Lecture 4: Sampling + Sampling Distributions

⇒ random sampling

- sampling with replacement
- sample contains N random variables which are all i.i.d. (independent and identically distributed)

⇒ stratified random sampling

- in US population:

54,0%	"Protestant" (i.e. non-Catholic, non-Orthodox)
23,9%	Catholic
0,6%	Orthodox
1,7%	Jewish
0,6%	Muslim
0,4%	Hindu
0,7%	Buddhist
2,0%	Other
16,1%	atheist, agnostic, nothing in particular

↑ bad definition because I'm including Coptic Christians + others in the Protestant category

- this classroom however is not a representative sample of the US population
- It looks rather Brooklyn to me.

- so suppose we polled this class: "Do you believe that a woman should have unrestricted access to abortion services?" (Yes = 1, No = 0)
- suppose that we have reason to believe that response depends on gender + religion
- In that case, we could poll this class, average the responses by gender + religion categories and then weight by proportion in US population
- would then have an estimate of US population mean

→ In population
- mean: μ
- variance: σ^2

→ In sample
- mean: $\bar{x}$
- variance: s^2

because mean + variance of a sample are not necessarily equal to population mean + variance

→ If we collect hundreds of equally sized samples from the same population & we calculated the mean of each sample, then the "mean of the means" would be:

$$E[\bar{x}] = \mu \quad \leftarrow \text{unbiasedness}$$

with: $\text{var}[\bar{x}] = \frac{1}{N}\sigma^2$ std. error: $\sigma/\sqrt{N} = \sigma_{\bar{x}}$

→ the expected value of the mean is the population mean

→ but each sample has a different mean so there is some variance in our estimates

- if we collect larger samples (high N)
- then there will be less variance in our estimate of the mean
- so std error falls as N increases

→ NOTE: The frequency distribution of sample means is approximately Normal even when the samples are taken from a distribution that is not normal in shape