F2
Beskrivning av ett datamaterial. Tabellering och val av diagram. Summatecknet

Tabellering av kvalitativ variabel

En kvalitativ variabel varierar över ett antal *kategorier*.

Antag att vi har observerat 300 personer och är intresserade av variabeln KÖN.

Ange antalet individer som tillhör varje kategori i en *frekvenstabell*.

Kön	**Antal personer**	**Andel personer (%)**
Män	120	40
Kvinnor	180	60
Totalt	300	100

Den *relativa frekvensen* för män ges av

$$\frac{120}{300} \times 100 = 40(\%).$$

I tabellen återges den *absoluta* och den *relativa(procentuella)* fördelningen.

Diagram över kvalitativ variabel: cirkeldiagram

Andelarna kan illustreras som tårtbitar eller *cirkelsektorer* med en viss *medelpunktsvinkel*.

Hela varvet i en cirkel utgör 360°. Vilken medelpunktsvinkel skall andelen män ha?

$$40\% \text{ av } 360^\circ = 0,40 \times 360 = 144^\circ.$$

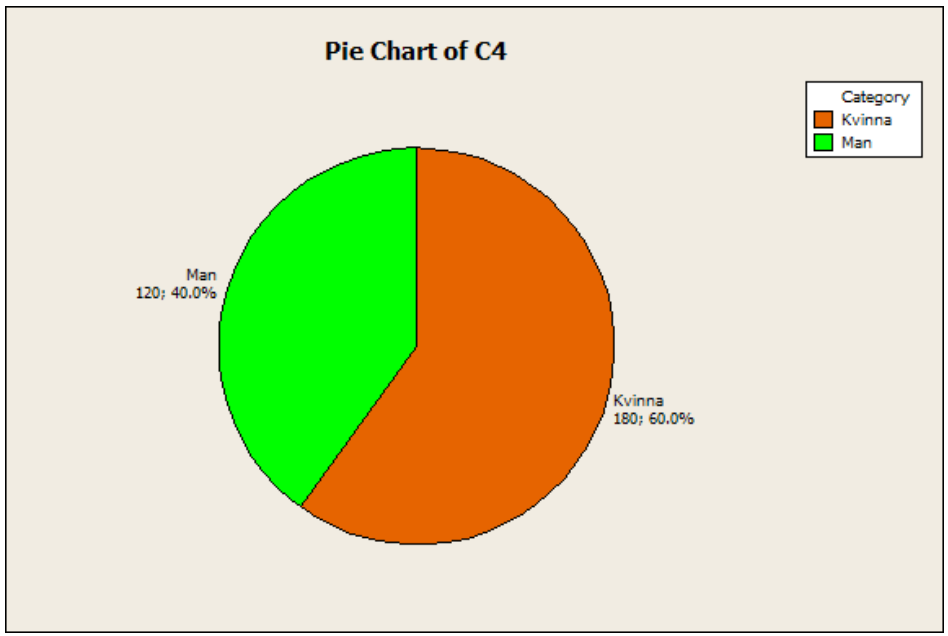
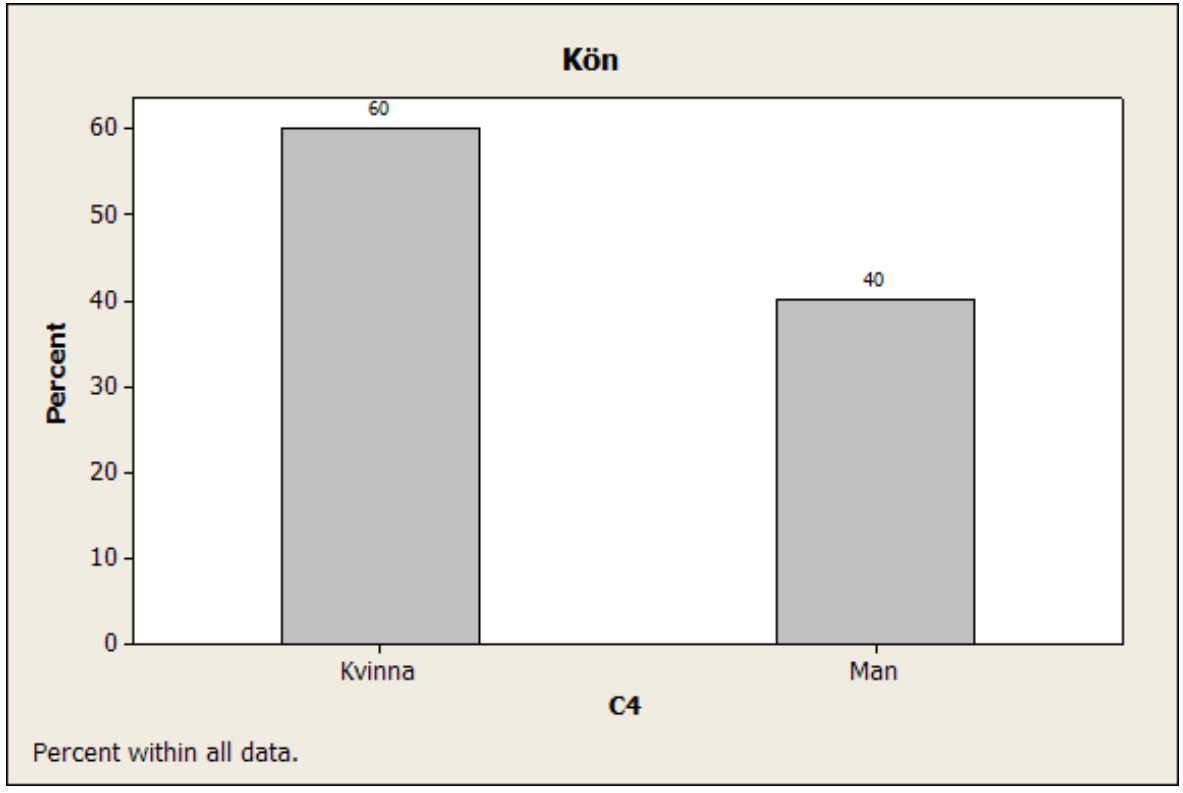


Diagram över kvalitativ variabel: stapeldiagram

Vi kan även rita staplar vars höjd motsvarar andelarna.



Tabellering av två kvalitativa variabler i kombination-korstabellering

Om vi har observationer på två variabler kan vi (förstås) göra två frekvenstabeller.

Om vi vill studera *samvariationen* mellan de två variablerna gör vi en *korstabell*.

Vi tabellerar *observationspar* istället för observationer.

Ange antalet individer som tillhör varje par av kategorier.

Exempel. Civilstånd och valdeltagande.
Individerna är antingen gifta(G) eller ej gifta(EG) och har röstat ($= 1$) eller ej röstat ($= 0$). Det ger fyra möjligheter

$$(\mathrm{G}, 0), (\mathrm{G}, 1), (\mathrm{EG}, 0), (\mathrm{EG}, 1).$$

Arbetstabell

Antag att de fyra första observationerna blir

$$(\mathrm{G}, 0), (\mathrm{EG}, 1), (\mathrm{G}, 1), (\mathrm{G}, 1).$$

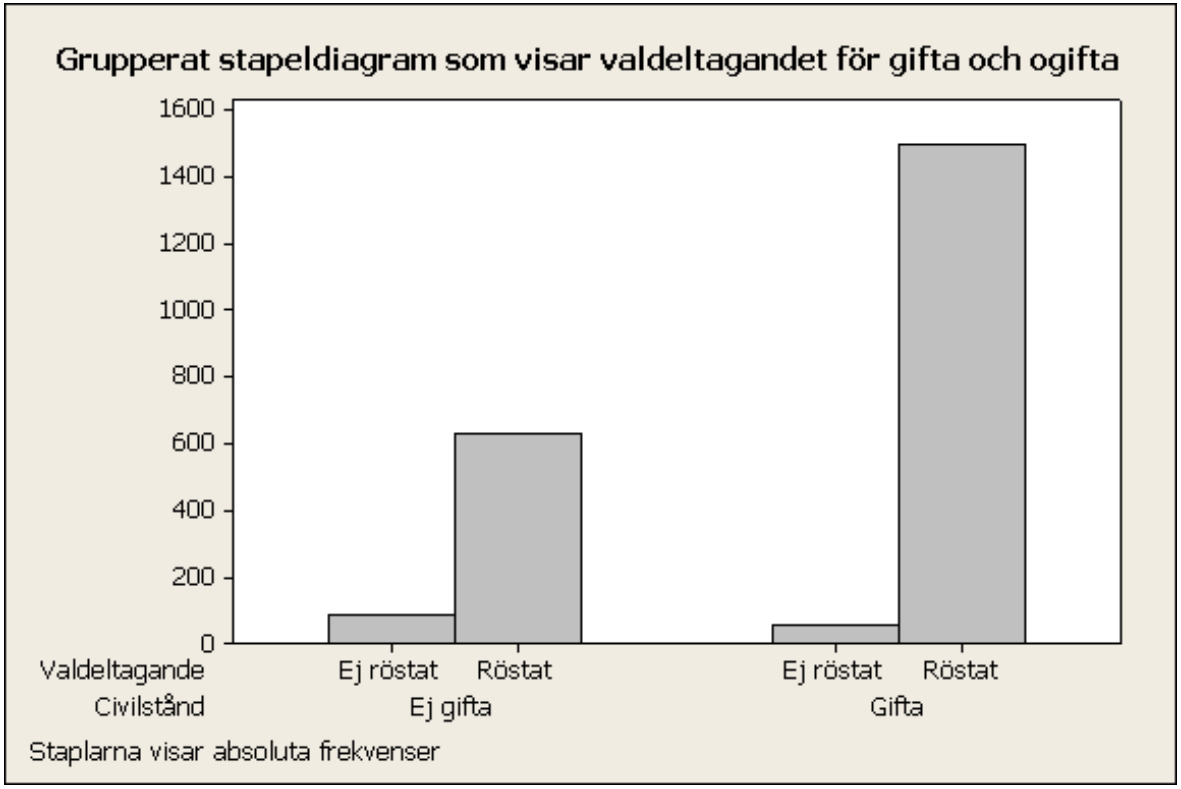
Civilstånd	**Valdeltagande**	
	0	**1**
G	/	//
EG		/

När alla observationer avprickats kan vi framställa korstabellen.

Civilstånd	**Valdeltagande**	
	Ej röstat	**Röstat**
Gifta	54	1496
Ej gifta	85	628

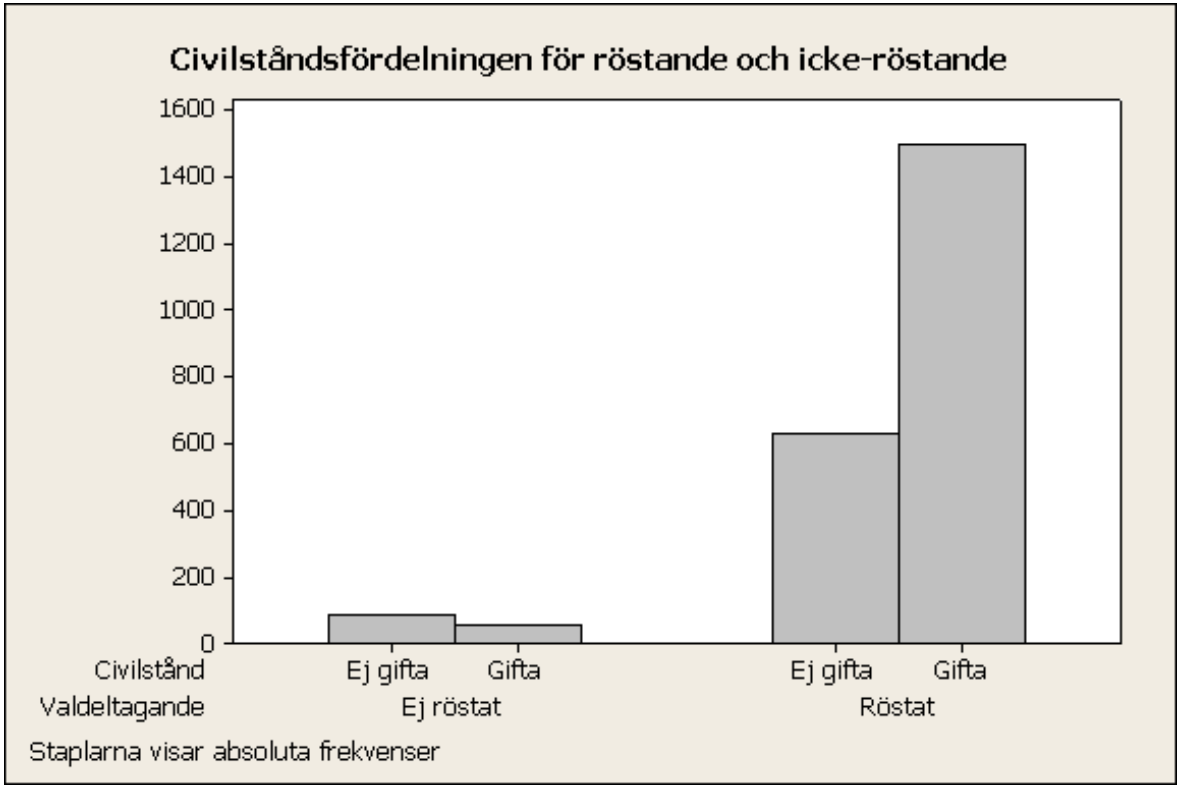
Korstabell eller grupperade staplar?

Civilstånd	Valdeltagande	
	Ej röstat	**Röstat**
Gifta	54	1496
Ej gifta	85	628



Korstabell eller grupperade staplar?

Civilstånd	Valdeltagande	
	Ej röstat	**Röstat**
Gifta	54	1496
Ej gifta	85	628



Korstabell-marginalfördelningar

Civilstånd	**Valdeltagande** **Ej röstat**	**Röstat**	**Summa**
Gifta	54	1496	1550
Ej gifta	85	628	713
Summa	139	2124	2263

Vi har summerat radvis och kolumnvis. T ex över gifta och ogifta som inte röstat

$$54 + 85 = 139.$$

Dessa fördelningar kallas *marginalfrekvenserna* eller *marginalfördelningarna*.

Marginalfördelningar

De bägge marginalfördelningarna kan framställas på vanligt vis. Den absoluta och den relativa fördelningen för CIVILSTÅND ges som

Civilstånd	**Antal personer**	**Andel personer (%)**
Gifta	1550	68,5
Ej gifta	713	31,5
Totalt	2263	100

Den absoluta och den relativa fördelningen för VALDELTAGANDE ges som

Valdeltagande	**Antal personer**	**Andel personer (%)**
Ej röstat	139	6,1
Röstat	2124	93,9
Totalt	2263	100

Dessa fördelningar kan avläsas ur korstabellen.

Procentuell uppdelning radvis

För att jämföra hur de gifta och de ogifta röstar, räknar vi om till *radprocent*.

	Valdeltagande		
Civilstånd	**Ej röstat**	**Röstat**	**Summa**
Gifta	54	1496	1550
Ej gifta	85	628	713
Summa	139	2124	2263

T ex

$$\frac{54}{1550} \times 100 = 3,5(\%).$$

och

$$\frac{1496}{1550} \times 100 = 96,5(\%).$$

Procentuell uppdelning radvis

Gör vi samma beräkningar för de ogifta, så får vi

	Valdeltagande		
Civilstånd	**Ej röstat**	**Röstat**	**Summa**
Gifta	3,5	96,5	100
Ej gifta	11,9	88,1	100
Summa	6,1	93,9	100

- Tabellen visar hur fördelningen av individer på variabeln VALDELTAGANDE är betingad av om individerna är gifta eller ogifta. Vi har två *betingade fördelningar*.
- Vi har beräknat procenten i horisontell riktning, men jämför procenttalen i de vertikala kolumnerna.
- I den första kolumnen ökar andelen icke-röstande när vi går från gifta till ogifta. I den andra tvärtom.

Samband mellan kvalitativa variabler

För att utröna om det finns ett samband mellan variablerna VALDELTAGANDE och CIVILSTÅND jämför man de betingade fördelningarna (som vi gjort).

Vi såg ovan att valdeltagandet skiljer sig åt mellan grupperna gifta och ogifta. Variabeln VALDELTAGANDE *beror* av variabeln CIVELSTÅND.

Hur skulle tabellen se ut om variablerna istället var *oberoende*?

Då skulle andelarna i de vertikala kolumnerna vara ungefär lika.

	Valdeltagande		
Civilstånd	**Ej röstat**	**Röstat**	**Summa**
Gifta	LIKA ↓	LIKA ↓	100
Ej gifta	↑ LIKA	↑ LIKA	100

Avslutande anmärkningar

- Vi kan beräkna kolumnprocent istället för radprocent.
- Vi har tittat på två egenskaper samtidigt, man kan gå vidare till tre eller fler egenskaper samtidigt.
  Antag att vi även är intresserade av skillnader mellan könen.
  Då har vi åtta olika kategorier

$$(\mathrm{G}, 0, Kv), (\mathrm{G}, 1, Kv), (\mathrm{EG}, 0, Kv), (\mathrm{EG}, 1, Kv)$$

$$(\mathrm{G}, 0, M), (\mathrm{G}, 1, M), (\mathrm{EG}, 0, M), (\mathrm{EG}, 1, M).$$

Gifta(G) eller ej gifta(EG) och har röstat ($= 1$) eller ej röstat ($= 0$)

Tabellering av kvantitativ variabel

Grundprincipen vid tabellering av kvantitativa variabler är att ange observationerna i storleksordning.

Antag att vi har n st observationer $x_1, \ldots, x_n$ som antar k stycken olika värden. Vi har fler observationer än värden på variabeln, så $n > k$.

Matematikbetyg Matematikbetyget för 25 elever (på den gamla goda tiden)

5 4 1 4 4 3 2 3 3 3 4 2 3 1 3 3 5 4 2 2 2 4 3 5 3.

När data framställs på detta vis kallar vi detta *ogrupperade data*. Låt

- $x_i =$ värdena på observationerna, $i = 1, 2, \ldots, n$
- $f_i =$ frekvensen för det i:te variabelvärdet, $i = 1, 2, \ldots, k$

Här är $n = 25$ och $k = 5$ (fem olika värden på variabeln MATEMATIKBETYG).

Tabellering av kvantitativ variabel

5 4 1 4 4 3 2 3 3 3 4 2 3 1 3 3 5 4 2 2 2 4 3 5 3.

Betyg (x_i)	**Avprickning**	**Frekvens** (f_i)
1	||	2
2	|||||	5
3	||||| ||||	9
4	||||| |	6
5	|||	3
		25

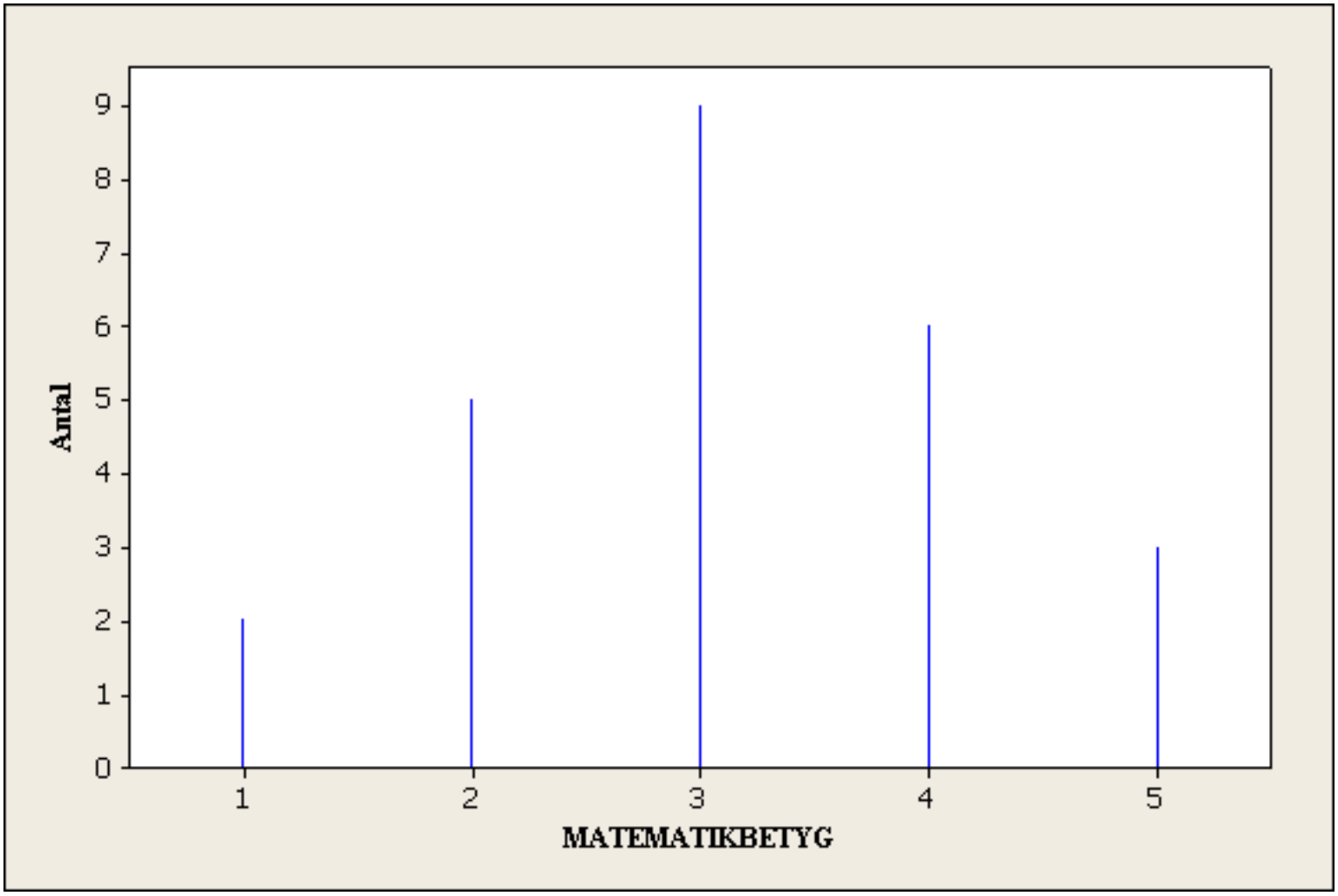
Notera att

$$\sum_{i=1}^{k} f_i = 25 = n.$$

En formel som alltid gäller. En frekvenstabell innebär att vi har *grupperat data*.

Stolpdiagram

Eftersom variabeln MATEMATIKBETYG är diskret och antar endast ett fåtal variabelvärden, så illustrerar vi fördelningen med ett *stolpdiagram*.



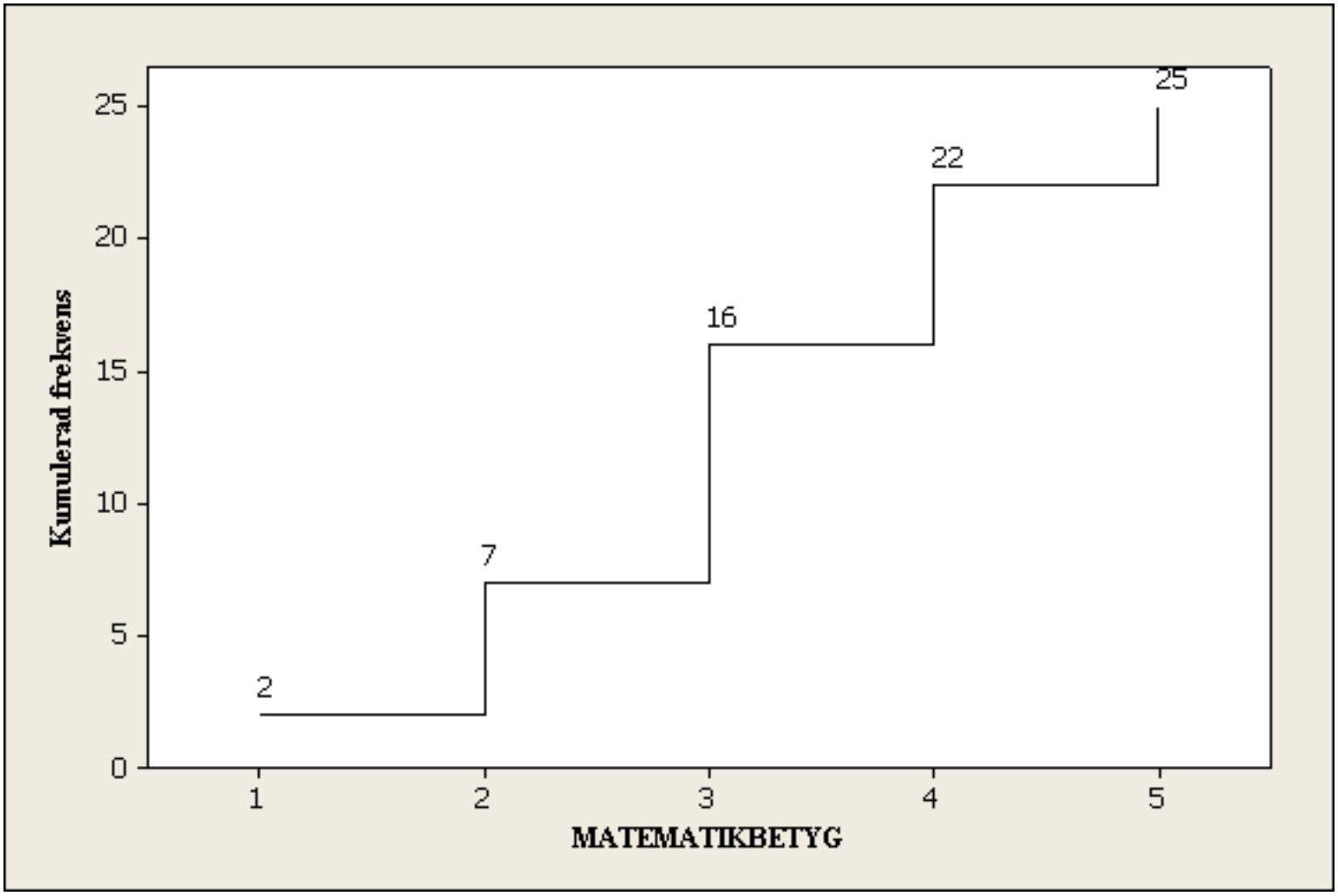
Tabellering av kvantitativ variabel-kumulerad frekvens

Den *kumulativa* frekvensen för ett visst variabelvärde visar hur många av observationerna som är *mindre än eller lika med detta variabelvärde*.

Betyg (x_i)	**Absolut frekvens** (f_i)	**Kumulativ frekvens** (F_i)
1	2	2
2	5	7
3	9	16
4	6	22
5	3	25

Trappstegsdiagram

Den kumulativa fördelningen för en diskret variabel som endast antar ett fåtal variabelvärden kan åskådliggöras i en *trappstegskurva*.



Tabellering av kvantitativ variabel-kumulerad relativ fördelning

Vi kan även beräkna relativtal för att underlätta jämförelser

Betyg (x_i)	**Relativ fördelning (%)**	**Kumulerad relativ fördelning**
1	8	8
2	20	28
3	36	64
4	24	88
5	12	100

Baby Ruth konfektyrer i styck

Ingredients: sugar, roasted peanuts, corn syrup, partially hydrogenated palm kernel and coconut oil, nonfat milk, cocoa, high fructose corn syrup, and less than 1% of glycerin, whey (from milk), dextrose, salt, monoglycerides, soy lecithin, soybean oil, natural and artificial flavors, carrageenan, TBHQ and citric acid (to preserve freshness), caramel color.

Tabellering av kvantitativ variabel-Babe Ruth konfektyrer i styck

Fyrtio Babe Ruth konfektyrer vägdes och vikterna sorterades i storleksordning.

20,5	20,7	20,8	21,0	21,0	21,4	21,5	22,0	22,1	22,5
22,6	22,6	22,7	22,7	22,9	22,9	23,1	23,3	23,4	23,5
23,6	23,6	23,6	23,9	24,1	24,3	24,5	24,5	24,8	24,8
24,9	24,9	25,1	25,1	25,2	25,6	25,8	25,9	26,1	26,7

VIKT är en kontinuerlig variabel, så vi måste klassindela observationerna. Vi väljer fem klasser som är 1.3 breda och startar i 20.4, d v s klassindelningen

20.4-21.6, 21.7-22.9, 23.0-24.2, 24.3-25.5, 25.6-26.9

Klassbredden är skillnaden mellan den övre och den undre gränsen för en klass. Den *undre gränsen* för klassen 20.4-21.6 är 20.35. Den *övre gränsen* är 21.65.

Tabellering av kvantitativ variabel-Babe Ruth konfektyrer i styck

Vi tabellerar datamaterialet i en frekvenstabell

Vikt/gram	**Avprickning**	**Frekvens**
20.4-21.6	|||||||	7
21.7-22.9	||||||||	9
23.0-24.2	||||||||	9
24.3-25.5	|||||||||	10
25.6-26.9	|||||	5

Eftersom data är ordnade i storleksordning är det lätt att finna frekvenserna.

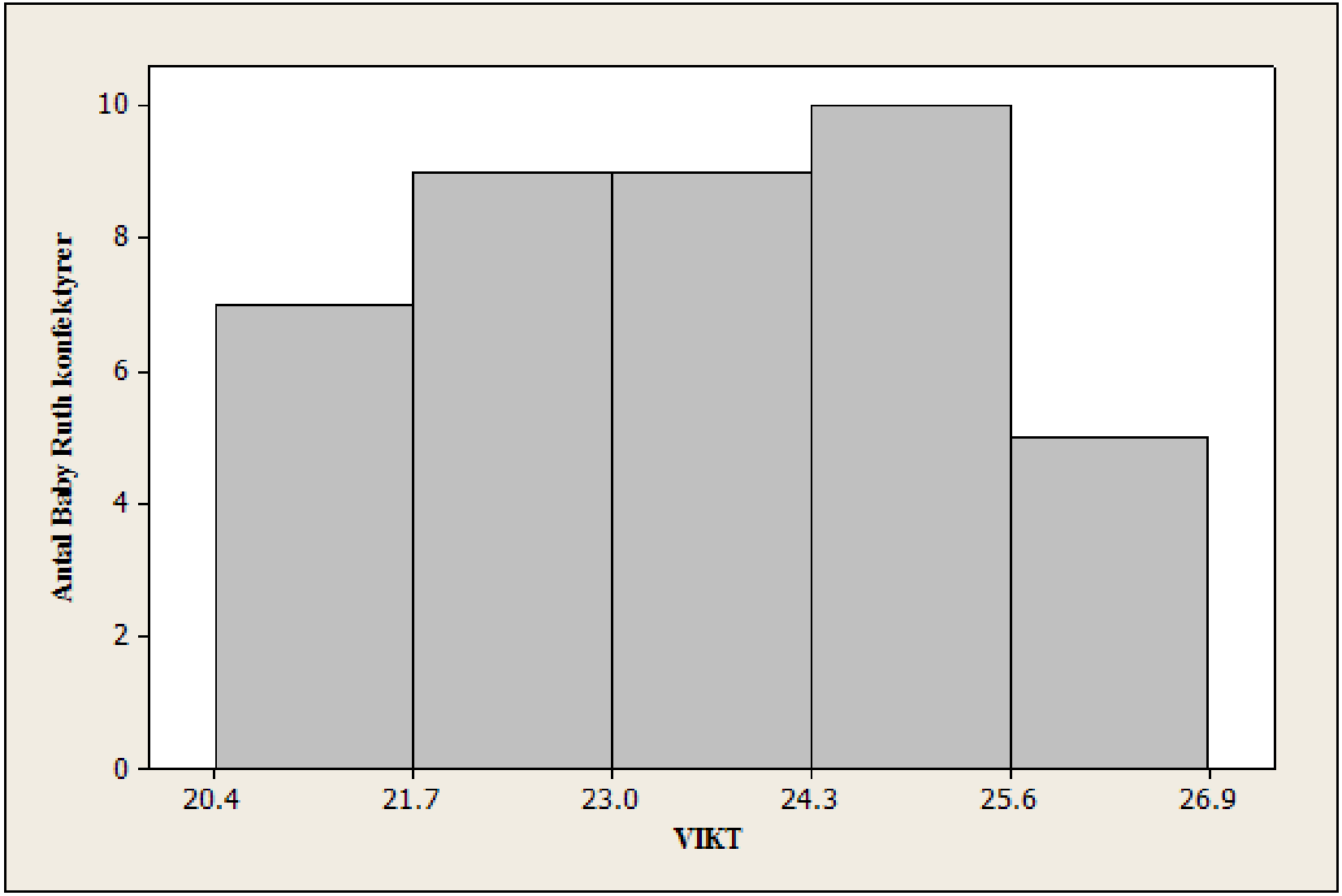
20.5	20.7	20.8	21.0	21.0	21.4	21.5	22.0	22.1	22.5
22.6	22.6	22.7	22.7	22.9	22.9	23.1	23.3	23.4	23.5
23.6	23.6	23.6	23.9	24.1	24.3	24.5	24.5	24.8	24.8
24.9	24.9	25.1	25.1	25.2	25.6	25.8	25.9	26.1	26.7

Tabellering av kvantitativ variabel-Babe Ruth konfektyrer i styck

För att få svar på hur många observationer som finns under ett visst värde beräknar vi kumulativa frekvenser.

Vikt/gram	**Absolut frekvens (f)**	**Kumulativ frekvens (F)**	**Relativ fördelning (%)**	**Kumulativ relativ fördelning (%)**
20.4-21.6	7	7	17,5	17,5
21.7-22.9	9	16	22,5	40,0
23.0-24.2	9	25	22,5	62,5
24.3-25.5	10	35	25,0	87.5
25.6-26.9	5	40	12,5	100
	40		100	

Babe Ruth konfektyrer i styck-histogram



Babe Ruth konfektyrer i styck-stamblad-diagram

20.5	20.7	20.8	21.0	21.0	21.4	21.5	22.0	22.1	22.5
22.6	22.6	22.7	22.7	22.9	22.9	23.1	23.3	23.4	23.5
23.6	23.6	23.6	23.9	24.1	24.3	24.5	24.5	24.8	24.8
24.9	24.9	25.1	25.1	25.2	25.6	25.8	25.9	26.1	26.7

Djup		
3	20	578
6	21	004
7	21	5
9	22	01
16	22	5667799
19	23	134
(5)	23	56669
16	24	13
14	24	558899
8	25	112
5	25	689
2	26	1
1	26	7

Exempel på följder av tal

Låt $x_1, x_2, \ldots, x_n$ vara n st tal.

x_1	x_2	x_3	x_4	x_5
1	2	3	4	5

eller

x_1	x_2	x_3	x_4	x_5
1	1	1	1	1

eller

x_1	x_2	x_3	x_4	x_5
1	2	4	8	16

Summatecknet

Summan av $x_1, x_2, \ldots, x_5$ skrivs som

$$x_1 + x_2 + x_3 + x_4 + x_5.$$

Hur ska vi skriva upp summor av väldigt många tal? Vi använder den grekiska bokstaven stora sigma, $\sum$, på följande sätt

$$x_1 + x_2 + x_3 + x_4 + x_5 = \sum_{i=1}^{5} x_i.$$

Symbolkombinationen utläses som ”summan av x_i, då i går från 1 till 5”. Σ kallas då *summatecknet*. Se KD sid 375 för mer information.

Summationsindex

Bokstaven i kallas *summationsindex* och kan väljas hur som helst!

$$1 + 2 + 3 + 4 + 5 = \sum_{j=1}^{5} j.$$

$$1 + 1 + 1 + 1 + 1 = \sum_{\nu=1}^{5} 1.$$

(Den grekiska bokstaven ν uttalas "ny").

$$1 + 2 + 4 + 8 + 16 = 2^0 + 2^1 + 2^2 + 2^3 + 2^4 = \sum_{m=0}^{4} 2^m.$$

Ytterligare exempel

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 = \sum_{n=1}^{5} x_n^2.$$

$$\begin{aligned} & (x_1 - \overline{x}) + (x_2 - \overline{x}) + (x_3 - \overline{x}) \\ + \quad & (x_4 - \overline{x}) + (x_5 - \overline{x}) = \sum_{k=1}^{5} (x_k - \overline{x}) = 0. \end{aligned}$$

Produkttecknet

Produkten av $x_1, x_2, \ldots, x_5$ skrivs som

$$x_1 x_2 \cdots x_5 = \prod_{i=1}^{5} x_i.$$

Här använder vi istället den grekiska bokstaven stora pi, $\prod$.
Produkttecknet fungerar på precis samma sätt som summatecknet.
Det används inte så ofta i statistiska sammanhang.