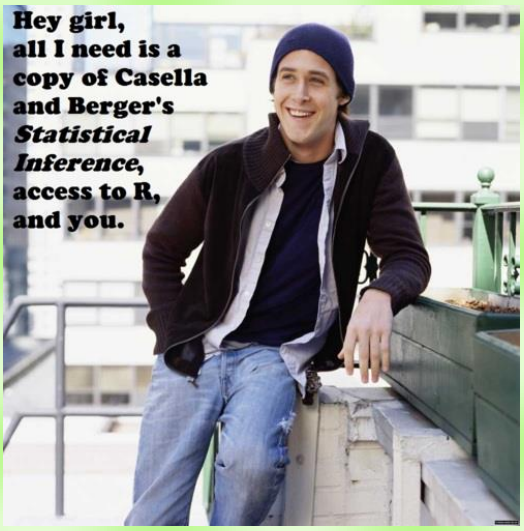
Ze skript Základy statistiky v prostředí R, přednášek, internetu, a vlastního úsudku sestavila Majda. Nedoporučuji číst biostatistikům.

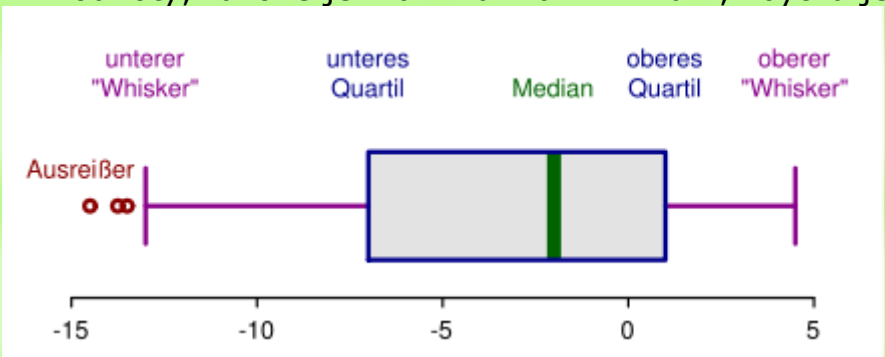
POPISNÁ STATISTIKA

- **Měřítka**
  - 0-1 – jen dvě varianty, př. pohlaví, kouří/nekouří
  - nominální – více nečíselných možností, tzv. faktor, př. město, odrůda
  - ordinální – uspořádaný faktor=uspořádané hodnoty nominálního měřítka, př. stupeň vzdělání
  - poměrové – srovnání s jednotkou – kolikrát je x větší než y, př. váha, výška, věk
  - intervalové – stejné vzdálenosti sousedních – o kolik je větší, př. rok narození
  - hrubší dělení měřítek
    - kvalitativní (diskrétní veličiny) – tam patří: 0-1, nominální, často ordinální – udávají se četnosti hodnot – kolikrát která nastala
    - kvantitativní (spojité veličiny) – tam patří: intervalové, poměrové, někdy ordinální, hodnoty jsou čísla



- **Míry polohy –** pomocí nich charakteristika vzorku jedním číslem – ukazují na velikost hodnot
  - Variační řada - hodnoty uspořádané od nejmenší po největší – z ní odvodíme min, max – ty jsou ale pro popis vzorku dost nevhodné. Variační řada se značí x a číslo v závorce – X(1), X(2),… - z ní odvodíme pořadí – X(1) má pořadí 1 atd… Když jsou čísla vedle sebe stejná, dostanou průměrnou hodnotu pořadí – např. 3,5, 3,5..; v řadě 3, 3, 3 je pořadí 2, 2, 2!
  - Průměr, vážený průměr (s využitím četností – vše vynásobíme vahami jednotlivých hodnot)- x s čárkou
  - Medián – dělí variační řadu na 2 stejné části – vetší a menší než medián, když je jich sudý počet, je medián průměr ze středních dvou, - x s vlnovkou
  - Minimum, maximum
  - Horní kvartil, dolní kvartil – mediány pro ¼, značka Q1, Q3, první decil
  - Výběrový kvantil x_p = p-tý percentil – odděluje p-té procento nejmenších hodnot od hodnot větších.. Medián je percentil $x_{0,5}$, kvartily x0,25 a x0,75!
  - Modus – x se stříškou – nejčastěji se vyskytující hodnota, stříškou se ve statistice také značí odhad

- **vlastnosti** – přičítání i násobení konstantou u každého prvku – pak i každá z měr polohy (obecná míra polohy – mí) stejná, jen změněná o konstantu
- graf – box-plot – červené divné slovo znamená odlehlé hodnoty, růžově je maximum a minimum, zbytku je rozumnět



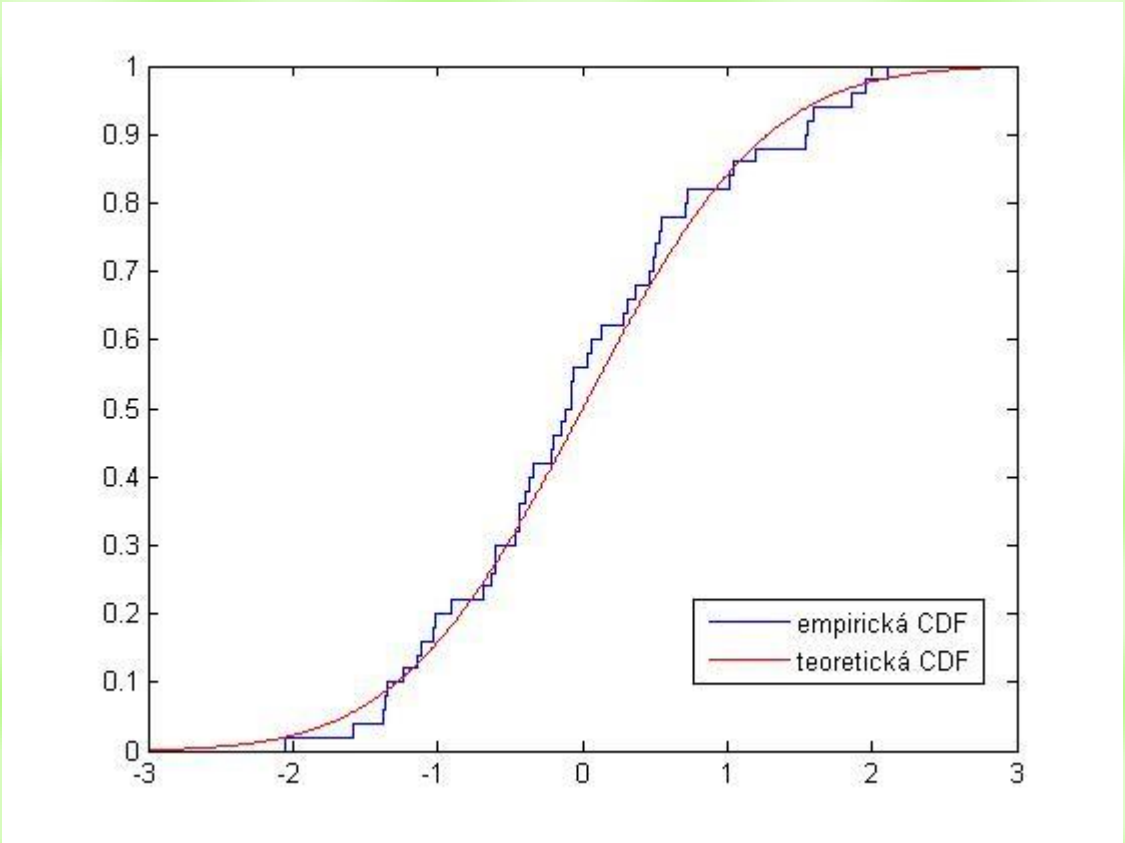
- **Míry variability**
  - číselně charakterizuje jinou vlastnost než míra polohy, ukazuje kolísání, variabilitu
  - **Podmínky**: Nesmí být měněny úrovně, ve které počítáme míry polohy, protože přičtu-li ke všem prvkům konstantu a, míry variability se nemění, vynásobím-li všechny prvky konstantou b (větší než 0), směrodatnou odchylku také vynásobím touto konstantou, rozptyl konstantou na druhou (viz níže) a ten zbytek radši ani nevědět..
  - směrodatná odchylka (SD, standard deviation) - sx – odmocnia rozptylu, platí pro ni obě podmínky.
  - (výběrový) rozptyl – sx^2 – téměř průměr druhých mocnin vzdáleností konkrétních hodnot od průměru – jak moc kolísají jednotlivé hodnoty kolem průměru, také se dá chápat jako míra vzájemné rozdílnosti hodnot x1-xn; platí pro něj pravidla o měrách variability, ale když násobím prvky konstantou b, musím roznásobit rozptyl b^2
  - rozpětí – R = xMax-Xmin, xn-x1
  - kvartilové rozpětí – Rq = Q3 - Q1 – lepší, není ovlivněno odlehlými hodnotami – interval, co obsahuje prostřední část dat
  - variační koeficient – Vx = sx/x prům. – nesplňuje podmínky nahoře – bud to kazí průměr, nebo se to vykrátí a nezbyde nic, pro co by se podmínka dělala ☺.. Používá se pro porovnání hodně polohou různých souborů dat

- **Z-skór**
  - založený na upravených původních číselných hodnotách, jeho hodnoty nezávisí na míře polohy ani variability
  - z-skor - zi říká, jak daleko je hodnota xi od průměru všech x. jednotka vzdálenosti je směrodatná odchylka sx. Kladná hodnota=xi je větší než průměr, záporná naopak
  - zi=xi-xprům/sx
  - používá se při určování normality – šikmost a špičatost

- průměr z-skóru = 0, směrodatná odchylka = 1

- **Šikmost a špičatost**
  - k posouzení, jestli číselná hodnota pochází z normálního rozdělení
  - šikmost – průměr 3.mocnin z-skórů, vyjadřuje symetrii rozložení pozorování kolem průměrné hodnoty
  - špičatost – průměr 4.mocnin z-skórů zmenšený o 3

- **Empirická distribuční funkce**
  - lze ji sestrojit pro ordinální a spojitou náhodnou proměnou, pro dané x je dána relativní četností hodnot menších nebo stejných než to x
  - ke každému x se zjistí, kolik hodnot je menších nebo stejně velkých jako x, a to číslo se dělí celkovým počtem hodnot.
  - Je odhadem distribuční funkce



TEORIE

- **Náhodné jevy**
  Náhodný pokus – dělá se za přesných podmínek a jeho výsledek je nejistý. S počtem opakování pokusu je ale četnost určitého výsledku stále stabilnější – např. stále stabilněji se rodí 50% holčiček

- libovolné tvrzení o výsledku náhodného pokusu – např. na kostce padne liché číslo
- skládají se z elementárních jevů – to je výčet všech výsledků, co můžou nastat. – vždy musí nastat jeden z nich a nesmí nastat 2 současně. Např. 6 možností na kostce; A, B, AB, 0 u krve..
- B – jev náhodný – podmnožina jevu jistého, Ω – jev jistý – nastane vždy, souhrn všech elementárních jevů – např. padne 1, nebo 2, nebo…6, Ø – jev nemožný – jediný není složen z elementárních jevů, je jím prázdná množina; podjev, jev opačný (s pruhem nahoře) – pokud nenastane jev A, nastane jev A´; Jevy neslučitelné – jejich průnikem je prázdná množina
- průnik jevů D a B – platí D a zároveň B, sjednocení jevů D a B – nastává alespoň jeden z jevů

- **Pravděpodobnost**
  - objektivní číselné vyjádření „naděje", že nastane daný jev, relativní četnost výsledku náhodného pokusu se stabilizuje při mnoha opakováních právě na hodnotě pravděpodobnosti
  - má stejné vlastnosti jako relativní četnost – je jejím idealizovaným protějškem
  - P(B)=mb/m mb…varianty vyhovující jevu B
    m…všechny varianty

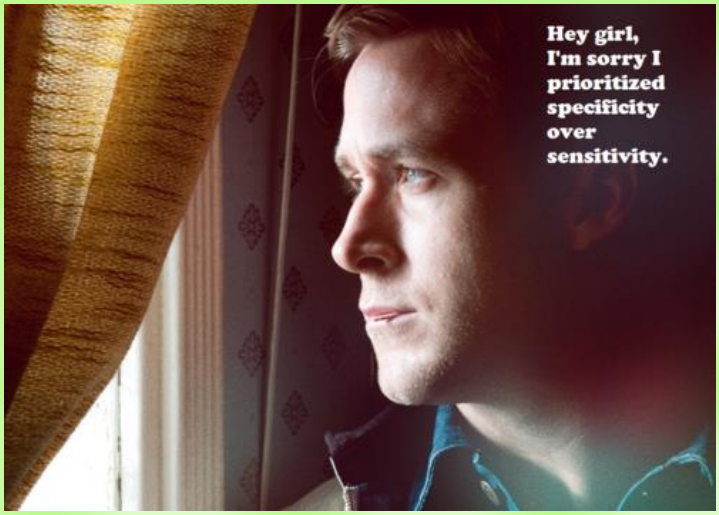
vlastnosti:
sčítání pravděpodobností: pravděpodobnost, že nastanou dva neslučitelné jevy, je rovna součtu jejich dílčích pravděpodobností
obecně $0 \leq P \leq 1$ – vždy nezáporná!, jev jistý má P=1 – A+nonA(jev opačný)=1

- **Podmíněná pravděpodobnost**
  - Máme kromě normální pravděpodobnosti ještě nějakou info o tom, že ta pravděpodobnost bude nějak vychýlená - pst jednoho jevu je ovlivněna pstí jevu jiného
  - když víme, že nastal jev A, pak podmíněná pst jevu B za podmínky A bude rovna relativní velikost B-průnik-A k velikosti A
  - P(B|A) = P(A-průnik-B) / P(A) – Vzorec pro násobení pravděpodobností nezávislých jevů
  - může být <,=,> než P(A)

- **Bayesův vzorec**

-umožňuje „obrátit" podmíněnu pst – místo P=A/B dovede určit P=B/A
-Máme nemoc, která se projevuje s určitou pravděpodobností – prevalencí. Vyšetřujeme jí testem, který nám dá výsledek. **Senzitivita** (citlivost) tohoto testu je pravděpodobnost, že správně určí nemocného, **specifita** je naopak pravděpodobnost, že na zdravého nezareaguje. Protože se ale může stát, že test určí zdravého jako nemocného, musíme

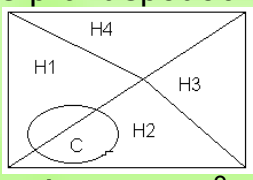
se ptát, jaká bude pravděpodobnost, že je určený nemocný skutečně nemocný a že zdravý s negativním výsledkem je skutečně zdravý. A to vše chceme zjistit z informace o existenci nebo neexistenci nemoci. Vypočítá nám to Bayesův vzorec:



P nemocného při pozitivním testu= senzitivita x prevalence/ senzitivita x prevalence+(1-P opravdu zdravého negativního x zbytek populace, co nemá prevalenci)

Obecně:

- P(A/B)=P(B/A)xP(A) / (B/A)xP(A)+P(B/nonA)xP(nonA)
- Hj – Rozdělení jevu jistého na x neslučitelných jevů, z nichž jeden musí platit. Těmto jevům se říká hypotézy.
- známe apriorní psti P(Hj)
- známe pravděpodobnost P(C|Hj)



Když spojíme hypotézy, Bayesův vzorec a úplnou pravděpodobnost pro počítání s hypotézami, dostaneme tvar: P(Hj/B)=P(B/Hj)xP(Hj) / suma od všech hypotéz P(B/Hi)xP(Hi)

Podmíněnou pravděpodobnost B/Hi známe, apriorní pravděpodobnost Hi také, a protože máme podmínku, že nastal jev B, dopočítáme aposteriorní pravděpodobnost (Hj/B) ☺ ☺ ☺

- **Úplná pravděpodobnost**
- vzorec pro úplnou pravděpodobnost klíčový pro Bayesův vzorec.
- Pravděpodobnost, s jakou náhodně zvolená osoba pozitivně reaguje na test – sjednocení pozitivních zdravých a pozitivních nemocných – jsou neslučitelné, takže je sečteme. Tento součet je úplná pravděpodobnost
    - P(C) je váženým průměrem podmíněných pravděpodobností

Ppozit.=senzitivita x prevalence+P zdravého, co reaguje pozitivně x ta část populace, kde není prevalence

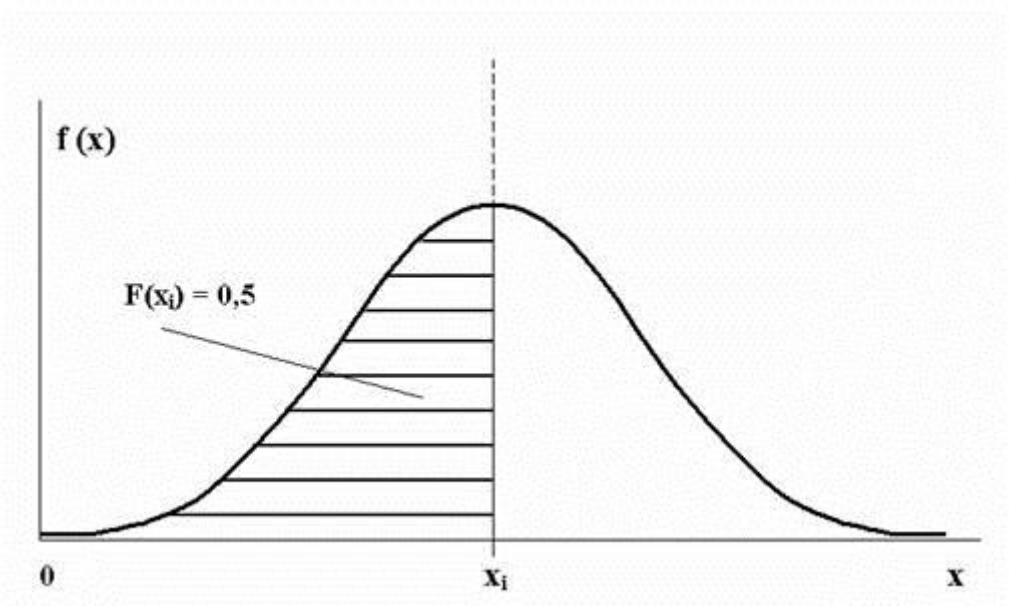
Když počítáme s více neslučitelnými jevy – s hypotézami – P(B)úplná pravděpodobnost=suma všech podmíněných pravděpodobností P(B/Hi)x apriorní pravděpodobnost P(Hi). Je to tedy vážený průměr podmíněných pravděpodobností, který je vážený apriorní pravděpodobností P(Hi).

- **Náhodná veličina**
  - číselně vyjádřený výsledek náhodného pokusu – např. jakou váhu má náhodně vybraný muž? Kolikrát padne šestka?
  - předem nevím, který výsledek vyjde, známe jen možné hodnoty a jejich pravděpodobnosti – ty musí mít dohromady součet 1
  - může mít diskrétní – vyjadřuje počet případů = četnost a pravděpodobnost těchto případů; nebo spojité rozdělení

- **Distribuční funkce**

můžeme pomocí ní popsat chování náhodné veličiny

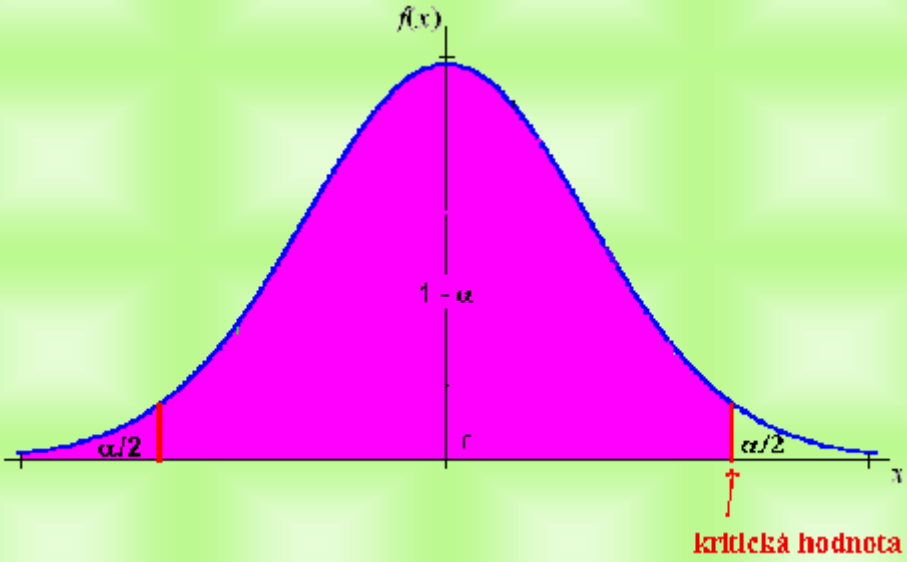
  - protějšek empirické distribuční funkce
  - distribuční funkce F(x) je nejúplnější popis pravděpodobnostního chování diskrétní nebo spojité náhodné proměnné X
  - $P(x1 < X \leq x2) = F(x2) - F(x1)$
  - Fx(x)=P(X<=x) – hodnota distribuční funkce v bodě x = pravděpodobnost, že náhodná veličina nepřekročí x
  - Když je náhodná veličina spojitá, pracujeme s hustotou jejího rozdělení (namísto pravděpodobnosti) – pravděpodobnost, že a<X<b je potom rovna velikosti plochy pod grafem na intervalu a až b, celá plocha pod grafem je rovna 1 a plocha od mínus nekonečno k x je rovna hodnotě distribuční funkce v bodě x (tedy Fx (x))



- **Střední hodnota mí náhodné veličiny X s diskrétním rozdělením (označení jako návod na použití – divné ale viz níže - E X)**
  - je to míra polohy, populační průměr (idealizovaný protějšek průměru)
  - počítá se z ní populační směrodatná odchylka a populační rozptyl

- je to jakési těžiště možných hodnot náhodné veličiny
  E X=suma všech hodnot vynásobených pravděpodobností vždy dané hodnoty; vážený průměr možných hodnot, kde vahami jsou pravděpodobnosti těchto hodnot
  Z toho odvodíme i populační medián – hodnota, pro kterou bude hodnota distribuční funkce Fx 1/2

- **Kvantily a kritické hodnoty**
  Kvantil (úplněji 100.p)% kvantil
  Zvolíme (podle toho, co chceme otestovat) si číslo p mezi 0 a 1 a pro dané číslo (náhodnou veličinu) X najdeme takové číslo x(p) - kvantil, pro které je distribuční funkce čísla X rovná p – pravděpodobnost, že je náhodná veličina menší nebo rovna kvantilu – taková konstanta, kterou s danou pravděpodobností náhodná veličina nepřekročí
  - kritická hodnota- taková konstanta, kterou s danou pravděpodobností náhodná veličina překročí – 90% kvantil=10% kritická hodnota



- **Definice nezávislosti dvou náhodných jevů**
- populační rozptyl - sigma^2 – to je značka, označení **var** je návod na to, že máme použít rozptyl, jeho odmocnina je směrodatná odchylka – splňují požadavky na míry variability stejně jako jejich nepopulační obdoby
  - pst jevu D nezávisí na tom, zda platí jev C
- P(D-průnik-C) = P(D) P(C)
  Náhodné veličiny X a Y jsou nezávislé, když jsou navzájem nezávislé všechny dvojice jevů A a B, kdy A je jev, který patří k X a B patří k Y - buď se jevy X a Y náhodou vyskytnou spolu = sdružené rozdělení – udává se pravděpodobností P=(X=xi, Y=yi) nebo každý zvlášť – s pravděpodobností P(X=xi) nebo P(Y=yj) – tomu se říká marginální rozdělení. Pokud se pravděpodobnost sdruženého rozdělení rovná násobku pravděpodobností marginálních rozdělení,

jsou náhodné veličiny nezávislé ☺ ☺ - tehdy můžeme rekonstruovat sdružené rozdělení z rozdělení marginálních. Pokud jsou náhodné veličiny závislé, už můžeme rekonstruovat jen marginální ze sdružených – to znám z chí kvadrát testu (viz níže) – ty uprostřed sdružené, když je posčítám na řádku/sloupci, vzniknou marginální

- **sdružené rozdělení** – popisuje společné chování veličin X a Y
- **marginální rozdělení** – chování jedné bez ohledu na druhou
- **podmíněné rozdělení** – popisuje chování Y při dané hodnotě X

- **Kovariance**
  - vyjadřuje vzájemnou závislost dvojice náhodných veličin

cov(X,Y)=střední hodnota ((xi-míx)(yi-míy))
platí, že cov (X,X)=var(X)
Jsou li náhodné veličiny X, Y nezávislé, pak je vždy cov (X,Y)=0 – kovariance je tedy ukazatelem možné závislosti!
Kovariance nám ale vůbec neukazuje těsnost nějakého vztahu, protože se její hodnoty mění s nastaveným měřítkem.. Pro vyjádření těsnosti vztahu se tedy používá populační korelační koeficient (je idealizovanou populační obdobou výběrového korelačního koeficientu – oba jsou bezrozměrné ukazatele), značí se ró.. (óooo, jaká podobnost s rozptylem!!)

Rozdělení

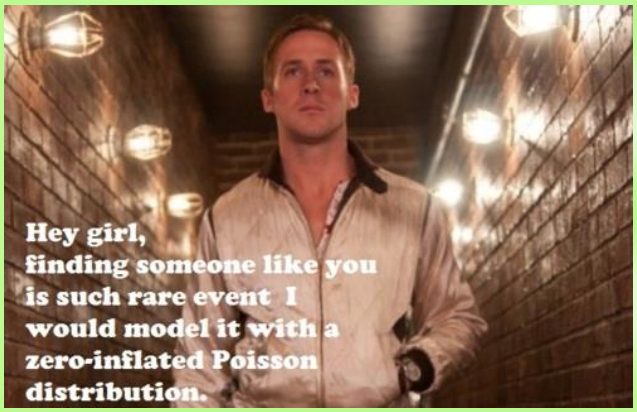
- **Hypergeometrické rozdělení**
  - Náhodná veličina Y, označující počet vybraných jednotek ze všech vykazujících sledovanou vlastnost, se řídí hypergeometrickým rozdělením. Např. 20 otázek na test, 5 z nich student umí (vykazují danou vlastnost), dostane 3, Y= jaká je pravděpodobnost, že si vytáhne alespoň jednu otázku, co umí?
  - je to rozdělení náhodné veličiny, kdy při opakování náhodného pokusu je výskyt sledovaného jevu závislý na výsledcích předchozích pokusů, jde tedy o pokusy, které jsou na sobě závislé
  - máme ryby v rybníku – je jich tam m, vylovíme jich a, ty označíme, pustíme a vylovíme n ryb. Nějaké z nich budou označené Y– bude jich více než 0, méně než n a méně než a, také musí být neoznačených méně než m-a, tzn Y=n-m+a – rozsah hodnot, jakých to může nabývat, potom s tímto rozsahem počítáme, kolika způsoby můžeme ryby vylovit – z toho vyplyne hypergeometrické rozdělení, z veličin a, m, n lze vypočítat i střední hodnotu a rozptyl Y, ze střední hodnoty pak vznikne odhad – m=na/Y

- **Binomické rozdělení**
  - zjišťujeme pouze výskyt či nevýskyt jevu B
  - pokusy jsou nezávislé (jejich počet n)

- pravděpodobnost výskytu B je v každém pokuse stejná pravděpodobnost 6 – 1/6, házíme 12x, počet padlých šestek je náhodná veličina Y, který nabývá hodnot od 0 do 12. Určí se pravděpodobnosti všech hodnot – tyto pravděpodobnosti charakterizují binomické rozdělení – Y má binomické rozdělení (Y=bi(n, pí)) – posloupnost n nezávislých pokusů, kdy každý může skončit zdarem nebo nezdarem. Náhodná veličina Y udává náhodný počet pokusů, v nichž nastal zdar.

- **Multinomické rozdělení**
  - zobecnění binomického rozdělení (multinomické rozdělení pro k=2 (2 možné situace) je binomické, stejně tak, když označíme jednu z marginálních pravděpodobností jedné situace jako zdar a zbytek jako nezdar, máme zase rozdělení binomické)
  - v dílčím pokusu máme *k* možných výsledků (jsou to jevy neslučitelné, dohromady dávají jev jistý)
    Máme n pokusů, v každém nastane jeden z k výsledků (A1-Ak), pravděpodobnosti toho, že nastane A1-Ak se v průběhu opakovaných pokusů nemění. Jako Y1-Yk označujeme počty pokusů, v kterých nastaly jevy A1-Ak. Hodnoty Y pak mají multinomické rozdělení. V každém případě nastane právě jeden z jevů A, proto: Y1+…+Yk=n. Y jsou závislé, když známe k-1 hodnot zjistíme všechny (z počtu n). Tzn. K-tice náhodných veličin Y má k-1 stupňů volnosti!
    -Velmi časté rozdělení – vždy, když třídíme četnosti znaku v nominálním měřítku – např. počty lidí s krevní skupinou
    -Má nějakou tajemnou spojitost s Chí-kvadrát rozdělením – kdo to dočte až dolů, tak mu to možná docvakne.

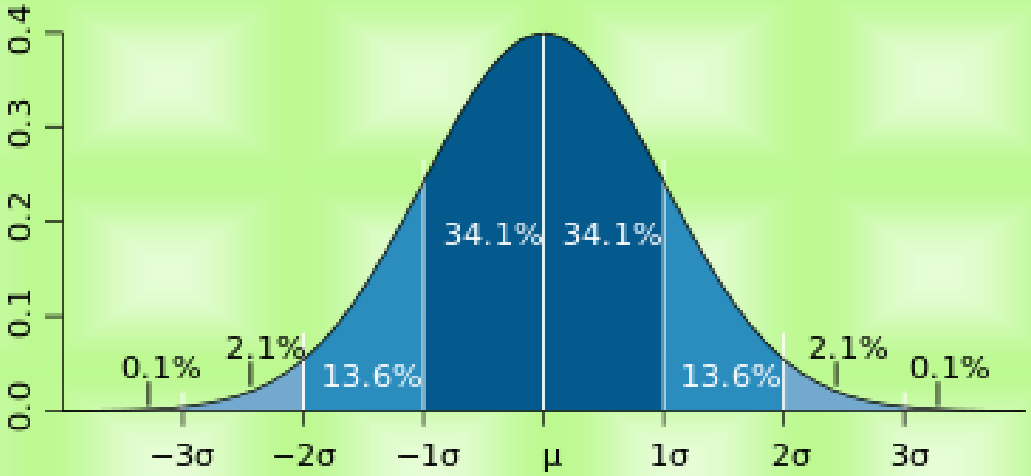
- **Poissonovo rozdělení**
  - =zákon řídkých jevů – aproximuje (přibližuje) nám binomické rozdělení, když jde o jevy, které se vyskytují jen zřídka – velký počet pokusů s malou pravděpodobností výskytu jevu



  - Střední hodnota i rozptyl náhodné veličiny s Poissonovým rozdělením jsou stejné
  - má ho náhodná veličina, která nabývá hodnot k = ,1,…(neurčeno), na rozdíl od binomického neznáme maximální hodnotu náhodné veličiny (u binomického to bylo n – počet členů) Y modeluje počet výskytu sledovaného jevu. Jevy zjišťujeme v oddělených časových úsecích, na různých plochách,... a musí být nezávislé. Parametr lambda, který je

s rozdělením spojený musí být úměrný velikosti úseku, na kterém jevy zkoumáme (např. plochy, času,..)

- **Normální rozdělení (Gaussovo)**
  - spojité rozdělení, symetrické okolo střední hodnoty



  - Graf – zvon, maximum pro x ve střední hodnotě, výška se rovná převrácené směrodatné odchylce – wooow
  - Plocha pod grafem je rozdělená na 6 částí, jejich základny jsou vždy dlouhé sigma, velikosti těchto částí jsou nezávislé na parametrech – procenta vždy stejná - zase wooow
  - Když veličinu násobíme nebo k ní něco připočítáváme, má stále normální rozdělení
  - Normované normální rozdělení – mí=1, sigma=0. Normovaná normální veličina Z z tohoto rozdělení je s 95% pravděpodobností v intervalu (-1,96;1,96) – vyplývá z výpočtu kvantilu z(0,975)=1,96
  - Normální rozdělení je velmi časté v důsledku centrální limitní věty (viz níže) a je prostě super.

-Náhodný výběr

Už byl náhodný jev a náhodný pokus, né že by se opakovalo pořád to samé dokola

Z x náhodných realizací veličiny X se náhodně vybraná n-tice nazývá náhodný výběr – např. při vážení téhož předmětu

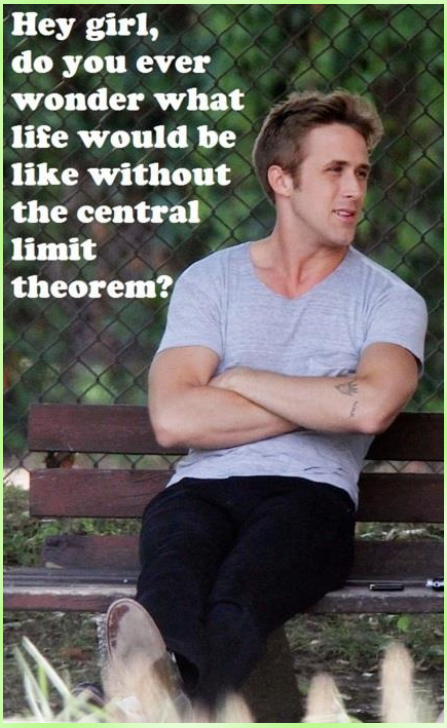
Střední hodnota mí z náhodně vybraných X zastupuje skutečnou hodnotu veličiny X, převrácená hodnota sigma (směrodatné odchylky) zase přesnost vážení. Mí v této situaci říkáme populační průměr (průměr X z náhodného výběru je roven parametru mí; je to tedy nestranný odhad parametru mí).

- **S.E.(X) Střední chyba průměru** *(To X má čárku nahoře – protože je to průměr a ne něco jiného..)*
- Směrodatná odchylka průměru v případě měření s náhodným výběrem z populace. Zjednodušeně řečeno je to číslo, které

označuje, jak moc se asi námi získaný průměr náhodného výběru liší od střední hodnoty základního souboru.

- S.E. (Xprům.) =sigma/odm z n
- Průměr tedy bude kolísat za použití náhodného výběru mnohem méně, než by kolísaly jednotlivé náhodné výběry
- Pod tímto pojmem se chápe i odhad střední chyby průměru, který se pak počítá S.E. (x prům.)= Směrodatná odchylka S/odm z n

- **Centrální limitní věta**
- s rostoucím n se rozdělení náhodné veličiny Z blíží k normovanému normálnímu rozdělení. To znamená, že průměr X má přibližně normální rozdělení
- máme hodnoty, co rozhodně nemají normální rozdělení, mají kladnou šikmost atd.. Když z hodnot náhodně vybereme tisíckrát nějakou a uděláme histogram, trochu se znormální. Když tisíckrát vybereme 10 hodnot, které zprůměrujeme a z těch tisíci průměrů uděláme histogram, bude ještě normálnější, když budeme průměrovat 100 hodnot, tak bude už úplně super.. (Rozptyl se bude neustále zmenšovat)
    - průměr má pro dost velká n normální rozdělení s rozptylem n-krát menším než jednotlivá pozorování, a to bez ohledu na výchozí rozdělení jednotlivých pozorování
    - to je často důvodem předpokladu o normálním rozdělení, výsledná hodnota je ovlivněna součtem velikého počtu nahodilých malých vlivů



STATISTICKÁ INDUKCE

- **95% interval spolehlivosti (konfidenční interval)**
- Rozmezí hodnot, ve kterých leží s 95% pravděpodobností **populační průměr**
- Máme výběrový průměr, vypočtený z X hodnot, ale neznáme směrodatnou odchylku sigma, takže si nemůžeme dopočítat z kvantilu normovaného normálního rozdělení, o kolik se hodnota našeho výběrového průměru s 95% pravděpodobností liší od skutečného průměru mí. Ale my si stejně poradíme, směrodatnou odchylku si odhadneme složitým vzorečkem – S=odm 1/n-1 x suma (odečtu každé konkrétní hodnoty od průměru) na druhou. Tímto jsme si všechno převedli z normálního rozdělení na studentovo, takže použijeme jeho kvantil Studentova rozdělení – pro z(0,975) je

to 2,064 (místo 1,960). Zvýšení kritické hodnoty je náš trest za to, že jsme neznali sigma a nahradili jsme ho jeho odhadem S.

- Teď už známe všechno, co potřebujeme, a tak konečně vypočítáme 95% interval spolehlivosti. Je to rozmezí, které s 95% pravděpodobností překryje skutečnou hodnotu mí. Intervaly spolehlivosti mohou být i jednostranné – v případě, že nás zajímá jen závislost jedním směrem – např. jestli jsou všichni větší než hodnota 1,5 metru..
- Co určuje velikost intervalu spolehlivosti (jeho délku)? Střední chyba průměru (S.E.X) – přesnost, s jakou výběrový průměr X odhaduje populační průměr mí. Dále pochopitelně požadovaná spolehlivost (1-alfa) a trošičku i počet stupnů volnosti (n-1)

- **Chyba prvního druhu** – zamítnutí nulové hypotézy, i když platí, tedy falešné potvrzení alternativní hypotézy

Padne nám ze 100 pokusů 21 krát šestka (což se může stát s 10% pravděpodobností) a my řekneme, že je kostka vadná. To ale dost riskujeme, že je kostka v pořádku (platí nulová hypotéza) a ty šestky padly náhodou. Když se chceme této chybě co nejvíce vyhnout, musíme si určit číslo, které náhodu připouští jen s (většinou) 5% pravděpodobností. Největší dovolené pravděpodobnosti chyby 1. druhu se říkáá…

- **…hladina testu - alfa (hladina pravděpodobnosti, významnosti)**
  - volí se před pokusem, nezávisle na jeho výsledku (ale dost svádí k tomu podvádět..)
  - bývá většinou 5% nebo 1%
  - definuje nám **kritický obor** – vypočítáme si hodnotu (kritickou hodnotu, k0), kde už začíná být pravděpodobnost, že se jednalo o náhodný úlet hodnot kamsi do dáli, menší než alfa. Všechny hodnoty vyšší než kritická nám spadnou do kritického oboru, a to je jako exploze, co zničí (zamítne) nulovou hypotézu, a my můžeme oslavovat ☺

  - Doplňkem kritického oboru je obor přijetí, když tam padnou hodnoty, H0 nezamítáme

- **Chyba druhého druhu** – nezamítnutí nulové hypotézy, i když neplatí – kostka je pošahaná, ale nám náhodou padne přesně v šestině pokusů šestka. Pravděpodobnost chyby II. druhu se říká beta
- pravděpodobnost této chyby se těžko ovládá – nejlepší způsob je pomocííí…
- **…síly testu - 1-beta**
  - pravděpodobnost zamítnutí neplatné nulové hypotézy
  - pst, s jakou prokážeme platnou „vědeckou hypotézu"

  Nejlepší způsob, jak zvýšit sílu testu, je přidat počet pozorování – je to jasné, čím více se na nějakou věc koukáme, tím větší je šance, že si na ní něčeho zajímavého všimneme.. Se zvýšením počtu pozorování (n) se také pochopitelně zvýší k0
  Síla testu nám přímo řekne, jakou procentuální šanci máme s našimi daty, a při naší hladině významnosti, na to, že nám test něco poví (že správně vyvrátíme nulovou hypotézu)

Tabulka 1. Platnost hypotéz o situaci v základním souboru a možná rozhodnutí na základě testování[1]

	Rozhodnutí	
Platí	H0	H1
H0	OK (P=1- α)	Chyba prvního druhu (P= α)
H1	Chyba druhého druhu (P= β)	OK (P= 1-β) Síla testu

- **Pojmy k populaci**
  - populace – všichni – prvky celé množiny, základní soubor
  - výběr – vzorek populace - podmnožina
  - reprezentativní výběr – podmnožina, pomocí které hodnotíme celý základní soubor, musí být základnímu souboru dostatečně podobná, musí věrně zachycovat poměry platné v celém souboru
  - parametr – neznámá konstanta, která nějakým způsobem určuje použitý pravděpodobnostní model.. aha..
  - odhad – statistika použitá k odhadu toho tajemného parametru – lepší definici nemám, ale dá se to vytušit z dalších kapitol..
  - statistika – je to vhodná funkce naměřených hodnot, která nám zpravidla pomáhá definovat kritický obor – např. výběrový průměr, chí kvadrát,..
  - populační průměr – mí (při normálním rozdělení) – je to průměr, který získáme měřením náhodného vzorku a bereme ho jako průměr celé populace

- **Dosažená hladina testu – p-hodnota**
  - nejmenší hladina, na které lze ještě z daných dat zamítnout nulovou hypotézu, pravděpodobnost, s jakou dostaneme výsledek, co bude stejně nebo více odporovat nulové hypotéze, p musí být větší než

hladina významnosti, abychom zamítli H0 – poslední část je fakt důležitá!

- **možností zkoumání závislosti dvojice znaků**
  - kvantitativní – kvantitativní
    - korelace, regrese
    - graficky – rozptylový (bodový) diagram
  - kvantitativní – kvalitativní
    - t-test, ANOVA
    - graficky – box-plot (krabicový diagram
  - kvalitativní – kvalitativní
    - chí-kvadrát test, Fischerův exaktní test
    - graficky – kontingenční tabulka

- **výběr testu**

rozdělení	normální	spojité
Jeden výběr	jednovýběrový t-test	jednovýběrový Wilcoxon
výběr dvojic	párový t-test	znaménkový, Wilcoxon
dva nezávislé výběry	dvouvýběrový t-tesT	Mann-Whitney (=dvouvýběrový Wilxocon)
k nezávislých výběrů	ANOVA	Kruskal-Wallis

- **neparametrické testy**

  - při neparametrickém testu testujeme jinou hypotézu o rozdělení základního souboru než je hypotéza o jeho parametru. Jejich řešení nezávisí na typu rozdělení základního souboru, takže na rozdíl od parametrických testů výsledky nejsou závislé na tom, zda jsme model rozdělení volili správně v souladu se skutečným rozdělením základního souboru.
  - Můžeme je pak použít i pro silně nenormální rozdělení, kdy parametrické testy předpokládající normální rozdělení selhávají. Jsou to ty testy, co mají v tabulce nahoře rozdělení spojité.

JEDEN VÝBĚR

- **T-test - jednovýběrový**
  - předpoklady – X1-Xn jsou nezávislé náhodné veličiny, mají stejné rozdělení (normální) s neznámými parametry mí a sigma na druhou (větší než nula)
  - srovnáváme mí a mí0, což je konstanta, která je daná (víme jí, zajímáme se o ni, porovnáváme s ní data); nulová hypotéza říká, že mí=mí0. K ní existují 3 alternativní: oboustranná alternativní hypotéza – mí nerovná se mí0 - a obě jednostranné (mí<mí0, mí>mí0) – ty ovšem volíme, jen pokud si to žádá náš výzkum nebo zkoumaná situace, ne když se nám k tomu hodí data – při oboustranném testu se počítá

s poloviční alfou, proto nás to svádí podvádět a použít jednostranné testy, kde je alfa celá. Jednostranný musí mít jeden směr, který vede od H0 k alternativě a žádný druhý směr není.
    - kritický obor dostaneme ze srovnání populačního průměru (který předpokládáme, mí0) a výběrového průměru, kterým se snažíme odhadnout skutečný populační průměr (T=X-mí0/S.E. (X)) – zajímá nás, jestli se liší – velikost rozdílu hodnotíme jeho vydělením S.E.(X) (tj. porovnáváme s přesností výběrového průměru). Vydělením vznikne bezrozměrná statistika T – záleží jen na rozsahu výběru n!!! Chytrá věc..
    - úpravou asi milionu vzorečků dojdeme k nezajímavé věci (skripta tvrdí opak), že interval spolehlivosti pro mí je množina všech středních hodnot mí0, které bychom nezamítli

Ověření předpokladů: tzv. robustním metodám nedodržení tolik nevadí, ale nemělo by se dít.
T - test má 2 předpoklady: nezávislost jednotlivých měření (např. když testuji výšku, neměla by být ve vzorku jednovaječná dvojčata) a normální rozdělení. To můžeme ověřit graficky pomocííí…

- **…pravděpodobnostního diagramu (probability plot, normal q-q plot)**
    - Grafické znázornění variační řady našeho pozorování proti hodnotám, které bychom ve stejném případě měli s normálním rozdělením = porovnání skutečných hodnot s jejich idealizovanými protějšky. Ideál je, aby byly všechny body na přímce se směrnicí sigma.

Test pro ověření normality je **Shapiro-Wilkův test**, dá se to dále udělat např. výběrovou šikmostí a špičatostí.

T-test – párový
Dvojice prvních a druhých měření (U, V), které jsou mezi sebou nezávislé; a rozdíly v obou měřeních (X=U-V) u všech hodnot mají normální rozdělení. Z nich testujeme hypotézu o shodě populačních průměrů U a V (míU=MíV). To je stejné, jako bychom testovali jednovýběrový t-test s mí=0, který se aplikuje na rozdíly U-V.

Znaménkový test

- lze jím prověřovat, zda jsou dva opačné jevy stejně pravděpodobné
- Používá se pro vyhodnocení párových pokusů v případech, kdy studovanou veličinu nemůžeme přesně měřit. V testu nepoužíváme žádné naměřené hodnoty, stačí nám rozhodnutí, zda pokusný zásah „A" zapůsobil více či méně než pokusný zásah „B". Pro svou jednoduchost se znaménkový test používá zejména k orientačnímu hodnocení předběžných pokusů, např. v mikrobiologii. Nepoužíváme hodnoty rozdílů, ale pouze jejich znaménka.
- Při platnosti H0 má jev X>0 pravděpodobnost ½

Párový Wilcoxonův test
Zdokonalení znaménkového testu – řešíme nejen, jaké je znaménko rozdílu, ale také vzdálenost rozdílu od nuly – ne přímo konkrétní číslo, jen pořadí hodnot absolutní hodnota U-V

Yatesova korekce
Využívá se pro malé hodnoty n – např. pár desítek. Čitatel se přiblíží o ½ k nule, spolehlivěji se tak dodrží zvolená hladina významnosti.

- řeší problém malých teoretických četností ve čtyřpolní tabulce pro test nezávislosti či homogenity (viz níže)

2 VÝBĚRY

- Srovnání 2 populací pomocí výběru z každé z nich, výběry nezávislé jak uvnitř populace, tak mezi populacemi
- H0 –oba populační výběry jsou stejné – založená na odhadu obou populačních průměrů

- **Dvouvýběrový t-test**
  - Předpoklady: všechny X a všechny Y nezávislé, s normálním rozdělením, shodné populační rozptyly sigma na druhou (to odhadneme pomocí váženého průměru výběrových rozptylů příšerným vzorečkem)
  - H0: míx=míy, rozhoduje se statistikou T=průměr x-průměr y(S.E.(průměr x-průměr y); stupně volnosti nx+ny-**2**
  -

- **Dvouvýběrový Wilcoxonův test (Mannův-Whitneyův)**
  pro 2 nezávislé výběry ze spojitého rozdělení, u kterých nepředpokládáme normální rozdělení=> porovnáváme 2 populace v kvantitativním znaku
  H0: rozdělení veličiny je v obou populacích stejné, populace se ve sledovaném znaku neliší=mají stejné populační mediány (rozdělení jsou shodná)
  Naměřené hodnoty nahradíme jejich pořadím, bez ohledu na to, jestli jsou v 1. Nebo 2. Populaci, za platnosti H0 budou pořadí podobná
  Sečteme všechny hodnoty pořadí u X a všechny u Y, jejich střední hodnoty budou dávat dohromady celý součet pořadí a to v poměru vůči velikostem souborů x a y – když platí H0

ANALÝZA ROZPTYLU

- **ANOVA**
  - analýza rozptylu **jednoduchého třídění**
  - Jako dvouvýběrový t-test, ale hodnotíme shodu středních hodnot několika nezávislých výběrů z normálního rozdělení, snaha prokázat závislost mezi znakem Y v kvantitativním měřítku a znakem A v nominálním – hodnoty A se dají zapsat čísly, ale mají slovní vyjádření a označují se jako faktor –

např. závislost mezi hodnotami cholesterolů u různých skupin – faktorem je skupina

- Máme *k* nezávislých výběrů z normálního rozdělení, *k* nám tedy říká, kolikrát máme soubor dat Y
- Předpoklad **homoskedasticity** – všechny skupiny mají stejný rozptyl, sigma na druhou 1=…=sigma na druhou *k*; opak je heteroskedasticita
- Pro k=2 je Anova stejná jako t-test, pro více k testujeme, jestli mí1=mí2=…=mí k, testujeme nestejnost většího počtu středních hodnot
- budeme to měřit jejich variabilitou (SS_A), tedy součet druhých mocnin jejich odchylek od celkového průměru (ten zjistíme jako vážený průměr průměrů jednotlivých výběrů), stupňů volnosti bude k-1
- problém je, že pokud H0 neplatí, tzn. střední hodnoty jsou rozdílné, zvětší to automaticky rozptyl. My ho pak nemůžeme správně odhadnout a dále s ním pracovat. Potřebujeme tedy jiný výpočet (odhad) rozptylu, kterému bude jedno, jestli jsou střední hodnoty ve shodě nebo ne. Takovým odhadem je reziduální rozptyl (je to vážený průměr výběrových rozptylů všech *k* výběrů - blablabla).
- Toto (to dole) je tabulka analýzy rozptylu, ve které jsou prý shrnuty všechny výpočty. Co z ní můžeme dobře odvodit je jev, nazývaný rozklad součtu čtverců. Tedy, že celkovou variabilitu SSt rozdělíme na 2 složky: SSa, což je variabilita daná různými středními hodnotami v různých výběrech; a SSe, což je reziduální variabilita – tedy variabilita uvnitř jednotlivých výběrů, kolísání kolem jejich jednotlivých průměrů.
- Z toho všeho vypočítáme příšernými vzorečky hodnotu statistiky F; čím je větší v porovnání s průměrným čtvercem (reziduálním rozptylem), tím máme větší šanci, že zamítneme H0. Je vzoreček na to, jak zjistit, kdy přesně už jsme překročili kritickou hodnotu a můžeme H0 zamítnout, ale lepší způsob je nechat si programem vypočítat p-hodnotu a pak se na ní podívat.

Variabilita	S	f	S/f	F	P
Výběry	S_A	$f_A=k-1$	S_A/f_A	F_A	p_A
Reziduální	S_E	$F_E=n-k$	S_E/f_E		
Celková	S_T	$f_T=n-1$			

S – součty čtverců
f – počet stupňů volnosti
S/f – průměrné čtverce
F – F-statistika
p – p-hodnota

- Máme asi milion metod pojmenovaných po milionech pánů, které nám pomůžou ověřit homoskedasticitu. Např. Bartlettův test, Levenův test

- Potom máme moc chytrou metodu pana Tukeya – Tukeyova metoda HSD, která nám poví, jaká byla příčina zamítnuté H0 (tedy které populační průměry jsou nestejné)

Kruskalův-Wallisův test
Tento test nám řeší stejný problém, jako řešil Wilcoxonův test v případě dvouvýběrového t-testu. Řeší tedy problém s nesplněním předpokladu o normálním rozdělení. Je to vlastně zobecnění Wilcoxonova testu.
Máme *k* nezávislých výběrů ze spojitého rozdělení (které mají všechny veličiny stejné, mohou se lišit jen posunutím). H0 říká, že jsou rozdělení v různých výběrech identická; tedy, že posunutí jsou nulová (tedy že jsou stejné populační průměry a mediány). Opět řešíme, jestli jsou průměrná pořadí ve výběrech podobná. Nepodobnosti průměrných pořadí měří statistika Q a H0 zamítáme podle chí kvadrát rozdělení.

Biostatistici rozhodně nezahálí, takže vymysleli i Dvojné třídění, kde se řeší nějaké interakce; když je to dvojné třídění, ale žádné interakce se neřeší, tak na to vymysleli Náhodné bloky. Když se to obojí chce počítat jako znaménkový test, jen na to Friedmanův test.

KORELACE A REGRESE

Korelace popisuje sílu vzájemné závislosti dvou veličin, regrese popisuje, jak vysvětlující veličina ovlivní veličinu vysvětlovanou

Výběrový (Pearsonův) korelační koeficient

- Populační korelační koeficient ró jsme řešili někde nahoře už dávno. Teď potřebujeme dostat jeho výběrový protějšek (jeho odhad, když počítáme s náhodným výběrem), který se značí r
- r=výběrová kovariance (S xy)/rozptyl X x rozptyl Y – nezapomenout, že z toho plyne, že je r bezrozměrné!
- Korelační koeficient nám ukazuje na lineární závislost dat, má hodnoty -1 až 1 a tyto hodnoty má v extrémním případě, kdy jsou všechny body na přímce
- Pokud je ró 0 znamená to, že jsou X a Y nezávislé. Ovšem protože je r jen odhad ró, vždy to alespoň trochu nenulové vyjde, i když jsou veličiny nezávislé. Musíme proto otestovat H0: veličiny X a Y (s normálním rozdělením) jsou nezávislé, tedy r je blízké nule. To prokážeme T statistikou.

- **Spearmanův korelační koeficient**

Hodnotí závislost spojitých znaků, když není dodrženo normální rozdělení. Místo hodnot znaků tedy hodnotíme pořadí těchto znaků. Místo Pearsonova r počítáme Spearmanovo r_S.
Výhody: citlivost na jakoukoliv závislost, nejen lineární; menší citlivost k výskytu odlehlých hodnot.

- **Regrese**
  Slouží k vysvětlení variability (kolísání) odezvy (závislé, vysvětlované proměnné, Y) na regresor (vysvětlující, nezávislá proměnná, x). Jednosměrná závislost Y na x. Velikost písmen x a Y je schválně jiná, aby nám došlo, že jsou jako opravdu zásadně odlišné.
  Pomůže nám: prokázat závislost odezvy (její střední hodnoty) na nezávislé proměnné, předpovídat (hmm.. to zní dobře, ale spíš odhadovat) střední hodnotu vysvětlované proměnné podle hodnot nezávislé proměnné.
  x známe přesně, není náhodná, Y náhodná je a skládá se ze dvou složek: deterministická složka je ta, co jí ji vnutí její závislost na x; a náhodná složka, která je daná vlastní vnitřní variabilitou Y. Např. výška synů je závislá na výšce otců (deterministická), ale i stejně vysocí otcové mají různě vysoké syny (stochastická složka).

Regresní přímka
Pro jednu spojitou nezávisle proměnnou.
Střední hodnoty nezávislé náhodné veličiny Y leží na přímce y=beta0 (absolutní člen=posunutí)+beta1 krát x (beta jedna je směrnice). Je to normální směrnicový tvar přímky se všemi pravidly jako nás učili na střední. Je dobré si všimnout, že na rozdíl od náhodného výběru tu nepředpokládáme, že budou všechny střední hodnoty stejné, chceme po nich jenom to, aby leželi na té své přímce. Pochopitelně stejně tak jako neznáme rozptyl, neznáme ani hodnoty regresních koeficientů beta 0 a beta 1. Beta 1 je extrémně důležitá hlavně proto, že nám ukazuje citlivost, s jakou reaguje střední hodnota Y na střední hodnotu x.
Y=beta0+beta1 krát x+E
beta0+beta1 krát x je vlastní závislost, E je modelem nevysvětlitelná variabilita (něco jako stochastická složka)
Neznámé regresní koeficienty odhadujeme **metodou nejmenších čtverců.** Je to fakt chytrá metoda, kdy máme body (ty značí naše měření, mají tedy souřadnice x a Y) a ty body posouváme po ose Y na regresní přímku. Délka posunu bodu se rovná délce jedné stany imaginárního čtverce (říkáme jí reziduum). My naší regresní přímku vedeme tak, aby byl součet ploch všech takto vzniklých čtverců (reziduální součet čtverců) co nejmenší.
Z reziduálního součtu čtverců můžeme vypočítat reziduální rozptyl, což je odhad rozptylu sigma na druhou.
Je sice hezké, že jsme si odhadli, jaká je závislost Y na x, ale ještě jsme si tu závislost ani neprokázali. To si prokážeme T statistikou, kdy H0: beta1=0 (ze střední odvodíme, že by pak přímka byla rovnoběžná s x a konstantní).
Když zvolíme beta1=0, úplně tím vlastně umažeme závislost Y na x. I pro tuto přímku si můžeme spočítat součet čtverců – tentokrát Celkový součet čtverců – a ten nám poví, jaká je variabilita hodnot závisle proměnné.
Když od Celkového součtu čtverců odečteme Reziduální součet,

dostaneme Regresní součet čtverců. Ten nám okomentuje variabilitu Y, kterou vysvětlíme vyšetřovanou závislostí Y na x. Všechny tyto dílčí úvahy nám hezky shrne…

… koeficient determinace

R^2=regresní součet čtverců/celkový (nebo 1 - reziduální/celkový)

Koeficient determinace nám tedy říká, jakou část variability závisle proměnné vysvětlíme díky její závislosti na x.

KONTINGENČNÍ TABULKY

Hodnotí četnosti nominálních hodnot

- **Chí-kvadrát test dobré shody**
  - při testování kvalitativní – kvalitativní, např. krevní skupiny a jejich četnosti
  - Aby byl výběr reprezentativní, musí být empirické četnosti shodné s % podílem v populaci

Očekávané (teoretické) četnosti – spočítáme je např. pomocí různých pravděpodobností, kombinatorikou a podobnými hrůzami

Empirické četnosti – experimentálně zjištěné, spočítáme je např. na prstech

H0 určí očekávané četnosti, říká, že očekávané jsou shodné s empirickými četnostmi

  - X^2 = suma (exper.četnost – teor.četnost)^2 / teor.četnost

Pokud vyjde chí kvadrát statistika velká a překročí kritickou hodnotu pro danou hladinu pravděpodobnosti a stupně volnosti, znamená to, že H0 zamítáme a výběr považujeme za nereprezentativní.. Krásné, elegantní, jednoduché

- **Kontingenční tabulka**

  Testuje nezávislost dvou znaků v nominálním měřítku.

  Zapíšeme četnosti kombinací vždy dvou znaků (sdružené četnosti), jejich součtem doplníme marginální četnosti v řádcích a sloupcích. V sloupcích jsou pak četnosti druhého znaku bez ohledu na první a v řádcích naopak. V rohu je celkový počet všech dvojic hodnot – n. Hodnoty prvního znaku – veličina X, hodnoty 1,..,r, druhého znaku – Y, hodnoty 1,…,s.

  - Sdružené četnosti se značí jako n_{RS}- tam se napíšou konkrétní čísla, o četnost na jakém místě zrovna půjde (např. B je n_{12}) a marginální četnosti se označí tak, že to, co sčítáme je číslo a to druhé + (např. b+d dole je $_{+2}$). Speciální případ pro r=c=2 je čtyřpolní tabulka:

a	B	a+b
c	D	c+d
a+c	b+d	N

Když testujeme hypotézu nezávislosti, normálně chí - kvadrát testem srovnáme očekávané a empirické četnosti. Výpočet jak očekávaných četností, tak testová statistika, jsou formálně shodné s chí - kvadrát testem nezávislosti, ale v tomto případě se mu říká **test homogenity**. Zajímá nás, zda jsou pravděpodobnosti jednotlivých možných výsledků dílčího pokusu ve všech populacích stejné.
Je to obdoba ANOVY, v případě testování homogenity dvou multinomických rozdělení dvouvýběrového t-testu (přestože se to počítá jako chí-kvadrát test).

McNemarův test

- obdoba testu párového (ale teď pro nominální hodnoty)

Porovnává výsledky dvojích měření na stejných objektech. Měření se musí týkat nominálního znaku. Vše je zapsáno v kontingenční tabulce (nutně čtvercové) a H0 říká, že pravděpodobnosti multinomického rozdělení popisujícího četnosti výsledků v prvním a ve druhém pokusu jsou stejné.