Základy statistického myšlení

Jiří Michálek

ČSJ 2011

Česká společnost pro jakost
Novotného lávka 5, 116 68 Praha 1

Základy statistického myšlení

RNDr.Jiří Michálek, CSc

Vydání 1. Praha 2011

Obsah publikace

1. Úvod

Tato publikace vznikla na základě dlouholetých zkušeností autora z výuky statistických metod pro řízení jakosti v kurzech pořádaných Českou společností pro jakost. Většina lidí z praxe nemá a ani nemůže mít dostatečné znalosti z teoretického pozadí matematické statistiky, které je dosti náročné a opírá se o řadu matematických disciplin, především o teorii pravděpodobnosti. Matematická statistika tak vlastně sedí na dvou židlích, jedna je ta teoretická, a druhá je ta praktická, kdy už máme nějaká data k dispozici a ptáme se, co s nimi. Dochází tak k nesrovnalostem se skutečností, které jsou způsobeny tím, že vlastní teorie stojí na předpokladech, které v praxi nemusí být splněny a nebo nejsou vůbec ověřovány. Pak interpretace závěrů získaných použitím nástrojů matematické statistiky může být zavádějící. Je nutné si uvědomit, že matematická statistika je mocný nástroj, ale má svá omezení, není všemocná. V současné době existuje celá řada různých statistických softwarů pro zpracování dat, což mnohdy vede k formálnímu využívání těchto prostředků, aniž by bylo ověřeno, zdali je správné ten či onen nástroj použít. Dalším, neméně častým problémem, je právě správná interpretace získaných výsledků.

Cílem této publikace je vysvětlit funkci základních nástrojů matematické statistiky a ukázat, že matematická statistika není pouze o zpracování dat, ale že je to vědecky podložený přístup k realitě, který se snaží pochopit náhodu, která nás každodenně doprovází, a umožňuje se s rolí náhody v našem životě vyrovnat, či

dokonce ji využívat. Nároky na čtenáře, který o statistice neví téměř nic, budou na úrovni střední školy a autor se bude snažit o co nejméně vzorečků a pokusí se pojmy nutné pro základy statistiky popsat spíše slovně. Matematická statistika má jedno výsadní postavení mezi matematickými disciplinami, že totiž využívá silně dedukci ve svém teoretickém základu, ale při vlastním zpracování informace obsažené v datech postupuje pomocí indukce, zobecňuje své závěry získané z dat na celou základní populaci. A to je základem statistického myšlení.

Čtenář, který o statistice prakticky nic nezná, by měl číst publikaci postupně od samého začátku, aby se pomalu seznamoval se základními pojmy. Naopak čtenář, který již něco zná, může si otevřít kapitolu, která ho právě zajímá, případně se může zorientovat v doporučené literatuře, jejíž seznam je na konci publikace. Bohužel česky psaných populárních knížek o statistice je velice málo na rozdíl od okolního světa, hlavně amerického, kde každoročně vycházejí nové tituly, v poslední době např. pod vlivem metodologie Six Sigma pro řízení jakosti, která je především postavena na aplikaci statistických nástrojů.

2. Co je statistické myšlení?

V našem světě se setkáváme s jevy, u nichž víme dopředu, jak jev dopadne, např. volný pád předmětu díky gravitaci. Tu skutečnost, že každý předmět spadne na zem, víme s jistotou, a proto hovoříme o jistém jevu. Naopak existují jevy, které nemohou nastat, jako např. že 6 a 5 jablek dohromady je 10 jablek. Takový jev nazýváme jevem nemožným. Existují ale jevy, u nichž jejich výsledek neznáme, i když podmínky, za nichž jev může či nemůže nastat, jsou stálé. Takovým jevům říkáme náhodné, jsou pod vlivem náhody. Lze je rozdělit do dvou kategorií, a to na individuální a hromadné. Individuální náhodný jev je takový, který se realizuje velice zřídka, např. že na mne spadne při chůzi podél domu rampouch. Hromadné náhodné jevy mají tendenci se opakovat, jako je např. výskyt neshodné jednotky při výrobě. A právě hromadnými jevy se zabývá teorie pravděpodobnosti a matematická statistika. K čemu je zapotřebí ta hromadnost? Ta je nutná k tomu, abychom měli na základě pozorování možnost posoudit pravděpodobnost takového jevu s pomocí sledování počtu výskytů sledovaného jevu. Těžko asi budeme zjišťovat pravděpodobnost spadnutí rampouchu na moji hlavu, když se mi to stalo zatím pouze jednou. Náhodné jevy můžeme různě kombinovat, např. sjednocovat či vytvářet společný průnik. Dva náhodné jevy nazveme disjunktní, když právě jejich průnik je nemožný jev.

Realizace sjednocení náhodných jevů představuje realizaci alespoň jednoho náhodného jevu ze sjednocení, naopak průnik představuje realizaci všech jevů tvořících

průnik. U disjunktních jevů není možné, aby se realizovaly současně.
V obvyklé mluvě se slova pravděpodobnost či pravděpodobné používají celkem běžně a každý intuitivně vnímá, co znamená, že něco je pravděpodobnější nežli něco jiného. Tím vlastně náhodným jevům přisuzujeme míru či váhu, která oceňuje šanci jejich realizace. Rovněž tak teorie pravděpodobnosti a tím i matematická statistika se snaží náhodné jevy vážit jejich pravděpodobností, což je obecně číslo mezi 0 a 1. Jakým způsobem to teorie pravděpodobnosti dělá, je poměrně velice náročné na matematický aparát. Matematická statistika na základě napozorovaných dat se snaží pravděpodobnosti náhodných jevů, které nás zajímají, odhadovat a porovnávat s našimi představami či požadavky o pravděpodobnostech těchto jevů. Rozhodně nás např. bude zajímat, jaká je pravděpodobnost výskytu neshodného kusu při stávajícím stavu výrobního procesu. Výskyt neshodného kusu je náhodný jev mající při výrobě hromadný charakter a my dopředu nevíme, zdali právě kontrolovaný kus je shodný či neshodný, je nutné si uvědomit, že pravděpodobnost výskytu neshodné jednotky závisí na stavu výrobního procesu a může se v čase měnit vlivem různých příčin, které mohou proces ovlivňovat.

Cílem majitele procesu je dostat ho do takového stavu, kdy tato pravděpodobnost bude co nejmenší či bude odpovídat požadavkům zákazníka a bude v čase stálá, což znamená, že „proces bude pod kontrolou“.

Výskyt neshodného kusu můžeme označit jedničkou, výskyt shodného kusu pak nulou. Tím náhodnému jevu přiřadíme číselnou charakteristiku, která má náhodný charakter, protože teprve po zkontrolování vyrobeného kusu zjistíme, zdali se jedná o 1 či o 0. Takové číselné charakteristice říkáme náhodná veličina, zde je zatím její nejjednodušší forma. Náhodnou veličinou ovšem složitější povahy je vlastně každý znak jakosti, který sledujeme na výrobku a který má číselný charakter. Pojem náhodné veličiny je jedním ze základních pojmů jak teorie pravděpodobnosti, tak matematické statistiky.
Náhodné veličiny mající číselnou povahu rozdělujeme do dvou kategorií. Jedny jsou diskrétní povahy a obvykle jejich naměřené hodnoty jsou z přirozených čísel a 0, takové veličiny se též nazývají atributivní a používají se k popisu počtu neshodných kusů či neshod. Druhou kategorii tvoří veličiny spojitého charakteru a ke zjištění jejich hodnot se obvykle potřebuje nějaké měřicí zařízení, a proto jsou též nazývány měřitelné. Jejich oborem hodnot jsou buď celá reálná přímka, polopřímka či konečný interval podle jejich povahy. Někdo by mohl namítnout, že na jakémkoliv měřidle lze odečíst pouze konečný počet různých hodnot, který je odvislý od přesnosti měřidla na kolik desetinných míst je možnost hodnotu určit, a tím vlastně i v tomto případě máme veličiny diskrétní povahy. Protože oněch různých hodnot odečitatelných ze stupnice měřidla může být i několik stovek či i více, s takovým souborem hodnot náhodné veličiny by se pracovalo velice krkolomně, proto se lépe matematika v tomto případě vyrovná se spojitým případem.

K tomu, abychom mohli využít patřičné statistické nástroje, potřebujeme data. V zásadě platí, čím více dat, tím lépe, ale nelze to brát jako přikázání, protože sběr dat se odehrává v čase a během odběru se mohou měnit podmínky, za nichž sběr dat probíhá. Změna podmínek obvykle vyvolává i změnu v chování dat, i když jsou náhodné povahy. O požadavcích na kvalitní data budeme hovořit později.

Nyní si na příkladu ukážeme, jak funguje statistické myšlení. Představme si, že chceme zjistit, jak vypadá rozdělení hmotnosti v mužské populaci v České republice u věku nad 18 let. Teoreticky by bylo možné zvážit každého mužského, ale to je z mnoha důvodů, např. časových a organizačních nemožné, proto musíme oslovit statistiku. Statistika udělá to, že navrhne vybrat pouze relativně malou část z mužské populace, která v tomto případě představuje tzv. základní soubor neboli populaci. Základní soubor bývá sice v mnoha situacích konečný, ale velice rozsáhlý, lze si představit i základní soubor někdy jako nekonečný. Skládá se ze všech jednotek, na nichž by se sledovaná náhodná veličina dala zjistit či změřit. V našem případě by náš základní soubor měl několik miliónů jedinců. Jakým způsobem vybrat onen vzorek a jak velký, na to statistika má nástroje, tzv. statistická šetření pro uspořádání výběrů a rozsah vzorku se také řídí požadavkem na přesnost našich závěrů. Intuitivně chápeme, že vybraný vzorek, neboli výběr by měl nějak rovnoměrně pokrývat celou populaci. Nejsprávnější řešení nabízí statistika ve formě tzv. prostého náhodného výběru, kdy je dána každé jednotce základního souboru stejná šance, že bude vybrána do výběru. To by se dalo zařídit v dnešní době třeba přes

databázi rodných čísel, kdyby se z množiny rodných čísel náhodně vybralo např. kolem dvou tisíc. Otázkou ale pak stejně zůstává, jak k těm vybraným rodným číslům sehnat ty jedince a změřit je. Ať už máme i jinak zorganizován sběr dat, např. systematický výběr založený na kontrole každého desátého kusu na výrobní lince, je nutno si uvědomit, že v každém případě máme v sebraných datech pouze neúplnou informaci o chování naší sledované náhodné veličině a my se snažíme na základě této omezené informace učinit celkovou představu o chování této veličiny, čili děláme závěr, jak to vypadá v celém základním souboru. Tedy usuzujeme z menšího na větší, a to je ten induktivní krok, který umožňuje statistika. Tento induktivní krok ale je spojen z rizikem, že naše rozšíření nebude odpovídat realitě v základním souboru. To může nastat, i když sběr dat byl proveden zodpovědně, byly odstraněny všechny příčiny a vlivy, které se daly odstranit, tak, aby sběr dat probíhal za přibližně stejných podmínek. Přesto závěry nemusí být v souladu s realitou. Proč? Protože do hry může zasáhnout náhoda a ta vybrala taková data, jaká jsme naměřili, i když jsme pro jejich analýzu použili ty správné nástroje, které statistika nabízí. S tímto faktem ale matematická statistika již od začátku analýzy dat počítá a jistí se tím, že si apriori onu míru rizika zvolí, např. na úroveň 5%. To ale znamená, že její závěry neplatí 100%ně, nemohou být černobílé, ale vždy s určitou mírou šedi, která je určena právě zvolenou mírou rizika. Abychom pochopili, jak to vypadá v základním souboru, musíme mít nějakou pomůcku, která nám to umožní, tedy udělat ten induktivní krok od výběru k základní populaci. Touto pomůckou je vhodný

matematicko-statistický model, který je vytvořen na základě našich sebraných dat. Matce přírodě je úplně jedno, jaký model použijeme, zde závisí pouze na naši šikovnosti, zkušenostech a vhodných nástrojích matematické statistiky pro posouzení vhodnosti toho či jiného modelu. V případě náhodné veličiny tímto modelem je vhodný tvar rozdělení pravděpodobnosti ve formě tzv. distribuční funkce.

V případě diskrétních náhodných veličin vystačíme obvykle s binomickým, Poissonovým či hypergeometrickým rozdělením, kdežto u spojitých náhodných veličin je situace komplikovanější, i když zde dominantní roli hraje normální rozdělení. Existují ale v praxi náhodné veličiny, které se nedají normálním rozdělením popsat. Co nám vhodně nalezený model dává? Bohužel nám neřekne, jaká hodnota sledovaného znaku jakosti se bude realizovat na dalším vyrobeném kuse, tak to v našem světě nefunguje, to by totiž někde musela existovat budoucnost a měli bychom možnost do ní nahlédnout. Ale jsme schopni třeba odhadnout pravděpodobnost výskytu neshodného kusu či určitého počtu neshod či s jakou pravděpodobností se spojitý znak jakosti může objevit mimo specifikace, a tím hodnotit stav výrobního procesu. Nalezení vhodného modelu to je maximum, které lze z dat o naší náhodné veličině získat.

Pokud máte k dispozici tvar rozdělení, tedy odpovídající distribuční funkci, tak z hlediska matematické statistiky máte vše, co potřebujete vědět o sledované náhodné veličině. Samozřejmě, že statistika má ve svém arzenálu nástroje pro volbu vhodného typu rozdělení pravděpodobnosti, kterými lze objektivně posoudit, který

typ rozdělení může přicházet v úvahu. Jedná se tak o tzv. testy dobré shody.
Někdo by mohl namítnout, že při 100%-ní kontrole, kdy je proměřován každý vyrobený kus, vlastně základní populaci známe celou, a to celou sérii vyráběných kusů jednoho typu. Problém je v tom, že pro hlubší analýzu, např. pro odhad pravděpodobnosti, s jakou se hodnota sledovaného znaku jakosti může vyskytnout mimo specifikace, pokud všechna naměřená data jsou uvnitř specifikačních mezí, je jedinou šancí nalezení vhodného modelu, s jehož pomocí je požadovaná pravděpodobnost odhadnuta. Zde tedy je nutné a i vhodnější základní populaci uvažovat zcela abstraktně, a to všechny možné hodnoty sledované veličiny, což může být i celá reálná přímka. Lze si situaci představit tak, že základní populace, zde číselná osa, je vlastně osudí, z něhož prostřednictvím měření nám Matka příroda vytahuje čísla s pomocí náhody a my pomocí vhodného modelu se snažíme popsat, jak to provádí. Pojem základního souboru obvykle dělá na začátku problém, protože si lidé mnohdy nedovedou ujasnit, co vzít za základní populaci. Abychom se tím dále netrápili, je možno se dohodnout, že v případě diskrétních náhodných veličin základním souborem či populací budou nula a přirozená čísla a u spojitých náhodných veličin celá reálná přímka bez jakéhokoliv vztahu k jednotkám, na nichž je měření či zjišťování hodnot příslušné náhodné veličiny prováděno. Budeme si tedy představovat, že realizaci hodnot sledované náhodné veličiny, tedy při jejím měření či sledování, jsou naměřené hodnoty vytaženy náhodně ze základního souboru a prvotním úkolem matematické

statistiky je pomocí volby vhodného modelu pochopit, jak se zjištěné hodnoty realizují vlivem náhody.

Sběr dat.
Již bylo dříve řečeno, že kvalita dat přímo ovlivňuje kvalitu závěrů učiněných z rozboru dat. Proto je nutno sběru dat věnovat patřičnou pozornost a jejich sběr velice dobře připravit. Sebraná data by měla být:

relevantní
reprezentativní
v dostatečném množství
vyplývat ze souvislostí.

Relevantnost dat znamená, že se musí vztahovat k problémům či otázkám, které máme pomocí dat objasnit či zodpovědět. Dostatečné množství je sice šalamounsky řečeno, ale neříká to, zdali je to 15 či 2000 hodnot. Bohužel nelze dát obecné doporučení, kolik naměřených či zjištěných hodnot je zapotřebí pro konkrétní problém. Dále to úzce souvisí i s podmínkami, za nichž jsou data sbírána. Někde je to snadné, máme automatické zápisy dat, jindy, např. při destrukčních zkouškách to bývá problém časový i finanční. Mnohdy za potřebná data musíme i zaplatit nemalé sumy. Reprezentativnost znamená, že data popisují důležité vlastnosti, které pomáhají pochopit daný problém a které nezkresleně tyto vlastnosti popisují. Spolu s daty, pokud je to možné, je vhodné zaznamenávat i veškeré změny, které během sběru dat nastaly. To nám pak pomůže třeba rozdělit data do skupin, které mají homogenní charakter vůči podmínkám, za nichž se data sebrala. To může znamenat např. rozdělení dat podle vedlejších příznaků,

jako je operátor, směna, vstupní materiál, zásah do procesu apod.
To vede k lepšímu pochopení, co všechno se může promítnout do chování dat. Pokud je dat skutečně malý počet, např. pod 10 hodnot, pak samozřejmě lze použít statistiku, vhodný software vždy něco vypočte, ale zjištěné závěry je nutno brát pouze orientačně a ne na nich bazírovat. Data by samozřejmě neměla být falšována či záměrně pozměňována. Velice důležitým momentem při sběru dat je ověření a analýza měřícího systému. To se týká jak použitých měřidel, tak i lidí, kteří budou měření provádět. V řeči MSA (Measurement Systém Analysis) to znamená provést ověření opakovatelnosti a reprodukovatelnosti jak u měřitelných veličin, tak i u atributivních veličin. Zjištěné či naměřené hodnoty by neměly být zaokrouhlovány, tím by se do dat vnesl další zdroj nežádoucího šumu. V případě spojitých veličin, by mezi naměřenými či zjištěnými hodnotami mělo být alespoň 6-7 navzájem různých hodnot. To úzce souvisí s rozlišovací schopností použitého měřidla. Můžete mít několik stovek dat, ale opakují se mezi nimi stále pouze tři různé hodnoty a veličina má evidentně spojitou povahu, a těžko budete hledat vhodný model pro popis takové veličiny. Je nutné pak vzít přesnější měřidlo, které provede měření alespoň o jedno desetinné místo navíc. Dále by mělo být důkladně zváženo, kam a v jaké formě se data budou zaznamenávat či zapisovat a kdo to bude popřípadě provádět. Je tedy vhodné mít již připravený např. formulář pro sběr dat ať už na papíře či v nějakém softwaru.
Pokud máme data sebrána, můžeme přistoupit k dalším krokům. Dalším důležitým krokem je ověření

věrohodnosti dat. O co se jedná? Jde o to, abychom mohli data považovat za spolehlivá, aby mezi daty nebyly nesmyslné hodnoty, které třeba vzniknou špatným zápisem. Pro takovou analýzu se např. velice dobře hodí průběhový diagram, v němž jsou data prostě zaznamenána v tom pořadí, jak byla naměřena. Tento nástroj bude podrobněji popsán v kapitole popisující grafické metody. Existují i další grafické nástroje, např. tzv. kapénkový diagram (v angličtině dot-plot). Může se někdy stát, že se některá data zdají, jako kdyby nepatřila do hlavního oblaku naměřených hodnot a jakoby ležela mimo. Takovým hodnotám říkáme odlehlé hodnoty a činí v praxi obvykle potíže. Ihned se naskýtá otázka, co s nimi? Máme s nimi v dalším počítat nebo je máme právo vyloučit, činí potíže již obvykle při hledání vhodného modelu pro sledovanou náhodnou veličinu. Když se nejedná evidentně o špatně zapsané údaje, pak by se mělo postupovat následovně. Pokud víme z dalších informací získaných během měření, že tato data byla získána za skutečně změněných podmínek, které nespíše ovlivnily sledovanou veličinu, pak máme právo odlehlé hodnoty vyloučit. Jestliže ale takové informace nemáme, může být riskantní taková data vyloučit, protože jejich vyloučení může silně změnit naše představy o sledované veličině, nalezený model nebude adekvátní a v budoucnu při dalších měřeních se můžeme dostat do potíží díky nesprávným závěrům učiněných na základě okleštěných dat. Zvláště u některých náhodných veličin, které mohou vykazovat nesymetrické chování, je toto nebezpečí značné. Jedná se o takové veličiny jako je házivost, ovalita, velikost trhací síly, měření úhlů, drsnost povrchu. Nevhodný model zkonstruovaný na základě „očištěných

dat“ pak může vést ke špatnému regulačnímu diagramu či ke špatnému hodnocení způsobilosti výrobního procesu. Matematická statistika má vhodné nástroje na testování odlehlých hodnot, ale tyto nástroje fungují za splněných určitých předpokladů. V některých softwarech jsou takové podezřelé hodnoty extra označeny, ale to neznamená, že je máme právo automaticky vyloučit. V matematické statistice po projití touto fází pak nastupují především grafické metody a popisná statistika, která vypočítává nejdůležitější číselné hodnoty získané z původních dat.

3. Popisná statistika

Popisná statistika umožňuje výpočet základních číselných charakteristik, které nám pak pomáhají co nejlépe využít informaci obsaženou v naměřených hodnotách. Představme si, že odebíráme data pravidelně z kontrolní skupiny třeba pěti výrobků každou hodinu. Za celou směnu tak máme několik desítek údajů třeba zapsaných v tabulkách v Excelu, ale i když se na tabulky budeme dívat sebedéle nic nám neřeknou a z dat nic nepoznáme.

My se budeme výhradně zabývat daty numerické povahy, i když existují i data tzv. kategoriální povahy, která nejsou vyjádřena čísly. Jedná se např. o zaměstnání, politickou příslušnost, pohlaví apod. Tato data se zpracovávají obvykle pouze grafickými metodami

Při grafickém zobrazení dat numerické povahy intuitivně cítíme, že v datech existuje jakási hodnota, která by se dala nazvat středem nebo nejočekávanější hodnota a pak další charakteristika, a to jejich rozptýlenost neboli variabilita. Aniž bychom něco blíže věděli o statistice, při

pohledu na dva soubory dat jsme schopni říci, že oba soubory mají asi stejné středy a stejnou variabilitu. Čím větší variabilita v datech, tím ve větší vzdálenosti se vyskytují naměřené hodnoty od onoho pomyslného středu. Popis dat pomocí těchto dvou charakteristik je náš subjektivní vjem nebo je to objektivní? Naměřené hodnoty se skutečně charakterizují těmito dvěma pojmy. Jedná se o pojmy parametr polohy a úroveň variability.

Se základní populací jsou tedy spojena dvě čísla, které obvykle neznáme, můžeme je z naměřených dat pouze odhadovat. Jedna charakteristika je parametr polohy, druhá je směrodatná odchylka. Hodnota parametru polohy představuje očekávanou hodnotu, která by byla naměřena, kdyby úroveň variability byla nulová. Směrodatná odchylka je mírou úrovně variability, čím je větší, tím také naměřené hodnoty vykazují větší rozptýlenost kolem hodnoty parametru polohy. Hodnota parametru polohy obecně může být jakékoliv reálné číslo, směrodatná odchylka je vždy kladná. Jakým způsobem čerpáme informace o těchto parametrech základní populace z dat?

Poloha dat na číselné ose se v praxi nejčastěji odhaduje výběrovým průměrem *(který označujeme* $\bar{x}$*)* či výběrovým mediánem *(který označujeme* $\tilde{x}$*)*. Výběrový průměr z dat spočítáme tak, že všechna data *(i-tou napozorovanou hodnotu označujeme* x_i*)* sečteme a podělíme jejich počtem (n), tedy

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,.$$

Výběrový průměr není nic jiného nežli aritmetický průměr z dat.

Pro spočítání výběrového mediánu potřebujeme data uspořádat podle velikosti a v případě lichého počtu dat vybereme prostřední hodnotu podle pořadí, tedy když bude k dispozici 27 hodnot, pak výběrový medián je právě čtrnáctá hodnota v pořadí. Pokud by dat byl sudý počet, pak vezmeme dvě prostřední hodnoty a uděláme z nich aritmetický průměr. Když budeme mít např. 30 hodnot, pak výběrový medián je aritmetický průměr z patnácté a šestnácté hodnoty v pořadí podle velikosti. Proč slovo výběrový? To je proto, že je počítán z dat, tedy z výběru. V základní populaci existuje totiž jeho protějšek, zvaný pouze medián, který rozděluje základní populaci na dvě poloviny. Pravděpodobnost, že naměřím hodnotu pod tímto mediánem je 50%, zrovna tak pravděpodobnost, že naměřím hodnotu nad mediánem. Ale tak jako neznáme hodnotu parametru polohy *(protějšek výběrového průměru v základním souboru je parametr polohy – střední hodnota)*, neznáme ani hodnotu mediánu základní populace, z dat ji můžeme pouze odhadnout, jak popsáno výše.

Někdo ihned namítne, kdy parametr polohy odhadovat pomocí výběrového průměru, a kdy pomocí výběrového mediánu. To silně závisí na povaze dat. Pokud data vykazují symetrické chování a lze očekávat, že i vhodný model pro popis sledované veličiny bude mít symetrickou skladbu, jako např. normální rozdělení, pak přichází v úvahu výběrový (aritmetický) průměr. Když ale data trpí asymetrií, pak je vhodnější výběrový medián. Vysvětleme si to na příkladu s hrubými platy v české republice. Podle statistických údajů se průměrný hrubý plat pohybuje někde za hranicí 22tisíc. Ten je spočítán skutečně jako aritmetický průměr ze získaných

dat, které dodávají firmy. Kdyby se spočítal ale mediánový plat, dojde se k hodnotě podstatně menší, odhad mezi 15 – 16 tisíci, a ten pak říká, že 50% pracovníků má hrubý příjem pod touto hodnotou a druhých 50% nad touto hodnotou. To lze vysvětlit tím, že medián není citlivý na odlehlá data, zde představovaná stotisícovými platy manažerů, kdežto aritmetický průměr je zaznamená.

Dále je nutné si uvědomit, že výběrový průměr i výběrový medián jsou vlastně další náhodné veličiny, které vstupují do hry, jsou totiž počítány z naměřených dat a kdybychom měli jiná data, jejich hodnoty by byly jiné. Kdežto hodnota parametru polohy se nemění, to je nenáhodné číslo. Jeho změna může být vyvolána jedině tím, že se změní podmínky při sběru dat, např. vlivem nějaké systematické příčiny, která se objevila. Data pak vykazují posun v parametru polohy. Proto je nesmírně důležité, pokud to jde, zamezit jakýmkoliv změnám v průběhu sběru dat, aby se podmínky v základní populaci neměnily a měli jsme možnost parametr polohy věrohodně odhadnout. Pokud se podmínky změní, a jsme schopni změnu zaznamenat, znamená to pro nás velice důležitou informaci, že nastala změna v základní populaci a data sebraná po změně se asi budou chovat jinak nežli data před změnou. Změna v parametru polohy se může pak projevit v tom, že aritmetický průměr z dat před změnou se může významně lišit od aritmetického průměru z dat získaných po změně a matematická statistika pak dovede odpovědět na otázku, zdali je tato změna statisticky významná nebo se to stalo pouze náhodou.

Druhou charakteristikou pro chování dat je úroveň variability. Ta se nejčastěji odhaduje na základě dat pomocí výběrové směrodatné odchylky, obvykle se označuje písmenem *s*, ale není to psané pravidlo. Vzorec pro výpočet je následující:

$$s = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})^2} \,.$$

V průmyslové praxi, např. u regulačních diagramů, se používá další číselná charakteristika pro úroveň variability, a to výběrové rozpětí. To se spočte jednoduše jako rozdíl maximální a minimální hodnoty z naměřených dat. Tato veličina se často označuje jako *R*, podle anglického názvu *range*. Tedy

$$R = Max(x_i) - Min(x_i).$$

Je nutné si opět uvědomit jako u parametru polohy, že k výběrové směrodatné odchylce, což je náhodná veličina spočtená z dat, podobně jako je výběrový (aritmetický) průměr, existuje pro základní populaci míra pro úroveň variability zvaná směrodatná odchylka, která charakterizuje variabilitu v základní populaci, nemění se a je nám obvykle neznáma. My ji dovedeme pouze odhadovat pomocí dat, např. pomocí výběrové směrodatné odchylky.
Samozřejmě se může někdo ptát, proč zrovna takové vzorečky pro odhady těchto dvou základních charakteristik chování získaných dat od sledované náhodné veličiny. Zatím právě stojí ta druhá strana

matematické statistiky, ta teoretická, která pomocí aparátu teorie pravděpodobnosti odvodí a zdůvodní, proč tyto vzorce. Jde o to, abychom vytáhli z dat co největší informaci o náhodné veličině a aby taky používané odhady měly rozumné vlastnosti. Např. si představme, že bychom měli možnost rozsah získaných dat neomezeně zvětšovat a současně bychom měli zajištěno, že podmínky sběru dat se v čase nemění. To by znamenalo, že též parametr polohy a směrodatná odchylka se v základní populaci nemění a my bychom se v limitě k těm neznámým hodnotám pomocí postupných výpočtů aritmetických průměrů a výběrových směrodatných odchylek libovolně blízko přiblížili a tak získali jejich přesné hodnoty. To ale samozřejmě v praxi nejde, máme k dispozici, jak už bylo řečeno, pouze neúplnou informaci o základní populaci, a proto naše odhady vykazují vždy nějakou chybu. Jak ta chyba je veliká, nám později řeknou tzv. konfidenční intervaly konstruované pro parametry základní populace.

Výběrová směrodatná odchylka i směrodatná odchylka pro základní populaci jsou vyjádřitelné ve stejných jednotkách jako původní naměřené hodnoty. Kvadrátům těchto charakteristik se říká výběrový rozptyl a rozptyl. Dosti často se ve statistické literatuře používá pro výběrové charakteristiky označení pomocí latinky a pro jejich protějšky ze základní populace odpovídající písmeno řecké abecedy, tedy pro ilustraci, my jsme označili výběrovou směrodatnou odchylku jako s, pak odpovídající označení směrodatné odchylky pro základní populaci je σ. U parametru polohy je trochu výjimka, dosti často se protějšek výběrového průměru označuje písmenem μ.

Dalšími charakteristikami pro popis dat jsou tzv. *percentily*. Pro jejich získání je nutno data vždy uspořádat podle velikosti od nejmenší hodnoty k největší. Zvolme číslo *k* mezi jedničkou a stem. Pak *k*-tý percentil znamená číslo, že *k* procent z naměřených hodnot je právě pod tímto percentilem, neboli *100-k* procent hodnot je nad tímto číslem. Jak takový percentil z dat vypočteme? Nechť *n* je celkový počet dat, tedy rozsah souboru. Vynásobíme *k* procent tímto *n* a podělíme 100. Pokud dostaneme celé číslo, pak najdeme v uspořádané řadě podle velikosti z našich dat hodnotu s tímto pořadím a vypočteme aritmetický průměr z této nalezené hodnoty a hodnoty bezprostředně následující. Z uspořádané řady čísel nesmíme vyloučit čísla se opakující. Pokud nedostaneme celé číslo, zaokrouhlíme ho směrem nahoru a odpovídajícím percentilem je hodnota s tímto pořadím v uspořádané řadě získaných hodnot. Ukažme si to na příkladě. Máme naměřené hodnoty
43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99.
Rozsah souboru je 25, tedy $n = 25$.. Chceme zjistit 90tý percentil. Pak $90*25/100 = 22,5$. Protože to není celé číslo, zaokrouhlíme nahoru a máme 23 a dvacátým třetím číslem v uspořádané řadě hodnot je 98, a to je vypočtený percentil. Kdybychom chtěli 20-tý percentil, tak $20*25/100 = 5$ a tedy aritmetický průměr z pátého a šestého čísla v pořadí hodnot je $(62+66)/2 = 64$. Výběrový medián je tedy 50-tý percentil.
Důležitá je správná interpretace percentilu. Percentil v žádném případě neznamená procento, ale dává odhad hranice, pod níž se právě vyskytuje to procento hodnot sledované veličiny. Proč odhad? Protože percentily jsou

počítané z dat a jiná data znamenají i jinou hodnotu téhož percentilu. Když se vrátíme k příkladu s hrubými příjmy v České republice, pak kdybychom měli k dispozici spolehlivý soubor hodnot příjmů, pak by nám jednotlivé percentily odhadovaly, jak např. vypadá výše příjmu , pod nímž má plat např. 30% pracovníků. Ta hranice by byla právě třicátým percentilem v souboru platů.

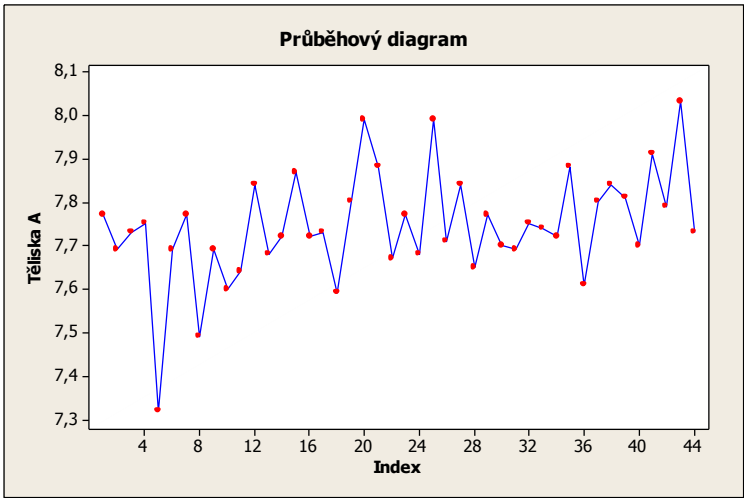
K čemu nám mohou percentily posloužit? Získáme pomocí nich přesnější představu o rozdělení hodnot sledované veličiny v základní populaci. Když se např. bude hodně lišit rozdíl výběrového mediánu a desátého percentilu od rozdílu mezi devadesátým percentilem a výběrovým mediánem, tak rozdělení sledované veličiny asi nebude symetrické. Tuto skutečnost bychom určitě odhalili v případu s hrubými příjmy.

Do popisné statistiky patří rozhodně následujících pět čísel napočítaných z dat. Jedná se o minimální hodnotu, maximální hodnotu, výběrový medián a 25-tý a 75-tý percentil. Těmto speciálním percentilům se říká kvartily, tedy dolní kvartil a horní kvartil a jak uvidíme dále, jejich hodnoty se používají pro tvorbu krabicového diagramu. V našem souboru dat výše uvedených pak minimální hodnota je 43, maximální je 99. Dolní kvartil, značí se obvykle Q1 má hodnotu 68, výběrový medián je 77 a horní kvartil Q3 = 89. Někdy se jako odhad pro míru variability používá tzv. kvartilové rozpětí, tedy rozdíl Q3 - Q1. V našem případě to dává hodnotu 21.

4. Grafy a diagramy

V této kapitole se opět zaměříme především na data s numerickými hodnotami. Grafické znázornění dat má tu velkou výhodu, že je velice názorné a dovede řadu věcí odhalit a vytvořit tak řadu hypotéz o chování sledované veličiny, na jejich ověření či vyvrácení se pak použijí další statistické nástroje. Do této kategorie spadají sloupcové diagramy a spojnicové diagramy, především průběhový diagram a histogram. Tyto dva grafické výstupy jsou používané v případě spojitých veličin, pro diskrétní veličiny se průběhový diagram rovněž hodí, ale místo histogramu se používá diagram četností. Jakmile máme k dispozici naměřené hodnoty, je doporučeno, sestrojit jejich průběhový diagram, který zaznamenává hodnoty v tom pořadí, jak byla data naměřena. Zachování pořadí, v jakém data byla zaznamenána, je velice důležitá informace, protože již z průběhu dat lze se ledacos dozvědět, aniž bychom zatím dělali nějakou podrobnou statistickou analýzu. Tak lze objevit např. periodicitu, nápadná seskupení nenáhodného charakteru, možné trendy, ulétlá pozorování apod. Na základě těchto pozorování lze pak vytvářet statistické hypotézy o chování sledované veličiny a ty pak testovat. Při porovnávání dvou či více průběhových diagramů třeba téže sledované veličiny ale z různých časových úseků, je nutno mít stejná měřítka na svislé ose, abychom skutečně srovnávali srovnatelné. Rovněž na vodorovné ose by měla být pro případ srovnání průběhů zachována stejná měřítka.

Dole na Obr.1 lze vidět takový průběhový diagram zaznamenávající naměřené hodnoty sledovaného znaku jakosti, zde délka osičky, získávané při aplikaci regulačního diagramu pro individuální hodnoty.



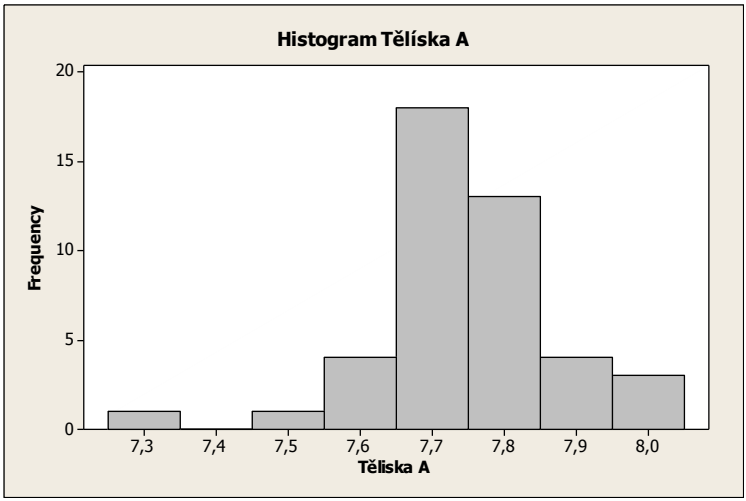
Obr. 1: Průběhový diagram

Při prohlídce tohoto průběhového diagramu asi každého napadne, zdali se v datech neskrývá nějaký trend. To je jedna hypotéza, ale odpověď na ni musí dát seriozní analýza, v tomto případě nejspíše lineární regrese, která se pokusí daty proložit přímku a její vhodnost otestovat. Jak již bylo řečeno, lze průběhový diagram použít jak na diskrétní dat, tak na data spojitého charakteru.
Dalším velice užitečným nástrojem je histogram. Základní myšlenka histogramu je rozdělení dat do vhodně zvolených číselných intervalů a sestrojení

sloupců nad jednotlivými intervaly, jejichž velikost odpovídá počtu hodnot, které připadly do toho kterého intervalu. Kolik získaných hodnot je zapotřebí mít? Obvykle se doporučuje, aby naměřených hodnot bylo nejméně 25-30. Pokud budeme mít hodnot několik stovek k dispozici, je samozřejmě možné udělat ze všech jeden histogram, ale mnohdy stojí za úvahu data rozdělit na více úseků a sestrojit nad každým úsekem histogram a podívat se, jak se budou případně lišit či podobat. Kolik intervalů máme uvažovat? Pokud máte k dispozici vhodný statistický software, ten to provede za Vás. Např. ale v Excelu je volba počtu intervalů a jejich šířky ponechána na uživateli. Nejmenší počet intervalů by se měl pohybovat mezi 7-8. To úzce souvisí z rozlišitelností měřícího zařízení, jak už bylo zmíněno dříve. Pro maximální počet intervalů se doporučuje 15-20, v literatuře lze najít jistá doporučení, např. že počet intervalů by měl být cca jako logaritmus z počtu dat. Výška sloupců nad jednotlivými intervaly může být jak v absolutních počtech dat z intervalů, tak i relativních četnostech. Je nutné si uvědomit, že tvorba histogramu bez nějakého softwaru je zcela subjektivní záležitost a dva jedinci ze stejných dat tedy mohou sestrojit dva různé histogramy. Na Obr.2 je vidět histogram sestrojený pomocí vhodného statistického softwaru z dat Tělíska A zobrazených v Obr.1.

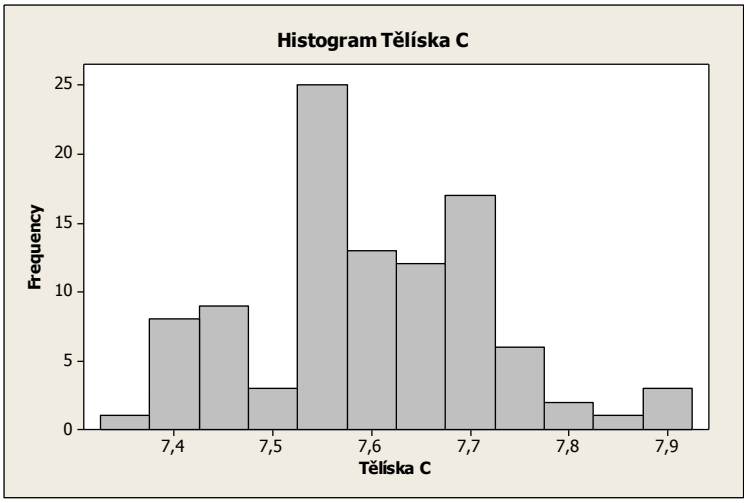
Co lze z histogramu odhalit? Především kde se asi dá očekávat Hodnota parametru polohy, jaká je úroveň variability dat a hlavně zdali data vykazují spíše symetrické či asymetrické chování kolem hodnoty parametru polohy.

Histogram je důležitý nástroj při hledání vhodného modelu pro popis chování sledované náhodné veličiny v základní populaci. Dále histogram může odhalit, zda data nepocházejí z více zdrojů, tedy z více základních populací, např. dvouvrcholový diagram na Obr.3.



Obr.2: Histogram z dat Tělíska A

Na Obr.2 jsou velikosti sloupců v absolutních četnostech, lze říci, že data vykazují celkem symetrické chování, vhodný model pro popis odpovídající náhodné veličiny délka tělíska by se hledal asi u normálního rozdělení.

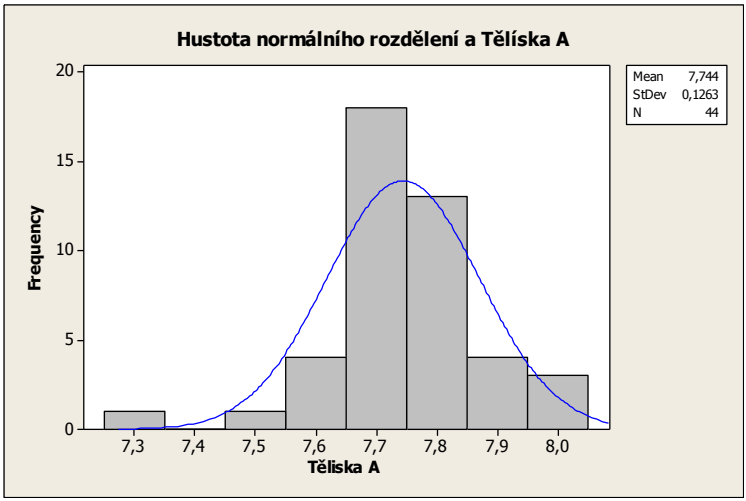


Obr.3: Histogram z dat Tělíska C

Jak vidno, data Tělíska C již nevykazují symetrické chování, to znamená, že nejspíše během tohoto měření veličiny délka tělíska došlo ke změně v parametru polohy, možná i úrovně variability, vlivem nějaké příčiny. Statistika nám ale neřekne, v čem je případná změna, ale pouze nás na tuto možnost upozorní.

Na dalším Obr. 4 je snaha najít vhodný model v normálním rozdělení, čili jak se říká „nafitovat“ vhodný model na data. Na toto statistika má vhodné nástroje, které patří do tzv. testů dobré shody, které objektivně posoudí, zdali uvažovaný model je vhodný či nikoliv. Zde důležitou roli hraje zkušenost, případně nějaká předchozí informace z dříve zpracovaných dat. Je nutné si uvědomit, že příroda žádné modely nepoužívá,

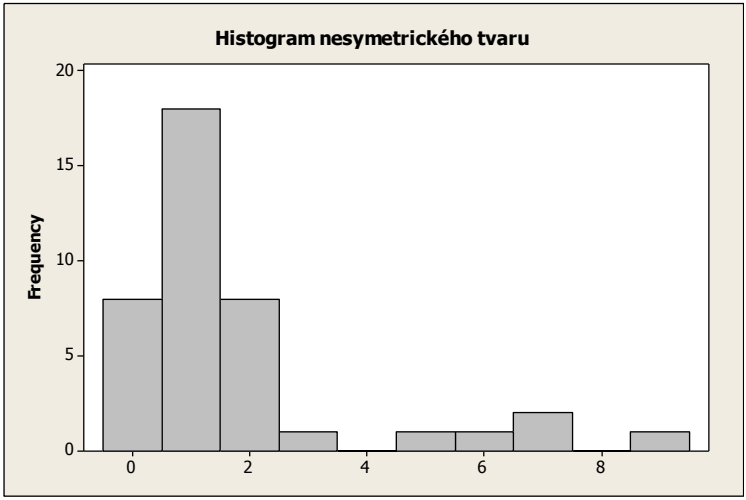
to jenom my se musíme pomocí nich vyrovnat s náhodou.
Abychom možnou změnu odhalili, je vhodné si vedle naměřených hodnot zaznamenávat vše, co se během sběru dat odehraje okolo.



Obr. 4: Proložení hustoty normálního rozdělení daty

Na tomto obrázku je vidět, jakým způsobem histogram slouží k nalezení vhodného modelu pro popis chování sledované veličiny. Samozřejmě bychom se mohli přít, zdali se tento model hodí či nikoliv. Naštěstí kvalifikovanou odpověď dá vhodný test dobré shody, který posoudí vlastně „vzdálenost“ mezi daty a uvažovaným modelem. Může se stát, že kandidátů na vhodnost je více. Pak se řídíme pravidlem „účel světí prostředky“ a obvykle volíme nejjednodušší řešení, pokud je mezi kandidáty normální rozdělení, pak tento

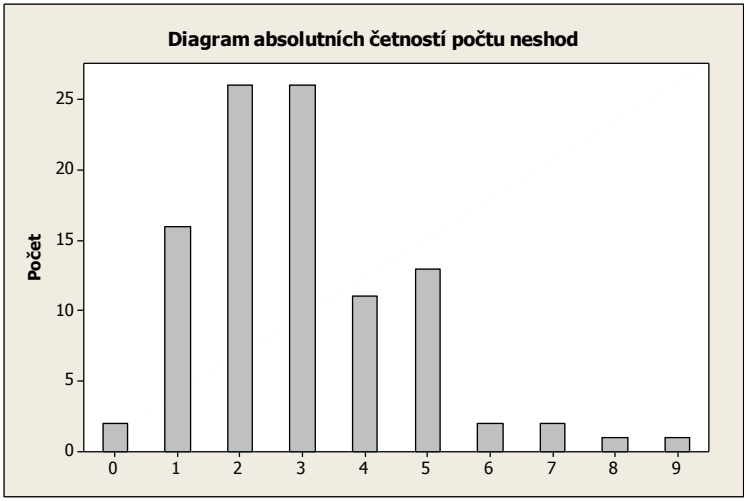
model. Důvod je zcela pragmatický, totiž s modelem normálního rozdělení se snadno pracuje a řada věcí se u něho dá explicitně spočítat, což u jiných modelů tak snadné není. Obecně platí, jakmile jsme nuceni opustit model normálního rozdělení, nastávají komplikace.
Pro úplnost ještě na dalším obrázku je histogram z dat, která evidentně nevykazují symetrické chování.



Obr. 5: Histogram z dat vygenerovaných z modelu lognormálního rozdělení

Na datech z Obr.5 je krásně vidět, jak by bylo možno data umístěná na histogramu v pravé části diagramu, považovat za ulétlé hodnoty. To ale není pravda. Jejich výskyt je vyvolán nesymetrickým chováním sledované veličiny a data byla získána z jiného modelu nežli je normální rozdělení. Takovýto model se může hodit pro popis např. měření házivosti.

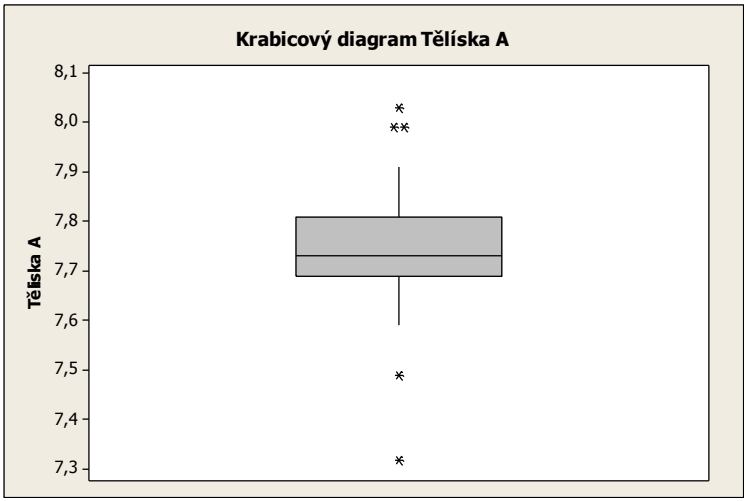
Nyní si ukážeme grafický výstup pro diskrétní dat. Při sledování diskrétních náhodných veličin, jako je např. počet nalezených neshod na výrobku, nás zajímá především zastoupení výskytu počtu nalezených neshod např. za určitou dobu (třeba směnu). Máme tedy znak jakosti, počet neshod na kontrolovaných výrobcích a potřebujeme si udělat představu, kolik výrobků nemělo vadu, kolik bylo s jednou vadou, kolik se dvěma vadami atd. K tomu slouží diagram s četnostmi.



Obr. 6: Rozdělení počtu neshod

Na Obr. 6 je diagram, který znázorňuje rozdělení počtu neshod, které byly zjištěny při kontrole 100 výrobků během jedné směny. Z diagramu je patrno, že mezi sto výrobky byly pouze dva bez vady, 16 výrobků mělo pouze 1 vadu, se dvěma vadami se vyskytlo 26 výrobků atd., dokonce se vyskytl i výrobek s devíti vadami.

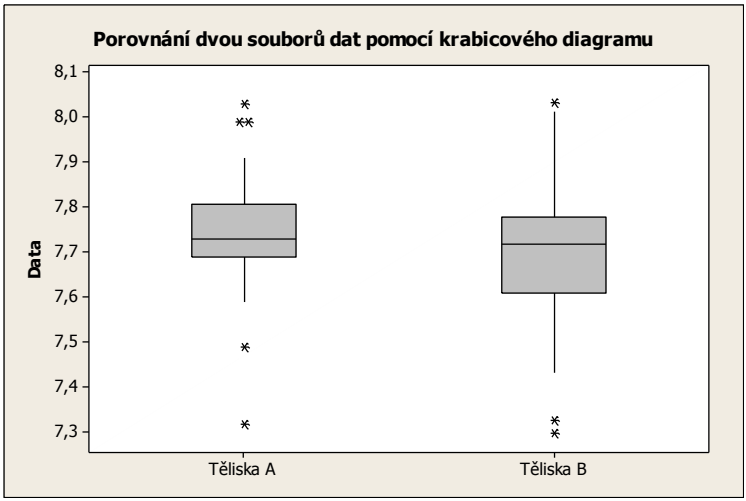
Samozřejmě je možno stupnici na svislé ose vyjádřit i v relativních četnostech či v procentech.
Dalším nástrojem pro grafické zpracování dat spojitého charakteru je tzv. krabicový diagram (box plot). Je zobrazen na Obr. 7.



Obr.7: Krabicový diagram Pro Tělíska A

Prostřední vodorovná čára v krabičce je výběrový medián z dat, horní hrana krabičky je horní výběrový kvartil, dolní hrana krabičky je dolní výběrový kvartil. Dále jsou na krabičce „vousy“, jejichž délka je dána podle toho, jestli jsou nalezeni kandidáti na odlehlá pozorování či nikoliv. Pokud ne, pak délka je dána rozdílem mezi maximální a minimální hodnotou z měření, pokud ano, pak délka „vousů“ se odvozuje od jeden a půl násobku kvartilového rozpětí *Q3 – Q1*. Je jasné, že v krabičce je namačkáno 50% naměřených hodnot z prostředku

oblaku dat, a tím je vlastně i zobrazena úroveň variability a odhadnut parametr polohy. Krabicový diagram je vhodné použít pro porovnání dvou či více situací, např. před zásahem do procesu a po zásahu. To je dobře vidět na Obr.8.

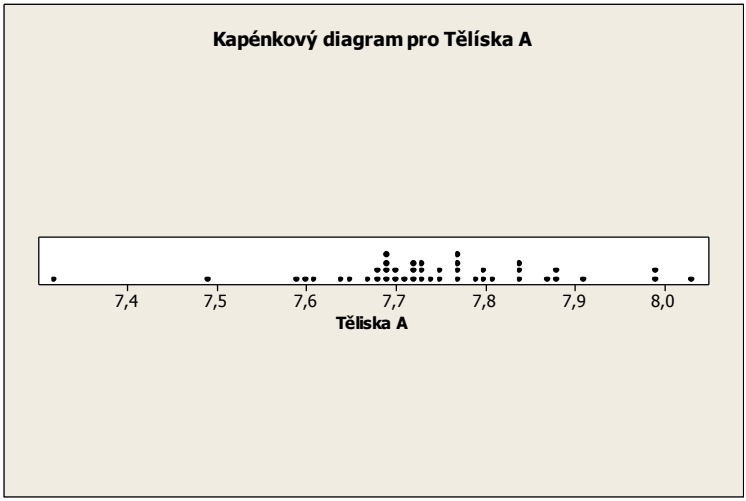


Obr. 8: Porovnání dvou souborů dat

Zde je na prvý pohled vidět, že oba soubory dat se nebudou lišit asi v hodnotě parametru polohy, ale mohou se lišit v míře variability, zdá se, že Tělíska B by mohla vykazovat větší úroveň variability nežli Tělíska A.
Je nutné upozornit na to, že větší krabička u krabicového diagramu neznamená větší počet dat, ale nejspíš větší variabilitu v datech.

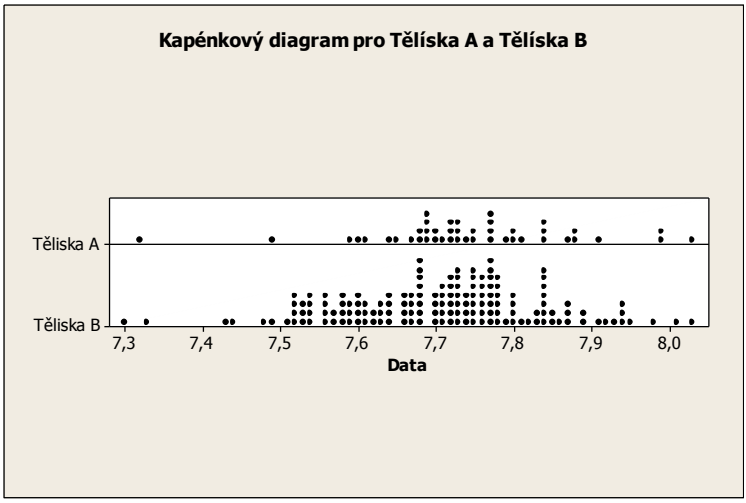
Podobnou roli jako histogram či diagram četností může splnit i kapénkový diagram (dot plot), který zobrazuje

vlastně na rozdíl od histogramu všechny naměřené hodnoty spolu s jejich četnostmi výskytu ve formě sloupců teček na jednotlivými hodnotami (viz Obr.9).



Obr. 9: Kapénkový diagram pro Tělíska A

Je to opět jeden ze šikovných a jednoduchých nástrojů pro porovnání dvou souborů dat. Obdobně jako porovnáním pomocí krabicových diagramů si analogickou představu můžeme udělat tvorbou dvou kapénkových diagramů (viz Obr.10)



Obr. 10: Kapénkový diagram pro dva soubory dat.

Z diagramu je vidět, že lze dojít zcela ke stejným závěrům ohledně souborů dat Tělíska A a Tělíska B jako jsme došli pomocí krabicových diagramů.

5. Binomické a Poissonovo rozdělení

Tato kapitola je věnována nejdůležitějším modelům pro popis diskrétních náhodných veličin. Oborem hodnot pro takovou veličinu jsou čísla 0,1,2,3,…….., teoreticky až do nekonečna. Z hlediska teorie pravděpodobnosti a tedy i matematické statistiky je chování takové veličiny dokonale popsáno, když budeme znát, s jakou pravděpodobností se budou tyto hodnoty při sledování takové veličiny vyskytovat, tedy nás zajímá, např. jaká je pravděpodobnost, že se bude realizovat hodnota 11. Obecně zapsáno, jedná se o pravděpodobnost náhodného jevu, že náhodná veličina, označme si ji *X,* nabude hodnoty *n,* v symbolech

$$\text{Pst}\{ X = n\} = p_n, \quad n = 0,1,2,3,\ldots\ldots,$$

kde Pst je zkratka pro pravděpodobnost. Obor hodnot diskrétní veličiny může být konečný či nekonečný. Obor hodnot náhodné veličiny spolu s odpovídajícími pravděpodobnostmi je tzv. rozdělení náhodné veličiny či její distribuce. Bohužel v praxi sice známe obor hodnot sledované veličiny, ale neznáme ony pravděpodobnosti. Jediná šance je tyto pravděpodobnosti na základě dat odhadnout a pomocí těchto odhadů najít vhodný model pro popis chování veličiny.

Pro praktické účely v případě diskrétních veličin, jako např. počet neshodných kusů nebo počet neshod zjištěných pro účely sledování stavu odpovídajícího výrobního procesu, obvykle vystačíme se dvěma modely, a to s binomickým rozdělením pro počet neshodných kusů a s Poissonovým rozdělením pro počet neshod.

Popis binomického rozdělení.

Toto rozdělení má konečný obor možných hodnot náhodné veličiny, a to {0,1,2,…….., N-1,N}, kde číslo N představuje např. celkový počet kontrolovaných výrobků na jejich funkčnost, obecně hovoříme o počtu pokusů, které byly provedeny. V našem případě pokus je překontrolování výrobku. Toto číslo je samozřejmě známo. Co obvykle známo není, je pravděpodobnost výskytu neshodného kusu, označme ji písmenem *p*. To je obecně číslo mezi 0 a 1. Obecně toto číslo znamená, že se pokus podařil či naopak, že pokus selhal, to už záleží na interpretaci konkrétního použití modelu binomického rozdělení. Tedy v našem označení, pokud veličina *X* se dá popsat modelem binomického rozdělení, pak

$$\text{Pst}\{\, X=n \,\} = \binom{N}{n} p^n (1-p)^{N-n},$$

kde $n = 0,1,2,3,\ldots\ldots..,N$. Pokud bychom chtěli znát pravděpodobnost, že v kontrolované dávce bude nejvýše třeba 10 neshodných kusů, pak musíme sečíst výše uvedené pravděpodobnosti pro $n=0,1,2,\ldots.,10$. Počítáme tak pravděpodobnost náhodného jevu $\{X \leq 10\}$, tedy obecně pravděpodobnost, že v kontrolované dávce bude nejvýše *k* neshodných kusů, bude

$$\text{Pst}\{X \leq k\} = \sum_{n=0}^{k} \binom{N}{n} p^n (1-p)^{N-n},$$

kde $k = 0,1,2,3,\ldots\ldots..,N$. Této funkci argumentu *k* se říká distribuční funkce, též funkce rozdělení

pravděpodobnosti, a zcela popisuje model binomického rozdělení. Parametr p se z dat nejlépe odhaduje relativní četností výskytu neshodného kusu, tedy když m bude zjištěný počet neshodných kusů mezi N kontrolovanými kusy, pak odhad parametru p bude:

$$\hat{p} = \frac{m}{N}.$$

Pokud jde o hodnotu parametru polohy, ten zde má hodnotu Np. Této hodnotě se též říká střední hodnota či očekávaný počet neshodných kusů a v modelu binomického rozdělení se spočítá podle vzorce

$$Np = \sum_{n=0}^{N} n \,\mathrm{Pst}\{X=n\},$$

což je vlastně obecný vzorec pro výpočet střední hodnoty pro jakýkoliv model diskrétní náhodné veličiny. Z toho ihned plyne, že nejlepším odhadem pro parametr polohy zde je přímo nalezený počet neshodných kusů m.

Pokud se týká úrovně variability, pak směrodatná odchylka vyjadřující variabilitu v celém základním souboru je rovna $\sigma = \sqrt{Np(1-p)}$. Toto se zjistí odmocněním odpovídajícího rozptylu, který se vypočte z obecného vzorce pro rozptyl

$$\sigma^2 = \sum_{n=0}^{N} (n - Np)^2 \,\mathrm{Pst}\{X=n\},$$

což v případě binomického rozdělení je $Np(1-p)$.
Vzorce pro výpočet základních parametrů v základním souboru si vlastně matematická statistika vypůjčila z fyziky pevného tělesa, protože parametr polohy, neboli střední hodnota, není nic jiného nežli těžiště, když si představíme rozdělení pravděpodobnosti jako rozdělení hmoty na oboru hodnot náhodné veličiny. Rozptyl pak není nic jiného, nežli druhý centrální moment. Jak se vypočítají příslušné hodnoty distribuční funkce, známe-li parametr p? Jedna možnost je pomocí tabulek, které se najdou v každé učebnici statistiky, a nebo pomocí vhodného softwaru, např. i pomocí Excelu.
Jaká nebezpečí mohou vzniknout při nesprávném použití modelu binomického rozdělení?

1. Počet prováděných pokusů N musí být dopředu znám.
2. Výsledek pokusu musí představovat pouze dvě vylučující se možnosti – úspěch či neúspěch
3. Pravděpodobnost p musí být neměnná během provádění pokusů
4. Pokusy mezi sebou musí být nezávislé, tedy výsledek jednoho pokusu nesmí ovlivňovat výsledek jiného pokusu.

Pokud nejsou tyto 4 požadavky splněny, nelze použít model binomického rozdělení. Je zřejmé, že nejhůře se bude v praxi ověřovat či zajišťovat požadavek na neměnnost parametru p. O tom je možno se přesvědčit třeba následujícím způsobem. Data byla sbírána po určitý časový úsek. Rozdělíme si tento časový úsek na několik menších podle nějaké apriorní informace či příznaků a spočítáme odhady parametru p pro každý takový úsek a pomocí nástrojů testování hypotéz otestujeme, zdali je

možno parametr *p* považovat za stejný ve všech kratších úsecích. Mnohdy stačí uvažovat pouze dva tři takové úseky.

Poissonovo rozdělení.

Představme si, že na kontrolované jednotce se může vyskytnout několik různých vad či neshod najednou. Dopředu tedy nevíme, kolik celkem přes všechny překontrolované jednotky bude všech neshod dohromady a nás samozřejmě zajímá, jaký bude průměrný počet neshod na jednotku. Toto číslo samozřejmě odráží stav výrobního procesu a sledovaná náhodná veličina, tedy počet neshod na jednotku, lze popsat modelem Poissonova rozdělení. Obor hodnot takové veličiny je obecně neohraničený, jedná se o {0,1,2,3,…….}a toto rozdělení má jeden parametr, obvykle značený λ, což je kladné číslo, představující právě očekávaný počet neshod na jednotku. Pravděpodobnost, že se vyskytne na jednotce právě *n* neshod, je dána vzorcem

$$\text{Pst}\{Y=n\} = \frac{\lambda^n}{n!}\exp(-\lambda),$$

kde $n = 0,1,2,3,\ldots\ldots\ldots$ a Y označuje náhodnou veličinu mající Poissonovo rozdělení.
Je nutné poznamenat, že zde pojem jednotka zdaleka nemusí představovat skutečně jeden kontrolovaný výrobek, ale jedná se o vhodně zvolenou jednotku, která může být třeba 10 kontrolovaných výrobků, které přicházejí ke kontrole každou hodinu. Dalším případem jednotky může být časový úsek a parametr λ pak může představovat očekávaný výskyt nějakého jevu za tuto

časovou jednotkou. Tímto příkladem může být uskutečněný počet telefonních spojení za časovou jednotku. Tento typ rozdělení pravděpodobnosti se v praxi dá aplikovat na řadu případů, důležitou vlastností tohoto rozdělení je, že binomické rozdělení, které de facto má parametry dva, a to N a p, se dá velice dobře aproximovat právě Poissonovým rozdělením, když N je hodně velké a na druhou stranu p naopak malé (stačí $p< 0{,}1$). Zde pak platí, že $\lambda = Np$.

Parametr λ představuje pro toto rozdělení jak střední hodnotu tak i rozptyl, tedy směrodatná odchylka pro základní populaci je pak $\sqrt{\lambda}$. Parametr λ samozřejmě znám není, máme k dispozici sebraná dat a je nutno tento parametr nějak odhadnout. Na základě teorie lze ukázat, že nejlepším odhadem pro něj je právě zjištěný počet výskytů sledovaných jevů na jednotce, nebo v případě více zkontrolovaných jednotek pak nejlepším odhadem je aritmetický průměr z celkového počtu zjištěných výskytů a z celkového počtu překontrolovaných jednotek.

Distribuční funkce Poissonova rozdělení má pak tvar

$$\mathrm{Pst}\{Y \le n\} = \sum_{j=0}^{n} \frac{\lambda^j}{j!} \exp(-\lambda),$$

pro $n = 0,1,2,3,\ldots\ldots$Jednotlivé pravděpodobnosti i hodnoty distribuční funkce lze buď získat pomocí tabulek, a nebo lze spočítat i v Excelu, tam je k dispozici vhodná funkce.

Jaké problémy se mohou vyskytnout při použití modelu Poissonova rozdělení? Opět největším problémem v praxi bývá ověření požadavku, že během sběru dat se

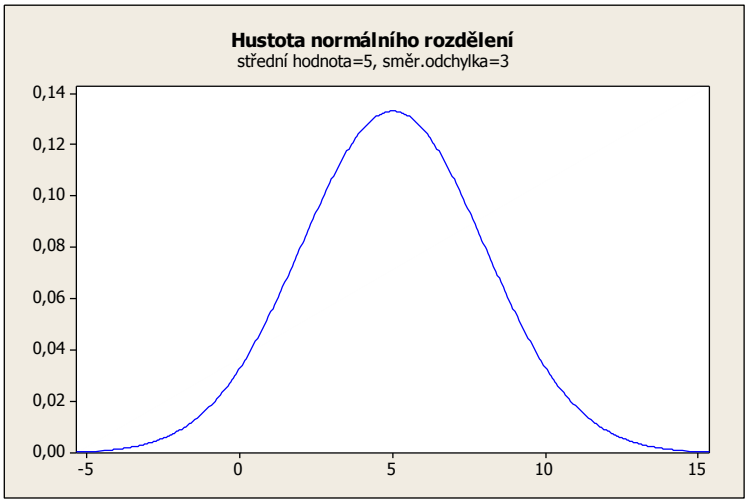
parametr základní populace λ nemění. Jakékoliv zásahy do výrobního procesu, ať už cílené či nežádoucí, mohou se promítnout do změny v parametru λ. Opět podobně, jak popsáno u binomického rozdělení, lze použít nástroje testování statistických hypotéz a podobným uspořádáním jako u binomického modelu otestovat neměnnost parametru λ během sběru dat.

5.Normální rozdělení

Pro spojité náhodné veličiny je normální rozdělení nejdůležitější model, jak z hlediska teoretického, tak i praktického. Tento model má zcela výsadní postavení i v tom, že řada věcí při jeho použití se dá explicitně spočítat a naopak řada nástrojů v matematické statistice je založena na předpokladu normality. Obor hodnot pro náhodnou veličinu popsatelnou normálním rozdělení je celá reálná přímka od $-\infty$ do $+\infty$. Na celém světě se model normálního rozdělení označuje $N(\mu,\sigma^2)$. Parametr μ představuje střední hodnotu, parametr σ je směrodatná odchylka. Model je popsán tzv. hustotou, což je spojitá křivka zvonovitého tvaru (viz Obr. 11), která teoreticky je všude kladná, ale velice rychle napravo i nalevo od střední hodnoty klesá k nule. Plocha pod hustotou má velikost 1 a od tvaru křivky je odvozeno rozdělení pravděpodobnosti.

Na Obr.11 je znázorněna křivka hustoty normálního rozdělení pro $\mu=5$ a $\sigma=3$. Je ihned vidět, že křivka je symetrická podél svislé osy procházející střední hodnotou, proto střední hodnota je i mediánem v základní

populaci. Když se mění parametr polohy μ, tak se nemění tvar křivky, který je dán hodnotou parametru σ. Obecně platí, čím větší σ, tím je křivka roztaženější a náhodná veličina popsatelná takovou křivkou bude vykazovat větší variabilitu. Parametr polohy μ je na hustotě normálního rozdělení ihned vidět, s parametrem σ je to složitější. Délka vodorovné úsečky spojující dva inflexní body na křivce je právě 2σ Speciální případ s μ=0 a σ=1 se nazývá normované či standardní normální rozdělení a značí se N(0,1). Má výsadní postavení, neboť jakákoliv normálně



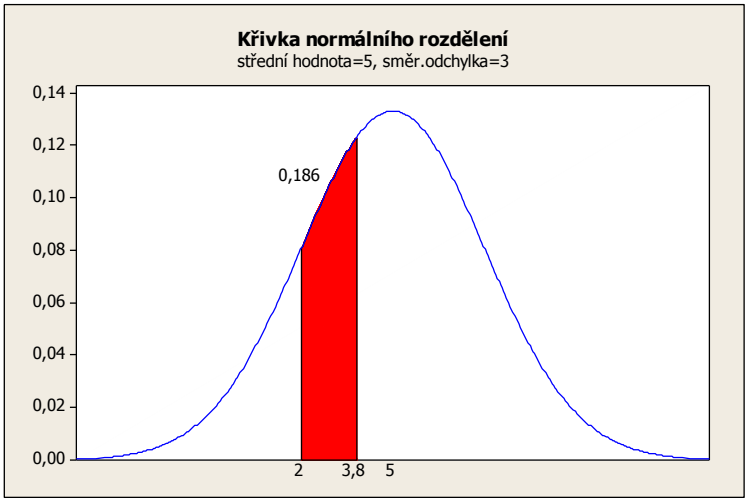
Obr. 11: Křivka hustoty normálního rozdělení.

rozdělená náhodná veličina X lineární transformací

$$Z = \frac{X - \mu}{\sigma},$$

se dá převést na náhodnou veličinu Z, která má rozdělení N(0,1).

Na dalším Obr.12 je znázorněno, jak se spočítá pravděpodobnost, s jakou se realizuje hodnota náhodné veličiny s daným normálním rozdělením v rámci zvoleného intervalu, zde (2; 3,8). Pravděpodobnost je dána plochou vymezenou zvoleným intervalem, pořadnicemi v jeho koncích a křivkou normálního rozdělení (viz vybarvená oblast na Obr.12).



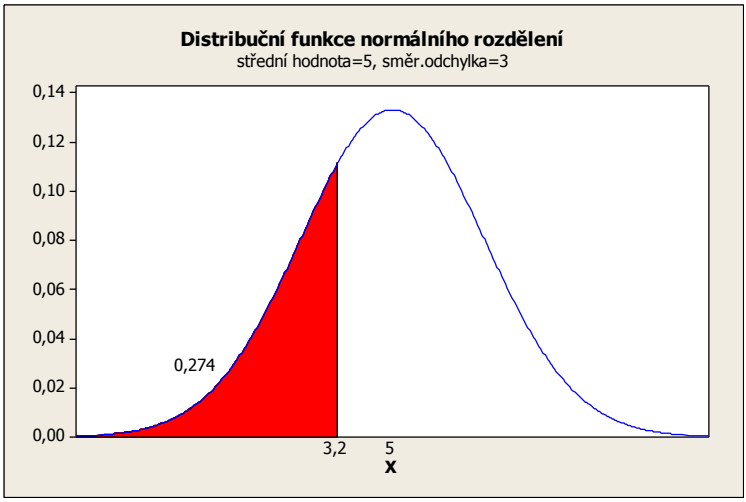
Obr.12: Výpočet pravděpodobnosti pomocí hustoty.

Vidíme, že pravděpodobnost padnutí naměřené hodnoty do intervalu (2; 3,8) je 0,186. Z toho je vidět, že pravděpodobnost realizace jediné zvolené hodnoty, např. 8,65, bude nulová, protože uvažovaná oblast pod křivkou se v tomto případě scvrkne na úsečku, a ta má nulovou plochu. I když takový jev má nulovou pravděpodobnost,

neznamená to, že hodnota 8,65 se nemůže naměřit. To je vlastnost všech spojitých náhodných veličin.
V případě aplikování modelu normálního rozdělení na danou veličinu, samozřejmě parametry μ a σ neznáme a jsme nuceni je z dat odhadnout. Pro odhad parametru μ se nejčastěji používá výběrový průměr, díky symetrii rozdělení se dá použít i výběrový medián, ale aritmetický průměr z dat je vydatnější a má vhodnější vlastnosti nežli výběrový medián. Pro odhad parametru σ je nejčastěji používá již známá výběrová směrodatná odchylka, ale např. v SPC v regulačních diagramech se používá dosti často výběrové rozpětí, které se vynásobí vhodnou konstantou závisející na počtu dat. Dokonce v případě malého počtu dat (do 5-6ti pozorování v rámci logické podskupiny při statistické regulaci), se doporučuje pracovat spíše s výběrovým rozpětím nežli s výběrovou směrodatnou odchylkou.
Zatím jsme hovořili o hustotě normálního rozdělení, nyní potřebujeme znát, jak vypadá distribuční funkce. Distribuční funkce, označme si ji *F(x)*, kde *x* je libovolné reálné číslo, je vlastně pravděpodobnost náhodného jevu, že hodnota náhodné veličiny bude menší nežli *x*, tedy

$$F(x)=\mathrm{Pst}\{ X < x\}.$$

Na Obr. 13 je vidět, jak se hodnota distribuční funkce vypočte z průběhu odpovídající hustoty. Bohužel výpočet hodnoty distribuční funkce je komplikovaný, protože pro distribuční funkci neexistuje explicitní vzorec pro stanovení její hodnoty. Je nutno buď použít tabulku nebo nějaký statistický software.



Obr. 13: Výpočet hodnoty distribuční funkce

Na Obr.13 je vidět, že hodnota distribuční funkce pro rozdělení N(5;9) pro x=3,2 se rovná 0,274. Je to velikost plochy pod hustotou teoreticky od -∞ po hodnotu 3,2.
Jak se s modelem normálního rozdělení pracuje? Představme si, že potřebujeme spočítat pravděpodobnost náhodného jevu, že hodnota sledované náhodné veličiny X se objeví v intervalu (a;b). Jde tedy o to spočítat

$$\mathrm{Pst}\{a<X<b)$$

za předpokladu, že veličina X má rozdělení normální s parametry μ a σ. Pro výpočet využijeme již výše uvedené transformace veličiny X na veličinu Z, která má rozdělení N(0,1). Zajisté platí, že

$$\text{Pst}\{a<X<b)=\text{Pst}\{\frac{a-\mu}{\sigma}<Z<\frac{b-\mu}{\sigma}\},$$

a tudíž pro výpočet potřebujeme znát hodnoty rozdělení N(0,1), a ty jsou tabelovány a dostupné v každé učebnici statistiky a nebo lze využít jakýkoliv statistický software, i v Excelu existuje funkce pro hodnoty rozdělení N(0,1). Abychom tedy tuto pravděpodobnost spočítali, potřebujeme znát pouze hodnoty distribuční funkce rozdělení N(0,1). Obvykle se tato distribuční funkce označuje v literatuře řeckým písmenem Φ. Takže výše uvedený vzorec lze doplnit následovně

$$\text{Pst}\{a<X<b)=\text{Pst}\{\frac{a-\mu}{\sigma}<Z<\frac{b-\mu}{\sigma}\}=$$

$$=\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})$$

Když bychom potřebovali znát pravděpodobnost, že hodnota normálně rozdělené veličiny bude větší nežli nějaká stanovená hodnota, např. horní specifikace, pak postupujeme tak, že nejdříve spočítáme pravděpodobnost opačného jevu, tedy, že hodnota bude pod horní specifikací, podle výše popsaného postupu, a tuto pravděpodobnost odečteme od 1. Platí totiž, že součet pravděpodobností náhodného jevu a jevu k němu opačnému, je právě 1.

Pomocí normálního rozdělení lze velice dobře aproximovat binomické rozdělení, pokud jeho parametry N a p splňují jednoduché podmínky, že $Np>10$ a současně $N(1-p)>10$. Jak provést aproximaci, ukážeme si na příkladě. Budeme 100x házet korunou a chceme vědět, jaká je pravděpodobnost toho jevu, že padne lícová strana více jak 60x. V tomto případě jsou

podmínky pro užití aproximace splněny. Dále potřebujeme znát střední hodnotu a směrodatnou odchylku pro binomickou veličinu, za předpokladu, že koruna není falešná a tedy $p = 0{,}5$. Dosazením do patřičných vzorečků dříve uvedených pro binomické rozdělení máme, že $\mu=50$ a $\sigma=5$. Označme si naši binomickou veličinu, představující počet líců při 100 hodech jako L a použijme transformaci

$$Z = \frac{L - 50}{5},$$

pro výpočet veličiny Z, která má přibližně normální rozdělení N(0,1). Stejně ale musíme přepočítat hranici 60 líců, získáme tak horní hranici pro veličinu Z, (60-50)/5=2. Takže jsme dospěli k výpočtu pravděpodobnosti náhodného jevu, že $Z>2$. Tato pravděpodobnost je v řeči distribuční funkce $1 - \Phi(2)$. S pomocí tabulek či softwaru zjistíme, že $\Phi(2)=0{,}9772$. Souhrnně, pravděpodobnost, že při 100 hodech se objeví více jak 60 lícových stran je asi 2,28%.

7. Rozdělení výběrových charakteristik a centrální limitní věta

V kapitole o popisné statistice jsme se seznámili již s některými výběrovými charakteristikami jako je výběrový průměr, výběrový medián či výběrová směrodatná odchylka. Říká se jim výběrové, protože se počítají z dat, tedy z výběru, a tak taky s nimi musíme zacházet. Jsou tedy funkcí od naměřených dat, a proto když naměřím jiná data, tak se hodnota uvažované

výběrové charakteristiky změní, což znamená, že variabilita obsažená v datech se přenese i do výběrové charakteristiky, která pak bude vykazovat svoji úroveň variability. Stejné je to i s parametrem polohy. Naměřená data se shromažďují kolem hodnoty parametru polohy a rovněž výběrová charakteristika bude mít svou hodnotu parametru polohy, kolem něhož se budou shlukovat hodnoty této naměřené charakteristiky odvozené od jednotlivých výběrů dat. Krátce řečeno, každá výběrová charakteristika je náhodnou veličinou, která má svoji vlastní distribuční funkci, jejíž tvar je dán jednak vzorcem pro výpočet charakteristiky z dat a jednak distribuční funkcí sledované náhodné veličiny, jejíž hodnoty získáváme sběrem dat. Odvození tvaru distribuční funkce pro tu kterou výběrovou charakteristiku je poměrně složitá záležitost vyžadující metody diferenciálního a integrálního počtu, mnohdy není možné dokonce odvodit explicitní vzorec pro hledanou distribuční funkci. Které charakteristiky počítat a proč, to je záležitost spadající do teoretického pozadí matematické statistiky.

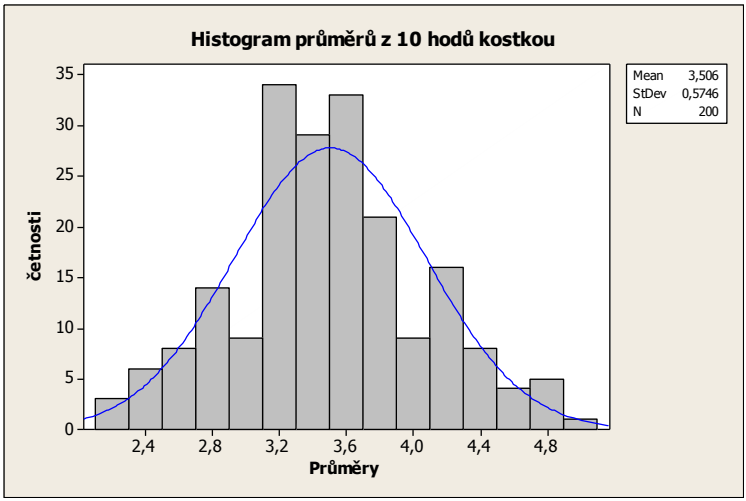
Nyní se musíme ještě zmínit o velice důležitém pojmu, bez něhož si nelze teorii pravděpodobnosti a matematickou statistiku představit a spousta věcí by se bez něho nedala vůbec odvodit. Jedná se o pojem stochastické nezávislosti. Je to vlastnost náhodných jevů odpozorovaná z reality, která má zásadní dopad na zpracování dat. Mějme dva náhodné jevy, označme si je A a B. Říkáme, že jsou stochasticky nezávislé (dále pro jednoduchost budeme říkat nezávislé), když pravděpodobnost jejich společného výskytu, tedy jejich

průniku, je rovna součinu jejich jednotlivých pravděpodobností. Tedy pomocí vzorce:

$$\mathrm{Pst}\{A \cap B\} = \mathrm{Pst}\{A\}\mathrm{Pst}\{B\}.$$

Rovněž o dvou náhodných veličinách lze hovořit o jejich nezávislosti. Mějme dvě náhodné veličiny X a Y. Říkáme, že jsou nezávislé, pokud jsou nezávislé náhodné jevy $\mathrm{Pst}\{X \leq x\}$ a $\mathrm{Pst}\{Y \leq y\}$ pro každé x a každé y.
Definice je sice celkem srozumitelná, ale jak zjistíme, jestli dva konkrétní náhodné jevy nebo dvě náhodné veličiny jsou nezávislé? Jedině zase na základě dat. Statistika má vhodné nástroje, kterými se dá otestovat hypotéza o nezávislosti jevů či veličin. Další možnost je skutečně v rámci použitého modelu dokázat výpočtem, že dva jevy či dvě náhodné veličiny jsou nezávislé. Tento postup lze např. použít, když se dokazuje nezávislost výběrového průměru a výběrové směrodatné odchylky počítaných z týchž dat za předpokladu, že data pocházejí od veličiny mající normální rozdělení.
Se závislostí je nutno např. počítat, když na daném výrobku se sleduje více znaků jakosti, které se samozřejmě mohou navzájem ovlivňovat. V praxi je celá řada případů, kdy se ví na základě zkušenosti, že nějaké veličiny jsou závislé či nezávislé. Správně by se ale tato skutečnost měla ověřit na sebraných datech, protože každá zkušenost má subjektivní charakter a člověk se může pouze domnívat, že to tak je. Více o nezávislosti, resp. o nekorelovanosti, náhodných veličin bude v kapitole věnované lineární regresi.
Nejdříve se budeme zabývat jednou z nejdůležitějších výběrových charakteristik, a to výběrovým průměrem.

Začneme příkladem. Provádíme hody kostkou, o níž budeme předpokládat, že je pravidelná, takže padnutí každé strany ze šesti možných je právě jedna šestina. Tím je dán model pro náhodnou veličinu, která představuje počet ok na té straně, která se při hodu objevila nahoře. Obor hodnot této veličiny je {1,2,3,4,5,6} a každý výsledek se realizuje s pravděpodobností 1/6. Provádíme sekvenci hodů a vždy po provedených deseti hodech spočítáme výběrový průměr, tedy aritmetický průměr z oné desítky pokusů. Průměry budeme zaznamenávat a řekněme, že jich máme 200. Z nich se sestavíme histogram a co uvidíme, je zobrazeno na Obr. 14.



Obr. 14: Histogram z průměrů deseti hodů kostkou

Když se podíváme na výše uvedený histogram ihned nás napadne, zdali by se chování průměrů z deseti hodů kostkou nedalo popsat normálním rozdělením. Podívejme se na základní výběrové charakteristiky. Výběrový průměr z dvou set průměrů je 3,508 a výběrová směrodatná odchylka je 0,5748. Nyní si spočteme, jaká je střední hodnota výchozí náhodné veličiny počítaná v rámci modelu. Zjistíme, že má hodnotu 3,5, totiž sečteme všechny hodnoty z oboru náhodné veličiny a podělíme 6. Tato hodnota se příliš neliší od aritmetického průměru 3,508. Je to náhoda? Není. Lze ukázat, že výběrový průměr má stejnou střední hodnotu jako výchozí náhodná veličina. Tato vlastnost platí obecně pro jakoukoliv náhodnou veličinu mající střední hodnotu, a proto se výběrový průměr používá často jako odhad pro střední hodnotu.
Nyní spočítáme směrodatnou odchylku výchozí veličiny počet ok na kostce. Opět použijeme vzorec pro výpočet směrodatné odchylky v modelu, čili

$$\sigma = \sqrt{\sum_{j=1}^{6} (j - 3{,}5)^2 \frac{1}{6}} \ .$$

Když toto spočítáme, tak dostaneme σ = 1,7078. My jsme průměry počítali vždy z deseti hodů kostkou, zkusme hodnotu pro σ podělit $\sqrt{10}$. Dospějeme tak k číslu 0,5401 a toto číslo porovnejme s hodnotou výběrové směrodatné odchylky pro průměry. Sice se liší více nežli aritmetický průměr od střední hodnoty, ale něco to přece jenom napovídá. Kdybychom měli ještě více průměrů k dispozici, tak by shoda byla ještě

patrnější. Opět lze obecně dokázat, že směrodatná odchylka výběrového průměru počítaného z n hodnot, má velikost směrodatné odchylky výchozí náhodné veličiny dělené $\sqrt{n}$. Lze totéž vyjádřit i v řeči rozptylů: Rozptyl výběrového průměru počítaného z n hodnot je n krát menší nežli je rozptyl výchozí náhodné veličiny.

Tato vlastnost platí pro směrodatnou odchylku výběrového průměru pouze za jednoho důležitého předpokladu. Výběr se musí skládat ze vzájemně nezávislých pozorování! Přesněji popsáno to znamená, že se na výběr o rozsahu n musíme dívat jako na posloupnost n nezávislých náhodných veličin, které mají totéž rozdělení pravděpodobnosti, tedy stejnou distribuční funkci. Tato stejnost v sobě odráží předpoklad, že během odběru dat se podmínky, za nichž sběr dat probíhá, nemění, tudíž není důvod měnit model a parametry modelu tím pádem jsou konstantní.

Nyní se vraťme k histogramu na Obr. 14. Jak uvidíme dále, nápaditá shoda histogramu z průměrů a hustotou normální rozdělení není náhodná, ale je to důsledek jednoho pilíře teorie pravděpodobnosti, tzv. Centrální limitní věty či teorému (CLT). Co tato věta tvrdí? Ať je rozdělení výchozí náhodné veličiny jakékoliv a my máme možnost neomezeně zvětšovat rozsah výběru n při zajištění nezávislosti mezi pozorováními a počítat výběrové průměry z nich, pak rozdělení pravděpodobnosti uvažovaných průměrů se přibližuje s rostoucím n normálnímu rozdělení. V případě, že rozdělení výchozí náhodné veličiny je normální $N(\mu,\sigma^2)$, pak výběrový průměr z n nezávislých pozorování má pro každé n normální rozdělení $N(\mu,\sigma^2/n)$.

Automaticky se nabízí otázka, kolik je zapotřebí pozorování, tedy jak velký by měl být rozsah výběru n, aby se při nenormálním rozdělení výchozí náhodné veličiny chování výběrových průměrů dalo již dobře popsat normálním rozdělením? Bohužel se dosti v praxi fungování CLT přeceňuje a normální rozdělení se pro popis chování průměrů automaticky použije už při malých rozsazích, n= 2,3,4…, kdy CLT ještě zdaleka nemusí platit (např. při použití regulačních diagramů pro podskupiny při aplikaci SPC). Obecně nelze dát přesnou odpověď, ta totiž silně závisí i na tvaru výchozí distribuční funkce. Lze ale říci, že na základě zkušeností statistiků, již pro $n \geq 25$ se dá očekávat dobrá shoda s normálním rozdělením.

Aproximace pomocí modelu normálního rozdělení nefunguje pouze pro výběrové průměry, ale i pro jiné veličiny, jako např. pro relativní četnosti. Představme si, že máme náhodnou veličinu popsatelnou binomickým rozdělením s parametrem p. Na základě náhodného výběru z N pokusů zjistíme hodnotu relativní četnosti výskytu sledovaného jevu, např. výskyt neshodné jednotky, tedy odhad parametru p

$$\hat{p} = \frac{m}{N},$$

kde m je počet zjištěných neshodných jednotek. Na základě CLT lze ukázat, že veličina

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{N}}}$$

má přibližně normální rozdělení N(0,1). Opět se můžeme ptát, kdy máme právo tuto aproximaci použít. Na základě zkušeností se dá hovořit o dobrém přiblížení standardnímu normálnímu rozdělení, když $Np \geq 10$ a současně i $N(1-p) \geq 10$.

K čemu se taková aproximace může hodit, ukazuje následující příklad. Na základě zkušeností z četnosti onemocnění dětí při chřipkové epidemii víme, že onemocní v průměru 25% žáků. Chceme vědět, jaká je pravděpodobnost za této situace, že procento nemocných překročí hranici 30%. Celkový počet žáků v konkrétní škole je 300. Samozřejmě, že vše by se dalo spočítat v modelu binomického rozdělení s p=0,25 a N=300. Snadnější je použít CLT a onu pravděpodobnost odhadnout pomocí normální aproximace. Ověříme, zdali má tato aproximace oprávnění. Np= 75, $N(1-p)$=225, čili bohatě jsou předpoklady pro aproximaci splněny. Nám se jedná o pravděpodobnost náhodného jevu, že relativní četnost nemocných žáků překročí hranici 0,3, čili

$$\text{Pst}\{ \hat{p} > 0{,}3\}.$$

Odhadneme pravděpodobnost opačného jevu a pak odečteme od 1.

$$\text{Pst}\{\ \hat{p} \le 0{,}3\} = \text{Pst}\{\frac{\hat{p} - 0{,}25}{\sqrt{\frac{0{,}25(1-0{,}25)}{300}}} \le \frac{0{,}3 - 0{,}25}{\sqrt{\frac{0{,}25(1-0{,}25)}{300}}}\ \}.$$

Po vypočtení dospějeme k aproximaci
Pst{$Z \le 2$} = $\Phi(2)$ = 0,97725. Tedy pravděpodobnost překročení nemocnosti nad 30% je 2,28%.
Ten závěr se dá použít ještě pro další úvahu. Předpokládejme, že procento nemocnosti se při chřipkové epidemii vyšplhá nad hranici 30%. Jestliže by platila hypotéza o očekávané úrovni 25%, pak by skutečně nastal jev mající pravděpodobnost 2,28%. Již na začátku jsme řekli, že statistika dělá závěry spojené vždy s rizikem. Zvolíme-li si v tomto případě riziko 5%, pak díky tomu, že toto riziko je větší nežli pravděpodobnost 2,28%, pak na základě tohoto faktu lze tvrdit, že pak hypotéza o 25% neplatí. Kdybychom riziko zvolili na úrovni 1%, pak by platila opačná nerovnost, a neměli bychom právo o platnosti hypotézy pochybovat. Trochu jsme těmito úvahami nakoukli do kapitoly týkající se testování statistických hypotéz.

8. Pomocná rozdělení

Tato kapitola je věnována třem rozdělením, která jsou odvozena od normálního rozdělení a jsou zvláštní tím, že tato rozdělení se obvykle nehodí jako modely pro popis chování nějaké sledované náhodné veličiny, mají totiž zcela teoretický charakter, ale statistika se bez nich nemůže obejít. Jedná se o tzv. χ^2-rozdělení (čte se chí-kvadrát), t-rozdělení a F-rozdělení.

χ^2-rozdělení.

V předchozí kapitole jsme se bavili o tom, že výběrový průměr je vlastně náhodná veličina, která má své vlastní rozdělení, které je odvozeno od výchozího rozdělení sledované náhodné veličiny. Rovněž tak výběrová směrodatná odchylka s má svoje rozdělení pravděpodobnosti a o tom si teď něco řekneme. Když vynásobíme výběrový rozptyl s^2 číslem $(n - 1)$, viz vzoreček pro výpočet s, dostaneme součet čtverců, totiž

$$(n- 1)\, s^2 = \sum_{j=1}^{n} (x_j - \bar{x})^2 ,$$

A když budeme znát parametr σ, pak obě strany rovnosti můžeme podělit číslem σ^2. Pak každý sčítanec v součtu čtverců na pravé straně je druhá mocnina náhodné veličiny, která má normální rozdělení. Tito sčítanci jsou navzájem závislí, protože ve všech se vyskytuje $\bar{x}$. Ale za předpokladu, že pozorování $x_1, x_2, \ldots\ldots, x_n$ budou nezávislá, lze ukázat, že onen součet kvadrátů se dá vyjádřit jako součet $(n - 1)$ kvadrátů nezávislých veličin, které mají rozdělení N(0,1). Tím jsme se dopracovali k rozdělení χ^2 o $(n - 1)$ stupních volnosti. Platí totiž, že náhodná veličina má rozdělení pravděpodobnosti χ^2 o k stupních volnosti, jestliže se dá vyjádřit jako součet k druhých mocnin nezávislých náhodných veličin majících rozdělení N(0,1). Stupně volnosti jsou jediným parametrem tohoto rozdělení a hodnota tohoto parametru je známa, protože se dá odvodit od počtu pozorování. Hodnoty distribuční funkce rozdělení χ^2 jsou tabelovány v každé učebnici statistiky, v každém statistickém softwaru, lze je spočítat i v Excelu.

t-rozdělení.

Toto rozdělení budeme potřebovat, až budeme testovat hypotézy o parametru polohy μ u normálního rozdělení. Z předchozího víme, že obvyklým odhadem parametru polohy je výběrový průměr. Mějme tedy dán výběrový průměr z nezávislých pozorování náhodné veličiny, která má rozdělení $N(\mu,\sigma^2)$. Z předchozí kapitoly víme, že výběrový průměr pak má rozdělení $N(\mu,\sigma^2)$. Když budeme uvažovat podíl $\frac{\bar{x}-\mu}{s}\sqrt{n}$, tak jsme vlastně použili již uvažovanou transformaci $Z=\frac{X-\mu}{o}$, která převádí normálně rozdělenou veličinu X na veličinu Z s rozdělením N(0,1). Zde za veličinu X jsme vzali výběrový průměr $\bar{x}$, a protože směrodatnou odchylku výběrového průměru $\sigma/\sqrt{n}$ neznáme, nahradili jsme ji výběrovou směrodatnou odchylkou $s/\sqrt{n}$. Nyní se využije již dříve zmíněná důležitá vlastnost normálního rozdělení, že výběrový průměr a výběrová směrodatná odchylka spočítané z nezávislých pozorování normálně rozdělené náhodné veličiny, jsou stochasticky nezávislé. Na základě tohoto faktu lze odvodit rozdělení pravděpodobnosti pro podíl $\frac{\bar{x}-\mu}{s}\sqrt{n}$ a tomuto rozdělení se říká t-rozdělení o (n-1) stupních volnosti. Počet stupňů volnosti (n–1) je dán stupni volnosti χ^2- rozdělení, které souvisí s výběrovou směrodatnou odchylkou s, jak bylo zmíněno výše. t-rozdělení se potřebuje při testování hypotéz o chování parametru polohy μ u normálního rozdělení. Distribuční funkce tohoto rozdělení je taktéž

tabelována v každé učebnici statistiky, v každém statistickém softwaru a lze ji spočítat i v Excelu.

F-rozdělení.
Toto rozdělení bylo odvozeno pro potřeby odpovědět na následující otázku. Pozorujeme náhodnou veličinu, např. znak jakosti na výrobku, a potřebujeme posoudit, zdali se po zásahu do procesu směrodatná odchylka skutečně zmenšila, jak bylo zásahem míněno. Jak to ověříme?

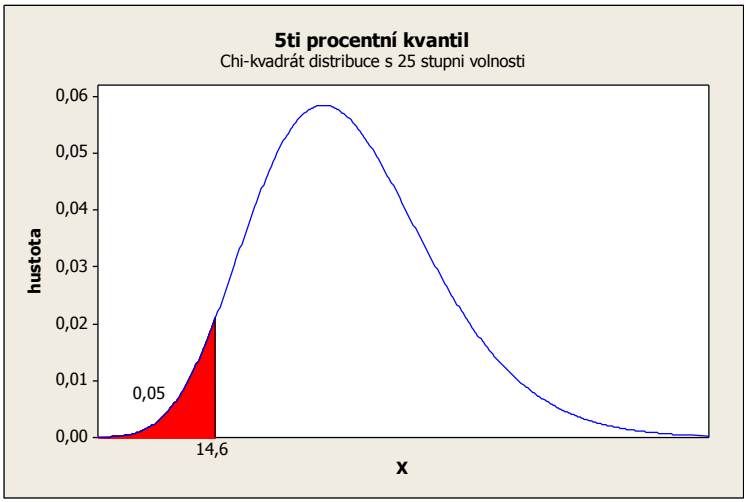
Musíme mít pozorování před opatřením a po opatření, označme si je $x_1, x_2, \ldots\ldots, x_n$ a $y_1, y_2, \ldots\ldots, y_m$. Rozsahy n, m se mohou lišit. Spočítáme výběrový rozptyl před opatřením, tedy $s^2(X)$ a po opatření $s^2(Y)$. Podíl těchto dvou náhodných veličin $\frac{s^2(X)}{s^2(Y)}$ za předpokladu, že jsou nezávislé, má pak tzv. F-rozdělení značené F(n-1, m-1) o (n-1, m-1) stupních volnosti v tomto pořadí. Nejdříve jsou stupně volnosti čitatele, a pak jmenovatele.
Stupně (n-1, m-1) jsou opět odvozeny od χ^2-rozdělení, které vystupují u rozdělení výběrových rozptylů. Opět hodnoty distribuční funkce F-rozdělení se najdou v každé učebnici statistiky, v každém statistickém softwaru a v Excelu je taktéž funkce pro výpočet F-rozdělení.
Spolu s těmito pomocnými rozděleními musíme zavést velice důležitý pojem, který ve statistice se vyskytuje velice často. Jedná se o pojem kvantilu, který lze zavést pro jakoukoliv distribuční funkci. Máme tedy nějakou náhodnou veličinu X, a jak již víme, její distribuční funkce F, je dána vzorcem

$$F(x) = \mathrm{Pst}\{X \leq x\},$$

kde x je reálné číslo. Představme si, že distribuční funkci známe a zvolme číslo α z intervalu (0,1). Hledejme takové reálné x_α , které bude splňovat rovnici

$$F(x_\alpha) = \mathrm{Pst}\{X \leq x_\alpha\} = \alpha.$$

Tato rovnice má jediné řešení x_α, které se nazývá 100α%-ní kvantil příslušné distribuční funkce. Blíže na Obr. 15, kde je zobrazen 5ti procentní kvantit od χ^2-rozdělení o 25 stupních volnosti.

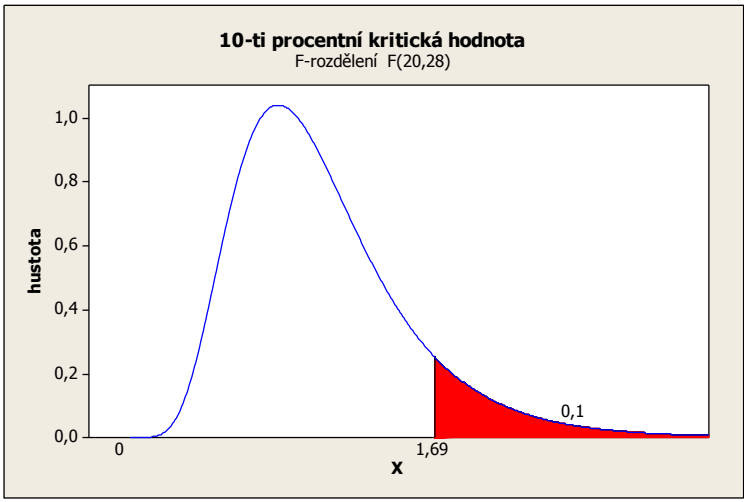


Obr.15: 5ti procentní kvantit χ^2-rozdělení o 25 stupních volnosti

Na Obr.15 je vidět, že 5ti procentním kvantilem je hodnota 14,6. To znamená, že pravděpodobnost výskytu náhodné veličiny mající χ^2-rozdělení o 25 stupních volnosti pod hodnotou 14,6 je právě 0,05.

Znalost kvantilů je, jak uvidíme dále, velice důležitá, a proto tabulky příslušných kvantilů pro normální rozdělení a pomocná rozdělení jsou k nalezení ve všech

učebnicích statistiky, statistických softwarech a rovněž pomocí patřičných funkcí v Excelu.
Obvykle současně s pojmem kvantilu se zavádí též pojem kritická hodnota, protože spolu bezprostředně souvisejí. Opět se na začátku zvolí číslo α z intervalu (0,1). Pak 100α%-ní kritická hodnota q_α není nic jiného nežli 100(1-α)%-ni kvantil, tedy $q_\alpha = x_{1-\alpha}$. Na Obr.16 je znázorněno graficky, co kritická hodnota určuje.



Obr.16: 10-ti procentní kritická hodnota F-rozdělení F(20,28)

Význam kritické hodnoty je následující. 100α%-ní kritická hodnota rozdělení pravděpodobnosti představuje hranici, nad níž se právě s pravděpodobností α vyskytne hodnota náhodné veličiny s uvažovaným typem rozdělení pravděpodobnosti. Kritické hodnoty se odvodí od příslušných hodnot kvantilů.

9. Konfidenční intervaly

Hodnotu parametru polohy či směrodatné odchylky v rámci celé základní populace neznáme a odhadujeme je z dat. Jako odhady nám slouží obvykle výběrový průměr a výběrová směrodatná odchylka. To jsou ale náhodné veličiny, čili hodnoty odhadů jsou náhodná čísla vypočítaná z dat, a tudíž se mění podle toho, jaká data jsme získali. Tedy máme sice odhad neznámého parametru, ale nevíme, jak daleko je od očekávané hodnoty parametru v rámci celé populace. Víme ale, že výběrové charakteristiky (říká se jim též statistiky) mají každá svoji střední hodnotu a směrodatnou odchylku a toto můžeme využít pro upřesnění informace o vzdálenosti mezi odhadem parametru a jeho očekávanou hodnotou. Velikost této vzdálenosti je právě vyjádřena délkou konfidenčního intervalu. Nejdříve si jeho odvození ukážeme na důležitém příkladu.

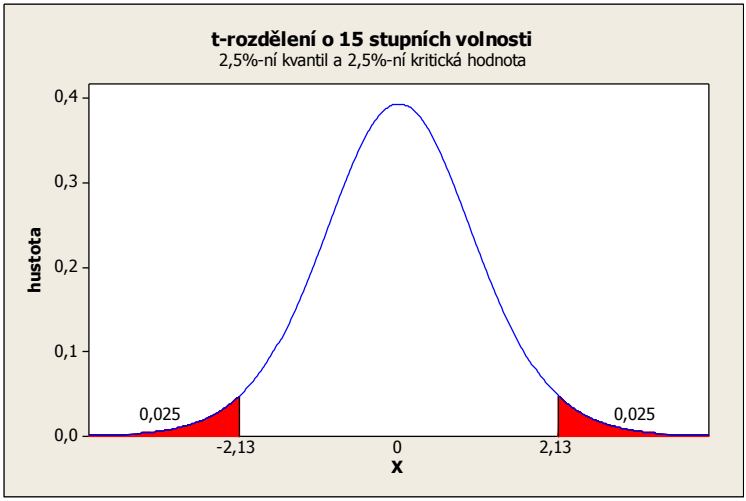
Mějme tedy náhodnou veličinu normálně rozdělenou s parametry μ a σ a odhadujeme střední hodnotu μ pomocí výběrového průměru. V kapitole o pomocných rozděleních jsme ukázali, že veličina

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n}$$

má t-rozdělení s $(n\text{-}1)$ stupni volnosti. Nyní si zvolíme úroveň rizika α, s nímž budeme pracovat. Obvykle se v praxi volí 0,05 či 0,01, ale není to pravidlo. Z t-rozdělení odvodíme jeho příslušný $100(\alpha/2)$-ní kvantil $x_{\alpha/2}$ a příslušnou $100(\alpha/2)$-ní kritickou hodnotu $q_{\alpha/2}$. Tyto hodnoty se vyznačují tím, že hodnota statistiky t se mezi nimi realizuje s pravděpodobností $1\text{-}\alpha$. Tedy pomocí vzorce

$$\text{Pst}\{\ x_{\alpha/2} \le \frac{\bar{x} - \mu}{s}\sqrt{n} \le q_{\alpha/2}\ \} = 1 - \alpha, \qquad (*)$$

jak je vidno na Obr.17.



Obr.17: Kvantil a kritická hodnota pro t-rozdělení.

Podle obrázku jsou 2,5%-ní hodnoty kvantilu a kritické hodnoty -2,13 a 2,13. To znamená, že s pravděpodobností 95% se hodnota náhodné veličiny mající t-rozdělení o 15 stupních volnosti objeví mezi -2,13 a 2,13.
Nyní upravíme ve vztahu (*) nerovnosti tak, aby se osamostatnil parametr μ. Hodnota pravděpodobnosti 1 – α se tím samozřejmě nezmění, protože jsme provedli pouze ekvivalentní úpravy v nerovnostech. Tím jsme dospěli k vyjádření

$$\text{Pst}\{\ \bar{x} - q_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} - x_{\alpha/2}\frac{s}{\sqrt{n}}\ \} = 1\text{-}\ \alpha.$$

Získali jsme tak dolní a horní hranici intervalu pro hodnoty parametru μ, který s pravděpodobností 1 – α pokrývá pro nás neznámou hodnotu tohoto parametru. Takovému intervalu se říká konfidenční interval. Číslo 1 – α se nazývá koeficient spolehlivosti. Jak je nutno rozumět této pravděpodobnosti? Představme si všechny takto zkonstruované konfidenční intervaly na základě dat. To je vlastně taky základní populace složená z těchto intervalů a 100(1-α)% těchto intervalů tu očekávanou hodnotu parametru μ pokrývá a 100α% ji nepokrývá. Takže když se realizují data, tak máme pravděpodobnost 1 – α, že vytáhneme z populace intervalů interval pokrývající správnou hodnotu μ a s opačnou nerovností interval nepokrývající tuto hodnotu.
Když se podíváme na vzorce pro výpočet mezí konfidenčního intervalu, tak vidíme, že závisí na třech veličinách. Jednak je to volba úrovně rizika α, pak na počtu pozorování a na úrovni směrodatné odchylky σ prostřednictvím jejího odhadu *s*. Někdo by si řekl, že si zvolí menší riziko volbou menší hodnoty pro α. Jak se to promítne to velikosti intervalu? Když chceš menší riziko, tak si konfidenční interval ale zvětšíš. Zmenšit jeho velikost se dá větším počtem pozorování, protože $\sqrt{n}$ vystupuje ve jmenovateli a nebo tím, že v základní populaci je menší úroveň variability a lze tedy očekávat i menší hodnotu výběrové směrodatné odchylky *s*.

Obdobným postupem lze odvodit i konfidenční interval pro rozptyl či pro směrodatnou odchylku v základní

populaci.. Z předchozího víme, že veličina $(n-1)\,s^2/\sigma^2$ má χ^2-rozdělení o n -1 stupních volnosti. Opět zvolíme koeficient spolehlivosti $1-\alpha$. Najdeme $100\alpha/2\%$-ní kvantil $x_{\alpha/2}$ a $100\alpha/2\%$-ní kritickou hodnotu $q_{\alpha/2}$. pro χ^2-rozdělení o n -1 stupních volnosti. Pak tedy platí

$$\text{Pst}\{\, x_{\alpha/2} \le (n-1)\, s^2/\sigma^2 \le q_{\alpha/2} \,\} = 1-\alpha.$$

Když provedeme podobné úpravy uvnitř závorek jako u výběrového průměru, tak získáme horní a dolní mez konfidenčního intervalu pro rozptyl, totiž

$$\text{Pst}\{\, (n-1)s^2/\, q_{\alpha/2} \le \sigma^2 \le (n-1)s^2/\, x_{\alpha/2} \,\}.$$

Konfidenční interval pro parametr σ získáme odmocněním mezí konfidenčního intervalu pro rozptyl. Pak tedy konfidenční interval pro směrodatnou odchylku σ má tvar

$$\left\langle \frac{s\sqrt{n-1}}{\sqrt{q_{\alpha/2}}}, \frac{s\sqrt{n-1}}{\sqrt{x_{\alpha/2}}} \right\rangle.$$

Vlastnosti konfidenčního intervalu pro parametr σ jsou stejné jako pro konfidenční interval parametru μ.
Je nutné připomenout, že výše odvozené konfidenční intervaly byly vypočteny za předpokladu, že výchozí náhodná veličina má normální rozdělení. Dále je dobré vědět, že pokud počet pozorování je větší nežli 30, pak kvantily t-rozdělení se dají velice dobře aproximovat kvantily normálního rozdělení N(0,1).

Dále je nutno si uvědomit, že v případě jiných rozdělení pravděpodobnosti výchozí veličiny, jako je např. log-normální či Weibullovo rozdělení, která také mají svoje parametry a které obvykle neznáme, je situace s odvozením příslušných konfidenčních intervalů pro tyto parametry daleko komplikovanější a mnohdy je nutno sáhnout pouze ke vhodné aproximaci.
Samozřejmě, že lze zkonstruovat konfidenční intervaly pro parametry binomického a Poissonova rozdělení.
V případě, že máme dostatečný počet pozorování, lze využít pro konstrukci intervalů aproximaci založenou na CLT. Pokud by počet pozorování nebyl dostatečný, tvary konfidenčních intervalů pro parametr p a parametr λ jsou komplikovanější a lze je najít v doporučené literatuře.

Z CLT víme, že náhodná veličina $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n}$ má při dostatečně velkém n přibližně normální rozdělení N(0,1).

Tato skutečnost se dá vhodně využít pro konfidenční interval parametru p na zvolené konfidenční úrovni 1-α.Pro tvar intervalu potřebujeme znát příslušné kvantily normovaného normálního rozdělení $z_{\alpha/2}$ a $z_{1-\alpha/2}$.

Díky symetrii normovaného normálního rozdělení platí, že $z_{\alpha/2} = - z_{1-\alpha/2}$.
Pak konfidenčbí interval pro parametr p má tvar:

$$\left\langle \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right\rangle.$$

Veličina $\hat{p}$ je odhad parametru p a je dán relativní četností výskytu neshodných jednotek. Využití konfidenčního intervalu si vysvětlíme na příkladě.

Představme si, že vyrábíme jednotky, u nichž se kontroluje pouze jejich funkčnost. Chceme zjistit, v jakém stavu se nachází náš výrobní proces z hlediska produkce neshodných jednotek. Z procesu odebereme např. 300 jednotek, a zjistíme, kolik jich funguje a kolik ne. Nechť počet nefunkčních jednotek je 4 kusy. Když spočítáme relativní četnost nefunkčních jednotek, máme odhad pravděpodobnosti výskytu nefunkční jednotky, tedy 4/300 = 0,0133, to je náš odhad $\hat{p}$. Zvolme úroveň spolehlivosti 95%. Potřebné kvantily od rozdělení N(0,1) jsou $z_{0,025}$ = -1,96 a $z_{0,975}$ =1,96. Po dosazení požadovaných hodnot do výše uvedeného vzorečku výsledný konfidenční interval je (0,0067, 0,0199). To znamená, že skutečná hodnota parametru p se s pravděpodobností 0,95 nachází mezi těmito dvěma mezemi. Pokud tedy by byl požadavek na výkonnost procesu na úrovni p = 0,0075, pak nemáme důvod to zpochybňovat. Pokud by byl požadavek p = 0,025, pak náš proces je dokonce lepší, pokud by se jednalo o p = 0,005, pak by náš proces toto nesplňoval.
Zcela analogicky lze využít pro posouzení stavu výrobního procesu konfidenční intervaly pro parametr polohy či pro úroveň variability v případě nějakého sledovaného znaku jakosti spojitého charakteru na výrobku. Pokud tedy konfidenční interval pro parametr polohy μ bude pokrývat jeho požadovanou hodnotu, např. prostředek specifikačního rozmezí, pak není důvod do procesu zasahovat, jestliže ale požadovaná hodnota bude mimo konfidenční interval, pak proces se zadanou úrovní konfidence nesplňuje požadavek kladený na jeho polohu. Tytéž úvahy lze provést i v případě úrovně variability. Když požadovaná hodnota pro parametr σ

bude větší nežli horní mez konfidenčního intervalu, proces resp. sledovaný znak jakosti vykazuje s velkou pravděpodobností menší úroveň variability, nežli je žádáno. Když ale požadovaná hodnota pro σ je menší nežli dolní mez konfidenčního intervalu, lze očekávat, že znak jakosti vykazuje větší variabilitu než je požadováno. V případě, že požadovaná hodnota pro σ je pokryta konfidenčním intervalem, tak nemáme důvod pochybovat o tom, že požadavek na parametr σ je splněn.

8. Základy testování statistických hypotéz

Testování hypotéz je oblast matematické statistiky, která umožňuje potvrdit či vyvrátit nějakou hypotézu o základní populaci. Hypotéza se může týkat nějakého předpokladu o chování hodnoty parametru jako je střední hodnota, medián či směrodatná odchylka, nebo se může týkat volby vhodného modelu, např. zda lze použít model normálního rozdělení či zda dva sledované znaky jakosti lze považovat za závislé či nezávislé. Hovoříme tak o parametrických nebo neparametrických testech. Statistická hypotéza má tedy charakter nějakého tvrzení o základní populaci, o jehož zamítnutí či nezamítnutí se rozhoduje na základě získaných dat pomocí statistických testů, což jsou postupy, které si statistika pro taková rozhodnutí připravila na základě teorie. Je nutné si hned na začátku připomenout, že vždy máme k dispozici pouze částečnou informaci z dat, a tedy závěry statistických testů mohou být chybné v tom smyslu, že neodpovídají skutečné situaci v základní populaci. To je způsobeno tím, že do hry vždy vstupuje náhoda. Teorie se snaží

navrhnout takové testy, u nichž by riziko chyby bylo co nejmenší.
Představme si, že sledujeme náhodnou veličinu, kterou je možno popsat binomickým rozdělením s parametrem p, který ale neznáme. Vyslovíme hypotézu, že parametr $p = 0{,}20$. Tato hypotéza se obvykle nazývá nulovou hypotézou. Proti nulové hypotéze musíme ale v každém případě vyslovit alternativní hypotézu, kterou přijmeme při zamítnutí nulové hypotézy. Nulová hypotéza a alternativní hypotéza se musí zásadně vylučovat, nemohou nastat současně. Na druhou stranu dohromady nemusí vyčerpat všechny možnosti. Nulová hypotéza tvaru $p = 0{,}20$ se nazývá jednoduchá, obdobně i alternativní hypotéza, např. že $p = 0{,}30$, ale může být i tzv. složená, např. že $p > 0{,}20$. Dosti často v literatuře se nulová hypotéza značí H_0 a alternativní hypotéza jako H_1 či A. Tedy např. H_0: $p = 0{,}20$ proti H_1: $p > 0{,}20$. Proti jednoduché hypotéze H_0: $p = 0{,}20$ tak může stát jedna ze tří možností složených alternativ. Buď tzv. oboustranná alternativa, že $p \neq 0{,}20$ nebo jedna z tzv. jednostranných alternativ $p > 0{,}20$ či $p < 0{,}20$. Volba alternativy zcela závisí na obsahu hypotézy, kterou chceme testovat a jakou odpověď potřebujeme znát. Obvykle při formulaci nulové hypotézy počítáme s tím, že nedošlo k žádné změně nebo hypotéza vyjadřuje to, co je očekáváno. Ale nemusí tomu být vždy tak, protože, jak uvidíme v dalším, zamítnutí hypotézy má vždy silnější váhu nežli její nezamítnutí, protože při nezamítnutí hypotéza nemusí platit. Obecně řečeno, věříme, že hypotéza platí, pokud nás data nedonutí ji zamítnout a přiklonit se k alternativě. Přijetí alternativy tedy znamená nalezení statisticky významného rozdílu proti nulové hypotéze.

Jaké chyby můžeme udělat při statistickém testu? Jednak se může stát, že zamítneme hypotézu, která ale platí. Této chybě se říká chyba 1.druhu. Pravděpodobnost výskytu takové chyby se nazývá riziko chyby 1. druhu a obvykle se značí α. Můžeme ale udělat i takovou chybu, že hypotéza ve skutečnosti neplatí, ale my ji nezamítneme. Hovoříme o chybě 2.druhu. Pravděpodobnost takové chyby se nazývá riziko chyby 2. druhu a značí se β. Samozřejmě bychom chtěli, aby teorie nabídla takové postupy, u kterých by bylo možno obě rizika minimalizovat. Leč, to nelze obecně udělat. Platí, že rizika α a β jsou jak spojité nádoby. Z tohoto důvodu statistika nabízí následující postup. Maximální velikost rizika α , se kterým jsme ochotni počítat, se zvolí apriori ještě před výsledkem testu. Tato hodnota se nazývá hladina významnosti testu. Teorie se pak snaží najít takové postupy, které minimalizují riziko β, neboli maximalizují tzv. sílu testu $1 - \beta$. Síla testu je tedy pravděpodobnost, s jakou já odhalím neplatnost nulové hypotézy. Bohužel ne vždy se podaří sestrojit takový test, který by měl maximální sílu ve třídě určitých testů. Proto je nutné vědět, jakou sílu má použitý test. Jeho síla jednak závisí na hladině významnosti a hlavně na počtu dat, která máme k dispozici. Čím větší počet dat, tím i větší síla testu.

Jak vypadá skladba statistického testu? Kromě zvolení hladiny významnosti každý test obsahuje testovou statistiku, což je teorií doporučená výběrová charakteristika, kterou použijeme v testu. Podle jejího chování se pak rozhodneme buď pro zamítnutí hypotézy nebo pro její nezamítnutí. Za předpokladu platnosti hypotézy se pak odvodí rozdělení pravděpodobnosti

zvolené testové statistiky a od úrovně hladiny významnosti se sestrojí pomocí tohoto rozdělení tzv. kritická oblast, do níž hodnota použité testové statistiky padne s pravděpodobností maximálně rovnou hladině významnosti α. Role kritické oblasti je ta, že pokud do ní hodnota testové statistiky padne, nulová hypotéza se zamítá ve prospěch alternativy. Síla testu je tedy pravděpodobnost, s níž padne hodnota testové statistiky do kritické oblasti za předpokladu, že alternativní hypotéza platí. Síla testu tedy měří sílu detekce změny vůči nulové hypotézy. Hranice kritické oblasti jsou obvykle určeny na základě kvantilů rozdělení testové statistiky za předpokladu platnosti nulové hypotézy. Výsledek testu je tedy dán tím, jestli hodnota testové statistiky padla do kritické oblasti či nikoliv. Když není v kritické oblasti, nemáme důvod nulovou hypotézu zamítnout, ale musíme si být vědomi i toho, že nemusí platit. Prostě nemáme dost dat na to, abychom nulovou hypotézu zamítli. V tom je zamítnutí nulové hypotézy silnější tvrzení nežli je její nezamítnutí. Pokud samozřejmě při nezamítnutí nulové hypotézy jiná možnost není, např. je-li nulová hypotéza jednoduchá, pak nulovou hypotézu přijímáme.

V konkrétním případě se může stát, že při volbě hladiny významnosti např. na úrovni 5% nulovou hypotézu zamítneme, ale kdyby hladina byla zvolena na úrovni 1%, tak nulová hypotéza testem projde. Bylo by zcela nesprávné volit výsledek testu až a posteriori a hýbat s hladinou významnosti podle výsledku testu a ho tím ovlivnit. Svědčí to ale o tom, že výsledek testu je tak říkaje „ na hraně“.

Prakticky v každém statistickém softwaru je výsledek testu zveřejněn ve formě tzv. *p-hodnoty*. Co toto číslo vyjadřuje? Toto číslo vyjadřuje pravděpodobnost, s jakou se může hodnota testové statistiky objevit nad (resp. pod, resp. nad i pod) vypočtenou hodnotou testové statistiky z konkrétních dat za platnosti nulové hypotézy. Možnosti „nad“, „pod“ či „nad i pod“ závisejí na tvaru kritické oblasti. Zní to složitě, ale interpretace je velmi jednoduchá. Pokud *p-hodnota* je menší nežli zvolená hladina významnosti, pak nulovou hypotézu zamítáme, pokud není menší, pak není důvod hypotézu zamítat. S *p-hodnotou* pracuje i Excel ve svých testech.
Nyní si probereme nejdůležitější testy pro binomické, Poissonovo a normální rozdělení.

Testy o parametru p binomického rozdělení.
Pro jednoduchost vzorečků budeme předpokládat, že lze použít CLT pro aproximaci chování relativní četnosti $\hat{p}$. Nechť nulová hypotéza zní $p=p_0$. Nejdříve budeme uvažovat oboustrannou alternativu A: $p \neq p_0$. Když bude platit hypotéza, pak podle CLT podíl $\dfrac{\hat{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$ má přibližně rozdělení N(0,1), kde n je počet pozorování. Zvolme hladinu významnosti α. Pak tedy

$$\mathrm{Pst}\{\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < z_{\alpha/2}\} = \alpha/2$$

a současně také

Pst{$z_{1-\alpha/2} < \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_{0)}}{n}}}$} = $\alpha/2$. Tím jsme zkonstruovali kritickou oblast $K = (-\infty, z_{\alpha/2}) \cup (z_{1-\alpha/2}, +\infty)$. Pokud tedy testová statistika

$$T(\hat{p}) = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_{0)}}{n}}}$$

padne do oblasti K, nulovou hypotézu zamítáme.

Zcela podobně lze testovat hypotézu o shodnosti parametru p v případě dvou náhodných výběrů ze dvou základních populací popsatelných binomickým rozdělením. Nulová hypotéza má tedy tvar H_0: $p_1=p_2$ proti oboustranné alternativní hypotéze A: $p_1 \neq p_2$. V případě platnosti nulové hypotézy má testová statistika

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

přibližně rozdělení N(0,1), kde $\hat{p}$ je relativní četnost výskytu sledovaného jevu počítaná z obou náhodných výběrů. Rozsah 1.výběru je n_1, 2. výběru pak n_2. Tato testová statistika je odvozena opět pomocí CLT a platí, pokud lze tuto aproximaci při dostatečných rozsazích výběrů použít. Jestliže platí nulová hypotéza, pak testová statistika má rozdělení aproximovatelné rozdělením N(0,1). Zvolíme tedy hladinu významnosti α a najdeme příslušné kvantily $z_{\alpha/2}$ a $z_{1-\alpha/2}$. Pokud hodnota testové statistiky se objeví mezi těmito kvantily, pak nulovou hypotézu nezamítáme.

Podobně bychom mohli uvažovat jednostranné hypotézy, jak je ukázáno použití testu v následujícím příkladě.

Pro ověření možných vedlejších účinků léku byl uspořádán experiment. První skupina o 374 osobách užívala lék, přičemž 26 si stěžovalo na vedlejší příznaky, druhá skupina o 210 osobách dostávala placebo a stěžovalo si 8 lidí. Nulová hypotéza zní, že není statisticky významného rozdílu mezi oběma skupinami, tedy H_0: $p_1=p_2$, zatímco alternativní hypotéza zde má jednostranný tvar A: $p_1 > p_2$. Zvolme hladinu významnosti na úrovni 5%. Spočítáme hodnotu testové statistiky na základě $\hat{p}_1 = 26/374 = 0{,}07$ a $\hat{p}_2 = 8/210 = 0{,}04$. Celková relativní četnost $\hat{p} = (26+8)/(210+374)=0{,}03$. Dosadíme do vzorečku pro testovou statistiku a získáme hodnotu 1,5. Tuto hodnotu můžeme porovnat s příslušným kvantilem rozdělení N(0,1) (zde 95%-ní) a nebo vypočítat odpovídající *p-hodnotu.* Ta je rovna pravděpodobnosti, s jakou se může veličina s rozdělením N(0,1) vyskytnout právě nad hodnotou 1,5. Tato pravděpodobnost je 6,68%, a protože je větší nežli hladina významnosti, nemáme důvod nulovou hypotézu zamítat. Lze tedy tvrdit, že experiment neprokázal statisticky významný rozdíl ve výskytu vedlejších příznaků při hladině významnosti 5%.

Nyní se probereme dva nejdůležitější testy týkající se parametrů normálního rozdělení.

F-test.

Pomocí tohoto testu porovnáváme úrovně variability u dvou základních populací s normálními rozděleními $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$. Nulová hypotéza je tedy H_0: $\sigma_1 = \sigma_2$, alternativa je oboustranná A: $\sigma_1 \neq \sigma_2$. Máme tedy dva soubory dat, první soubor je o rozsahu n_1, druhý soubor

má rozsah n_2. Vypočteme z dat příslušné výběrové rozptyly a testová statistika má tvar jejich podílu, tedy

$$F = s_1^2 / s_2^2 ,$$

Za předpokladu rovnosti rozptylů statistika F má F-rozdělení o $n_1 - 1, n_2 - 1$ stupních volnosti, značíme $F(n_1 - 1, n_2 - 1)$. Na pořadí stupňů volnosti záleží, nejdříve jsou stupně volnosti čitatele, pak jmenovatele. Pokud statistika bude mít hodnotu buď malou či zase naopak velkou, lze spíše očekávat zamítnutí nulové hypotézy. Tomu odpovídá i kritická oblast testu. Zvolme hladinu významnosti α. Najdeme odpovídající kvantily F-rozdělení $f_{\alpha/2}$ a $f_{1-\alpha/2}$. Jestliže hodnota statistiky F bude menší nežli kvantil $f_{\alpha/2}$ nebo větší nežli kvantil $f_{1-\alpha/2}$., pak nulovou hypotézu zamítáme na hladině významnosti α.
Tímto způsobem lze tedy testovat shodnost směrodatných odchylek ve dvou základních souborech, které lze popsat modelem normálního rozdělení.
V případě, že základní soubor nelze popsat normálním rozdělením, tak statistika má k dispozici testy, které jsou avšak založeny na jiných testových statistikách.
Když jsme otestovali rozptyly, můžeme přistoupit k testování shody parametrů polohy μ_1 a μ_2.

Dvouvýběrový t-test.
Nulová hypotéza t-testu má tvar H_0: $\mu_1 = \mu_2$. Alternativa je oboustranná, neboť se nám jedná buď o shodu parametrů polohy nebo o jejich rozdílnost. Samozřejmě by bylo možné uvažovat i jednostranné alternativy podle povahy problému, který bychom řešili.

Dvouvýběrový t-test užívá různé testové statistiky podle výsledku F-testu o rovnosti rozptylů, který má předcházet před provedením t-testu. Jestliže F-test nezamítne rovnost rozptylů, pak společný rozptyl pro obě populace je odhadován najednou z obou náhodných výběrů, a tím se zvětší informace o základních populacích. Jestliže ale F-test rovnost rozptylů zamítne, je nutno každý rozptyl odhadovat pouze z dat příslušného náhodného výběru. Při rovnosti rozptylů se používá testová statistika ve tvaru $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s^2}{n_1}+\frac{s^2}{n_2}}}$,

kde s^2 je odhad společného rozptylu $s^2=\frac{(n_1-1)s_x^2+(n_2-1)s_y^2}{n_1+n_2-2}$, kde s_x^2 a s_y^2 jsou výběrové rozptyly z jednotlivých výběrů. Při zvolené hladině významnosti α najdeme příslušné kvantily $t_{\alpha/2}(n_1+n_2-2)$ a $t_{1-\alpha/2}(n_1+n_2-2)$, protože za platnosti nulové hypotézy testová statistika má t-rozdělení o n_1+n_2-2 stupních volnosti. Bude-li testová statistika v absolutní hodnotě větší nežli kvantil $t_{1-\alpha/2}(n_1+n_2-2)$, pak máme právo nulovou hypotézu zamítnout.

Pokud nám F-test zamítne rovnost rozptylů, pak testová statistika má tvar $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n_1}+\frac{s_y^2}{n_2}}}$, kde s_x^2 a s_y^2 jsou příslušné výběrové rozptyly a za platnosti nulové hypotézy se rozdělení testové statistiky aproximuje t-rozdělením o

stupních volnosti, které vypočtou jako nejbližší celé číslo k podílu $$\frac{(s_x^2/n_1 + s_y^2/n_2)^2}{(s_x^2/n_1)^2/(n_1-1) + (s_y^2/n_2)/(n_2-1)}.$$

Rozhodovací kritérium je pak stejné jako u případu rovnosti rozptylů založené na příslušných kvantilech t-rozdělení.

Existuje ještě speciální verze t-testu zvaná párový t-test, který vyžaduje sběr dat uspořádaný do párů. Co je tím míněno, je nejlépe vysvětlit na příkladě. V kravíně máme 50 krav a máme údaje o jejich dojivosti před a po použití nových krmných směsí. Potřebujeme zjistit, zdali se použitím nových směsí dojivost zvýšila či nikoliv. Tedy pro každou krávu máme dvojici údajů, jednak před a jednak po novém krmení. Pozorování jsou uspořádána v párech, proto párový test.

Jak t-test , tak F-test jsou dostupné v Excelu mezi statistickými funkcemi, výsledek testů je dán prostřednictvím *p-hodnoty*.

Použití párového testu si ukážeme na následujícím příkladu. Na 10 jednotkách byla provedena měření před a po provedení doporučeného opatření. Zde jsou měření:

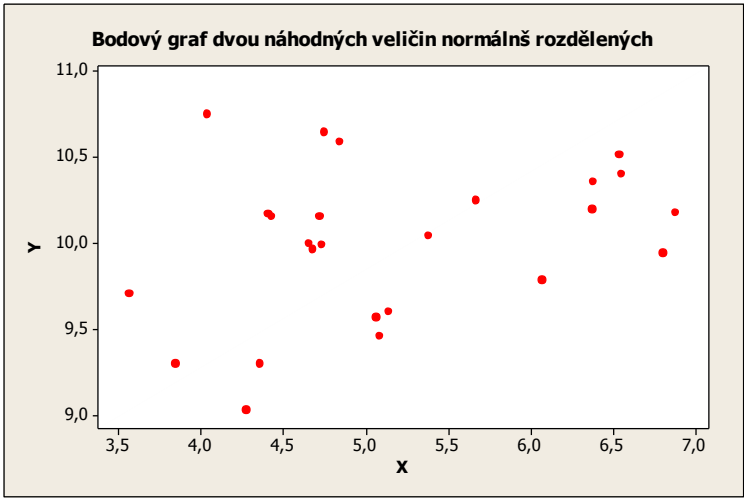
Před	***Po***	***Rozdíl***
85	80	5
80	80	0
95	88	7
87	90	-3
78	72	6
82	79	3
57	50	7
69	73	-4
73	78	-5
98	95	3

Nulová hypotéza H_0: $\mu_1=\mu_2$ se testuje proti jednostranné alternativě A: $\mu_1 \neq \mu_2$ na hladině významnosti 5%. Pokud bude nulová hypotéza platit, tak výběrový průměr od veličiny ***Rozdíl*** podělený svou výběrovou směrodatnou odchylkou a vynásobený $\sqrt{10}$ by měl mít t-rozdělení o 9ti stupních volnosti. Výběrová směrodatná odchylka veličiny ***Rozdíl*** má hodnotu 4,64. Pak hodnota testové statistiky je $\frac{2}{4,64}\sqrt{10} =$ 1,36. Protože alternativa je zvětšení střední hodnoty, hypotézu budeme zamítat tím více, čím bude hodnota testové statistiky větší. Kritická oblast je tedy nad příslušným kvantilem odvozeným od hladiny významnosti.Hodnotu testové statistiky porovnáme s 95%-ním kvantilem t-rozdělení o 9ti stupních volnosti, který je 1,83. Protože tato hodnota je větší nežli hodnota testové statistiky, pak nulovou hypotézu nemáme právo zamítnout.

11. Korelace a regrese

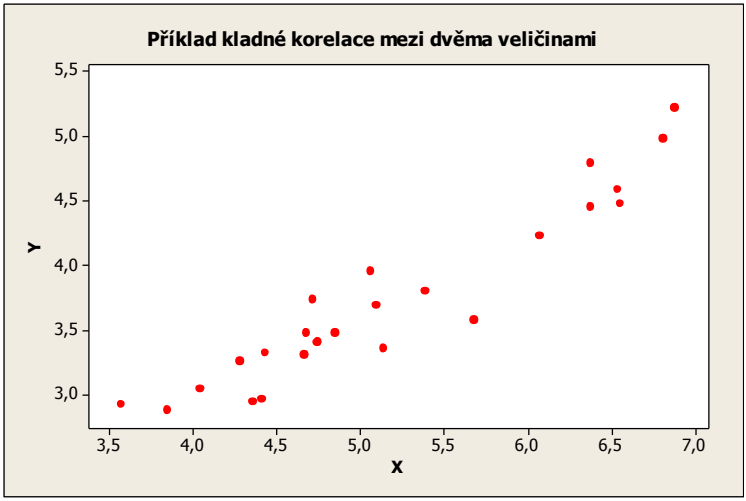
Obvykle na výrobku nebývá pouze jeden znak jakosti, ale několik najednou, které se mohou různě ovlivňovat. Jak je silná jejich vazba, na to statistika používá nástroj zvaný regresní analýza. V této kapitole se budeme zabývat pouze nejjednodušším případem, a to je lineární regrese. Budeme studovat vazbu mezi dvěma náhodnými veličinami *X* a *Y,* veličina *Y* je spojitého charakteru. Lze též uvažovat regresi pro atributivní znaky, ale ten případ je komplikovanější a zde zmíněn nebude. Mnohdy lze tento problém vyřešit tak, že atributivní znak, třeba počet neshod, převedeme do řeči procent, a ta již mají spojitý charakter. Abychom mohli vyšetřovat regresi mezi dvěma veličinami, potřebujeme mít data uspořádaná do

dvojic. Ve dvojici jeden údaj se týká veličiny X , druhý veličiny Y, tedy máme k dispozici obecně n dvojic (x_i,y_i), $i = 1,2,\ldots\ldots,n$. V prvém kroku je doporučeno si udělat grafickou analýzu dat ve formě tzv. bodového grafu, kde na vodorovnou osu je nanášena veličina X, a na druhou osu veličina Y. Toto jednoduché grafické zobrazení již ledacos může napovědět o vzájemném vztahu mezi oběma veličinami. Především jestli jsou tzv. korelované nebo nikoliv. Na jedné straně jsou veličiny, které jsou stochasticky nezávislé, ty jsou samozřejmě nekorelované, na druhé straně jsou veličiny funkčně závislé, kdy každé hodnotě veličiny X odpovídá pouze jediná hodnota veličiny Y. Na Obr. 18 jsou znázorněny hodnoty dvou veličin, které jsou stochasticky nezávislé.

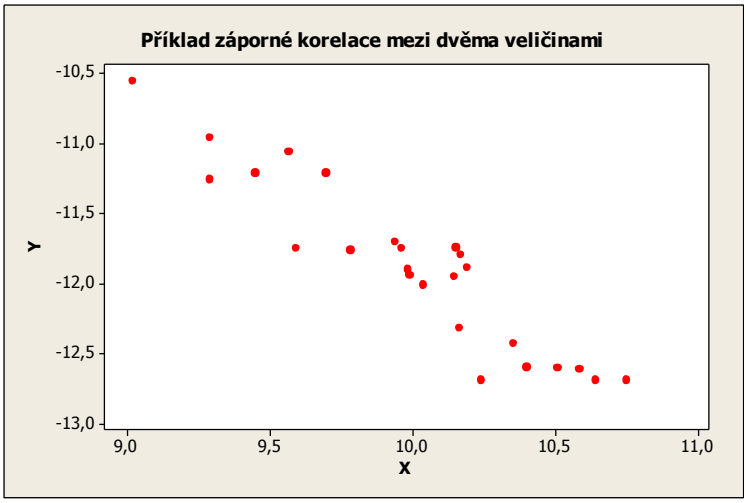


Obr.18: Bodový graf

Z grafu na Obr.18 je ihned patrno, že obě veličiny jsou nejspíše nekorelované, protože zakreslené dvojice hodnot veličin zaplňují zcela chaoticky plochu obrázku. Tato hypotéza by samozřejmě musela být potvrzena či vyvrácena statistickou analýzou.
Na Obr.19 je naopak ihned vidět, že obě veličiny se nejspíše ovlivňují a s růstem veličiny *X* je spojen i růst veličiny *Y*. V tomto případě říkáme, že mezi oběma veličinami existuje kladná korelace.



Obr.19: Příklad kladné korelace
Opačně lze najít případ dvou veličin, kdy s růstem jedné je spojen pokles druhé. Tato situace je uvedena na Obr.20. Pak hovoříme o záporné korelaci mezi veličinami.



Obr. 20: Příklad záporné korelace

Podstatou regresní analýzy je snaha najít vhodný model, který by vysvětloval vztah mezi sledovanými náhodnými veličinami. Tímto modelem se snažíme pochopit, jak silná je vazba mezi veličinami a v dalším ho použít pro další výpočty jako je např. odhadnout veličinu Y i pro jiné hodnoty veličiny X, než pro které byla měření provedena. Nejjednodušším příkladem takového modelu je model lineární regrese, kde se snažíme popsat vztah mezi veličinami pomocí rovnice

$$Y = aX + b + E\,,$$

kde a a b jsou nějaké konstanty, parametry modelu, veličina X je tzv. vysvětlující veličina, veličina Y je vysvětlovaná veličina a náhodná veličina E představuje náhodné chyby, o nichž se obvykle předpokládá, že mají normální rozdělení $N(0,\sigma^2)$, kde rozptyl σ^2 bývá neznámý. Konstanty a,b rovněž nebývají známy a je nutno je z naměřených dat odhadnout. Vhodnost modelu

se ověřuje na základě tzv. reziduí, což jsou odchylky naměřených hodnot veličiny Y od hodnot vypočtených dosazením odpovídajících hodnot veličiny X do odhadnutého modelu. Rezidua jsou vlastně odhady náhodných chyb.

Pokud lze vztah mezi veličinami X a Y vysvětlit pomocí modelu lineární regrese, míra korelace mezi X a Y se posuzuje tzv. výběrovým koeficientem korelace $\hat{r}$, který se vypočte podle následujícího vzorce

$$\hat{r} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y},$$

kde s_x, s_y jsou výběrové směrodatné odchylky. Hodnoty výběrového koeficientu korelace se vyskytují vždy mezi -1 a +1. Čím je koeficient $\hat{r}$ v absolutní hodnotě blíže k +1, tím je větší míra korelace mezi X a Y. Bude-li $\hat{r} = \pm 1$, pak existuje dokonce funkční lineární vztah mezi X a Y. Naopak čím bude v absolutní hodnotě koeficient $\hat{r}$ menší a blíže k nule, tím bude i menší korelace mezi X a Y. Znamení koeficientu $\hat{r}$ vyjadřuje i povahu korelace. Když je hodnota $\hat{r}$ kladná, lze usuzovat na kladnou korelaci mezi veličinami, pokud bude hodnota koeficientu $\hat{r}$ záporná, svědčí to o záporné korelaci mezi X a Y. Výběrový koeficient korelace $\hat{r}$ slouží jako základ pro testovou statistiku při testování nulové hypotézy o nekorelovanosti mezi veličinami proti obecné alternativě, že veličiny jsou vzájemně korelovány. Čím bude absolutní hodnota $\hat{r}$ menší, tím spíše nebudeme nulovou hypotézu zamítat, čím ale bude větší, tím spíše se budeme přiklánět k platnosti alternativní hypotézy.

Je nutno upozornit, že formální použití výběrového koeficientu korelace bez ověření lineárního modelu pro

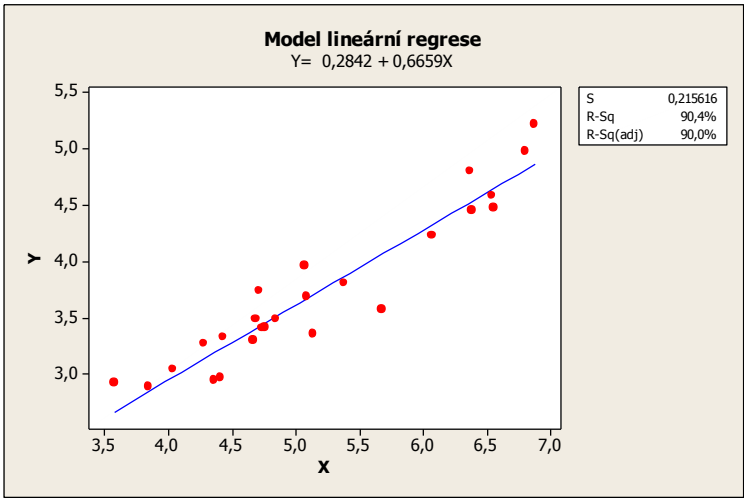
uvažované veličiny může vést k mylnému závěru. Snadno lze najít taková data, která evidentně nesplňují model lineární regrese a přesto vykazují malou absolutní hodnotu koeficientu $\hat{r}$. Z tohoto důvodu je žádoucí sestrojit alespoň bodový graf před výpočtem koeficientu $\hat{r}$ a testováním úrovně korelace.

Na Obr.21 je provedena regresní analýza pomocí lineárního modelu. Posouzení shody modelu s naměřenými daty je provedeno ve statistice často používanou tzv. metodou nejmenších čtverců, kdy je daty proložena taková přímka, která dává minimální součet kvadrátů odpovídajících reziduí. Vhodnost modelu je jednak posuzována podle chování reziduí, jednak hodnotou R-Sq, která v % vyjadřuje, jaký podíl variability je možno vysvětlit modelem. V případě uvedeném na Obr.21 se jedná o 90%, což značí velice dobrou shodu modelu s daty.

Metoda nejmenších čtverců dává odhad směrnice proložené přímky $\hat{a}$=0,6659 a odhad pro absolutní člen $\hat{b}$=0,2842. Pokud bychom testovali nulovou hypotézu H_0: $a = 0$ proti alternativě A: $a \neq 0$, pak bychom museli nulovou hypotézu na zvolené hladině významnosti 5% zamítnout. Lze tedy souhrnně říci, že v tomto případě je model lineární regrese vhodným modelem pro chování dat veličiny *X* a veličiny *Y*. Výběrový koeficient korelace má hodnotu 0,9509, což značí velkou míru korelovanosti mezi veličinami.

Je ale velice nutné zdůraznit jeden důležitý fakt. I když jsme objevili vhodný regresní model, nikterak to nemusí znamenat, že mezi oběma veličinami existuje příčinný vztah ve skutečnosti. To samozřejmě statistika odhalit neumí. Z modelu lze tedy pouze vyčíst, že se změnou

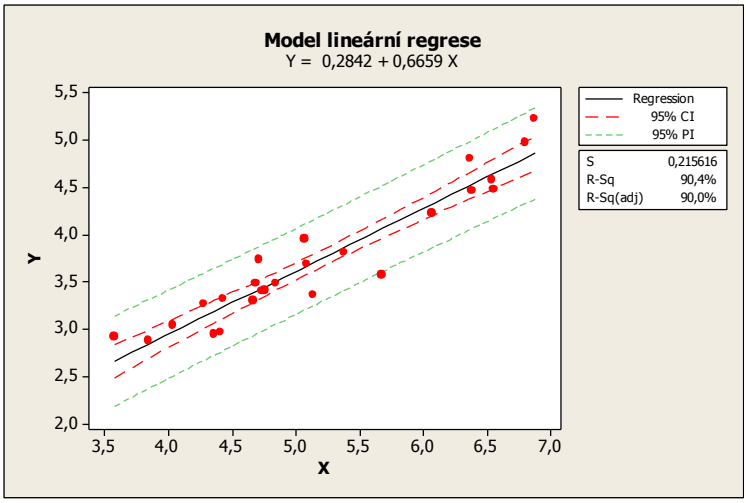
veličiny *X* lze s velkou pravděpodobností očekávat lineární změnu ve veličině *Y*, ale důvod těchto změn může být vyvolána zcela jinou veličinou, která do modelu vůbec zahrnuta není. Např. toto veličinou může být teplota, která jasně působí kladně na spotřebu piva, ale rovněž způsobuje větší spotřebu ochranných krémů proti spálení pokožky a kdybychom si vynesli graf týdenní spotřeby piva a krémů od jara do konce léta, s velkou pravděpodobností bychom z grafu tušili kladnou korelaci, z níž by těžko někdo vyvozoval, že spotřeba piva vyvolává i větší spotřebu krémů na opalování.



Obr.21: Lineární regrese – vhodný model

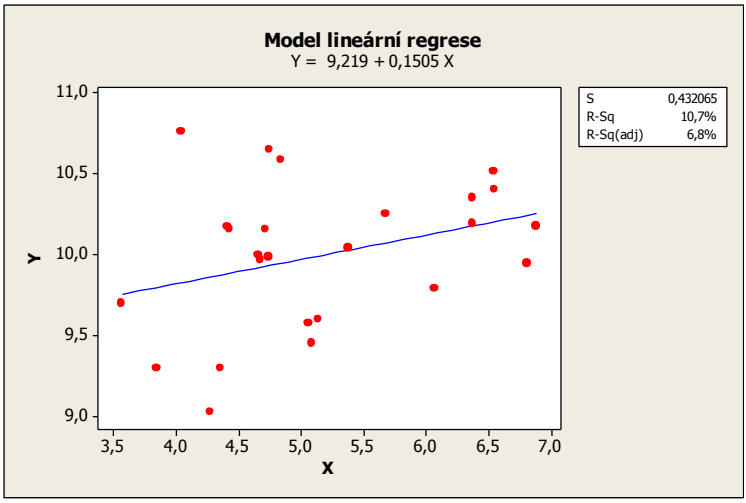
Toto nebezpečí mylné interpretace je velké hlavně v těch případech, když jsou data pouze pozorována či sbírána a nejsou získána v podmínkách plánovaného experimentu, kdy vliv známých doprovodných veličin je znám a pokud možno eliminován.

Tak jako bodový odhad např. parametru polohy sám neříká nic o vzdálenosti od skutečné hodnoty parametru polohy, ale tuto informaci nám dá teprve příslušný konfidenční interval, tak podobně i proložená přímka daty nic neříká, jak je „daleko“ od skutečné přímky vyjadřující vztah mezi *X* a *Y*. Tuto informaci nám poskytnou konfidenční meze mající hyperbolický průběh a proložená přímka je jejich osou symetrie, které jsou sestrojeny s předem zvolenou úrovní spolehlivosti. Na Obr. 22 jsou těmi užšími a říkají, že jakákoliv přímka ležící celá svými hodnotami mezi těmito mezemi může být tou skutečnou přímkou vyjadřující správnou relaci mezi veličinami *X* a *Y*. Jinými slovy řečeno, oblast mezi konfidenčními mezemi je oblastí, která s danou spolehlivostí pokrývá očekávané, tedy střední hodnoty veličiny *Y*. Jestliže bychom chtěli ale vědět, kde s předem zadanou pravděpodobností se může realizovat hodnota náhodné veličiny *Y* v závislosti na vybrané hodnotě veličiny *X*, pak musíme zkonstruovat příslušné tzv. predikčmí meze, které vymezují oblast, v níž se s onou zvolenou pravděpodobností může hodnota od veličiny *Y* objevit. Tyto meze jsou širší nežli konfidenční meze, jak je vidět na Obr. 22.



Obr. 22: Konfidenční a predikční meze lineární regrese

Jak je vidět z následujícího obrázku Obr.23, tak sice lineární model byl nalezen, ale vysvětluje pouze necelých 11% variability, zbytek je „schován“ v reziduích, což by při vhodném modelu mělo být přesně naopak. Statistický software vždycky něco vypočte, ale je nutno výsledek ověřit a správně interpretovat. V tomto případě při testování nulové hypotézy o směrnici *a* dojdeme totiž k závěru, že hypotézu nelze zamítnout, tedy ji musíme respektovat a ta nám říká, že data nejsou korelována, i když by se zdálo, že hodnota výběrového koeficientu korelace je dostatečně velká, je totiž 0,3276.
Na tomto obrázku je tedy uveden případ, kdy nelze hovořit o vhodnosti lineárního modelu, i když by se zdálo, že metoda nejmenších čtverců proložila daty vhodnou přímku.



Obr. 23: Lineární regrese – nevhodný model

Další varování se týká tzv. predikce, tedy předpovídání chování veličiny *Y* tím způsobem, že využijeme nalezený model a lidově řečeno jeho platnost protáhneme na další hodnoty veličiny *X*, jednoduše tak, že nové hodnoty veličiny *X* dosadíme do nalezeného modelu a získáme tak nové hodnoty pro *Y*. Jednak je nutno si vždy uvědomit, že do měření vstupuje náhoda a dále, podmínky, za nichž měření probíhala, se mohou změnit a protažení modelu zdaleka nemusí již platit pro jiné hodnoty veličiny *X*.

12.Nejčastější chyby

Největší chybou bývá špatná příprava sběru dat a vlastní sběr. Je nutné si uvědomit, že sebe lepší statistické metody nezlepší informaci, která je v datech obsažena. Když již na začátku bude informace chudá, tak bude chudý i výsledek statistické analýzy a nebude konzistentní se skutečností. Především je nutné nezapomenout na zapisování všech změn, které se odehrály během sběru dat, pokud jejich výskytu neumíme zabránit a předem odstranit, protože tato vedlejší informace nám výrazně pomůže pochopit chování dat a rovněž správně naformulovat statistické hypotézy. Dále je nutné připomenout, že závěry jakékoliv statistické analýzy nejsou nikdy stoprocentně pravdivé, existuje vždy riziko, že data řekla něco jiného, nežli skutečně je. Ale statistika je tak férová, že si je jednak vždy vědoma tohoto rizika a jeho míru si nastaví vždy předem.
Při grafickém zpracování dat je nutné dbát na měřítka, především na svislé ose, abychom např. při srovnávání dvou histogramů neměli u každého jiné měřítko. To se obecně týká jakýchkoliv grafů, softwary většinou umožňují zadat stejné měřítko u několika grafů najednou. U histogramu je nutné dodržovat doporučený počet intervalů podle počtu dat, rozhodně by počet intervalů neměl klesnout pod 6-7. Při použití histogramu pro identifikaci vhodného modelu, je žádoucí, aby sloupce histogramu byly vyjádřeny v relativních četnostech. Dalším častým problémem bývá počet dat. Jestliže není problémem data nasbírat, čím více dat ,tím samozřejmě

lépe. Když ale data se získávají obtížně, je jich málo, tak sice lze statistickou analýzu provést, protože software vždy něco vypočte, ale výsledek je nutno brát velice opatrně a pouze jako informativní.
Je nutné neustále si opakovat, že správné použití statistiky vyžaduje ověření předpokladů, které jsou nutné pro použití toho či jiného statistického nástroje. Často se v praxi např. ignoruje ověření normality sledované náhodné veličiny a výsledek se bere zcela vážně.
Při interpretaci závěrů založených na testování statistických hypotéz se často zcela automaticky přijímá platnost nulové hypotézy, pokud není přijata alternativa, a tím nulová hypotéza zamítnuta. Tato interpretace je možná pouze tehdy, když nulová hypotéza a její alternativa dohromady vyčerpávají všechny možnosti týkající se testovaného parametru, např. nulová hypotéza je, že parametr polohy je roven 5 proti alternativě, že je různý od 5. Zde jiná možnost pro hodnoty parametru polohy není. Něco jiného bude tehdy, když nulová hypotéza bude opět 5, ale alternativa bude pouze 6. Pokud budeme testovat takovou nulovou hypotézu proti takové alternativě a na základě výsledku testu hypotézu nezamítneme, zdaleka nemusí být pravda, že nulová hypotéza musí platit. My nemáme v tomto případě dostatek informace pro její zamítnutí, takže ji zamítnout nemůžeme, ale ona platit nemusí. Třeba parametr polohy je ve skutečnosti 5,1. K tomu, abychom rozlišili hodnotu 5 parametru polohy od hodnoty 5,1, bychom potřebovali více dat nežli máme k dispozici.

13. Závěr a literatura

Čtenář, který dospěl až k této poslední kapitole, si snad odnesl něco, co mu pomohlo pochopit, jak statistika funguje a jaké jsou její základní principy. Matematická statistika patří do matematiky, a tedy mezi přírodní vědy, přesto má v jednom výjimečné postavení. Slouží jako nástroj pro zpracování dat pro ostatní přírodní vědy, ale nejenom pro ně, a ačkoliv její teoretické zázemí je postaveno na deduktivních základech, přesto při vlastním zpracování dat využívá induktivní postup v tom, že zevšeobecňuje z náhodného výběru na celou základní populaci.

V textu byly zmíněny pouze některé z důležitých částí matematické statistiky, v seznamu literatury pak jsou uvedeny další prameny, kde je možno se důkladněji seznámit s dalšími nástroji. Napsání publikace bylo vedeno snahou lidsky přiblížit lidem, kteří potřebují data zpracovávat a většinou na to nejsou dostatečně připraveni, že řada postupů a nástrojů statistiky se dá pochopit tzv. selským rozumem a že není třeba se statistiky obávat.

Seznam použité a doporučené literatury:

[1] Anděl J.:*Matematická statistika*, SNTL/Alfa Praha 1978
[2] Dupač V., Hájek J.: *Pravděpodobnost ve vědě a technice*, Nakldatelství ČSAV, Praha 1962
[3] Rumsey D.: *Statistics Essentials for Dummies*, Wiley Publishing, Indianopolis, 2010
[4] Fabian F., Horálek V., Král J., Křepela J.,Michálek J.: *Využití podpory Microsoft Excel při aplikaci základních statistických metod*, Vydavatelství ČSJ, Praha 2008
[5] Michálek J. a kolektiv: *Statistické metody řízení jakosti*, Vydavatelství ČSJ, Praha 2007
[6] Drdla M., Karpíšek Z.: *Statistické metody,* VUT Brno, 1999

Použitý software Minitab 16, akademická licence

Název: Základy statistického myšlení
Autor: RNDr. Jiří Michálek, CSc
Vydání: 1. vydání
Počet stran: 95
Formát: A5
Tisk:

Vydala: Česká společnost pro jakost,
Novotného lávka 5, 116 68 Praha 1
www.csq.cz

ISBN: