

Biostatistika

V čem se liší biostatistika od statistiky? Nebiostatistiku by použil například analytický chemik, který chce zjistit jaká je koncentrace vybrané sloučeniny v kapalném vzorku. Má k dispozici kapalý vzorek, který je zcela homogenní. Analytický chemik ze vzorku odebere několik vzorků a ty analyzuje, například chromatograficky. Koncentrace pro každý vzorek mu vyjde trochu jinak, protože chromatografie není perfektní. Z těchto hodnot je možné vypočítat průměr, směrodatnou odchylku atd a tak se dostat k odhadu skutečné koncentrace. Pokud by byl počet měření velmi vysoký, pak by se přiblížil skutečné hodnotě koncentrace.

Biostatistika

Naproti tomu, biostatistik by mohl zjišťovat, jestli lék na snížení krevního tlaku skutečně snižuje tlak. Změřil by tlak stovce lidí, kteří lék užívají, a srovnat jejich tlak se známým průměrným tlakem v populaci (tento postup byl vybrán pro názornost, skutečné testování účinnosti léku by bylo složitější). Podobně jako nebiostatistik by biostatistik spočítal průměr a další statistiky. Hlavní rozdíl mezi bio a nebiostatistikou je v tom, že různí pacienti mají skutečně různý krevní tlak. Odchyly nejsou způsobeny (alespoň z větší části) nepřesností měření.

Biostatistik si musí představit “průměrného” pacienta, který má průměrný tlak. Reální pacienti se od takového průměrného pacienta liší v genetice a v tom jak byli během života vystaveni různým vlivům.

Inferenční statistika

V takzvané inferenční (nebo též indukční) statistice (*inferential statistics*) testujeme statistické hypotézy na základě teorie pravděpodobnosti. Obvykle rozeznáváme dva druhy protichůdných hypotéz, nulovou a alternativní hypotézu. Ve vědecké praxi se snažíme objevit něco nového co změní současný stav světa.

Nulová hypotéza (*null hypothesis*, H_0) spočívá (obvykle) v tom, že se nic nemění, nic se neděje, nic jsme neobjevili, stav poznání je takový jaký byl před provedením experimentu. V případě testování nového léčiva je nulovou hypotézou to, že nové léčivo nefunguje.

Alternativní hypotéza (*alternative hypothesis*, H_A) je (obvykle) opakem nulové hypotézy, tedy že se něco děje, mění, že stav poznání se experimentem změnil. V případě testování nového léčiva je alternativní hypotézou to, že léčivo funguje.

Inferenční statistika

Dále můžeme rozeznávat biologickou a statistickou nulovou a alternativní hypotézu.

Biologickou nulovou hypotézou může být: „testovaná sloučenina nesnižuje krevní tlak“

Biologickou alternativní hypotézou může být: „testovaná sloučenina snižuje krevní tlak“

Statistickou nulovou hypotézou může být: „průměrný krevní tlak pacientů užívajících testovanou sloučeninu je stejný nebo vyšší než pacientů užívajících placebo“

Statistickou alternativní hypotézou může být: „průměrný krevní tlak pacientů užívajících testovanou sloučeninu je nižší než pacientů užívajících placebo“

Inferenční statistika

Pomocí experimentů a jejich statistické analýzy je možné, na určité hladině pravděpodobnosti, **vyvrátit nulovou hypotézu a přijmout alternativní hypotézu.**

Nulové hypotéza se nikdy nepřijímá. Představte si například léčivo, které snižuje systolický krevní tlak, ale jen o jeden milimetr rtuti. Pokud by kontrolní a experimentální skupina měly obě po deseti pacientech, pak je velmi nepravděpodobné, že by byl rozdíl pouhého jednoho milimetru rtuti takovýmto experimentem odhalen, když směrodatná odchylka systolického krevního tlaku je přibližně 20 mmHg. Nepomohlo by, kdyby kontrolní a experimentální skupina měly po stovce pacientů. Pokud by ale obě skupiny obsahovaly po tisíci pacientech, pak už existuje slušná šance na zamítnutí nulové hypotézy. Pokud bychom v prvních dvou případech přijali nulovou hypotézu a tvrdili, že léčivo nesnižuje krevní tlak, pak by to bylo špatně. Správným závěrem v prvních dvou případech je, **že nemáme dostatek dat, abychom zamítli nulovou hypotézu a že nulovou hypotézu nezamítáme** (ani ji nepřijímáme, ani nezamítáme, ani nepřijímáme alternativní hypotézu).

Chyby

V inferenční statistice rozeznáváme dva druhy chyby, **chybu I. a chybu II. typu** (*type I error a type II error*). Pokud platí nulová hypotéza (léčivo nefunguje) a my nulovou hypotézu nezamítneme, pak je vše v pořádku. Pokud platí alternativní hypotéza a my zamítneme nulovou hypotézu a přijmeme alternativní, pak je také vše v pořádku. Chyba I. typu nastane v okamžiku, když platí nulová hypotéza, ale my ji zamítneme, jinými slovy že zjistíme, že léčivo funguje, i když ve skutečnosti nefunguje. Chyby II. typu se dopustíme v okamžiku, když platí alternativní hypotéza, ale my nezamítneme nulovou hypotézu, jinými slovy že nezjistíme, že léčivo funguje, i když funguje.

Hladina spolehlivosti

Před každým testem hypotézy je nutné zvolit **hladinu spolehlivosti** (*significance level*), která se obvykle označuje symbolem α . Volba této hodnoty bude diskutována později. Ve stručnosti, pokud chceme minimalizovat pravděpodobnost chyby I. typu, pak zvolíme velmi nízkou hodnotu α , například 0,01 nebo menší. Naopak, pokud chceme minimalizovat pravděpodobnost chyby II. typu, pak zvolíme velmi vysokou hodnotu α , například 0,05. Při volbě hodnoty α je nutné vzít v potaz dopady toho, že nefunkční léčivo může být podáváno pacientů, nebo naopak že funkční léčivo může být z testů vyřazeno jako nefunkční.

P-hodnota

Dalším klíčovým konceptem statistického testování hypotéz je takzvaná **p-hodnota** (*p-value*). Jedná se o hodnotu pravděpodobnosti, že, za předpokladu platnosti nulové hypotézy, získáme statistiku (například rozdíl mezi průměry měřených hodnot mezi kontrolní a experimentální skupinou) stejný nebo vyšší než jsou skutečné výsledky experimentu. V případě testování léku na snížení krevního tlaku je význam p-hodnoty následující. Představme si na chvíli, že testované léčivo nefunguje, tedy že platí nulové hypotéza. Pak budou mít pacienti v kontrolní a experimentální skupině přibližně stejný krevní tlak. Nejpravděpodobnější výsledek je ten, že průměrný systolický krevní tlak v experimentální a kontrolní skupině bude stejný. Protože se jedná o náhodné veličiny, existuje relativně vysoká pravděpodobnost, že se průměrné krevní tlaky budou lišit o 1 mmHg, nižší, že se budou lišit o 10 mmHg a tak dále. Tato pravděpodobnost se nazývá **p-hodnota**.

P-hodnota

P-hodnota v tomto případě závisí na statistice, tedy například rozdílu průměrných krevních tlaků mezi kontrolní a experimentální skupinou, na směrodatných odchylkách v obou skupinách a na počtu pacientů v obou skupinách. Je vysoká pravděpodobnost, že rozdíl krevních tlaků mezi kontrolní a experimentální skupinou bude nízký (například 5 mmHg, tedy $\frac{1}{4}$ směrodatné odchylky), zvláště pokud je pokus proveden na malém počtu pacientů. Naopak, je velmi nízká pravděpodobnost, že rozdíl krevních tlaků mezi kontrolní a experimentální skupinou bude vysoký (například 20 mmHg, tedy hodnota směrodatné odchylky), zvláště pokud je pokus proveden na velkém počtu pacientů.

Testování hypotéz

Praktickým provedením statistické analýzy dat je následující. Nejprve je nutné naplánovat experiment s ohledem na zpracování výsledných dat. Design experimentu bude dále diskutován. Pro daný experiment dále formulujeme nulovou a alternativní hypotézu a hladinu spolehlivosti. Po provedení experimentu vypočteme vhodným softwarem p-hodnotu a srovnáme ji s hladinou spolehlivosti. Pokud je p-hodnota nižší než hladina spolehlivosti, pak zamítáme nulovou hypotézu a přijímáme alternativní hypotézu. Pokud je p-hodnota vyšší než hladina spolehlivosti, pak nezamítáme nulovou hypotézu.

Testování hypotéz

Praktickým provedením biostatistické analýzy dat je následující (1):

1. Specifikujte biologickou otázku, která má být zodpovězena.
2. Formulujte biologickou nulovou a alternativní hypotézu.
3. Formulujte biologickou otázku ve formě statistické nulové a alternativní hypotézy.
4. Zjistěte které veličiny jsou relevantní k biologické otázce.
5. Zjistěte o jaký typ veličin se jedná (spojité nebo kategorické).
6. Podle počtu, typu a předpokládaného rozdělení a podle testované hypotézy vyberte vhodný statistický test a navrhnete experiment.
7. Proveďte experiment.
8. Otestujte data jestli splňují předpoklady zamýšlených testů (normalita, homoskedasticita, případně jiné). Pokud neplatí, vyberte vhodnější testy.
9. Proveďte statistické testy a interpretujte jejich výsledky.
10. Prezентуйте své výsledky.

Testování hypotéz

Test normálního rozdělení:

Normální rozdělení je základním předpokladem řady statistických testů, jako je t-test nebo ANOVA. Proto je před provedením těchto testů potřeba otestovat zdali mají normální rozdělení. Často používaným testem je Shapirův-Wilkův test. Nulovou hypotézou je, že hodnoty mají normální rozdělení, alternativní hypotézou je že normální rozdělení nemají. V případě, že je zamítnuta nulová hypotéza, je nutné použít neparametrické metody, které nevyžadují normální rozdělení.

R: `shapiro.test(vyber)`

Testování hypotéz

Jednovýběrový t-test:

Pokud chceme porovnat výběr s nějakou přesnou hodnotou, pak můžeme provést jednovýběrový t-test. Jednovýběrový t-test může být oboustranný nebo jednostranný.

Nulovou hypotézou oboustranného testu je, že průměr výběru je rovný srovnávané hodnotě. Nulovou hypotézou jednostranného testu je, že průměr výběru je menší nebo rovný nebo větší nebo rovný srovnávané hodnotě. Existuje řada veličin jejichž přesná hodnota pro kontrolní skupinu je v podstatě známa (např. průměrné IQ je 100). V biologických a lékařských disciplínách ale častěji nemáme přesnou hodnotu k dispozici a srovnáváme dva výběry, tedy provádíme dvouvýběrový t-test.

R: `t.test(vyber)`

Testování hypotéz

Dvouvýběrový t-test:

Pokud chceme porovnat dva výběry (například krevní tlak pacientů, kteří užívali a neužívali léčivo), pak můžeme provést dvouvýběrový t-test. Dvouvýběrový t-test existuje jako párový a nepárový. Párový t-test bychom použili, pokud máme od každého pacienta jedno měření tlaku před podáním léčiva a druhé po podání. Tato data jsou spárována na základě identity pacientů. Pokud by byla skupina pacientů, kteří by dostali léčivo a skupina, která by nedostala nic, nebo dostala placebo, pak bychom použili nepárový t-test. Nepárový t-test existuje ve dvou variantách (viz dále). T-test obecně vyžaduje normální rozdělení výběrů.

R: `t.test(vyber1, vybr2)`

Testování hypotéz

Nepárový dvouvýběrový t-test:

Nepárový t-test existuje ve dvou variantách, podle toho, jestli mají porovnávané výběry stejné rozptyly. To je možné otestovat pomocí F-testu pro srovnání rozptylu výběrů (v R funkce `var.test`). Nulovou hypotézou tohoto testu je, že výběry mají stejné rozptyly. Pokud nezamítneme nulovou hypotézu, pak použijeme jednu variantu t-testu (v R funkce `t.test` s volbou `var.equal=TRUE`), pokud zamítneme tak druhou (v R funkce `t.test` s volbou `var.equal=FALSE`)

```
R: var.test(vyber1, vyber2)
t.test(vyber1, vyber2, var.equal=TRUE)
t.test(vyber1, vyber2, var.equal=FALSE)
```

Testování hypotéz

Neparametrické jedno a dvouvýběrové testy:

Pokud nemají data normální rozdělení, je možné použít neparametrické metody. Je možné použít jedno- nebo dvouvýběrový Wilcoxonův test (dvouvýběrová varianta je známa jako Mannův-Whitneyův test). Existuje nepárová a párová varianta dvouvýběrového testu. Použití je velmi podobné jako v případě t-testu.

```
R: wilcox.test(vyber)
   wilcox.test(vyber1, vyber2)
   wilcox.test(vyber1, vyber2, paired=T)
```


Testování hypotéz

Analýza rozptylu (ANOVA):

Analýza rozptylu se používá v okamžiku, kdy chceme porovnat více než dva výběry, například už v okamžiku, kdy máma tři výběry. Například můžeme mít k dispozici hodnoty krevního tlaku pacientů z kontrolní skupiny, kterým nebylo podáno ani léčivo ani placebo, pacientů kterým bylo podáno placebo a pacientů, kterým byl podán testovaný lék.

Nulovou hypotézou je, že průměrný krevní tlak ve všech třech skupinách je stejný. Alternativní hypotézou je, že se průměrný krevní tlak ve skupinách liší, buď že se průměr jednoho výběry výrazně liší od ostatních, nebo že se všechny průměry navzájem výrazně liší. Pro provedení ANOVy existují předpoklady, které je nutné splnit.

R: `anova(lm(vysledky~faktor, data=vybery))`

Testování hypotéz

Předpoklady pro použití analýzy rozptylu:

Aby bylo možné použít analýzu rozptylu, je nutné aby data splňovala následující podmínky. Za prvé, hodnoty musí mít normální rozdělení. Pokud tomu tak není, je možné nahradit analýzu rozptylu pomocí Kruskalova-Wallisova testu. Za druhé, všechny výběry musí mít stejné rozptyly. Pokud tomu tak není, je možné se pokusit data transformovat tak, aby rozptyly byly stejné, například je logaritmovat. Za třetí, chyby výběrů musí být vzájemně nezávislé.

R: shapiro.test
bartlett.test
fligner.test

Testování hypotéz

Vícefaktorová ANOVA:

ANOVA může být provedena i vícefaktorově. Například nás zajímá jak závisí nějaká hodnota, která vystihuje zdravotní stav, na tom, jestli člověk kouří, pije, na jeho příjmech, jestli sportuje a podobně. Postupovali bychom tak, že do jednotlivých sloupečků bychom zapsali informaci zda člověk kouří, pije, případně kolik toho vykouří a kolik toho vypije, a do posledního sloupečku bychom zapsali veličinu charakterizující zdravotní stav. ANOVou bychom mohli zjistit které faktory tuto veličinu ovlivňují. Dále bychom mohli využít ANOVu s interakcemi, čímž bychom mohli například zjistit, že samotné kouření nebo samotné pití alkoholu veličinu neovlivňuje, ale kombinace kouření a pití ano. Vícefaktorová ANOVA má další podmínky pro správné provedení.

Testování hypotéz

Vícenásobné porovnávání:

Pokud srovnáváme mnoho výběrů mezi sebou, je nutné nějakým způsobem kompenzovat skutečnost, že s každým porovnáním roste riziko nesprávného zamítnutí nulové hypotézy. Pokud změříme pomocí real-time PCR koncentrace vybrané mRNA v pěti vzorcích zdravé tkáně a pěti vzorcích nemocné tkáně, pak můžeme použít dvouvýběrový t-test a zjistit jestli se koncentrace mRNA liší. Pokud zvolíme $\alpha=0,05$, pak máme 5% šanci, že vyjde, že se liší, i když se neliší. Pokud ale něco podobného uděláme pro 10 000 genů změřených pomocí DNA čipu nebo RNA seq, pak máme pravděpodobnost $(1-0,05)^{10000} = 5,2 \cdot 10^{-178}$, že pro všechny geny správně vyjde že se koncentrace jejich mRNA neliší, tedy máme prakticky jisté $(1-5,2 \cdot 10^{-178})$, že alespoň jednou vyjde, že se liší.