Repetisjon av grunnleggende begreper i statistikk

- En **populasjon** består av alle mulige observasjoner vi kan gjøre.
- En populasjon beskrives som regel gjennom en fordeling f_X. I dette tilfellet kaller vi populasjonen en **f_X-populasjon**, og vi mener at å trekke en tilfeldig verdi fra populasjonen er det samme som å trekke en tilfeldig verdi fra fordelingen f_X.

  NB: Vi kaller en $\mathcal{N}(\mu, \sigma^2)$-populasjon en **normalpopulasjon**.
- En delmengde av populasjonen kalles et **utvalg**.
- For å være nyttig må utvalget være **representativt** for populasjonen. I vårt kurs oppnås det ved et **tilfeldig utvalg** hvor n verdier velges tilfeldig (med en uniform fordeling) fra populasjonen. Med et tilfeldig utvalg av størrelse n fra en f_X-populasjon mener vi at $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f_X$.
- En **observator** er en observerbar funksjon av en eller flere av de stokastiske variablene som utgjør utvalget. Typiske eksempler et empirisk middelverdi og empirisk varians.
- Observatorer er stokastiske variabler under gjentatte tilfeldige utvalg fra populasjonen. Denne sannsynlighetsfordelingen kalles en **utvalgsfordeling**.
- **Ekstremt viktig:** En lineærkombinasjon av uavhengige normalfordelte stokastiske variabler har en normalfordeling. Spesielt betyr dette at hvis $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, så har vi $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(0, \sigma^2/n)$.
- **Sentralgrenseteoremet** sier at hvis $\bar{X}_n$ er det empiriske middlet i et tilfeldig utvalg av størrelse n fra en populasjon med forventningsverdi μ og endelig varians $\sigma^2 < \infty$. Så er har vi

  $$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0, 1)$$

  når $n \to \infty$.

  NB: Vi antar at for $n \geq 30$ er $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ en god tilnærming.
- **Ekstremt viktig:** Dette betyr at for en vilkårlig populasjon med endelig varians vet vi at utvalgsfordelingen til $\bar{X}$ for et tilfeldig utvalg av størrelse n er tilnærmet normal med forventningsverdi μ og varians σ^2/n hvis n er stor nok.

- Hvis $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ er empirisk varians for et tilfeldig utvalg av størrelse n fra en normalpopulasjon med varians σ^2 har vi at

$$\frac{n-1}{\sigma^2}S^2 = \sum_{i=1}^{n}\frac{(X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}.$$

  NB: Vi har mistet en frihetsgrad fordi vi alltid har $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ og det bare er $n-1$ dimensjoner hvor variasjon er mulig.

- **Ekstremt viktig:**
  - S^2 er **ikke** χ^2_{n-1}-fordelt! $\frac{n-1}{\sigma^2}S^2$ er χ^2_{n-1}-fordelt!
  - Dette gjelder **bare** for en normalpopulasjon. Det finnes ingen sentralgrenseteorem som gjør at dette er gyldig for andre populasjoner for store n!