PSY2014

Forelesning 6

Kategorisk data

- **Kategorisk data:** Data der verdiene utgjør et bestemt antall distinkte kategorier.
  - Nominelle variabler: Ingen rangering av kategoriene (f.eks. bilmerker).
  - Ordinale variabler: Distinkte kategorier, men de har en rangering (f.eks. vektklasser i boksing).

- Regresjon er egnet når den *avhengige* variabelen er kvantitativ.
  - Vi så på forelesning 5 hvordan vi kan inkludere kategoriske uavhengige variabler gjennom rekoding til dummy-variabler.

- Hva kan du gjøre dersom du er interessert i å se på sammenhengen mellom to kategoriske variabler?
  - Kjøpes det ulike bilmerker i Oslo, Bergen og Trondheim?

I dag

- En introduksjon til grunnleggende sannsynlighetsregning
  - (Agresti 4.1).
  - Hva er utfallsrom, sannsynlighet, uavhengighet?
  - Fundamentale regler for sannsynlighetsregning (additiv, multiplikativ).

- Kji-kvadrat statistikken og dens bruk i analyse av hvorvidt vi kan konkludere at to variabler er *uavhengige*.
  - (Agresti 8.1-8.3)

Grunnleggende sannsynlighetsregning/regler

Agresti 4.1

Stokastisk (Random)

- Et fenomen er **stokastisk** dersom individuelle utfall ikke er gitt, men dersom utfall følger en bestemt fordeling over gjentatte repetisjoner.

  F.eks. Jeg vet ikke utfallet av et enkelt kast av en rettferdig mynt, men jeg vet at det vil være 50/50 kron og mynt på tvers av mange kast.

- **Utfallsrom:** Settet av alle mulige utfall av et stokastisk fenomen.
  - Hva er utfallsrommet tll ett myntkast?
    - S = {Kron, Mynt}
  - Hva er utfallsrommet tll ett terningkast?
    - S={ 1, 2, 3, 4, 5, 6 }
  - Hva er utfallsrommet tll *tre* myntkast?
    - S = { KKK, KKM, KMK, KMM, MKK, MKM, MMK, MMM}

- En hendelse (*event*) er et bestemt subsett av utfallsrommet.
  - I statistisk litteratur refereres disse ofte til med en bokstav.

Husk utfallsrommet til tre myntkast

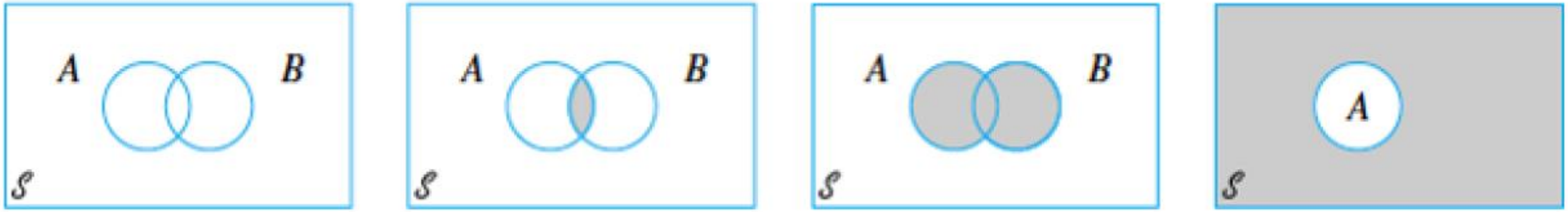
- S = { KKK, KKM, KMK, KMM, MKK, MKM, MMK, MMM }

- Hvilket subset av S utgjør hendelsen *A*, at alle kastene er mynt?
  - S = {KKK, KKM, KMK, KMM, MKK, MKM, MMK, **MMM**}
  - A = {**MMM**}

- Hvilket subset av S utgjør hendelsen *B*, at du får nøyaktig to kron på tre kast?
  - S = {KKK, **KKM**, **KMK**, KMM, **MKK**, MKM, MMK, MMM}
  - B = {**KKM, KMK, MKK**}

Venn diagram

- Et venn-diagram viser alle mulige logiske relasjoner mellom samlinger av forskjellige sett.
- Brukes ofte som en uformell representasjon, og eksakte størrelser/former av regioner er ikke viktige.



(a) Venn diagram of events A and B (b) Shaded region is $A \cap B$ (c) Shaded region is $A \cup B$ (d) Shaded region is A'

(Devore, 2012)

Sannsynlighet

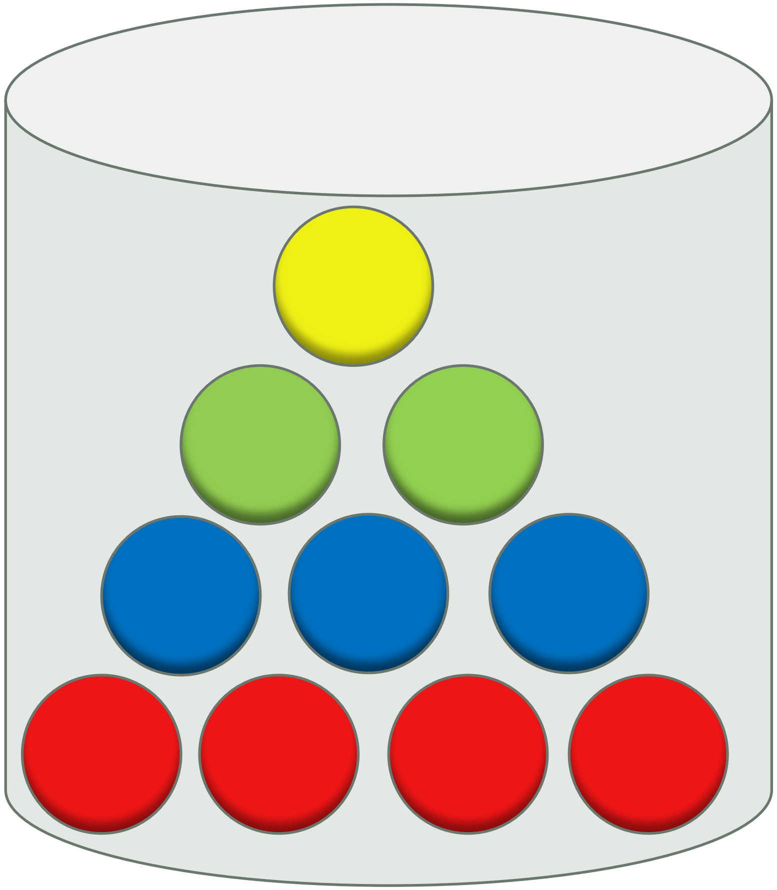
- Dersom alle utkommene av et stokastisk fenomen forekommer like ofte, definerer vi *sannsynligheten* for utfallet A som:

$$P(A) = \frac{Antall\ elementer\ i\ A}{Antall\ elementer\ i\ S}$$

- S = { KKK, KKM, KMK, KMM, MKK, MKM, MMK, MMM}
  - Hva er sannsynligheten for å få tre mynt ved kast på tre kast av en rettferdig mynt?
    - $\frac{1}{8}$
  - Hva er sannsynligheten for å få nøyaktig to kron på tre kast?
    - $\frac{3}{8}$

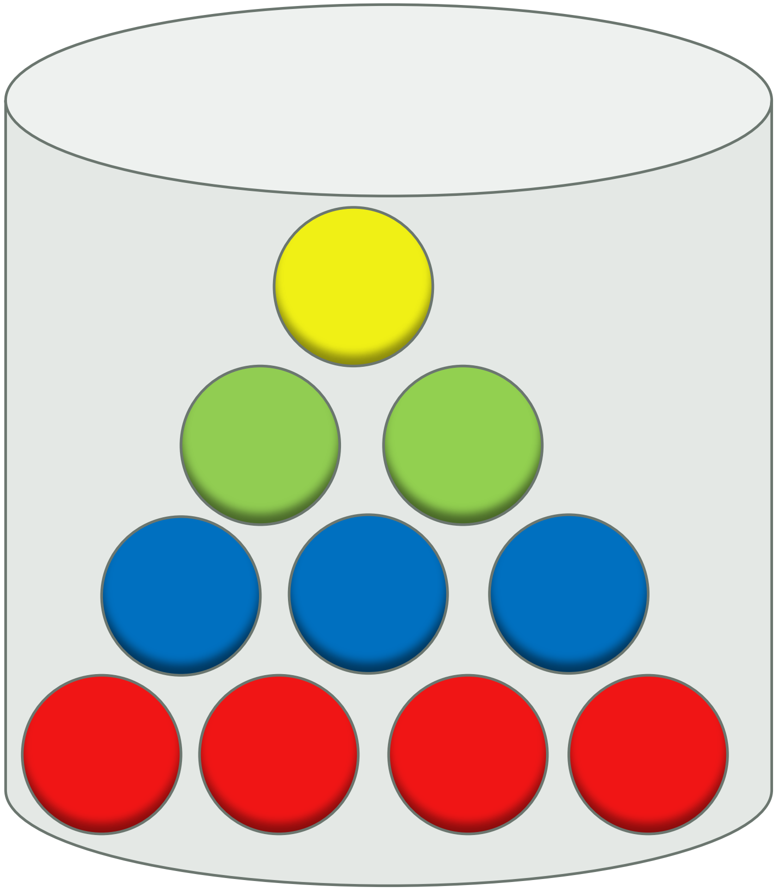
A er utfallet at ballen er **grønn**. Hva er sannsynligheten for A?

1. P(A) = 0.0
2. P(A) = 0.1
3. P(A) = 0.2
4. P(A) = 0.3
5. P(A) = 0.4



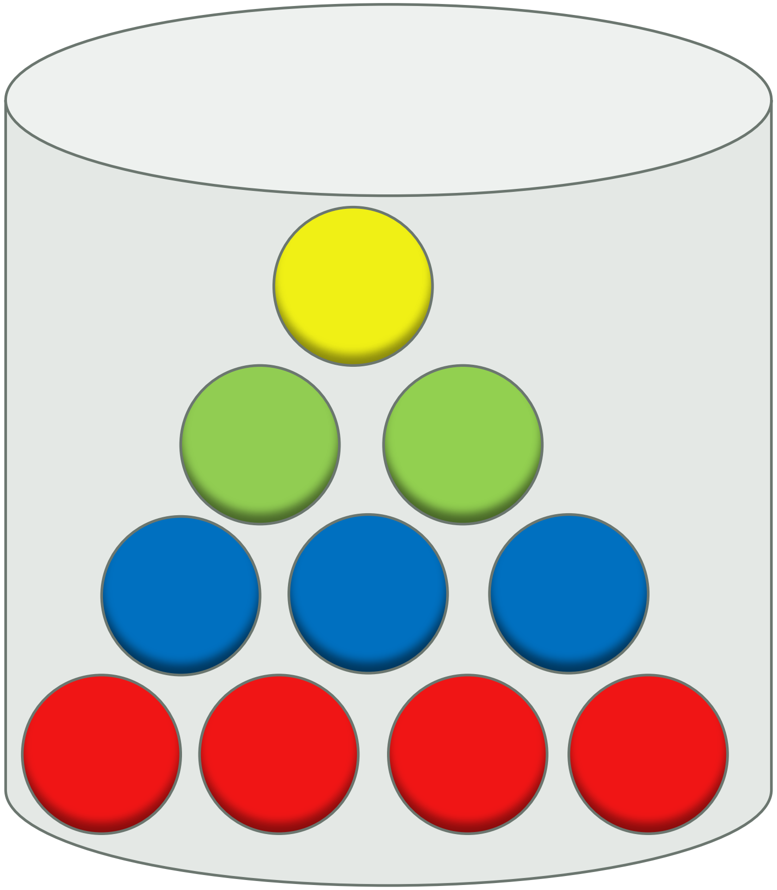
B er utfallet at ballen er **rød**. Hva er sannsynligheten for A?

1. P(B) = 0.0
2. P(B) = 0.1
3. P(B) = 0.2
4. P(B) = 0.3
5. P(B) = 0.4



A er utfallet at ballen er utfallet at ballen er **rød.** B er utfallet at ballen er **grønn.** Hva er P(A eller B)?

1. P(A eller B) = 0.0
2. P(A eller B) = 0.2
3. P(A eller B) = 0.4
4. P(A eller B) = 0.6
5. P(A eller B) = 0.8



Additiv regel (Regel 2, A 4.2)

- To utfall er *gjensidig utelukkende (mutually exclusive)* dersom det at ett skjer utelukker det andre.

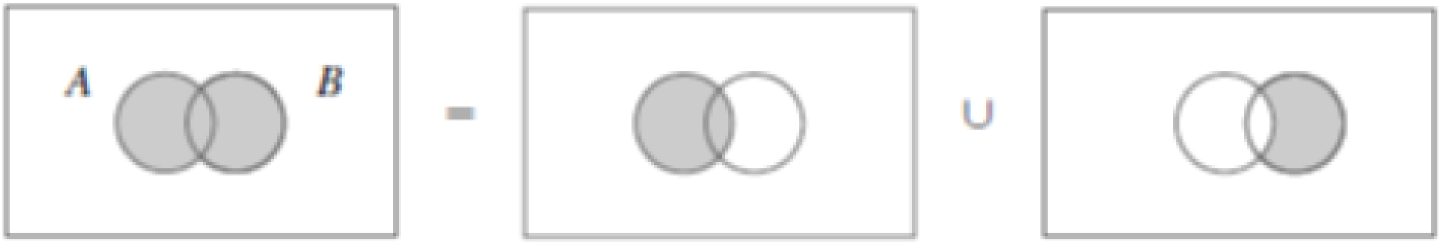
(e) Mutually exclusive events

- S = {KKK, KKM, KMK, KMM, MKK, MKM, MMK, MMM}
- Er disse utfallene gjensidig utelukkende?
  - A: Få bare kron, B: Får bare mynt
    - {KKK, KKM, KMK, KMM, MKK, MKM, MMK, MMM}
  - A: Få nøyaktig to kron, B: Få nøyaktig to mynt
    - {KKK, KKM, KMK, KMM, MKK, MKM, MMK, MMM}

Additiv regel for sannsynlighet: Dersom A og B er gjensidig utelukkende er sannsynligheten for enten A *eller* B:

$$P(A\ eller\ B) = P(A) + P(B).$$

Additiv regel (2)



- Hvis A og B er to hendelser, så er sannsynligheten.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- For gjensidig utelukkende sett forenkles dette til

$$P(A \cup B) = P(A) + P(B)$$

Simultan sannsynlighet (Joint probability)

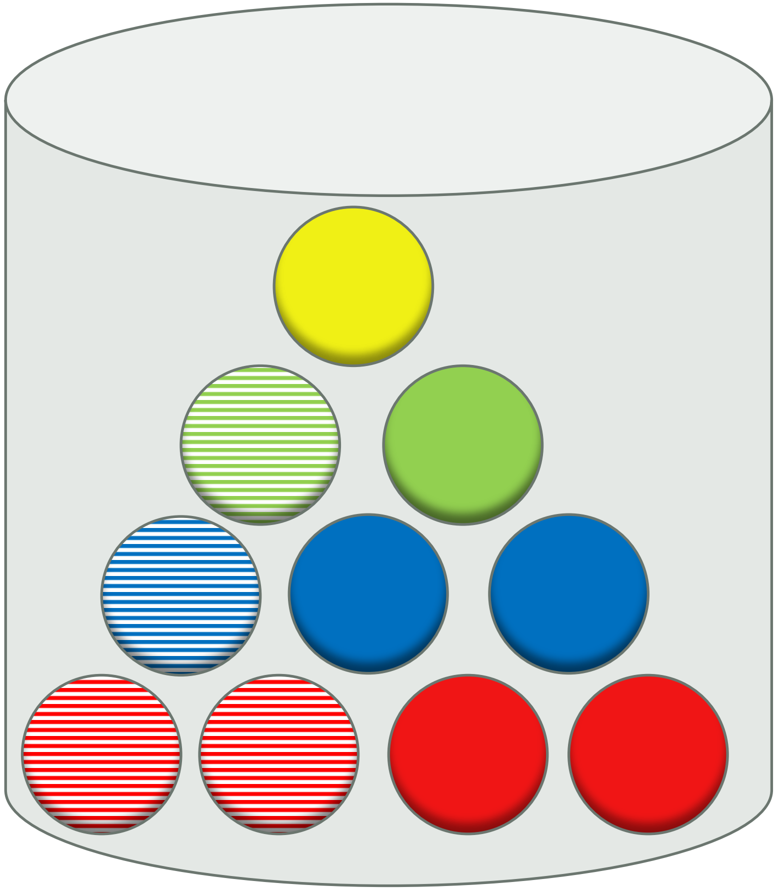
- **Simultan sannsynlighet:** Sannsynligheten for at flere utfall inntreffer.
  - Gitt to utfall, A og B, den simultane sannsynligheten skrives ofte som $P(A, B), eller\ P(A\ og\ B)$.

A er utfallet at ballen er **rød**.

B utfallet at ballen er **stripet.**

Hva er P(A og B)?

1. P(A og B) = 0.00
2. P(A og B) = 0.10
3. P(A og B) = 0.20
4. P(A og B) = 0.40
5. P(A og B) = 0.80
6. Umulig å si

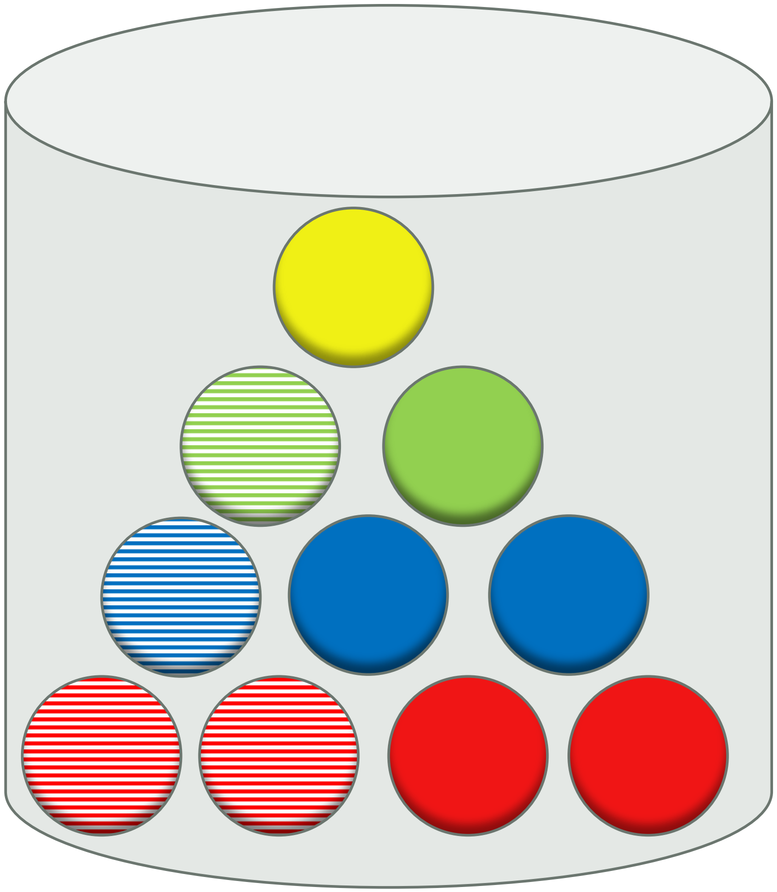


A er utfallet at ballen er **gul**.

B utfallet at ballen er **stripet.**

Hva er P(A og B)?

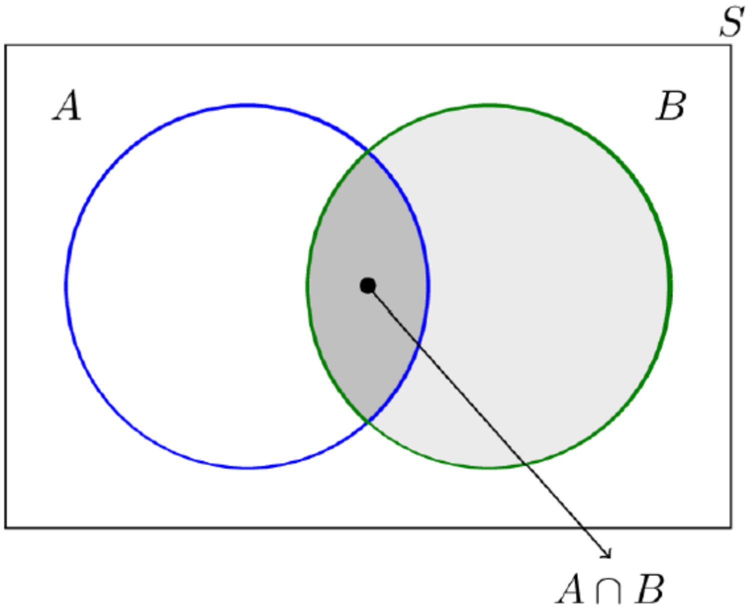
1. P(A og B) = 0.00
2. P(A og B) = 0.05
3. P(A og B) = 0.08
4. P(A og B) = 0.15
5. P(A og B) = 0.24
6. Umulig å si



Betinget sannsynlighet

- For enhver to utfall A og B, der P(B) ikke er lik 0, er den *betingede* sannsynligheten av *A gitt B* definert som:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A|B) \cdot P(B)$$

Uavhengighet

- To utfall er *uavhengige* dersom

$$P(A|B) = P(A)$$

Multiplikasjonsregelen (Regel 4, A 4.2)

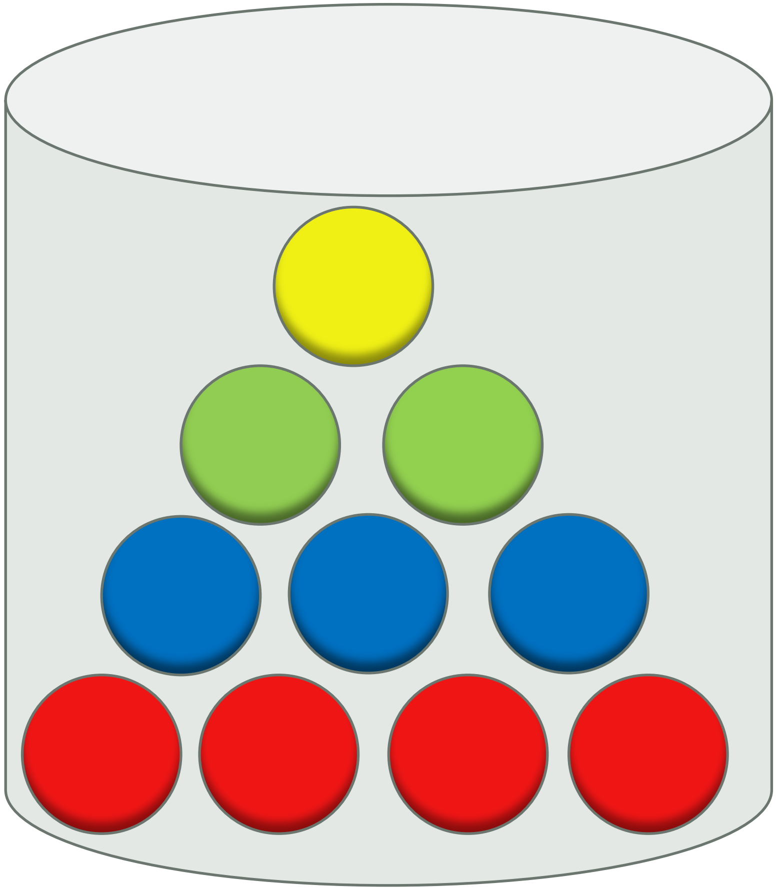
$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A|B) \cdot P(B)$$

Multiplikasjonsregelen: Dersom A og B er *uavhengige* utfall, så er den simultane sannsynligheten (sannsynligheten for begge) gitt ved

$$P(A\ og\ B) = P(A) \times P(B).$$

Velg en ball, legg den tilbake, og velg en til. A utfallet at den *første ballen er rød*. B er utfallet at den *andre ballen er grønn*. Hva er P(A og B)?

1. P(A og B) = 0.00
2. P(A og B) = 0.05
3. P(A og B) = 0.08
4. P(A og B) = 0.15
5. P(A og B) = 0.24



Marginal sannsynlighet

Table 5.2 The relationship between parenthood and seat belt use

Parenthood	Wear Seat belt	Do Not Wear Seat belt	Total
Children			20
No children			80
Total	30	70	100

Hva er sannsynligheten for å bruke setebelte? $\frac{30}{100} = 0.3$

Hva er sannsynligheten for å være en forelder? $\frac{20}{100} = 0.2$

Rad og kolonne sannsynlighetene kalles ofte de *marginale sannsynlighetene.*

Simultan sannsynlighet

Table 5.2 The relationship between parenthood and seat belt use

Parenthood	Wear Seat belt	Do Not Wear Seat belt	Total
Children	15	5	20
No children	15	65	80
Total	30	70	100

- Hva er den simultane sannsynligheten for hendelsene «Bruke setebelte» og «ha barn i bilen»?
- $\frac{15}{100} = 0.15$

- Hva er den simultane sannsynligheten for hendelsene «Bruke setebelte» og «ikke ha barn i bilen»?
  - $\frac{15}{100} = 0.15$

Table 5.2 The relationship between parenthood and seat belt use

Parenthood	Wear Seat belt		Do Not Wear Seat belt		Total
Children		6		14	20
No children		24		56	80
Total	30		70		100

Svart: Den observerte frekvensen

Rød: Den forventede frekvensen dersom variablene var uavhengige..

La oss spørre: «Hva ville forventede cellefrekvenser være dersom bruk av setebelte og å ha barn i bilen var uavhengige variabler?

If they were independent, we know from the multiplicative law of probability:

Sannsynligheten av å bruke belte OG ha barn i bilen? $0.3 \times 0.2 = 0.06$

Sannsynligheten av ikke å bruke belte OG ha barn i bilen? $0.7 \times 0.2 = 0.14$

Sannsynligheten av ikke å bruke belte OG ikke ha barn i bilen? $0.3 \times 0.8 = 0.24$

Sannsynligheten av å bruke belte OG ikke ha barn i bilen? $0.7 \times 0.8 = 0.56$

Hva har vi lært om setebeltebruk?

- De observerte frekvensene for bruk av setebelte/ha barn i bilen later til å være forskjellig fra de vi ville forvente dersom variablene var uavhengige.
  - Dette antyder at det er en sammenheng mellom det å bruke setebelte, og det å ha barn i bilen.
- Men, stopper du 100 biler til vil tallene være noe forskjellig. Kanskje den observerte sammenhengen kanskje bare er en følge av sampling?
- Vi ønsker en måte å teste de følgende hypotesene:
  - H_0: Variablene er statistisk uavhengige
  - H_a: Variablene er ikke statistisk uavhengige

Kji-kvadrat tester

Agresti 8.1-8.3

Vi vil se på to andendelser

1. Kji-kvadrat *goodness of fit* test
   - Teste i hvilken grad verdiene på en observert variabel er i overenstemmelse med en bestemt fordeling.

2. Kji-kvadrat test av uavhengighet
   - Test på i hvilken grad to variabler er (statistisk) uavhengige.

- Skillet viser ulik *bruk* av kji-kvadrat statistikken. Beregningen er den samme i begge tilfeller.
- Testen på uavhengighet vil være den viktigste for oss.

Kji-kvadrat statistikken.

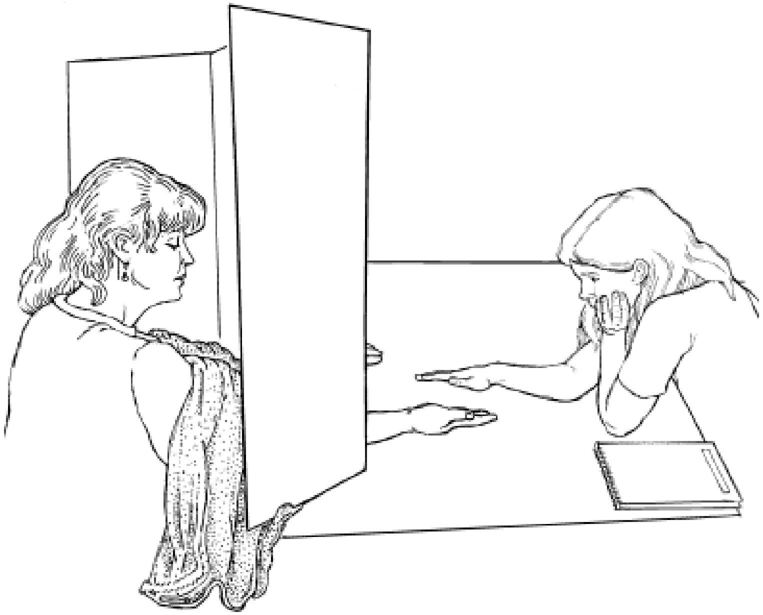
$$\sum_{over\ alle\ cellene} \frac{(observert\ frekvens - forventet\ frekvens)^2}{forventet\ frekvens}$$

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- Legg merke til:
  - Denne summen blir liten når de observerte frekvensene har verdier som er lik de forventede.
  - Vi deler på forventet antall, fordi en forskjell mellom observert og forventet antall må sees i forhold til hvor mange man forventer.
    - Det er mindre *relativ forskjell* på 1000og 1001, enn på 10 og 11, selv om forskjellen i begge tilfeller er «1».

Eksempel: Terapeutisk touch

- Holistisk sykepleie er en spesialisering I USA.
  - Av ulike behandlinger som tilbys er "terapeutis touch" (TT).
- De som praktiserer TT hevder de kan føle det "menneskelige energifeltet".
- Emily Rosa er den yngste personen til å få en artikkel publisert I JAMA (journal of the American Medical Association).
  - http://www.youtube.com/watch?feature=player_embedded&v=mNoRxCRJ-Y0#!



1. Goodness-of-fit test

	Correct	Incorrect	Total
Observed	123	157	280

Goodness of fit testen lar oss sammenlikne de observerte frekvensene med de vi ville forventet dersom de var trukket fra en bestemt fordeling.

Her ønsker jeg å sammenlikne de observerte frekvensene med de vi ville forventet om deltagere bare gjettet.

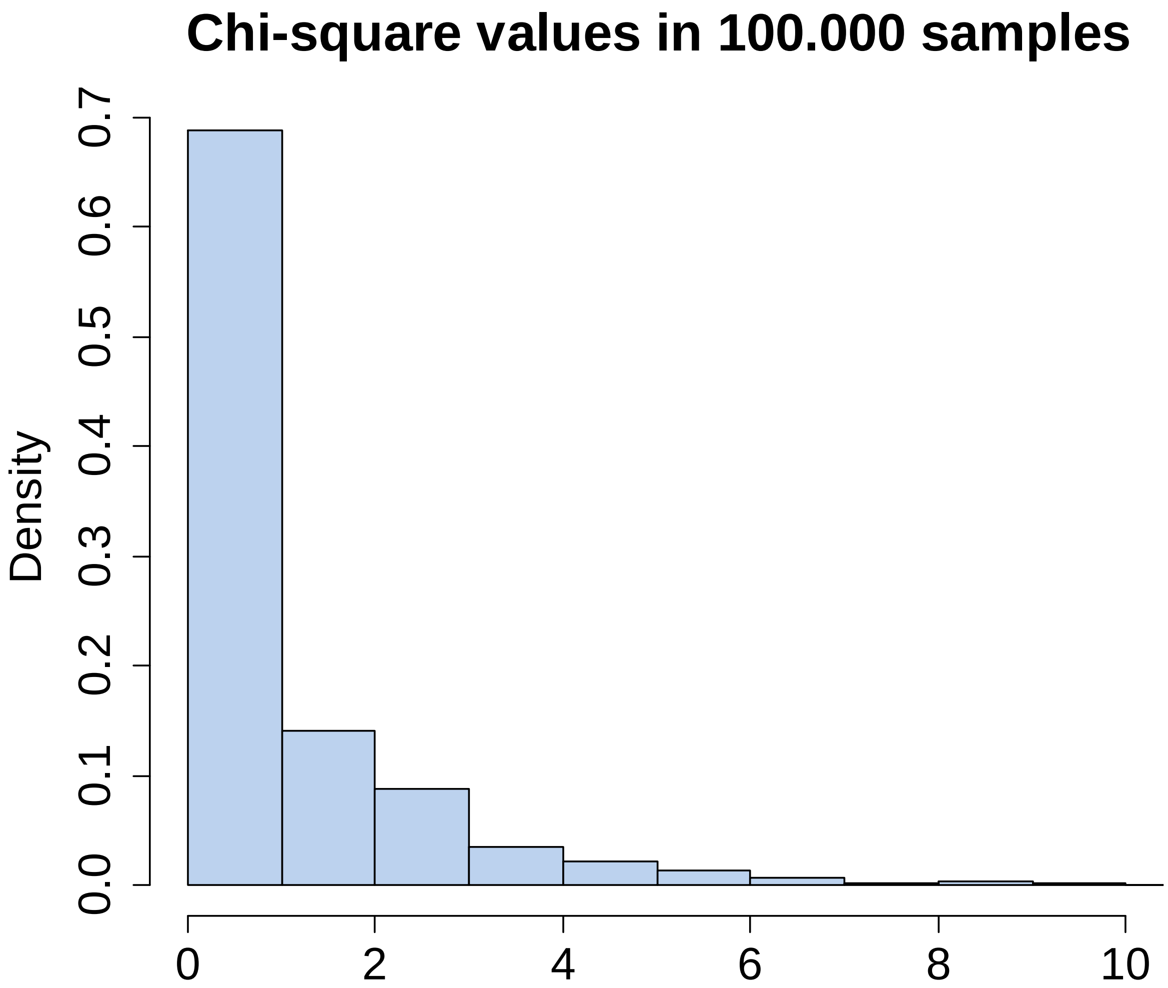
H_0: The observerte dataene data skyldes ren gjetting.
H_1: The observed dataene skyldes *ikke* ren gjetting.

I et sett av 280 forsøk, hvor mange korrekte og gale responser ville du forventet dersom det var helt tilfeldig dersom folk bare gjettet?

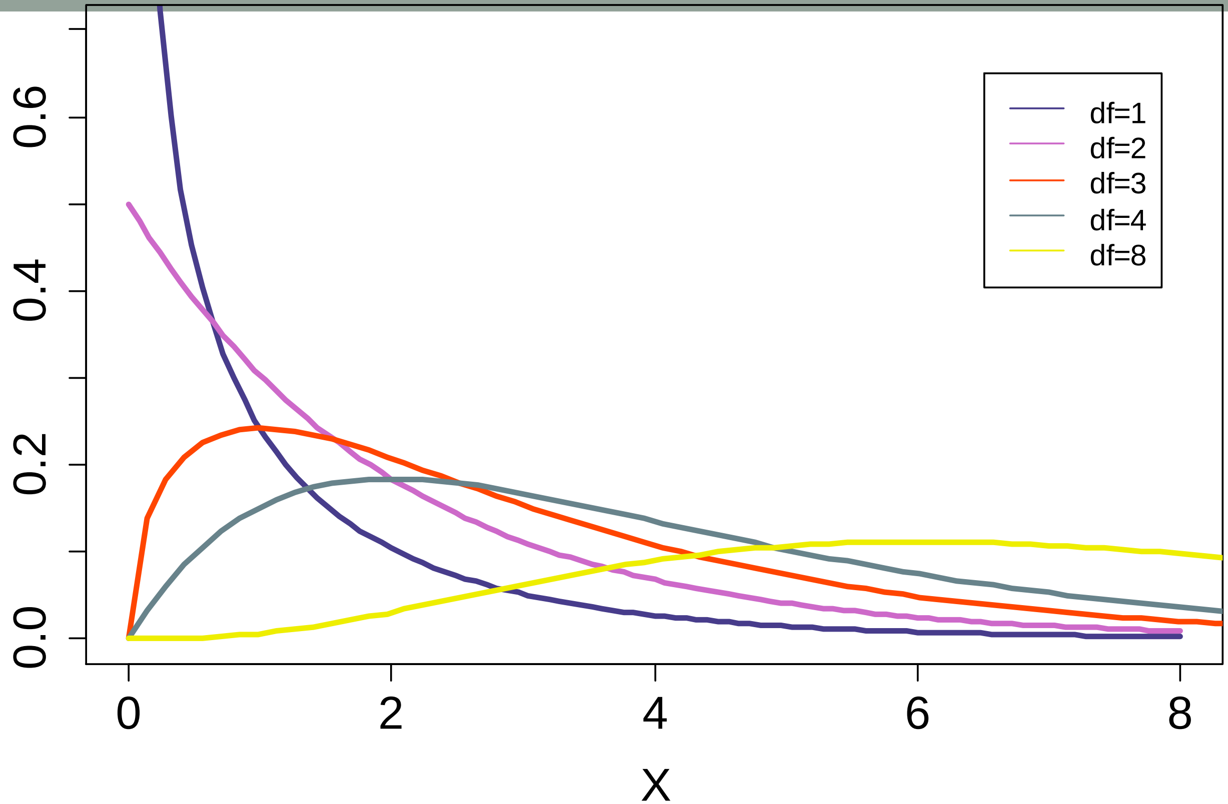
$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(123-140)^2}{140} += \frac{(157-140)^2}{140} = 4.129$$

Er 4.129 mye? Igjen må vi vite hva samplingfordelingen er under H_0.

- Dersom du samlet 280 observasjoner, men deltagere bare *gjettet,* hva slags verdier ville vi se på kji-kvadrat statistikkene på tvers av ulike utvalg?
- Utvalg 1:(Korrekt: 129, Feil: 151) $= \frac{(129-140)^2}{140} + \frac{(151-140)^2}{140} = 1.7285$
- Utvalg 2: (Korrekt :132, Feil : 148) $= \frac{(132-140)^2}{140} + \frac{(146-140)^2}{140} = 0.914$
- Utvalg 3: (Korrekt :134, Feil : 146) $= \frac{(134-140)^2}{140} + \frac{(146-140)^2}{140} = 0.514$
- Etc…



Samplingfordelingen til kji-kvadrat statistikken under H_0

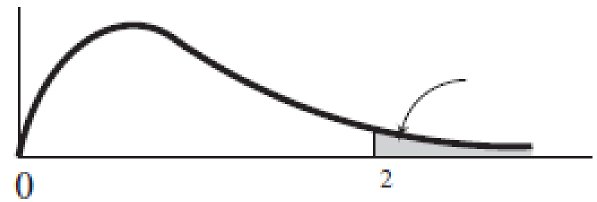


Under H_0 følger den ovserverte verdien for kji-kvadrat statistikken en kji-kvadrat fordeling.

- Formen på en kji-kvadrat fordeling bestemmes av ett tall (antall frihetsgrader).
- Vi må altså vite antall frihetsgrader. I vårt eksempel er df=1. Vi kan da finne den kritiske verdien i en kji-kvadrat tabell.

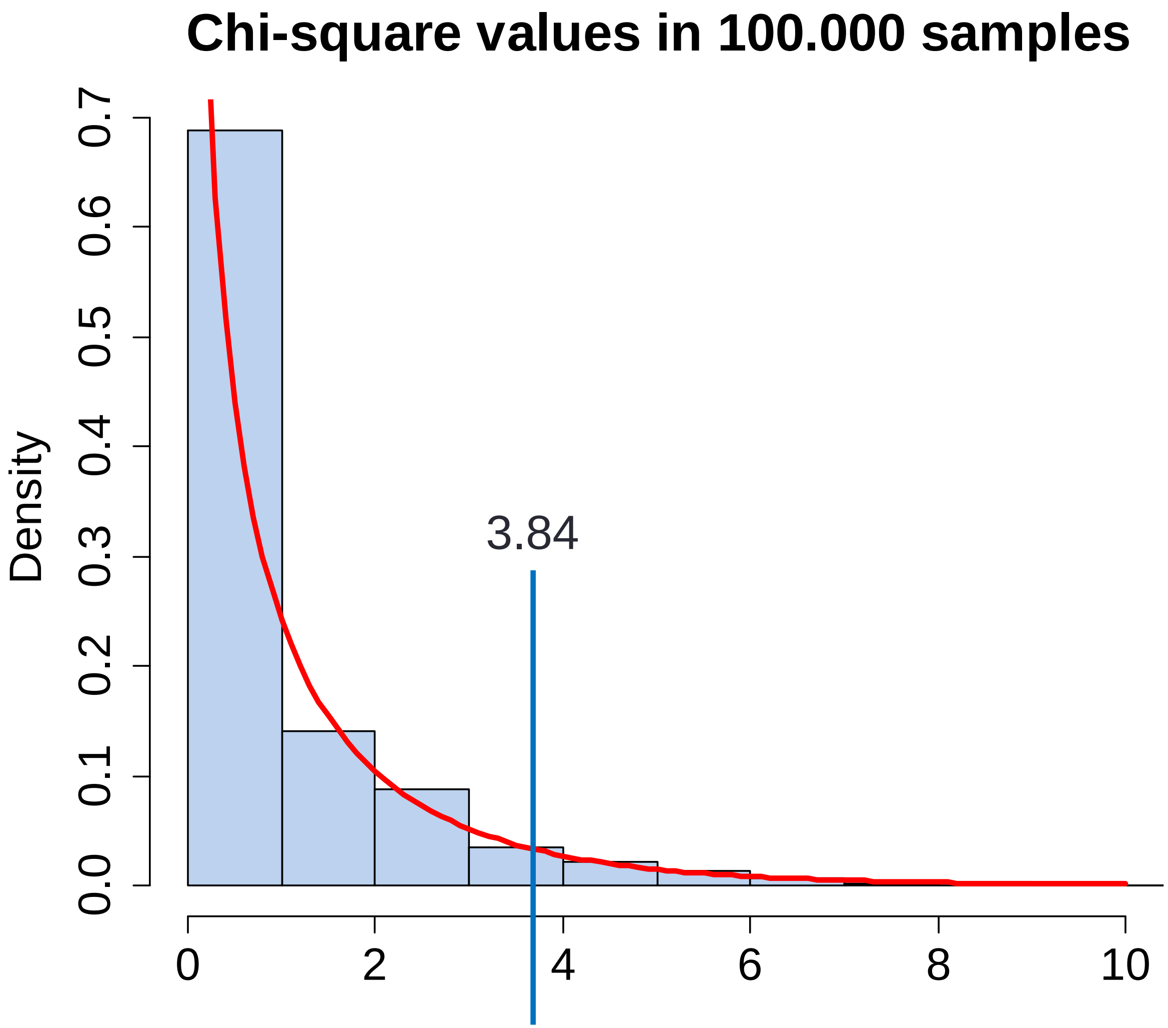
Chi-kvadrat tabell

Appendix χ^2: Upper Percentage Points of the χ^2 Distribution



df	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.82
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92
10	2.15	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03

Den kritiske verdien for en kji-kvadratfordeling med en frihetsgrad er 3.84



The sampling distribution follows a *Chi-square distribution*, and 95% of the area of this distribution lies to the left of 3.84.

2. Chi-square test av uavhengighet

	Party			
Gender	**Democrats**	**Independent**	**Republican**	Total
Women	573	516	422	1511
Men	386	475	399	1260
Total	959	991	821	2771

Gitt en krysstabell som den over er det naturlig å spørre; «Er rader og kollonner uavgengige?» Her, «er kjønn og politisk tilhørighet uavhengige variabler?».

H_0: Kjønn og politisk tilhørighet er (statistisk) uavhengige variabler.
H_a: Kjønn og politisk tilhørighet er (statistisk) avhengige variabler.

	Col_1	Col_2	…	Col_c	Totalt
Row_1	X_{11}	X_{12}	…	X_{1c}	R_1
Row_2	X_{21}	X_{22}	…	X_{2c}	R_2
…	…	…	…	…	…
Row_r	X_{r1}	X_{r2}	…	X_{rc}	R_r
Total	C_1	C_2	…	C_c	n

Hva er sannsynligheten for å være i rad 2? $\frac{R_2}{n}$

Hva er sannsynligheten for å være i kolonne 1? $\frac{C_1}{n}$

Hva er den simultane sannsynligheten for å være *både* i rad 2 og kolonne 1? $\frac{x_{21}}{n}$

Hva *ville* den simultane sannsynligheten for å være både i rad 2 og kolonne 1 vært *om de var uavhengige*? $\frac{R_2}{n} \times \frac{C_1}{n}$

Det jeg har nå er en *sannsynlighet under H_0. For å få forventet frekvens i celle E_{21} må jeg gange med n.* $\frac{R_2 \times C_1}{n}$

	Parti			
Kjønn	**Democrats**	**Independent**	**Republican**	Totalt
Kvinner	573	516	422	1511
Menn	386	475	399	1260
Totalt	959	991	821	2771

Under H_0 hvor mange demokratiske menn forventer vi?

$$\frac{R_2 \times C_1}{n} = \frac{1260 \times 959}{2771} = 436.07$$

Under H_0 hvor mange republikanske kvinner forventer vi?

$$\frac{R_2 \times C_1}{n} = \frac{1511 \times 821}{2771} = 447.68$$

	Parti			
Kjønn	**Democrats**	**Independent**	**Republican**	Totalt
Kvinner	573 (522.93)	516 (540.38)	422 (447.68)	1511
Menn	386 (436.07)	475 (450.62)	399 (373.32)	1260
Totalt	959	991	821	2771

De svarte tallen er de observerte frekvensene, mens de røde er de forventede frekvensene under H_0.

Hva betyr det at de røde og svarte tallene er forskjellige?

	Parti			
Kjønn	**Democrats**	**Independent**	**Republican**	Totalt
Kvinner	573 (522.93)	516 (540.38)	422 (447.68)	1511
Menn	386 (436.07)	475 (450.62)	399 (373.32)	1260
Totalt	959	991	821	2771

$$\sum_{over\ alle\ cellene} \frac{(observert - forventet)^2}{forventet} = \sum_{Over\ alle\ cellene} \frac{(O-E)^2}{E}$$

$$\frac{(573-522.93)^2}{522.93} + \frac{(516-540.38)^2}{540.38} + \frac{(422-447.68)^2}{447.68} + \frac{(386-436.07)^2}{436.07} + \frac{(475-450.62)^2}{450.62} + \frac{(399-373.32)^2}{373.32}$$

16.20

$$\sum_{Over\ alle\ cellene} \frac{(X-E)^2}{E} = 16.2$$

Hvordan tolke dette. Er det mye eller lite?

Hvilke verdier ville vi forvente å se dersom kjønn og politisk tilhørighet (rader og kolonner) virkelig var uavhengige?

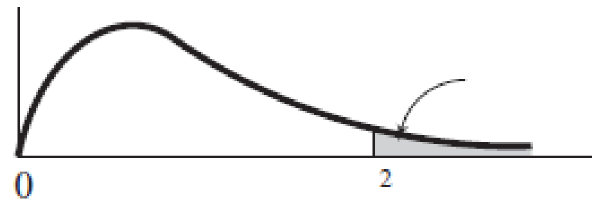
Under H_0 (rader og kolonner er uavhengige) følger kji-kvadrat statistikken over ulike utvalg en kji-kvadrat fordeling med $(r-1)\times(c-1)$ frihetsgrader.

Hvor mange frihetsgrader har den da i vårt tilfelle?

$$(r-1)\times(c-1) = (2-1)\times(3-1) = 2$$

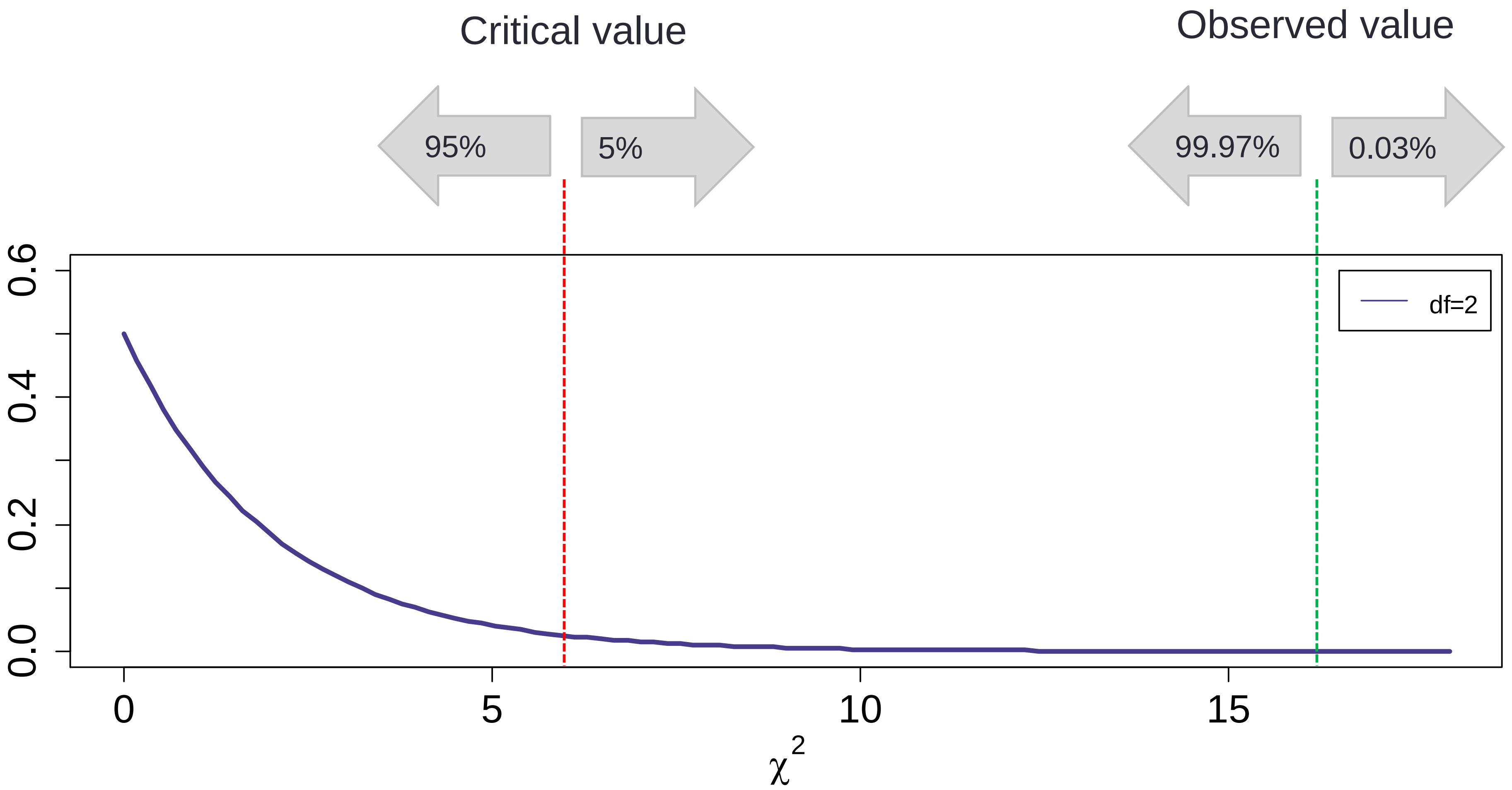
The Chi-square table (H. p. 691)

Appendix χ^2: Upper Percentage Points of the χ^2 Distribution



df	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.82
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92
10	2.15	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03

Den kritiske verdien for en kji-kvadrat fordeling med 2 frihetsgrader er 5.99



Teststatistikken har en verdi på16.20, den kritiske verdien er 5.99. Hva kan vi konkludere med vedrørende kjønn og poløitisk tilhørighet?

Et ganske annet eksempel

- Frekvensen av ord i Jane Austens bøker.
  - Etter hennes død forsøkte en annen forfatter å fullføre Sandition (Austen skrev selv de første 11 kapitlene før hun ble syk).
  - Klarte han å skrive i en stil uadskillelig fra Jane Austen?

Ord	Sense and sensibility	Emma	Sandition I	Sandition II
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94	105	37	22
with	59	74	28	43
without	18	10	10	4
Total	375	440	202	196

Test av kolonner 1-3: `12.2714, df = 10, p-value = 0.2673`

Test of kolonner 1-4: `45.5775, df = 15, p-value = 6.205e-05`

Analyse av residualene

- Sammenlikning av de observerte og forventede frekvensene lar oss teste hypotesen om uavhengighet.
- Forskjellen $f_o - f_e$, differansen mellom observer frekvens of forventet frekvens kalles et residual.
- Hvor stort må et residual være før det er av betydning? Også her standardiserer vi residualene.

Standardized Residual

The ***standardized residual*** for a cell equals

$$z = \frac{f_o - f_e}{se} = \frac{f_o - f_e}{\sqrt{f_e(1 - \text{row proportion})(1 - \text{column proportion})}}.$$

Here, *se* denotes the standard error of $f_o - f_e$, presuming H_0 is true. The standardized residual is the number of standard errors that $(f_o - f_e)$ falls from the value of 0 that we expect when H_0 is true.

Kjønn	**Democrats**	**Independent**	**Republican**	Totalt
Kvinner	573 (522.93)	516 (540.38)	422 (447.68)	1511
Menn	386 (436.07)	475 (450.62)	399 (373.32)	1260
Totalt	959	991	821	2771

- Residual for *mannlige demokrater:* 386-436.07=-50.07

- For å finne det *standardiserte residualet*deler vi på SE.
  - Row proportion = (1260/2771)
  - Column proportion = (959/2771)

  - Se= $\sqrt{((436.07)*(1-\text{Row proportion})*(1-\text{Column proportion}))}$ =12.47

  - Standardiserte residual, Z=(386-436.07)/(−4.015)=-4.01

TABLE 8.8: Standardized Residuals (in Parentheses) for Testing Independence between Party ID and Gender

	Party Identification		
Gender	Democrat	Independent	Republican
Female	573 (4.0)	516 (−1.9)	422 (−2.1)
Male	386 $\boxed{(-4.0)}$	475 (1.9)	399 (2.1)

Avsluttende om Kji-hvadrat analysene

- Begrenset innsikt følger av å forkaste H_0.
  - Selv om du kan forkaste hypotesen om uavhengighet, sier Kji-kvadrat tester sier i seg selv ikke noe om typen eller styrken på sammenhengen.
- Testen er ikke egnet for små utvalg (typisk kreves det at det forventede antallet er minst 5 i alle cellene).
- For valide resultater kreves det at observasjonene er uavhengige av hverandre.