5. THE BASICS OF EXPERIMENTAL DESIGN

5.1. *To plan experiments*

- Exps are ***expensive*** [1].
- Many biol. exps are ***comparative*** [2].
- **ED** is aiming to gather as ***much*** relevant info. as possible, at the ***lowest*** possible cost [3].
- Usually one method may give the same ***degree*** of evidence as the other, but at a ***lower*** cost [4].
- An ethical aspect: efficient exps can ***reduce*** the ***need*** for animals in animal exps.

[1] each measurement carries with it a cost, in ***time*** or in ***money***;

[2] you want to ***compare*** different ***treatments***, e.g. varieties, feeding schemes etc.;

[3] ED can be defined as ***observations*** under ***controlled*** conditions;

[4] if there is a choice between two ***methods*** of designing, you would prefer the ***cheaper*** alternative.

What factors affect the choice **ED**? [1]

- many believe: it is the ***number*** of ***observations***.
- other factors: e.g. ***blocking*** [2].

https://thepigsite.com/

- How to make an inefficient exp.?:
- You can spend ***large*** resources on measurements where the variation is ***small*** [3].
- So, ***observations*** have a different "***value***", depending on the **ED**.
- **ED** therefore influences the analysis of data.

[1] the ***precision*** of an exp.;

[2] i.e. we compare the treatments within ***blocks*** of units that are as ***similar*** as ***possible***, e.g. different feeding schemes for pigs may be compared within groups of pigs that are litter mates: the influence of biological variation can be controlled, and we get a more efficient exp.;

[3] This is because ***errors*** in chemical analyses are often ***small***, while there may be ***large*** ***variation*** between soil samples.

[3] e.g., it is often ***less*** ***efficient*** to take 2 soil samples and to make 10 chemical analyses on each sample, than to take 10 soil samples and make 2 analyses on each.

5.2 *Terms and definitions*


https://www.feednavigator.com/Article/2017/09/01/Is-tree-bark-the-remedy-for-diarrhea-in-calves

- In exp. you expose a number of ***objects*** to different ***treatments*** (***factors***) [1].
- ***Factors*** can be denoted with capital letters like ***A***.
- The different ***values*** or ***classifications*** of a factor are called ***levels***.
- The number of ***levels*** of a factor **A** will be denoted **a** [2].
- An *exp.* ***unit*** is the ***smallest*** unit that is given an individual treatment [3].

[1] e.g. different fertilizers, temperatures etc.;

[2] **Example 1**. Three feeding schemes for calves: the factor **A**=feed, with **a=3** levels, $\mathbf{A}_1$=high protein level; $\mathbf{A}_2$=normal protein level, and $\mathbf{A}_3$=low protein level.
Example 2. You want to compare four ecol. fertilizers on tomatoes: the factor **A**=fertilizer has **a**=4 levels: $\mathbf{A}_1$=Compost; $\mathbf{A}_2$=Sawdust;$\mathbf{A}_3$=Grass; and $\mathbf{A}_4$=Control;

[3] from which data will be collected.
Example 3: The calves are fed individually: the ***calf*** is the exp. unit; all calves in a box are fed together: the ***box*** is the exp. unit.
Example 4: Each tomato plant is fertilized individually: the plant is the exp. unit;.you fertilize a box with e.g. 10 plant: the box is the exp. unit.

- Exp. can simultaneously contain ***many*** factors (*factorial* *exps.*) [1].

- A factor may be *fixed* (**FF**) or *random* (**RF**) [2]:

- **FF** if you selected the *levels* of the factor, you are only *interested* [3].

- **RF** if the *levels* are (or can be regarded as) a *random* sample from a population of levels [4].

[1] e.g. "gender" might be a factor with two levels "male" and "female" and "diet" might be a factor with three levels "low", "medium" and "high" protein.

[2] depending on how the ***levels*** of the ***factor*** were selected;

[3] e.g. you are only ***interested*** in ***three*** levels of feeding: A_1; A_2; A_3.

[4] e.g. some agricultural exps are undertaken on farms, so the farm may be regard as a **RF**, representing a large population of farms. Typical **RF** are farms, herds and stands (a group of growing plants of a specified kind, especially trees).

5.3. *Experimental design: replication*

- E.g. how to study which of *two varieties* gives the highest yield?
- You plant variety **A** on one field, and variety **B** on another field.
- *Result*: variety **A** gives the highest *yield*.
- Can you generalize the result? [1]
- Probably not: one field alone is *not enough* to draw conclusions.
- You need *more data* to give a *general* recommendation.

1

A

B

2

[1] i.e could you safely advice to use variety A?

Population Variance	Sample Variance
$\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	$s^2 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size

- You need more *observations* (*replication*).
- Replication makes it possible to *estimate* the exp. *error* [1]:
- If you *increase* the *number* of *replicates* you get a more precise result [2].
- What are ”real” replicates?
- Two *fields* divided in two *plots*: variety **A** is planted on *both plots* on one field, variety **B** on *both plots* on another field.
- Do we now have replicates? [3]

[1] This can be done by calculating the ***variance*** among all ***units*** that have the same ***treatment***. It (σ^2)measures how far each number in the set is from the mean: it is calculated by taking the differences between each number in the set and the mean, squaring the differences and dividing the sum of the squares by the number of values in the set. The **SD** (σ) = square root of ***variance***.

[2] because the variance of a sample mean is $\boldsymbol{\sigma^2/n}$, so $\boldsymbol{\sigma^2}$ become smaller when **N** increases.

[3] yes.

- E.g., we want to estimate an ***average*** **OM** content in soil.

- How many “real” replicates do we have?:

- ***one*** soil sample from ***one*** location, 200 analyses [1];

- ***one*** soil sample from each of 10 locations, 20 analyses of each sample [2];

- select 10 areas, 10 samples from each area, 2 analyses of each sample [3];

- select 10 areas, 10 samples from each area (mix them in one), 2 analyses of each mixed sample [4].

[1] No; [2] No; [3] Yes; [4] No;

5.4. *Experimental design: local control*

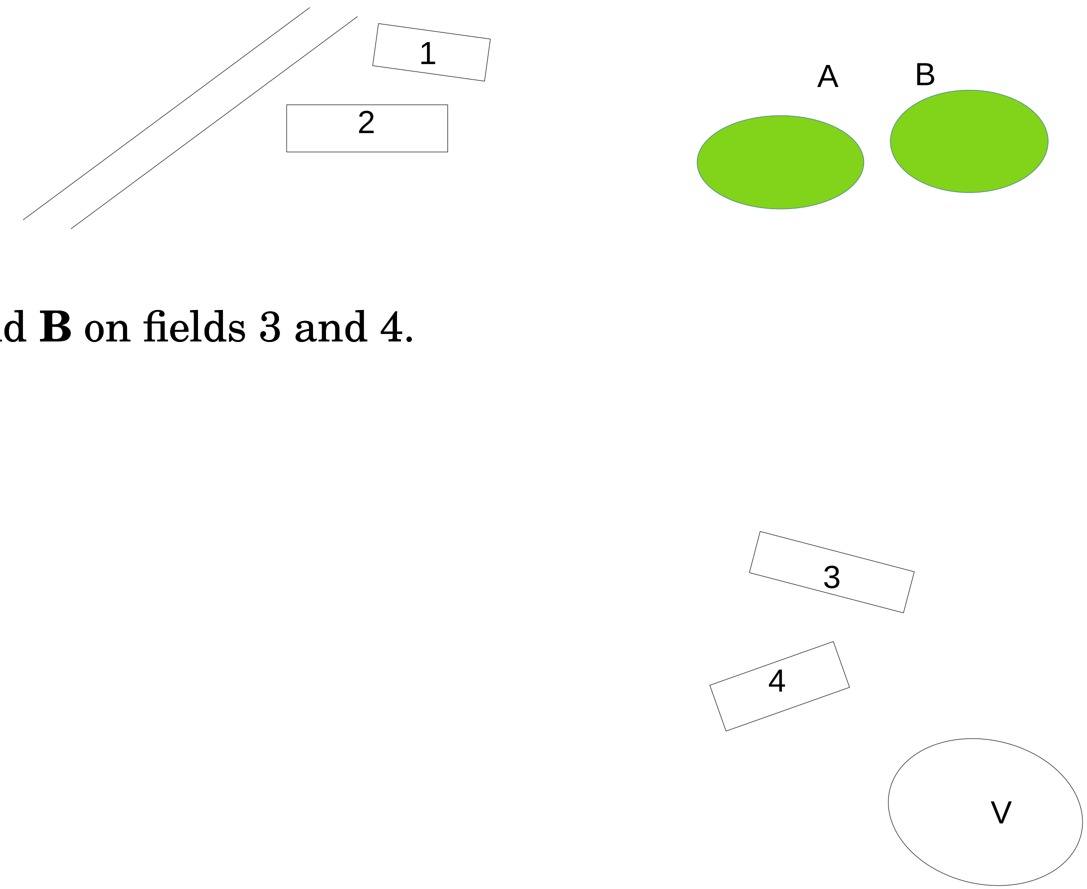
- E.g. two varieties **A** and **B** are ***compared***.

- There are ***four fields***: 1, 2 are located together close to the road, and 3, 4 together some distance from the village.

- ***Where*** to plant ***variety*** **A** and **B**?

- One ***approach***: **A** on fields 1 and 2, and **B** on fields 3 and 4.

- Is this approach ***good*** enough?



- Result: **A** gives the *highest* yield.

- Question: the result occurred because of the *soil*, or because **A** was *better*?

- A better alternative: **A** on one of the fields (1 or 2), and **B** on the other. The same with fields 3 and 4.

A

- The *comparison* between **A** and **B** is less affected by possible differences in soil conditions.

- This would give a fair *comparison* between the two varieties (“local control”).

5.5. *Experimental design: randomization*

- E.g. we decided to plant **A** on *<u>either</u>* field 1 or 2.
- How to *<u>decide</u>* which field to chose?[1]
- Assume you just selected *<u>one</u>* of them.
- Is this *<u>random</u>* selection? No.
- Why? We all have *<u>biases</u>*: even we are not always *<u>conscious</u>* /ˈkɒnʃəs/ of them.
- The only way to assign **A** and **B** to fields 1 and 2 by *<u>process</u>* that **cannot** be *<u>controlled</u>* by the researcher.

[1] at first glance this does not seems to be important question.

How to do this?

https://www.donnaonthebeach.com

- Toss a coin [1].
- Mix the chits of paper with names and draw them from a bowl.
- Use table of random numbers.
- Numbers can be generated by computer [2].

[1] e.g. if “head” turns up, use **A**. Only if you have two treatments (аверс-реверс);

[2] Microsoft Excel has a function to produce random numbers: =RAND(). E.g. you want random numbers from 1 to 200, you could enter: =INT(200*RAND())+1.

The INT eliminates the digits after the decimal, the 200* creates the range to be covered, and the +1 sets the lowest number in the range.

How to use: [1]

1. E.g. we want to randomly sample 5 units of a population of 200 units, each has been assigned a number from 1 to 200 [2].

2. Without looking, point to a starting spot in the table, e.g. 75635.

3. So, 756 (too large, > 200), we choose the next number in that column: 407 (too large), etc.

4. We arrive at 070 (07015), 038 (03811), 045 (04594), 055 (05542), and 194 (19428).

7. RESULT: Units 38, 45, 55, 70, and 194 will be used for our random sample.

75635

07015

03811

04594

Table of Random Numbers

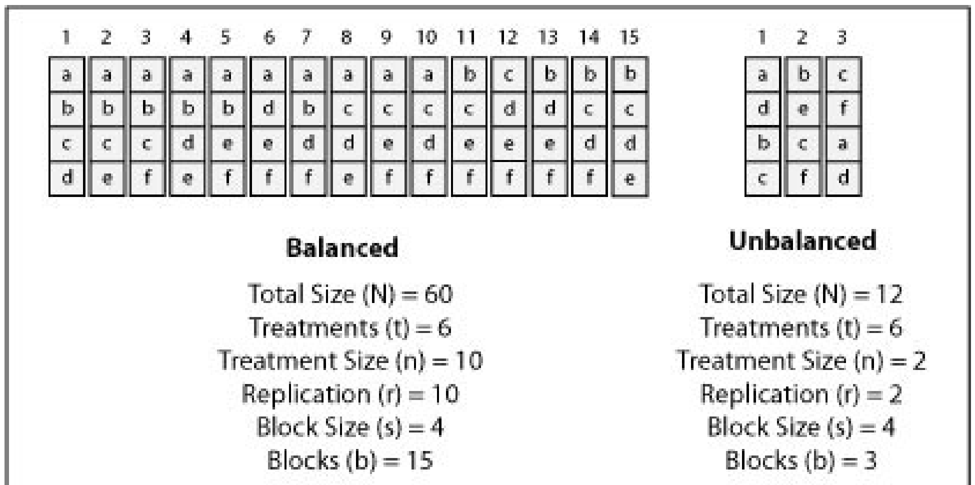
36518	36777	89116	05542	29705	83775	21564	81639	27973	62413	85652	62817	57881
46132	81380	75635	19428	88048	08747	20092	12615	35046	67753	69630	10883	13683
31841	77367	40791	97402	27569	90184	02338	39318	54936	34641	95525	86316	87384
34180	93793	64953	51472	65358	23701	75230	47200	78176	85248	90589	74567	22633
78435	37586	07015	98729	76703	16224	97661	79907	06611	26501	93389	92725	68158
41859	94198	37182	61345	88857	53204	86721	59613	67494	17292	94457	89520	77771
13019	07274	51068	93129	40386	51731	44254	66685	72835	01270	42523	45323	63481
82448	72430	29041	59208	95266	33978	70958	60017	39723	00606	17956	19024	15819
25432	96593	83112	96997	55340	80312	78839	09815	16887	22228	06206	54272	83516
69226	38655	03811	08342	47863	02743	11547	38250	58140	98470	24364	99797	73498
25837	68821	66426	20496	84843	18360	91252	99134	48931	99538	21160	09411	44659
38914	82707	24769	72026	56813	49336	71767	04474	32909	74162	50404	68562	14088
04070	60681	64290	26905	65617	76039	91657	71362	32246	49595	50663	47459	57072
01674	14751	28637	86980	11951	10479	41454	48527	53868	37846	85912	15156	00865
70294	35450	39982	79503	34382	43186	69890	63222	30110	56004	04879	05138	57476
73903	98066	52136	89925	50000	96334	30773	80571	31178	52799	41050	76298	43995
87789	56408	77107	88452	80975	03406	36114	64549	79244	82044	00202	45727	35709
92320	95929	58545	70699	07679	23296	03002	63885	54677	55745	52540	62154	33314
46391	60276	92061	43591	42118	73094	53608	58949	42927	90993	46795	05947	01934
67090	45063	84584	66022	48268	74971	94861	61749	61085	81758	89640	39437	90044
11666	99916	35165	29420	73213	15275	62532	47319	39842	62273	94980	23415	64668
40910	59068	04594	94576	51187	54796	17411	56123	66545	82163	61868	22752	40101
41169	37965	47578	92180	05257	19143	77486	02457	00985	31960	39033	44374	28352
76418												

[1] one of variety of methods;

[2] since the population size is a three-digit number, we will use the first three digits of the numbers listed in the table.

5.6. *Experimental design: balance and completeness*

- ***Balanced*** exp.: there is an *equal number* of observations for each *treatment* combination [1].
- If some treatment combinations are *missing* the exp. is *incomplete*.



Balanced and Unbalanced Incomplete Block Designs: Constrained by Six Treatments and a Block Size of Four

https://articles.extension.org/pages/69751/introduction-to-randomization-and-layout

- **BE** are *easier* to analyze and to interpret.

[1] otherwise it is not. It is not necessarily ***essential*** but it is ***simpler*** than unbalanced and ***desirable*** if an equal accuracy or confidence interval width for treatment comparisons is ***important***.

- An *unbalanced* exp. add *complexity* to the analysis of the data [1].

- An *unbalanced* *incomplete* block design can be used for a smaller experiment [2, 3].

[1] E.g. some levels are expected to produce ***greater*** variation than others and so more units are ***assigned*** to those levels;

[2] not every ***treatment*** pair occurs together equal amounts of time;

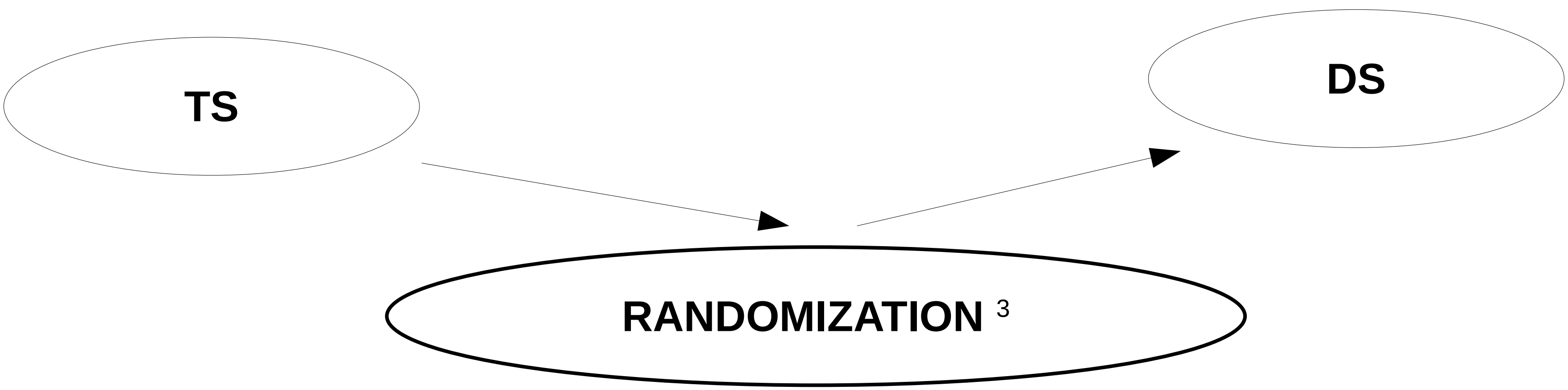
[3] is ***small*** enough to repeat up to four times and still remain smaller than the balanced design.

5.7. *Treatment structure and design structure*

- **TS**: the *set* of *treatments* (treatment combinations) that the experimenter has selected to study and/or compare.

- E.g. if compare *varieties*, the *varieties* are the *treatments* [1].

- **DS**: the *grouping* of the exp. units into homogeneous *groups* or *blocks* [2].



[1] ***treatments*** may be amounts of fertilizer used; types of feeding schemes, etc;

[2] tells you in what way you attempt to achieve ***local control***;

[3] to assign ***treatments*** of TS to experimental ***units*** DS

- Question: why **DS** is important?

- Answer: to ensure a fair *comparison* when the number of exp. units is "*small*"

- Commonly used **DS**:

- completely randomized design (**CRD**);
- randomized complete block design (**RCBD**);
- Latin square design (**LSD**), etc.

5.8. *Completely randomized design*

- **CRD**: the treatments are assigned *completely* at random [1].

- **CRD** is appropriate *only* for exp. with *homogeneous* exp. *units* [2].

- For field exp., (generally *large variation* among exp. plots, e.g. in soil), the **CRD** is *rarely* used.

1. One factor.

- E.g.: three varieties **A**, **B** and **C**; field - may be divided in 12 plots:

B	A	A
A	C	C
C	B	B
C	A	B

[1] so that each ***experimental unit*** has the ***same chance*** of receiving any one treatment: any difference among experimental units receiving the same treatment is considered as experimental error;

[2] e.g. laboratory experiments, where environmental ***effects*** are relatively ***easy*** to control.

1 2 3 4

2. *Two* or more factors.

- E.g.: two feeding schemes, $\mathbf{A_1}$, $\mathbf{A_2}$; and two cows varieties $\mathbf{B_1}$ and $\mathbf{B_2}$; i.e. four treatment *combinations*: $\mathbf{A_1B_1}$, $\mathbf{A_1B_2}$, $\mathbf{A_2B_1}$ and $\mathbf{A_2B_2}$; i.e. four cows of each variety.

- How to design the exp.?

- *One way*: randomly select two *cows* of each variety for $\mathbf{A_1}$, and two for $\mathbf{A_2}$.

- *More* than two *factors*: e.g.: the level of **NPK** as *factors*; each has three sub-levels [1] = 27 *treatment* combinations (3x3x3); *two* replications = 54 plots [2, 3].

[1] e.g. ***low***, ***medium***, ***high***;

[2] e.g. how to keep the number of treatment ***small***: e.g. try to reduce the number of levels = 8 (2x2x2) combinations;

[3] Why include so ***many*** ***factors*** in the same exp.? One reason is that it allows to study ***interactions*** between the factors.

5.9. *Randomized block design*

- **RBD**: subjects divided into ***subgroups*** (blocks)[1], then, subjects within each block are ***randomly*** assigned to ***treatment*** conditions.

- **RBD**: widely used in ***forestry*** research and suites well for field exp..

- So, **RBD** presented by ***blocks*** of equal ***size***, each of which contains all the treatments.

- E.g. a field has a slope [2].

- We can improve **ED** so that it ***catches*** the ***differences*** in e.g. soil fertility [3].

1	2	3
B	A	C
A	C	B
C	B	A
C	A	B

[1] the variability ***within*** blocks is ***less*** than the variability ***between*** blocks;

[2] in the direction of the arrow;

[3] each treatment is present once in each row.

- E.g. we want to study how yield of leek depends on the time for adding compost (0, 2 or 4 weeks) and distance between the rows (50 and 70 cm).

- **RBD** may be organized as follows:

Block 1		
0/50	4/70	4/50
2/70	0/70	2/50

Block 2		
4/70	0/50	4/50
2/50	0/70	2/70

Block 3		
2/70	4/70	0/70
2/50	4/50	0/50

5.10. *Latin square design*

- **LSD**: very *efficient* designs allowing two *blocking* factors [1].

- E.g. a plot of land where the fertility might *change* in both *directions*, both rows and columns can be used as *blocking factors* [2].

- E.g. to compare four varieties **A**, **B**, **C**, **D**, the field can be divided in 16 plots, each variety can be planted on four plots (replication).

- The field has *uneven* conditions, gradually changing from the top left to the lower right corner.

- Distribute the treatments *randomly* each only once in each *row* and *column.*

B	A	C	D
D	B	A	C
A	C	D	B
C	D	B	A

[1] you can simultaneously control (or eliminate) ***two*** sources of ***variability***;

[2] LSD can be used in ***many*** other exp..

5.11. *Split plot design*

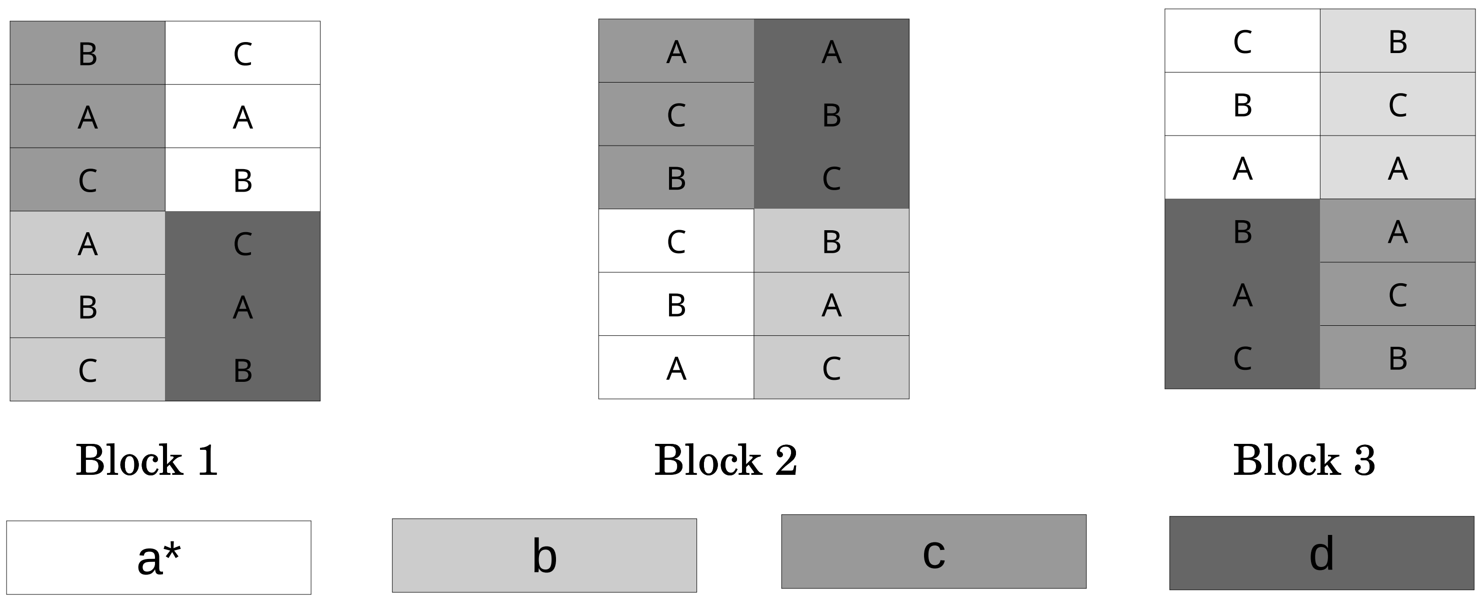
- **SPD**: consists of *two* exp. with *different* exp. units of different "**size**".

- E.g. sometimes you need to use *different* *blocking* *schemes* for different variables.

- E.g. we have *variety* (**A**; **B**; and **C**) and method of *tilling* (1; 2; 3; and 4) as factors [1].

[1] the different tilling ***methods*** can only be used on rather large areas, while plants can grow on small plots.

- **SPD**: the tilling *methods* are applied on *main* plots, while the *varieties* on *split* plots within the main.



* Tilling method