	Hypothesis FAIR		Hypothesis BIASED
Prior:	2	:	1
Likelihood of heads:	0.5	:	0.4
Posterior:	1	:	0.4

equivalently 10:4 or $\frac{10}{14} : \frac{4}{14}$

Results of the flips	Hypothesis FAIR		Hypothesis BIASED
	2	:	1
H	0.5	:	0.4
T	0.5	:	0.6
T	0.5	:	0.6
H	0.5	:	0.4
T	0.5	:	0.6
Posterior	0.0625	:	0.03456
equivalently:	~64%	:	~36%

Results of the flips	Hypothesis FAIR		Hypothesis BIASED
	2^{1}	:	2^{0}
H	2^{-1}	:	$2^{-1.32}$
T	2^{-1}	:	$2^{-0.74}$
T	2^{-1}	:	$2^{-0.74}$
H	2^{-1}	:	$2^{-1.32}$
T	2^{-1}	:	$2^{-0.74}$
Posterior	$2 \cdot 2^{-5}$	:	$1 \cdot 2^{-4.85}$
equivalently:	~64%	:	~36%

Table 1

	🚶‍♀️	🚲	🚌
☀️	20%	30%	20%
☁️	0%	0%	30%

Table 2

	🚶‍♀️	🚲	🚌
☀️	10%	20%	40%
☁️	10%	10%	10%

Table 3

	🚶‍♀️	🚲	🚌
☀️	15%	10%	45%
☁️	5%	20%	5%

Independent Table

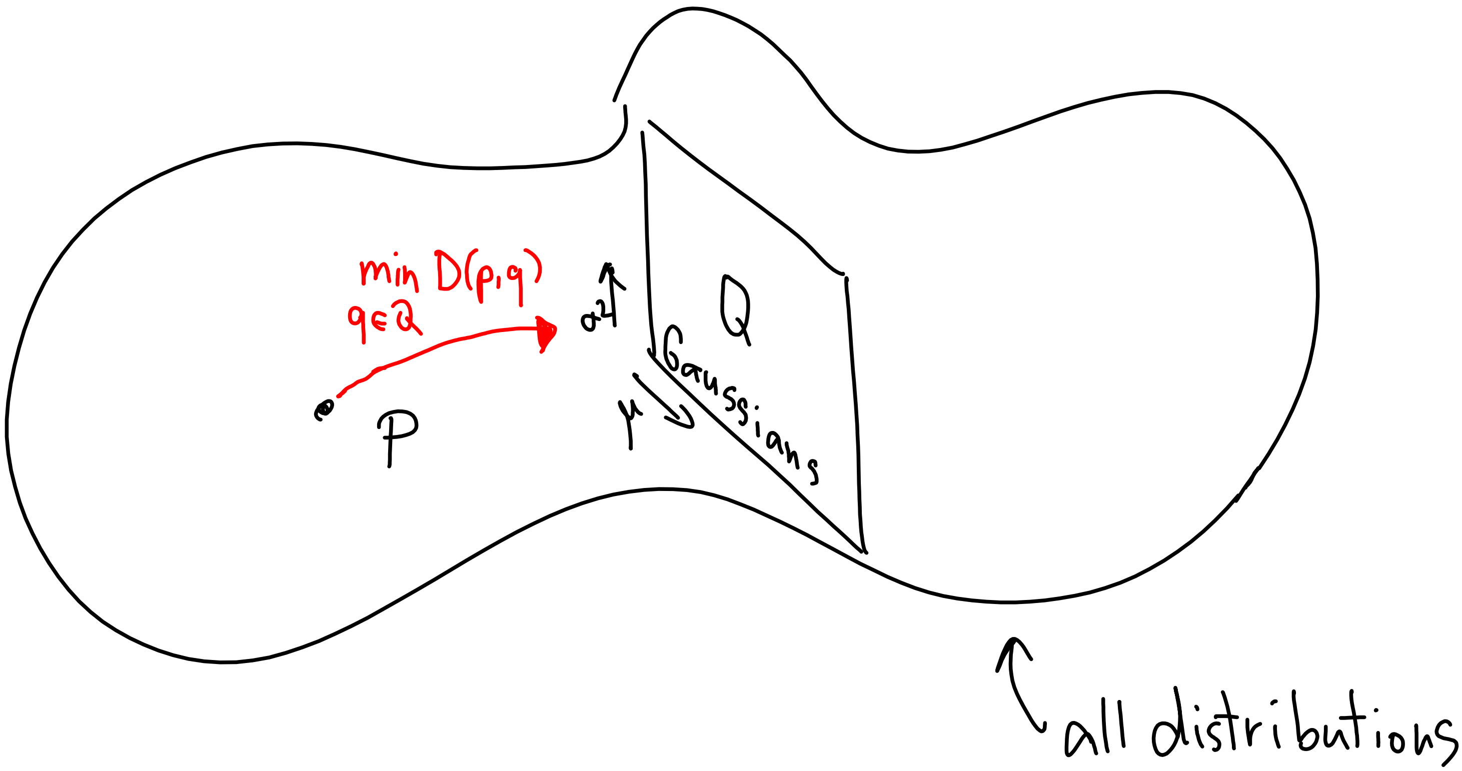
	🚶‍♀️	🚲	🚌
☀️	14%	21%	35%
☁️	6%	9%	15%

$$\underbrace{\sum_{i=1}^{n} p_i \lg \frac{p_i}{q_i}}_{\text{KL divergence}} = \underbrace{\sum_{i=1}^{n} p_i \lg \frac{1}{q_i}}_{\text{crossentropy}} - \underbrace{\sum_{i=1}^{n} p_i \lg \frac{1}{p_i}}_{\text{entropy}}$$

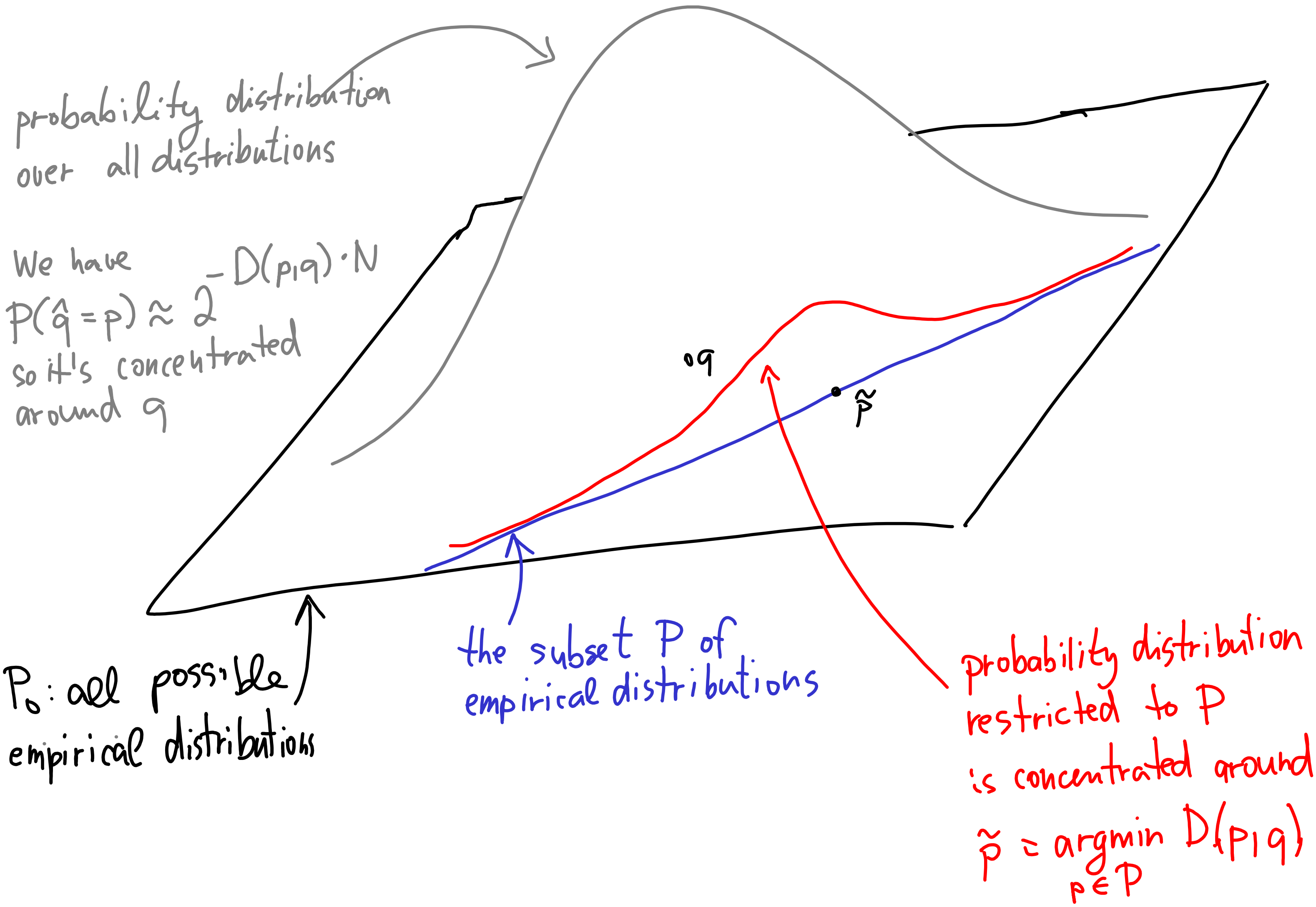
also known as relative entropy

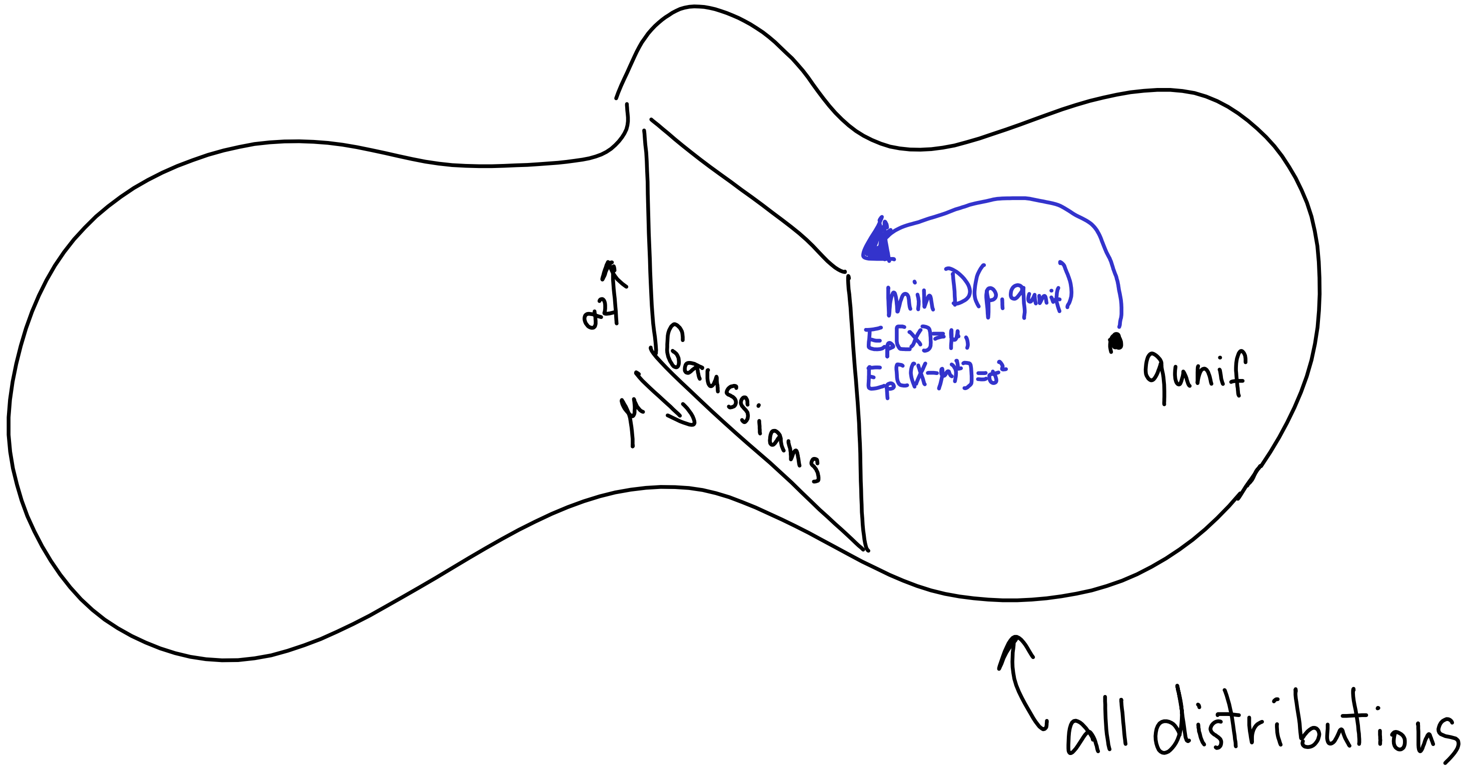
Results of the flips	Hypothesis FAIR		Hypothesis BIASED
	2^{1}	:	2^{0}
H	2^{-1}	:	$2^{-1.32}$
T	2^{-1}	:	$2^{-0.74}$
T	2^{-1}	:	$2^{-0.74}$
H	2^{-1}	:	$2^{-1.32}$
T	2^{-1}	:	$2^{-0.74}$
Posterior	2^{-4}	:	$2^{-4.85}$
equivalently:	~64%	:	~36%

Results of the flips	Hypothesis FAIR		Hypothesis BIASED
H	1		1.32
T	1		0.74
T	1		0.74
H	1		1.32
T	1		0.74
Sum	5		4.85
Posterior	2^{1-5}	:	$2^{0-4.85}$



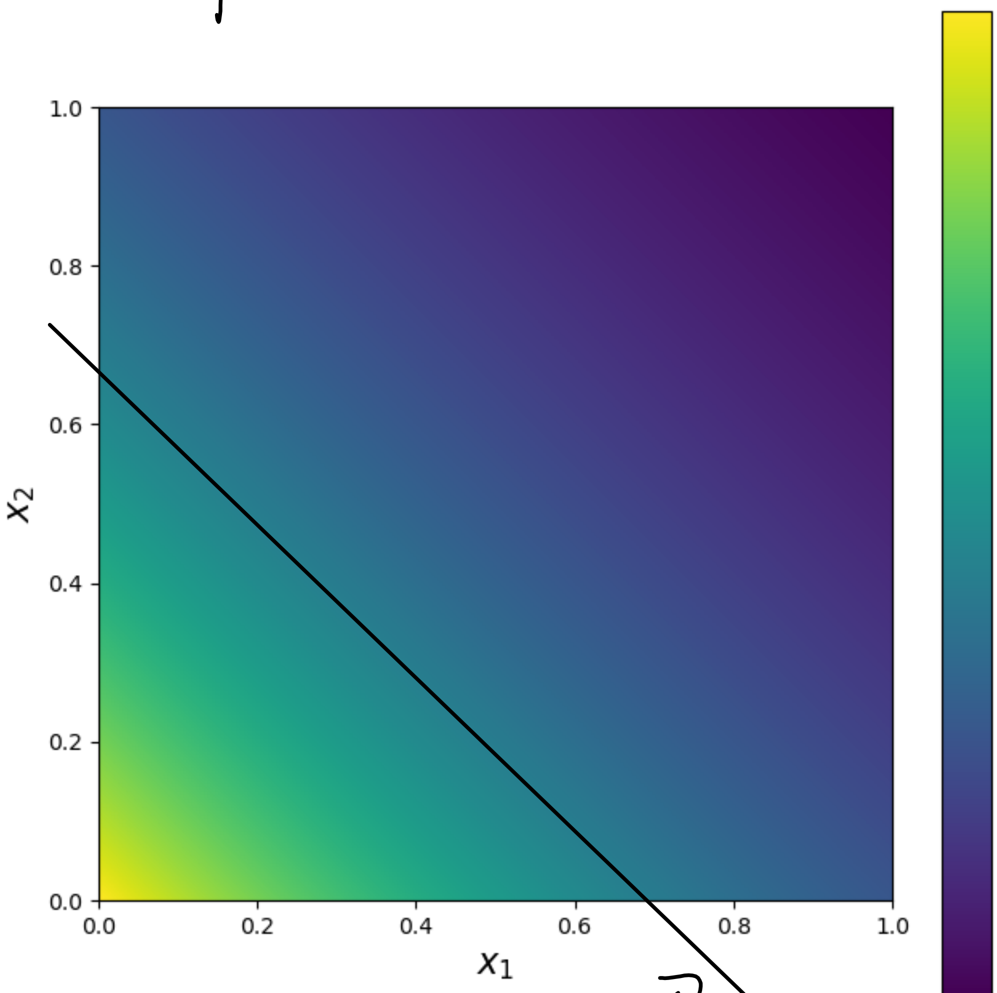
$\min_{q \in Q} D(p,q)$
σ^2
μ
Q
Gaussians
P
all distributions


probability distribution over all distributions
We have $P(\hat{q}=p) \approx 2^{-D(p\|q)\cdot N}$
so it's concentrated around q
q
$\tilde{p}$
P_0: all possible empirical distributions
the subset P of empirical distributions
probability distribution restricted to P
is concentrated around
$\tilde{p} = \operatorname{argmin}_{p \in P} D(p\|q)$

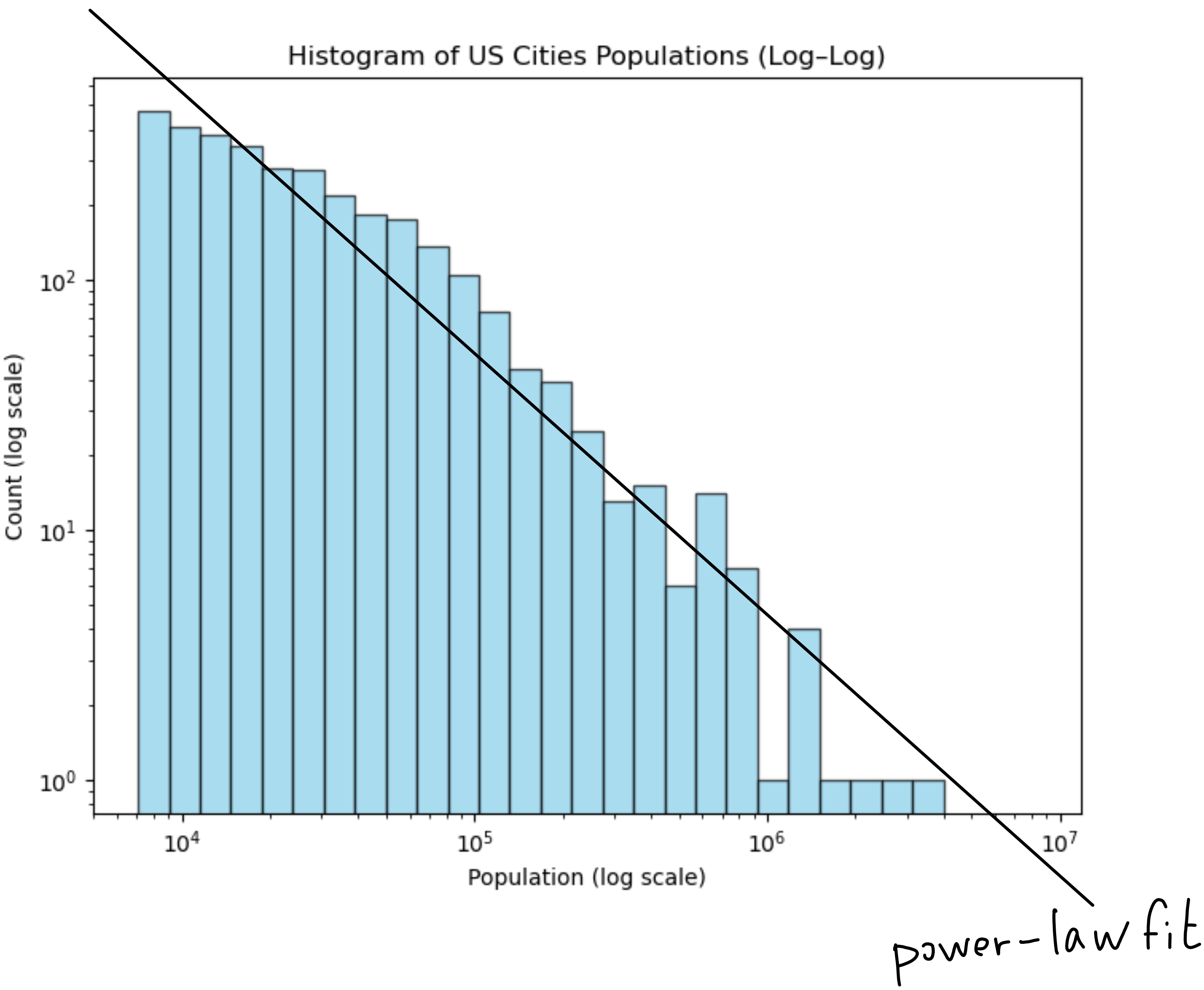


σ^2
μ
Gaussians
$\min D(p, q_{unif})$
$E_p[X]=\mu,$
$E_p[(X-\mu)^2]=\sigma^2$
qunif
all distributions

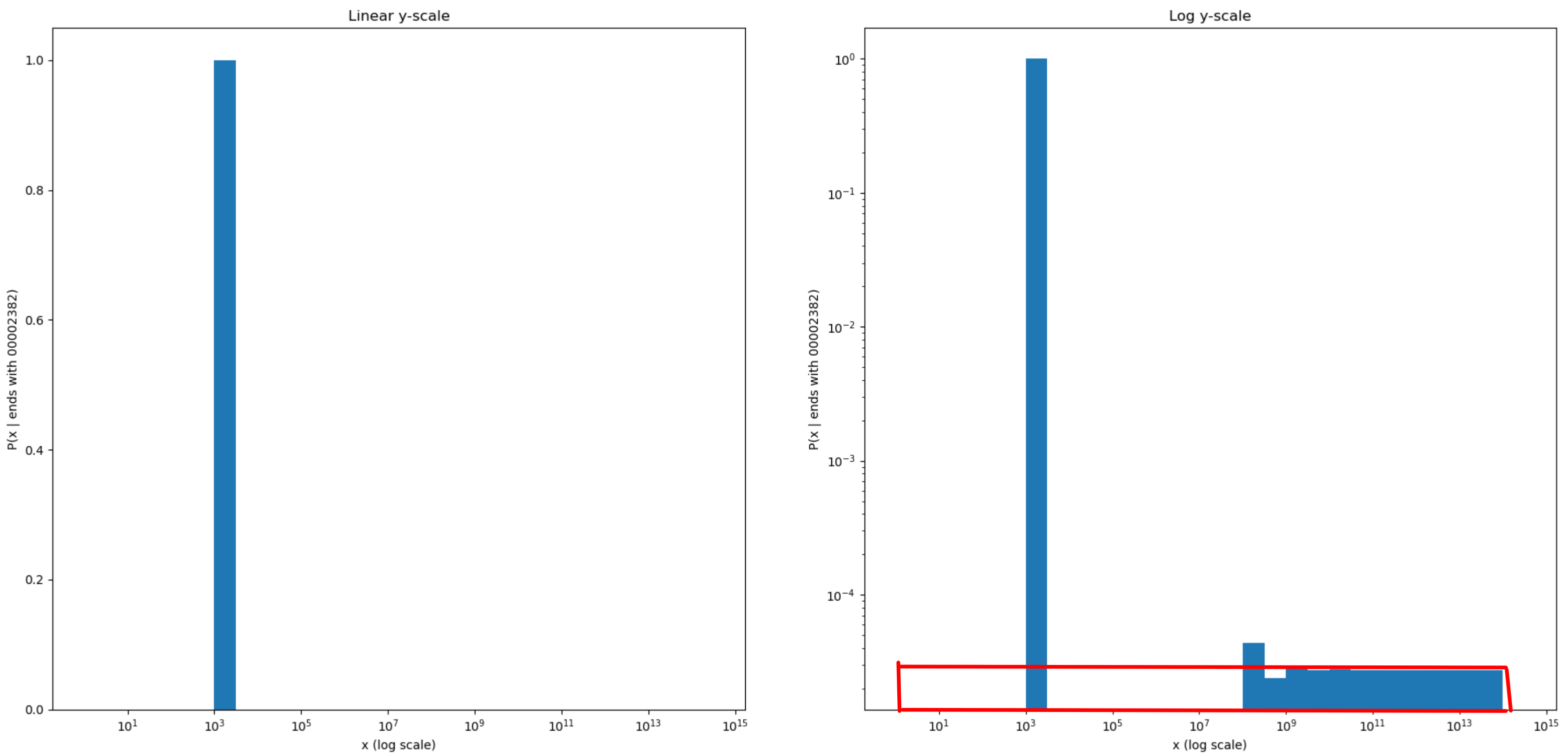
$$p(x_1, x_2) \propto e^{-x_1 - x_2}$$

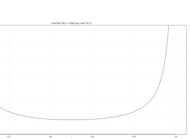


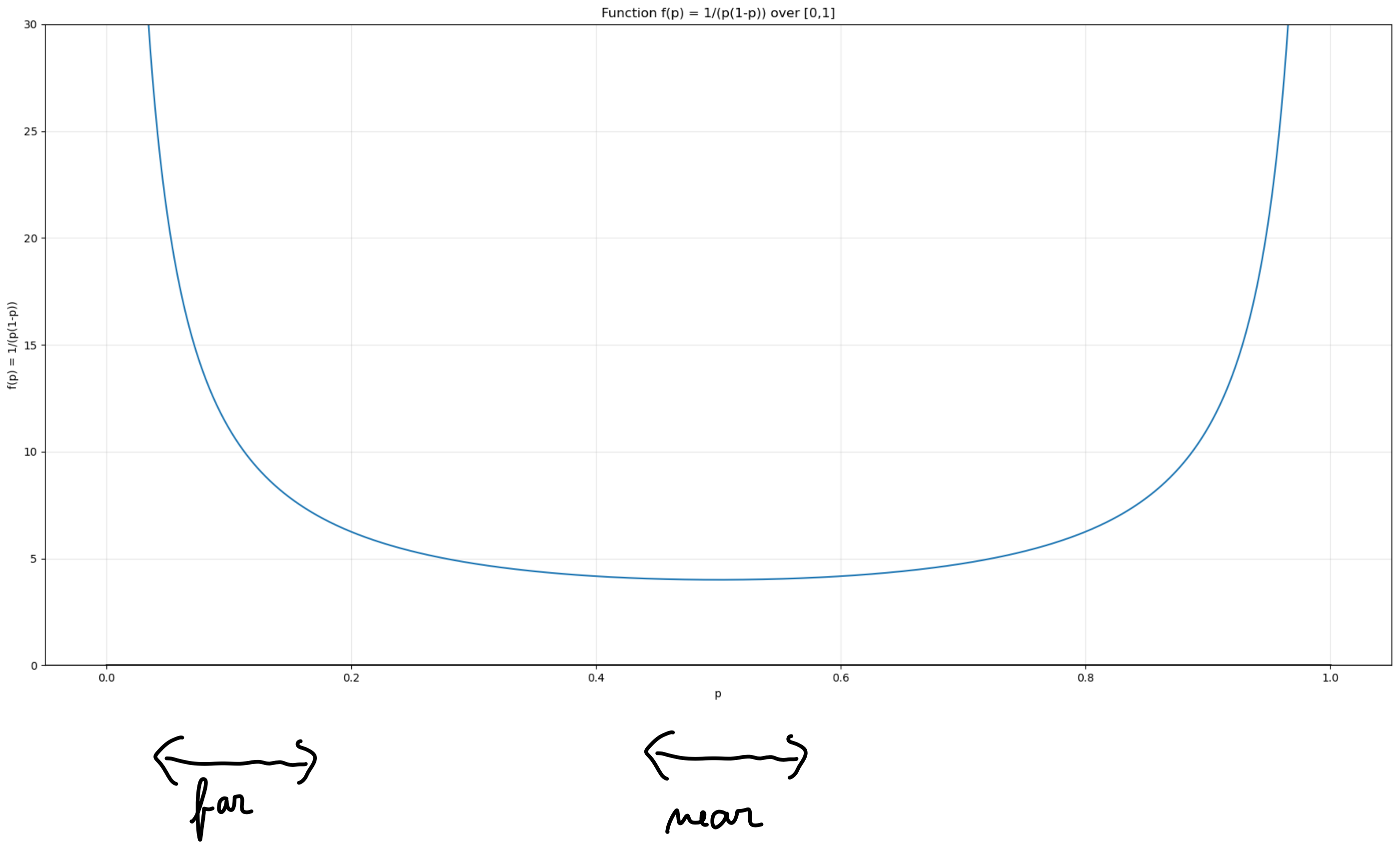
uniform density here



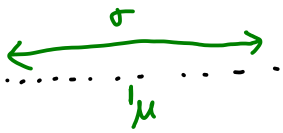
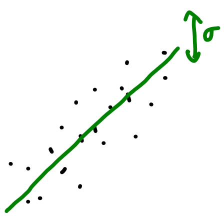
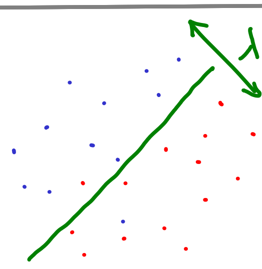
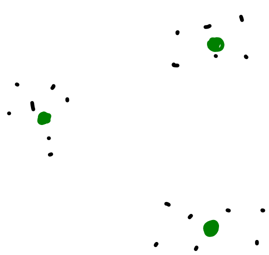
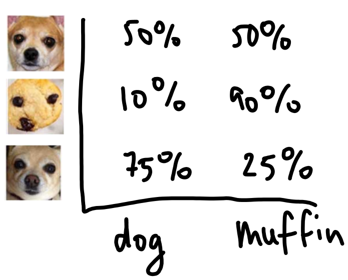
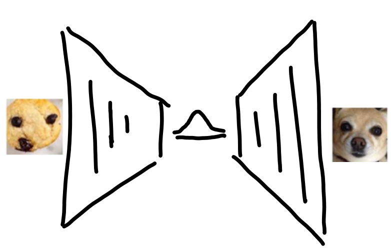
Histogram of US Cities Populations (Log–Log)
Count (log scale)
10^2
10^1
10^0
10^4
10^5
10^6
10^7
Population (log scale)
power-law fit

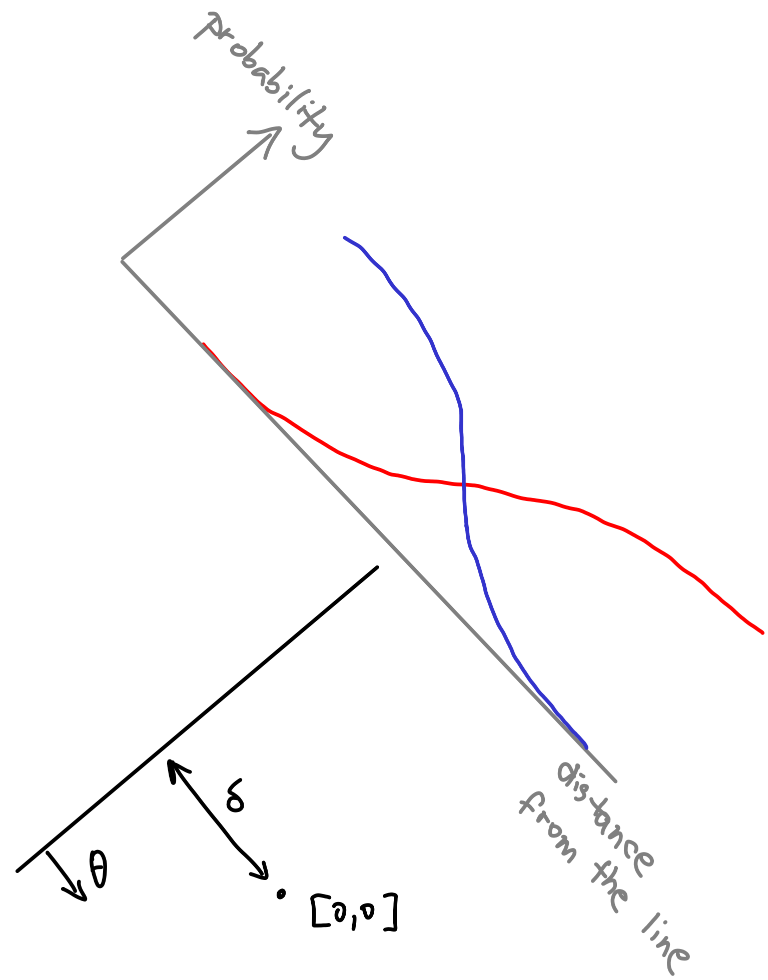

Linear y-scale
P(x | ends with 00002382)
1.0
0.8
0.6
0.4
0.2
0.0
10^1
10^3
10^5
10^7
10^9
10^{11}
10^{13}
10^{15}
x (log scale)
Log y-scale
P(x | ends with 00002382)
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^1
10^3
10^5
10^7
10^9
10^{11}
10^{13}
10^{15}
x (log scale)



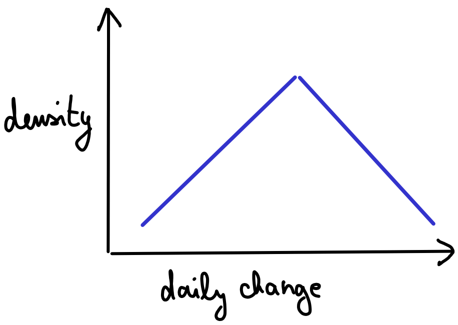

Function f(p) = 1/(p(1-p)) over [0,1]
f(p) = 1/(p(1-p))
p
far
near

Rough picture $\xrightarrow{\text{max entropy}}$ Probabilistic model $\xrightarrow{\text{max likelihood}}$ Loss function

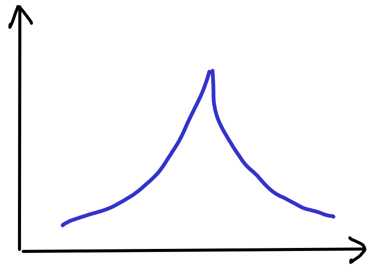
	Rough picture	Probabilistic model	Loss function
mean + variance		$p(x) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\hat{\mu} = \operatorname{argmin}_{\mu} \sum_{i=1}^{n} (x_i-\mu)^2$
linear regression		$p(y\|x) \propto e^{\frac{-(y-(ax+b))^2}{2\sigma^2}}$	$\hat{a},\hat{b} = \operatorname{argmin}_{a,b} \sum_{i=1}^{n} (y_i-(ax_i+b))^2$
logistic regression		$p(\bullet \| (x,y)) \propto \sigma(\lambda(\theta\cdot(x,y)+\delta))$	too horrible to write here
k-means clustering		$p(x\|\mu_j) \propto e^{-\frac{\\|x-\mu_j\\|^2}{2\sigma^2}}$	$\hat{\mu}_1,\dots,\hat{\mu}_k = \operatorname{argmin}_{\mu_1,\dots,\mu_k} \sum_{i=1}^{n} \min_{j=1}^{k} \\|x-\mu_j\\|^2$
classifying images		$p_j(X) \propto e^{-\lambda NN_j(X)}$	$\operatorname{argmin}_{\text{weights}} \sum_{i=1}^{m} \lg \frac{1}{p_{\ell_i}(x_i)}$
Variational Autoencoder		$p(x,y) \nearrow p(x)\,p'(y\|x)$, $p(x,y) \searrow q(y)\,q(x\|y)$	too horrible to write here


probability
distance
from the line
δ
θ
[0,0]

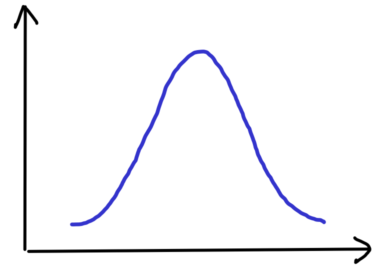
A

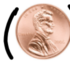


B



C



$p($$) = p($$) = \frac{1}{2}$

model: $q($$) = \frac{1}{2}$

$q($$) = \frac{1}{2}$

model: $q($$) = 0.999$

$q($$) = 0.001$

Medium surprise

Large surprise

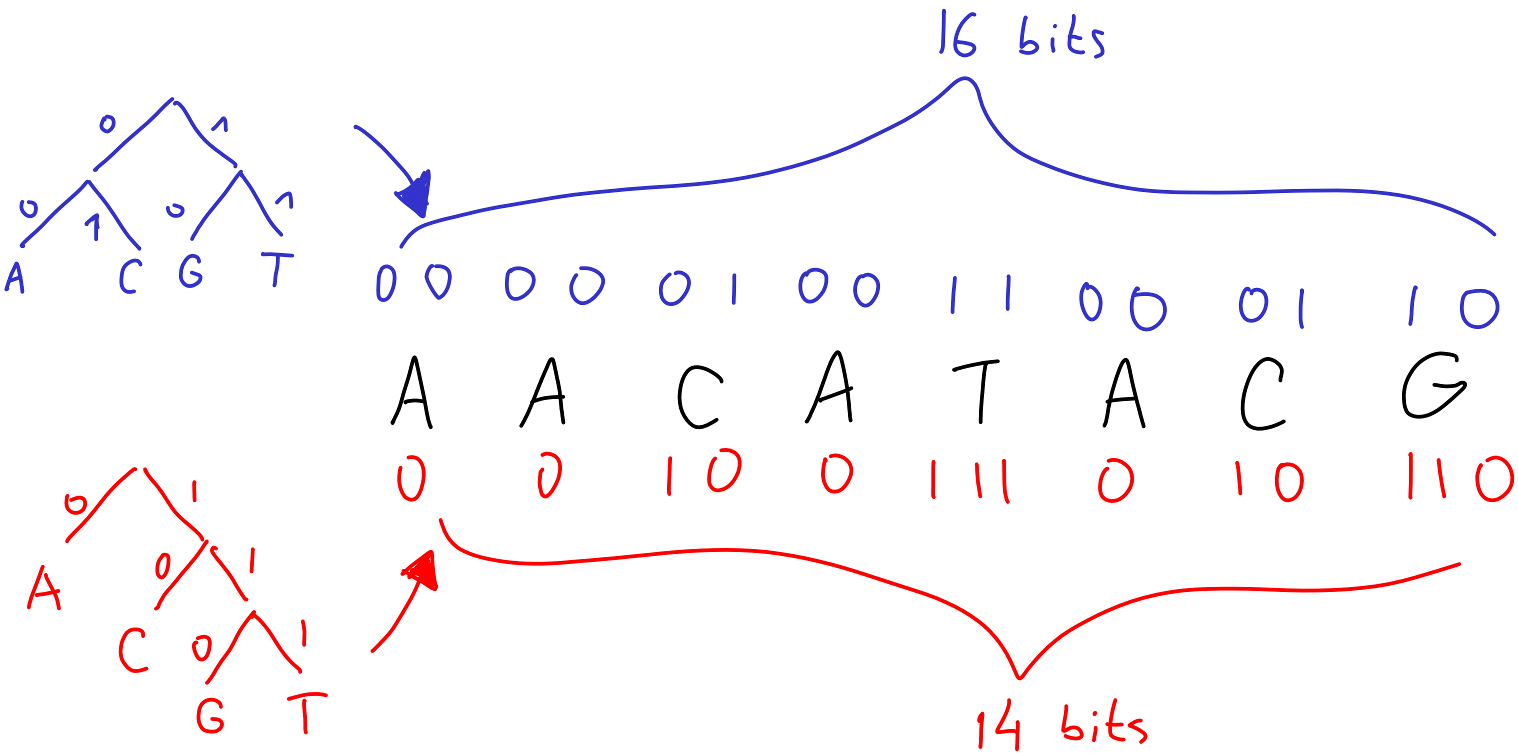
p() = $\frac{1}{2}$

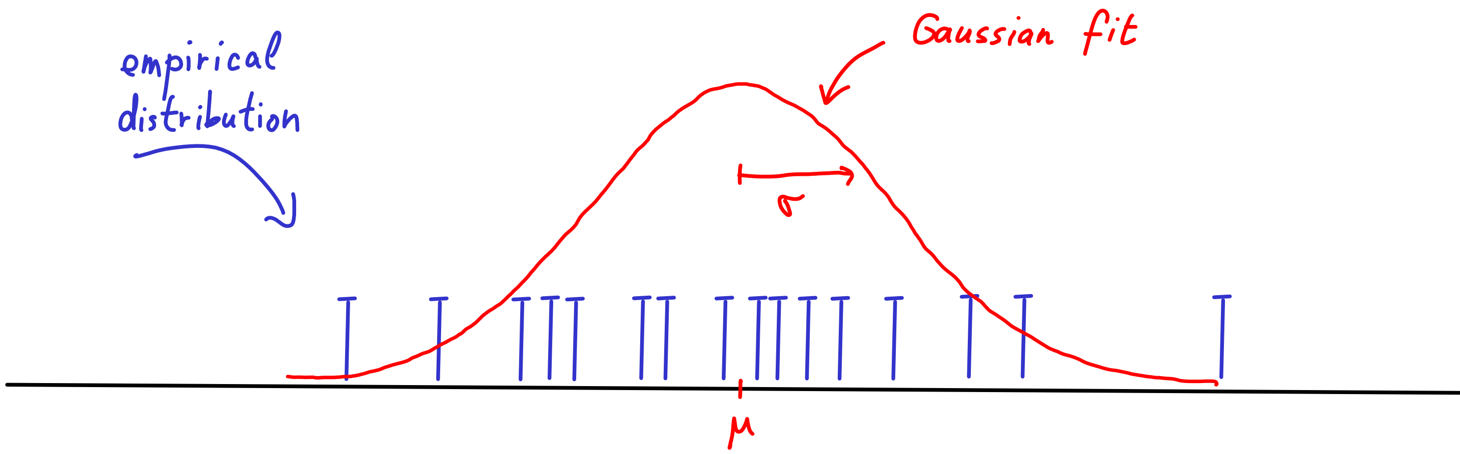
Medium surprise

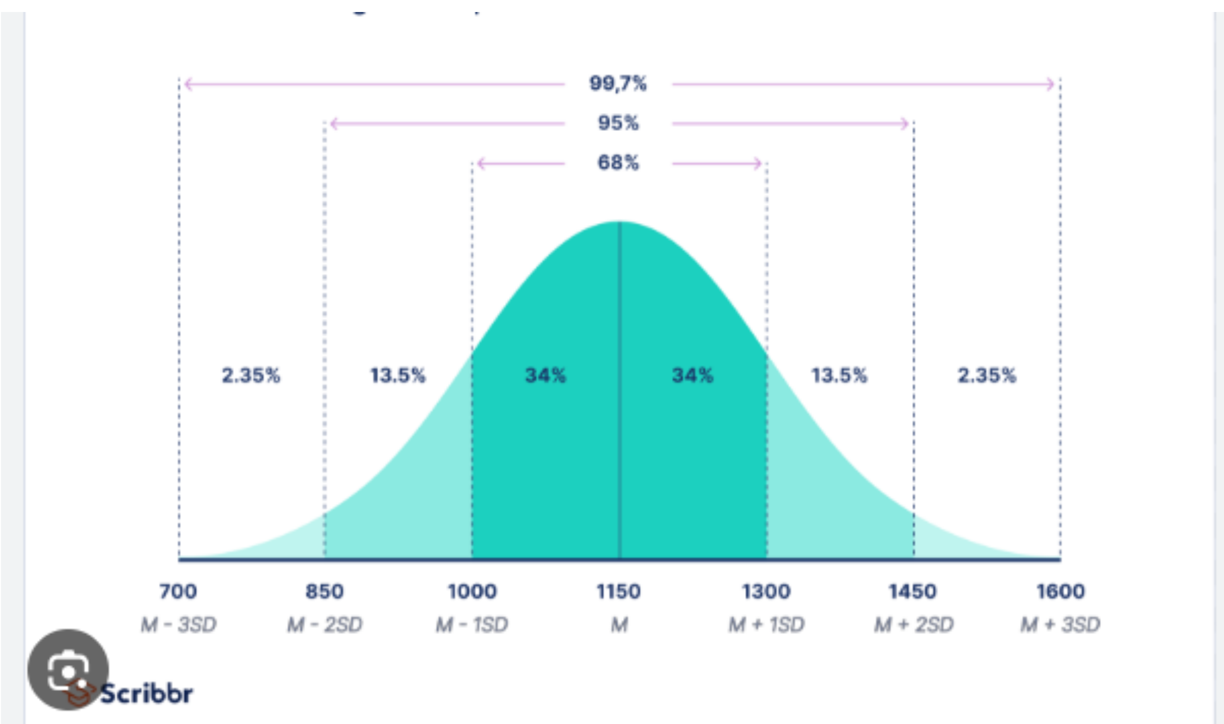
p() = 0.001

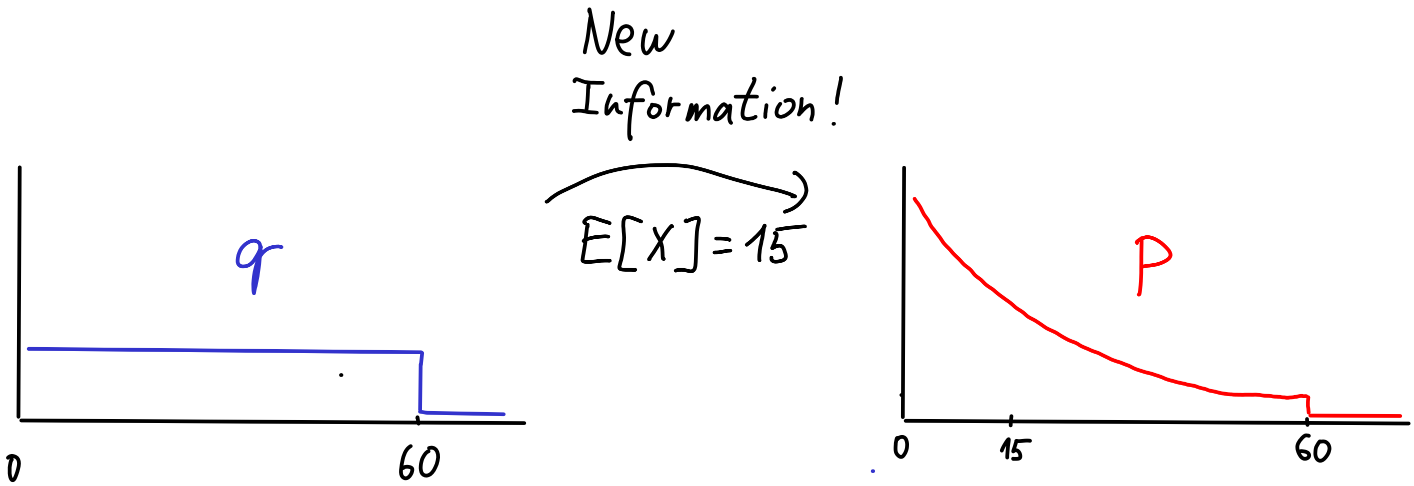
Small average surprise

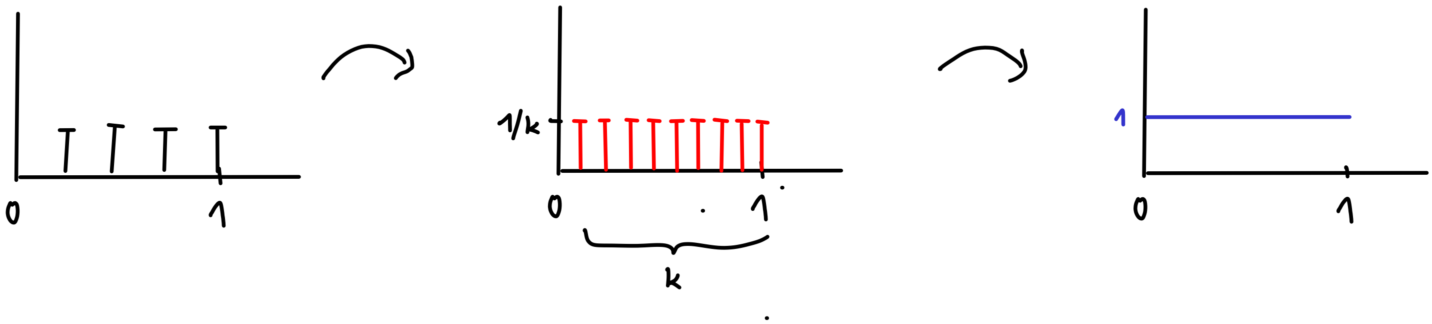
(though  $\Rightarrow$ large surprise)


16 bits
A C G T
00 00 01 00 11 00 01 10
A A C A T A C G
0 0 10 0 111 0 10 110
A C G T
14 bits


empirical distribution
Gaussian fit
σ
μ


99,7%
95%
68%
2.35%
13.5%
34%
34%
13.5%
2.35%
700
M - 3SD
850
M - 2SD
1000
M - 1SD
1150
M
1300
M + 1SD
1450
M + 2SD
1600
M + 3SD
Scribbr


New
Information!
E[X] = 15
q
0
60
p
0
15
60



0
1
1/k
0
1
k
1
0
1

assumption
of independence

	🚶‍♀️	🚲	🚌	
☀️	?	?	?	70%
☁️	?	?	?	30%
	20%	30%	50%	

we only know
the marginals

	🚶‍♀️	🚲	🚌
☀️	14%	21%	35%
☁️	6%	9%	15%