Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 8: Probabilistic models: Naive Bayes

Feb 9, 2017

elements of probability in many of these methods

Random variable

- A variable that can take values within a fixed set (discrete) or within some range (continuous).

$$X \in \{1, 2, 3, 4, 5, 6\}$$

$$X \in \{the, a, dog, cat, runs, to, store\}$$

$$P(X = x)$$

Probability that the random variable X takes the value x (e.g., 1)

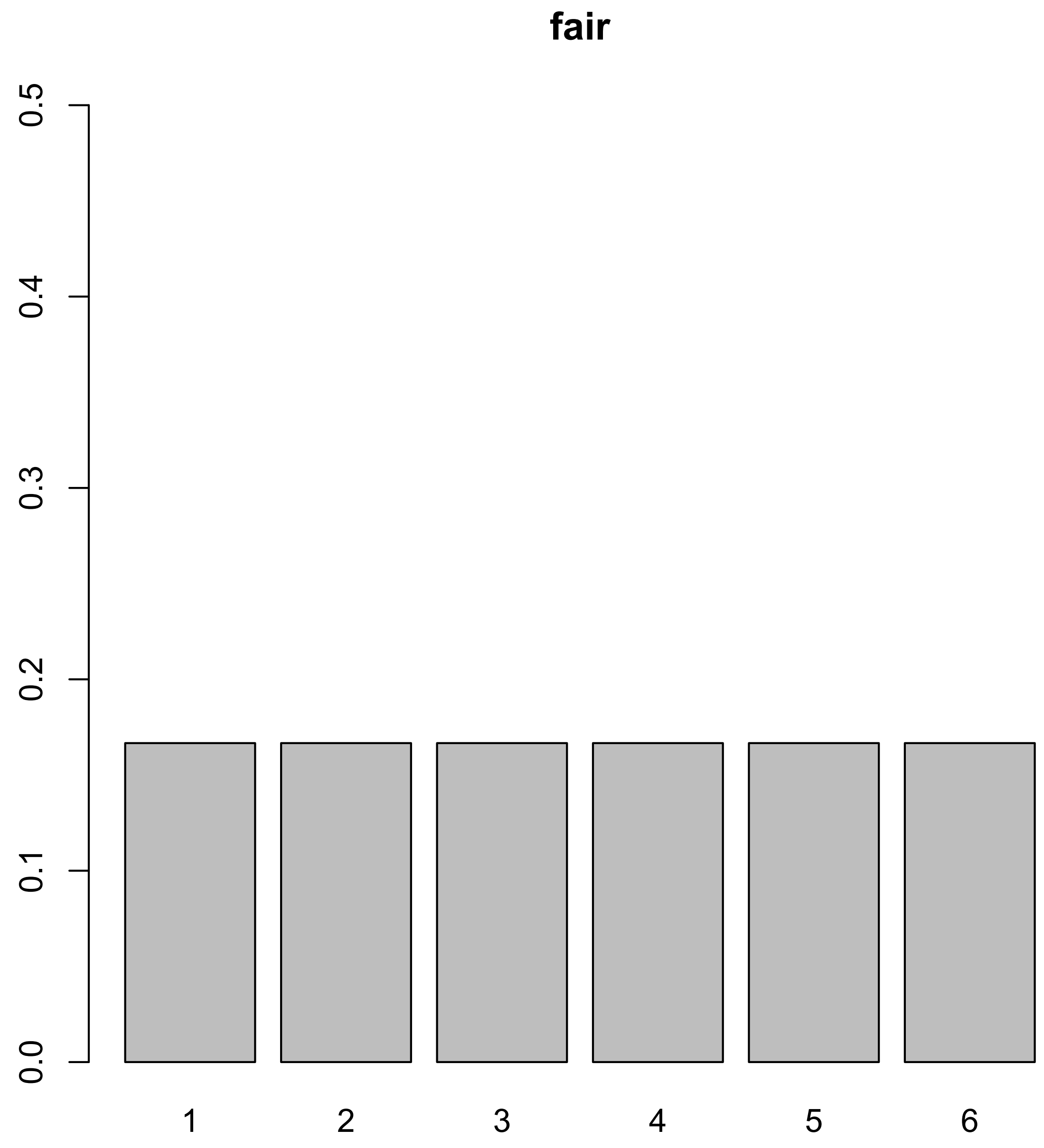
$$X \in \{1, 2, 3, 4, 5, 6\}$$

Two conditions:

1. Between 0 and 1: $0 \leq P(X = x) \leq 1$
2. Sum of all probabilities = 1 $\sum_x P(X = x) = 1$

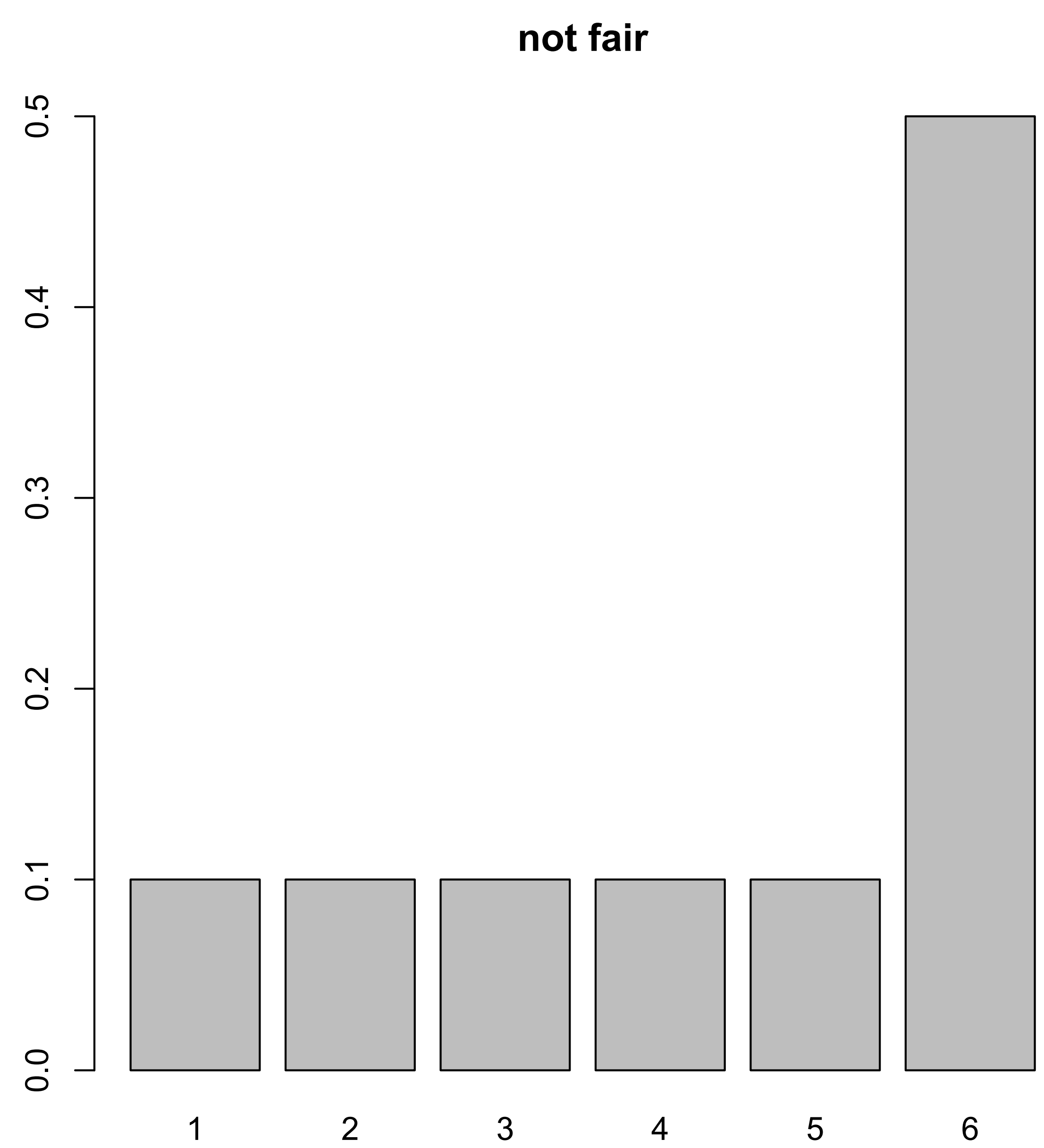
Fair dice

$$X \in \{1, 2, 3, 4, 5, 6\}$$



Weighted dice

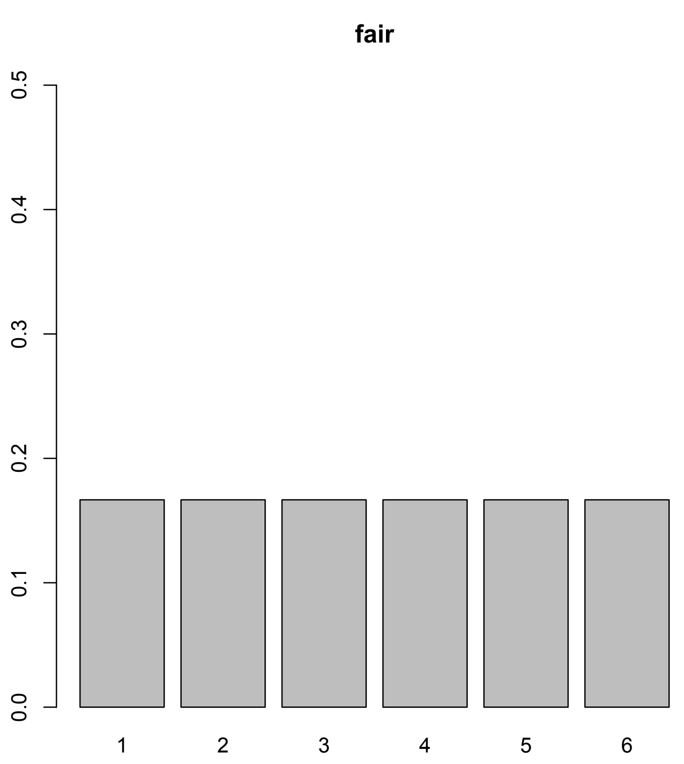
$$X \in \{1, 2, 3, 4, 5, 6\}$$



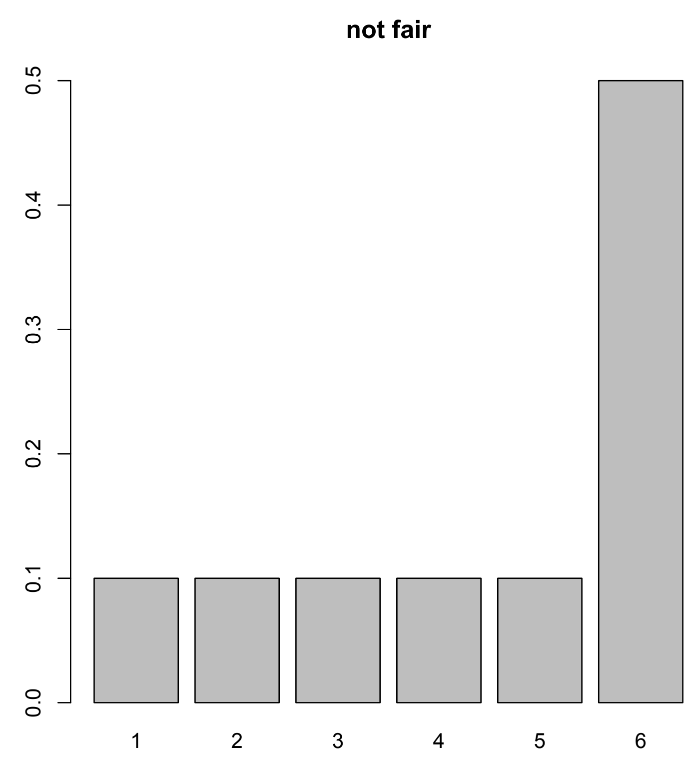
Inference

$$X \in \{1, 2, 3, 4, 5, 6\}$$

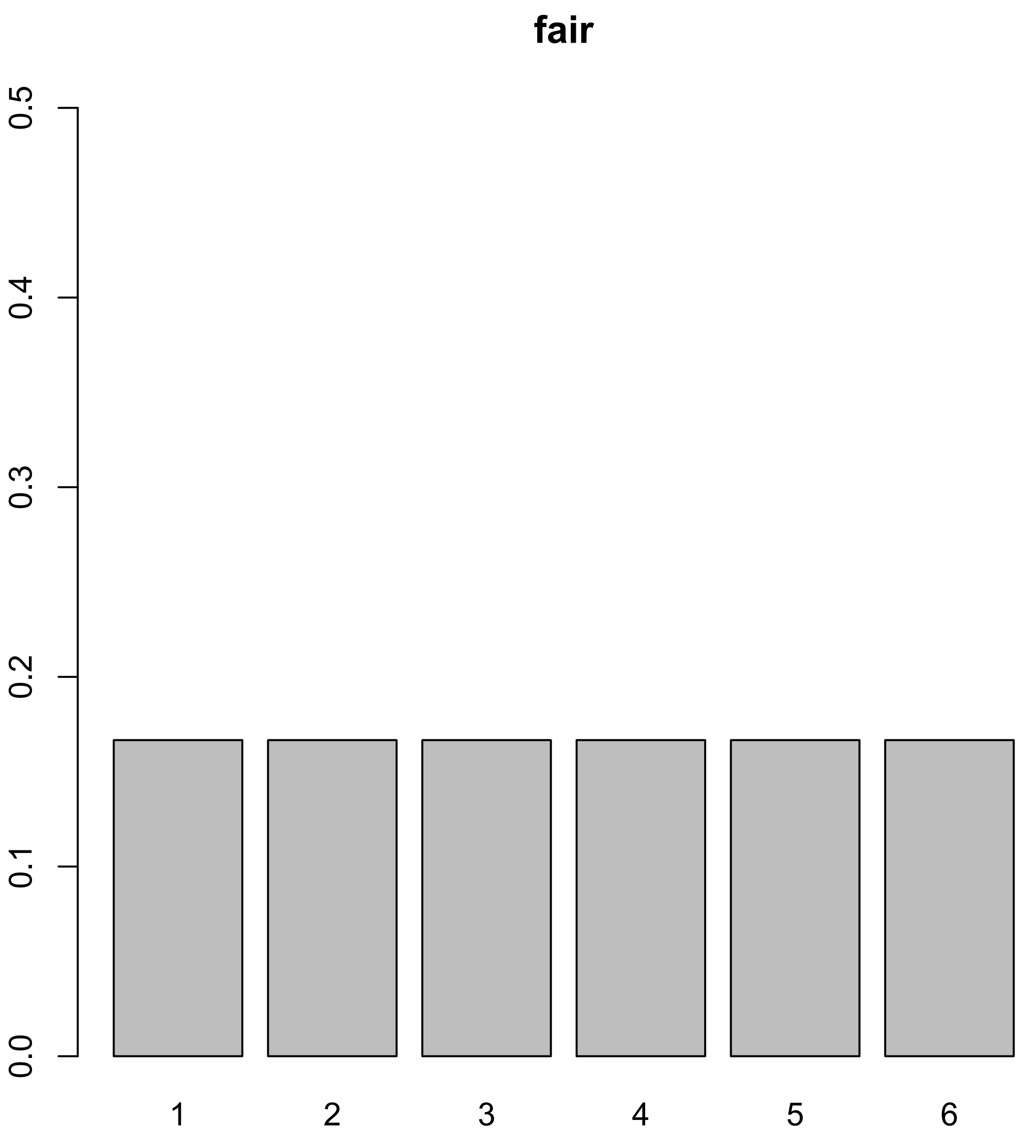
We want to *infer* the probability distribution that generated the data we see.

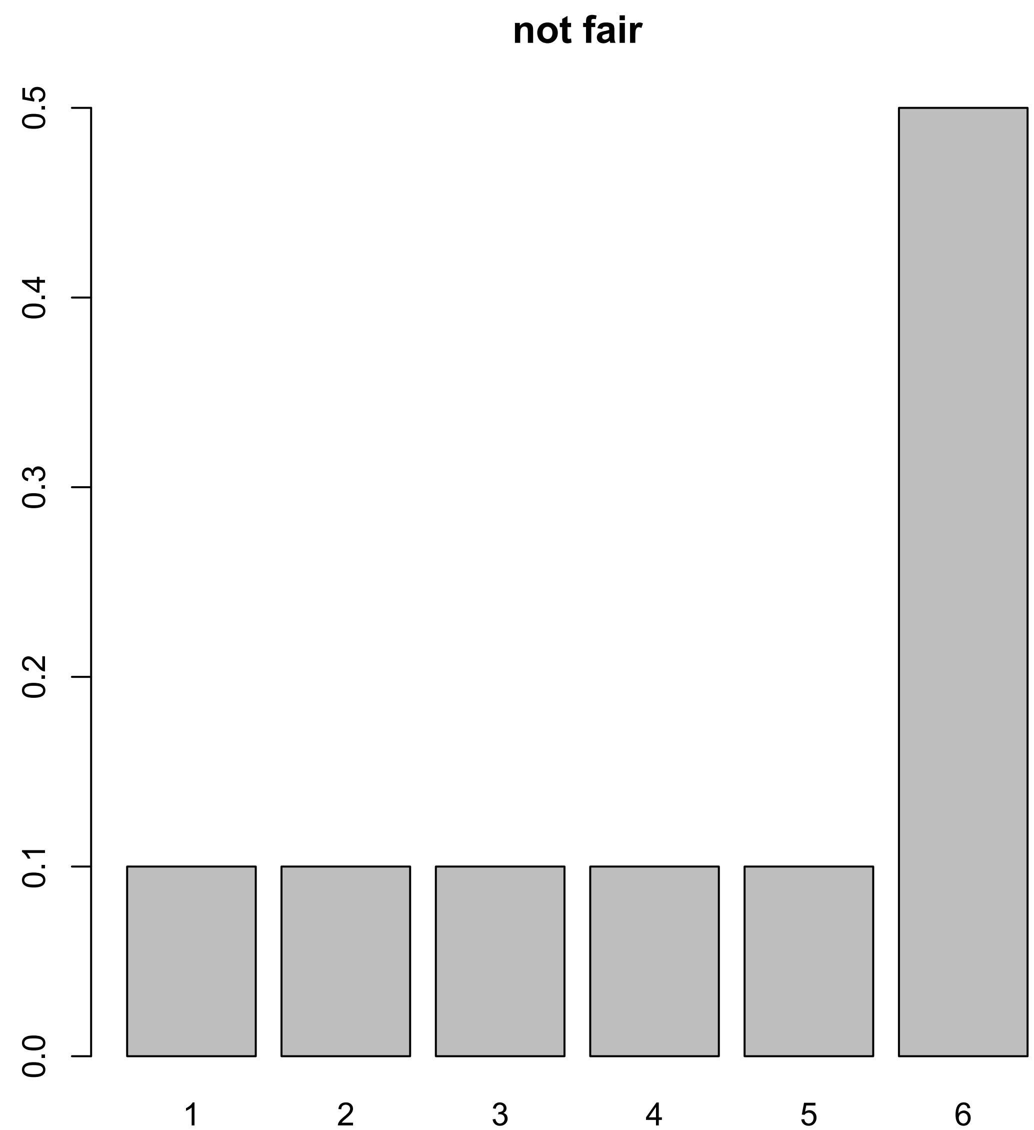


?

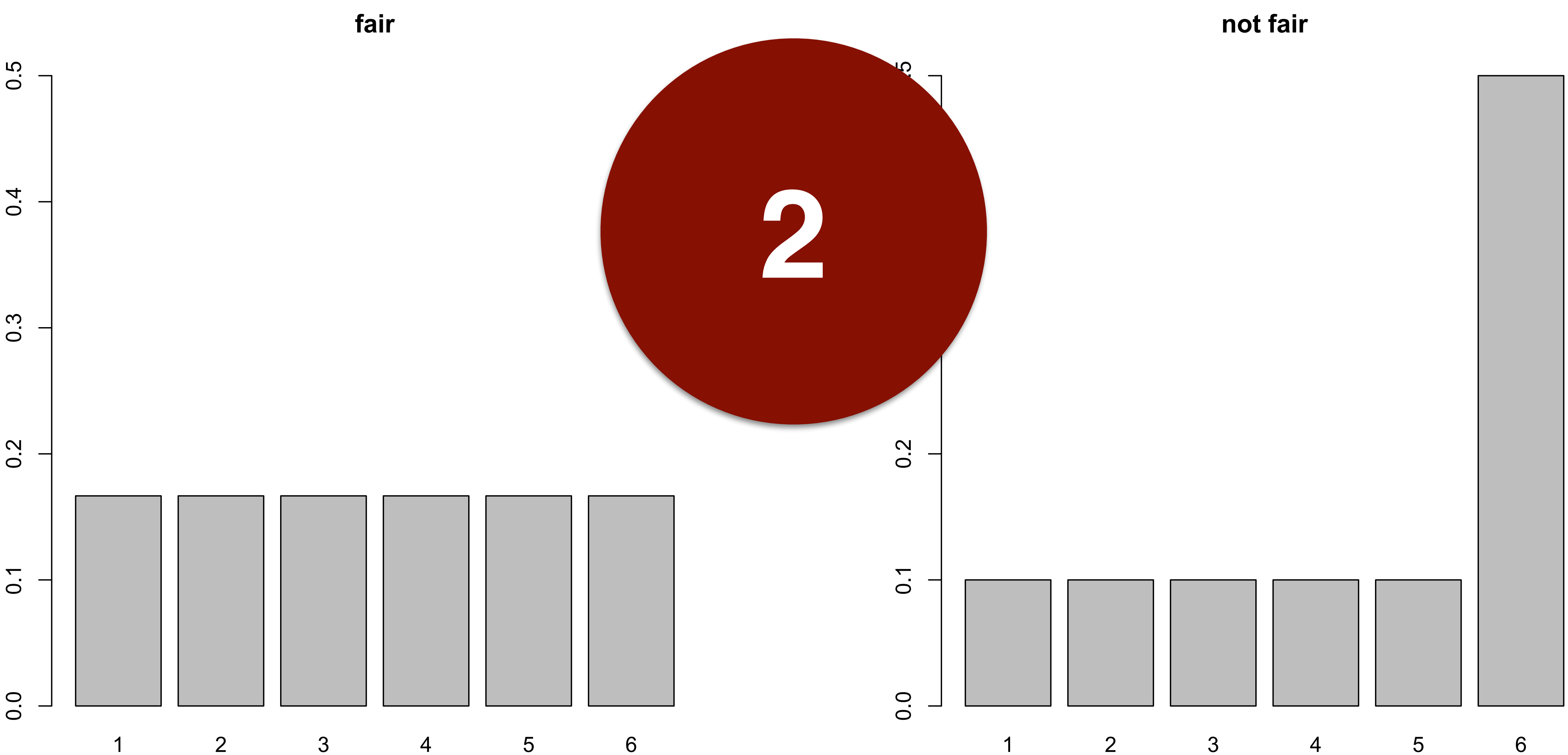


Probability

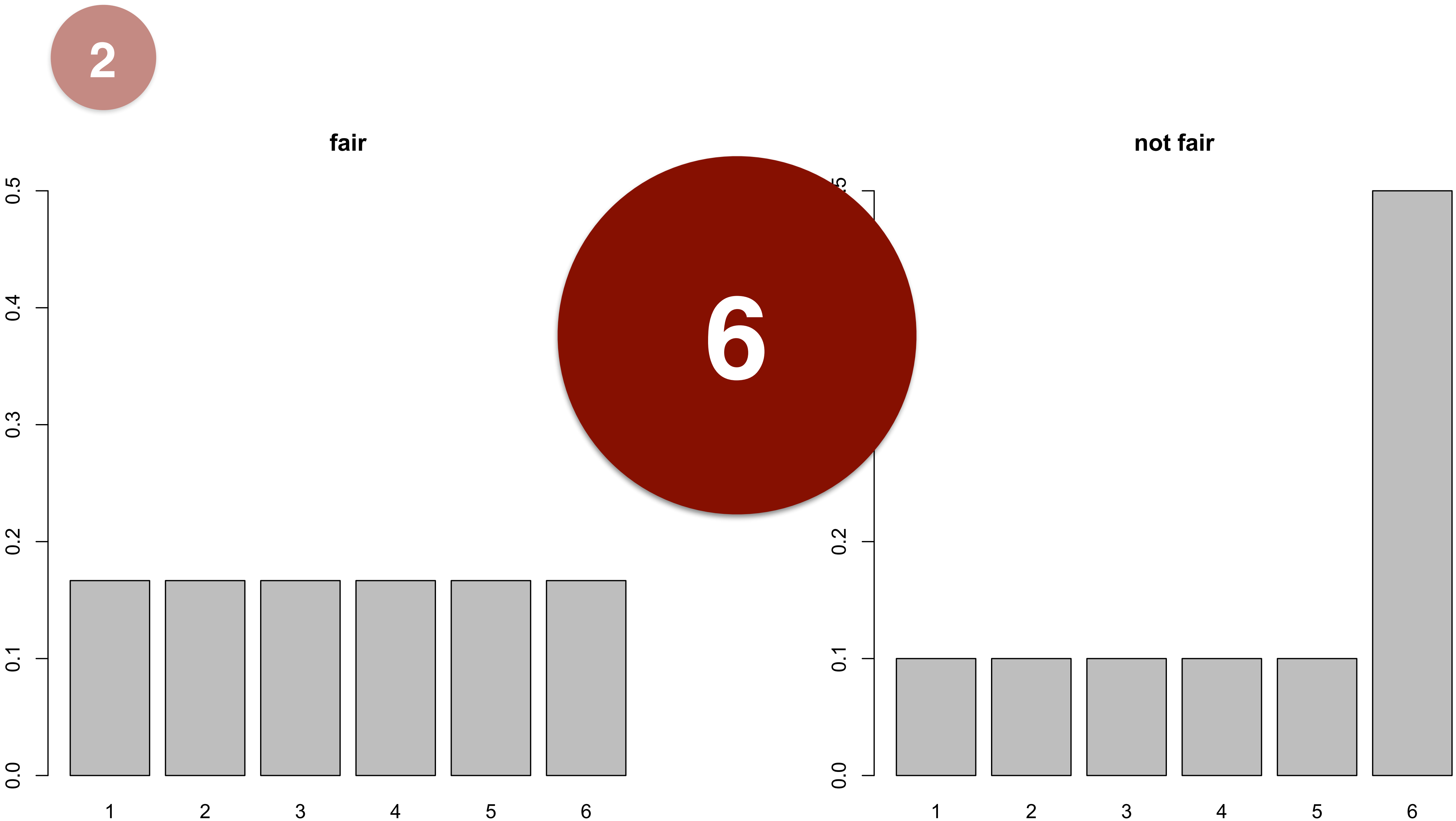




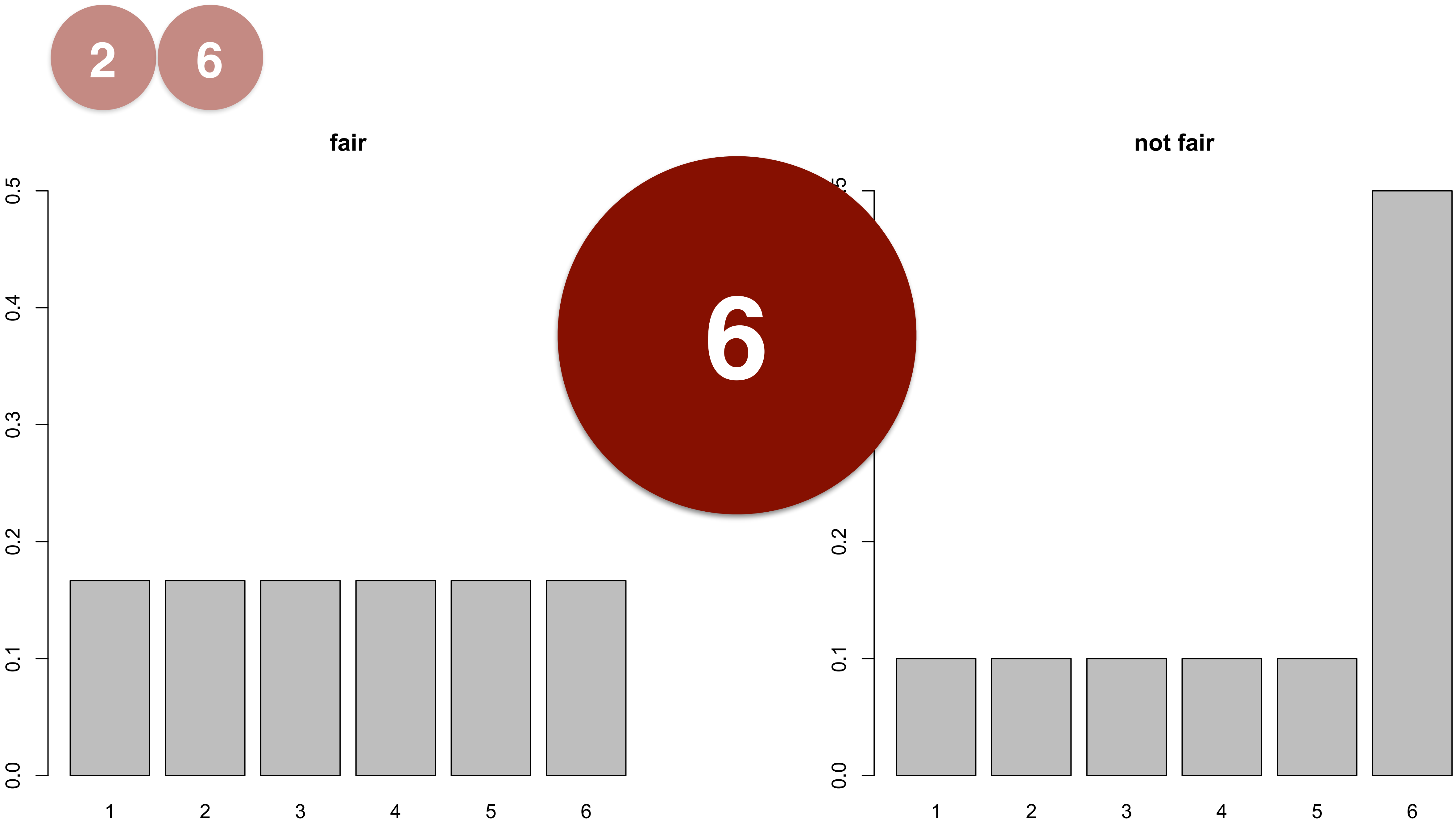
Probability



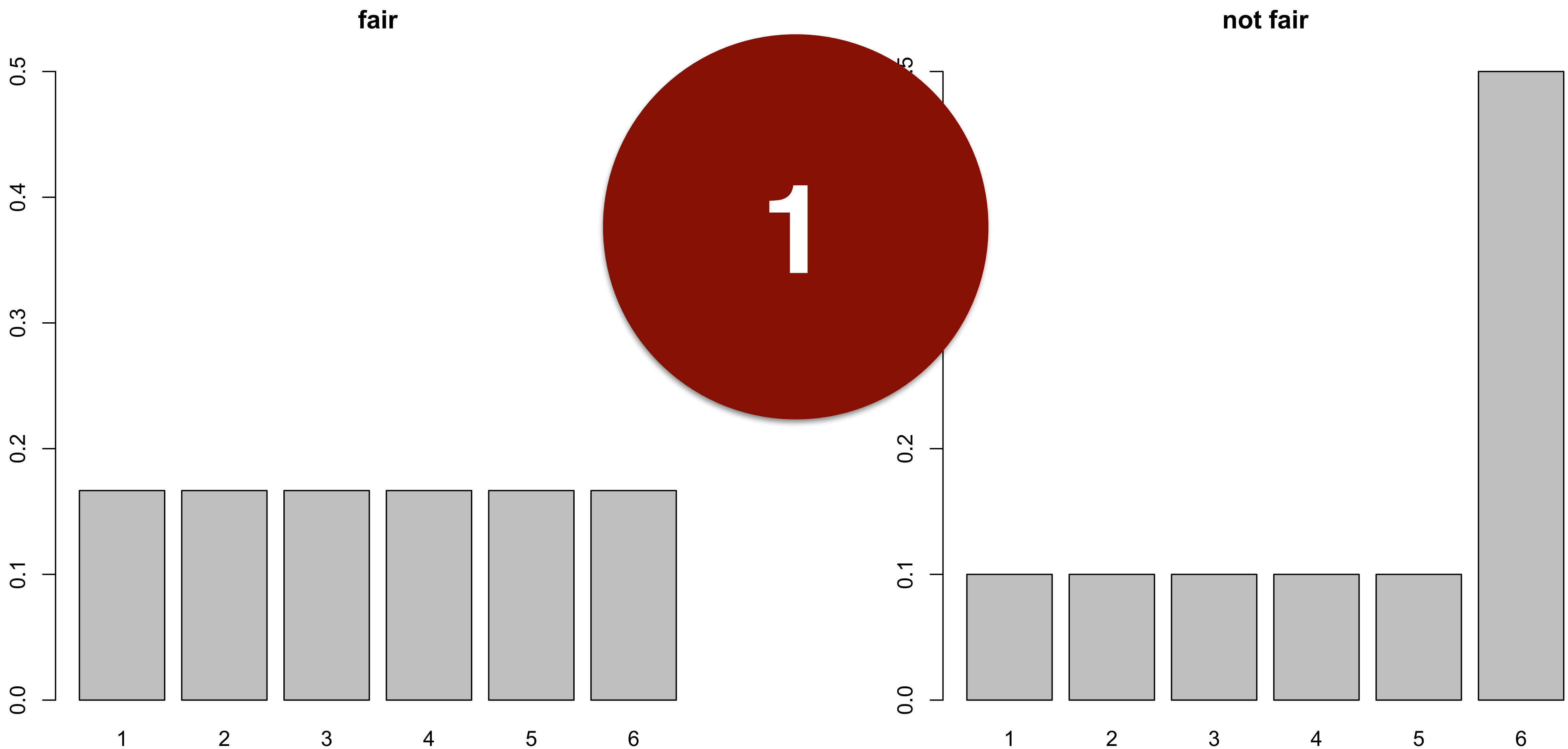
Probability



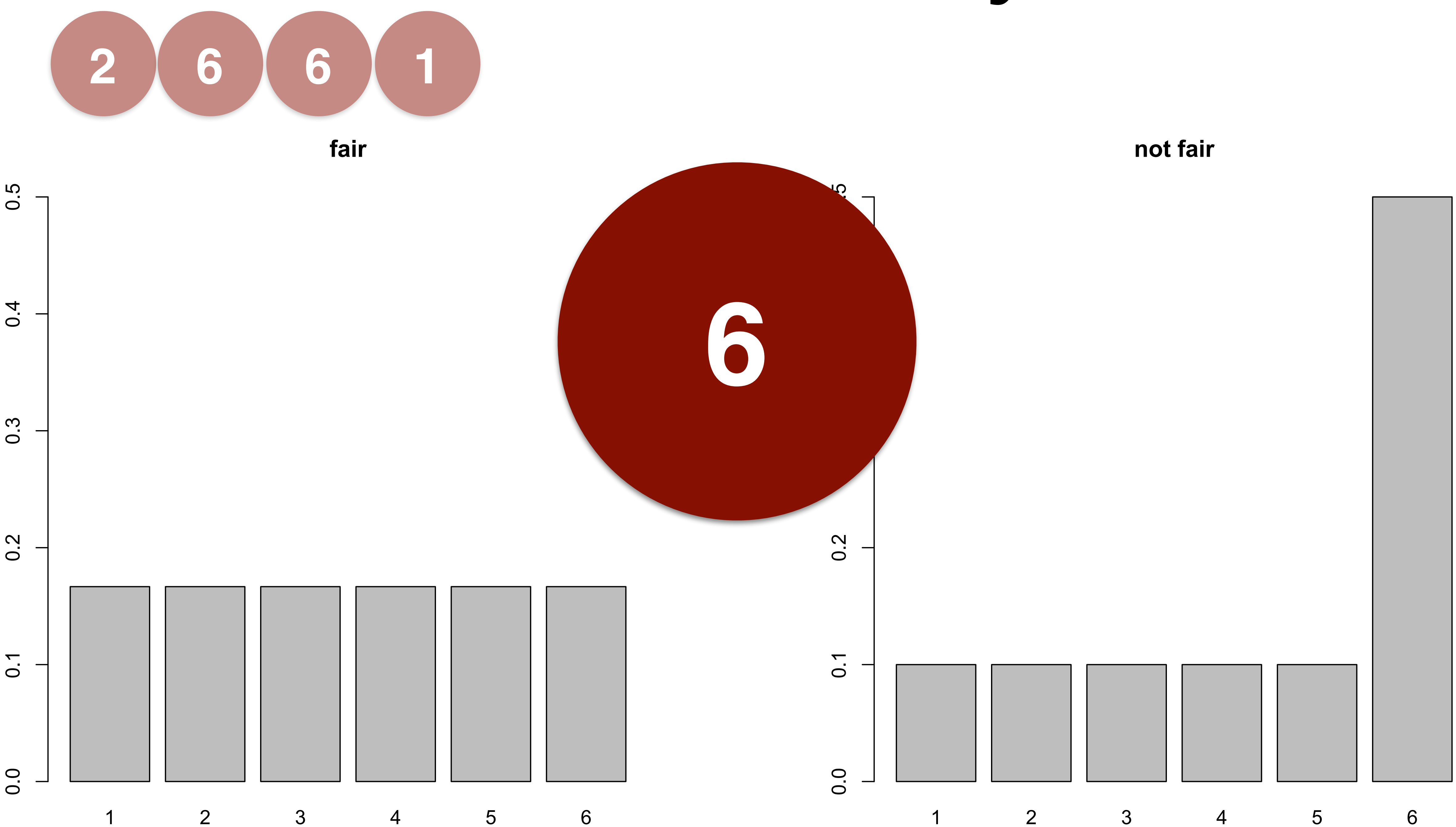
Probability



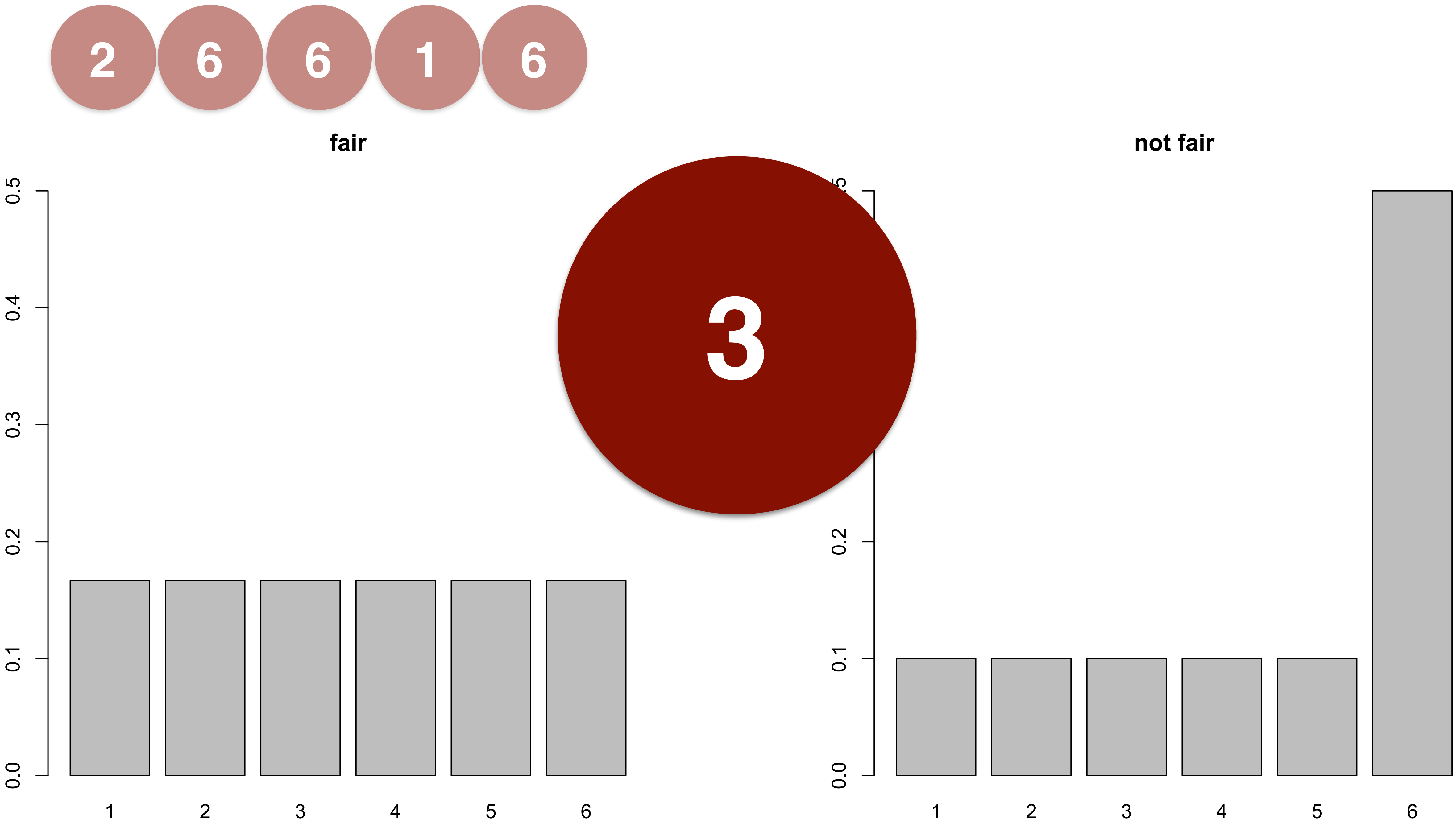
Probability



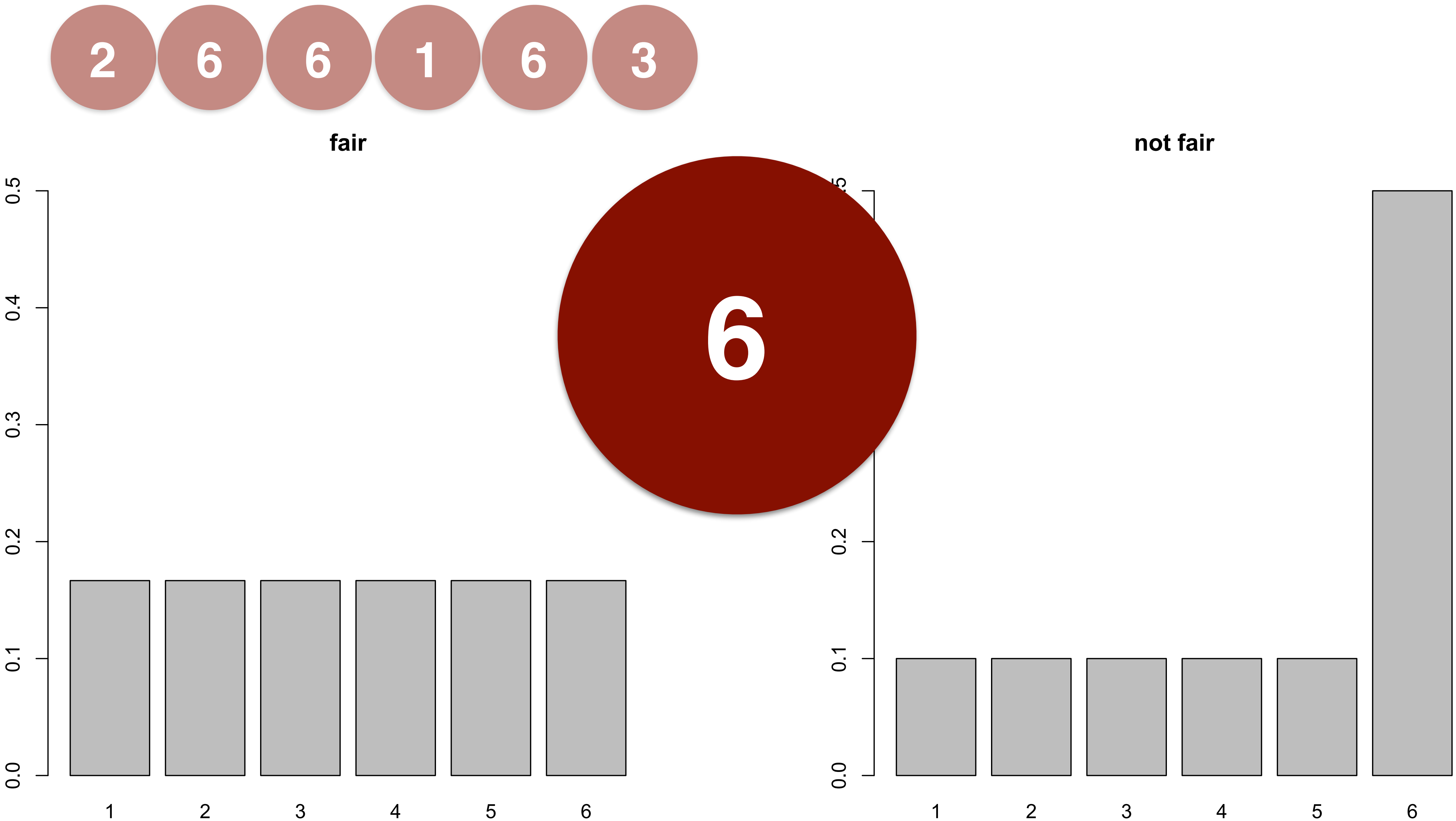
Probability



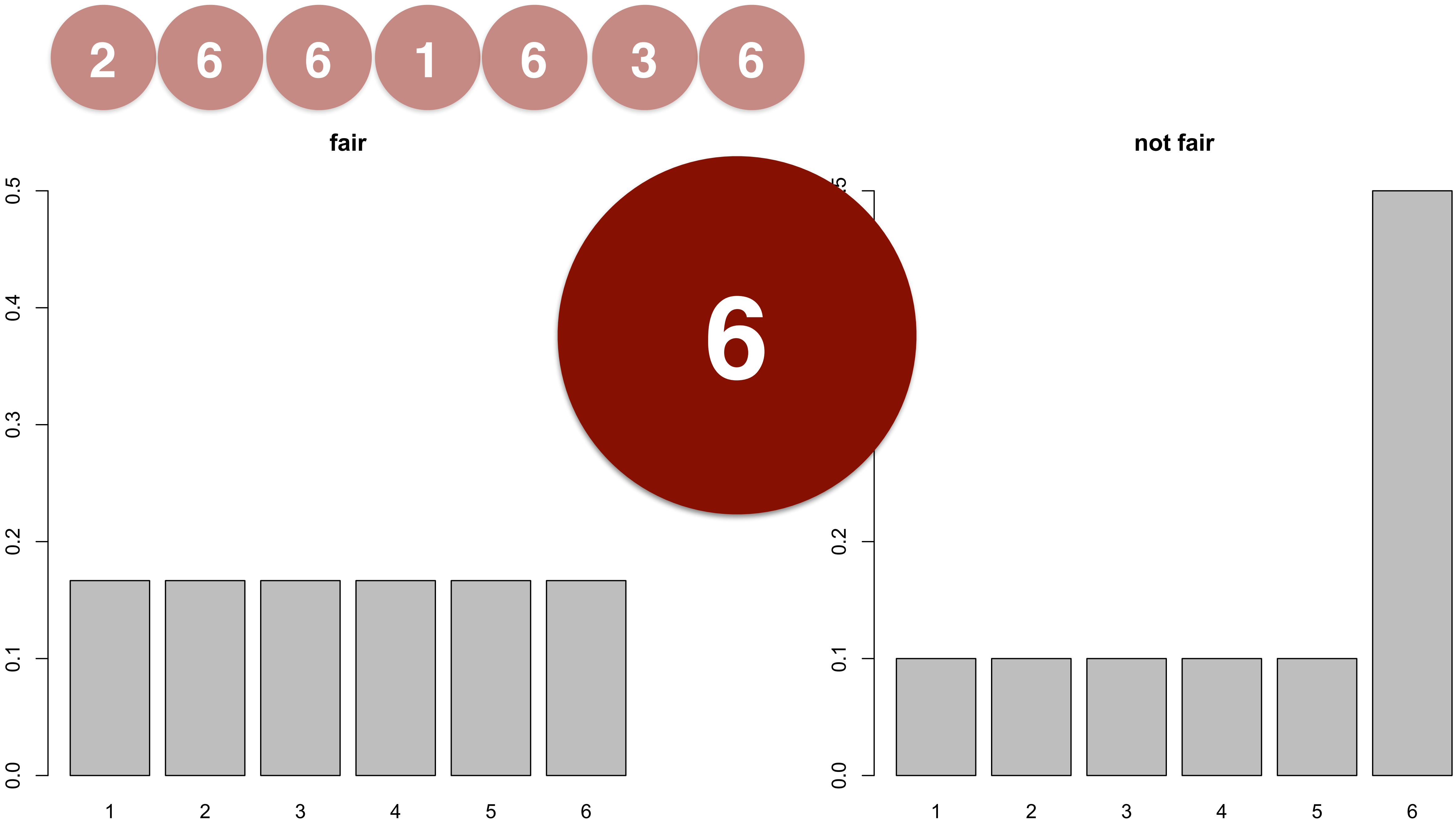
Probability



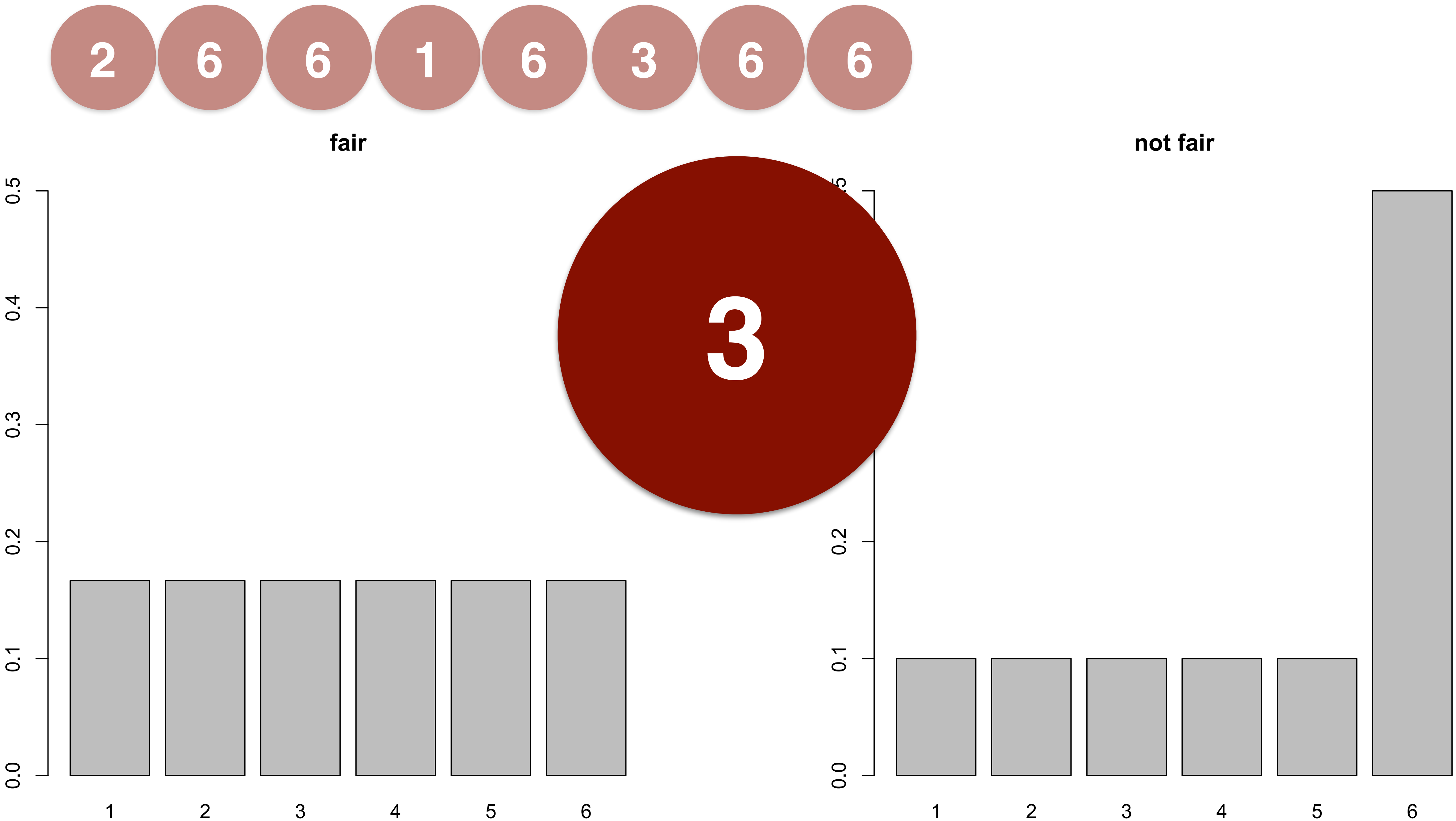
Probability



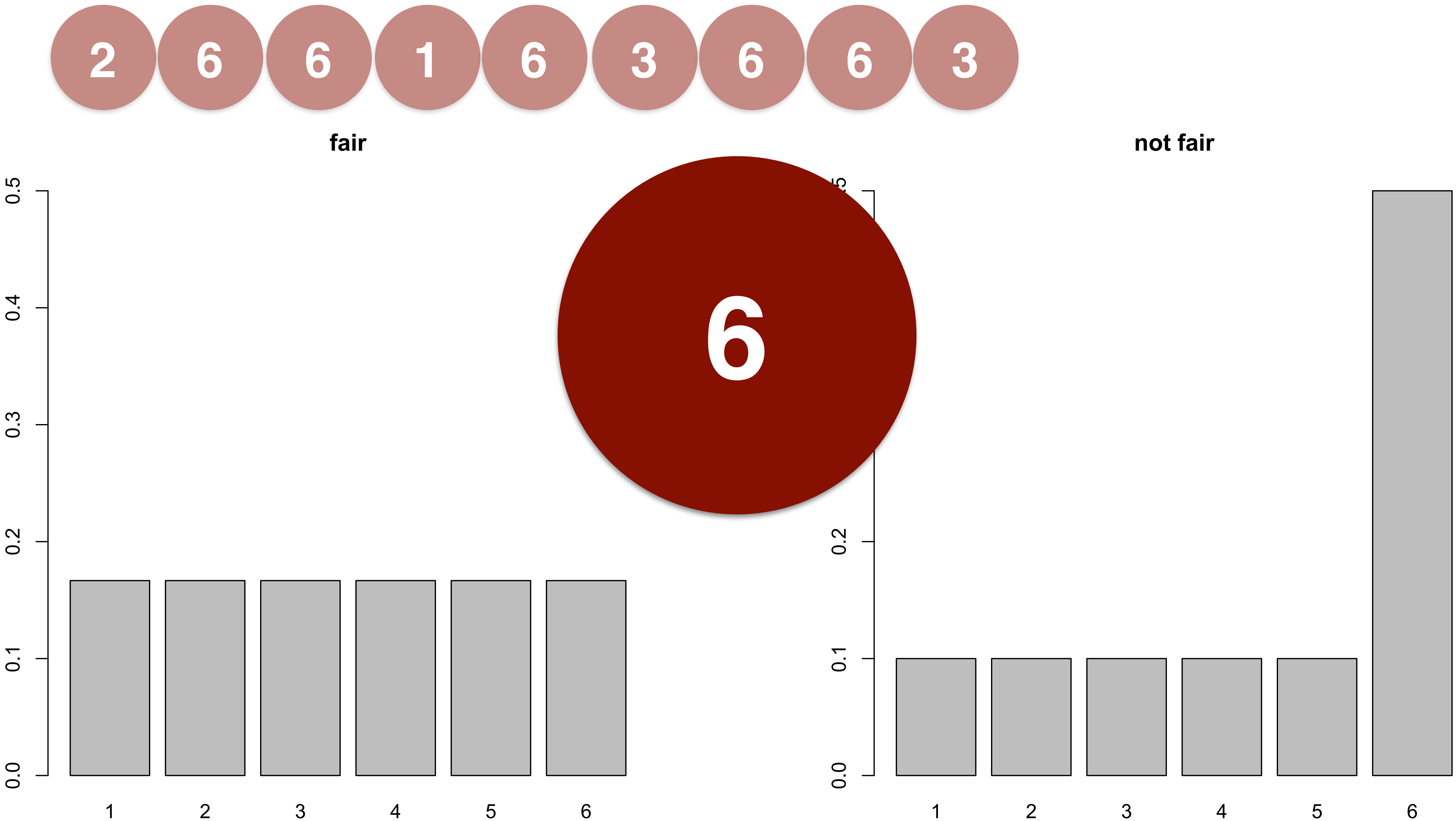
Probability



Probability

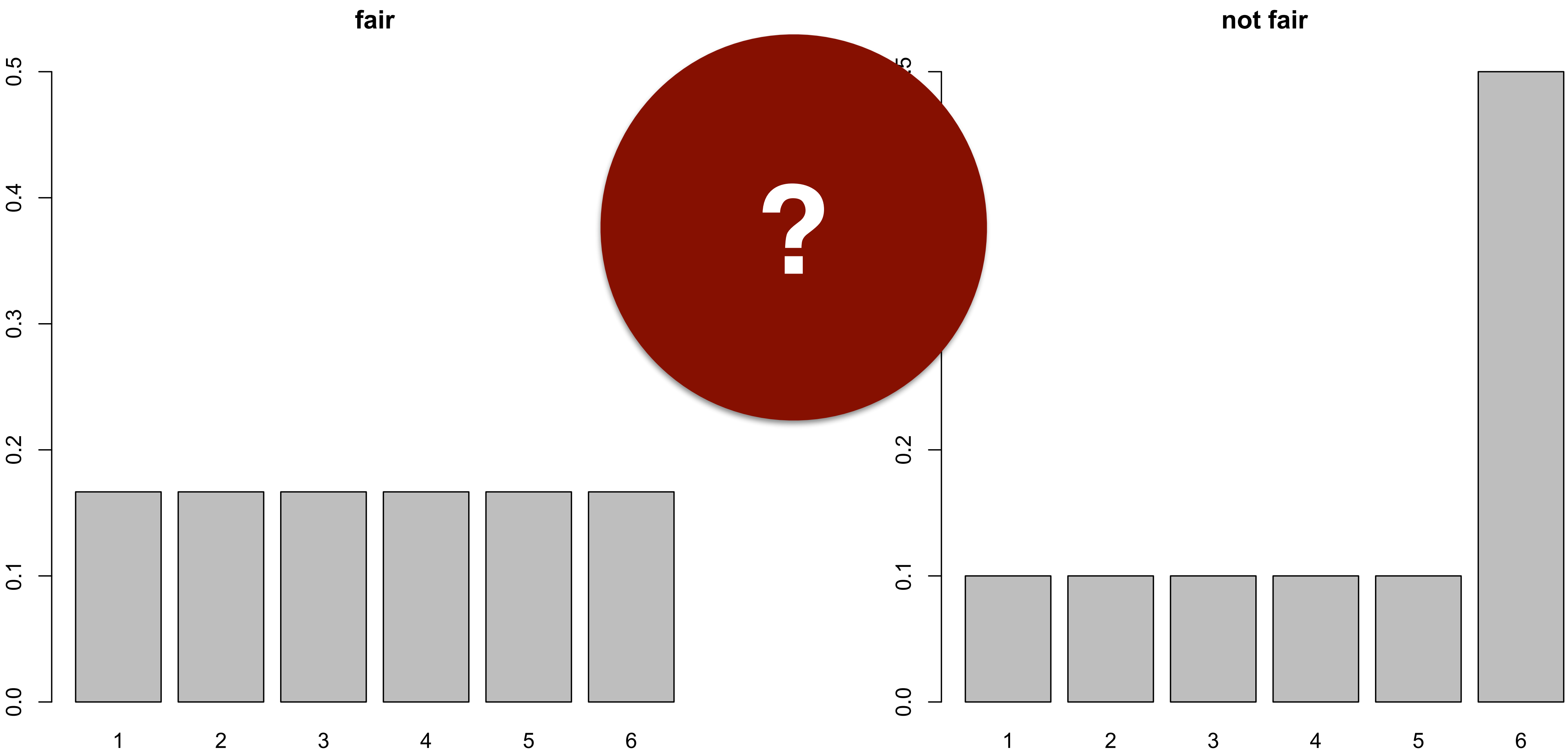


Probability



Probability

2 6 6 1 6 3 6 6 3 6



1

15,625

Independence

- Two random variables are independent if:

$$P(A, B) = P(A) \times P(B)$$

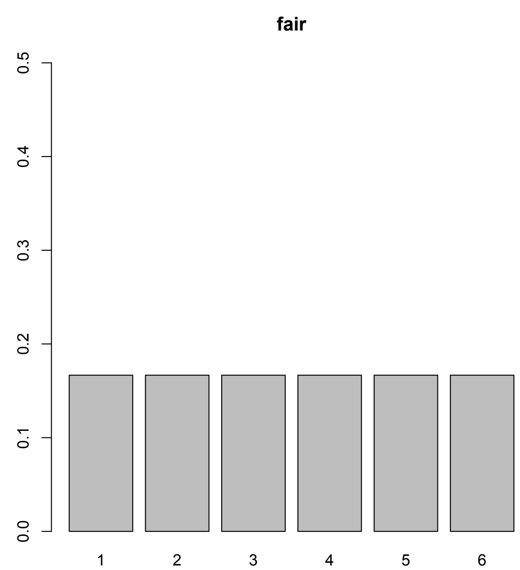
- In general:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{N} P(x_i)$$

- Information about one random variable (B) gives no information about the value of another (A)

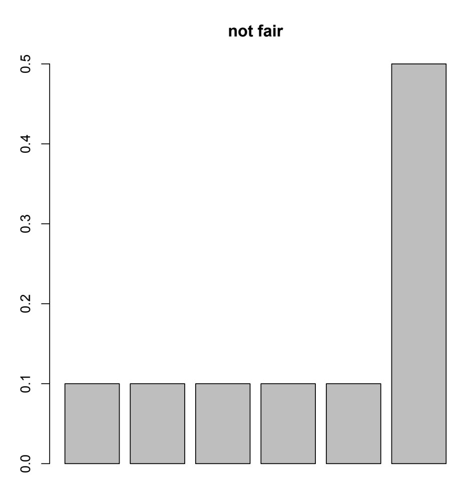
$$P(A) = P(A \mid B) \qquad P(B) = P(B \mid A)$$

Data Likelihood

P(2 6 6 | 

)

$= .17 \times .17 \times .17$
$= 0.004913$

P(2 6 6 | 

)

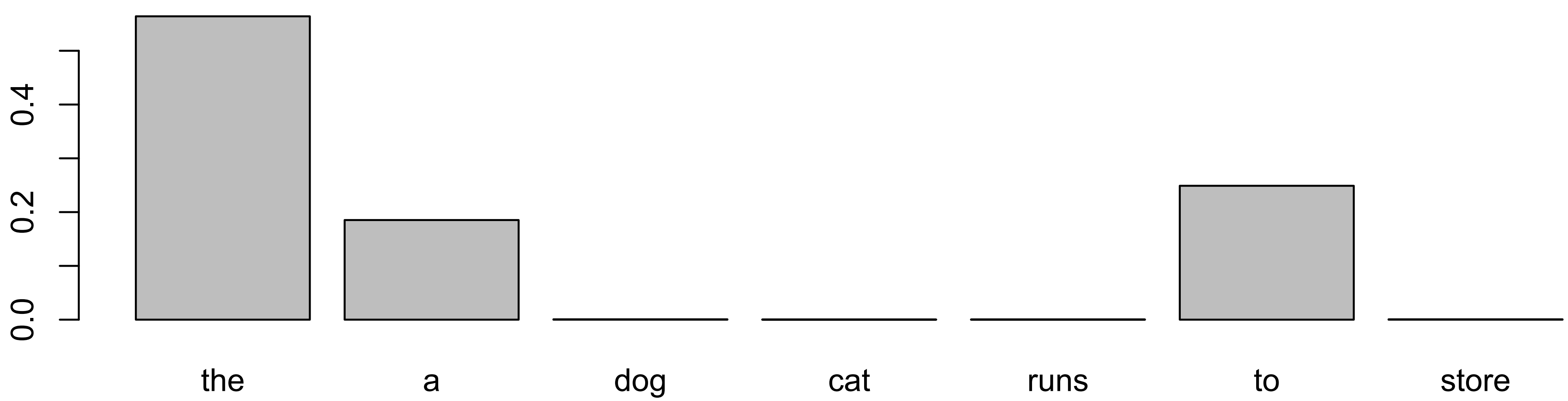
$= .1 \times .5 \times .5$
$= 0.025$

Data Likelihood

- The likelihood gives us a way of discriminating between possible alternative parameters, but also a strategy for picking a single best* parameter among all possibilities

Unigram probability

$$X \in \{the, a, dog, cat, runs, to, store\}$$



How do we calculate this?

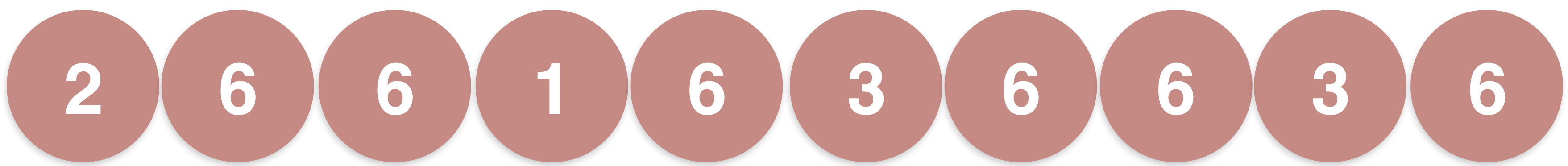
days Mr. Bingley returned Mr. Bennet's visit, and sat about ten minutes with him in his library. He had
ined hopes of being admitted to a sight of the young ladies, of whose beauty he had heard much; but
ly the father. The ladies were somewhat more fortunate, for they had the advantage of ascertaining fro
window that he wore a blue coat, and rode a black horse. An invitation to dinner was soon afterwards
ched; and already had Mrs. Bennet planned the courses that were to do credit to her housekeeping, w
wer arrived which deferred it all. Mr. Bingley was obliged to be in town the following day, and, conseq
to accept the honour of their invitation, etc. Mrs. Bennet was quite disconcerted. She could not imagi
usiness he could have in town so soon after his arrival in Hertfordshire; and she began to fear that he
ays flying about from one place to another, and never settled at Netherfield as he ought to be. Lady Lu
d her fears a little by starting the idea of his being gone to London only to get a large party for the ball
soon followed that Mr. Bingley was to bring twelve ladies and seven gentlemen with him to the assemb
ls grieved over such a number of ladies, but were comforted the day before the ball by hearing, that in
ve he brought only six with him from London, his five sisters and a cousin. And when the party entered
bly room it consisted of only five , the husband of the eldest, and
r young man. Mr. Bingley was g d a pleasant countenance, and
cted manners. His sisters were f ion. His brother-in-law, Mr. Hurs
looked the gentleman; but his f of the room by his fine, tall per
ome features, noble mien, and t on within five minutes after his
ce, of his having ten thousand a year. The gentlemen pronounced him to be a fine figure of a man, the
ed he was much handsomer than Mr. Bingley, and he was looked at with great admiration for about ha
g, till his manners gave a disgust which turned the tide of his popularity; for he was discovered to be p
bove his company, and above being pleased; and not all his large estate in Derbyshire could then sav
aving a most forbidding, disagreeable countenance, and being unworthy to be compared with his frie
gley had soon made himself acquainted with all the principal people in the room; he was lively and
rved, danced every dance, was angry that the ball closed so early, and talked of giving one himself at
field. Such amiable qualities must speak for themselves. What a contrast between him and his friend!
danced only once with Mrs. Hurst and once with Miss Bingley, declined being introduced to any other
ent the rest of the evening in walking about the room, speaking occasionally to one of his own party. H
ter was decided. He was the proudest, most disagreeable man in the world, and everybody hoped th
never come there again. Amongst the most violent against him was Mrs. Bennet, whose dislike of his

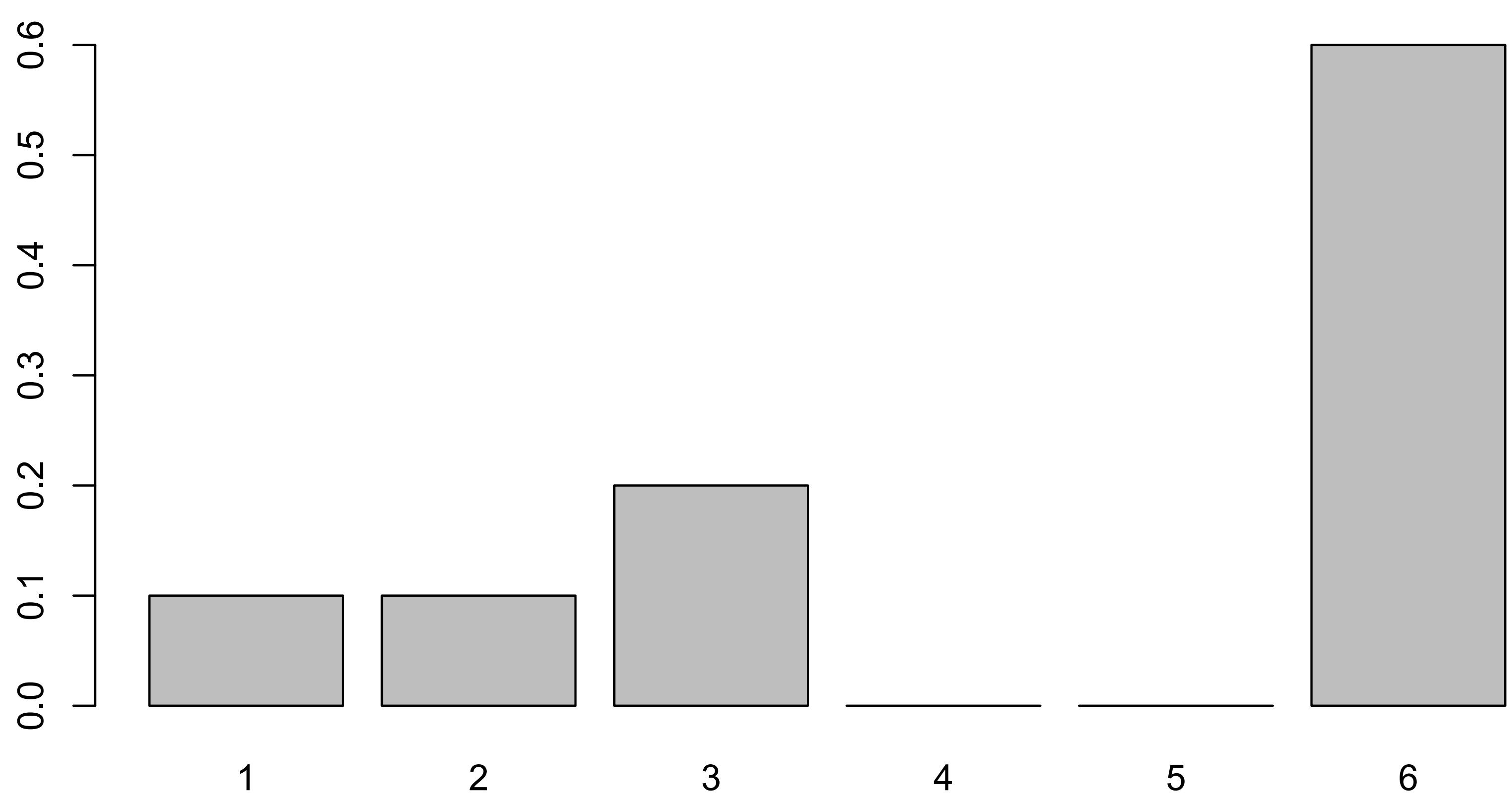
$P(X=\text{“the”}) = 28/536 = .052$

Maximum Likelihood Estimate

- This is a maximum likelihood estimate for P(X); the parameter values for which the data we observe (X) is most likely.

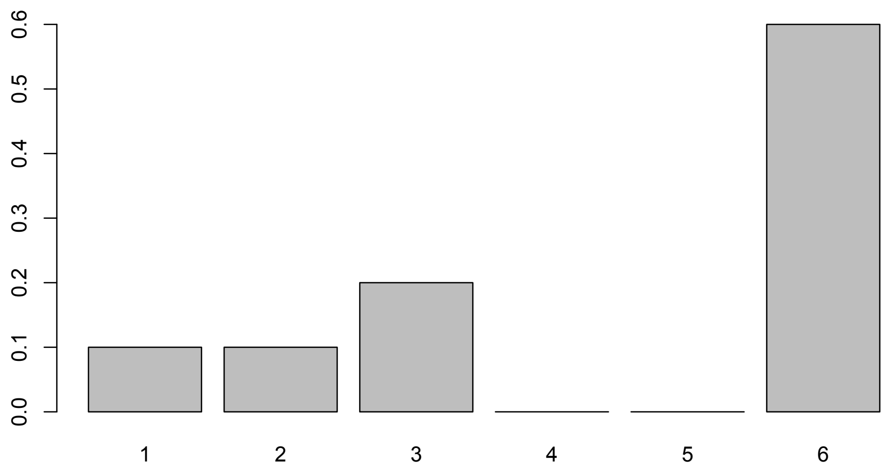
Maximum Likelihood Estimate





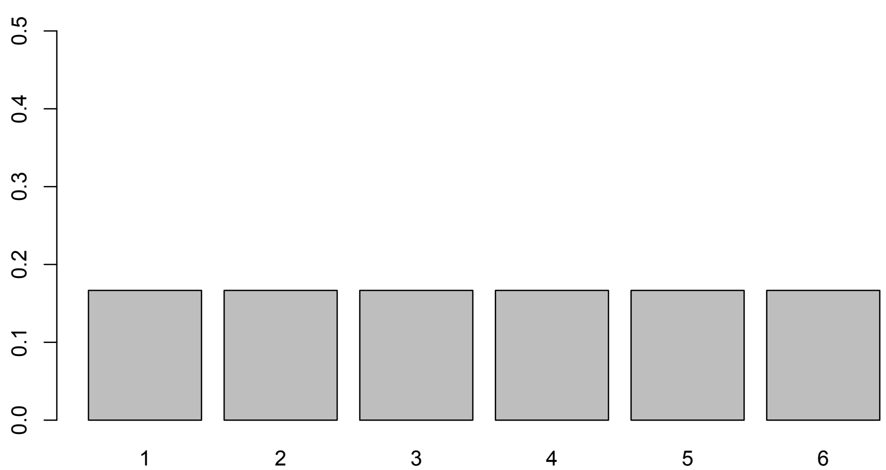
2 6 6 1 6 3 6 6 3 6

θ_1



$$P(X \mid \theta_1) = 0.0000311040$$

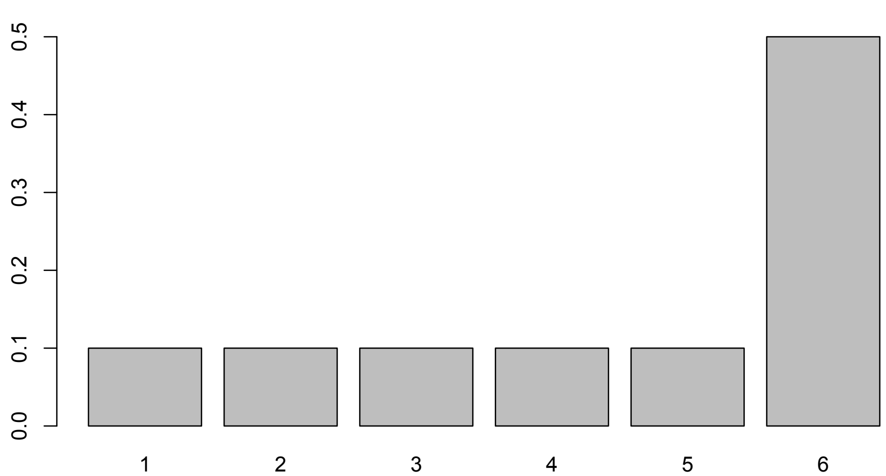
θ_2



$$P(X \mid \theta_2) = 0.0000000992$$

(313x less likely)

θ_3



$$P(X \mid \theta_3) = 0.0000031250$$

(10x less likely)

Conditional Probability

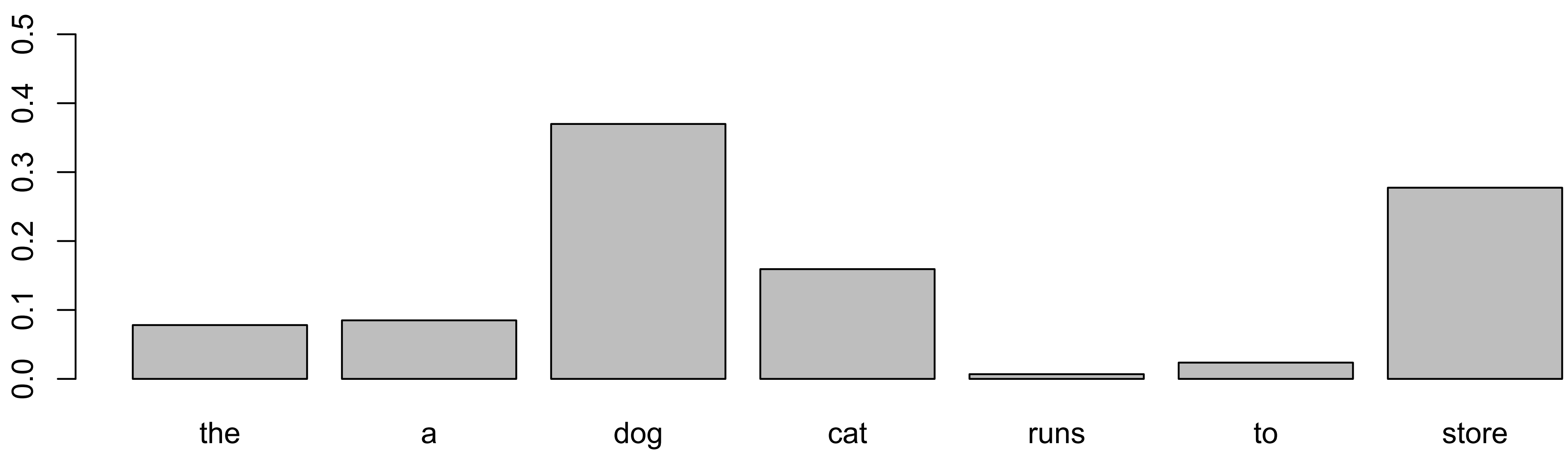
$$P(X = x|Y = y)$$

- Probability that one random variable takes a particular value *given* the fact that a different variable takes another

$$P(X_i = dog|X_{i-1} = the)$$

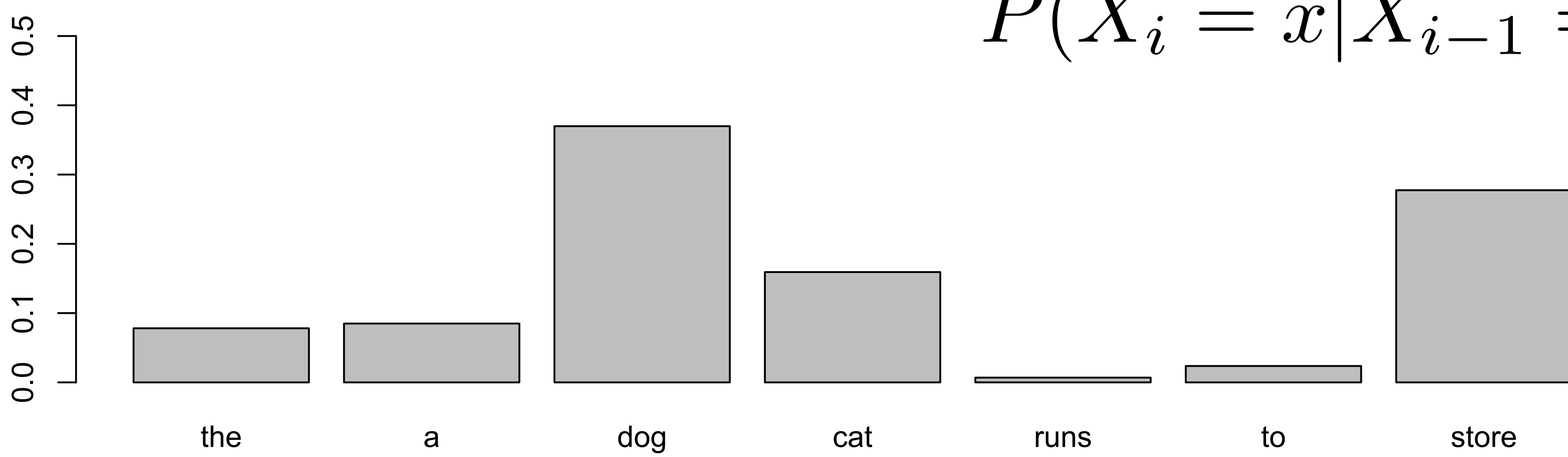
Conditional Probability

$$P(X_i = dog | X_{i-1} = the)$$

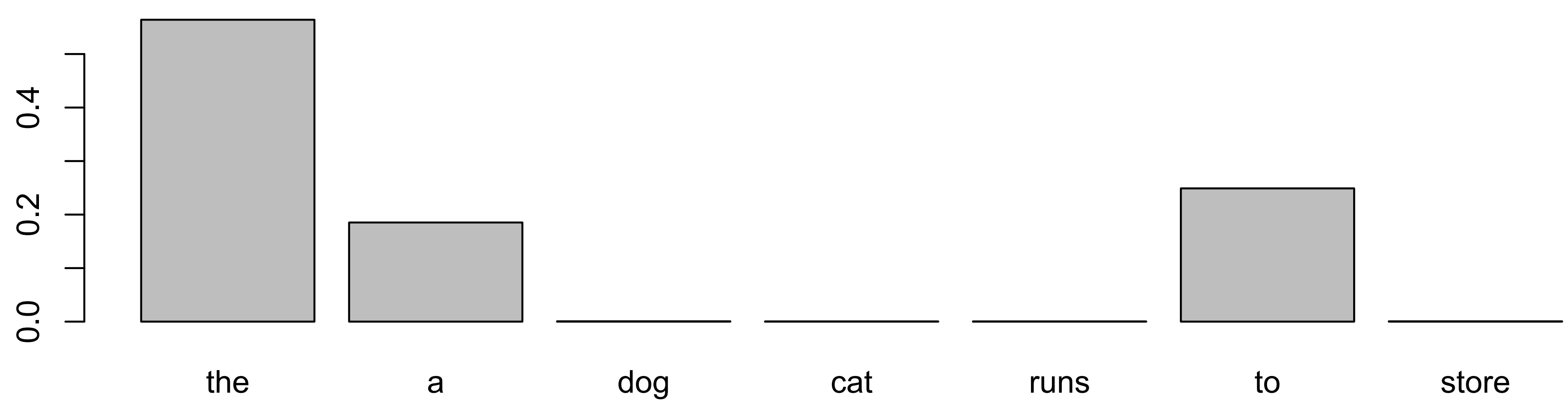


Conditional Probability

$$P(X_i = x | X_{i-1} = the)$$



$$P(X_i = x)$$



ined hopes of being admitted to a sight of the **young** ladies, of whose beauty he had heard much; bu
ly the **father**. The **ladies** were somewhat more fortunate, for they had the **advantage** of ascertaining f
window that he wore a blue coat, and rode a black horse. An invitation to dinner was soon afterwards
ched; and already had Mrs. Bennet planned the **courses** that were to do credit to her housekeeping, v
wer arrived which deferred it all. Mr. Bingley was obliged to be in town the **following** day, and,
quently, unable to accept the **honour** of their invitation, etc. Mrs. Bennet was quite disconcerted. She
gine what business he could have in town so soon after his arrival in Hertfordshire; and she began to
might be always flying about from one place to another, and never settled at Netherfield as he ought
ucas quieted her fears a little by starting the **idea** of his being gone to London only to get a large part
l; and a report soon followed that Mr. Bingley was to bring twelve ladies and seven gentlemen with him
sembly. The **girls** grieved over such a number of ladies, but were comforted the **day** before the **ball** by
g, that instead of twelve he brought only six with him from London--his five sisters and a cousin. And w
rty entered the **assembly** ro gley, his two sisters, the husb
eldest, and another young m manlike; he had a pleasant
nance, and easy, unaffectec n air of decided fashion. His
-in-law, Mr. Hurst, merely loo soon drew the **attention** of th
y his fine, tall person, hands ch was in general circulation
ive minutes after his entranc **tlemen** pronounced him to b
ure of a man, the **ladies** declared he was much handsomer than Mr. Bingley, and he was looked at wi
dmiration for about half the evening, till his manners gave a disgust which turned the **tide** of his popu
was discovered to be proud; to be above his company, and above being pleased; and not all his large
in Derbyshire could then save him from having a most forbidding, disagreeable countenance, and bei
hy to be compared with his friend. Mr. Bingley had soon made himself acquainted with all the **princip**
in the **room**; he was lively and unreserved, danced every dance, was angry that the **ball** closed so e
ked of giving one himself at Netherfield. Such amiable qualities must speak for themselves. What a co
en him and his friend! Mr. Darcy danced only once with Mrs. Hurst and once with Miss Bingley, decline
ntroduced to any other lady, and spent the **rest** of the **evening** in walking about the **room**, speaking
onally to one of his own party. His character was decided. He was the **proudest**, most disagreeable m
rld, and everybody hoped that he would never come there again. Amongst the **most** violent against hi
ennet, whose dislike of his general behaviour was sharpened into particular resentment by his having

$P(X_i=\text{"room"} \mid X_{i-1}=\text{"the"}) = 2/28 = .071$

Conditional Probability

$$P(X = vampire) \text{ vs. } P(X = vampire|Y = horror)$$

$$P(X = manners|Y = austen) \text{ vs. } P(X = whale|Y = austen)$$

0.00036 0

$$P(X = manners|Y = austen) \text{ vs. } P(X = manners|Y = dickens)$$

0.00036 = 6.7x times more than 0.000053

Authorship Attribution

"Mr. Collins was not a sensible man"

Independence Assumption

"Mr. Collins was not a sensible man"

$$P(x_i = \text{Mr.}, x_2 = \text{Collins}) = P(x_i = \text{Mr.}) \times P(x_2 = \text{Collins})$$

This is certainly untrue in this case, because the presence of Mr. makes Collins more likely (they are dependent)

Independence Assumption

"Mr. Collins was not a sensible man"

We will assume the features are independent:

$$P(x_1, x_2, x_3, x_4, x_6, x_7 \mid c) = P(x_1 \mid c) P(x_2 \mid c) \ldots P(x_7 \mid c)$$

$$P(x_i \ldots x_n \mid c) = \prod_{i=1}^{N} P(x_i \mid c)$$

A simple classifier

"Mr. Collins was not a sensible man"

Austen		Dickens	
P(X=Mr. | Y=Austen)	0.0084	P(X=Mr. | Y=Dickens)	0.00421
P(X=Collins | Y=Austen)	0.00036	P(X=Collins | Y=Dickens)	0.000016
P(X=was | Y=Austen)	0.01475	P(X=was | Y=Dickens)	0.015043
P(X=not | Y=Austen)	0.01145	P(X=not | Y=Dickens)	0.00547
P(X=a | Y=Austen)	0.01591	P(X=a | Y=Dickens)	0.02156
P(X=sensible | Y=Austen)	0.00025	P(X=sensible | Y=Dickens)	0.00005
P(X=man | Y=Austen)	0.00121	P(X=man | Y=Dickens)	0.001707

A simple classifier

"Mr. Collins was not a sensible man"

P(X = "Mr. Collins was not a sensible man" | Y = Austen)

= P("Mr" | Austen) × P("Collins" | Austen) ×
P("was" | Austen) × P("not" | Austen) …
= 0.000000022507322 (**≈ 2.3 × 10^{-8}**)

P(X = "Mr. Collins was not a sensible man" | Y = Dickens)

P("Mr" | Dickens) × P("Collins" | Dickens) ×
P("was" | Dickens) × P("not" | Dickens) …
= 0.000000002078906 (**≈ 2.1 × 10^{-9}**)

A simple classifier

- The classifier we just specified is a maximum likelihood classifier, where compare the likelihood of the data under each class and choose the class with the highest likelihood

Likelihood: probability of data (here, under class y)

$$P(X = x_i \ldots x_n \mid Y = y)$$

Prior probability of class y

$$P(Y = y)$$

Bayes' Rule

Prior belief that Y = y
(before you see any data)

Likelihood of the data
given that Y=y

$$P(Y = y|X = x) = \frac{P(Y = y)P(X = x|Y = y)}{\sum_y P(Y = y)P(X = x|Y = y)}$$

Posterior belief that Y=y given that X=x

Bayes' Rule

Prior belief that Y = Austen (before you see any data)

Likelihood of "Mr. Collins was not a sensible man" given that Y= Austen

$$P(Y = y|X = x) = \frac{P(Y = y)P(X = x|Y = y)}{\sum_y P(Y = y)P(X = x|Y = y)}$$

Posterior belief that Y=Austen given that X="Mr. Collins was not a sensible man"

This sum ranges over y=Austen + y=Dickens (so that it sums to 1)

Likelihood: probability of data (here, under class y)

$$P(X = x_i \ldots x_n \mid Y = y)$$

Prior probability of class y

$$P(Y = y)$$

Posterior belief in the probability of class y after seeing data

$$P(Y = y \mid X = x_i \ldots x_n)$$

Naive Bayes Classifier

$$\frac{P(Y = Austen)P(X = \text{“}Mr...\text{”}|Y = Austen)}{P(Y = Austen)P(X = \text{“}Mr...\text{”}|Y = Austen) + P(Y = Dickens)P(X = \text{“}Mr...\text{”}|Y = Dickens)}$$

Let's say P(Y=Austen) = P(Y=Dickens) = 0.5
(i.e., both are equally likely a priori)

$$= \frac{0.5 \times (2.3 \times 10^{-8})}{0.5 \times (2.3 \times 10^{-8}) + 0.5 \times (2.1 \times 10^{-9})}$$

$$P(Y = Austen|X = \text{“}Mr...\text{”}) = 91.5\%$$

$$P(Y = Dickens|X = \text{“}Mr...\text{”}) = 8.5\%$$

Taxicab Problem

“A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

- 85% of the ca   Blue.
- A witness ide   d the reliability of the witness u   ted on the night of the accident   ly identified each one of the two colors 80% of the time and failed 20% of the time.

“Base rate fallacy”
Don’t ignore prior information!

What is the probability that the cab involved in the accident was Blue rather than Green knowing that this witness identified it as Blue?”

(Tversky & Kahneman 1981)

Prior Belief

- Now let's assume that Dickens published 1000 times more books than Austen.

  - P(Y= Austen) = 0.000999
  - P(Y = Dickens) = 0.999001

$$\frac{0.000999 \times (2.3 \times 10^{-8})}{0.000999 \times (2.3 \times 10^{-8}) + 0.999001 \times (2.1 \times 10^{-9})}$$

$$P(Y = Austen|X) = 0.011$$

$$P(Y = Dickens|X) = 0.989$$

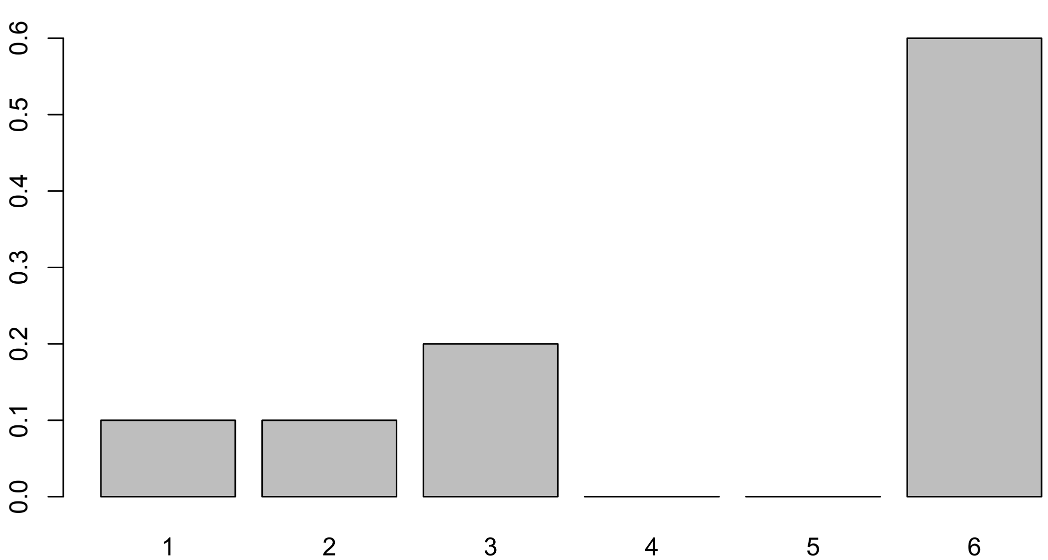
Priors

- Priors can be informed (reflecting expert knowledge) but in practice, but priors in Naive Bayes are often simply estimated from training data

$$P(Y = \text{Austen}) = \frac{\text{\# of Austen texts}}{\text{\# of total texts}}$$

Smoothing

- Maximum likelihood estimates can fail miserably when features are never observed with a particular class.



What's the probability of:

Smoothing

- One solution: add a little probability mass to every element.

maximum likelihood estimate

$$P(x_i \mid y) = \frac{n_{i,y}}{n_y}$$

$n_{i,y}$ = count of word i in class y
n_y = number of words in y
V = size of vocabulary

smoothed estimates

$$P(x_i \mid y) = \frac{n_{i,y} + \alpha}{n_y + V\alpha}$$

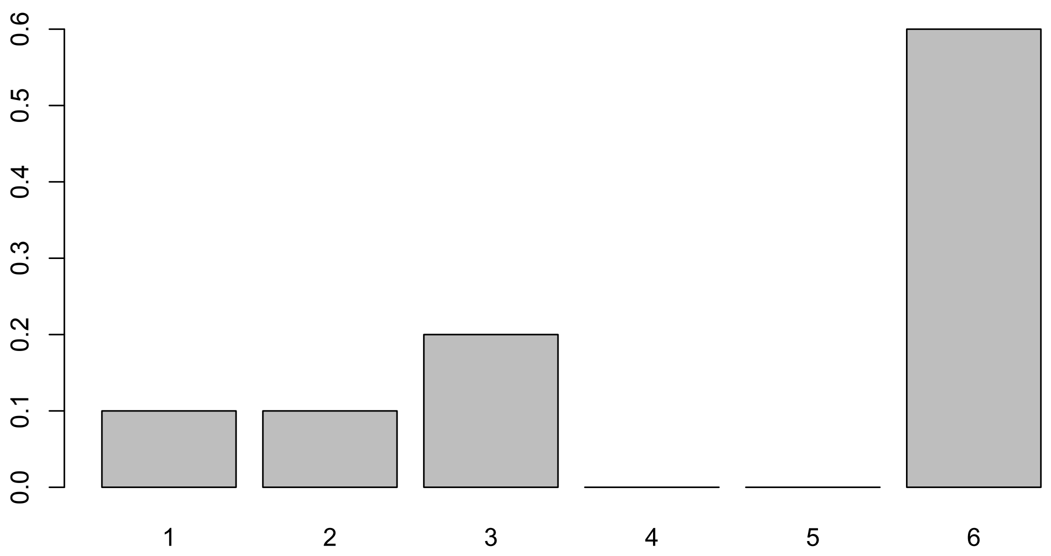
same α for all x_i

$$P(x_i \mid y) = \frac{n_{i,y} + \alpha_i}{n_y + \sum_{j=1}^{V} \alpha_j}$$

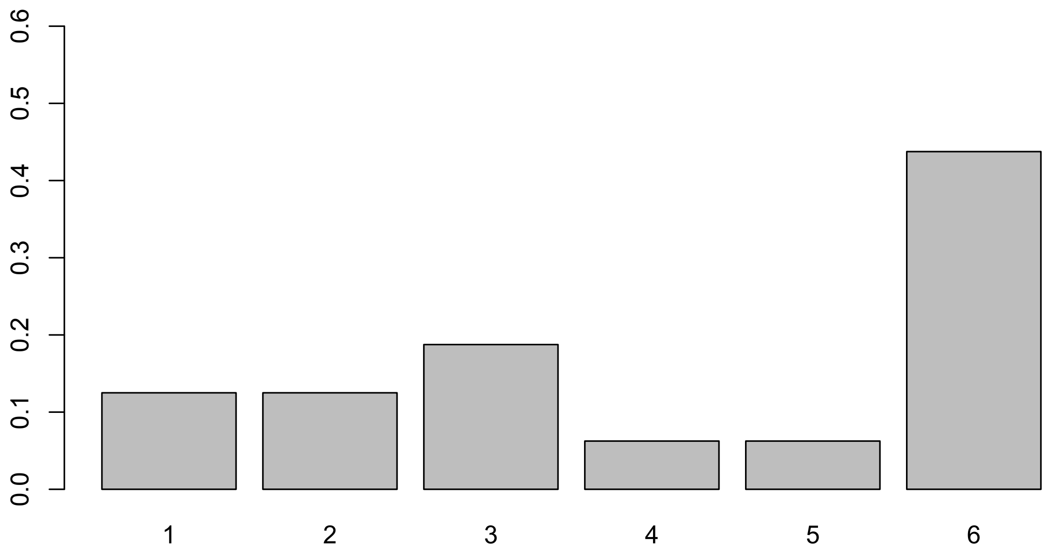
possibly different α for each x_i

Smoothing

MLE



smoothing with $\alpha = 1$



Naive Bayes training

Training a Naive Bayes classifier consists of estimating these two quantities from training data for all classes y

$$P(Y=y|X=x) = \frac{P(Y=y)P(X=x|Y=y)}{\sum_y P(Y=y)P(X=x|Y=y)}$$

At test time, use those estimated probabilities to calculate the posterior probability of each class y and select the class with the highest probability

Naive Bayes

- We’ve just described Naive Bayes with a multinomial distribution, but any probability distribution can be modeled as well.

Probability distributions

Normal

Gamma

Poisson

Geometric

Exponential

Multinomial

Bernoulli

Beta

Binomial

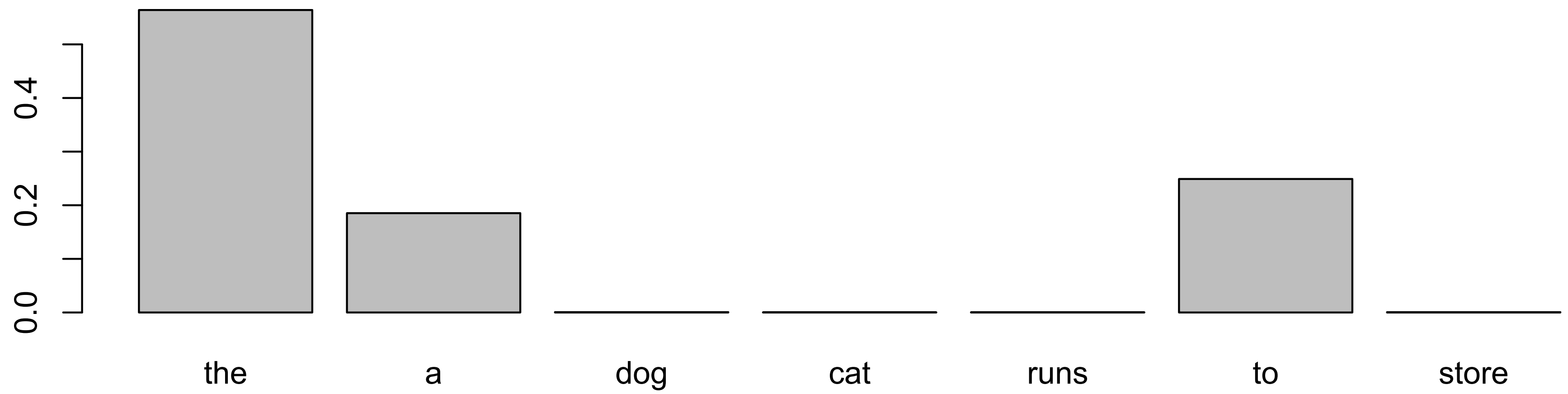
Uniform

Dirichlet

Multinomial

Discrete distribution for modeling count data (e.g., word counts; single parameter θ

$\theta =$



the	a	dog	cat	runs	to	store
3	1	0	1	0	2	0
531	209	13	8	2	331	1

Multinomial

Maximum likelihood parameter estimate

$$\hat{\theta}_i = \frac{n_i}{N}$$

	the	a	dog	cat	runs	to	store
count n	531	209	13	8	2	331	1
θ	0.48	0.19	0.01	0.01	0.00	0.30	0.00

Bernoulli

- Binary event (true or false; {0, 1})
- One parameter: p (probability of an event occurring)

$$P(x = 1 \mid p) = p$$

$$P(x = 0 \mid p) = 1 - p$$

Examples:

- Probability of a particular feature being true (e.g., self-reported location = Berkeley)

$$\hat{p}_{mle} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Bernoulli

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	p_{MLE}
f_1	1	0	0	0	1	1	0	0	0.375
f_2	0	0	0	0	0	0	1	0	0.125
f_3	1	1	1	1	1	0	0	1	0.750
f_4	1	0	0	1	1	0	0	1	0.500
f_5	0	0	0	0	0	0	0	0	0.000

Bernoulli

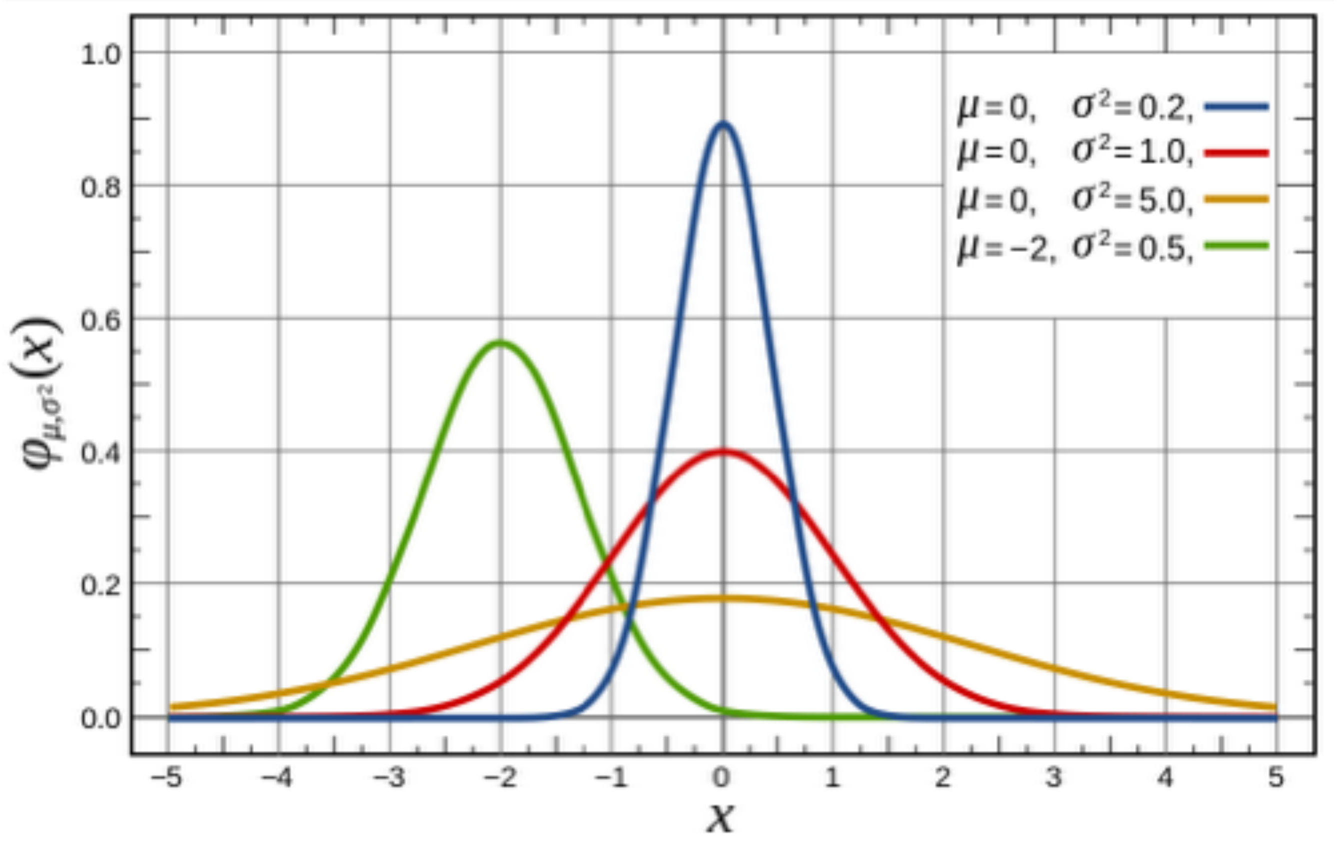
	Republican				Democrat					
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	$p_{MLE,R}$	$p_{MLE,D}$
f_1	1	0	0	0	1	1	0	0	0.25	0.50
f_2	0	0	0	0	0	0	1	0	0.00	0.25
f_3	1	1	1	1	1	0	0	1	1.00	0.50
f_4	1	0	0	1	1	0	0	1	0.50	0.50
f_5	0	0	0	0	0	0	0	0	0.00	0.00

Normal

- continuous (-∞, ∞)
- μ (mean) (-∞, ∞)
- σ^2 (variance) > 0

Examples:

- Age
- Height



$$P(x = -2 \mid \mu = -2, \sigma^2 = 0.5) = 0.56$$

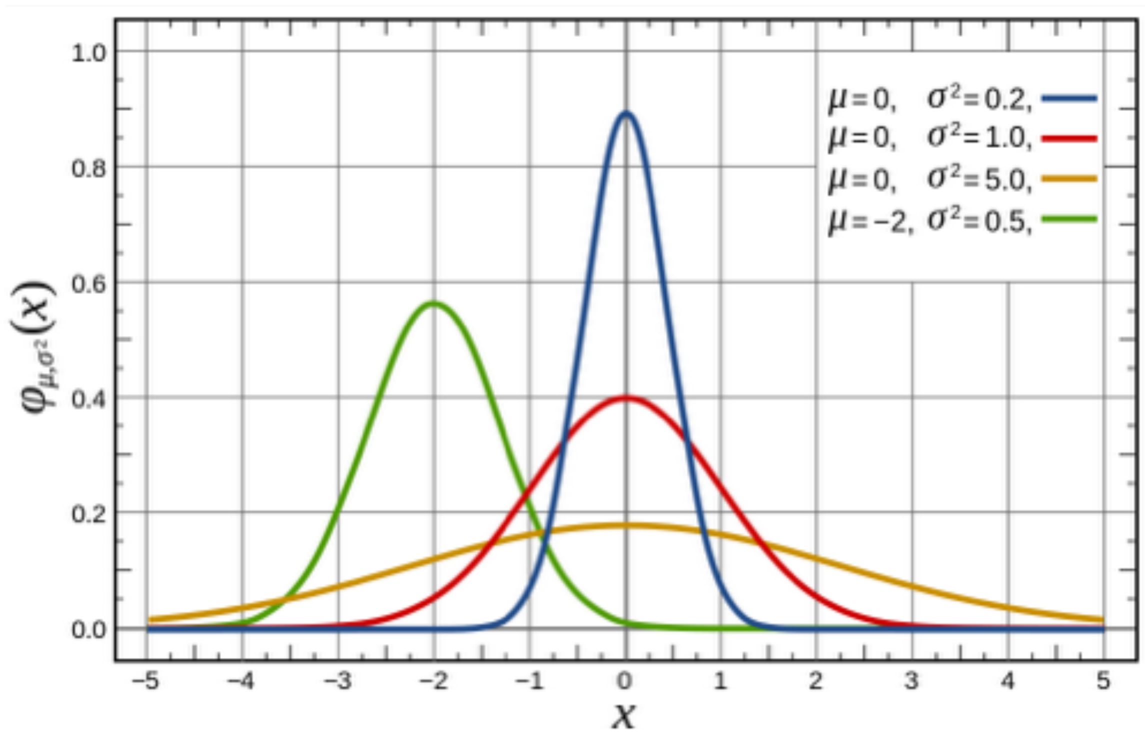
$$P(x = -2 \mid \mu = 0, \sigma^2 = 1) = 0.05$$

Normal

Maximum likelihood parameter estimates

$$\hat{\mu}_{mle} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2_{mle} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2$$



Normal

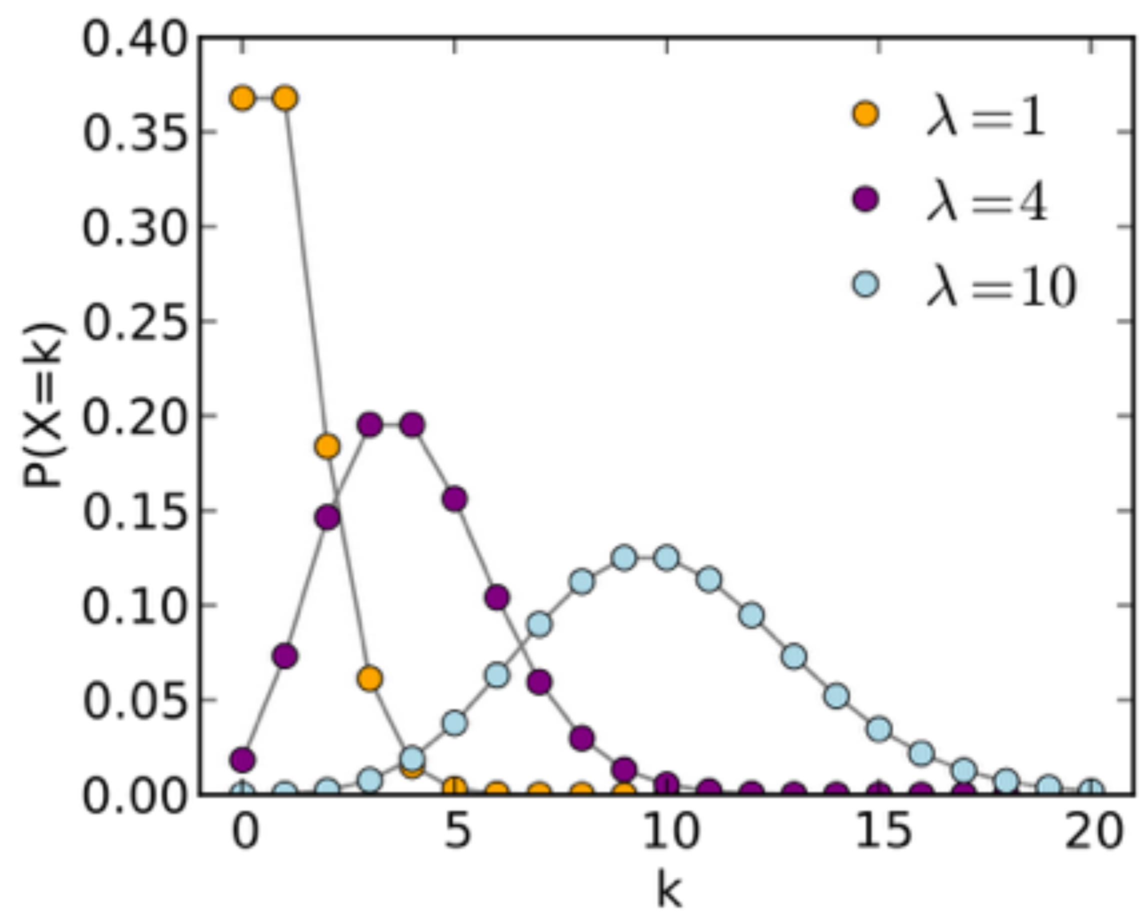
	Republican				Democrat					
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	$\mu_{MLE,R}$	$\mu_{MLE,D}$
f_1	3.4	-2.1	5.2	7.6	11.6	9.1	9.7	10.8	3.5	10.3
f_2	-0.3	8.5	5.6	11.5	5.4	6.2	3.1	12.7	6.3	6.8
f_3	-0.6	3.7	1.2	5.6	3.4	-4.4	8.0	6.2	2.5	3.3
f_4	2.5	6.7	0.5	2.6	13.2	6.1	13.7	7.7	3.1	10.2
f_5	7.0	5.0	5.6	16.3	15.4	14.9	2.3	6.3	8.5	9.7

Poisson

- discrete (0, 1, 2, …)
- λ > 0
- Models the number of events within a fixed interval of time

Examples:

- Number of emails in one hour
- Number of children in family



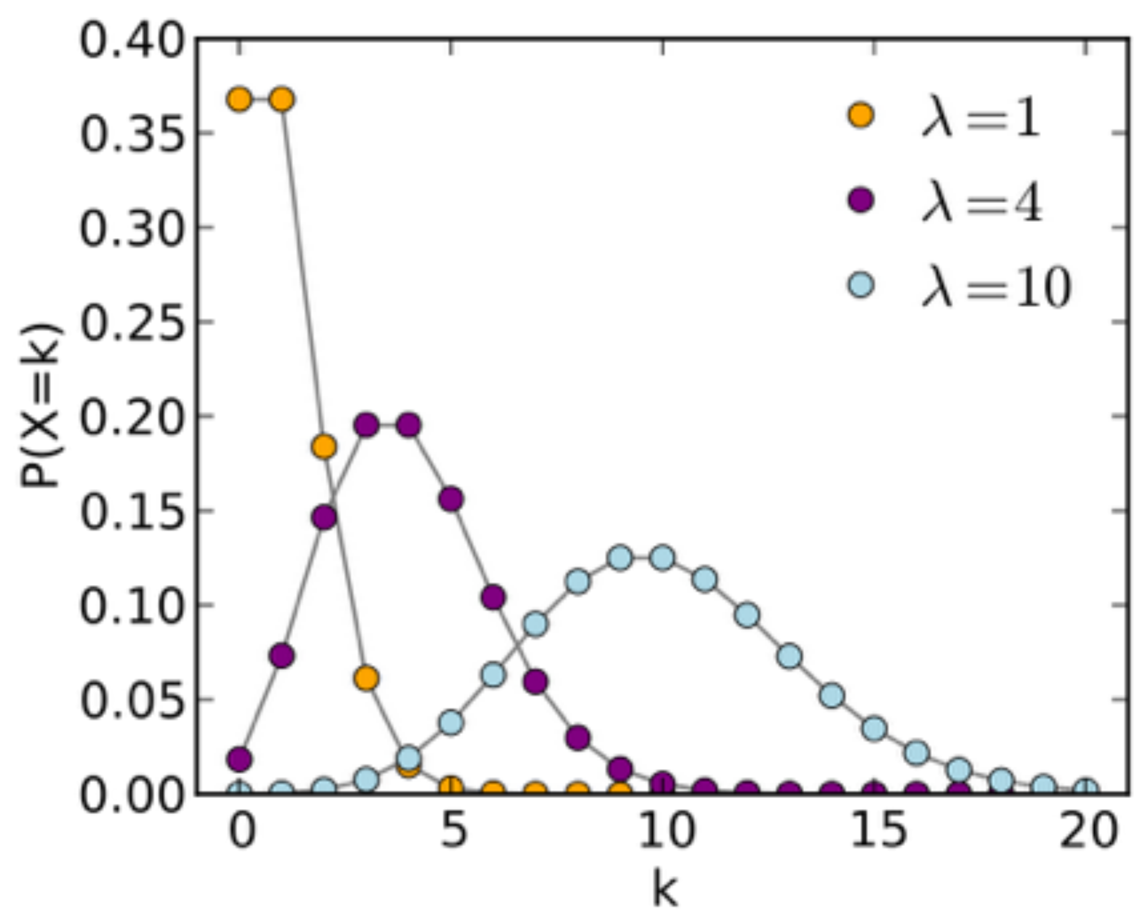
$$P(x = 4|\lambda = 10) = 0.02$$

$$P(x = 4|\lambda = 4) = 0.20$$

Poisson

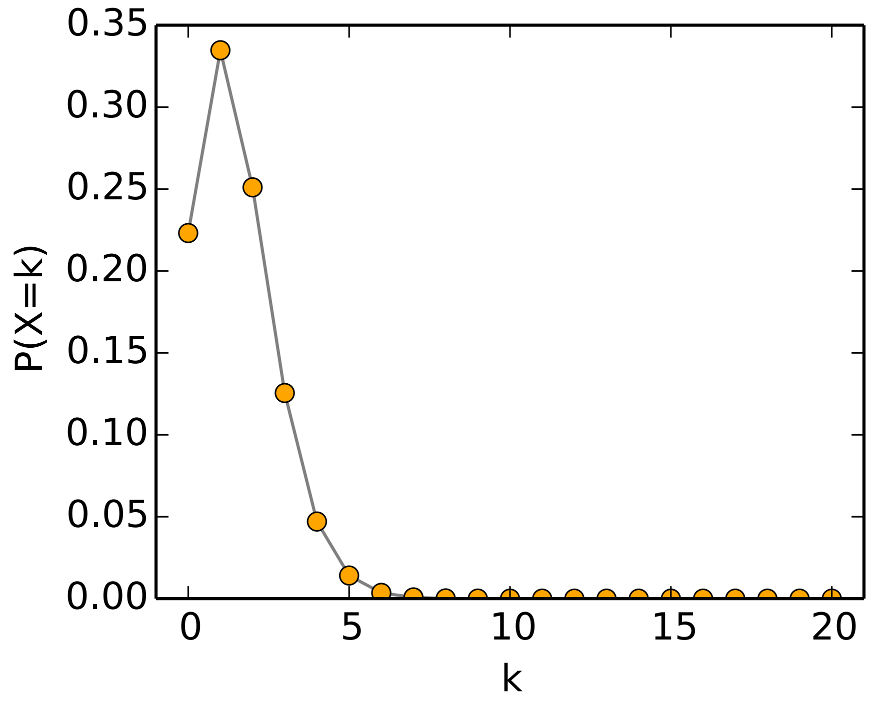
Maximum likelihood parameter estimate

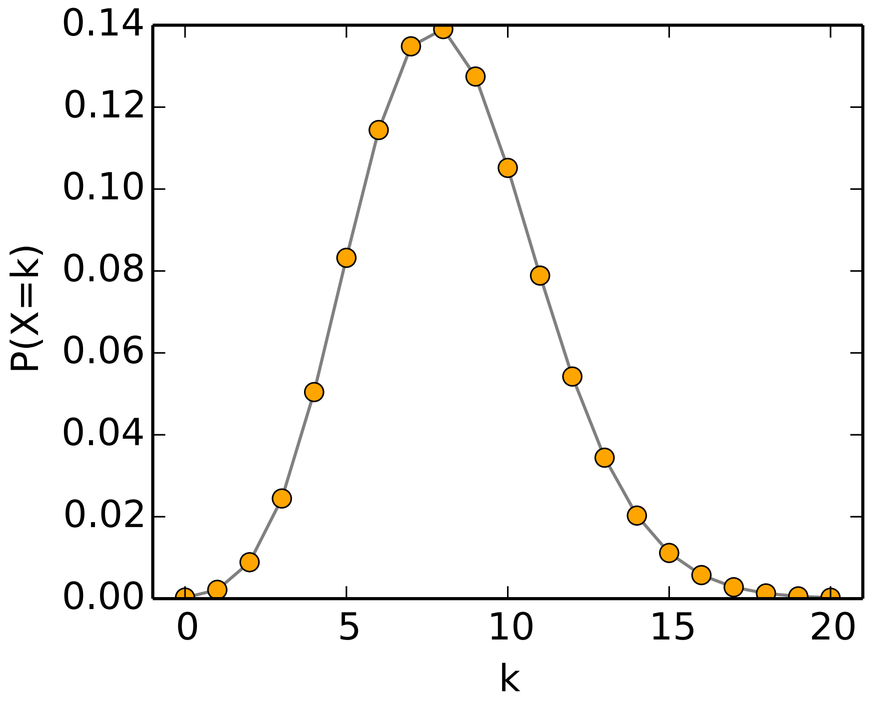
$$\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} x_i$$



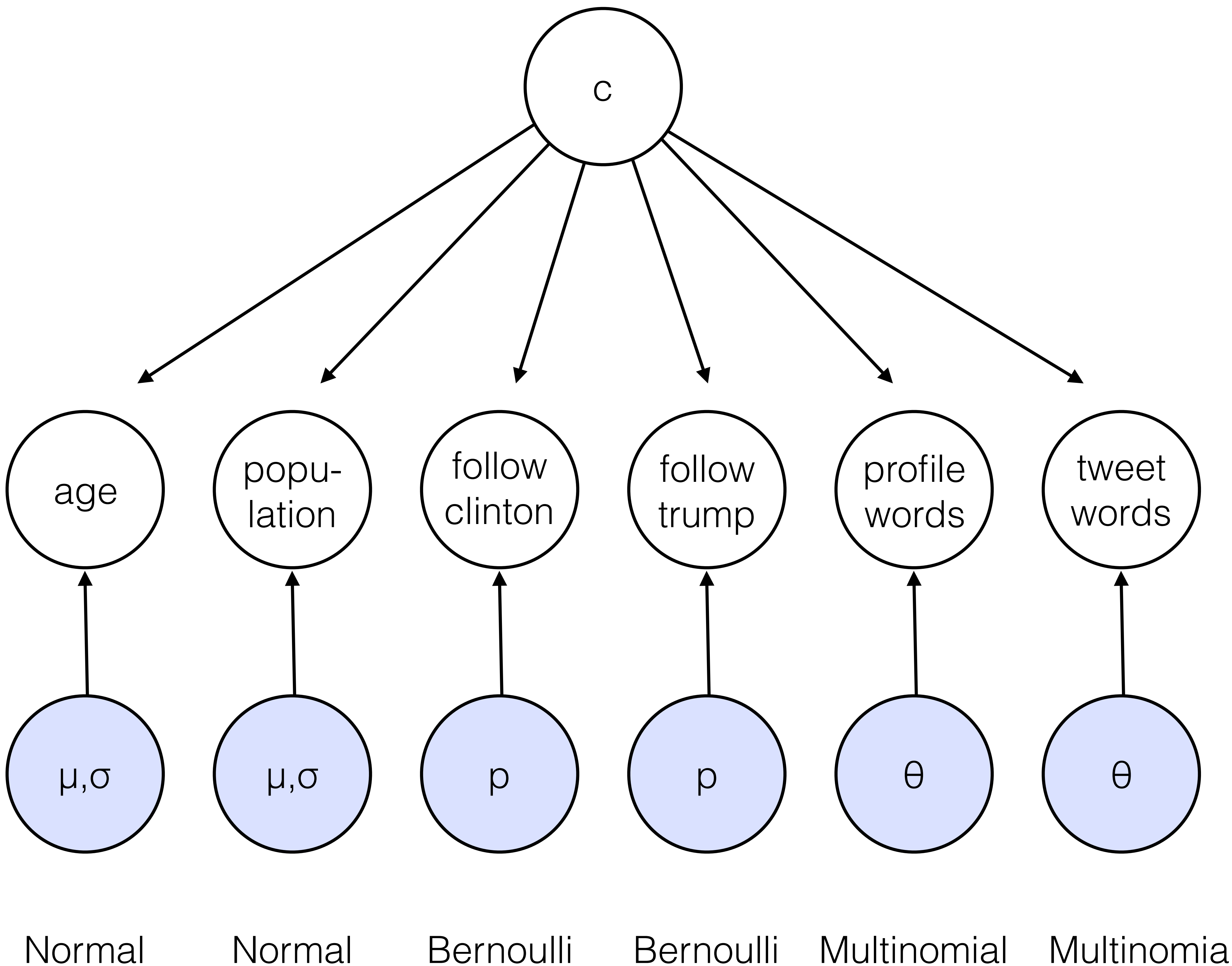
Poisson

	Republican				Democrat					
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	$\lambda_{MLE,R}$	$\lambda_{MLE,D}$
f_1	1	2	2	1	6	10	8	9	1.5	8.25





Feature	Value	Distribution?
follow clinton	0	
follow trump	0	
age	24	
word counts in profile	Berkeley, liberal, runner	
word counts in profile	the, election, a, data, movies	
population size of your city	116,000	


c
age
popu-
lation
follow
clinton
follow
trump
profile
words
tweet
words
μ,σ
μ,σ
p
p
θ
θ
Normal
Normal
Bernoulli
Bernoulli
Multinomial
Multinomial

$$P(X \mid c = \text{Dem}) = \prod_{i=1}^{N} P(X_i \mid c = \text{Dem})$$

$$= \text{Norm}(age \mid \mu_{age,dem}, \sigma^2_{age,dem})$$
$$\times\ \text{Norm}(population \mid \mu_{population,dem}, \sigma^2_{population,dem})$$
$$\times\ \text{Bernoulli}(followClinton \mid p_{followClinton,dem})$$
$$\times\ \text{Bernoulli}(followTrump \mid p_{followTrump,dem})$$
$$\times\ \text{Multinomial}(w_{profile} \mid \theta_{profile,dem})$$
$$\times\ \text{Multinomial}(w_{tweets} \mid \theta_{tweets,dem})$$

$$P(c = \text{Dem} \mid X) = \frac{P(c = \text{Dem}) \times P(X \mid c = \text{Dem})}{P(c = \text{Dem}) \times P(X \mid c = \text{Dem}) + P(c = \text{Rep}) \times P(X \mid c = \text{Rep})}$$

Authorship Attribution

Koppel et al. (2009), Computational Methods in Authorship Attribution (JASIST)

Representation

FW	A list of 512 function words, including conjunctions, prepositions, pronouns, modal verbs, determiners, and numbers (purely stylistic)
POS	Thirty-eight part-of-speech unigrams and 1,000 most common bigrams using the Brill (1992) part-of-speech tagger (purely stylistic)
SFL	All 372 nodes in SFL trees for conjunctions, prepositions, pronouns, and modal verbs (purely stylistic)
CW	The 1,000 words with highest information gain (Quinlan, 1986) in the training corpus among the 10,000 most common words in the corpus
CNG	The 1,000 character trigrams with highest information gain in the training corpus among the 10,000 most common trigrams in the corpus (cf. Keselj, 2003)

Models

NB	WEKA's implementation (Witten & Frank, 2000) of Naïve Bayes (Lewis, 1998) with Laplace smoothing
J4.8	WEKA's implementation of the J4.8 decision tree method (Quinlan, 1986) with no pruning
RNW	Our implementation of a version of Littlestone's (1988) Winnow algorithm, generalized to handle real-valued features and more than two classes (Schler, 2007)
BMR	Genkin et al.'s (2006) implementation of Bayesian multiclass regression
SMO	Weka's implementation of Platt's (1998) SMO algorithm for SVM with a linear kernel and default settings

Accuracy

TABLE 2. Accuracy on test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the e-mail corpus.

Features/learner	NB (%)	J4.8 (%)	RMW (%)	BMR (%)	SMO (%)
FW	60.2	58.7	66.1	68.2	63.8
POS	61.0	59.0	66.1	66.3	67.1
FW + POS	65.9	61.6	68.0	67.8	71.7
SFL	57.2	57.2	65.6	67.2	62.7
CW	67.1	66.9	74.9	78.4	74.7
CNG	72.3	65.1	73.1	80.1	74.9
CW + CNG	73.2	68.9	74.2	83.6	78.2

TABLE 4. Accuracy test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the blog corpus.

Features/learner	NB (%)	J4.8 (%)	RMW (%)	BMR (%)	SMO (%)
FW	38.2	30.3	51.8	63.2	63.2
POS	34.0	30.3	51.0	63.2	60.6
FW + POS	47.0	34.3	62.3	70.3	72.0
SFL	35.4	36.3	61.4	69.2	71.7
CW	56.4	51.0	62.9	72.5	70.5
CNG	65.0	48.9	67.1	80.4	80.9
CW + CNG	69.9	51.6	75.4	86.1	85.7