SSNAP - Statistical Analysis

Part II - Statistical Inference

Duke UNIVERSITY

Dr. Mine Çetinkaya-Rundel
Duke University

slides at **bit.ly/ssnap-2022**

frequentist inference

example: Paul the octopus

- Prediction of 2010 World Cup winners:
  - Presented with 2 clear plastic boxes, each containing food and marked with flag of a team.
  - Winner: Box which Paul opened first to eat its contents.
- Accurately predicted the outcome of 8 games!

https://www.youtube.com/watch?v=Ya85knuDzp8

example: Paul the octopus

Paul the Octopus predicted 8 World Cup games, and predicted them all correctly.

Does this provide convincing evidence that Paul actually has psychic powers, i.e. that he does better than just randomly guessing?

two competing claims

null hypothesis

"There is nothing going on"

alternative hypothesis

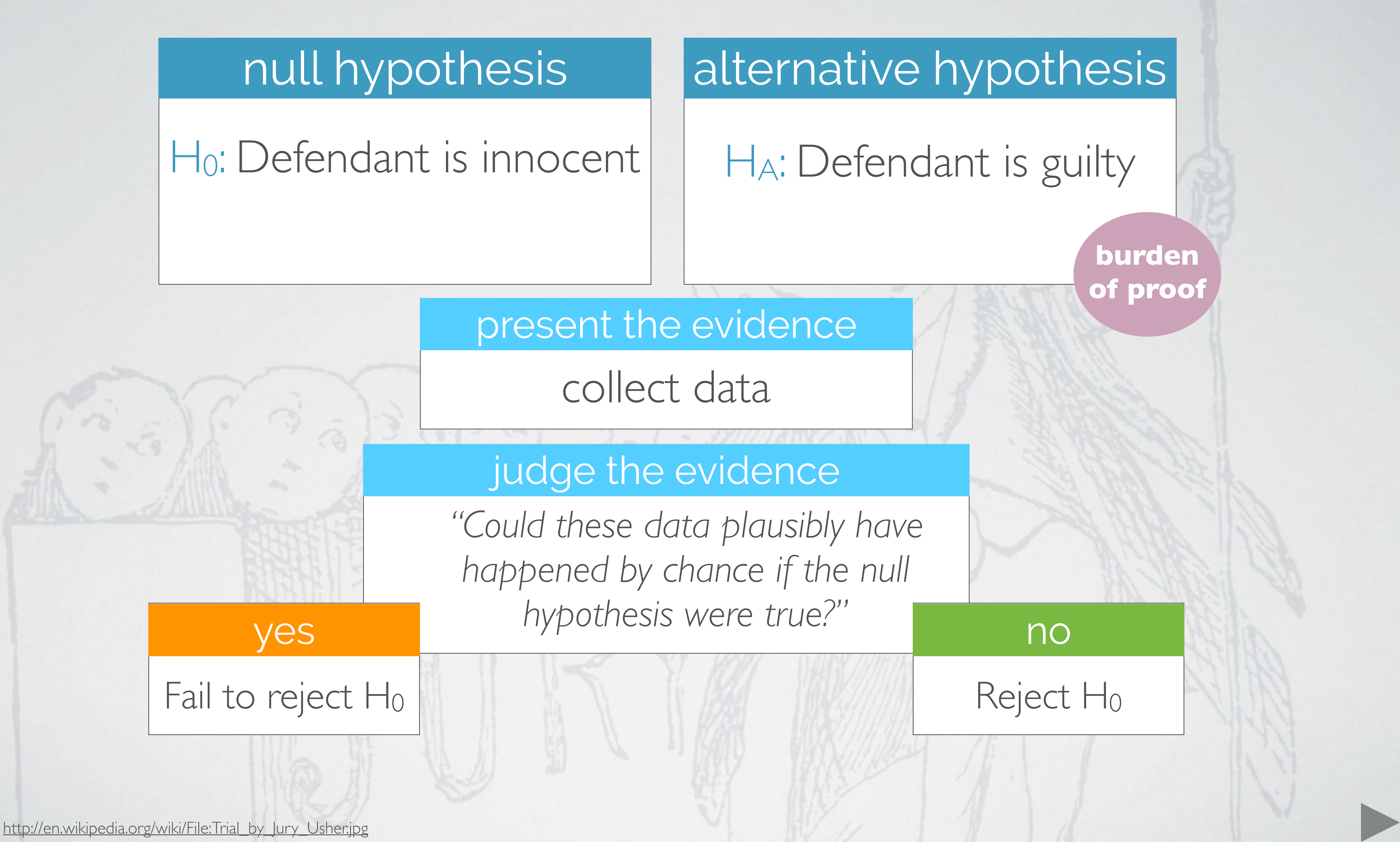
"There is something going on"

In context of Paul's predictions, which of the following does the null hypothesis of "there is nothing going on" maps to?

a. Paul does no better than random guessing.
b. Paul does better than random guessing.
c. Paul predicts all games correctly.
d. Paul predicts none of the games correctly.
e. Paul predicts 50% of the games correctly.

In context of Paul's predictions, which of the following does the null hypothesis of "there is nothing going on" maps to?

a. **Paul does no better than random guessing.**
b. Paul does better than random guessing.
c. Paul predicts all games correctly.
d. Paul predicts none of the games correctly.
e. Paul predicts 50% of the games correctly.



null hypothesis

H_0: Defendant is innocent

alternative hypothesis

H_A: Defendant is guilty

burden of proof

present the evidence

collect data

judge the evidence

"Could these data plausibly have happened by chance if the null hypothesis were true?"

yes

Fail to reject H_0

no

Reject H_0

Image source: http://en.wikipedia.org/wiki/File:Trial_by_Jury_Usher.jpg

hypothesis testing framework

Which of the following is not a component of the hypothesis testing framework?

a. Start with a null hypothesis that represents the status quo
b. Set an alternative hypothesis that represents the research question, i.e. what we're testing for
c. Conduct a hypothesis test under the assumption that the altertnative hypothesis is true
d. If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, stick with the null hypothesis
e. If the test results suggest that the data do provide convincing evidence for the alternative hypothesis, then reject the null hypothesis in favor of the alternative

Which of the following is not a component of the hypothesis testing framework?

a. Start with a null hypothesis that represents the status quo
b. Set an alternative hypothesis that represents the research question, i.e. what we're testing for
c. **Conduct a hypothesis test under the assumption that the altertnative hypothesis is true**
d. If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, stick with the null hypothesis
e. If the test results suggest that the data do provide convincing evidence for the alternative hypothesis, then reject the null hypothesis in favor of the alternative

Which of the following is the best set of hypotheses associated with the following two claims: "Paul does no better than random guessing" and "Paul does better than random guessing"?

a. $H_0: p = 0$; $H_A: p > 0$
b. $H_0: p = 1/8$; $H_A: p > 1/8$
c. $H_0: p < 0.5$; $H_A: p = 0.5$
d. $H_0: p = 0.5$; $H_A: p > 0.5$
e. $H_0: p = 0.5$; $H_A: p = 1$

Which of the following is the best set of hypotheses associated with the following two claims: "Paul does no better than random guessing" and "Paul does better than random guessing"?

a. $H_0: p = 0$; $H_A: p > 0$
b. $H_0: p = 1/8$; $H_A: p > 1/8$
c. $H_0: p < 0.5$; $H_A: p = 0.5$
d. $H_0: p = 0.5$; $H_A: p > 0.5$
e. $H_0: p = 0.5$; $H_A: p = 1$

two competing claims

null hypothesis	alternative hypothesis
"There is nothing going on"	"There is something going on"
Paul does no better than random guessing.	Paul does better than random guessing.
$H_0: p = 0.5$	$H_A: p > 0.5$

example: Paul the octopus

Paul the Octopus predicted 8 World Cup games, and predicted them all correctly. Does this provide convincing evidence that Paul actually has psychic powers, i.e. that he does better than just randomly guessing?

- Use a fair coin, and label head as success (correct guess)
- One simulation: flip the coin 8 times and record the proportion of heads (correct guesses)
- Repeat the simulation many times, recording the proportion of heads at each iteration
- Calculate the percentage of simulations where the simulated proportion of heads is at least as extreme as the observed proportion

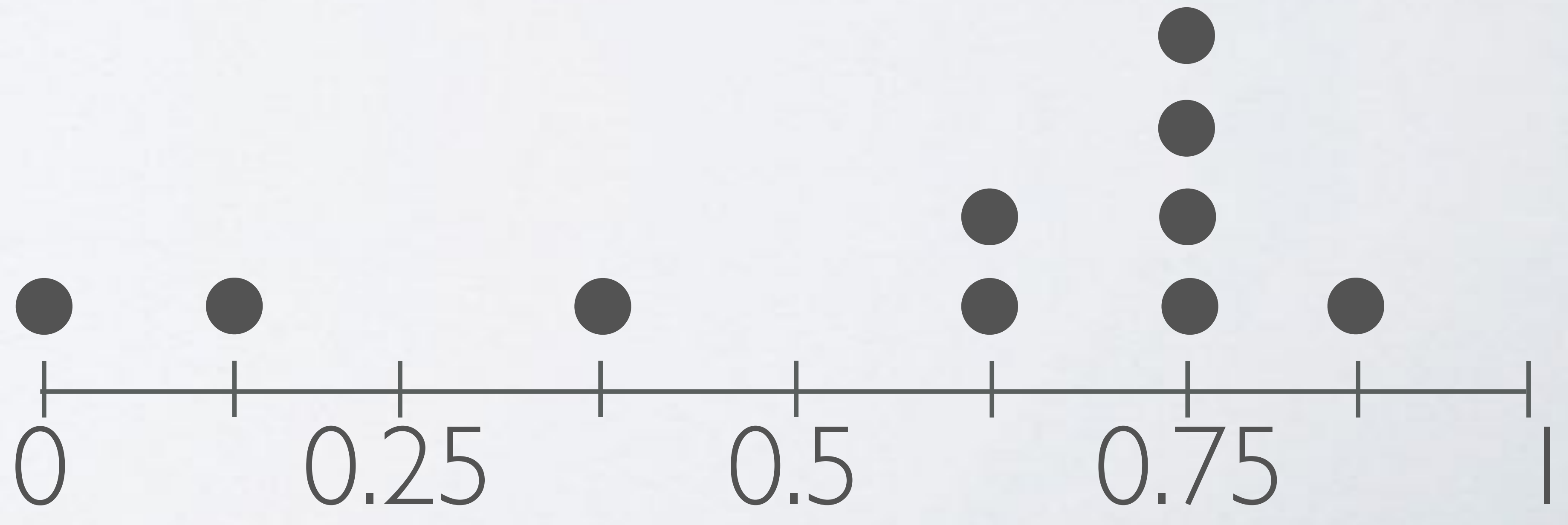
$H_0: p = 0.5$

$H_A: p > 0.5$

simulating Paul

									$\hat{p}$
simulation 1:	H	H	H	T	H	H	H	H	7 / 8 = 0.875
simulation 2:	T	H	H	T	H	T	T	T	3 / 8 = 0.375
simulation 3:	T	T	H	H	H	H	T	H	5 / 8 = 0.625
…									…
simulation 10:	T	H	T	H	H	H	H	H	6 / 8 = 0.75

What proportion of simulations yielded a proportion of success at least as extreme as Paul's?



conclusion of the test

Based on the probability that you just calculated, which of the following is the best conclusion of this hypothesis test?

a. It is **likely** to predict 8 or more games correctly if randomly guessing, hence the data suggest that Paul is doing **no better** than randomly guessing.
b. It is **likely** to predict 8 or more games correctly if randomly guessing, hence the data suggest that Paul is doing **better** than randomly guessing.
c. It is **very unlikely** to predict 8 or more games correctly if randomly guessing, hence the data suggest that Paul is doing **no better** than randomly guessing.
d. It is **very unlikely** to predict 8 or more games correctly if randomly guessing, hence the data suggest that Paul is doing **better** than randomly guessing.
e. None of the above.

Based on the probability that you just calculated, which of the following is the best conclusion of this hypothesis test?

a. It is **likely** to predict 8 or more games correctly if randomly guessing, hence the data suggest that Paul is doing **no better** than randomly guessing.
b. It is **likely** to predict 8 or more games correctly if randomly guessing, hence the data suggest that Paul is doing **better** than randomly guessing.
c. It is **very unlikely** to predict 8 or more games correctly if randomly guessing, hence the data suggest that Paul is doing **no better** than randomly guessing.
d. **It is** very unlikely **to predict 8 or more games correctly if randomly guessing, hence the data suggest that Paul is doing** better **than randomly guessing.**
e. None of the above.

making a decision

- **Hypotheses:**
  - $H_0: p = 0.5$ - Paul does no better than random guessing
  - $H_A: p > 0.5$ - Paul does better than random guessing

- **Data:** Paul predicted 8 out of 8 games correctly

- **Results:** Assuming H_0 is true, the probability of obtaining results at least as extreme as Paul's is almost 0.

- **Decision:** Since this probability is low (lower than 5%), we reject H_0 in favor of H_A.
  - This doesn't mean we proved the alternative hypothesis, just that the data provide convincing evidence for it.

example: sex discrimination

"Are individuals who identify as female discriminated against in promotion decisions made by their managers who identify as male?"

- study considered sex roles, and only allowed for options of "male" and "female." We should note that the identities being considered are not gender identities and that the study allowed only for a binary classification of sex.
- 48 male bank supervisors given the same personnel file, asked to judge whether the person should be promoted
- identical files, except that half of them indicated the candidate identified as male and the other half indicated the candidate identified as female
- files randomly assigned to managers
- 35 / 48 promoted
- are females are unfairly discriminated against?

example: sex discrimination

		promotion		
		promoted	not promoted	total
sex	male	21	3	24
	female	14	10	24
	total	35	13	48

% of males promoted = 21/24 ≈ 88%
% of females promoted = 14/24 ≈ 58%

two competing claims

null hypothesis	alternative hypothesis
"There is nothing going on"	"There is something going on"
promotion and gender are independent, no gender discrimination, observed difference in proportions is simply due to chance	promotion and gender are dependent, there is gender discrimination, observed difference in proportions is not due to chance.

simulation scheme

[use a deck of playing cards to simulate this experiment]

1. face card: not promoted, non-face card: promoted
   - set aside the jokers, consider aces as face cards
   - take out 3 aces → 13 face cards left in the deck (face cards: A, K, Q, J)
   - take out a number card → 35 number (non-face) cards left in the deck (number cards: 2-10)

Step 1:

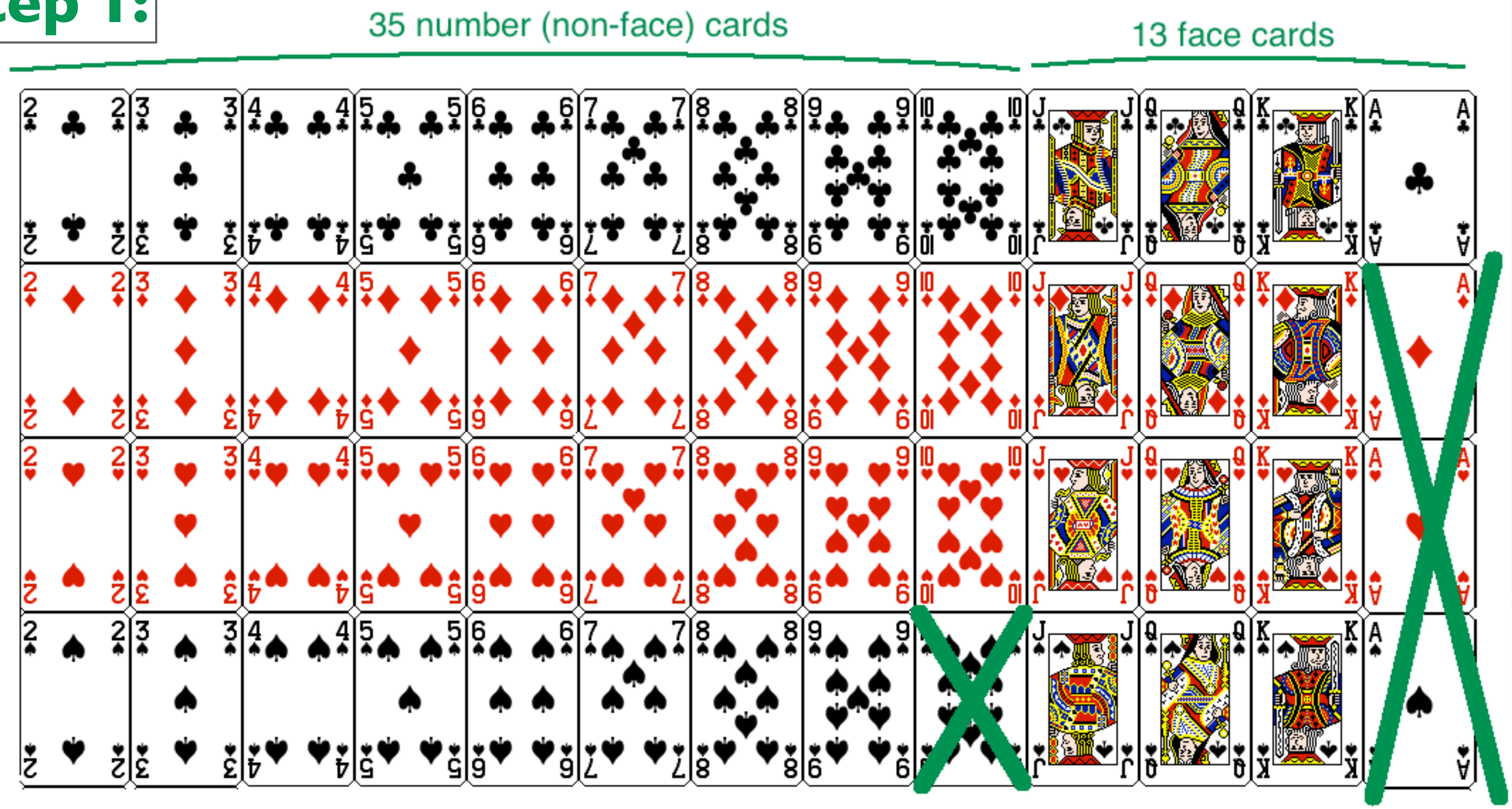


Image source: http://www.jfitz.com/cards/

simulation scheme

[use a deck of playing cards to simulate this experiment]

1. face card: not promoted, non-face card: promoted
   - set aside the jokers, consider aces as face cards
   - take out 3 aces → 13 face cards left in the deck (face cards: A, K, Q, J)
   - take out a number card → 35 number (non-face) cards left in the deck (number cards: 2-10)
2. shuffle the cards, deal into two groups of size 24, representing males and females

Step 2:

Shuffle and split into two groups of 24

(males and females)

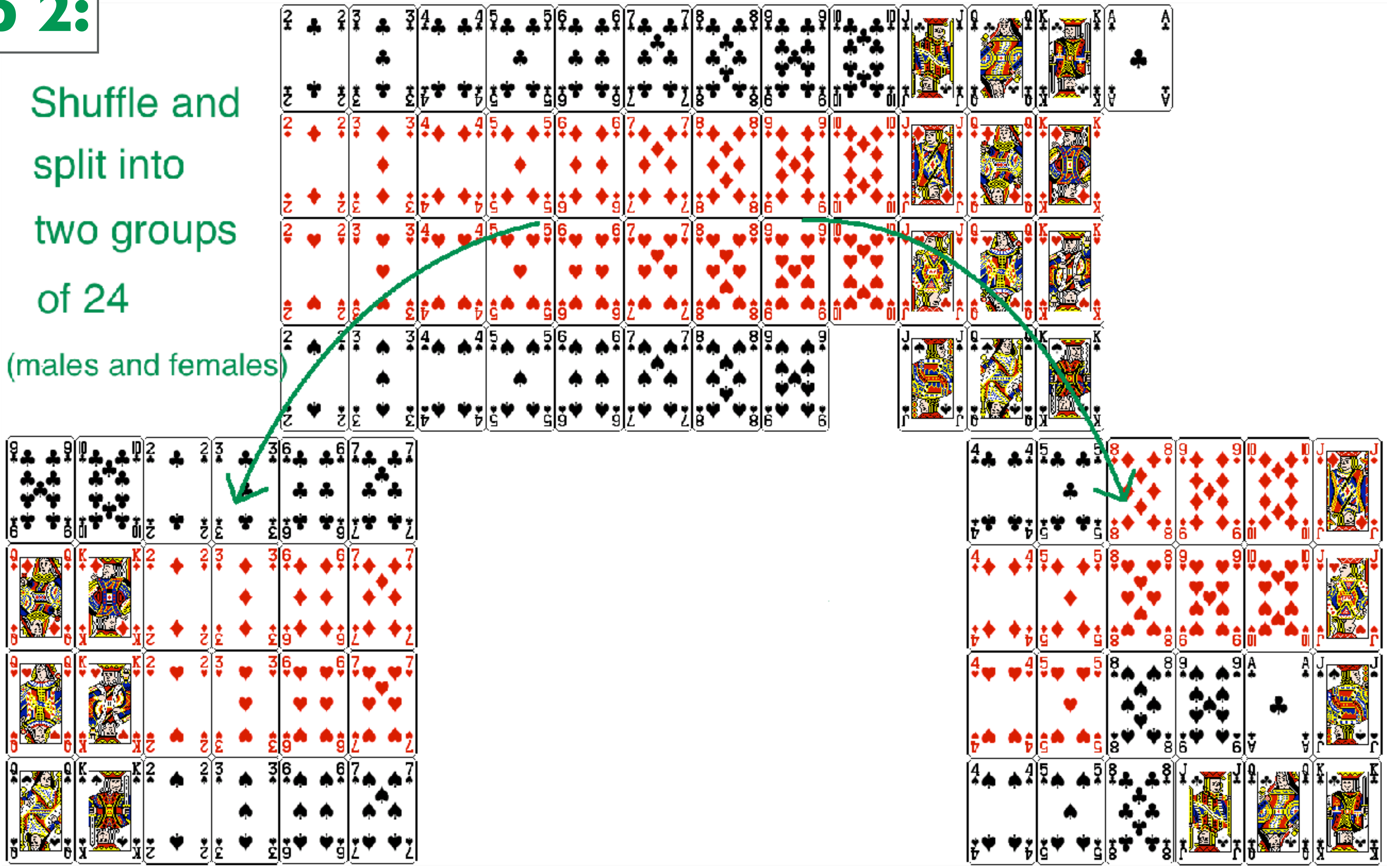


Image source: http://www.jfitz.com/cards/

simulation scheme

[use a deck of playing cards to simulate this experiment]

1. face card: not promoted, non-face card: promoted
   - set aside the jokers, consider aces as face cards
   - take out 3 aces → 13 face cards left in the deck (face cards: A, K, Q, J)
   - take out a number card → 35 number (non-face) cards left in the deck (number cards: 2-10)
2. shuffle the cards, deal into two groups of size 24, representing males and females
3. count how many number cards are in each group (representing promoted files)
4. calculate the proportion of promoted files in each group, take the difference (male - female), and record this value

Steps 3&4:

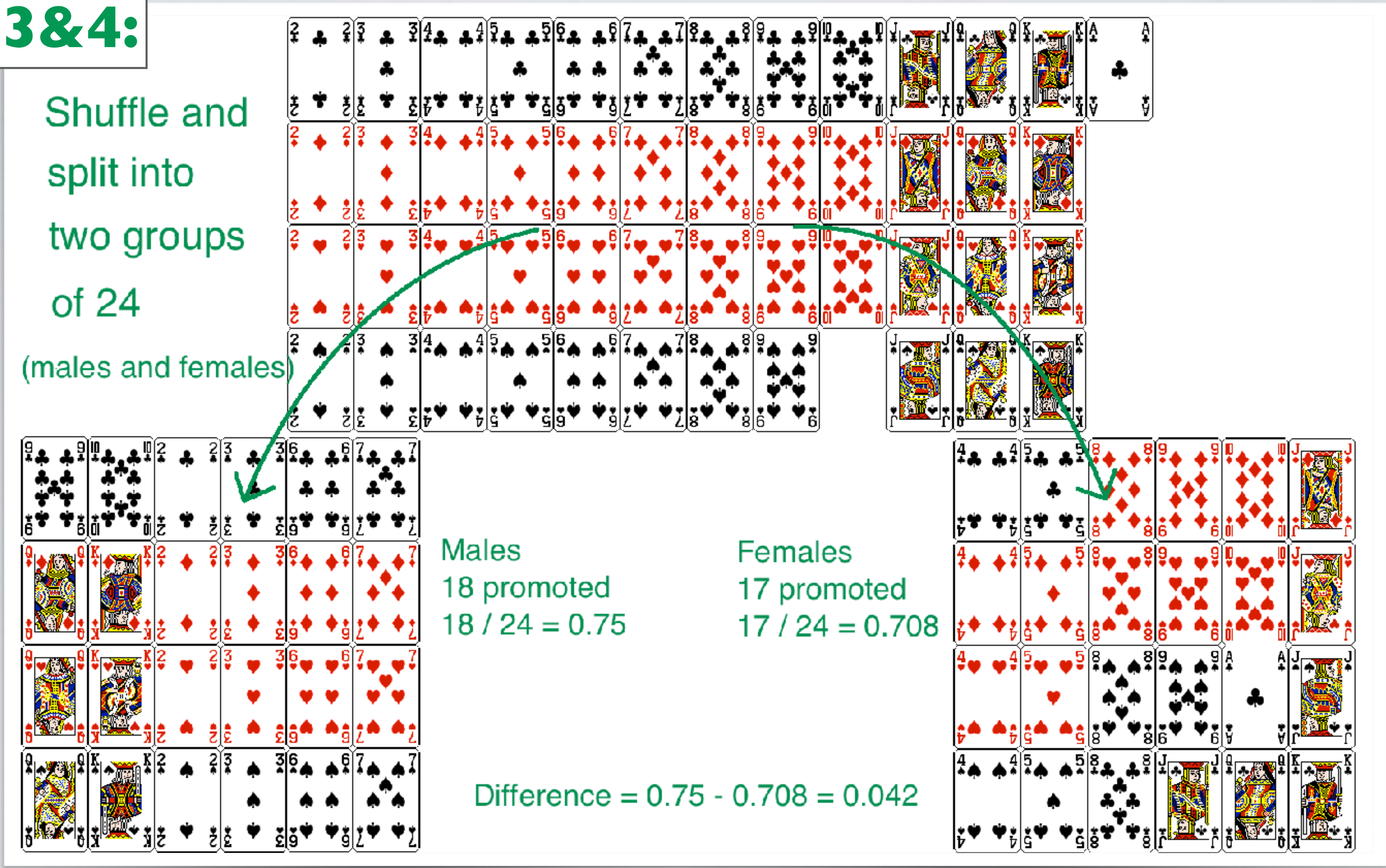
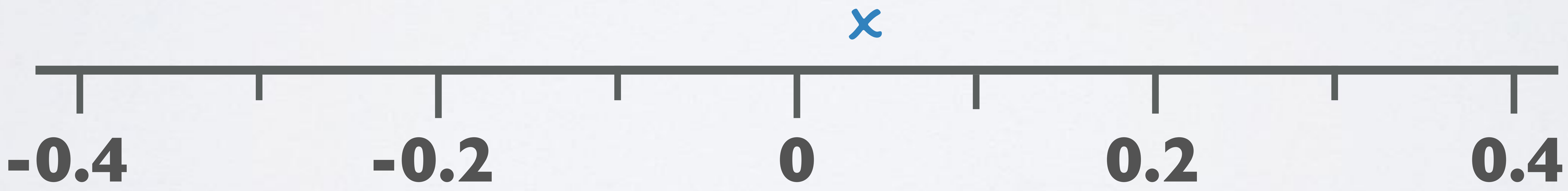


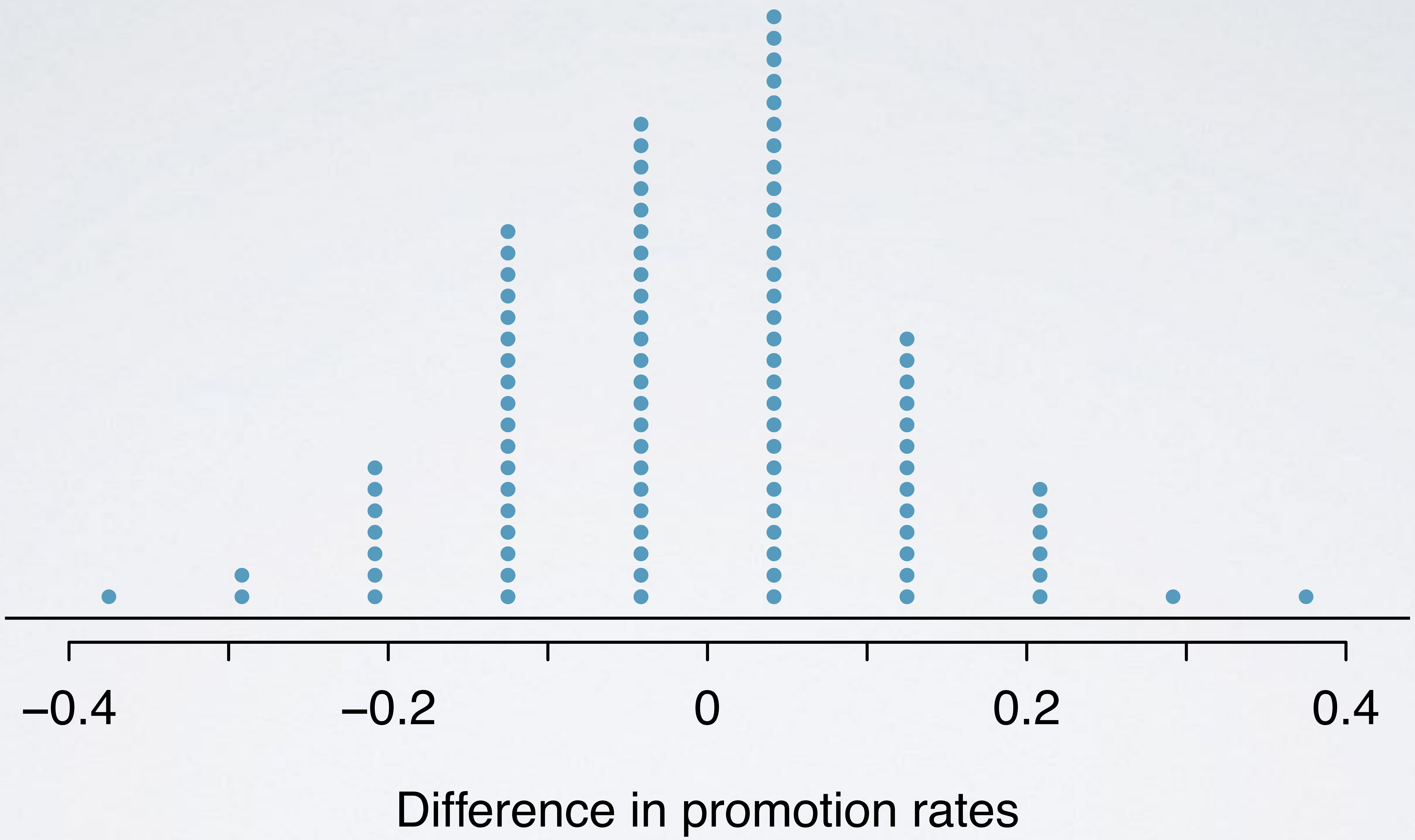
Image source: http://www.jfitz.com/cards/


x
-0.4
-0.2
0
0.2
0.4

simulation scheme

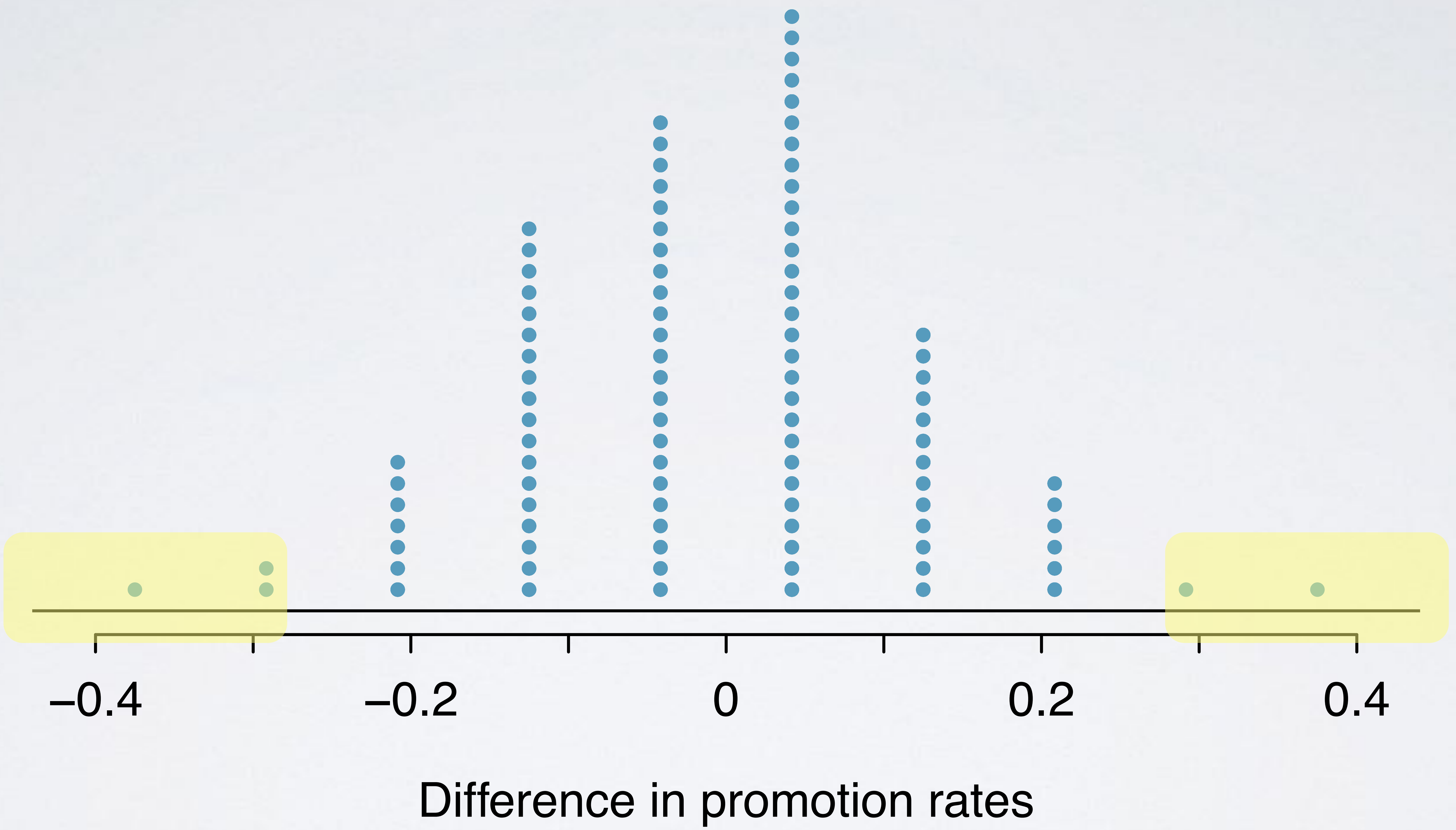
[use a deck of playing cards to simulate this experiment]

1. face card: not promoted, non-face card: promoted
   - set aside the jokers, consider aces as face cards
   - take out 3 aces → 13 face cards left in the deck (face cards: A, K, Q, J)
   - take out a number card → 35 number (non-face) cards left in the deck (number cards: 2-10)
2. shuffle the cards, deal into two groups of size 24, representing males and females
3. count how many number cards are in each group (representing promoted files)
4. calculate the proportion of promoted files in each group, take the difference (male - female), and record this value
5. repeat steps 2 - 4 many times


–0.4
–0.2
0
0.2
0.4
Difference in promotion rates

making a decision

- Results from the simulations look like the data → the difference between the proportions of promoted files between males and females was due to chance (promotion and sex are independent)
- Results from the simulations do not look like the data → the difference between the proportions of promoted files between males and females was <u>not</u> due to chance, but due to an actual effect of gender (promotion and sex are dependent)



–0.4
–0.2
0
0.2
0.4
Difference in promotion rates

Summary

- set a null and an alternative hypothesis
- simulate the experiment assuming that the null hypothesis is true
- evaluated the probability of observing an outcome at least as extreme as the one observed in the original data

p-value

- and if this probability is low, reject the null hypothesis in favor of the alternative

bayesian inference

example: early HIV testing in the military

- First screen with ELISA
- If positive, two more rounds of ELISA
- If either positive, two western blot assays
- Only if both positive, determine HIV infection

example: early HIV testing in the military

ELISA

- Sensitivity (true positive): 93%
- Specificity (true negative): 99%

$P(+ \mid HIV) = 0.93$

$P(- \mid no\ HIV) = 0.99$

Western blot

- Sensitivity: 99.9%
- Specificity: 99.1%

Prevalance: 1.48 / 1000

$P(HIV) = 0.00148$

$$P(has\ HIV \mid ELISA\ +) =?$$

Sources:

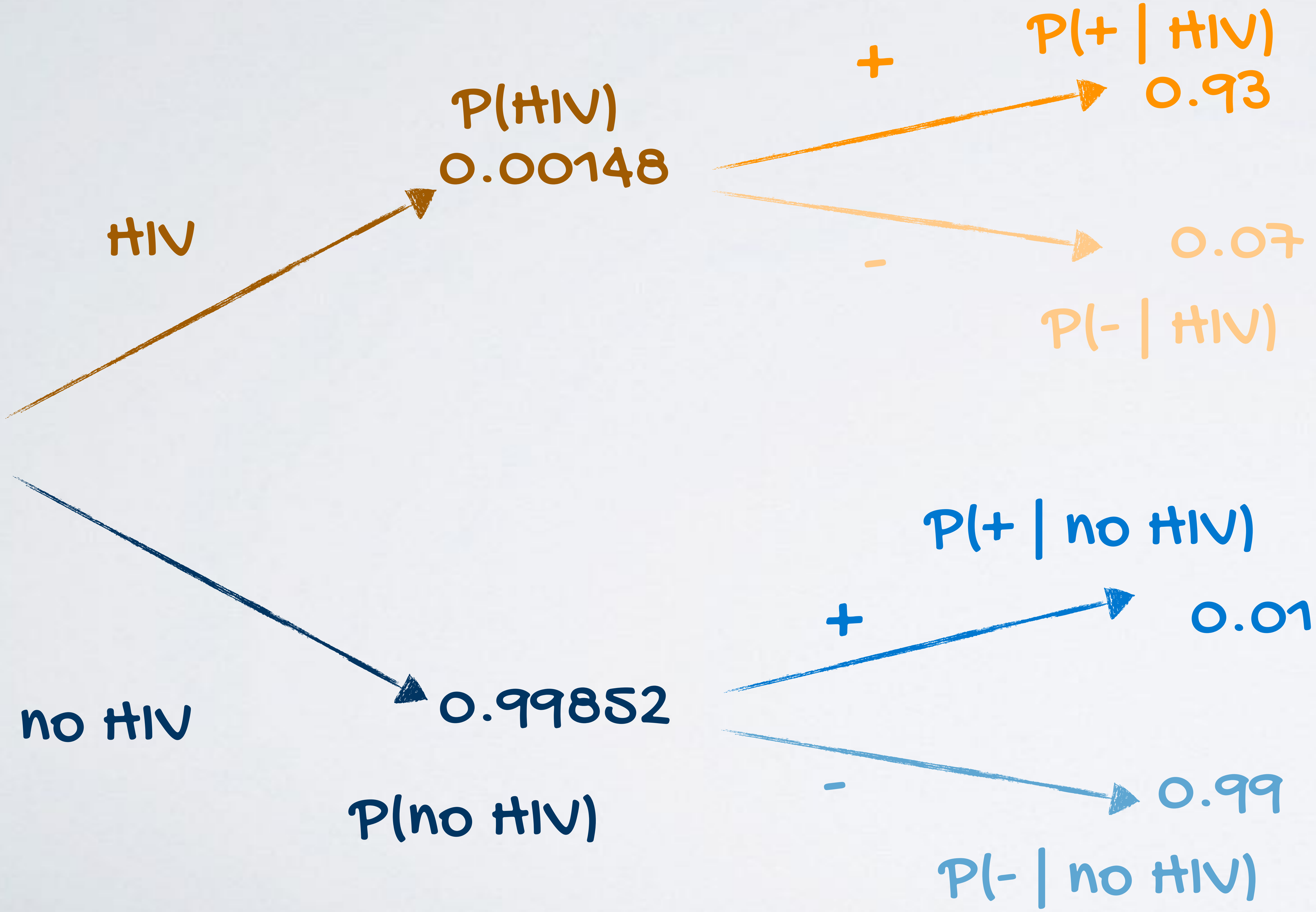
- Petricciani (1985). Licensed tests for antibody to human T-lymphotropic virus type III: sensitivity and specificity. Annals of internal medicine, 103(5), 726-729.
- Burke et. al. (1987). Diagnosis of human immunodeficiency virus infection by immunoassay using a molecularly cloned and expressed virus envelope polypeptide: comparison to Western blot on 2707 consecutive serum samples. Annals of internal medicine, 106(5), 671-676.
- Burke et. al. (1987). Human immunodeficiency virus infections among civilian applicants for United States military service, October 1985 to March 1986. New England Journal of Medicine, 317(3), 131-136.

Prior to any testing, what probability should be assigned to a recruit having HIV?

$P(HIV) = 0.00148$

Posterior probability

When a recruit goes through HIV screening there are two competing claims: recruit has HIV and recruit doesn't have HIV. If the ELISA yields a positive result, what is the probability this recruit has HIV?



P(HIV and +)

$0.00148 \times 0.93 = 0.0013764$

$$P(HIV \mid +) = \frac{P(HIV \text{ and } +)}{P(+)}$$

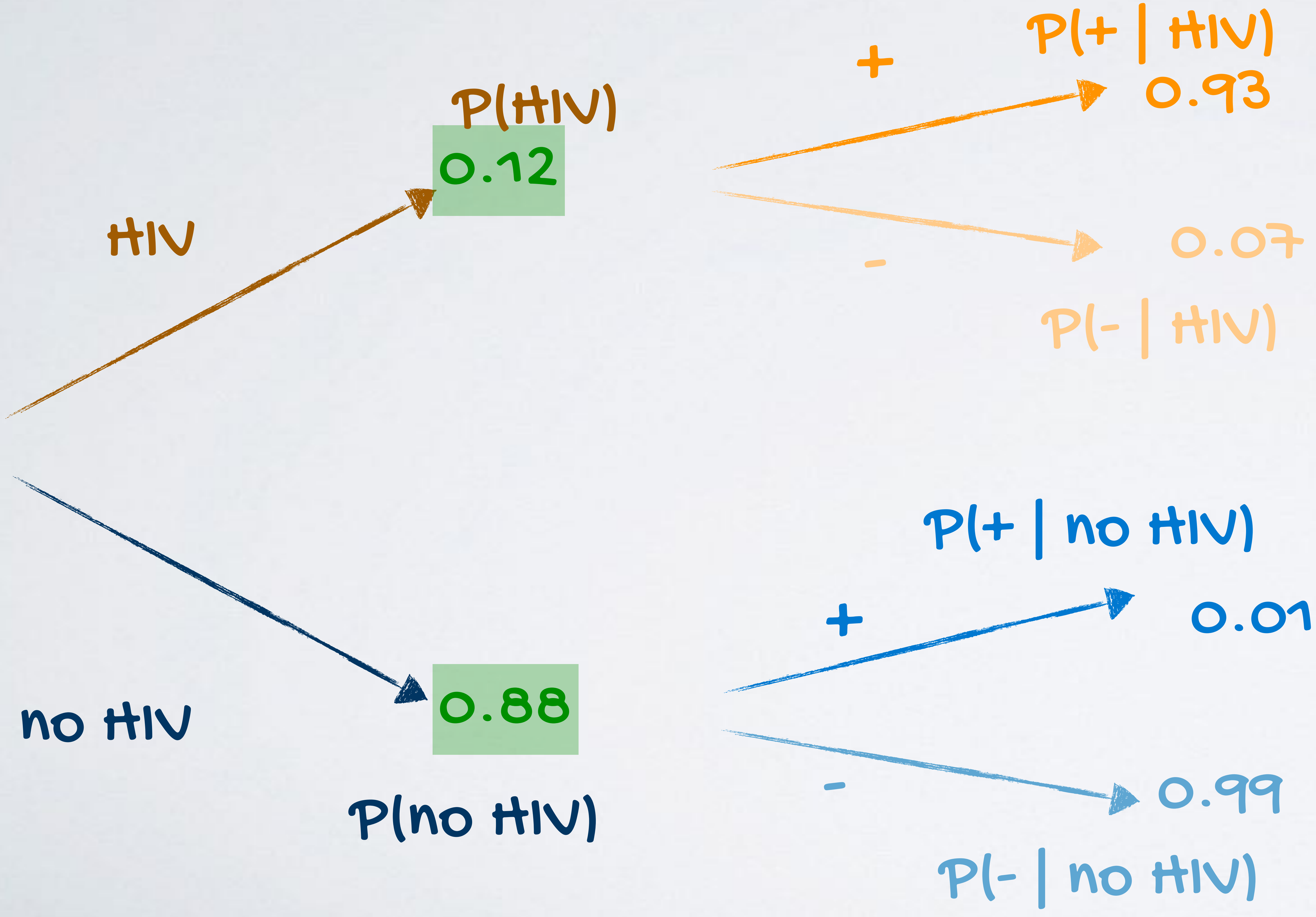
$$= \frac{0.0013764}{0.0013764 + 0.0099852} \approx 0.12$$

$0.99852 \times 0.01 = 0.0099852$

P(no HIV and +)

Bayesian updating

Since a positive outcome on the ELISA doesn't necessarily mean that the recruit actually has HIV, they are retested. What is the probability of having HIV if this second ELISA also yields a positive result?



P(HIV and +)

$0.12 \times 0.93 = 0.1116$

$$P(HIV \mid +) = \frac{P(HIV \text{ and } +)}{P(+)}$$

$$= \frac{0.1116}{0.1116 + 0.0088} \approx 0.93$$

$0.88 \times 0.01 = 0.0088$

P(no HIV and +)

recap

- Individual vs. group diagnostics
- Updating only the prior vs. also updating sensitivity and specificity
- Bayesian updating

bayesian & frequentist definitions of probability

frequentist definition

$$P(E) = \lim_{n \to \infty} \frac{n_E}{n}$$

bayesian definition

- Indifferent between winning
  - $1 if event E occurs, or
  - winning $1 if you draw a blue chip from a box with $1{,}000 \times p$ blue chips $+ 1{,}000 \times (1-p)$ white chips
- Equating the probability of event E, $P(E)$, to the probability of drawing a blue chip from this box, p

$$P(E) = p$$

Example: Based on a 2022 Pew Research poll on 5,074 Adults: *"We are 95% confident that 68% to 72% of Americans think inflation is the biggest problem facing the country."*

- 95% of random samples of 5,074 adults will produce confidence intervals for the proportion of Americans who think inflation is the biggest problem facing the country.
- Common misconceptions:
  - There is a 95% chance that this confidence intervals includes the true population proportion.
  - The true population proportion is in this interval 95% of the time.

Source: https://www.pewresearch.org/fact-tank/2022/05/12/by-a-wide-margin-americans-view-inflation-as-the-top-problem-facing-the-country-today/

credible intervals

- Allows us to describe the unknown true parameter not as a fixed value but with a probability distribution
- This will let us construct something like a confidence interval, except we can make probabilistic statements about the parameter falling within that range.
  - **Example:** *"The posterior distribution yields a 95% credible interval of 68% to 72% for the proportion of Americans who think inflation is the biggest problem facing the country."*
- These are called credible intervals.

Source: http://www.pewsocialtrends.org/2016/02/04/most-americans-say-government-doesnt-do-enough-to-help-middle-class/

inference for a proportion: frequentist approach

example: morning after pill

- **research question:** Is RU-486 an effective "morning after" contraceptive?
- **participants:** 40 women who came to a health clinic asking for emergency contraception
- **design:** Random assignment to RU-486 or standard therapy (20 in each group)
- **data:**
  - 4 out of 20 in RU-486 (treatment) became pregnant
  - 16 out of 20 in standard therapy (control) pregnant
- **question:** How strongly do these data indicate that the treatment is more effective than the control?

framework

- simplification: one proportion
  - consider the 20 total pregnancies
  - question: How likely is it that 4 pregnancies occur in the treatment group?
- if treatment and control are equally effective + sample sizes for the two groups are the same

$$\text{P(pregnancy comes from treatment group)} = p = 0.5$$

hypotheses

$p =$ probability that a given pregnancy comes from the treatment group

$H_0 : p = 0.5$ - No difference, a pregnancy is equally likely to come from the treatment or control group

$H_A : p < 0.5$ - Treatment is more effective, a pregnancy is less likely to come from the treatment group

p-value

- $k = 4$ and $n = 20$ - since there are 20 pregnancies total, and 4 occur in the treatment group
- $p = 0.5$ - assuming H_0 is true
- p–value = $P(k \leq 4)$

```r
sum(dbinom(0:4, size = 20, p = 0.5))
```

```
## [1] 0.005908966
```

inference for a proportion: bayesian approach

- consider the 20 total pregnancies
  - **question:** How likely is it that 4 pregnancies occur in the treatment group?
- if treatment and control are equally effective + sample sizes for the two groups are the same

$$\text{P(pregnancy comes from treatment group)} = p = 0.5$$

hypotheses, i.e. models

- delineate plausible models:
  - assume p could be

    10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, or 90%
- consider 9 models, instead of 1 as in the frequentist paradigm
  - $p = 20\%$: Given a pregnancy occurs, there is a 2:8 or 1:4 chance that it will occur in the treatment group

specifying the prior

- prior probabilities reflect state of belief prior to the current experiment
- incorporate information learned from all relevant research up to the current point in time, but *not* incorporate information from the current experiment
- suppose my prior probability for each of the 9 models is as presented below:

Model (p)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
Prior	0.06	0.06	0.06	0.06	0.52	0.06	0.06	0.06	0.06	1

  - benefit of treatment is symmetric — equally likely to be better or worse than the standard treatment
  - 52% chance that there is no difference between the treatments

likelihood

- calculate $P(data \mid model)$ for each model considered.
- this probability is called the **likelihood**:

$$P(data \mid model) = P(k = 4 \mid n = 20, p)$$

calculating the likelihood

```r
p <- seq(from = 0.1, to = 0.9, by = 0.1)
prior <- c(rep(0.06, 4), 0.52, rep(0.06, 4))
likelihood <- dbinom(4, size = 20, prob = p)
```

Model (p)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
Prior, $P(model)$	0.06	0.06	0.06	0.06	0.52	0.06	0.06	0.06	0.06	1
Likelihood, $P(data \mid model)$	0.0898	0.2182	0.1304	0.035	0.0046	0.0003	0	0	0	

use Bayes' rule to calculate the posterior probability, i.e. *P(model | data)*

$$P(model \mid data) = \frac{P(model \ \& \ data)}{P(data)}$$

$$= \frac{P(data \mid model) \times P(model)}{P(data)}$$

calculating the posterior

```
numerator <- prior * likelihood
denominator <- sum(numerator)
posterior <- numerator / denominator
sum(posterior)
```

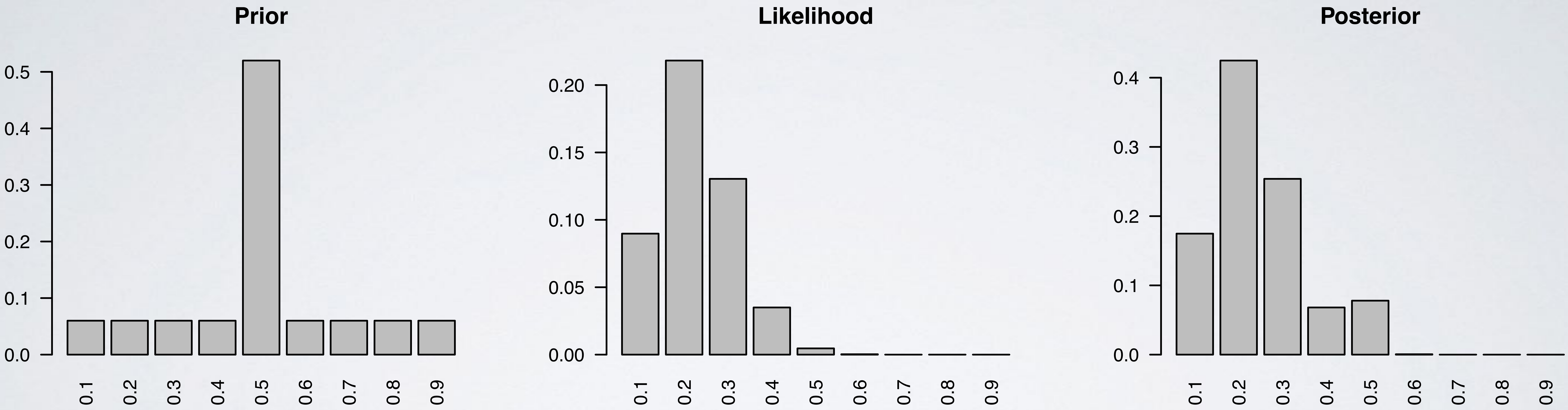
```
## [1] 1
```

Model (p)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
Prior, *P(model)*	0.06	0.06	0.06	0.06	0.52	0.06	0.06	0.06	0.06	1
Likelihood, *P(data | model)*	0.0898	0.2182	0.1304	0.035	0.0046	0.0003	0	0	0	
P(data|model) x P(model)	0.0054	0.0131	0.0078	0.0021	0.0024	0	0	0	0	0.0308
Posterior, *P(model|data)*	0.1748	0.4248	0.2539	0.0681	0.0780	0.0005	0	0	0	1

decision making

- posterior probability that $p = 0.2$ is 42.48%
  - this model has the highest posterior probability
- calculation of the posterior incorporated prior information and likelihood of data observed
  - data "at least as extreme as observed" plays no part in the Bayesian paradigm
- note that probability that $p = 0.5$ dropped from 52% in the prior to about 7.8% in the posterior
  - this demonstrates how we update our beliefs based on observed data

prior, likelihood, and posterior



- Bayesian paradigm allows us to make direct probability statements about our models
- we can also calculate the probability that RU-486 (the treatment) is more effective than the control
  - this is the sum of the posteriors of the models where $p < 0.5$

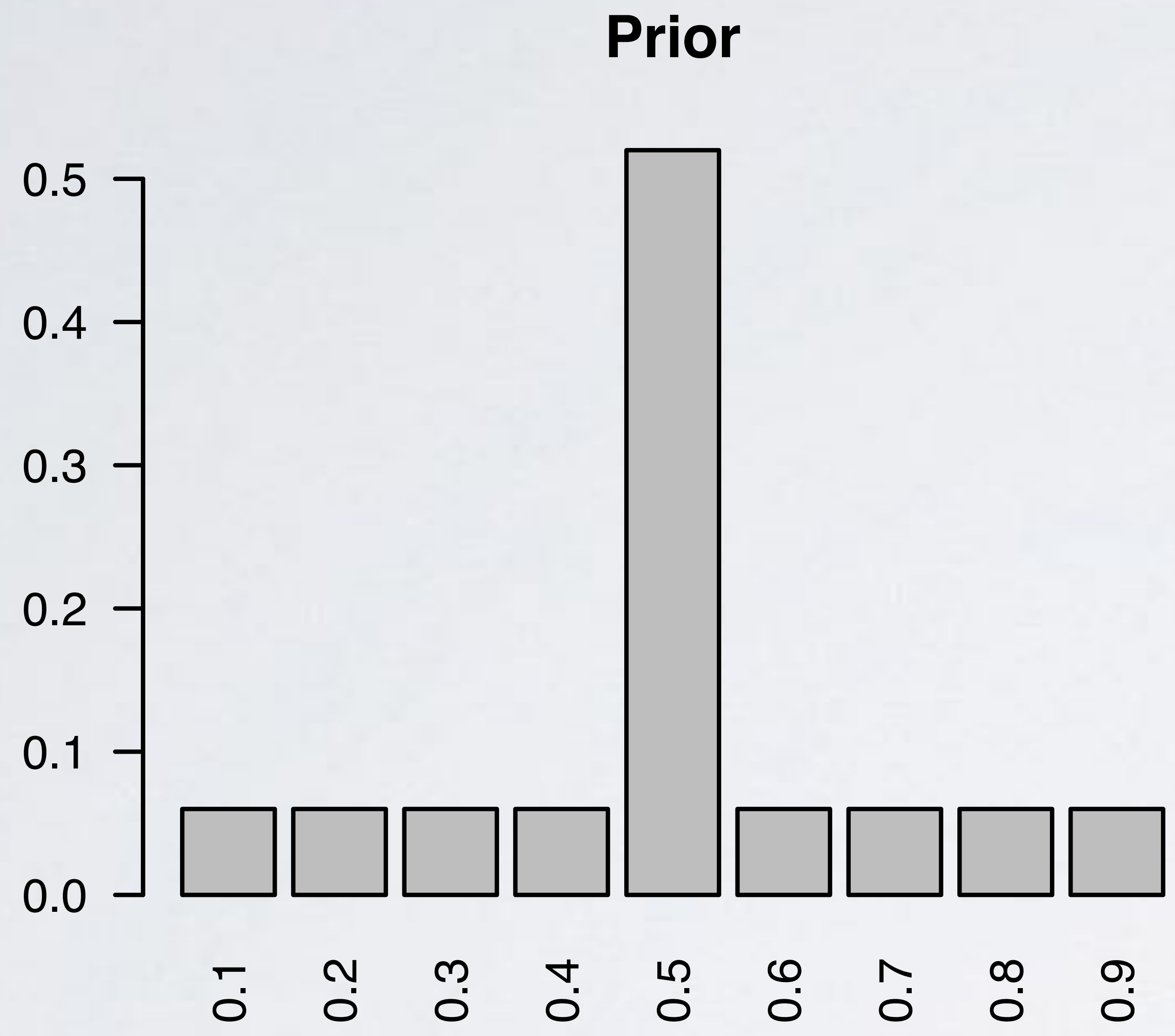
Model (p)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
Posterior, $P(model \mid data)$	0.1748	0.4248	0.2539	0.0681	0.0780	0.0005	0	0	0	1

0.9216

effect of sample size

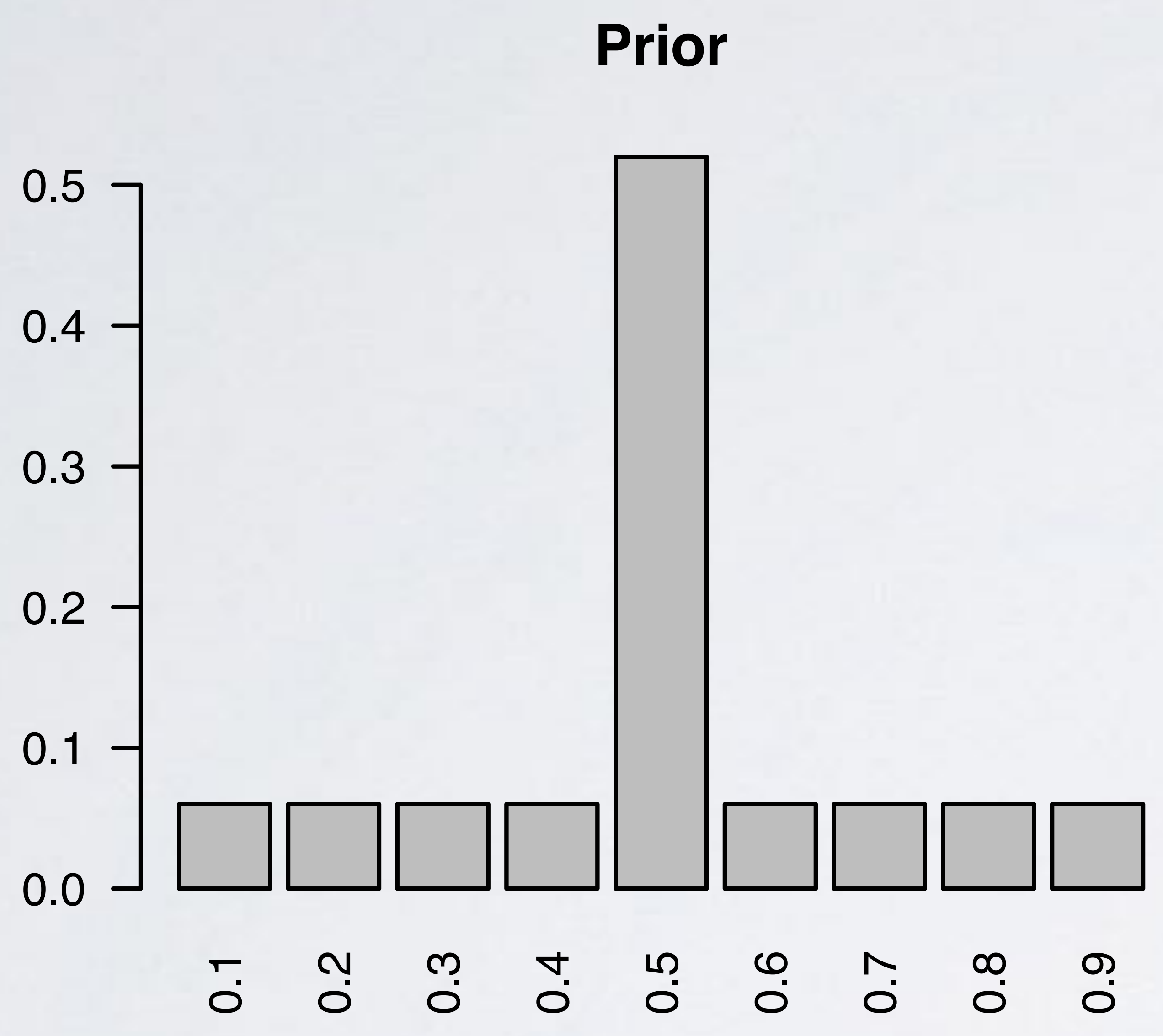
original results

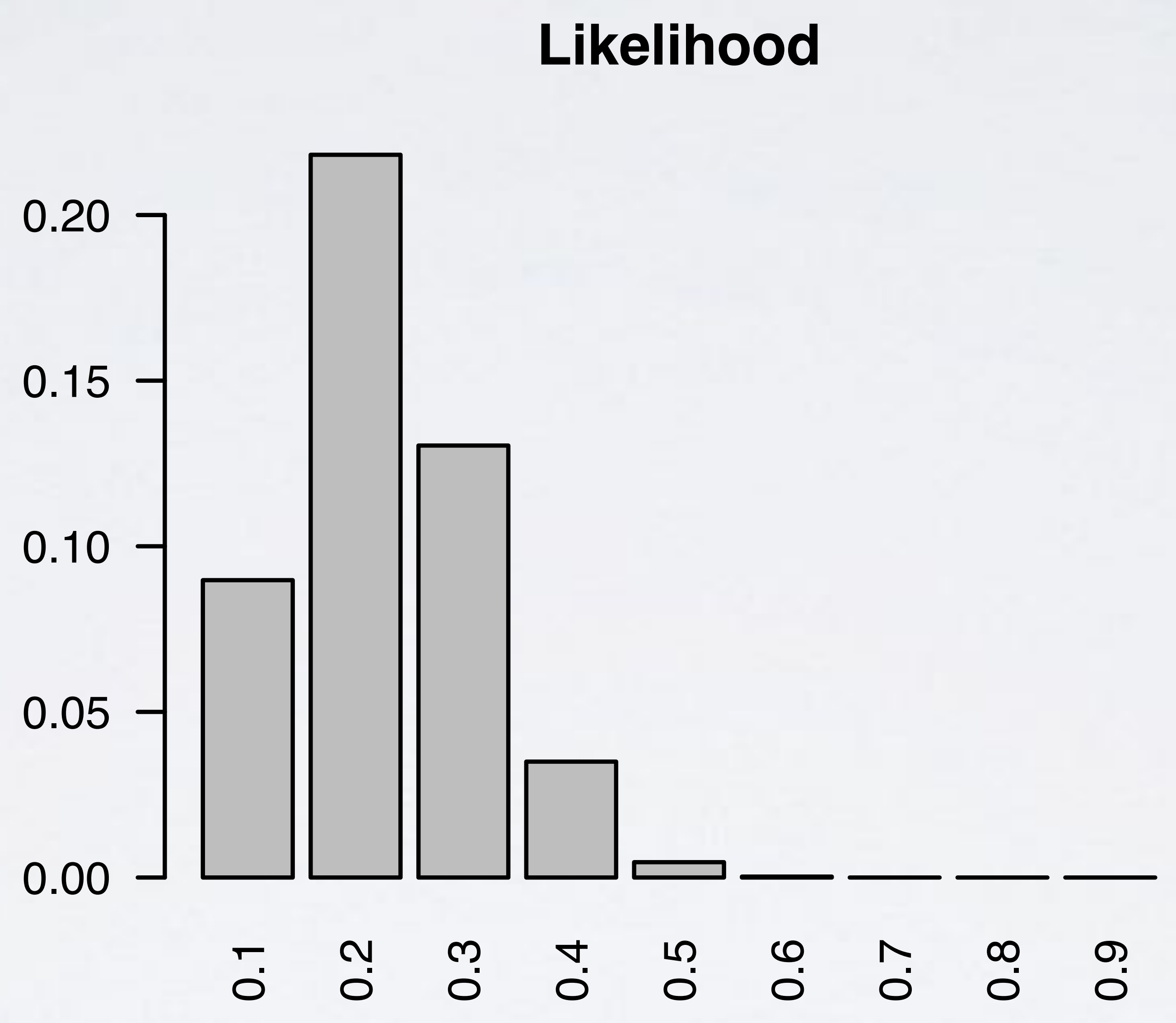
$n = 20, k = 4$



original results

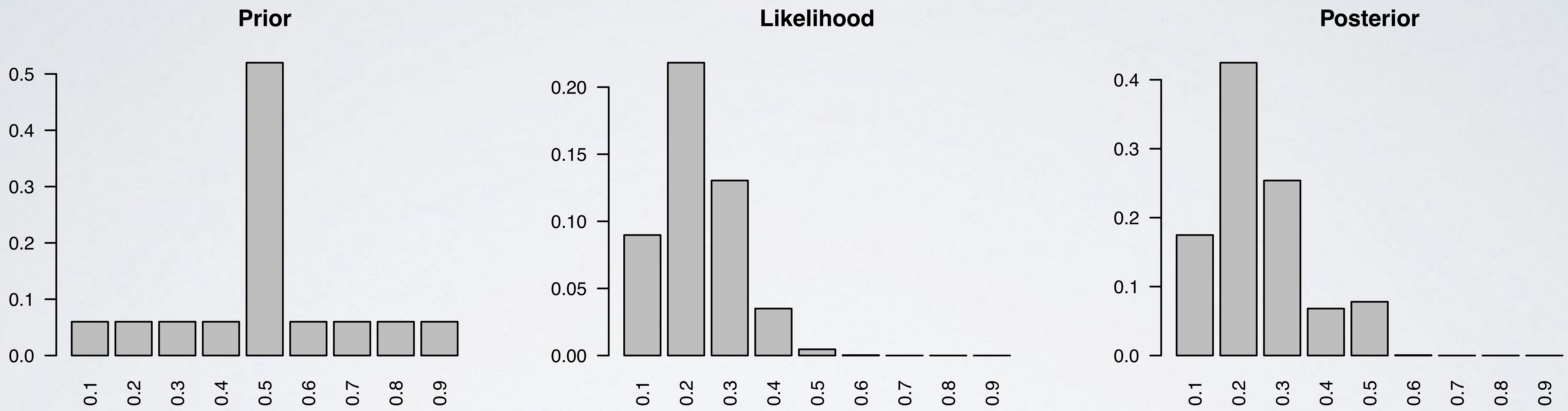
$n = 20, k = 4$





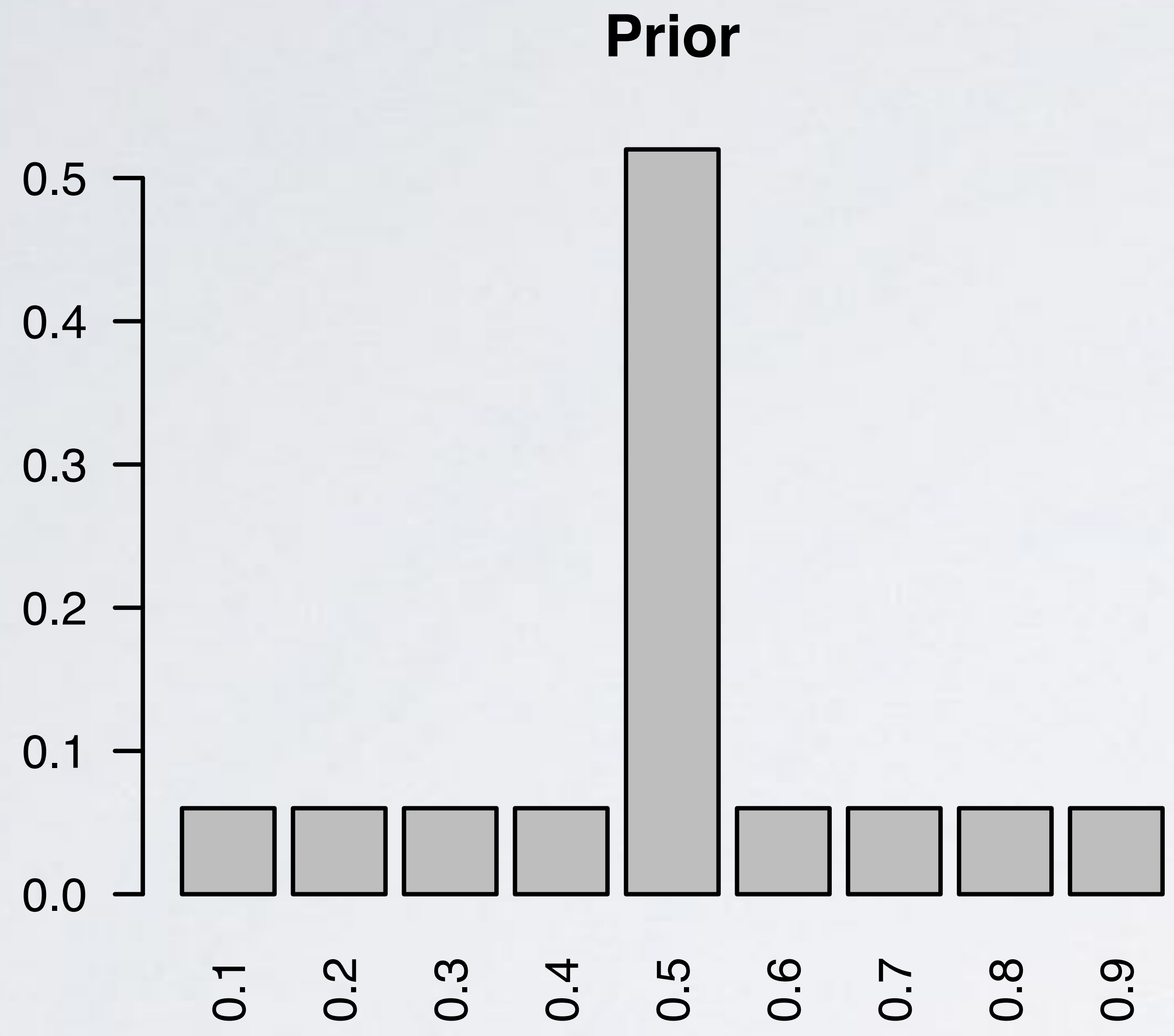
original results

$n = 20, k = 4$



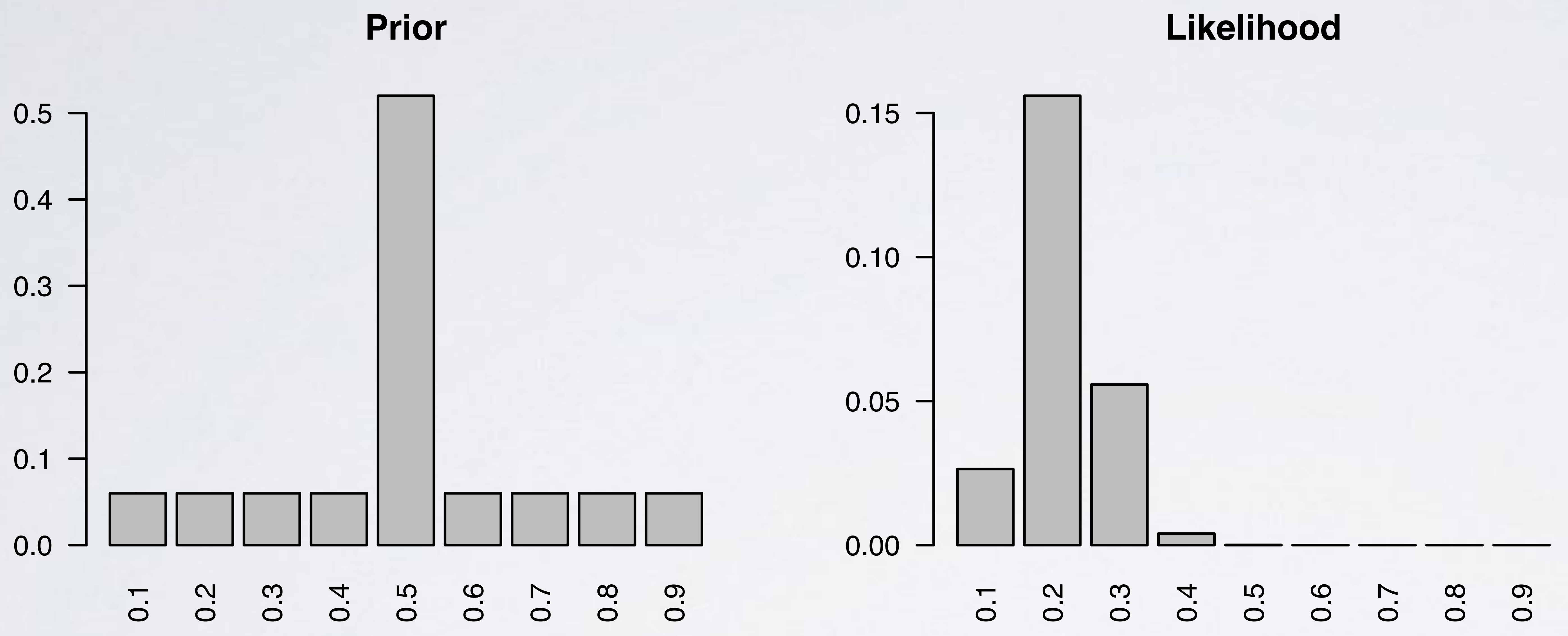
what if we had more data

$n = 40, k = 8$



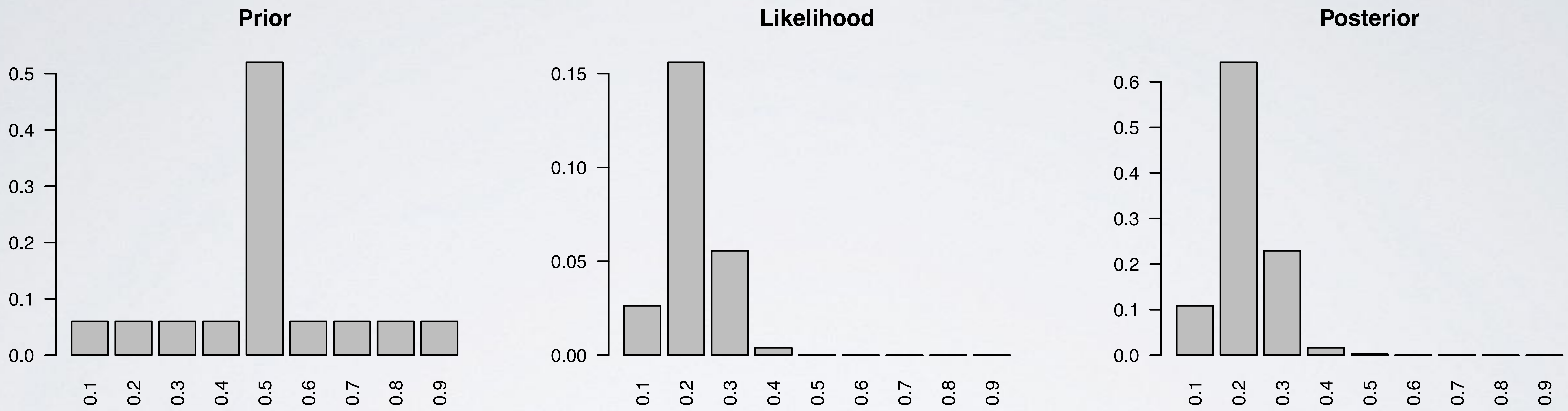
what if we had more data

$n = 40, k = 8$



what if we had more data

$n = 40, k = 8$



what if we had more data

$n = 200, k = 40$