Je mavrica res enakomerna?

Matej Vrankar

28. september 2017

1 Uvod

Odločil sem se narediti statistično analizo bombonov Skittles. Idejo za to sem dobil, ko sem na internetu videl slike, ko so ljudje razvrščali barvne bombone v stolpce in tako primerjali število barv v posamezni vrečki. Pri nekaterih primerih so se višine stolpcev razlikovale za nezanemarljiv faktor, tako sem dobil idejo, da s preverjanem statistične domnev ugotovim, če so barve bombonov enakomerno porazdeljene.

Prevejanjana, ki jih bom opravljal so preverjanja za statistični nadzor kvalitete (Statistical quality control), s katerim se je Walter A. Shewhart začel ukvarjati že v zgodnjih 20 letih prejšnjega stoletja. Že takrat so poznali moč vzorčenja in uporabo statističnih metod za možnost napovedovanja izdelkov glede na vrednosti vzorca. O tem je pisal tudi G. S. Radford v svoji knjigi The Control of Quality in Manufacturing [1]. Med drugo svetovno vojno se je potreba za kvaliteto še bolj dvignila, saj bi kakršnakoli napaka lahko bila usodna. Po koncu vojne so predstavniki Ameriške zveze za nadzor kvalitete odšli na Japonsko z namenom, da bi pomagali uvesti bolšji nadzor kvalitete. Kmalu za tem so Japonci njihov model nadgradili tako, da so poleg meritev izdelkov vpeljali še meritve delavcev in še bolj izboljšali kvaliteto končnega izdelka.

Taka preverjanja se v proizvodnjah dogajajo zelo pogosto, saj podjetje hoče čim bolj enakomerno porazdeljene bombone in maksimirati profit, kar pomeni, da morajo vrečke biti težje od 125g, da se izognejo kaznim raznih inšpekcij in tožbam, seveda pa nočejo imeti prevelikih izgub z preveč napolnjenimi vrečkami. SQC temelji na diagramu poteka, ki na grobo izgleda tako, da vzamemo naključni vzorec, z uporabo statističnih metod preverimo njegove lastnosti, ter napravo ponovno kalibriramo, ter začnemo na novo z zajemom novega vzorca.

V naši študiji na prvi pogled izgleda, da gre bolj za zabavo vendar se podobna preverjanja uporabljajo npr. v proizvodnji zdravil, kjer morajo biti deleži kemijskih sestavin zelo natančno odmerjeni, saj bi sicer lahko prišlo do nezaželeni učnikov.

Izračunal sem tudi povprečno število bombonov, povprečno neto maso vrečke, delež bombonov posamezne barve, izrisal porazdelitev mase vrečk. Preveril sem še 5 domnev in izračunal 6 intervalov zaupanja in korelacijo med maso vrečke in številom bombonov.

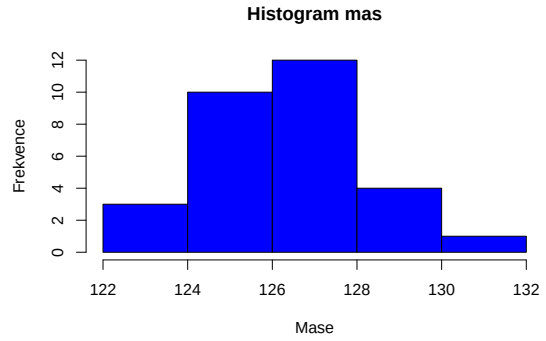
Na začetku predstavlmo, kako smo prišeli do vzorca. Sledi analiza masa vrečk, primerjava mas vrečk med vzorci ter še preverjanje domneve, da je povprečna masa vrečke enaka od 125g. Nato bom prikažemo še deleže vseh barv bombonov, ter naredil 3 različne teste za preverjanje enakomerne porazdelitve barv. Na koncu izračunamo še korelacijo med maso vrečke in številom bombonov.

2 Zbiranje podatkov

Kupil sem po 10 vrečk iz 3 različnih proizvodnih obdobij, tako sem dobil testni vzorec velikosti 30. Ker so bili bomboni kupljeni v isti trgovini in imeli enako oznako za proizvodno linijo ter časovni žig, lahko verjetno privzamemo, da je vzorec enak kot, če bi v tovarni naključno vzeli 10 vrečk iz proizvodne. Za vsako vrečko sem izmeril neto maso bombonov, preštel posamezne barve, ter skupno število bombonov, to sem naredil za vsako od proizvodnih obdobij posebej. Povprečno število bombonov v vrečki je bilo 117,2, povprečna masa vrečke pa 126,7g.

3 Porazdelitev mas vrečk

Za posamezne mase sem preštel število vrečk, ki so imela to maso, tako sem prišel do histograma, glej sliko 1.



Slika 1: Histogram, ki prikazuje frekvence mas vrečk.

Ker je število vzorca veliko, torej je večje oziroma enako 30, bi morala biti masa bombonov porazdeljena približno zvonasto. Če ne veste zakaj ali pa ste radovedni in bi radi izvedeli več o statistični obdelavi podatkov, si lahko preberete knjigo For All Practical Purposes [3] od ameriškega statistka Moora. Za vajo lahko preverimo, če v tem primeru res velja izkustveno pravilo za verjetnosti znotraj 2σ, 4σ, 6σ. Te verjetnosti bi morale biti enake približno 68%, 95% ter 99%. Dejanske verjetnosti sem izračunal tako, da sem izračunal vzorčno povprečje $\bar{x}$ in vzorčni odklon s (ki ju uporabim kot oceni za povprečje populacije ter njen odklon):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 126,7 \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = 2,003$$

Na intervalu $[i * s - \mu, i * s + \mu]$, kjer i je število odmikov, preštel vse vrednosti, ki so znotraj intervala. Na koncu pa sem delil še z 30, da sem dobil verjetnosti. Izkazalo se je, da je verjetnost, da je masa znotraj $2s$ enaka 76,76%, za $4s$ 93,33%, pri $6s$ pa so bile v intervalu vse vrednosti. Vidimo, da so verjetnosti zelo podobne, ter lahko sklepamo, da so mase vrečk porazdeljeno zvonasto.

Interval zaupanja za povprečno maso vrečke in preverjanje domneve, da je povprečna masa vrečke različna od $126g$***, pri stopnji zaupanja*** 95%***, in*** $n = 30$***:***

$$\bar{x} \pm t_{1-\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}} \qquad H_0 : \mu = 126 \qquad \text{TS} = \frac{\bar{x} - \mu}{s} \cdot \sqrt{n}, z \text{ porazdelitev}$$

$$= 126,7 \pm 1,59 \qquad H_a : \mu \neq 126 \qquad \text{TS} = 1,91$$

$$= [125,11, 128,29] \qquad P\text{-vrednost} = 0,065$$

S stopnjo tveganja 5% lahko rečemo, da je povprečna masa bombonov na intervalu med 124,0 in 127,2. Za ničelno hipotezo H_0 rečemo, da je povprečna masa enaka 126g, za alternativno hipotezo $H_{/alpha}$ pa da povprečna masa ni enaka 126g. Za izbiro hipoteze si pomagamo z testno statistiko TS. Če je TS v kritičnem območju K_α sprejmemo hipotezo H_α sicer pa sprejememo hipotezo H_0. P-vrednost je najmanjša stopnja značilnosti, pri kateri še zavrnemo ničelno domnevo. Iz podatkov zgoraj lahko razberemo, da H_0 pri $\alpha = 0,05$ sprejmemo, za $\alpha = 0,10$ pa ničelno domnevo ne moremo zavrniti.

Interval zaupanja za razliko povprečij mas iz prvega in drugega proizvodnega obdobja, pri stopnji tveganja 5%***:***

$$df = \left\lfloor \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1 - 1} \cdot (\frac{s_1^2}{n_1})^2 + \frac{1}{n_2 - 1} \cdot (\frac{s_2^2}{n_2})^2} \right\rfloor, \qquad \bar{x}_1 - \bar{x}_2 \pm t_{1-\alpha/2}(df) \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df = 14 \qquad 0,5 \pm 1,63$$

S stopnjo zaupanja 95% dobimo rezultat, da je interval za razliko prvega in drugega obdobja med $-1,13$ ter 2,13. Ker je območje intervala obsega negativne in prav tako pozitivne vrednosti, ne moremo trditi, da ima povprečna vrečka iz enega obdobja večjo maso, kot vrečka iz drugega obdobja. Če bi bil interval obsegal samo negativne vrednosti, bi pa lahko sklepali, da je povprečna masa prve vrečke manjša od druge. To se seveda da izračunati na bolj eleganten način z preverjanjem domneve o razliki povprečij.

$$H_0 : \mu_1 - \mu_2 = 0 \qquad \text{TS} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, t \text{ porazdelitev}$$

$$H_a : \mu_1 > \mu_2 \neq 0 \qquad \text{TS} = 0.65$$

$$P\text{-vrednost} = 0.264$$

Ker je P-vrednosti tako velika, na osnovi danih vzorcev ne moremo zaključiti, da je povprečna masa vrečke iz prvega proizvodnega obdobja večja od drugega.

Preverjanje domneve, da je povprečna masa večja od mase, ki je zapisana na embalaži $(125g)$***, pri stopnji zaupanja*** 95%***:***

$$H_0 : \mu = 125 \qquad \text{TS} = \frac{\bar{x} - \mu}{s} \cdot \sqrt{n}, z \text{ porazdelitev}$$

$$H_a : \mu > 125 \qquad \text{TS} = 4,64$$

$$K_\alpha = [1,645, \infty] \qquad P\text{-vrednost} = 0.00001$$

Testna statistika je znotraj kritičnega območja, zato H_0 zavrnemo, kar pomeni, da je pri stopnji zaupanja 95% povprečna masa večja od 125g.

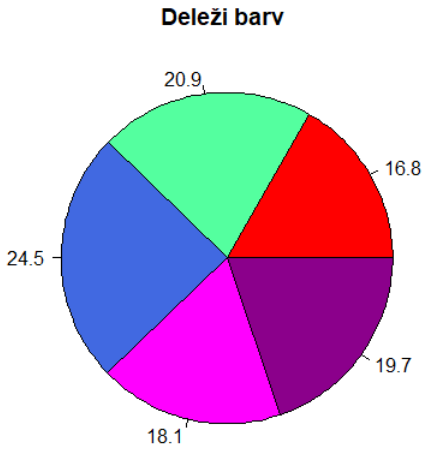
4 Deleži barv

V vrečki je 5 različnih barv, in sicer rdeča, vijolična, roza, modra in zelena. Barve naj bi bile po besedah proizvajalca porazdeljene enakomerno. Za posamezno bravo sem seštel vse bombone ter izračunal njihove deleže, in prišel do naslednjih podatkov, tabela 1.

Tabela 1: Število bombonov posamezne barve.

Barva	Število bombonov	Delež
Rdeča	590	16,8
Vijolična	694	19,7
Roza	636	18,1
Modra	862	24,5
Zelena	734	20,9
Skupaj	3516	100

Razvidno je, da je največ modrih bombonov, najmanj pa rdečih, ostale 3 barve so dokaj podobno zastopane. Na grafu 2, so prikazani tudi deleži posamezne barve.



Slika 2: Krožni diagram prikazuje deleže barv v vrečkah bombonov Skittles.

Intevali zaupanja za deleže barv v vrečki pri stopnji zaupanja $95\,\%$***:***

$$\hat{p} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$$

Primer izračuna za rdečo barvo

$$0,168 \pm z_{0,975} \cdot \sqrt{\frac{0,1689 \cdot 0,832}{3516}} =$$
$$= 0,168 \pm 0,012.$$

Interval zaupanja za delež rdeče barve pri stopnji tveganja 5% je od 0,146 do 1,80. Izračunal sem še za ostale barve in prišel do naslednjih rezultatov: zelena [0,196, 0,222] modra [0,231, 0,259], roza [0,169, 0,193] in vijolična [0,184, 0,210].

Preverjanje domneve, da sta deleža zelene in vijolične barve enaka, pri stopnji zaupanja $90\,$***%:***

$$H_0 : \pi_1 = \pi_2 \qquad \hat{p} = \frac{y_1 + y_2}{n_1 + n_2},$$

$$H_a : \pi_1 \neq \pi_2 \qquad \text{TS} = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p} \cdot \hat{q} \cdot (\frac{1}{n_1} + \frac{1}{n_2})}}, z \text{ porazdelitev}$$

$$K_\alpha = [-\infty, -1,645] \cup [1,645, \infty] \qquad \text{TS} = 1,26$$

$$P\text{-vrednost} = 0,104$$

Testna statistika ni znotraj kritičnega območja, zato ničelno domnevo H_0 ne zavrnemo.

Preverjanje domneve, da je razlika deležev modrih in zelenih bombonov manjša od $5\,$***%, pri stopnji tveganja*** $1\,$***%:***

$$H_0 : \pi_1 - \pi_2 = 0,05 \qquad \text{TS} = \frac{\hat{p_1} - \hat{p_2} - p_0}{\sqrt{\hat{p} \cdot \hat{q} \cdot (\frac{1}{n_1} + \frac{1}{n_2})}}, z \text{ porazdelitev}$$

$$H_a : \pi_1 - \pi_2 < 0,05 \qquad \text{TS} = -1,3$$

$$K_\alpha = [-\infty, -2,33] \qquad P\text{-vrednost} = 0,096$$

Testna statistika ni znotraj kritičnega območja, torej ničelne domeneve H_0 ne moremo zavrniti.

Preverjanje domneve, da je porazdelitev bombonov enakomerna, pri stopnji tveganja 5%:
Za izračun porazdelitve je najprej potrebno pridobiti frekvence f_o posameznih barv, te lahko preberemo iz tabele 1. Nato je na vrsti izračun teoretičnih frekvenc f_t, ker preverjamo, če je porazdelitev enakomerna so frekvence vseh kategorij enake. Izračunal sem jo tako, da sem delil število vseh bombonov z število kategorij (barv), torej $f_t = \frac{3516}{5} = 703,2$.

$$H_0 : \text{porazdelitev je enakomerna} \qquad \text{TS} = \sum_{i=1}^{k} \frac{(f_o - f_t)^2}{f_t}, \chi^2 \text{ porazdelitev}$$

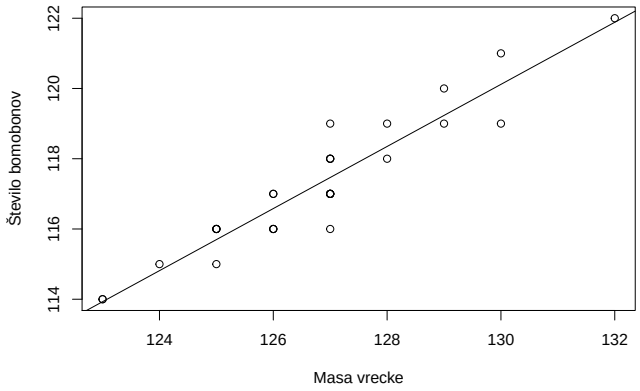
$$H_a : \text{porazdelitev ni enakomerna} \qquad \text{TS} = 61,97$$

$$K_\alpha = [9,49, \infty] \qquad P\text{-vrednost} = 0,00001$$

Testna statistika je v kritičnem območju, zato ničelno domnevo H_0 zavrnemo, kar pomeni, da barve bombonov niso porazdeljene enakomerno.

5 Povezava med maso vrečke in številom bombonov

Že brez, da bi sploh videli podatke, lahko sklepamo, da bo Pearsonov korelacijski koeficient zelo blizu 1, saj je jasno, da bo vrečka, ki je težja imela tudi večje število bombonov. Za prikaz linearnosti sem v programskem jeziku R s funkcijo plot() izrisal razsevni graf, glej sliko 3, ter s funkcijo abline() dodal še regresijsko premico.



Slika 3: Razsevni graf in regrecijska premica.

Da se prepričamo, da je Pearsonov korelacijski koeficient res blizu 1, ga še izračunajmo. Za to potrebujemo povprečno vrednost obeh parametrov. Naj bo x masa vrečke in y število bombonov, potem je $\bar{x} = 117,2$ in $\bar{y} = 126,7$.

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

$$r_{X,Y} = 0,94$$

Prišli sem do rezultata, da je Pearsonov korelacijski koeficient enak 0,94, kar je res blizu 1.

Na koncu pa še z metodo preverjanja domneve o povezanosti dveh spremenljivk preverimo, če sta masa vrečke in število bombonov linearno povezani:

$H_0 : \rho = 0$ (nista linearno povezani)	TS $= \dfrac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1 - r_{xy}^2}}$, z porazdelitev
$H_a : \rho \neq 0$ (sta linearno povezani)	TS $= 14,62$
$K_\alpha = [-\infty, -2,76] \cup [2,76, \infty]$	P-vrednost $\doteq 0$

Testna statistika je znotraj kritičnega območja, zato zavremo ničelno domnevo H_0, kar pomeni, da sta masa vrečke in število bombonov pri stopnji zaupanja 99% linearno povezani.

6 Zaključek

Pri 30ih vrečkah bombonov Skittles je bila povprečna masa vrečke $\bar{x}$ enaka 127,3g, vzorčni odklon s pa 2,0003. Verjetnosti, da ima naključno izbrana vrečka maso med $[i*s-\mu, i*s+\mu]$ za i-je 1-3, oziroma, da je masa znotraj $2s$, $4s$ in $6s$ so bile 0,768, 0,933 ter 1. S preverjanjem domnev sem prišel do zaključka, da je povprečna masa nekoliko večja od zapisane na embalaži. Za preverjanje enakomerne porazdelitve barv bombonov sem opravil 3 različne preverjanja, saj je tako napoved bolj točna. Pri prvih dveh testih nimamo dovolj podatkov, da bi lahko rekli, da barve niso enakomerno porazdeljene, ko pa naredimo še preverjanje z χ^2, lahko rečemo, da porazdelitev barv ni enakomerna. Ugotovili smo tudi, da sta masa vrečke in število bombonov linearno povezani.

Literatura

[1] George S. Radford, The Control of Quality in Manufacturing, New York, Ronald Press Co.,1992

[2] W. Mendenhall in T. Sincich, Statistics for Engineering and the Sciences, 4. izdaja, Prentice Hall, 1995

[3] D. S. Moore, For All Practical Purposes, 9. izdaja, Purdue University, COMAP, 2011