

Lecture 3

karl.sigfrid@stat.su.se

Frekvenstabeller - en variabel

Class	Count
First	324
Second	285
Third	710
Crew	889

Table 2.2

A frequency table of the *Titanic* passengers.

Class	Percentage (%)
First	14.67
Second	12.91
Third	32.16
Crew	40.26

Table 2.3

A relative frequency table for the same data.

- ▶ En frekvenstabell redovisar antalet observationer i varje kategori.
- ▶ En relativ frekvenstabell visar andel istället för antal.
- ▶ Har vi flera kategoriska variabler kan vi göra en separat frekvenstabell för var och en. Tabeller över enskilda variabler visar dock inte *samband* mellan variabler.

Korstabeller - två variabler

- ▶ En **korstabell (contingency table)** visar samband mellan två variabler.
- ▶ Korstabellen nedan visar resultat ur en svensk studie som undersökte om det finns samband mellan prostatacancer och hur ofta en person äter fisk.

		Prostate Cancer		Total
		No	Yes	
Fish Consumption	Never/Seldom	110	14	124 (2.0%)
	Small Part of Diet	2420	201	2621 (41.8%)
	Moderate Part	2769	209	2978 (47.5%)
	Large Part	507	42	549 (8.8%)
	Total	5806 (92.6%)	466 (7.4%)	6272 (100%)

Korstabeller - två variabler

- ▶ Den ena variabeln delar in deltagarna i **fyra kategorier baserat på diet**: De som aldrig/sällan äter fisk, de som äter lite fisk, de som äter måttligt med fisk, och de som äter mycket fisk.
- ▶ Den andra variabeln delar in deltagarna i **två kategorier**: De som diagnostiserades med prostatacancer under studieperioden och de som inte diagnostiserades med prostatacancer.

		Prostate Cancer	
		No	Yes
Fish Consumption	Never/Seldom	110	14
	Small Part of Diet	2420	201
	Moderate Part	2769	209
	Large Part	507	42
	Total	5806 (92.6%)	466 (7.4%)
		Total	
		124 (2.0%)	
		2621 (41.8%)	
		2978 (47.5%)	
		549 (8.8%)	
		6272 (100%)	

Korstabeller - simultana fördelningar

- ▶ I det **gula fältet** ser vi en **simultanfördelning (joint distribution)**. En simultanfördelning delar in observationerna i grupper baserat på två eller fler variabler.
- ▶ Om en variabel har 4 kategorier och en annan variabel har 2 kategorier får vi sammanlagt $4 \cdot 2 = 8$ kategorier, en för varje möjlig kombination.

		Prostate Cancer		Total
		No	Yes	
Fish Consumption	Never/Seldom	110	14	124 (2.0%)
	Small Part of Diet	2420	201	2621 (41.8%)
	Moderate Part	2769	209	2978 (47.5%)
	Large Part	507	42	549 (8.8%)
	Total	5806 (92.6%)	466 (7.4%)	6272 (100%)

Korstabeller - simultana fördelningar

- ▶ Om vi lägger samman talen i det gula fältet blir summan 6272, som är det totala antalet observationer (dvs antalet deltagare i studien).
- ▶ Den simultana fördelningen visar exempelvis att det fanns det 110 deltagare i studien som sällan/aldrig åt fisk **och** som inte fick prostatacancer.

		Prostate Cancer		Total
		No	Yes	
Fish Consumption	Never/Seldom	110	14	124 (2.0%)
	Small Part of Diet	2420	201	2621 (41.8%)
	Moderate Part	2769	209	2978 (47.5%)
	Large Part	507	42	549 (8.8%)
	Total	5806 (92.6%)	466 (7.4%)	6272 (100%)

Korstabeller - marginalfördelningar

- ▶ **Marginalfördelningen** av en kategorisk variabel visar antalet observationer per kategori utan att vi tar någon hänsyn till den andra variabeln.
- ▶ I den högra marginalen ser vi marginalfördelningen av variabeln för hur ofta deltagarna åt fisk. Där ser vi bland annat att totalt 124 deltagare sällan eller aldrig åt fisk.

		Prostate Cancer		Total
		No	Yes	
Fish Consumption	Never/Seldom	110	14	124 (2.0%)
	Small Part of Diet	2420	201	2621 (41.8%)
	Moderate Part	2769	209	2978 (47.5%)
	Large Part	507	42	549 (8.8%)
	Total	5806 (92.6%)	466 (7.4%)	6272 (100%)

Korstabeller - marginalfördelningar

- ▶ I bottenmarginalen ser vi marginalfördelningen för variabeln prostatacancer. Totalt 466 testdeltagare diagnostiserades med prostatacancer under studien.
- ▶ Varje marginalfördelning summerar till det totala antalet observationer: $124 + 2621 + 2978 + 549 = 5806 + 466 = 6272$

		Prostate Cancer		Total
		No	Yes	
Fish Consumption	Never/Seldom	110	14	124 (2.0%)
	Small Part of Diet	2420	201	2621 (41.8%)
	Moderate Part	2769	209	2978 (47.5%)
	Large Part	507	42	549 (8.8%)
	Total	5806 (92.6%)	466 (7.4%)	6272 (100%)

Korstabeller - marginalfördelningar

- ▶ Marginalfördelningen för en kategorisk variabel är samma fördelning som den vi ser i frekvenstabellen när vi bara inkluderar en variabel.
- ▶ För att bekräfta det skriver vi först ut en korstabell i R.
- ▶ Sedan skriver vi ut frekvenstabellerna för de två variablerna var för sig.

```
tally(~ diet + cancer, data=fish, format="count", margins=T)
```

diet	cancer		
	No	Yes	Total
Never	110	14	124
Small	2420	201	2621
Moderate	2769	209	2978
Large	507	42	549
Total	5806	466	6272

Korstabeller - marginalfördelningar

```
tally(~ diet + cancer, data=fish, format="count", margins=T)
```

diet	cancer		Total
	No	Yes	
Never	110	14	124
Small	2420	201	2621
Moderate	2769	209	2978
Large	507	42	549
Total	5806	466	6272

```
tally(~ diet, data=fish, format="count", margins=T)
```

diet					Total
	Never	Small	Moderate	Large	
	124	2621	2978	549	6272

```
tally(~ cancer, data=fish, format="count", margins=T)
```

cancer			Total
	No	Yes	
	5806	466	6272

Korstabeller - relativa fördelningar

- ▶ Genom att dela varje frekvens med det totala antalet observationer inom en grupp och multiplicera med 100 kan vi få en **relativ frekvenstabell**.
- ▶ Antalet som åt mycket fisk och som inte fick cancer var exempelvis 507, så motsvarande andel blir

$$\frac{507}{6272} \cdot 100\% = 8.0835\%.$$

Korstabeller - relativa fördelningar

I praktiken räknar vi sällan ut andelarna för hand. Genom att sätta format="percent" kan vi skapa en korstabell med relativa frekvenser med tally-funktionen i R.

```
tally(~ diet + cancer, data=fish, format="percent", margins=T)
```

	cancer		
diet	No	Yes	Total
Never	1.7538265	0.2232143	1.9770408
Small	38.5841837	3.2047194	41.7889031
Moderate	44.1485969	3.3322704	47.4808673
Large	8.0835459	0.6696429	8.7531888
Total	92.5701531	7.4298469	100.0000000

Vi har dock fortfarande inte sett någon tabell som enkelt låter oss se samband mellan variablerna, vilket var syftet med korstabellen.

Korstabeller - betingade fördelningar

- ▶ För att kunna se samband mellan variablerna introducerar vi begreppet **betingad fördelning (conditional distribution)**.
- ▶ En betingad fördelning (conditional distribution) är fördelningen av en variabel **givet** ett värde av en annan variabel.
- ▶ Vi kan exempelvis vilja undersöka sambandet mellan prostatacancer och fisk i dieten genom att ställa frågorna:
 - ▶ Hur stor andel av de som *aldrig/sällan* åt fisk fick prostatacancer?
 - ▶ Hur stor andel av de som åt *lite fisk* fick prostatacancer?
 - ▶ Hur stor andel av de som åt *måttligt med fisk* fick prostatacancer?
 - ▶ Hur stor andel av de som åt *mycket fisk* fick prostatacancer?

Korstabeller - betingade fördelningar

- ▶ När vi ställer de här frågorna är vi intresserade av risken för prostatacancer **betingat på** en viss diet.
- ▶ För att få svaren tittar vi på en diet-kategori i taget, där en diet-kategori definieras av hur ofta en person äter fisk.
- ▶ Genom att jämföra personer som tillhör olika diet-kategorier kan vi se om det finns samband.

Korstabeller - betingade fördelningar

- ▶ När vi räknar ut betingade fördelningar intresserar vi oss bara för den kategori som vi betingar på.
- ▶ Vill vi ha fördelningen av variabeln prostatacancer betingat på att deltagaren i studien aldrig/sällan äter fisk ignorerar vi dem som inte tillhör den kategorin.

		Prostate Cancer	
		No	Yes
Fish Consumption	Never/Seldom	110	14
	Small Part of Diet	2420	201
	Moderate Part	2769	209
	Large Part	507	42
	Total	5806 (92.6%)	466 (7.4%)
		Total	
		124 (2.0%)	
		2621 (41.8%)	
		2978 (47.5%)	
		549 (8.8%)	
		6272 (100%)	

Korstabeller - betingade fördelningar

- ▶ När vi räknar ut betingade fördelningar intresserar vi oss bara för den kategori som vi betingar på.
- ▶ Vill vi ha fördelningen av variabeln prostatacancer betingat på att deltagaren i studien aldrig/sällan äter fisk ignorerar vi dem som inte tillhör den kategorin.

		Prostate Cancer		Total
		No	Yes	
Fish Consumption	Never/Seldom	110	14	124 (2.0%)

Korstabeller - betingade fördelningar

- ▶ Vi har 14 deltagare fick prostatacancer och 110 som inte fick det.

		Prostate Cancer		Total
		No	Yes	
Fish Consumption	Never/Seldom	110	14	124 (2.0%)

Korstabeller - betingade fördelningar

- ▶ Vi har 14 deltagare fick prostatacancer och 110 som inte fick det.
- ▶ Vi räknar fram den relativa frekvensen på samma sätt som när vi bara har en variabel. Andelen med cancer är $14/124 = 0.1129$ och andelen utan cancer är $110/124 = 0.8871$. Detta är vår betingade fördelning.

		Prostate Cancer		Total
		No	Yes	
Fish Consumption	Never/Seldom	110	14	124 (2.0%)

Korstabeller - betingade fördelningar

- ▶ Med hjälp av R skapar vi en tabell för fördelningen av variabeln prostatacancer betingat på variabeln diet.
- ▶ Vi anger fördelningen i procent.

```
# Tillägget |> t() vänder på tabellen. Testa både med och utan.  
tally(~cancer|diet,data=fish,format="percent",margins=T) |> t()
```

diet	cancer		
	No	Yes	Total
Never	88.709677	11.290323	100.000000
Small	92.331171	7.668829	100.000000
Moderate	92.981867	7.018133	100.000000
Large	92.349727	7.650273	100.000000

Korstabeller - betingade fördelningar

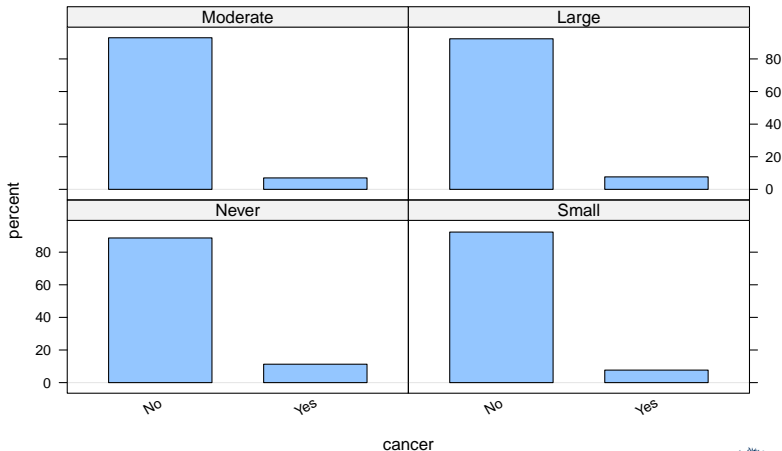
- ▶ Notera att **varje rad** summerar till 100 procent. Så ska det vara eftersom varje rad representerar en av de kategorier som vi betingar fördelningen på. Varje rad kan alltså ses som en egen självständig fördelning.
- ▶ Nu kan vi se skillnader mellan grupperna. Bland dem som aldrig eller sällan åt fisk hade drygt 11 procent diagnosticerats med prostatacancer. Bland övriga grupper är motsvarande siffra lite över 7 procent.

diet	cancer		
	No	Yes	Total
Never	88.709677	11.290323	100.000000
Small	92.331171	7.668829	100.000000
Moderate	92.981867	7.018133	100.000000
Large	92.349727	7.650273	100.000000

Kategoriska variabler och samband

Vi kan ge en snabbare överblick av samma information med en graf.

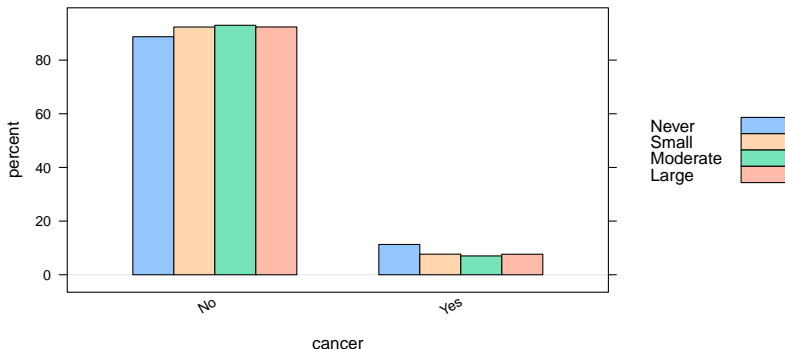
```
bargraph(~cancer|diet, data=fish, type="percent")
```



Kategoriska variabler och samband

Vi kan också gruppera efter variabeln cancer. Färgerna indikerar diet. Två staplar av samma färg är tillsammans 100 procent. Vi ser att cancer förekom något mer bland de deltagare som aldrig eller sällan åt fisk.

```
bargraph(~cancer, groups=diet, data=fish, type="percent")
```



Kategoriska variabler och samband

- ▶ Givet en viss diet kan vi nu säga hur stor andel som har diagnosticerats med cancer.
- ▶ Betyder det att vi även kan säga hur stor andel av dem som diagnosticerat med cancer som har en viss diet? **Nej, inte utan att räkna ut nya betingade värden.**
- ▶ Den här gången vill vi veta hur ofta personer äter fisk
 - ▶ betingat på att en person har diagnosticerats med cancer.
 - ▶ betingat på att en person inte har diagnosticerats med cancer.

Kategoriska variabler och samband

Vi tar den här gången fram en tabell som är betingad på cancer-variabeln.

```
tally(~ diet | cancer, data=fish, format="percent", margins=T)
```

diet	cancer	
	No	Yes
Never	1.894592	3.004292
Small	41.681020	43.133047
Moderate	47.692043	44.849785
Large	8.732346	9.012876
Total	100.000000	100.000000

Räkneexempel: Säg att vi vill räkna ut andelen som aldrig/sällan åt fisk betingat på att de fick cancer. Vi *ignorerar* då alla deltagare som *inte* fick cancer. Antalet deltagare som fick prostatacancer var 466. Antalet av dessa som aldrig/sällan åt fisk var 14. Den procentuella andelen av de som fick cancer som aldrig/sällan åt fisk var alltså $(14/466) * 100\% = 3.004292\%$.

Kategoriska variabler och samband

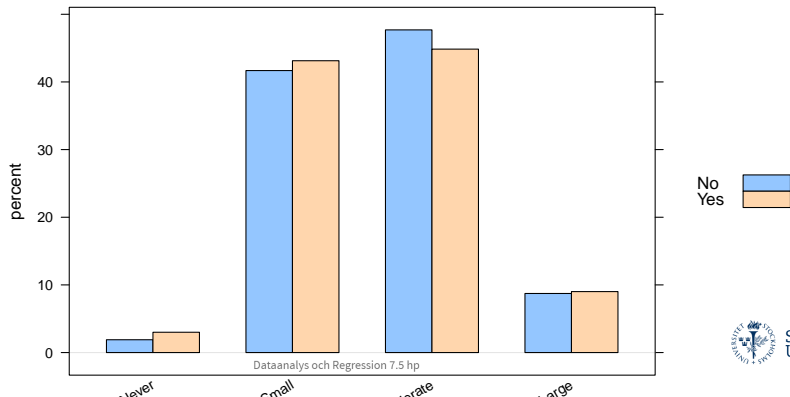
- ▶ Notera att det nu är **kolumnerna** som summerar till 100. Det beror på att vi nu har gjort en separat frekvenstabell för varje kolumn.
- ▶ Något som framgår här, och som vi inte såg när vi betingade fördelningen på diet, är att det är få av deltagarna som aldrig eller sällan äter fisk.

	cancer	
diet	No	Yes
Never	1.894592	3.004292
Small	41.681020	43.133047
Moderate	47.692043	44.849785
Large	8.732346	9.012876
Total	100.000000	100.000000

Kategoriska variabler och samband

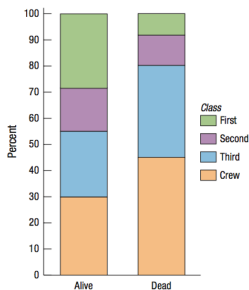
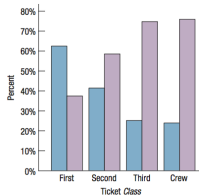
- ▶ Vi gör ett stapeldiagram grupperat på cancer-variabeln.
- ▶ De blå staplarna summerar till 100 procent, och staplarna i beige summerar också till 100 procent.
- ▶ Vi ser att bland de som aldrig eller sällan äter fisk är cancerdiagnoserna något överrepresenterade.

```
bargraph(~diet, groups=cancer, data=fish, type="percent")
```



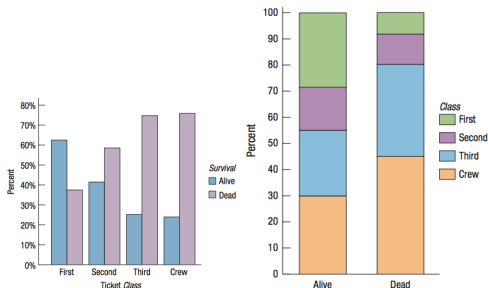
Kategoriska variabler och samband - fler exempel

- ▶ Stapeldiagrammet till vänster är betingat på variabeln *Class*. Det hjälper oss att besvara frågan om hur stor andel som överlevde inom varje biljettklass.
- ▶ Stapeldiagrammet till höger är betingat på variabeln *Survived*. Det hjälper oss att besvara frågan om hur stor andel av dem som överlevde, respektive av dem som inte överlevde, som reste i en viss klass.



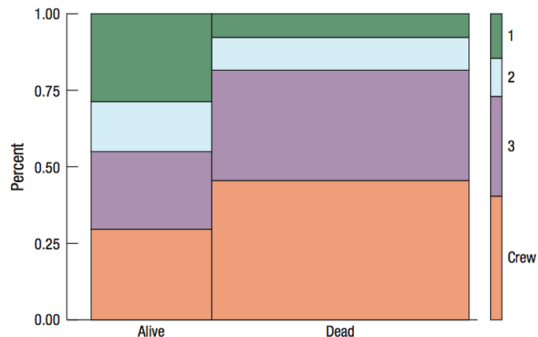
Kategoriska variabler och samband - fler exempel

- ▶ Stapeldiagrammet till vänster ger en bild av andelen som överlevde katastrofen, men ingen information om hur många som ingick i varje klass.
- ▶ Stapeldiagrammet till höger ger en bild av hur många som reste i respektive klass, men ingen information om andelen som överlevde.



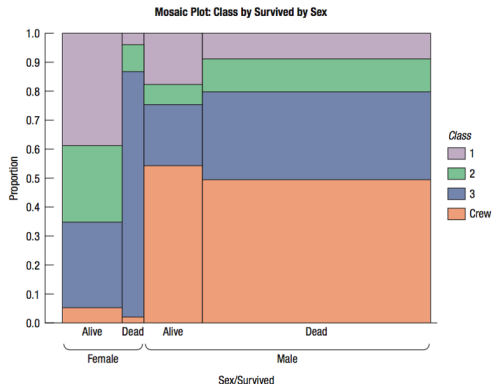
Kategoriska variabler och samband - mosaic plot

En **mosaic plot** ger en mer komplett bild av en simultan fördelning än ett vanligt stapeldiagram. Varje ruta har en **area** som motsvarar andelen observationer som rutan representerar.



Kategoriska variabler och samband - mosaic plot

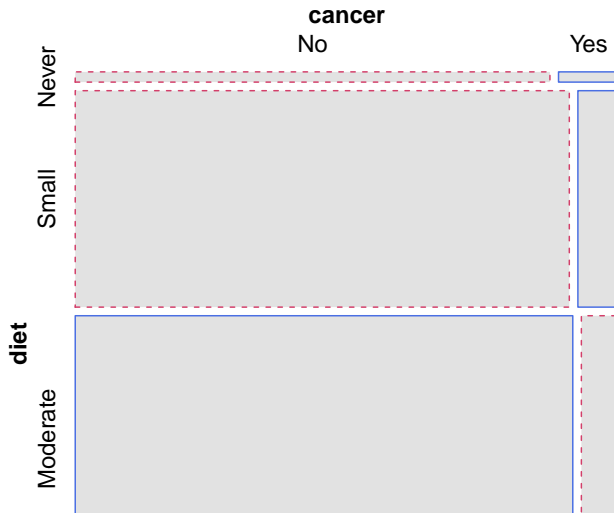
En mosaic plot kan dessutom visa fler än två variabler i samma bild. Den här grafen visar variablerna *Class*, *Survived* och *Gender*.



Kategoriska variabler och samband - mosaic plot

Mosaic plot över deltagarna i studien om fisk och prostatacancer.

```
mosaic(~diet + cancer, data=fish, shade=TRUE,  
       gp=shading_Friendly2, legend=FALSE) #Kräver paketet vcd
```



Kategoriska variabler och samband - fisk och cancer

- ▶ Det är tydligt att deltagare i studien som diagnosticerades med cancer var överrepresenterade bland dem som aldrig åt fisk.
- ▶ Betyder det att vi har hittat ett samband? Ja och nej.
 - ▶ Ja, **bland deltagarna** i studien finns ett samband.
 - ▶ Nej, vi kan inte utan vidare säga att det finns ett samband som gäller för **hela befolkningen**.
- ▶ Det finns bara 14 deltagare i studien som fick prostatacancer och som aldrig/sällan äter fisk. Det är ett litet underlag om vi vill dra slutsatser som gäller hela befolkningen.
- ▶ Det kan vara slumpen som gör att vi ser ut att ha ett samband mellan två variabler.

Kategoriska variabler och samband - mardrömmar

- ▶ När vi försöker avgöra om ett samband mellan två variabler i vårt datamaterial beror på slumpen eller om det beror på att de faktiskt finns ett mer generellt samband, då är vi inne på ett statistikområde som kallas **inferens**.
- ▶ Formella metoder för inferens ingår i del 2 av kursen, men redan nu kan vi ge en mer allmän bild av vad det är.

Kategoriska variabler och samband - mardrömmar

- ▶ I kursboken (sid 99-100) beskrivs en studie som handlar om mardrömmar och vilken sida vi sover på. Tabellen nedan visar en korstabell med de 63 deltagarna i studien.
- ▶ Anta att vi vill bedöma om det finns ett samband mellan de två variablerna. Det är svårt att se i vår korstabell.

		Side		
		Right	Left	Totals
Dreams	Nightmares	6	9	15
	Sweet Dreams	35	13	48
	Totals	41	22	63

Kategoriska variabler och samband - mardrömmar

- ▶ För att se skillnaden mellan de som sover på höger sida och de som sover på vänster sida skapar vi en tabell som vi betingar på variabeln *side*.
- ▶ Den betingade tabellen besvarar frågorna:
 - ▶ Hur stor andel av dem som sover på höger sida har mardrömmar?
 - ▶ Hur stor andel av dem som sover på vänster sida har mardrömmar?
- ▶ Vi skapar vår betingade tabell med hjälp av funktionen *tally* i R.

```
tally(~dream | side, data=nightmares,  
      format="percent", margins=TRUE)
```

	side	
dream	Right	Left
Nightmares	14.63415	40.90909
Sweet dreams	85.36585	59.09091
Total	100.00000	100.00000

Kategoriska variabler och samband - mardrömmar

- Drygt 40 procent av dem som sover på vänster sida har mardrömmar, men bara knappt 15 procent av dem som sover på höger sida.
- Uppenbarligen finns en stor skillnad mellan de två deltagargrupperna. Är det slumpen eller ett samband som finns i hela befolkningen?

	side	
dream	Right	Left
Nightmares	14.63415	40.90909
Sweet dreams	85.36585	59.09091
Total	100.00000	100.00000

Kategoriska variabler och samband - mardrömmar

- ▶ En vanlig metod för att avgöra om det finns ett verkligt samband mellan två variabler är att ställa upp en **hypotes** om att det **inte** finns något samband mellan variablerna.
- ▶ I det här fallet skulle hypotesen vara att det saknas samband mellan mardrömmar och vilken sida vi sover på.
- ▶ Om hypotesen stämmer borde mardrömmarna vara fördelade slumpvis, utan hänsyn till vem som sover på vänster sida.

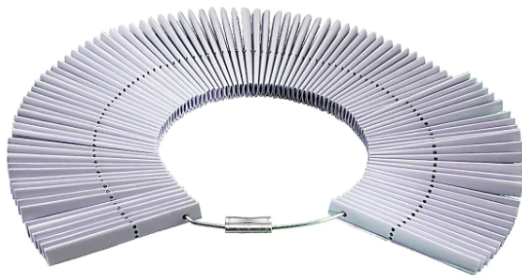
Kategoriska variabler och samband - mardrömmar

- ▶ Anta att vi, som i studien, har 41 personer som sover på höger sida och 22 som sover på vänster sida. 15 deltagare har mardrömmar och 48 deltagare sover gott.
- ▶ Om vi fördelar de 15 mardrömmarna slumpvis skulle vi exempelvis kunna få utfallet som visas i korstabellen till höger.

		Side		
		Right	Left	Totals
Dreams	Nightmares	6	9	15
	Sweet Dreams	35	13	48
	Totals	41	22	63

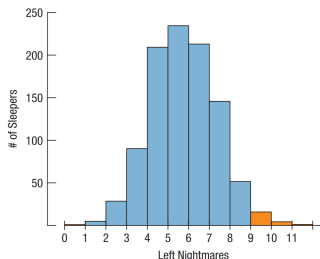
		Side		
		Right	Left	Totals
Dreams	Nightmares	11	4	15
	Sweet Dreams	30	18	48
	Totals	41	22	63

Kategoriska variabler och samband - mardrömmar



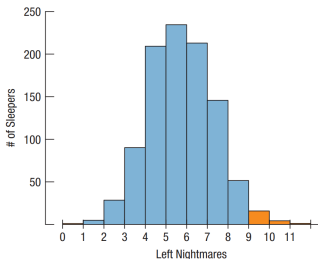
Kategoriska variabler och samband - mardrömmar

- ▶ Vi upprepar vår procedur att fördela mardrömmarna slumpmässigt 1000 gånger! För varje gång noterar vi hur många av de 22 som sover på vänster sida som fick en mardrömslott.
- ▶ När proceduren upprepats 1000 gånger har vi en lista med 1000 värden.
- ▶ Vi kan illustrera fördelningen i vår lista som ett histogram.



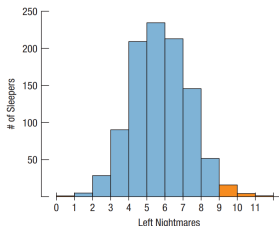
Kategoriska variabler och samband - mardrömmar

- ▶ Varje stapel i histogrammet visar hur många gånger proceduren resulterade i att gruppen som sover på vänster sida fick ett visst antal mardrömmar.
- ▶ Den högsta stapeln visar exempelvis att proceduren ungefär 250 gånger resulterade i att 5 personer som sover på vänster sida fick en mardrömslott.



Kategoriska variabler och samband - mardrömmar

- ▶ De **gula** staplarna till höger visar hur många gånger den slumpmässiga proceduren resulterade i att 9 eller fler av dem som sover på vänster sida fick mardrömmar.
- ▶ Om hypotesen är sann, dvs om mardrömmar är oberoende av vilken sida vi sover på, så var resultatet i forskningsstudien ett sällsynt utfall. Det talar emot att hypotesen skulle vara riktig.



Kategoriska variabler och samband - kausalitet

- ▶ Anta att vi har konstaterat att det finns ett samband mellan att sova på vänster sida och att drömma mardrömmar, eller att det finns ett samband mellan att äta fisk och att drabbas av prostatacancer. Betyder det att sovande på vänster sida **orsakar** mardrömmar och att fisk i kosten minskar risken för cancer?
- ▶ **Nej, ett samband är inte samma sak som kausalitet!**
- ▶ Det kan finnas **andra orsaker** till sambandet än att den ena variabeln påverkar den andra. Kanske är det så att människor som är hälsomedvetna oftare äter fisk, samtidigt som de också äter mer grönsaker och ägnar sig mer åt fysisk aktivitet.

Kategoriska variabler och samband - kausalitet

Att fundera över

- ▶ Om det finns ett samband mellan **fler tv-apparater** i ett land och **högre medellivslängd**, beror det på att tv-tittande är nyttigt eller finns det någon annan bakomliggande variabel som spelar in?
- ▶ När du tittar på ett dataset, var medveten om att det alltid finns **dolda variabler (lurking variables)**, det vill säga variabler som inte är inkluderade i datamaterialet.

Kategoriska variabler och samband - Simpson's paradox

Simpson's paradox innebär att ett samband mellan två variabler kan försvinna när datamaterialet **delas in i olika grupper**. På sid 106 i kursboken hittar vi en korstabell som tycks peka på att män hade lättare än kvinnor att bli antagna som doktorand på UC Berkeley.

	Admit	Reject	%Admit
Men	1158	1493	43.7%
Women	557	1278	30.4%

Över 40 procent av männen som söktes blev antagna men bara 30 procent av kvinnorna. Det ser ut som att kvinnor blev negativt särbehandlade.

Kategoriska variabler och samband - Simpson's paradox

När vi ser samma siffror nedbrutna per fakultet (school) blir bilden en annan.

School	Male Admits	Female Admits	Male%	Female%
A	512	89	62.1%	82.4%
B	313	17	60.2%	68.0%
C	120	202	36.9%	34.1%
D	138	131	33.1%	34.9%
E	53	94	27.7%	23.9%
F	22	24	5.9%	7.0%

På fyra av de sex fakulteterna var det en större andel av de sökande kvinnorna än av de sökande männen som kom in. Hur går det ihop?

Kategoriska variabler och samband - Simpson's paradox

School	Male Admits	Female Admits	Male%	Female%
A	512	89	62.1%	82.4%
B	313	17	60.2%	68.0%
C	120	202	36.9%	34.1%
D	138	131	33.1%	34.9%
E	53	94	27.7%	23.9%
F	22	24	5.9%	7.0%

- ▶ Fler män sökte till school A och B, där det var lättare att komma in.
- ▶ Fler kvinnor sökte till school E och F, där det var svårare att komma in.

Kvinnor hade lägre antagningsgrad på grund av att de sökte program som var svårare att komma in på.