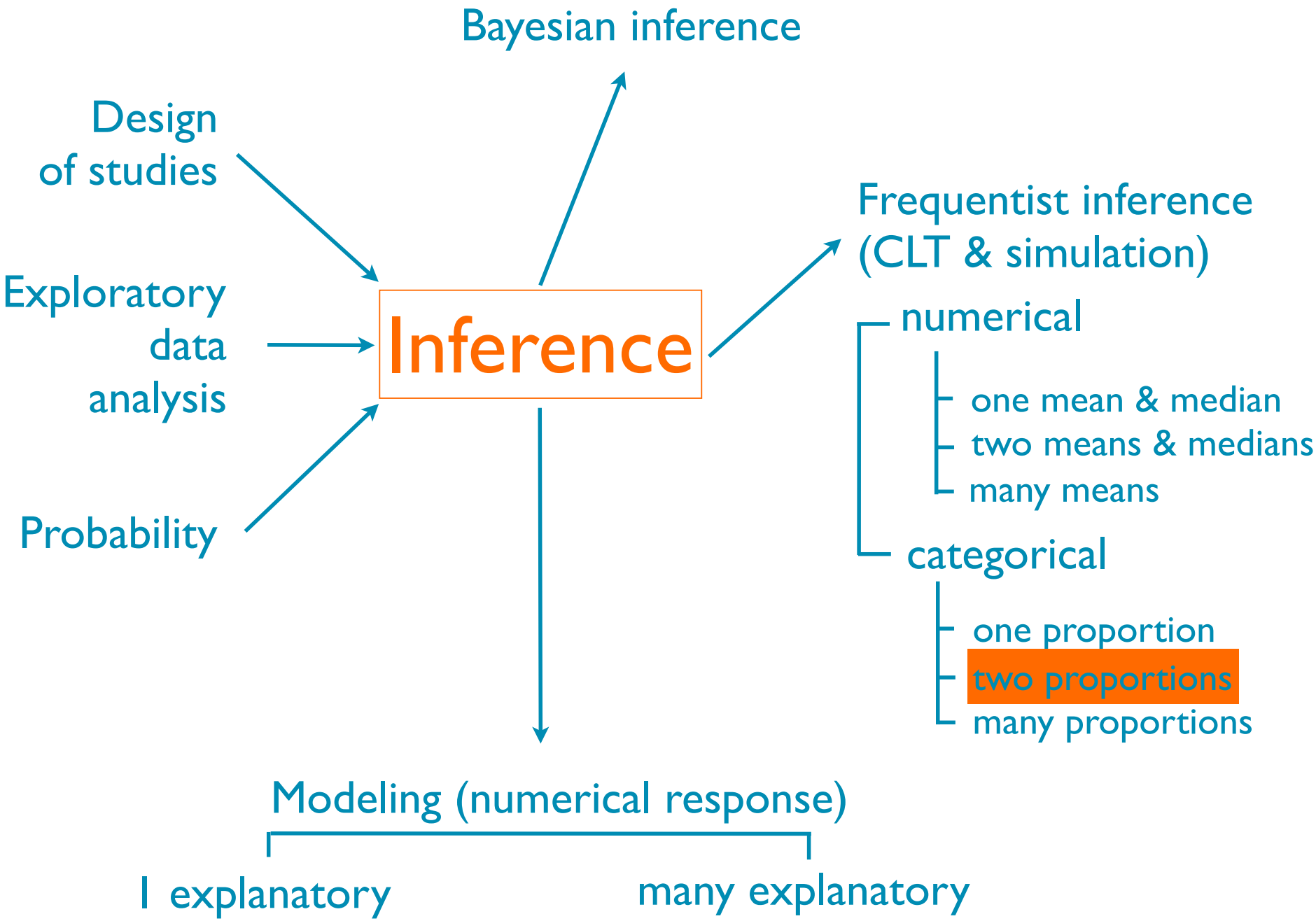
$$(\overline{X}_n - \overline{Y}_m) \pm t_{n+m-2,\alpha/2}\, s \sqrt{1/n + 1/m}$$

What can we do if the population variances of the groups are not equal?

- Well, if the sample sizes n and m are big enough, one option is (as usual) approximating σ_1^2 and σ_2^2 by s_1^2 and s_2^2 and reporting the interval

$$(\overline{X}_n - \overline{Y}_m) \pm z_{\alpha/2} \sqrt{\sigma_1^2/n + \sigma_2^2/m}$$

  as an approximate 95% confidence interval for $\mu_1 - \mu_2$.
- If the sample sizes aren't big enough, most statistical software packages have implementations of appropriate intervals. The formulas are awkward and we won't cover them here, but you should know that intervals exist if you ever need them.


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Proportions

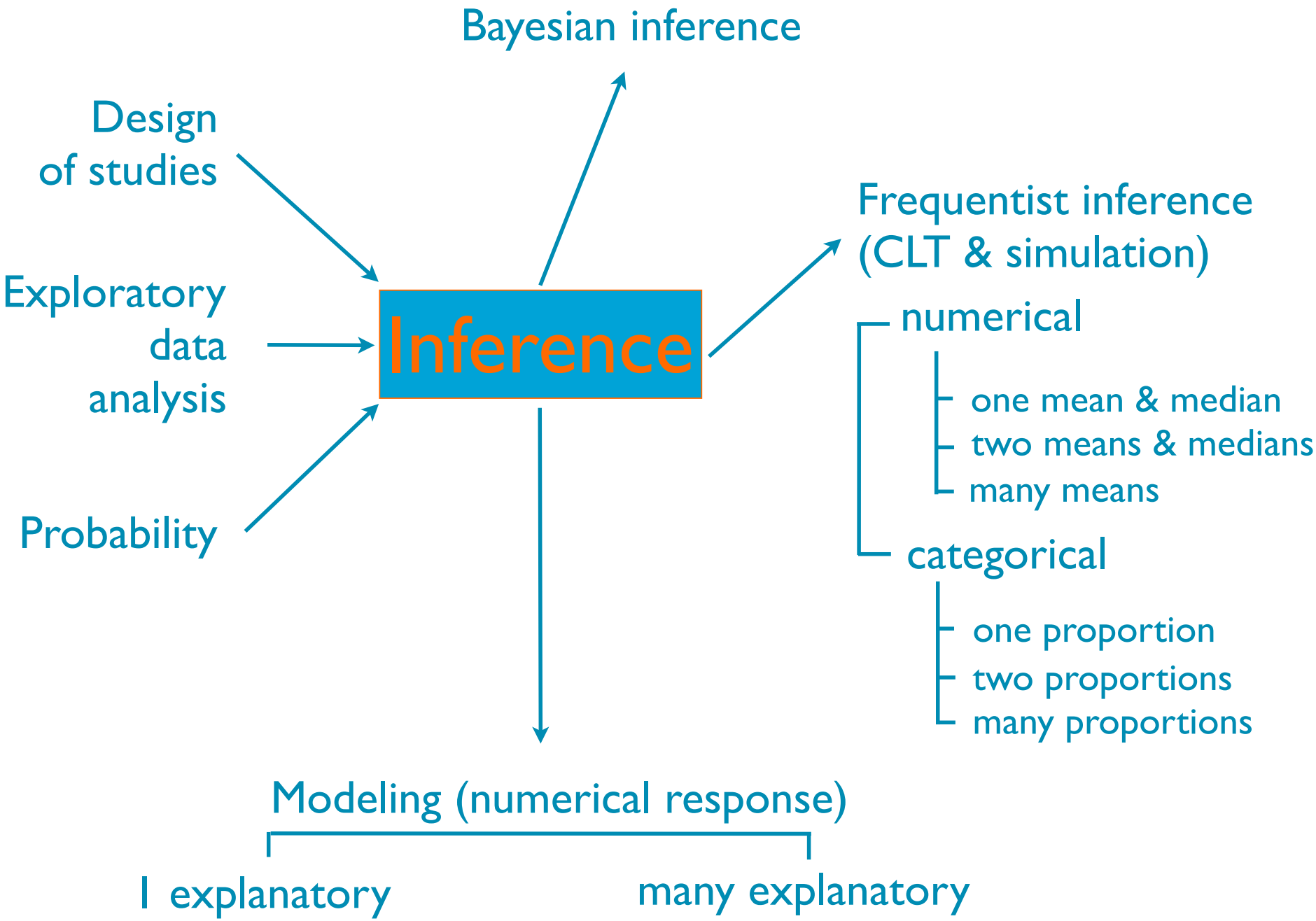
Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p_1)$ and $Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} \text{Bernoulli}(p_2)$. As usual, assume that $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_m)$ are independent. Let $\widehat{p}_1 = \overline{X}_n$ and $\widehat{p}_2 = \overline{Y}_m$ be the sample proportions in the two groups. By CLT, we have

$$\widehat{p}_1 - \widehat{p}_2 \approx \text{Normal}(p_1 - p_2, p_1(1-p_1)/n + p_2(1-p_2)/m)$$

so an approximate $(1-\alpha)$ confidence interval for the difference $p_1 - p_2$ is

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2} \sqrt{\widehat{p}_1(1-\widehat{p}_1)/n + \widehat{p}_2(1-\widehat{p}_2)/m}$$

We could use the same idea for Poisson and find an interval for the difference of λs as needed.


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Steps for Testing Hypotheses

1. Set the hypotheses.
2. Check assumptions and conditions.
3. Calculate a *test statistic* and a p-value.
4. Make a decision, and interpret it in context of the research question.

p-values

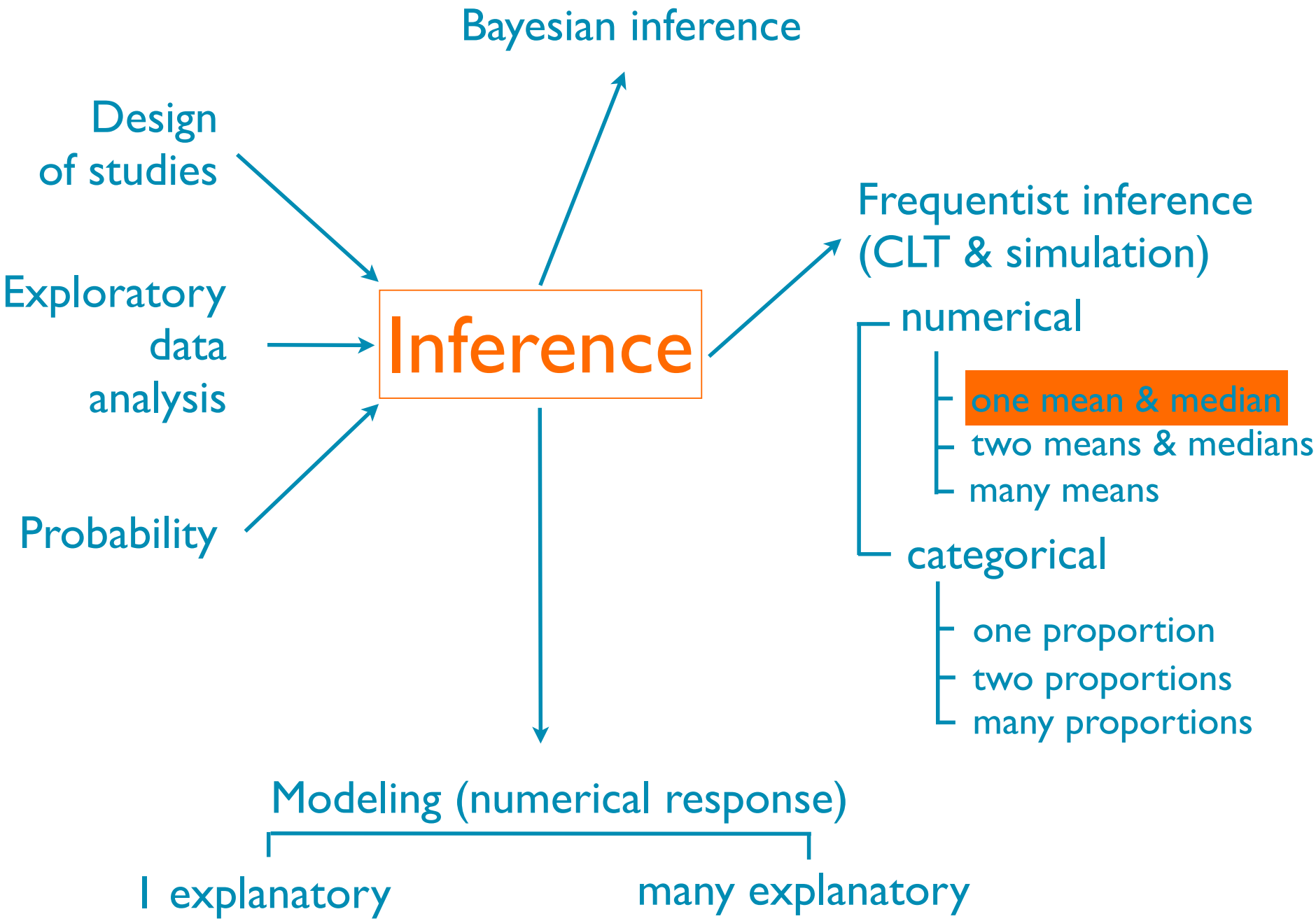
- We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.

p-values

- We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is *low* (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* H_0.

p-values

- We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is *low* (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* H_0.
- If the p-value is *high* (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence *do not reject* H_0.


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Hypothesis testing for a population mean

1. Set the hypotheses
   - $H_0 : \mu = \textit{null value}$
   - $H_A : \mu <$ or $>$ or $\neq \textit{null value}$
2. Calculate the point estimate
3. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Normality: nearly normal population or $n \geq 30$, no extreme skew – or use the t distribution
4. Calculate a *test statistic* and a p-value (draw a picture!)

$$Z = \frac{\overline{X} - \mu}{\text{SE}}, \textit{ where } \text{SE} = \frac{s}{\sqrt{n}}$$

5. Make a decision, and interpret it in context
   - If p-value $< \alpha$, reject H_0, data provide evidence for H_A
   - If p-value $> \alpha$, do not reject H_0, data do not provide evidence for H_A

Inference using the t-distribution

- If σ is unknown, use the t-distribution with SE $= \frac{s}{\sqrt{n}}$.

Inference using the t-distribution

- If σ is unknown, use the t-distribution with SE $= \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
  - no extreme skew

Inference using the t-distribution

- If σ is unknown, use the t-distribution with SE $= \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
  - no extreme skew
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{\text{SE}}, \text{ where } df = n - 1$$

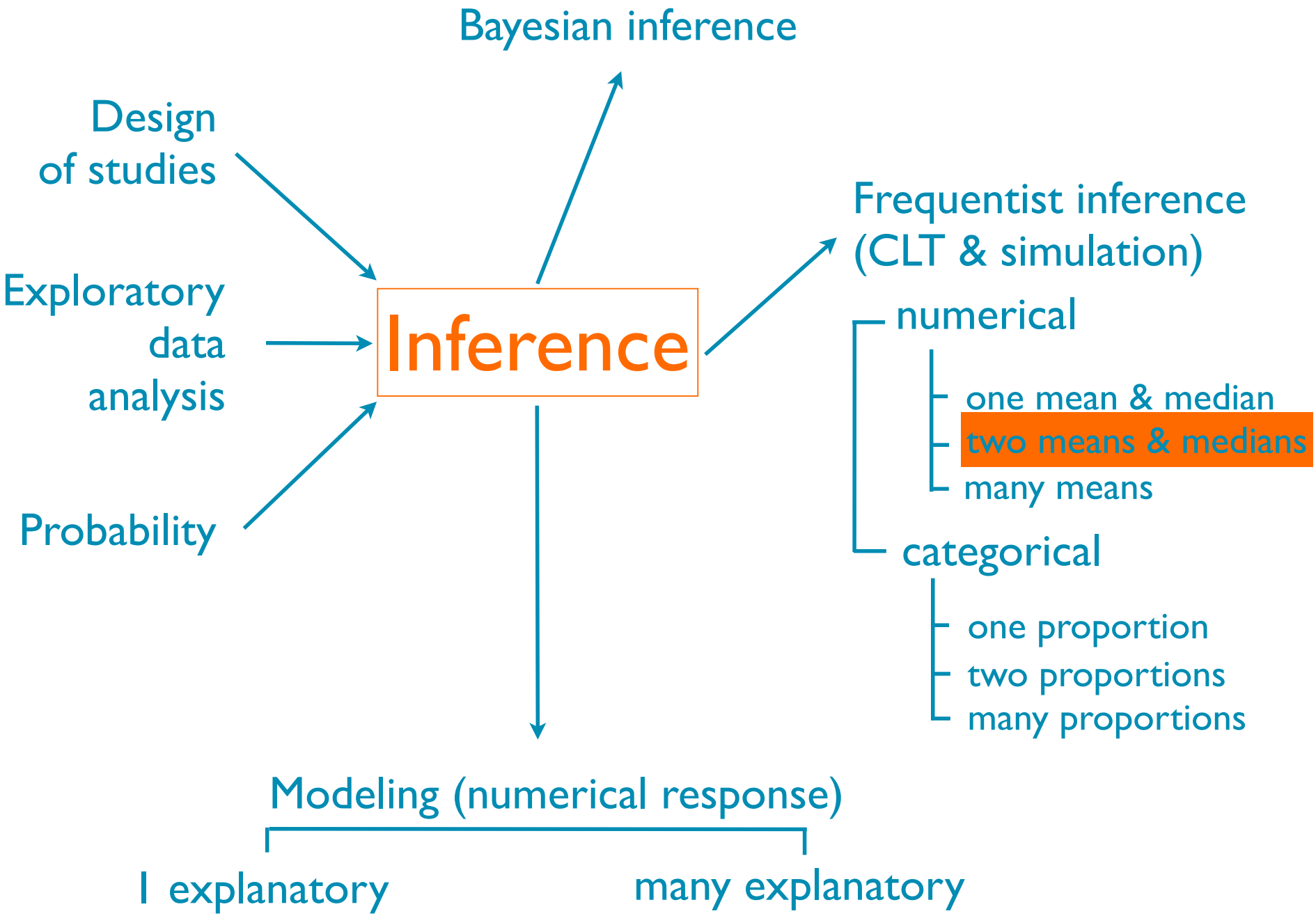
Inference using the t-distribution

- If σ is unknown, use the t-distribution with SE $= \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
  - no extreme skew
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{\text{SE}}, \text{ where } df = n - 1$$

- Confidence interval:

$$\text{point estimate} \pm t^{\star}_{df} \times \text{SE}$$


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a t-distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a t-distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population) and between groups
  - no extreme skew in either group

Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a t-distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population) and between groups
  - no extreme skew in either group
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = min(n_1 - 1, n_2 - 1)$$

Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a t-distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population) and between groups
  - no extreme skew in either group
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = min(n_1 - 1, n_2 - 1)$$

- Confidence interval:

$$\text{point estimate} \pm t^{\star}_{df} \times SE$$

One mean:

$df = n - 1$

HT:

$H_0 : \mu = \mu_0$

$T_{df} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

CI:

$\bar{x} \pm t^\star_{df} \frac{s}{\sqrt{n}}$

Paired means:

$df = n_{diff} - 1$

HT:

$H_0 : \mu_{diff} = 0$

$T_{df} = \frac{\bar{x}_{diff} - 0}{\frac{s_{diff}}{\sqrt{n_{diff}}}}$

CI:

$\bar{x}_{diff} \pm t^\star_{df} \frac{s_{diff}}{\sqrt{n_{diff}}}$

Independent means:

$df = min(n_1 - 1, n_2 - 1)$

HT:

$H_0 : \mu_1 - \mu_2 = 0$

$T_{df} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

CI:

$\bar{x}_1 - \bar{x}_2 \pm t^\star_{df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		*Type 1 Error, α*
	H_A true		

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	

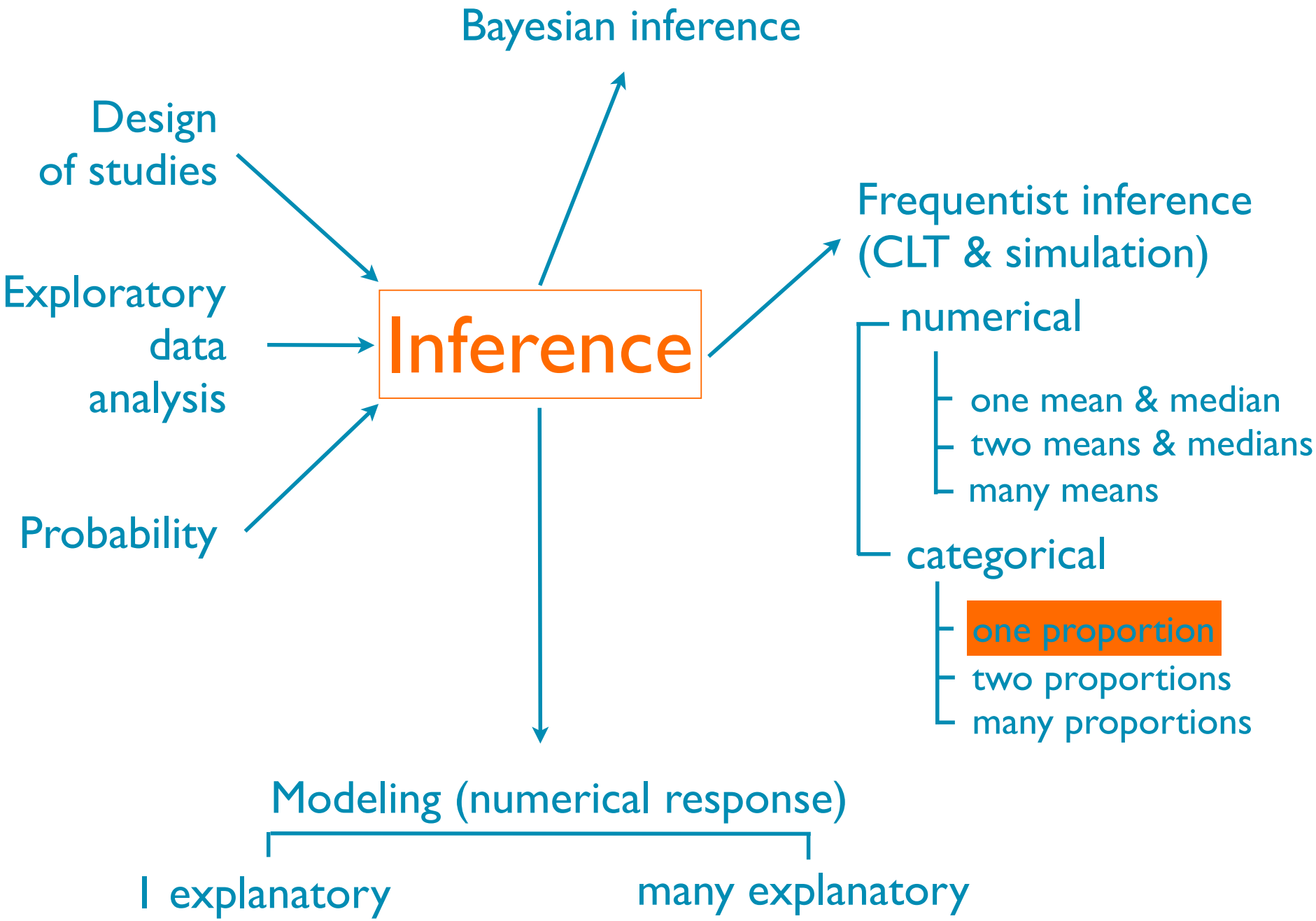
- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	*Power,* $1-\beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Inference for one proportion

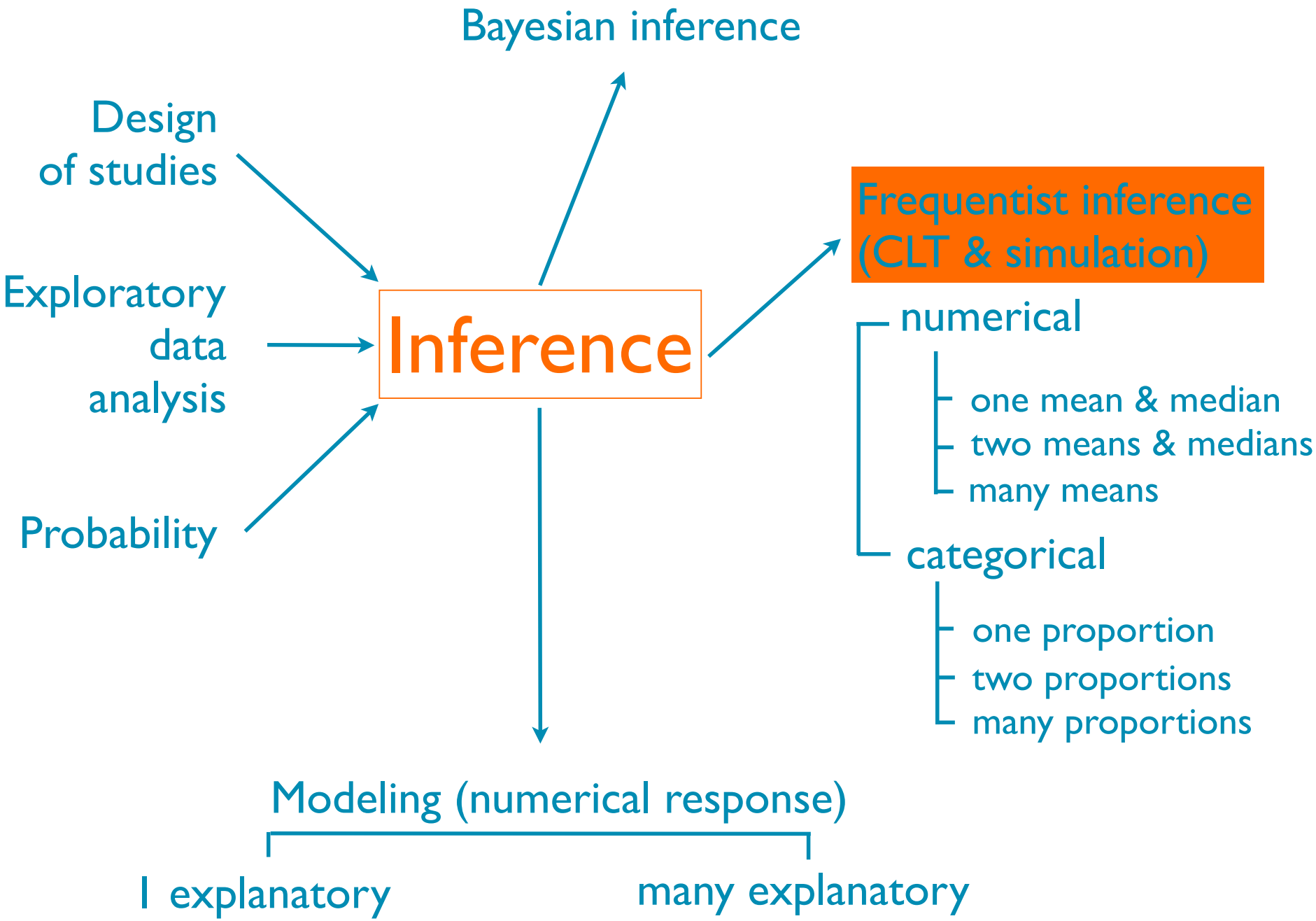
- Population parameter: p, point estimate: $\hat{p}$

Inference for one proportion

- Population parameter: p, point estimate: $\hat{p}$
- Conditions:
  - independence
    - random sample and 10% condition
  - at least 10 successes and failures
    - if not $\rightarrow$ randomization

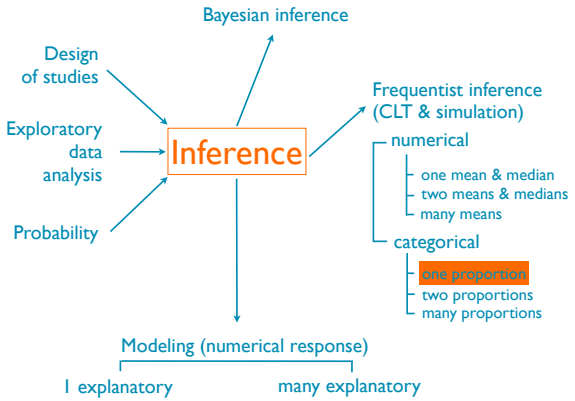
Inference for one proportion

- Population parameter: p, point estimate: $\hat{p}$
- Conditions:
  - independence
    - random sample and 10% condition
  - at least 10 successes and failures
    - if not $\rightarrow$ randomization
- Standard error: $SE = \sqrt{\frac{p(1-p)}{n}}$
  - for CI: use $\hat{p}$
  - for HT: use p_0


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

$n = 50$ and $\hat{p} = 0.80$. Hypotheses: $H_0 : p = 0.82; H_A : p < 0.82$. We use a randomization test because the sample size isn't large enough for $\hat{p}$ to be distributed nearly normally ($50 * 0.82 = 41 > 10; 50 * 0.18 = 9 < 10$)

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 $\hat{p}_{sim} \leq 0.82$.

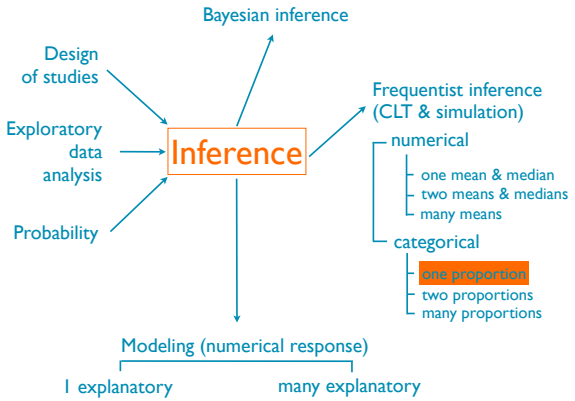
(b) Place 82 red and 18 blue chips in a bag. Sample, without replacement, 50 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.80$.

(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 $\hat{p}_{sim} \leq 0.80$.

(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 $\hat{p}_{sim} \leq 0.80$.

$n = 50$ and $\hat{p} = 0.80$. Hypotheses: $H_0 : p = 0.82; H_A : p < 0.82$. We use a randomization test because the sample size isn't large enough for $\hat{p}$ to be distributed nearly normally ($50 * 0.82 = 41 > 10; 50 * 0.18 = 9 < 10$)

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.



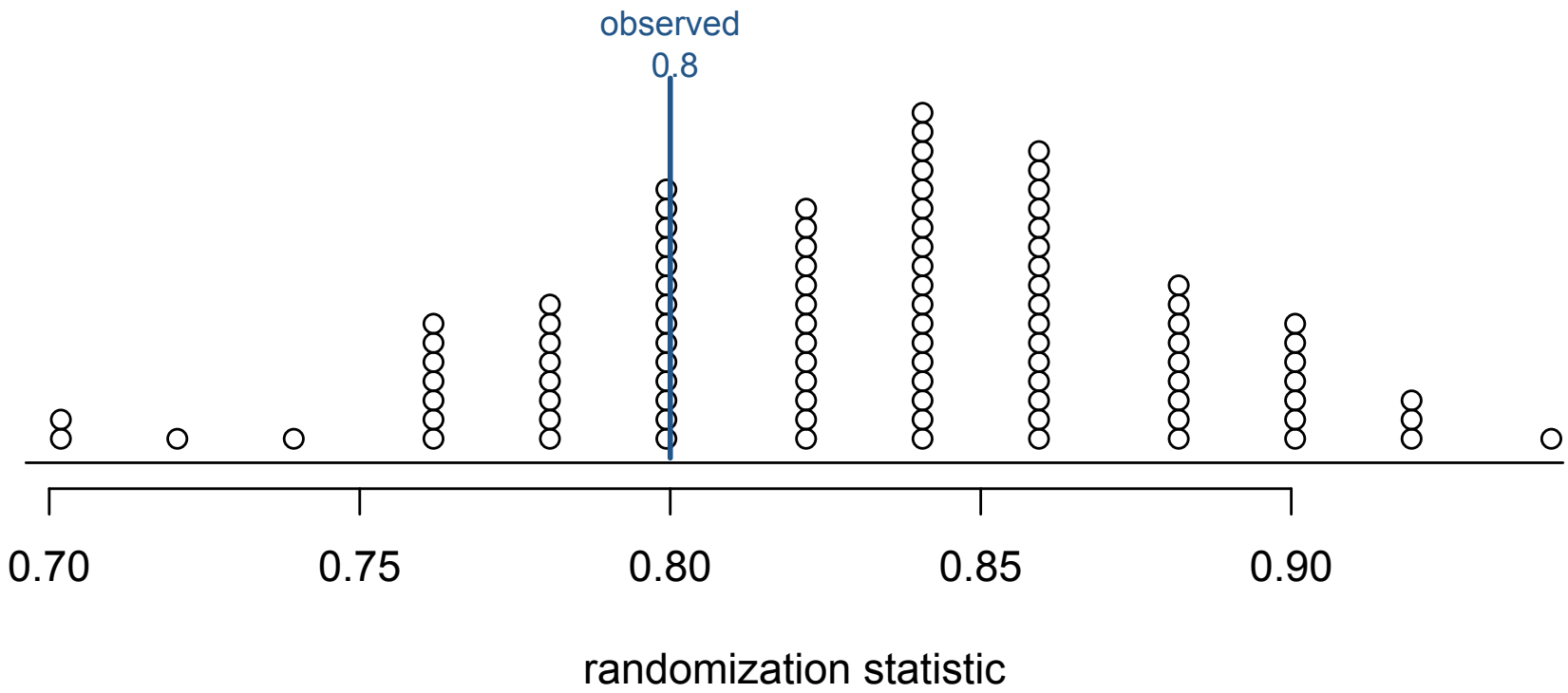
(a) Place 80 red and 20 blue chips in a bag. Sample, with replacement, 50 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.82$.

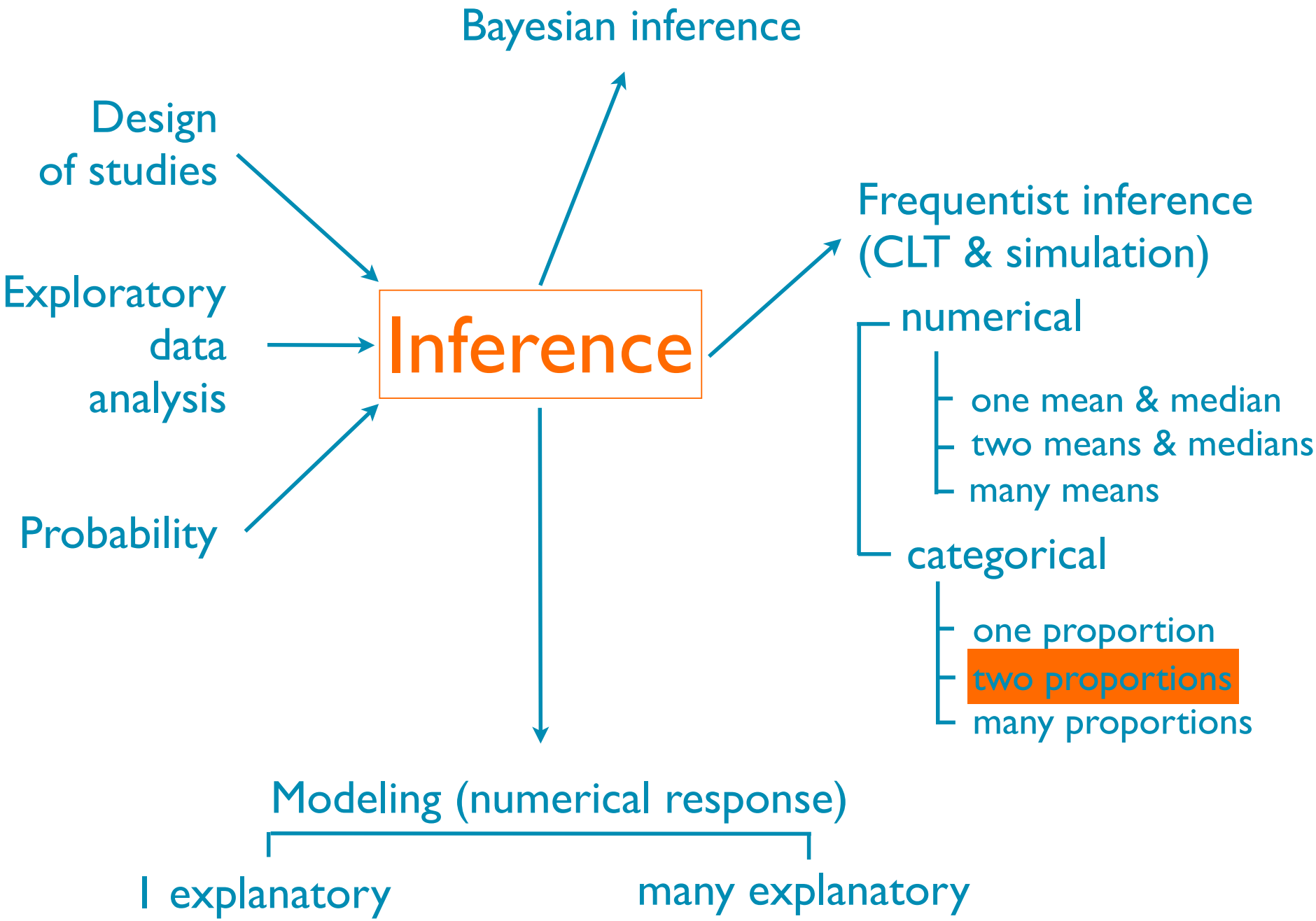
(b) Place 82 red and 18 blue chips in a bag. Sample, without replacement, 50 chips and calculate the proportion of reds. Repeat this many times and calculate the proportion of simulations where $\hat{p}_{sim} \leq 0.80$.

(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 $\hat{p}_{sim} \leq 0.80$.*

(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 $\hat{p}_{sim} \leq 0.80$.

What is the center of the randomization distribution? What is the result of the hypothesis test?




Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$

Comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- Conditions:

Comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- Conditions:
  - independence within groups
    - random sample and 10% condition met for both groups
  - independence between groups
  - at least 10 successes and failures in each group
    - if not $\rightarrow$ randomization (Section 6.4)

Comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- Conditions:
  - independence within groups
    - random sample and 10% condition met for both groups
  - independence between groups
  - at least 10 successes and failures in each group
    - if not $\rightarrow$ randomization (Section 6.4)
- $SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - for CI: use $\hat{p}_1$ and $\hat{p}_2$
  - for HT:
    - ★ when $H_0 : p_1 = p_2$: use $\hat{p}_{pool} = \frac{\#\ suc_1 + \#suc_2}{n_1 + n_2}$
    - ★ when $H_0 : p_1 - p_2 =$ *(some value other than 0)*: use $\hat{p}_1$ and $\hat{p}_2$
      - this is pretty rare

Reference - standard error calculations

	one sample	two samples
mean	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Reference - standard error calculations

	one sample	two samples
mean	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

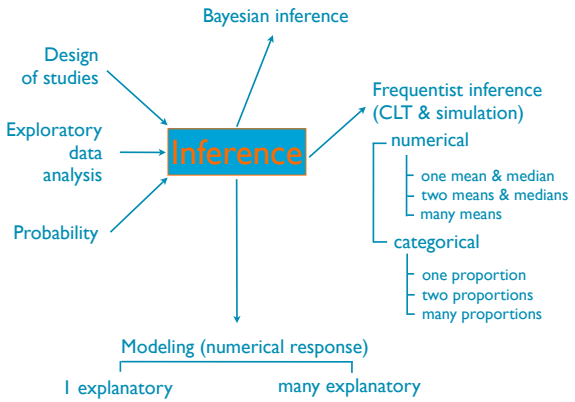
- When working with means, it's very rare that σ is known, so we usually use s.

Reference - standard error calculations

	one sample	two samples
mean	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

- When working with means, it's very rare that σ is known, so we usually use s.
- When working with proportions,
  - if doing a hypothesis test, p comes from the null hypothesis
  - if constructing a confidence interval, use $\hat{p}$ instead

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



(a) chi-square test of independence

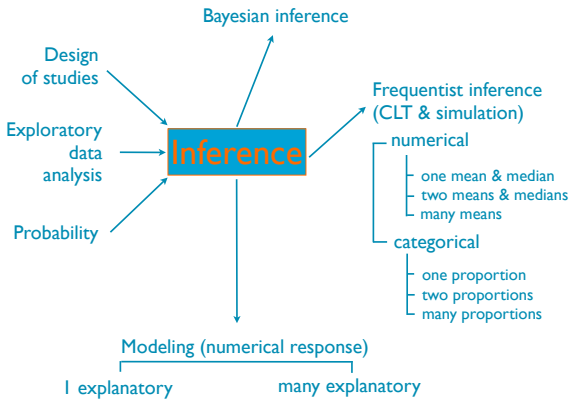
(b) chi-square test of goodness of fit

(c) anova

(d) linear regression

(e) t-test

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) linear regression

(e) t-test

Chi-square test

Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

Chi-square test

Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

H_0: There is no inconsistency between the observed and the expected counts.
The observed counts follow the same distribution as the expected counts.

Chi-square test

Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

H_0: There is no inconsistency between the observed and the expected counts. *The observed counts follow the same distribution as the expected counts.*

H_A: There is an inconsistency between the observed and the expected counts. *The observed counts do not follow the same distribution as the expected counts.*

χ^2 statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \qquad \text{where } k = \text{total number of cells}$$

When conducting a goodness of fit test to evaluate how well the observed data follow an expected distribution, the degrees of freedom are calculated as the number of cells (k) minus 1. $df = k - 1$

Conditions for the chi-square test

1. *Independence:* Each case that contributes a count to the table must be independent of all the other cases in the table.

Conditions for the chi-square test

1. *Independence:* Each case that contributes a count to the table must be independent of all the other cases in the table.
2. *Sample size:* Each particular scenario (i.e. cell) must have at least 5 *expected* cases.

Conditions for the chi-square test

1. *Independence:* Each case that contributes a count to the table must be independent of all the other cases in the table.
2. *Sample size:* Each particular scenario (i.e. cell) must have at least 5 *expected* cases.
3. *df > 1:* Degrees of freedom must be greater than 1.

Conditions for the chi-square test

1. *Independence:* Each case that contributes a count to the table must be independent of all the other cases in the table.
2. *Sample size:* Each particular scenario (i.e. cell) must have at least 5 *expected* cases.
3. *df > 1:* Degrees of freedom must be greater than 1.

Failing to check conditions may unintentionally affect the test's error rates.

Chi-square test of independence

The test statistic is

$$ts = \sum_{\textbf{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

The O_{ij} is the observed count for the cell in row i, column j.

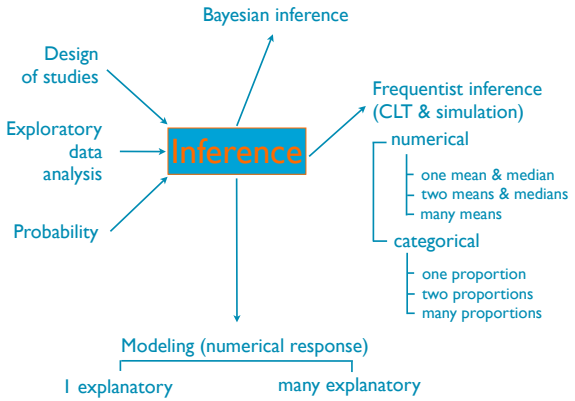
The E_{ij} uses the following formula:

$$E_{ij} = \frac{(\textbf{ith row sum}) * (\textbf{jth column sum})}{\textbf{total}}$$

Degrees of freedom equal to

$$\textbf{(number of rows - 1)} * \textbf{(number of columns - 1)}.$$

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

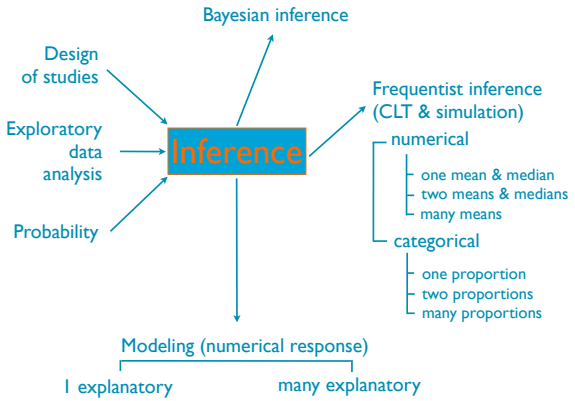
(b) chi-square test of goodness of fit

(c) anova

(d) linear regression

(e) t-test

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



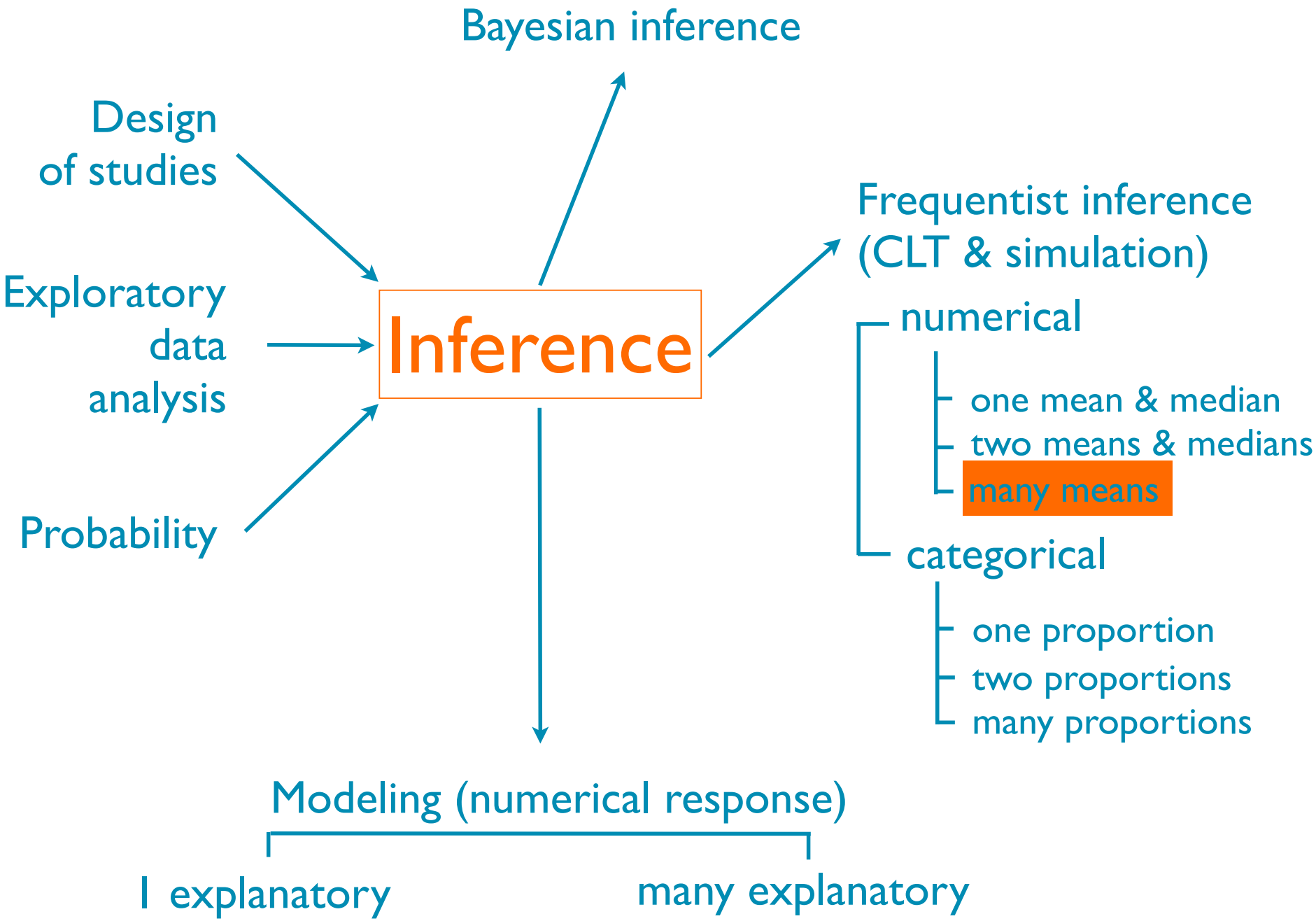
(a) z-test

(b) chi-square test of goodness of fit

(c) *anova*

(d) *linear regression*

(e) t-test


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable.

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable.

H_0 : The mean outcome is the same across all categories,

$$\mu_1 = \mu_2 = \cdots = \mu_k,$$

where μ_i represents the mean of the outcome for observations in category i.

H_A : At least one mean is different than others.

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied.
   - Carefully consider whether the data may be independent (e.g. no pairing).
   - Always important, but sometimes difficult to check.

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied.
   - Carefully consider whether the data may be independent (e.g. no pairing).
   - Always important, but sometimes difficult to check.
2. The observations within each group should be nearly normal.
   - Especially important when the sample sizes are small.

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied.
   - Carefully consider whether the data may be independent (e.g. no pairing).
   - Always important, but sometimes difficult to check.
2. The observations within each group should be nearly normal.
   - Especially important when the sample sizes are small.
3. The variability across the groups should be about equal.
   - Especially important when the sample sizes differ between groups.

z/*t* test vs. ANOVA - Purpose

z/*t test*

Compare means from *two* groups to see whether they are so far apart that the observed difference cannot reasonably be attributed to sampling variability.

$$H_0 : \mu_1 = \mu_2$$

ANOVA

Compare the means from *two or more* groups to see whether they are so far apart that the observed differences cannot all reasonably be attributed to sampling variability.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

z/t test vs. ANOVA - Method

z/t test

Compute a test statistic (a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE[\bar{x}_1 - \bar{x}_2]}$$

ANOVA

Compute a test statistic (a ratio).

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

z/t test vs. ANOVA - Method

z/t test

Compute a test statistic (a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE[\bar{x}_1 - \bar{x}_2]}$$

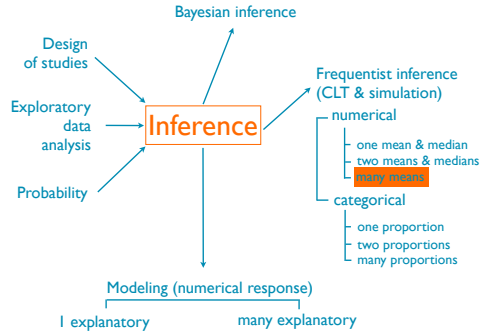
ANOVA

Compute a test statistic (a ratio).

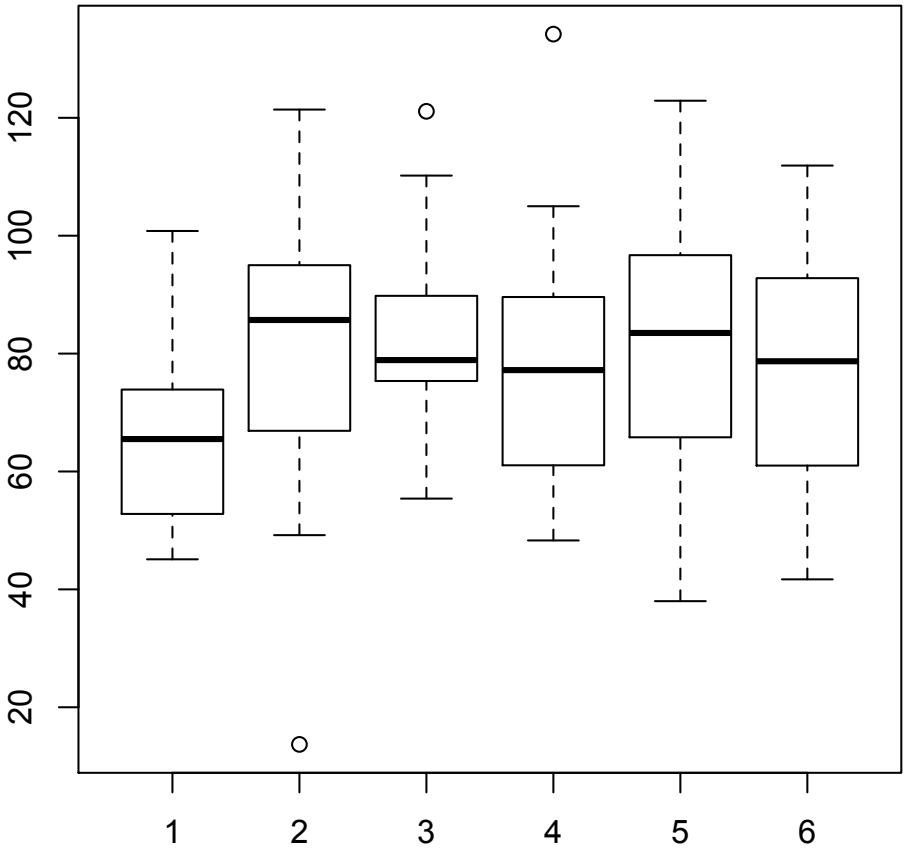
$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

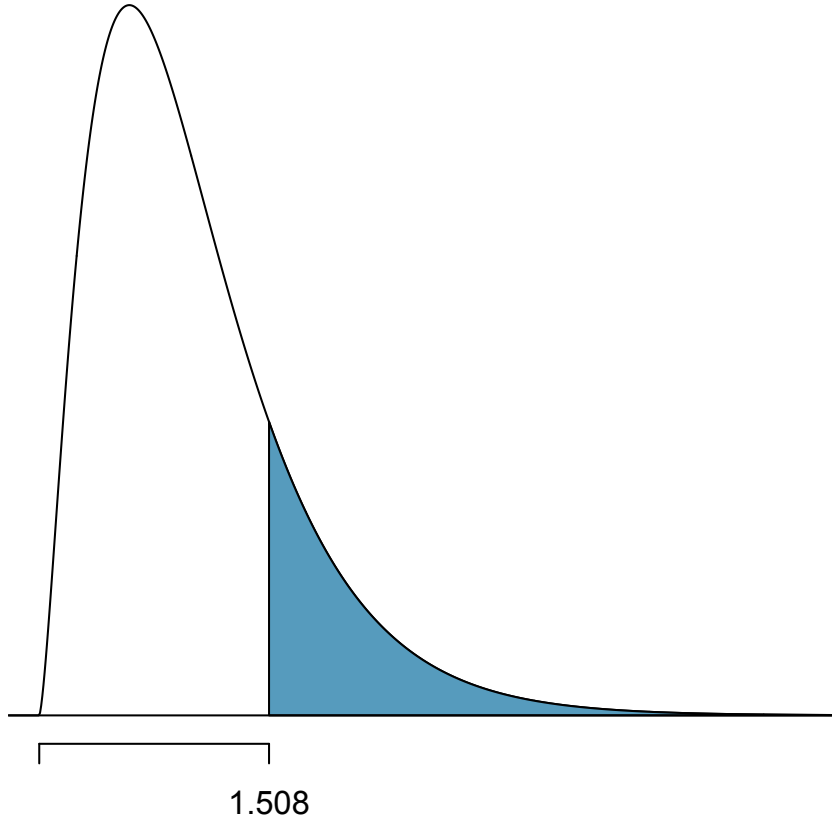
- Large test statistics lead to small p-values.
- If the p-value is small enough H_0 is rejected, we conclude that the population means are not equal.

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

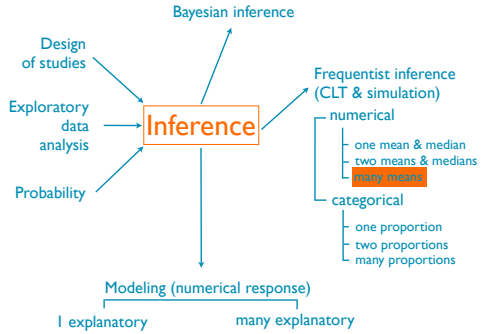



```
Response variable: numerical, Explanatory variable: categorical
ANOVA

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
```

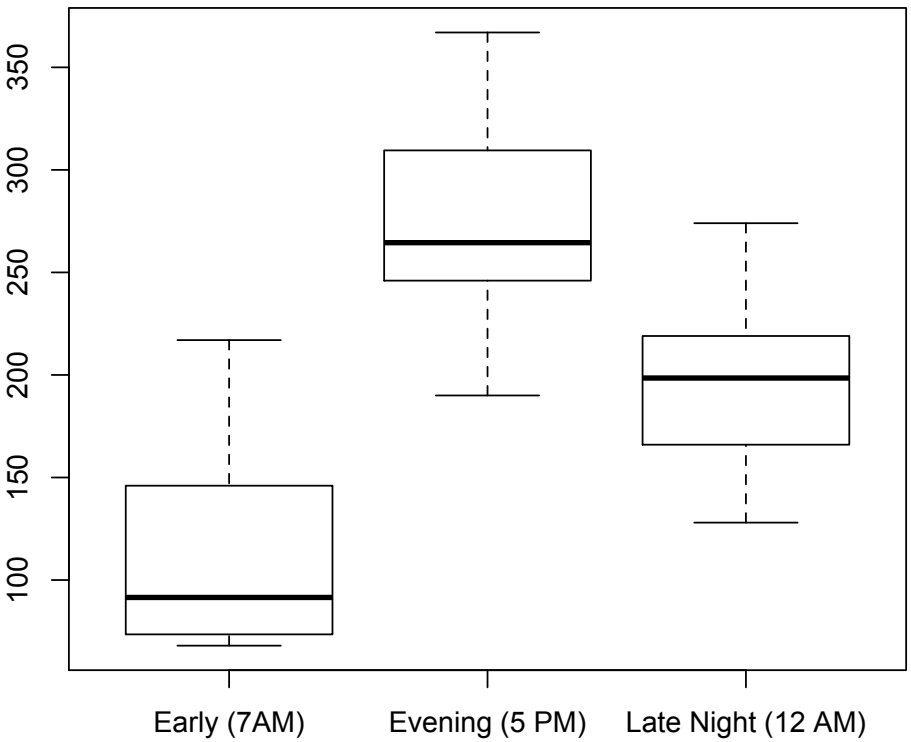

120
100
80
60
40
20
1
2
3
4
5
6

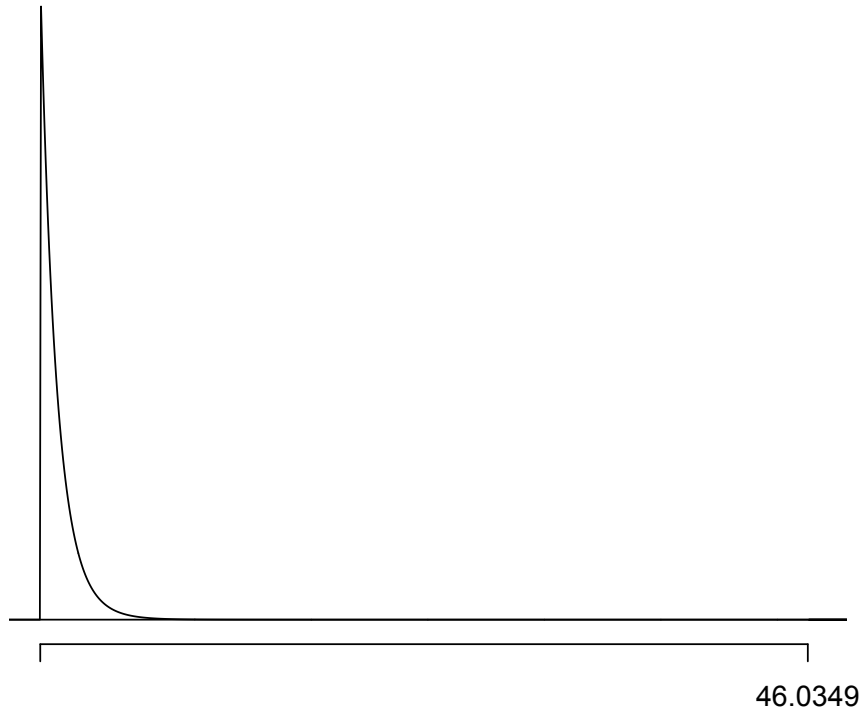

1.508

Data are collected on download times at three different times during the day. Do average download times vary by time of day?




```
Response variable: numerical, Explanatory variable: categorical
Summary statistics:
n_Early (7AM) = 16, mean_Early (7AM) = 113.375, sd_Early (7AM) = 47.6541
n_Evening (5 PM) = 16, mean_Evening (5 PM) = 273.3125, sd_Evening (5 PM) = 52.1929
n_Late Night (12 AM) = 16, mean_Late Night (12 AM) = 193.0625, sd_Late Night (12 AM)

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      2 204641  102320  46.035 1.306e-11
Residuals 45 100020    2223

Pairwise tests: t tests with pooled SD
                   Early (7AM) Evening (5 PM)
Evening (5 PM)               0             NA
Late Night (12 AM)           0              0
```





Use modified α, $\alpha^\star = \frac{0.05}{3}$, for pairwise tests.

Degrees of freedom associated with ANOVA

- groups: $df_G = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_G$

If the ANOVA assumption of equal variability across groups is satisfied, we can make the t-distribution approach slightly more precise by using a *pooled standard deviation*:

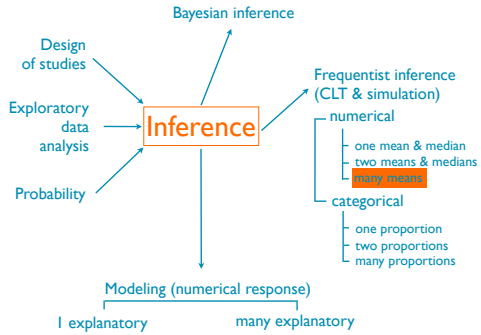
- The pooled standard deviation is a way to **use data from all groups** to better estimate the standard deviation from each group
- By pooling all the data, we can use a **larger degree of freedom** for the t-distribution
- Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ if the standard deviations of the groups are equal

Pooled standard deviation estimate from ANOVA

- The standard deviation of each group is estimated as $s_{pooled} = \sqrt{MSE}$
- Use the error degrees of freedom, $n_1 + n_2 - k$, for t-distributions
- The standard error of test statistic

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

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
```

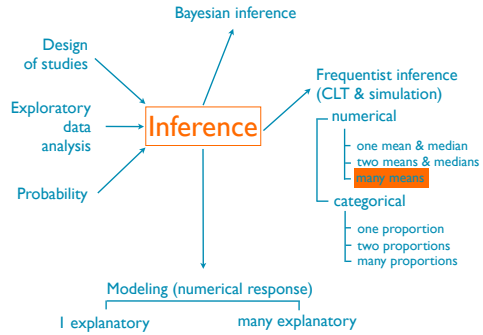
(a) $\frac{204641}{204641+100020}$

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

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

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

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



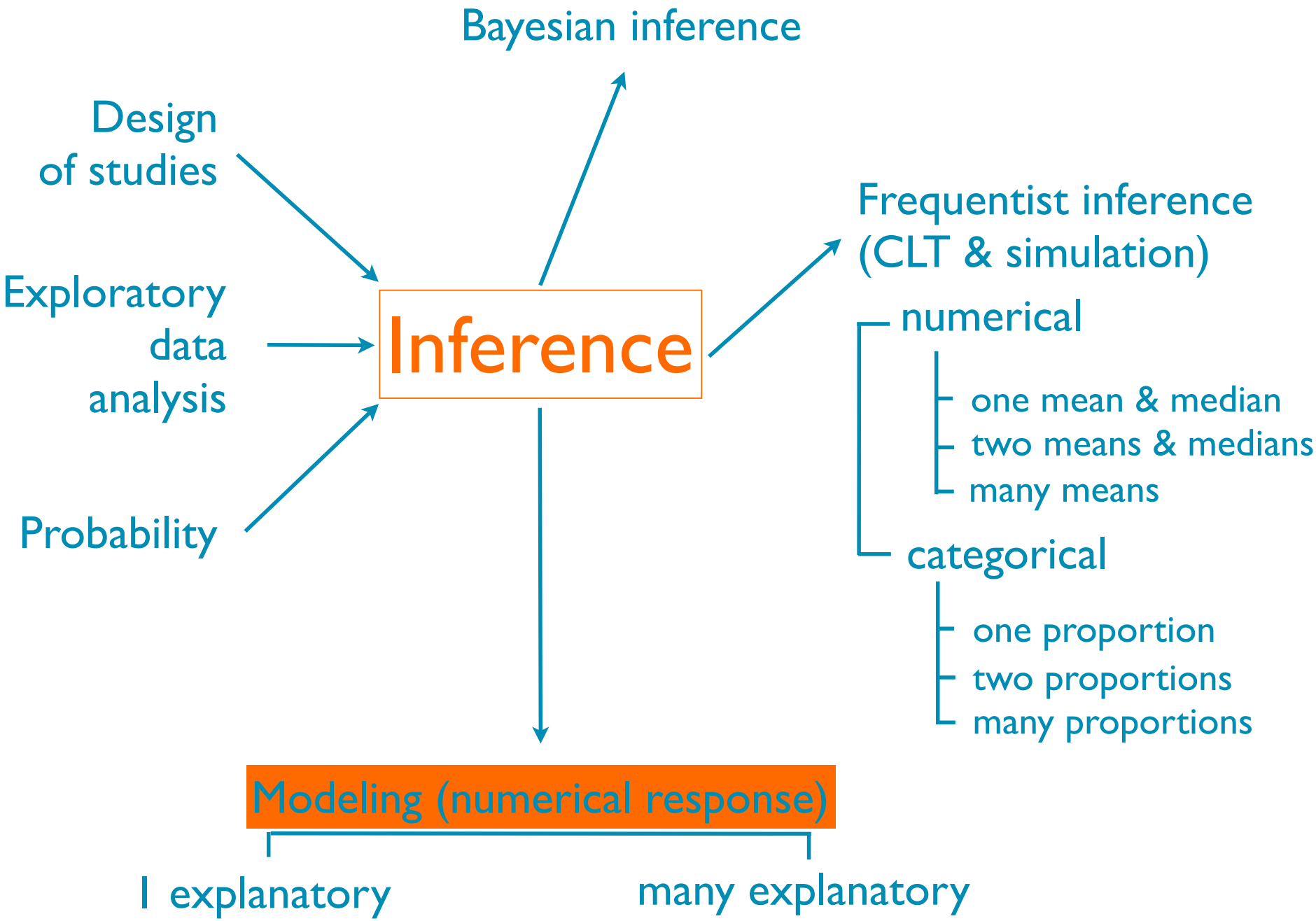

```
Response: data
          Df Sum Sq Mean Sq F value    Pr(>F)
group      2 204641  102320  46.035 1.306e-11
Residuals 45 100020    2223
```

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

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

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

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


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Sample Correlation

To estimate the true correlation coefficient, define

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

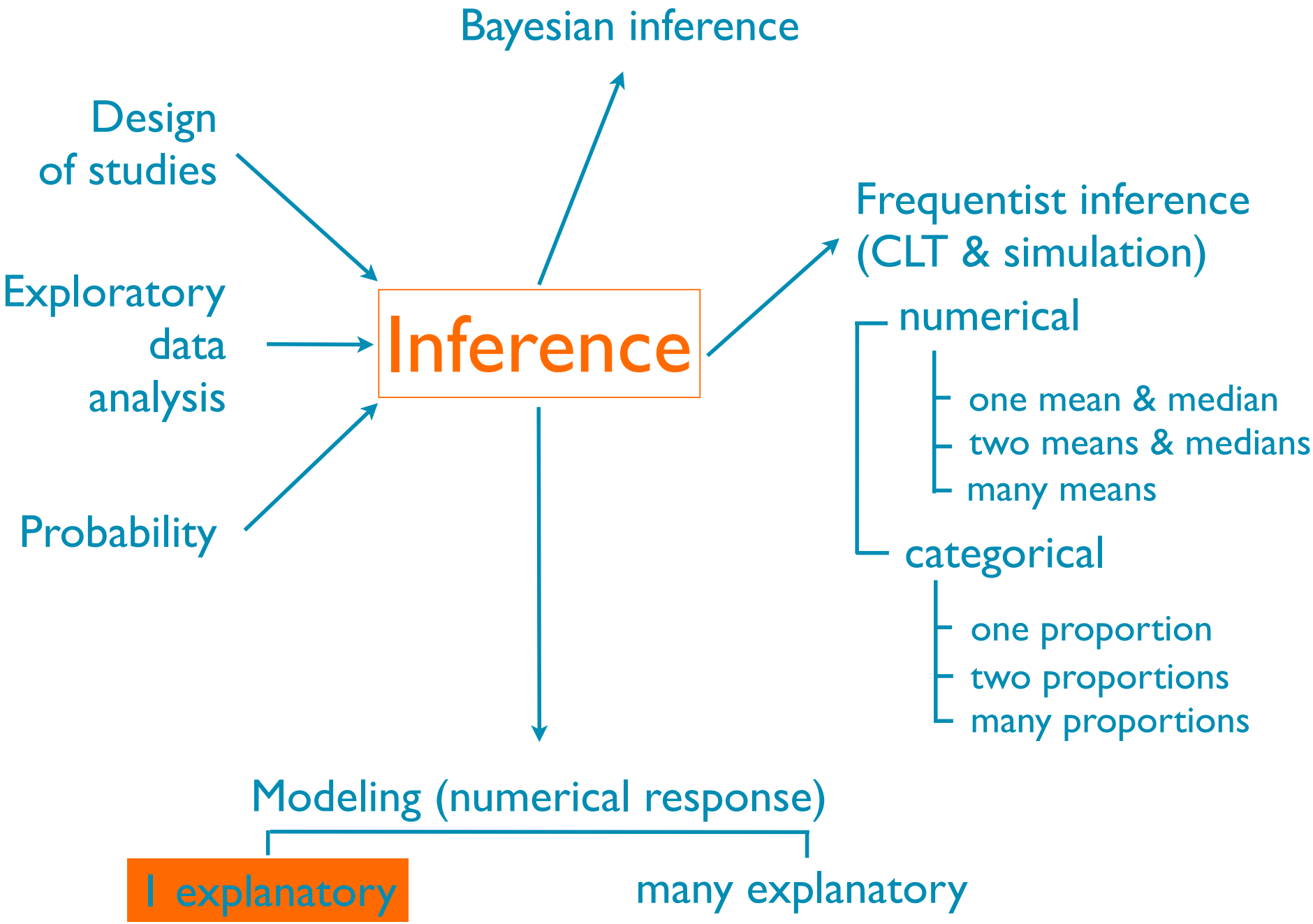
$$SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

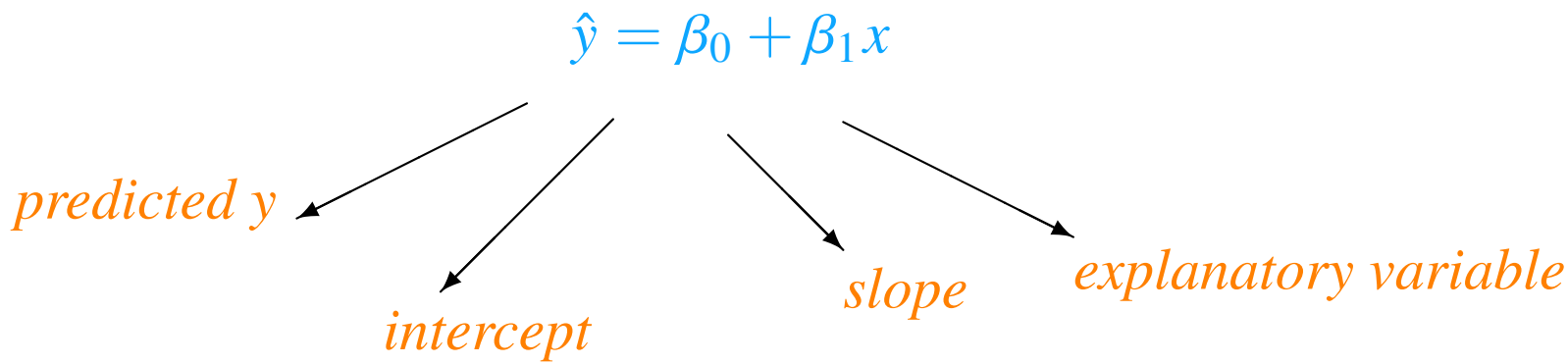
Note: observe that if divided by $n-1$, these are the sample versions of the variances and the covariance. So there's no need to memorize.

Then the sample correlation is

$$r = \frac{\hat{Cov}(x,y)}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{SS_{xy}/n-1}{\sqrt{(SS_{xx}/n-1)(SS_{yy}/n-1)}} = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}.$$


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

The least squares line



Notation:

- Intercept:
  - Parameter: β_0
  - Point estimate: $\hat{\beta_0} = b_0$
- Slope:
  - Parameter: β_1
  - Point estimate: $\hat{\beta_1} = b_1$

Method of Least Squares

Thus, to find the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the coefficients in the equation of a line, we need to get the values that minimize the sum of the squared vertical distances. The sum of the squared vertical distances is

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

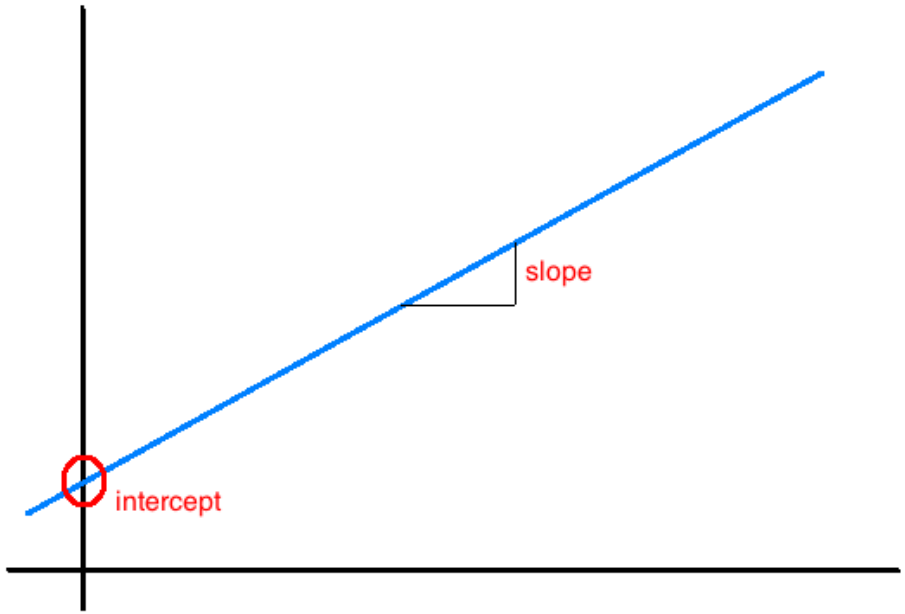
So take the derivative of $f(\beta_0, \beta_1)$ with respect to β_0 and β_1, set these equal to zero, and solve. One finds that:

$$b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}; \quad \textbf{and} \quad b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}} = r\frac{s_y}{s_x}$$

Thus, the line defined by $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is the least squares line.

Interpretation of slope and intercept

- *Intercept:* When $x = 0$, y is expected to equal the intercept.
- *Slope:* For each unit in x, y is expected to increase / decrease on average by the slope.



Note: *These statements are not causal, unless the study is a randomized controlled experiment.*

Conditions for the least squares line

When fitting a least square line, we generally require

1. **Linearity:** The data should show a linear trend

Conditions for the least squares line

When fitting a least square line, we generally require

1. **Linearity:** The data should show a linear trend
2. **Nearly normal residuals:** the residuals must be nearly normally distributed

Conditions for the least squares line

When fitting a least square line, we generally require

1. **Linearity:** The data should show a linear trend
2. **Nearly normal residuals:** the residuals must be nearly normally distributed
3. **Constant variability:** the variability of points around the least squares line remains roughly constant
4. **Independence observations:** depends on data collection method, often violated for time-series data. (We suspect that order of data collection may influence the outcome.)

The variability in response

- Total sum of squares

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Regression sum of squares

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Residual sum of squares

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Coefficient of determination

$$r^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

Confidence interval for average values

A confidence interval for the average (expected) value of y, $E(y)$, evaluated at given $x^\star$ is,

$$\hat{y} \pm t^\star_{n-2} s \sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$

where s is standard deviation of the residuals, calculated as

$$\frac{1}{n-2} \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

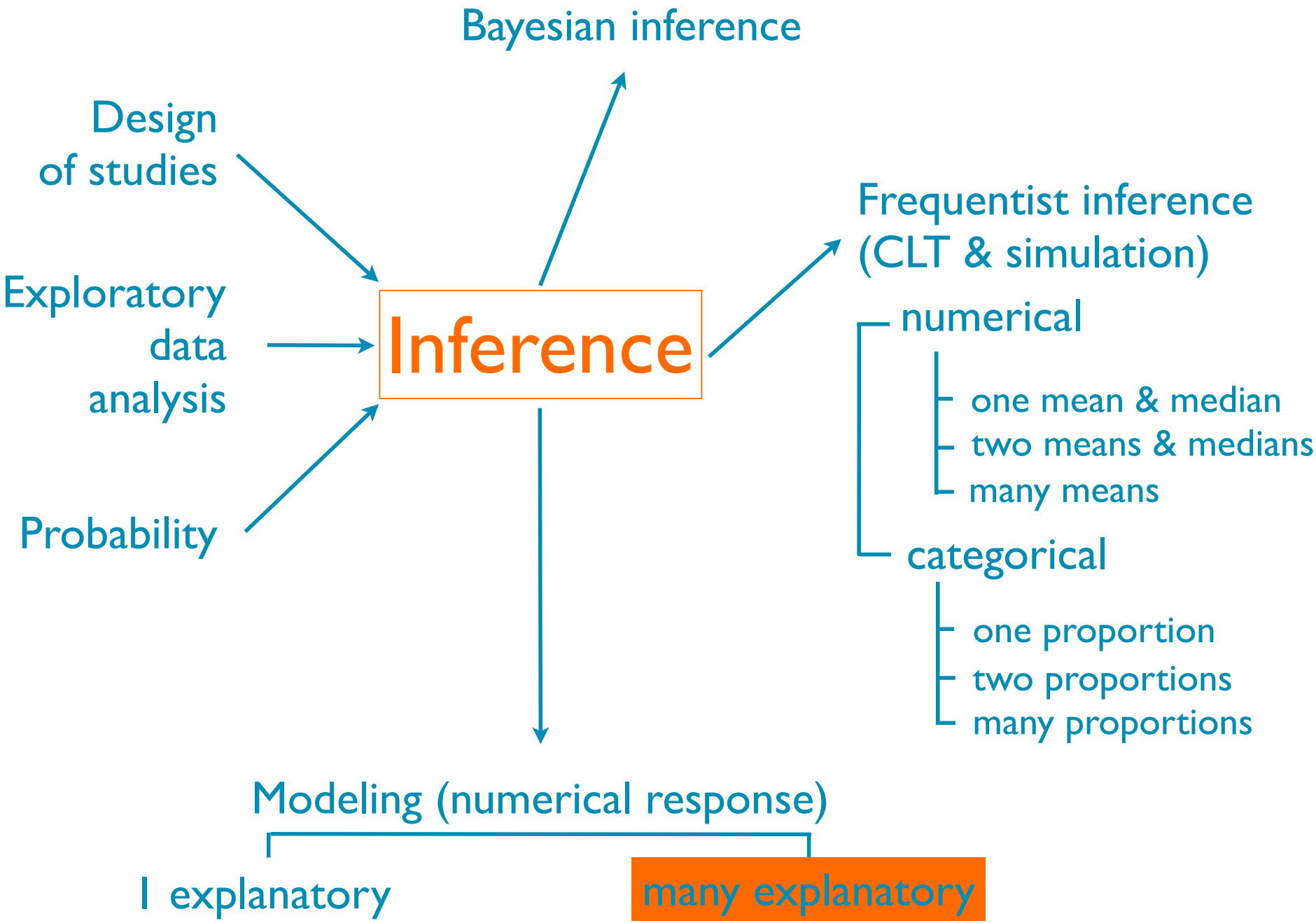
Prediction interval for a future value

A **prediction interval** for y for a given $x^\star$ is

$$\hat{y} \pm t^\star_{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$

where s is the standard deviation of the residuals.

- The formula is very similar, except the variability is higher since there is an added 1 in the formula
- Predicted level: if we repeat the study of obtaining a regression data set many times, each time forming a $XX\%$ prediction interval at $x^\star$, and wait to see what the future value of y is at $x^\star$, then roughly $XX\%$ of the prediction intervals will contain the corresponding actual value of y


Bayesian inference
Design of studies
Exploratory data analysis
Probability
Inference
Frequentist inference (CLT & simulation)
numerical
one mean & median
two means & medians
many means
categorical
one proportion
two proportions
many proportions
Modeling (numerical response)
1 explanatory
many explanatory

Multiple Linear Regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

The model depends on the following conditions

1. Residuals are nearly normal (primary concern relates to residuals that are outliers)
2. Residuals have constant variability
3. Independence of observations (and hence residuals)
4. Each variable is linearly related to the outcome
5. Also important to make sure that your explanatory variables are not collinear

We often use graphical methods to check the validity of these conditions, which we will go through in detail in the following slides.