STAT 391

4/11/23

Lecture 5

$\geq$ next Thursday : Q1

HW 1 due tomorrow

Quiz 1:
all material until $<$ Continuous Sample spaces
At begining of class

Lecture Notes II – Maximum Likelihood Estimation for Discrete Distributions

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

April, 2021

Max Likelihood Principle ✓

ML estimation for arbitrary discrete distributions ✓

Other ML estimation examples ✓

ML estimate as a random variable ←

Reading: Ch. 4.1, 4.2

A random experiment

Coin toss $S = \{0,1\}$ $n=10 \Rightarrow S = \{0, 0.1, 0.2, ..0.5.. 1.0\}$ Θ^{ML}

$= \{\frac{n_1}{10}\}_{n_1 \in \{0,..10\}}$

$\theta_1^{true} = 0.5$

$\theta^{true} \in \{\theta^{ML} \text{ observed}\}$

n_1	θ_1^{ML}
10 trials 4	0.4
10 trials 7	0.7
6	0.6
6	0.6
6	0.6

independent

GUESSES

$\theta^{ML} = 0.55$ possible? No

- values near θ^{true} are more likely

$n_1^* = 29 \Rightarrow \theta^{*ML} = 0.58$

$n^* = 50$ indep trials

take mode $\theta^{mode} = 0.6$

average θ^{ML}'s = Modern Robust Method

median θ^{ML}'s

resample ...

~~ML estimation for arbitrary discrete distributions~~

Candy sampling

θ^{true}_{Green} unknown

n	n_{Green}	$\hat{\theta}_{Green}$
2	0	0
5	1	0.2
2	1	0.5
4	1	0.25
2	1	0.5
6	3	0.5
5	2	0.4

independent

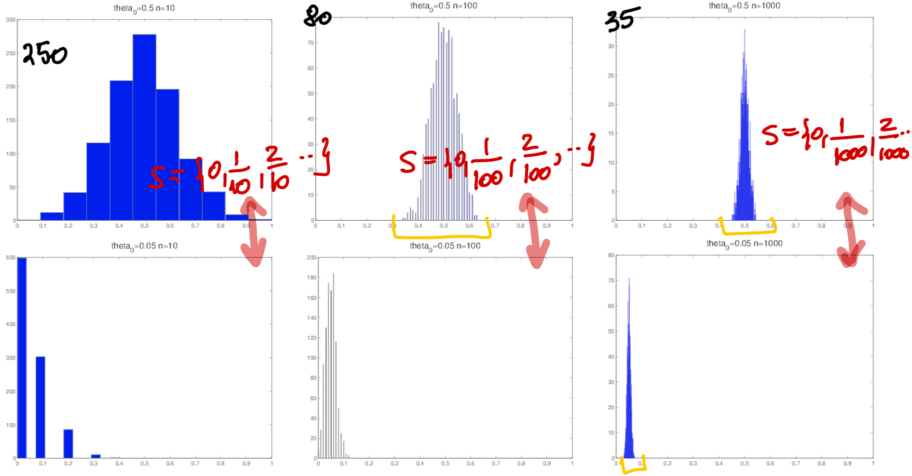
$n^* = 26 \quad n^*_G = 9 \quad \theta^{*ML} = \frac{9}{26} \simeq 0.31$

ML estimate as a random variable $\times$ 1000 experiments

$n=10$ $n=100$ $n=1000$

true $\theta_0 = 0.5$

true $\theta_0 = 1/20 = 0.05$



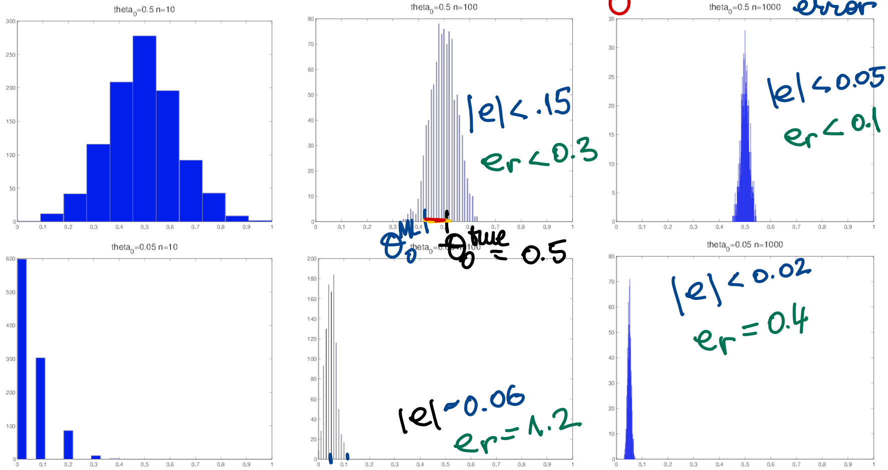
distribution of $\hat{\theta}_0^{ML}$ concentrates $\Leftrightarrow$ interval that contains observed $\hat{\theta}_0^{ML}$ shrinks

ML estimate as a random variable

$\mathrm{Var}\,\mathrm{Bern}(p) = p(1-p)$

$e = \theta^{true} - \theta^{ML} \lessgtr 0$ absolute error

$e_r = \dfrac{|\theta^{true} - \theta^{ML}|}{\theta^{true}}$ relative error



$|e| < .15$

$e_r < 0.3$

$|e| < 0.05$

$e_r < 0.1$

θ_0^{ML}

$\theta_0^{true} = 0.5$

$|e| < 0.02$

$e_r = 0.4$

$|e| \sim 0.06$

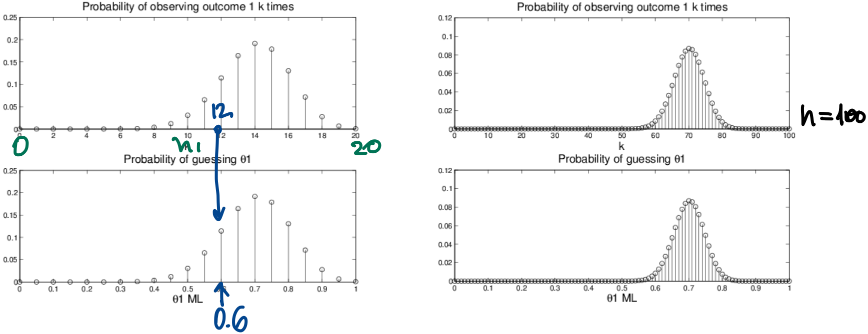
$e_r = 1.2$

$\theta^{true} = 0.05$

ML estimate as a random variable

$n = 20$
$\theta_1 = 0.7$ $\Rightarrow n_1 \sim \mathrm{Binom}(0.7, 20)$

$$\binom{n}{n_0\ n_1\ \ldots n_{m-1}} = \frac{n!}{n_0!n_1!\ldots n_{m-1}!}$$



$$\theta_1^{ML} = \frac{n_1}{20}$$

$$P\left[\theta_1^{ML} = \frac{n_1}{n} \,\middle|\, n=20\right] = P\left[n_1 \mid n=20\right] = \mathrm{Binom}(\ldots)$$

Lecture Notes III: Discrete probability in practice – Small Probabilities

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

April, 2021

The problem with estimating small probabilities

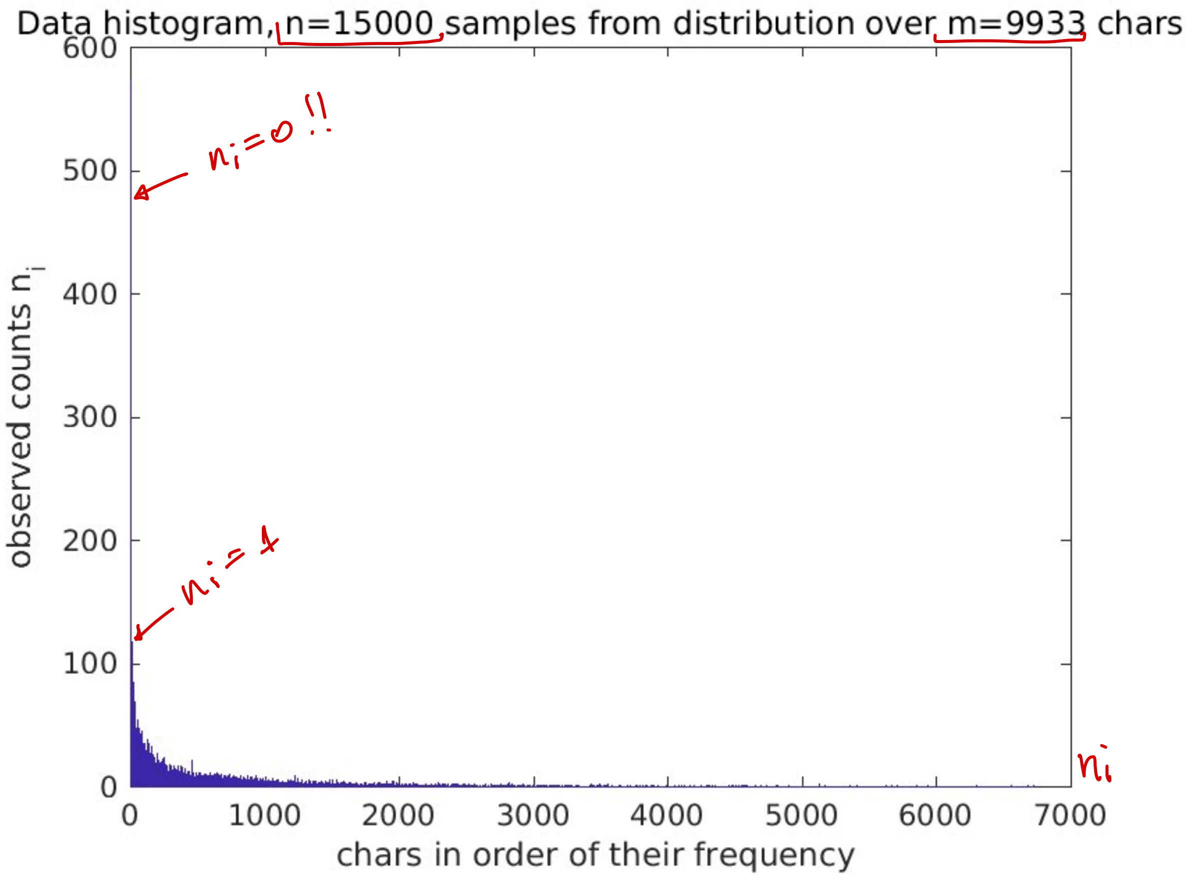
Definitions and setup

Additive methods (Laplace, Dirichlet, Bayesian, ELE)

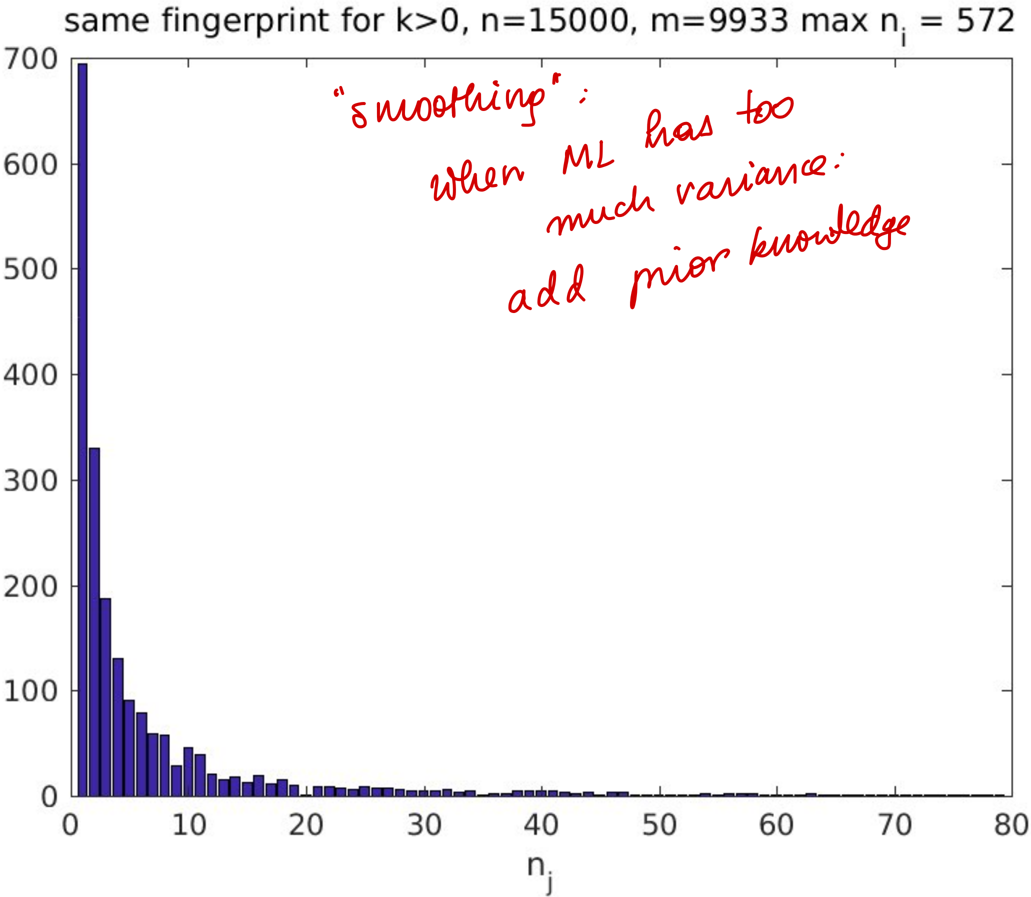
Discounting (Ney-Essen)

Multiplicative smoothing: Estimating the next outcome (Witten-Bell, Good-Turing)

Back-off or shrinkage – mixing with simpler models



Data histogram, n=15000 samples from distribution over m=9933 chars
observed counts n_i
chars in order of their frequency
0
100
200
300
400
500
600
1000
2000
3000
4000
5000
6000
7000
$n_i = 0$!!
$n_i = 1$
n_i


same fingerprint for k>0, n=15000, m=9933 max n_i = 572
"smoothing":
when ML has too
much variance:
add prior knowledge
0
100
200
300
400
500
600
700
0
10
20
30
40
50
60
70
80
n_j

Definitions and setup

We will look at estimating categorical distributions from samples, when the number of outcomes m is large.

- Let $S = \{1, \dots m\}$ be the sample space, and $P = (\theta_1, \dots \theta_m)$ a distribution over S.
- We draw n independent samples from P, obtaining the **data set** $\mathcal{D}$
- Define **the counts** $\{n_j = \#j \text{ appears in } \mathcal{D},\ i = 1, \dots n\}$. The counts are also called **sufficient statistics** or **histogram**.
- Define the **fingerprint** (or **histogram of histogram**) of $\mathcal{D}$ as the counts of the counts, i.e $\{\, r_k = \#\text{counts}\, n_j = k,\ \text{for } k = 0, 1, 2 \dots\}$

  Example $m = 26$ alphabet letters

Data	**Counts** n_i	**Fingerprint** r_k
	$n_j = 0$: `a,b,g,j,l,m,n,` `p,v,w,y,z`	$r_0 = 12 = \lvert\{$`a,b,g,...,y,z`$\}\rvert$
`the red fox is quick`		$r_1 = 12 = \lvert\{$`c,d,f,h,...,u,x`$\}\rvert$
$n = 16$ letters	$n_j = 1$: `c,d,f,h,k,o,q,r,s,t,u,x`	$r_2 = 2 = \lvert\{$`e,i`$\}\rvert$
	$n_j = 2$: `e,i`	$r_3 = \dots r_n = 0$
	$n_j = 0$: `a,b,c...,x,z`	$r_0 = 26 - 6 - 1 - 1 - 1 = 17$
	$n_j = 1$: `f,i,n,r,t,w`	$r_1 = 6 = \lvert\{$`f,i,n,r,t,w`$\}\rvert$
`ho ho who s on first`	$n_j = 2$: `s`	$r_2 = 1 = \lvert\{$`s`$\}\rvert$
$n = 15$ letters	$n_j = 3$: `h`	$r_3 = 1 = \lvert\{$`h`$\}\rvert$
	$n_j = 4$: `o`	$r_4 = 1 = \lvert\{$`o`$\}\rvert$

- It is easy to verify that $n_j \in 0 : n$, hence $r_{0:n}$ may be non-zero (but $r_{n+1,n+2,\dots} = 0$), and that

$$m \;=\; r_0 + r_1 + \dots r_n \quad n \;=\; 0 \times r_0 + 1 \times r_1 + \dots k \times r_k + \dots \tag{1}$$

Smoothing on an example

$m = |S|$

i-th letter in sentence

j-th letter in S

- **the counts** $\{n_j = \#j$ appears in $\mathcal{D}$, $i = 1, \ldots n\}$ (or **sufficient statistics** or **histogram**)
- **fingerprint** (or **histogram of histogram**) of $\mathcal{D}$ as the counts of the counts $\{\, r_k = \#\text{counts}\, n_j = k,\ \text{for } k = 0, 1, 2 \ldots\}$, and $R_k = \{\, j,\ n_j = k, \}$

Histogram of hist

Histogram of n_j's

Example $\boxed{m = 26 \text{ alphabet letters}}$

Data

`the red fox is quick`

$n = 16$ letters

Counts n_i

$n_j = 0$: `a,b,g,j,l,m,n,p,v,w,y,z`

$n_j = 1$: `c,d,f,h,k,o,q,r,s,t,u,x`

$n_j = 2$: `e,i`

Fingerprint r_k

$r_0 = 12 = |\{$`a,b,g,...,y,z`$\}|$

$r_1 = 12 = |\{$`c,d,f,h,...,u,x`$\}|$

$r_2 = 2 = |\{$`e,i`$\}|$

$r_3 = \ldots r_n = 0$

need counts

$\theta_a^{ML} = 0 = \theta_b^{ML}$

$\theta_c^{ML} = \frac{1}{16}$ . . .

BAD!! NEED SMOOTHING

Principle: if $n_j = n_{j'} \Rightarrow \theta_j = \theta_{j'}$ after smoothing

$j, j' \in S$