PRIMER sa obrađenim podacima u Excel-u (2007)

U tabeli je data raspodela mase slučajno odabranih studenata nekog fakulteta:

Masa (kg)	[60, 63)	[63, 66)	[66, 69)	[69, 72)	[72, 75)
Broj studenata	5	18	42	27	8

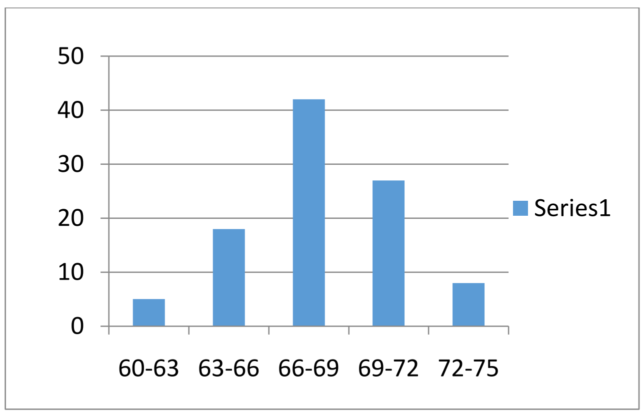
a) **Predstaviti podatke grafički u vidu histograma i poligona frekvencija i kumulative.**

Histogram frekvencija je stepenasta figura koja se sastoji od pravougaonika čija je osnova dužina intervala (u koje su mase grupisane), a "visina" učestanost (broj studenata). Ako spojimo sredine gornjih strana pravougaonika dobićemo krivu koja predstavlja statistički analog za funkciju gustine obeležja (mase studenata) – poligon učestanosti.

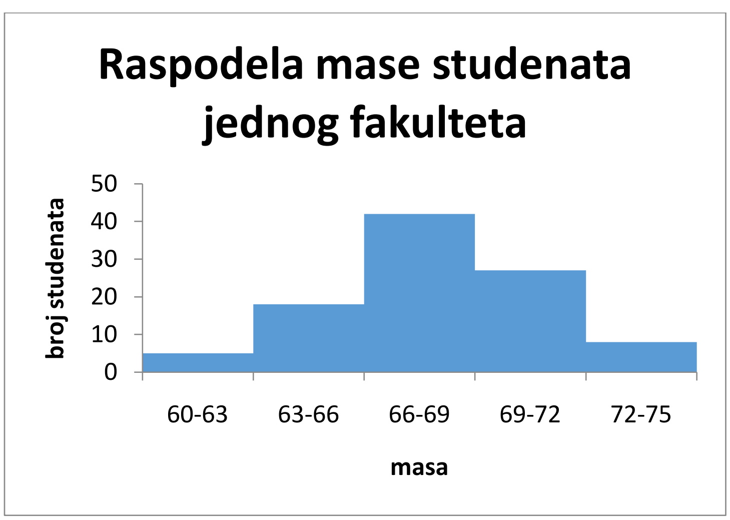
Grafički ćemo prikazati naše podatke u Excel-u koristeći funkciju *Insert Charts*.

Zapišimo date podatke u Excel-u, označimo ih i odaberimo opciju *Insert – Charts – Column* (*2-D Column*)

Masa	Učestanost
60-63	5
63-66	18
66-69	42
69-72	27
72-75	8



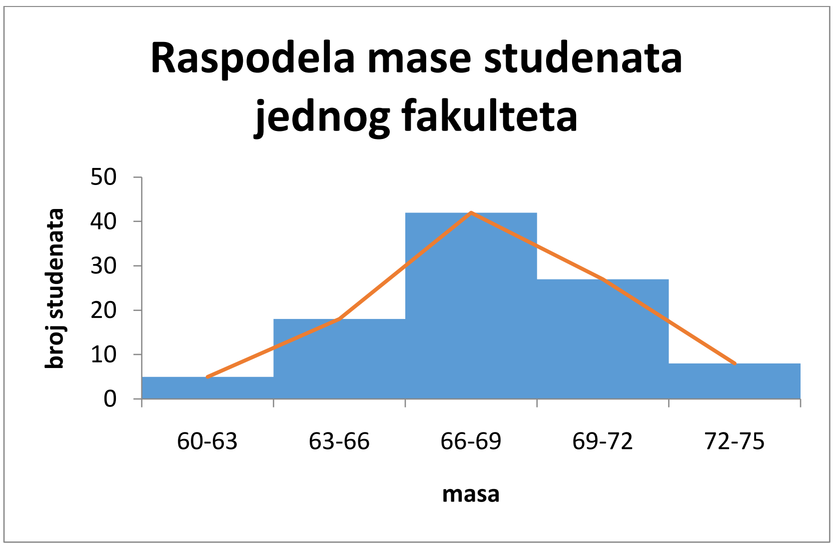
Da bi smo ovaj *chart* pretvorili u histogram moramo spojiti "pravougaonike podataka". Desnim klikom na njih dolazimo do opcije *Format Data Series*, gde treba *Gap Width* smanjiti na nulu. Takođe, možemo srediti histogram, imenovati odgovarajuće ose kroz opciju *Chart Layout* koja nam se nudi selektovanjem samog grafika.



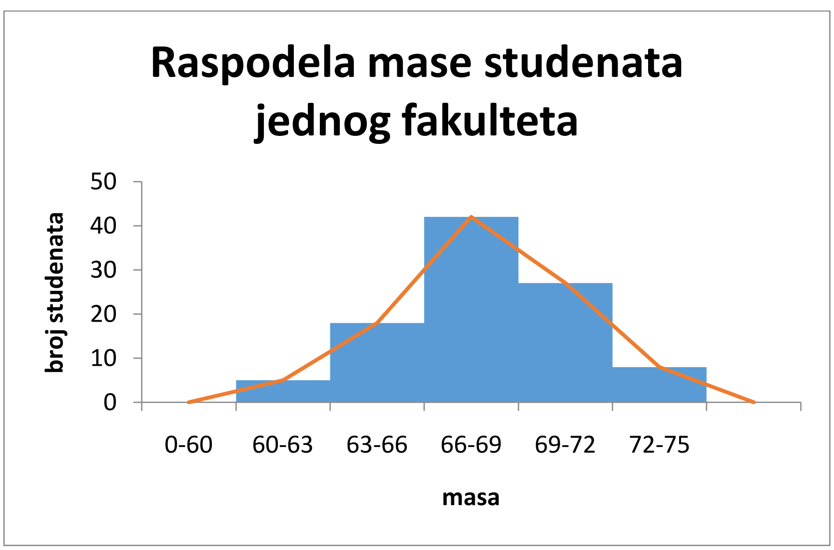
Da bismo formirali poligon učestanosti intervalni varijacioni niz zamenjujemo sa diskretnim: sredinu i-tog intevrala uzimamo za predstavnika tog interval kome dodeljujemo intervalnu učestanost. Prethodno formiranoj tabeli u excel-u dodajemo još jednu kolonu – sredine intervala, koristeći funkciju *AVERAGE*.

Masa	Učestanost	Sredine intervala
60-63	5	61.5(=AVERAGE(60,63))
63-66	18	64.5
66-69	42	67.5
69-72	27	70.5
72-75	8	73.5

Na istom grafiku možemo predstaviti i poligon učestanosti, dodavanjem nove serije podataka kroz opciju *Select Data – Add*, čime dodajemo novu seriju podataka ("pravougaonika") na već postojeći histogram. Za *Series name* selektovati sredine intervala, a *Series values* kolonu učestanosti. Da bismo napravili poligon potrebno je još samo promeniti *Chart Type* tih novih, obično drugom bojom obojenih "pravogaonika" selektovanjem istih i klikom na opciju *Change Chart Type – Line*. Na taj način dolazimo do sledećeg prikaza raspodele mase studenata:



Ili, uz male izmene serije podataka, histogram i poligon učestanosti mase studenata jednog fakulteta se može prikazati:



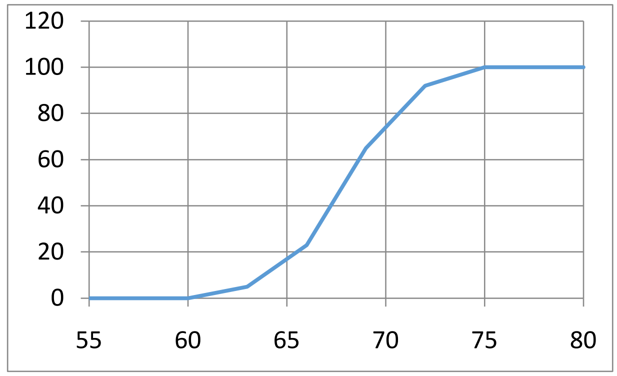
Primećujemo da dobijena kriva liči na Gausovu (normalnu) gustinu raspodele, pa bi se mogla postaviti hipoteza o normalnoj raspodeli našeg obeležja – mase studenata.

Kumulativa se konstruiše, u slučaju intervalno datih podataka, tako što se na apscisu nanose vrednosti koje čine intervale,a na ordinatu kumulirane (zbirne) vrednosti učestanosti koje vežemo za krajeve intervala.

Na osnovu početne tabele (kolone A i B) formiramo tabelu vrednosti potrebnih za kreiranje grafika kumultive (kolone C i D), zatim selektujemo kolone C i D i biramo opciju *Insert – Charts – Scatter* – poslednji u nizu ponuđenih.

	A	B	C	D
1	0-60	0	55	0
2	60-63	5	60	0
3	63-66	18	63	5
4	66-69	42	66	23
5	69-72	27	69	65
6	72-75	8	72	92
7			75	100
8			80	100

=SUM(B$2:B2)
Ostale vrednosti možemo dobiti kopiranjem prethodne ćelije



b) **Naći: srednju vrednost, medijanu i mod mase studenata; uzoračku disperziju i standardnu devijaciju (odstupanje).**

c) **Naći 95% interval poverenja za matematičko očekivanje mase studenata (prosečnu masu studenata).**

Pogledati fajl [Tačkaste i intervalne ocene parametara raspodele] .

Srednju vrednost mase studenata nalazimo prema formuli $\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n}$, gde su x_i sredine inetrvala, n_i odgovarajuće učestanosti, k broj intervala, a n obim uzorka, $n = \sum_{i=1}^{k} n_i$.

	A	B	C	
1	60-63	5	61.5	(=AVERAGE(60,63))
2	63-66	18	64.5	
3	66-69	42	67.5	
4	69-72	27	70.5	
5	72-75	8	73.5	
6	n (= SUM(b1:b5))	100		
7	Srednja vrednost	67.95	(=SUMPRODUCT(C1:C5,B1:B5)/B6)	

Medijanu lako možemo uočiti na grafiku kumulative, ili u tabeli sa kumulativnim učestanostima. Vidimo da ona pripada 3. intervalu ("66-69"), odnosno možemo reći da je medijana ovog uzorka masa od 67,5 kg.

Po definiciji, mod,ili tačnije modalna klasa, je interval sa najvećom učestanošću, što bi bio 3.interval.

Uočavamo da srednja vrednost, medijana i mod mase studenata pripadaju istom intervalu, što ukazuje na simetričnost raspodele obeležja.

Uzoračku disperziju mase studenata nalazimo prema formuli $s^2 = \frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i}{n}$ ili $s^2 = \overline{x^2} - \bar{x}^2$, gde je $\overline{x^2} = \frac{\sum_{i=1}^{k} x_i^2 n_i}{n}$, dok je standardna devijacija (odstupanje) $s = \sqrt{s^2}$.

	A	B	C	D	
1	60-63	5	61.5	41.6025	=(C2-B$7)^2
2	63-66	18	64.5	11.9025	Povlačenjem/kopiranjem
3	66-69	42	67.5	0.2025	=(C3-B$7)^2
4	69-72	27	70.5	6.5025	B$7 – srednja vrednost
5	72-75	8	73.5	30.8025	=(C5-B$7)^2

	A	B	C	D	
6	n	100			
7	$\bar{x}$	67.95			
8	s^2	8.5275			=SUMPRODUCT(D1:D5,B1:B5)/B6
9	s	2.90188			=SQRT(B8)

95% interval poverenja za prosečnu masu studenata nalazimo iz uslova $P\left\{\left|\frac{(\bar{X}-a)\sqrt{n-1}}{S}\right| < z\right\} = 0{,}95$ i da statistika $\frac{(\bar{X}-a)\sqrt{n-1}}{S}$ ima približno $\mathrm{N}(0{,}1)$ raspodelu (tablice normalne raspodele postoje u Excel-u).

95% interval poverenja za prosečnu masu studenata je oblika $\left(\bar{X} - z\frac{S}{\sqrt{n-1}}, \bar{X} + z\frac{S}{\sqrt{n-1}}\right)$. Na osnovu uzorka smo već našli neophodne tačkaste ocene, samo je na osnovu nivoa poverenja $\beta = 0{,}95$ i uslova potrebno naći konstantu z. U Excel-u postoji funkcija *NORMSINV* koja predstavlja inverznu funkciju raspodele slučajne veličine sa $\mathrm{N}(0{,}1)$ raspodelom. ($z = F^{-1}(\beta + \frac{1-\beta}{2})$)

10	z	1.959964			= NORMSINV (0.975)
11	$\frac{S}{\sqrt{n-1}}$	0.29349			uzoračko standardno odstupanje srednje vrednosti
12	Donja granica	67.37447			=B7-A10*B11
13	Gornja granica	68.52523			=B7+A10*B11

Sa pouzdanošću od 95% zaključujemo da se prosečna masa studenata nalazi u intervalu $(67{,}37477, 68{,}52523)$.

d) **Testirati hipotezu $\boldsymbol{H_0}: \boldsymbol{a = 69}$ protiv alternative $\boldsymbol{H_1}: \boldsymbol{a} < 69$ za prag značajnosti $\boldsymbol{\alpha = 0.01}$.**

Pogledati fajl Testiranje statističkih hipoteza. Parametarski testovi . Na realizovanom uzorku izračunavamo vrednost test statistike $U = \frac{(\bar{X}-a_0)\sqrt{n-1}}{S}$, i na osnovu datog praga značajnosti, pretpostavke da test statistika ima Studentovu raspodelu sa $n-1 = 99$ stepeni slobode (da je H_0 tačna) i oblika alternativne hipoteze $H_1: a < 69$, određujemo kritičnu oblast testa i dolazimo do zaključka.

Koristeći prethodno izračunate statistike, odmah možemo izračunati vrednost test statistika ovog testa:

14	u	-3.57764			= (B7-69)/B11
15	t	2.364606			=TINV(0.01*2,99)

TINV(verovatnoća,broj stepeni slobode) vraća onu konstantu t za koju važi $P\{|X| > t\} = \text{verovatnoći}$. Zbog toga se se kod jednostranih testova (alternative oblika “<” ili “>”) kritična vrednost t nalazi za verovatnoću $2 * \alpha$.

Na osnovu nađene kritične vrednosti i oblika alternativne hipoteze, kritična oblast je $C = (-\infty, -2{,}364606]$. Realizovana vrednost test statistike $u \in C$, tako da hipotezu $H_0: a = 69$ odbacujemo.

16	$H_1: a < 69$	Odbacujemo nultu hipotezu	=IF(B14<-B15,"Odbacujemo nultu hipotezu","Ne odbacujemo nultu hipotezu")

Napomena: Da smo imali drugačije alternativne hipoteze kako bismo sproveli testiranje:

17	$H_1: a > 69$	Ne odbacujemo nultu hipotezu	=IF(B14>B15,"Odbacujemo nultu hipotezu","Ne odbacujemo nultu hipotezu")
18	$H_1: a \neq 69$	t =2.626405 (=TINV(0.01,99))	Ne odbacujemo nultu hipotezu (=IF(OR(B14<-2.626405,B14>2.626405),"Odbacujemo nultu hipotezu","Ne odbacujemo nultu hipotezu"))

e) **Za isti prag značajnosti testirati hipotezu $H_0: \sigma^2 = 8,6$ protiv alternative $H_1: \sigma^2 \neq 8,6$.**

Pogledati fajl Testiranje statističkih hipoteza. Parametarski testovi .

Na realizovanom uzorku izračunavamo vrednost test statistike $U = \frac{nS^2}{{\sigma_0}^2}$, koristeći prethodno izračunatu statistiku:

19	u	99.15698			= B6*B8/8.6

Na osnovu datog praga značajnosti $\alpha = 0.01$, ako je $H_1: \sigma^2 \neq 8{,}6$ iz uslova $P\{U \leq c_1\} = \frac{\alpha}{2}$ i $P\{U \geq c_2\} = \frac{\alpha}{2}$, ako U ima χ^2 raspodelu sa $n - 1 = 99$ stepeni slobode (pod pretpostavkom da je H_0 tačna), nalazimo vrednosti c_1 i c_2.
CHIINV(verovatnoća,broj stepeni slobode) u Excel-u vraća onu konstantu c za koju važi $P\{X \geq c\} =$ verovatnoći.

20	c_1	66.51011			= CHIINV(1-0.01/2,99)
21	c_2	138.9868			= CHIINV(0.01/2,99)

Kritična oblast je $C = [0, c_1] \cup [c_2, +\infty)$. Realizovana vrednost test statistike $u \notin C$, tako da hipotezu $H_0: \sigma^2 = 8{,}6$ ne odbacujemo.

22	$H_1: \sigma^2 \neq 8{,}6$	Ne odbacujemo nultu hipotezu	=IF(OR(B19<B20,B19>B21),"Odbacujemo nultu hipotezu","Ne odbacujemo nultu hipotezu")

Napomena: Da smo imali drugacije alternativne hipoteze kako bismo sproveli testiranje:

23	$H_1: \sigma^2 > 8{,}6$	kritična vrednost c=134.6416 (=CHIINV(0.01,99))	$C = [c, +\infty) \Rightarrow u \notin C \Rightarrow$ Ne odbacujemo nultu hipotezu
24	$H_1: \sigma^2 < 8{,}6$	kritična vrednost c=69.22989 (=CHIINV(1-0.01,99))	$C = [0, c] \Rightarrow u \notin C \Rightarrow$ Ne odbacujemo nultu hipotezu

f) **Ispitati hipotezu da masa studenata ima normalnu raspodelu sa pragom značajnosti $\alpha = 0,05$.**

Pogledati fajl Hi-kvadrat test.

Da bismo izračunali na osnovu uzorka vrednost test statistike hi-kvadrat testa $U = \sum_{k=1}^{r} \frac{(N_k - np_k)^2}{np_k}$, formiramo u Excel-u sledeću tabelu:

	A	B	C	D	E	F
1	I_k		n_k	p_k	$n * p_k$	$\frac{(n_k - n * p_k)^2}{n * p_k}$
2	I klasa	$(-\infty, 63)$	5	0.043921	4.392064	0.084149
3	II klasa	$(63,66]$	18	0.20674	20.67395	0.345847
4	III klasa	$(66,69]$	42	0.39069	39.06896	0.219893
5	IV klasa	$(69,72]$	27	0.277376	27.7376	0.019615
6	V klasa	$(72, +\infty)$	8	0.081274	8.127421	0.001998
7			$n = 100$ (=SUM(C2:C6))	1 (=SUM(D2:D6))	E2=C$7*D2	0.6715 (=SUM(F2:F6))

Prvo smo formirali $r = 5$ disjunktnih klasa na osnovu intervalno datog uzorka, uz obaveznu proveru da frekvencija svake klase mora biti bar 5. U kolonu C uneli učestanosti, u ćeliji C7 izračunali obim uzorka.

U koloni D smo računali teorijske verovatnoće pod pretpostavkom da je nulta hipoteza tačna, da masa studenata ima normalnu raspodelu. Samo je pre toga potrebno, ako nisu dati, oceniti parametre normalne raspodele: $\hat{a} = \bar{x}$ i $\hat{\sigma}^2 = S^2$, što smo mi učinili kroz prethodne etape primera ($l = 2$). Koristeći dobijene vrednosti ($\bar{x} = 67{,}95$ i $s = 2{,}9$) i funkciju *NORMDIST* u Excel-u koja odgovara funkciji raspodele slučajne veličine sa $N(a, \sigma^2)$ raspodelom, dolazimo do “verovatnoća svake klase”.

D	
p_k	
0.043921	=NORMDIST(63,67.95,2.9, TRUE)
0.20674	=NORMDIST(66,67.95,2.9, TRUE)-NORMDIST(63,67.95,2.9, TRUE)
0.39069	=NORMDIST(69,67.95,2.9, TRUE)-NORMDIST(66,67.95,2.9, TRUE)
0.277376	=NORMDIST(72,67.95,2.9, TRUE)-NORMDIST(69,67.95,2.9, TRUE)
0.081274	=1-NORMDIST(72,67.95,2.9, TRUE)
1	=SUM(D2:D6)

U koloni F merimo odstupanje teorijskih od empirijskih frekvencija svake klase (F2=(C2-E2)^2/E2). Zbirno odstupanje predstavlja realizovanu vrednost test statistike u (F7). Ako je nulta hipoteza tačna, realizovana vrednost test statistike ne bi smela biti velika tj. trebalo bi da bude u granicama dozvoljenog odstupanja.

Za dovoljno veliko n raspodela statistike U se može aproksimirati χ^2 raspodelom sa $r - l - 1 = 2$ stepeni slobode, tako da je za zadati prag značajnosti $\alpha = 0{,}05$ najveća dozvoljena vrednost ove statistike konstanta c , za koju je $P\{U \geq c\} = 0{,}05$.

8	c	5.991465	=CHIINV(0.05,2)			

Uočavamo da $u \notin C = [c, +\infty)$, tako da nultu hipotezu, da je raspodela mase studenata normalna, ne odbacujemo.