# Experimenty a jejich statistické vyhodnocení I

Biologická technika

# Odborné (neučitelské) studium

- Bi5040 Biostatistika – základní kurz
- Bi5980 Statistické hodnocení biodiverzity
- Bi7920 Zpracování biologických dat
- Bi9001 Statistická analýza experimentálních dat

# Experiment vs. pozorování

- rozdíl?
- oba postupy mohou produkovat data (statisticky vyhodnotitelné)!
- kauzální příčinnost?
  - korelační analýza
  - zejména v případě sběru dat za jednotlivá období (roky)

# Základní pojmy

- **Pokusná (experimentální) jednotka:** objekt, na kterém provádíme měření sledovaných znaků (proměnných, angl. *variable*)
- **Počet opakování:** počet pokusných jednotek v jednotlivé dílčí variantě experimentu.
- **Populace:** soubor pokusných jednotek, z něhož náhodným výběrem vybíráme pokusné jednotky, na kterých posléze provádíme vlastní experiment nebo měření. Její velikost může být konečná i (prakticky) nekonečná.
- **Náhodný výběr, randomizace:** proces, při němž z populace vybíráme pokusné jednotky, případně těmto jednotkám přiřazujeme plánované pokusné zásahy či v případě rostlin (které se většinou aktivně nepohybují z místa na místo ) jejich pozici na pěstební ploše. **Velmi důležitý krok realizace experimentu!!!** Člověk nemusí být nestranný např. při přiřazování pokusných zásahů jednotlivým pokusným jednotkám (podvědomě může stranit některé variantě). **Jiný než náhodný výběr a uspořádání při designu pokusu kde se randomizace předpokládá nemá relevantní vypovídací schopnost!**
  - generátor náhodných čísel (např. Excel, funkce RANDBETWEEN, nutno v nabídce Nástroje → Doplňky označit volbu Analytické nástroje; použijte F9 pro přepočtení dat).

# Příklad na randomizaci

- Vaším cílem je realizovat plně znáhodněný (randomizovaný) experiment, kdy budete sledovat růst rostlin ve skleníku v závislosti na 5 koncentracích hydroponického živného roztoku. K dispozici máte 150 předklíčených rostlin. $n = 10$.
- popište postup a pomocí MS Excel proveďte adekvátní randomizace!

# Typy dat

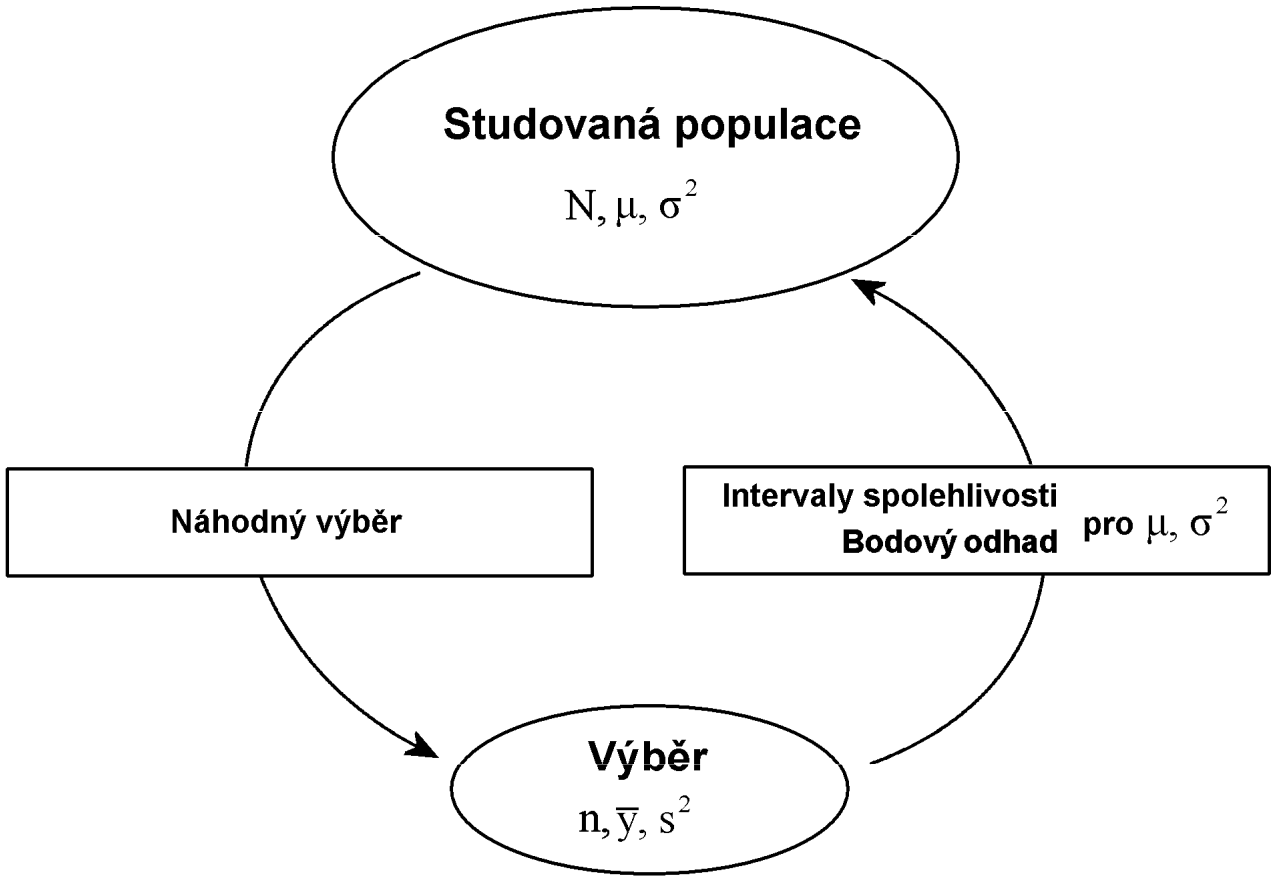
- **nominální:** jsme schopni určit jejich rovnost, případně nerovnost
- **ordinální:** jsme schopni je seřadit podle pořadí
- **kardinální**
- **intervalové:** jsme schopni interpretovat rozdíl dvou hodnot příslušného znaku
- speciální případ: data na kruhové stupnici (azimut, časové údaje...)
- **poměrové:** jsme schopni navíc interpretovat i jejich poměr; musí být smysluplná nula!

- **Znaky spojité:** mohou nabývat všech hodnot v teoreticky možném intervalu pro daný znak
- **Znaky diskrétní:** nabývají pouze několika hodnot
- **Znaky alternativní:** nabývají pouze dvou hodnot.
- **Znaky odvozené:** získávají se výpočtem z několika měřených proměnných, tj. nejsou přímo měřeny na pokusném objektu. Použití těchto znaků, např. poměrů či procent, přináší většinou komplikovanější postupy při statistické analýze.

# **Příklady:** zařaďte níže uvedená data do příslušného typu dat!

- školní klasifikace 1 až 5
- označení pokusných rostlin v experimentu
- teplota vyjádřená ve stupních Celsia
- teplota vyjádřená v Kelvinech
- datum rozkvětu rostlin
- listová plocha rostlin změřená v experimentu
- obsah chlorofylů v listech pokusných rostlin
- nominální, ordinální a kardinální typ dat dle jejich informační hodnoty
- výška rostlin
- kolonizace kořenů mykorhizními houbami

# Statistické šetření



# *Výběrový aritmetický průměr (arithmetic mean)*

- je bodovým odhadem populačního průměru

$$\bar{y} = \frac{\sum y_i}{n}$$

# *Výběrový rozptyl (variance)*

- je bodovým odhadem populačního rozptylu. Popisuje variabilitu dat, s růstem počtu opakování (*n*) se nijak zásadně nemění

$$VAR = s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

# *Výběrová směrodatná odchylka (standard deviation)*

- je bodovým odhadem populační směrodatné odchylky . Vypočteme jej jako druhou odmocninu výběrového rozptylu. Tak jako VAR, se vzrůstajícím počtem opakování se její hodnota výrazně nemění

$$SD = s = \sqrt{s^2}$$

# *Výběrová standardní chyba průměru (standard error of mean; s.e., s.e.m., SE, SEM)*

- je charakteristika přesnosti odhadu průměru (tedy alternativa k intervalu spolehlivosti). Její velikost klesá se zvyšujícím se počtem opakování (*n*)

$$SE = \frac{SD}{\sqrt{n}}$$

# *Variační koeficient (coefficient of variation; CV)*

- Je smysluplný pouze pro kardinální data na poměrové stupnici. Slouží pro porovnání variability skupin, jejichž průměry se liší.

$$CV = \frac{SD}{\bar{y}}$$

# *Další statistiky…*

- ***Rozpětí (range)***
  - Rozdíl mezi nejvyšší a nejnižší naměřenou hodnotou.
- ***Medián (median)***
  - *Md*, je prostřední hodnota z rozpětí sledované veličiny – v případě symetrického normálního rozložení stejný počet měření leží pod a nad mediánem.
- ***Modus (modus)***
  - *Mo*, nejvíce frekventovaná, tedy nejčastěji se vyskytující hodnota sledované veličiny.
- ***Kvantily, kvartily***
  - Čtvrtina dat leží pod spodním kvartilem, čtvrtina dat nad horním kvartilem.

# Příklad

- vypočtěte výše uvedené (aritmetický průměr až *CV*) popisné statistiky pro následující soubor měření.

| | | | |
|---|---|---|---|
| 2,18 | 2,89 | 2,08 | 2,49 |
| 2,28 | 1,83 | 1,21 | 1,53 |
| 2,07 | 1,60 | 2,93 | 1,16 |
| 1,68 | 4,25 | 2,41 | 3,40 |

# Prezentace výsledků – základní statistiky

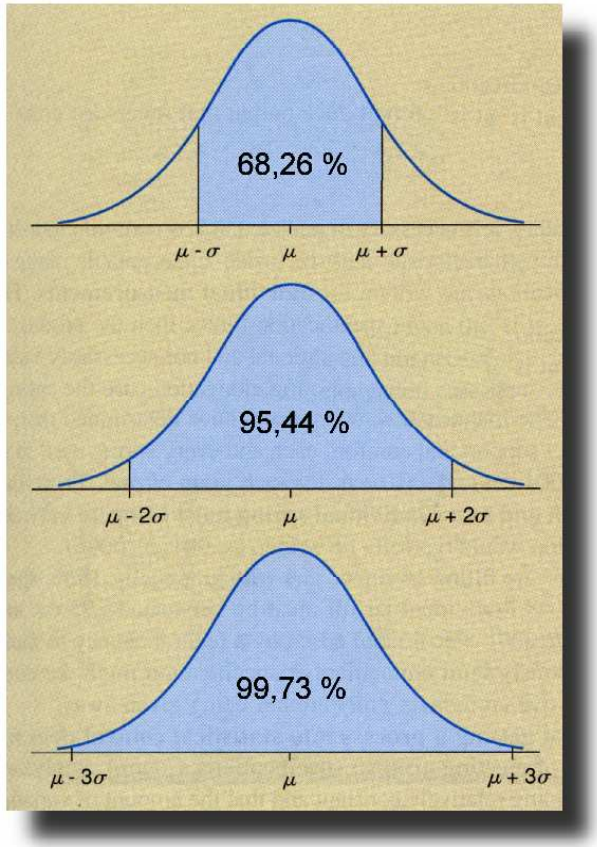
- Ať již prezentujete výsledky formou tabulky nebo grafu, vždy nezapomeňte dodržet následující „minimální“ pravidla:
  - vždy uvést, jaké statistiky prezentujeme:
    - průměr a *SD* – v případě, že chceme referovat o variabilitě dat
    - průměr a *SE*, popřípadě interval spolehlivosti – v případě, že chceme referovat o přesnosti našeho odhadu populačního průměru
  - vždy uvést počet opakování *n*, na nichž byly statistiky vypočteny; s pomocí *n* lze dopočítat ostatní základní statistiky

# *Normální rozložení*

- 68,26 % všech měření leží v intervalu mean ± *SD*
- 95,44 % všech měření leží v intervalu mean ± 2 *SD*
- 99,73 % všech měření leží v intervalu mean ± 3 *SD*

- 50 % všech měření leží v intervalu mean ± 0,674 *SD*
- 95 % všech měření leží v intervalu mean ± 1,960 *SD*
- 99 % všech měření leží v intervalu mean ± 2,576 *SD*

- *má v analýze dat výsadní postavení, neboť je předpokladem většiny parametrických statistických testů.*



# *Testování normality dat*

- **Shapiro-Wilkův W test**

V současnosti zřejmě nejpoužívanější test normality dat. Pokud vypočtená statistika *W* je významná (*P*<0,05), zamítáme hypotézu normálního rozložení testovaných dat.

- **Kolmogorov-Smirnovův test normality (příp. Lilienforsova modifikace)**

Druhý velmi často používaný test normality. Pokud vypočtená statistika *D* je významná (*P*<0,05), zamítáme hypotézu normálního rozložení testovaných dat.

- **šikmost (skewness) a špičatost (kurtosis)**

Hodnoty těchto statistik mimo rozpětí –2 až +2 indikují významné odchylky od normality.

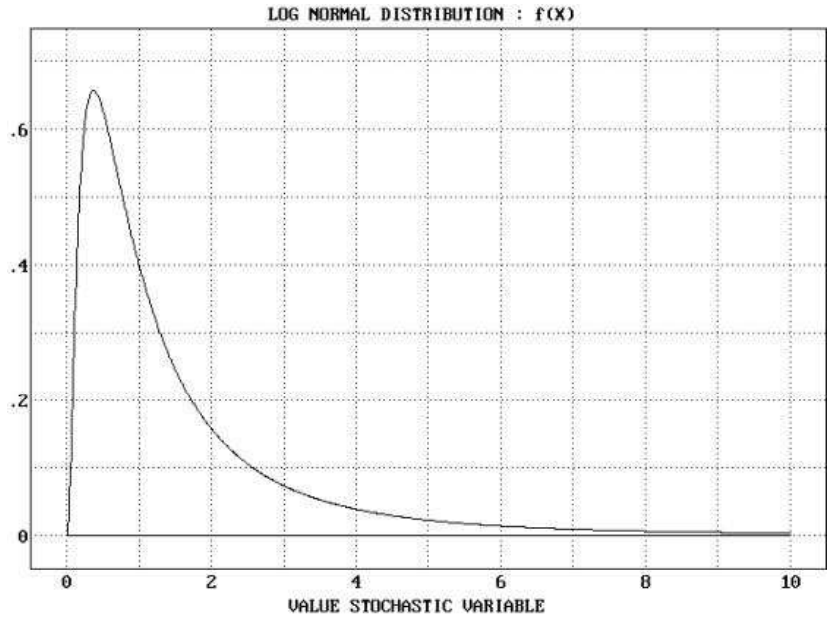
- **grafická znázornění rozložení dat**

Histogram, Leaf and Stem, grafy Box & Whisker a Normal Probability.

# *Lognormální rozložení*

- Biologická data v důsledku multiplikativních efektů často mají nikoli normální, ale tzv**. lognormální ditribuci**. Jedná se o rozložení zešikmené **doprava**. Proměnná s lognormálním rozložením má po zlogaritmování (logaritmické transformaci) normální rozložení.

- u dat z lognormálního rozložení (ale nejen u něj!) pozitivně koreluje velikost výběrového průměru a výběrového rozptylu – rychlý indikátor pro aplikaci **logaritmické transformace**



# Teorie testování hypotéz

- Statistické postupy testují pravděpodobnost platnosti tzv. **nulové hypotézy**. Tato hypotéza předpokládá, že rozdíly mezi průměry (či jinými parametry daného rozložení sledované veličiny) jsou nulové, jinými slovy, že tyto průměry jsou shodné. Matematicky lze výše uvedenou nulovou hypotézu zapsat ve tvaru:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_n$$

# Teorie testování hypotéz

- Vzhledem k faktu, že pracujeme pouze s náhodným vzorkem, můžeme na základě statistického testu dospět buď ke správnému rozhodnutí, nebo se můžeme dopustit jedné ze dvou chyb:

| **Objektivní realita** | **Naše rozhodnutí na základě statistického testu** | |
|---|---|---|
| | Zamítáme hypotézu $H_0$ | Nezamítáme hypotézu $H_0$* |
| $H_0$ je pravdivá | Chyba I. druhu ($\alpha$, $p$, $P$) | Správné rozhodnutí |
| $H_0$ je nepravdivá | Správné rozhodnutí | Chyba II. druhu ($\beta$) |

* Ve statistické mluvě nepoužíváme obrat „akceptujeme $H_0$“, nýbrž obrat „nezamítáme $H_0$“.

# Teorie testování hypotéz

- V praxi se často vypočtené hladiny významnosti statistických testů prezentují ve formě: n.s. (zkratka not significant) $P > 0,05$, * $P \leq 0,05$, ** $P < 0,01$, ***$P < 0,001$. Označení výsledku * však může znamenat, že konkrétní vypočtená P mohla být rovna 0,049, ale také 0,011. Obdobně, n.s. může znamenat výsledek testu $P = 0,051$, ale také $P = 0,95$! **Je nesporně lepší vždy uvádět konkrétní vypočtenou *P* namísto hvězdičkové symboliky!** Výsledek testu $P = 0,055$, $n = 8$ jednoznačně napovídá, že při větším počtu opakování by vypočtená hladina významnosti mohla klesnout pod 5%.

# Vztahy mezi hladinou významnosti P, velikostí a variabilitou výběru a velikostí rozdílu výběrových (populačních) průměrů

- Na následujících čtyřech snímcích budou pouze kvalitativně definovány vztahy mezi veličinami, se kterými se setkáváme při statistických výpočtech.
- Pro zjednodušení budeme uvažovat nejjednodušší možný případ, kdy testujeme statistickou významnost rozdílu mezi výběrovými průměry dvou souborů dat.

- **Mají-li oba soubory dat konstantní velikost a variabilitu, potom se zvyšujícím se rozdílem jejich výběrových průměrů se snižuje hodnota *P*.**

- **Hladina významnosti *P* je při objektivně existujícím nenulovém rozdílu srovnávaných populačních průměrů nepřímo úměrná velikosti výběru.**

- **Je-li rozdíl populačních průměrů malý, může být na základě statistických testů označen za průkazný ($P \leq 0,05$) pouze v případě testování na velkých výběrových vzorcích (velké $n$).**

- **Nulovou hypotézu nelze na základě statistického testu NIKDY potvrdit.**

# Srovnání dvou souborů dat (*t*-testy)

- dva soubory dat, vzniklé buď dvěma výběry, nebo dvěma pokusnými zásahy
- plně znáhodněný vs. párový experiment
  - u **plně znáhodněného experimentu** není možné data z obou souborů uspořádat do dvojic (párů), u kterých bychom mohli považovat chybovou složku za nulovou. Jinými slovy, nelze vytvořit dvojice experimentálních jednotek např. se stejným genetickým založením. Vše je plně randomizováno!
  - u **párového uspořádání** jsou buď 1) data (resp. experimentální jednotky, na kterých byla tato data naměřena) uspořádána po dvojicích, u kterých víme, že jejich genetické založení je zcela shodné či alespoň velmi podobné, popř. 2) měříme opakovaně na jedné pokusné jednotce – opakovaně buď v průběhu času, nebo lépe v jeden moment na různých částech téhož objektu. Hodnotu chybové složky pak považujeme pro oba členy každého páru za stejnou. Kromě obecné závislosti hodnot v párech je vše ostatní rovněž randomizováno!

# Srovnání dvou souborů dat (*t*-testy)

- vstupní podmínkou pro aplikaci parametrických testů (*F*-test, *t*-test) je **normální distribuce vstupních dat (resp. normalita rozdílů spárovaných dat v případě párového designu)**. Pokud data tuto podmínku nesplňují, je s výjimkou velkých souborů dat nutné aplikovat neparametrické testy, popřípadě lze data transformovat. Pro aplikaci příslušného *t*-testu u plně znáhodněného experimentu se navíc testuje i **homogenita rozptylů** (pomocí *F*-testu).

# Srovnání dvou souborů dat (*t*-testy) – postup s využitím MS Excel

- rozhodnutí, zda se jedná o párový nebo plně znáhodněný design
  - párový design:
    - spočítat rozdíl dat v jednotlivých párech, tento rozdíl otestovat na normalitu (Analýza dat – Popisná statistika; šikmost a špičatost v intervalu <-2; 2>)
    - spočítat pomocí nástroje Analýza dat – Dvouvýběrový párový t-test na střední hodnotu

# Srovnání dvou souborů dat (*t*-testy) – postup s využitím MS Excel

- plně znáhodněný design:
  - pomocí nástroje Analýza dat – Popisná statistika otestovat normalitu dat ve skupinách (šikmost a špičatost v intervalu <-2; 2>)
  - pomocí nástroje Analýza dat – Dvouvýběrový F-test pro rozptyl otestovat homogenitu rozptylů
  - podle výsledků F-testu spočítat Dvouvýběrový t-test s rovností/nerovností rozptylů

# Jednocestná ANOVA

- ANOVA – analysis of variance, analýza rozptylu
- porovnání průměrů více než dvou skupin současně
- např. kultivační experiment ve cvičením z Fyziologie rostlin - vliv deficience vybraných živin (N, P, Fe, Ca) na růstové charakteristiky kukuřice seté

# Jednocestná ANOVA

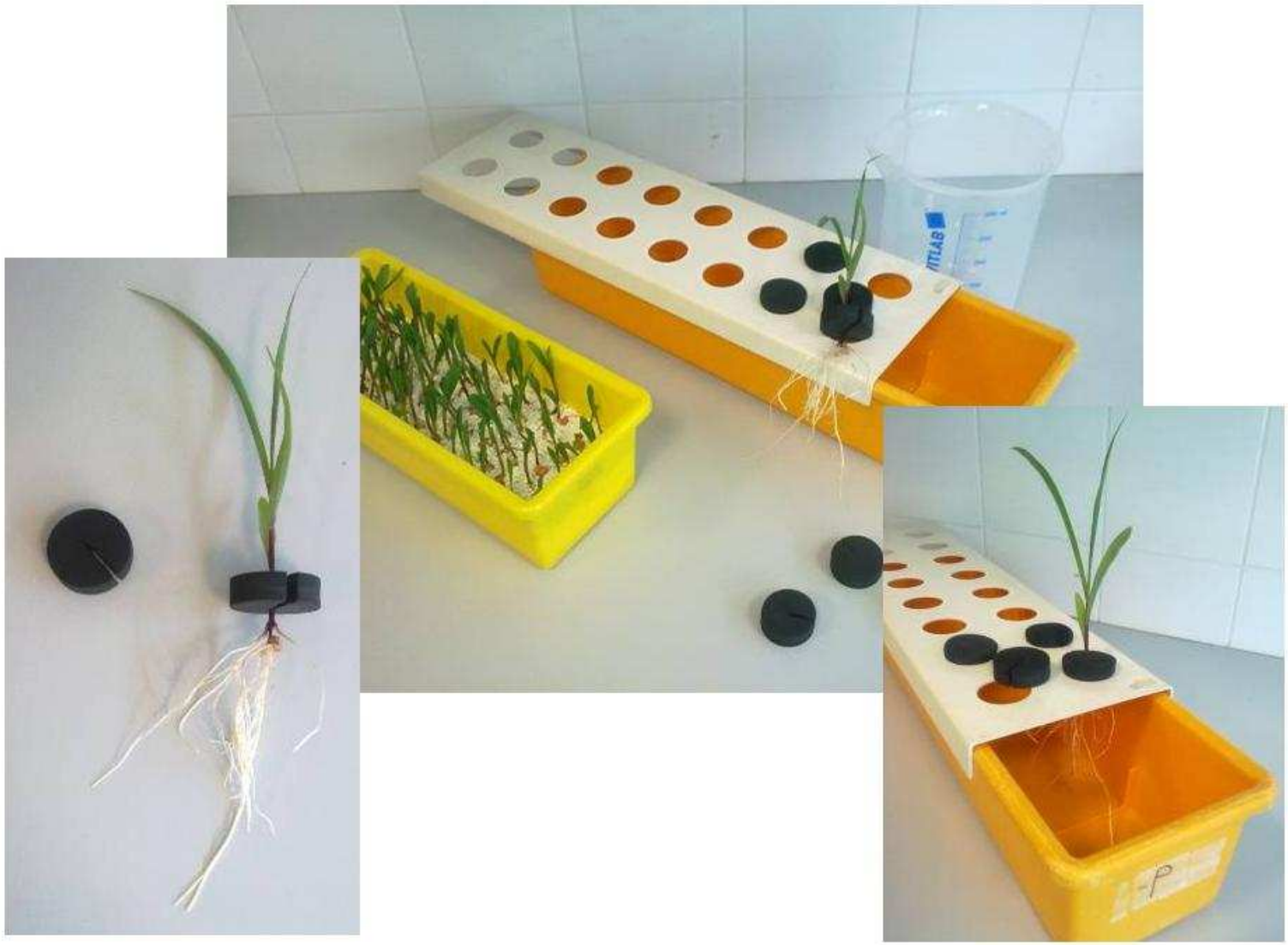
# Jednocestná ANOVA

- proč je lépe použít jednu analýzu rozptylu než sérii t-testů?
  - kolik dílčích t-testů bychom museli spočítat pro výše zobrazený kultivační experiment s 5 variantami?
  - jakou maximální chybu I. druhu připouštíme v každém dílčím  t-testu?
  - jaká by byla celková pravděpodobnost chyby I. druhu při porovnání všech variant t-testy?

# Jednocestná ANOVA - předpoklady

- související s designem experimentu, tj. randomizací:
  - náhodnost jednotlivých opakování uvnitř skupin i mezi skupinami
  - nezávislost chyb
- související s vlastními získanými daty
  - homogenita rozptylů (shodnost variability jednotlivých skupin)
  - normalita chyb (reziduí)

# Jednocestná ANOVA - předpoklady



# Jednocestná ANOVA - předpoklady

# Jednocestná ANOVA - předpoklady