Eksperimenter 6:
Statistik og sandsynlighed

Dette er et crash-kursus i de mest basale begreber i statistik og sandsynlighed. Vi har allerede brugt en del statistik og sandsynlighedsregning i forbindelse med entropibegrebet og den statistiske termodynamik. De metoder vi benytter er dog helt generelle, og anvendes som bekendt også for eksempel til vurdering af usikkerheder ved eksperimenter eller undersøgelse af signifikans i forbindelse med stikprøver.

For at illustrere begreberne vil vi her udelukkende benytte os af terningerne som vores udstyr. Denne øvelse laver vi i plenum, hvor hver studerende har 10 terninger.

Som det første vil vi introducere lidt begreber:

Terningekast nummer "i" benævnes X_i og kaldes en *stokastisk variabel*, det faktiske *udfald* af den stokastisk variabel X_i benævnes x_i (altså lille x).

Den stokastiske variabel kan antage værdier i *udfaldsrummet* $\Omega = \{1, ..., 6\}$. Det er en tradition i matematikken at benævne udfaldsrummet med Ω. Den tradition følger vi uden at forveksle med multipliciteten, som vi her vil benævne ν_j for det j-te udfald. Vores udfald benævner vi ϵ_j, som i vores tilfælde simpelthen er $\epsilon_j = j$ for $j = 1, 2, 3, 4, 5, 6$. (For at sammenligne med den statistiske fysik, kan vi kalde udfaldene ϵ_j makrotilstande, hvor hver makrotilstand indeholder ν_j mikrotilstande.)

Til hvert udfald ϵ_j tilknytter vi en *sandsynlighed* $p_j = p(\epsilon_j)$, hvor $0 \leq p_j \leq 1$. (Igen for at knytte an til fysikken: Der er den fundamentale antagelse at $p_j = \nu_j / \nu_{\text{total}}$, hvor $\nu_{\text{total}} = \sum_j \nu_j$ er samtlige mikrotilstande).

Til enhver delmængde af Ω (for eksempel $A = \{1\} \cup \{4\} = \{1, 4\}$) er der tilknyttet en sandsynlighed $p(A)$. Om sandsynligheden skal der gælde de simple regneregler at $p(A \cup B) = p(A) + p(B) - p(A \cap B)$ og $P(\Omega) = 1$.

1 Vis på en tegning at regnereglen ovenfor er rimelig.

2 Vis at den passer ved at udregne sandsynlighederne i følgende tilfælde: $A = \{\text{øjenantallet er lige}\}$, $B = \{\text{øjenantallet er større end 3}\}$

3 Vis ud fra de to regler ovenfor at komplementærmængden $(\Omega \setminus A)$ til mængde A har sandsynlighed $p(\Omega \setminus A) = 1 - p(A)$.

Vi vil nu definere:

Middelværdi eller *gennemsnit*: $\mu = \sum_j p(\epsilon_j)\epsilon_j$

Varians: $\sigma^2 = \sum_j p(\epsilon_j)(\epsilon_j - \mu)^2 = \sum_j p(\epsilon_j)\epsilon_j^2 - \mu^2$

Bemærk at σ har samme dimension som ϵ_j. σ kaldes *spredning*.

4 Vis det sidste lighedstegn i definitionen på variansen.

5 Udregn middelværdi og varians for et terningekast, når vi antager at terningen er "sand", altså at alle udfald er lige sandsynlige.

En serie af udfald $x_1, x_2, ..., x_n$ kaldes en *stikprøve* af længde n.

Ud fra vores stikprøve kan vi definere:

Stikprøvemiddelværdi: $m = \frac{1}{n}\sum_{i=1}^{n} x_i$

Stikprøvevarians: $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - m)^2$.

Læg mærke til at vi benyttede græske bogstaver til definitionerne med sandsynligheder og latinske bogstaver til definitionerne for stikprøver. Læg ligeledes mærke til at summerne som indgår i de to forskellige sæt af definitioner er forskellige (hvordan?). Om lidt skal vi se hvorfor der står $(n-1)$ i nævneren for stikprøvevariansen og ikke, som man måske kunne tro, n.

Lad os nu antage at vi ikke kender noget til om terningerne er "sande". Vi kender altså ikke p_j.

6 Hvordan skal vi bestemme p_j for en given "mistænkelig terning?

Vi bestemmer nu *fordelingen* $\tilde{p}_j$ eksperimentelt ved kast med $n = 900$ terninger. Vi definerer (selvfølgelig) $\tilde{p}_j = n_j/n$, hvor n_j er antallet af udfald $x_i = j$.

7 Vis at stikprøvemiddelværdien og den "sande" middelværdi er identiske hvis $\tilde{p}_j = p_j$ for alle j.

8 Hjemmeopgave for de flittige: Udregn spredningen omkring 150 for antallet af ettere ud af 900 kast. Vink: Benyt binomialformlen for ettere og ikke-ettere, Stirlings formel og $\ln(1 + x) \approx x$.

For at gøre de følgende udledninger mere gennemskuelige indfører vi nu følgende notation for den "sande" middelværdi: $\langle f(x)\rangle = \sum_j p_j f(x_j)$ og dermed $\mu = \langle x_i\rangle$ samt $\sigma^2 = \langle x^2\rangle - \langle x\rangle^2$.

Vi forestiller os nu at udtage en stikprøve af længde n, for eksempel ved at gentage et eksperiment n gange. Det vi ønsker at bestemme er den "sande" middelværdi μ, mens det vi får er stikprøvemiddelværdien m. Vi kan altså betragte m som en måling af μ. Vi er interesseret i usikkerheden på denne måling, således at vi kan bestemme hvor stor en stikprøve vi bliver nød til at tage.

Vi bestemmer først middelværdien for m (ud af en række stikprøver):

$$\langle m\rangle = \langle\frac{1}{n}\sum_{i=1}^{n} x_i\rangle = \frac{1}{n}\sum_{i=1}^{n}\langle x_i\rangle = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

Vi får altså i gennemsnit det rigtige resultat, der er altså intet *bias* (systematisk fejl). Dernæst bestemmer vi variansen:

$$\begin{aligned}
\sigma(n)^2 = \langle m^2\rangle - \mu^2 &= \langle(\frac{1}{n}\sum_{i=1}^{n} x_i)^2\rangle - \mu^2 && (1)\\
&= \frac{1}{n^2}\langle\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j\rangle - \mu^2\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle x_i x_j\rangle - \mu^2\\
&= \frac{1}{n^2}\Big(\sum_{i=1}^{n}\langle x_i^2\rangle + \sum\sum_{i\neq j}\langle x_i x_j\rangle\Big) - \mu^2\\
&= \frac{1}{n^2}(n\langle x^2\rangle + (n^2-n))x\langle^2) - \mu^2\\
&= \frac{1}{n}(\langle x^2\rangle - \mu^2) = \frac{\sigma^2}{n}
\end{aligned}$$

Vi har her benyttet $\langle x_i x_j\rangle = \langle x_i\rangle\langle x_j\rangle$, som gælder når X_i og X_j er *uafhængige* stokastiske variable. (For uafhængige variable gælder per definition at $p(x_i, x_j) = p(x_i)p(x_j)$, hvorfra vi får at udtrykket ovenfor gælder). Altså får vi spredningen $\sigma(n) = \sigma/\sqrt{n}$ for middelværdien bestemt ud fra n udfald. Det resultat kommer ikke som nogen overraskelse!

9 Vi har en mistanke om at der kastes plat og krone med en "falsk" mønt, som har p(krone)=0,51. I et spil taber vi en krone ved "plat" og vinder en krone ved "krone". Bestem μ og σ for denne falske mønt. Hvor mange gange skal vi kaste mønten for at bestemme om den er falsk?

Som lovet kan vi nu forstå faktoren $(n-1)$ i formlen for stikprøvevariansen:

10 Udregn gennemsnittet for stikprøvevariansen. Vink: indsæt udtrykket for m og udregn (pas på dobbeltsummerne!).