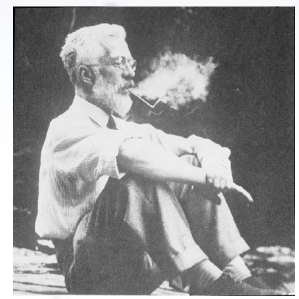
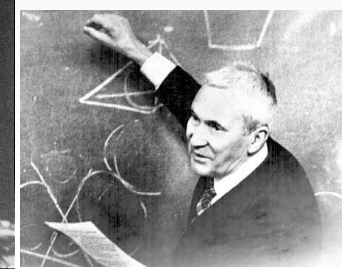
Applied Statistics

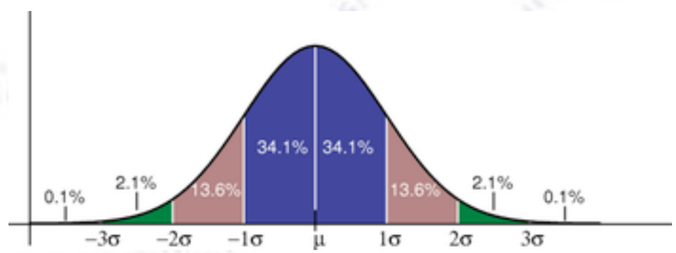
Stratification





Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

Stratified sampling

Suppose you want to measure the average height of students at KU.

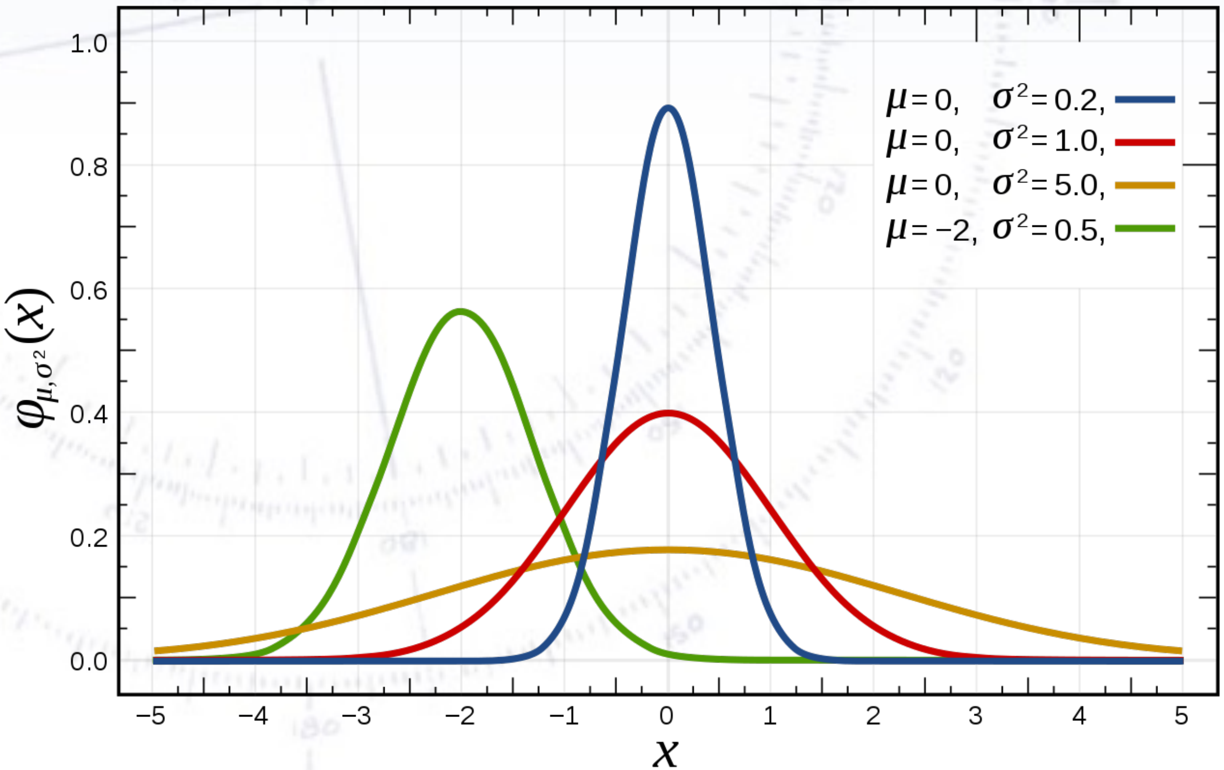
The simplest method is done by sampling N measurements, calculating the mean μ and standard deviation σ, and quoting the result as $\mu \pm \sigma/\mathrm{sqrt}(N) = \mu \pm \sigma_\mu$

Stratified sampling

Suppose you want to measure the average height of students at KU.

The simplest method is done by sampling N measurements, calculating the mean μ and standard deviation σ, and quoting the result as $\mu \pm \sigma/\mathrm{sqrt}(N) = \mu \pm \sigma_\mu$

However, one can do better!



Stratified sampling

Suppose you want to measure the average height of students at KU.

The simplest method is done by sampling N measurements, calculating the mean μ and standard deviation σ, and quoting the result as $\mu \pm \sigma/\mathrm{sqrt}(N) = \mu \pm \sigma_\mu$

Since it is known that students come in (at least) two types known to have different height distributions, the above uncertainty can be reduced if we know the fraction of each type (from other sources).

By separately determining the height of women and men, we avoid two sources of uncertainty:

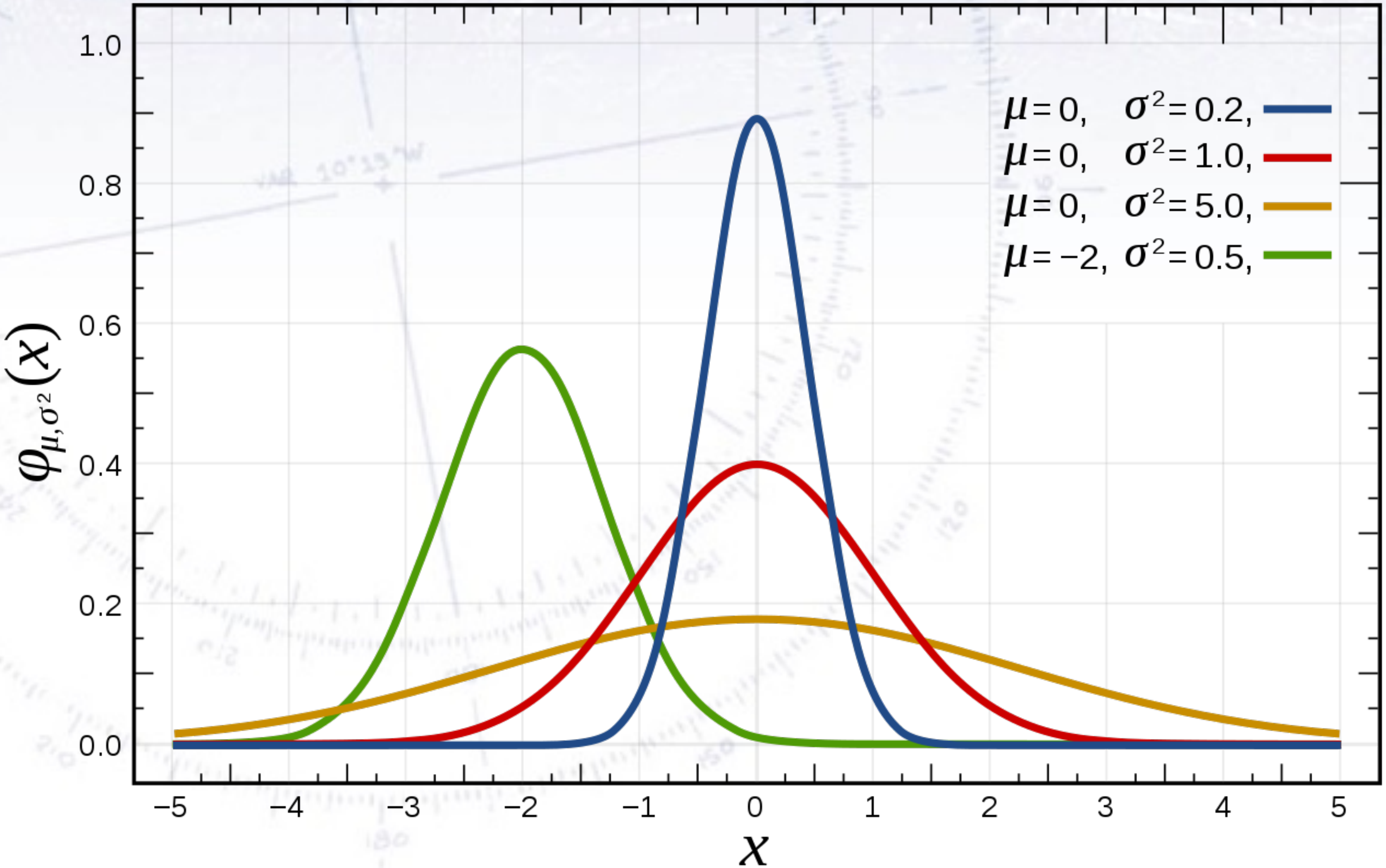
1. The variation due to our random fraction sampling of women and men.
2. The enlarged standard deviation from mixing two samples.

This can be particularly important, when you only have low statistics sampling.

And this is used in anything from costumer characterisation to political surveys.

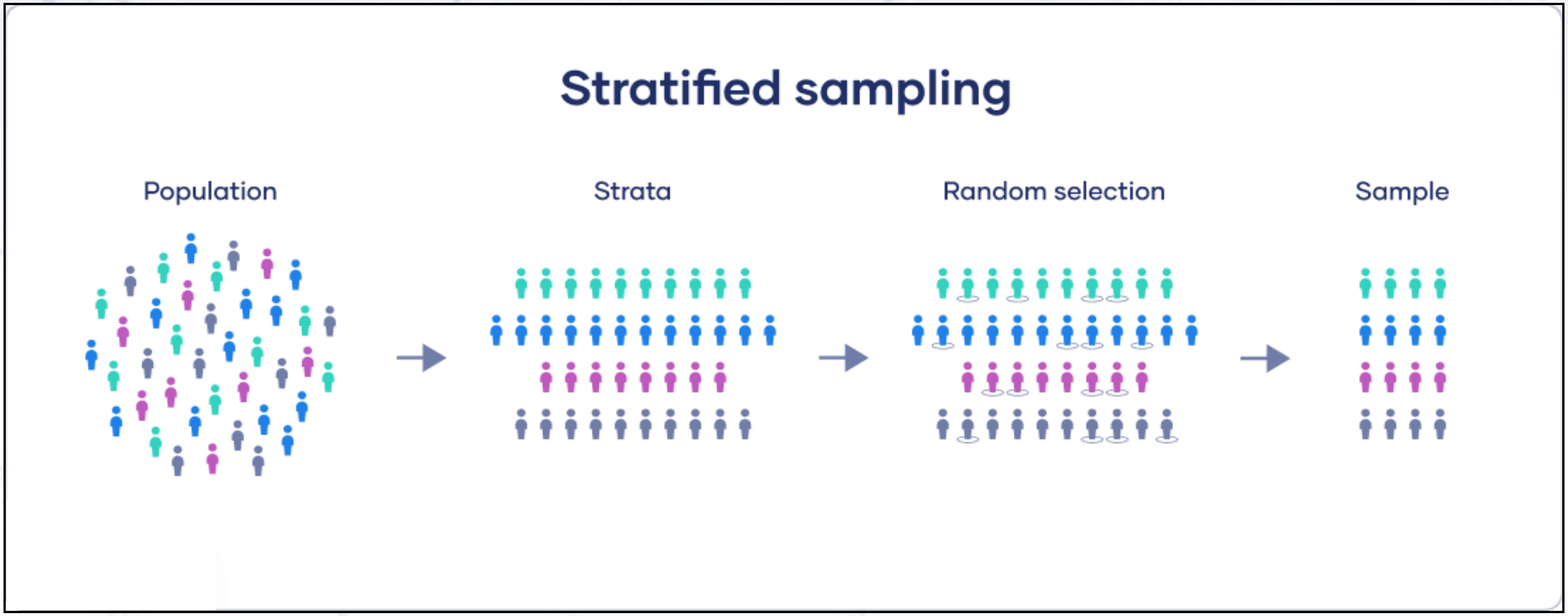
How about variations?

If one sample turns out to have a (much) larger variation (i.e. std.) than the other, then it pays to sample this group more…. proportionally to their std. [Barlow, p.95]



Stratified Sampling

In stratified sampling, you try to get the best out of the small sample you make your estimates from. The optimal way is to divide the sample (into strata), and sample equally in each of these.



If the strata do not have the same standard deviation, then one should select fractions of each strata proportionally to the Std.

An old proof

More generally, one should avoid mixing "good" data with "poor" data. Without dividing it, the poor tends to dilute the good.

Rather, do your analysis for the good and poor separately - getting the full power of both (but especially the good) - and combine these afterwards.

One tagging category: $$D \equiv \frac{1}{N}\sum_{i}^{N}(1-2\langle\omega\rangle_i), \tag{14.1}$$

Two tagging categories: $$D_1 \equiv \frac{1}{N_1}\sum_{i\in N_1}(1-2\langle\omega\rangle_i), \quad D_2 \equiv \frac{1}{N_2}\sum_{i\in N_2}(1-2\langle\omega\rangle_i). \tag{14.2}$$

Obviously, $ND = N_1D_1 + N_2D_2$. The effective tagging efficiency Q for each analysis is then:

$$Q_{one} \equiv \varepsilon D^2 = \varepsilon\left(\frac{N_1D_1+N_2D_2}{N}\right)^2 = \frac{1}{\varepsilon}(\varepsilon_1^2D_1^2+\varepsilon_2^2D_2^2+2\varepsilon_1\varepsilon_2D_1D_2), \tag{14.3}$$

$$Q_{two} \equiv \varepsilon_1D_1^2+\varepsilon_2D_2^2. \tag{14.4}$$

The claim is that Q_{two} is greater than Q_{one}, which can be proven as follows:

$$\begin{aligned} Q_{two}-Q_{one} &= \epsilon_1D_1^2+\epsilon_2D_2^2-\frac{1}{\epsilon}(\epsilon_1^2D_1^2+\epsilon_2^2D_2^2+2\epsilon_1\epsilon_2D_1D_2) \\ &= \frac{1}{\epsilon}[\epsilon_1(\epsilon-\epsilon_1)D_1^2+\epsilon_2(\epsilon-\epsilon_2)D_2^2-2\epsilon_1\epsilon_2D_1D_2] \\ &= \frac{\epsilon_1\epsilon_2}{\epsilon}(D_1^2+D_2^2-2D_1D_2) = \frac{\epsilon_1\epsilon_2}{\epsilon}(D_1-D_2)^2 \geq 0. \quad \square \end{aligned} \tag{14.5}$$

Bonus slides