

JMP LAB 3, MET 3431

Korrelasjon, regresjon og χ^2 -kvadrattesten, dvs. Kap 10 og 11.

Datafil: klassens data alle ny (den ligger i Blackboard).

DEL 1. Korrelasjon og regresjon

Er det noen sammenheng mellom sigarettforbruket i et land og dødeligheten av hjertesykdommer i landet? Hvis vi måler x =gjennomsnittlig sigarettforbruk (per voksen per år) og y =dødelighet av hjertesykdom (per 100 000 per år) blir resultatet som vist i tabellen nedenfor.

Land	Sigarettforbruk	Dødelighet
Usa	3900	257
Canada	3350	211
Australia	3220	238
England	2790	194
Sveits	2780	124
Irland	2770	187
Island	2290	110
Finland	2160	233
Tyskland	1890	150
Mexico	1680	32
Frankrike	1410	60
Norge	1090	136

1. Legg inn dataene i en ny JMP fil, med tre variabler: *Land*, *Sigarettforbruk* og *Dødelighet*. Variablene er henholdsvis av type nominal, kontinuerlig og kontinuerlig. Lag et plot av Sigarettforbruk mot Dødelighet (y). *Analyze > Fit Y by X*. Legg Sigarettforbruk på X og Dødelighet på Y. Ser du en lineær trend? Hva er trenden?
2. Regn ut *korrelasjonskoeffisienten* r . JMP: Velg den røde diamanten på scatterplott og gå til *Density Ellipse*, der du velger f.eks. 0.95. Du får en utskrift nederst i output vinduet, *Correlation*. Hva blir den forklarte variansen r^2 ? Gi en tolkning av r^2 .

3. Kjør t-testen for lineær samvariasjon. Test om det er en positiv sammenheng mellom røyking og hjertesykdommer. Bruk symbolet ρ for korrelasjonskoeffisienten i populasjonen til å skrive opp nullhypotesen og alternativhypotesen på symbolsk form. Test på $\alpha=0.05$ nivået. Testobservatoren er: Hva blir t, og hva er den kritiske verdien? Konkluder i et lett forståelig språk.

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

4. Få JMP til å beregne regresjonslinja: Velg rød diamant og *Fit Line*. Skriv opp regresjonsformelen. Hva er stigningstallet? Tolk stigningstallet. Anta at for Sverige vet vi at Sigarettforbruket er 1480 men at vi ikke har data for hjertesykdommer i Sverige. Bruk regresjonsformelen til å beregne dødelighet fra hjertesykdommer i Sverige.
5. Åpne fila *klassens data alle ny* som ligger i Blackboard. Sjekk om det er en lineær korrelasjon mellom alder og timer brukt på studier for studentene i Drammen. Skriv opp hypotesene på symbolsk form, og konkluder i et lettfattelig språk. Bruk $\alpha=0.05$.

DEL 2. Kji-kvadrattesten

Denne testen brukes for samvariasjon mellom *kategoriske (nominal)* variabler.

6. Du jobber i et plateselskap. Sjefen lurer på om det er en sammenheng mellom musikk-sjanger og salgskanal (butikk eller nedlasting over nett). Salgstallene for sist uke er gitt i en kryss-tabellen:

	Nedlastet	Butikk	
Pop	180	290	470
Klassisk	15	73	88
Folkemusikk	35	87	122
	230	450	680

Bruk Kji-kvadrat testen og avgjør om det er avhengighet mellom salgskanal og sjanger. Skriv opp nullhypotesen og alternativhypotesen. Bruk $\alpha=0.01$. Regn for hånd ut de forventede verdiene E og beregn kji-kvadrat testobservatoren: Finn kritisk verdi i tabell A-4. Konkluder i lettfattelig språk.

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

7. *klassens data alle ny*: Test på $\alpha=0.05$ nivået om det er sammenheng mellom røyking og holdning til fotball på tv. Se på alle byene sammen. Velg Analyze>Fit Y by X og legg *Fotball* på X og *Røyker* på Y. Hva er forventet antall E for røykere som er helt uenige? Hva er verdien på testobservatoren? Frihetsgrader? Hva er p-verdien? Konkluder.