Think Stats: Pravděpodobnost a statistika pro programátory

Verze 1.6.0

Think Stats

Pravděpodobnost a statistika pro programátory

Verze 1.6.0

Allen B. Downey

Green Tea Press

Needham, Massachusetts

Green Tea Press
9 Washburn Ave
Needham MA 02492

Originální formou této knihy je zdrojový kód LaTeX. Kompilací tohoto kódu vzniká zobrazení učebnice, které je nezávislé na konkrétním zařízení a může být převedeno do jiných formátů a vytištěno.

Zdrojový kód LaTeX pro tuto knihu je dostupný na `http://thinkstats.com`.

K vytvoření obálky této knihy byla použita fotografie Paula Friela (`http://flickr.com/people/frielp/`), který ji poskytl v souladu s podmínkami licence Creative Commons Attribution. Originání fotografie je k dispozici na `http://flickr.com/photos/frielp/11999738/`.

Předmluva

Proč jsem napsal tuto knihu

Think Stats: Pravděpodobnost a statistika pro programátory je učebnice pro nový typ kurzu poskytující úvod do pravděpodobnosti a statistiky. Důraz je kladen na využití statistických metod při práci s velkými soubory dat. Východiskem je výpočetní přístup, který má hned několik předností:

- Psaní programů slouží studentům jako nástroj, jak rozvíjet a testovat porozumění probírané látce. Píší například funkce pro výpočet metody nejmenších čtverců, reziduí a determinačního koeficientu. Psaní a testování tohoto kódu se neobejde bez porozumění příslušným konceptům a implicitně také koriguje případné nepochopení.

- Studenti provádějí experimenty, jejichž cílem je otestovat statistické chování. Například prozkoumávají centrální limitní větu tím, že generují vzorky z různých rozdělení. Ve chvíli, kdy vidí, že součet hodnot z Paretova rozdělení nekonverguje k normálnímu rozdělení, si uvědomí předpoklady, na nichž je centrální limitní věta založena.

- Některé myšlenky, které je obtížné uchopit matematicky, je snadné pochopit na základě simulace. Provádíme například aproximaci p-hodnot pomocí simulací Monte Carlo, čímž narůstá význam p-hodnoty.

- Díky použití spojitých rozdělení a výpočtů je možné představit také témata jako například bayesovský odhad, která nejsou běžně součástí úvodních kurzů. V jednom cvičení jsou studenti například požádáni, aby vypočítali aposteriorní rozdělení pro „problém německého tanku“, což je složité v rámci analytického přístupu, ale překvapivě jednoduché, použijeme-li výpočetní přístup.

- Vzhledem k tomu, že studenti pracují v univerzálním programovacím jazyku (Python), jsou schopni importovat data téměř z jakéhokoliv zdroje. Nemusí se omezit pouze na data, která byla očištěna a naformátována pro konkrétní statistický nástroj.

Kniha je vhodná pro projektový přístup. V mém kurzu pracují studenti na semestrálním projektu, v rámci kterého si mají položit statistickou otázku, najít soubor dat, který jim na ni může dát odpověď, a aplikovat každou z probíraných technik na jejich vlastní data.

Jako ukázka typu analýzy, jaký od svých studentů očekávám, slouží případová studie, která se prolíná celou knihou. Tato případová studie využívá data ze dvou zdrojů:

- Národní šetření růstu rodin (National Survey of Family Growth – NSFG), prováděné Americkými centry pro kontrolu a prevenci nemocí (U. S. Centers for Disease Control and Prevention – CDC), jehož cílem je shromáždit „informace o rodinném životě, sňatcích a rozvodech, těhotenstvích, neplodnosti, užívání antikoncepce a zdraví mužů a žen“. (Viz `http://cdc.gov/nchs/nsfg.htm`.)
- Systém sledování rizikových faktorů chování (Behavioral Risk Factor Surveillance System – BRFSS), prováděný Národním centrem pro prevenci chronických onemocnění a podporu zdraví (National Center for Chronic Disease Prevention and Health Promotion) za účelem „sledování zdravotních podmínek a rizikového chování ve Spojených státech“. (Viz `http://cdc.gov/BRFSS/`.)

V ostatních příkladech jsou využívána data zpřístupněná Daňovou správou USA (IRS), Americkým úřadem pro sčítání lidu a Bostonským maratonem.

Jak jsem napsal tuto knihu

Když někdo píše novou učebnici, obvykle začne tím, že přečte stohy starých učebnic. Ve výsledku pak většina učebnic obsahuje stejný materiál v prakticky stejném pořadí. Často se vyskytují fráze a chyby, které se šíří od jedné knihy k další. Stephen Jay Gould upozornil na jeden příklad ve své eseji: „The Case of the Creeping Fox Terrier (Případ plíživého foxteriéra)[1].“

Já jsem takto nepostupoval. Vlastně jsem v průběhu psaní této knihy nepoužil téměř žádné tištěné materiály, a to z několika důvodů:

[1] Psí plemeno zhruba poloviční velikosti Hyracotheria (viz `http://wikipedia.org/wiki/Hyracotherium`).

- Mým cílem bylo prozkoumat nový přístup k tomuto materiálu, a tak jsem nechtěl být příliš vystaven existujícím přístupům.
- Protože tuto knihu zpřístupňuji pod volnou licencí, chtěl jsem si být jistý tím, že žádná její část nebude zatížena autorskoprávními omezeními.
- Mnozí čtenáři mých knih nemají přístup ke knihovnám s tištěnými materiály, a tak jsem se snažil odkazovat na zdroje, které jsou volně dostupné na internetu.
- Zastánci starých médií si myslí, že výlučné využívání elektronických zdrojů je znakem lenosti a nespolehlivosti. Možná mají pravdu, pokud jde o to první, ale myslím si, že se mýlí v tom druhém bodě, a tak jsem chtěl otestovat svoji teorii.

Zdroj, který jsem využíval víc než kterýkoliv jiný, je Wikipedie, postrach knihovníků všude na světě. Obecně mohu říci, že články, které jsem si přečetl o statistických tématech, byly velmi dobré (i když jsem v průběhu psaní provedl několik drobných změn). Odkazy na stránky na Wikipedii uvádím na mnoha místech své knihy a doporučuji vám se na tyto odkazy podívat. V řadě případů uvedená stránka na Wikipedii pokračuje tam, kde jsem se svým výkladem skončil. Termíny a způsob zápisu používané v této knize jsou obecně konzistentní s Wikipedií, až na případy, kdy jsem měl dobrý důvod pro odchýlení se.

Další užitečné zdroje, na které jsem narazil, jsou Wolfram MathWorld a (samozřejmě) Google. Také jsem použil dvě knihy, dílo Davida MacKaye *Information Theory, Inference, and Learning Algorithms*, což je kniha, která mě přivedla k bayesovské statistice, a dále dílo Press et al. *Numerical Recipes in C*. Obě knihy jsou ale dostupné také online, a tak jsem bez obav.

Allen B. Downey
Needham MA

Allen B. Downey je profesorem počítačové vědy na Franklin W. Olin College of Engineering.

Seznam přispěvatelů

Máte-li nějakou připomínku nebo návrh na opravu, kontaktujte mě prosím prostřednictvím e-mailu `downey@allendowney.com`. Jestliže na základě vaší

zpětné vazby provedu nějakou změnu, přidám vaše jméno na seznam přispěvatelů (pokud mě nepožádáte, abych vaše jméno neuváděl).

Když ve zprávě uvedete alespoň část věty, ve které se chyba objevuje, usnadníte mi tím hledání. Číslo strany a oddílu jsou také dobré, ale nepracuje se s nimi tak snadno. Díky!

- Lisa Downey a June Downey si přečetly počáteční verzi a provedly řadu oprav a poskytly mi spoustu připomínek.
- Steven Zhang objevil několik chyb.
- Andy Pethan a Molly Farison mi pomohli odladit některá řešení a Molly si všimla několika překlepů.
- Andrew Heine našel chybu v mé chybové funkci.
- Dr. Nikolas Akerblom ví, jak velké je Hyracotherium.
- Alex Morrow vyjasnil jeden z příkladů kódu.
- Jonathan Street zachytil chybu právě včas.
- Gábor Lipták našel překlep v knize a řešení štafetového zápasu.
- Velké díky patří Kevinu Smithovi a Timu Arnoldovi za jejich práci na plasTeXu, který jsem použil ke konverzi této knihy na DocBook.
- George Caplan mi poslal několik návrhů pro lepší srozumitelnost.
- Julian Ceipek našel chybu a několik překlepů.
- Stijn Debrouwere, Leo Marihart III, Jonathan Hammler a Kent Johnson našli chyby v prvním tištěném vydání.
- Dan Kearney našel překlep.
- Jeff Pickhardt našel nefunkční odkaz a překlep.
- Jörg Beyer našel v knize překlepy a provedl řadu oprav v dokumentačních řetězcích (docstrings) doprovodného kódu.
- Tommie Gannert poslal opravný soubor s řadou oprav.
- Alexander Gryzlov navrhl objasnění v jednom cvičení.
- Martin Veillette mi nahlásil chybu v jednom ze vzorců pro Pearsonovu korelaci.
- Christoph Lendenmann mě upozornil na několik tiskových chyb.

Obsah

Předmluva **v**

1 Statistické myšlení pro programátory **1**

1.1 Rodí se prvorozené děti se zpožděním? 2

1.2 Statistický přístup 3

1.3 Národní šetření růstu rodin 4

1.4 Tabulky a záznamy 6

1.5 Významnost 9

1.6 Glosář 10

2 Popisná statistika **13**

2.1 Střední hodnoty a průměry 13

2.2 Rozptyl 14

2.3 Rozdělení 15

2.4 Zobrazování histogramů 16

2.5 Grafické znázornění histogramů 17

2.6 Zobrazování pravděpodobnostních funkcí (PMFs) 19

2.7 Grafické znázornění pravděpodobnostních funkcí (PMFs) . . 21

2.8 Odlehlé hodnoty 22

2.9 Další vizualizace 22

2.10 Relativní riziko 23
2.11 Podmíněná pravděpodobnost 24
2.12 Referování o výsledcích 25
2.13 Glosář 25

3 Distribuční funkce **27**
3.1 Paradox počtu studentů v kurzu 27
3.2 Limity pravděpodobnostních funkcí (PMFs) 29
3.3 Percentily 30
3.4 Distribuční funkce 32
3.5 Zobrazení distribučních funkcí (CDFs) 33
3.6 Zpátky k datům z šetření 34
3.7 Podmíněná rozdělení 35
3.8 Náhodná čísla 36
3.9 Souhrnné statistické charakteristiky podruhé 37
3.10 Glosář 37

4 Spojitá rozdělení **39**
4.1 Exponenciální rozdělení 39
4.2 Paretovo rozdělení 42
4.3 Normální rozdělení 44
4.4 Normální pravděpodobnostní graf 47
4.5 Logaritmicko-normální rozdělení 49
4.6 Proč model? 51
4.7 Generování náhodných čísel 52
4.8 Glosář 52

5 Pravděpodobnost **55**

5.1 Pravidla pravděpodobnosti 57

5.2 Monty Hall 58

5.3 Poincaré 60

5.4 Další pravidlo pravděpodobnosti 61

5.5 Binomické rozdělení 62

5.6 Série a exponovaná místa 63

5.7 Bayesova věta 65

5.8 Glosář 68

6 Operace s rozděleními **71**

6.1 Šikmost 71

6.2 Náhodné proměnné 73

6.3 Hustota pravděpodobnosti (PDFs) 74

6.4 Konvoluce 76

6.5 Proč normální? 78

6.6 Centrální limitní věta 79

6.7 Struktura rozdělení 80

6.8 Glosář 81

7 Testování hypotéz **83**

7.1 Testování rozdílu v průměrech 84

7.2 Výběr hladiny významnosti 86

7.3 Definování zjištění 87

7.4 Interpretace výsledku 88

7.5 Křížová validace 89

7.6 Vyjadřování bayesovských pravděpodobností 90

7.7 Chí-kvadrát test 91
7.8 Účinný resampling 93
7.9 Síla 94
7.10 Glosář 95

8 Odhadování **97**

8.1 Hra na odhad 97
8.2 Odhadněte rozptyl 99
8.3 Pochopení chyb 99
8.4 Exponenciální rozdělení 100
8.5 Intervaly spolehlivosti 101
8.6 Bayesovský odhad 101
8.7 Implementace bayesovských odhadů 103
8.8 Cenzurovaná data 105
8.9 Problém s lokomotivou 106
8.10 Glosář 109

9 Korelace **111**

9.1 Standardní skóre 111
9.2 Kovariance 112
9.3 Korelace 113
9.4 Vytváření bodových grafů v pyplot 114
9.5 Spearmanova pořadová korelace 118
9.6 Metoda nejmenších čtverců 119
9.7 Dobrá shoda 122
9.8 Korelace a kauzalita 123
9.9 Glosář 126

Kapitola 1

Statistické myšlení pro programátory

Tato kniha se zabývá tím, jak přetavit data ve znalosti. Data jsou levná (tedy alespoň relativně). Znalosti se získávají obtížněji.

Představím zde tři oblasti, které spolu souvisejí:

Pravděpodobnost zkoumá náhodné jevy. Většina lidí intuitivně rozumí tomu, co jsou stupně pravděpodobnosti, a proto můžeme používat slova jako "pravděpodobně"a "nepravděpodobný", aniž bychom prošli speciální přípravou. My však budeme mluvit o tom, jak tyto stupně kvantitativně vyjádřit.

Statistika je obor, který na základě vzorků dat vyvozuje určitá tvrzení o populacích. Většina statistických analýz vychází z pravděpodobnosti, a tak jsou tyto dvě oblasti obvykle prezentovány společně.

Výpočty jsou vhodným nástrojem pro kvantitativní analýzu, ke zpracování statistických údajů se přitom často využívají počítače. Výpočetní experimenty jsou také užitečné při zkoumání konceptů v oblasti pravděpodobnosti a statistiky.

Tato kniha vychází z teze, že pokud umíte programovat, můžete této dovednosti využít k pochopení pravděpodobnosti a statistiky. Tato témata jsou často prezentována z pohledu matematiky a tento postup některým lidem vyhovuje. S některými důležitými myšlenkami v této oblasti je obtížné pracovat matematicky, ale je poměrně snadné uchopit je výpočetně.

Zbytek této kapitoly je věnován případové studii motivované otázkou, kterou jsem zaslechl, když jsme s mojí ženou čekali naše první dítě, a sice: Rodí se prvorozené děti se zpožděním?

1.1 Rodí se prvorozené děti se zpožděním?

Když zadáte tento dotaz do Googlu, najdete spoustu diskusí na toto téma. Někteří lidé tvrdí, že je to pravda, jiní zase, že je to mýtus, a někteří tvrdí, že je to naopak: že se prvorozené děti rodí předčasně.

V těchto diskusích se také lidé snaží podepřít svoje tvrzení daty. Našel jsem spoustu takovýchto příkladů:

> „Dvě mé kamarádky, které nedávno porodily své první dítě, OBĚ téměř 2 týdny přenášely, než u nich začaly porodní bolesti nebo jim je doktoři vyvolali.“
>
> „Moje první dítě se narodilo o 2 týdny později. Teď to ale vypadá, že se druhé narodí o dva týdny před termínem!!“
>
> „Nemyslím si, že by to mohla být pravda, protože moje sestra, která je prvorozená, se mé matce narodila předčasně a stejně tomu bylo i u mnoha mých bratranců a sestřenic. “

Takovéto zprávy se označují jako **anekdotické důkazy**, protože se zakládají na údajích, které jsou nepublikované a obvykle mají osobní charakter. Na takovýchto historkách není nic špatného, pokud se objevují v běžné konverzaci, a tak se nechci do lidí, které jsem tady citoval, nijak navážet.

My bychom ale mohli chtít přesvědčivější důkazy a spolehlivější odpovědi. Tomuto standardu anekdotické důkazy většinou nedostojí, a to z následujících důvodů:

Malý počet pozorování: Jestliže těhotenství trvá v případě prvorozených dětí déle, rozdíl bude pravděpodobně malý v porovnání s přirozenou variabilitou. V takovém případě bychom zřejmě museli porovnat velký počet těhotenství, abychom si mohli být jisti, že takový rozdíl opravdu existuje.

Selektivní zkreslení: Lidé, kteří se zapojují do diskuse o této otázce, by mohli být zainteresovaní, protože jejich první dítě se narodilo se zpožděním. V takovém případě by výsledky byly zkresleny procesem výběru dat.

Konfirmační zkreslení: U lidí, kteří tomuto tvrzení věří, bychom mohli očekávat větší pravděpodobnost, že přispějí nějakými potvrzujícími příklady. U lidí, kteří o tomto tvrzení pochybují, zase existuje větší pravděpodobnost, že budou uvádět protichůdné příklady.

Nepřesnost: Historky jsou často osobní příběhy a jako takové si je lidé často nepřesně pamatují, nepřesně vyprávějí a opakují atd.

Takže jak bychom na to mohli jít lépe?

1.2 Statistický přístup

Jako způsob, jak se vypořádat s omezeními historek, použijeme statistické nástroje, mimo jiné:

Sběr dat: Použijeme data z velkého národního šetření, které bylo navrženo speciálně s cílem dospět ke statisticky validním závěrům ohledně americké populace.

Popisná statistika: Vygenerujeme statistické charakteristiky, které výstižně shrnou data, a vyhodnotíme různé způsoby vizualizace dat.

Explorační analýza dat: Budeme hledat vzory, rozdíly a další znaky, které nám pomohou najít odpovědi na otázky, jež nás zajímají. Zároveň se budeme mít na pozoru před nekonzistentností a budeme si uvědomovat omezení.

Testování hypotéz: V případě pozorovaných zjištění, jako například rozdílu mezi dvěma skupinami, vyhodnotíme, zda je takové zjištění skutečné, nebo jestli k němu mohlo dojít náhodně.

Odhad: Použijeme data z výběrového souboru k odhadu vlastností obecné populace.

Provedeme-li tyto kroky pečlivě, tak abychom se vyhnuli různým úskalím, můžeme dospět k závěrům, které jsou mnohem lépe podložené a u kterých existuje větší pravděpodobnost, že jsou správné.

1.3 Národní šetření růstu rodin

Od roku 1973 provádějí Americká centra pro kontrolu a prevenci nemocí (U.S. Centers for Disease Control and Prevention – CDC) Národní šetření růstu rodin (National Survey of Family Growth – NSFG), jehož cílem je shromáždit „informace o rodinném životě, sňatcích a rozvodech, těhotenstvích, neplodnosti, užívání antikoncepce a zdraví mužů a žen. Výsledky tohoto šetření se využívají... k plánování zdravotnických služeb a programů zdravotního vzdělávání a také k provádění statistických studií rodin, plodnosti a zdraví.“[1]

Data shromážděná v rámci tohoto šetření použijeme k prozkoumání otázky, zda se prvorozené děti opravdu rodí se zpožděním, a dalších otázek. Abychom byli schopni používat tato data efektivně, je nezbytné rozumět designu uvedené studie.

Šetření NSFG představuje **průřezovou** studii, což znamená, že zachycuje stav určité skupiny v konkrétním okamžiku. Nejběžnější alternativou je **longitudinální** studie, která se věnuje pozorování jisté skupiny opakovaně po určitou dobu.

Šetření NSFG bylo provedeno sedmkrát; jednotlivé realizace tohoto šetření se nazývají **cykly**. My budeme používat data z Cyklu 6, který probíhal od ledna 2002 do března 2003.

Cílem tohoto šetření je vyvodit závěry ohledně konkrétní **populace**; v případě šetření NSFG jsou cílovou populací lidé žijící ve Spojených státech ve věku 15–44 let.

Lidé účastnící se šetření se nazývají **respondenti**; skupina respondentů, která má společný charakteristický znak (věk, pohlaví, vzdělání atp.), se označuje jako **kohorta**. Obecně se dá říci, že průřezové studie by měly být **reprezentativní**, což znamená, že každý člen cílové populace má stejnou šanci se zúčastnit. Tento ideál je pochopitelně v praxi obtížně realizovatelný, ale lidé provádějící šetření se mu snaží maximálně přiblížit.

Šetření NSFG není reprezentativní. Určité skupiny jsou v něm záměrně nadreprezentovány (**oversampled**). Při přípravě studie byly tři skupiny – Hispánci, Afroameričané a teenageři – zastoupeny více, než jaké je jejich zastoupení v populaci USA. Důvodem pro takovéto nadreprezentování (oversampling) bylo zajistit, aby počet respondentů z každé z těchto skupin byl dostatečně velký pro vyvození validních statistických závěrů.

[1] Viz `http://cdc.gov/nchs/nsfg.htm`.

Nevýhodou nadreprezentování pochopitelně je, že na základě statistických údajů zjištěných v rámci šetření již není tak snadné činit závěry o obecné populaci. K tomuto se ještě vrátíme později.

Cvičení 1.1 Přestože bylo šetření NSFG provedeno sedmkrát, nejedná se o longitudinální studii. Na toto téma si přečtěte stránky na Wikipedii `http://wikipedia.org/wiki/Cross-sectional_study` a `http://wikipedia.org/wiki/Longitudinal_study`, abyste se ujistili, že rozumíte, proč tomu tak je.

Cvičení 1.2 V tomto cvičení si stáhnete data z NSFG. Tato data budeme využívat v celé knize.

1. Běžte na `http://thinkstats.com/nsfg.html`. Přečtěte si podmínky užívání těchto dat a klikněte na „Souhlasím s těmito podmínkami“ (za předpokladu, že souhlasíte).

2. Stáhněte si soubory s názvem `2002FemResp.dat.gz` a `2002FemPreg.dat.gz`. První soubor je věnovaný respondentům a obsahuje jeden řádek po každou ze 7 643 respondentek. Druhý soubor obsahuje jeden řádek pro každé těhotenství, o kterém respondentka poskytla údaje.

3. Online documentace šetření je k dispozici na `http://www.icpsr.umich.edu/nsfg6`. Projděte si nabídku v levém navigačním panelu, abyste si vytvořili představu o tom, jaká data jsou zahrnuta. Můžete si také přečíst dotazníky na `http://cdc.gov/nchs/data/nsfg/nsfg_2002_questionnaires.htm`.

4. Webová stránka věnovaná této knize nabízí kód ke zpracování datových souborů z šetření NSFG. Stáhněte si `http://thinkstats.com/survey.py` a spusťte jej ve stejném adresáři, do kterého jste uložili datové soubory. Měl by přečíst datové soubory a zobrazit počet řádků v každém z nich:

```
Number of respondents 7643
Number of pregnancies 13593
```

5. Projděte si kód, ať zjistíte, co umí. V následující části se budeme zabývat tím, jak to funguje.

1.4 Tabulky a záznamy

Básník a filozof Steve Martin jednou řekl:

> „Oeuf“ znamená vejce, „chapeau“ znamená klobouk. Vypadá to, že ti Francouzi mají jiné slovo úplně pro všechno.

Jako Francouzi, také databázoví programátoři mluví trochu jiným jazykem, a protože my také pracujeme s databází, potřebujeme se naučit nějaká slovíčka.

Každý řádek v souboru věnovaném respondentům obsahuje údaje o jednom respondentovi. Tyto údaje se nazývají **záznam**. Proměnné, které tvoří záznam, se označují jako **pole**. Soubor záznamů se nazývá **tabulka**.

Jestliže si přečtete `survey.py`, narazíte na definice tříd pro `Record`, což je objekt reprezentující záznam, a `Table`, který reprezentuje tabulku.

Existují dvě podtřídy `Record` – `Respondent` a `Pregnancy` – a ty obsahují záznamy z tabulek o respondentech a těhotenstvích. Prozatím jsou tyto třídy prázdné. Především nemáme žádnou inicializační metodu (init method) pro inicializaci jejich atributů. Namísto toho použijeme`Table.MakeRecord` ke konverzi řádku textu na objekt `Record`.

Máme také dvě podtřídy `Table`: `Respondents` a `Pregnancies`. Inicializační metoda (init method) v každé třídě uvádí implicitní název datového souboru a typ záznamu, který má být vytvořen. Každý objekt `Table` má atribut s názvem `records`, což je seznam objeků `Record`.

Metoda `GetFields` vrátí pro každý `Table` seznam entic, které specifikují pole ze záznamu, která budou uložena jako atributy v každém objektu `Record`. (Možná by bylo dobré si poslední větu přečíst dvakrát.)

Například zde je `Pregnancies.GetFields`:

```
    def GetFields(self):
        return [
            ('caseid', 1, 12, int),
            ('prglength', 275, 276, int),
            ('outcome', 277, 277, int),
            ('birthord', 278, 279, int),
            ('finalwgt', 423, 440, float),
            ]
```

První entice říká, že pole `caseid` je ve sloupcích 1 až 12 a že se jedná o celé číslo. Každá entice obsahuje následující údaje:

pole (field): Název atributu, kde bude uloženo pole. Většinou používám název uvedený v číselníku NSFG převedený na samá malá písmena.

začátek (start): Index počátečního sloupce pro toto pole. Například index začátku pro `caseid` je 1. Tyto indexy si můžete vyhledat v číselníku NSFG na `http://nsfg.icpsr.umich.edu/cocoon/WebDocs/NSFG/public/index.htm`.

konec (end): Index konečného sloupce pro toto pole. Například index konce pro `caseid` je 12. Na rozdíl od Pythonu je zde index konce uváděn *včetně*.

konverzní funkce (conversion function): Funkce, která přijme řetězec a převede jej na vhodný typ. Můžete použít vestavěné funkce, jako například `int` a `float`, nebo funkce definované uživateli. Pokud se konverze nezdaří, obdrží atribut hodnotu řetězce `'NA'`. Pokud nechcete provést konverzi pole, můžete zadat identitní funkci nebo použít `str`.

Pro záznamy o těhotenstvích extrahujeme následující proměnné:

caseid je celočíselná identifikace respondenta.

prglength je celočíselná délka těhotenství vyjádřená v týdnech.

outcome je celočíselný kód vyjadřující výsledek těhotenství. Kód 1 znamená porod živého dítěte.

birthord je celé číslo vyjadřující pořadí porodu pro každý porod živého dítěte. Například kód pro prvorozené dítě je 1. V případě jiného výsledku než živě narozené dítě je pole prázdné.

finalwgt je statistická váha spojená s respondentem. Jedná se o hodnotu s pohyblivou řádovou čárkou, která vyjadřuje počet lidí v americké populaci, který daný respondent zastupuje. Členové nadměrně zastoupených (oversampled) skupin mají nižší váhu.

Pokud si pozorně pročtete knihu obsahující dokumentaci případů (casebook), zjistíte, že většina těchto proměnných je **rekódována (recodes)**, což znamená, že nejsou součástí **surových dat** sebraných během šetření, ale že jsou vypočteny na základě surových dat.

Například `prglength` pro porody živých dětí se rovná původní proměnné `wksgest` (počet týdnů těhotenství), pokud je tento údaj dostupný. V ostatních případech je odhadnut pomocí `mosgest * 4.33` (počet měsíců těhotenství vynásobený průměrným počtem týdnů v měsíci).

Rekódování často vychází z logiky, která hlídá konzistentnost a přesnost dat. Obecně platí, že rekódované proměnné je dobré použít, pokud neexistuje nějaký pádný důvod pro to, abyste surová data zpracovali sami.

Můžete si také všimnout toho, že `Pregnancies` disponuje metodou `Recode`, která provádí dodatečnou kontrolu a rekódování.

Cvičení 1.3 V tomto cvičení napíšete program, který prozkoumá data v tabulce Pregnancies.

1. V adresáři, kam jste si uložili `survey.py` a datové soubory, vytvořte soubor s názvem `first.py` a napište nebo vložte následující kód:

   ```
   import survey
   table = survey.Pregnancies()
   table.ReadRecords()
   print 'Number of pregnancies', len(table.records)
   ```

   Výsledek by měl být 13593 těhotenství.

2. Napište smyčku, která zopakuje `table` a spočítá počet živě narozených dětí. Najděte dokumentaci `outcome` a ujistěte se, že váš výsledek je v souladu se shrnutím v dokumentaci.

3. Upravte smyčku tak, abyste rozdělili porody živých dětí do dvou skupin, jednu pro prvorozené děti a druhou pro ostatní. Opět si přečtěte dokumentaci `birthord`, abyste se ujistili, že jsou vaše výsledky konzistentní.

   Když pracujete s novým souborem dat, je taková kontrola užitečná pro identifikaci případných chyb a nekonzistentností v datech, odhalení chyb ve vašem programu a pro kontrolu, že rozumíte tomu, jak jsou tato pole kódována.

4. Vypočtěte průměrnou délku těhotenství (v týdnech) pro prvorozené děti a ostatní. Zjistili jste mezi těmito dvěma skupinami nějaký rozdíl? Jak velký?

Řešení tohoto cvičení si můžete stáhnout na `http://thinkstats.com/first.py`.

1.5 Významnost

V předchozím cvičení jste porovnali délku těhotenství pro prvorozené děti a ostatní. Pokud vše fungovalo, jak mělo, zjistili jste, že prvorozené děti se, v průměru, rodí přibližně o 13 hodin později.

Takovýto rozdíl označujeme jako **pozorované zjištění**; tedy jako situaci, kdy se může něco dít, ale nejsme si tím zatím jistí. Stále nám ještě zbývá položit si několik otázek:

- Jestliže mají obě skupiny odlišné průměry, co ostatní **souhrnné statistické charakteristiky**, jako například medián a rozptyl? Můžeme rozdíl mezi oběma skupinami vyjádřit s větší přesností?
- Je možné, že rozdíl, který jsme pozorovali, by mohl vzniknout náhodně, i kdyby byly skupiny, které jsme porovnávali, ve skutečnosti stejné? Pokud ano, vyvodili bychom z toho závěr, že zjištění nebylo **statisticky významné**.
- Může být pozorované zjištění výsledkem selektivního zkreslení nebo jiné chyby v nastavení experimentu? Pokud ano, pak bychom mohli dospět k závěru, že zjištění je ve skutečnosti **artefakt**, tedy něco, co jsme (náhodně) vytvořili, spíše než zjistili.

Zodpovězení těchto otázek nám zabere většinu zbývajícího prostoru této knihy.

Cvičení 1.4 Nejlepší způsob, jak se seznámit se statistikou, je pracovat na nějakém projektu, který vás zajímá. Existuje nějaká otázka jako „Rodí se prvorozené děti se zpožděním?“, kterou byste chtěli prozkoumat?

Přemýšlejte o otázkách, které vám osobně připadají zajímavé, nebo o obecně vžitých názorech, anebo o kontroverzních tématech, která mají politické důsledky, a zjistěte, jestli dokážete formulovat otázku, která by se hodila ke statistickému zkoumání.

Podívejte se po datech, která by vám pomohla vaši otázku zodpovědět. Užitečným zdrojem mohou být vlády, protože data z veřejných výzkumů jsou často volně dostupná[2].

[2] V den, kdy jsem napsal tento odstavec, rozhodl soud ve Velké Británii, že zákon o svobodném přístupu k informacím (Freedom of Information Act) se vztahuje na data z vědeckých výzkumů.

Dalším způsobem, jak najít data, je Wolfram Alpha, což je spravovaná sbírka kvalitních souborů dat dostupná na `http://wolframalpha.com`. Výsledky z Wolfram Alpha podléhají autorskoprávním omezením. Je dobré si přečíst podmínky, než se k něčemu zavážete.

Google a další vyhledávače vám také mohou pomoci s hledáním dat, ale v tomto případě může být obtížnější vyhodnotit kvalitu zdrojů na webu.

Pokud máte pocit, že někdo už vaši otázku zodpověděl, podívejte se pořádně, jestli je odpověď odůvodněná. Data nebo analýza mohou trpět nějakými nedostatky, v důsledku nichž je závěr nespolehlivý. V takovém případě můžete provést jinou analýzu se stejnými daty, nebo se poohlédnout po lepším zdroji dat.

Jestliže najdete publikovaný článek, který se zabývá vaší otázkou, měli byste být schopni získat nezpracovaná data. Mnoho autorů zveřejňuje svá data na webu, ale pokud jde o citlivé údaje, bude nejspíš potřeba obrátit se na příslušné autory, poskytnout jim informace o tom, k jakému účelu chcete data použít, nebo souhlasit s konkrétními podmínkami užití. Buďte vytrvalí!

1.6 Glosář

anekdotický důkaz (anecdotal evidence): Důkaz, často osobní povahy, který je získán neformální cestou namísto dobře naplánované studie.

populace, základní soubor (population): Skupina, kterou máme zájem zkoumat, často skupina lidí, ale tento pojem se používá i pro zvířata, rostliny a minerály[3].

průřezová studie (cross-sectional study): Studie, která shromažďuje data o určité populaci v určitém okamžiku.

longitudinální studie (longitudinal study): Studie, která sleduje určitou populaci v průběhu času. Data jsou tedy opakovaně sbírána od téže skupiny.

respondent: Osoba, která se účastní šetření.

kohorta (cohort): Skupina nebo soubor respondentů, kteří mají nějaký společný charakteristický znak – například věk, pohlaví, vzdělání apod.

[3]Pokud tuto frázi nepoznáváte, podívejte se na `http://wikipedia.org/wiki/Twenty_Questions`.

výběr, vzorek, výběrový soubor (sample): Podmnožina populace použitá ke sběru dat.

reprezentativní (representative): Vzorek je reprezentativní, jestliže každý člen populace má stejnou šanci být zahrnut do vzorku.

nadreprezentování (oversampling): Technika navýšení zastoupení části populace za účelem eliminace chyb v důsledku malého rozsahu výběru.

záznam (record): V databázi, soubor informací o jedné osobě nebo jiném objektu zkoumání.

pole (field): V databázi, jedna z pojmenovaných proměnných, která tvoří záznam.

tabulka (table): V databázi, soubor záznamů.

surová, nezpracovaná data (raw data): Hodnoty získané a zaznamenané bez nebo s malou mírou kontroly, výpočtu nebo interpretace.

rekódovaná proměnná (recode): Hodnota, která je vygenerovaná výpočtem nebo jinou logickou operací aplikovanou na surová data.

souhrnná statistická charakteristika (summary statistic): Výsledek výpočtu, který redukuje soubory dat na jediné číslo (nebo alespoň na malou množinu čísel), které(á) vystihuje(í) nějakou charakteristickou vlastnost dat.

pozorované zjištění, pozorovaný výsledek (apparent effect): Zjištění nebo souhrnná statistická charakteristika, která naznačuje, že se děje něco zajímavého.

statisticky významný (statistically significant): Pozorované zjištění je statisticky významné, jestliže je nepravděpodobné, aby k němu došlo náhodně.

artefakt (artifact): Pozorované zjištění, které je způsobeno zkreslením, chybou měření nebo jiným druhem chyby.

Kapitola 2

Popisná statistika

2.1 Střední hodnoty a průměry

V předchozí kapitole jsem se zmínil o třech souhrnných statistických charakteristikách – průměru, rozptylu a mediánu – aniž bych vysvětlil, o co se jedná. Než se tedy pustíme do dalšího výkladu, pojďme se na ně blíže podívat.

Máme-li vzorek x_i o n hodnotách, pak průměr, μ, je součet hodnot vydělený počtem hodnot; neboli

$$\mu = \frac{1}{n} \sum_i x_i$$

- Průměr vzorku je souhrnná statistická charakteristika, kterou vypočteme pomocí výše uvedeného vzorce.
- Střední hodnoty v tomto textu chápeme jako obecné označení pro statistické charakteristiky, které popisují typické hodnoty vzorku, jako je průměr, modus nebo medián. (Ve statistice se pak setkáte se střední hodnotou ve smyslu váženého průměru náhodného rozdělení; rovněž bývá používána jako synonymum pro medián).

V některých případech se průměr k popisu souboru hodnot dobře hodí. Například jablka bývají přibližně stejně velká (alespoň ta v supermarketech). Jestliže si tedy koupím 6 jablek, jejichž celková hmotnost je 3 libry, pak můžu vyvodit přiměřený závěr, že každé z nich váží zhruba půl libry.

Naproti tomu u dýní je třeba počítat s větší rozmanitostí. Předpokládejme, že ve své zahrádce pěstuji několik druhů dýní a jednoho dne sklidím tři

okrasné dýně, z nichž každá váží 1 libru, dvě dýně, které se hodí na pečení koláčů, po 3 librách každá a jednu obří dýni odrůdy Atlantic Giant® , která váží 591 liber. Průměr tohoto vzorku je 100 liber, ale kdybych vám řekl: „Průměrná dýně v mojí zahradě váží 100 liber," nebyl by takový výrok pravdivý, nebo by byl přinejmenším zavádějící.

V tomto příkladu neexistuje žádná smysluplná střední hodnota, protože neexistuje žádná typická dýně.

2.2 Rozptyl

Jestliže neexistuje jedno číslo, které by poskytovalo souhrnnou informaci o hmotnostech dýní, mohou nám lépe posloužit čísla dvě: průměr a **rozptyl**.

Stejně jako průměr slouží k popisu centrální tendence, rozptyl slouží k popisu **variability**. Rozptyl souboru hodnot je definován vzorcem

$$\sigma^2 = \frac{1}{n}\sum_i (x_i - \mu)^2$$

Složka x_i-μ se nazývá „odchylka od průměru", takže rozptyl představuje střední kvadratickou odchylku, a proto se označuje jako σ^2. Druhá odmocnina z rozptylu, σ, se nazývá **směrodatná odchylka**.

Pokud vezmeme rozptyl jako takový, je obtížně interpretovatelný. Jedním z problémů je, že pracuje se zvláštními jednotkami. V našem příkladu jsme jako měrnou jednotku použili libru, rozptyl je proto vyjádřen jako druhá mocnina libry. Jako smysluplnější se jeví směrodatná odchylka, jejíž jednotkou jsou v tomto případě libry.

Cvičení 2.1 K cvičením v této kapitole si stáhněte materiály dostupné na `http://thinkstats.com/thinkstats.py`, které obsahují univerzální funkce, jež budeme používat při práci s touto knihou. Dokumentaci k těmto funkcím si můžete přečíst zde: `http://thinkstats.com/thinkstats.html`.

Napište funkci s názvem `Pumpkin`, která využívá funkce z `thinkstats.py` k výpočtu průměru, rozptylu a směrodatné odchylky vah dýní z předešlé části.

Cvičení 2.2 Znovu použijte kód z `survey.py` a `first.py` a vypočtěte směrodatnou odchylku délky těhotenství pro prvorozené děti a ostatní. Zdá se, že variabilita je pro obě skupiny stejná?

Jak velký je rozdíl mezi průměry v porovnání s těmito směrodatnými odchylkami? Co z tohoto srovnání vyplývá ohledně statistické významnosti daného rozdílu?

Pokud už máte nějaké předchozí zkušenosti, možná jste se setkali se vzorcem pro rozptyl, který měl ve jmenovateli výraz n - 1 , namísto n. Tato statistická charakteristika se nazývá „výběrový rozptyl“ a používá se k odhadu rozptylu v populaci na základě výběrového souboru. K tomuto se opět vrátíme v Kapitole 8.

2.3 Rozdělení

Souhrnné statistické charakteristiky jsou výstižné, ale skrývají v sobě jisté nebezpečí, protože zastírají data. Alternativou je sledování **rozdělení** dat, které ukazuje, jak často se která hodnota vyskytuje.

Nejběžnějším znázorněním rozdělení je **histogram**, což je graf zobrazující četnost nebo pravděpodobnost každé hodnoty.

V tomto kontextu znamená **četnost** (frequency) počet výskytů hodnoty v souboru dat – nemá to nic společného s výškou zvuku nebo laděním rádiového signálu. **Pravděpodobnost** je četnost vyjádřená jako podíl rozsahu výběru , n.

Efektivním postupem pro výpočet četností v Pythonu je použití slovníku. Uvažujme posloupnost hodnot, `t`:

```
hist = {}
for x in t:
    hist[x] = hist.get(x, 0) + 1
```

Výsledkem je slovník, kde jsou hodnotám přiřazeny četnosti. Abychom se od četností dostali k pravděpodobnostem, provedeme dělení n, které se označuje jako **normalizace**:

```
n = float(len(t))
pmf = {}
for x, freq in hist.items():
    pmf[x] = freq / n
```

Normalizovaný histogram se označuje jako **PMF** („probability mass function“), tedy pravděpodobnostní funkce. Jedná se o funkci, kde jsou

hodnotám přiřazeny pravděpodobnosti (pojem „mass“ (hmotnost) objasním v Oddílu 6.3).

To, že nazýváme slovník v jazyce Python funkcí, může být poněkud matoucí. V matematice se pojmem funkce označuje zobrazení z jedné množiny hodnot do jiné. V Pythonu *obvykle* zobrazujeme matematické funkce prostřednictvím objektů funkcí, avšak v tomto případě používáme slovník (slovníky se také nazývají „zobrazení“, jestli vám to pomůže k lepšímu pochopení).

2.4 Zobrazování histogramů

Vytvořil jsem pythonovský modul s názvem `Pmf.py`, který obsahuje definice tříd pro Hist objekty, které zobrazují histogramy, a Pmf objekty, které zobrazují pravděpodobnostní funkce (PMFs). Dokumentaci si můžete projít na `thinkstats.com/Pmf.html` a kód si stáhnout z `thinkstats.com/Pmf.py`.

Funkce `MakeHistFromList` přijme seznam hodnot a vrátí nový Hist objekt. Můžete to vyzkoušet v pythonovském interaktivním režimu:

```
>>> import Pmf
>>> hist = Pmf.MakeHistFromList([1, 2, 2, 3, 5])
>>> print hist
<Pmf.Hist object at 0xb76cf68c>
```

`Pmf.Hist` znamená, že tento objekt je členem Hist třídy, definované v modulu Pmf. Obecně platí, že používám velká písmena pro názvy tříd a funkcí a malá písmena pro proměnné.

Hist objekty poskytují metody, jak vyhledat hodnoty a jejich pravděpodobnosti. `Freq` přijme hodnotu a vrátí její četnost:

```
>>> hist.Freq(2)
2
```

Jestliže vyhledáte hodnotu, která se nikdy nevyskytla, pak je četnost 0.

```
>>> hist.Freq(4)
0
```

`Values` vrátí nesetříděný seznam hodnot v Hist:

```
>>> hist.Values()
[1, 5, 3, 2]
```

Chcete-li si projít setříděné hodnoty, můžete použít zabudovanou funkci sorted:

```
for val in sorted(hist.Values()):
    print val, hist.Freq(val)
```

Pokud máte v plánu vyhledat všechny četnosti, pak bude efektivnější použít Items, která vrátí nesetříděný seznam párů hodnota-četnost:

```
for val, freq in hist.Items():
     print val, freq
```

Cvičení 2.3 Modus rozdělení je hodnota s největší četností (viz http://wikipedia.org/wiki/Mode_(statistics)). Napište funkci nazvanou Mode, která přijme Hist objekt a vrátí hodnotu s největší četností.

Jako trochu pokročilejší verzi napište funkci s názvem AllModes, která přijme objekt Hist a vrátí seznam párů hodnota-četnost v sestupném pořadí četnosti. Nápověda: modul operator nabízí funkci nazvanou itemgetter, kterou můžete přenést jako klíč do sorted.

2.5 Grafické znázornění histogramů

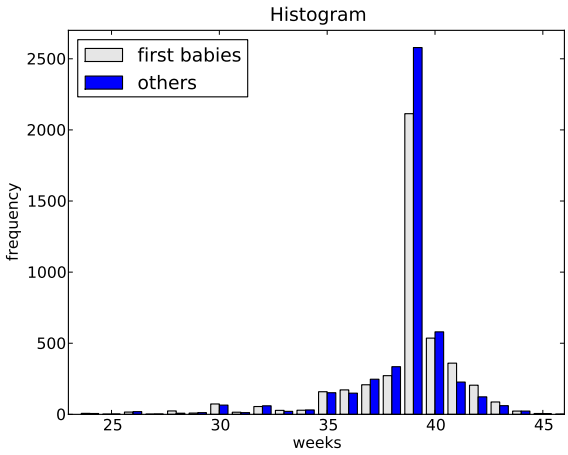
Existuje celá řada pythonovských balíků k vytváření obrázků a grafů. My se zde podíváme na pyplot, který je součástí balíku matplotlib dostupného na http://matplotlib.sourceforge.net.

Tento balík je součástí mnoha instalací Pythonu. Abyste zjistili, jestli jej máte, spusťte Python překladač a proveďte:

```
import matplotlib.pyplot as pyplot
pyplot.pie([1,2,3])
pyplot.show()
```

Jestliže máte matplotlib, měl by se vám zobrazit jednoduchý kruhový graf; v opačném případě jej budete muset nainstalovat.

Histogramy a pravděpodobnostní funkce (PMFs) jsou nejčastěji graficky znázorňovány jako sloupcové grafy. Funkce, která se použije v pyplot k nakreslení grafu, je bar. Hist objekty umožňují metodu nazvanou Render, která vrátí setříděný seznam hodnot a seznam příslušných četností, což je formát, který bar předpokládá:



Obrázek 2.1: Histogram délek těhotenství.

```
>>> vals, freqs = hist.Render()
>>> rectangles = pyplot.bar(vals, freqs)
>>> pyplot.show()
```

Napsal jsem modul nazvaný `myplot.py`, který poskytuje funkce pro grafické znázornění histogramů, pravděpodobnostních funkcí (PMFs) a dalších objektů, s nimiž se záhy seznámíme. Související dokumentaci si můžete přečíst na `thinkstats.com/myplot.html` a kód si stáhnout z `thinkstats.com/myplot.py`. Nebo můžete použít `pyplot` přímo, dle vlastního uvážení. V obou případech je dokumentace pro `pyplot` k dispozici na webu.

Obrázek 2.1 ukazuje histogramy délek těhotenství pro prvorozené děti a ostatní.

Histogramy jsou užitečné, protože jsou z nich bezprostředně patrné následující vlastnosti:

Modus: Nejčastěji se vyskytující hodnota v rozdělení se nazývá **modus**. Na obrázku 2.1 je zřejmý modus v 39. týdnu. V tomto případě je modus souhrnnou statistickou charakteristikou, která nejlépe vystihuje typickou hodnotu.

Tvar: Kolem modu je rozdělení asymetrické; napravo klesá rychle, zatímco nalevo je pokles pomalejší. Z medicínského hlediska to dává smysl. Děti se často narodí předčasně, ale zřídka později než ve 42. týdnu. Pravá strana rozdělení je však useknutá také kvůli tomu, že doktoři často po 42. týdnu zasáhnou.

Odlehlé hodnoty: Hodnoty vzdálené od modu se nazývají **odlehlé hodnoty**. Některé z nich představují neobvyklé případy, jako například děti narozené v 30. týdnu. Mnohé z nich jsou ale způsobeny chybami, k nimž dojde při vykazování nebo zaznamenávání dat.

Přestože histogramy ozřejmí některé vlastnosti, obvykle se nedají příliš dobře použít pro srovnání dvou rozdělení. V tomto příkladu je méně „prvorozených“ než „ostatních“ dětí, a tak jsou některé zjevné rozdíly v histogramech způsobeny rozsahem výběru. Tento problém můžeme vyřešit pomocí pravděpodobnostních funkcí (PMFs).

2.6 Zobrazování pravděpodobnostních funkcí (PMFs)

`Pmf.py` poskytuje třídu s názvem `Pmf`, která zobrazuje pravděpodobnostní funkce (PMFs). Její zápis může být matoucí, ale je opravdu takovýto: `Pmf` je název modulu a také název třídy, takže celý název třídy je `Pmf.Pmf`. Často používám `pmf` jako název proměnné. Konečně pak v textu používám zkratku PMF pro označení obecného konceptu pravděpodobnostní funkce (probability mass function), nezávisle na mém způsobu implementace.

Pro vytvoření Pmf objektu, použijte `MakePmfFromList`, která přijme seznam hodnot:

```
>>> import Pmf
>>> pmf = Pmf.MakePmfFromList([1, 2, 2, 3, 5])
>>> print pmf
<Pmf.Pmf object at 0xb76cf68c>
```

Pmf a Hist objekty si jsou v mnoha ohledech podobné. Metody `Values` a `Items` fungují stejně pro oba typy. Největší rozdíl je v tom, že Hist přiřadí k hodnotám počitadla celých čísel, zatímco Pmf přiřadí k hodnotám pravděpodobnosti v režimu pohyblivé řádové čárky.

K vyhledání pravděpodobnosti spojené s konkrétní hodnotou, použijte `Prob`:

```
>>> pmf.Prob(2)
0.4
```

Stávající Pmf můžete modifikovat zvýšením pravděpodobnosti spojené s hodnotou:

```
>>> pmf.Incr(2, 0.2)
>>> pmf.Prob(2)
0.6
```

Nebo můžete pravděpodobnost vynásobit činitelem:

```
>>> pmf.Mult(2, 0.5)
>>> pmf.Prob(2)
0.3
```

Jestliže modifikujete Pmf, výsledek nemusí být normalizovaný, tzn. že pravděpodobnosti již nemusí být v součtu rovny 1. Pro kontrolu můžete volat `Total`, která vrátí součet pravděpodobností:

```
>>> pmf.Total()
0.9
```

Pro renormalizaci volejte `Normalize`:

```
>>> pmf.Normalize()
>>> pmf.Total()
1.0
```

Pmf objekty umožňují metodu `Copy`, takže si můžete vytvořit a modifikovat kopii, aniž by tím byl dotčen originál.

Cvičení 2.4 Jak uvádí Wikipedie: „Analýza přežití je statistická disciplína, která se zabývá smrtí biologických organismů a selháním mechanických systémů;“ viz `http://wikipedia.org/wiki/Survival_analysis`.

V rámci analýzy přežití je často užitečné vypočíst např. zůstatkovou životnost mechanické součástky. Pokud známe rozdělení životností a stáří součástky, jsme schopni spočítat rozdělení zůstatkových životností.

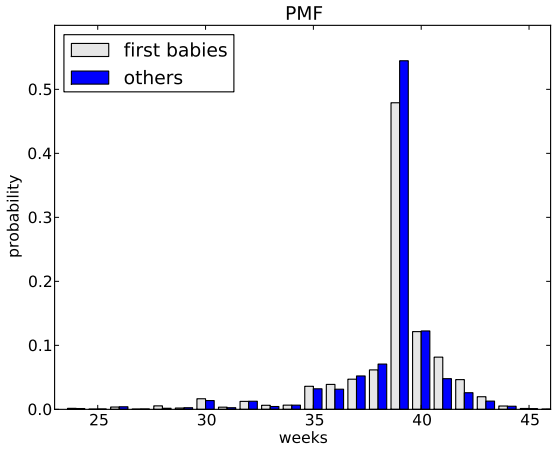
Napište funkci nazvanou `RemainingLifetime`, která přijme Pmf životností a věku a vrátí novou Pmf, která zobrazuje rozdělení zůstatkových životností.

Cvičení 2.5 V Oddílu 2.1 jsme spočítali průměr výběru sečtením prvků a vydělením n. Jestliže znáte PMF, můžete stále vypočítat průměr, ovšem postup bude trochu jiný:

$$\mu = \sum_i p_i x_i$$

kde x_i jsou jedinečné hodnoty v PMF a p_i=PMF(x_i). Obdobně můžete vypočíst rozptyl pomocí následujícího vzorce:

$$\sigma^2 = \sum_i p_i (x_i - \mu)^2$$



Obrázek 2.2: PMF délek těhotenství.

Napište funkce nazvané `PmfMean` a `PmfVar`, které přijmou Pmf objekt a vypočtou průměr a rozptyl. K otestování těchto metod zkontrolujte, jestli jsou konzistentní s metodami `Mean` a `Var` v `Pmf.py`.

2.7 Grafické znázornění pravděpodobnostních funkcí (PMFs)

Běžně se můžeme setkat se dvěma způsoby grafického znázornění PMFs:

- Pro znázornění Pmf pomocí sloupcového grafu můžete použít `pyplot.bar` nebo `myplot.Hist`. Sloupcové grafy se nejlépe hodí, jestliže jsou číslo nebo hodnoty v Pmf malé.
- Pro znázornění Pmf jako přímky můžete použít `pyplot.plot` nebo `myplot.Pmf`. Spojnicové grafy jsou nejvhodnější, jestliže máme velké množství hodnot a Pmf je vyrovnaná.

Obrázek 2.2 ukazuje PMF délek těhotenství jako sloupcový graf. Prostřednictvím PMF jsme schopni zřetelněji vidět, kde se rozdělení liší. Zdá se, že u prvorozených dětí je menší pravděpodobnost, že se narodí načas (v 39. týdnu) a větší pravděpodobnost, že se narodí se zpožděním (v 41. a 42. týdnu).

Kód generující obrázky v této kapitole je dostupný na `http://thinkstats.com/descriptive.py`. K jeho spuštění budete potřebovat moduly, které importuje, a data z NSFG (viz Oddíl 1.3).

Poznámka: `pyplot` umožňuje funkci nazvanou `hist`, která přijme posloupnost hodnot, spočítá histogram a graficky jej znázorní. Protože používám `Hist` objekty, obvykle nepoužívám `pyplot.hist`.

2.8 Odlehlé hodnoty

Odlehlé hodnoty jsou hodnoty, které jsou vzdálené od střední hodnoty. Odlehlé hodnoty mohou být způsobeny chybami při sběru nebo zpracování dat, nebo se může jednat o správné ale neobvyklé výsledky měření. Je vždy rozumné provést kontrolu s ohledem na odlehlé hodnoty a přitom je někdy také užitečné a vhodné je odstranit.

V seznamu délek těhotenství pro porody živého dítěte je následujících 10 nejnižších hodnot {0, 4, 9, 13, 17, 17, 18, 19, 20, 21}. Hodnoty pod 20 týdnů jsou zcela jistě chyby, hodnoty nad 30 týdnů jsou pravděpodobně legitimní. Avšak hodnoty mezi tím se těžko interpretují.

Naproti tomu nejvyšší hodnoty jsou následující:

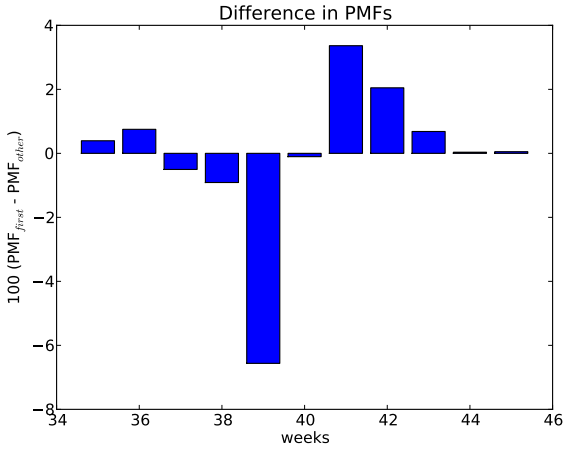
```
weeks  count
43     148
44     46
45     10
46     1
47     1
48     7
50     2
```

Opět platí, že některé hodnoty jsou téměř jistě chyby, není ale snadné získat o tom absolutní jistotu. Jednou možností je provést **ořezání** dat tím, že odstraníme část nejvyšších a nejnižších hodnot (viz `http://wikipedia.org/wiki/Truncated_mean`).

2.9 Další vizualizace

Histogramy a pravděpodobnostní funkce (PMFs) se hodí pro explorační analýzu dat; jakmile máte představu o tom, co se děje, je často užitečné navrhnout vizualizaci, která se zaměří na pozorované zjištění.

V rámci dat NSFG se největší rozdíly v rozdělení objevují blízko modu. Má proto smysl přiblížit si tuto část grafu a provést transformaci dat ke zvýraznění rozdílů.



Obrázek 2.3: Rozdíl v procentech, po týdnech.

Obrázek 2.3 ukazuje rozdíl mezi PMFs pro týdny 35–45. Provedl jsem násobení 100, abychom získali rozdíly v procentních bodech.

Na tomto obrázku je dobře patrný vzor: u prvorozených dětí je menší pravděpodobnost, že se narodí v 39. týdnu, a poněkud větší pravděpodobnost jejich narození v 41. nebo 42. týdnu.

2.10 Relativní riziko

Začali jsme otázkou: „Rodí se prvorozené děti se zpožděním?“ Abychom to trochu upřesnili, řekněme, že dítě se narodí předčasně, pokud přijde na svět v průběhu 37. týdne nebo dříve, načas, pokud se narodí v průběhu 38., 39. nebo 40. týdne, a se zpožděním, pokud se narodí v průběhu 41. týdne nebo později. Rozpětí jako tato, která se používají k seskupování dat, se nazývají **třídy**.

Cvičení 2.6 Vytvořte soubor s názvem `risk.py`. Napište funkce nazvané `ProbEarly`, `ProbOnTime` a `ProbLate`, které přijmou PMF a vypočítají podíl narození, která spadají do jednotlivých tříd. Nápověda: Napište zobecněnou funkci, kterou tyto funkce volají.

Vytvořte tři PMFs, jednu pro prvorozené děti, jednu pro ostatní a jednu pro všechny porody živých dětí. Pro každou PMF vypočtěte pravděpodobnost předčasného narození, narození načas a narození se zpožděním.

Jedním ze způsobů jak sumarizovat takováto data je pomocí **relativního rizika**, které představuje poměr dvou pravděpodobností. Například prav-

děpodobnost, že se prvorozené dítě narodí předčasně, je 18,2 %. Pro ostatní děti je tato pravděpodobnost 16,8 % a relativní riziko je tedy 1,08. To znamená, že u prvorozených dětí je přibližně o 8 % větší pravděpodobnost, že se narodí předčasně.

Napište kód, který potvrdí tento výsledek, a poté vypočtěte relativní riziko předčasného narození a narození se zpožděním. Řešení si můžete stáhnout z `http://thinkstats.com/risk.py`.

2.11 Podmíněná pravděpodobnost

Představte si, že nějaká vaše známá je těhotná a právě začíná 39. týden jejího těhotenství. Jaká je šance, že se dítě narodí v následujícím týdnu? Jak se odpověď změní, jestliže se jedná o první dítě?

Odpověď na tuto otázku můžeme najít prostřednictvím výpočtu **podmíněné pravděpodobnosti**, což je (ehm!) pravděpodobnost, která závisí na podmínce. V tomto případě je touto podmínkou skutečnost, že víme, že se dítě nenarodilo v 0.–38. týdnu.

Zde je jeden způsob, jak to provést:

1. Na základě známé PMF vytvořte falešnou kohortu o 1000 těhotenstvích. Pro každý počet týdnů, x, je počet těhotenství s délkou x 1000 PMF(x).
2. Z kohorty odstraňte všechna těhotenství, která trvají méně než 39 týdnů.
3. Vypočtěte PMF zbývajících délek; výsledkem je podmíněná PMF.
4. Vypočtěte podmíněnou PMF, jestliže $x = 39$ týdnů.

Tento algoritmus je koncepčně jasný, ale není příliš účinný. Jednoduchou alternativou je odstranit z rozdělení hodnoty pod 39 a pak provést renormalizaci.

Cvičení 2.7 Napište funkci, která uplatní kterýkoli z těchto algoritmů a vypočte pravděpodobnost, že se dítě narodí v průběhu 39. týdne, za předpokladu, že se nenarodilo před začátkem 39. týdne.

Zobecněte tuto funkci tak, abyste vypočítali pravděpodobnost, že se dítě narodí během týdne x, za předpokladu, že se nenarodilo před začátkem

týdne x, a to pro všechna x. Proveďte grafické znázornění této hodnoty jako funkce x pro prvorozené děti a ostatní.

Řešení tohoto úkolu si můžete stáhnout zde: `http://thinkstats.com/conditional.py`.

2.12 Referování o výsledcích

Dospěli jsme do bodu, kdy jsme prozkoumali data a zaznamenali jsme několik pozorovaných zjištění. Pro tuto chvíli předpokládejme, že jsou tato zjištění skutečná (nezapomeňme však, že jde o předpoklad). Jak bychom měli o těchto výsledcích referovat?

Odpověď by se mohla odvíjet od toho, kdo otázku položil. Například vědec by se mohl zajímat o jakýkoli (skutečný) výsledek, bez ohledu na jeho velikost. Doktor by se mohl zajímat pouze o ty výsledky, které jsou **klinicky významné**; totiž takové, které mají vliv na rozhodnutí o léčbě. Těhotnou ženu by mohly zajímat výsledky, které jsou pro ni relevantní, jako například podmíněné pravděpodobnosti, jimiž jsme se zabývali v předchozí části.

Způsob referování o výsledcích závisí také na vašich cílech. Jestliže se snažíte ukázat významnost nějakého zjištění, pak byste mohli zvolit souhrnné statistické charakteristiky, jako například relativní riziko, které zdůrazňují rozdíly. Jestliže je vaším cílem uklidnit pacienta, možná byste raději sáhli po statistických charakteristikách, které tyto rozdíly zasadí do kontextu.

Cvičení 2.8 Na základě výsledků z předešlých cvičení předpokládejte, že vás někdo požádal, abyste shrnul/a, co jste zjistil/a o tom, jestli se prvorozené děti rodí se zpožděním.

Které souhrnné statistické charakteristiky byste použil/a, pokud byste chtěl/a dostat nějaký příběh do večerních zpráv? A které byste použil/a, kdybyste chtěl/a uklidnit nervózního pacienta?

Na závěr si představte, že jste Cecil Adams, autor *The Straight Dope* (`http://straightdope.com`), a vaším úkolem je odpovědět na otázku: „Rodí se prvorozené děti se zpožděním?“ Napište odstavec, kterým na tuto otázku odpovíte s pomocí výsledků z této kapitoly jasně, přesně a výstižně.

2.13 Glosář

centrální tendence (central tendency): Charakteristika vzorku populace; intuitivně se jedná o nejprůměrnější hodnotu.

variabilita (spread): Charakteristika vzorku populace; intuitivně se jedná o popis velikosti variability.

rozptyl (variance): Souhrnná statistická charakteristika, která se často používá ke kvantifikaci variability.

směrodatná odchylka (standard deviation): Druhá odmocnina z rozptylu, také se používá jako míra variability.

četnost (frequency): Počet výskytů hodnoty ve výběrovém souboru.

histogram: Zobrazení rozložení četností jednotlivých hodnot nebo graf, který toto zobrazení znázorňuje.

pravděpodobnost (probability): Četnost vyjádřená jako podíl rozsahu výběru.

normalizace (normalization): Proces dělení četnosti rozsahem výběru za účelem získání pravděpodobnosti.

rozdělení (distribution): Souhrn hodnot, které se vyskytují ve výběru a četnost, nebo pravděpodobnost, každé z nich.

pravděpodobnostní funkce (PMF) (probability mass function): Zobrazení rozdělení jako funkce, kde jsou hodnotám přiřazeny pravděpodobnosti.

modus (mode): Hodnota s největší četností ve výběru.

odlehlá hodnota (outlier): Hodnota vzdálená od střední hodnoty.

ořezat (trim): Odstranit ze souboru dat odlehlé hodnoty.

třída (bin): Rozpětí používané k seskupování hodnot, které se vyskytují blízko sebe.

relativní riziko (relative risk): Poměr dvou pravděpodobností, často používaný k měření rozdílu mezi rozděleními.

podmíněná pravděpodobnost (conditional probability): Pravděpodobnost vypočtená za předpokladu platnosti určité podmínky.

klinicky významný (clinically significant): Výsledek, jako například rozdíl mezi skupinami, který je relevantní pro praxi.

Kapitola 3

Distribuční funkce

3.1 Paradox počtu studentů v kurzu

Na řadě amerických vysokých škol a univerzit je poměr studentů k fakultnímu sboru 10:1. Studenti jsou však často překvapeni, když zjistí, že průměrný počet studentů v kurzu je větší než 10. Za touto nesrovnalostí stojí dva důvody:

- Studenti si obvykle zvolí 4–5 kurzů za semestr, ale učitelé často učí pouze 1 nebo 2 kurzy.
- Není mnoho studentů, kteří by si libovali v kurzech, do kterých je zapsáno málo studentů, ale počet studentů v kurzu, který si zvolí velké množství studentů, je (ehm!) velký.

První efekt je zřejmý (minimálně poté, co je na něj poukázáno), zatímco ten druhý tak evidentní není. Podívejme se tedy na příklad. Předpokládejme, že vysoká škola nabízí v rámci jednoho semestru 65 kurzů, přičemž jejich rozdělení podle počtu studentů v kurzu (velikosti kurzu) je následující:

velikost	počet
5- 9	8
10-14	8
15-19	14
20-24	4
25-29	6
30-34	12
35-39	8
40-44	3
45-49	2

Pokud se zeptáte na průměrný počet studentů v kurzu děkana, bude postupovat tak, že stanoví PMF, vypočte průměr a uvede, že průměrný počet studentů v kurzu je 24.

Pokud se ale budete dotazovat skupiny studentů na to, kolik studentů je v jejich kurzu, a vypočtete průměr, pak dospějete k závěru, že průměrný počet studentů v kurzu je větší.

Cvičení 3.1 Určete PMF těchto dat a vypočtěte průměr z pohledu děkana. Vzhledem k tomu, že data byla uspořádána do tříd, můžete použít střed každé třídy.

Nyní zjistěte rozdělení počtu studentů v kurzu z pohledu studentů a vypočtěte průměr.

Předpokládejme, že chcete zjistit rozdělení počtu studentů v kurzech na vysoké škole, ale nemůžete získat spolehlivá data od děkana. Alternativou je vybrat náhodný vzorek studentů a zeptat se jich na počet studentů v každém z kurzů, který navštěvují. Pak můžete vypočíst PMF na základě jejich odpovědí.

Výsledek by byl zkreslený nadreprezentováním kurzů s větším počtem studentů, ale skutečné rozdělení počtu studentů v kurzech byste mohli odhadnout na základě vhodné transformace pozorovaného rozdělení.

Napište funkci nazvanou `UnbiasPmf`, která přijme PMF pozorovaných hodnot a vrátí nový Pmf objekt, který vyjadřuje odhad rozdělení počtu studentů v kurzech.

Řešení tohoto problému si můžete stáhnout na `http://thinkstats.com/class_size.py`.

Cvičení 3.2 Ve většině běžeckých závodů startují všichni účastníci zároveň. Jestliže máte rychlé tempo, pak obvykle na začátku závodu předběhnete spoustu lidí, ale po několika mílích už běží všichni kolem vás zhruba stejnou rychlostí.

Když jsem běžel dálkový štafetový závod (209 mílí) poprvé, všiml jsem si zvláštního jevu: Když jsem předběhl jiného běžce, byl jsem obvykle výrazně rychlejší, a když jiný běžec předběhl mě, byl obvykle výrazně rychlejší.

Nejdřív jsem si myslel, že rozdělení rychlostí by mohlo být bimodální, totiž že se závodu účastnilo hodně pomalých běžců a hodně rychlých běžců, ale málo běžců s podobnou rychlostí, jakou běhám já.

Pak jsem si uvědomil, že jsem se stal obětí zkreslení výběru. Závod byl neobvyklý hned ve dvou ohledech: Jednak se startovalo postupně, takže jednotlivé týmy začínaly různě, a jednak byla řada týmů složená z běžců na různé úrovni.

Ve výsledku tak byli jednotliví běžci rozptýleni po trati, přičemž mezi jejich rychlostí a místem, kde se nacházeli, nebyl žádný jednoznačný vztah. Když jsem zahájil svůj úsek, představovali běžci kolem mě (do velké míry) náhodný vzorek běžců účastnících se závodu.

Odkud se tedy bere to zkreslení? V době, kdy běžím na trati závodu, je pravděpodobnost, že předběhnu jiného běžce nebo že on předběhne mě, proporcionální k rozdílu v našich rychlostech. Abyste si uvědomili, proč tomu tak je, zamyslete se nad extrémy. Jestliže jiný běžec běží stejně rychle jako já, pak ani jeden z nás nepředběhne toho druhého. Jestliže někdo běží tak rychle, že za dobu, kdy já běžím, uběhne celou trať závodu, pak je jisté, že mě takový běžec předběhne.

Napište funkci nazvanou `BiasPmf`, která přijme Pmf představující skutečné rozdělení rychlostí běžců a rychlost běžícího pozorovatele a vrátí novou Pmf představující rozdělení rychlostí běžců, jak je vnímá pozorovatel.

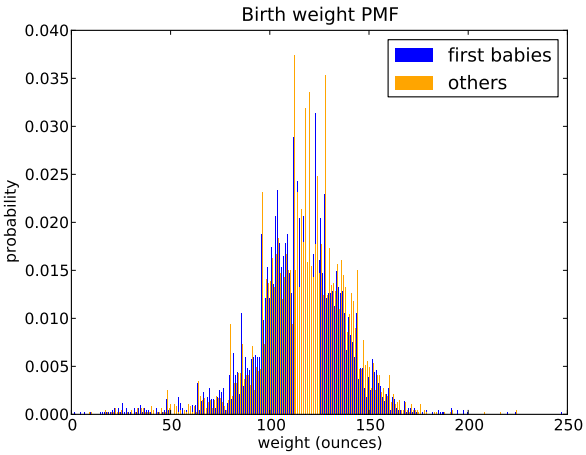
K otestování vaší funkce získejte rozdělení rychlostí z normálního běžeckého závodu (ne štafety). Vytvořil jsem program, který přečte výsledky závodu James Joyce Ramble 10K v Dedhamu ve státě Massachusetts a převede tempo každého běžce na míle za hodinu (MPH). Stáhněte si jej z: `http://thinkstats.com/relay.py`. Spusťte jej a podívejte se na PMF rychlostí.

Nyní vypočtěte rozdělení rychlostí, jaké byste pozoroval/a, pokud byste běžel/a štafetový závod rychlostí 7,5 MPH s touto skupinou běžců. Řešení si můžete stáhnout na: `http://thinkstats.com/relay_soln.py`

3.2 Limity pravděpodobnostních funkcí (PMFs)

Pravděpodobnostní funkce fungují dobře při malém počtu hodnot. Jakmile ale počet hodnot stoupne, pravděpodobnost spojená s každou z hodnot se sníží a narůstá vliv náhodného šumu.

Mohli bychom se například zajímat o rozdělení porodních hmotností. V rámci souboru dat NSFG je proměnná `totalwgt_oz` záznamem váhy při narození vyjádřené v uncích. Obrázek 3.1 znázorňuje PMF těchto hodnot pro prvorozené děti a ostatní.



Obrázek 3.1: PMF porodních hmotností. Tento obrázek znázorňuje omezení PMFs: obtížně se porovnávají.

Celkově připomínají tato rozdělení známé „normální rozdělení“ s velkým počtem hodnot blízko průměru a malým počtem hodnot o mnoho výše nebo níže.

Části tohoto grafu se ale obtížně interpretují. Je tam velké množství prudkých vzestupů a propadů a některé zjevné rozdíly mezi rozděleními. Je obtížné určit, které z těchto rysů jsou významné. Také není snadné identifikovat celkové vzory. Například které rozdělení má podle vás vyšší průměr?

Tyto problémy je možné zmírnit tím, že data uspořádáme do tříd, tedy rozdělíme oblast hodnot do intervalů, které se vzájemně nepřekrývají, a spočítáme počet hodnot v každé třídě. Rozdělení do tříd může být užitečné, ale není vůbec snadné správně nastavit velikost tříd. Budou-li dostatečně velké, aby odfiltrovaly šum, pak mohou zároveň odfiltrovat také užitečné informace.

Alternativou, která se těmto problémům vyhne, je **distribuční funkce** označovaná také jako **CDF**. Než se k ní ale dostaneme, je potřeba se podívat na percentily.

3.3 Percentily

Pokud jste vykonali nějaký standardizovaný test, pravděpodobně jste obdrželi výsledky ve formě hrubého skóre a **percentilového pořadí**. V tomto

kontextu představuje percentilové pořadí podíl osob, které dosáhly horšího výsledku než vy (nebo stejného). Takže pokud jste se umístili v „90. percentilu," dosáhli jste stejného nebo lepšího výsledku než 90 % všech, kdo zkoušku vykonali.

Zde je postup, jak můžete vypočítat percentilové pořadí hodnoty, `your_score`, relativně vůči skóre v rámci řady `skóre`:

```
def PercentileRank(scores, your_score):
    count = 0
    for score in scores:
        if score <= your_score:
            count += 1

    percentile_rank = 100.0 * count / len(scores)
    return percentile_rank
```

Například jestliže by skóre v řadě byla 55, 66, 77, 88 a 99 a vy byste dosáhl/a výsledku 88, pak by vaše percentilové pořadí bylo `100 * 4 / 5`, tedy 80.

Jestliže máte hodnotu, pak je snadné zjistit její percentilové pořadí, opačný postup je o něco obtížnější. Znáte-li percentilové pořadí a chcete zjistit odpovídající hodnotu, jednou z možností je uspořádat hodnoty a hledat tu, kterou chcete zjistit:

```
def Percentile(scores, percentile_rank):
    scores.sort()
    for score in scores:
        if PercentileRank(scores, score) >= percentile_rank:
            return score
```

Výsledkem tohoto výpočtu je **percentil**. Například 50. percentil je hodnota s percentilovým pořadím 50. V rozdělení výsledků zkoušky odpovídá 50. percentilu hodnota 77.

Cvičení 3.3 Tento způsob využití `Percentile` není příliš efektivní. Lepší postup je použít percentilové pořadí k výpočtu indexu odpovídajícího percentilu. Napište verzi `Percentile`, která využívá tohoto algoritmu.

Řešení si můžete stáhnout z `http://thinkstats.com/score_example.py`.

Cvičení 3.4 Volitelné: Pokud chcete vypočíst pouze jeden percentil, pak setřídění skóre příliš nepomůže. Lepší možností je zvolit selekční algoritmus, o kterém si můžete přečíst na `http://wikipedia.org/wiki/Selection_algorithm`.

Vytvořte (nebo najděte) provedení selekčního algoritmu a použijte jej k napsání efektivní verze `Percentile`.

3.4 Distribuční funkce

Teď, když už rozumíme percentilům, jsme připraveni začít se zabývat distribuční funkcí (CDF). Distribuční funkce (CDF) je funkce, která hodnotám přiřazuje jejich percentilové pořadí v rámci rozdělení.

CDF je funkcí x, kde x je jakákoliv hodnota, která se může vyskytnout v rozdělení. Abychom získali CDF(x) konkrétní hodnoty x, vypočteme podíl hodnot ve výběru, které jsou menší (nebo rovné) x.

Zde je vyjádření téhož jako funkce, která přijme výběr `t` a hodnotu `x`:

```
def Cdf(t, x):
    count = 0.0
    for value in t:
        if value <= x:
            count += 1.0

    prob = count / len(t)
    return prob
```

Tato funkce by vám měla připadat povědomá; je téměř identická s `PercentileRank`, až na to, že výsledkem je pravděpodobnost v rozpětí 0–1, namísto percentilového pořadí v rozpětí 0–100.

Jako příklad předpokládejme, že výběr je tvořen hodnotami {1, 2, 2, 3, 5}. Zde je několik hodnot z jeho distribuční funkce (CDF):

CDF(0) = 0

CDF(1) = 0,2

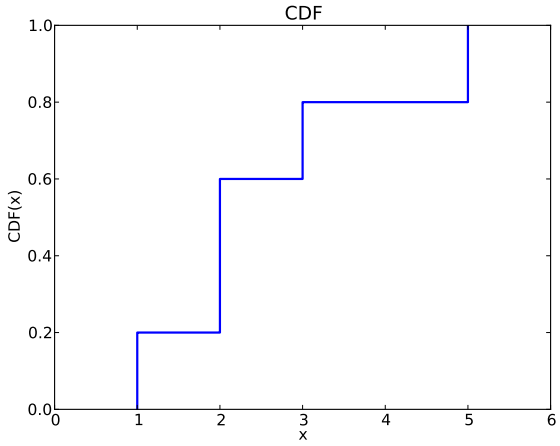
CDF(2) = 0,6

CDF(3) = 0,8

CDF(4) = 0,8

CDF(5) = 1

Můžeme vypočítat distribuční funkci pro jakoukoliv hodnotu, kterou může x nabývat, nejen hodnoty, které se vyskytují ve výběru. Jestliže je x menší



Obrázek 3.2: Příklad CDF.

než nejnižší hodnota ve výběru, pak CDF(x) je 0. Jestliže je x větší než nejvyšší hodnota, pak CDF(x) je 1.

Obrázek 3.2 je grafickým znázorněním této CDF. CDF výběru je skoková funkce. V následující kapitole se zaměříme na rozdělení, jejichž distribuční funkce jsou spojité funkce.

3.5 Zobrazení distribučních funkcí (CDFs)

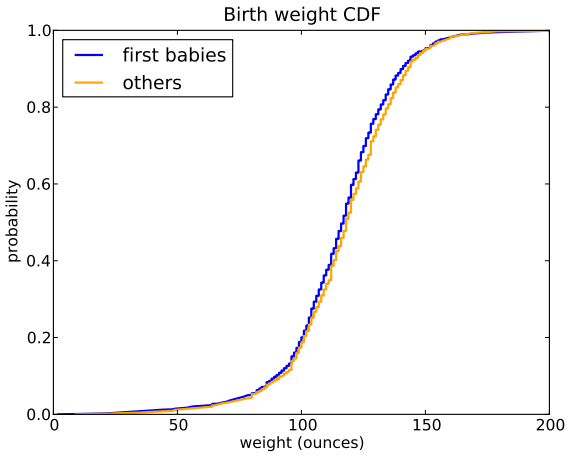
Vytvořil jsem modul nazvaný `Cdf`, který obsahuje třídu s názvem `Cdf`, která zobrazuje CDFs. Dokumentaci k tomuto modulu si můžete přečíst na `http://thinkstats.com/Cdf.html` a můžete si ji stáhnout zde: `http://thinkstats.com/Cdf.py`.

Cdfs jsou prováděny se dvěma setříděnými seznamy: `xs`, který obsahuje hodnoty, a `ps`, který obsahuje pravděpodobnosti. Nejdůležitější metody, které Cdfs nabízejí, jsou následující:

`Prob(x)`: Na základě hodnoty x vypočte pravděpodobnost p = CDF(x).

`Value(p)`: Na základě pravděpodobnosti p vypočte odpovídající hodnotu x; to znamená inverzní CDF p.

Protože `xs` a `ps` jsou setříděné, tyto funkce mohou využít bisekční algoritmus, který je efektivní. Doba běhu je proporcionální k logaritmu počtu hodnot; viz `http://wikipedia.org/wiki/Time_complexity`.



Obrázek 3.3: CDF porodních hmotností.

Cdfs také umožňují funkci `Render`, která vrátí dva seznamy, `xs` a `ps`, jež jsou vhodné ke grafickému znázornění CDF. Protože CDF je skoková funkce, tyto seznamy obsahují dva prvky pro každou jedinečnou hodnotu v rámci rozdělení.

Modul Cdf nabízí několik funkcí pro vytváření distribučních funkcí, včetně `MakeCdfFromList`, která přijme řadu hodnot a vrátí jejich Cdf.

Konečně pak `myplot.py` skýtá funkce nazvané `Cdf` a `Cdfs`, které graficky zobrazí distribuční funkce jako přímky.

Cvičení 3.5 Stáhněte si `Cdf.py` a `relay.py` (viz Cvičení 3.2) a vygenerujte graf, který zobrazí CDF běžeckých rychlostí. Která funkce vám poskytne lepší představu o tvaru rozdělení, PMF nebo CDF? Řešení si můžete stáhnout z `http://thinkstats.com/relay_cdf.py`.

3.6 Zpátky k datům z šetření

Obrázek 3.3 ukazuje CDFs porodních hmotností pro prvorozené děti a ostatní v souboru dat NSFG.

Díky tomuto obrázku získáme mnohem jasnější představu o tvaru rozdělení a rozdílech mezi nimi. Je patrné, že prvorozené děti jsou mírně lehčí napříč celým rozdělením, s větší odchylkou nad průměrem.

Cvičení 3.6 Jaká byla vaše váha při narození? Pokud nevíte, zavolejte svojí matce nebo někomu jinému, kdo to ví. Za použití úhrnných dat (všechny

porody živých dětí) vypočtěte rozdělení porodních hmotností a využijte je ke stanovení vašeho percentilového pořadí. Jestliže jste prvorození, zjistěte svoje percentilové pořadí v rozdělení pro prvorozené děti. Jinak použijte rozdělení pro ostatní děti. Jestliže jste v 90. percentilu nebo výše, zavolejte svojí matce a omluvte se jí.

Cvičení 3.7 Předpokládejte, že spolu se svými spolužáky počítáte percentilové pořadí vašich porodních hmotností, a pak vypočtěte CDF percentilových pořadí. Jak si myslíte, že bude vypadat? Nápověda: Jaká část vaší třídy bude podle vás nad mediánem?

3.7 Podmíněná rozdělení

Podmíněné rozdělení je rozdělení podmnožiny dat, která je zvolena podle určité podmínky.

Například pokud máte nadprůměrnou váhu, ale zároveň jste výrazně nadprůměrně vysoký/á, pak můžete mít relativně nízkou váhu na svoji výšku. Následuje postup, jak můžete takové tvrzení zpřesnit.

1. Zvolte kohortu lidí, kteří jsou stejně vysocí jako vy (v rámci určitého rozpětí).
2. Najděte CDF hmotnosti pro tyto lidi.
3. Najděte percentilové pořadí vaší váhy v tomto rozdělení.

Percentilová pořadí jsou užitečná pro porovnání výsledků z různých testů nebo testů použitých pro různé skupiny.

Například lidé účastnící se běžeckých závodů jsou obvykle rozděleni do skupin podle věku a pohlaví. Abyste mohli porovnat osoby z různých skupin, můžete převést časy, za které závod zaběhli, na percentilová pořadí.

Cvičení 3.8 Nedávno jsem běžel závod James Joyce Ramble 10K v Denhamu ve státě Massachusetts. Výsledky jsou dostupné na `http://coolrunning.com/results/10/ma/Apr25_27thAn_set1.shtml`. Běžte na uvedené stránky a najděte moje výsledky. Doběhl jsem 97. mezi 1633 závodníky, takže jaké je moje percentilové pořadí mezi závodníky?

V rámci mé sekce (M4049 znamená „muži ve věku mezi 40 a 49 lety“) jsem doběhl 26. z 256. Jaké je moje percentilové pořadí v mé sekci?

Pokud budu za 10 let stále ještě běhat (a já doufám, že budu), budu zařazen do sekce M5059. Za předpokladu, že se mé percentilové pořadí v mé sekci nezmění, jaké zpomalení bych měl očekávat?

Mezi mnou a mou studentkou, která je zařazena do sekce F2039, panuje přátelská rivalita. Jak rychle by musela běžet v dalším závodě 10K, aby mě "porazila", pokud jde o percentilové pořadí?

3.8 Náhodná čísla

Distribuční funkce se dobře hodí pro generování náhodných čísel na základě známého rozdělení. Podívejme se, jak se to dělá:

- Vyberte si náhodnou pravděpodobnost v rozpětí 0–1.
- Použijte `Cdf.Value` k nalezení hodnoty v rozdělení, jež odpovídá pravděpodobnosti, kterou jste si vybrali.

Nemusí být úplně zřejmé, proč toto funguje, ale protože je snazší to provést než vysvětlit, pojďme to zkusit.

Cvičení 3.9 Napište funkci s názvem `Sample`, která přijme Cdf a celé číslo n a vrátí seznam n hodnot náhodně zvolených z Cdf. Nápověda: Použijte `random.random`. Řešení tohoto cvičení najdete v `Cdf.py`.

S využitím rozdělení porodních hmotností ze souboru NSFG vytvořte náhodný vzorek o 1000 prvcích. Vypočtěte CDF vzorku. Sestavte graf, který zobrazuje původní CDF a CDF náhodného vzorku. Pro velké hodnoty n by měla být rozdělení stejná.

Tento proces generování náhodného vzorku na základě měřeného vzorku se nazývá **resampling**.

Existují dva způsoby, jak vybrat vzorek ze souboru: s opakování nebo bez opakování. Když si představíte, že taháte koule z urny[1], „opakování (vracení)“ znamená, že koule v průběhu vrátíte (a zamícháte), takže soubor zůstává při každém tahu stejný. „Bez opakování (vracení)“ znamená, že každá koule může být vytažena pouze jednou, takže zbývající soubor je po každém tahu jiný.

[1] Scénář s taháním koulí z urny představuje standardní model pro procesy náhodného výběru (viz `http://wikipedia.org/wiki/Urn_problem`).

V Pythonu může být výběr s opakováním proveden prostřednictvím `random.random`, která provede výběr percentilového pořadí, nebo `random.choice`, která slouží k výběru prvku z řady. K výběru bez opakování slouží `random.sample`.

Cvičení 3.10 Předpokládá se, že čísla vygenerovaná na základě `random.random` budou rovnoměrně rozdělená mezi 0 a 1, neboli že každá hodnota v rámci tohoto rozpětí bude mít stejnou pravděpodobnost.

Vytvořte 1000 čísel pomocí `random.random` a graficky znázorněte jejich PMF a CDF. Dokážete říci, jestli jsou rovnoměrně rozdělená?

O rovnoměrném rozdělení si můžete přečíst více na `http://wikipedia.org/wiki/Uniform_distribution_(discrete)`.

3.9 Souhrnné statistické charakteristiky podruhé

Jakmile jste vypočetli CDF, je snadné vypočíst také ostatní souhrnné statistické charakteristiky. Medián je přesně 50. percentil[2]. 25. a 75. percentil se často používají k ověření, zda je rozdělení symetrické. Jejich rozdíl, který se nazývá **mezikvartilové rozpětí**, pak měří variabilitu.

Cvičení 3.11 Napište funkci s názvem `Median`, která přijme Cdf a vypočte medián, a funkci nazvanou `Interquartile`, která vypočte mezikvartilové rozpětí.

Vypočtěte 25., 50. a 75. percentil CDF porodních hmotností. Naznačují tyto hodnoty, že se jedná o symetrickou distribuci?

3.10 Glosář

percentilové pořadí (percentile rank): Procento hodnot v rozdělení, které jsou menší nebo rovné určené hodnotě.

distribuční funkce (CDF) (cumulative distribution function): Funkce, která přiděluje hodnotám jejich percentilové pořadí.

[2]Můžete narazit na jiné definice mediánu. Jde zejména o to, že některé zdroje uvádějí, že jestliže máte sudý počet prvků ve výběru, je medián průměrem prostředních dvou prvků. Toto je ale zvláštní případ, který není nutné znát a který má navíc zvláštní efekt spočívající ve vytvoření hodnoty, jež není ve výběru. Pokud já vím, medián je 50. percentil. Tečka.

percentil (percentile): Hodnota spojená s konkrétním percentilovým pořadím.

podmíněné rozdělení (conditional distribution): Rozdělení vypočtené na základě platnosti určité podmínky.

resampling: Proces generování náhodného výběru z rozdělení, které bylo vypočteno z výběru.

opakování, vracení (replacement): Při pořizování výběru znamená „opakování“ to, že soubor je při každém výběru stejný. „Bez opakování“ znamená, že každý prvek může být vybrán pouze jednou.

mezikvartilové rozpětí (interquartile range): Míra variability, rozdíl mezi 75. a 25. percentilem.

Kapitola 4

Spojitá rozdělení

Rozdělení, s nimiž jsme se dosud setkali, se označují jako **empirická rozdělení**, protože vychází z empirických pozorování, tedy výběrových souborů, které jsou nevyhnutelně konečné.

Alternativou je **spojité rozdělení**, které se vyznačuje distribuční funkcí (CDF), jež je spojitou funkcí (v protikladu ke skokové funkci). Spojitá rozdělení mohou posloužit k přiblížení celé řady jevů reálného světa.

4.1 Exponenciální rozdělení

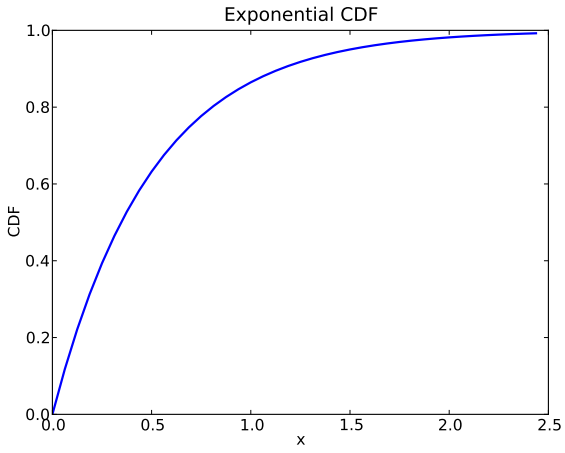
Začnu exponenciálním rozdělením, protože se s ním dobře pracuje. V reálném světě narazíme na exponenciální rozdělení, když se podíváme na sled jevů a změříme čas mezi výskytem těchto jevů, který označujeme jako **inter-arrival time**. Jestliže pro jevy platí, že se stejnou pravděpodobností mohou nastat kdykoliv, pak rozdělení časových intervalů mezi výskytem jevů má tendenci k exponenciálnímu rozdělení.

CDF exponenciálního rozdělení je následující:

$$CDF(x) = 1 - e^{-\lambda x}$$

Parametr λ určuje tvar rozdělení. Obrázek 4.1 ukazuje, jak vypadá tato CDF, jestliže $\lambda = 2$.

Obecně platí, že průměr exponenciálního rozdělení je $1/\lambda$, takže průměr tohoto rozdělení je 0,5. Medián je $\ln(2)/\lambda$, což je zhruba 0,35.



Obrázek 4.1: CDF exponenciálního rozdělení.

Jako příklad rozdělení, které je přibližně exponenciální, se podíváme na časový interval mezi jednotlivými narozeními. 18. prosince 1997 se v porodnici v Brisbane v Austrálii[1] narodilo 44 dětí. Časy narození všech 44 dětí byly zveřejněny v místních novinách. Data si můžete stáhnout z `http://thinkstats.com/babyboom.dat`.

Obrázek 4.2 ukazuje CDF časových intervalů mezi narozeními v minutách. Zdá se, že má obecný tvar exponenciálního rozdělení, ale jak toto rozdělení poznáme?

Jednou z možností je graficky znázornit komplementární CDF, 1 - CDF(x) pomocí logaritmické stupnice na ose y. Pro data z exponenciálního rozdělení je výsledkem přímka. Podívejme se na to, proč toto funguje.

Pokud graficky znázorníte komplementární CDF (CCDF) souboru dat, u kterého předpokládáte exponenciální charakter, pak očekáváte funkci jako:

$$y \approx e^{-\lambda x}$$

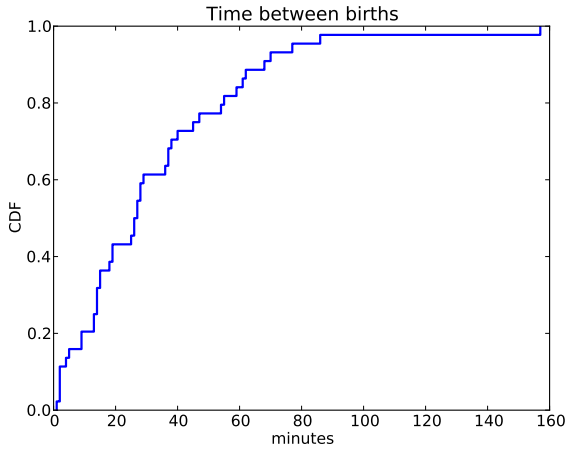
Jestliže použijeme logaritmickou stupnici na obou stranách, dostaneme:

$$\log y \approx -\lambda x$$

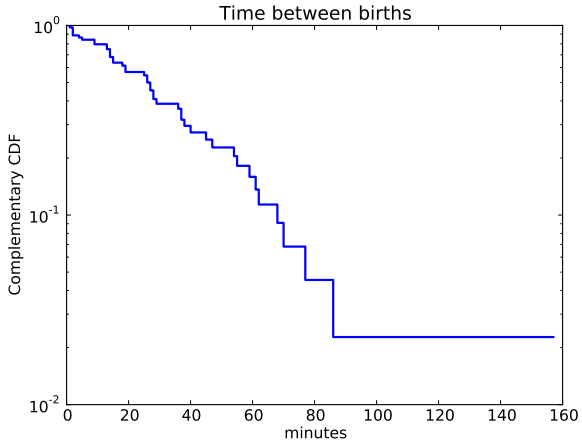
Takže na logaritmické stupnici na ose y je CCDF přímka se sklonem $-\lambda$.

Obrázek 4.3 ukazuje CCDF časových intervalů mezi narozeními na logaritmické stupnici na ose y. Nejedná se zcela o přímku, což naznačuje, že

[1] Tento příklad vychází z informací a dat uvedených v Dunn, „A Simple Dataset for Demonstrating Common Distributions,“ Journal of Statistics Education v.7, n.3 (1999).



Obrázek 4.2: CDF časových intervalů mezi narozeními.



Obrázek 4.3: CCDF časových intervalů mezi narozeními.

exponenciální rozdělení je pouhou aproximací. Předpoklad, ze kterého vycházíme, – že narození je stejně pravděpodobné v kterýkoliv okamžik dne – neodpovídá s největší pravděpodobností zcela skutečnosti.

Cvičení 4.1 U malých hodnot n neočekáváme, že se empirické rozdělení bude přesně shodovat se spojitým rozdělení. Jedním ze způsobů, jak vyhodnotit kvalitu souladu, je vytvořit výběr ze spojitého rozdělení a sledovat, jak dobře odpovídá datům.

Funkce `expovariate` v modulu `random` generuje náhodné hodnoty z exponenciálního rozdělení se stanovenou hodnotou λ. Použijte ji k vygenerování 44 hodnot z exponenciálního rozdělení s průměrem 32,6. Graficky znázorněte CCDF na logaritmické stupnici na ose y a porovnejte ji s Obrázkem 4.3.

Nápověda: Ke grafickému znázornění pomocí logaritmické stupnice na ose y můžete použít funkci `pyplot.yscale`.

Nebo jestliže použijete `myplot`, funkce `Cdf` přijme booleovskou hodnotu parametru `complement`, která určí, zda graficky znázorňovat CDF nebo CCDF, a řetězcové hodnoty parametrů `xscale` a `yscale`, které definují, jak budou osy vypadat. Pro grafické znázornění CCDF pomocí logaritmické stupnice na ose y:

```
myplot.Cdf(cdf, complement=True, xscale='linear', yscale='log')
```

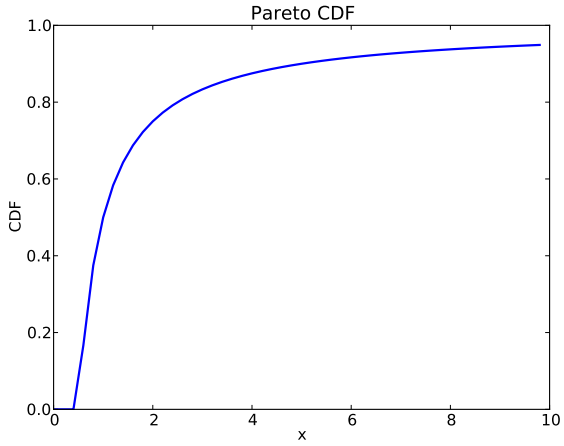
Cvičení 4.2 Shromážděte údaje o datech narození studentů ve vašem kurzu, setřiďte je a vypočítejte časový interval mezi daty narození ve dnech. Graficky znázorněte CDF časových intervalů mezi daty narození a CCDF na logaritmické stupnici na ose y. Vypadá to jako exponenciální rozdělení?

4.2 Paretovo rozdělení

Paretovo rozdělení je pojmenováno po ekonomovi Vilfredu Paretovi, který jej použil k popisu rozdělení bohatství (viz `http://wikipedia.org/wiki/Pareto_distribution`). Od té doby se používá k popisu jevů v přírodních i společenských vědách, včetně velikosti měst a obcí, částic prachu a meteoritů, lesních požárů a zemětřesení.

CDF Paretova rozdělení je následující:

$$CDF(x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha}$$



Obrázek 4.4: CDF Paretova rozdělení.

Parametry x_m a α určují umístění a tvar rozdělení. x_m představuje minimální možnou hodnotu. Obrázek 4.4 ukazuje CDF Paretova rozdělení s parametry $x_m = 0{,}5$ a $\alpha = 1$.

Medián tohoto rozdělení je $x_m 2^{1/\alpha}$, tedy 1, ale 95. percentil je 10. Naproti tomu v případě exponenciálního rozdělení s mediánem 1 má 95. percentil hodnotu pouze 1,5.

Existuje jednoduchý vizuální test k určení toho, zda se empirické rozdělení shoduje s Paretovým rozdělením: Pokud pro obě osy použijeme logaritmickou stupnici, CCDF má podobu přímky. Jestliže graficky znázorníte CCDF výběru z Paretova rozdělení na lineární stupnici, očekáváte funkci, která bude vypadat takto:

$$y \approx \left(\frac{x}{x_m}\right)^{-\alpha}$$

Když použijeme logaritmickou stupnici na obou stranách, získáme:

$$\log y \approx -\alpha \, (\log x - \log x_m)$$

Jestliže tedy provedete grafické znázornění $\log y$ versus $\log x$, výsledkem by měla být přímka se sklonem $-\alpha$ a konstantou $\alpha \log x_m$.

Cvičení 4.3 Modul `random` nabízí funkci `paretovariate`, která generuje náhodné hodnoty z Paretova rozdělení. Přijme parametr pro α, ale nikoliv x_m. Implicitní hodnota x_m je 1; rozdělení s jiným parametrem můžete vygenerovat vynásobením hodnotou x_m.

Napište wrapper funkci s názvem `paretovariate`, která přijme α a x_m jako

parametry použije `random.paretovariate` k vygenerování hodnot z dvouparametrického Paretova rozdělení.

Použijte svoji funkci k vygenerování výběru z Paretova rozdělení. Vypočtěte CCDF a graficky ji znázorněte pomocí logaritmické stupnice na obou osách. Je výsledkem přímka? Jaký je sklon?

Cvičení 4.4 Abyste získali určitý cit pro Paretova rozdělení, představte si, jak by vypadal svět, pokud by rozdělení výšek lidí mělo podobu Paretova rozdělení. Na základě parametrů $x_m = 100$ cm a $\alpha = 1{,}7$ získáme rozdělení s přiměřeným minimem 100 cm a mediánem 150 cm.

Vygenerujte 6 miliard náhodných hodnot z tohoto rozdělení. Jaký je průměr tohoto vzorku? Jaká část populace je nižší než průměr? Jakou výšku má nejvyšší člověk ve světě, kterému vládne Paretovo rozdělení?

Cvičení 4.5 Zipfův zákon je založen na pozorování četnosti užití různých slov. Nejběžnější slova mají velmi vysokou četnost. Na druhé straně existuje řada neobvyklých slov, jako například „hapax legomenon“, která se vyskytují jen několikrát. Zipfův zákon předpokládá, že v určitém souboru textů označovaném jako „korpus“ bude rozdělení četnosti výskytu slov zhruba odpovídat Paretovu rozdělení.

Najděte velký korpus pro jakýkoliv jazyk, který je dostupný v elektronické podobě. Spočítejte, kolikrát se v něm vyskytují jednotlivá slova. Zjistěte CCDF četnosti výskytu slov a graficky ji znázorněte pomocí logaritmické stupnice na obou osách. Platí Zipfův zákon? Jaká je, přibližně, hodnota α?

Cvičení 4.6 Weibullovo rozdělení je zobecněním exponenciálního rozdělení, které se vyskytuje v analýze poruch (viz `http://wikipedia.org/wiki/Weibull_distribution`). CDF tohoto rozdělení je:

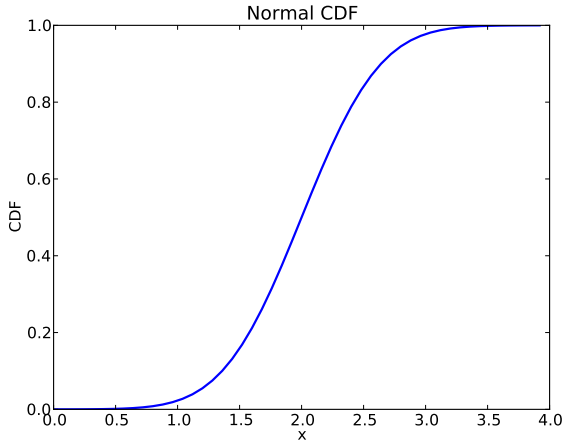
$$CDF(x) = 1 - e^{-(x/\lambda)^k}$$

Najdete transformaci, jejíž pomocí získá Weibullovo rozdělení podobu přímky? Co naznačují směrnice a konstantní člen/konstanta přímky?

Použijte `random.weibullvariate` k vygenerování výběru z Weibullova rozdělení a použijte jej k otestování vaší transformace.

4.3 Normální rozdělení

Normální rozdělení, nazývané také Gaussovo, se používá nejčastěji, protože jeho prostřednictvím lze popsat velké množství jevů, tedy alespoň při-



Obrázek 4.5: CDF normálního rozdělení.

bližně. Ukazuje se, že pro jeho všudypřítomnost existuje velmi dobrý důvod, k němuž se dostaneme v Oddílu 6.6.

Normální rozdělení se vyznačuje řadou vlastností, které jej činí vhodným pro analýzu, ale CDF k nim nepatří. Na rozdíl od ostatních rozdělení, jimiž jsme se zde zabývali, pro normální distribuční funkci neexistuje žádné vyjádření s uzavřenou formou; nejčastěji využívanou alternativou je její zápis prostřednictvím **chybové funkce**, což je speciální funkce zapsaná jako erf(x):

$$CDF(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$

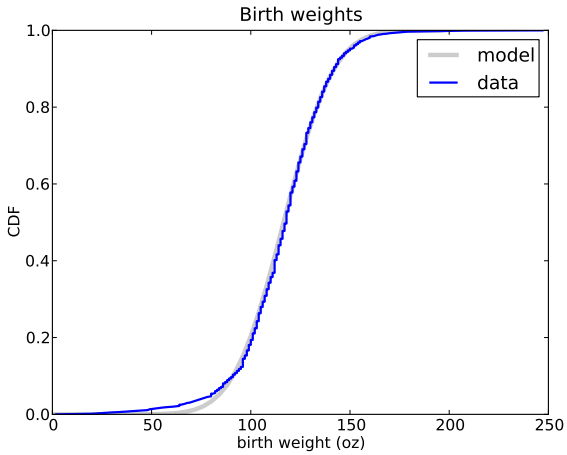
Parametry μ a σ určují průměr a směrodatnou odchylku rozdělení.

Jestli vás z těchto vzorců bolí oči, nezoufejte, protože je snadné je provést v Pythonu[2]. Existuje celá řada rychlých a přesných způsobů, jak přiblížit erf(x). Jeden z nich si můžete stáhnout na `http://thinkstats.com/erf.py`, kde najdete funkce nazvané `erf` a `NormalCdf`.

Obrázek 4.5 ukazuje CDF normálního rozdělení s parametry $\mu = 2{,}0$ a $\sigma = 0{,}5$. Signoidní tvar křivky je znakem, podle něhož je možné poznat normální rozdělení.

V předchozí kapitole jsme se zabývali rozdělením porodních hmotností v souboru dat NSFG. Obrázek 4.6 ukazuje empirickou CDF hmotností všech

[2]Od verze Python 3.2 je to dokonce ještě snazší, protože `erf` je obsažena v modulu `math`.



Obrázek 4.6: CDF porodních hmotností s normálním modelem.

živě narozených dětí a CDF normálního rozdělení se stejným průměrem a rozptylem.

Normální rozdělení je dobrým modelem pro tento soubor dat. **Model** představuje užitečné zjednodušení. V tomto případě je užitečný, protože celé rozdělení můžeme shrnout pomocí pouhých dvou čísel $\mu = 116{,}5$ a $\sigma = 19{,}9$, přičemž výsledná chyba (rozdíl mezi modelem a daty) je malá.

Pod 10. percentilem dochází k rozporu mezi daty a modelem; je zde více dětí s nízkou hmotností, než bychom očekávali v normálním rozdělení. Pokud chceme studovat předčasně narozené děti, je důležité zachytit tuto část rozdělení správně. Z tohoto důvodu by normální model nemusel být vhodný.

Cvičení 4.7 Wechslerova škála inteligence dospělých je test, který je určen k měření inteligence[3]. Výsledky jsou transformovány tak, že rozdělení skóre v obecné populaci je normální s $\mu = 100$ a $\sigma = 15$.

Použijte `erf.NormalCdf` k prozkoumání četnosti jevů, které se v normálním rozdělení vyskytují řídce. Jaká část populace má IQ vyšší než průměr? Jaká část má IQ nad 115? 130? 145?

Jev, který můžeme označit jako „six sigma“, je hodnota přesahující průměr o 6 směrodatných odchylek, takže six sigma IQ je 190. Kolik předpokládáte, že na světě se 6 miliardami lidí žije jedinců s IQ 190 nebo vyšším[4]?

[3]Kolem otázky, zda ji opravdu měří či neměří, se vedou fascinující spory, jejichž zkoumání stojí za to věnovat svůj čas.

[4]V souvislosti s tímto tématem by vás mohlo zajímat toto: `http://wikipedia.org/wiki/Christopher_Langan`.

Cvičení 4.8 Graficky znázorněte CDF délek těhotenství pro všechny živě narozené děti. Má výsledek podobu normálního rozdělení?

Vypočtěte průměr a směrodatnou odchylku výběru a graficky znázorněte normální rozdělení se stejnými parametry. Představuje normální rozdělení dobrý model pro tato data? Pokud byste měli shrnout toto rozdělení dvěma statistickými charakteristikami, které byste zvolili?

4.4 Normální pravděpodobnostní graf

Pro exponenciální, Paretovo a Weibullovo rozdělení existují jednoduché transformace, které můžeme použít, chceme-li otestovat, jestli spojité rozdělení představuje vhodný model pro konkrétní soubor dat.

Pro normální rozdělení žádná takováto transformace k dispozici není, ale existuje alternativní možnost označovaná jako **normální pravděpodobnostní graf**. Ten je založen na **rankits**: jestliže vygenerujete n hodnot z normálního rozdělení a setřídíte je, pak k-tý rankit je průměr rozdělení pro k-tou hodnotu.

Cvičení 4.9 Napište funkci s názvem `Sample`, která vygeneruje 6 vzorků z normálního rozdělení, kde $\mu = 0$ a $\sigma = 1$. Setřiďte a vraťte hodnoty.

Napište funkci s názvem `Samples`, která volá `Sample` 1000x a vrátí seznam 1000 seznamů.

Jestliže na tento seznam seznamů aplikujete funkci `zip`, výsledkem bude 6 seznamů, každý o 1000 hodnotách. Vypočtěte průměr každého z těchto seznamů a vytiskněte výsledky. Můj odhad je, že vám vyjde něco jako:

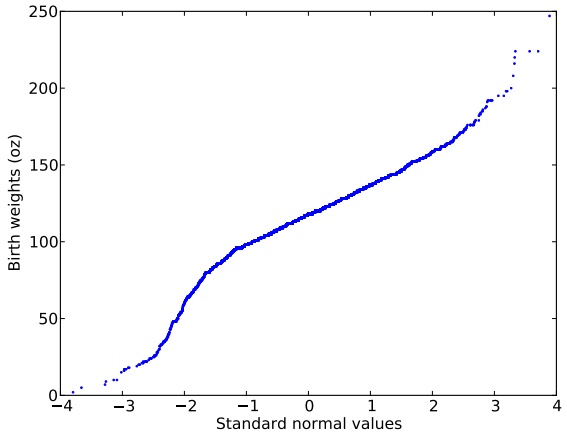
{-1,2672, -0,6418, -0,2016, 0,2016, 0,6418, 1,2672}

Pokud zvýšíte počet volání funkce `Sample`, výsledky by měly konvergovat k těmto hodnotám.

Přesně vypočíst rankits je mírně obtížné, ale existují numerické metody, jak je přiblížit. A existuje také rychlá a hrubá metoda, kterou je ještě jednodušší použít:

1. Z normální distribuce, kde $\mu = 0$ a $\sigma = 1$, vytvořte výběr o stejném rozsahu jako váš soubor dat a setřiďte jej.

2. Setřiďte hodnoty v souboru dat.



Obrázek 4.7: Normální pravděpodobnostní graf porodních hmotností.

3. Graficky znázorněte setříděné hodnoty z vašeho souboru dat versus náhodné hodnoty.

Tato metoda funguje dobře pro velké soubory dat. Pro menší soubory dat je možné ji vylepšit vygenerováním $m(n+1)$ - 1 hodnot z normálního rozdělení, kde n je rozsah souboru dat a m je multiplikátor. Pak vyberte každý m-tý prvek, počínaje m-tým.

Tato metoda funguje také pro další rozdělení za předpokladu, že víte, jak vygenerovat náhodný výběr.

Obrázek 4.7 představuje rychlý a hrubý normální pravděpodobnostní graf pro data porodních hmotností.

Zakřivení tohoto grafu naznačuje určité odchylky od normálního rozdělení, ale přesto se jedná o (dostatečně) dobrý model pro celou řadu účelů.

Cvičení 4.10 Napište funkci s názvem `NormalPlot`, která přijme řadu hodnot a vygeneruje normální pravděpodobnostní graf. Řešení si můžete stáhnout z `http://thinkstats.com/rankit.py`.

Použijte běžecké rychlosti z `relay.py` k vygenerování normálního pravděpodobnostního grafu. Je normální rozdělení dobrým modelem pro tato data? Řešení si můžete stáhnout z `http://thinkstats.com/relay_normal.py`.

4.5 Logaritmicko-normální rozdělení

Jestliže logaritmy množiny hodnot vykazují normální rozdělení, tyto hodnoty mají **logaritmicko-normální** (také **lognormální**) rozdělení. CDF logaritmicko-normálního rozdělení je stejné jako CDF normálního rozdělení, kde x je nahrazeno $\log x$.

$$\mathrm{CDF}_{lognormal}(x) = \mathrm{CDF}_{normal}(\log x)$$

Parametry logaritmicko-normálního rozdělení se obvykle označují jako μ a σ. Nezapomeňte ale, že tyto parametry *nejsou* průměr a směrodatná odchylka. Průměr logaritmicko-normálního rozdělení je $\exp(\mu + \sigma^2/2)$ a směrodatná odchylka je nepěkná[5].

Ukazuje se, že rozdělení hmotností dospělých je přibližně logaritmicko-normální [6].

Národní centrum pro prevenci chronických onemocnění a podporu zdraví (National Center for Chronic Disease Prevention and Health Promotion) provádí každoroční šetření jako součást Systému sledování rizikových faktorů chování (Behavioral Risk Factor Surveillance System – BRFSS)[7]. V roce 2008 se do šetření zapojilo 414 509 respondentů, kteří byli dotázáni na jejich demografické údaje, zdraví a zdravotní rizika.

Shromážděná data zahrnují hmotnosti 398 484 respondentů vyjádřené v kilogramech. Obrázek 4.8 ukazuje rozdělení log ; kde ˚ je hmotnost v kilogramech, spolu s normálním modelem.

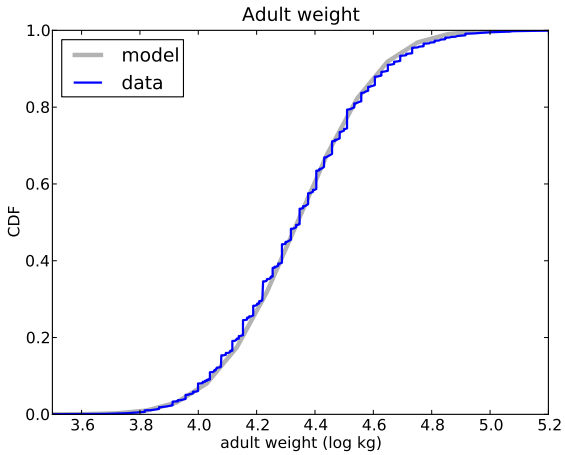
Normální model vyjadřuje dobrou shodu s daty, přestože nejvyšší hmotnosti přesahují to, co bychom očekávali od normálního modelu dokonce i po provedení logaritmické transformace. Protože rozdělení log ˚ odpovídá normálnímu rozdělení, vede nás to k závěru, že ˚ odpovídá logaritmicko-normálnímu rozdělení.

Cvičení 4.11 Stáhněte si data z BRFSS na `http://thinkstats.com/CDBRFS08.ASC.gz` a můj kód vytvořený k jejich přečtení na `http://`

[5] Viz `http://wikipedia.org/wiki/Log-normal_distribution`.

[6] Na tuto možnost mě upozornil komentář (bez citace) na `http://mathworld.wolfram.com/LogNormalDistribution.html`. Následně jsem našel článek, ve kterém je navržena logaritmická transformace a také je zde naznačena příčina: Penman and Johnson, „The Changing Shape of the Body Mass Index Distribution Curve in the Population,“ Preventing Chronic Disease, 2006 July; 3(3): A74. Online at `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1636707`.

[7] Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2008.



Obrázek 4.8: CDF hmotností dospělých (logaritmická transformace).

`thinkstats.com/brfss.py`. Spusťte `brfss.py` a potvrďte tisk souhrnných statistických charakteristik pro několik proměnných.

Napište program, který přečte hmotnosti dospělých z BRFSS a vygeneruje normální pravděpodobnostní grafy pro ° a log ♀ Řešení si můžete stáhnout z `http://thinkstats.com/brfss_figs.py`.

Cvičení 4.12 Rozdělení počtu obyvatel pro města a obce bylo uvedeno jako příklad jevu z reálného světa, který lze popsat pomocí Paretova rozdělení.

Americký úřad pro sčítání lidu (U.S. Census Bureau) zveřejňuje údaje o počtu obyvatel každého města a obce ve Spojených státech, které má vlastní samosprávu. Vytvořil jsem malý program, který si stáhne tato data a uloží je do souboru. Můžete si jej stáhnout na `http://thinkstats.com/populations.py`.

1. Pročtěte si tento program, ať víte, co umí. Pak jej spusťte, abyste stáhli a zpracovali data.

2. Napište program, který spočítá a graficky znázorní rozdělení počtu obyvatel 14 593 měst a obcí v souboru dat.

3. Graficky znázorněte CDF na lineární stupnici a logaritmické stupnici na ose x, tak abyste získali představu o tvaru tohoto rozdělení. Pak graficky znázorněte CCDF na logaritmické stupnici pro obě osy, abyste zjistili, jestli má tvar charakteristický pro Paretovo rozdělení.

4. Vyzkoušejte ostatní transformace a grafy v této kapitole, abyste zjistili, jestli pro tato data existuje lepší model.

K jakému závěru jste dospěli ohledně rozdělení velikosti měst a obcí? Řešení si můžete stáhnout z `http://thinkstats.com/populations_cdf.py`.

Cvičení 4.13 Daňová správa USA (The Internal Revenue Service of the United States – IRS) uvádí data o daních z příjmů na `http://irs.gov/taxstats`.

Jeden ze souborů tohoto úřadu obsahující údaje o příjmech fyzických osob za rok 2008 je dostupný na `http://thinkstats.com/08in11si.csv`. Převedl jsem jej do textového formátu označovaného zkratkou CSV, což znamená „hodnoty oddělené čárkou“ (comma-separated values). Můžete si jej přečíst pomocí modulu `csv`.

Extrahujte rozdělení příjmů z tohoto souboru dat. Představuje některé ze spojitých rozdělení uvedených v této kapitole dobrý model pro tato data? Řešení si můžete stáhnout z `http://thinkstats.com/irs.py`.

4.6 Proč model?

Na začátku této kapitoly jsem řekl, že spojitá rozdělení lze použít k modelování mnoha jevů reálného světa. „Takže,“ mohli byste se zeptat, „kterých?“

Stejně jako všechny modely, jsou také spojitá rozdělení abstrakcí, což znamená, že vynechávají určité detaily, jež nejsou považovány za důležité. Například pozorované rozdělení může obsahovat chyby měření nebo zvláštnosti specifické pro daný výběr. Spojité modely přitom tyto idiosynkrazie vyhlazují.

Spojité modely jsou zároveň formou komprese dat. Jestliže nějaký model dobře přiléhá k souboru dat, pak je možné pomocí malé množiny parametrů souhrnně popsat velký objem dat.

Někdy může být skutečnost, že data týkající se určitého přírodního jevu odpovídají spojitému rozdělení, překvapivá, avšak takováto pozorování mohou vést k novým vhledům do fyzických systémů. Někdy jsme schopni vysvětlit, proč pozorované rozdělení má určitou formu. Například Paretovo rozdělení je často výsledkem generativních procesů s kladnou zpětnou vazbou (tzv. procesů preferenčního připojování (preferential attachment processes): viz `http://wikipedia.org/wiki/Preferential_attachment`.).

Spojitá rozdělení jsou vhodná pro matematickou analýzu, jak si ukážeme v Kapitole 6.

4.7 Generování náhodných čísel

Spojité CDFs jsou také vhodné pro generování náhodných čísel. Jestliže existuje efektivní způsob, jak vypočíst inverzní CDF, ICDF(p), můžeme vygenerovat náhodné hodnoty s vhodným rozdělením na základě výběru z rovnoměrného rozdělení od 0 do 1, a následnou volbou

x = ICDF(p)

Například CDF exponenciálního rozdělení je

$$p = 1 - e^{-\lambda x}$$

Řešením pro x získáme:

x= -log (1 - p) / λ

V Pythonu tedy můžeme napsat

```
def expovariate(lam):
    p = random.random()
    x = -math.log(1-p) / lam
    return x
```

Parametr jsem nazval `lam`, protože `lambda` je klíčové slovo v Pythonu. Většina provedení funkce `random.random` může vrátit 0, ale ne 1, takže 1 - p může nabývat hodnotu 1, ale ne 0, což je dobré, protože log 0 není definovaný.

Cvičení 4.14 Napište funkci s názvem `weibullvariate`, která přijme `lam` a `k` a vrátí náhodnou hodnotu z Weibullova rozdělení s uvedenými parametry.

4.8 Glosář

empirické rozdělení (empirical distribution): Rozdělení hodnot ve výběru.

spojité rozdělení (continuous distribution): Rozdělení popsané prostřednictvím spojité funkce.

časový interval mezi výskytem jevů (interarrival time): Čas, který uplyne mezi výskytem dvou jevů.

chybová funkce (error function): Speciální matematická funkce, jejíž název je odvozen od toho, že se vyskytuje při studiu chyb měření.

normální pravděpodobnostní graf (normal probability plot): Graf setříděných hodnot ve výběru versus očekávaná hodnota každé z nich za předpokladu jejich normálního rozdělení.

rankit: Očekávaná hodnota prvku v setříděném seznamu hodnot z normálního rozdělení.

model: Užitečné zjednodušení. Spojitá rozdělení často představují dobrý model složitějších empirických rozdělení.

korpus (corpus): Soubor textů sloužící jako vzorek jazyka.

hapax legomenon (hapaxlegomenon): Slovo, které se v korpusu vyskytuje pouze jednou. V této knize má, zatím, dva výskyty.

Kapitola 5

Pravděpodobnost

V kapitole 2 jsem uvedl, že pravděpodobnost je četnost vyjádřená jako podíl rozsahu výběru. To je jedna definice pravděpodobnosti, ale není jediná. Vlastně se dá říci, že otázka definování pravděpodobnosti je do jisté míry sporná.

Začneme proto částmi, které nejsou kontroverzní, a postupně se budeme propracovávat dále. Panuje obecná shoda na tom, že pravděpodobnost je reálná hodnota mezi 0 a 1 a že má sloužit jako kvantitativní ukazatel odpovídající kvalitativnímu uchopení skutečnosti, že některé věci jsou pravděpodobnější než jiné.

„Věci“, kterým přidělujeme pravděpodobnosti, se označují jako **jevy**. Jestliže E představuje určitý jev, pak $P(E)$ představuje pravděpodobnost výskytu E. Situace, kdy E může a nemusí nastat, se nazývá **pokus**.

Jako příklad si představte, že máte standardní kostku o šesti stranách a chcete znát pravděpodobnost toho, že vám padne 6. Každý hod představuje pokus. Pokaždé, když padne 6, je to považováno za **úspěch**; ostatní pokusy jsou považovány za **neúspěch**. Tyto pojmy se používají i v situacích, kdy „úspěch“ je špatný a „neúspěch“ je dobrý.

Máme-li konečný počet n pokusů a zaznamenáme s úspěchů, pak pravděpodobnost úspěchu je s/n. Je-li množina pokusů nekonečná, je definování pravděpodobností o něco ošemetnější, ale většina lidí je ochotná akceptovat pravděpodobnostní tvrzení o hypotetické řadě identických pokusů, jako například hod mincí nebo vrh kostkou.

Potíže nastávají ve chvíli, kdy hovoříme o pravděpodobnostech jedinečných jevů. Mohli bychom například chtít znát pravděpodobnost toho, že

určitý kandidát zvítězí ve volbách. Každé volby jsou ale jedinečné, a tak neexistuje žádná řada identických pokusů, o kterých bychom mohli uvažovat.

O takovýchto případech někteří lidé tvrdí, že se na ně pravděpodobnost nevztahuje. Tento postoj se někdy označuje jako **frekventistický přístup**, protože definuje pravděpodobnost na základě četnosti. Jestliže neexistuje žádná množina identických pokusů, nemůžeme mluvit ani o pravděpodobnosti.

Frekventistický přístup je z filozofického hlediska bezpečný, ale zároveň je frustrující, protože omezuje oblast pravděpodobnosti na fyzické systémy, které jsou buď náhodné (jako například rozpad atomů), nebo natolik nepředvídatelné, že je modelujeme jako náhodné (například kutálející se kostka). Cokoliv, co souvisí s lidmi, je do značné míry mimo hru.

Alternativou je **bayesovský přístup**, který pravděpodobnost definuje jako stupeň přesvědčení o tom, že nastane určitý jev. Z této definice vyplývá, že pojem pravděpodobnosti může být aplikován prakticky na jakoukoliv situaci. Určitá potíž s bayesovským pojetím pravděpodobnosti spočívá v tom, že závisí na stavu poznání konkrétního člověka. Lidé s různými informacemi mohou být o stejném jevu přesvědčeni do různé míry. Z tohoto důvodu se mnoho lidí domnívá, že pravděpodobnost v bayesovském pojetí je subjektivnější než pravděpodobnost založená na četnosti.

Jako příklad si můžeme položit otázku: Jaká je pravděpodobnost, že Thaksin Shinawatra je thajským premiérem? Zastánce frekventistického přístupu by řekl, že pro tento jev neexistuje žádná pravděpodobnost, protože neexistuje žádná množina pokusů. Thaksin buď je, nebo není premiér; není to otázka pravděpodobnosti.

Naproti tomu stoupenec bayesovského přístupu by byl ochoten přiřadit tomuto jevu určitou pravděpodobnost na základě znalostí, kterými disponuje. Například jestliže si vzpomínáte, že v Thajsku došlo roku 2006 k převratu, a jste si téměř jistí, že Thaksin byl v té době premiérem, který byl sesazen, mohli byste přidělit pravděpodobnost např. 0,1, která zohledňuje možnost, že si tuto událost nepamatujete správně, nebo že se Thaksin opět vrátil do funkce.

Pokud se podíváte na Wikipedii, zjistíte, že Thaksin není thajským premiérem (v okamžiku, kdy toto píšu). Na základě těchto informací můžete revidovat svůj odhad pravděpodobnosti na 0,01, který bere do úvahy možnost, že se Wikipedie mýlí.

5.1 Pravidla pravděpodobnosti

V rámci pravděpodobnosti založené na četnosti můžeme odvodit pravidla pro vyjádření vztahu mezi pravděpodobností různých jevů. Zřejmě nejznámější z těchto pravidel je

$P(A \text{ and } B) = P(A)\, P(B)$ Pozor: Ne vždy je to pravda!

kde $P(A \text{ and } B)$ je pravděpodobnost, že nastanou oba jevy, A i B . Tento vzorec se snadno pamatuje. Jediný problém je, že *ne vždy je to pravda*. Tento vzorec platí pouze, jestliže A a B jsou **nezávislé**, což znamená, že vím-li, že nastal jev A, nemá to vliv na pravděpodobnost toho, že nastane B, a naopak.

Například, jestliže A je, že mi při hodu mincí padne panna, a B je, že mi při vrhu kostkou padne 1, pak A a B jsou nezávislé jevy, protože hod mincí mi neřekne nic o vrhu kostkou.

Jestliže ale vrhnu dvě kostky a A je to, že mi padne alespoň jedna šestka, a B je to, že mi padnou dvě šestky, pak A a B nejsou nezávislé jevy, protože vím-li, že nastal jev A, zvyšuje se pravděpodobnost jevu B a vím-li, že nastal jev B, pak je pravděpodobnost A rovna 1.

V případě, že A a B nejsou nezávislé, je často užitečné vypočítat podmíněnou pravděpodobnost, $P(A \mid B)$, což je pravděpodobnost, že nastane A za předpokladu, že víme, že nastalo B:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Z toho můžeme odvodit obecný vztah

$P(A \text{ and } B) = P(A)\, P(B \mid A)$

Toto už asi nebude tak snadno zapamatovatelné, ale když si to přeložíte do češtiny, tak by to mělo dávat smysl: „Pravděpodobnost, že nastanou obě věci, je pravděpodobnost, že nastane první z nich a pak druhá, za předpokladu té první.“

Na pořadí jevů nijak zvlášť nezáleží, takže bychom mohli napsat také

$P(A \text{ and } B) = P(B)\, P(A \mid B)$

Tyto vztahy platí, bez ohledu na to, zda jsou A a B nezávislé či nikoliv. Jsou-li nezávislé, pak $P(A \mid B) = P(A)$, čímž se dostáváme tam, kde jsme začali.

Protože všechny pravděpodobnosti se pohybují od 0 do 1, je snadné ukázat, že

$$P(A \text{ and } B) \leq P(A)$$

Pro lepší představu uvažujme, že nějaký klub přijímá pouze osoby, které splní určitý požadavek A. Nyní předpokládejme, že přidali nový požadavek na členství, který označíme jako B. Zdá se zřejmé, že počet členů klubu se zmenší, nebo zůstane stejný, pokud všichni členové splní podmínku B. Přesto ale existují situace, kdy si lidé při tomto typu analýzy vedou překvapivě špatně. Příklady a diskuse ohledně tohoto jevu viz `http://wikipedia.org/wiki/Conjunction_fallacy`.

Cvičení 5.1 Jestliže hodím dvě kostky a součet hodnot je 8, jaká je pravděpodobnost, že jeden z hodů je 6?

Cvičení 5.2 Jestliže hodím 100 kostek, jaká je pravděpodobnost, že mi padnou samé šestky? Jaká je pravděpodobnost, že nepadne ani jedna šestka?

Cvičení 5.3 Následující otázky jsou převzaty z Mlodinow, *The Drunkard's Walk*.

1. Jestliže má rodina dvě děti, jaká je pravděpodobnost, že má dvě děvčata?
2. Jestliže má rodina dvě děti a víme, že alespoň jedno z nich je děvče, jaká je pravděpodobnost, že mají dvě děvčata?
3. Jestliže má rodina dvě děti a víme, že starší z nich je děvče, jaká je pravděpodobnost, že mají dvě děvčata?
4. Jestliže má rodina dvě děti a víme, že alespoň jedno z nich je děvče, která se jmenuje Florida, jaká je pravděpodobnost, že mají dvě děvčata?

Můžete předpokládat, že pravděpodobnost toho, že dítě je ženského pohlaví, je 1/2 a že děti v rodině jsou nezávislé pokusy (více než v jednom smyslu). Můžete také předpokládat, že procento děvčat, která se jmenují Florida, je malé.

5.2 Monty Hall

Monty Hallův problém je dobrým adeptem na nejkontroverznější otázku v historii pravděpodobnosti. Scénář je velmi jednoduchý, ale správná odpověď jde natolik proti přirozené intuici, že ji spousta lidí prostě nedokáže

přijmout a mnoho inteligentních lidí si dokonce utrhlo ostudu nejen tím, že to nepochopili, ale ještě na veřejnosti s vervou obhajovali své nesprávné přesvědčení.

Monty Hall je jméno původního moderátora soutěžní show *Let's Make a Deal*. Monty Hallův problém se zakládá na jedné z pravidelných soutěží, které jsou součástí této televizní show. Pokud byste se účastnili této show, scénář by byl následující:

- Monty vám ukáže troje zavřené dveře a řekne vám, že za každými z nich je nějaká výhra: jednou z nich je auto a druhé dvě jsou méně hodnotné výhry v podobě arašídového másla a umělých nalepovacích nehtů. Umístění výher za dveřmi je náhodné.
- Cílem hry je uhodnout, za kterými dveřmi je auto. Když uhodnete, auto je vaše.
- Takže si vyberete některé dveře, které označíme jako dveře A. Zbylé dveře označíme jako dveře B a C.
- Před tím, než Monty otevře dveře, které jste si vybrali, rád ještě zvyšuje napětí tím, že otevře buď dveře B, nebo C, podle toho, za kterými není auto. (Pokud je auto za dveřmi A, může Monty s klidem otevřít dveře B nebo C, takže jedny z nich náhodně vybere).
- Pak vám Monty nabídne možnost, abyste buď zůstali u svojí původní volby, nebo si zvolili druhé dveře, které zůstaly zavřené.

Otázka zní, je lepší ponechat si, nebo změnit volbu, nebo je to jedno?

Většině lidí jejich intuice velmi důrazně napovídá, že je to jedno. Jejich argumentace je taková, že zbývají dvoje dveře, takže pravděpodobnost, že auto je za dveřmi A, je 50 %.

To je ale špatně. Ve skutečnosti je vaše šance na výhru, pokud zůstanete u dveří A, pouze 1/3; pokud svoji volbu změníte, vaše šance je 2/3. Vysvětlím proč, ale neočekávám, že mi budete věřit.

Klíčem k této úloze je uvědomit si, že existují tři možné scénáře: Auto je za dveřmi A, B, nebo C. Protože jsou výhry rozmístěny náhodně, pravděpodobnost každého ze scénářů je 1/3.

Pokud je vaše strategie ponechat si volbu dveří A, pak vyhrajete pouze ve scénáři A, jehož pravděpodobnost je 1/3.

Pokud zvolíte strategii změny volby, pak vyhrajete buď ve scénáři B, nebo C, takže celková pravděpodobnost výhry je 2/3.

Pokud vás tento argument zcela nepřesvědčil, jste v dobré společnosti. Když můj kamarád představil toto řešení Paulu Erdősovi, jeho reakce byla takováto: „Ne, to není možné. Neměl by v tom být žádný rozdíl.[1]“

Nenechal se přesvědčit žádnými argumenty. Nakonec ho přesvědčila až počítačová simulace.

Cvičení 5.4 Napište program, který simuluje Monty Hallův problém a použijte jej k odhadu pravděpodobnosti výhry, jestliže zůstanete u své volby a jestliže ji změníte.

Pak si přečtěte diskusi ohledně tohoto problému na `http://wikipedia.org/wiki/Monty_Hall_problem`.

Co vám připadá víc přesvědčivé, simulace nebo argumentace, a proč?

Cvičení 5.5 Pro porozumění Monty Hallovu problému je důležité uvědomit si, že tím, že Monty rozhodne, které dveře otevřít, vám dává informace. Abyste si uvědomili, proč na tom záleží, představte si případ, kdy Monty neví, kde je jaká výhra, a tak náhodně vybere dveře B nebo C.

Pokud otevře dveře, za nimiž se nachází auto, je po hře, vy jste prohrál/a a nedostanete šanci rozhodnout se, jestli si volbu ponecháte nebo ji změníte.

Jinak je ale výhodnější volbu změnit nebo zůstat u té původní?

5.3 Poincaré

Henri Poincaré byl francouzský matematik, který učil na Sorbonně kolem roku 1900. Následující historka o něm je pravděpodobně smyšlená, ale představuje zajímavý pravděpodobnostní problém.

Poincaré údajně podezříval místní pekárnu z toho, že prodává bochníky chleba, které mají ve skutečnosti nižší váhu, než uváděný 1 kg. A tak si každý den v roce koupil bochník chleba, přinesl jej domů a zvážil. Na konci roku graficky znázornil rozdělení svých výsledků a ukázalo se, že odpovídá normálnímu rozdělení s průměrem 950 g a směrodatnou odchylkou 50 g. Svá zjištění předložil chlebové policii, která udělila pekaři výstrahu.

[1] Viz Hoffman, *The Man Who Loved Only Numbers*, s. 83.

Další rok Poincaré pokračoval ve svém každodenním vážení chleba. Na konci roku zjistil, že průměrná váha byla 1 000 g, přesně jak by tomu mělo být, ale opět si stěžoval chlebové policii a ta tentokrát dala pekaři pokutu.

Proč? Protože tvar rozdělení byl asymetrický. Na rozdíl od normálního rozdělení bylo sešikmené doprava, což potvrzuje hypotézu, že pekař dál pekl bochníky chleba o 950 g, ale Poincarému dával záměrně ty těžší.

Cvičení 5.6 Napište program, který simuluje pekaře, který vybere n bochníků z rozdělení s průměrem 950 g a směrodatnou odchylkou 50 g a dá Poincarému ten nejtěžší. Při jaké hodnotě n získáme rozdělení s průměrem 1 000 g? Jaká je směrodatná odchylka?

Porovnejte toto rozdělení s normálním rozdělením se stejným průměrem a stejnou směrodatnou odchylkou. Je rozdíl ve tvaru rozdělení dostatečně velký na to, aby přesvědčil chlebovou policii?

Cvičení 5.7 Půjdete-li na taneční zábavu, kde jsou partneři přidělováni náhodně, jaké bude procento párů tvořených tanečníky opačného pohlaví, kde žena je vyšší než muž?

V souboru BRFSS (viz Oddíl 4.5) je rozdělení výšek přibližně normální s parametry $\mu = 178$ cm a $\sigma^2 = 59{,}4$ cm u mužů a $\mu = 163$ cm a $\sigma^2 = 52{,}8$ cm u žen.

Teď trochu odbočím, ale můžete si také všimnout, že směrodatná odchylka je větší u mužů, a můžete přemýšlet nad tím, jestli je výška mužů více variabilní. K porovnání variability mezi skupinami je vhodné vypočíst **variační koeficient**, což je směrodatná odchylka jako podíl průměru, σ/μ. Podle tohoto ukazatele vykazuje výška žen o něco větší variabilitu.

5.4 Další pravidlo pravděpodobnosti

Jestliže se dva jevy **vzájemně vylučují**, znamená to, že může nastat pouze jeden z nich, takže podmíněná pravděpodobnost je 0:

$$P(A \mid B) = P(B \mid A) = 0$$

V tomto případě je snadné vypočíst pravděpodobnost kteréhokoliv z jevů:

$$P(A \text{ or } B) = P(A) + P(B) \quad \text{Pozor: Ne vždy je to pravda.}$$

Nezapomeňte ale, že toto platí, jen pokud se oba jevy vzájemně vylučují. Obecně platí, že pravděpodobnost A nebo B nebo obou je:

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Důvod, proč musíme odečíst $P(A \text{ and } B)$, je ten, že bychom ji jinak započítali dvakrát. Například když hodím dvě mince, pravděpodobnost, že mi padne alespoň jednou orel je 1/2 + 1/2 - 1/4. Musím odečíst 1/4, protože jinak počítám možnost panna-panna dvakrát. Problém se ještě více osvětlí, když hodím tři mince.

Cvičení 5.8 Jestliže hodím dvě kostky, jaká je pravděpodobnost, že hodím alespoň jednu 6?

Cvičení 5.9 Jaký je obecný vzorec pro pravděpodobnost A nebo B ale ne obou?

5.5 Binomické rozdělení

Jestliže hodím 100 kostek, pak je pravděpodobnost, že mi padnou samé šestky $(1/6)^{100}$. A pravděpodobnost, že mi nepadne ani jedna šestka, je $(5/6)^{100}$.

Tyto případy jsou snadné, ale v obecnější rovině by nás mohlo zajímat, jaká je pravděpodobnost, že nám padne k šestek, pro všechny hodnoty k od 0 do 100. Odpovědí je **binomické rozdělení**, které má následující PMF:

$$\text{PMF}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

kde n je počet pokusů, p je pravděpodobnost úspěchu a k je počet úspěšných pokusů.

Binomický koeficient se čte „n nad k“ a lze jej přímo vypočíst takto:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Nebo rekurzivně takto:

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

se dvěma základními případy: jestliže $n = 0$, výsledek je 0; jestliže $k = 0$, výsledek je 1. Když si stáhnete `http://thinkstats.com/thinkstats.py`, uvidíte funkci s názvem `Binom`, která je efektivním nástrojem pro vypočet binomického koeficientu.

Cvičení 5.10 Jestliže hodíte minci 100krát, očekáváte, že vám padne panna zhruba 50krát, jaká je ale pravděpodobnost, že vám padne panna přesně 50krát?

5.6 Série a exponovaná místa

Lidé nemají příliš dobrou intuici, pokud jde o náhodné procesy. Když někoho požádáte o vygenerování „náhodných" čísel, výsledkem je obvykle náhodně vypadající řada čísel, která je ale více uspořádaná než skutečné náhodné řady. Naopak, když lidem ukážete skutečné náhodné řady, mají tendenci vidět vzory tam, kde nejsou.

Příkladem druhého jevu je fakt, že mnoho lidí věří na „série" ve sportu: o hráči, kterému se v poslední době daří, se říká, že má „šťastnou ruku," zatímco o hráči, kterému se nedaří, se říká, že má „smolnou sérii."

Statistici otestovali tyto hypotézy v řadě sportů a shodně dospěli k závěru, že nic takového jako série neexistuje[2]. Předpokládáte-li, že každý pokus je nezávislý na těch předešlých, budete moci vypozorovat občasné dlouhé řetězce úspěchů nebo neúspěchů. Tyto pozorované série však nejsou dostatečným důkazem o tom, že mezi po sobě jdoucími pokusy existuje nějaký vztah.

S tím souvisí také iluze seskupování, což je tendence vidět shluky v prostorových vzorech, které jsou ve skutečnosti náhodné (viz `http://wikipedia.org/wiki/Clustering_illusion`).

K otestování pravděpodobnosti toho, že pozorovaný shluk má nějaký význam, můžeme simulovat chování náhodného systému, abychom zjistili, nakolik je pravděpodobné, že vytvoří podobný shluk. Tento proces se nazývá simulace **Monte Carlo**, protože generování náhodných čísel připomíná hry v kasinu (a Monte Carlo se proslavilo právě svými kasiny).

Cvičení 5.11 Hraje-li v basketbalovém zápase 10 hráčů a každý z nich vystřelí v průběhu hry 15krát a každý výstřel má 50% pravděpodobnost, že padne koš, jaká je pravděpodobnost, že v daném zápase uvidíte alespoň jednoho hráče hodit 10 košů v řadě? Sledujete-li sezónu o 82 zápasech, jaká je pravděpodobnost, že uvidíte alespoň jednu sérii 10 vstřelených košů nebo neúspěšných střel?

[2]Viz například Gilovich, Vallone and Tversky, „The hot hand in basketball: On the misperception of random sequences," 1985.

Tento problém ukazuje některé silné a slabé stránky simulace Monte Carlo. Silnou stránkou je, že je často snadné a rychlé napsat simulaci, aniž by to vyžadovalo nějaké rozsáhlé znalosti pravděpodobnosti. Slabou stránkou je, že odhadnutí pravděpodobnosti vzácných jevů může trvat velmi dlouho! Trochu analýzy nám může ušetřit spoustu počítání.

Cvičení 5.12 V roce 1941 se Joeovi DiMaggio podařilo alespoň jednou úspěšně odpálit v řadě 56 zápasů za sebou[3]. Pro mnohé fanoušky baseballu je tato série tím největším úspěchem v historii sportu vůbec, protože dosažení takového výsledku bylo tolik nepravděpodobné.

Použijte simulaci Monte Carlo k odhadu pravděpodobnosti, že se v příštím století nějakému hráči v hlavní baseballové lize podaří úspěšně odpálit v 57 nebo více zápasech za sebou.

Cvičení 5.13 Centra pro kontrolu nemocí (Centers for Disease Control – CDC) definují zvýšený výskyt rakoviny (cancer cluster) jako „větší než očekávaný počet případů rakoviny, které se vyskytnou ve skupině lidí v určité zeměpisné oblasti v určitém časovém období.[4]“

Mnoho lidí si zvýšený výskyt rakoviny vysvětluje jako důkaz rizika spojeného s životním prostředím, avšak mnoho vědců a statistiků považuje zkoumání zvýšeného výskytu rakoviny za ztrátu času[5]. Proč? Jedním z (několika) důvodů je to, že identifikace zvýšeného výskytu rakoviny je klasický případ situace označované jako klam texaského ostrostřelce (viz `http://wikipedia.org/wiki/Texas_sharpshooter_fallacy`).

Nicméně jestliže někdo ohlásí zvýšený výskyt rakoviny, mají CDC povinnost to prověřit. Podle jejich webových stránek:

> „Vyšetřovatelé vypracují definici ‘případu’, určí příslušné časové období a populaci, která je riziku vystavena. Pak vypočítají očekávaný počet případů a porovnají jej s pozorovaným počtem. Zvýšený výskyt je potvrzen, jestliže poměr pozorovaných případů k očekávaným případům je větší než 1,0 a rozdíl je statisticky významný.“

1. Předpokládejme, že konkrétní typ rakoviny má výskyt 1 případ na tisíc obyvatel za rok. Budete-li sledovat konkrétní kohortu 100 lidí po dobu 10 let, budete očekávat přibližně 1 případ. Jestliže byste se setkali

[3] Viz `http://wikipedia.org/wiki/Hitting_streak`.

[4] Z `http://cdc.gov/nceh/clusters/about.htm`.

[5] Viz Gawande, „The Cancer Cluster Myth,“ *New Yorker*, Feb 8, 1997.

se dvěma případy, moc by vás to nepřekvapilo, ale výskyt většího počtu případů než dva už by byl velmi neobvyklý.

Napište program, který vytvoří simulaci velkého počtu kohort po dobu 10 let a odhadne rozdělení celkového počtu případů.

2. Pozorování je považováno za statisticky významné, jestliže jeho pravděpodobnost založená čistě na náhodě, označovaná jako p-hodnota, je nižší než 5 %. Uvažujeme-li kohortu o 100 osobách během období 10 let, kolik případů by se muselo vyskytnout, aby bylo splněno toto kritérium?

3. Teď si představte, že rozdělíte populaci 10 000 osob do 100 kohort a sledujete je po dobu 10 let. Jaká je pravděpodobnost, že alespoň jedna z těchto kohort vykáže „statisticky významný“ zvýšený výskyt? Co se stane, budeme-li požadovat, aby p-hodnota byla 1 %?

4. Teď si představte, že 10 000 lidí rozdělíte do mřížky 100 $\times$100 a budete je sledovat po dobu 10 let. Jaká je pravděpodobnost, že se vyskytne alespoň jeden blok 10 $\times$10 kdekoliv v mřížce se statisticky významným zvýšeným výskytem?

5. Nakonec si představte, že sledujete mřížku 10 000 lidí po dobu 30 let. Jaká je pravděpodobnost, že se kdykoliv v průběhu tohoto období na libovolném místě mřížky vyskytne 10letý interval s blokem 10 $\times$10 se statisticky významným zvýšeným výskytem?

5.7 Bayesova věta

Bayesova věta vyjadřuje vztah mezi podmíněnými pravděpodobnostmi dvou jevů. Podmíněná pravděpodobnost, často zapisovaná jako $P(A \mid B)$, je pravděpodobnost výskytu jevu A, za předpokladu, že víme, že nastal jev B. Bayesova věta zní:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Abychom se přesvědčili o její pravdivosti, pomůže nám, když zapíšeme $P(A \text{ and } B)$, což je pravděpodobnost toho, že nastane A a B

$$P(A \text{ and } B) = P(A)\ P(B \mid A)$$

Také ale platí

$$P(A \text{ and } B) = P(B)\ P(A \mid B)$$

Takže

$$P(B)\ P(A \mid B) = P(A)\ P(B \mid A)$$

Provedeme-li dělení $P(B)$, dostaneme Bayesovu větu[6].

Bayesova věta je často interpretována jako výrok o tom, jak soubor dokladů E ovlivňuje pravděpodobnost hypotézy H:

$$P(H|E) = P(H)\frac{P(E|H)}{P(E)}$$

Vyjádřeno slovy, tato rovnice říká, že pravděpodobnost H poté, co jste zaznamenali E, je součinem $P(H)$, což je pravděpodobnost H před tím, než jste zaznamenali důkaz, a poměru $P(E \mid H)$, tedy pravděpodobnosti zaznamenání důkazu za předpokladu, že H je pravdivá, a $P(E)$, tedy pravděpodobnosti zaznamenání důkazu za jakýchkoli okolností (bez ohledu na to, zda je H pravdivá či nikoliv).

Tento způsob čtení Bayesovy věty se označuje jako „diachronní" interpretace, protože popisuje to, jak je pravděpodobnost hypotézy v průběhu času **aktualizována**, obvykle ve světle získaných důkazů. V tomto kontextu se $P(H)$ označuje jako **apriorní** pravděpodobnost a $P(H \mid E)$ se označuje jako **aposteriorní** pravděpodobnost. $P(E \mid H)$ představuje **věrohodnost** důkazu a $P(E)$ je **normalizační konstanta**.

Klasickým případem užití Bayesovy věty je interpretace klinických testů. Například rutinní testování na užívání nelegálních drog je stále častěji používáno na pracovištích i ve školách (Viz `http://aclu.org/drugpolicy/testing`.). Společnosti, které provádí tyto testy, uvádí, že tyto testy jsou senzitivní, což znamená, že by měly vykázat pozitivní výsledek v případě, že je ve vzorku přítomna droga (nebo metabolity), a specifické, což znamená, že by měly vykázat negativní výsledek v případě nepřítomnosti drog.

Studie v Journal of the American Medical Association[7] odhadují, že senzitivita v případě běžných drogových testů je kolem 60 % a specificita se pohybuje kolem 99 %.

Teď předpokládejte, že jsou tyto testy použity na pracovní kolektiv, kde je skutečná míra užívání drog 5 %. Kolik ze zaměstnanců, kterým vyšel pozitivní test, opravdu užívá drogy?

[6] Viz `http://wikipedia.org/wiki/Q.E.D.`!

[7] Tato čísla jsem získal z Gleason and Barnum, „Predictive Probabilities In Employee Drug-Testing," at `http://piercelaw.edu/risk/vol2/winter/gleason.htm`.

V Bayesovském pojetí chceme vypočítat pravděpodobnost užívání drog v případě pozitivního testu, $P(D \mid E)$. Použijeme-li Bayesovu větu:

$$P(D|E) = P(D)\frac{P(E|D)}{P(E)}$$

Apriorní pravděpodobnost, $P(D)$, je pravděpodobnost užívání drog před tím, než se dozvíme výsledek testu, což je 5 %. Věrohodnost, $P(E \mid D)$, je pravděpodobnost pozitivního testu za předpokladu užívání drog, což představuje senzitivitu.

Normalizační konstanta, $P(E)$, se počítá o poznání obtížněji. Musíme vzít do úvahy dvě možnosti, $P(E \mid D)$ a $P(E \mid N)$, kde N je hypotéza, že subjekt testování neužívá drogy:

$$P(E) = P(D)\ P(E \mid D) + P(N)\ P(E \mid N)$$

Pravděpodobnost falešně pozitivního výsledku, $P(E \mid N)$, je doplněním specificity, neboli 1 %.

Když to dáme všechno dohromady, dostaneme

$$P(D|E) = \frac{P(D)P(E|D)}{P(D)P(E|D) + P(N)P(E|N)}$$

Dosazením příslušných hodnot získáme $P(D \mid E)$ = 0,76, což znamená, že z lidí, kterým vyjde pozitivní test, je přibližně 1 ze 4 nevinný.

Cvičení 5.14 Napište program, který přijme skutečnou míru užívání drog a senzitivitu a specificitu testu a použije Bayesovu větu k výpočtu $P(D \mid E)$.

Předpokládejme, že stejný test je použit na populaci, kde je skutečná míra užívání drog 1 %. Jaká je pravděpodobnost, že člověk s pozitivním testem opravdu užívá drogy?

Cvičení 5.15 Toto cvičení je z `http://wikipedia.org/wiki/Bayesian_inference`.

> „Předpokládejme, že máme dvě misky se sušenkami. Miska č. 1 obsahuje 10 sušenek s čokoládovými kousky a 30 obyčejných sušenek, zatímco miska č. 2 obsahuje 20 kusů od každého druhu. Náš přítel Fred vybere náhodně jednu misku a z ní vybere náhodně sušenku. Ukáže se, že sušenka je obyčejná. Jaká je pravděpodobnost toho, že ji Fred vzal z misky č. 1?“

Cvičení 5.16 Bonbony modré barvy se staly součástí balíčku M&Ms v roce 1995. Do té doby byl mix barev v balíčku čokoládových M&Ms následující: 30 % hnědá, 20 % žlutá, 20 % červená, 10 % zelená, 10 % oranžová, 10 % světle hnědá. Následně bylo složení takovéto: 24 % modrá, 20 % zelená, 16 % oranžová, 14 % žlutá, 13 % červená, 13 % hnědá.

Můj kamarád má dva balíčky M&Ms a řekl mi, že jeden je z roku 1994 a druhý z roku 1996. Neprozradí mi, který je který, ale dá mi jeden bonbon M&M z každého balíčku. Jeden je žlutý a druhý zelený. Jaká je pravděpodobnost, že žlutá M&M je z balíčku z roku 1994?

Cvičení 5.17 Toto cvičení je převzato z MacKay, *Information Theory, Inference, and Learning Algorithms*:

Elvis Presley měl bratra-dvojče, který zemřel při porodu. Podle článku na Wikipedii o dvojčatech:

> „Podle odhadu tvoří dvojčata přibližně 1,9 % světové populace, jednovaječná dvojčata přitom tvoří 0,2 % celkové populace a 8 % všech dvojčat.“

Jaká je pravděpodobnost, že Elvis byl z jednovaječných dvojčat?

5.8 Glosář

jev (event): Něco, co může a nemusí nastat, s určitou pravděpodobností.

pokus (trial): Jedna z řady příležitostí, kdy může nastat určitý jev.

úspěch (success): Pokus, při němž nastane určitý jev.

neúspěch (failure): Pokus, při němž jev nenastane.

frekventistický přístup (frequentism): Striktní interpretace pravděpodobnosti, která se vztahuje pouze na řadu identických pokusů.

bayesovský přístup (Bayesianism): Obecnější interpretace, která využívá pravděpodobnost k vyjádření subjektivního stupně přesvědčení.

nezávislé (independent): Dva jevy jsou nezávislé, jestliže výskyt jednoho nemá vliv na pravděpodobnost toho druhého.

variační koeficient (coefficient of variation): Statistický ukazatel, který měří variabilitu, normalizovanou střední hodnotou, a slouží k porovnání mezi rozděleními s různými průměry.

simulace Monte Carlo (Monte Carlo simulation): Metoda výpočtu pravděpodobností na základě simulace náhodných procesů (viz `http://wikipedia.org/wiki/Monte_Carlo_method`).

aktualizace (update): Proces využívání dat k revidování pravděpodobnosti.

apriorní pravděpodobnost (prior): Pravděpodobnost před bayesovskou aktualizací.

aposteriorní pravděpodobnost (posterior): Pravděpodobnost vypočtená na základě bayesovské aktualizace.

věrohodnost důkazu (likelihood of evidence): Jeden z pojmů Bayesovy věty, pravděpodobnost důkazu podmíněná hypotézou.

normalizační konstanta (normalizing constant): Jmenovatel Bayesovy věty používaný k normalizaci výsledku tak, aby vyjadřoval pravděpodobnost.

Kapitola 6

Operace s rozděleními

6.1 Šikmost

Šikmost je statistický ukazatel, který měří asymetrii rozdělení. Máme-li řadu hodnot x_i, šikmost výběrového souboru je:

$$g_1 = m_3 / m_2^{3/2}$$

$$m_2 = \frac{1}{n} \sum_i (x_i - \mu)^2$$

$$m_3 = \frac{1}{n} \sum_i (x_i - \mu)^3$$

V m_2 můžete poznat střední kvadratickou odchylku (známou také jako rozptyl); m_3 je pak třetí mocnina odchylky od průměru.

Záporná šikmost znamená, že rozdělení je „zešikmeno doleva;“ tedy, že je vychýlené víc doleva než doprava. Kladná šikmost značí to, že rozdělení je zešikmeno doprava.

V praxi není vypočítání šikmosti výběrového souboru obvykle dobrý nápad. Vyskytují-li se totiž v souboru odlehlé hodnoty, mají disproporční vliv na g_1.

Dalším způsobem, jak vyhodnotit asymetrii rozdělení, je podívat se na vztah mezi průměrem a mediánem. Extrémní hodnoty mají větší vliv na průměr než na medián, takže v rozdělení, které je zešikmeno doleva, je průměr menší než medián.

Pearsonova míra šikmosti je alternativní ukazatel šikmosti, který explicitně zachycuje vztah mezi průměrem, μ, a mediánem, $\mu_{1/2}$:

$$g_p = 3(\mu - \mu_{1/2})/\sigma$$

Tuto statistickou charakteristiku můžeme označit za **robustní**, což znamená, že méně podléhá vlivu odlehlých hodnot.

Cvičení 6.1 Napište funkci s názvem `Skewness`, která vypočte g_1 pro výběrový soubor.

Vypočtěte šikmost pro rozdělení délky těhotenství a porodní hmotnosti. Jsou výsledky konzistentní s tvarem těchto rozdělení?

Napište funkci s názvem `PearsonSkewness`, která vypočítá g_p pro tato rozdělení. Jak vypadá g_p v porovnání s g_1?

Cvičení 6.2 „Efekt jezera Wobegon“ je humorná přezdívka[1] pro **iluzi nadřazenosti**, což je lidský sklon přeceňovat své schopnosti v porovnání s ostatními. Například podle některých šetření je více než 80 % respondentů přesvědčeno, že jsou lepší než průměrný řidič (viz `http://wikipedia.org/wiki/Illusory_superiority`).

Jestliže interpretujeme „průměrný“ ve smyslu mediánu, pak je tento výsledek logicky nemožný, ale jestliže „průměrným“ myslíme průměr, pak je výsledek možný, i když nepravděpodobný.

Jaké procento populace má větší než průměrný počet nohou?

Cvičení 6.3 Daňová správa USA (The Internal Revenue Service of the United States – IRS) zpřístupňuje data o daních z příjmů a další statistické ukazatele na adrese `http://irs.gov/taxstats`. Pokud jste udělali cvičení 4.13, už jste s těmito daty pracovali; v opačném případě postupujte podle v něm uvedených instrukcí k získání rozdělení příjmů z tohoto souboru dat.

Jaká část populace uvádí, že má zdanitelné příjmy nižší než průměr?

Vypočtěte medián, průměr, šikmost a Pearsonovu míru šikmosti dat týkajících se příjmů. Vzhledem k tomu, že jsou data rozdělena do tříd, budete muset provést určitá přiblížení.

Giniho koeficient slouží jako ukazatel nerovnosti příjmů. Přečtěte si o něm na `http://wikipedia.org/wiki/Gini_coefficient` a napište funkci nazvanou `Gini`, která jej vypočítá pro rozdělení příjmů.

[1] Pokud vám to nic neříká, podívejte se na `http://wikipedia.org/wiki/Lake_Wobegon`.

Nápověda: Použijte PMF pro výpočet relativní střední diference (viz `http://wikipedia.org/wiki/Mean_difference`).

Řešení tohoto cvičení si můžete stáhnout z `http://thinkstats.com/gini.py`.

6.2 Náhodné proměnné

Náhodná proměnná představuje proces, který generuje náhodné číslo. Náhodné proměnné se většinou označují velkými písmeny, jako například X. Když uvidíte náhodnou proměnnou, měli byste si pomyslet „hodnota vybraná z rozdělení.“

Například formální definice distribuční funkce je:

$$\text{CDF}_X(x) = P(X \leq x)$$

Tomuto zápisu jsem se zatím vyhýbal, protože je tak příšerný, ale tady je jeho vysvětlení: CDF náhodné proměnné X, vypočtená pro konkrétní hodnotu x, je definována jako pravděpodobnost, že určitá hodnota vygenerovaná náhodným procesem X je menší nebo rovná x.

Jako informatik považuji za užitečné uvažovat o náhodné proměnné jako o objektu, který poskytuje metodu, kterou nazvu `generate`, jež využívá náhodný proces ke generování hodnot.

Například zde je definice pro třídu, která představuje náhodné proměnné:

```
class RandomVariable(object):
    """Parent class for all random variables."""
```

A zde je náhodná proměnná s exponenciálním rozdělením:

```
class Exponential(RandomVariable):
    def __init__(self, lam):
        self.lam = lam

    def generate(self):
        return random.expovariate(self.lam)
```

Inicializační metoda (init method) přijme parametr λ a uloží jej jako atribut. Metoda `generate` vrátí náhodnou hodnotu z exponenciálního rozdělení s uvedeným parametrem.

Pokaždé, když vyvoláte `generate`, dostanete jinou hodnotu. Hodnota, kterou dostanete, se nazývá **náhodná hodnota (random variate)**, což je důvod, proč řada názvů funkcí v modulu `random` obsahuje slovo „variate.“

Kdybych generoval pouze exponenciální veličiny, neobtěžoval bych se s definováním nové třídy a použil bych `random.expovariate`. Ale pro jiná rozdělení by mohlo být užitečné použít objekty RandomVariable. Například Erlangovo rozdělení je spojité rozdělení s parametry λ a k (viz `http://wikipedia.org/wiki/Erlang_distribution`).

Jeden ze způsobů generování hodnot z Erlangova rozdělení je provést součet k hodnot z exponenciálního rozdělení se stejným λ. Zde je provedení:

```
class Erlang(RandomVariable):
    def __init__(self, lam, k):
        self.lam = lam
        self.k =
k~self.expo = Exponential(lam)

    def generate(self):
        total = 0
        for i in range(self.k):
            total += self.expo.generate()
        return total
```

Inicializační metoda (init method) vytvoří objekt Exponential s daným parametrem; následně jej použije `generate`. Obecně platí, že inicializační metoda může přijmout jakoukoliv množinu parametrů a funkce `generate` může provést jakýkoliv náhodný proces.

Cvičení 6.4 Napište definici třídy, která představuje náhodnou proměnnou s Gumbelovým rozdělením (viz `http://wikipedia.org/wiki/Gumbel_distribution`).

6.3 Hustota pravděpodobnosti (PDFs)

Derivace CDF se nazývá **hustota pravděpodobnosti**, nebo také PDF. Například PDF exponenciálního rozdělení je

$$\mathrm{PDF}_{expo}(x) = \lambda e^{-\lambda x}$$

PDF normálního rozdělení je

$$\text{PDF}_{normal}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Vyhodnocení PDF pro konkrétní hodnotu x nám většinou příliš nepomůže. Výsledkem není pravděpodobnost, ale *hustota* pravděpodobnosti.

Ve fyzice se pod pojmem hustota rozumí hmotnost na jednotku objemu; abychom získali hmotnost, musíme násobit objemem, nebo, pokud hustota není konstantní, integrovat přes objem.

Obdobně hustota pravděpodobnosti měří pravděpodobnost na jednotku x. K získání hmotnosti pravděpodobnosti[2] musíme integrovat přes x. Například jestliže x je náhodná proměnná, jejíž PDF je PDF_X, můžeme spočítat pravděpodobnost, že hodnota z X se bude nacházet mezi -0,5 a 0,5:

$$P(-0.5 \le X < 0.5) = \int_{-0.5}^{0.5} \text{PDF}_X(x)dx$$

Nebo, protože CDF je integrál PDF, můžeme napsat

$$P(-0.5 \le X < 0.5) = \text{CDF}_X(0.5) - \text{CDF}_X(-0.5)$$

Pro některá rozdělení můžeme CDF vypočítat explicitně, takže bychom použili druhou možnost. Jinak musíme obvykle integrovat PDF numericky.

Cvičení 6.5 Jaká je pravděpodobnost, že hodnota vybraná z exponenciálního rozdělení s parametrem λ se bude nacházet mezi 1 a 20? Svoje řešení vyjádřete jako funkci λ. Výsledek si nechte někde po ruce, protože s ním budeme dál pracovat v Oddíle 8.8.

Cvičení 6.6 V rámci šetření BRFSS (viz Oddíl 4.5), je rozdělení výšek přibližně normální s parametry μ = 178 cm a σ^2 = 59,4 cm u mužů a μ = 163 cm a σ^2 = 52,8 cm u žen.

Abyste se mohli stát členem Blue Man Group, musíte být muž s výškou mezi 5'10" a 6'1" (viz `http://bluemancasting.com`). Jaké procento americké populace spadá do tohoto rozpětí? Nápověda: viz Oddíl 4.3.

[2]Posuneme-li analogii o krok dál, pak průměr rozdělení představuje jeho těžiště a rozptyl jeho moment setrvačnosti.

6.4 Konvoluce

Předpokládejme, že máme dvě náhodné proměnné, X a Y, s rozděleními CDF_X a CDF_Y. Jaké je rozdělení součtu $Z = X + Y$?

Jednou možností je napsat objekt RandomVariable, který vygeneruje tento součet:

```
class Sum(RandomVariable):
    def __init__(X, Y):
        self.X = X
        self.Y = Y

    def generate():
        return X.generate() + Y.generate()
```

Známe-li RandomVariables, X a Y, můžeme vytvořit objekt Sum, který představuje Z. Pak můžeme použít výběr ze Z k aproximování CDF_Z.

Tento přístup je jednoduchý a univerzální, ale nepříliš efektivní. K správnému odhadu CDF_Z je potřeba vygenerovat velký vzorek, a ani tak není odhad přesný.

Jestliže jsou CDF_X a CDF_Y vyjádřeny jako funkce, někdy se nám podaří najít CDF_Z přesně. Zde je postup:

1. Na úvod předpokládejme, že konkrétní hodnota X je x. Pak $\text{CDF}_Z(z)$ je
$$P(Z \leq z \mid X = x) = P(Y \leq z - x)$$
Pojďme se na to společně podívat. Levá strana je „pravděpodobnost, že součet je menší než z, za předpokladu, že první člen je x.“ Jestliže je tedy první člen x a součet musí být menší než z, pak druhý člen musí být menší než z - x.

2. Abychom zjistili pravděpodobnost toho, že Y je menší než z - x, vypočteme CDF_Y.
$$P(Y \leq z - x) = \text{CDF}_Y(z - x)$$
Toto vyplývá z definice CDF.

3. Zatím dobré? Pojďme dále. Protože ale neznáme hodnotu x, musíme vzít do úvahy všechny hodnoty, které by mohlo nabývat, a integrovat přes ně:
$$P(Z \leq z) = \int_{-\infty}^{\infty} P(Z \leq z \mid X = x)\, \text{PDF}_X(x)\, dx$$

Integrand je „pravděpodobnost, že Z je menší nebo rovno z, za předpokladu, že $X = x$, krát pravděpodobnost, že $X = x$.“

Substitucí z předchozích kroků získáme

$$P(Z \leq z) = \int_{-\infty}^{\infty} \mathrm{CDF}_Y(z - x)\ \mathrm{PDF}_X(x)\ dx$$

Levá strana je definice CDF_Z, takže vyvodíme závěr:

$$\mathrm{CDF}_Z(z) = \int_{-\infty}^{\infty} \mathrm{CDF}_Y(z - x)\ \mathrm{PDF}_X(x)\ dx$$

4. K získání PDF_Z proveďte oboustrannou derivaci vzhledem k z. Výsledkem je

$$\mathrm{PDF}_Z(z) = \int_{-\infty}^{\infty} \mathrm{PDF}_Y(z - x)\ \mathrm{PDF}_X(x)\ dx$$

Pokud jste studovali signály a systémy, možná tento integrál budete znát. Jedná se o **konvoluci** PDF_Y a PDF_X, označenou pomocí operátoru $\star$.

$\mathrm{PDF}_Z = \mathrm{PDF}_Y \star \mathrm{PDF}_X$

Rozdělení součtu je tudíž konvoluce rozdělení. Viz `http://wiktionary.org/wiki/booyah`!

Jako příklad uvažujme, že X a Y jsou náhodné proměnné s exponenciálním rozdělením s parametrem λ. Rozdělení $Z = X + Y$ je:

$$\mathrm{PDF}_Z(z) = \int_{-\infty}^{\infty} \mathrm{PDF}_X(x)\ \mathrm{PDF}_Y(z - x)\ dx = \int_{-\infty}^{\infty} \lambda e^{-\lambda x}\ \lambda e^{-\lambda(z-x)} dx$$

Musíme mít na paměti, že PDF_{expo} je 0 pro všechny negativní hodnoty, ale s tím se vypořádáme tak, že přizpůsobíme integrační meze:

$$\mathrm{PDF}_Z(z) = \int_{0}^{z} \lambda e^{-\lambda x}\ \lambda e^{-\lambda(z-x)}\ dx$$

Nyní můžeme kombinovat členy a přesunout konstanty mimo integrál:

$$\mathrm{PDF}_Z(z) = \lambda^2 e^{-\lambda z} \int_{0}^{z} dx = \lambda^2 z\ e^{-\lambda z}$$

Vidíme, že se jedná o PDF Erglangova rozdělení s parametrem $k = 2$ (viz `http://wikipedia.org/wiki/Erlang_distribution`). Takže konvo-

luce dvou exponenciálních rozdělení (se stejným parametrem) je Erlangovo rozdělení.

Cvičení 6.7 Jestliže X má exponenciální rozdělení s parametrem λ a Y má Erlangovo rozdělení s parametry k a λ, jaké je rozdělení součtu $Z = X + Y$?

Cvičení 6.8 Předpokládejme, že vezmu dvě hodnoty z nějakého rozdělení, jaké je rozdělení větší z hodnot? Vyjádřete svoji odpověď prostřednictvím PDF nebo CDF rozdělení.

S narůstajícím počtem hodnot konverguje rozdělení maxima k jednomu z rozdělení extrémních hodnot; viz `http://wikipedia.org/wiki/Gumbel_distribution`.

Cvičení 6.9 Znáte-li Pmf objekty, můžete vypočítat rozdělení součtu vyčíslením všech párů hodnot:

```
for x in pmf_x.Values():
    for y in pmf_y.Values():
        z += x + y
```

Napište funkci, která přijme PMF_X a PMF_Y a vrátí novou Pmf, která představuje rozdělení součtu $Z = X + Y$.

Napište obdobnou funkci, která vypočte PMF $Z = \max(X, Y)$.

6.5 Proč normální?

Dříve jsem uvedl, že normální rozdělení se dobře hodí pro analýzu, ale neřekl jsem proč. Jedním z důvodů je fakt, že jsou při lineární transformaci a konvoluci uzavřená. K vysvětlení toho, co to znamená, bude dobré představit určité způsoby zápisu.

Jestliže je rozdělení náhodné proměnné X normální s parametry μ a σ, můžete zapsat

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

kde symbol $\sim$ znamená „je rozděleno“ a psací písmeno $\mathcal{N}$ značí „normální.“

Lineární transformace X je něco jako $X' = aX + b$, kde a a b jsou reálná čísla. Množina rozdělení je při lineární transformaci uzavřená, jestliže X' je ve stejné množině jako X. Normální rozdělení má tuto vlastnost; jestliže $X \sim \mathcal{N}(\mu, \sigma^2)$,

$X' \sim \mathcal{N}(a\mu + b, a^2\,\sigma^2)$

Normální rozdělení jsou uzavřená také při konvoluci. Jestliže $Z = X + Y$a $X \sim \mathcal{N}(\mu_X, {\sigma_X}^2)$ a $Y \sim \mathcal{N}(\mu_Y, {\sigma_Y}^2)$, pak

$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Ostatní rozdělení, s nimiž jsme se zde setkali, tyto vlastnosti nemají.

Cvičení 6.10 Jestliže $X \sim \mathcal{N}(\mu_X, {\sigma_X}^2)$ a $Y \sim \mathcal{N}(\mu_Y, {\sigma_Y}^2)$, jaké je rozdělení $Z = aX + bY$?

Cvičení 6.11 Podívejme se, co se stane, když sečteme hodnoty z jiných rozdělení. Vyberte si dvojici rozdělení (jakékoliv dvě z nabídky exponenciálního, normálního, lognormálního a Paretova rozdělení) a vyberte takové parametry, díky nimž budou jejich průměr a rozptyl podobné.

Vygenerujte náhodná čísla z těchto rozdělení a vypočítejte rozdělení jejich součtů. Použijte testy z Kapitoly 4, abyste zjistili, zda lze součet modelovat pomocí spojitého rozdělení.

6.6 Centrální limitní věta

Dosud jsme dospěli k následujícím zjištěním:

- Jestliže sečteme hodnoty vybrané z normálních rozdělení, rozdělení součtu je normální.
- Jestliže sečteme hodnoty vybrané z jiných rozdělení, součet obecně nevykazuje jedno ze spojitých rozdělení, která jsme zde popsali.

Ukazuje se ale, že pokud provedeme součet velkého počtu hodnot z téměř jakéhokoliv rozdělení, rozdělení součtu konverguje k normálnímu.

Konkrétněji pak lze říci, že jestliže rozdělení hodnot má průměr a směrodatnou odchylku μ a σ, je rozdělení součtu přibližně $\mathcal{N}(n\mu, n\sigma^2)$.

Toto se označuje jako **centrální limitní věta**. Je to jeden z nejužitečnějších nástrojů statistické analýzy, ale má také určitá omezení:

- Hodnoty musí být vybrány nezávisle.
- Hodnoty musí pocházet ze stejného rozdělení (ačkoliv tento požadavek může být zmírněn).

- Hodnoty musí být zvoleny z rozdělení s konečným průměrem a rozptylem, takže většina Paretových rozdělení nepřichází v úvahu.
- Počet hodnot potřebný k tomu, aby nastala konvergence, závisí na šikmosti rozdělení. Součty z exponenciálního rozdělení konvergují v případě malých rozsahů výběru. U součtů z lognormálního rozdělení tomu tak není.

Centrální limitní věta vysvětluje, alespoň částečně, převahu normálních rozdělení v přírodním světě. Většina vlastností zvířat a ostatních forem života je ovlivněna velkým počtem genetických faktorů a faktorů spojených s životním prostředím, jejichž působení je aditivní. Vlastnosti, které měříme, jsou součtem velkého počtu malých vlivů, takže jejich rozdělení má tendenci být normální.

Cvičení 6.12 Jestliže vyberu vzorek, x_1 .. x_n, a sice nezávisle z rozdělení s konečným průměrem μ a rozptylem σ^2, jaké je rozdělení průměru vzorku:

$$\bar{x} = \frac{1}{n} \sum x_i$$

Jak se n zvyšuje, co se stane s rozptylem průměru vzorku? Nápověda: Znovu si projděte Oddíl 6.5.

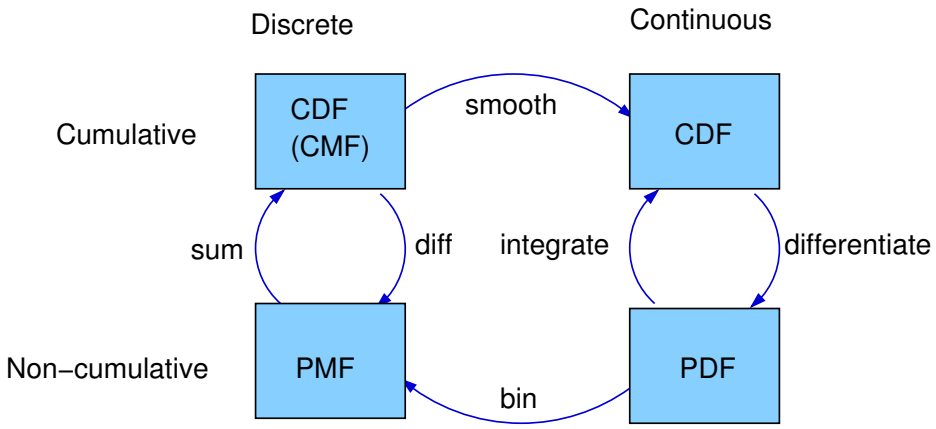
Cvičení 6.13 Vyberte si rozdělení (exponenciální, lognormální nebo Paretovo) a zvolte hodnoty pro parametr(y). Vygenerujte výběrové soubory o rozsahu 2, 4, 8 atd. a vypočtěte rozdělení jejich součtů. Použijte normální pravděpodobnostní graf, abyste zjistili, jestli je rozdělení přibližně normální. Kolik členů budete muset sečíst, abyste dosáhli konvergence?

Cvičení 6.14 Místo rozdělení součtů vypočítejte rozdělení součinů. Co se stane, když se počet členů zvýší? Nápověda: Podívejte se na rozdělení logaritmů součinů.

6.7 Struktura rozdělení

Dosud jsme se setkali s PMFs, CDFs a PDFs; pojďme si je krátce připomenout. Obrázek 6.1 ukazuje, jaké jsou mezi těmito funkcemi vzájemné vztahy.

Začali jsme pravděpodobnostními funkcemi (PMFs), které znázorňují pravděpodobnosti pro množinu diskrétních hodnot. Abychom se od PMF propracovali k distribuční funkci (CDF), vypočítali jsme kumulativní součet.



Obrázek 6.1: Struktura vzájemných souvislostí distribučních funkcí.

Abychom byli konzistentní, diskrétní CDF by se měla nazývat kumulativní funkce (cumulative mass function – CMF), ale pokud je mi známo, tak tento pojem nikdo nepoužívá.

Abyste se dostali od CDF k hustotě pravděpodobnosti (PMF), můžete vypočítat rozdíly v kumulativních pravděpodobnostech.

Podobně pak PDF je derivací spojité CDF; nebo ekvivalentně CDF je integrálem PDF. Nezapomeňte ale, že PDF přiděluje hodnotám hustoty pravděpodobnosti. Abyste získali pravděpodobnost, musíte integrovat.

Abyste se od diskrétního rozdělení dostali ke spojitému, můžete provést různé druhy vyrovnání dat. Jednou z forem vyrovnání dat je předpokládat, že data pocházejí z analytického spojitého rozdělení (jako je exponenciální nebo normální rozdělení) a odhadnout parametry takového rozdělení. Přesně tomu se věnuje Kapitola 8.

Jestliže rozdělíte PDF do souboru tříd, můžete vygenerovat PMF, která je alespoň aproximací PDF. Tuto techniku používáme v Kapitole 8 k získání bayesovského odhadu.

Cvičení 6.15 Napište funkci s názvem `MakePmfFromCdf`, která přijme objekt Cdf a vrátí odpovídající objekt Pmf.

Řešení tohoto cvičení najdete v `thinkstats.com/Pmf.py`.

6.8 Glosář

šikmost (skewness): Vlastnost rozdělení; intuitivně, se jedná míru asymetričnosti rozdělení.

robustní (robust): Statistická charakteristika je robustní, jestliže je relativně imunní vůči působení odlehlých hodnot.

iluze nadřazenosti (illusory superiority): Lidský sklon představovat si, že jsem lepší než průměr.

náhodná proměnná (random variable): Objekt, který představuje náhodný proces.

náhodná hodnota (random variate): Hodnota vygenerovaná náhodným procesem.

hustota pravděpodobnosti (PDF) (probability density function): Hustota pravděpodobnosti je derivátem spojité CDF.

konvoluce (convolution): Operace, která vypočítá rozdělení součtu hodnot ze dvou rozdělení.

centrální limitní věta (Central Limit Theorem): „Nejvyšší zákon Chaosu“, podle Sira Francise Galtona, průkopníka statistiky.

Kapitola 7

Testování hypotéz

Při průzkumu dat ze souboru NSFG jsme narazili na řadu „pozorovaných zjištění“, včetně několika rozdílů mezi prvorozenými a ostatními dětmi. Tato zjištění jsme dosud brali za bernou minci, v této kapitole je ale, konečně, podrobíme zkoušce.

Základní otázkou, na kterou chceme najít odpověď, je, zda jsou tato zjištění skutečná. Například jestliže vypozorujeme rozdíl v průměrné délce těhotenství u prvorozených a ostatních dětí, chceme zjistit, jestli je daný rozdíl skutečný, nebo jestli k němu došlo náhodně.

Ukazuje se, že tuto otázku není lehké zodpovědět přímo, a tak budeme postupovat ve dvou krocích. Nejprve otestujeme, zda je dané zjištění **významné**, a poté se pokusíme výsledek interpretovat jako odpověď na naši původní otázku.

V kontextu statistiky se pojem „významný“ používá v technickém významu, který se liší od způsobu užití tohoto slova v běžném jazyce. Jak jsme uvedli v dřívější definici, pozorované zjištění je statisticky významné, jestliže je nepravděpodobné, že je dílem náhody.

Abychom to upřesnili, je potřeba zodpovědět tři otázky:

1. Co máme na mysli pod pojmem „náhoda“?
2. Co máme na mysli pod pojmem „nepravděpodobné“?
3. Co myslíme na mysli pod pojmem „zjištění“?

Všechny tři tyto otázky jsou těžší, než se zdá. Nicméně existuje obecná struktura, kterou lidé používají k otestování statistické významnosti:

Nulová hypotéza: **Nulová hypotéza** představuje model systému vycházející z předpokladu, že pozorované zjištění bylo ve skutečnosti dílem náhody.

p-hodnota: **P-hodnota** je pravděpodobnost pozorovaného zjištění při nulové hypotéze.

Interpretace: Na základě p-hodnoty dospějeme k závěru, že zjištění buď je, nebo není statisticky významné.

Tento proces se nazývá **testování hypotéz**. Opírá se o obdobnou logiku jako důkaz sporem. Abychom ověřili správnost matematického výroku A, předpokládáme dočasně, že výrok A není pravdivý. Jestliže tento předpoklad vede ke sporu, vyvodíme z toho, že výrok A musí být ve skutečnosti pravdivý.

Obdobně pak, abychom otestovali hypotézu jako například "Toto zjištění je skutečné ,,, předpokládáme dočasně, že není. Tento předpoklad představuje nulovou hypotézu. Na základě tohoto předpokladu vypočteme pravděpodobnost pozorovaného zjištění, tedy p-hodnotu. Je-li p-hodnota dostatečně nízká, vyvodíme z toho závěr, že je nepravděpodobné, aby nulová hypotéza byla pravdivá.

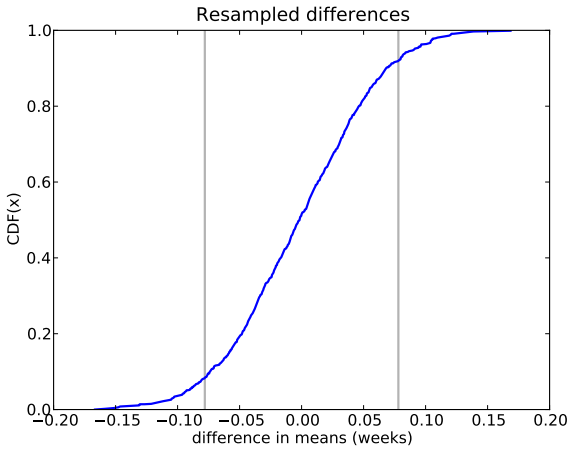
7.1 Testování rozdílu v průměrech

Jednou z nejsnadněji otestovatelných hypotéz je pozorovaný rozdíl v průměru mezi dvěma skupinami. V rámci dat v souboru NSFG jsme zjistili, že průměrná délka těhotenství u prvorozených dětí byla mírně větší a průměrná porodní hmotnost mírně menší. Teď se podíváme na to, zda jsou tato zjištění významná.

Pro tyto příklady je nulová hypotéza tvrzení, že rozdělení pro tyto dvě skupiny jsou totožná a že pozorovaný rozdíl je dílem náhody.

Pro účely výpočtu p-hodnot zjistíme celkové rozdělení pro všechny porody živých dětí (prvorozené a ostatní děti), vygenerujeme náhodné výběrové soubory o stejném rozsahu jako pozorované výběrové soubory a vypočítáme rozdíl v průměrech při nulové hypotéze.

Jestliže vygenerujeme velký počet výběrových souborů, můžeme spočítat, jak často je rozdíl v průměrech (v důsledku náhody) stejně velký nebo větší než skutečně pozorovaný rozdíl. Tento podíl představuje p-hodnotu.



Obrázek 7.1: CDF rozdílu v průměru pro resamplovaná data.

Pokud jde o délku těhotenství, pozorovali jsme n = 4413 prvorozených dětí a m = 4735 ostatních dětí, rozdíl v průměru byl přitom δ = 0,078 týdnů. Abych docílil aproximace p-hodnoty tohoto zjištění, sloučil jsem rozdělení, vygeneroval jsem výběrové soubory o rozsahu n a m a vypočítal jsem rozdíl v průměru.

Toto je další příklad procesu zvaného resampling, protože vybíráme náhodný vzorek ze souboru dat, který je sám o sobě výběrem z obecné populace. Vypočítal jsem rozdíly pro 1000 výběrových párů. Obrázek 7.1 ukazuje jejich rozdělení.

Rozdíl v průměru se blíží 0, jak se dá očekávat u vzorků ze stejného rozdělení. Svislé čáry představují oříznuté okraje, kde x = -δ nebo x = δ.

Z 1000 výběrových párů bylo 166, kdy byl rozdíl v průměru (kladný nebo záporný) stejně velký nebo větší než δ, takže p-hodnota je přibližně 0,166. Jinými slovy, očekáváme, že se setkáme se zjištěním, které bude stejně velké jako δ přibližně v 17 % případů, dokonce i když je skutečné rozdělení pro obě skupiny stejné.

Pozorované zjištění tudíž není příliš pravděpodobné, ale je dostatečně nepravděpodobné? Této otázce se budu věnovat v další části.

Cvičení 7.1 V souboru dat NSFG je rozdíl v průměrné hmotnosti u prvorozených dětí 2,0 unce. Vypočtěte p-hodnotu tohoto rozdílu.

Nápověda: U tohoto druhu resamplingu je důležité provádět výběr s opakováním, proto byste měli použít `random.choice` spíše než `random.sample` (viz Oddíl 3.8).

Můžete začít kódem, který jsem použil k vygenerování výsledků v tomto oddílu. Můžete si jej stáhnout z `http://thinkstats.com/hypothesis.py`.

7.2 Výběr hladiny významnosti

Při testování hypotéz se musíme mít na pozoru před dvěma druhy chyb.

- Chyba I. druhu, označovaná také jako **falešně pozitivní**, nastává, když přijmeme hypotézu, která je ve skutečnosti nepravdivá; neboli jestliže považujeme nějaké zjištění za významné, když je ve skutečnosti dílem náhody.
- Chyba II. druhu, označovaná také jako **falešně negativní**, nastává, když zamítneme hypotézu, která je ve skutečnosti pravdivá; neboli jestliže nějaký atribut připíšeme náhodě, když je skutečný.

Nejběžnější přístup k testování hypotéz spočívá v tom, že se zvolí hladina významnosti[1], α, pro p-hodnotu a jako významná jsou přijata jakákoliv zjištění s p-hodnotou menší než α. Běžnou volbou v případě α je 5 %. Na základě tohoto kritéria není pozorovaný rozdíl v délce těhotenství u prvorozených dětí významný, zatímco rozdíl v hmotnosti je.

U tohoto typu testování hypotéz můžeme vypočítat pravděpodobnost chyby I. druhu explicitně: Jedná se o α.

Abyste pochopili, proč tomu tak je, uvažujte o definici chyby I. druhu – pravděpodobnost přijetí hypotézy, která je nepravdivá – a o definici p-hodnoty – pravděpodobnost vygenerování naměřeného zjištění, jestliže je hypotéza nepravdivá.

Když dáme tyto dvě definice dohromady, můžeme si položit otázku: Je-li hypotéza nepravdivá, jaká je pravděpodobnost vygenerování naměřeného zjištění, které bude považováno za významné s hladinou α? Odpověď zní α.

Pravděpodobnost chyby I. druhu můžeme snížit snížením hladiny významnosti. Například jestliže je hladina významnosti 1 %, existuje pouze 1% pravděpodobnost chyby I. druhu.

Není to však zcela zadarmo: Snížení hladiny významnosti znamená zvýšení důkazního standardu, což zvyšuje pravděpodobnost zamítnutí platné hypotézy.

[1] Známá také jako „kritérium významnosti“.

Obecně existuje mezi chybami I. a II. druhu vzájemné vyvažování. Jediný způsob, jak snížit obě zároveň, je zvýšit rozsah vzorku (nebo, v některých případech, snížit chybu měření).

Cvičení 7.2 K prověření vlivu rozsahu výběru na p-hodnotu, uvažujte o tom, co se stane, když odstraníte polovinu dat ze souboru NSFG. Nápověda: Použijte `random.sample`. Co když odstraníte tři čtvrtiny dat, a tak dále?

Jaký je nejmenší rozsah výběru, u něhož je rozdíl v průměrné porodní hmotnosti stále ještě významný při $\alpha = 5$ %? O kolik větší musí být rozsah výběru při $\alpha = 1$ %?

Můžete začít kódem, který jsem použil k vygenerování výsledků v tomto oddílu. Je ke stažení zde: `http://thinkstats.com/hypothesis.py`.

7.3 Definování zjištění

Stane-li se něco neobvyklého, lidé často říkají něco jako: „Teda! Jaká byla pravděpodobnost, že se *tohle* stane?“ Tato otázka dává smysl, protože intuitivně vnímáme, že některé věci jsou pravděpodobnější než jiné. Při podrobnějším přezkoumání tato intuice ale vždy neobstojí.

Předpokládejme například, že hodím 10krát mincí a po každém hodu zapíšu P, když padne panna, a O, když padne orel. Bude-li výsledkem sekvence OPPOPOOOPP, asi to nikoho příliš nepřekvapí. Jestliže bude ale výsledek vypadat takto: PPPPPPPPPP, asi byste reagovali slovy jako „Teda! Jaká byla pravděpodobnost, že se *tohle* stane?“

V tomto příkladu je ale pravděpodobnost obou sekvencí stejná: jedna ku 1024. A to stejné platí pro jakoukoliv jinou sekvenci. Takže ptáme-li se: „Jaká byla pravděpodobnost, že se *tohle* stane?“, musíme si dát pozor na to, co myslíme tím „tohle“.

Pro data ze souboru NSFG jsem definoval zjištění jako „rozdíl v průměru (kladný nebo záporný) stejný nebo větší než δ.“ Tím, že jsem provedl tuto volbu, jsem se rozhodl hodnotit velikost rozdílu bez ohledu na znaménko.

Takovýto test se nazývá **dvoustranný** test, protože bereme do úvahy obě strany (kladnou a zápornou) v rozdělení na Obrázku 7.1. Použitím dvoustranného testu testujeme hypotézu o tom, že existuje významný rozdíl mezi rozděleními, aniž bychom specifikovali znaménko příslušného rozdílu.

Alternativou je pak použití **jednostranného** testu, který se ptá, zda průměr u prvorozených dětí je významně *vyšší* než průměr u ostatních dětí. Protože je tato hypotéza konkrétnější, je p-hodnota nižší – v tomto případě je přibližně poloviční.

7.4 Interpretace výsledku

V úvodu této kapitoly jsem uvedl, že otázka, na niž chceme najít odpověď, je, zda je pozorované zjištění skutečné. Začali jsme definicí nulové hypotézy, označované jako H_0, což je hypotéza o tom, že zjištění není skutečné. Následně jsme definovali p-hodnotu, tedy $P(E \mid H_0)$, kde E je zjištění stejné nebo větší než pozorované zjištění. Poté jsme vypočetli p-hodnoty a porovnali jsme je s hladinou α.

To je užitečný krok, ale nepřináší odpověď na původní otázku, tedy na to, zda je zjištění skutečné. Existuje několik způsobů interpretace výsledku testu hypotézy:

Klasický: Při klasickém testování hypotéz platí, že je-li p-hodnota menší než α, můžeme říci, že zjištění je statisticky významné, ale nemůžeme říci, že je skutečné. Taková formulace je dostatečně opatrná, abychom se vyhnuli ukvapeným závěrům, ale je hluboce neuspokojivá.

Praktický: V praxi lidé nepostupují takto formálně. Ve většině vědeckých časopisů badatelé uvádí p-hodnoty bez rozpaků a čtenáři si je vykládají jako důkaz o tom, že pozorované zjištění je skutečné. Čím nižší je p-hodnota, tím větší je důvěra v takový závěr.

Bayesovský: To, co nás skutečně zajímá, je $P(H_A \mid E)$, kde H_A je hypotéza o tom, že zjištění je skutečné. Za použití Bayesovy věty

$$P(H_A \mid E) = \frac{P(E \mid H_A)\, P(H_A)}{P(E)}$$

kde $P(H_A)$ je apriorní pravděpodobnost H_A před tím, než jsme zaznamenali příslušné zjištění, $P(E \mid H_A)$ je pravděpodobnost zpozorování E, za předpokladu, že zjištění je skutečné, a $P(E)$ je pravděpodobnost zpozorování E při jakékoliv hypotéze. Vzhledem k tomu, že zjištění buď je, nebo není skutečné:

$$P(E) = P(E \mid H_A)\, P(H_A) + P(E \mid H_0)\, P(H_0)$$

Jako příklad vypočítám $P(H_A \mid \mathrm{E})$ pro délky těhotenství v rámci NSFG. Už jsme vypočítali $P(E \mid H_0) = 0{,}166$, takže nám zbývá jen vypočítat $P(E \mid H_A)$ a zvolit hodnotu pro apriorní pravděpodobnost.

Pro účely výpočtu $P(E \mid H_A)$ předpokládáme, že je zjištění skutečné – tedy že rozdíl v průměrné délce trvání, δ, je ve skutečnosti roven námi vypozorovanému výsledku 0,078. (Tento způsob formulace H_A je trochu podvodný. Tento problém vysvětlím a nápravu zjednám v dalším oddílu.)

Vygenerováním 1000 výběrových párů, vždy jednoho z každého rozdělení, jsem odhadl, že $P(\mathrm{E} \mid H_A) = 0{,}494$. Při apriorní pravděpodobnosti $P(H_A) = 0{,}5$ je aposteriorní pravděpodobnost H_A rovna 0,748.

Jestliže apriorní pravděpodobnost H_A je 50 %, aktualizovaná pravděpodobnost, se zohledněním důkazů z tohoto souboru dat, je téměř 75 %. Dává smysl, že aposteriorní pravděpodobnost je vyšší, neboť data poskytují určitou podporu pro danou hypotézu. Mohlo by se ale zdát překvapivé, že ten rozdíl je tak velký, zejména v situaci, kdy jsme zjistili, že rozdíl v průměrech nebyl statisticky významný.

Ve skutečnosti není metoda, kterou jsem použil v tomto oddílu, tak docela správná a projevuje se u ní tendence přeceňovat význam důkazů. V dalším oddílu provedu korekci této tendence.

Cvičení 7.3 Na základě dat z NSFG určete, jaká je aposteriorní pravděpodobnost, že rozdělení porodních hmotností se liší u prvorozených a ostatních dětí.

Můžete začít kódem, který jsem použil k vygenerování výsledků v tomto oddílu. Můžete si jej stáhnout zde: `http://thinkstats.com/hypothesis.py`.

7.5 Křížová validace

V předchozím příkladu jsme použili k formulaci hypotézy H_A soubor dat a pak jsme tentýž soubor dat použili k jejímu otestování. To není dobrý nápad; může se totiž velmi snadno stát, že vygenerujeme zavádějící výsledky.

Problém spočívá v tom, že i když je nulová hypotéza pravdivá, je pravděpodobné, že mezi libovolnými dvěma skupinami existuje nějaký rozdíl, δ, čistě náhodně. Použijeme-li zjištěnou hodnotu δ k formulaci hypotézy, je pravděpodobné, že $P(H_A \mid E)$ bude vysoká, i když je H_A nepravdivá.

S tímto problémem se můžeme vypořádat s pomocí **křížové validace**, která používá jeden soubor dat k výpočtu δ a *jiný* soubor dat k vyhodnocení H_A. První soubor dat se nazývá **trénovací množina**, druhý pak **testovací množina**.

V rámci studie jako je NSFG, která v každém cyklu sleduje jinou kohortu, můžeme použít jeden cyklus pro trénování a jiný pro testování. Nebo můžeme data rozdělit do podmnožin (náhodně), a pak použít jednu pro trénování a druhou pro testování.

Já jsem použil druhý přístup, kdy jsem data z Cyklu 6 rozdělil zhruba na dvě poloviny. Provedl jsem test několikrát s různým náhodným rozdělením. Průměrná aposteriorní pravděpodobnost byla $P(H_A \mid E) = 0{,}621$. Podle očekávání je dopad důkazů menší, částečně v důsledku menšího rozsahu vzorku v testovací množině, ale také v důsledku toho, že už nepoužíváme stejný soubor dat pro trénování i testování.

7.6 Vyjadřování bayesovských pravděpodobností

V předešlé části jsme zvolili apriorní pravděpodobnost $P(H_A) = 0{,}5$. Máme-li množinu hypotéz a nemáme přitom žádný důvod domnívat se, že jedna z nich je pravděpodobnější než ostatní, je obvyklé, že každé z nich přidělíme stejnou pravděpodobnost.

Někteří lidé mají vůči bayesovským pravděpodobnostem výhrady, protože jsou závislé na apriorních pravděpodobnostech a může se stát, že se nenajde shoda ohledně toho, jaká apriorní pravděpodobnost je správná. Pro ty, kdo očekávají, že vědecké výsledky budou objektivní a univerzální, je tato vlastnost hluboce znepokojivá.

Jednou možnou odpovědí na tuto výhradu je to, že v praxi mají silné důkazy tendenci zastínit efekt apriorní pravděpodobnosti, takže lidé, kteří začnou s různými apriorními pravděpodobnostmi, nakonec konvergují ke stejné aposteriorní pravděpodobnosti.

Další možností je uvádět pouze **věrohodnostní poměr**, $P(\mathrm{E} \mid H_A) \;/\; P(E \mid H_0)$, namísto aposteriorní pravděpodobnosti. Tímto způsobem si čtenáři mohou dosadit jakoukoliv apriorní pravděpodobnost a vypočítat svoje vlastní aposteriorní pravděpodobnosti. Věrohodnostní poměr se někdy označuje jako Bayesův faktor (viz `http://wikipedia.org/wiki/Bayes_factor`).

Cvičení 7.4 Jestliže jste určili apriorní pravděpodobnost hypotézy H_A jako 0,3 a objeví se nové důkazy, v jejichž světle vyjde věrohodnostní poměr na 3 ve vztahu k nulové hypotéze H_0, jaká je v tomto případě aposteriorní pravděpodobnost pro H_A?

Cvičení 7.5 Toto cvičení je převzato z MacKay, *Information Theory, Inference, and Learning Algorithms*:

> Na místě činu zanechali stopy své krve dva lidé. Oliver, který je podezřelý, je podroben testu, který určí jeho krevní typ jako 0. Ukáže se, že ony dvě stopy krve odpovídají typu 0 (v místní populaci se jedná o běžný typ, s 60% četností) a dále typu AB (vzácný typ, s četností 1 %). Poskytují tato data (krevní typy nalezené na místě činu) důkaz svědčící ve prospěch tvrzení, že Oliver byl jedním ze dvou lidí, jejichž krev byla nalezena na místě činu?

Nápověda: Vypočítejte věrohodnostní poměr pro tento důkaz; jestliže je větší než 1, pak tento důkaz svědčí ve prospěch uvedeného tvrzení. Řešení a diskusi najdete na straně 55 MacKayovy knihy.

7.7 Chí-kvadrát test

V Oddílu 7.2 jsme dospěli k závěru, že pozorovaný rozdíl v průměrné délce těhotenství u prvorozených a ostatních dětí nebyl významný. Avšak v Oddílu 2.10, když jsme spočítali relativní riziko, jsme zjistili, že u prvorozených dětí existuje větší pravděpodobnost, že se narodí předčasně, menší pravděpodobnost, že se narodí načas, a větší pravděpodobnost, že se narodí se zpožděním.

Možná mají tedy tato rozdělení stejný průměr a různý rozptyl. Mohli bychom otestovat významnost rozdílu v rozptylu, ale rozptyl je méně robustní charakteristika než průměr a testování hypotéz s ohledem na rozptyl obvykle nepřináší kýžené výsledky.

Alternativou je otestovat hypotézu, která odráží konkrétní zjištění tak, jak je pozorováno, příměji. Neboli hypotézu o tom, že u prvorozených dětí existuje větší pravděpodobnost, že se narodí předčasně, menší pravděpodobnost, že se narodí načas, a větší pravděpodobnost, že se narodí se zpožděním.

Postup spočívá v pěti jednoduchých krocích:

1. Definujeme množinu kategorií, označovaných jako **políčka**, do nichž by každé z dětí mohlo spadat. V tomto příkladu máme šest políček, protože existují dvě skupiny (prvorozené a ostatní děti) a tři třídy (předčasně, načas a se zpožděním narozené).

   Použiji definice z Oddílu 2.10: Dítě se narodí předčasně, jestliže přijde na svět v průběhu 37. týdne nebo dříve, načas, jestliže se narodí v průběhu 38., 39. nebo 40. týdne, a se zpožděním, jestliže se narodí v průběhu 41. týdne nebo později.

2. Spočítáme počet dětí očekávaných v každém políčku. Při nulové hypotéze předpokládáme, že rozdělení jsou pro obě skupiny stejná, takže můžeme vypočítat úhrnné pravděpodobnosti: P(předčasně), P(načas) a P(se zpožděním).

   Pro prvorozené děti máme n = 4413 vzorků, takže při nulové hypotéze očekáváme n P(předčasně) prvorozených dětí, které se narodí předčasně, n P(načas), které se narodí načas atd. Obdobně pak máme m = 4735 ostatních dětí, takže očekáváme m P(předčasně) ostatních dětí, které se narodí předčasně atd.

3. Pro každé políčko spočítáme odchylku, neboli rozdíl mezi zjištěnou hodnotou, O_i, a očekávanou hodnotou, E_i.

4. Vypočteme určitou míru celkové odchylky; tato veličina se nazývá **testová statistika**. Nejběžnější volbou je statistika chí-kvadrát:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

5. K výpočtu p-hodnoty můžeme použít simulaci Monte Carlo, což představuje pravděpodobnost zjištění statistiky chí-kvadrát, která bude stejně vysoká jako hodnota zjištěná při nulové hypotéze.

Použijeme-li statistiku chí-kvadrát, tento proces se označuje jako **chí-kvadrát test** nebo také **test dobré shody**. Jedním ze znaků chí-kvadrát testu je, že rozdělení testové statistiky lze vypočítat analyticky.

Za použití dat ze souboru NSFG jsem vypočítal χ^2 = 91,64, přičemž takovýto výsledek by se náhodně vyskytl přibližně v jednom případě z 10 000. Vyvozuji z toho, že tento výsledek je statisticky významný, s jedním upozorněním – opět jsme použili stejný soubor dat pro exploraci i testování. Bylo by dobré potvrdit tento výsledek za použití jiného souboru dat.

Kód, který jsem použil v tomto oddílu, si můžete stáhnout z `http://thinkstats.com/chi.py`.

Cvičení 7.6 Představte si, že provozujete kasino a máte podezření na jednoho ze zákazníků, že kostku, kterou dává k dispozici kasino, vyměnil za vlastní „podvrženou kostku;“ tedy takovou, se kterou bylo manipulováno tak, aby jedna ze stran padala častěji než ostatní. Zadržíte tedy domnělého podvodníka a zabavíte příslušnou kostku, teď ale musíte prokázat, že byla podvržená.

Hodíte kostkou 60krát a dostanete následující výsledky:

Hodnota	1	2	3	4	5	6
Četnost	8	9	19	6	8	10

Jaká je statistika chí-kvadrát pro tyto hodnoty? Jaká je pravděpodobnost, že se takto vysoká hodnota chí-kvadrát vyskytne náhodně?

7.8 Účinný resampling

Čtenář této knihy s předchozí znalostí statistiky se zřejmě zasmál, když uviděl Obrázek 7.1, protože jsem použil velké výpočetní síly k simulaci něčeho, na co jsem mohl přijít analyticky.

Je zřejmé, že matematická analýza není hlavním předmětem této knihy. Jsem ochoten používat počítače k tomu, abych na věci přicházel „hloupým“ způsobem, protože jsem přesvědčen o tom, že pro začátečníky je jednodušší porozumět simulacím a je také jednodušší demonstrovat, že jsou správné. A tak pokud provedení simulace netrvá příliš dlouho, nemám žádné výčitky kvůli tomu, že jsem přeskočil analýzu.

Nicméně existují situace, kdy trochu analýzy může člověku ušetřit hodně výpočtů a Obrázek 7.1 je jedním z těchto případů.

Určitě si vzpomínáte, že jsme testovali pozorovaný rozdíl v průměrné délce těhotenství pro $n = 4413$ prvorozených dětí a $m = 4735$ ostatních dětí. Dospěli jsme k úhrnnému rozdělení pro všechny děti, vybrali jsme vzorky o rozsahu n a m a vypočítali jsme rozdíl ve výběrových průměrech.

Místo toho jsme mohli přímo vypočítat rozdělení rozdílu ve výběrových průměrech. Pro začátek uvažujte o tom, co znamená výběrový průměr: vybereme n vzorků z rozdělení, sečteme je a výsledek vydělíme n. Jestliže má vzorek průměr μ a rozptyl σ^2, pak dle centrální limitní věty víme, že součet vzorků je $\mathcal{N}(n\mu, n\sigma^2)$.

Abychom zjistili rozdělení výběrových průměrů, musíme se odvolat na jednu z vlastností normálního rozdělení: jestliže X je $\mathcal{N}(\mu, \sigma^2)$,

$$aX + b \sim \mathcal{N}(a\mu + b, a^2\,\sigma^2)$$

Když provedeme dělení n, $a = 1/n$ a $b = 0$, pak

$$X/n \sim \mathcal{N}(\mu/n, \sigma^2/\,n^2)$$

Rozdělení výběrového průměru je tedy $\mathcal{N}(\mu, \sigma^2/n)$.

Abychom získali rozdělení rozdílu mezi dvěma výběrovými průměry, odvoláme se na další vlastnost normálního rozdělení: jestliže X_1 je $\mathcal{N}(\mu_1, {\sigma_1}^2)$ a X_2 je $\mathcal{N}(\mu_2, {\sigma_2}^2)$,

$$aX_1 + bX_2 \sim \mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

Pak jako speciální případ:

$$X_1 - X_2 \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

Dáme-li to vše dohromady, dospějeme k závěru, že vzorek na Obrázku 7.1 je vybrán z $\mathcal{N}(0, f\sigma^2)$, kde $f = 1/n + 1/m$. Po dosazení $n = 4413$ a $m = 4735$ očekáváme, že rozdíl ve výběrových průměrech bude $\mathcal{N}(0, 0{,}0032)$.

Můžeme použít `erf.NormalCdf` k výpočtu p-hodnoty pozorovaného rozdílu v průměrech:

```
delta = 0.078
sigma = math.sqrt(0.0032)
left = erf.NormalCdf(-delta, 0.0, sigma)
right = 1 - erf.NormalCdf(delta, 0.0, sigma)
```

Součet levé a pravé strany je p-hodnota, 0,168, což je velmi blízko našemu odhadu na základě resamplingu, který byl 0,166. Kód, který jsem použil v tomto oddílu, si můžete stáhnout z `http://thinkstats.com/hypothesis_analytic.py`

7.9 Síla

Je-li výsledek testování hypotézy negativní (tj. zjištění není statisticky významné), můžeme z toho vyvodit závěr, že zjištění není skutečné? To záleží na síle testu.

Statistická **síla** je pravděpodobnost, že test bude pozitivní, jestliže je nulová hypotéza nepravdivá. Obecně závisí síla testu na rozsahu vzorku, velikosti zjištění a hladině α.

Cvičení 7.7 Jaká je síla testu v Oddílu 7.2, za použití $\alpha = 0{,}05$ a předpokladu, že skutečný rozdíl mezi průměry je 0,078 týdnů?

Sílu můžete odhadnout vygenerováním náhodných vzorků z rozdělení s daným rozdílem v průměru, otestováním pozorovaného rozdílu v průměru a spočítáním počtu pozitivních testů.

Jaká je síla testu s $\alpha = 0{,}10$?

Jedním ze způsobů, jak uvádět sílu testu, vedle negativního výsledku, je tvrzení na způsob tohoto: „Pokud by bylo pozorované zjištění stejně velké jako x, tento test by zamítl nulovou hypotézu s pravděpodobností p.“

7.10 Glosář

významný (significant): Zjištění je statisticky významné, jestliže je nepravděpodobné, aby k němu došlo náhodně.

nulová hypotéza (null hypothesis): Model systému založený na předpokladu, že k pozorovanému zjištění došlo v důsledku náhody.

p-hodnota (p-value): Pravděpodobnost toho, že by zjištění mohlo vzniknout náhodně.

testování hypotéz (hypothesis testing): Proces určování, zda je pozorované zjištění statisticky významné.

chyba I. druhu (type I error) (falešně pozitivní (false positive)): Závěr, že zjištění je skutečné, když tomu tak není.

chyba II. druhu (type II error) (falešně negativní (false negative)): Závěr, že zjištění vzniklo v důsledku náhody, když tomu tak není.

dvoustranný test (two-sided test): Test, který se ptá: „Jaká je pravděpodobnost zjištění, které je stejně velké jako pozorované zjištění, ať už je kladné nebo záporné?“

jednostranný test (one-sided test): Test, který se ptá: „Jaká je pravděpodobnost zjištění, které je stejně velké jako pozorované zjištění a má také stejné znaménko?“

křížová validace (cross-validation): Proces testování hypotéz, který využívá jeden soubor dat pro explorační analýzu dat a jiný soubor dat pro testování.

trénovací množina (training set): Soubor dat používaný k formulování hypotézy pro testování.

testovací množina (testing set): Soubor dat používaný k testování.

testová statistika (test statistic): Statistika používaná k měření odchylky pozorovaného zjištění od toho, co je očekáváno náhodně.

chí-kvadrát test, test dobré shody (chi-square test): Test využívající chí-kvadrát statistiku jako testovou statistiku.

věrohodnostní poměr (likelihood ratio): Poměr $P(E \mid A)$ k $P(E \mid B)$ pro dvě hypotézy A a B, což je způsob, jak referovat o výsledcích bayesovské analýzy bez spoléhání se na apriorní pravděpodobnost.

políčko (cell): Jedná se o kategorie v chí-kvadrát testu, do kterých jsou údaje rozděleny.

síla (power): Pravděpodobnost, že test zamítne nulovou hypotézu, jestliže je nepravdivá.

Kapitola 8

Odhadování

8.1 Hra na odhad

Pojďme si zahrát hru. Budu si myslet rozdělení a vy musíte uhodnout, o které jde. Začneme něčím jednoduchým a postupně se bude obtížnost zvyšovat.

Myslím si rozdělení. Dám vám dvě nápovědy: Jedná se o normální rozdělení a zde je náhodný vzorek vybraný z tohoto rozdělení:

{-0,441, 1,774, -0,101, -1,138, 2,975, -2,138}

Jaký si myslíte, že je průměr základního souboru, μ, tohoto rozdělení?

Jednou možností je použít k odhadu μ výběrový průměr. Dosud jsme používali symbol μ jak pro výběrový průměr, tak pro průměr základního souboru/průměr jako parametr, nyní je však odliším tak, že pro výběrový průměr použiji $\bar{x}$. V tomto příkladu je $\bar{x}$ rovno 0,155, takže by bylo rozumné odhadnout, že $\mu = 0{,}155$.

Tento proces se označuje jako **odhadování** a statistika, kterou jsme použili (výběrový průměr) se nazývá **odhad**.

Použití výběrového průměru pro odhadnutí μ je natolik zřejmé, že je obtížné představit si jakoukoliv odpovídající alternativu. Představte si ale, že hru změníme zapojením odlehlých hodnot.

Myslím si rozdělení. Jedná se o normální rozdělení a zde je vzorek sestavený nespolehlivým měřičem, který občas umístí špatně desetinnou čárku.

{-0,441, 1,774, -0,101, -1,138, 2,975, -213,8}

Jaký je nyní váš odhad μ? Použijete-li výběrový průměr, váš odhad bude -35,12. Je toto nejlepší možná volba? Jaké jsou alternativy?

Jednou možností je určit a vyřadit odlehlé hodnoty a pak spočítat výběrový průměr zbytku. Další možností je použít jako odhad medián.

To, který odhad nám poslouží nejlépe, závisí na okolnostech (např. zda jsou přítomny odlehlé hodnoty) a také na tom, jaký je účel. Snažíte se minimalizovat chyby, nebo maximalizovat vaši šanci na nalezení správné odpovědi?

Jestliže se v souboru nevyskytují žádné odlehlé hodnoty, výběrový průměr minimalizuje **střední kvadratickou chybu (mean squared error – MSE)**. Jestliže budeme hru mnohokrát opakovat a pokaždé vypočítáme chybu $\bar{x}$ - μ, výběrový průměr minimalizuje

$$MSE = \frac{1}{m}\sum(\bar{x} - \mu)^2$$

Kde m je počet opakování hry na odhad (nezaměňovat s n, což je rozsah výběrového souboru použitého k výpočtu $\bar{x}$).

Minimalizace MSE je pěkná vlastnost, ale není to vždy ta nejlepší strategie. Představte si například situaci, kdy odhadujeme rozdělení rychlostí větru na staveništi. Pokud bude náš odhad příliš nadsazený, může se stát, že konstrukce bude zbytečně naddimenzovaná, čímž se zvýší náklady. Pokud ale bude náš odhad příliš nízký, budova by mohla spadnout. Protože náklady jako funkce chyby jsou asymetrické, minimalizace MSE není nejlepší strategií.

Jako další příklad uvažujte situaci, kdy hodím tři šestistranné kostky a požádám vás o předpověď součtu. Jestliže se přesně trefíte, získáte výhru, jinak nedostanete nic. V tomto případě je hodnota, která minimalizuje MSE, 10,5, ale to by byl příšerný odhad. Pro tuto hru potřebujete odhad, který má největší šanci na správnou odpověď, a tím je **maximálně věrohodný odhad (maximum likelihood estimator – MLE**). Jestliže zvolíte 10 nebo 11, vaše šance na výhru je 1 ku 8, a to je také nejlepší pozice, jakou můžete mít.

Cvičení 8.1 Napište funkci, která vybere 6 hodnot z normálního rozdělení s $\mu = 0$ a $\sigma = 1$. K odhadu μ použijte výběrový průměr a vypočtěte chybu $\bar{x}$ - μ. Proveďte funkci 1000krát a vypočtěte MSE.

Nyní program modifikujte tak, aby použil jako odhad medián. Znovu vypočtěte MSE a porovnejte s MSE pro $\bar{x}$.

8.2 Odhadněte rozptyl

Myslím si rozdělení. Jedná se o normální rozdělení a zde je (známý) vzorek:

{-0,441, 1,774, -0,101, -1,138, 2,975, -2,138}

Jaký si myslíte, že je rozptyl, σ^2, mého rozdělení? Opět platí, že zřejmou volbou je použití výběrového rozptylu jako odhadu. K označení výběrového rozptylu budu používat S^2, abych jej odlišil od neznámého parametru σ^2.

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

Pro velké výběrové soubory představuje S^2 adekvátní odhad, ale pro malé výběrové soubory bývá příliš nízký. Vzhledem k této nešťastné vlastnosti se označuje jako **zkreslený** odhad.

Odhad je **nezkreslený**, jestliže očekávaná celková (nebo střední) chyba po mnoha opakováních hry na odhad je 0. Naštěstí existuje další jednoduchá statistika, která je nezkresleným odhadem σ^2:

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

Největší problém s tímto odhadem je, že jeho název a symbol se používají nekonzistentně. Název „výběrový rozptyl“ může odkazovat buď k S^2 nebo S^2_{n-1} a symbol S^2 se používá pro kterýkoliv z nich nebo pro oba.

Zajímá-li vás vysvětlení, proč je S^2 zkreslený, a důkaz o tom, že S^2_{n-1} je nezkreslený, podívejte se na `http://wikipedia.org/wiki/Bias_of_an_estimator`.

Cvičení 8.2 Napište funkci, která vybere 6 hodnot z normálního rozdělení s $\mu = 0$ a $\sigma = 1$. Použijte výběrový rozptyl k odhadu σ^2 a vypočtěte chybu S^2 - σ^2. Proveďte funkci 1000krát a vypočtěte střední chybu (ne kvadratickou).

Nyní program modifikujte, aby použil nezkreslený odhad S^2_{n-1}. Znovu vypočtěte střední chybu a podívejte se, zda konverguje k nule s tím, jak se zvyšuje počet opakování hry.

8.3 Pochopení chyb

Než budeme pokračovat, je důležité vyjasnit si jeden častý zdroj nedorozumění. Vlastnosti jako MSE a zkreslení jsou dlouhodobá očekávání založená na mnoha opakováních hry na odhad.

Zatímco hru hrajete, neznáte chyby. Tím mám na mysli to, že když vám dám vzorek a požádám vás o odhad parametru, můžete vypočítat hodnotu odhadu, ale nemůžete vypočítat chybu. Pokud byste mohli, pak byste nepotřebovali odhad!

Důvod, proč hovoříme o chybě odhadu, je, abychom popsali chování různých odhadů z dlouhodobého hlediska. V této kapitole provádíme experimenty, jejichž smyslem je prozkoumat toto chování. Tyto experimenty jsou umělé v tom smyslu, že známe skutečné hodnoty parametrů, a tak můžeme vypočítat chyby. Při práci se skutečnými daty ale tyto hodnoty neznáte, a tak nemůžete chyby vypočítat.

Nyní ale zpět ke hře.

8.4 Exponenciální rozdělení

Myslím si rozdělení. Jedná se o exponenciální rozdělení a zde je vzorek:

{5,384, 4,493, 19,198, 2,790, 6,122, 12,844}

Jaký si myslíte, že je parametr λ tohoto rozdělení?

Obecně platí, že průměr exponenciálního rozdělení je $1/\lambda$, takže budeme-li postupovat opačně, mohli bychom zvolit

$$\hat{\lambda} = 1 \ / \ \bar{x}$$

Je běžnou praxí používat pro odhady zápis se stříškou, takže $\hat{\lambda}$ je odhad λ. A ne ledajaký odhad; je to také odhad MLE[1]. Chcete-li tedy maximalizovat svoji šanci na přesný odhad λ, pak $\hat{\lambda}$ je tou správnou volbou.

Víme ale, že $\bar{x}$ není robustní, jestliže jsou přítomny odlehlé hodnoty, takže očekáváme, že u $\hat{\lambda}$ nastane stejný problém.

Možná bychom mohli najít alternativu, která by se opírala o výběrový medián. Jak si vzpomínáte, medián exponenciálního rozdělení je $\ln(2) \ / \ \lambda$, z čehož můžeme opět vyvodit definici odhadu

$$\hat{\lambda}_{1/2} = \ln(2)/\mu_{1/2}$$

kde $\mu_{1/2}$ je výběrový medián.

Cvičení 8.3 Proveďte experiment, abyste zjistili, který z $\hat{\lambda}$ a $\hat{\lambda}_{1/2}$ vede k nižší MSE. Otestujte, zda je některý z nich zkreslený.

[1] Viz `http://wikipedia.org/wiki/Exponential_distribution#Maximum_likelihood`.

8.5 Intervaly spolehlivosti

Dosud jsme se zabývali odhady, které generují jednotlivé hodnoty, známé jako **bodové odhady**. Pro mnohé problémy by se nám ale lépe hodil interval specifikující horní a dolní hranici neznámého parametru.

Nebo obecněji bychom mohli chtít celé rozdělení, tedy rozpětí hodnot, kterých by parametr mohl nabývat, a dále pro každou hodnotu v rámci daného rozpětí také údaj o její pravděpodobnosti.

Začněme **intervaly spolehlivosti**.

Myslím si rozdělení. Jedná se o exponenciální rozdělení a zde je vzorek:

{5,384, 4,493, 19,198, 2,790, 6,122, 12,844}

Chci od vás rozpětí hodnot, o kterém si myslíte, že pravděpodobně obsahuje neznámý parametr λ. Konkrétněji chci 90% interval spolehlivosti, což znamená, že pokud budeme hrát tuto hru znovu a znovu, váš interval bude obsahovat λ v 90 % případů.

Ukazuje se, že tato verze hry je obtížná, a tak vám prozradím odpověď a na vás bude pouze ji otestovat.

Intervaly spolehlivosti se zpravidla popisují pomocí rizika chyby, α, takže pro 90% interval spolehlivosti je riziko chyby $\alpha = 0{,}1$. Interval spolehlivosti pro parametr λ exponenciálního rozdělení je

$$\left(\hat{\lambda}\frac{\chi^2(2n, 1-\alpha/2)}{2n}, \hat{\lambda}\frac{\chi^2(2n, \alpha/2)}{2n}\right)$$

kde n je rozsah výběrového souboru, $\hat{\lambda}$ je odhad založený na průměru z předchozího oddílu a $\chi^2(k, x)$ je CDF chí-kvadrát rozdělení s k stupni volnosti, hodnocené pro x (viz `http://wikipedia.org/wiki/Chi-square_distribution`).

Obecně se dá říci, že je obtížné vypočítat intervaly spolehlivosti analyticky, ale je relativně snadné je odhadnout pomocí simulace. Nejprve se ale musíme zmínit o bayesovských odhadech.

8.6 Bayesovský odhad

Jestliže provedete výběr a vypočtete 90% interval spolehlivosti, je lákavé říci, že existuje 90% pravděpodobnost, že se skutečná hodnota parametru

nachází uvnitř intervalu. Z frekventistického pohledu to ale není správné, protože parametr je neznámá, ale pevně daná hodnota. Buď se nachází, nebo nenachází v intervalu, který jste vypočetli, takže frekventistická definice pravděpodobnosti se zde neuplatní.

Pojďme tedy zkusit jinou verzi naší hry.

Myslím si rozdělení. Jedná se o exponenciální rozdělení a λ jsem vybral z rovnoměrného rozdělení mezi 0,5 a 1,5. Zde je vzorek, který označím jako X:

{2,675, 0,198, 1,152, 0,787, 2,717, 4,269}

Na základě tohoto vzorku řekněte, jakou hodnotu λ si myslíte, že jsem vybral.

V této verzi hry λ *je* náhodná veličina, takže můžeme oprávněně hovořit o jejím rozdělení a můžeme ji snadno vypočíst na základě Bayesovy věty.

Zde jsou jednotlivé kroky:

1. Rozdělte rozpětí (0,5, 1,5) do množiny tříd o stejné velikosti. Pro každou třídu definujeme H_i, což je hypotéza, že skutečná hodnota λ spadá do i- té třídy. Protože λ byla vybrána z rovnoměrného rozdělení, apriorní pravděpodobnost, $P(H_i)$, je stejná pro všechny i.

2. Pro každou hypotézu vypočítáme pravděpodobnost $P(X \mid H_i)$, což je pravděpodobnost, že vybereme vzorek X za předpokladu H_i.

$$P(X \mid H_i) = \prod_j \text{expo}(\lambda_i, x_j)$$

   kde expo(λ, x) je funkce, která vypočte PDF exponenciálního rozdělení s parametrem λ, hodnoceným pro x.

$$PDF_{expo}(\lambda, x) = \lambda e^{-\lambda x}$$

   Symbol $\prod$ vyjadřuje součin řady (viz `http://wikipedia.org/wiki/Multiplication#Capital_Pi_notation`).

3. Pak na základě Bayesovy věty je aposteriorní rozdělení

   $P(H_i \mid X) = P(H_i)\ P(X \mid H_i)\ / f$

   kde f je normalizační faktor

$$f = \sum_i P(H_i) P(X \mid H_i)$$

Známe-li aposteriorní rozdělení, je snadné vypočítat interval spolehlivosti. Například abyste vypočetli 90% interval spolehlivosti, můžete použít 5. a 95. percentil aposteriorního rozdělení.

Bayesovské intervaly spolehlivosti se někdy nazývají **credible intervals**. O tom, v čem se liší, si můžete přečíst zde: `http://wikipedia.org/wiki/Credible_interval`.

8.7 Implementace bayesovských odhadů

Ke znázornění apriorního rozdělení bychom mohli použít Pmf, Cdf, nebo jakékoliv jiné zobrazení rozdělení, ale vzhledem k tomu, že chceme hypotéze přidělit konkrétní pravděpodobnost, je Pmf přirozenou volbou.

Každá hodnota v Pmf představuje hypotézu; například hodnota 0,5 znamená hypotézu, že λ je 0,5. V apriorním rozdělení mají všechny hypotézy stejnou pravděpodobnost. Apriorní pravděpodobnost proto můžeme konstruovat takto:

```
def MakeUniformSuite(low, high, steps):
    hypos = [low + (high-low) * i / (steps-1.0) for i in range(steps)]
    pmf = Pmf.MakePmfFromList(hypos)
    return pmf
```

Tato funkce vytvoří a vrátí Pmf, která představuje soubor souvisejících hypotéz, označovaný jako **suite**. Každá hypotéza má stejnou pravděpodobnost, takže se jedná o **rovnoměrné** rozdělení.

Argumenty `low` a `high` specifikují rozpětí hodnot; `steps` je počet hypotéz.

K provedení aktualizace vezmeme soubor hypotéz (suite) a soubor důkazů (evidence):

```
def Update(suite, evidence):
    for hypo in suite.Values():
        likelihood = Likelihood(evidence, hypo)
        suite.Mult(hypo, likelihood)
    suite.Normalize()
```

Pro každou hypotézu v suite vynásobíme apriorní pravděpodobnost věrohodností důkazu. Pak suite normalizujeme.

V této funkci musí být `suite` Pmf, avšak `evidence` může být jakéhokoliv typu, za předpokladu, že `Likelihood` ví, jak jej interpretovat.

Zde je věrohodnostní funkce (likelihood function):

```
def Likelihood(evidence, hypo):
    param = hypo
    likelihood = 1
    for x in evidence:
        likelihood *= ExpoPdf(x, param)

    return likelihood
```

V `Likelihood` předpokládáme, že `evidence` je výběr z exponenciálního rozdělení a vypočteme součin z předchozího oddílu.

`ExpoPdf` vyhodnotí PDF exponenciálního rozdělení pro `x`:

```
def ExpoPdf(x, param):
    p = param * math.exp(-param * x)
    return p
```

Když to vše dáme dohromady, zde je kód, který vytvoří apriorní pravděpodobnost a vypočítá aposteriorní pravděpodobnost

```
evidence = [2.675, 0.198, 1.152, 0.787, 2.717, 4.269]
prior = MakeUniformSuite(0.5, 1.5, 100)
posterior = prior.Copy()
Update(posterior, evidence)
```

Kód používaný v tomto oddílu si můžete stáhnout z: `http://thinkstats.com/estimate.py`.

Když uvažuji o bayesovském odhadu, představuji si místnost plnou lidí, kde každý člověk má jinou domněnku ohledně něčeho, co se snažíte odhadnout. Takže v tomto příkladu má každý z nich jinou domněnku o správné hodnotě λ.

Na začátku každý člověk přisuzuje své vlastní hypotéze určitý stupeň spolehlivosti. Poté, co jsou konfrontováni s důkazy, provede každý aktualizaci této spolehlivosti na základě $P(E \mid H)$, pravděpodobnosti důkazu za předpokladu jejich hypotézy.

Věrohodnostní funkce (likelihood function) většinou vypočítá pravděpodobnost, která může být maximálně 1, takže na začátku se každému spolehlivost sníží (nebo zůstane stejná). Pak ale provedeme normalizaci, čímž se spolehlivost každému zvýší.

Čistý výsledek pak je, že některým spolehlivost stoupne a jiným klesne, v závislosti na relativní pravděpodobnosti jejich hypotézy.

8.8 Cenzurovaná data

Následující problém je uveden v Kapitole 3 knihy Davida MacKaye *Information Theory, Inference and Learning Algorithms*, kterou si můžete stáhnout z: `http://www.inference.phy.cam.ac.uk/mackay/itprnn/ps/`.

> Nestabilní částice jsou emitovány ze zdroje a rozpadají se ve vzdálenosti x, což je reálné číslo, které má exponenciální pravděpodobnostní rozdělení s [parametrem] λ. Případy, kdy dojde k rozpadu, je možné pozorovat, pouze pokud k nim dojde v okně sahajícím od x = 1 cm do x = 20 cm. Na místech $\{ x_1, \ldots, x_N\}$ je pozorováno n případů rozpadu. Jaká je hodnota λ?

Toto je příklad problému s odhadem, kde se vyskytují **cenzurovaná data**; neboli když víme, že některá data jsou systematicky vyloučena.

Jednou z předností bayesovského odhadu je, že si umí s cenzurovanými daty relativně snadno poradit. Můžeme použít metodu z předchozího oddílu s jedinou změnou – musíme nahradit PDF_{expo} podmíněným rozdělením:

$$\text{PDF}_{cond}(\lambda, x) = \lambda e^{-\lambda x} / Z(\lambda)$$

pro $1 < x < 20$, a jinak 0, s

$$Z(\lambda) = \int_1^{20} \lambda e^{-\lambda x}\, dx = e^{-\lambda} - e^{-20\lambda}$$

Možná si na Z(λ) vzpomenete z Cvičení 6.5. Říkal jsem vám, abyste si výsledek nechali někde po ruce.

Cvičení 8.4 Stáhněte si `http://thinkstats.com/estimate.py`, která obsahuje kód z předešlého oddílu, a vytvořte kopii nazvanou `decay.py`.

Upravte `decay.py`, aby vypočítala aposteriorní rozdělení λ pro vzorek X = {1,5, 2, 3, 4, 5, 12}. Pro apriorní rozdělení můžete použít rovnoměrné rozdělení mezi 0 a 1,5 (vyjma 0).

Řešení tohoto problému si můžete stáhnout zde: `http://thinkstats.com/decay.py`.

Cvičení 8.5 V senátních volbách ve státě Minnesota roku 2008 byl konečný výsledek hlasování 1 212 629 hlasů pro Ala Frankena a 1 212 317 hlasů pro Norma Colemana. Franken byl prohlášen za vítěze, ale jak upozornil Charles Seife v *Proofiness*, rozdíl v počtu získaných hlasů byl mnohem menší než tolerance chyby, a tak měla být jako výsledek vyhlášena rovnost hlasů.

Budeme-li předpokládat, že existuje možnost, že se hlas může ztratit nebo být započten dvakrát, jaká je pravděpodobnost, že Coleman ve skutečnosti získal více hlasů?

Nápověda: Pro modelování chybového procesu budete muset doplnit některé údaje.

8.9 Problém s lokomotivou

Problém s lokomotivou je klasický problém odhadu známý také pod názvem „Problém německého tanku“. Zde uvádím verzi, která se objevuje v Mostellerově knize *Fifty Challenging Problems in Probability*:

> „Železnice čísluje své lokomotivy v pořadí 1..N. Jednoho dne uvidíte lokomotivu s číslem 60. Odhadněte, kolik lokomotiv má daná železnice.“

Než si přečtete zbytek tohoto oddílu, pokuste se zodpovědět tyto otázky:

1. Jaká je u konkrétního odhadu, $\hat{N}$, věrohodnost důkazu, $P(E \mid \hat{N})$? Jaký je maximálně věrohodný odhad?

2. Jestliže uvidíme vlak i, zdá se rozumné odhadovat nějaký násobek i, proto předpokládejme $\hat{N} = ai$. Jaká hodnota a minimalizuje střední kvadratickou chybu?

3. Budete-li nadále předpokládat, že $\hat{N} = ai$ Jste schopni určit hodnotu a, která zajistí, aby $\hat{N}$ fungoval jako nezkreslený odhad?

4. Pro jakou hodnotu N je 60 průměrná hodnota?

5. Jaké je bayesovské aposteriorní rozdělení za předpokladu apriorního rozdělení, které je stejnoměrné od 1 do 200?

K dosažení těch nejlepších výsledků byste měli nad těmito otázkami strávit nějaký čas, než budete pokračovat.

Pro konkrétní odhad $\hat{N}$ je pravděpodobnost spatření vlaku i $1/\hat{N}$, jestliže $i \leq \hat{N}$, a jinak 0. Takže MLE je $\hat{N} = i$. Jinými slovy, pokud uvidíte vlak 60 a chcete maximalizovat svoje šance na správné zodpovězení otázky, pak byste měli odhadovat, že vlaků je 60.

Tento odhad si ale nevede příliš dobře, pokud jde o MSE. Uděláme lépe, když zvolíme $\hat{N} = ai$; zbývá nám jen najít vhodnou hodnotu pro a.

Předpokládejte, že vlaků je ve skutečnosti N. Pokaždé, když hrajeme tuto hru na odhad, uvidíme vlak i a odhadneme ai, takže kvadratická chyba je $(ai - N)^2$.

Pokud budeme hrát hru Nkrát a uvidíme každý vlak jednou, střední kvadratická chyba je

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(ai - N)^2$$

Abychom minimalizovali MSE, provedeme derivaci vzhledem k a:

$$\frac{dMSE}{da} = \frac{1}{N}\sum_{i=1}^{N}2i(ai - N) = 0$$

A vyřešíme pro a.

$$a = \frac{3N}{2N+1}$$

Na první pohled se toto nejeví jako příliš užitečné, protože N se objevuje na pravé straně, z čehož vyplývá, že potřebujeme znát N, abychom mohli zvolit a, avšak kdybychom znali N, vůbec bychom nepotřebovali odhad.

Nicméně pro velké hodnoty N konverguje optimální hodnota pro a k $3/2$, takže bychom mohli zvolit $\hat{N} = 3i/\,2$.

Abychom nalezli nezkreslený odhad, můžeme vypočítat střední chybu (mean error – ME):

$$ME = \frac{1}{N}\sum_{i=1}^{N}(ai - N)$$

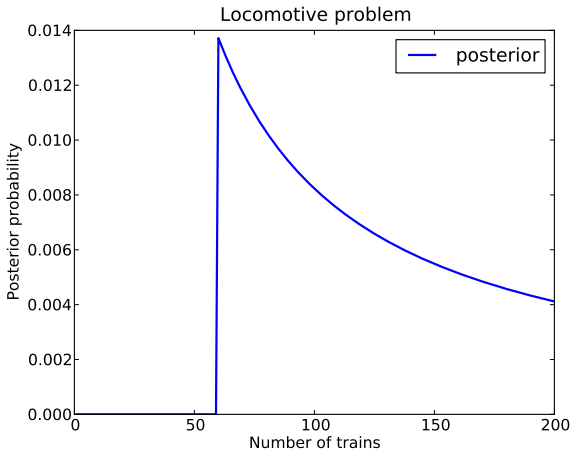
A najít hodnotu a, při níž ME = 0, tedy

$$a = \frac{2N}{N+1}$$

Pro velké hodnoty N konverguje a k 2, takže bychom mohli zvolit $\hat{N} = 2i$.

Dosud jsme vygenerovali tři odhady, i, $3i/2$ a $2i$, jejichž vlastnosti jsou maximalizace věrohodnosti, minimalizace kvadratické chyby a nezkreslenost.

Přesto existuje ještě další způsob, jak vygenerovat odhad, spočívající ve výběru hodnoty, při níž je průměr základního souboru roven výběrovému



Obrázek 8.1: Aposteriorní rozdělení počtu vlaků.

průměru. Jestliže uvidíme vlak i, je výběrový průměr přesně i; základní soubor vlaků, který má stejný průměr, je $\hat{N} = 2i - 1$.

Konečně pak, abychom vypočetli bayesovské aposteriorní rozdělení, vypočteme

$$P(H_n \mid i) = \frac{P(i \mid H_n)P(H_n)}{P(i)}$$

Kde H_n je hypotéza, že existuje n vlaků, a i je důkaz: viděli jsme vlak i. Opět $P(i \mid H_n)$ je $1/n$, jestliže $i < n$, a jinak 0. Normalizační konstanta, $P(i)$, je součet čitatelů pro každou hypotézu.

Jestliže je apriorní rozdělení rovnoměrné od 1 do 200, začneme s 200 hypotézami a vypočteme pravděpodobnost každé z nich. Provedení si můžete stáhnout z: `http://thinkstats.com/locomotive.py`. Obrázek 8.1 ukazuje, jak vypadá výsledek.

90% credible interval pro toto aposteriorní rozdělení je [63, 189], což je stále dost široký interval. To, že jsme viděli jeden vlak, neposkytuje silný důkaz pro žádnou z hypotéz (i když vylučuje hypotézy s $n < i$).

Začneme-i s jiným apriorním rozdělením, je aposteriorní rozdělení odlišné, což pomáhá vysvětlit, proč jsou ostatní odhady tak různorodé.

Jedním ze způsobů, jak uvažovat o různých odhadech, je to, že jsou implicitně založeny na odlišných apriorních rozděleních. Jestliže existují dostatečné důkazy, které zastíní apriorní rozdělení, pak mají všechny odhady tendenci konvergovat; jinak, jako v tomto případě, neexistuje žádný odhad, který by v sobě spojoval všechny vlastnosti, které bychom potřebovali.

Cvičení 8.6 Zobecněte `locomotive.py`, abyste se vypořádali s případem, kdy uvidíte více než jeden vlak. Změna by se měla dotknout pouze několika řádek kódu.

Zjistěte, jestli byste byli schopni zodpovědět ostatní otázky týkající se případu, kdy uvidíte více než jeden vlak. Diskusi o tomto problému a několik řešení najdete zde: `http://wikipedia.org/wiki/German_tank_problem`.

8.10 Glosář

odhadování (estimation): Proces vyvození parametrů rozdělení ze vzorku.

odhad (estimator): Statistika používaná k odhadu parametru.

střední kvadratická chyba (mean squared error): Míra chyby odhadu.

maximálně věrohodný odhad (maximum likelihood estimator): Odhad, který vypočte bodový odhad s nejvyšší věrohodností.

zkreslení (bias): Tendence odhadu být nad nebo pod skutečnou hodnotou parametru, při zprůměrování pro opakované vzorky.

bodový odhad (point estimate): Odhad vyjádřený jako jediná hodnota.

interval spolehlivosti (confidence interval): Odhad vyjádřený jako interval s konkrétní pravděpodobností, že obsahuje skutečnou hodnotu parametru.

credible interval: Jiný název pro bayesovský interval spolehlivosti.

cenzurovaná data (censored data): Soubor dat vybraný způsobem, který systematicky vylučuje některá data.

Kapitola 9

Korelace

9.1 Standardní skóre

V této kapitole se budeme zabývat vztahy mezi proměnnými. Například tušíme, že výška souvisí s váhou. Vyšší lidé obvykle také více váží. **Korelace** popisuje tento typ vztahu.

Měření korelace je ztíženo tím, že proměnné, které chceme porovnávat, nemusí být vyjádřeny ve stejných jednotkách. Například výška může být uvedena v centimetrech, zatímco váha v kilogramech. A i když jsou vyjádřeny ve stejných jednotkách, nejsou ze stejných rozdělení.

Tyto problémy se běžně řeší dvěma způsoby:

1. Transformací všech hodnot na **standardní skóre**. Výsledkem je Pearsonův korelační koeficient.

2. Transformací všech hodnot na jejich percentilové pořadí. Výsledkem je Spearmanův koeficient.

Jestliže X je řada hodnot, x_i, můžeme provést transformaci na standardní skóre odečtením průměru a vydělením směrodatnou odchylkou: $z_i = (x_i - \mu) / \sigma$.

Čitatel představuje odchylku: vzdálenost od průměru. Vydělení směrodatnou odchylkou σ **normalizuje** tuto odchylku, takže hodnoty Z jsou bezrozměrné (bez jednotek) a jejich rozdělení má průměr 0 a rozptyl 1.

Má-li X normální rozdělení, pak ho má také Z; jestliže je však X sešikmené nebo obsahuje odlehlé hodnoty, pak totéž platí i pro Z. V těchto případech

je robustnější použít percentilová pořadí. Obsahuje-li R percentilová pořadí hodnot v X, je rozdělení R stejnoměrné mezi 0 a 100, bez ohledu na rozdělení X.

9.2 Kovariance

Kovariance vyjadřuje míru tendence dvou proměnných ke společné variabilitě. (Lze použít u proměnných, které mají stejné jednotky. Kovariance standardizovaná na variabilitu dat je pak korelací.) Máme-li dvě řady, X a Y, jejich odchylky od průměru jsou

$$dx_i = x_i - \mu_X$$

$$dy_i = y_i - \mu_Y$$

kde μ_X je průměr X a μ_Y průměr Y. Jestliže se X a Y mění společně, jejich odchylky mají tendenci ke stejnému znaménku.

Jestliže je společně vynásobíme, součin je kladný, mají-li odchylky stejné znaménko, a záporný, mají-li opačné znaménko. Sečtením součinů tak získáme míru jejich tendence k tomu, aby se měnili společně.

Kovariance je průměr těchto součinů:

$$Cov(X, Y) = \frac{1}{n} \sum dx_i dy_i$$

kde n je délka příslušných dvou řad (musí být stejně dlouhé).

Kovariance je užitečná u některých výpočtů, ale zřídkakdy se uvádí jako souhrnná statistická charakteristika, protože je obtížné ji interpretovat. Mimo jiné jsou její jednotky součinem jednotek X aY. Takže kovariance váhy a výšky by mohla být vyjádřena v jednotkách kilogram-metr, což nemá příliš velký význam.

Cvičení 9.1 Napište funkci s názvem `Cov`, která přijme dva seznamy a vypočítá jejich kovarianci. K otestování vaší funkce vypočítejte kovarianci seznamu se sebou samým a ověřte, že Cov(X, X) = Var(X).

Řešení si můžete stáhnout z `http://thinkstats.com/correlation.py`.

9.3 Korelace

Jedním možným řešením je vydělit odchylky směrodatnou odchylkou σ, čímž získáme standardní skóre, a vypočíst součin standardních skóre:

$$p_i = \frac{(x_i - \mu_X)}{\sigma_X} \frac{(y_i - \mu_Y)}{\sigma_Y}$$

Průměr těchto součinů je

$$\rho = \frac{1}{n} \sum p_i$$

Nebo můžeme přepsat ρ vytknutím σ_X a σ_Y:

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

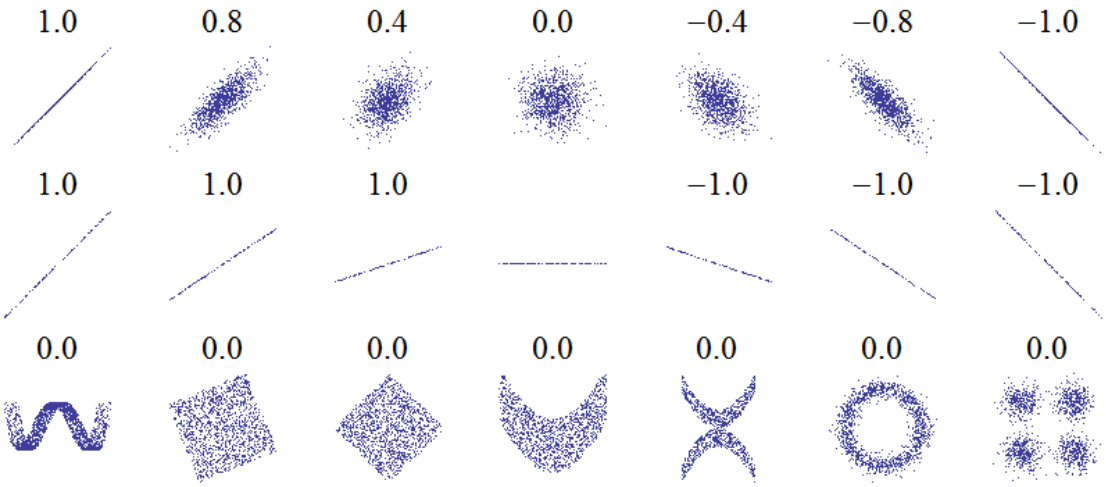
Tato hodnota se označuje jako **Pearsonova korelace** a je pojmenována po Karlu Pearsonovi, významném statistikovi, který se podílel na formování této vědní disciplíny. Tento ukazatel se snadno počítá a snadno interpretuje. Protože standardní skóre jsou bezrozměrná, stejně tak je tomu i v případě ρ.

Pearsonova korelace se vždy pohybuje mezi -1 a +1 (včetně obou). Její velikost ukazuje na sílu korelace. Jestliže ρ = 1, proměnné dokonale korelují, což znamená, že pokud znáte jednu z nich, můžete dokonale předpovědět druhou z nich. Totéž platí, jestliže ρ = -1. To znamená, že mezi proměnnými existuje negativní (nepřímá) korelace, avšak pro účely predikce je negativní korelace právě tak dobrá, jako pozitivní (přímá) korelace.

Většina korelací v reálném světě není absolutní, ale přesto je tento vztah užitečný. Například jestliže znáte něčí výšku, mohli byste odhadnout váhu daného člověka. Možná se netrefíte úplně přesně, ale pořád bude váš odhad lepší, než kdybyste vůbec neznali jeho výšku. Pearsonova korelace je mírou toho, o kolik lepší je takový odhad.

Jestliže tedy platí ρ = 0, znamená to, že mezi proměnnými neexistuje žádný vztah? Bohužel nikoliv. Pearsonova korelace měří pouze *lineární* vztahy. Existuje-li mezi proměnnými nelineární vztah, ρ nevyjadřuje plně sílu dané závislosti.

Obrázek 9.1 je převzatý z `http://wikipedia.org/wiki/Correlation_and_dependence`. Ukazuje bodové grafy a korelační koeficienty pro několik pečlivě sestavených souborů dat.



Obrázek 9.1: Příklady souborů dat se škálou korelací.

Horní řada ukazuje lineární vztahy se škálou korelací. Na základě této řady si můžete vytvořit představu o tom, jak vypadají různé hodnoty ρ. Druhá řada ukazuje dokonalé korelace s různými sklony, což dokládá, že korelace nemá žádný vztah ke sklonu (o odhadování sklonu se záhy zmíním). Třetí řada ukazuje proměnné, které jsou evidentně vzájemně provázané, ale protože jejich vztah je nelineární, korelační koeficient se rovná 0.

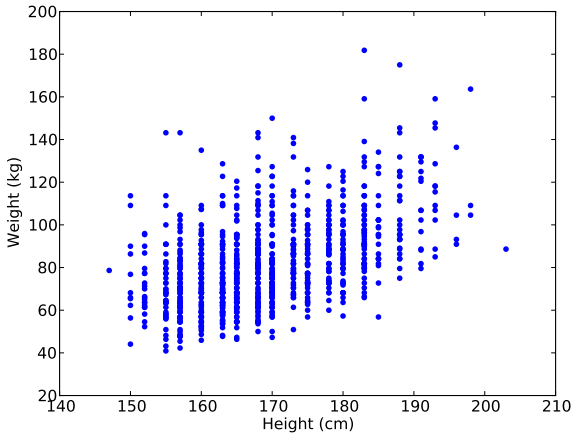
Ponaučení, které z toho plyne, je, že byste se měli vždy podívat na bodový graf vašich dat před tím, než slepě vypočítáte korelační koeficient.

Cvičení 9.2 Napište funkci nazvanou `Corr`, která přijme dva seznamy a vypočítá jejich korelaci. Nápověda: Použijte `thinkstats.Var` a funkci `Cov`, kterou jste vytvořili v předešlém cvičení.

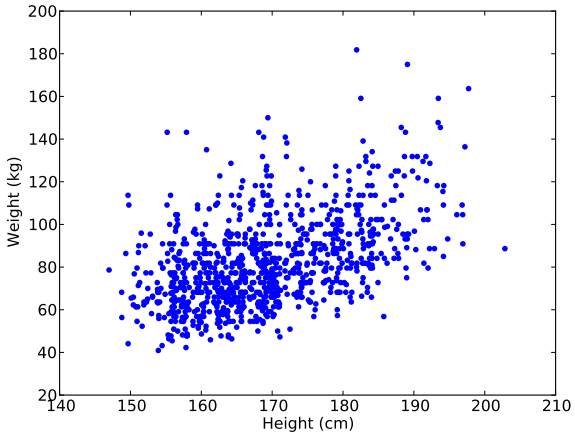
K otestování vaší funkce vypočtěte kovarianci seznamu se sebou samým a ověřte, že Corr(X, X) je 1. Řešení si můžete stáhnout z `http://thinkstats.com/correlation.py`.

9.4 Vytváření bodových grafů v pyplot

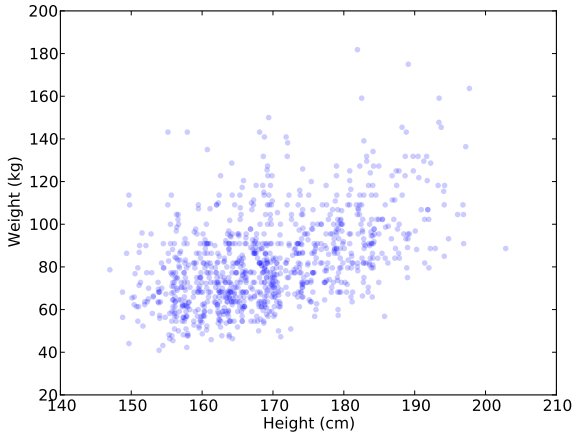
Nejjednodušší způsob, jak prověřit vztah mezi dvěma proměnnými, je sestavit bodový graf. Vytvoření dobrého bodového grafu ale není vždy snadné. Jako příklad vytvořím graf zobrazující vztah mezi váhou a výškou u respondentů z šetření BRFSS (viz Oddíl 4.5). `pyplot` obsahuje funkci nazvanou `scatter`, která vytváří bodové grafy:



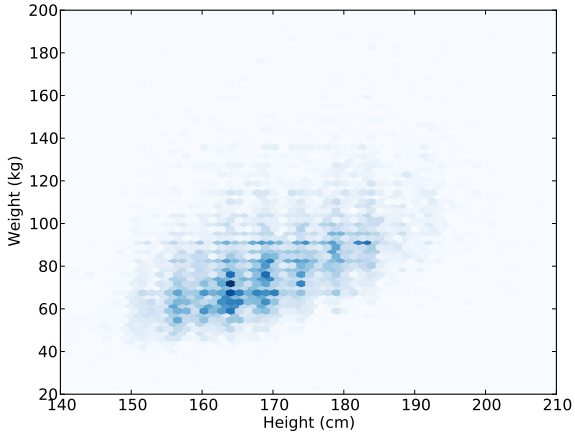
Obrázek 9.2: Jednoduchý bodový graf znázorňující vztah mezi váhou a výškou u respondentů v rámci BRFSS.



Obrázek 9.3: Bodový graf s daty, na která byla aplikována metoda jitter.



Obrázek 9.4: Bodový graf, ve kterém byla uplatněna metoda jitter a transparency.



Obrázek 9.5: Bodový graf s daty rozdělenými do tříd za použití `pyplot.hexbin`.

```
import matplotlib.pyplot as pyplot
pyplot.scatter(heights, weights)
```

Na Obrázku 9.2 je vidět výsledek. Není překvapením, že ukazuje na pozitivní korelaci – vyšší lidé jsou obvykle také těžší. Není to ale zrovna nejlepší způsob zobrazení dat, protože data vytváří sloupcové shluky. Problém spočívá v tom, že údaje o hmotnosti byly zaokrouhleny na nejbližší palec, pak převedeny na centimetry a znovu zaokrouhleny. V průběhu tohoto procesu se některé informace ztratily.

Tyto informace už nedostaneme zpátky, ale dopad na bodový graf můžeme minimalizovat pomocí metody **jitter**, aplikované na data. Podstatou této metody je dodání náhodného šumu, abychom vyvážili vliv zaokrouhlení. Protože byly naměřené hodnoty zaokrouhleny na nejbližší palec, mohou být nepřesné až o 0,5 palce nebo 1,3 cm. Proto jsem dodal stejnoměrný šum v rozpětí -1,3 až 1,3:

```
jitter = 1.3
heights = [h + random.uniform(-jitter, jitter) for h in heights]
```

Na Obrázku 9.3 je zobrazen výsledek. Na základě použití metody jitter je tvar vztahu zřejmější. Obecně byste měli k uplatnění metody jitter přistoupit pouze pro účely vizualizace a datům upraveným pomocí metody jitter se vyhnout, pokud je chcete analyzovat.

Dokonce ani když použijete metodu jitter, není to ten nejlepší způsob zobrazení dat. V grafu je spousta bodů, které se překrývají, což vede k tomu, že data v hustě pokrytých částech jsou skrytá a naopak nepřiměřenou pozornost strhávají odlehlé hodnoty.

Toto můžeme vyřešit pomocí parametru `alpha`, díky němuž se stanou body částečně transparentními:

```
pyplot.scatter(heights, weights, alpha=0.2)
```

Na Obrázku 9.4 je zobrazen výsledek. Překrývající se datové body vypadají tmavší, takže tmavost je proporcionální k hustotě. V této verzi grafu je patrný pozorovaný artefakt – horizontální přímka poblíž 90 kg nebo 200 liber. Vzhledem k tomu, že data vychází z údajů v librách, které o sobě poskytli respondenti sami, jako nejpravděpodobnější se nabízí vysvětlení, že některé hodnoty byly zaokrouhleny (dost možná směrem dolů).

Použití metody transparency funguje dobře u nepříliš velkých datových souborů, ale na tomto obrázku je zobrazeno pouze prvních 1000 záznamů v rámci BRFSS, z celkových 414509.

Pro větší soubory dat je jednou z možností hexbin graf, který rozdělí graf do hexagonálních tříd a každou třídu vybarví podle toho, kolik datových bodů do ní patří. `pyplot` obsahuje funkci s názvem `hexbin`:

```
pyplot.hexbin(heights, weights, cmap=matplotlib.cm.Blues)
```

Obrázek 9.5 ukazuje výsledek za použití mapy modré barvy. Předností hexbin grafu je, že dobře znázorňuje tvar vztahu a je efektivní při práci s velkými soubory dat. Nevýhodou je, že odlehlé hodnoty se stávají neviditelnými.

Ponaučení, které z toho plyne, je, že není snadné vytvořit bodový graf, který není potenciálně zavádějící. Kód pro uvedené obrázky si můžete stáhnout z `http://thinkstats.com/brfss_scatter.py`.

9.5 Spearmanova pořadová korelace

Pearsonova korelace funguje dobře, jestliže mezi proměnnými existuje lineární vztah a jestliže se jedná o proměnné s přibližně normálním rozdělením. Není ale robustní, jsou-li přítomny odlehlé hodnoty.

Toto zjištění dobře dokládá Anscombeho kvartet, který obsahuje čtyři soubory dat se stejnou korelací. V jednom souboru existuje lineární vztah s náhodným šumem, druhý představuje nelineární vztah, další je dokonalý vztah s odlehlou hodnotou a u posledního neexistuje žádný vztah kromě artefaktu způsobeného odlehlou hodnotou. O tomto kvartetu si můžete přečíst více na `http://wikipedia.org/wiki/Anscombe's_quartet`.

Spearmanova pořadová korelace představuje alternativu, která zmírňuje působení odlehlých hodnot a sešikmených rozdělení. K výpočtu Spearmanovy korelace musíme spočítat **pořadí** každé hodnoty, které je jejím indexem v rámci setříděného vzorku. Například ve vzorku {7, 1, 2, 5} má hodnota 5 pořadí 3, protože pokud jednotlivé členy seřadíme, pak se nachází na třetím místě. Pak spočítáme Pearsonovu korelaci pro pořadí.

Alternativou ke Spearmanově korelaci je použít transformaci, která data více přiblíží normálnímu rozdělení, a pak vypočítat Pearsonovu korelaci pro transformovaná data. Například mají-li data přibližně logaritmicko-normální rozdělení, můžete vzít logaritmus každé hodnoty a vypočítat korelaci logaritmů.

Cvičení 9.3 Napište funkci, která přijme řadu a vrátí seznam uvádějící pořadí jednotlivých členů. Například pro řadu {7, 1, 2, 5} bude výsledek { 4, 1, 2, 3}.

Vyskytuje-li se stejná hodnota víckrát, striktně korektní řešení by bylo přiřadit každé z nich průměr jejich pořadí. Pokud to však budeme ignorovat a pořadí jim přiřadíme v náhodném sledu, chyba obvykle bývá jen malá.

Napište funkci, která přijme dvě řady (o stejné délce) a vypočte jejich Spearmanův pořadový koeficient. Řešení si můžete stáhnout z `http://thinkstats.com/correlation.py`.

Cvičení 9.4 Stáhněte si `http://thinkstats.com/brfss.py` a `http://thinkstats.com/brfss_scatter.py`. Spusťte je a ujistěte se, že jste schopni číst data z BRFSS a generovat bodové grafy.

Porovnáte-li bodové grafy s Obrázkem 9.1, jakou hodnotu Pearsonovy korelace očekáváte? A jaká hodnota vám vyšla?

Protože rozdělení váhy dospělých je logaritmicko-normální, vyskytují se v něm odlehlé hodnoty, které ovlivňují korelaci. Pokuste se graficky znázornit vztah log(váha) versus výška a vypočíst Pearsonovu korelaci pro transformovanou proměnnou.

Nakonec pak vypočtěte Spearmanovu pořadovou korelaci pro váhu a výšku. Který koeficient je podle vás nejlepším ukazatelem síly tohoto vztahu? Řešení si můžete stáhnout zde: `http://thinkstats.com/brfss_corr.py`.

9.6 Metoda nejmenších čtverců

Korelační koeficienty měří sílu a znaménko vztahu, ale ne sklon. Pro odhad sklonu existuje několik možností. Nejčastěji se používá **lineární regrese metodou nejmenších čtverců**. „Lineární regrese“ je proložení dat přímkou, která slouží k modelování vztahu mezi proměnnými. Metoda „nejmenších čtverců“ je metoda, která minimalizuje střední kvadratickou chybu (MSE) mezi přímkou a daty[1].

Předpokládejme, že máme řadu bodů Y, kterou chceme vyjádřit jako funkci jiné řady X. Jestliže mezi X a Y existuje lineární vztah s konstantou α a sklonem β, očekáváme, že každé y_i bude přibližně $\alpha + \beta\, x_i$.

Pokud se ale nejedná o dokonalou korelaci, je tato predikce pouze přibližná. Odchylka, neboli **reziduum** je

$$\varepsilon_i = (\alpha + \beta x_i) - y_i$$

[1] Viz `http://wikipedia.org/wiki/Simple_linear_regression`.

Reziduum může být způsobeno náhodnými faktory, jako například chybou měření, nebo nenáhodnými faktory, které jsou neznámé. Například pokoušíme-li se predikovat váhu jako funkci výšky, neznámými faktory mohou být způsob stravování, cvičení a tělesný typ.

Jestliže použijeme nesprávné parametry α a β, rezidua budou větší, a tak intuitivně dává smysl, že by požadované parametry měly být takové, které minimalizují rezidua.

Jako obvykle bychom mohli minimalizovat absolutní hodnotu reziduí, nebo jejich druhých mocnin, nebo třetích mocnin atd. Nejběžnější volbou je minimalizovat součet kvadratických reziduí

$$\min_{\alpha,\beta} \sum \varepsilon_i^2$$

Proč? Existují pro to tři dobré a jeden špatný důvod:

- Umocnění na druhou má zřejmý efekt – s kladnými i zápornými rezidui je nakládáno stejně, což většinou chceme.
- Umocněním na druhou získávají větší rezidua na váze, avšak ne natolik, aby největší reziduum vždy dominovalo.
- Jestliže jsou rezidua nezávislá na x, náhodná a mají normální rozdělení s $\mu = 0$ a konstantní (ale neznámou) σ, pak je metoda nejmenších čtverců také maximálně věrohodným odhadem α a β.[2]
- Hodnoty $\hat{\alpha}$ a $\hat{\beta}$, které minimalizují kvadratická rezidua, mohou být efektivně vypočteny.

Poslední důvod byl opodstatněný ve chvíli, kdy byla výpočetní účinnost důležitější než volba nejvhodnější metody pro řešený problém. To už ale neplatí, a tak stojí za to zvážit, jestli kvadratická rezidua jsou opravdu tím, co je potřeba minimalizovat.

Jestliže například používáte hodnoty X k predikování hodnot Y, příliš vysoký odhad by mohl být lepší (nebo horší), než příliš nízký odhad. V takovém případě by se vám mohlo hodit vypočíst nákladovou funkci, cost(ε_i), a minimalizovat celkové náklady.

Nicméně výpočet metodou nejmenších čtverců je rychlý, snadný a často také dostatečně dobrý. Podívejme se tedy na postup:

[2] Viz Press et al., *Numerical Recipes in C*, Chapter 15 at `http://www.nrbook.com/a/bookcpdf/c15-1.pdf`.

1. Vypočtěte výběrové průměry, $\bar{x}$ a $\bar{y}$, rozptyl X a kovarianci X a Y.
2. Odhadovaný sklon je
$$\hat{\beta} = \frac{Cov(X,Y)}{Var(X)}$$
3. A konstanta je
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Pokud vás zajímá, jak je toto odvozeno, přečtěte si `http://wikipedia.org/wiki/Numerical_methods_for_linear_least_squares`.

Cvičení 9.5 Napište funkci nazvanou `LeastSquares`, která přijme X a Y a vypočítá $\hat{\alpha}$ a $\hat{\beta}$. Řešení si můžete stáhnout zde: `http://thinkstats.com/correlation.py`.

Cvičení 9.6 Opět za použití dat z šetření BRFSS vypočtěte lineární regresi metodou nejmenších čtverců pro log(váha) versus výška. Řešení si můžete stáhnout z `http://thinkstats.com/brfss_corr.py`.

Cvičení 9.7 Rozdělení rychlosti větru na konkrétním místě určuje hustotu větrné energie, což představuje horní hranici průměrného množství energie, kterou je větrná turbína na daném místě schopna vygenerovat. Podle některých zdrojů k modelování empirických rozdělení rychlosti větru dobře slouží Weilbullovo rozdělení (viz `http://wikipedia.org/wiki/Wind_power#Distribution_of_wind_speed`).

Pro posouzení toho, zda je konkrétní místo vhodné pro umístění větrné turbíny, můžete na místě instalovat anemometr a měřit po určitou dobu rychlost větru. Je však obtížné změřit přesně chvost rozdělení rychlosti větru, protože, jak vyplývá z povahy věci, jevy na chvostu nenastávají příliš často.

Jedním ze způsobů, jak si s tímto problémem poradit, je použít měření k odhadu parametrů Weilbullova rozdělení a pak integrovat přes spojitou PDF za účelem výpočtu hustoty větrné energie.

K odhadu parametrů Weilbullova rozdělení můžeme použít transformaci ze Cvičení 4.6 a pak použít lineární regresi k nalezení sklonu a konstanty transformovaných dat.

Napište funkci, která přijme vzorek z Weilbullova rozdělení a odhadne jeho parametry.

Nyní napište funkci, která přijme parametry Weilbullova rozdělení rychlosti větru a vypočte průměrnou hustotu větrné energie (zřejmě bude nutné provést nějaký průzkum, abyste se orientovali v této části).

9.7 Dobrá shoda

Jestliže jsme nalezli shodu mezi lineárním modelem a daty, mohlo by nás zajímat, o jak dobrou shodu se jedná. To ale záleží na tom, jakému má sloužit účelu. Jedním možným způsobem hodnocení modelu je jeho predikční síla.

V kontextu predikce se veličina, kterou se pokoušíme odhadnout, nazývá **závisle proměnná** a veličina, kterou používáme k odhadu, se nazývá **nezávisle proměnná**.

Abychom změřili prediktivní sílu modelu, můžeme vypočítat **determinační koeficient**, běžněji známý jako „R na druhou“:

$$R^2 = 1 - \frac{Var(\varepsilon)}{Var(Y)}$$

K pochopení toho, co znamená R^2, uvažujte (znovu) situaci, kdy se pokoušíte odhadnout váhu nějakého člověka. Pokud byste o dané osobě nevěděli vůbec nic, nejlepší strategií by bylo odhadnout $\bar{y}$; v tom případě by střední kvadratická chyba (MSE) vašich odhadů byla Var(Y):

$$MSE = \frac{1}{n}\sum(\bar{y} - y_i)^2 = Var(Y)$$

Kdybych vám ale řekl výšku daného člověka, odhadli byste $\hat{\alpha} + \hat{\beta}\ x_i$; v tom případě by vaše MSE byla Var(ε).

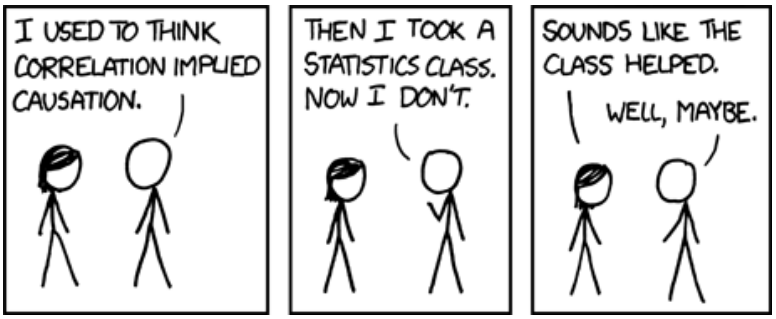
$$MSE = \frac{1}{n}\sum(\hat{\alpha} + \hat{\beta}x_i - y_i)^2 = Var(\varepsilon)$$

Výraz Var(ε)/Var(Y) tedy představuje poměr střední kvadratické chyby s nezávisle proměnnou a bez ní, což je podíl variability, kterou model nechává bez vysvětlení. Komplement, R^2, je pak podíl variability, kterou model vysvětluje.

Jestliže v modelu vychází, že $R^2 = 0{,}64$, mohli byste říci, že model vysvětluje 64 % variability, nebo pokud byste chtěli být přesnější, že redukuje střední kvadratickou chybu (MSE) vašich predikcí o 64 %.

V kontextu modelu lineární regrese metodou nejmenších čtverců se ukazuje, že mezi determinačním koeficientem a Pearsonovým korelačním koeficientem ρ existuje jednoduchý vztah:

$R^2 = \rho^2$



Obrázek 9.6: Převzato z `xkcd.com` autora Randalla Munroa.

Viz `http://wikipedia.org/wiki/Howzzat`!

Cvičení 9.8 Wechslerova škála inteligence dospělých (WAIS) slouží k měření inteligence. Skóre jsou kalibrována tak, že průměr a směrodatná odchylka v obecné populaci jsou 100 a 15.

Předpokládejme, že chcete předpovědět něčí WAIS skóre na základě skóre, kterého daný člověk dosáhl ve standardizovaném testu SAT. Podle jedné studie existuje mezi celkovým SAT skóre a WAIS skóre Pearsonova korelace rovná 0,72.

Pokud byste váš prediktor aplikovali na velký vzorek, jakou průměrnou kvadratickou chybu (MSE) vašich predikcí byste očekávali?

Nápověda: Jaká je MSE, jestliže budete vždy odhadovat 100?

Cvičení 9.9 Napište funkci s názvem `Residuals`, která přijme X, Y, $\hat{\alpha}$ a $\hat{\beta}$ a vrátí seznam ε_i.

Napište funkci nazvanou `CoefDetermination`, která přijme ε_i a Y a vrátí R^2. K otestování vašich funkcí, ověřte, že $R^2 = \rho^2$. Řešení si můžete stáhnout z `http://thinkstats.com/correlation.py`.

Cvičení 9.10 Za použití dat o výšce a váze z šetření BRFSS (ještě jednou) vypočtěte $\hat{\alpha}$, $\hat{\beta}$ a R^2. Kdybyste se pokoušeli odhadnout něčí váhu, jak moc by vám pomohlo, pokud byste znali výšku dané osoby? Řešení si můžete stáhnout z `http://thinkstats.com/brfss_corr.py`.

9.8 Korelace a kauzalita

Webový komiks `xkcd` poukazuje na to, jak obtížné je vyvodit příčinný vztah:

Obecně můžeme říci, že existence vztahu mezi dvěma proměnnými nevypovídá nic o tom, zda jedna způsobuje druhou, nebo naopak, nebo jestli obě způsobuje něco úplně jiného.

Toto pravidlo se dá shrnout větou „Korelace neimplikuje kauzalitu“, která je natolik obsažná, že má svoji vlastní stránku na Wikipedii: `http://wikipedia.org/wiki/Correlation_does_not_imply_causation`.

Takže co můžete udělat, abyste získali důkaz o kauzalitě?

1. Využijte čas. Jestliže A předchází B, pak A může způsobovat B, ale ne naopak (alespoň v souladu s obecně přijímaným chápáním kauzality). Pořadí jevů nám může pomoci vyvodit směr kauzality, ale nevylučuje možnost, že jak A, tak B je způsobeno něčím jiným.

2. Využijte nahodilosti. Jestliže rozdělíte velkou populaci náhodně do dvou skupin a spočítáte průměry téměř jakékoliv proměnné, očekáváte, že rozdíl bude malý. Toto je důsledek centrální limitní věty (takže podléhá stejným požadavkům).

   Jestliže jsou skupiny téměř identické ve všech proměnných až na jednu, můžete eliminovat nepravé vztahy.

   Toto funguje, i pokud nevíte, které jsou relevantní proměnné. Ale funguje to ještě lépe, pokud to víte, protože pak můžete zkontrolovat, že dané skupiny jsou identické.

Těmito myšlenkami je motivován **randomizovaný kontrolovaný test**, ve kterém jsou subjekty náhodně rozřazeny do dvou (nebo více) skupin: **experimentální** skupiny, která je podrobena nějaké intervenci, jako například podání nového léku, a **kontrolní skupiny**, která není podrobena žádné intervenci, nebo dostává jinou léčbu, jejíž účinky jsou známé.

Randomizovaný kontrolovaný test je nejspolehlivějším způsobem, jak demonstrovat příčinný vztah, a tvoří také základnu vědecky podložené medicíny (viz `http://wikipedia.org/wiki/Randomized_controlled_trial`).

Bohužel kontrolované testy jsou možné pouze v laboratorních vědách, medicíně a několika dalších disciplínách. V sociálních vědách jsou kontrolované experimenty vzácné, obvykle proto, že je nemožné je provést nebo jsou neetické.

Jednou z alternativ je hledání **přirozeného experimentu**, kdy jsou různé formy „léčby“ aplikovány na skupiny, které jsou jinak podobné. Přirozené experimenty s sebou nesou nebezpečí spočívající ve skutečnosti, že skupiny

se mohou navzájem lišit způsoby, které nejsou zřejmé. Více o tomto tématu si můžete přečíst na `http://wikipedia.org/wiki/Natural_experiment`.

V některých případech lze příčinný vztah vyvodit za použití **regresní analýzy**. Lineární regrese metodou nejmenších čtverců je jednoduchá forma regrese, která vysvětluje závisle proměnnou pomocí jedné nezávisle proměnné. Existují podobné techniky, které pracují s libovolným počtem nezávisle proměnných.

Těmito technikami se zde nebudu zabývat, ale existují také jednoduché způsoby, jak získat kontrolu nad nepravými vztahy. Například v rámci šetření NSFG jsme zjistili, že prvorozené děti mají tendenci k nižší hmotnosti než ostatní děti (viz Oddíl 3.6). Porodní hmotnost koreluje ale také s věkem matky a matky prvorozených dětí bývají obvykle mladší než matky ostatních dětí.

Je proto možné, že prvorozené děti mají nižší hmotnost, protože jejich matky jsou mladší. Abychom korigovali vliv věku, mohli bychom rozdělit matky do věkových skupin a porovnat porodní hmotnosti prvorozených dětí a ostatních dětí v každé věkové skupině.

Jestliže bude rozdíl mezi prvorozenými dětmi a ostatními dětmi v každé věkové skupině stejný jako byl v rámci úhrnných dat, můžeme z toho vyvodit závěr, že rozdíl není vázaný na věk. Pokud žádný rozdíl nezjistíme, můžeme otázku uzavřít s tím, že výsledek je zcela způsoben věkem. Nebo pokud bude rozdíl menší, můžeme kvantifikovat, jaký podíl má na výsledku věk.

Cvičení 9.11 Data v rámci NSFG zahrnují proměnnou nazvanou `agepreg`, která zaznamenává věk matky v okamžiku porodu. Vytvořte bodový graf, který zachytí věk matky a hmotnost dítěte pro každý porod živého dítěte. Pozorujete mezi nimi nějaký vztah?

Vypočtěte lineární regresi metodou nejmenších čtverců pro tyto proměnné. Jaké jsou jednotky odhadovaných parametrů $\hat{\alpha}$ a $\hat{\beta}$? Jak byste shrnuli tyto výsledky jednou nebo dvěma větami?

Vypočtěte průměrný věk matek prvorozených dětí a průměrný věk ostatních matek. Jaký rozdíl v průměrné porodní hmotnosti očekáváte na základě věkového rozdílu mezi skupinami? Jaká část skutečného rozdílu v porodní hmotnosti je vysvětlena věkovým rozdílem?

Řešení si můžete stáhnout z `http://thinkstats.com/agemodel.py`. Jste-li zvědaví na vícerozměrnou regresi, můžete zkusit spustit `http://thinkstats.com/age_lm.py`, která ukazuje, jak využít balík R pro statistické výpočty z Pythonu. To už je ale úplně jiná kniha.

9.9 Glosář

korelace (correlation): Popis závislosti mezi proměnnými.

normalizovat (normalize): Transformovat množinu hodnot tak, aby jejich průměr byl 0 a rozptyl 1.

standardní skóre (standard score): Hodnota, která byla normalizována.

kovariance (covariance): Míra tendence dvou proměnných ke společné variabilitě.

pořadí (rank): Index toho, kde se člen nachází v uspořádaném seznamu.

metoda nejmenších čtverců (least squares fit): Model souboru dat, který minimalizuje součet čtverců reziduí.

reziduum (residual): Ukazatel odchylky skutečné hodnoty od modelu.

závisle proměnná (dependent variable): Proměnná, kterou se snažíme predikovat nebo vysvětlit.

nezávisle proměnná (independent variable): Proměnná, kterou používáme k predikci závisle proměnné. Také se označuje jako vysvětlující proměnná.

determinační koeficient (coefficient of determination): Ukazatel toho, o jak dobrou shodu lineárního modelu se jedná.

randomizovaný kontrolovaný test (randomized controlled trial): Experimentální design, kdy jsou subjekty náhodně rozděleny do dvou skupin a různým skupinám je poskytnuta různá léčba.

léčba (treatment): Změna nebo intervence aplikovaná jedné skupině v kontrolovaném testu.

kontrolní skupina (control group): Skupina v kontrolovaném testu, která neobdrží žádnou léčbu, nebo obdrží léčbu, jejíž účinek je známý.

přirozený experiment (natural experiment): Experimentální design, který využívá přirozeného rozdělení subjektů do skupin způsoby, které jsou přinejmenším přibližně náhodné.