Introduction to Probability with MATLAB
Spring 2014

Lecture 2 / 12

Jukka Kohonen
Department of Mathematics and Statistics
University of Helsinki

Probabilities in an equiprobable space

To get *P(A)*, there are two things to do:

1. What are the **elementary events**? How many are they?
   - can you perhaps **list** them all? Or,
   - can you **imagine** a process that lists them all
     (for example, as in the **rule of product** we saw last time)?
2. **Which** elementary events **belong to A**?
   (We may call those the ”**favorable** elementary events”)
   How many are they?

Once you know these (and assume equiprobability), then

$$P(A) = \frac{n(A)}{n(\Omega)}$$

Example. Two dice, P(both are even)

- **Elementary events** = ordered pairs out of the set {1,...,6}. There are **36** of them (rule of product): {(1,1),(1,2),...,(6,5),(6,6)}
- **Favorable elementary events:** first die must be one of {2,4,6}, likewise the second, so the favorable outcomes are {(2,2),(2,4),(2,6),...,(6,6)}
  - **Rule of product**, there are $3 \cdot 3 = 9$ of them

- Probability = **9 / 36** = 1/4

Example. Two dice, P(X+Y=6)

- **36** elementary events
- **Favorable** are:
  {(1,5),(2,4),(3,3),(4,2),(5,1)}
  that is **5** outcomes
- Probability = **5** / **36**

Let's experiment...

```
>> n=1e6;
>> x=dice(n);
>> y=dice(n);

>> sum(x+y==6) / n
ans =
    0.1385

>> 5/36
ans =
    0.1389
```

Seems close!

Subsets of given size ”combinations”

4 persons ABCD shake hands. How many handshakes occur? (subsets of 2 persons)

If we list all ordered pairs, $4 \cdot 3 = 12$

AB, AC, AD,

BC, BA, BD,

CA, CB, CD,

DA, DB, DC

The red ones are the same set
{A,B} = {B,A}

The blue ones are the same set
{A,C} = {C,A}

And so on. Every subset has been listed twice, so the number of subsets is 12 / 2 = 6

Number of combinations

In a set of *n* elements, the **number of *k*-element subsets** (*k*-combinations) can be computed thus:

- Count all ordered *k*-sequences: $(n)_k$
- Each *k*-combination corresponds to *k*! different ordered sequences, thus the number of *k*-combinations is

$$\frac{(n)_k}{k!} = \frac{n(n-1)\dots(n-k+1)}{k!} = \frac{n!}{(n-k)!k!} = \binom{n}{k}$$

- Also known as the binomial coefficient, ”*n* choose *k*”.
- MATLAB: **`nchoosek(n, k)`**

Listing all cases (Matlab)

Ordered sequences

>> **perms**('ABCD')

```
ans =
DCBA
DCAB
DBCA
DBAC
DABC
DACB
CDBA
CDAB
CBDA
CBAD
CABD
CADB
BCDA
BCAD
BDCA
BDAC
BADC
BACD
ACBD
ACDB
ABCD
ABDC
ADBC
ADCB
```

Combinations

>> **nchoosek**('ABCD', 2)

```
ans =
AB
AC
AD
BC
BD
CD
```

>> **nchoosek**(4, 2)

```
ans =
    6
```

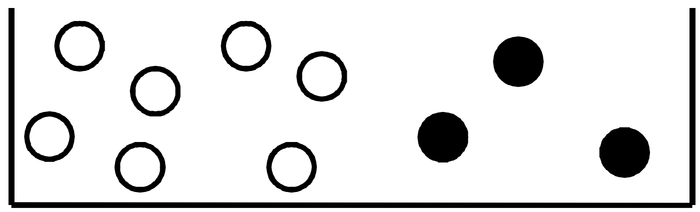
P(Four of a kind) in poker

see e.g. Wikipedia: Poker probability

- 52 cards (4 suits, 13 cards each)
- A hand is a 5-element subset (combination), so there are nchoosek(52, 5) different hands
- Counting the favorable hands (via rule of product):
  - The four cards must have the same value. This value is one of **13** possible values. The hand then contains all four cards of that value (no choice here).
  - The fifth card can be any one of the remaining **48** cards.
  - Rule of product: **13 · 48 = 624** favorable hands
- Probability is 624 / nchoosek(52, 5) ≈ 0.000240

How to try that in Matlab?

- We could list **all** nchoosek(52, 5) = about **2.6 million** hands of five cards,
  - **Pick out** those that are ”four of a kind” (”favorable”)
  - **Count** them one by one, and then compute #A / #Ω
- Or, we could generate **some random hands** of five cards, check **how often** we got ”four of a kind”, and compute relative frequency.
- How do these approaches differ?

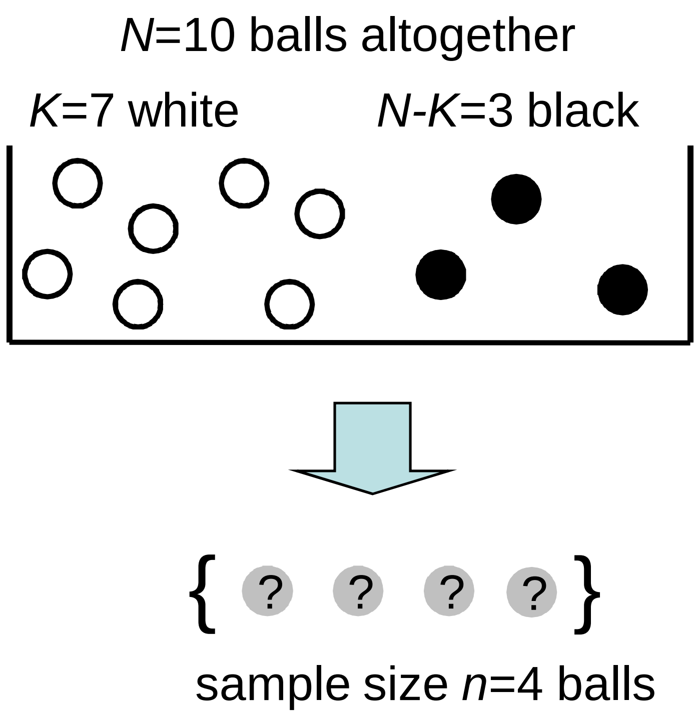


SAMPLING WITH OR WITHOUT REPLACEMENT

Without replacement

- A population of *N* elements (balls, people etc.), of two kinds (”white”, ”black”)
- A **sample** = subset of size *n* is chosen at random (equiprobably)

- Define event $\boldsymbol{A_k}$ **= ”there are *k* white balls in the sample”**
- It contains many elementary events (<u>which</u> white balls, <u>which</u> black balls are in the sample)



Counting favorable outcomes

k white balls chosen from *K* — *n*-*k* black balls chosen from *N*-*K*

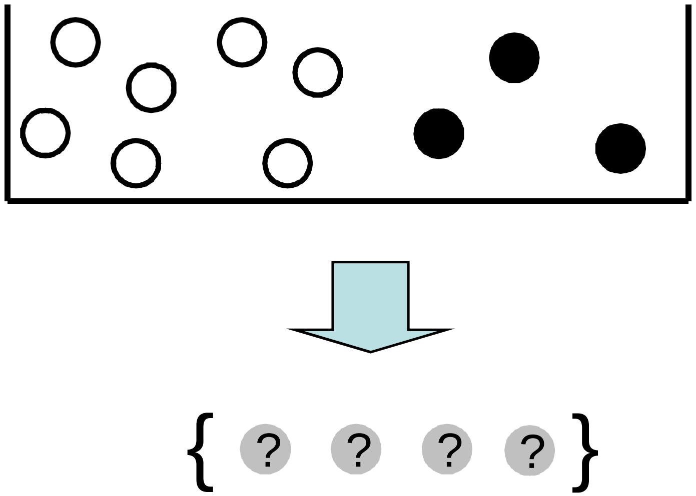
$$P(A_k) = \frac{\binom{K}{k} \times \binom{N-K}{n-k}}{\binom{N}{n}}$$

Counting all outcomes (*n*-subsets of *N*)

Without replacement (example)

N=10 balls altogether

K=7 white *N*-*K*=3 black



sample size *n*=4 balls

$$P(A_2) = \frac{\binom{7}{2} \times \binom{3}{2}}{\binom{10}{4}} = \frac{21 \times 3}{210} = 0.3$$

Similarly

P(”0 white”) = 0 (why?)

P(”1 white”) ≈ 0.033

P(”2 white”) = 0.300

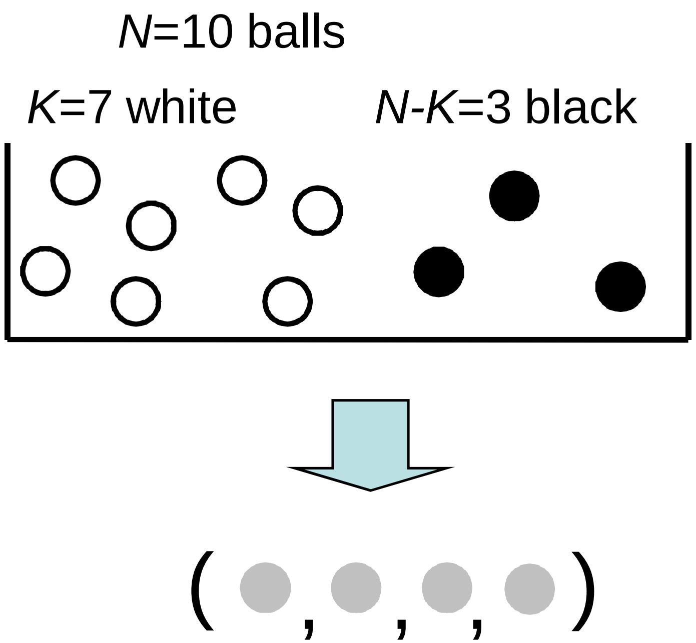
P(”3 white”) = 0.500

P(”4 white”) ≈ 0.167

With replacement

- Same population. A ball is picked at random, **placed back into the population**, again a ball is picked at random and so on.
- Not necessarily physical ”replacing” (placing back into the population. The point is that **each time, a ball is picked from the same population** (e.g. pick people from phone directory)

- Thus the same ball can appear again! A subset is not a good model for the sampling.
- Instead the sample is an ordered *n*-sequence <u>where same element can occur again</u> (just like in die-tossing)



Favorable outcomes

Location of the k whites within sample of n

k white out of *K*

n-k black out of *N-K*

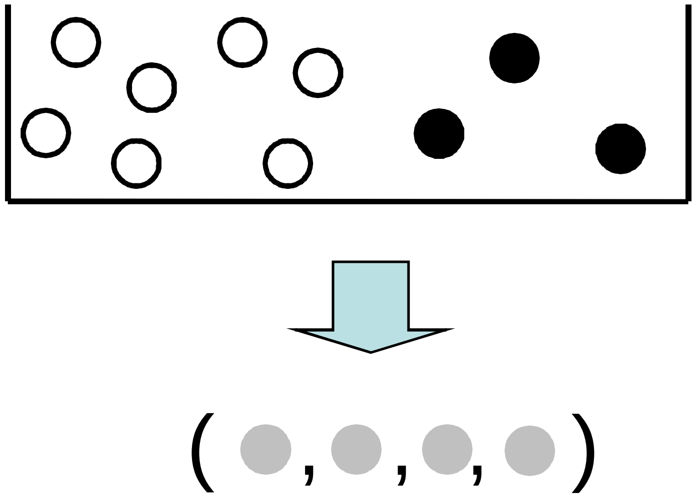
$$P(A_k) = \binom{n}{k} \frac{K^k (N-K)^{n-k}}{N^n}$$

Elementary events = ordered sequences of *n* elements out of *N*

With replacement (example)

N=10 balls

K=7 white N-K=3 black



Favorable outcomes

Location of the 2 whites within sample of 4

2 white out of 7

2 black out of 3

$$P(A_2) = \binom{4}{2} \frac{7^2 \times 3^2}{10^4} \approx 0{,}265$$

Elementary events = ordered sequences of 4 elements out of 10

Comparison: without *vs.* with

7 white and 3 black, sample of n=4

k	P(”k white”) **without** replacement	P(”k white”) **with** replacement
0	0,000	0,008 (why > 0 ?)
1	0,033	0,076
2	0,300	0,265
3	0,500	0,412
4	0,167	0,240

Example: A very large population

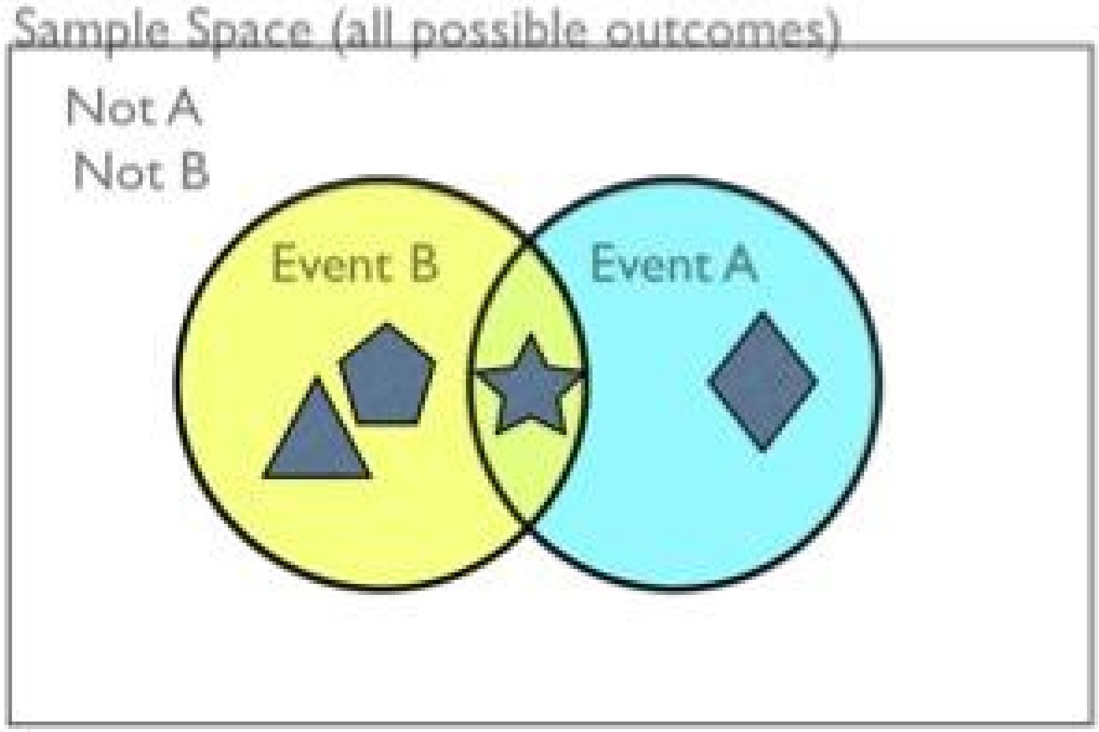
- N = 5 000 000 (people living in Finland)
- K = 500 000 (people living in Helsinki)
- n = 3 (sample) $n << K$ ja $n << N-K$

- P(sample contains exactly one person from Helsinki)?
- Without replacement

$$P(A_1) = \frac{\binom{500\,000}{1} \times \binom{4\,500\,000}{2}}{\binom{5\,000\,000}{3}} \approx 0{,}243\,000\,094$$

- With replacement

$$P(A_1) = \binom{3}{1} \frac{500\,000^1 \times 4\,500\,000^2}{5\,000\,000^3} = 0{,}243\,000\,000$$



NON-EQUIPROBABLE PROBABILITY SPACE

Finite sample space (cf. G&S book Chapter 1)

- As before, Ω = sample space = set of all possible alternatives (elementary events, outcomes): **exactly one of them will be true.**
- **Outcomes ω_i** (where i=1,...,n) need not be equally probable, each one has **some probability $P(\{\omega_i\}) = p_i$**
- We still maintain **additivity of probability** (G&S p. 19 and 22). The probability of an event is **defined to be the sum** of the elementary event probabilities
- Unlike the equiprobable space, we no longer care much about ”counts” of elementary events. Instead we must add up their elementary probabilties.

- Eg. pin tossing (2 outcomes, eg. P(down)=0.7, P(up)=0.3)
- Eg. gender of child (2 outcomes, eg. P(boy)=0.51)
- Eg. loaded die (several elementary events with arbitrary probabilities)

Useful properties (rules)

- Even if the outcomes are not equiprobable, probability has properties similar to what we saw in Exercise set 1.
  See G&S book p. 22.

- Additivity (for disjoint sets).
- Monotonicity
- Boundedness: $0 \leq P(A) \leq 1$
- Complement rule: $P(A^c) = 1 - P(A)$
- Why? These follow from how probability was defined, and from the properties of addition

BERNOULLI TRIAL

Tossing a coin n times

- For each toss we have
  P(heads) = ½
  P(tails) = ½
- What is the probability of getting exactly k times the result ”heads”? Surely $0 \le k \le n$.
- We can take elementary events to be the **sequences** of n results (1=heads, 0=tails)
- We can even assume them **equiprobable**
- There are 2^n such sequences, so...

Tossing a <u>pin</u> *n* times

- Suppose
  P(up) = p
  P(down) = $q = 1 - p$
- What is the probability of getting exactly k times the result ”up” ?
- We can still take elementary events to be the **sequences** of n results (1=up, 0=down)
- But we can **no longer assume them equiprobable**

Tossing a pin *n* times...

- What is the probability of getting, for example, the ordered sequence

  (up, up, down) ?

- Appealing to a frequency interpretation of probability, this seems to be

  $$p \cdot p \cdot q$$

- We will accept this for a while. (We shall later learn a formal concept of ”independence”.)

Tossing a pin...

- But the event ”2 times up” allows the following possibilities:
  (up, up, down)
  (up, down, up)
  (down, up, up)
- Each of those has the same probability *ppq* (why?), so by additivity
  P(2 times up) = 3*ppq*

Binomial probability

- More generally: if we have
  - n tosses of the pin, and each time
  - probability p of pin landing ”up”, and $q = 1-p$ of landing ”down”
- Then the probability of exactly k ”up” results is nchoosek(n, k) $\cdot p^k \cdot q^{n-k}$

- (see e.g. G&S p. 96—98)