Empiiriline jaotus

Andmed, sagedustabel, empiiriline jaotus, suhteline sagedus, näitlikustamine, histogramm, tulpdiagramm, kahe tunnuse ühisjaotus, risttabel

Põhikool

Andmete kogumisele järgneb tavaliselt andmete korrastamine. Andmete korrastamisel on enamasti sobiv esimese sammuna luua iga tunnuse jaoks **sagedustabel**, millesse märgitakse kõik tunnuse väärtused ja iga väärtuse sagedus, mis näitab mitu korda see väärtus andmestikus esineb.

Näide. Olgu klassis 25 õpilast. Küsiti, mitu õde või venda igaühel neist on. Tulemused on järgmises tabelis:

Õdede ja vendade arv	0	1	2	3	Kokku
Õpilaste arv, kellel on nii palju õdesid-vendi	7	12	4	2	25

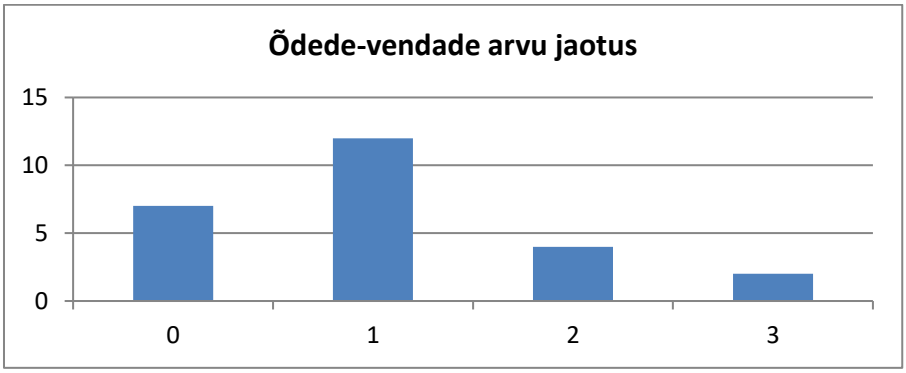
Sagedustabeli põhjal koostatakse tunnuse **empiiriline jaotus**. Sõna „empiiriline“ tähistab siin seda, et andmed on saadud reaalsete katsete või mõõtmiste, mitte teoreetiliste arutelude tulemusena. Kui see on selge, võib selle sõna ära jätta.

Jaotust kirjeldab sagedustabeli sarnane tabel, kus igale tunnuse väärtusele vastab tema **suhteline sagedus (osakaal).** Suhteline sagedus leitakse, jagades iga väärtuse sageduse vaatluste koguarvuga. Sageli esitatakse suhtelised sagedused protsentidena. Tunnuse väärtuste loetelu koos nende suhteliste sagedustega moodustab tunnuse jaotuse. Suhteliste sageduste summa on alati 1 või 100%.

Leiame nüüd ka õdede ja vendade arvu jaotuse

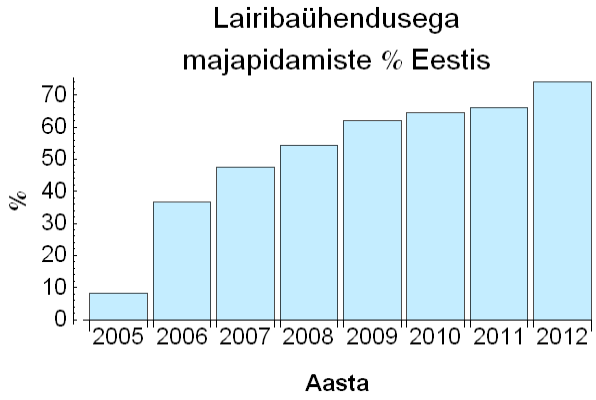
Õdede ja vendade arv	0	1	2	3	Kokku
Õpilaste arv, kellel on nii palju õdesid-vendi	7	12	4	2	25
Suhtelised sagedused	28%	48%	16%	8%	100%

Jaotuse **näitlikustamiseks** on sobiv kasutada **histogramm**i või **tulpdiagrammi,** kuid on ka muid võimalusi. Oluline on see, et esitatav diagramm illustreerib jaotust adekvaatselt ja ei tekita mõne optilise illusiooni tõttu vaatajates ekslikku muljet.

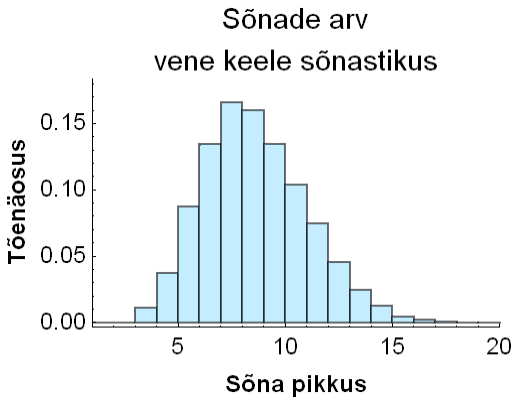


Kui tunnusel on erinevaid väärtusi väga palju, siis on jaotuse esitamisel otstarbekas väärtusi rühmade kaupa kokku võtta, et anda graafikule või tabelile kompaktsem kuju, kuigi sellega kaasneb teatav infokadu.

Tunnuse empiirilise jaotuse graafilise esituse näiteid



Esitus tulpdiagrammi abil



Esitus histogrammi abil

Gümnaasium

Mitme tunnuse ühisjaotus

Tavaliselt sisaldab andmestik mitme tunnuse andmeid. Kahe tunnuse ühist empiirilist jaotust esitab kõige paremini nn **risttabel**. Näitena esitame liiklusõnnetustel kannatanute andmeid esitava tabeli, milles on kaks kahe väärtusega tunnust: kannatanu staatus (juht/kaasreisija) ja kannatanu olukord (hukkunud/ vigastatud, kuid elus)

Kannatanu staatus	Kannatanu olukord		
	Hukkunud	Vigastatud, elus	Kokku
Juht	16	1416	1432
Kaasreisija	6	593	599
Kokku	22	2009	2031

Nende andmete põhjal saab arvutada tunnuste empiirilise ühisjaotuse, mis annab tabeli iga lahtri jaoks suhtelise sageduse. Selle leidmiseks jagatakse kõik tabelid olevad sagedused andmete koguarvuga, mis on tabeli parempoolses alumises nurgas, praegu on see 2031.

Tavaliselt esitatakse suhtelised sagedused protsentidena:

Kannatanu staatus	Kannatanu olukord		
	Hukkunud	Vigastatud, elus	Kokku
Juht	0,79	69,72	70,51
Kaasreisija	0,29	29,20	29,49
Kokku	1,08	98,92	100

Kahe tunnuse ühisjaotuse tabeli viimane rida ja viimane veerg esitavad kummagi tunnuse (ühemõõtmelise) jaotuse.

Kannatanu staatus:
Juht ……70,51
Kaasreisija ….29,49.

Kannatanu olukord
Hukkunud……..1,08
Vigastatud, elus….98,92

Millele pöörata tähelepanu

Vahel on mõned andmepunktid teistest oluliselt erinevad ja paiknevad ülejäänutest eemal. Neid nimetatakse **erinditeks**. Erindite puhul tuleb selgitada, kas on tegemist vaatlusvigadega, objektidega, mis ei kuulu uuritavasse üldkogumisse või on need õiged ja antud üldkogumisse kuuluvad objektid. Erindite kaasamine või väljajätmine võib uurimistulemusi märgatavalt mõjutada. Kui erind uurimistulemuste hulgast kõrvaldatakse, tuleb seda analüüsimisel tingimata kajastada: missugune oli erindi väärtus ja missugusel kaalutlusel see välja jäeti.

Seotud mõisted:

Tõenäosus, tõenäosusjaotus, andmed, mudel