StatSoft

# Popisná statistika – kvalitativní veličiny

V tomto bloku se budeme věnovat popisné statistice, a to konkrétně té části, která je vhodná pro popis kvalitativních veličin. V tomto díle tedy nebudeme posuzovat objektivitu a kvalitu naměřených hodnot, nebudeme testovat statistickou významnost apod., ale pouze vhodně popíšeme náš datový soubor tak, abychom získali lepší představu o charakteru a průběhu dat.

Úkolem popisné statistiky je graficky a číselně popsat datový soubor, aby byly dobře patrné jeho statistické vlastnosti a eventuálně byl také srovnatelný s ostatními datovými soubory. Popisné charakteristiky (míry polohy a variability) nám kromě popisu našich konkrétních experimentálních dat dají také základ pro odhadování populačních charakteristik. Než začneme s učebnicovým výkladem, pojďme se podívat na příklad použití popisné statistiky, tentokrát na akademické půdě:

Student oboru bankovnictví se dostavil do počítačové učebny, kde ho čekal krátký průběžný zápočtový test z aplikované statistiky. O právě probírané látce nevěděl téměř nic, statistika je pro něj nutné zlo, v bance bude totiž pracovat nejméně ve středním managementu. Ke zkoušce si přinesl velice sofistikovaný tahák, který vycházel z loňských variant těchto testů, jeho postup bychom mohli nazvat studentskou klasikou. Cílem testu bylo prokázat praktické znalosti statistiky a vhodně je aplikovat ve statistickém softwaru, který má univerzita k dispozici. Po otevření datového souboru CreditScoring se v zadání dočetl: „Aplikujte popisnou statistiku na všechna klientská data v souboru, výstupy ze softwaru okomentujte a uložte pod Vaším jménem na sdílené úložiště."

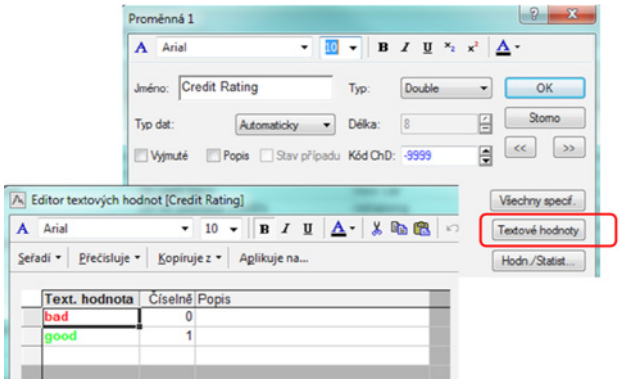
| | 1 Credit Rating | 2 Balance of Current Account | 3 Duration of Credit | 4 Payment of Previous Credits | 5 Purpose of Credit | 6 Amount of Credit | 7 Value of Savings | 8 Employed by Current Employer for | 9 Installment in % of Available Income | 10 Marital Status | 11 Gender | 12 Living in Current Household for | Most Valu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | bad | no running account | 36 | no problems with current credits | retraining | $3 003,00 | no savings | 5-8 years | 25-35 | single | male | < 1 year | life insurance |
| 2 | good | no balance | 48 | hesistant | retraining | $17 085,60 | >1400 | 1-5 years | 25-35 | single | male | 1-5 years | life insurance |
| 3 | bad | >$300 | 36 | no previous credits | used car | $15 363,60 | no savings | unemployed | < 15 | divorced/living ap | female | 1-5 years | life insurance |
| 4 | good | no running account | 24 | paid back | new car | $8 986,60 | no savings | > 8 years | 25-35 | divorced/living ap | female | >8 years | ownership of |
| 5 | good | >$300 | 24 | no previous credits | retraining | $1 761,20 | no savings | 5-8 years | < 15 | single | male | < 1 year | no assets |
| 6 | good | no balance | 12 | no previous credits | retraining | $1 451,80 | <140 | 5-8 years | 15- 25 | single | male | >8 years | no assets |
| 7 | bad | no running account | 30 | no previous credits | used car | $4 351,20 | no savings | <1 year | 25-35 | divorced/living ap | male | >8 years | car |
| 8 | good | no balance | 15 | paid back | furniture | $2 151,80 | >1400 | > 8 years | < 15 | single | male | >8 years | no assets |

Na test absolutně nepřipraven, začal pozvolna lovit svůj tahák. Učitel, v tu chvíli zaneprázdněn studentkou, která se nedokázala, jak se později ukázalo, kvůli aktivnímu CapsLocku správně „zalogovat", mu dal čas k bezpečnému nahlédnutí do připraveného taháku. V něm u popisné statistiky byla hesla *míry polohy, míry variability (průměr, medián, modus, rozptyl, směrodatná odchylka atd.)*. Po načtení dat v softwaru v dialogu *vybrat proměnnou* zvolil *vybrat vše* a na panelu deskriptivní statistika zaškrtnul *průměr, rozptyl* a další míry, které si během doktorandovy nepozornosti stihl z přečteného papírku zapamatovat a potvrdil výpočet. Průměr z proměnné Credit Rating je 0,7, napsal do protokolu, když komentoval výsledek. Průměr 1,3 u proměnné pohlaví se mu příliš nezdál, nechal ho proto raději bez komentáře. Na konci uvedl vše na pravou míru větou: Proměnné, které zůstaly neokomentované, se neinterpretují. Ve slabé chvilce učitele ještě mobilem vyfotil zadání testu a poté opustil třídu. Za necelou hodinu na Facebooku napsal: Test byl celkem v pohodě, pořád stejný, zde posílám zadání. V komentářích se během chvíle objevily hlášky typu: „A máš k tomu i řešení?"

Příklad ze života ukazuje na špatnou aplikaci deskriptivních statistik. Zmiňovaná osoba sice správně zvolila klíčové vzorce, ale aplikovala je plošně na všechna data v souboru.

Jak mohl software vypočítat průměr 0,7 z binární textové proměnné s variantou znaku ***bad/good***? V softwaru *STATISTICA* je každá textová hodnota reprezentována číslem, se kterým software pracuje. *STATISTICA* tedy sečetla všechny hodnoty 0 a 1. Klientů s ratingem „good" je v souboru 700 a velikost celého souboru je 1000.

V tomto případě je kódování v softwaru 0 a 1, hodnota 0,7 tak bude po vynásobení číslem 100 reprezentovat procento případů, které mají v odpovědi „good". Pod textovými hodnotami kategorických veličin se však mohou skrývat i jiné číselné reprezentace, typicky např. 101, 102 atd.




## Grafický a číselný popis kvantitativních veličin

Jak vystihnout vlastnosti nominálních a ordinálních proměnných tak, aby měl příjemce k dispozici přehledné informace, které však zároveň vystihují důležité rysy souborů? Obecně vždy záleží na počtu kategorií daného znaku a typu kategorie. Základním výpočtem je však téměř vždy určení četností, tedy kolik dat mám v jednotlivých kategoriích, výsledek je potom zobrazen v tzv. tabulce četností.

## Tabulka četností

*(V softwaru STATISTICA:* ***Statistiky -> Základní statistiky/tabulky -> Tabulky četností****)*

Tabulka rozdělení četností podává informaci o výskytu jednotlivých variant znaku v souboru z hlediska jejich počtu. V tabulce níže je vidět, že v náhodně vybraném vzorku studentů Oxfordu je 50 % těch, kteří studují společenské vědy, tedy 6 studentů.

| | Tabulka četností:divisions (Oxford.sta) | | | |
|---|---|---|---|---|
| Kategorie | Četnost | Kumulativní četnost | Rel.četnost | Kumulativní rel.četnost |
| Humanities | 3 | 3 | 25,00000 | 25,0000 |
| Medical Sciences | 2 | 5 | 16,66667 | 41,6667 |
| Social Sciences | 6 | 11 | 50,00000 | 91,6667 |
| Physical & Life Sciences, Mathematical | 1 | 12 | 8,33333 | 100,0000 |
| ChD | 0 | 12 | 0,00000 | 100,0000 |

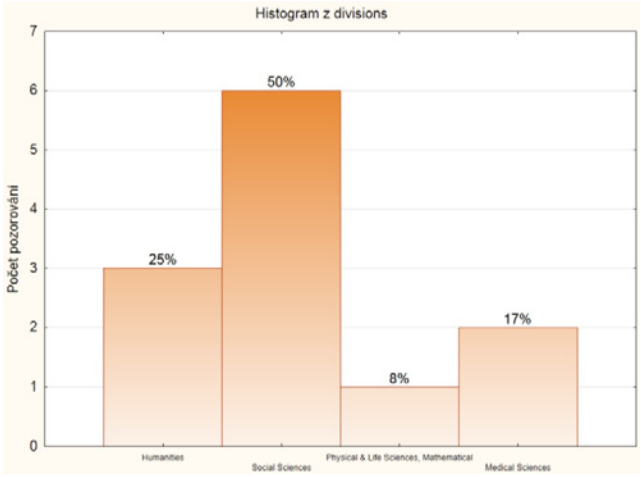
Tři studenti v našem vzorku studují humanitní obor, což v našem vzorku tvoří 25 %. 91 % respondentů studuje buď humanitní obor, společenské vědy nebo medicínu. Kumulativní relativní četnost zde nemá až takový význam, význam relativní kumulativní četnosti ukazuje soubor počet dětí, které uváděli respondenti v anonymním dotazníku v rámci marketingového šetření:

| | Tabulka četností:Počet dětí (marketing.sta) | | | |
|---|---|---|---|---|
| Kategorie | Četnost | Kumulativní četnost | Rel.četnost | Kumulativní rel.četnost |
| 0 | 6 | 6 | 18,75000 | 18,7500 |
| 1 | 10 | 16 | 31,25000 | 50,0000 |
| 2 | 11 | 27 | 34,37500 | 84,3750 |
| 3 | 3 | 30 | 9,37500 | 93,7500 |
| 4 | 1 | 31 | 3,12500 | 96,8750 |
| 5 | 1 | 32 | 3,12500 | 100,0000 |
| ChD | 0 | 32 | 0,00000 | 100,0000 |

V souboru, kde je 32 odpovědí, převažuje jedno a dvě děti (31,25 % a 34,3 %) a 93,7 % respondentů má maximálně tři děti, resp. tři a méně dětí. Pouze 2 jedinci měli v souboru více než 3 děti. Chybějící hodnoty proměnná počet dětí neobsahuje (**ChD** = 0). Pro porovnání různých rozdělení četností, která se liší svými rozsahy, je vhodné používat relativní četnosti místo absolutních. Relativní četnosti $p_i$ je podíl jednotlivých absolutních četností k celkovému rozsahu souboru. Počet dětí je sama o sobě proměnná diskrétní, obecně kvantitativní. Průměrný počet dětí u respondentů v konkrétním souboru tak má svůj smysl. V tomto případě však tuto proměnnou bereme jako ordinální a počet dětí značí danou kategorii. Obecně proměnné, které jsou diskrétní a mají malý počet obměn (přibližně cca do 10), mohou být, pokud je to žádoucí, považovány za ordinální.

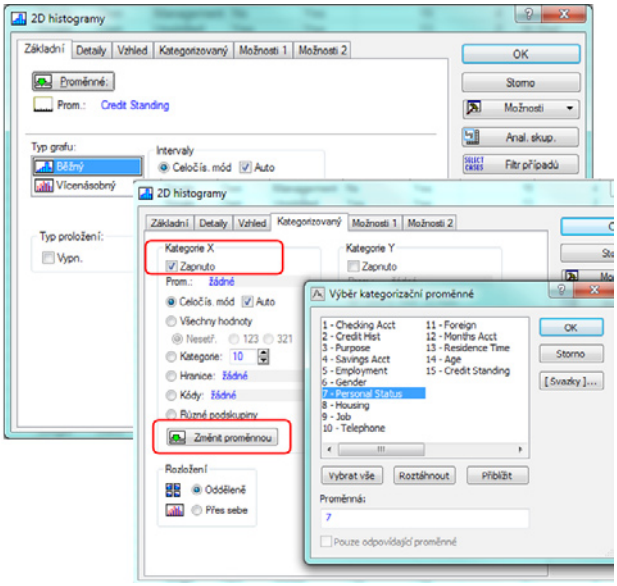
## Histogram

Grafickým znázorněním tabulky četností je histogram četností (**Grafy -> Histogramy**), kde základny mají délku zvolených intervalů a výšky velikost příslušných třídních četností, alternativně relativních četností:

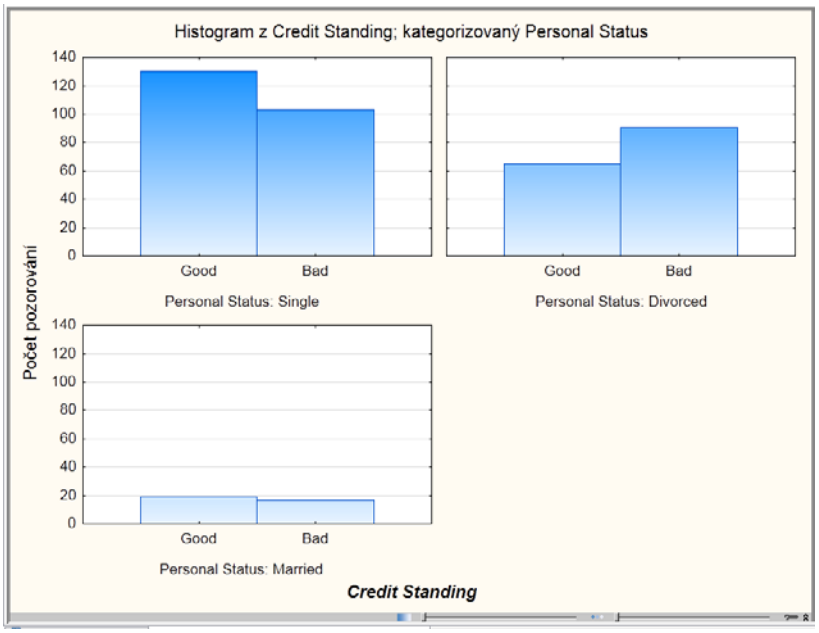



V praxi je však pouhé zobrazení četností obměn jednoho znaku nedostatečné a obměny znaku je nutné rozlišit podle další kategorické proměnné, například podle pohlaví, vzdělání, použitého polotovaru, odrůdy, typu stroje atd.

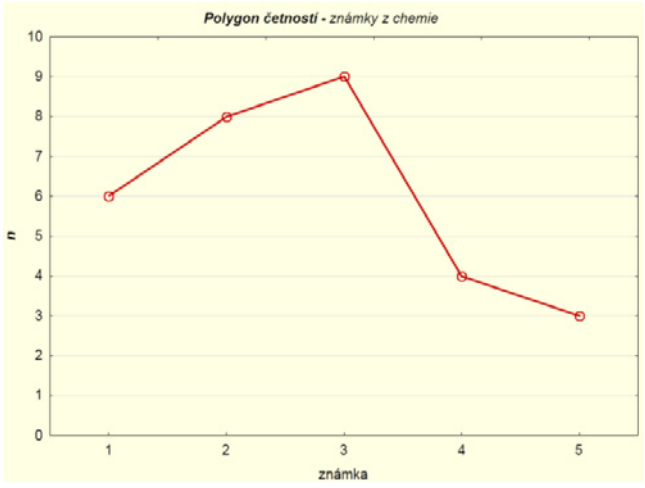
*(V softwaru STATISTICA:* ***Grafy -> Histogramy -> záložka Kategorizovaný -> Kategorie X -> Změnit proměnnou****)*



Obrázek níže ukazuje četnosti klientů z hlediska schopnosti splácet úvěr, kteří jsou dále rozděleni podle svého osobního stavu (ženatý, rozvedený, svobodný). Dle rozložení jednotlivých kategorií bychom neusuzovali na nějakou významnější závislost mezi rodinným stavem a schopností splácet úvěr, případně pouze na velmi slabou.




O testování hypotéz nad kvalitativními znaky bude řeč v nějakém z příštích článků. U spojité veličiny bychom museli jednotlivé výskyty znaku zařadit do kategorií, v takovém případě potom hovoříme o intervalovém rozdělení četností. Ve výše uvedeném případě naopak mluvíme o prostém rozdělení četností. Alternativou k histogramu je polygon četností, někdy nazýván četnostní funkce, který opět na ose X zachycuje hodnoty znaku (Xi) a na ose Y jim odpovídající četnosti (ni). Graf četnostní funkce pro známky s chemie:




V příští části našeho seriálu budeme pokračovat s aplikací popisné statistiky také na spojité veličiny a postupně si ukážeme všechny nejdůležitější metriky a grafy, která se na spojité veličiny v praxi aplikují.

StatSoft STATISTICA

**StatSoft CR s.r.o.**
Ringhofferova 115/1, 155 21 Praha 5 – Zličín
**t** +420 233 325 006, **f** +420 233 324 005
**e** info@statsoft.cz