2 MLE: Multinomial
9 Bayesian Statistics
23 Flipping Coins, Revisited
31 Conjugate Distributions
47 Extra: MLE Derivation

## 21: Bayesian Statistics and Beta

Jerry Cain

February 28, 2024

Lecture Discussion on Ed


MLE: Multinomial

## Okay, just one more MLE with the Multinomial

Consider a sample of $n$ iid random variables where:

- Each element is drawn from one of $m$ outcomes. $P$ (outcome $i$ ) $=p_{i}$, where $\sum_{i=1}^{m} p_{i}=1$
- $X_{i}=\#$ of trials with outcome $i$, where $\sum_{i=1}^{m} X_{i}=n$ )
this is the dassic description
of a Multinmial distribution



## Okay, just one more MLE with the Multinomial

Consider a sample of $n$ iid random variables where:

- Each element is drawn from one of $m$ outcomes.

$$
P(\text { outcome } i)=p_{i}, \text { where } \sum_{i=1}^{m} p_{i}=1
$$

- $X_{i}=\#$ of trials with outcome $i$, where $\sum_{i=1}^{m} X_{i}=n$


Example: Suppose each RV is outcome of 6-sided die.

- Roll the dice $n=12$ times.
- Observe data: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

$$
\begin{aligned}
& X_{1}=3, X_{2}=2, X_{3}=0, \\
& X_{4}=3, X_{5}=1, X_{6}=3
\end{aligned} \quad \text { check: } X_{1}+X_{2}+\cdots+X_{6}=12
$$

## Okay, just one more MLE with the Multinomial

Consider a sample of $n$ id random variables where:

- Each element is drawn from one of $m$ outcomes. $P\left(\right.$ outcome $i$ ) $=p_{i}$, where $\sum_{i=1}^{m} p_{i}=1$
- $X_{i}=\#$ of trials with outcome $i$, where $\sum_{i=1}^{m} X_{i}=n$

1. What is the likelihood of observing the sample $\left(X_{1}, X_{2}, \ldots, X_{m}\right)$, given the probabilities $p_{1}, p_{2}, \ldots, p_{m}$ ?

if $p_{1}, i p_{2}, p_{3}$, etc are unknmm, then they are the parameters. In any MLE pablum, we try tochoose
B. $p_{1}^{X_{1}} p_{2}^{X_{2}} \cdots p_{m}^{X_{m}}$
C. $\frac{n!}{X_{1}!X_{2}!\cdots X_{m}!} X_{1}^{p_{1}} X_{2}^{p_{2}} \cdots X_{m}^{p_{m}}$

## Okay, just one more MLE with the Multinomial

Consider a sample of $n$ iid random variables where:

- Each element is drawn from one of $m$ outcomes. $P\left(\right.$ outcome $i$ ) $=p_{i}$, where $\sum_{i=1}^{m} p_{i}=1$
- $X_{i}=\#$ of trials with outcome $i$, where $\sum_{i=1}^{m} X_{i}=n$

$$
\begin{aligned}
& \text { here, } \theta=\left(p_{1}, p_{2}, p_{3}, \ldots, p_{m}\right) \\
& \stackrel{\downarrow}{L(\theta)}=\frac{n!}{X_{1}!X_{2}!\cdots X_{m}!} p_{1}^{X_{1}} p_{2}^{X_{2}} \cdots p_{m}^{X_{m}} \\
& \text { recall that }
\end{aligned}
$$

1. What is the likelihood of observing the sample $\left(X_{1}, X_{2}, \ldots, X_{m}\right)$, given the probabilities $p_{1}, p_{2}, \ldots, p_{m}$ ?
2. What is $\theta_{M L E}$ ?
$L L(\theta)=\log (n!)-\sum_{i=1}^{m} \log \left(X_{i}!\right)+\sum_{i}^{m} X_{i} \log \left(p_{i}\right)$, such that $\sum_{i=1}^{m} p_{i}=1$

Optimize with
Lagrange multipliers in extra slides

$$
\longrightarrow \theta_{M L E}: p_{i}=\frac{X_{i}}{n} \quad \begin{aligned}
& \text { Intuitively, probability } \\
& p_{i}=\text { proportion of outcomes }
\end{aligned}
$$

## When MLEs attack!

Consider a 6-sided die.

- Roll the dice $n=12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

What is $\theta_{M L E}$ ?

## When MLEs attack!

$$
\begin{aligned}
\text { MLE for } \\
\text { Multinomial: }
\end{aligned} p_{i}=\frac{X_{i}}{n}
$$

## Consider a 6-sided die.

- Roll the dice $n=12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes
$\theta_{M L E}:$

$$
\begin{aligned}
& p_{1}=3 / 12 \quad \text { - MLE say you just never roll threes. } \\
& p_{2}=2 / 12 \\
& p_{3}=0 / 12 \quad \text { ! } \\
& p_{4}=3 / 12 \\
& p_{5}=1 / 12 \\
& p_{6}=3 / 12 \\
& \text { - Do you really believe that? } \\
& \text { But what if you cannot } \\
& \text { observe anymore rolls? }
\end{aligned}
$$

# Bayesian Statistics 

## Starting Today!

Today we are going to learn something unintuitive, beautiful, and useful!

## We are going to think of probabilities as random variables.

## A new definition of probability

Flip a coin $n+m$ times, produce $n$ heads. We don't know the probability $\theta$ that the coin comes up heads.


The world's first coin

## Frequentist

$\theta$ is a single value.

$$
\theta=\lim _{n+m \rightarrow \infty} \frac{n}{n+m} \approx \frac{n}{n+m}
$$

Bayesian
$\theta$ is a random variable.
$\theta$ 's continuous support: $(0,1)$

## Let's play a game

Roll 2 dice. If neither roll is a 6 , you win (event $W$ ). Else, I win (event $W^{C}$ ).

- Before you play, what's the probability that you win?
- Play once. What's the probability that you win?
- Play three more times. What's the probability that you win?

Frequentist


Bayesian statistics: Constantly update your prior beliefs.

## Bayesian probability

Bayesian statistics: Probability represents our everevolving understanding of the world.

Mixing discrete and continuous random variables, combined with Bayes' Theorem, allows us to reason about probabilities as random variables.

## Mixing discrete and continuous

Let $X$ be a continuous random variable, and $N$ be a discrete random variable.

Bayes'
Theorem:

$$
f_{X \mid N}(x \mid n)=\frac{p_{N \mid X}(n \mid x) f_{X}(x)}{p_{N}(n)}
$$

Intuition: $\quad P(X=x \mid N=n)=\frac{P(N=n \mid X=x) \overparen{P(X=x})}{P(N=n)}$
$\approx f_{x}(x) \epsilon_{x}$

$$
f_{X \mid N}(x \mid n) \varepsilon_{X}=\frac{P(N=n \mid X=x) f_{X}(x) \varepsilon_{X X}}{P(N=n)} \Rightarrow f_{X \mid N}(x \mid n)=\frac{p_{N \mid X}(n \mid x) f_{X}(x)}{p_{N}(n)}
$$

## Bayes' Theorem: All Flavors

Let $X, Y$ be continuousand $M, N$ be discrete random variables.

Original Bayes:

$$
\left.p_{M \mid N}(m \mid n)=\frac{p_{N \mid M}(n \mid m) p_{M}(m)}{p_{N}(n)}\right\} \begin{gathered}
\text { this is our origival } \\
\text { Bages' theorem from } \\
\text { Lecture } 4
\end{gathered}
$$

$$
\begin{aligned}
& \left.f_{X \mid N}(x \mid n)=\frac{p_{N \mid X}(n \mid x) f_{X}(x)}{p_{N}(n)}\right\} \text { from prior slide } \\
& \left.p_{N \mid X}(n \mid x)=\frac{f_{X \mid N}(x \mid n) p_{N}(n)}{f_{X}(x)}\right\} \begin{array}{c}
\text { arathor mixed form } \\
\text { not unlike the } \\
\text { one above it. }
\end{array}
\end{aligned}
$$

$$
\left.f_{X \mid Y}(x \mid y)=\frac{f_{Y \mid X}(y \mid x) f_{X}(x)}{f_{Y}(y)}\right\} \begin{aligned}
& \text { Jacoband Kann tanght } \\
& \text { this form when they } \\
& \text { guest lectund for } \\
& \text { Lecture 17. }
\end{aligned}
$$

Mix Bayes \#1:

Mix Bayes \#2:

All continuous:

## Mixing discrete and continuous

Let $\theta$ be a random variable for the probability your coin comes up heads, and $N$ be the number of heads you observe in an experiment.

$$
\begin{aligned}
& \text { posterior } \\
& f_{\theta \mid N}(x \mid n)=\frac{p_{N \mid \theta}(n \mid x) f_{\theta}(x)}{p_{N}(n)}
\end{aligned}
$$

normalization constant

- Prior belief of parameter $\theta$
- Likelihood of $N=n$ heads, given parameter $\theta=x$.
- Posterior updated belief of parameter $\theta$.

$$
\begin{aligned}
& f_{\theta}(x) \\
& p_{N \mid \theta}(n \mid x) \\
& f_{\theta \mid N}(x \mid n)
\end{aligned}
$$

Beta RV

## Beta random variable

def A Beta random variable $X$ is defined as follows:

$$
\begin{aligned}
& X \sim \operatorname{Beta}(a, b) \\
& a>0, b>0 \\
& \text { Support of } X \text { : }(0,1) \\
& \text { where } B(a, b)=\int_{0}^{1} x^{a-1}(1-x)^{b-1} d x \text {, normalizing constant }
\end{aligned}
$$

Expectation $E[X]=\frac{a}{a+b} \quad$ Variance $\operatorname{Var}(X)=\frac{a b}{(a+b)^{2}(a+b+1)}$ $B(a, b)$ is chosen si that $\int_{0}^{1} f(x) d x=1$

$$
\frac{1}{B(a, b)} \int_{0}^{1} x^{a-1}(1-x)^{b-1} d x=1 \Rightarrow B(a, b)=\int_{0}^{1} x^{a-1}(1-x)^{b-1} d x
$$

## Beta RV with different $a, b$

$X \sim \operatorname{Beta}(a, b) \quad$ PDF $\quad f(x)=\frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1}$

$$
a>0, b>0
$$

Support of $X:(0,1)$
where $B(a, b)=\int_{0}^{1} x^{a-1}(1-x)^{b-1} d x$, normalizing constant

 + a third case (next slide)

Note: PDF symmetric when $a=b$

## Beta RV with different $a, b$

$$
X \sim \operatorname{Beta}(a, b)
$$

Match PDF to distribution:

$a, b$ positive
integers
geaterthar $\{\{$
A. Beta $(5,5)$

3.0
B. Beta (2,8)

In CS109, we focus on Beta functions where $a, b$ are both positive integers.

## Beta random variable

def A Beta random variable $X$ is defined as follows:


Beta can be a distribution of probabilities.

## Beta can be a distribution of probabilities.



Beta parameters $a, b$ are determined by the outcome of an experiment.

But which experiment?

## Flipping a coin with unknown probability

## Flip a coin with unknown probability

Flip a coin $n+m$ times, observe $n$ heads.

- Before our experiment, $\theta$ (the probability that the coin comes up heads) is equally like to be any probability in ( 0,1 ).
- Let $N=$ number of heads.
- Given $\theta=x$, coin flips are independent.

What is our updated belief of $\theta$ after we observe $N=n$ ?

## What are reasonable distributions of the following?

1. $\theta$
2. $N \mid \theta=x \quad$ Likelihood $N \mid \theta=x \sim \operatorname{Bin}(n+m, x)$
3. $\theta \mid N=n \quad$ Bayesian posterior. Use Bayes'!

## Flip a coin with unknown probability

Flip a coin $n+m$ times, observe $n$ heads.

- Before our experiment, $\theta$ (the probability that the coin comes up heads) is equally like to be any probability in ( 0,1 ).
- Let $N=$ number of heads.
- Given $\theta=x$, coin flips are independent.

What is our updated belief of $\theta$ after we observe $N=n ? \quad$ Posterior: $f_{\theta \mid N}(\theta \mid n)$

$$
\begin{aligned}
f_{\theta \mid N}(x \mid n)= & \frac{p_{N \mid \theta}(n \mid x) f_{\theta}(x)}{p_{N}(n)}=\frac{\binom{n+m}{n} x^{n}(1-x)^{m} \cdot 1}{p_{N}(n)} \\
& =\frac{\binom{n+m}{n}}{p_{N}(n)} x^{n}(1-x)^{m}=\frac{1}{c} x^{n}(1-x)^{m}, \text { where } c=\int_{0}^{1} x^{n}(1-x)^{m} d x
\end{aligned}
$$

constant with respect to $x$,

## Let's try it out

1. Start with a $\theta \sim \operatorname{Uni}(0,1)$ over probability that a coin lands heads.
2. Flip a coin 8 times. Observe $n=7$ heads and $m=1$ tail
3. What is our posterior belief of the probability $\theta$ ?

## Beta RV with different $a, b$

$$
\begin{gathered}
X \sim \operatorname{Beta}(a, b) \quad \text { PDF } \quad f(x)=\frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1} \\
\begin{array}{c}
a>0, b>0 \\
\text { Support of } X:(0,1)
\end{array} \quad \text { where } B(a, b)=\int_{0}^{1} x^{a-1}(1-x)^{b-1} d x, \text { normalizing constant }
\end{gathered}
$$

$$
\begin{gathered}
f_{\theta \mid N}(x \mid n)=\frac{1}{c} x^{7}(1-x)^{1} \quad \text { is the PDF for } \operatorname{Beta}(8,2)! \\
c \text { normalizes to valid PDF }
\end{gathered}
$$

## Let's try it out

1. Start with a $\theta \sim \operatorname{Uni}(0,1)$ over probability that a coin lands heads.
2. Flip a coin 8 times. Observe $n=7$ heads and $m=1$ tail
3. What is our posterior belief of the probability $\theta$ ?


$$
f_{\theta \mid N}(x \mid n)=\frac{1}{c} x^{7}(1-x)^{1}
$$

c normalizes to valid PDF

## 3. What is our posterior belief of the probability $\theta$ ?

- Start with a $\theta \sim$ Uni $(0,1)$ over probability
- Observe $n=7$ successes and $m=1$ failures
- Your new belief about the probability of $\theta$ is:

$$
f_{\theta \mid N}(x \mid n)=\frac{1}{c} x^{7}(1-x)^{1}, \text { where } c=\int_{0}^{1} x^{7}(1-x)^{1} d x
$$

Posterior belief, $\theta \mid N$ :

$$
\operatorname{Beta}(a=8, b=2)
$$

$$
f_{\theta \mid N}(x \mid n)=\frac{1}{c} x^{8-1}(1-x)^{2-1}
$$

$$
\operatorname{Beta}(a=n+1, b=m+1)
$$



## CSiog focus: Beta where $a, b$ both positive integers $\quad x \sim \operatorname{Beta}(a, b)$


summary
MLE:
Bayesian
view point:
$\theta$ is a value aka print estivate
$\theta$ is a vardum vaviabl.
$\Theta$ expueses
Beta:

$$
\begin{gathered}
a=\text { "successes" }+1 \\
b=\text { "failures" }+1
\end{gathered}
$$

Beta parameters $a, b$ are determined by the outcome of an experiment.

- Beta (in CS109) models the randomness of the probability of experiment success.
- Beta parameters depend on our data and our prior.


# Conjugate distributions 

## A note about our prior

1. Start with a $\theta \sim \operatorname{Uni}(0,1)$ over probability that a coin lands heads.
 heads and $m=1$ tail
$f_{\theta \mid N}(x \mid n)=\frac{1}{c} x^{7}(1-x)^{1}$ c normalizes to valid PDF

## Beta RV with different $a, b$

$$
\begin{gathered}
X \sim \operatorname{Beta}(a, b) \quad \text { PDF } \quad f(x)=\frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1} \\
\quad \begin{array}{l}
a>0, b>0 \\
\text { Support of } X:(0,1)
\end{array} \quad \text { where } B(a, b)=\int_{0}^{1} x^{a-1}(1-x)^{b-1} d x, \text { normalizing constant }
\end{gathered}
$$



Note: PDF symmetric when $a=b$

## A note about our prior

1. Start with a $\theta \sim \operatorname{Uni}(0,1)$ over probability that a coin lands heads. Beta(1,1)


Check this out. $\operatorname{Beta}(a=1, b=1)$ :

$$
\begin{aligned}
& f(x)= \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1} \\
&=\frac{1}{\int_{0}^{1} 1 d x} \\
&=1 \quad \begin{array}{l}
\text { where } 0<x<1
\end{array} \\
&\left.\begin{array}{c}
\text { comfirms that } \\
\text { Beta }(1,1) \text { came os Uni } \\
\text { Stanford University } \\
34
\end{array}\right)
\end{aligned}
$$

## Beta is a conjugate distribution for Bernoulli

Beta is a conjugate distribution for Bernoulli, meaning:

- Prior and posterior parametric forms are the same


## Beta is a conjugate distribution for Bernoulli

Beta is a conjugate distribution for Bernoulli, meaning:

1. If our prior belief of the parameter is Beta, and
2. Our experiment is Bernoulli, then
(observe $n$ successes, $m$ failures)
3. Our posterior is also Beta.

Proof: $\quad \theta \sim \operatorname{Beta}(a, b) \quad N \mid \theta \sim \operatorname{Bin}(n+m, x)$

$$
\begin{aligned}
f_{\theta \mid N}(x \mid n) & =\frac{p_{N \mid \theta}(n \mid x) f_{\theta}(x)}{p_{N}(n)}=\frac{\binom{n+m}{m} x^{n}(1-x)^{m} \cdot \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1}}{p_{N}(n)} \\
\begin{array}{c}
\text { constants that } \\
\text { don't depend on } x
\end{array} & =C \cdot x^{n}(1-x)^{m} \cdot x^{a-1}(1-x)^{b-1} \\
& =C \cdot x^{n+a-1}(1-x)^{m+b-1} \square \Rightarrow \begin{array}{r}
\text { Beta }(a, b) \text { be came } \\
\text { Beta }(a+n, b+m) \\
\text { Stanford University }
\end{array}
\end{aligned}
$$

## Beta is a conjugate distribution for Bernoulli

This is the main takeaway of Beta.
Beta is a conjugate distribution for Bernoulli, meaning:

- Prior and posterior parametric forms are the same $\rightarrow$ both ave Beta
- Practically, conjugate means easy update:

Add number of "heads" and "tails" seen to Beta parameters.
You can invent a prior to express how biased you believe the coin is a priori:

- $\theta \sim \operatorname{Beta}(a, b)$ : pretend you've conducted $(a+b-2)$ imaginary trials, where ( $a-1$ ) trials produced a head and $(b-1)$ produced a
- Choosing Beta $(1,1)=$ Uni $(0,1)$ means you don't hold any prior beliefs

$$
\text { Prior } \operatorname{Beta}\left(a=n_{i m a g}+1, b=m_{\text {imag }}+1\right)
$$



Experiment Observe $n$ successes and $m$ failures $\leftarrow$ inheuntly | Binmial/Bermmilli |
| :---: |

Posterior $\operatorname{Beta}\left(a=n_{\text {imag }}+n+1, b=m_{\text {imag }}+m+1\right)$

## Medicinal Beta

- Before being tested, a medicine is believed to "work" 80\% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn’t work" for 6.

What is your new belief that the drug "works"?

## Frequentist

Let $p$ be the probability your drug works.

$$
p \approx \frac{14}{20}=0.7
$$

POV: prior beliefs ivrelerant in prior/expert belief about probability.

## Medicinal Beta

- Before being tested, a medicine is believed to "work" 80\% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn’t work" for 6.

What is your new belief that the drug "works"?

## Frequentist

Let $p$ be the probability your drug works.

$$
p \approx \frac{14}{20}=0.7
$$

## Bayesian

Let $\theta$ be the probability your drug works.
$\theta$ is a random variable.
POV: history matters and shonidn't
be ignored.

## Medicinal Beta

$$
\begin{aligned}
\text { Prior } & \operatorname{Beta}\left(a=n_{i m a g}+1, b=m_{\text {imag }}+1\right) \\
\text { Posterior } & \operatorname{Beta}\left(a=n_{i m a g}+n+1, b=m_{\text {imag }}+m+1\right)
\end{aligned}
$$

- Before being tested, a medicine is believed to "work" 80\% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn't work" for 6.

What is your new belief that the drug "works"?
What is the prior distribution of $\theta$ ? (select all that apply)
A. $\quad \theta \sim \operatorname{Beta}(1,1)=\operatorname{Uni}(0,1)$
B. $\theta \sim \operatorname{Beta}(81,101)$
C. $\theta \sim \operatorname{Beta}(80,20)$
D. $\theta \sim \operatorname{Beta}(81,21)$
E. $\quad \theta \sim \operatorname{Beta}(5,2)$

## Medicinal Beta

$$
\begin{aligned}
\text { Prior } & \operatorname{Beta}\left(a=n_{\text {imag }}+1, b=m_{\text {imag }}+1\right) \\
\text { Posterior } \operatorname{Beta}(a & \left.=n_{\text {imag }}+n+1, b=m_{\text {imag }}+m+1\right)
\end{aligned}
$$

- Before being tested, a medicine is believed to "work" 80\% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn’t work" for 6.

What is your new belief that the drug "works"?
(Bayesian interpretation)
What is the prior distribution of $\theta$ ? (select all that apply)
A. $\quad \theta \sim \operatorname{Beta}(1,1)=\operatorname{Uni}(0,1)$
B. $\theta \sim \operatorname{Beta}(81,101)$
C. $\theta \sim \operatorname{Beta}(80,20)$
(D) $\theta \sim \operatorname{Beta}(81,21) \rightarrow$ Interpretation: 80 successes / 100 imaginary trials
(E.) $\theta \sim \operatorname{Beta}(5,2) \longrightarrow 4$ successes $/ 5$ imaginan trials (you can choose either based on how strongly you believe in prior data. We choose E on next slide)

## Medicinal Beta

$$
\begin{aligned}
\text { Prior } & \operatorname{Beta}\left(a=n_{i m a g}+1, b=m_{\text {imag }}+1\right) \\
\text { Posterior } \operatorname{Beta}(a & \left.=n_{\text {imag }}+n+1, b=m_{\text {imag }}+m+1\right)
\end{aligned}
$$

- Before being tested, a medicine is believed to "work" 80\% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn’t work" for 6.

What is your new belief that the drug "works"?
(Bayesian interpretation)
Prior

$$
\begin{array}{ll}
\text { Prior: } & \\
\text { Posterior: } & \\
& \\
& \\
& \sim \operatorname{Beta}(a=5, b=2) \\
& \sim \operatorname{Beta}(a=5+14, b=2+6) \\
\end{array}
$$



## Medicinal Beta

$$
\begin{aligned}
\text { Prior } & \operatorname{Beta}\left(a=n_{i m a g}+1, b=m_{\text {imag }}+1\right) \\
\text { Posterior } \operatorname{Beta}(a & \left.=n_{\text {imag }}+n+1, b=m_{\text {imag }}+m+1\right)
\end{aligned}
$$

- Before being tested, a medicine is believed to "work" 80\% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn’t work" for 6.

What is your new belief that the drug "works"?
Prior: $\quad \theta \sim \operatorname{Beta}(a=5, b=2)$
Posterior: $\quad \theta \sim \operatorname{Beta}(a=5+14, b=2+6)$

$$
\sim \operatorname{Beta}(a=19, b=8)
$$

What do you report to pharmacists?
A. Expectation of posterior
B. Mode of posterior
C. Distribution of posterior
D. Nothing
(Bayesian interpretation)


## Medicinal Beta

$$
\begin{aligned}
\text { Prior } & \operatorname{Beta}\left(a=n_{i m a g}+1, b=m_{\text {imag }}+1\right) \\
\text { Posterior } & \operatorname{Beta}\left(a=n_{i m a g}+n+1, b=m_{\text {imag }}+m+1\right)
\end{aligned}
$$

- Before being tested, a medicine is believed to "work" 80\% of the time.
- The medicine is administered to 20 patients.
- It "works" for 14, "doesn’t work" for 6.

What is your new belief that the drug "works"?
Prior: $\quad \theta \sim \operatorname{Beta}(a=5, b=2)$
Posterior: $\quad \theta \sim \operatorname{Beta}(a=5+14, b=2+6)$

$$
\sim \operatorname{Beta}(a=19, b=8)
$$

What do you report to pharmacists?

$$
\begin{aligned}
& E[\theta]=\frac{a}{a+b}=\frac{19}{19+8} \approx 0.70 \\
& \underbrace{\operatorname{mode}(\theta)=\frac{a-1}{a+b-2}}=\frac{18}{18+7} \approx 0.72
\end{aligned}
$$ "most likely" parameter given the data.

we will prove later on

## Food for thought

In this lecture:
$X \sim \operatorname{Ber}(p)$

If nothing is known about the parameter $p$, Bayesian statisticians will:

- Treat the parameter as a random variable $\theta$ with a Beta prior distribution
- Conduct experiments
- Based on the outcomes of those experiments, update the posterior distribution of $\theta$

Food for thought:
Any parameter for a "parameterized" random variable can be thought of as a random variable.

$$
Y \sim \mathcal{N}\left(\mu, \sigma^{2}\right)
$$

## Estimating our parameter directly

(our focus so far)

Maximum
Likelihood
Estimator
(MLE)

What is the parameter $\theta$ that maximizes the likelihood of our observed data $\left(x_{1}, x_{2}, \ldots, x_{n}\right)$ ?

$$
\begin{gathered}
L(\theta)=f\left(X_{1}, X_{2}, \ldots, X_{n} \mid \theta\right) \\
=\prod_{i=1}^{n} f\left(X_{i} \mid \theta\right) \\
\theta_{M L E}=\underset{\theta}{\arg \max } f\left(X_{1}, X_{2}, \ldots, X_{n} \mid \theta\right) \\
\text { likelihood of data }
\end{gathered}
$$

Observations:

- MLE maximizes probability of observing data given a parameter $\theta$. It's fitting the curve to match the data.
- If we are estimating $\theta$, shouldn't we maximize the probability of $\theta$ directly? SAT word: adumbrate


# Extra: MLE: Multinomial derivation 

## Okay, just one more MLE with the Multinomial

Consider a sample of $n$ i.i.d. random variables where

- Each element is drawn from one of $m$ outcomes. $P$ (outcome $i$ ) $=p_{i}$, where $\sum_{i=1}^{m} p_{i}=1$
- $X_{i}=\#$ of trials with outcome $i$, where $\sum_{i=1}^{m} X_{i}=n$

1. What is the likelihood of observing the sample $\left(X_{1}, X_{2}, \ldots, X_{m}\right)$, given the probabilities $p_{1}, p_{2}, \ldots, p_{m}$ ?

$$
L(\theta)=\frac{n!}{X_{1}!X_{2}!\cdots X_{m}!} p_{1}^{X_{1}} p_{2}^{X_{2}} \cdots p_{m}^{X_{m}}
$$

2. What is $\theta_{M L E}$ ?
$L L(\theta)=\log (n!)-\sum_{i=1}^{m} \log \left(X_{i}!\right)+\sum_{i=1}^{m} X_{i} \log \left(p_{i}\right)$, such that $\sum_{i=1}^{m} p_{i}=1$

$$
\theta_{M L E}: p_{i}=\frac{X_{i}}{n} \quad \begin{aligned}
& \text { Intuitively, probability } \\
& p_{i}=\text { proportion of outcomes }
\end{aligned}
$$

## Optimizing MLE for Multinomial

$\theta=\left(p_{1}, p_{2}, \ldots, p_{m}\right)$
$\theta_{M L E}=\underset{\theta}{\arg \max } L L(\theta)$, where $\sum_{i=1}^{m} p_{i}=1$

## Use Lagrange multipliers to account for constraint

Lagrange multipliers:

$$
A(\theta)=L L(\theta)+\lambda\left(\sum_{i=1}^{m} p_{i}-1\right)=\sum_{i=1}^{m} X_{i} \log \left(p_{i}\right)+\lambda\left(\sum_{i=1}^{m} p_{i}-1\right) \begin{aligned}
& \left(\begin{array}{l}
\text { drop } \\
\text { non- } p_{i} \\
\text { terms })
\end{array}\right.
\end{aligned}
$$

Differentiate w.r.t.
each $p_{i}$, in turn:

$$
\frac{\partial A(\theta)}{\partial p_{i}}=X_{i} \frac{1}{p_{i}}+\lambda=0 \Rightarrow p_{i}=-\frac{X_{i}}{\lambda}
$$

Solve for $\lambda$, noting
$\sum_{i=1}^{m} X_{i}=n, \sum_{i=1}^{m} p_{i}=1:$

$$
\sum_{i=1}^{m} p_{i}=\sum_{i=1}^{m}-\frac{X_{i}}{\lambda}=1 \quad \Rightarrow 1=-\frac{n}{\lambda} \quad \Rightarrow \lambda=-n
$$

Substitute $\lambda$ into $p_{i}$

$$
p_{i}=\frac{X_{i}}{n}
$$

