UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

KONŠTRUKCIA SELEKČNÝCH ÚMRTNOSTNÝCH TABULIEK

Semestrálny projekt z predmetu Demografická štatistika

2023

Adam Martinka

UNIVERZITA KOMENSKÉHO V BRATISLAVE

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

KONŠTRUKCIA SELEKČNÝCH ÚMRTNOSTNÝCH TABULIEK

Semestrálny projekt z predmetu Demografická štatistika

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1113 Matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky

Bratislava, 2023

Adam Martinka

Abstrakt

V tomto semestrálnom projekte sa pozrieme bližšie na selekčné úmrtnostné tabuľky a ich konštrukciu. Hlavným rozdielom oproti klasickej úmrtnostnej tabuľke, je v tom, že v tej selekčnej sa predpovedajú úmrtia a mnohé iné javy s úmrtiami spojené aj na základe iných charakteristík ako iba veku a pohlavia. Takých charakteristík je veľké množstvo ale pre upresnenie samotného názvu selekcie v názve projektu sa môže napríklad jednať o to či je daná osoba fajčiar alebo či daná osoba žije v meste alebo na vidieku, alebo podľa toho aké rizikové povolanie daná osoba vykonáva. V teoretickej časti si popíšeme ich definíciu a porovnáme ich s definíciou úmrtnostných tabuliek, z ktorých boli odvodené. Následne spíšeme par praktických úvah v čom spočíva ich výhoda v porovnaní s klasickými úmrtnostnými tabuľkami a vytvorime viacero konceptov tvorby tabuliek, či už podľa zdrojov alebo podľa vlastných nápadov. V praktickej časti potom s využitím teoretickej časti vytvoríme viacero tabuliek, ktoré by mohli reprezentovať selekčný charakter.

Kľúčové slová: selekčné úmrtnostné tabuľky, úmrtnostné tabuľky, konštrukcia úmrtnostných tabuliek, konštrukčný koncept selekčných úmrtnostných tabuliek.

Abstract

In this term project, we will take a closer look at selection mortality tables and their construction. The main difference from a traditional mortality table is that the selection mortality table predicts deaths and many other death-related phenomena based on characteristics other than age and sex. There are a large number of such characteristics, but to clarify the very name of the selection in the title of the project, it may be, for example, whether a person is a smoker, or whether a person lives in an urban or rural area, or according to what risky occupation a person is in. In the theoretical part we will describe their definition and compare them with the definition of the mortality tables from which they were derived. We will then write a couple of practical considerations on what their advantage is compared to the classical mortality tables and we will develop several concepts for the creation of the tables, either according to the sources or according to our own ideas. In the practical part we will then use the theoretical part to create several tables that could represent selection.

Keywords: selection mortality tables, mortality tables, construction of mortality tables, construction concept of selection mortality tables.

Obsah

Úvod

Prvá časť projektu je zameraná na teoretické objasnenie a porovnanie selekčným úmrtnostných tabuliek, spolu s ich možným významom a aj frekvenciou výskytu vo forme voľného webového vyhľadávania. Vytvoríme aj tri koncepty metodík na tvorbu selekčných tabuliek spolu s nutnými predpokladmi pre možnosť ich použitia. Následne ich v druhej časti, ktorá je praktického zamerania aplikujeme a vytvoríme nejaké konkrétne tabuľky.

1 Teória

1.1 História [1]

História úmrtnostných tabuliek siaha až do 17. storočia v Anglicku a Francúzsku. Hlavnou úlohou tabuliek bolo určovanie výšky životného poistenia. Zrejme sa od tej doby ukázalo, že určovanie poistného na základe úmrtnostných tabuliek je dobrou a cestou a v dnešnej dobe sa používajú naďalej na určovanie životného poistenia ako aj na výpočet dôchodku.

1.2 Definícia selekčných úmrtnostných tabuliek

Najprv si zadefinujeme niečo čo už nám je známe, tj. úmrtnostné tabuľky. Úmrtnostné tabuľky sú zhotovované už z historických dát zomretých jedincov za nejaké časové obdobie , spravidla za 1 alebo 1/2 roka. Teda zaznamenávame napr. koľko jedincov umrelo z celkovej populácie za dané časové obdobie. Na základe takto skonštruovaných údajov sa do tabuliek pridávajú vypočítané odhady rôznych pravdepodobností ako napr. pravdepodobnosť, že sa daný jedinec dožije nejakú časovú jednotku. Z praktického hľadiska pravdepodobnosť úmrtia jedného jedinca v priebehu jedného roka je veľmi nešpecifická, pretože intuitívne vieme povedať, že nejakí jedinci majú väčšiu pravdepodobnosť úmrtia ako druhí. Práve kvôli tomuto faktu je zaužívané, že úmrtnostné tabuľky vytvorené z historických relevantných dát sú triedené podľa veku a pohlavia. Príklad veľmi jednoduchej (na ilustratívne účely) vymyslenej tabuľky je možné vidieť v Tabuľka 1, kde dáta v tabuľke, v druhom a treťom stĺpci určujú, aký pomer ľudí s danými charakteristikami zomrelo v priebehu jedného roka.

Vek	Muži	Ženy
0-18	0.02	0.03
19-30	0.02	0.03
>31	0.1	0.1

Tabuľka 1

Selekčné úmrtnostné tabuľky by sme mohli neformálne nazvať ako rozšírené úmrtnostné tabuľky. Ak by sme chceli zostrojiť selekčnú úmrtnostnú tabuľku je potrebné aby vstupné dáta z ktorých sa selekčná konštruuje mali viacej charakteristík ako len vek a pohlavie a zároveň aby všetky charakteristiky selekčná úmrtnostná tabuľka mala v sebe vo finálnom zobrazení zahrnuté. Dané charakteristiky môžu byť demografického charakteru ako napr. štátna príslušnosť, úplná/neúplná rodina, rodinný stav, atď [2] alebo charakter, ktorý je získaný formou životného skúmania jedincov (napr. prostredníctvom ankety) napríklad či je daný človek fajčiar, na akej stupnici rizikovosti sa nachádza jeho povolanie, aké ma stravovacie návyky atď. Intuitívne je cítiť, že takto zostrojené tabuľky majú väčšiu výpovednú hodnotu, napr. pri určovaní životného poistenia pre konkrétneho jedinca. Tým pádom poistné spoločnosti vedia formou vstupnej ankety lepšie 'umiestniť' a aj určiť napr. odhad pravdepodobnosti úmrtia daného jedinca a následne mu určiť adekvátnu výšku splátky. Tiež je ale vhodné poznamenať, že zatiaľ čo úmrtnostné tabuľky majú v sebe zahrnuté nejaké odhady prevdepodobností (tie si uvedieme v jednej z podkapitól) pozostávajúcich z historických dát,(obsahujúce vek, v ktorom jedinec zomrel a aj jeho pohlavie), tak výsledná úmrtnostná tabuľka sa dá zobraziť v dvojdimenzionalnóm zobrazení ako je napr. vidieť na Tabuľka 1, tak pri selekčných by to mohol byť problém. Buď by mohol pribudnúť jeden stĺpec navyše s danou hodnotou charaktretistiky tj. fajčiar/nefajčiar a teda z pôvodných 3 riadkov by sa počet riadkov v Tabuľka 1 musel zdvojnásobiť alebo by sme vytvorili dve tabuľky zvlášť pre fajčiara a nefajčiara tj. Tabuľka 1 by sme prekonvertovali na trojrozmernú.

1.3 Ako je to v realite

V realite pri bežnom webovom vyhľadávaní o úmrtnostných tabuľkách (ďalej už UT) nájdeme kvantum informácií. Od histórie, ktorú sme aj krátko spomenuli v prvej časti, cez samotné historické dáta UT pre veľký počet rôznych štátov až po samotnú metodiku konštrukcie UT pomocou historických dát o počte mŕtvych jedincov. Teda na základe vlastnenia takýchto predpokladov ako sú samotné dáta o počte mŕtvych a me-

todika tvorby UT, je možné si samostatne skonštruovať UT. Je dôležité poznamenať, že metodika tvorby UT nie je všade jednotná a však býva väčšinou veľmi podobná a majú ju zväčša na starosti ústredné štatistické úrady. Napríklad na Slovensku má na starosti tvorbu UT Štatistický úrad Slovenskej republiky a pod ním prislúchajúcej úrady. Konkrétnu metodiku, ktorú používa náš štatistický úrad vieme nájsť na [4] (a časť z nej bude použitá v ďalšej podkapitole).

Ak však chceme nájsť podobné informácie aj k selekčným úmrtnostným tabuľkám (ďalej už SUT), tak dochádzame k problému. Jediný prienik s naším cieľom vypracovania zadania je len definícia samotnej SUT a teda čím sa odlišuje od UT. Ani po obšírnejšom vyhľadávaní v rôznych jazykoch sa nenašli žiadne informácie typu ako pri UT, teda už vypracované SUT ani metodika vypracovania SUT a ani žiadne dáta. Možné zapríčinenie si môžeme vysvetliť následovným príkladom. Aktuálne žijeme v dobe veľkého technologického progresu, kde sa určite javy (napr. firmy predpovedajú správanie zákazníka v ich aplikácií) predpovedajú na základe už skonštruovaného modelu, ktorý je závislý od rôznych signifikantných charakteristík napr. koľko ma zákazník rokov, aký je jeho finančný príjem, či je ovplyvnený najnovšími trend technológiami atď. Všetky tieto dáta na konštrukciu modelu si firma pozbierala sama za účelom čo najlepšieho pochopenia správania zákazníka a tým chcú maximalizovať svoj zisk. Je takisto zrejmé, že firma by bola nerada aby takýto model bol voľne dostupný všetkým lebo s tým nastáva aj riziko, že sa model môže dostať až ku konkurentovi a ten môže nejakým spôsobom ovplyvniť zisk pôvodnej firmy (napr. by model začal tiež používať). To čo sme tu teraz spomínali s firmou bolo za účelom bližšieho objasnenia SUT. Tak ako firmy používajú modely na predikcie, je viac menej jasné, že SUT sú používané napr. poisťovňami, ktoré chcú čo najlepšie nastaviť výšku poistného pre svojho zákazníka, tak aby boli v profite a zákazník bol ochotný za poistku platiť. Poisťovne by takisto neboli rady keby ich SUT, boli voľne dostupné, práve kvôli tomu, že by sa mohli dostať do rúk konkurenta tak ako pri príklade firmy. Teda SUT poisťovní by sa dali chápať ako ich interné tajomstvo a je zrejmé, že metodika tvorby SUT môže byť v niečom odlišná pre každú poisťovňu (rôzne charakteristiky alebo

iné metodiky spracovania tabuliek). Podľa [5] poisťovne tvoria SUT, podla homogénnych skupín ľudí, teda napr. osobitne pre fajčiarov a osobitne pre nefajčiarov alebo osobitne pre rôzne druhy zamestnania. A takisto podľa toho istého zdroja v USA rozlišuje od roku 1986 zvlášť UT pre fajčiarov a nefajčiarov. Ak by však chcela poisťovňa poistiť napr. človeka, ktorý je fajčiar a zároveň jeho pracovným zameraním je baníctvo, tak by bolo primerane použiť nejaký prienik týchto dvoch UT a nie iba jednú z nich alebo pri najhoršom by použila tú, kde je nižšia pravdepodobnosť úmrtia jedinca v identickom časovom horizonte, čo je nevýhodnejšie pre výšku splátky jedinca a výhodnejšie pre poisťovňu. V [7] sa nám podarilo nájsť približné pravidlo ako funguje pri konštrukcií tabuliek zvlásť pre rôzne charakteristiky a ten takisto popíšeme v následujúcej podkapitole. Na záver podkapitoly je ešte vhodne podotknúť, že poisťovne nie sú jediné, ktoré využívajú SUT ale taktiež ich môžu využívať demografi na zisťovanie očakávaného veku závislosti od ostatných demografických charakteristík alebo napr. klinickí štatistici, ktorí zaznamenávajú dĺžku života v závislosti od choroby a v akom štádiu ju pred smrťou mali.

1.4 Konštruovanie selekčných úmrtnostných tabuľiek

V tejto podkapitole vytvoríme 3 metodiky konštrukcie SUT. Prvá bude najjednoduchšia bez nejakej veľkej aplikácie nadstavby voči UT. Druhá metodika bude nadstavbou prvej podľa [7], ktorá je v istých bodových prípadoch podobná prvej. A nakoniec skonštruujeme poslednú metodiku, ktorá by sa mohla považovať aj za najkomplexnejšiu.

1.4.1 Predpoklady

Ešte pred samotnou tvorbou si zadefinujeme označenia podľa [1] a [7] spolu s platnými vzťahmi, ktoré sú odvodené v referenciách, ktoré budú spoločné pre všetky tri metodiky a spolu k ním aj zhrnieme nutné predpoklady vstupných dát pre možnosť použitia metodík. Najprv definujme premenné následovne.

- l_x: počet jedincov, ktorý prežili do veku x rokov. Typicky platí, že $l_0 = 10^5$, čo

predstavuje štandardizovaný fixný počet žijúcich jedincov.

- ${}_{t_1|t_2}q_x = (l_{x+t_1} - l_{x+t_1+t_2}){l_x}^{-1}$: pravdepodobnosť (ďalej už pp.), že jedinec vo veku x rokov sa dožije t_1 rokov a potom v ľubovoľný okamih do t_2 rokov zomrie. Zaužívané značenie je $q_x = {}_{0|1}q_x$.

- ${}_tp_x = \frac{l_{x+t}}{l_x}$: komplement ku ${}_{0|t}q_x$, teda pp., že sa dožije $x+t$ rokov. Zaužívané značenie je $p_x = {}_1p_x$.

- $d_x = l_x - l_{x+1} = l_x q_x$: počet jedincov, ktorí zomreli v dokončenom veku x rokov.

- $e^x = \sum_{i=1}^{\infty} {}_tp_x$: očakávaný počet rokov, ktorých sa dožije už x ročný jedinec.

Predpoklad na dáta pre využitie z hocijakých daných metodík je následovný: Dáta musia byť historického charakteru záznamoch o úmrtiach jedincov v nejakej konkrétnej populácií, z ktorých je možne podľa [1] vytvoriť UT (obsah pohlavia nie je nutnosťou). Taktiež musí platiť, že pre každé jedno dáto záznamu o smrti je k dispozícií aj charakteristika danej osoby, podľa ktorej chceme skonštruovať SUT. Doterajšia časť podmienok je postačujúcou podmienkou pre prvú a tretiu metodiku a však nutnou pre druhú. Príklad takých dát môžeme vidieť v Tabuľka 2. Je vidieť, že predpoklady sú splnené na

Vek [0-120]	Pohlavie [M/Z]	Fajčiar [1/0]	Rizik. povolania [0-3]	# havárií [>=0]
23	M	0	0	1
23	M	1	2	4
23	Z	0	1	2
58	Z	0	0	1
40	M	1	3	2

Tabuľka 2: Vymyslené záznamy o smrti jedincoch.

konštrukciu UT (máme záznamy o smrtiach ľudí spolu s ich vekom), ďalej ak nepočítame pohlavie ako charakteristiku, tak máme celkovo 3 rôzne charakteristiky. Z toho vyplýva, že na takéto dáta by bolo možné aplikovať metodiku 1 alebo 3.

Navyše predpoklad pre možnosť využitia metodiky dva (nie podľa referencie ale podľa vlastného uváženia), sú dodatočné informácie o veku osoby a to pre ten vek, kedy nastal posledný stav zmeny danej charakteristiky (post. podmienkou je aspoň jeden stĺpec). Teda ak by sme chceli použiť opätovne Tabuľka 2, tak takto upravená tabuľka pre použitie v druhej metodike by vyzeralo následovne Tabuľka 3.

Z takto vytvorených dát sa dá už vytvoriť SUT podľa metodiky dva. Sú splnené nutné podmienky tak ako v predch. tabuľke a zároveň dodatočná podmienka toho kedy nastal, posledný stav zmeny charakteristiky, napr. v 3. stĺpci vidíme, v akom veku dané osoby začali alebo prestali fajčiť (druhý číselný údaj) ale takisto vidíme, ktorý z nich nikdy v živote neboli fajčiari, v poslednom stĺpci zas vidíme v akom veku spôsobili poslednú nehodu za volantom. Ak by nejakému dátu chýbal tento vek započatia/skončenia

Vek [0-120]	Pohl. [M/Z]	Fajčiar [1/0],[v]	Rizik. povol. [0-3],[v]	# havárií [>=0],[v]
23	M	0,0	0,18	1,18
23	M	1,22	2,18	4,23
23	Z	0,0	1,18	2,22
58	Z	0,20	0,22	1,22
40	M	1,17	3,22	2,25

Tabuľka 3: Vymyslené záznamy o smrti jedincoch vol.2 .

charakteristiky, pre jednoduchosť by bolo vyfiltrované preč. Ak by zas, nejaký stĺpec nezaznamenával hodnoty veku započatia/skončenia charakteristiky, tak by bol stĺpec odstránený. Minimálne však musí byť aspoň jeden. Dôležitým dodatkom je, že charakteristika nemusí mať vek kedy charakteristika naposledy začala/skončila, v prípade fajčenia je to zavádzajúci údaj ak napr. 60 ročná osoba má napísané, že je nefajčiar od 59 roku, tj. nevieme históriu pred tým vekom. V takom prípade sa môže použiť celkové historické trvanie charakteristiky fajčenia za život.

1.4.2 Prvá metodika

Spočíva iba v tom, že pre každú jednu charakteristiku sa vytvorí zvlášť klasická UT. Zoberieme prvú charakteristiku a jej prislúchajúcu hodnotu. Ak je hodnota typu boolean tak je postup veľmi jednoduchý a identický s postupom konštrukcie UT v [1], s tým rozdielom, že sa postup nepoužije na všetky dáta ale najprv sa dáta prefiltrujú na tie kde je hodnota danej charakteristika rovná true/1 a na tie sa následne aplikuje daný postup, tj. vypočíta sa celkový súčet jedincov zomretých v daných rôznych rokoch, normovaním určime l_x pre všetky x, následne určime aj ostatné hodnoty premenných (prípadne nejaké dodatočné) z definície vyššie a tým pádom máme skonštruovanú UT pre jedincov, ktorých charakteristika bola naplnená hodnotou true. Analogický postup by bol aplikovaný aj pre tých kde hodnota charakteristiky bola false.

Ak by sme však mali UT pre nejakú z hora neohraničenú charakteristiku alebo typu int. Teda majme dve charakteristiky ch_1, ch_2, kde ch_1 nadobúda hodnoty od 0 po n_{small}

a ch_2 nadobúda hodnoty od 0 po n_{big}. Ak je hodnota n_{small} malá tak sa môže vykonať identicky postup ako pri predch. variante. V ďalšom prípade máme n_{big}, ktoré predstavuje vysoké číslo. V takom prípade máme dve alternatívy, ktoré môžeme vykonať. Ak nechceme vytvoriť celkovo n_{big} UT, tak môžeme hodnoty $\{0, 1, 2, \ldots, n_{big}\}$ roztriediť na nový systém množín $\{sep(b_0, b_1), sep(b_1, b_2), sep(b_2, b_3), \ldots, sep(b_{k-1}, b_k)\}$, kde:

$$sep(i, j) = \{\forall i, j : 0 \geq i \geq j | \exists k \in \{0, 1, 2, \ldots, n_{big}\} : i \geq k \geq j\} \subseteq \{0, 1, 2, \ldots, n_{big}\}.$$

Príklad $A = \{1, 2, 3, 4, 6, 7, ..., 200\}$ potom $sep(2, 7) = \{2, 3, 4, 6, 7\}$. Každé $sep(i, j)$ nadobúda hodnoty medzi i,j , ktoré sú zároveň súčasťou pôvodnej množiny, na ktorú sep aplikujeme. Potom pre každú získanú $sep(i, j)$ sa pôvodné dáta prefiltrujú podľa toho aby daná charakteristika nadobúdala hociktorú z hodnôt $sep(i, j)$ a na tie dáta sa aplikuje konštrukcia UT. Celkovo vznikne toľko UT aký je počet $\{sep(b_0, b_1), sep(b_1, b_2), sep(b_2, b_3), \ldots, sep(b_{k-1}, b_k)\}$. Ďalšia alternatíva v prípade, že n_{big} je veľké je jednoduchá. Za predpokladu, že je to dôležitá charakteristika a skupinovou konštrukciou UT (tj. prech. alternatíva) by sme mohli veľa stratiť, tak takisto pre každú hodnotu vytvoríme UT zvlášť. Teraz uvedieme nejaké príklady ako by sme sa v rámci prvej metodiky správali ako poisťovňa, ktorá ma poistiť klienta na základe nejakej charakteristiky.Keby máme poistiť klienta fajčiara, tak by sme sa pozerali na nami vytvorenú UT pre fajčiarov a na základe nej určovali výšku poistného. Keby sme poisťovali napríklad klienta, s rizikovým povolaním stupňa 2 z celkových 4 stupňov [0-3], tak si pozrieme prislúchajúcu UT pre rizikové povolanie s hodnotou rovnou dvom a podľa toho určujeme poistné. Ak by prišiel klient, ktorý celkovo zmenil 50krát pracovnú pozíciu a túto charakteristiku neberieme s takou veľkou váhou aby sme pre každú hodnotu konštruovali zvlášť jednu UT. Tak sme si miesto 50 hodnôt vytvorili 5 množín tj. od 0 po 10 zmien prac. pozície, od 10 po 20, ... ,od 40 po 50 a pre každú z nich mame spravenú UT. Klient zapadá do kategórie od 40 po 50 zmien prac. pozície, kde si zoberieme prislúchajúcu UT a podľa nej mu stanovíme výšku poistného. Posledným prípadom je, že by klient, ktorého tu spomíname spĺňal všetky 3 charakteristiky naraz (fajčiar, rizi. pov. 2, celkový počet zmien prác. pozícií je 50) a teraz nevieme, ktorú UT na neho použiť pri určovaní poistného. Ak by sme chceli byť ako poisťovňa čo najviac v zisku

tak vyberieme tú UT, kde je jeho očakávaný vek zvyšného života najmenší, za predpokladu, že to je ochotný kúpiť. Ďalšou voľbou je, že zoberieme UT, ktorej charakteristika je podľa nás dôležitejšia pri predpovedaní očakávaného veku zvyšného života v porovnaní s ostatnými charakteristikami. Ak by sme neboli ani s jednou alternatívou spokojní tak siahneme po tretej metodike.

1.4.3 Druhá metodika [7]

Tento typ konštrukcie je podobný tomu prvému a však je spoľahlivejší. Ako krátku motiváciu uvedieme skutočnosť podľa [7]. Poisťovne si všimli na ich historických dátach, že ak zoberieme človeka vo veku 50 rokov, ktorý v tom istom roku uzatvoril životné poistenie za normálnu cenu, tak má očakávanú dĺžku života väčšiu ako očakávaná dĺžka života náhodne vybratého 50 ročného človeka z populácie. To dáva zmysel z hľadiska vykonanej vstupnej lekárskej kontroly (health-check). Takisto si všimli, že ak prejde od tohto momentu napríklad 6 rokov tak o danom človeku už nevedia v akom zdravotnom stave je, aj keď 6 rokov do zadu to vedeli. Teda aktuálne je daný človek poistený vo veku 56 rokov a od posledného poistenia ubehlo 6 rokov. Akurát v ten čas sa poisti nový 56 ročný človek, ktorého podľa historických dát je očakávaná dĺžka života opäť väčšia ako 56 ročného človeka, ktorý uzatvoril zmluvu pred 6 rokmi. Poisťovne si to začali všímať a vytvorila sa definícia, ktorá je v [7] na pár riadkov a voľným slovom znamená, že sa oplatí pozorovať nie len aktuálny stav alebo charakteristiku jedinca ale aj čas kedy naposledy bol zmenený, preto sme si vytvorili taký predpoklad pre dáta ako sú napr. v Tabuľka 3.

V referencii taktiež nie je popísaný presný postup ako zostrojiť SUT ale miesto toho môžeme povedať, že tam je dobre rozšírenie pre prvú metodiku, pomocou ktorého vieme lepšie umiestniť klienta tam kde patrí a teda aj lepšie určiť očakávanú dĺžku života. O to s tým bude teraz viacej práce a to kvôli tomu, že sa UT nebudú tvoriť iba pre rôzne hodnoty charakteristiky ale aj pre rôzne hodnoty poslednej zmeny charakteristiky. Nič nám však nebraní opätovné použiť $sep(i, j)$ na "zgrupenie"nejakých dat do jedného. Netreba zabudnúť, že niekedy miesto posledného započatia/skončenia charakteristiky je

dobre použiť celkové trvanie charakteristiky za život, napr. pre fajčenie.

V takomto prípade keby sme si zobrali príklad klienta z predch. metodiky. Bol fajčiar, a fajčil celkovo 30 rokov v živote. Zoberieme si zostrojenú UT pre fajčiara, ktorý fajčí už 30 rokov a tam sa pozrieme podľa jeho aktuálneho veku aká očakávaná dĺžka života mu prislúcha. Ak by aktuálne vykonávala rizikovú prácu stupňa 2, tak by nás mohlo zaujímať ako dlho ju už vykonáva (posledná zmena charakteristiky), lebo by mohlo platiť, že čím dlhšie ju vykonáva tým je v nej zručnejší a tým je menšia pravdepodobnosť, že sa mu niečo zle prihodí. Alebo sa miesto toho môžeme rovno pozrieť koľko rokov celkovo vykonával práce s rizikovosťou rovnou dvom, keďže vieme, že už 50krát menil prácu, tak je to možno aj relevantnejší údaj. To koľkokrát menil prácu by sme pomocou druhej metodiky nevyhodnocovali. Ak by sme sa chceli rozhodnúť opäť na základe všetkých troch charakteristík naraz, tak je lepšie siahnuť po tretej metodike.

1.4.4 Tretia metodika

Ako sme videli v predošlých dvoch prípadoch, tak oba zaostávali ak by sme chceli klientovi nájsť UT podľa jeho viacerých charakteristík naraz. Môžeme si uvedomiť, že ak by sme pôvodné historické dáta o úmrtiach dokázali istým spôsobom filtrovať podľa zadaných charakteristík napr. klienta (fajčiar, ako dlho už fajči, akú rizikovú robotu vykonáva), tak by nám po filtrácií ostali dáta len o úmrtiach jedincov s tými istými charakteristikami. Následne by sa z takých dát dala vytvoriť opätovne klasická UT. Veľmi neformálne by sa to dalo interpretovať ako prierez cez viac-dimenzionálnú (pre každú rôznu hodnotu charakteristiky sa počítajú tie iste hodnoty $l_x, q_x, p_x...$) SUT hodnotami charakteristík.Tým pádom dostávame iba dáta mŕtvych s tými istými charakteristikami. Po tom ako sme získali takýto prierez - UT, môžeme klienta čo najlepšie umiestniť podľa veku a pozorovať vyššie spomenuté premenne, na základe ktorých sa môžeme rozhodovať o výške poistného.

Zostrojenie takého filtra je záležitosť pár hodín, ak sa rozprávame a vybudovaní mierne komplexnej aplikácie s grafickým rozhraním, ktorej primárnou úlohou by bolo efektívne spracovanie historických dát na vstupe, následne prefiltrovanie alebo tzv. spomínaný

prierez cez SUT podľa zadaných charakteristík a finálne vytvorenie UT (za predpokladu, že dáta s takými char. existujú), ktorá by bola exportovateľná v dátovej forme napr. .csv súboru. Z časového hľadiska sme si to uľahčili a tento filter vytvorili inakším spôsobom, ktorý popíšeme v praktickej časti.

2 Praktická časť

Teraz si ukážeme ako sme pomocou spomínaných troch metodík vytvorili viacero rôznych UT selekčného charakteru. UT selekčného charakteru budú obsahovať iba základné premenné definované v teoretickej časti.

2.1 Generovanie dát

Najprv sme si vygenerovali nejaké pseudonáhodné dáta. Generovali sme ich pre celkovo 8 rôznych premenných a to: vek:int, pohlavie:[1 | 0], fajčiar:[1 | 0], koľko rokov v živote bol už fajčiarom:int, rizikovosť práce, ktorú daný človek vykonával:[0 | 1 | 2 | 3], koľko rokov dozadu naposledy bola zmenená rizikovosť práce:int, koľkých dopravných bol daný človek účastníkom. Celá generácia dát až po export do .csv súboru prebiehala v súbore src.r. Každá hodnota bola generovaná z iného rozdelenia, tak aby aspoň trochu odpovedali skutočnosti ako napr. aby deti v mladom veku neboli fajčiari alebo aby niekto za život nefajčil viacej ako je jeho dosiahnutý vek atď. Ešte by sme mohli podotknúť, že premenné boli generované nezávisle od generácie hodnôt ostatných premenných. To znamená, že v konštruovaných UT nebudeme musieť vidieť skutočnosti, ktoré by mali byť zrejmé ako napr. že fajčiari vo veku x rokov sa dožívajú menej ako nefajčiari vo veku x rokov (tých prípadov tam je viacej). Extrakt dát je možné vidieť na Obrázok 1. Celkové dáta sa nachádzajú v data.csv, kde ich počet sme nastavili na 10^5.

age	gender	smokers	totalTimeSmokers	riskWorkRate	lastAgeChange	carAccidents	totalTimeFromLastAgeChange
49	0	0	0	1	49	0	0
47	1	0	0	1	47	0	0
50	0	0	0	1	45	0	5
12	1	0	0	0	12	0	0
60	0	0	0	1	52	0	8
52	1	0	0	1	52	0	0
88	0	0	0	2	88	0	0
72	1	0	0	0	66	3	6
112	1	0	0	2	110	0	2
67	1	0	0	3	58	2	9
37	1	0	0	2	27	1	10

Obr. 1: Grafické znázornenie samplu dát s ktorými budeme pracovať

2.2 Prvá metodika

Všetko ohľadom konštrukcie SUT na UT podla rôznych charakteristík sa nachádza v súbore metodika1.xlsx. Tak ako sme v teoretickej časti spomínali, tak tu sme pre každú charakteristiku typu boolean, tj. fajčiar a nefajčiar vytvorili zvlášť dve UT, ktoré su v hárkoch utNefajciari a utFajciari. UT pre fajčiarov sme zostrojili tak ako sme spomínali, najprv filtráciou dát čisto pre fajčiarov a potom následným postupom určovania premenných z definície. Obdobne sme to spravili aj pre rizikovosť povolaní, kde sme tento krát vytvorili 4 ďalšie hárky, pre každú hodnotu jeden hárok zvlásť s názvami utPovolanieR[0,1,2,3]. Posledný prípad bolo vytvorenie UT pre počet nehôd. Kedze najpocetnejsii pocet nehod malo v dátach hodnotu 12, tak sme sa rozhodli, ze nebudeme konštruovať 12 UT tabuliek ale rozdelíme si ich podľa spomínaných množín $sep(i, j)$ do 3 kategórií: nemajú žiaden záznam o nehodách, hárok: uTnehody(0-0), mierne nehodoví, hárok: utNehody(0-6) a nakoniec vysoko nehodoví, hárok utNehody(7-12).

2.3 Druhá metodika

Všetko ohľadom konštrukcie SUT na UT podla rôznych charakteristík sa nachádza v súbore metodika2.xlsx. Po otvorení si všimneme podobnú schému ako v prvom súbore a však môžeme vidieť aj spomínaný rozdiel časových hodnotách započatia/skončenia alebo celkového času za život pre rôzne charakteristiky. Pre fajčiarov nemáme iba jednu tabuľku, v dátach sme si všimli, ze najdlhší záznam o mŕtvom fajčiarovi je taky, ktorý fajčil celkovo

10 rokov svojho života. Mohli by sme urobiť 10 UT tabuliek pre každý jeden rok ale miesto toho opäť pomocou $sep(i, j)$ množín si dáta rozdelíme na kategórie slabší fajčiari, tí ktorí fajčia od 1 do 5 rokov a tuhí fajčiari, tí od 6 do 10 rokov. Dáta mame opäť prefiltrované a v hárkoch utFajciari(1-5) a utFajciari(6-10) sme pre nich vytvorili zvlásť UT tabuľky. Podobne sme opäť postupovali aj pri rizikovosti roboty, kde sme začali brat v úvahu počet rokov od poslednej zmeny rizikovosti roboty, a však tu sme to museli rozdeliť pre každú rizikovosť od 1 po 3 zvlásť. To znamená, ze napríklad tých s rizikovosťou 1 sme rozdelili na dve skupiny, krátkodobo v danej rizikovosti (nedávno vykonávali inakšiu rizikovú robotu) alebo dlhodobo v danej rizikovosti. Týmto činom nám vzniklo 6 ďalších hárkov. Konštruované UT pre počet nehôd sme vynechali.

2.4 Tretia metodika

Všetko ohľadom konštrukcie SUT na UT podla rôznych charakteristík sa nachádza v súbore metodika3.xlsx. Tu sme na rozdiel od predošlých dvoch využili filter, ktorému na vstupe povieme aké sú charakteristiky daného klienta. Tie sú napr. už predvyplnené konkrétne v danom súbore, v hárku: dataFilter v časti N4:T5 hodnotami:
(age:any,gender:any,smoker:1,riskWork:2,carAccidents:any,time smoking:5,time risk working:any), teda klient je fajčiar s 5 ročnou históriou, aktuálne pracuje v rizikovej robote s hodnotou 2 a viacej o ňom nevieme. Tieto hodnoty (alebo aj nové a iné) sa vyplnia do spomínanej tabuľky. Následne v zelenej bunke sa vygeneruje formula, ktorú skopírujeme a ako textovú hodnotu prilepíme do bunky A2. Musíme sa uistiť, že sme hodnotu prilepili ako textovú, ak ani v takom prípade nedostávame žiadne dáta na zobrazenie je možne, ze medzi dátami neboli také osoby ktoré mali rovnaké charakteristiky ako boli zadane na vstupe. V opačnom prípade vyskočia v stĺpcoch A:H nejaké dáta už mŕtvych osob, ktoré mali rovnaké charakteristiky pred smrťou ako klient. Ak chceme z daných dát vidieť skonštruovanú UT tabuľku tak klikneme na hárok sut, ktorá sa vždy dynamicky mení podľa prefiltrovaných dát v hárku dataFilter. Excelovým smerom sme sa vybrali lebo sme chceli ušetriť čas a zároveň sme ho primárne používali počas semestra.

Záver

V semestrálnom projekte sme sa snažili priblížiť SUT. Začali sme tým, z čoho mohli byť odvodené, následne sme ich porovnávali s UT a zostrojili nejaké definície na základe získaných útržkov z vyhľadávania. Ukázali sme ich praktické využitie, ktorým sme veľakrát interpretovali nejaké praktické príklady alebo aj samotnú praktickú časť tvorby SUT (hráme sa na poisťovňu). Vytvorili sme podla vlastného zváženia dve metodiky a jednu vysvetlili podľa poskytnutých referencií. A potom sme sa v praktickej časti snažili na základe týchto metodik vytvoriť SUT alebo ich aspoň interpretovať v UT zobrazení.

Zoznam použitej literatúry

[1] Szűcs, G.: *Prednášky z demografickej štatistiky*. Interný výučbový materiá, Dostupné na adrese: https://liveuniba.sharepoint.com/sites/FMFIUK_DemStat_2022/Uebn%20materily/DemStat%20-%20AAA%20-%20Prednasky%20-%202022.pdf?CT=1672411715961&OR=ItemsView

[2] Jurčová, D.: Slovník demografických pojmov. INFOSTAT – Inštitút informatiky a štatistiky, Výskumné demografické centrum, Edícia: Akty, Bratislava, 2005, ISBN 80-85659-40-9. [cit. 30.12.2022] Dostupné na adrese: http://www.infostat.sk/vdc/pdf/slovnik_2verdd.pdf.

[3] SEKEROVÁ, V., BILÍKOVÁ, M.: Poistná matematika. Bratislava: Ekonóm, 2002. s. 33 [cit. 30.12.2022] Dostupné na adrese: http://maag.euba.sk/documents/Umrtnostnetabulky.pdf.

[4] ŠÚ SR. [online]. Metodické vysvetlivky ku konštrukcii úmrtnostných tabuliek pre Slovenskú republiku. Webové sídlo Štatistického úradu Slovenskej republiky, Štatistiky, Obyvateľstvo a migrácia, Ukazovatele, Tabuľky života. [cit.31.12.2022] Dostupné na adrese: https://slovak.statistics.sk/wps/wcm/connect/2b1db2ea-7f14-41df-82aa-24a5302d9715/Metodika_UT.pdf?MOD=AJPERES.

[5] SEKEROVÁ, V., BILÍKOVÁ, M.: Poistná matematika. Bratislava: Ekonóm, 2002. s. 33

[6] HMD [online]. *The Human Mortality Database*. [cit. 02.11.2020] Dostupné na adrese: https://www.mortality.org/

[7] David C. M. Dickson, Mary R. Hardy, Howard R. Waters: Actuarial Mathematics for Life Contingent Risks. Cambridge University Press.

[8] Microsoft Corporation [online]. *Excel help & learning*. [cit. 01.01.2023] Dostupné na adrese: https://support.microsoft.com/en-us/excel.

[9] R Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. [cit. 01.01.2023] Dostupné na adrese: https://www.R-project.org/.

Prílohy

GitHub príloha

Všetký prílohy, na ktoré sme sa v projekte odkazovali nájdeme na:
`https://github.com/devAdam117/schDemPro`