AAV – přednášející Mgr. Patrik Galeta

2. října 2008 - 2. přednáška + cvičení
- test - vypisovací odpovědi, pokaždé v odpovědi bude postup a výsledek, pokud budu mít jen jedno z toho, potom dostanu jen část bodů

- čím dál tím víc se ve společenských disciplínách uplatňují kvantitativní metody
- cílem kurzu je dovědět se něco o popisu dat (souhrnné informace, přehledná prezentace) a statistické inferenci (jak zobecnit informace z výběrových šetření na celou populaci)

GAISE
- důraz na statistickou gramotnost a myšlení
- používání reálných dat
- pochopení konceptu, ne znalost postupu výpočtu
- aktivní učení na výuce
- užití nových technologií

Kvantitativní a kvalitativní výzkumy

Kvantitativní výzkum
- testování hypotéz
- teorie → hypotézy → vydedukování závěrů → výběr proměnných → sběr dat → testování hypotéz → přijetí či zamítnutí hypotéz → výběr závěru podle výsledku testování
- deduktivní
- několik proměnných na velkém množství jedinců
- sbírají se jen ta data, která potřebujeme
- můžeme zde zobecnit problémy

Kvalitativní výzkum
- porozumění, vytváření teorie
- pozorování → sběr dat → analýza → nalezení pravidelnosti mezi daty → vytvoření generalizujících výroků, předběžných závěrů → konfrontace s dalšími daty → vytvoření dalších předběžných závěrů → v bodě nasycenosti systému (kdy nová data nepřinášejí žádnou novou informaci) vytvoření teorie
- induktivní
- mnoho proměnných na několika jedincích
- sbírají se úplně všechna data
- nemůžeme zde zobecnit výsledky

Základní pojmy

Jednotka výzkumu
- objekt, na kterém provádíme měření
- např. člověk, rodina, třída, město, billboard, jazyk

Proměnná
- vlastnost jednotky výzkumu
- má více než jednu hodnotu
- např. u člověka je to národnost, pohlaví nebo třeba věk, u billboardu barevnost nebo velikost písma
- hodnoty proměnné - u proměnné pohlaví to je muž či žena, u proměnné věk potom škála
- proměnné měříme

a) nominální (kategoriální)
- *hodnoty nelze seřadit podle velikosti*
- *rozdíl mezi přilehlými hodnotami není konstantní*
- *poměry hodnot a nula nemají smysl*
- např. barva očí, bydliště, živočišný druh, obor studia

- ty nominální hodnoty, které nabývají jen dvou hodnot (muž-žena, ano-ne), nazýváme **<u>binární proměnné</u>**

b) ordinální (pořadové)
- *lze je seřadit podle velikosti*
- *rozdíl mezi přilehlými hodnotami není konstantní*
- *poměry hodnot a nula nemají smysl*
- např. malý-střední-vysoký, dětství-dospívání-dospělost-stáří, výborně-velmi dobře-dobře-vyhověl-nevyhověl
- mezi těmi jednotlivými hodnotami ale nesmí být stejný rozdíl

c) intervalové (rozdílové)
- lze je seřadit podle velikosti
- existuje konstantní rozdíl mezi přilehlými hodnotami
- poměry hodnot a nula nemají smysl
- např. IQ, stupně teploty, datum
- mezi jednotlivými hodnotami musí být počitatelné stupně

d) poměrové (podílové)
- lze je seřadit podle velikosti
- existuje konstantní rozdíl mezi přilehlými hodnotami
- poměry a hodnot a nula mají smysl
- např. výška postavy v centimetrech, věk v číslech, barva očí v pixelech nebo RGB
- dá se s nimi dělat nejvíc matematických operací, jsou nejzajímavější

- z poměrové proměnné lze získat všechny ostatní, ale neplatí to obráceně - z nominální proměnné neudělám třeba proměnnou intervalovou (protože mi na to chybí potřebné informace)
- nominální a ordinální jsou **<u>kvalitativní proměnné</u>** a intervalové a poměrové jsou **<u>kvantitativní proměnné</u>**
- kvantitativní proměnné se často ještě rozdělují na **<u>diskrétní</u>** a **<u>spojité</u>** - diskrétní mohou nabývat jen celočíselných hodnot, spojité jsou ty ostatní (třeba ta čísla mezi jedničkou a dvojkou) - diskrétní je třeba počet dětí, počet aut na ulici, počet obyvatel ve městě, spojité jsou třeba věk, výška, výdělek

kvalitativní → nominální	kvantitativní → intervalové	kvantitativní → diskrétní
↓	↓	↓
ordinální	poměrové	spojité

<u>Tabulka původních dat</u>

- přepsaná data z protokolů
- obsahují úplně všechny sebrané informace
- do prvního řádku se píšou názvy proměnných a do prvního sloupce kódy jedinců (pořadová čísla)
- do řádku se zapisuje jedinec (v podobě kódu nebo pořadového čísla) a do sloupce jeho proměnné (věk, počet dětí, stupeň dosaženého vzdělání,…)
- na co si dát pozor - seřadit jednotky podle nějakého klíče (abeceda, inventární číslo)
- proměnné se dají zapisovat slovy, číselnými kódy - u počtů je to jasné (věk, počet dětí), ostatní hodnoty si musíme do čísel převést (např. u vzdělání bude základní = 1, střední = 2, vysokoškolské = 3)
- při kódování hodnot nominálních proměnných s nimi pořád musím pracovat jako s nominálními hodnotami (pracovat s čísly jako se slovy), to samé platí u ordinálních proměnných
- kódování hodnot diskrétních proměnných se vždy píše bez desetinného místa (tedy 1 a ne 1,0)
- kódování hodnot spojitých proměnných se zase vždy píše s desetinnými místy a ta musí mít pokaždé stejný počet (tedy 23,4 35,0 28,2 a ne 23,423 35 28,24)

9. října 2008 - 3. přednáška + cvičení

Popisná statistika - tabulkové a grafické souhrny

- jak zjednodušit množství informací z původních tabulek
- souhrn dat, přehledná informace o struktuře dat, ztráta informace

a) tabulkové souhrny - četnostní tabulka
b) grafické souhrny - sloupcový graf, histogram
c) číselné souhrny - míry polohy a variability - viz přednáška

- tabulky, grafy a číselné souhrny se odlišují podle typu proměnné

Nominální proměnné

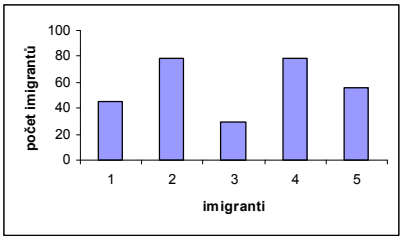
a) četnostní tabulka
- tři sloupce - v prvním popisky, v druhém absolutní četnost a ve třetím relativní četnost
- v posledním řádku vždy součet oněch hodnot

věk	absolutní četnost	relativní četnost
nedospělí	45	0,45
dospělí	56	0,55
celkem	101	1,00

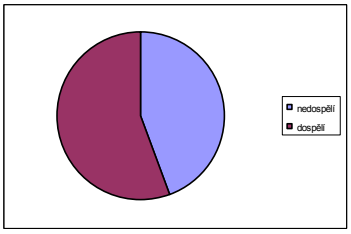
- **absolutní četnost** - např. nedospělých je celkem 45
- **relativní četnost** - procentuální vyjádření (absolutní četnost : celkem x 100 = relativní četnost), s tím, že se to píše s desetinnou čárkou, tedy nenapíšu 45%, ale 0,45; celkem musí vždycky relativní četnost vyjít 1,00

b) sloupcový graf
- jde do něj převést četnostní tabulka
- v grafu se musí doplnit popisky os
- vynechat čáry v pozadí a graf udělat co nejjednodušší černobílý
- sloupky řadit podle logického klíče, jako je to v četnostní tabulce (neřadit sloupce podle velikosti)
- vybrat vhodně podrobnou škálu (např. pokud jsou hodnoty do 800, nebudu dělat škálu až do 2000, udělám ji kratší, ale podrobnější)
- zkrátka udělat graf podle tabulky, jak jsme se učili na gymplu
- může mít více variant - může být vertikální nebo horizontální (podle toho, aby to dobře vypadalo na stránce a taky se to na ní vešlo, pokud je třeba hodně moc proměnných, potom je lepší horizontální položení)
- variantou může být také koláčový graf - výhodou je, že se hned vidí, jaká je to část z celku, nevýhodou je relativní nepřehlednost

sloupcový graf



koláčový graf

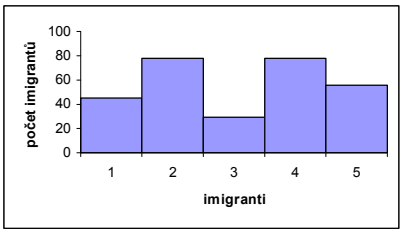


Diskrétní proměnné
a) četnostní tabulka
b) sloupcový graf

Spojité proměnné
a) četnostní tabulka
b) histogram
- sloupcový graf, který nemá oddělené sloupce, není mezi jednotlivými sloupci mezera

- u histogramu nesmí mít sloupce různé barvy, u sloupcového grafu mohou, ale je lepší, když jsou všechny sloupce stejné barvy

histogram



Rozdělení proměnné
- způsob vyjádření proměnné - někdy tabulkou, někdy grafem

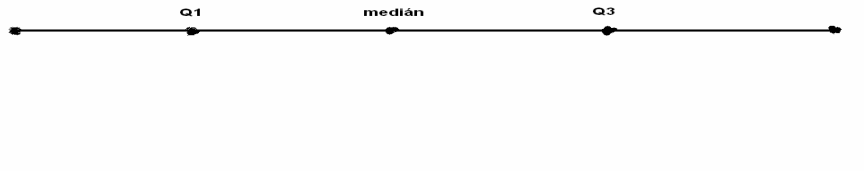
Číselné souhrny
- cílem je vybrat jednoho zástupce z množiny zástupců, který by byl průměrný

a) velikost souboru
- počet jedinců v souboru
- označuje se velkým, nebo malým N
- kolik jedinců zastupuje vybraný vzorek

b) míry polohy
- jaká je typická hodnota v souboru?
- jakého řádu jsou hodnoty?
- reprezentativní hodnota pro soubor
- nepopisuje tvar rozdělení

$$\bar{y} = \frac{y1 + y2 + y3}{n}$$

- **aritmetický průměr** - nejčastěji (y je značka pro jedince, y s čarou je průměrný jedinec)
- **medián** - hodnota proměnné, pro kterou platí, že polovina hodnot proměnné v souboru je větších a polovina menších, pokud je počet hodnot sudá, vypočítá se průměr z dvou středních hodnot, u nominálních proměnných medián počítat nelze (nejde to ani u ordinální proměnné)
- **odlehlá hodnota** - hodnota, která je daleko od mediánu
- medián není průměr (viz příklad průměrný plat v ČR - průměrný plat v ČR je dnes asi 21 000 Kč, ale medián je asi od 2 3 tisíce menší, protože je daleko více lidí, kteří berou menší plat, než lidí, kteří berou plat nadprůměrný)
- **kvartily** - rozděluji rozdělení hodnoty proměnné na čtvrtiny (spodní Q1, střední Q2 a horní kvartil Q3 - střední kvartil = medián), kvartily se zase nedají počítat u nominálních a ordinálních proměnných (jen u těch proměnných, které se dají seřadit)



- **kvantily** - obecně rozdělují hodnoty na nějaké části (decily na 10 částí, percentily na 100 částí), je to hodnota, pod níž leží definovaná část údajů (pokud budu chtít vědět 40. percentil, tak to bude ta hodnota, která má pod sebou 40% a nad sebou 60% hodnot)

• **modus** - nejčastější hodnota proměnné v souboru, používá se především u nominálních a ordinálních proměnných, ale lze jej vypočítat i u intervalových a poměrových proměnných (v naší třídě je nejčastější modus žena)

c) míry rozptylu

- popisují variabilitu hodnot
- **rozsah** - rozdíl mezi největší a nejmenší hodnotou v souboru, rozsah = max - min
- **mezikvartilová rozpětí** - odečtení třetího a prvního kvartilu, mezikvartilové rozpětí = Q3 - Q1, není ovlivněn odlehlou hodnotou, protože nepočítá s extrémy
- **rozptyl** - průměrná hodnota druhé mocniny odchylky hodnot proměnné od průměru proměnné

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

- **směrodatná odchylka** - SD, druhá odmocnina rozptylu, označuje se jako s, čím větší odchylka, tím je větší variabilita, je ovlivněna odlehlou hodnotou

16. října 2008 - 4. přednáška + cvičení

Diskrétní proměnné a binomické rozdělení

- např. Aboriové mají takový zvyk, že ženich musí zaplatit obrovskou sumu rodině nevěsty, za to, že ji odvádí z domu, proto je výhodnější, aby se v rodině narodila holčička, než chlapeček → preferují Aboriové při porodu holčičky?

- objekt - rodina (rodina s pěti dětmi, budeme brát v úvahu 100 rodin)
- proměnná - počet dívek nebo chlapců (budeme měřit počet dívek - min. 0 dívek - tedy 5 chlapců, max. 5 dívek v rodině - tedy 0 chlapců v rodině)

→ uděláme tabulku, která bude mít 100 řádků (pro každou rodinu) a jeden sloupec (pro počet dívek v té rodině)

rodina č.	počet F
1	1
2	4
3	3
4	5

→ potom uděláme četnostní tabulku a sloupcový graf

- jenže z grafu nepoznáme, jestli výsledky jsou přirozené nebo je to ta preference → mohli bychom výsledky porovnat s výsledky u populace, u které neexistuje žádná preference, nebo tyto neutrální výsledky teoreticky odhadnout, pokud existuje nějaký teoretický odhad…

- existuje - říká se tomu binomické rozdělení

Binomické rozdělení

- teoretické rozdělení proměnné
- je vždy souměrné (jako Gausova křivka)
- toto rozdělení lze použít pro binomické proměnné (ano-ne, pohlaví)
- předpoklady binomického rozdělení - jevy jsou na sobě nezávislé, jevy jsou vyčerpávající, P(F) = 0,5 (P je pravděpodobnost toho, co je v závorce, F jako female → pravděpodobnost, že se narodí dívka je 0,5)
- čím jsou způsobeny odlišnosti? - nedodržením předpokladů (jevy nejsou na sobě nezávislé, jevy nejsou vyčerpávající, vzorek nebyl náhodný, pravděpodobnost je jiná) nebo náhodou (kvantitativní výzkumy s ní pracují, ale kvalitativní ne)
- jak zjistit, jestli jsou rozdíly způsobené náhodně nebo preferencí? - testováním hypotéz

30. října 2008 - 5. přednáška + cvičení

- funkce v excelu
  - směrodatná odchylka - smodch.výběr, nebo =odmocnina(souřadnice)
  - rozptyl - var.výběr

Koeficient variace
- cv, v
- míra variability

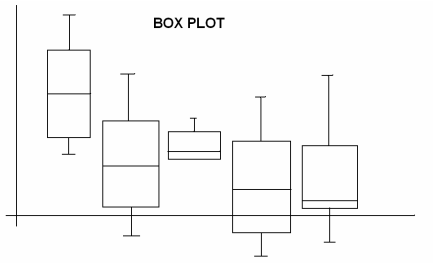
$$CV = \frac{SD \times 100}{\bar{y}}$$

- koeficient variace je ovlivněn pouze variabilitou

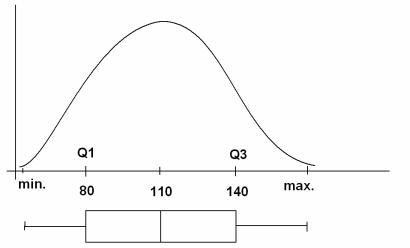
	kg	g
Alechandro	226	226 000
Benjamin	10	10 000
Daniel	78	78 000
Gullermo	46	46 000
Humberto	145	145 000
Jaime	185	185 000
průměr	115	115 000
SD	84	83 995
CV	73,04	73,04

Aplety
- grafické znázornění pro internet, nejčastěji v javascriptu

Box plot
- alternativa histogramu pro spojitou proměnnou



- pro rozložení IQ ve společnosti bude box plot vypadat následovně:
  a) nakreslíme si osy, na nich graf (v tomto případě Gausovu křivku, protože rozložení IQ ve společnosti je rovnoměrné)
  b) určíme střední hodnotu (110 - takové IQ má 50% populace), dolní (80 - podle nás 25% populace nemá IQ ani 80) a horní (140 - tady podle nás má 25% populace IQ vyšší než 140, HAHA) kvartil
  c) nakreslíme box plot, který bude mít střední čárku v úrovni 110, krajní strany budou v úrovni 80 a 140, fousky budou sahat k minimu a k maximu → pokud jsme to dělali na Gausově křivce, bude potom box plot souměrný



- výhody box plotu oproti histogramu - do jedné osové soustavy se vejde víc box plotů (tedy více histogramů), u histogramů bychom potřebovali pro každý histogram jednu osovou soustavu, z box plotu vyčteme více informací než z histogramů

Příklad 01:
Sisal se vyrábí z agáve. V 50. a 60. letech se brazilští Aboriové dostali do problémů, protože prales se začal vytěžovat a těžaři se snažili Aborie vystrnadit. Aboriové nakonec přesídlili na náhorní plošinu, což je savana s ostrůvky lesa. Aboriové nemohli sbírat potravu jako v pralese, proto opustili lov a sběr a dali se na pěstování (hlavně kukuřice a fazolí). Mimo to začali také pěstovat agáve a hodně rodin se vzdalo pěstování tradičních plodin kvůli agáve. Ale na výrobu sisalu z agáve je třeba velkých a drahých strojů. Proto se agáve vozilo do center, kde ty stroje byly a tím začala strukturalizace moci (mezi těmi, co pěstovali agáve, a těmi, co ho zpracovávali, ale také mezi muži a ženami, kdy ženy nemohly obsluhovat stroje). Často ale docházelo k nehodám, protože obsluhovat tyto stroje bylo velice těžké a nebezpečné, potom už muži nemohli tyto stroje obsluhovat (třeba když přišli o ruce) a rodiny se dostali do bídy. Někteří Aboriové ovšem zůstali u pěstování tradičních plodin (kukuřice a fazole) a neupadali tolik do chudoby.

Otázka:
Jsou na tom z hlediska chudoby lépe Aboriové, kteří pěstují tradiční plodiny, než ti, co přešli na pěstování agáve?

A toto jsou výsledky výzkumu:

	trad. plod.	agáve
dospělí	3620 kal	3941 kal
děti	2868 kal	2905 kal

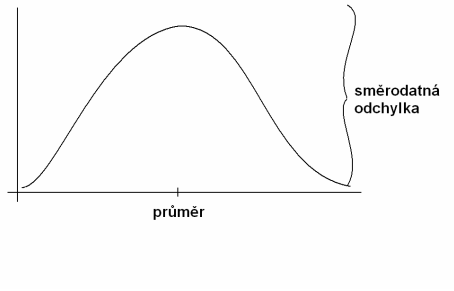
U těch, co pěstují tradiční plodiny, je u dospělých průměrný příjem 3620 kal, u dětí 2868 kal. U těch, co pěstují agáve, je průměrný příjem u dospělých 3941 kal a u dětí 2905 kal.

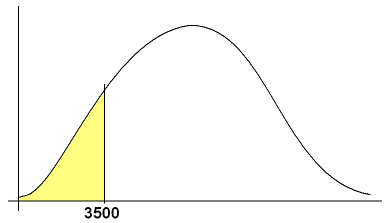
Kolik procent populace dospělých má denní příjem nižší než je norma OSN 3500 kal?

→ normální rozdělení

Normální rozdělení

- symetrické (histogram)
- je pouze pro spojité proměnné
- má pouze dva parametry - průměr a směrodatnou odchylku
- šířka histogramu určuje variabilitu (směrodatnou odchylku), nejvyšší bod grafu bude průměr





- jak vypočítat obsah pod křivkou?
- obsah pod křivkou je přímoúměrný četnosti hodnot
- budeme k tomu potřebovat **standardizované normální rozdělení**
  průměr = 0
  SD = 1

Když se vrátíme k původnímu příkladu se sisalem, kdy 3500 kal je norma OSN a kde se ptáme, kolik lidí je pod touto normou → kolikátý kvantil je hodnota 3500? → použijeme **tabulku kvantilů standardizovaného normálního rozdělení** (stáhnout z netu), kdy 0 = 50%

Standardizace

- každé normální rozdělení lze standardizací převést na standardizované normální rozdělení

$$z_i = \frac{y_i - \bar{y}}{SD}$$

y_i - zadaná hodnota, která nám vytyčuje hranici (v našem případě to je 3500, protože se ptáme, kolik lidí má menší kalorický příjem než je tato hodnota)

Pokud tento vzorec aplikujeme na náš příklad, potom tedy $\mathbf{z_i = 3500}$ (y_i, norma OSN) **- 3941** (což je průměrný kalorický příjem na jedince) a to vše **vydělím 310** (spočítaná směrodatná odchylka) → vyjde nám číslo **1,4225** → kouknu do tabulky kvantilů standardizovaného rozdělení a tam zjistím, že **hodnota 1,42 odpovídá číslu 0,078**, což převedeno na procenta je **7,8%** → to všechno znamená, že **7,8% lidí populace Aboriů, kteří se živí pěstováním agáve, má nižší kalorický příjem než je norma OSN.**

- toto všechno jde ale použít jen pro normální rozdělení, které je symetrické!!!

Jak zjistím, jestli je dané rozdělení normální?

a) kontrola histogramu (pohledem poznám, jestli je to asi symetrické)
b) srovnání mediánu a průměru (medián musí být roven průměru)
c) výpočet šikmosti a špičatosti
d) testování

Šikmost

- g_1
- číslo, které se dá vypočítat
- u normálního rozdělení je $g_1 = 0$
- zešikmené rozdělení doprava, když g_1 je větší než 0, zešikmené rozdělení doleva, když g_1 je menší než 0

Špičatost

- g_2
- zase číslo, které charakterizuje tvar rozdělení
- u normálního rozdělení má hodnotu 0

- **laptokurtické** rozdělení (spíše špičaté), kdy g_2 je větší než 0, **platykurtické** rozdělení (spíš zploštělé), kdy g_2 je menší než 0

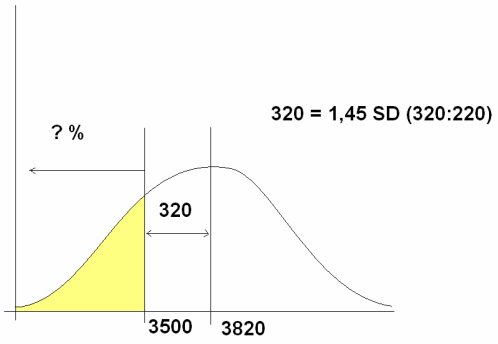
13. listopadu 2008 - 6. přednáška + cvičení

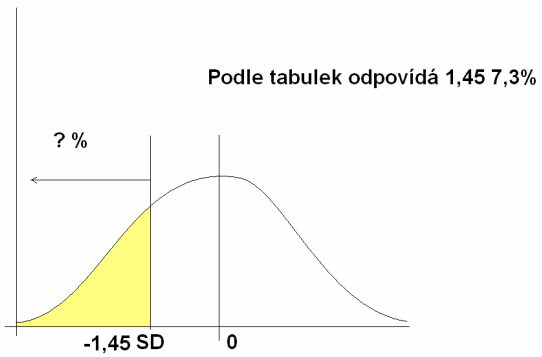
- **hodnota kvantilu** - daná hodnota, čárka na ose, kde je ten kvantil
- **hladina kvantilu** - to, co je menší než hodnota kvantilu, všechno, co je pod kvantilem

---

Příklad 02:
Kolik dospělých Aboriů má nedostatečný kalorický příjem?

dospělí	trad. plod.	sisal
průměr	3820	3941
SD	220	310
N	120	122
norma	3500	3500





Odpověď:
7,3% Aboriů, kteří pěstují tradiční plodiny, má nedostatečný kalorický příjem (počítali jsme to pro ty, co pěstují tradiční plodiny).

Jak se to počítalo?
Nakreslím si graf normálního rozdělení, kdy uprostřed vynesu průměrnou hodnotu (v našem případě průměrný příjem kalorií na jednoho Aborie) → Doleva potom vynesu hodnotu normy (v našem případě je norma daná OSN) → chci vypočítat, co je pod tou normou → zjistím rozdíl mezi průměrem a normou (odečtu normu od průměru) → tento rozdíl vydělím SD (směrodatnou

odchylkou) a tak získám hodnotu, kterou si najdu v tabulce kvantilů standardizovaného normálního rozdělení → tato hodnota podle tabulky odpovídá nějakým procentům → a tyto procenta jsou náš výsledek! (prakticky ten hnusný vzoreček vůbec nepotřebuju)

Testy SAT a ACT
- jsou to standardizované testy, které vyplňují studenti, kteří se hlásí v USA na VŠ

Příklad 03:
Čísla pro ACT: Průměr - 18 bodů, SD - 6 bodů, dosažené maximum - 36 bodů → kolika bodů dosáhne 30% nejlepších studentů?

Výpočet:
Tady budeme postupovat obráceně
1) nakreslit graf normálního rozdělení (uprostřed bude 18, napravo **x** - to je tentokrát náš výsledek)
2) od toho x nalevo je 70%, od x napravo je 30%
3) v tabulce najdu kvantil, který odpovídá hodnotě 30%
4) zjištěný kvantil vynásobím SD
5) tím vypočítám rozdíl mezi průměrem (18) a hledaným x
6) x jednoduše zjistím tak, že sečtu průměr (18) a ten rozdíl (mělo by vyjít 21,… bodů)

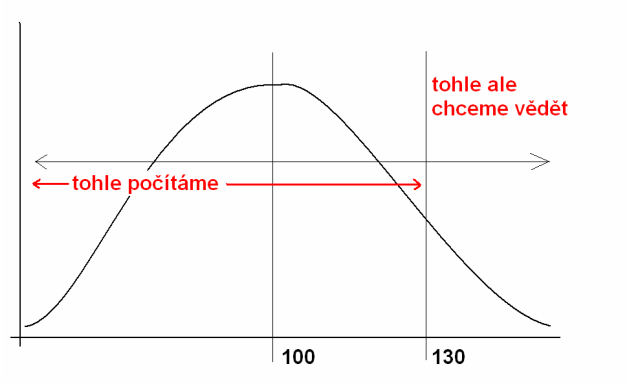
Pokud chceme porovnat dva studenty, kdy jeden dělal test SAT a měl 720 bodů a druhý dělal ACT a měl 21,… bodů, musíme jejich výsledky převést na procenta → čili počítáme to samé jako u Aboriů (v grafu použijeme průměr a dosaženou hodnotu studenta, která je graficky umístěna napravo → chceme vypočítat vše, co je od dosaženého počtu bodů nalevo - ano, jde to přes průměr - a pak už klasicky spočítáme rozdíl mezi průměrem a dosaženými body, tento rozdíl vydělíme SD, získanou hodnotu najdeme v tabulce a tak zjistíme, kolik procent vlastně onen student získal) → takhle to vypočítáme pro oba studenty a potom je můžeme podle procent srovnat (zjistíme tak vlastně, kolik procent studentů bylo horší než oni).

(Čísla pro SAT: Průměr - 500 bodů, SD - 100 bodů, dosažené maximum - 800)

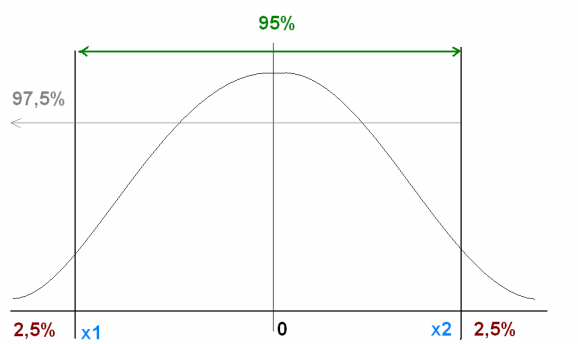
Příklad 04:
Průměrné IQ - 100, SD - 16, kolik lidí může být v Mense, když dolní hranice je IQ 130?

Výpočet:
1) graf normálního rozdělení (uprostřed 100, napravo 130)
2) 130 - 100 = 30
3) 30 : 16 = 1,88
4) 1,88 podle tabulky odpovídá kvantilu 97% → 97% se do Mensy nedostane

Odpověď:
3% lidí se dostane do Mensy. (Teoreticky, když je nás 10 000 000, by Mensa mohla mít 300 000 lidí.)



Příklad 05:
Od jakého do jakého kvantilu je rozpětí 95%?



97,5% je podle tabulek 1,96 → x1 = -1,96, x2 = 1,96

Inferenční statistika

Statistická chyba
- informuje nás o nejistotě odhadu
- Obama:McCain 49:44%, statistická chyba - 2,9% → Obama se pohybuje v intervalu 46-52%, McCain v intervalu 41-47% → intervaly se překrývají ⇄ výsledky se ještě můžu zvrátit
- výzkum veřejného mínění - tři agentury se ve stejný čas ptaly stejné populace občanů ČR na jednu stejnou otázku (na předvolební preference ODS a ČSSD), ale výsledky výzkumu se značně lišily (jedna agentura odhadla těsné vítězství ČSSD s vysokým počtem hlasů, druhá jasné vítězství ODS s průměrným počtem hlasů, třetí vítězství ČSSD s velice nízkým počtem hlasů) → proč se odhady tolik lišily? Může za to prostě jen náhoda, jiné vysvětlení není

Princip testování hypotéz
- zajímají nás názory celé populace
- měříme jen názory náhodně vybraných jedinců

Základní soubor (populace)
- soubor všech jedinců
- populace není biologická, ale statistická
- nekonečná nebo rozsáhlá velikost

Náhodný výběr
- reprezentativní část populace náhodně vybraná z celé populace
- omezená velikost souboru

Populace vs. soubor
- populace - podíl voličů ODS - 35% (počítá se z výsledku voleb) → vlastnosti populace jsou **parametry**
- soubor - podíl voličů ODS - 20-50% (odhaduje se z předvolebních výzkumů) → vlastnosti souboru jsou **statistiky**

Parametry
- vlastnosti populace
- řecká písmena
- např. parametrický podíl voličů ODS

Statistiky
- vlastnosti náhodného výběru
- latinská písmena
- např. výběrový podíl voličů ODS

	parametr	statistika
průměr	μ (mí)	y s čarou
SD	σ (sigma)	s
rel. četnost	π (pí)	p

- zajímá nás parametrický podíl podpory ODS (výsledek voleb) → parametrický podíl nelze získat (leda až u voleb), lze ale získat výběrový podíl podpory ODS (předvolební průzkum)

Interval spolehlivosti
- způsob jak odhadnout z předvolebních průzkumů výsledky voleb

Konstrukce intervalu spolehlivosti

Příklad 06:
populace - studenti univerzity (n = 20 000)
proměnná s normálním rozdělením - výška postavy
parametry - průměr = 172 cm, SD = 7 cm

Výpočet:
1) z populace náhodně vybereme 5 studentů (náhodný výběr n = 5) a to samé uděláme 1000x → máme 1000 náhodných výběrů, kdy n = 5.
2) u každého výběru uděláme průměr → máme 1000 výběrových průměrů
3) vytvoříme histogram rozdělení průměrů náhodných výběrů (není to rozdělení výšky postavy, ale rozdělení průměrů těch každých pěti náhodně vybraných studentů) = **výběrové rozdělení**

Centrální limitní věta - rozdělení průměrů náhodných výběrů z populace s normálním rozdělením je samo normálně rozdělené a to bez ohledu na velikost výběru, s rostoucí velikostí souboru bude rozdělení průměrů výběrů z populace o jakémkoliv rozdělení dosahovat normální rozdělení = je jedno, jaké bylo na začátku rozdělení hodnot, teď to výběrové bude normální) → zkrátka rozdělení průměrů náhodných výběrů bude vždy normální rozdělení

- **průměr rozdělení** - μ
- **směrodatná odchylka průměru** = **směrodatná chyba** ($\sigma_{y \text{ s pruhem}}$)

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

Směrodatná odchylka
- míra variability proměnné (výšky postavy) - SD, σ

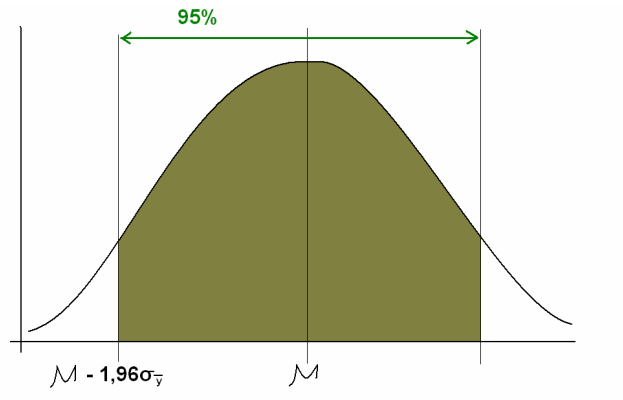
Směrodatná chyba
- míra variability průměru souborů - SE, $\sigma_{y \text{ s pruhem}}$

---

4) zde použijeme ono rozpětí 95% (příklad 05) - 95% bude v intervalu, zbytek budou extrémy (moc malý, nebo moc velký)

Interval spolehlivosti se počítá tak, že si vybereme jeden náhodný výběr o určité velikosti, který bude mít určitý výběrový průměr, což je střed intervalu spolehlivosti → spočítáme směrodatnou chybu = rozpětí intervalu spolehlivosti

například: pokud si náhodně vyberu hodnotu 170 cm, zakreslím ji doprostřed grafu, nakonec pomocí vzorce zjistím, že směrodatná chyba je třeba 12 → rozpětí bude 164-176 (dolní hodnota = 170 - 6, horní hodnota = 170 + 6)



Jak souvisí interval spolehlivosti se statistickou chybou?

20. listopadu 2008 - 7. přednáška + cvičení

IS : σ + (-) z x sigma σ

IS - interval spolehlivosti - odhad parametrického průměru
z - kvantil standardizovaného normálního rozdělení - 1,96

- problém při výpočtu IS? - nelze spočítat, neznáme parametrickou odchylku, ale statistickou Sσ = $\frac{S}{\text{odmocnina z n}}$

proto je třeba pozměnit vzorec na statistickou odchylku IS : σ + (-) t x σ
t - kvantil studentova rozdělení

Studentovo rozdělení - symetrické, tvar je ovlivněn stupněm volnosti (degree of freedom **df = n – 1**)

3 způsoby, jak ovlivnit IS?
a) ovlivnit spolehlivost - určí se hledaná procenta (užší interval = snížit spolehlivost)
b) ovlivnit směrodatnou chybu - na základě směrodatné odchylky (snížit směrodatnou chybu = zúžit interval) (zúžit interval = snížit odchylku)
c) ovlivnit směrodatnou chybu - na základě velikosti souboru (zvýšit počet členů = zúžit interval)

- nemůžeme ovlivnit odchylku - můžeme tedy měnit procenta a zvyšovat velikost souboru, ale to je všechno

Interval spolehlivosti pro relativní četnost
- relativní četnost - pravděpodobnost úspěchu - 60% lidí podporuje trest smrti
  - parametrická četnost - 0,6 – π
  - výběrová četnost - to, co my zjistíme

Výběrové rozdělení četnosti
= normální rozdělení
= směrodatná chyba S_p / SE_p– výběrových relativních četností

n = 20
$S_p^2 = \frac{p\,(1-p)}{n-1}$ p = 0,5,0,58 . . . . výběrové rozdělení (čestnost), je normální
π = hledáme

IS: p + (-) z (hodnota kvantilu, 1.96) x S_p

27. listopadu 2008 - 9. přednáška + cvičení

Testování hypotéz

Příklad 07:
Jsou muži v průměru vyšší než ženy?

Řešení:
1) vezmu soubor třeba 20 žen a stejně velký soubor mužů
2) změřím jejich výšku postavy
3) vypočítám průměr a zjistím třeba, že ženy jsou v průměru vysoké 166 cm a muži 173 cm
→ muži jsou v průměru vyšší než ženy - ale pouze v tomto souboru mužů a žen, neplatí to hned pro celou populaci, protože jsme srovnali průměry náhodných výběrů, výběrové průměry jsou náhodná čísla a srovnávat dvě náhodná čísla nemá logiku

parametrický průměr	
F	165
M	170

→ teď už to můžeme stáhnout na celou populaci, protože se to týká parametrického průměru
→ 20 žen je jeden soubor → my jsme těchto souborů udělali celkem třeba 10 000, to samé jsme udělali se souborem mužů → pokaždé jsme porovnali průměry u obou souborů → někdy vyšlo, že byly ženy v souboru vyšší než muži
Z populací žen a mužů s parametrickými průměry 165 cm a 170 cm můžeme náhodně vybrat soubory tak, že
a) průměr souboru žen je větší než průměr souboru mužů
b) průměr souboru žen je roven průměru souboru mužů
c) průměr souboru žen je menší než průměr souboru mužů

Kdy můžeme zamítnout hypotézu, že průměr souboru žen je větší než průměr souboru mužů?
Pokud tato hypotéza neplatí, potom těchto variant bude málo → pokud těchto variant bude málo, potom jsou zanedbatelné a my můžeme směle tvrdit opak, můžeme tvrdit, že tato hypotéza asi neplatí, prostě ji zamítneme

Metodologie zamítnutí hypotézy

a) hypotézu nelze dokázat
- shoda dat s hypotézou ještě neznamená, že je hypotéza pravdivá

b) hypotézu lze pouze vyvrátit
- hypotézy platí pouze do té doby, než se nám ji podaří vyvrátit

Testování hypotéz

Mayové
- voda se tradičně získává ze studní, pro vodu chodí jen ženy → nošení vody na dlouhou vzdálenost → potom mexická vláda zavedla vodovod a ženy už nemusely tak daleko
- důsledek - po zavedení vodovodů do vesnic výrazně klesl energetický výdej žen
- lze předpokládat, že ušetřenou energii ženy investovaly do výživy plodu
- otázka je, jestli tělesná zátěž matky ovlivňuje délku novorozence
- mámě dvě populace novorozenců
  a) novorozenci, jejichž matky jsou vystaveny vysoké zátěži (chodí daleko pro vodu)
  b) novovrozenci, jejichž matky nejsou vystaveny vysoké zátěži (vodu berou z vodovodu)

a)
- normální rozdělení
- průměr μ = ?
- SD σ = ?
- máme náhodný soubor 15 novorozenců

b)
- normální rozdělení
- průměr $\mu_0 = 50$

- SD $\sigma_0 = 2$

Otázka:
Je parametrický průměr populace novorozenců matek s vysokou zátěží stejný jako parametrický průměr populace novorozenců matek s nízkou zátěží?
Úroveň zátěže matky namá vliv na průměrnou délku novovrozence.

Nulová hypotéza ($H_{0\text{ - nula}}$)

H_0: $\mu = 50{,}0$ (cm)

$H_0: \mu = \mu_0$

μ - průměr populace s vysokou populací
μ_0 - průměr populace s nízkou populací

Pro nulovou hypotézu platí, že ji kladu pro to, abych ji vyvrátila, konstatuje nulový rozdíl.

Alternativní hypotéza (H_A)

- je v rozporu s nulovou hypotézou, negace nulové hypotézy

$H_A: \mu \neq \mu_0$

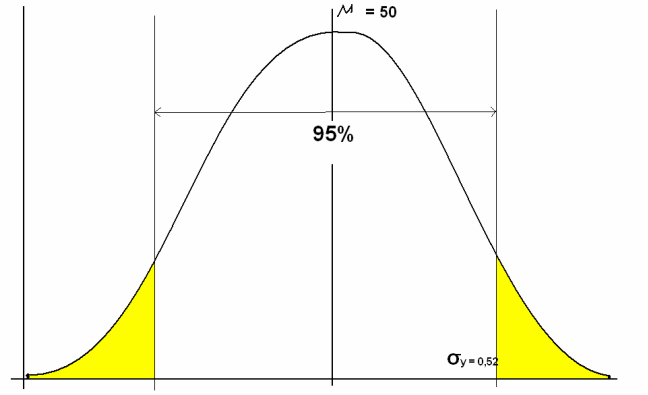
Délka novorozence
- normální rozdělení
- n = 15
- průměr $\mu = 50$
- SD $\sigma = 2$

- uděláme 1000 výběrových průměrů (jeden zahrnuje 15 jedinců)

Sample distribution - rozdělení souboru - průměr proměnné
Sampling distribution - rozdělení výběrových průměrů (nebo obecně čehokoliv) **- výběrové rozdělení**

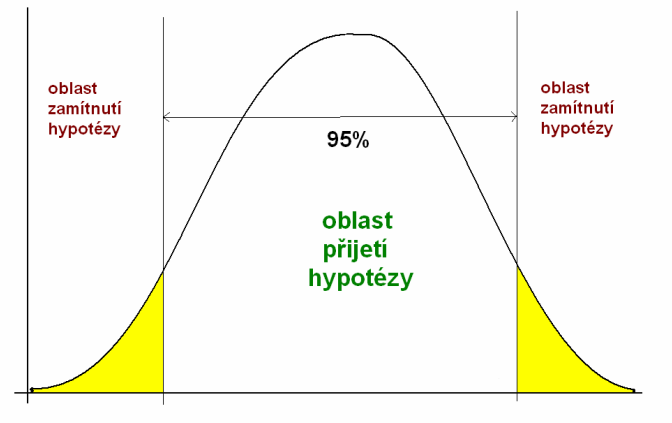
→ rozdělení průměru souborů (n = 15)
- rozdělení bude normální
- průměr bude roven parametrickému průměru (v našem případě to bude 50 cm)
- míra variability bude SE (směrodatná chyba)

→ vypočítáme SE podle vzorce (viz výš) → SE = 0,52 → míra variability je 0,52



A počítáme jako předminulou přednášku kalorický příjem Aboriů… Akorát teď dopočítáváme obě krajní hodnoty
Výsledkem je rozmezí 48,9 až 51,1 (pro soubor s vysokou zátěží)

→ zjistíme, že výběrový průměr je 49,8 cm (délka novorozenců žen s velkou zátěží je 49,8 cm)
Ovšem tohle nemůžeme zobecnit pro celou populaci, protože jsme to počítali jen pro náhodný výběr - nemůžeme tedy jednoznačně říct, že délka novorozenců žen s velkou zátěží je menší



Chyba I a II druhu

1) nulová hypotéza v populaci platí
   a) na základě testování ji přijmeme → správné tvrzení
   b) na základě testování ji zamítneme → chyba I druhu

2) nulová hypotéza v populaci neplatí
   a) na základě testování ji přijmeme → chyba II druhu
   b) na základě testování ji zamítneme → správné tvrzení

- chyba I druhu (alfa) - zamítnutí platné nulové hypotézy
- chyba II druhu (beta) - přijetí neplatné nulové hypotézy

- velikost chyby I druhu si nastavujeme sami, standartně je to 0,05 (popř. 0,01)

- průměr výběru je v oblasti přijetí hypotézy
  - s chybou I druhu alfa = 0,05 přijímáme H_0
  - na 5% hladině významnosti přijímáme H_0

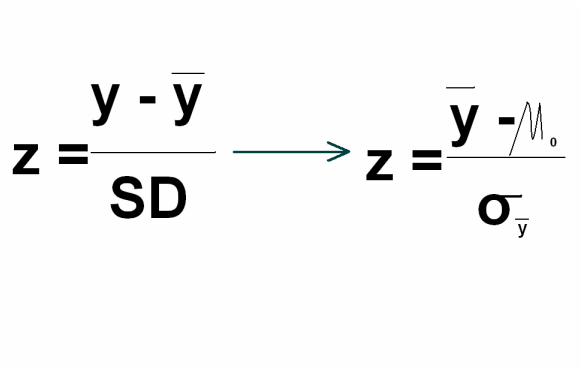
Jednovýběrový t-test (one-sample t-test)
- jednovýběrový proto, že pracuju s jedním výběrem (u jedné populace)
- testovací kritérium - studentovo t

Když se vrátíme k původnímu příkladu u Mayů…
y = 47,25 cm
s = 4,10 cm
n = 15
Nulová hypotéza: Ptáme se, jestli byl jedinec vybrán z počtu jedinců blížících se 50 cm?

1) stanovíme nulovou hypotézu (H_0: $\mu = \mu_0$) → uděláme všechno, co před chvílí → máme graf s dolní hranicí 48,9 a horní 51,1

→ naše y je 47,25, což je mimo rozptyl → to znamená, že vybraný soubor asi není vybrán z populace o průměru 50 cm ⇉ hypotézu zamítnu
Tohle ale není správný postup! Budeme používat ten následující…

Použijeme vzorec pro standardizaci, který ale zobecníme



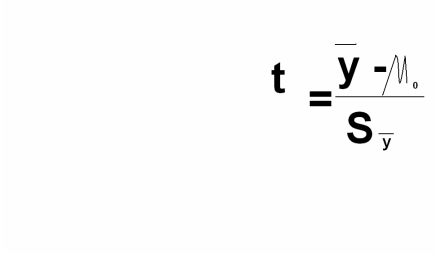
$$z = \frac{y - \bar{y}}{SD} \longrightarrow z = \frac{\bar{y} - \mu_0}{\sigma_{\bar{y}}}$$

$$z = \frac{-2,75}{0,516} = -5,3$$

→ ze vzorce nám vyjde číslo 5,3 - to je to, jak moc se to blíží k tomu průměru (50 cm)
-1,96 a 1,96 jsou srovnávací hodnoty
-5,3 a 5,3 jsou extrémní hodnoty → je malá pravděpodobnost, že byl jedinec vybrán z průměru

Zamítnutí hypotézy se nerozohoduje podle skutečné hodnoty průměru, ale podle standardizované hodnoty průměru
Je to testové kritérium

Tohle ale platí pro normální rozdělení
Standardizované skóre pro studentovo rozdělení se počítá podle vzorce



$$t = \frac{\bar{y} - \mu_0}{S_{\bar{y}}}$$

$$t = \frac{47,25 - 50}{1,06} = -2,6$$

Takže pokud nám vyjde extrém, potom bude hypotéza nepravdivá a zamítneme ji, pokud nám vyjde normál, potom je hypotéza potvrzena

Snažíme se zjistit, jak moc je pravděpodobné vybrat soubor, který jsme vybrali

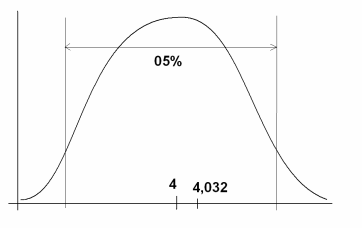
---

Příklad 08:

Obecná politická orientace (na škále od 1 do 7, kdy 1 = liberál, 4 = neutrál, 7 = konzervativec)
n = 627
y s čarou = 4,032 (výběrový průměr)
s = 1,257 (odchylka)

1) nulová hypotéza: Obecný politický názor je neutrální.
   H_0: $\mu = 4$

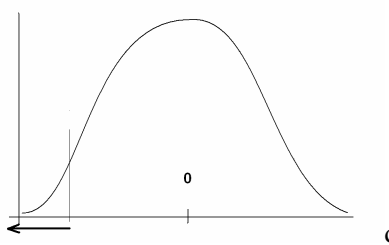
2) graf pro normální rozdělení



3) vzorec pro testovém kritérium

$$t = \frac{y - \mu_0}{S_{\bar{y}}}$$

4) graf pro studentovo rozdělení



df = 626

5) spočítáme $S_{v\ s\ pruhem}$ (směrodatná odchylka)

$$S_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = 0{,}05$$

6) dopočítáme t (testové kritérium) → t = 0,64

Hodnota P
- jiný způsob vyjádření odlehlosti našeho t
- přesnější informace

4. prosince 2008 - 10. přednáška + cvičení

- hodnota P vyjadřuje četnost testových statistik (např. t), které stejně nebo ještě více odporují H_0 (jsou ještě více extrémnější) než pozorovaná hodnota testové statistiky

---

Příklad 09:

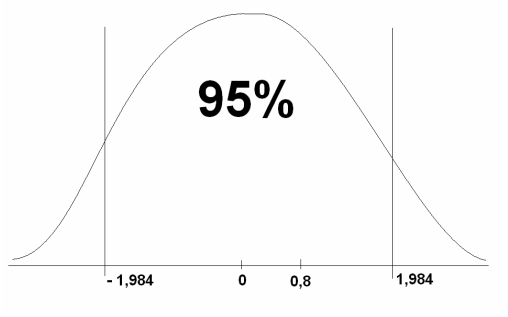
Přijímací test na SŠ
průměr dosažených bodů je 500
testujeme 100 studentů (n = 100)

y s čarou - 508
s = 100
máme dva druhy studentů - americké (μ_a) a zahraniční (μ_b)
Mají zahraniční studenti jiné průměrné skóre?

1) stanovíme nulovou hypotézu
H_0: $\mu_b = 500$
H_A: $\mu_b \neq 500$ (alternativní hypotéza)

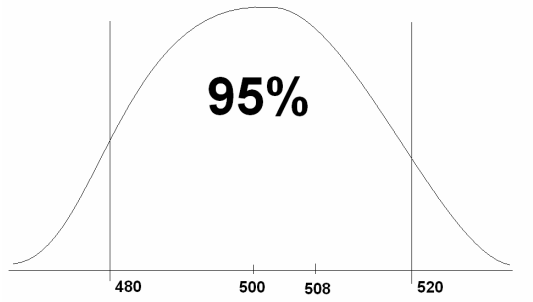
2) vypočítáme testové kritérium t
t = (y s čarou - μ) / $s_{y\ s\ čarou}$
$s_{y\ s\ čarou}$ = s / odmocnina z n = 100 / odmocnina ze 100 = 10
t = (508 - 500) / 10 = 0,8

3) z grafu zjistíme, jestli můžeme hapotézu zamítnout, čili jestli mají zahraniční studenti jiný průměr bodů z testů než američtí studenti.



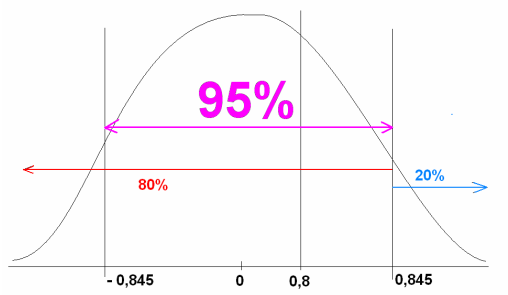
ptáme se: je 0,8 zamítnutelné kritérium?
podíváme se do tabulky studentova rozdělení, tam najdeme hodnotu 1,984 při df = 99 ve sloupečku 0,975 (proč 99? To se odvíjí od velikosti souboru n = 100, ale vždy od tohoto počtu musíme odečíst 1. A proč 0,975? Protože se pohybujeme v rozpětí 95%, do 100% nám zbývá 5%, jenže my počítáme jen jednu stranu, proto od 100 odečteme 2,5 a jsme na těch 97,5, čili 0,975)
0,8 se nachází v oblasti 95%, protože je menží než zjištěných 1,984
Z toho se odvíjí, že kritérium nemůžeme zamítnout, protože je v onom rozmezí 95%
Proto můžeme z 95% pravděpodobností říct, že zahraniční studenti mají stené průměrné skóre jako američtí studenti (protože nemůžeme zamítnout nulovou hypotézu, nulová hypotéze v tomto případě platí)

4) zjistit, jestli jedinec, který má skóre 508 (y s čarou) patří do námi testované skupiny



hodnota 500 je důsledek centrální limitní věty - tedy pokud vycházíme z n = 100 a pokud bychom všechny zprůměrovali, vyjde nám za daných podmínek průměr 500 (μ_a), SE = 10 (SE = SD / odmocnina z n)
1,984 . SE = 1,984 . 10 = 2 → 20 (proč? Nevím, je to podle nějakého vzorce)
a dolní hranice bude 480 (protože 500 - 20) a horní hranice bude 520 (protože 500 + 20)
→ s 95% pravděpodobností můžeme říct, že student, který dosáhl výsledku 508, byl vybrán z naší testovací skupiny

5) vypočítat hodnotu P



zjistíme t, když máme df = 100 a 80% (sloupeček 0,8)
kritická hodnota P je 0,05, protože je to tak vždycky - to si pamatuj
P = 0,4256, protože máme 20% od 80% nahoru a 20% od 80% dolů → 0,2 + 0,2 = 0,4 (jak to spočítal přesně, to prý nepotřebujeme znát), nám stačí těch 0,4, i když je to nepřesně, protože už z toho víme, že to je větší než 0,05
→ nezamítáme hypotézu, protože P je větší než jeho kritická hodnota (a my právě chceme teď znát ten extrém, ale tohle není extrémnější, takže nezamítáme)

Výsledky výběrových testů
P = 0,8 → nemůžeme zamítnout
P = 0,06 → tohle je na hraně
P = 0,04 → můžeme zamítnout, ale je to na hraně
P = 0,0001 → s čistým svědomím můžeme zamítnout
→ hranice je vždy 0,05
u 2. a 3. P je tak na hraně, že stačí jen jinak zaokrouhlit a už máme špatně celou hypotézu

Všechna P, která jsou větší než 0,05 jsou extrémy, a proto je nezamítáme!
P se v excelu počítá podle funkce TDIST

Jestli se testové kritérium zvětšuje, P se snižuje → budeme hypotézu více zamítat
Jakoukoliv hypotézu můžeme zamítnout, pokud dostatečně zvětšíme soubor
(to plyne z výpočtů ze vzorce SE = SD / odmocnina z n → navýšení n vede k navýšení testového kritéria)
No jo, ale to si potom můžu se statistikama hrát, jak se mi zachce, a můžu ovlivňovat její výsledky?!
↓↓↓↓↓↓↓↓↓↓↓↓↓↓

Statistická významnost
- statistický rozdíl
- prostě jsme při přepočítání došli k jiným výsledkům, proto zamítneme původní hypotézu jako chybnou, protože nám to statisticky nesedí

Praktická významnost
- je to statistický význam, který pro normální lidi není významný
- znamená to, že rozdíl mezi výsledky je tak malinký (v řádu desetin), že to pro nás nemá žádný význam (jen pro hnidopichy)