Statistika poprvé - popis dat

Představíme si základní pojmy tzv. popisné statistiky, tedy způsobu kterým lze výstižně popsat získaná data. Nejvhodnější způsob prezentace dat je zobrazení jejich grafu a uvedení vybraných popisných statistik. Které vybrat, jak je získat a užívat ukazuje tato kapitola. [vysvětlit p-value]

Typy dat

Ať už provádíme výzkum přírodovědný nebo společenskovědní, vždy se setkáváme s třemi typy dat, které se liší způsoby jak je zpracovávat i analyzovat.

intervalová - jsou to číselná data, se kterými můžeme provádět běžné matematické operace jako je sčítání, odčítání atp. Příkladem je věk žáků, procenta či body z testu, roky školní docházky žáka (nikoli ročník!)

ordinální - hodnoty takovýchto veličin jsou uspořádané, ale i v případě že jsou uváděny jako čísla, nejsou u nich uplatnitelné ony „běžné matematické operace“. Můžeme říci že jedna hodnota je vyšší/větší než druhá, ale nemůžeme určit kolikrát přesně je vyšší. Příkladem je nejvyšší dosažené vzdělání nebo školní ročník, ale i školní prospěch hodnocený známkou 1-5 či stupni A-F. Kdybychom jako prospěch chápali pouze výsledky testů v podobě bodového či procentuelního ohodnocení, pak by se jednalo o intervalová data - ale takto plochou klasifikaci věřím nikdo z vyučujících nepoužívá. Jde tedy o jasná data ordinálního typu a obecně rozšířené „počítání průměru“ na konci pololetí (či prospěchové stipendium na základě průměru výsledků zkoušek) je striktně matematicky vzato chybné.

nominální – buďte opatrní abyste je nezaměňovali s předchozí kategorií. Tato data nám říkají že se objekty liší; nemůžeme ale tuto diferenci hodnotit, tedy ani data řadit. Nejjasnějším příkladem jsou pohlaví, náboženská či státní příslušnost, paralelky ročníku [6] nebo barva vlasů.

Ukažme si tedy, jak postupovat u jednotlivých typů dat.

[6] Skupiny pro výuku jazyků rozřazené na základě výsledků testů jsou nominální kategorie v případě kdy u žáků sledujete zmíněnou barvu vlasů, tedy nepředpokládáte závislost na jejich jazykových dovednostech; pokud sledujete úspěšnost v druhém jazyce, jedná se o ordinální kategorie.

Intervalová data

Rozložení dat

Pro veškerou následnou práci s intervalovými daty potřebujeme vědět zda mají data tzv. normální rozložení. Pojem normální zde představuje *terminus technicus* používaný pro rozložení, které je teoreticky velmi dobře prozkoumané a díky znalosti jeho vlastností lze taková data velmi dobře používat.

Pro odhad nakolik se naše data blíží tomuto teoretickému optimu můžeme využít již použitý příkaz stem(trida8A.vek).

Robustnější řešení představuje grafický výpis pomocí příkazů

```
> hist(trida8A.vek, freq=F)
> lines(density(trida8A.vek))
```

Ovšem na příkladu věku žáků není graf příliš dobře čitelný a lépe nám poslouží již použitý boxplot.

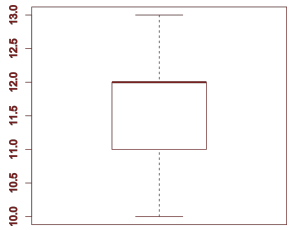
```
> boxplot(trida8A.vek)
```

[vysvětlit normáln rozložení]

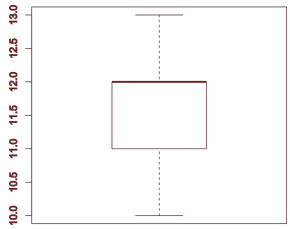
Z těchto grafů získáme hrubou představu. Abychom byli schopni posoudit rozložení exaktně, nestačí náš dojem. Využijeme tedy náš první statistický test, jmenuje se Shapiro-Wilk test normality a to jednoduchým příkazem

```
> shapiro.test(trida8A.vek)

        Shapiro-Wilk normality test

data:  trida8A.vek
W = 0.8508, p-value = 0.0228
```

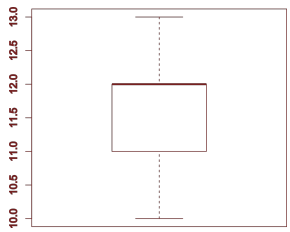
Pamatujete si ještě Exupéryho hada který snědl slona?



Obrázek 2: slona prosím.



Obrázek 3: Silná čára udává medián, vnější zarážky určují 95% dat souboru, v boxu leží druhý a třetí kvartil dat.



Obrázek 4: Silná čára udává medián, vnější zarážky určují 95% dat souboru, v boxu leží druhý a třetí kvartil dat.

Ve výstupu skriptu zjistíme p-value[7] tohoto testu - pokud je větší než 0.05, pak jsou data rozložena NORMÁLNĚ, parametricky, pokud je menší než uvedená hodnota, jsou rozložena neparametricky. Parametrická data se snadněji zpracovávají a jděte to oslavit. Neparametrická si velmi pravděpodobně vynutí konzultaci s někým zkušenějším, možná řešení naznačuje kapitola xxx.

Pozorný čtenář může být zaražen výsledkem našeho zjištění – proč nemá věk dětí v ročníku normální rozložení? První důvod je matematický a spočívá v zaokrouhlování hodnot na celé roky. V takovém okamžiku se totiž dopouštíme nepřiměřeného zaokrouhlování; celý soubor dat (osmáci) jsou převážně děti ± 6 měsíců, takové nelze měřit s přesností pouze na celé roky, respektive takové měření nemá vypovídací hodnotu.

Druhý důvod souvisí s již uvedeným výběrem dětí do ročníku. Do první třídy jsou přijímány především děti s datem 1. března daného roku ±6 měsíců - jedná se tedy o zcela nenáhodný výběr. Aby pak vykázala takováto skupina normální rozložení, musela by celorepubliková porodnost sama o sobě mít normální rozložení s maximem právě kolem 1. března. [dáte sem porodnost čr] Budeme-li tedy měřit nepřesně (na celé roky), získáme ostrý, nic neříkající pík; budeme-li měřit přesně (na měsíce) pak získáme opět nic neříkající pás víceméně konstantní porodnosti. Další zvyšování přesnosti měřením na konkrétní den pak získáme falešně nepravidelné rozložení - narazíme na problém malých čísel.

Normalita nebo nenormalita rozložení dat ovlivňuje veškerou další práci s intervalovými data a to nejenom výběr korektních statistických testů, ale třeba i tak zdánlivě jednoduchou otázku, jakou je popisná statistika. Popisná statistika je druhý krok v naší analýze – zjistili jsme jaké rozložení mají naše data, nyní pro ně spočítáme popisné hodnoty které výstižně celý soubor dat charakterizují.

[7] Koncept p-value je vysvětlen v sekci věnované statistickým testům, v této chvíli se omezme na konstatování že právě z ní rozpoznáme výsledek testování.

Zvažte dospělé lidi na váze pro kamiony - vaším výzkumným zjištěním bude, že hmotnostní rozdíly mezi lidmi jsou zcela zanedbatelné a že můžeme s klidem říci že všichni váží stejně = 100kg.

Z mého pohledu je fascinující, že při sledování jednoho atomu radioaktivního zářiče naprosto nejsme schopni predikovat jeho rozpad. Máme-li ale byť drobounké zrnko zářiče (a tedy velmi mnoho atomů), dovedeme velmi přesně říci kolik se jich za daný časový úsek rozpadne. Problému se věnujeme později v kapitole Kolik dat potřebujeme...

Popisná statistika

Existují různé statistické veličiny – notoricky známými je nejnižší a nejvyšší hodnota, aritmetický průměr; v předchozích odstavcích diskutované rozložení dat patří rovněž mezi jednu z nejvýznamnějších charakteristik popisné statistiky. V následující sekci je stručně představíme a popíšeme způsob jak je používat a interpretovat, nejprve uveďme jejich tabelární přehled, v témže pořadí jsou poté probírány v textu.

symbol	popis
min	nejnižší zjištěná hodnota, minimum
max	nejvyšší zjištěná hodnota, maximum
rozsah	rozdíl mezi minimem a maximem
n	celkový počet vzorků/měření
Σ	suma; součet všech hodnot
$\overline{x}$	aritmetický průměr
$\tilde{x}$	medián
$\hat{x}$	modus
$Q1$	první kvartil
$Q1$	směrodatná odchylka
$Q1$	standardní chyba
$Q1$	rozptyl

Tabulka 1: Veličiny popisující datový soubor, horizontální čára odděluje první čtyři veličiny použitelné i u ordinálních dat; u dat intervalového typu připadají v úvahu všechny zde uvedené.

[modus nejde u nominálních?]

MINIMUM zjistíme funkcí min(data) a představuje nejnižší naměřenou hodnotu.

MAXIMUM zjistíme funkcí max(data).

ROZSAH odvodíme max(data)-min(data). Tyto tři základní veličiny v podstatě nemají vliv na interpretaci dat a není nutné (a často ani vhodné) je uvádět.

CELKOVÝ POČET VZORKŮ je naopak mimořádně významný a nikdy jej nezapomeňte uvádět - nestačí se ale omezit na sdělení v textu „Dotazník jsem rozdal 1234 žákům, 875 jich dotazník

Zjistí-li Česká školní inspekce že extrémy v počtu neomluvených hodin žáků ZŠ daného školního roku jsou 0 a 317 hodin, tak co to vypovídá o celém souboru? Vůbec nic. Nula je logická, druhý, horní extrém je vhodný buď pro novinářský titulek nebo pro intervenci u dané osoby; naše porozumění fenoménu záškoláctví ale potřebuje vidět data v jiném úhlu pohledu.

vyplnilo, po vyloučení nekompletních zbyl výsledný počet 829 zpracovaných dotazníků“ (jakkoli je takové sdělení v úvodním popisu nutné), ale tento údaj budete opakovaně a neustále uvádět u všech statistický veličin dále v textu či grafech – tento počet měření je totiž, ze statistického pohledu, nesmírně významný a bez něj není žádný údaj kompletní.

Průměrné skóre otázky č. 15 bylo 25 bodů (n=829)

V čem spočívá rozdíl mezi sto korunami z první republiky a dnešními? Dnes je v oběhu řádově vyšší objem korun a tudíž identická částka nese menší kupní sílu (tedy pravděpodobně, auto srovnatelných parametrů s dnešními by bylo na začátku století dražší oproti dnešku i v absolutních číslech; neplechu zde působí faktor zvaný *pokrok*), říká se že je devalvovaná. Podobně tak číselně identická statistická veličina má vždy svým způsobem „jiný význam/sílu“ při odlišném počtu vzorků.

Suma je veličina kterou zpravidla neuvádíme, ovšem používá se pro řadu statistických výpočtů. Znak se jmenuje sigma a je to písmeno velké S řecké abecedy. xxxcheck

Artimetický průměr je mimořádně oblíbenou veličinou, často ovšem užitou nevhodně. Nejprve pojmenování – přívlastek aritmetický je dobré vždy výslovně uvádět, neboť pojem průměr nese v různých kontextech odlišný význam. Počítá se jako podíl sumy a počtu vzorků a je vhodný *výhradně pro parametrická* data.

Ještě jednou – než v textu či analýze dat vaší práce použijete aritmetický průměr, ujistěte se statistickým testem (strana xxx) že data mají parametrické rozložení nebo alespoň velmi blízké parametrickému. Pokud má Shapiro test p-value menší než 0,05, nemůžeme jej pro popis dat použít.

Medián je střední hodnota, tedy seřadíme-li si prvky vzestupně, tak medián udává hodnotu uprostřed této řady [xxxsudý počt rpvků se průměruje?]. Je základní veličinou pro popis neparametricky rozložených dat, kde nahrazuje aritmetický průměr.

Kvartily je označení pro seřazené prvky ve vzestupném pořadí, tedy úzce souvisí s mediánem – první kvartil tedy sděluje že 25% vzorků má hodnotu menší nebo rovnu dané hodnotě. Druhý kvartil je roven mediánu, třetí kvartil 75% vzorků a čtvrtý kvartil je roven maximu.

```
quantile(data)
```

Jméno funkce napovídá že kvartil je specifický případ obecnějšího kvantilu. Chceme-li tedy zjistit nejvyšší hodnotu první pětiny datového souboru, voláme

quantile(data, 0.2)
Medián zjistíme voláním

```
quantile(data, 0.5)
```

Modus představuje nejčastější hodnotu v datovém souboru. [co když jich je tam více stejných?]

SMĚRODATNÁ ODCHYLKA

STANDARDNÍ CHYBA

ROZPTYL

Příklad špatné praxe Nejpopulárnějším příkladem pro demonstraci problému popisu neparametrického rozložení dat je hrubá měsíční mzda (dále jen mzda). Popularita asi vychází jednak ze srozumitelnosti příkladu, ale také pro překvapivě důsledné ignorování korektní popisné statistiky při prezentování těchto dat veřejnosti.

Rozložení výše mezd je typicky neparametrické, výrazně zašikmené. Existence minimální mzdy (v roce 2015 8500.- Kč) a její oblíbenost u zaměstnavatelů v kombinaci s velmi vysokými platy vysokých manažerských funkcí vytváří zcela specifické rozložení. Zejména vlivem extrémně vysokých platů pobíraných cca 2% zaměstnanců je ovlivněn aritmetický průměr natolik, že nepodává informaci, kterou od něj implicitně očekáváme – vždyť 2/3 zaměstnanců mají mzdu menší než je tento průměr!

xxx graf

Jak již bylo uvedeno v předchozí kapitole, u neparametrického rozložení je průměr nevhodný a manipulativní. Pro popis dat s neparametrickým rozdělením je třeba užít medián či modus. Modus (nejčastější hodnota) by byl v tomto případě prezentoval data ještě hůře vzhledem k tomu, že odpovídá minimální mzdě, omezit ho na střední pásmo rozsahu by rovněž nebylo korektní; zbývá tedy medián - střední hodnota. Polovina zaměstnanců má plat vyšší než je tato částka, polovina menší, tedy sdělení které nejtěsněji souzní s obecně vnímaným pojetím „průměrné mzdy" a je tedy *vhodným* popisem dat[8]. V médiích se ale bohužel se-

Plným jménem HRUBÁ MĚSÍČNÍ NOMINÁLNÍ MZDA NA PŘEPOČTENÉ ZAMĚSTNANCE - tedy bez zohlednění daňových benefitů, ale především přepočtena na úvazky. Nelze tedy zaměňovat s měsíčními příjmy - zisky jiné než plat, popřípadě více zaměstnání jednoho člověka. V roce 2011 novináři z MfDNES odhalili ve fakultní nemocnici na Vinohradech lékaře s celkovým úvazkem 6,2 – koneckonců i česká politika je v nabalování funkcí a pozic v dozorčích radách chronicky churá.

[8] V roce 2013 byl medián 22 557Kč, průměr 26 444Kč (Statistická ročenka ČR 2014).

tkáváme téměř výhradně s uváděním průměru, který je vyšší, tudíž líbivější než korektní medián - proto je *manipulativní*, neboť příčina jeho užívání netkví v neznalosti či nepochopení, ale populismu.

Zkusme tedy v našich pracích užívat korektní přístup k popisné statistice, v němž nehledáme cesty k podpoření stanovené hypotézy či úspěšné prezentaci své práce, nýbrž usilujeme o výstižný popis dat a zjištěných výsledků takových, jaká ve skutečnosti jsou.

Nepředstavujme si ale zároveň, že teoretický ideál plně objektivního popisu reality je možný – metodika sběru dat, zpracování, analýza i interpretace jsou vždy zatíženy našimi (byť i nevědomými) paradigmaty. Pro zájemce doporučuji výbornou a čtenářsky přístupnou knihu cite: sedláček ekono dobra azla, str 148: *„Fakta nefungují bez racionálně vnímajícího, tedy jistého racionálního rámce, ve kterém získávají interpretaci, jména a smysl. Jak píše Cadwell, nic takového jako syrová fakta*[9] *neexistuje.“*

Standardní chyba aneb jak to napsat do diplomky [nominální data!?]

[9] Rozuměj fakta bez interpretačního rámce, tedy bez vysvětlující teorie

Zautomatizování testu normality a popisných charakteristik

Protože u každých intervalových dat, která chceme zpracovávat, musíme nejprve zkontrolovat jejich rozložení a vytvořit popisnou statistiku, nabízí se další zobecnění našeho skriptu, tedy si tyto prvky (popis dat, vizualizace rozložení dat a test jejich normality) zabalíme do jediné funkce - ta už je trochu složitější a není nutné abyste jí rozuměli, důležité je abyste ji používali.

[podmínečné vytvoření adresáře, , kontrola if exist, výpis do texťáku summary a výsledku shapiro? - nebo lépe udělat layout kde je plovoucí průměr, histogram+density a výsledky summary, dále podle shapiro jak to zapsat do práce]

```
#funkce má dva parametry, první jsou vlastní data, druhý je jejich slovní popis podle kter
analyzuj(trida8A.vek, "věk žáků třídy 8.A")
```

Nominální a ordinální data

Ve srovnání s možnostmi, které se nám nabízejí při zpracování intervalových dat, mají nominální a ordinální typy daleko skrovnější nástroje. Přesto se jim není možné, obzvláště v pedagogickém výzkumu, vyhnout a věnujeme jim tuto část

symbol	popis
min	nejnižší zjištěná hodnota, minimum
max	nejvyšší zjištěná hodnota, maximum
rozsah	rozdíl mezi minimem a maximem
n	celkový počet vzorků/měření

Tabulka 2: Veličiny popisující datový soubor ordinálních a nominálních dat.