Pravděpodobnost a statistika - Domácí úloha

Banka

Vypracoval: Petr Dvořák

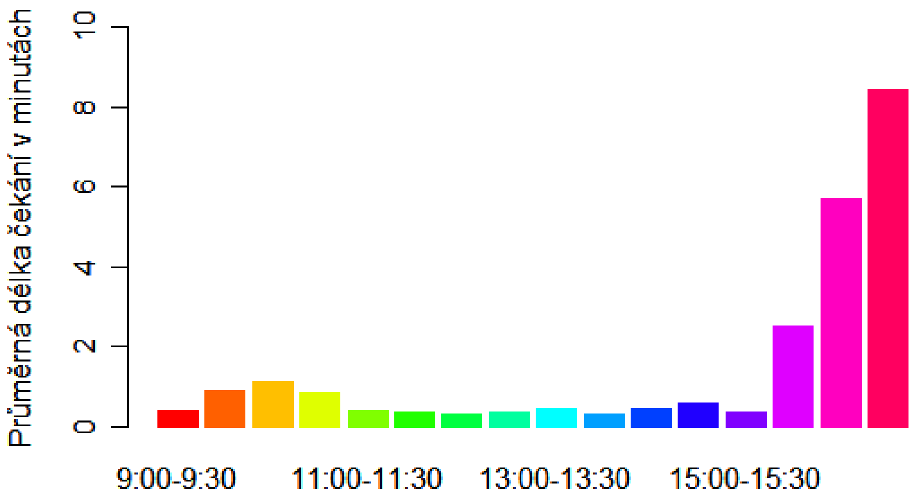
a) Pro každého klienta si musíme uložit čas jeho příchodu a dobu, po kterou klient čekal ve frontě (tu lze dopočítat odečtením času odchodu od času, kdy se klient dostal na řadu - to jsou výstupy které nám dá funkce banka.jedenden). Dále si pro každý odsimulovaný den uložíme celkový počet obsloužených klientů a abychom mohli řešit úlohu c), pro každý den si uložíme i čas ukončení obsluhy posledního klienta. Tato data poté použijeme na vypočtení průměrné doby čekání v jednotlivých denních dobách, na vypočtení střední hodnoty počtu obsloužených zákazníku za den a na určení střední hodnoty času ukončení obsluhy posledního klienta.

b) Chceme graficky znázornit průměrnou dobu čekání v závislosti na denní době. V souvislosti s tím je ale nejprve nutné určit, na jak dlouhé intervaly budeme dělit otvírací dobu. Tento krok je důležitý - na zobrazení průměrné doby čekání po jednotlivých minutách by bylo potřeba mnohem více měření (např. pro jednu konkrétní minutu by se nám při malém počtu vstupních dat mohlo stát, že budeme počítat průměr z dat jediného klienta). Naopak, brát celou otvírací dobu jako jediný interval a zobrazit průměrnou dobu čekaní všech klientů (de fakto bez ohledu na denní dobu) bychom sice mohli, nicméně tento statistický výsledek by nám mnoho nenapověděl například v situaci, kdy bychom chtěli optimalizovat využití přepážek v závislosti na denní době. Spočteme si proto nyní optimální délku intervalu pro zobrazení průměrné doby čekání.

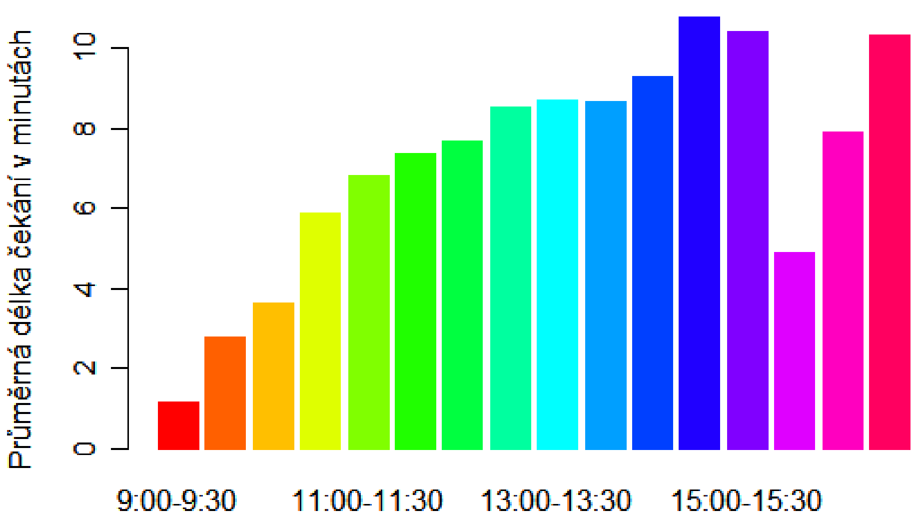
Na základě výstupu které jsme si uložili víme, že průměrný počet klientů obsloužených za den je zhruba 253. Provedli jsme simulace pro 250 dní. Můžeme proto předpokládat, že máme zhruba $n \doteq 250 \times 253 = 63250$ naměřených hodnot. Podle Sturgesova pravidla $m = 1 + 3.3 \times log_{10}n \doteq 16.84$ je optimální počet intervalů, otvírací doba je 480 minut, tedy ideální délka intervalu je $480/16.84 \doteq 28.5$ minut. Nejbližši rozumně velký interval (zohledníme-li logické členění otvírací doby a členění jednotlivých úseků otvírání a zavírání přepážek po půlhodinách) je tedy 30 minut.

Oba grafické výstupy potvrdili (triviální) doměnku, že při uzavření některých přepážek si klient počká v průměru o něco déle. Pro ilustraci uvádím následující příklady. Jsou-li otevřeny všechny přepážky, průměrná doba čekání (myšlen je zde celodenní průměr) je něco málo přes 2 minuty. Jsou-li některé přepážky zavřené, vzroste průměrná doba čekání na více než 7 minut.

Obrázek 1: Průměrná doba čekání v jednotlivých půlhodinových intervalech, jsou-li otevřeny všechny přepážky.



Obrázek 2: Příslušné hodnoty průměrné doby čekání za situace, kdy jsou některé přepážky uzavřeny (viz. zadání).



c) Studujme nyní čas ukončení obsluhy posledního klienta. Máme $n = 250$ naměřených hodnot, jsme schopni spočítat střední hodnotu $\bar{X} \doteq 502.257$ (jako aritmetický průměr naměřených hodnot) i (výběrový) rozptyl σ^2. Dle Centrální limitní věty můžeme pro dostatečně velká n použít aproximaci normálním rozdělením. Časy ukončení obsluhy jsou nezávislé váhodné veličiny, jsou stejně rozdělené. Zároveň pro ně můžeme použít CLV, neboť jsou splněny její předpoklady.

Tedy můžeme aproximaci normálním rozdělením použít (na histogramu si můžeme také povšimnout kladné šikmosti, pozn. autor). 90% interval spolehlivosti navrhneme tak, jako by se jednalo o výběr z normálního rozdělení, tj.

$$(\bar{X} - \frac{\sigma}{\sqrt{n}} z(\alpha/2), \bar{X} + \frac{\sigma}{\sqrt{n}} z(\alpha/2)).$$

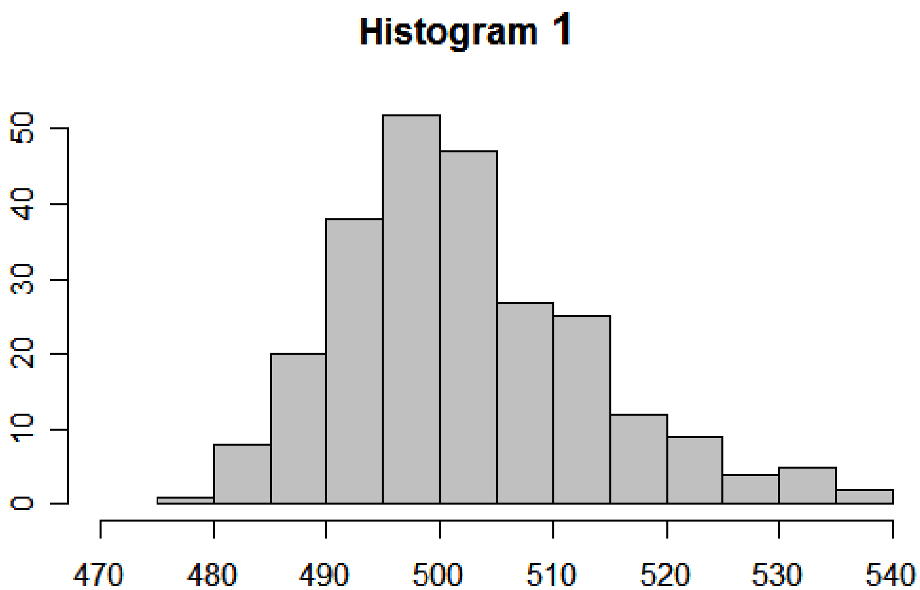
Dosazením příslušných hodnot (zavoláním funkce t.test v prostředí R) dostaneme konfidenční interval (501.09,503.43) - to jsou minuty po otvírací době, což není moc lidsky čitelné. Prakticky to znamená, že reálná střední hodnota času obsluhy posledního klienta bude na 90% ležet v časovém intervalu mezi 17:21'05" a 17:23'30". Vizuálním porovnáním námi naměřených hodnot však zjistíme, že velká část hodnot leží mimo tento interval (tedy že střední hodnota sice může v námi určeném intervalu ležet, ale spousta klientů je obsloužená dříve a občas se stane, že je někdo obsloužen o hodně později)... Velký statistický význam bych tedy tomuto výsledku nepřikládal.

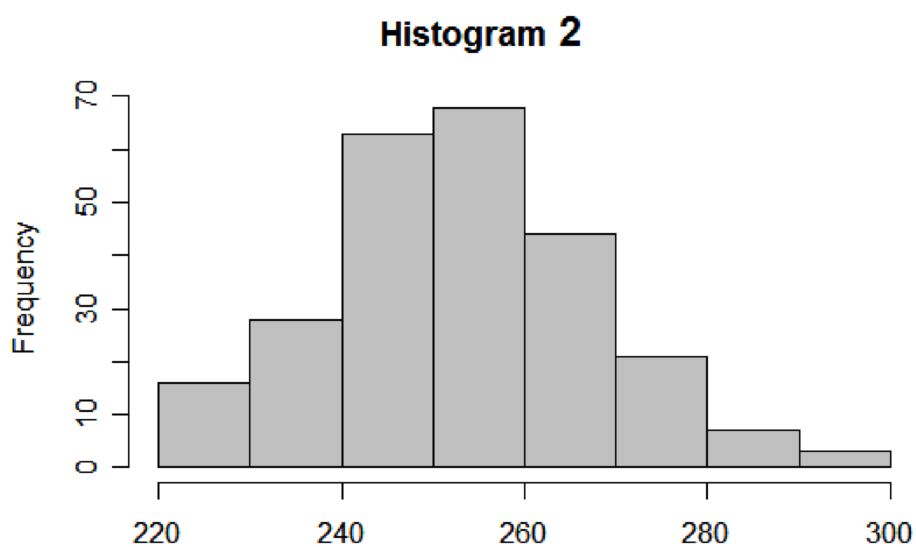
d) Z histogramu 2 je vidět, ze průměrný počet klientů obsloužených za den se řídí přibližně normálním rozdělením. Můžeme tedy použít jednovýběrový t-test. T-kvantily jsou pro $n = 250$ přibližně schodné s normálními. Označme

$$T = (\bar{X} - \mu) \cdot \frac{\sqrt{n}}{\sigma},$$

kde μ je testovaná hodnota (tedy 253), $\bar{X} \doteq 253.396$ je aritmetický průměr námi zjištěných hodnot, n je počet dní, který je v našem příkladě 250, a $\sigma^2 \doteq 14.72^2$ je (výběrový) rozptyl. Nulovou hypotézu *"Průměrný počet klientů obsloužených za jeden den je 253."* zamítneme na hladině $\alpha = 5\%$ tehdy, když $|T| \geq z(\alpha)$. Můžeme snadno dopočítat, že $z(0.05) \doteq 1.645$, $|T| \doteq 0.43$.

Dosazením hodnot které nám vygeneroval program tedy zjistíme, ze hypotézu na 5% hladině významnosti nezamítneme. Pro upřesnění můžeme dotat, že p-hodnota (kterou nám vrátila funkce t.test programu R) je v našem případě rovna $p \doteq 0.67$.





Vysázeno systémem LaTeX