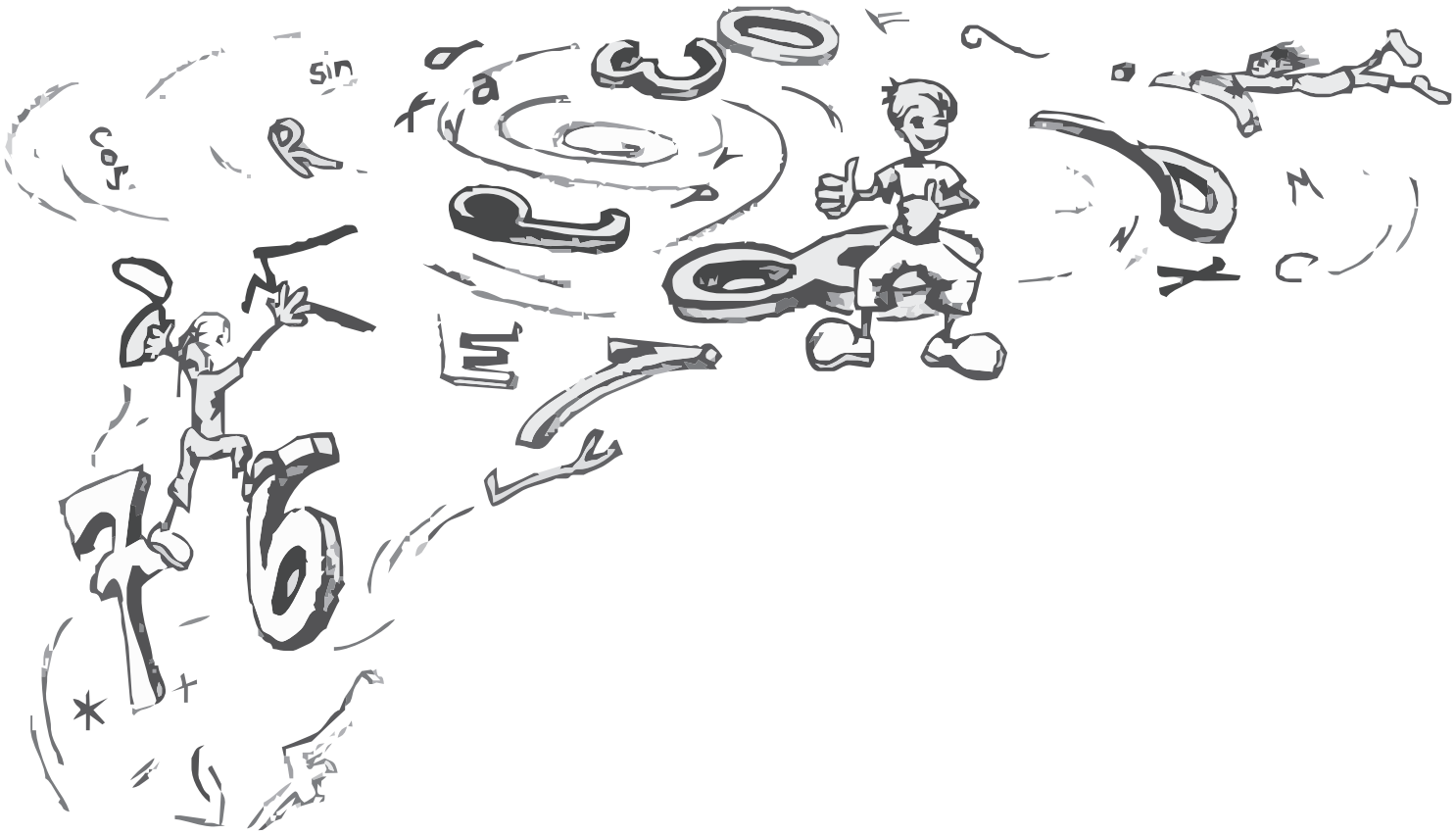
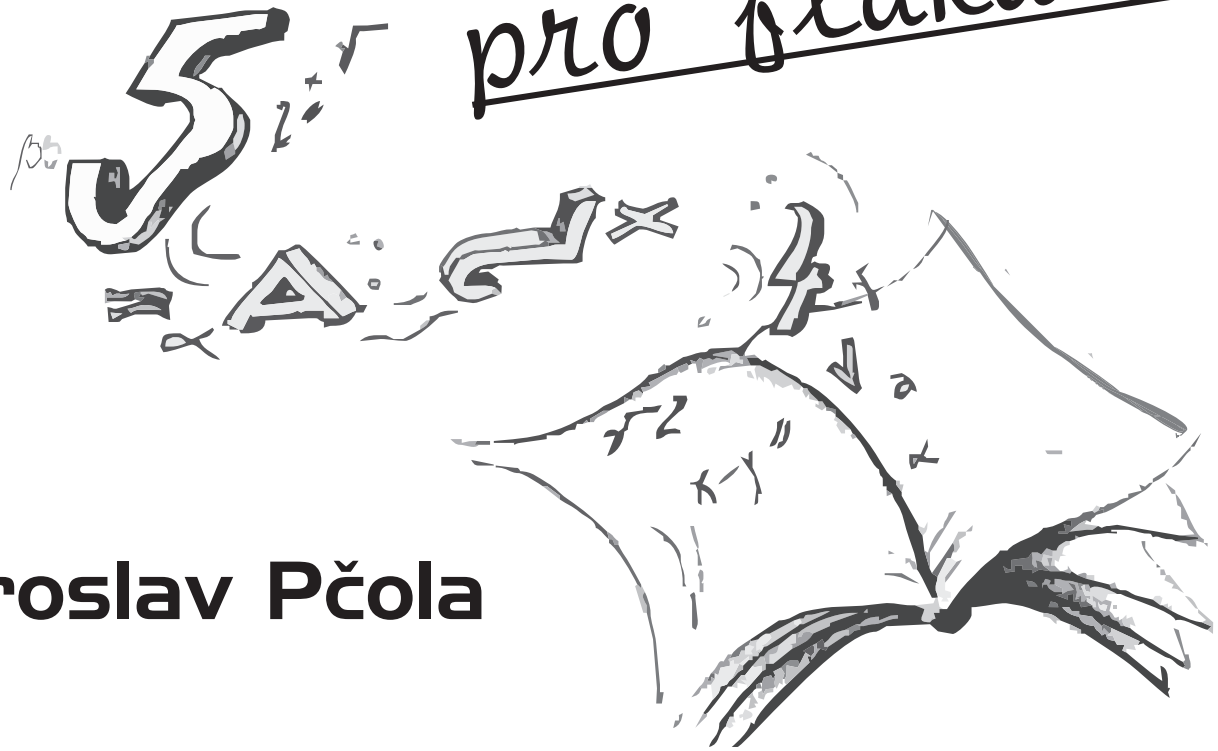


Statistika

pro fláka



Jaroslav Pčola

NOVÁ

STATISTIKA PRO FLÁKAČE

K DOSTÁNÍ NA WEBU:

WWW.PROFLAKACE.CZ

- ☑ **V ČEŠTINĚ!**
- ☑ **SROZUMITELNÝ VÝKLAD**
- ☑ **VŠECHNY TYPOVÉ PŘÍKLADY**

  PODLE OFICIÁLNÍCH POŽADAVKŮ NA ZÁVĚREČNÝ TEST 4ST201
- ☑ **2,5 KRÁT VÍCE TEXTU**,

  PROPRACOVANÝ A INOVOVANÝ VÝKLAD, VÝSTUPY ZE SASU,INDEXY, ČASOVÉ ŘADY A MNOHO NOVÉHO

S TÍMTO SKRIPTEM TO POCHOPÍTE. OBJEDNEJTE SI HO A NETRAPTE SE;)

Toto je český preklad pôvodnej SPF od našej neznámej spolužiačky ktorej touto cestou ďakujem. Verím že Ti pomôže, aj keď to je už trocha zastaralý matroš a je v ňom veľa chýb. Ak máš záujem o novú SPF, teda skriptum, ktoré je podobné tomuto, ale je tam toho skoro 3x viac a bez chýb potom navštív web proflakace.cz. Autor je samozrejme stále rovnaký ☺ Veľa zdaru! Jaro P.

PS.: Aby som neuploadoval len reklamu, prikladam ako bonus k tejto verzii aj kapitolu Anova z novej SPF, najdes ju na strane 24.

Statistika pro flákače

Úvodem bych asi měla upozornit, že se tu nedovíte nic víc než co je v originálu statistiky pro flákače ve slovenštině, ptž je to podstatě doslovný překlad. Přepsala sem to do češtiny pro případ, že by někomu čtení českého materiálu přišlo snazší a hlavně rychlejší než slovenský originál.:)

Základní pojmy:

Elementární spravování statistických údajů

Držím se knihy, kde od strany 17 po str. 29 existuje kapitola s tímto názvem. Dozvíte se, že pokud měříme nějakou veličinu (např. výška dětí ve třídě), můžeme tyto naměřené hodnoty různě rozdělit. Použijeme tabulku rozdělení četnosti, kam si napíšeme pod sebe jaké různé výšky jsme naměřili a kolikrát.

167 … 1
172 … 5
176 … 4
180 … 6

To kolikrát jsme naměřili danou veličinu se nazývá **absolutní četnost** a to jaký podíl má daná výška na celku se nazývá **relativní četnost** (tedy kolik procent lidí ve třídě má výšku např. 172 = 5/(1+5+4+6) = 31,25%).

Veličiny můžeme rozdělit i do intervalů, které si určíme, např. od 150 do 160, od 160 do170 a od 170 do 180. Kdybychom měli 1000 dětí tak je rozdělení přehlednější než u tabulky rozdělení četností.

Míry úrovně – Polohy

Poloha vlastně znamená, že pokud si představíme číselnou osu a chtěli by jsme na ni naznačit to hafo našich čísel (př. ty výšky dětí) jen jediným bodem, kam by jsme dali značku. Je to vlastně jaká si střední hodnota. Možností je víc:
Průměr – průměrů existuje několik. Nás zajímá hlavně aritmetický – tzn. klasika, sčítám všechny hodnoty jsme naměřili a vydělím je počtem měření.
Medián - je hodnota středního člena, kdybychom postavili děti podle výšky vedle sebe, výška toho, kdo bude ve středu bude medián. Kdyby byly 2, medián by byl průměr dvou hodnot ve středu.
Modus – je nejčastěji se vyskytující hodnota. Tedy z čísel 2, 2, 3, 4, 5, 5, 5, 6 by byl modus 5, protože je tam 3krát.

Míry variability

Variabilita je vzdálenost našich naměřených hodnot od střední hodnoty. Můžeme ji vyjádřit různě, např. máme 10 dětí a průměr jejich výšek je 170. Tím jsme vyjádřili polohu a tak můžeme říct, že jejich výšky se pohybují od 162 do 185, tak jsme vyjádřili variabilitu. Konkrétně tomuto vyjádření se říká **variační rozpětí**., kdy jednoduše odčítáme nejmenší hodnotu od nejvyšší. Tzn. 185–162= 23. Tento způsob je velmi náchylný na extrémy, pokud by jsme měli jen jediného 130cm trpaslíka, naše variační rozpětí by bylo 53. Proto chytří lidé vymysleli **rozptyl**. Když počítáme rozptyl tak od každé naměřené hodnoty odečteme průměr všech naměřených hodnot. Vyjde nám tedy odchylka od střední hodnoty. Tu potom umocníme na druhou, abychom neměli záporná čísla. Pak všechny tyto odchylky umocněné na druhou spočítáme a potom vydělíme je počtem (zprůměrujeme), výsledek je ten náš rozptyl.

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Rozptyl nám, ale o skutečné variabilitě moc nepoví. Všechny odchylky jsou tam umocněné na druhou, takže při třech dětech o výšce 150, 160 a 170cm by nám vyšel průměr 160.
Odchylky na druhou: **(**150-160)^2 + (160-160)^2 +(170-160)^2 = 100 + 100 = 200
Vydělíme třemi a vyjde nám rozptyl 66,6 cm na druhou.

Proto se používá odmocnina z rozptylu a ta se nazývá **směrodatná odchylka**, v našem případě je odmocnina z 66,6cm 8,165 cm. A toto číslo přibližně vyjadřuje, že „víc jak 50% naměřených hodnot (výšek) se neodchyluje od průměru v obou směrech o víc než 8,165“. V případě tří hodnot to úplně super nevychází, ale přepočtěte si to s převrácenými hodnotami a bude to fajn☺
Směrodatnou odchylku značíme stejně jako rozptyl, jenomže ne na druhou.

A nakonec vychytávka, vzorců na rozptyl je několik, ale jednoznačně nejpoužívanější je asi tzn. **výpočtový tvar rozptylu**. Jeho vzorec je:

$$s_x^2 = \overline{x^2} - \bar{x}^2$$

Tedy zprůměrujeme druhé mocniny naměřených hodnot a od toho odečteme průměr hodnot umocněný na druhou. How simple:-P Když víte jak spočítat průměr, rozptyl, odmocnit ho a dostat směrodatnou odchylku, můžeme pokročit dál.

Pravděpodobnost

Definice

Pravděpodobnost nám říká přibližně kolik příznivých výsledků dostaneme z množství pokusů, tedy např. : když se ptáme jaká je pravděpodobnost, že nám padne na kostce 6, tak to, že nám padne 6 je příznivý výsledek a všechny ostatní jsou nepříznivé.

(Velmi dobrý materiál k základům pravděpodobnosti je Studentův průvodce pravděpodobností.)

Náhodná veličina

Výsledky pokusů (činnosti při, kterých můžeme dostat vícero výsledků) jsou často čísla, např. hod kostkou, počet poruch za směnu, atd. Tato čísla můžeme nazvat náhodnou veličinou. Tato veličina může nabývat různé hodnoty, v případě kostky našla náhodná veličina X (takhle se značí) může nabývat hodnoty 1, 2 ,3 ,4 ,5 nebo 6. Počet nehod by se mohl pohybovat od 0 až do nekonečna a naše náhodná veličina by mohla být jakékoliv číslo.

Každá konkrétní hodnota náhodné veličiny X má i svou pravděpodobnost, tedy jak na kostce může nabývat X hodnoty od 1 do 6, tak můžeme určit pravděpodobnost pro X=1, X=2,…Pravděpodobnost, že padne jedno číslo ze 6ti je 1/6 a tedy o pravděpodobnost X=1 bude 1/6, zapisujeme P(X=1)=1/6, a to jistě bude platit i pro ostatní hodnosty X, ptz každé číslo na kostce má stejnou pravděpodobnost. Právě jsme přiřadili každé možné hodnotě X pravděpodobnost, definovali jsme tzv. **pravděpodobnostní funkci**. Vypadá takto:

$P(x=1) = 1/6$
$P(x=2) = 1/6$
$P(x=3) = 1/6$
…
$P(x=6) = 1/6$

Z této funkce můžeme lehce odvodit druhou funkci a to **distribuční**, která nám udává pravděpodobnost, že náhodná veličina X nabude hodnotu **menší** než nějaké číslo. Pro naší kostku by to vypadalo takhle:

$F(x) = 0$ pro $x<1$
$= 1/6$ pro $1 \leq x < 2$
$= 2/6$ pro $2 \leq x < 3$
$= 3/6$ pro $3 \leq x < 4$
$= 4/6$ pro $4 \leq x < 5$
$= 5/6$ pro $5 \leq x < 6$
$= 1$ pro $x \geq 6$

Tedy znamená to, že když chceme vědět jaká je pravděpodobnost, že padne číslo menší než X, tak si jen dosadíme hodnotu, podíváme se v kterém intervalu je naše x a příslušející pravděpodobnost. Pro X = 4 by to bylo 4/6. Pro x>6 samozřejmě stoprocentní pravděpodobnost, ptž je jasné že určitě padne číslo menší.
Co je potřeba vědět: v testu může být akorát tak zjistit z distribuční fce pravděpodobnost a nakopak, vysvětlit její hodnotu v nějakém bodě, případně zakreslit graf.

Příklady:
1.) Veličina X nabývá hodnot 1,2 nebo 3. Známe pravděpodobnost P(1)= 0,2; P(2)= 0,5. Určete chybějící pravděpodobnost P(3) . Dále vypočítejte a interpretujte hodnotu distribuční funkce v bodě 2.

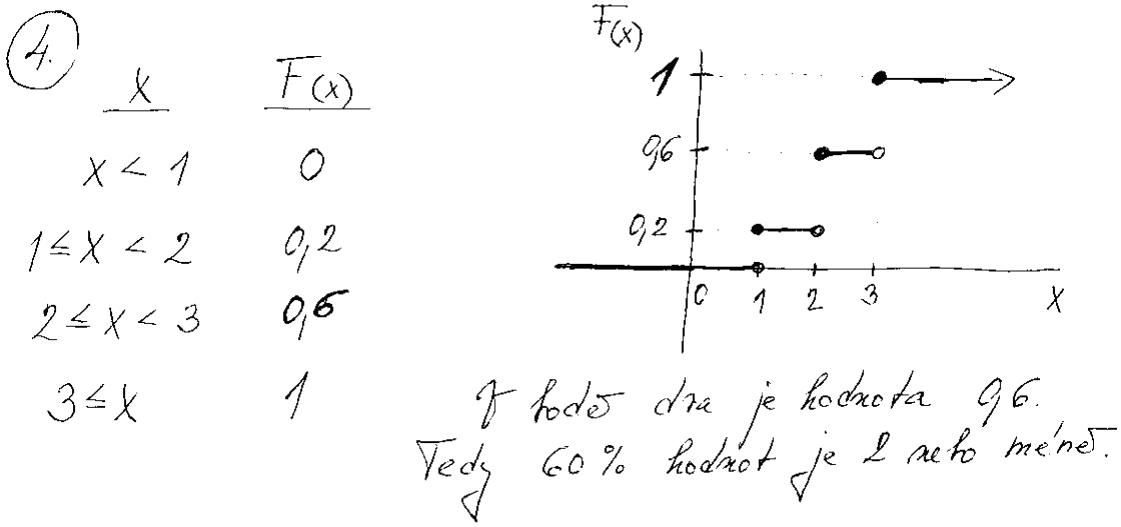
4.

X	1	2	3
P	0,2	0,5	0,3

$F(2) = 0,7$ --- 70% padlých hodnot kvůli 1 nebo 2.

Pravděpodobnost, že padne 3 je to co chybí do 100%, tzn 0,3. Distribuční fce v bodě 2 říká jaká pravděpodobnost, že padne číslo menší nebo rovno 2.

2.) 20% rodin má v domě jednu místnost, 50% jich má dvě, 30% má tři. Pro veličinu počet místností načrtněte graf distribuční funkce. Jakou má hodnotu v bodě 2?Co tato hodnota znamená?



Při tvorbě grafu a celkové distribuční fce stále dáváme uzavřené intervaly na levou stranu. Proč graf vypadá tak jak vypadá nebudeme rozebírat, to vám může být jedno, každý graf na distribuční fci vypadá takto a je to snad jediný graf ve statistice tak si to zapamatujte☺. Pro hodnoty x< je pravděpodobnost nulová, ptz nikdo nemá méně než 1 pokoj.

Rozdělení náhodné veličiny

Náhodné veličiny, tedy výsledky těch našich náhodných pokusů mají různé číselné výsledky. Tyto výsledky se dají také chápat jako statistický soubor, tedy na nich můžeme použít některé základní charakteristiky. Teda například můžeme definovat jejich střední hodnotu – jakýsi průměr nebo jejich rozptyl, resp. směrodatnou odchylku. Dále existují různá „rozdělení náhodných veličin“ díky kterým dokážeme pravděpodobnost různých jevů. Tato rozdělení jsou rovnice, které se používají tak, že do nich dosadíme hodnoty proměnných (střední hodnota, rozptyl, počet měření atd.) a ptáme se na pravděpodobnost že nastane nějaký jev, tedy většinou na nějakou pravděpodobnost, že nastane nějaká náhodná veličina X, resp. že bude X>5 a podobně. Je to podobné jako u distribuční funkce. Nic nevysvětlí lépe než praktický příklad, tak přejděme rovnou k rozdělení.

Binomické rozdělení - je rozdělení, které musíme dosadit dvěma proměnnými a to proměnnou n, která značí počet nezávislých náhodných pokusů (např. počet hodů kostkou) a proměnnou p, která značí pravděpodobnost jevu, který sledujeme (pravděpodobnost, že padne

6tka). Je tu ještě i proměnná q, která se však jen dopočítá jako (1-p) a je to teda pravděpodobnost, že daný jev nenastane. Binomické rozdělení nám dokáže vypočítat pravděpodobnost, že se v sérii pokusů (n) bude vyskytovat jev, který má nějakou pravděpodobnost (p) právě X krát. A to X to je ta naše náhodná veličina, kterou si můžeme zvolit. Pro úplnost vzorec:

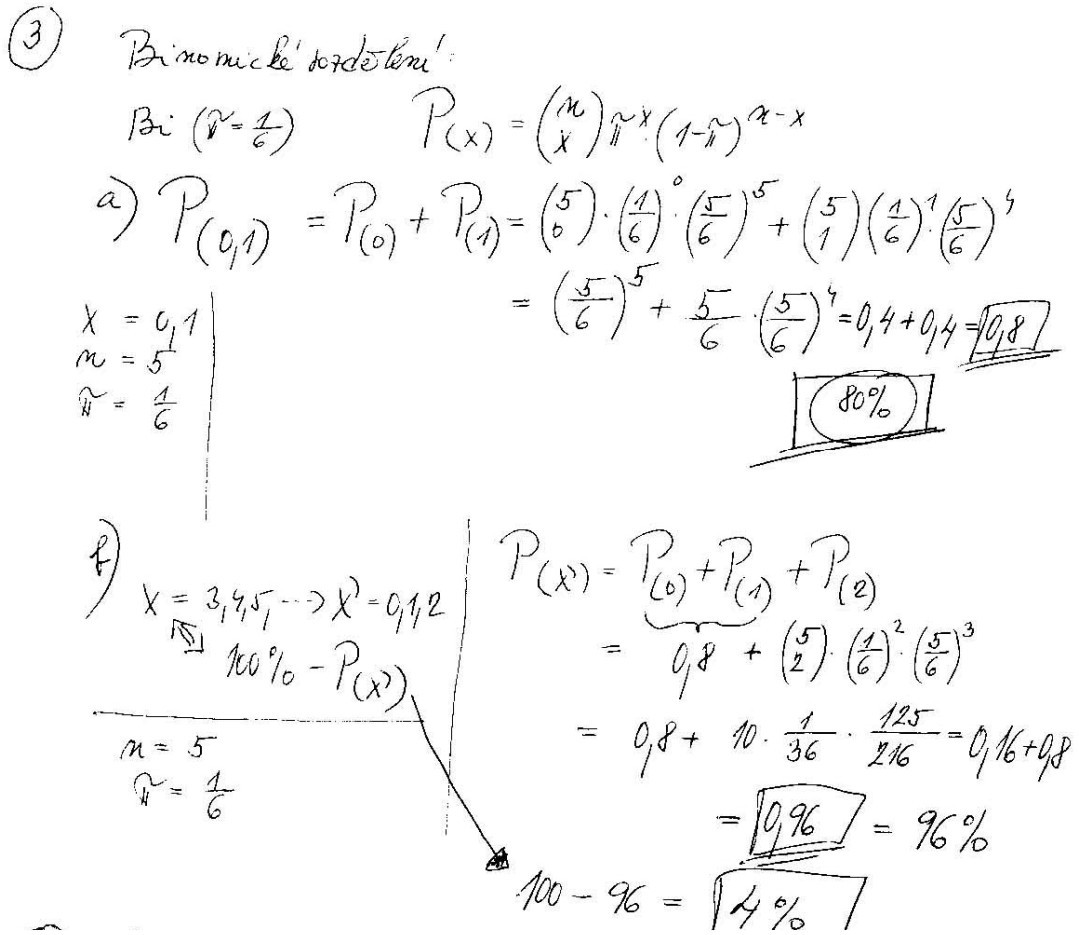
$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

Při čem to „n nad x“ je kombinační číslo, které se počítá jako n!/((n-x)!*x!)

Příklady:

1.) *Jaká je pravděpodobnost, že v pěti hodech kostkou padne 6 nanejvýš jednou a jaká je pravděpodobnost, že padne aspoň třikrát?*

Teda typické, série pokusů, je jich n, pravděpodobnost,že padne konkrétní číslo na kostce je jasná- tedy 1/6. Třeba si uvědomit, že zjišťujeme pravděpodobnost náhodné veličiny, kterou je „počet padnutí šestky v pěti hodech.“ Tato může nabývat hodnoty od 0 až do 5. Když chceme zjistit pravděpodobnost, že padne nanejvýš jednou, teda 0 nebo jedenkrát, bude se toto p rovnat součtu P(X=0) a P(X=1). Obdobně pro „alespoň třikrát“, teda 3, 4 a 5 je to buď součet pravděpodobností P(3)+P(4)+P(5) nebo i 1-(P(0)+P(1)+P(2)), protože pak by jsme sčítali pravděpodobnost od 0 až po 5 tak nám musí vyjít 100%, že jedna z nich nastane. Můžeme teda od celku (100% = 1) odečíst co chceme a vyjde nám pravděpodobnost, že nastane to co jsme odečetli. Pravděpodobnost (p) značit jako pí btw.



2.) *náhodná veličina má pravděpodobnostní fci P(x)= (3 nad x) *0,1 na x*0,9na 3-x pro X= 0, 1, 2, 3 jinak je pravděpodobnost rovna nule. Udělejte tabulku pravděpodobnostní fce a a vypočítejte modus, střední hodnotu a rozptyl této veličiny.*

Vypočítejte teda jednotlivé pravděpodobnosti, ta co má největší je modus, střední hodnota se u náhodných veličin počítá tak, že hodnoty, které nabývá vynásobíme pravděpodobnost toho, že nastanou a potom je sčítáme dohromady.

4.

a) $P_{(0)} = \binom{3}{0} \cdot 0,1^0 \cdot 0,9^{3-0} = 1 \cdot 1 \cdot 0,9^3 = 0,729$

$P_{(1)} = \binom{3}{1} \cdot 0,1^1 \cdot 0,9^2 = 3 \cdot 0,1 \cdot 0,81 = 0,243$

$P_{(2)} = \binom{3}{2} \cdot 0,1^2 \cdot 0,9^1 = 3 \cdot 0,01 \cdot 0,9 = 0,027$

$P_{(3)} = \binom{3}{3} \cdot 0,1^3 \cdot 0,9^0 = 1 \cdot 0,001 \cdot 1 = 0,001$

x	0	1	2	3
P(x)	0,729	0,243	0,027	0,001

f) modus = $\hat{x}$ = nejčetnější, nejpravděpodobnější hodnota

$\hat{x} = 0$

g) $E_{(x)} = \sum P_{(x)} \cdot X = 0 \cdot 0,729 + 1 \cdot 0,243 + 2 \cdot 0,027 + 3 \cdot 0,001 =$

$= 0,3$

Poissonovo rozdělení - je podobné binomickému, akorát ho používáme v případě, že počet prvků je tedy ***n*** je víc než 30 a pravděpodobnost je malá, prakticky menší než 0,1 tzn. 10%. Má jeden parametr **λ** - lambda, který se rovná p-krát n. Pravděpodobnost x se vypočítá pomocí rovnice:

$$P(X) = \frac{\lambda^x}{x!} e^{-\lambda}$$

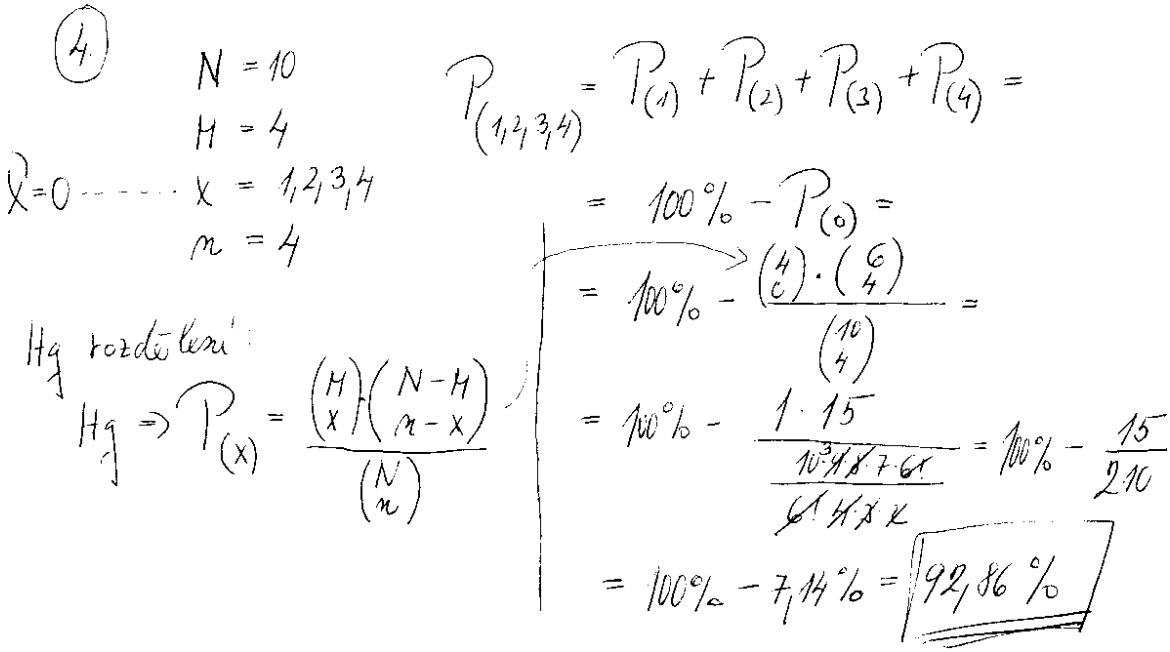
V testech se moc tyto příklady nevyskytují, dávám proto odkaz na příklad 2.19 v učebnici, resp. i v aplikacích, ale tam jsou poměrně složité a tak do hloubky to podle mne není třeba, akorát vás to vystraší.

Hypergeometrické rozdělení – používáme při výběru bez vracení, a zároveň když máme pomíchané dva druhy něčeho (př. černé a bílé koule, nahnilá a zdravá jablka, přičemž po vytažení z pytlíku to nevracíme zpět a taháme dál). Parametry tohoto rozdělení jsou: **N** – počet všech prvků, **M** – počet prvků s nějakou specifickou vlastností, **n** – počet prvků kolik

taháme, a konečně naše náhodná veličina, které pravděpodobnost hledáme je **x** a označuje počet prvků, z kterých jsme vytáhli ty, které mají specifickou vlastnost, tedy např. kolik z jablek je nahnilých. Vzorec je buď níže v příkladě nebo na straně 83.

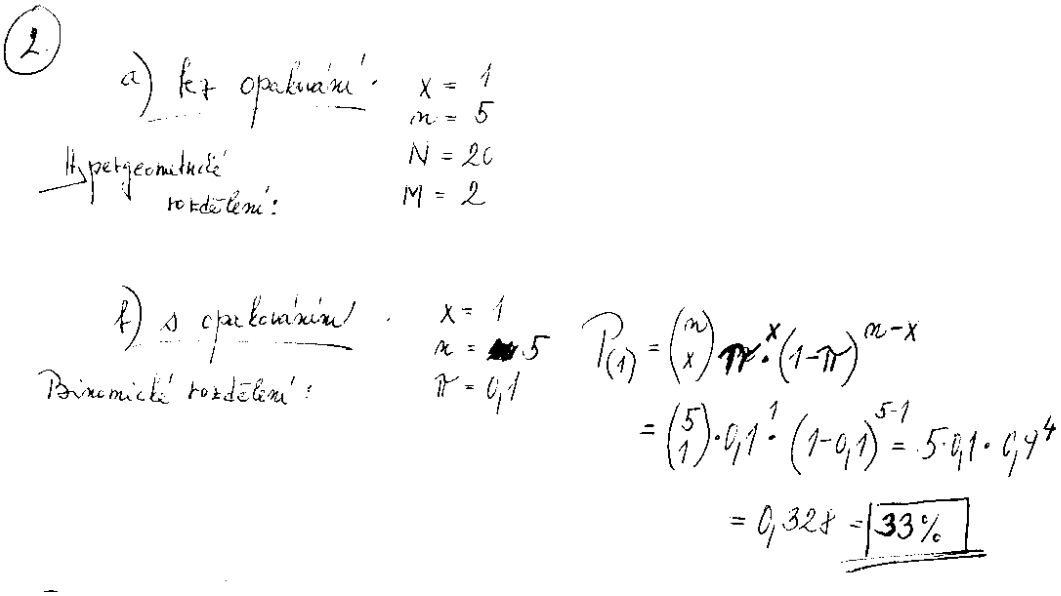
Příklady:

1.) *Máme 10 výrobků a 4 z nich jsou vadné, vytahujeme 4 bez vracení, jaká je pravděpodobnost, že aspoň jeden z nich bude vadný?*



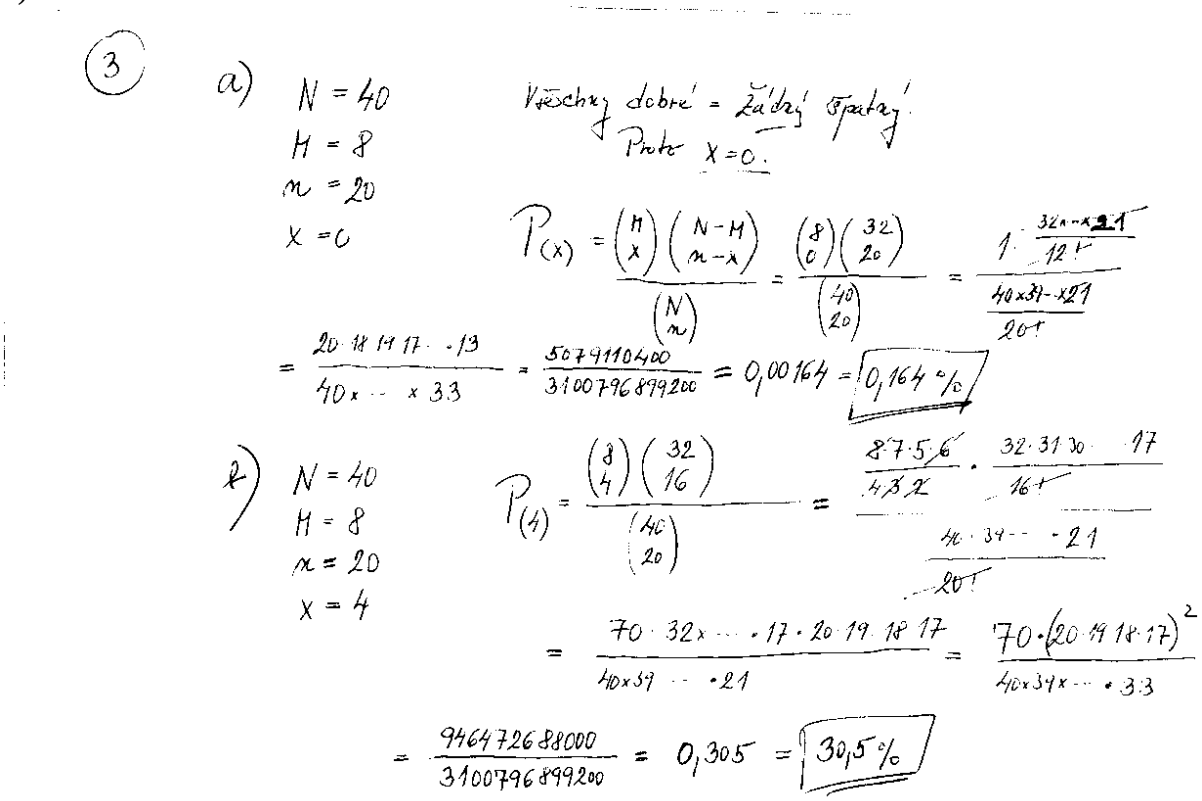
Stačí nám zjistit jaká je pravděpodobnost,že bude 0 vadných výrobků, to odečteme od 100% a máme výsledek. Vadné dosadíme za M, celkový počet výrobků je N, taháme z nich 4 teda n…mělo by to být jasné, ne?

2.) *V zásilce 20ti výrobků jsou 2 zmetky, náhodně taháme 5 kusů. Jaká je pravděpodobnost , vytáhneme jeden zmetek když taháme s vracením a jaká je pravděpodobnost, když bez vracení?*



I trocha opakování, výsledek při hypergeometrickém rozdělení by měl být 39,47%.

3.) Pěstovatel nakoupil 40 sazenic jabloní. Špatný skladováním došlo k tomu, že 8 z nich uschlo. Jaká je pravděpodobnost, že při náhodném výběru 20 sazenic (bez vracení) budou:
a) všechny dobré?
b) 4 uschlé?



Normální rozdělení – Má dva parametry a to μ - mí které je totožné s průměrem a δ^2 - delta na druhou, které je totožné s rozptylem. Tedy pokud budeme mít v případě zadaný rozptyl a průměr, automaticky je budeme považovat v případě normálního rozdělení za mí a deltu. Což je prakticky vždy, ptž je sakra málo příkladů na rozdělení, kde jsou zadané tyto proměnné a nepočítá se to přes normální rozdělení.

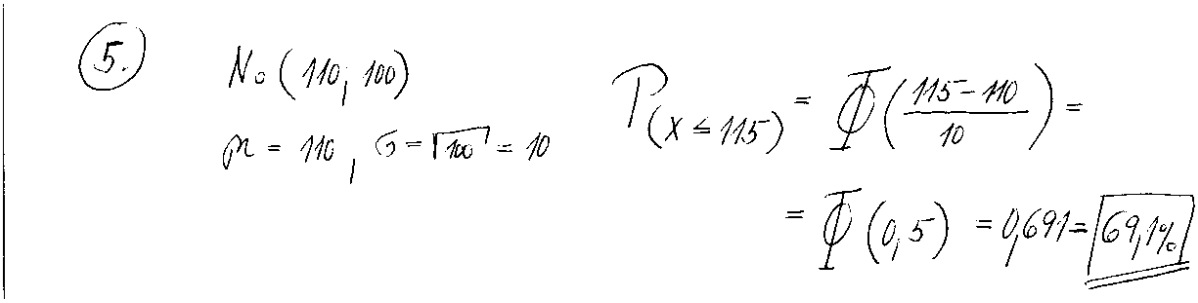
Pojďme tedy k praxi. Jestli si ještě pamatujete na distribuční funkcí, která vlastně vypovídá o tom, jaká je pravděpodobnost, že náhodná veličina X nabude hodnoty menší než nějaké číslo, tak také normální rozdělení má tuto funkci. Pomocí jí by jsme mohli zjistit jaká je pravděpodobnost, že rozměr součástky bude menší než nějaké číslo podobné. Tento vzorec na distribuční fci normálního rozdělení je však poměrně složitý, proto se zavedla tzv. normovaná veličina **U**. Její vzorec obsahuje průměr μ, odmocninu z rozptylu = δ , a X, tedy nějakou hodnotu náhodné veličiny.

$$U = \frac{X - \mu}{\delta}$$

Nemusíme tedy při počítání příkladů, kdy se nás ptají na distribuční fci (jaká je pravděpodobnost, že x bude menší/větší než X co zadáváme do vzorce…) používat složitý vzorec, stačí že vypočítáme tuto normovanou veličinu U, potom se podíváme do tabulek , kde na základě toho kolik nám ta veličina U vyšla zjistíme příslušnou hodnotu distribuční fce, značíme $\Phi(U)$, teda hledanou pravděpodobnost. Na závěr zdůrazním, distribuční fce udává P

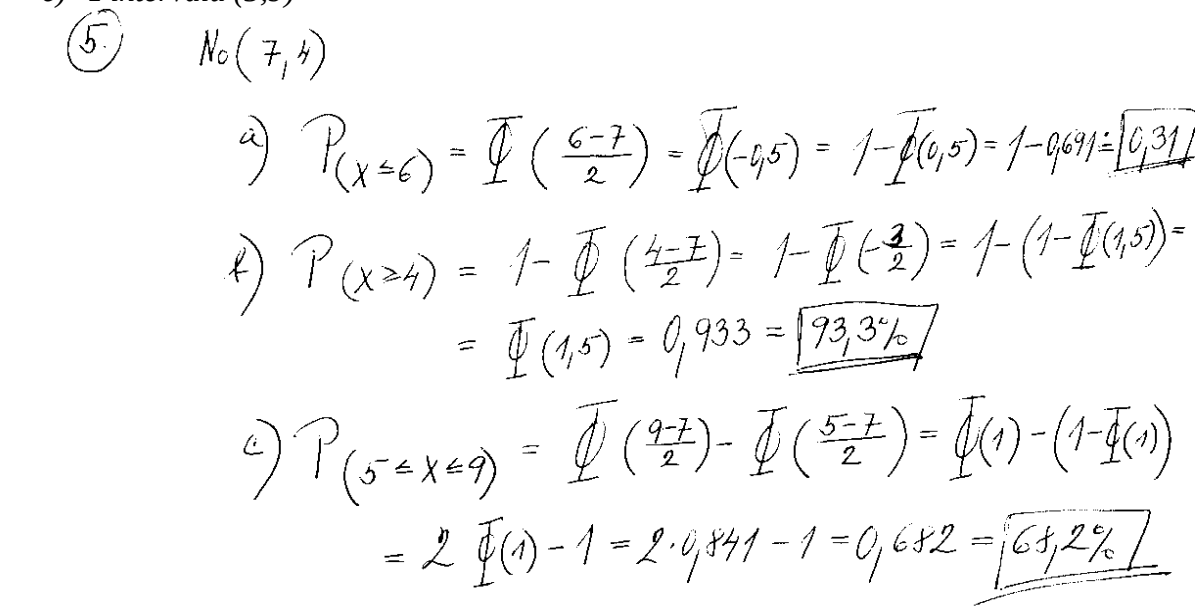
že X bude menší než nějaké číslo, tedy pokud se ptají na P že X bude větší, logicky musíme tu hodnotu distribuční fce příslušející našemu U odečíst od 1 => $(1-\Phi(U))$.

Příklady:
1.) Hmotnost vyráběných součástek je normálně rozdělená veličina se střední hodnotou 110 gramů a rozptylem 100. S jakou pravděpodobností bude hmotnost součástky menší než 115 gramů?



Tedy - ptají se na P, když hmotnost bude menší než 115 gramů, z toho vyplívá, že se ptají na distribuční fci a že za X dosadíme 115. Rozptyl je 100, za deltu dosadíme jeho odmocninu tedy 10, za μ dosadíme 110, teda střední hodnotu. Vypočítáme U a potom se už jen podíváme do tabulek a příslušná hodnota pravděpodobnosti k našemu vypočítanému U je výsledek.

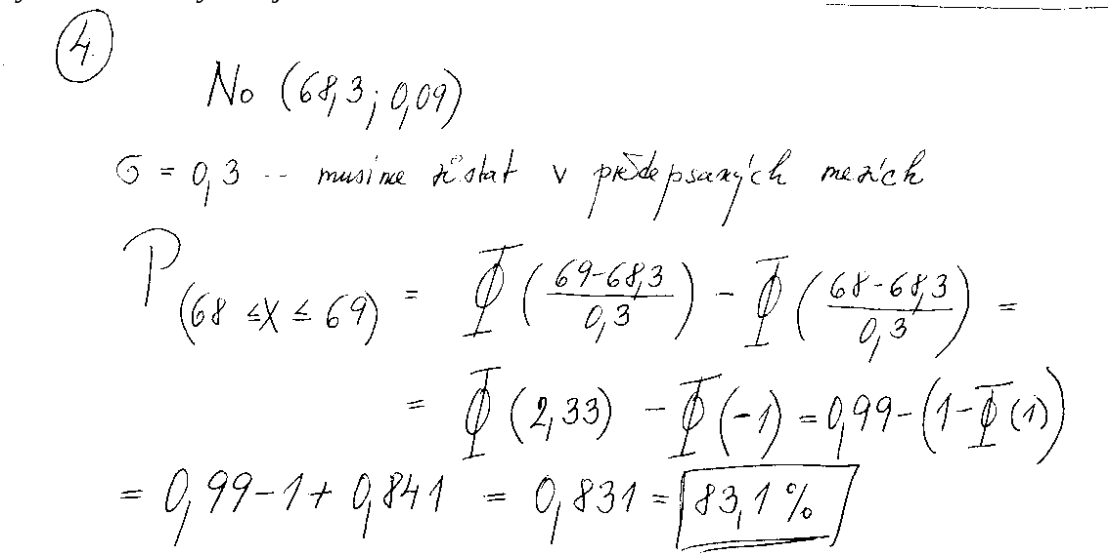
2.)Náhodná veličina X má normální se střední hodnotou 7 a rozptylem 4. Určete, že nám tato náhodná veličina nebude hodnot:
a) maximálně 6
b) alespoň 4
c) z intervalu (5,9)



Za a.) by to mělo být jasné, jen dosadíme hodnoty, za b.) je to to $(1-\Phi(U)$, a za c.) je to pravděpodobnost, že to bude menší než 9 oproštěná o pravděpodobnost, že to bude méně než 5, čímž nám vznikne pravděpodobnost intervalu od 5 do 9.

3.) Hmotnost výrobku je vyhovující pokud je v rozmezí 68-69gramů. Za standardních podmínek má hmotnost přibližné normální rozdělení se střední hodnotou μ =68,3gramů a

směrodatnou odchylku v předepsaných mezích. Jaká je pravděpodobnost, že hmotnost výrobku bude vyhovující?



Pokud je směrodatná odchylka v předepsaných mezích, smí být maximálně 0,3, ptž jinak by ve směru dolů překročila limit (68 < střední hodnota ± směrodatná odchylka < 69). Hledáme pravděpodobnost, že hmotnost bude v mezích, tedy že P bude menší než 69 a zároveň musíme odečíst pravděpodobnost, že bude menší než 68. výsledek je tedy pravděpodobnost, že hmotnost je v intervalu (68,69).

Závěr k rozdělení a Co je potřeba vědět: Když se držím jen testových příkladů, tak to jsou nejčastěji rozdělení, nebudeme tu rozebírat (aspoň zatím) Fischerovo, či logaritmicko normální, ani limitní věty, pokud máte zájem, je to v knize☺, nebudem se prozatím zabývat ani rozdělení chí-kvadrát, to budem rozebírat potom. Je třeba je vědět všechno co je tady, jsou to hodně konkrétní věci, pokud tedy nechápete z celé této kapitoly alespoň polovinu, tak ani nepokračujte dál☺, ale v podstatě jde o princip, všechny vzorce jsou v tabulkách, stačí proto jen vědět, co kam dosadit.

Zpracování dat z výběrových šetření

Úvod

Když statisticky zjišťujeme nějaký jev, často nastává situace, že rozsah souboru je tak velký, že je velmi obtížné zjistit skutečný stav. Tedy když zjišťujeme preference politických stran je samozřejmé, že se nebudeme ptát každého občana koho bude volit. Obdobně u testování součástek nebudeme testovat každou z nich zvlášť, ale jen nějaký vzorek, tzv. výběrový soubor. Příkladů ze života si i sami vymyslíte spoustu. Pro nás je důležité to, že **základní soubor** (celá populace, všechny součástky) má vlastní statistiky tak jako rozptyl a průměrná hodnota. Tím pádem má nějaký průměr resp. rozptyl i výběrový soubor (tedy ten vzorek) a my ve valné většině případů chceme na základě dat, které máme z výběrového souboru určit průměr nebo rozptyl základního souboru, přesněji řečeno, určit interval, ve kterém se tyto statistiky nacházejí.

Odhad parametru

Nebudeme moc vrtat do teorie. Spíš se zaměříme na to podstatné - jak se to počítá. Když si otevřeme vzorce na straně 4, všechno co je podstatné máme právě před sebou. Jde nám o to

zjistit buď rozptyl, nebo střední hodnotu základního souboru, když víme hodnoty nějakého výběru n prvků. Aby jsme věděli jak tyto vzorce použít, je nutné pochopit 2 věci:

Věc 1: Nebudeme to vysvětlovat, a tedy budeme to brát jako fakt, že když zjišťujeme rozptyl, nebo střední hodnotu základního souboru (ZS) na základě nějakého výběru z něho, udáváme výsledek v intervalu, protože není možné určit přesnou hodnotu (když se například ptáme 1000 lidí na jejich výšku a průměr nám vyjde 170 nemůžeme jednoduše říct, že i z milionu lidí bude průměr výšky 170). Je ale možné říct,že např. „průměr základního souboru se bude na 100% nacházet v intervalu 150 – 200cm.“ V praxi však často nastane problém, že když určujeme interval, ve kterém se procentně bude nacházet zjišťovaná neznámá, je to interval tak široký, že je nám to na nic. Ale když děláme rozbor preferencí na vzorku 1000 lidí a jako výsledek uvedeme preference Demokratické strany u celé populace jsou na 100% v intervalu 10-20%, je to použitelné. Proto se tyto intervaly uvádějí na přesnost menší než 100%, a to obvykle na 95%, která nám už poskytuje užší interval, při málo změněné věrohodnosti. Tuto přesnost nazýváme **konfidenční interval** a značíme ho jako 1- α = 0,95, přičemž α znamená vlastně možnou chybu odhadu. V případě 1- α = 0,95 je možná chyba 5% Všechny vzorce, které se budou používat budou cca ve tvaru:

P(X<zjišťovaná hodnota<Y)= 1-α

A tedy: S pravděpodobností rovno 1-α se bude zjišťovaná hodnota souboru (základního souboru = průměr anebo rozptyl) nacházet v intervalu (X;Y).

Věc 2: Co je „X“ a „Y“, které jsou ve vzorci výše?

$$P\left(\overline{x} - u_{1-(\alpha/2)}\frac{\delta}{\sqrt{n}} < \mu < \overline{x} + u_{1-(\alpha/2)}\frac{\delta}{\sqrt{n}}\right) = 1 - \alpha$$

Levou a pravou hranici intervalu, ve kterém se bude nacházet zjišťovaná hodnota <u>základního souboru</u> (v případě střední hodnoty značíme μ) tvoří skoro stále korespondující hodnota <u>výběru ze základního souboru</u> (tedy pokud chceme vědět střední hodnotu ZS-používá se tam střední hodnota výběru), od kterého je na levé straně odečtená a na pravé zase přičtena jakási **chyba odhadu**, ve kterém je vždy zakomponovaný kvantil **nějakého rozdělení**, o kterém nemusíme nic moc vědět, akorát to, že ho najdeme v tabulkách kde jakou jeho hodnoty uspořádané podle pravděpodobnosti se kterou interval určujeme teda podle 1-α.

Shrnutí: Máme nějaký statistický soubor, který je tak veliký ,že z něho vybereme několik exemplářů, poznačíme si kolik jsme jich vybrali a zjistíme průměr a rozptyl. My chceme,ale zanalyzovat průměr základního souboru a tedy pomocí toho, že jsme naměřili hodnoty té vybrané skupiny dosadíme naměřená čísla, a pár čísel které najdeme v tabulkách do vzorce a ten nám vypočítá v jakém intervalu se nachází námi hledaná hodnota ZS. Tento výsledek bude přesný podle toho, jaký zvolíme konfidenční interval. Čím vyšší chceme přesnost tím bude interval širší. Proto většinou volíme přesnost 95%, což je kompromis mezi šířkou intervalu a přesností.

Tedy abychom se věnovali příkladům, je jich několik typů:

1. **Zjišťujeme střední hodnotu**
    a. malý výběr prvků ze ZS (n<30)
        aa. poznáme jejich rozptyl
        ab. Nepoznáme jejich rozptyl

b. velký výběr (n>30) a nepoznáme rozptyl

2. zjišťujeme průměr (tím se nebudeme zaobírat)

3. odhadujeme relativní četnosti ZS

V případě, že zjišťujeme střední hodnotu a náš výběr zahrnuje méně než 30 hodnot (např. jsme testovali méně než 30 součástek), použijme vzorce, uvedené na str. 4 pod nadpisem Normální rozdělení, přičemž se řídíme tím zda poznáme anebo nepoznáme rozptyl našeho ZS:

1.aa Zjišťujeme střední hodnotu, malý výběr, poznáme rozptyl,

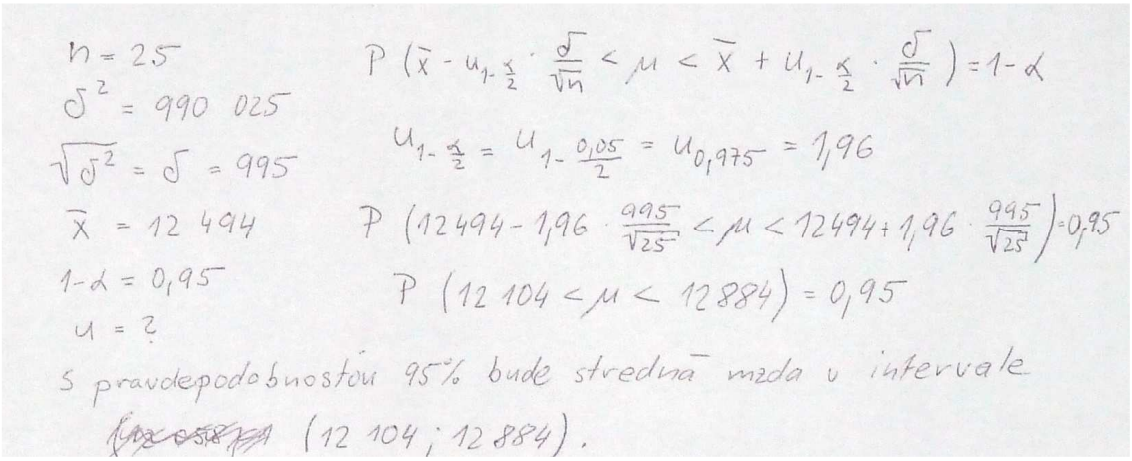
Použijeme vzorec:

$$P\left(\bar{x} - u_{1-(\alpha/2)}\frac{\delta}{\sqrt{n}} < \mu < \bar{x} + u_{1-(\alpha/2)}\frac{\delta}{\sqrt{n}}\right) = 1 - \alpha$$

Příklad:

1. Zjišťujeme střední hodnotu mezd všech absolventek zdravotní školy, přičemž pomocí předešlého zkoumání víme, že její rozptyl je 990 025. Vybrali jsme náhodně 25 absolventek, u kterých jsme zjistili,průměrnou mzdu 12 494 Kč. Sestrojte interval střední mzdy absolventek s přesností 95%.

Řešení: Protože absolventek je jenom 25 je to malý výběr. Celá naše práce se skládá jenom z dosazování do vzorce. Za $\bar{x}$ dosadíme průměrnou mzdu 12 494Kč. Za u musíme dosadit hodnotu příslušného kvantilu, který najdeme v tabulkách (tabulka č. IV. Kvantity normovaného normálního rozdělení). Nejdřív, ale musíme vědět s jakou pravděpodobností počítáme. V zadaní chtějí přesnost 95% a to znamená, že 1- α se bude rovnat 0,95 a alfa 0,05. ve vzorci se ale píše, že dosadíme kvantit $u_{1-(\alpha/2)}$ proto musíme spočítat kolik je 1- (alfa/2). Je to 0,975 a v tabulkách najdeme hodnotu kvantilu příslušející pravděpodobnosti 0,975 a tou je 1,96. Tuto hodnotu dosadíme do výrazu $u_{1-(\alpha/2)}$. Za deltu dosadíme směrodatnou odchylku což je odmocněný rozptyl, tedy 995. A nakonec ještě dosadíme 25 za n. Spočítáme a vidíme v jakých intervalech se bude μ základního souboru nacházet – to je náš výsledek. V případě, že bychom měli rovnou zadanou směrodatnou odchylku, tak samozřejmě rovnou dosadíme za deltu.



1. **ab Zjišťujeme střední hodnotu, malý výběr, nepoznáme rozptyl**,

Když se podíváme do vzorců, zjistíme, že vzorce pro počítání se známým a neznámým rozptylem jsou podobné, nebudeme proto uvádět příklady jen upozorníme na podstatné rozdíly.

1. Když neznáme rozptyl, musíme na jeho místo dosadit něco jiného. Říká se tomu výběrový rozptyl. Značí se s'_x a je to vlastně rozptyl toho našeho výběru. Jsou 2 možnosti, buď si ho musíme spočítat ze zadaných hodnot podle prvního vzorce ze str. 4 ve vzorcích. V případě sestřiček z předešlého příkladu bychom museli znát mzdu každé z nich a pak bychom to dělili podle toho vzorce podobně jako rozptyl na začátku. Příklad takového počítání najdete v příkladu níže. Druhá možnost je, že bude už zadaná a basta.:-)

2. Namísto rozdělení normovaného normálního rozptylu se používá kvantil **t** Studentova rozdělení. Pro nás to znamená, že namísto toho abychom nalistovali v tabulkách při počítání stranu 5, nalistujeme stranu 8, kde hledáme hodnoty tohoto kvantilu přičemž musíme dát pozor, že při tomto rozdělení záleží i na velikosti výběru – **n,** a teda musíme najít správný řádek podle velikosti n (ten sloupec označený v).

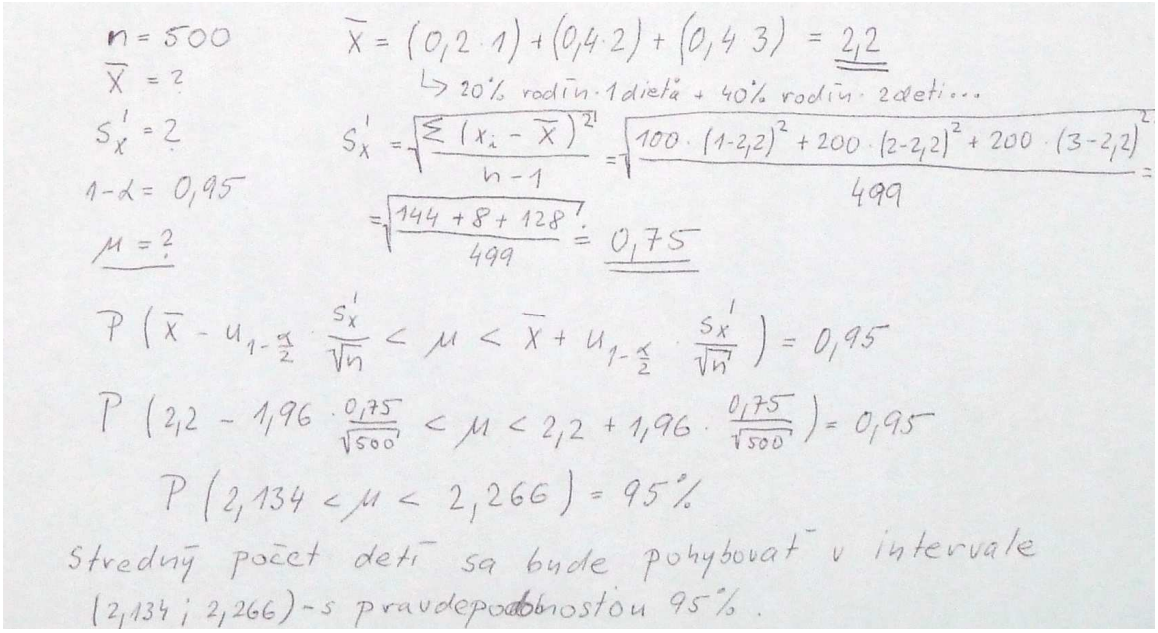
1. b Zjišťujeme střední hodnotu, velký výběr, neznáme rozptyl

Opět je to velmi podobné jako v předešlých příkladech, akorát že výběr je velký, teda je v něm víc než 30 členů. V dalším příkladě zároveň musíme spočítat výběrový rozptyl, i když často se vyskytují příklady kde je rovnou dané čemu se rovná s'_x**.**

Příklad:

1. Na základě uvedené tabulky zaznamenávající pro 500 sledovaných rodin počet dětí a počet místností sestrojte 95% oboustranný interval spolehlivosti pro střední počet dětí.

Počet dětí v rodině	1	1	2	2	2	3	3
Počet místností	1	2	1	2	3	2	3
Početnost rodin (%)	10	10	10	20	10	20	20



Nejdřív si spočítáme střední hodnotu výběru (výběr je těch 500 rodin = n), která znamená kolik „dětí připadá na jednu rodinu“. Dále spočítáme výběrový rozptyl, jako je to

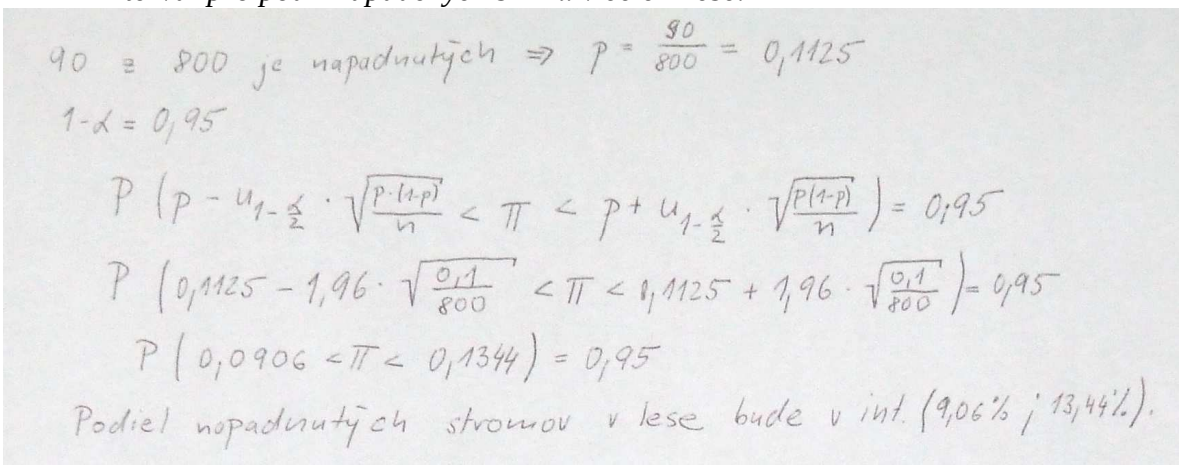
v demonstrovaném řešení. V tabulkách najdeme kvantil pro pravděpodobnost 0,975. Dosadíme do vzorce a hotovo.

3. Odhadujeme relativní četnost základního souboru

Relativní četnost můžeme odhadovat když je náhodná veličina rozdělená alternativně. Není to žádná věda, prakticky to znamená, že ***může nabývat jen dva stavy***, tedy buď je jablko zdravé nebo zkažené, buď se narodí kluk nebo holka, atd. Zároveň je daná pravděpodobnost s jakou jedna nebo druhá možnost nastanou (je jasné, že „pravděpodobnost že nastane jeden jev“ + “pravděpodobnost že nastane druhý“ = 1; protože jeden z nich určitě nastane).
Odhad relativní četnosti znamená že v zadání je řečeno jaká část souboru má nějakou vlastnost (př. je zkažená) a my z toho máme vypočítat kolik % základního souboru bude mít tuto vlastnost, samozřejmě s nějakou pravděpodobností, většinou 95%.
V příkladech tedy opět nejde o nic jiného, jen správně identifikovat to, že je to právě tento typ příkladu - použít správný vzorec a správně do něj dosadit.

Příklad:

1. *Průzkumem se zjistilo, že 90 z 800 smrků je napadených kůrovcem. Zjistěte 95% interval pro podíl napadených smrků v celém lese.*

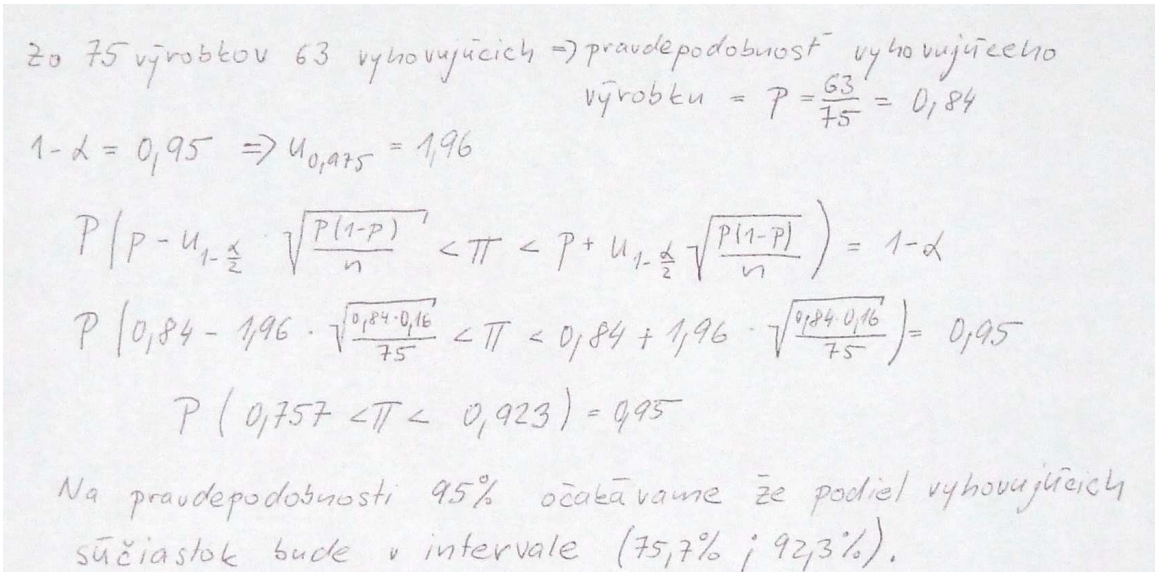


Nejprve si musíme ujasnit pravděpodobnost, jsou 2 možnosti, buď je strom napadený nebo ne. Pravděpodobnost, že je napadený vyplívá z našeho výběru a tedy 90/800=0,1125. Teď se podívejme na vzorec.

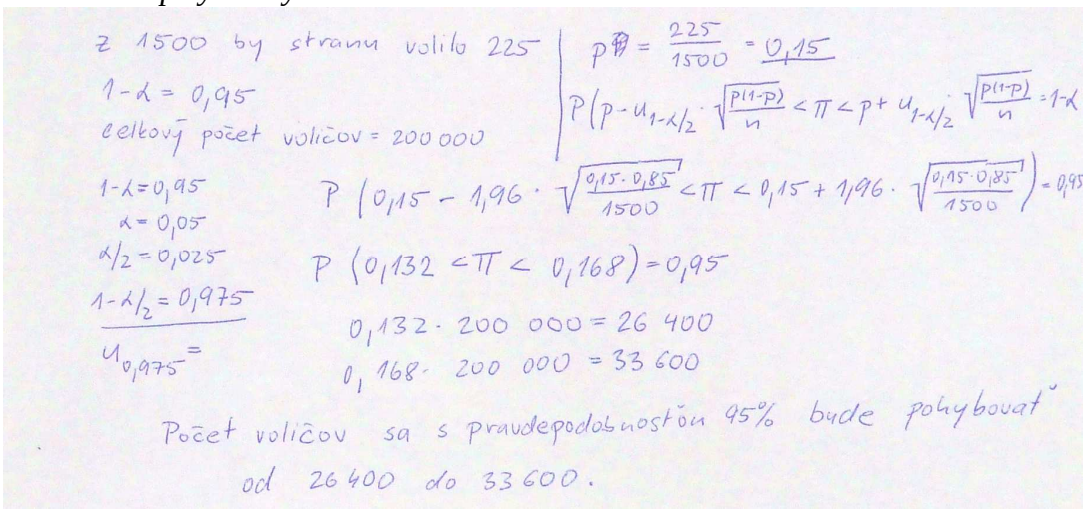
$$P\left(p - u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \pi < p + u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right) = 1-\alpha$$

Jak vidíme, dosadíme jen „p“ což je pravděpodobnost, že je strom napadený, potom příslušný kvantil „u“, který najdeme v tabulkách a „n“, tedy počet členů výběru. Je třeba si uvědomit, že když by se ptali na podíl **ne**napadených stromů v lese,tak by jsme museli vypočítat pravděpodobnost, že je strom nenapadnutý a ti potom dosadit za p.

2. Mezi 75 kontrolovanými výrobky mělo 63 vyhovující jakost. Sestrojte 95% interval spolehlivosti pro podíl vyhovujících výrobků.

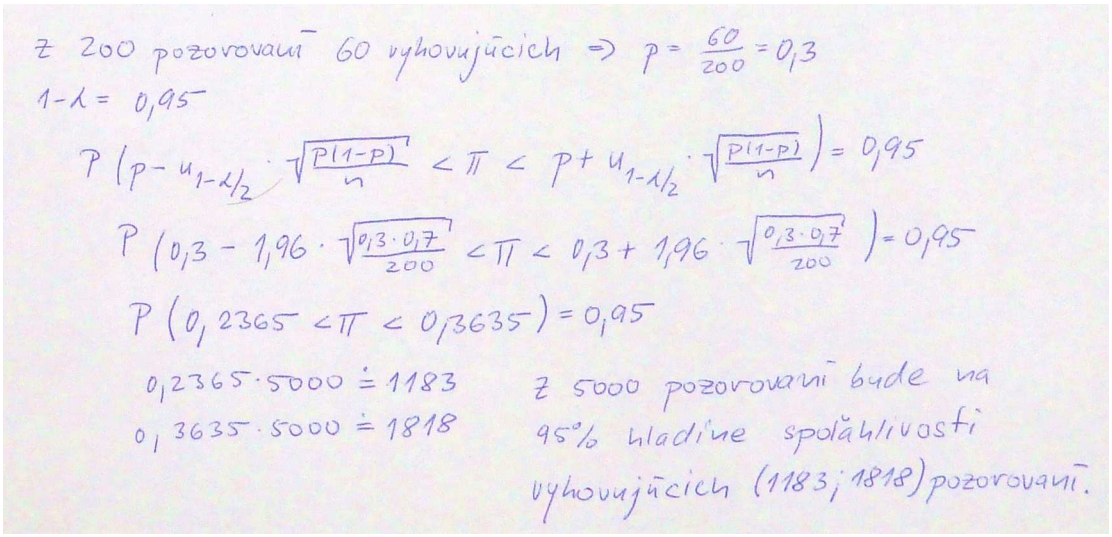


3.Z 1500 náhodně dotazovaných dospělých obyvatel města by určitou stranu volilo 225. Odhadněte se spolehlivostí 0,95 počet potenciálních voličů této strany ve městě, ve kterém žije 200 000 dospělých obyvatel.



Princip zůstává stále stejný akorát na závěr, vynásobíme počet obyvatel %, které nám vyšli. Obdobný příklad z novějších testů:

4.Z 200 pozorování bylo 60 vyhovujících. Sestrojte oboustranný interval a s 95% přesností odhadněte kolik bude vyhovujících pozorování z 5000.



Testování hypotéz

Co vlastně je to testování hypotéz je vcelku srozumitelně vysvětleno v knížce na str. 133, doporučuji si to přečíst, protože my to tu proběhneme dost zjednodušeně a povrchně.

Hypotéza je předpoklad o něčem. Věta „Zítra v poledne bude 22 stupňů celsia“ by se také dala považovat za hypotézu. Naším úkolem ve statistice je danou hypotézu ověřit a vyjádřit její pravděpodobnost (otestovat ji). Proto máme stále zadanou (nebo si musíme sami vymyslet podle zadání příkladu) tzv. nulovou hypotézu, značíme H0. Touto nulovou hypotézou je hypotéza, kterou testujeme, tedy o které chceme rozhodnout zda je pravdivá. Proti této hypotéze postavíme hypotézu H1, která musí původní hypotézu popírat.
Např. když výrobce lentilek garantuje, že v balíčku jich je 40, bude to naše H0. Proti ní musíme postavit nějakou jinou, která to popírá. Nejčastěji je to jednoduché popření H0, tedy „V balíčku není 40 lentilek“. Takže je to prakticky negace H0 značíme „H1=nonH0“. Když však vezmeme v úvahu specifikum příkladu a tedy není špatně když je v balíčku lentilek víc, můžeme sestavit i jinou hypotézu H1 a to že „lentilek je v balíčku méně než 40“.

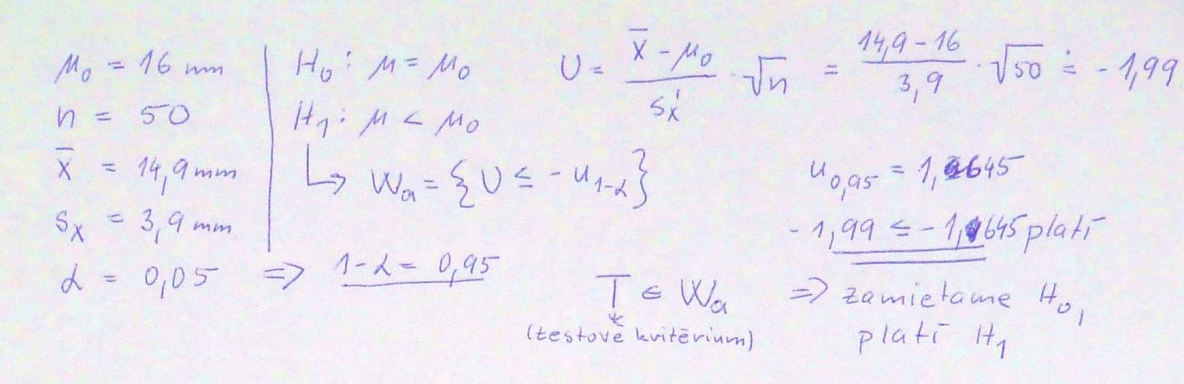
To by bylo něco Nového na úvod, ale když chceme tyto věci prakticky spočítat, je to skoro opakování. Hlavní je pochopit systém, který spočívá v práci se vzorci a tabulkami. Otevřeme vzorce na str. 5. Na ní najdeme tabulky, velmi rozumně rozdělené na 3 sloupce, přičemž:

1. V prvním sloupci vždy najdeme **hypotézu H0**, kterou chceme testovat. Jenže ve statistice budeme testovat jen věci typu: „výrobce udává takovouto střední hodnostu blabla…“ a „rozptyl mezd je takový a takový…a je jaký je teď?“, a tedy to bude stále nějaký předpoklad o velikosti střední hodnoty či rozptylu a naše H0 bude stále velikost **je** taková jako udává výrobce, že **je** taková jak se předpokládá. Hned vedle si vybereme jednu z formulací H1, tedy buď úplně zamítneme H0, řekneme že střední hodnota se nebude rovnat té kterou předpokládáme ($\mu \neq \mu_0$), jde napsat i „non H0“, nebo jen řekneme, že bude vyšší či nižší.
2. Ve druhém sloupci máme **testované kritérium**. To je rovnice ve které máme na levé straně nějakou neznámou, kterou potřebujeme vypočítat (př. U,t) a do pravé strany by jsme měli dosázet proměnné podle zadání, popř. z tabulek. Když vypočítáme proměnnou na levé straně můžeme přikročit ke 3. sloupci.
3. Ve 3. sloupci se nachází **Kritický bod (KB)**. Označuje se **Wa.** Pokud **nepatří proměnná**, kterou jsme vypočítali ve sloupci 2 **do tohoto kritického bodu** potom

je platná hypotéza H0. Pokud naopak **proměnná patří, tak hypotéza H0 není platná a platí H1**. Jak vidíme kritické body jsou většinou 3 (3rovnice Wa=něco), ty jsou pod sebou seřazené podle toho jakou hypotézu H1 jsme zvolili. Pokud jsme použili první H1, koukneme se na 1 kritický bod. Pokud 2.H1 tak na 2 KB a ostatní nás nezajímají. KB vždy porovnává vypočítané proměnné ze 2. sloupce s nějakým kvantilem, který najdeme v tabulkách. Když nerovnost ve složených závorkách platí, značí to, že proměnná patří do kritického oboru = je neplatná hypotéza H0.

Určitě to bude absolutně jasné po příkladu:

1. *Odchylka délky součástek od normy je průměrně 16mm, po změně technologie bylo náhodně vybraných 50 a zjištěná odchylka byla 14,9mm se směrodatnou odchylkou 3,9mm. Otestuj na hladině významnosti 5% hypotézu, že technologie snižuje průměrnou velikost odchylky*



Nejdřív si musíme uvědomit co vlastně počítáme. Když porovnáváme střední hodnotu odchylky součástek od normy před a po technologické změně, použijme buď první a nebo druhou tabulku ze vzorců na straně 6. Vybereme si druhou, protože první se používá pro výběr < 30 a my jsme vybrali 50 součástek. Pak zformujeme nulovou hypotézu. Možnost, jak je vidět z tabulky, máme jenom jednu: $\mu = \mu_0$, přičemž za μ_0 vždy považujeme původní střední hodnotu, našem případě tu před změnou technologie. Tedy naše H0 říká, že nový průměr μ (po změně technologie) těch odchylek od normy je stejný jako byl původně. My máme zjistit, zda technologie snižuje průměrnou odchylku. Proto naší alternativní hypotézou H1 bude, že průměrná velikost odchylek po změně bude menší než předtím. Tedy H1 = $\mu < \mu_0$. Teď přejdeme do 2. sloupce a dosadíme do vzorce. $\bar{x}$ je vždy průměr po změně (je to průměr výběru, zatím co μ, které zjišťujeme, je průměr celého souboru po změně technologie) μ_0 zas průměr původního souboru. Za výběrový rozptyl dosadíme směrodatnou odchylku a za n počet prvků výběru. Vypočítáme U (výsledek tohoto testovaného kritéria se často označuje všeobecně T) a potom ještě najdeme v tabulkách správný kvantil pro pravděpodobnost 0,95 a po porovnání vidíme že uvedená nerovnost (2. shora, ptz jsme vybrali druhou H0) platí. Když nerovnost platí, jsme v kritickém oboru. To znamená, že platí hypotéza H1. Střední hodnota odchylek po změně technologie je skutečně nižší než před změnou.

2. velmi podobný příklad najdete v Aplikacích na str. 132/pr. 6
3. Dalším ukázkovým příkladem na tuto problematiku je př. č. 8 také na str. 135, kde musíme zase použít vzorec ze 3. tabulky, ptz nám jde o výpočet rozptylu. Výrobce

tvrdí, že směrodatná odchylka je 0,9, my chceme dokázat, že tento údaj není pravdivý a tedy, že směrodatná odchylka je větší (menší nám nevadí)

Chí - kvadrát test dobré shody

Tato úloha je tak rozšířená v testech, že jí věnujeme vlastní nadpis☺. K jejímu řešení nám bude pomáhat tabulka na str. 6 ve vzorcích. Celý test spočívá v tom, že na začátku jsou zadané nějaké předpoklady, tedy většinou pravděpodobnosti jak se něco odhaduje (př. prodej CD Kryštofa se odhaduje 30% objemu mužům a 70% ženám). Potom máme nějaká konkrétní čísla (př. o prodeji) a my v tomto testu porovnáváme předpoklad a skutečnost a rozhodneme zda odchylka od předpokladu vznikla vinou chybného předpokladu nebo či je jen náhodná.

Ukázkový příklad:

1. *Kostkou jsme házeli 30krát a výsledky jsou zaznamenány v tabulce. Rozhodněte na pravděpodobnost 95% , zda je kostka spolehlivá.*

X	1	2	3	4	5	6
N	4	6	7	2	5	6
o	5	5	5	5	5	5

V prvním řádku je X, číslo, které padlo na kostce. Ve 3. řádku je o – předpoklad, předpokládá se totiž, že každé jedno číslo padne ze 30ti pokusů právě 5krát, to by bylo ideálně přesně podle pravděpodobnosti. Ve 2. řádku je reálná hodnota, kolikrát které číslo padlo. Jak vidíme ideální to není, ale naším úkolem je zjistit, zda je to jen náhodná odchylka nebo je kostka nespolehlivá.

Když se teď podíváme na vzorec, nejprve si všimneme hypotézy. Nulovou hypotézou bude stále že $\pi = \pi_0$, tedy že předpoklad se splnil (odchylka je náhodná), druhou hypotézou - H1 je „non H0“ a tedy že předpoklad se nesplnil (odchylka má nějakou příčinu, anebo byl špatný odhad). Hypotézy jsou tedy dané a můžeme vypočítat testové kritérium, které je v tomto případě chí-kvadrát. Nejedná se o nic zvláštního, akorát že výsledek porovnáváme podle nerovnice ze 3. sloupce s kvantilem z jiné tabulky. Vzorec trochu zjednodušeně, prakticky je to to stejné, ale líp se na to kouká:

$$\chi^2 = \sum \frac{(n-o)^2}{o}$$

Přičemž **„n“ je skutečnost** a **„o“ je dohad**. Teď už jen doplníme čísla, spočítáme všechny výsledky dohromady a máme výsledek, který můžeme porovnat s tabulkami na str. 7 . **Pozor**, pro chí-kvadrát je specifické, že řádek v tabulkách vybereme podle toho, kolik máme $\mathbf{x_i}$.

(tedy kolik je možností, možných odpovědí…) avšak musíme od toho odčítat jednotku. My máme 6 možností, které mohou padnout na kostce, podíváme se tedy do 5. řádku. Zase postupujeme tak, že když naše testované kritérium patří do Wa, zamítneme hypotézu H0.

1.

x_i	1	2	3	4	5	6
n_i	4	6	7	2	5	6
o_i	5	5	5	5	5	5

1. H_0: kostka je spravedlivá (shoda)
2. H_1: není spravedlivá

3. $T = \sum \frac{(n_i - o_i)^2}{o_i} =$

$$\frac{(4-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(7-5)^2}{5} + \frac{(2-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(6-5)^2}{5} = \frac{1+1+4+9+0+1}{5} = \frac{16}{5}$$

$$= 3{,}2$$

4. $W = \left(\chi^2_{1-\alpha}(5), \infty\right) = \left(11{,}07; \infty\right)$

Jak vidíme, testové kritérium vyšlo 3,2 avšak 0,95 kvantil chí-kvadrát rozdělení je 11,07 a podmínka je, aby testové kritérium vyšlo vyšší než kvantit. Proto nejsme v kritickém oboru a můžeme potvrdit hypotézu H0 a kostku označit za spolehlivou, protože její odchylka je jen náhodná.

Příklady:

1. *Marketingový plán tvrdil, že záznam koncertu skupiny Stupid Kids se prodá v průměru 50% CD ku 30% DVD a 20% kazety. Za měsíc se skutečně prodá 2552 kusů CD, 923DVD a 384ks kazet tohoto koncertu. Ověřte zda byl předpoklad marketingového plánu správný.*

1.

Nosič	CD	DVD	MC
x_i	2552	923	384
0_i	1930	1158	771
$x_i - a_i$	622	-235	-387

Celkem: 2552 + 923 + 384 = 3859

1. H_0: předpoklad byl správný (shoda)
2. H_1: předpoklad nebyl správný

3. $T = \sum \frac{(x_i - a)^2}{0_i} = \frac{622^2}{1930} + \frac{(-235)^2}{1158} + \frac{(-387)^2}{771} =$

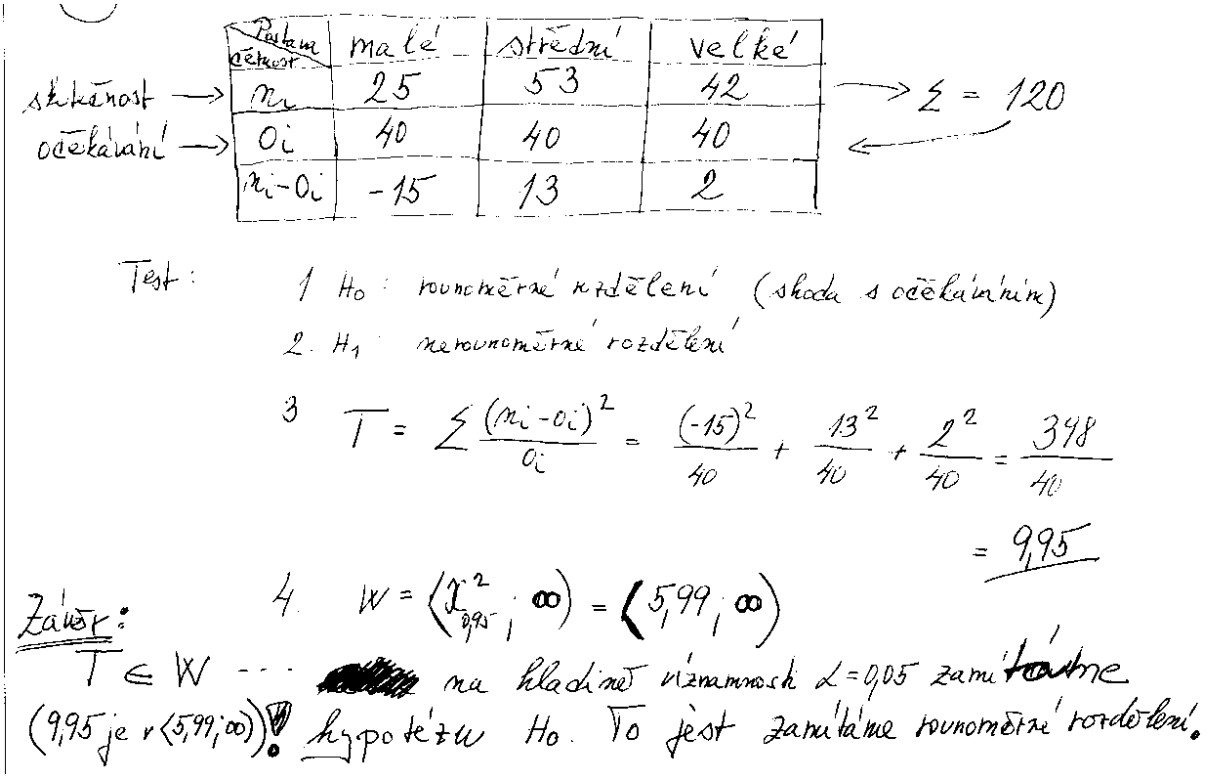
$= 200{,}46 + 47{,}69 + 194{,}25 = 442{,}4$

4. $W = \left(\chi^2_{(2)}_{0,95}; \infty\right) = (5{,}99; \infty)$

Závěr: $T \in W$ -- zamítáme na $\alpha = 0{,}05$ hypotézu H_0, tj. zamítáme, že by předpoklad byl správný.

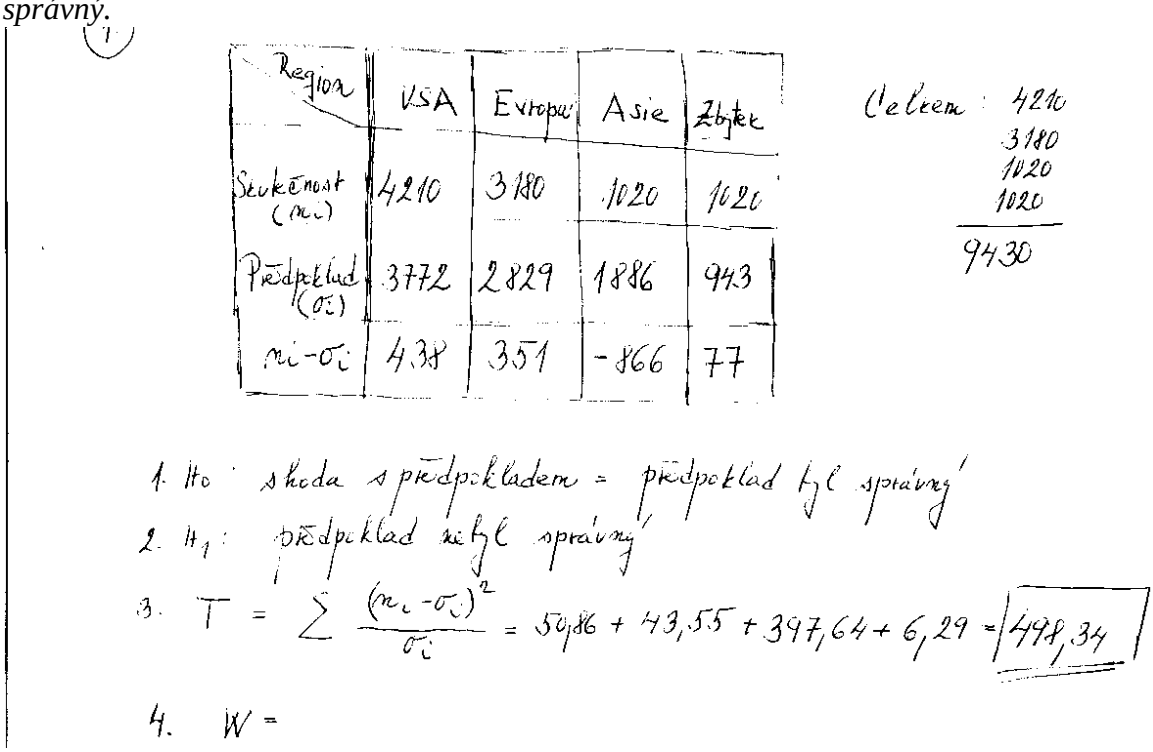
Nejprve si sestavíme tabulku, v 1. řádku je skutečnost kolik kusů různých nosičů bylo prodáno. Když spočítáme celkový prodej a vynásobíme příslušným procentem, které bylo uváděno v marketing. plánu, vyjde nám, jaký byl předpoklad v kusech , tj. druhý řádek tabulky. Ve 3. je už jen vypočítaný rozdíl pro lepší počítání. Dosadíme čísla do vzorce, vypočítáme testové kritérium. Když budeme hledat kvantit chí-kvadrát rozdělení v tabulkách, nesmíme zapomenou, že koukáme do druhého řádku, protože máme tři možné nosiče. Pro jednoduchost můžeme říci, že řádek chí-kvadrátu odvodíme podle počtu sloupců v tabulce minus 1. Pravděpodobnost není daná, tak požíváme standardní 0,95.

2. Při náhodném průzkumu bylo 25 lidí označených jako osoby malé postavy, 53 jako osoby střední postavy a 42 velké postavy. Ověřte na 5% hladině tvrzení o rovnoměrném podílu velikosti.



Tady zase tabulka, skutečnost je jasná, předpoklad, že jsou lidé podle velikosti rovnoměrně rozděleni, tedy ze 120 lidí by jsme měli mít po 40 z každé velikosti

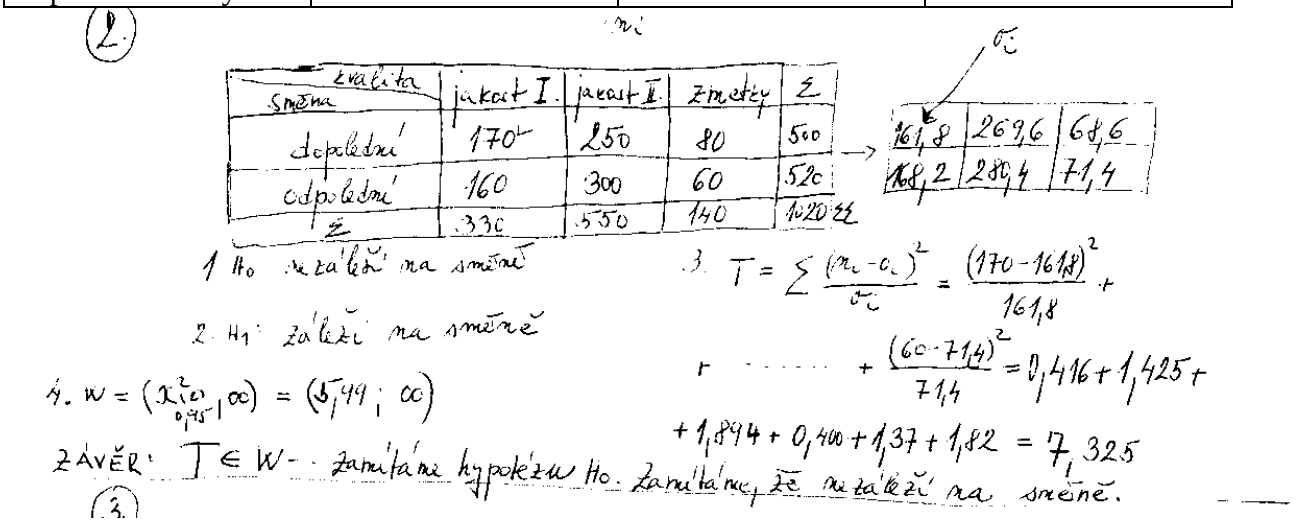
3. Marketingový plán tvrdil, že pračky DW půjdou na odbyt v poměru 40% USA, 30% Evropa, 20% Asie a 10% zbytek světa. Po 1. týdnu je v USA prodaných 4210 praček, v E. 3180, v Asii 1020 a ve zbytku světa také 1020. Otestujte, zda je původní předpoklad marketingového plánu správný.



Kvantil a výsledek si můžete dopočítat sami…☺ Čtyři sloupce mínus 1 je 3.

4. Na základě údajů z tabulky určete,zda kvalita výrobků závisí na tom která směna je vyráběla.

počty výrobků	jakost I.	jakost II	zmetky
vyrobených během dopolední směny	170	250	80
vyrobených během odpolední směny	160	300	50



Prakticky stejná úloha jako ty předešlé, akorát je potřeba dát pozor na to jak spočítáme odhadované počty výrobků za předpokladu, že směny vyrábějí výrobky stejné kvality. Tyto výpočty uděláme tak, že např. celkový počet výrobků první jakosti (330) vydělíme celkovým počtem vyrobených výrobků (1020) , tím pádem dostaneme podíl I. Jakosti na vyrobených součástkách. Tento podíl potom vynásobíme počtem všech součástek vyrobených první směnou (500) a dostaneme odhadovaný počet výrobků první jakosti pro první směnou. Podobně vynásobíme tento podíl počtem všech součástek druhé směny (520). A potom pokračujeme na druhou jakost atd.

5. Otestuj na 5% hladině významnosti předpoklad o nezávislosti odpovědí na pohlaví.

Pohlaví	Ano	Ne
Muž	25	40
Žena	35	40

Když by byly odhady nezávislé na pohlaví musíme to zase přepočítat stejným způsobem jako v předešlém případě. Tedy odpovědi ANO dohromady (25+35=60), vydělit celkovým počtem odpovědí (140) a nakonec vynásobit celkovým počtem mužů…atd.,atd. Dva sloupce, používáme tedy první řádek chí-kvadrátu z tabulek.

(2)

Odpověď / Pohl.	ANO (n_i)	NE	Σ
Muž	25	40	65
Žena	35	40	75
Σ	60	80	140 (ΣΣ)

⇒

(o_i)	
27,86	37,14
32,14	42,86

1. H_0: odpověď nezávisí na pohlaví
2. H_1: závisí na pohlaví
3. $T = \sum \frac{(n_i - o_i)^2}{o_i} = \frac{(25-27{,}86)^2}{27{,}86} + \dots + \frac{40-42{,}86}{42{,}86} = 0{,}294 + 0{,}220 + 0{,}254 + 0{,}191 \doteq 0{,}96$
4. $W = \left(\chi^2_{0{,}95}(1); \infty\right) = (3{,}841; \infty)$

Závěr: $T \notin W$ … na $\alpha = 0{,}05$ nelze zamítnout hypotézu H_0, tj. nelze zamítnout, že odpověď nezávisí na pohlaví.

6. Rozdělíme firmy podle několika kategorií podle jejich velikosti na velké, střední a malé. Průzkumem bylo zjištěno, že mezi 40 velkými firmami jich exportuje 15, 75 středními exportuje 25 a ze 60ti malých exportuje přesně polovina. Rozhodněte vhodným testem (na 5% hladině) zda je závislost mezi velikostí firmy a tím zda exportuje nebo ne.

(2)

Odpověď / Pohl.	ANO (n_i)	NE	Σ
Muž	25	40	65
Žena	35	40	75
Σ	60	80	140 (ΣΣ)

⇒

(o_i)	
27,86	37,14
32,14	42,86

1. H_0: odpověď nezávisí na pohlaví
2. H_1: závisí na pohlaví
3. $T = \sum \frac{(n_i - o_i)^2}{o_i} = \frac{(25-27{,}86)^2}{27{,}86} + \dots + \frac{40-42{,}86}{42{,}86} = 0{,}294 + 0{,}220 + 0{,}254 + 0{,}191 \doteq 0{,}96$
4. $W = \left(\chi^2_{0{,}95}(1); \infty\right) = (3{,}841; \infty)$

Závěr: $T \notin W$ … na $\alpha = 0{,}05$ nelze zamítnout hypotézu H_0, tj. nelze zamítnout, že odpověď nezávisí na pohlaví.

Když export není závislý na velikosti bude každá firma exportovat stejně, bez ohledu na to do jaké kategorie spadá. Musíme proto opět přepočítat odhad tak,aby vyjadřoval export bez ohledu na velikost, jen s ohledem na poměr v jakém se zúčastnili statistického měření.

ANALÝZA ROZPTYLU – ANOVA (kapitola z novej SPF)

Analýzu rozptylu využíváme v případě, že máme **víc druhů testovaného předmětu, a zároveň každý druh testujeme vícekrát**. Například máme tři druhy benzínu a testujeme jeho spotřebu v pěti jízdách pro každá druh. Je jasné, že každá jízda se bude ve spotřebě lišit, protože spotřebu paliva ovlivňuje mnoho faktorů. Každopádně se bude spotřeba v pěti jízdách pro jeden druh benzínu točit okolo nějaké střední hodnoty a bude mít nějaký rozptyl. Tento rozptyl spotřeby benzínu jednoho druhu v pěti jízdách, rozptyl jedné skupiny, se nazývá **vnitroskupinový rozptyl** - $S_{y,v}$. Jednotlivé skupiny (druhy benzínů) mají svůj celkový průměr, který jsme vypočítali z těch pěti měření. Ale i tyto průměry se budou u každé skupiny pravděpodobně lišit a můžeme stanovit jejich střední hodnotu a rozptyl. Tento rozptyl, protože nám něco říká o kolísání hodnot mezi jednotlivými druhy benzínu, mezi jednotlivými skupinami se nazývá **meziskupinový rozptyl** - $S_{y,m}$. A nakonec **celkový rozptyl** – tedy mezi skupinami i uvnitř skupin se vypočítá jako jednoduchý součet: $S_y = S_{y,v} + S_{y,m}$. Pokud počítáme příklady, jedná se většinou o **testování hypotézy, že střední hodnoty skupin se rovnají**. Takže nulová hypotéza bude předpokládat, že nejsou rozdíly mezi jednotlivými druhy, proti čemuž postavíme alternativní hypotézu, která popírá nulovou – tedy, že tam nějaký rozdíl je, že alespoň jedna střední hodnota se odlišuje a není to jen statistická odchylka. Testovým kritériem je F-test, jehož výsledek zase jen porovnáme s kvantilem z tabulek a podle toho, zda proměnná patří do kritického oboru, zamítneme jednu z hypotéz.

1. *Zkoumali jsme tři druhy benzínu a u každého jsme udělali 5 měření spotřeby. Meziskupinový rozptyl jsme vyčíslili na 0,25 a vnitroskupinový se rovná 0,08. Ověřte hypotézu, že se spotřeby u těchto třech druhů rovnají.*

$S_{y,m} = 0,25 \qquad S_{y,v} = 0,08 \qquad k = 3 \qquad n = 15$

$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad W_a = \{F \geq F_{1-\alpha}\}$

$H_1 : non\ H_0 \qquad F_{0,95}(2;12) = 3,885$

$$F = \frac{\frac{S_{y,m}}{k-1}}{\frac{S_{y,v}}{n-k}} = \frac{\frac{0,25}{2}}{\frac{0,08}{12}} = 18,75 \qquad 18,75 \geq 3,885 \Rightarrow \textit{platí}$$

Zamítáme hypotézu H_0.

Proměnnou *k* je počet skupin – v našem případě 3 – a proměnnou *n* je počet měření. Pozor, proměnná „počet měření" je myšlená jako **součet počtu měření v každé skupine**. V našem případě v každé ze skupin bylo 5 měření, tedy celkový počet měření (n) je 5+5+5=15. Jediný problém snad může nastat při hlednání hodnoty kvantilu v tabulkách, protože F-rozdělení má stupně volnosti. Sloupec v tabulce určíme vypočítáním $\upsilon_1 = k - 1$. Řádek, ve kterém najdeme hledanou hodnotu, určíme vypočítáním $\upsilon_2 = n - k$.

2. *Zkoumali jsme odrůdy brambor a udělali dohromady 28 měření. Z nich jsme zjistili, že vnitroskupinový rozptyl je dvakrát větší než meziskupinový a testové kritérium nám vyšlo F=4. Zjistěte, kolik odrůd jsme zkoumali na hladině pravděpodobnosti $\alpha = 0,05$ zamítneme nulovou hypotézu.*

$$F = 4 \qquad n = 28 \qquad S_{y,v} = 2 \cdot S_{y,m} \qquad k = ?$$

$$F = \frac{\frac{S_{y,m}}{k-1}}{\frac{S_{y,v}}{n-k}} \qquad 4 = \frac{\frac{S_{y,m}}{k-1}}{\frac{2 \cdot S_{y,m}}{28-k}}$$

$$4 = \frac{S_{y,m}}{k-1} \cdot \frac{28-k}{2 \cdot S_{y,m}}$$

$$4 = \frac{1}{k-1} \cdot \frac{28-k}{2} = \frac{28-k}{2k-2}$$

$$8k - 8 = 28 - k$$

$$9k = 36$$

$$k = 4$$

Jednoduchými matematickými úpravami a substitucí vyplývající ze vztahu meziskupinového a vnitroskupinového rozptylu jsme přišli k počtu zkoumaných odrůd. Nyní můžeme potvrdit nebo vyvrátit hypotézu. Testové kritérium máme zadané, stačí jen najít správný kvantil F-rozdělení.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \qquad W_a = \{F \le F_{1-\alpha}\}$$

$$H_1 : non\ H_0$$

$$F_{0,95}(k-1; n-k) = F_{0,95}(3; 24) = 3,008$$

$$4 \le 3,008 \Rightarrow neplatí$$

Hodnota testového kritéria nepatří do kritického oboru – zamítám hypotézu H_1. Rozdíl ve středních hodnotách naměřených veličin se neprokázal.

3. *Dopočítejte chybějící hodnoty do tabulky, udělejte test včetně zapsání hypotéz a vyvoďte závěry, vhodným ukazatelem změřte sílu závislosti a závěr též okomentujte.*

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	456, 1333333	228, 0666	?	<.0001
Error	12	23, 33333333	?		
Corrected Total	14	?			

Co se týká doplňování hodnot do tabulky, pomůže nám vysvětlující tabulka:

Zdroj	Stupně volnosti	Součet čtverců	Průměr čtverců	Testové kritérium F	P - hodnota
Model	k-1	Sy,m	Sy,m/(k-1)	F	P - hodnota
Error	n-k	Sy,v	Sy,v/(n-k)		
Corrected Total	n-1	Sy			

Když víme, které číslo co znamená, neměl by být problém doplňovat čísla. První otazník (zleva) doplníme na základě vztahu $S_y = S_{y,v} + S_{y,m}$ = 479,466. Druhý otazník dostaneme vydělením součtu čtverců stupni volnosti = 23,3333/12 = 1,9444. Třetí otazník je testové kritérium, které musíme vypočítat:

$$F = \frac{\frac{S_{y,m}}{k-1}}{\frac{S_{y,v}}{n-k}} = \frac{\frac{456,1333}{2}}{\frac{23,3333}{12}} = \frac{228,0666}{1,9444} = 117,294$$

Dále máme udělat test, neboli stanovit hypotézy a porovnat vypočítanou hodnotu testového kritéria F s kritickým oborem:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad F = 117,294$$
$$H_1 : non\ H_0 \qquad W_a = \{F \geq F_{1-\alpha}\}$$
$$F_{1-\alpha}[k-1; n-k] = F_{0,95}(2;12) = 3,885$$
$$117,294 \geq 3,885 \Rightarrow \textit{platí}$$

Testové kritérium patří do kritického oboru, proto zamítáme nulovou hypotézu a konstatujeme, že střední hodnoty zkoumaných předmětů se liší. Změřit sílu závislosti můžeme koeficientem determinace (podrobněji viz strana 44):

$$P^2 = \frac{S_{y,m}}{S_y} = \frac{456,1333}{479,4666} = 0,95133$$

Dokázali jsme silnou závislost a model můžeme označit za kvalitní.

REGRESNÍ PŘÍMKA

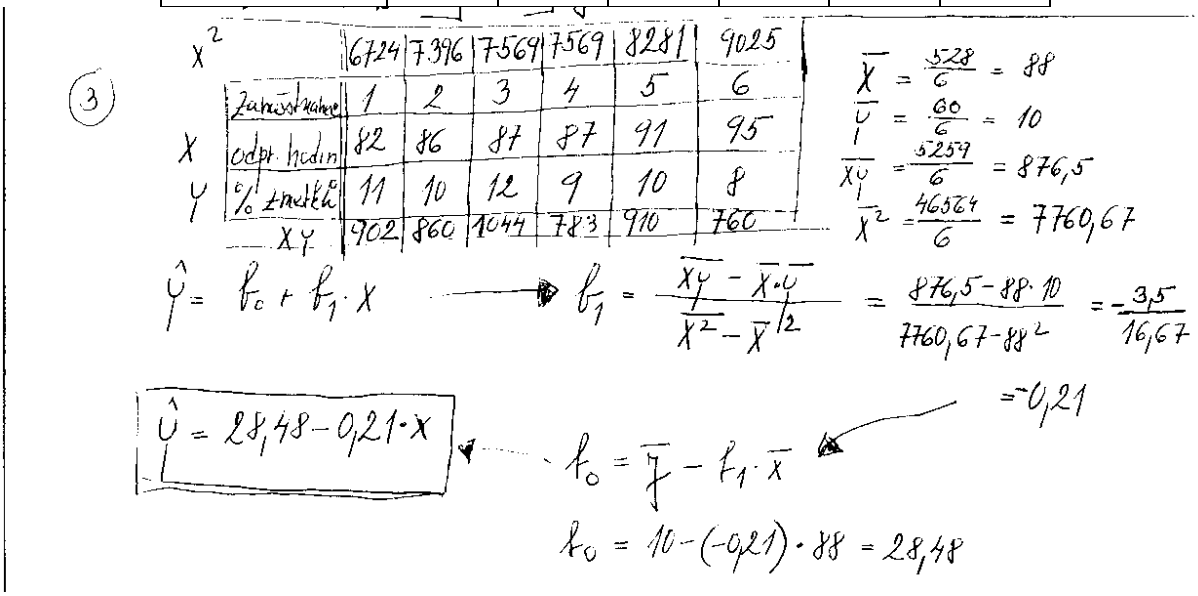
Určení regresní přímky

Co se regrese týká, pro testy je prakticky použitelná akorát tak regresní přímka a korelační koeficient, proto se budeme zaobírat jen jimi. Regresní přímka udává závislost Y na X, tedy např. závislost ceny auta na jeho věku, závislost prodeje zimní obuvi na množství napadaného sněhu a podobně. Tato závislost může být buď přímá (čím více sněhu, tím víc prodané obuvi) nebo i nepřímá (čím vyšší věk, tím nižší cena auta). Když chceme matematicky vyjádřit lineární regresní přímku používáme rovnici $y = \beta_0 + \beta_1 x$, která má dva parametry beta 0 a beta 1, přičemž vidíme, že B0 je jakási pevná složka (tedy auto bude stále něco stát, ať bude jakkoliv staré) a B1 se bude měnit v závislosti od x. Zároveň je jasné, že když bude B1 kladná tak se zvyšujícím se x se bude zvyšovat i y, půjde tedy o přímou závislost, naopak když B1 bude záporná, půjde o závislost nepřímou. Otázka tedy je, jak se dopracovat k hodnotám parametrů B0 a B1. Nejlépe je to vidět na příkladu:

1. Zaměstnanci firmy se zapracovávají na nové výrobní lince. Pro 6 zaměstnanců je zaznamenávaný počet dosud odpracovaných hodin (veličina X) a zjištěný procentuální podíl

chybných výrobků (Y). Určete regresní přímku (tj. hodnoty jejích parametrů) závislosti Y na X, interpretujte co nejvýstižněji hodnotu směrnice.

zaměstnanec č.	1	2	3	4	5	6
odprac. hodin	82	86	87	87	91	95
% zmetků	11	10	12	9	10	8



Vidíme, že počet chybných výrobků by měl záviset na odpracovaných hodinách na nové lince, označíme tedy počet odpracovaných hodin jako X a počet chyb jako Y (Y je vždy závislá proměnná, tedy ta které velikost závisí na velikosti X). No a teď si vybereme nějaký vzorec ze strany 8 na výpočet B1. Autor řešení používá např. $\frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - \overline{x}^2}$, tedy stále vypočítáme průměry X,Y, součin jejich průměrů, průměr druhé mocniny a druhou mocninu průměru, které potom dosadíme do vzorce. V případě B0 (první vzorec na str. 9) jen dosadíme už vypočítaná čísla. **Směrnice regresivní přímky** se označuje jako parametr B1, přičemž na základě toho, zda je znamínko před ním plus nebo mínus můžeme určit, zda je závislost přímá nebo nepřímá. V tomto případě je před směrnicí znaménko **minus**, což značí že **závislost je nepřímá.** Tedy čím více hodin pracovníci na pracovní lince odpracují, tím méně chyb udělají. Jako výsledek zapíšeme hotovou rovnici regresní přímky ve tvaru $y = \beta_0 + \beta_1 x$ a vyjádříme se o typu závislosti.

Pokud, ale chceme změřit i sílu jakou jsou X a Y na sobě závislé, musíme použít tzv. **korelační koeficient**. Jeho vzorec najdeme ve vzorcích na straně 7:

$$r_{yx} = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sqrt{\left(\overline{x^2} - \overline{x}^2\right)\left(\overline{y^2} - \overline{y}^2\right)}}$$

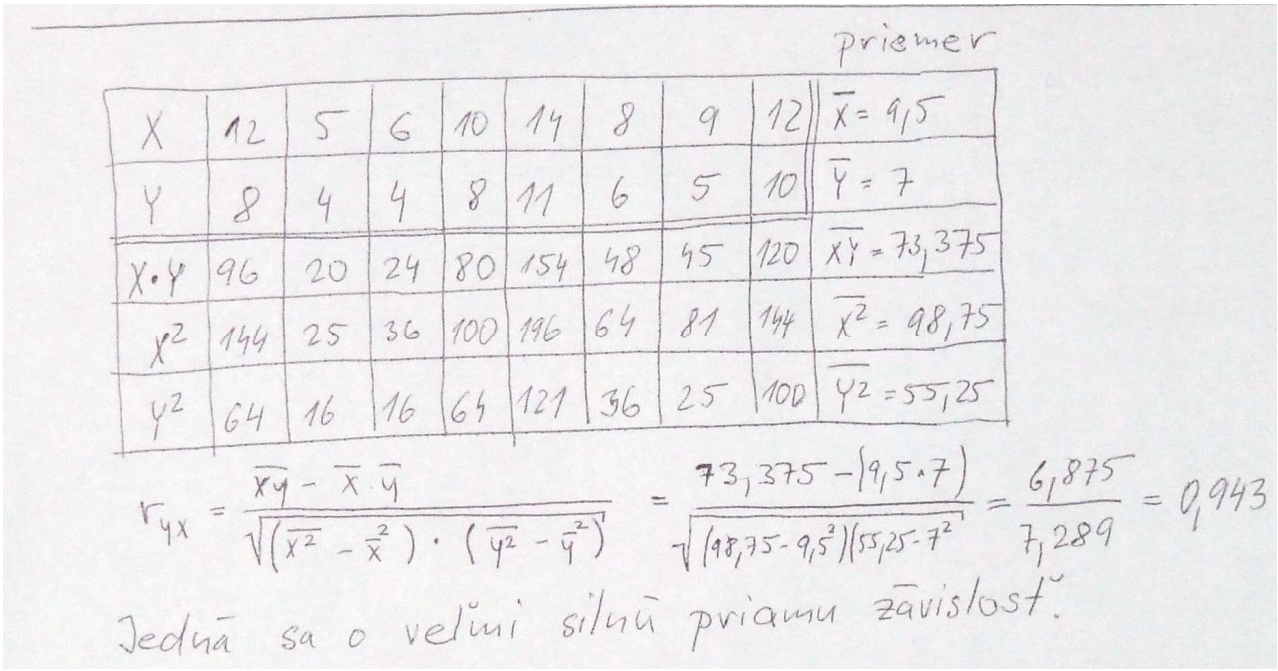
Tento koeficient **nabývá hodnot od -1 do 1** a podobně jako B1 nám znaménko říká zda je závislost přímá nebo nepřímá. Zároveň, ale čím je bližší nule, tím je závislost slabší, v případě nuly není mezi X a Y žádná lineární závislost.

Další příklady:

1. Určete druh a sílu závislosti proměnné X a Y, když jsme naměřili tyto hodnoty.

x	12	5	6	10	14	8	9	12
y	8	4	4	8	11	6	5	10

Stačí nám spočítat korelační koeficient, ze kterého vypočítáme i druh a sílu závislosti. Když je kladný závislost je přímá. Je velmi blízko 1, což značí skoro dokonalou závislost.



2. Během 10ti letních dní bylo zaznamenáváno kolik hodin denně svítilo slunce a počet litrů zmrzliny (=Y). Vypočítejte hodnotu korelačního koeficientu závislosti a na x a interpretujte jeho význam. Z dat byly vypočítány údaje v následující tabulce:

veličina	x	y	x^2	y^2	xy
Součet hodnot	25	140	72	2550	460

(3.)

$$r = \frac{\bar{xy} - \bar{x}\cdot\bar{y}}{\sqrt{(\bar{x^2} - \bar{x}^2)\cdot(\bar{y^2} - \bar{y}^2)}} = \frac{46 - 2,5\cdot 14}{\sqrt{(7,2 - 2,5^2)(255 - 14^2)}} = \frac{11}{\sqrt{0,95\cdot 59}}$$

$$= \frac{11}{\sqrt{56,05}} = 1,47$$

$\bar{xy} = \frac{460}{10} = 46$

$\bar{x} = \frac{25}{10} = 2,5$

$\bar{y} = \frac{140}{10} = 14$

$\bar{x^2} = \frac{72}{10} = 7,2$

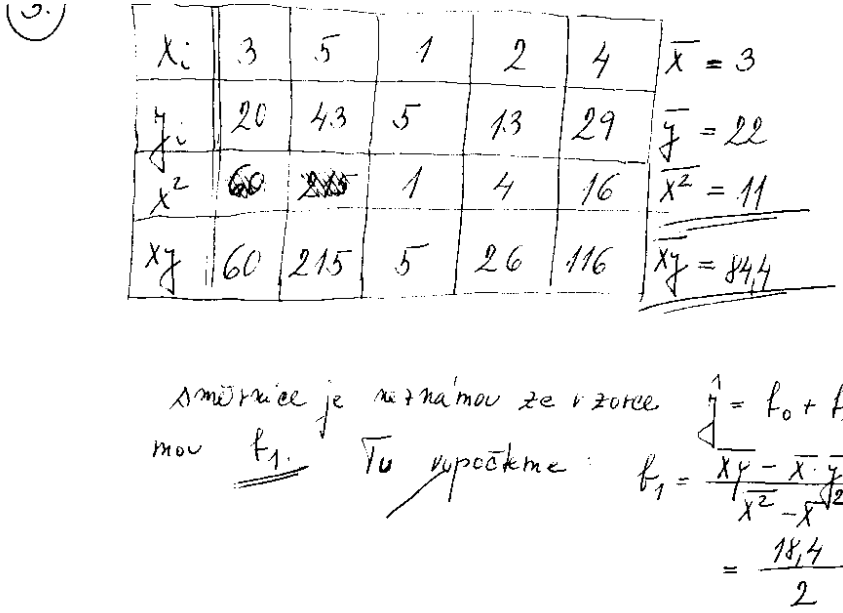
$\bar{y^2} = \frac{2550}{10} = 255$

hodnota je vyšší než 1 což nemůže být. Údaje byly špatně zadané.

r se musí pohybovat v intervalu $\langle -1, 1 \rangle$.

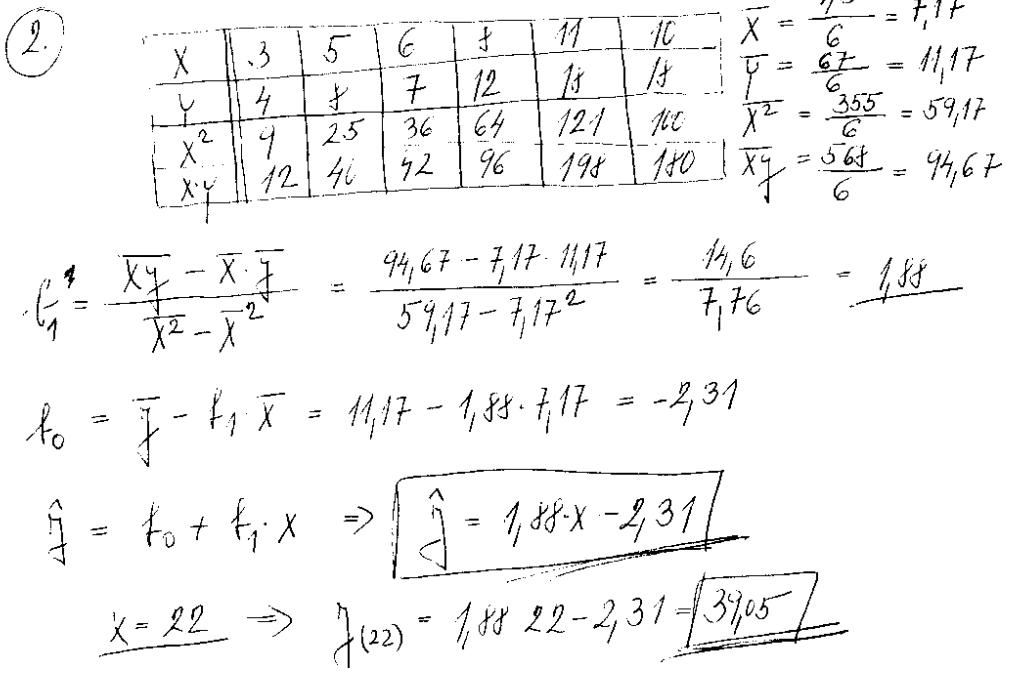
3. Vypočítejte a $\bar{xy}, \bar{x^2}$ hodnotu směrnice

x_i	3	5	1	2	4
y_i	20	43	5	13	29



4. Pro X a Y jsme zaznamenali hodnoty v tabulce. Pomocí regresní přímky zjistěte střední hodnotu proměnné Y, v případě když X=22.

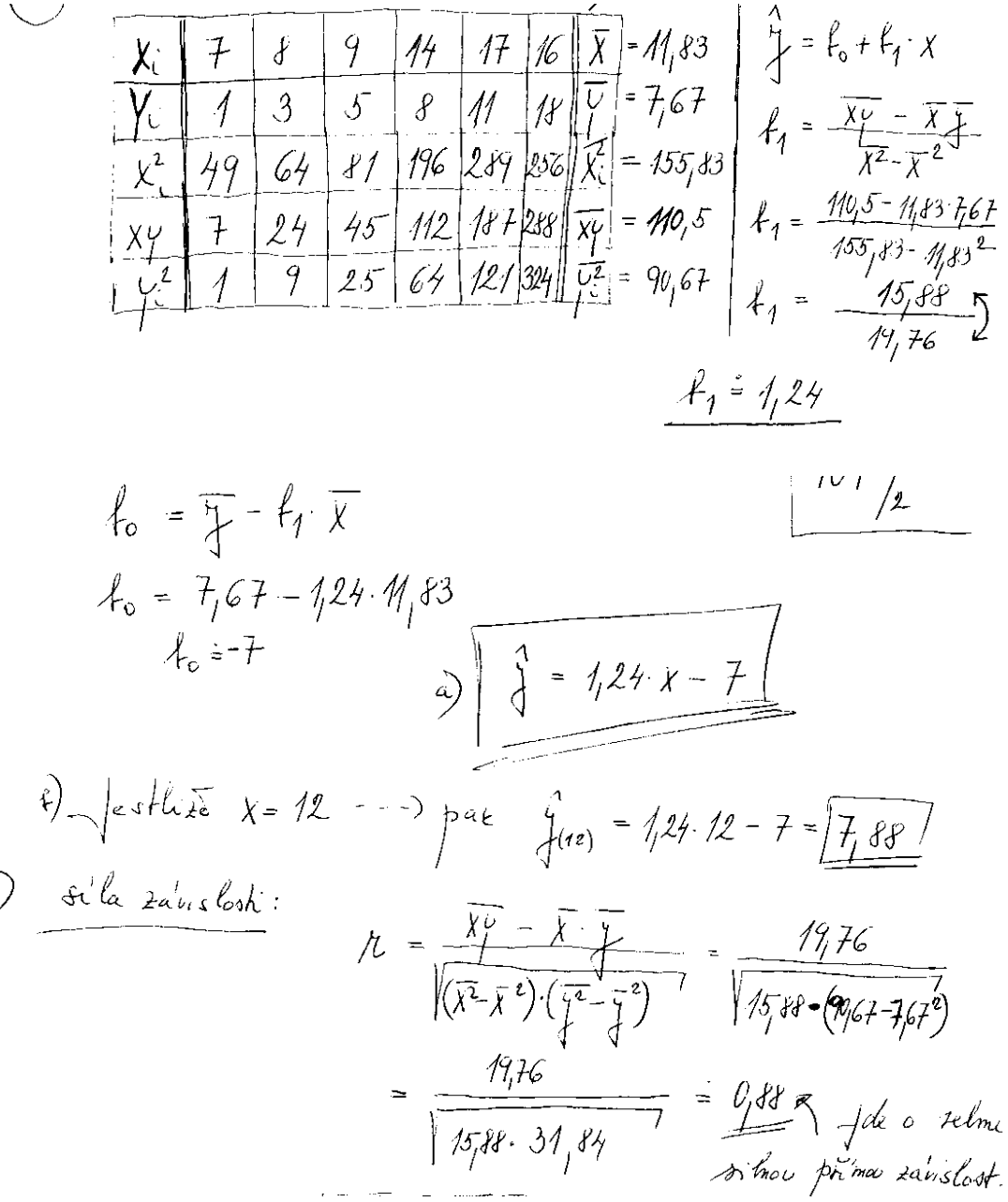
x	3	5	6	8	11	10
y	4	8	7	12	18	18



Musíme si uvědomit, že tím, že vypočítáme rovnici konkrétní regresní přímky, dostali jsme vlastně nástroj, který nám na základě vztahu, který mezi proměnnými je, přiřadí každému X příslušné Y a nebo i naopak. V tomto případě bylo tedy nutné vypočítat regresní přímku, potom do ní dosadit X=22 a tak zjistit hodnotu Y.

5. závislost vysvětlované proměnné Y na vysvětlované proměnné X vyrovnejte přímkou, stanovte odhady Y, když X=12 a změřte sílu závislosti.

x_i	7	8	9	14	17	16
yi	1	3	5	8	11	18



Tedy vypočítat rovnici regresní přímky , dosadit do něj za X dvanáct a zjistit tak Y a potom ještě spočítat i korelační koeficient. Lehké, ne?:)

Indexy a časové řady

Časové řady

Časová řada je nějaká jednoduchá posloupnost čísel sepsaných vedle sebe, přičemž my analyzujeme jak tyto čísla stoupají a podobně. Tedy např. roční růst HDP za 10 let je časová řada. Podmínkou u těchto řad jsou akorát logické věci, tedy např. aby čísla byly ve stejných jednotkách a aby byly porovnávány v určitých stálých intervalech atd. Jediné co je podstatné u časových řad je vyznat se v tom která proměnná co znamená. Proto na začátku pojmenujeme většinu proměnných a vysvětlíme jejich význam.

1. $\mathbf{Y_t}$ je hodnota časové řady v nějakém bodě t, toto t se bere zleva doprava, podle toho jak máme zapsané údaje, tedy jak máme řadu:

2 4 7 13 18

Tak Yt1=2; Yt2=4… V následujících příkladech bude použita tato časová řada.

2. $\bar{y}$ je průměrné $\mathbf{Y_t}$, tedy průměrná velikost člena časové řady, vzorec je vzorcích ,ale logicky, sečteme všechny členy časové řady a vydělíme je počtem. Např.
$\bar{y}$ =(2+4+7+13+18)/5=8,8

3.$\mathbf{k_t}$ je tempo růstu (koeficient růstu) časové řady. Vydělíme hodnotu jednoho období hodnotou období minulého. Podle toho, jestli je výsledek větší nebo menší než 1, můžeme říci jak a o kolik (%) se nám sledovaná hodnota zvětšila nebo zmenšila. Např. : $\mathbf{k_2}$ = 4/2= 2 (výsledek „2“ hovoří o tom, že hodnota řady ve druhém období(4) je 200% hodnota řady v období prvním(2), nárůst je tedy o 100%, viz bod 6).

4. $\bar{k}$ je průměrné tempo růstu (koeficient růstu) časové řady. Pomocí tohoto koeficientu umíme vypočítat průměrné tempo růstu celé časové řad, přičemž nám stačí vědět první a poslední hodnotu. Jak vyplývá ze vzorce je to T-1 odmocnina podílu posledního a prvního člena řady, kterou zkoumáme. To „T-1“ znamená, že spočítáme kolik členů má časová řada, odečteme od toho 1 a výsledkem to budeme odmocňovat. Např.: V naší řadě máme 5 členů, odmocňujeme tedy 4 odmocninou z podílu posledního (18) a prvního (2) člena = 1,732. Naše řada tedy roste s každým obdobím průměrně o 73%.

5. $\mathbf{r_t}$ se nazývá relativní přírůstek a vypovídá o kolik procent se změnila hodnota, prakticky se vypočítá jak $\mathbf{k_t-1}$, tedy v případě $\mathbf{k_2}$=2 je přírůstek 2-1=1=100%. Rozdíl mezi relativním přírůstkem a tempem růstu je ten, že relativní přírůstek hovoří o procentuálním nárůstu zatímco tempo růstu vypovídá o tom o kolik procent minulého období bychom museli mít, aby výsledkem bylo období běžné.

6. $\mathbf{d_t}$ se nazývá absolutní přírůstek(ve vzorcích Δ y) a říká, o kolik jednotek se změnila hodnota časové řady, Tedy v našem případě $\mathbf{d_2}$ je 4-2=2; $\mathbf{d_3}$ je 7-4=3. Pokud po nás chtějí průměrný absolutní přírůstek vypočítáme rozdíl posledního a prvního člena období, který ale ještě musíme vydělit počtem období -1.

To by byla teorie a teď příklad:

1. *V následujících časových řadách dopočítejte celkem 5 chybějících údajů, výsledek je nutné doložit písemně (alespoň náznak výpočtu). Zjištěné údaje XXX nejsou požadované.*

a)

Hodnota časové řady (y_t)		156	175
Koeficient růstu (k_t)	XXX	0,9750	

b)

Hodnota časové řady (y_t)	XXX	XXX
Koeficient růstu (k_t)	XXX	
Relativní přírůstek (r_t)	XXX	−0,3200

c)

Hodnota časové řady (y_t)	25	
Absolutní přírůstek (d_t)	XXX	
Relativní přírůstek (r_t)	XXX	0,2800

a)

Hodnota y_t		156	175
Koeficient k_t	XXX	0,975	1,122

b)

Hodnota (y_t)	XXX	XXX
Koeficient růstu (k_t)	XXX	0,68
Relat. přírůstek (r_t)	XXX	-0,32

$k_t = r_t + 1$

c)

Hodnota (y_t)	25	32
ab. přírůstek (d_t)	XXX	7
Relat. přírůstek (r_t)	XXX	0,28

Za a) vidíme, že hodnota 156 je 0,975 násobkem předešlé hodnoty, tedy 156/ 0,975=160; zároveň chceme spočítat tento koeficient pro čísla 156 a 175= 175/156=1,122
Za b) víme, že tempo růstu je vždy o 1 větší než relativní přírůstek.
Za c) když relativní přírůstek je 28%, převedeme procenta na jednotky, tedy 28% z 25 je 7, tedy následující člen bude o 7 větší a absolutní přírůstek bude analogicky 7.