Lecture 03:

EDA I and Sampling Methods

Markus Hohle

University California, Berkeley

Data Science for Scientific Computing

MSSE 277A, 3 Units

Berkeley MSSE
MOLECULAR SCIENCE & SOFTWARE ENGINEERING

EDA I and Sampling Methods

Outline

Probabilities

Entropy and Information

Sampling Methods

- Random Sampling
- Stratified Sampling
- Systematic Sampling
- Cluster Sampling
- Bootstrap Sampling
- Oversampling & Undersampling

EDA I and Sampling Methods

Outline

Probabilities

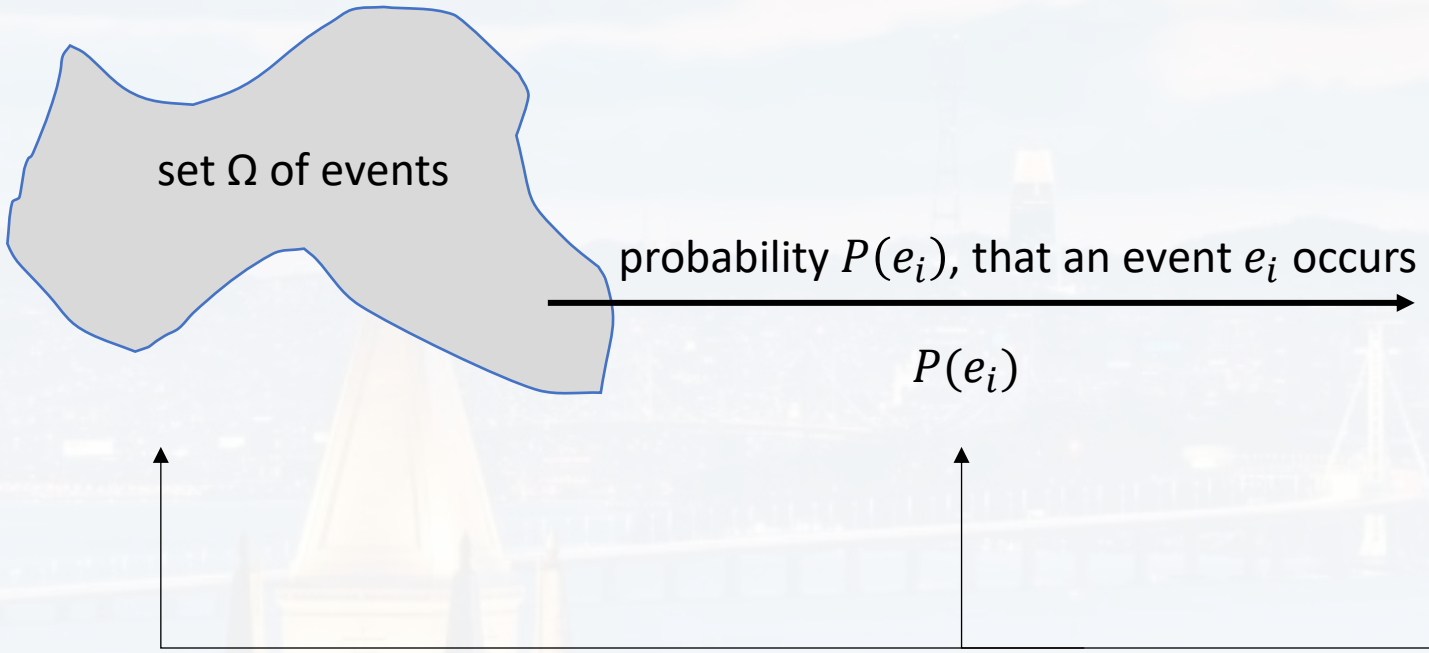
Entropy and Information

Sampling Methods

- Random Sampling
- Stratified Sampling
- Systematic Sampling
- Cluster Sampling
- Bootstrap Sampling
- Oversampling & Undersampling

here: **heuristic explanation** → more mathematical rigorous: see **Cox's theorem**

axioms of probability
important quantities
the sampling error

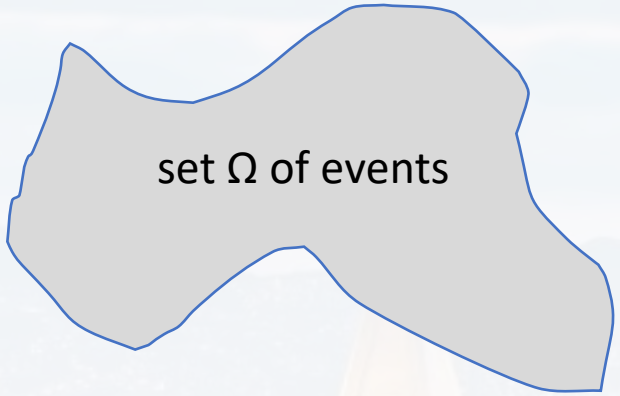


from data/observations → deriving Ω

from data/observations → understanding process
→ finding $P(e_i)$

here: **heuristic explanation** → more mathematical rigorous: see **Cox's theorem**

axioms of probability
important quantities
the sampling error

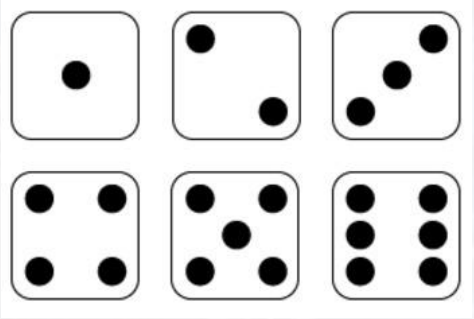


from data/observations → deriving Ω

1 2 2 1 2 2 1 2 1 1 2 1 1 1 2 observations

Is the observation 3 just rare **(and that's why we haven't observed it)**, or is $3 \notin \Omega$?

events can be ***discrete***

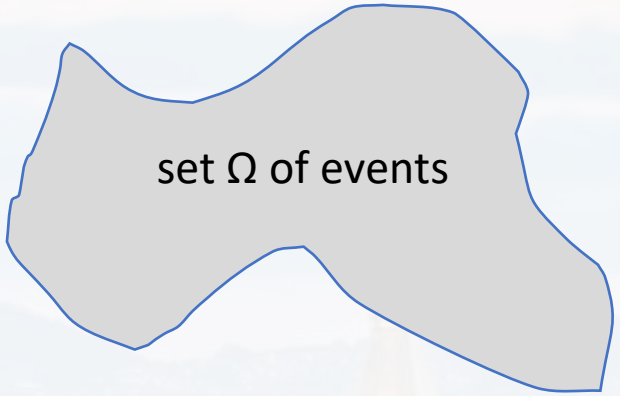


or ***continuous***

- car speed measured in a speed trap
- a person's weight, etc

axioms of probability
important quantities
the sampling error

$P(e_i)$, that an event e_i occurs



1st axiom: $P(e_i)$ is a **non-negative, real number**

2nd axiom: the probability that at **least one** of the events in the entire sample space will occur is **1**
if events are **collectively exhaustive**

from 1st and 2nd : $P(e_i) = [0, 1]$ for any e_i

If events are **mutually exclusive**: **3rd axiom:** $P(\bigcup_{i=1}^{\infty} e_i) = \sum_{i=1}^{\infty} P(e_i)$ ∪: **"union"**

$\bigcup_{i=1}^{\infty} e_i$ means e_1 **or** e_2 **or** …. e_∞

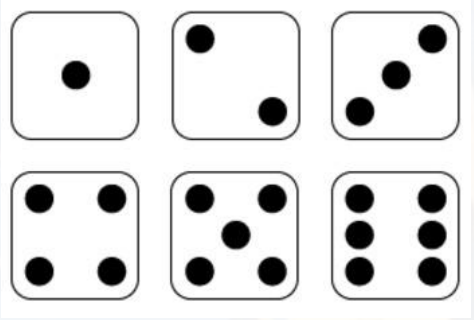
axioms of probability
important quantities
the sampling error

$P(e_i)$, that an event e_i occurs

If events are **mutually exclusive**: **3rd axiom:** $P(\bigcup_{i=1}^{\infty} e_i) = \sum_{i=1}^{\infty} P(e_i)$

∪: **"union"**

$\bigcup_{i=1}^{\infty} e_i$ means e_1 **or** e_2 **or** …. e_∞



The probability that we roll a **4 or a 6** equals…

$$P\left(e_4 \bigcup e_6\right) =$$

…the **probability** that we roll a **4 plus** the **probability** that we roll a **6**

$$P(e_4) + P(e_6)$$

"or" equals addition!

axioms of probability
important quantities
the sampling error

$P(e_i)$, that an event e_i occurs

"or" equals addition!

$$P(\cup_{i=1}^{\infty} e_i) = \sum_{i=1}^{\infty} P(e_i)$$

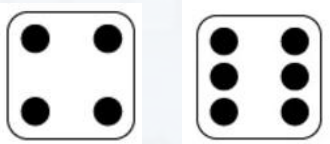
∪: **"union"**

"and" equals multiplication!

$$P\left(\bigcap_{i=1}^{\infty} e_i\right) = \prod_{i=1}^{\infty} P(e_i)$$

two dice: The probability that we roll a **4 and a 6** equals…

$$P\left(e_4 \bigcap e_6\right) =$$



∩: **"intersection"**

…the **probability** that we roll a **4 times** the **probability** that we roll a **6**

$$P(e_4)P(e_6)$$

axioms of probability
important quantities
the sampling error

$P(e_i)$, **that an event** e_i **occurs**

"or" equals addition!

$$P(\bigcup_{i=1}^{\infty} e_i) = \sum_{i=1}^{\infty} P(e_i)$$

∪: **"union"**

∩: **"intersection"**

"and" equals multiplication!

$$P\left(\bigcap_{i=1}^{\infty} e_i\right) = \prod_{i=1}^{\infty} P(e_i)$$

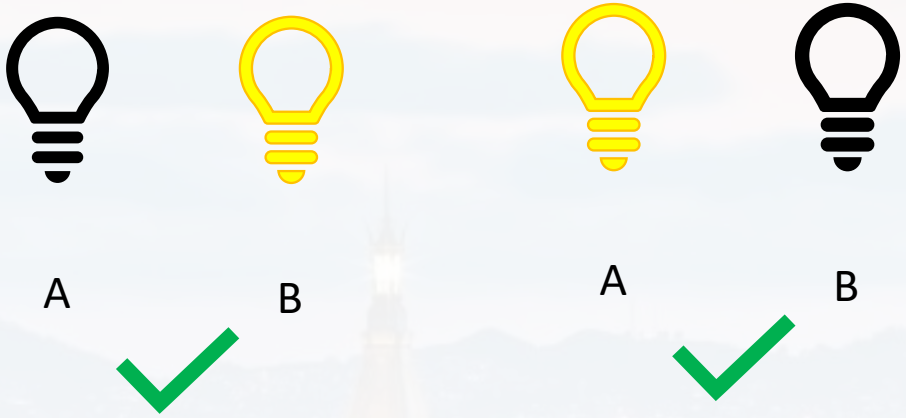
Be carful if events are not mutually exclusive (like a **set of events** or a **sequence of events**)

two light bulbs A and B:

A B

What is the probability
that A or B is turned on?

two light bulbs A and B:



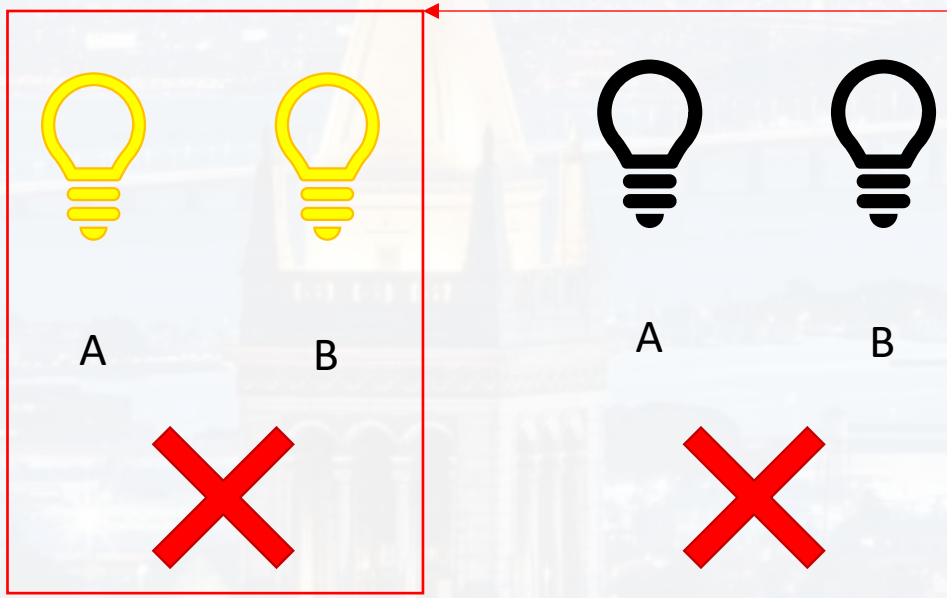
What is the probability that A **or** B is turned on?

axioms of probability
important quantities
the sampling error

$P(A) + P(B)$

∪: "union"

∩: "intersection"



$P(A)P(B)$

$$P\left(A \bigcup B\right) = P(A) + P(B) - P\left(A \bigcap B\right)$$

$$= P(A) + P(B) - P(A)P(B)$$

axioms of probability
important quantities
the sampling error

$P(e_i)$, that an event e_i occurs

"or" equals addition!

$$P(\bigcup_{i=1}^{\infty} e_i) = \sum_{i=1}^{\infty} P(e_i)$$

∪: **"union"**

∩: **"intersection"**

"and" equals multiplication!

$$P\left(\bigcap_{i=1}^{\infty} e_i\right) = \prod_{i=1}^{\infty} P(e_i)$$

$$P\left(e_4 \bigcup e_6\right) = P(e_4) + P(e_6) - P(e_4)P(e_6)$$

inclusion - exclusion principle

$$P\left(A \bigcup B\right) = P(A) + P(B) - P\left(A \bigcap B\right)$$

$$P\left(A \bigcup B \bigcup C\right) = P(A) + P(B) + P(B) + P(A)P(B)P(C) - P(A)P(B) - P(A)P(C) - P(C)P(B)$$

complement probability for not A, $\bar{A}$:

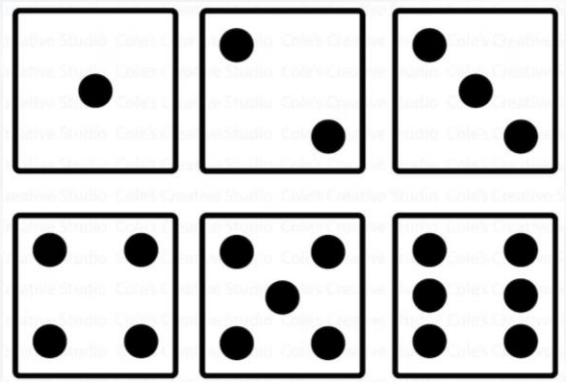
$$P(\bar{A}) = 1 - P(A)$$

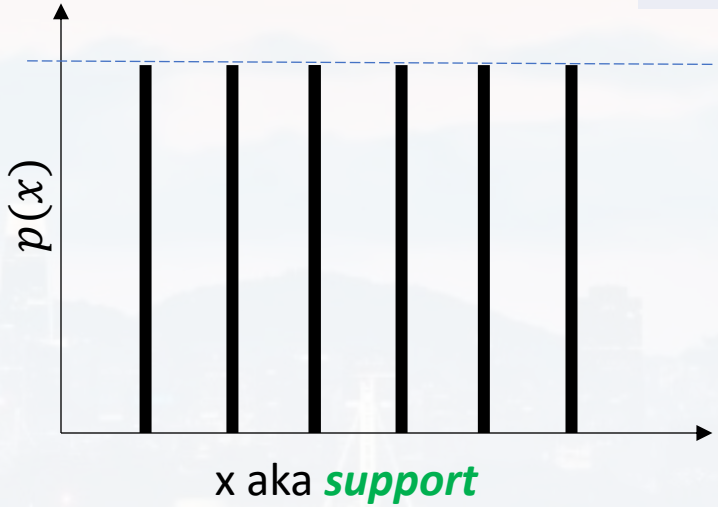
because: $P(e_i) = [0, 1]$ for any e_i

$p(x)$, **that an event x occurs**

axioms of probability
important quantities
the sampling error

discrete (= countable)





fair die

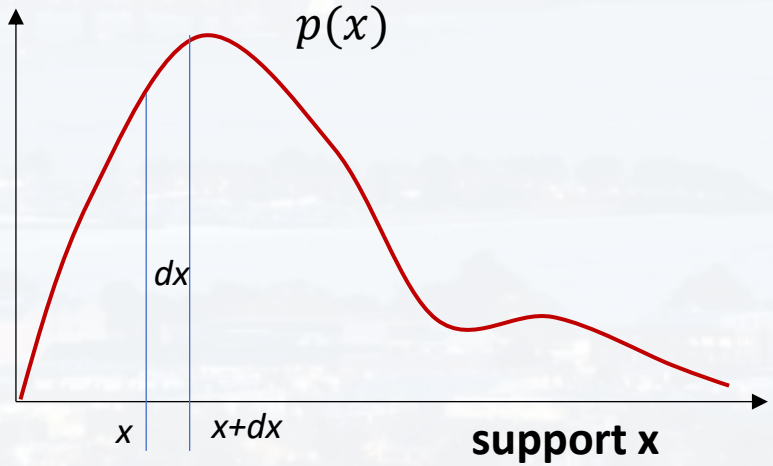
continuous

[a ≤ x ≤ b]

p(x) doesn't make sense

→ *p(x) dx*

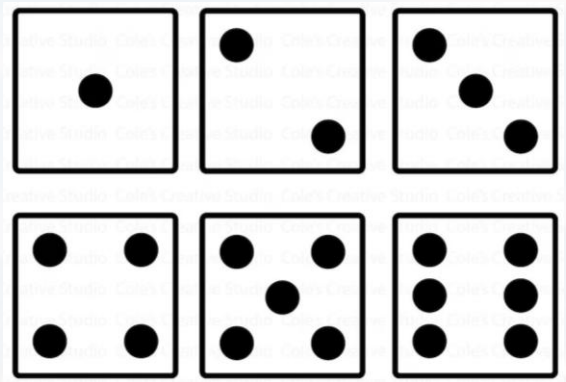
probability **d**ensity **f**unction

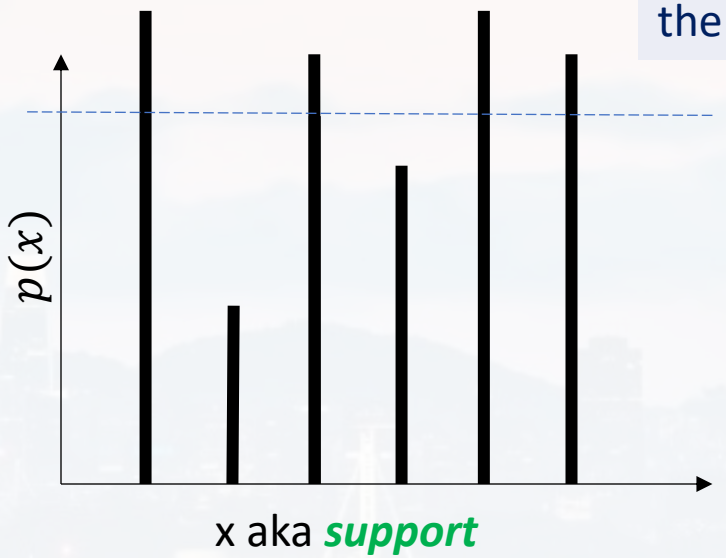


$p(x)$, **that an event x occurs**

axioms of probability
important quantities
the sampling error

discrete (= countable)





biased die

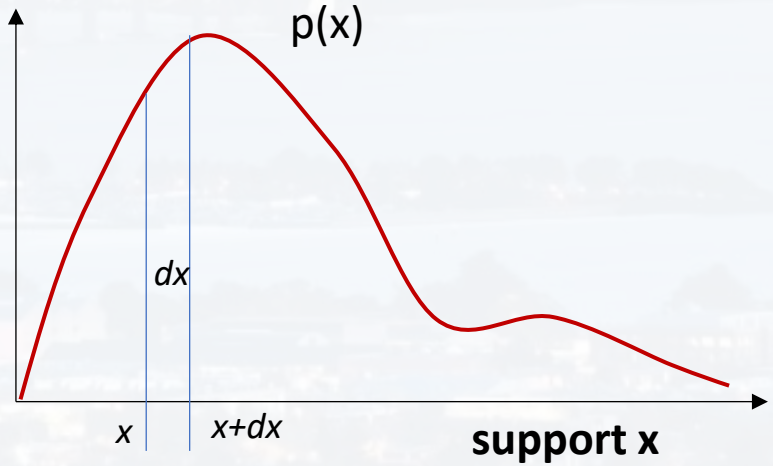
continuous

[a ≤ x ≤ b]

p(x) doesn't make sense

→ *p(x) dx*

probability **d**ensity **f**unction



EDA I and Sampling Methods: Probabilities

$p(x)$, that an event x occurs

axioms of probability
important quantities
the sampling error

	the mean μ (barycenter)	**the variance σ^2** (natural scatter)
discrete (= countable)	$\mu = E(x) = \sum_i x_i \, p(x_i)$	$\sigma^2 = var(x) = \sum_i (x_i - \mu)^2 \, p(x_i)$
continuous	$\mu = E(x) = \int x \, p(x) \, dx$	$\sigma^2 = var(x) = \int (x - \mu)^2 \, p(x) \, dx$

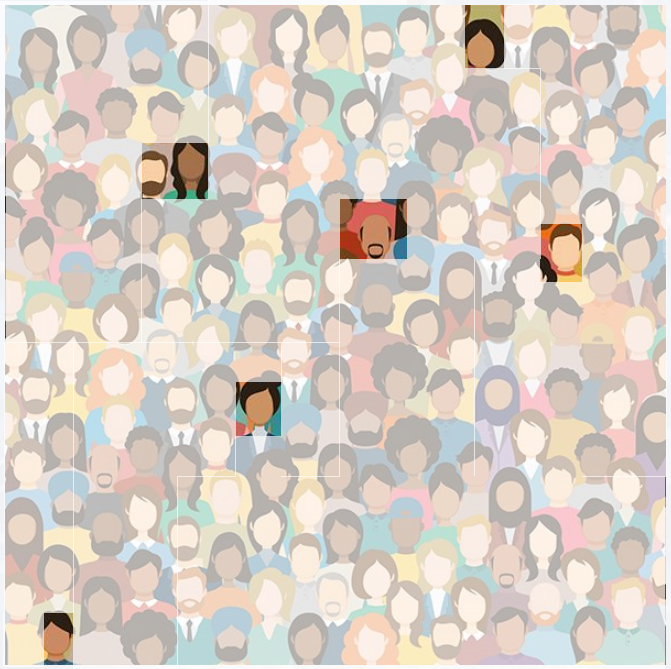
EDA I and Sampling Methods: Probabilities

axioms of probability
important quantities
the sampling error

task: you want to find out which **fraction q** of the population doesn't like vanilla ice cream

problem: it is not possible, to ask everyone, so you need to estimate a value for q, $\overline{q}$.

solution: you can estimate the error of $\overline{q}$, $\varepsilon_{\overline{q}}$ as a function of number n of people you asked



all N people

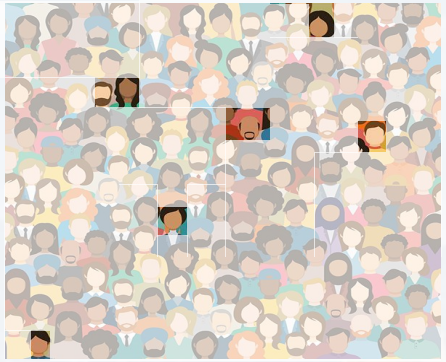
those who don't like vanilla ice cream

N: population size
n: sample size

EDA I and Sampling Methods: Probabilities

task: you want to find out which **fraction q** of the population doesn't like vanilla ice cream

axioms of probability
important quantities
the sampling error



each time you interview a person, (s)he

- doesn't like vanilla ice cream with a probability of q
- does like vanilla ice cream with a probability of $1 - q$

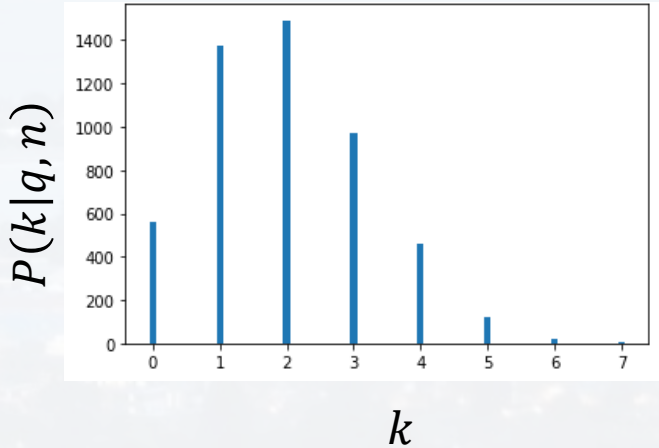
N: population size
n: sample size

What is the probability $P(k|q,n)$ to find k people who don't like vanilla ice cream if you randomly ask n people?

answer: it is a **binomial problem**

$$P(k|q,n) = \binom{n}{k} q^k (1-q)^{n-k}$$

binomial distribution



axioms of probability
important quantities
the sampling error

task: you want to find out which **fraction q** of the population doesn't like vanilla ice cream

What is the probability $P(k|q,n)$ to find $\boldsymbol{k}$ people who don't like vanilla ice cream if you randomly ask $\boldsymbol{n}$ people?

N: population size
n: sample size

answer: it is a **binomial problem**

$$P(k|q,n) = \binom{n}{k} q^k (1-q)^{n-k}$$

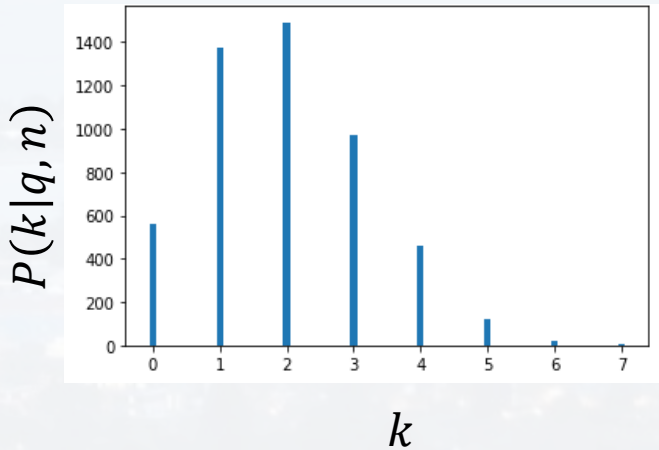
binomial distribution

Each time we'd ask n randomly chosen people, k would be different!

variance of k (binomial)

$$var(k) = \sum_{k=0}^{n} (k - qn)^2 \binom{n}{k} q^k (1-q)^{n-k} = \boldsymbol{qn(1-q)}$$

$$\boldsymbol{var(x) = \sum_i (x_i - \mu)^2 p(x_i)}$$



axioms of probability
important quantities
the sampling error

task: you want to find out which **fraction q** of the population doesn't like vanilla ice cream

$$P(k|q,n) = \binom{n}{k} q^k (1-q)^{n-k}$$

binomial distribution

N: population size
n: sample size

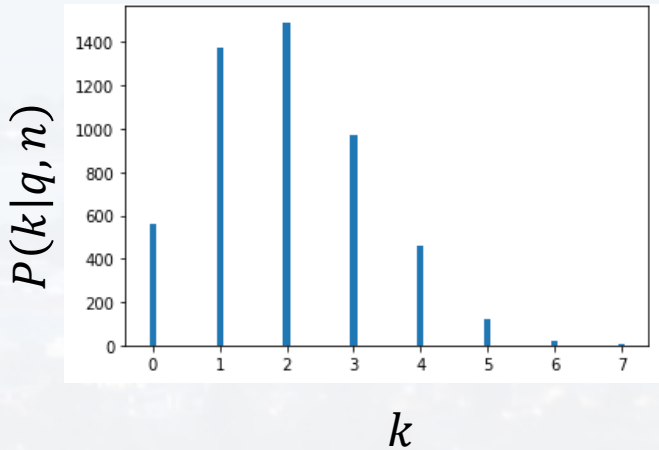
Each time we'd ask n randomly chosen people, k would be different!

$$var(k) = \boldsymbol{qn(1-q)} \qquad \sigma_k(k) = \sqrt{qn(1-q)}$$

our **estimate** for q would be $\bar{q} = \frac{k}{n}$

i. e. the **uncertainty** for q (**error propagation!**) would be

$$\varepsilon_{\bar{q}} = \frac{\sigma_k(k)}{n} = \frac{\sqrt{qn(1-q)}}{n} = \frac{\sqrt{q(1-q)}}{\sqrt{n}} \sim \frac{1}{\sqrt{n}}$$



EDA I and Sampling Methods: Probabilities

axioms of probability
important quantities
the sampling error

task: you want to find out which **fraction q** of the population doesn't like vanilla ice cream

the **estimate** for q $\quad \bar{q} = \frac{k}{n}$

N: population size
n: sample size

the **uncertainty** for q would be $\quad \varepsilon_{\bar{q}} = \frac{\sqrt{q(1-q)}}{\sqrt{n}} \sim \frac{1}{\sqrt{n}}$

note:

- equation only applies for $N = \infty$
- one can show for $N \neq \infty$ that $\varepsilon_{\bar{q}} = \frac{\sqrt{q(1-q)}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$ Finite Population Correction
- population mean: $\mu = \frac{1}{N}\sum_i^N x_i$ sample mean: $\bar{x} = \frac{1}{n}\sum_i^n x_i$
- population variance: $\sigma^2 = \frac{1}{N}\sum_i^N (x_i - \mu)^2$ sample variance: $S^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$
- better way for estimating q or other quantities of the population: Bayesian Parameter Estimation

EDA I and Sampling Methods

Outline

Probabilities

Entropy and Information

Sampling Methods

- Random Sampling
- Stratified Sampling
- Systematic Sampling
- Cluster Sampling
- Bootstrap Sampling
- Oversampling & Undersampling

idea: gain of information = **"degree of surprise"**
if something happens that we expected already → no surprise
→ no information

$p(x_i)$: probability of event x_i

↓

$h(x_i)$: function that measures "information"

$$\log[p(x_i)\,p(x_j)] = \log[p(x_i)] + \log[p(x_j)]$$

1) $h(x_i)$ should be additive

2) $h(x_i)$ should be monotonic

3) if x_i and x_j are independent, then

$$h(x_i, x_j) = h(x_i) + h(x_j)$$

$$p(x_i, x_j) = p(x_i)p(x_j)$$

idea: gain of information = **"degree of surprise"**
if something happens that we expected already → no surprise
→ no information

$p(x_i)$: probability of event x_i

$h(x_i)$: function that measures "information"

1) $h(x_i)$ should be additive $h(x_i, x_j) = h(x_i) + h(x_j)$

2) $h(x_i)$ should be monotonic

3) if x_i and x_j are independent, then $p(x_i, x_j) = p(x_i)p(x_j)$

$\log[p(x_i)\, p(x_j)] = \log[p(x_i)] + \log[p(x_j)]$ $p(x_i) \leq 1$

$h(x_i) = -\log[p(x_i)]$ information is ***positive***
low $p(x_i)$ → "great surprise"

idea: gain of information = **"degree of surprise"**
if something happens that we expected already → no surprise
→ no information

$h(x_i) = -\log[p(x_i)]$

$p(x_i)$: **probability of event x_i**

The event x_i is randomly drawn from $p(x_i)$ → ***average*** amount of information

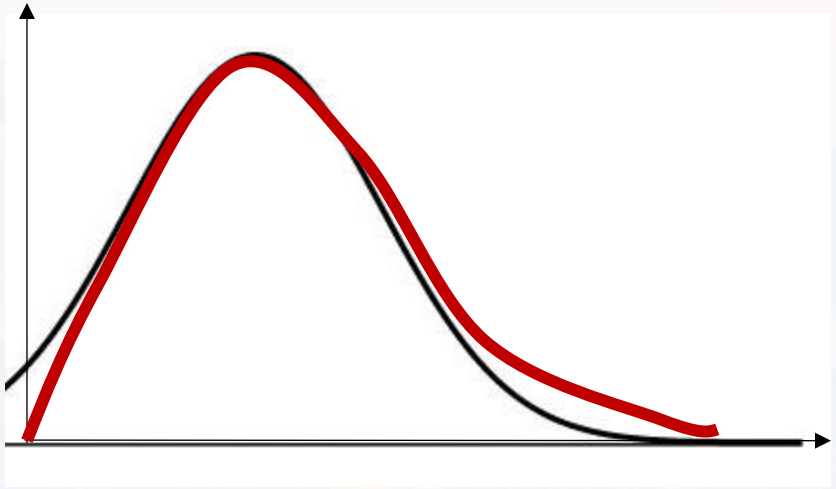
Entropy S

$$S = -\sum_{i=1}^{I} p(x_i) \log[p(x_i)] \quad \text{(discrete)}$$

$$S = -\int p(x) \log[p(x)]\, dx \quad \text{(continuous or } \textit{differential}\text{)}$$

note:
- the **base** of log is **arbitrary** (often 2 or e)
- $\lim_{p \to 0} (p \log p) = 0$
- S is large → no information
- S is zero → all information
- continuous entropy **can** be negative
- continuous entropy is **not** exactly equivalent to discrete entropy (see LDDP)

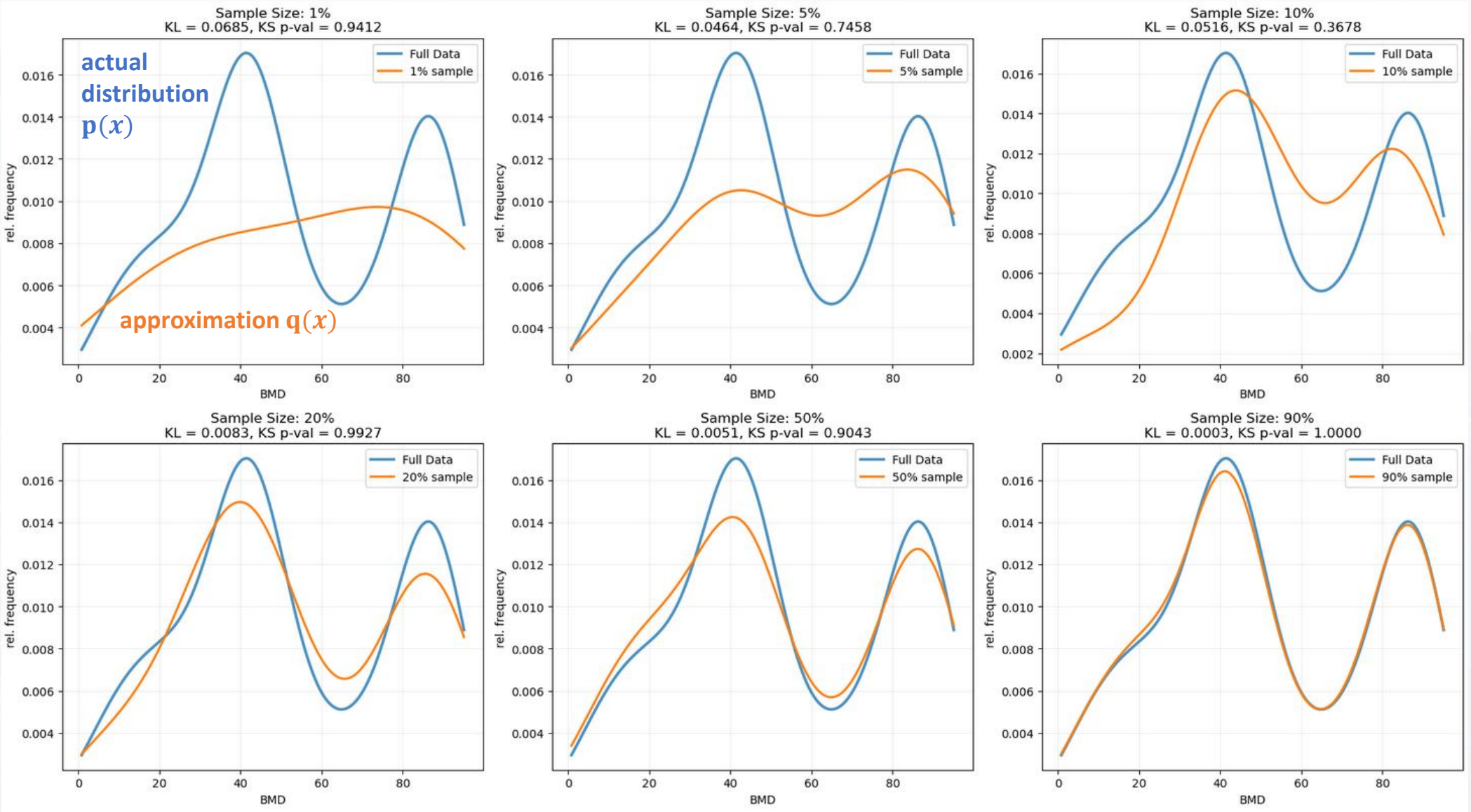
often, we (need to) approximate $\boldsymbol{p(x)}$ by some other distribution $\boldsymbol{q(x)}$



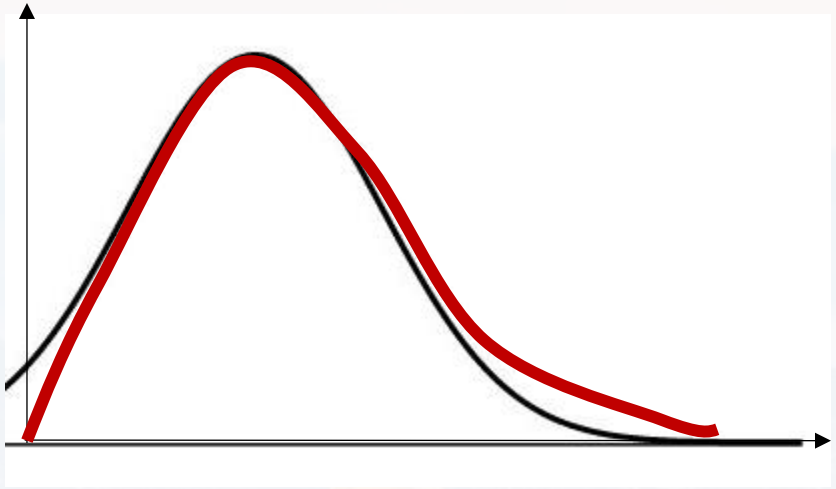
How much is our information "off" if we work with the approximation $\boldsymbol{q(x)}$?

EDA I and Sampling Methods: Entropy and Information

How much is our information "off" if we work with the approximation $\boldsymbol{q(x)}$?



often, we (need to) approximate $\boldsymbol{p(x)}$ by some other distribution $\boldsymbol{q(x)}$



How much is our information "off" if we work with the approximation $\boldsymbol{q(x)}$?

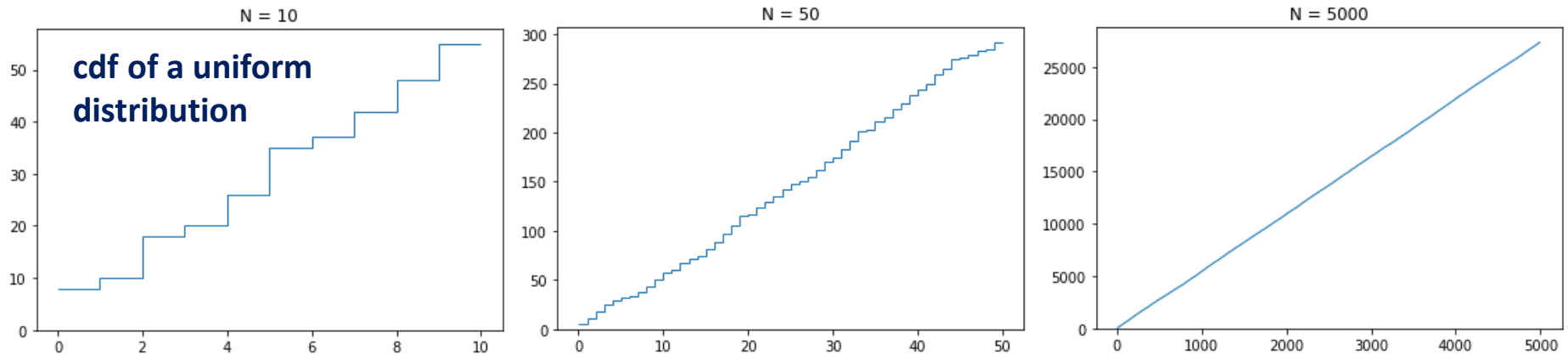
$$-\int p(x)\log[\boldsymbol{q(x)}]\,dx - \left[-\int p(x)\log[\boldsymbol{p(x)}]\,dx\right] = -\int p(x)\log\left[\frac{\boldsymbol{q(x)}}{\boldsymbol{p(x)}}\right]dx = KL(p||q)$$

KL or ***Kullback-Leiber divergence***

- is **not** a distance! $KL(p||q) \neq KL(q||p)$
- ranges from zero (both distributions are equal) to ∞

another way to compare distributions is the **Kolmogorov – Smirnov – test (KS test)**

comparing the **C**umulative **d**ensity **f**unctions $P(x) = \int_a^x p(x)\,dx$



cdf and pdf of a normal distribution

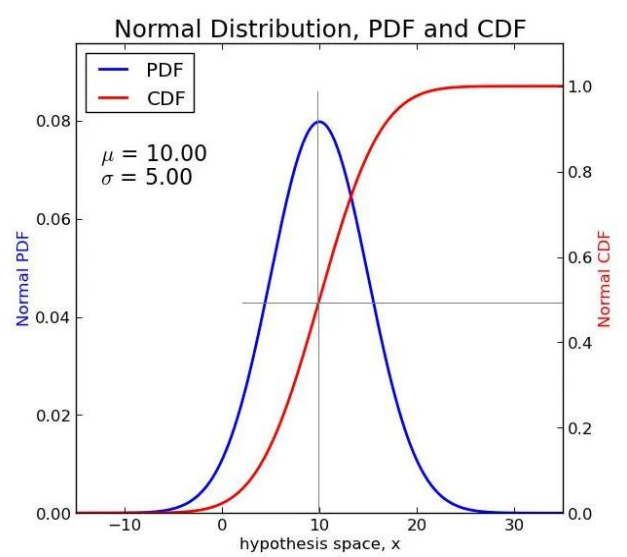
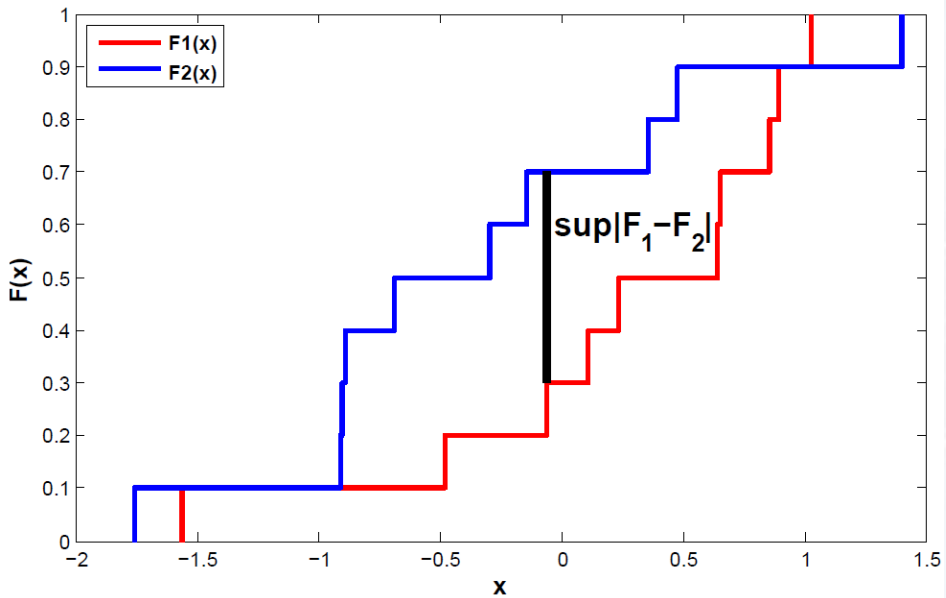


image curtesy

another way to compare distributions is the **Kolmogorov – Smirnov – test (KS test)**

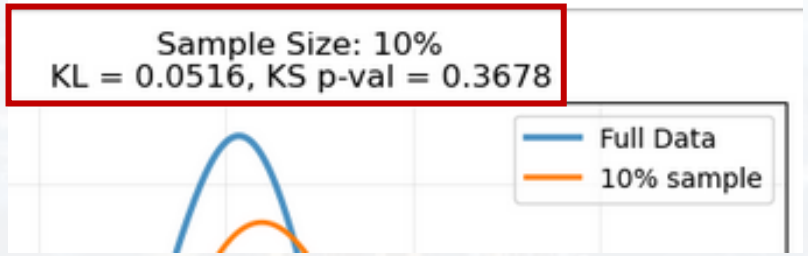
comparing the **C**umulative **d**ensity **f**unctions $P(x) = \int_a^x p(x)\,dx$



KS test compares the cdfs F_1 and F_2 of two pdfs by looking at the largest difference D

→ returns a p-value:

*"The probability that we obtain such a **value (or larger) for D**, assuming **F_1 and F_2** are obtained from the **identical** pdf!"*



EDA I and Sampling Methods

Outline

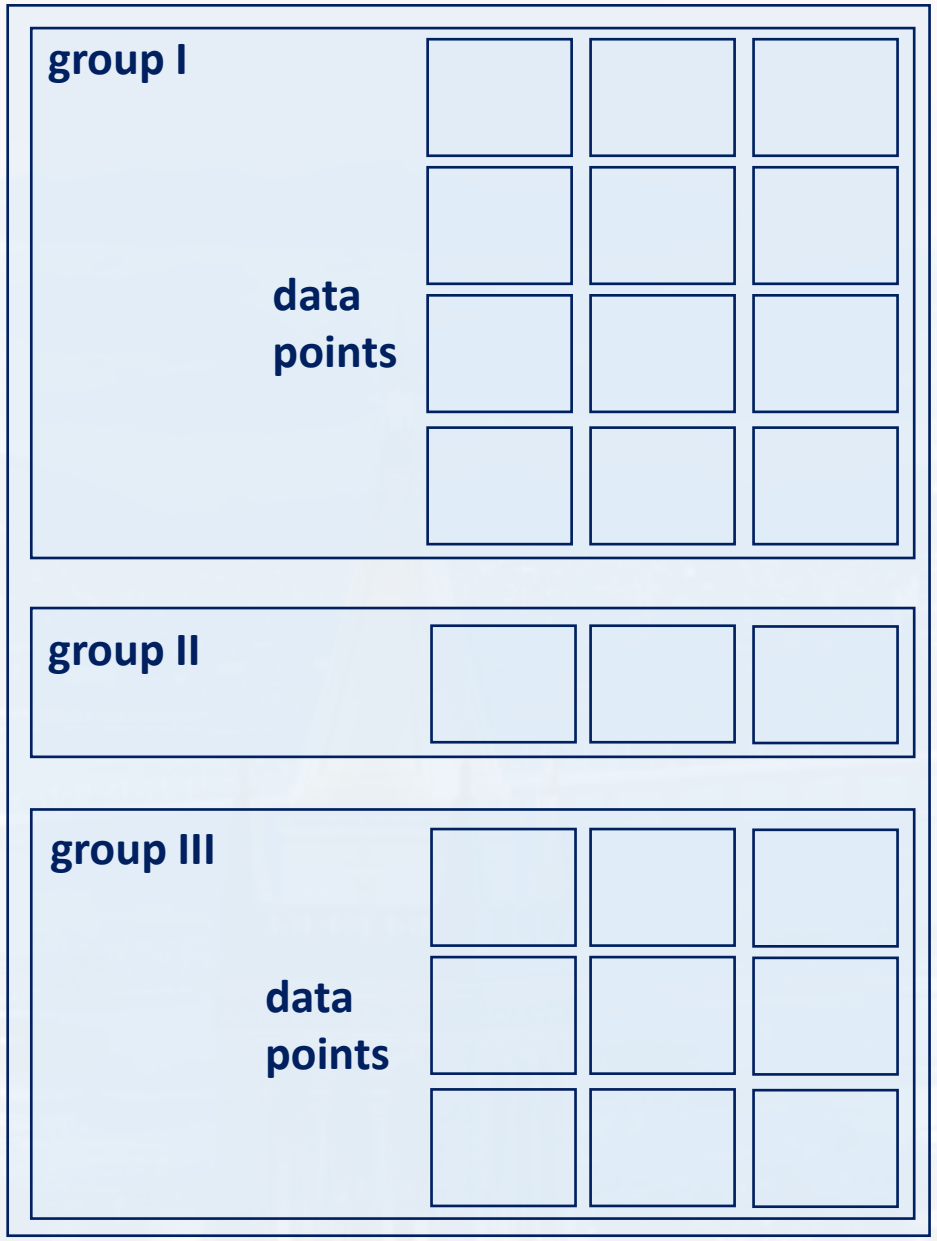
Probabilities

Entropy and Information

Sampling Methods

- Random Sampling
- Stratified Sampling
- Systematic Sampling
- Cluster Sampling
- Bootstrap Sampling
- Oversampling & Undersampling

EDA I and Sampling Methods:



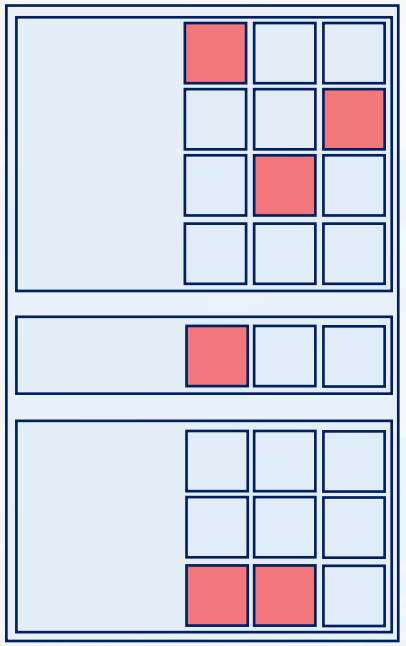
- groups can differ in sizes
- data is usually sign. different between groups

Random Sampling
Stratified Sampling
Systematic Sampling
Cluster Sampling
Bootstrap Sampling
Oversampling & Undersampling

	TARGET_LEVEL	HIT_CALL	BMD
group I	gene	0.988987	46.472969
	gene	0.962960	42.798135
	gene	0.917757	41.108881
group II	cell	0.960401	78.170916
	cell	0.957572	89.467940

see: `SamplingMethods.ipynb`

EDA I and Sampling Methods:

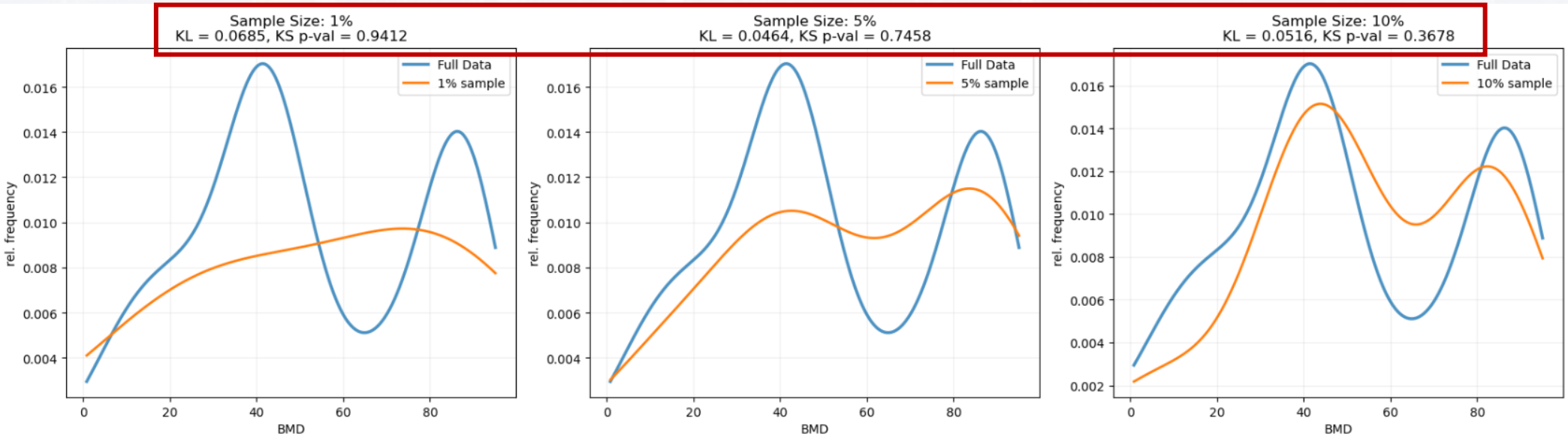


how: - sampling regardless of groups and their proportions

pros:
- simple
- unbiased
- automatically maintains *approx.* proportion of groups

cons:
- needs large sample size, especially if many groups with different proportions
- list of groups need to be complete

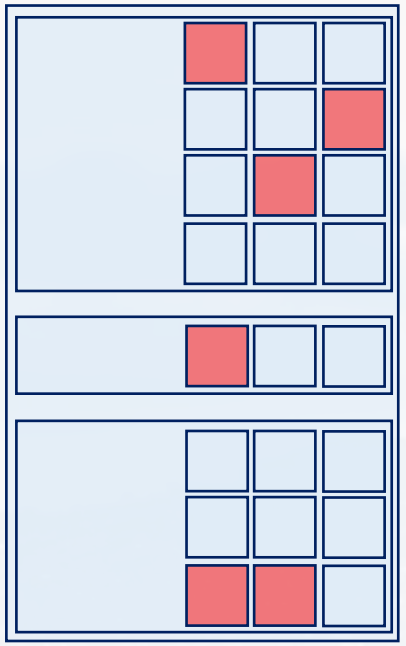
Random Sampling
Stratified Sampling
Systematic Sampling
Cluster Sampling
Bootstrap Sampling
Oversampling & Undersampling



N = 420

see: `SamplingMethods.ipynb`

EDA I and Sampling Methods:

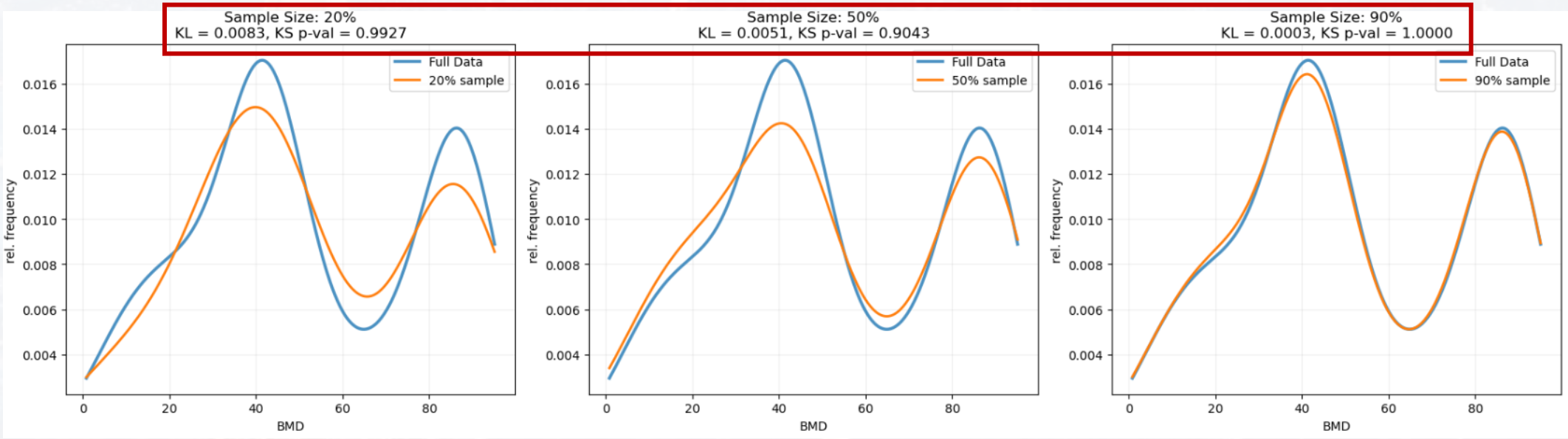


how: - sampling regardless of groups and their proportions

pros:
- simple
- unbiased
- automatically maintains *approx.* proportion of groups

cons:
- needs large sample size, especially if many groups with different proportions
- list of groups need to be complete

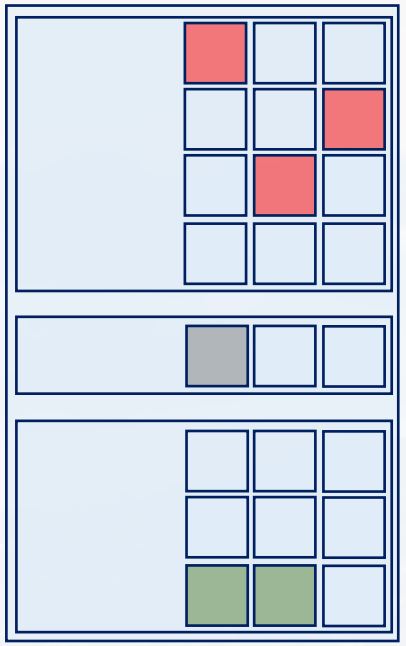
Random Sampling
Stratified Sampling
Systematic Sampling
Cluster Sampling
Bootstrap Sampling
Oversampling & Undersampling



N = 420

see: `SamplingMethods.ipynb`

EDA I and Sampling Methods:



how: - sampling within groups (aka **strata**)
- population is exhaustively partitioned into **disjoint subgroups**

pros: - maintains *exact.* proportion of groups
- useful when groups are expected to follow different distributions

cons: - don't apply if the population cannot be exhaustively partitioned into disjoint subgroup

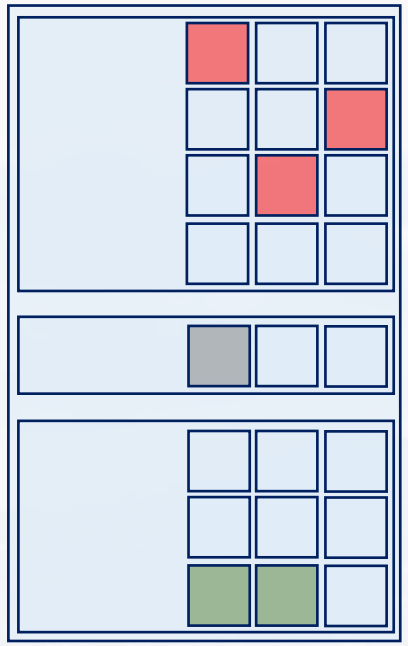
TARGET_LEVEL	HIT_CALL	BMD
gene	0.988987	46.472969
gene	0.962960	42.798135
gene	0.917757	41.108881
cell	0.960401	78.170916
cell	0.957572	89.467940

```
L = list(File[col_groups])
for g in Groups:
    ct = L.count(g)
    print(g + ": " + str(ct) + " appearances" )
```

```
organ: 102 appearances
none: 3 appearances
tissue: 18 appearances
gene: 202 appearances
cell: 61 appearances
chemical: 14 appearances
random: 20 appearances
```

see: `SamplingMethods.ipynb`

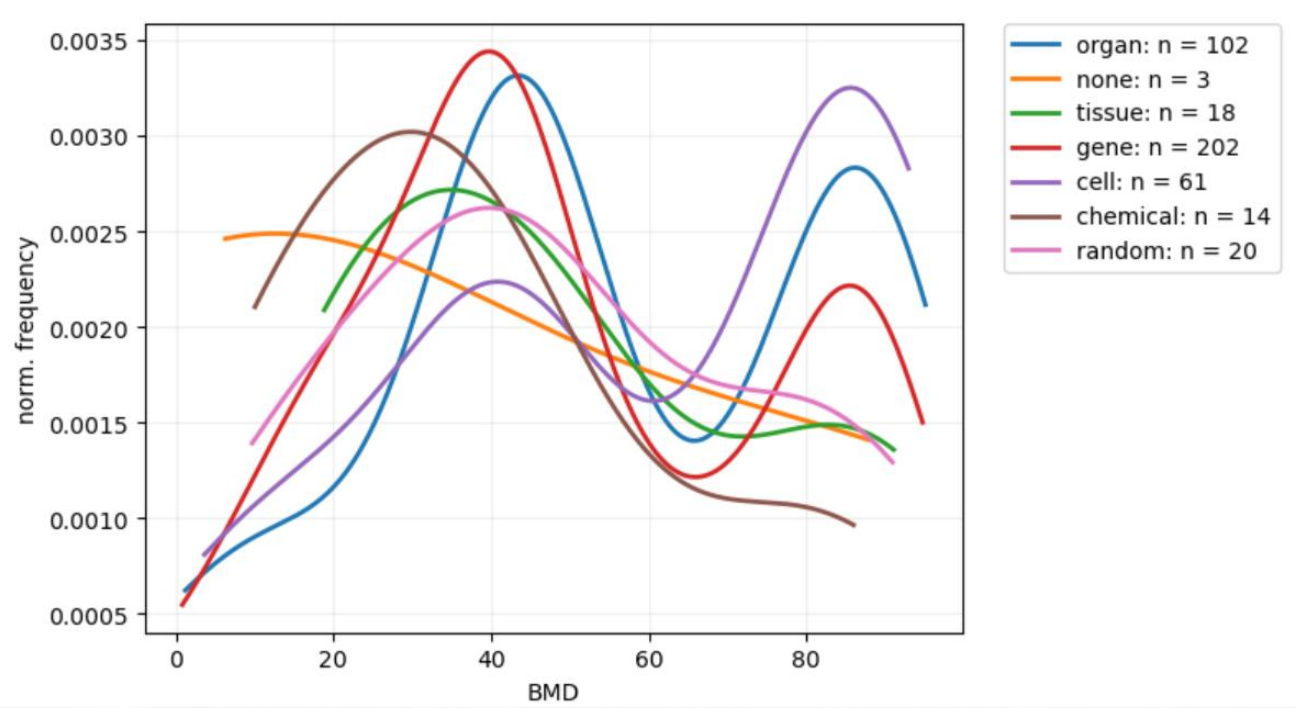
EDA I and Sampling Methods:



how: - sampling within groups (aka **strata**)
- population is exhaustively partitioned into **disjoint subgroups**

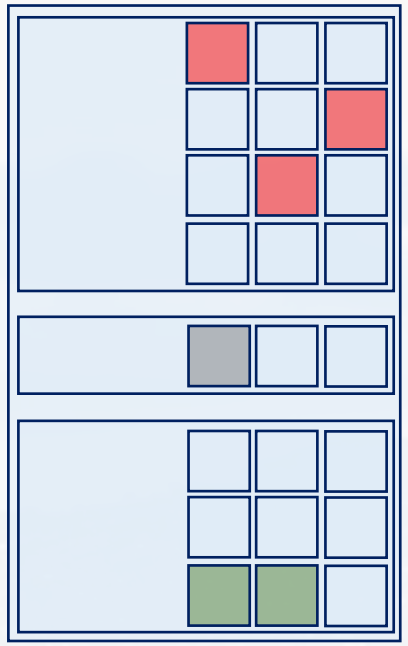
pros: - maintains *exact.* proportion of groups
- useful when groups are expected to follow different distributions

cons: - don't apply if the population cannot be exhaustively partitioned into disjoint subgroup



see: `SamplingMethods.ipynb`

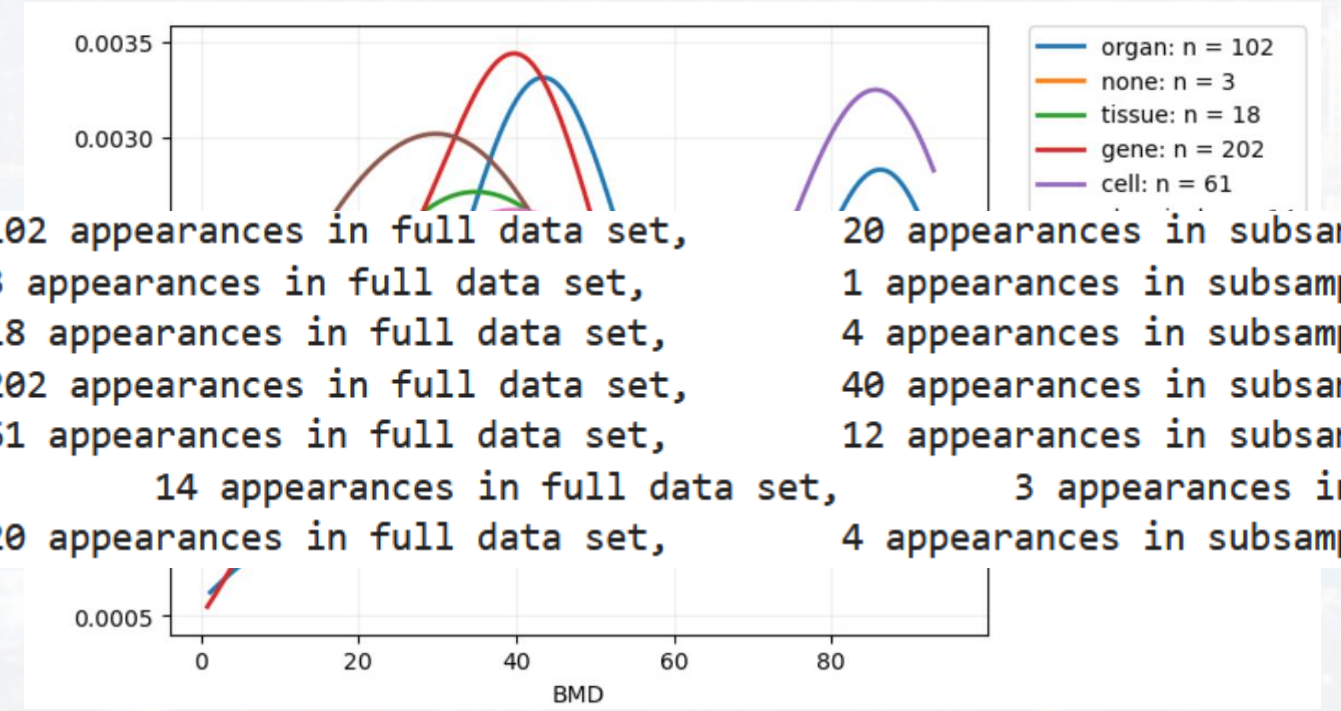
EDA I and Sampling Methods:



how: - sampling within groups (aka **strata**)
- population is exhaustively partitioned into **disjoint subgroups**

pros: - maintains *exact.* proportion of groups
- useful when groups are expected to follow different distributions

cons: - don't apply if the population cannot be exhaustively partitioned into disjoint subgroup



drawing 20% from each group

```
organ:    102 appearances in full data set,      20 appearances in subsample.    Ratio =  0.20
none:     3 appearances in full data set,        1 appearances in subsample.     Ratio =  0.33
tissue:   18 appearances in full data set,       4 appearances in subsample.     Ratio =  0.22
gene:     202 appearances in full data set,      40 appearances in subsample.    Ratio =  0.20
cell:     61 appearances in full data set,       12 appearances in subsample.    Ratio =  0.20
chemical:       14 appearances in full data set,       3 appearances in subsample.    Ratio =  0.21
random:   20 appearances in full data set,       4 appearances in subsample.     Ratio =  0.20
```

see: `SamplingMethods.ipynb`

how: - sampling each k-th element

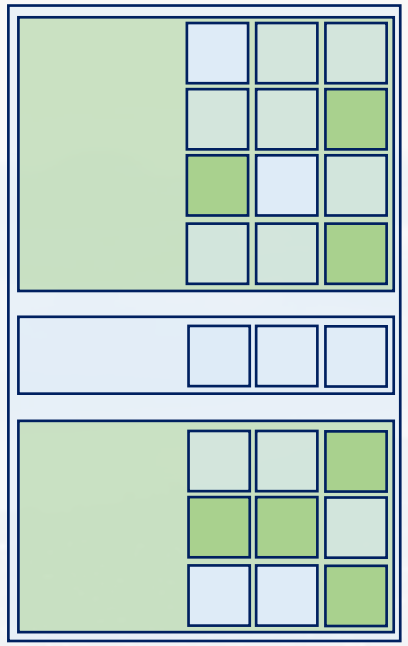
pros: - simple
- samples evenly over population

cons: - imposes bias if there are hidden pattern in the data

0	46.472969
20	88.885569
40	63.151552
60	26.627395
80	35.520452
100	19.830881
120	38.801088
140	43.063503
160	89.690644
180	85.789285
200	89.592299
220	86.492331
240	41.470968
260	13.242978
280	85.977758
300	62.225787
320	85.612638
340	16.862983
360	85.480457
380	27.196121
400	22.218282

see: `SamplingMethods.ipynb`

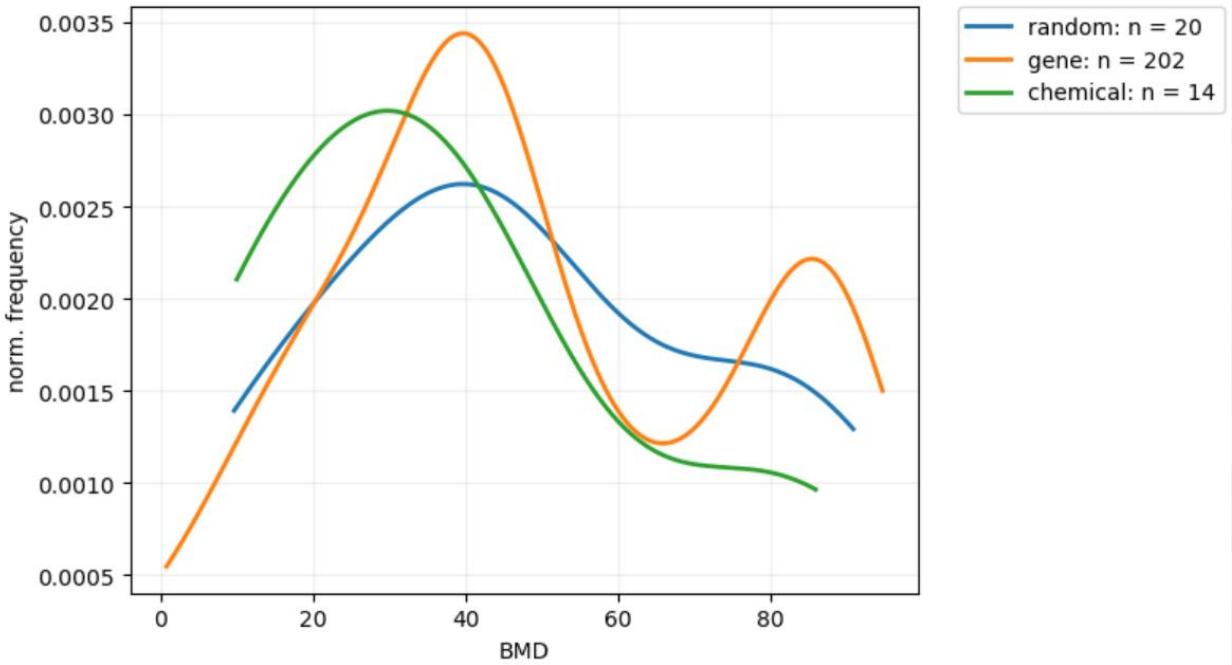
EDA I and Sampling Methods:



how: - randomly select the groups (aka cluster)
- sample from those cluster

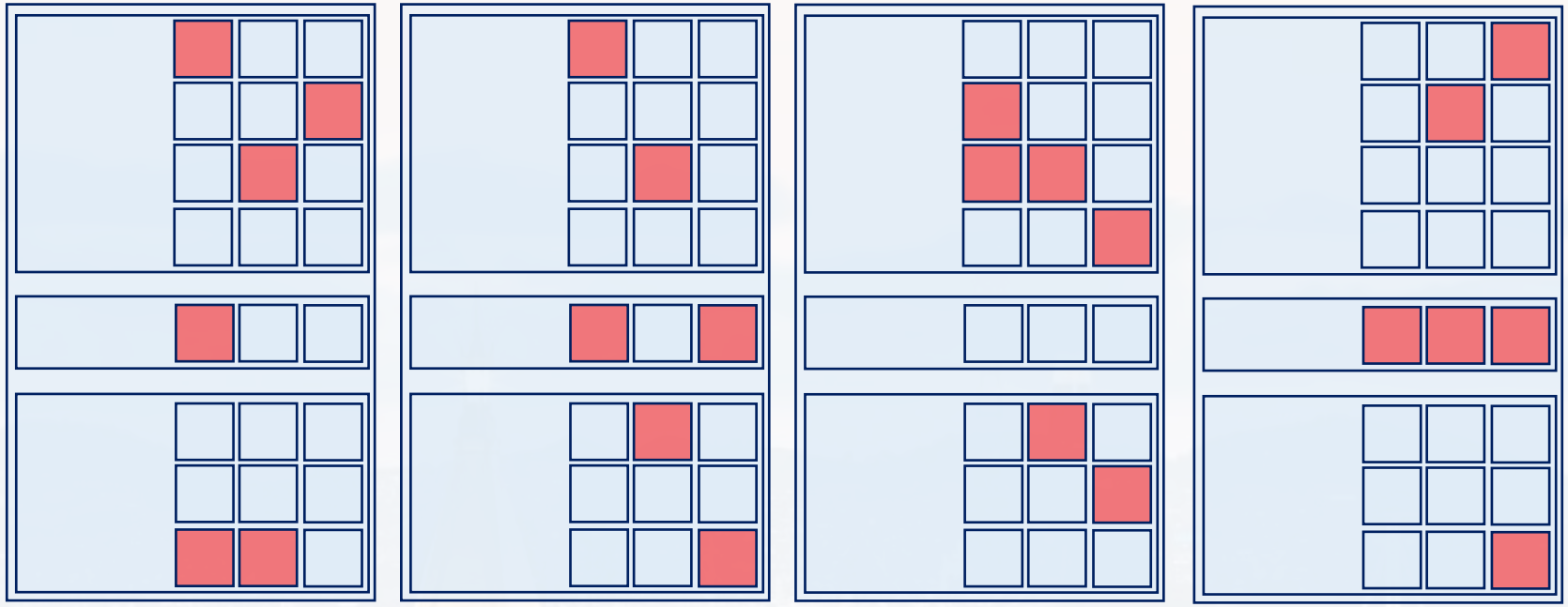
pros: - efficient **for large data sets** (eg. polls)
- if cluster not expected to differ a lot

cons: - prone to bias
- prone to sample errors
- only use if no other choice!



- randomly select three groups
- draw 20% from each group

see: `SamplingMethods.ipynb`



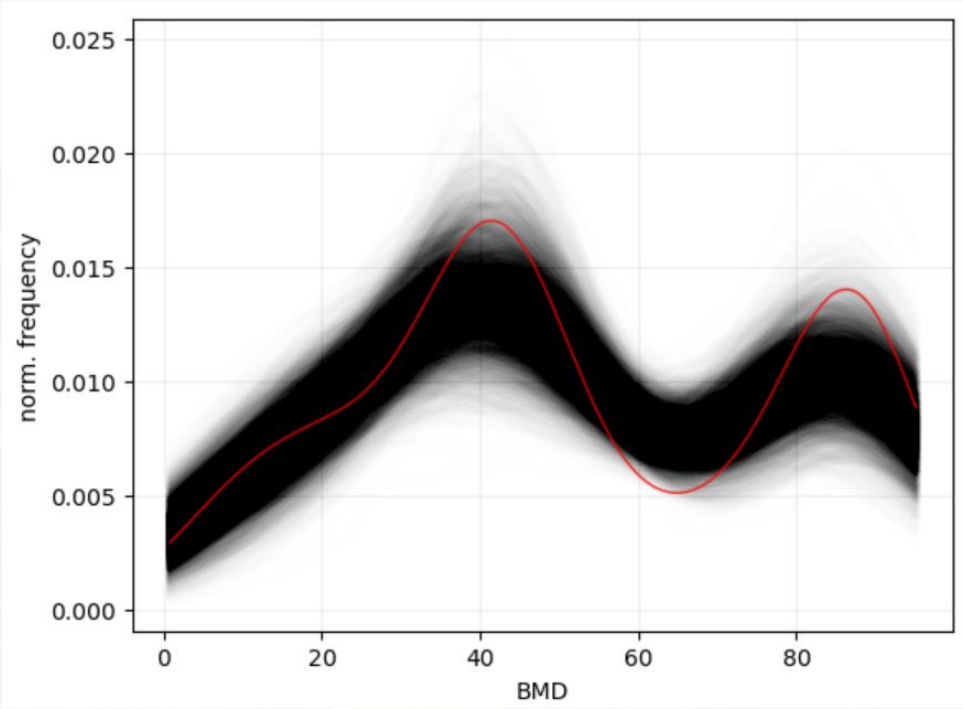
how: - sampling with replacement, N_{boot} times

pros: - simple
- for estimating sample error empirically → percentiles

cons: - $N_{boot} \gg n$

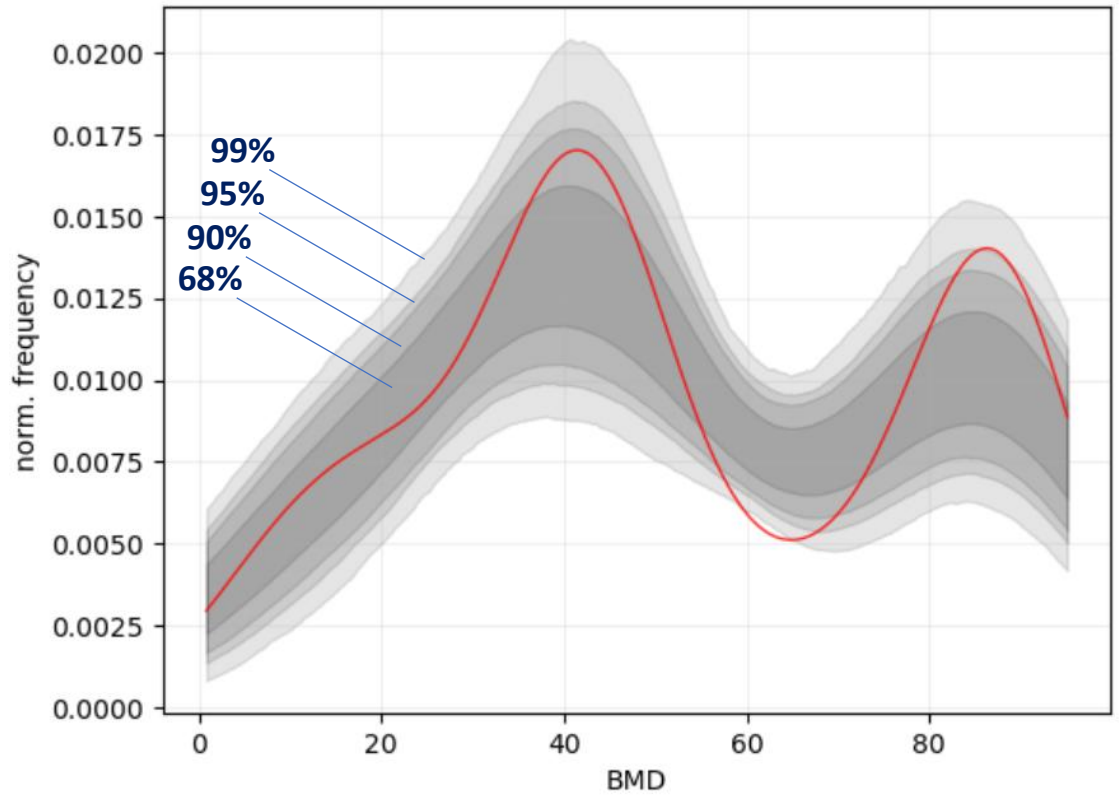
see: `SamplingMethods.ipynb`

EDA I and Sampling Methods:



- $n = 50$
- $N_{boot} = 5000$

Random Sampling
Stratified Sampling
Systematic Sampling
Cluster Sampling
Bootstrap Sampling
Oversampling & Undersampling



see: `SamplingMethods.ipynb`

EDA I and Sampling Methods:

how: - sampling according to group imbalance

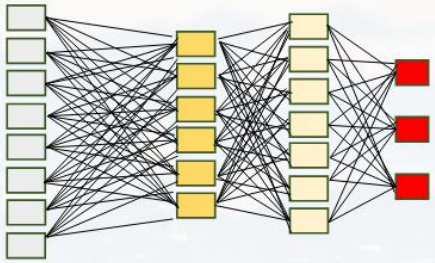
pros: - compensates sample imbalance

cons: - repetitions

9,900 dog images




100 cat images



99% accuracy

see: `SamplingMethods.ipynb`

EDA I and Sampling Methods:

how: - sampling according to group imbalance

pros: - compensates sample imbalance

cons: - **repetitions**

9,900 dog images




100 cat images

	x	y
178	-1.680904	cat
113	-0.442641	cat
37	0.606229	cat
178	-1.680904	cat
160	-1.026721	cat
...	...	...
66	0.415702	cat
66	0.415702	cat
83	2.124732	cat
101	-0.686784	cat
83	2.124732	cat

see: `SamplingMethods.ipynb`

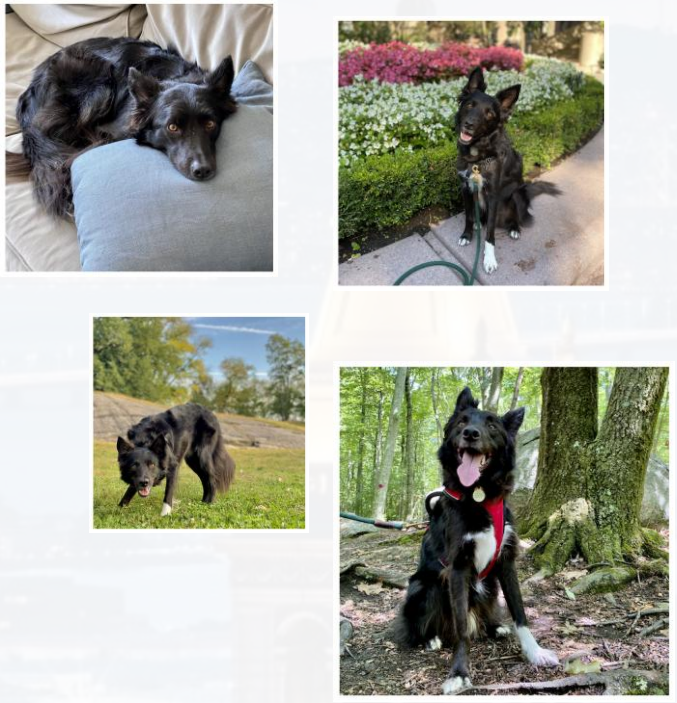
EDA I and Sampling Methods:

how: - sampling according to group imbalance

pros: - compensates sample imbalance

cons: - **repetitions**

9,900 dog images



100 cat images

data augmentation for reducing repetitions!

see: `SamplingMethods.ipynb`

EDA I and Sampling Methods

Thank you very much for your attention!