CONFIDENCE INTERVALS

Assume some familiarity with

- Normal distribution/Standard Normal Distribution
- Standard Deviation
- Central Limit Theorem

We take a sample of data from dataset with any kind of distribution

- And takes its mean $\hat{\mu}$
- Let's go away from this for a bit and focus on Normal Distributions and the Central Limit Theorem

Central Limit Theorem

- Lets say we take N different samples
  - Sample 1 (a group of observations)
  - Sample 2 (a group of observations)
  - ….
- The CLT says that the means of these samples surround the population mean $\boldsymbol{\mu}$ and the means of these samples follow a **normal distribution** with standard deviation $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$ where **n**="size of the sample" and $\boldsymbol{\sigma}$ is the Global STD. (With bootstrapping we can estimate that with the sample STD)

Normal distribution

- Approximately 66 % of the data is 1 STD away from mean
- Approximately 95 % of the data is 2 STD away from mean
- Approximately 99 % of the data is 2.5.. STD away from mean

Coming back .. We take a sample of data from a random dataset with any kind of distribution and assume a 95 % confidence interval

- And take its mean $\widehat{\mu}$
- We know this $\hat{\mu}$ is part of a distribution over many $\hat{\mu} - s$ from different imaginary samples.
- Based on the characteristics of Normal distribution (95 % of data 2 STD from mean) we can say with 95 % confidence that $\hat{\mu}$ is somewhere in the interval 2 STDs away from the mean OR we can say with 99 % confidence that its 2.5.. STDs interval from the Global mean.
- Based on Central Limit Theorem we can say the STD= $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$
- **Putting these 2 things together we can say that our observation $\hat{\mu}$ lies with 95 % confidence In area $\boldsymbol{\mu}$** +/- 2 $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$
- If we are 95 % sure that our observation is 2 $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$ **from global mean then we are saying that we are 95 % sure that global mean is +/-** 2 $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$ **from our observation.**
  - **If it would not**
    - **The observation would have to be further than** 2 $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$ **from the global mean**
  - **AKA we have confidence interval of 95 % that global mean $\boldsymbol{\mu}$ is in area of +/-** 2 $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$ **from** $\hat{\mu}$
  - OR
  - **With same logic we have confidence interval of 99 % that global mean $\boldsymbol{\mu}$ is in area of +/-** 2.5 $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$ **from** $\hat{\mu}$