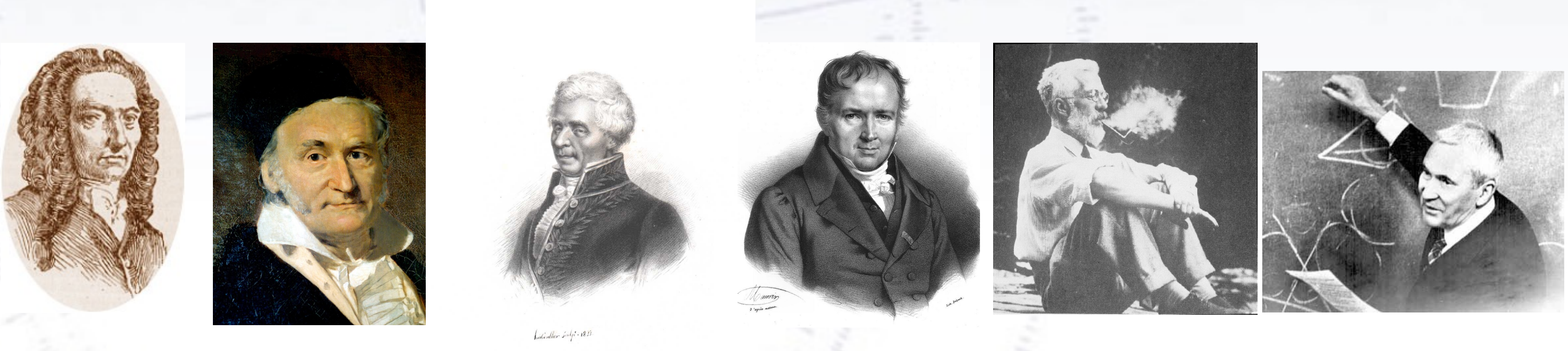
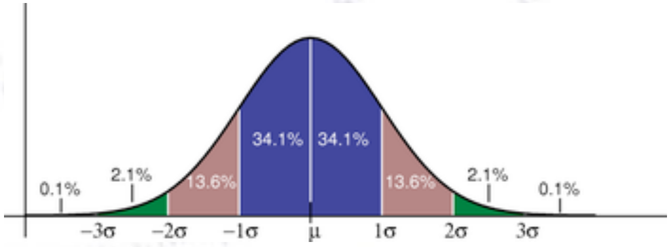
Applied Statistics

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.

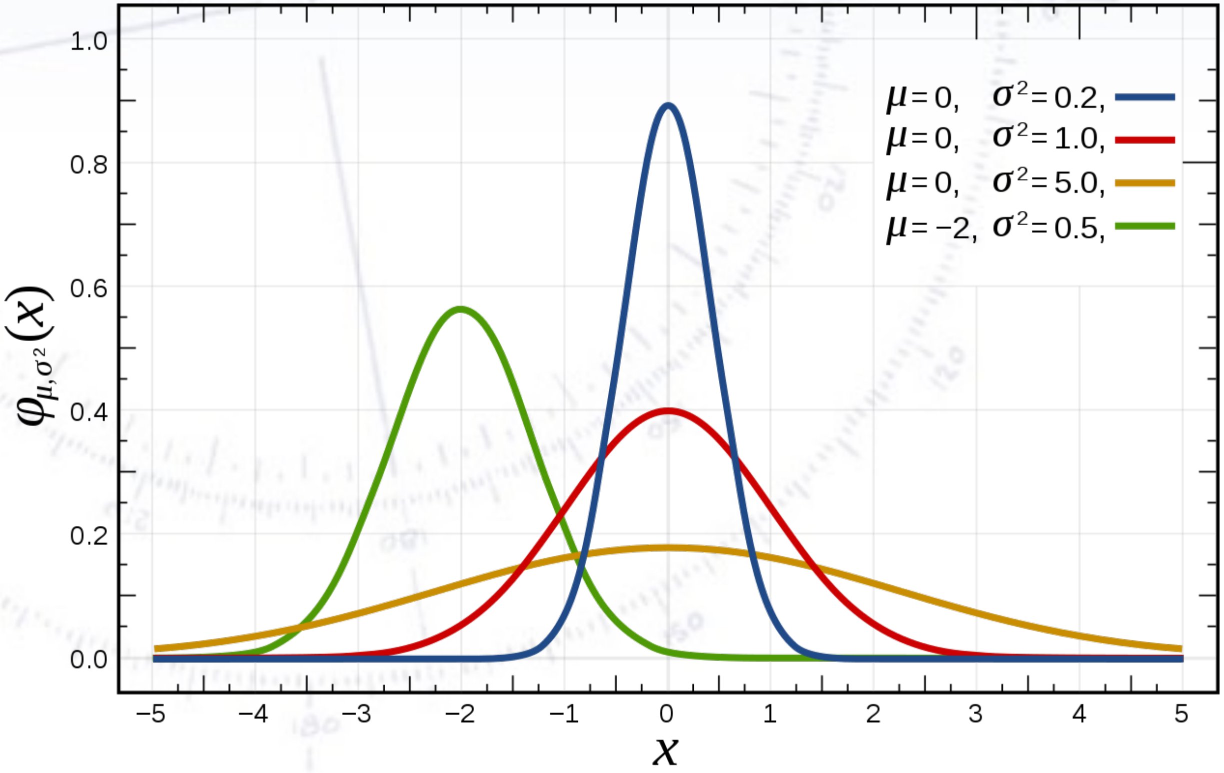
The simplest method is done by sampling N measurements, calculating the mean μ and standard deviation σ, and quoting the result as $\mu \pm \sigma/\text{sqrt}(N) = \mu \pm \sigma_{\mu}$

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).

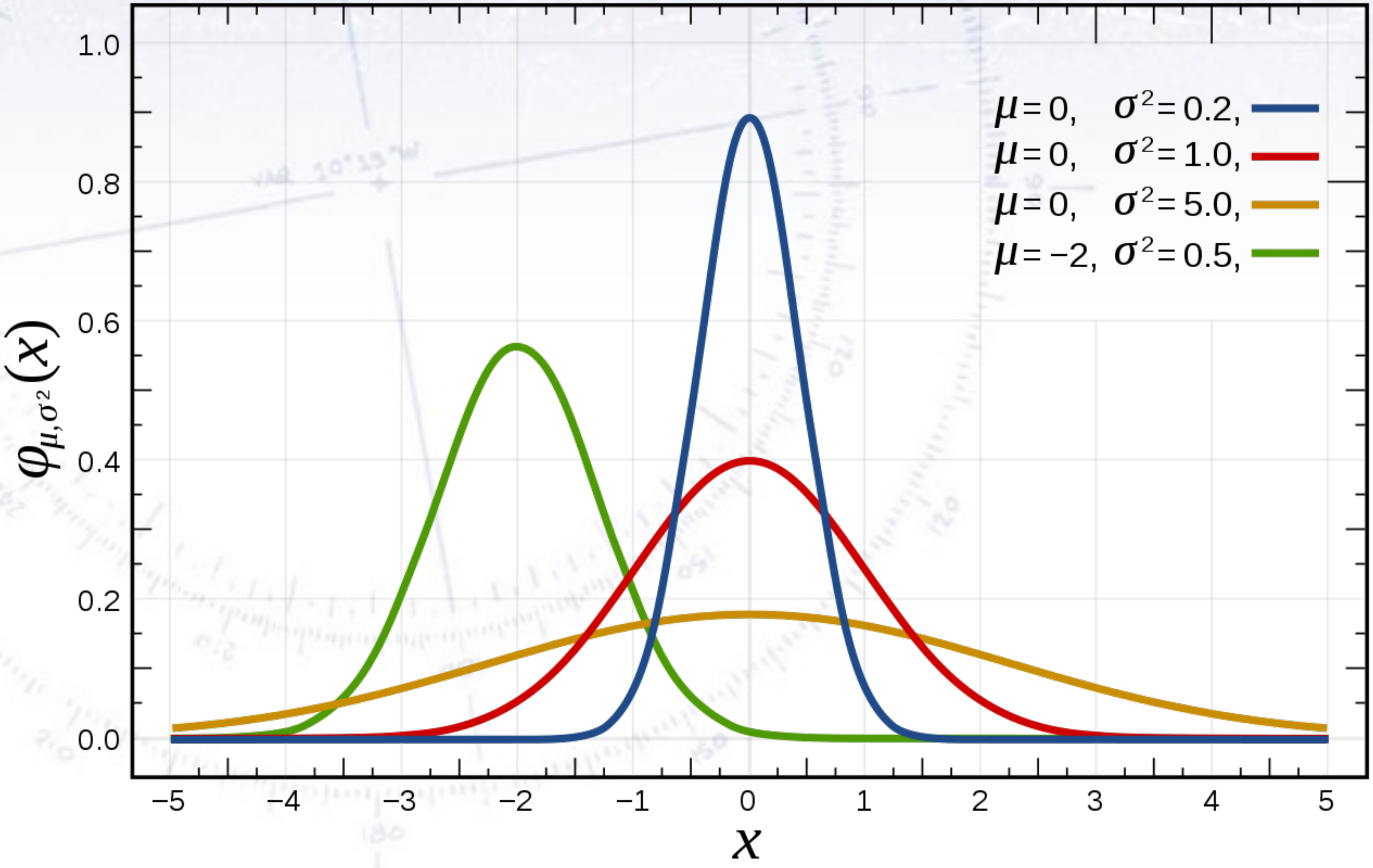
By separately determining the height of women and men, we avoid two sources of uncertainty:
1. The enlarged standard deviation from mixing two samples.
2. The variation due to our random fraction sampling between types.

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

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

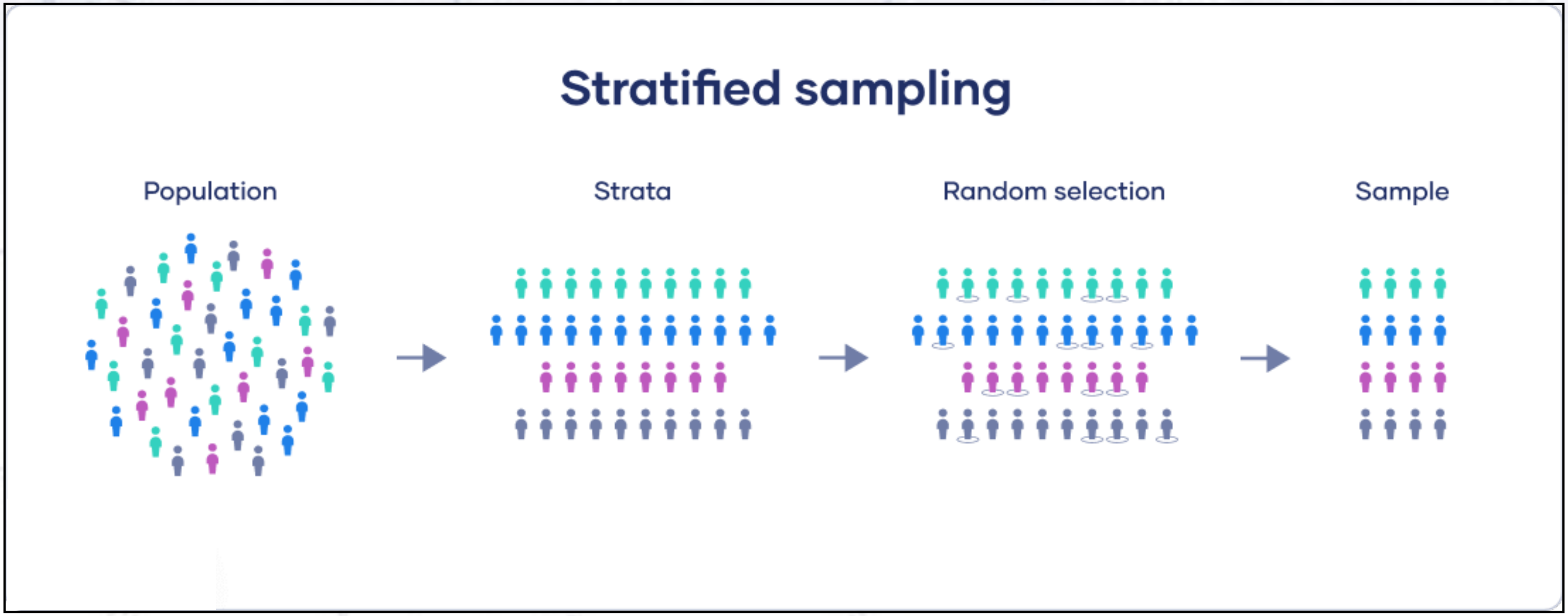
How about variations?

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

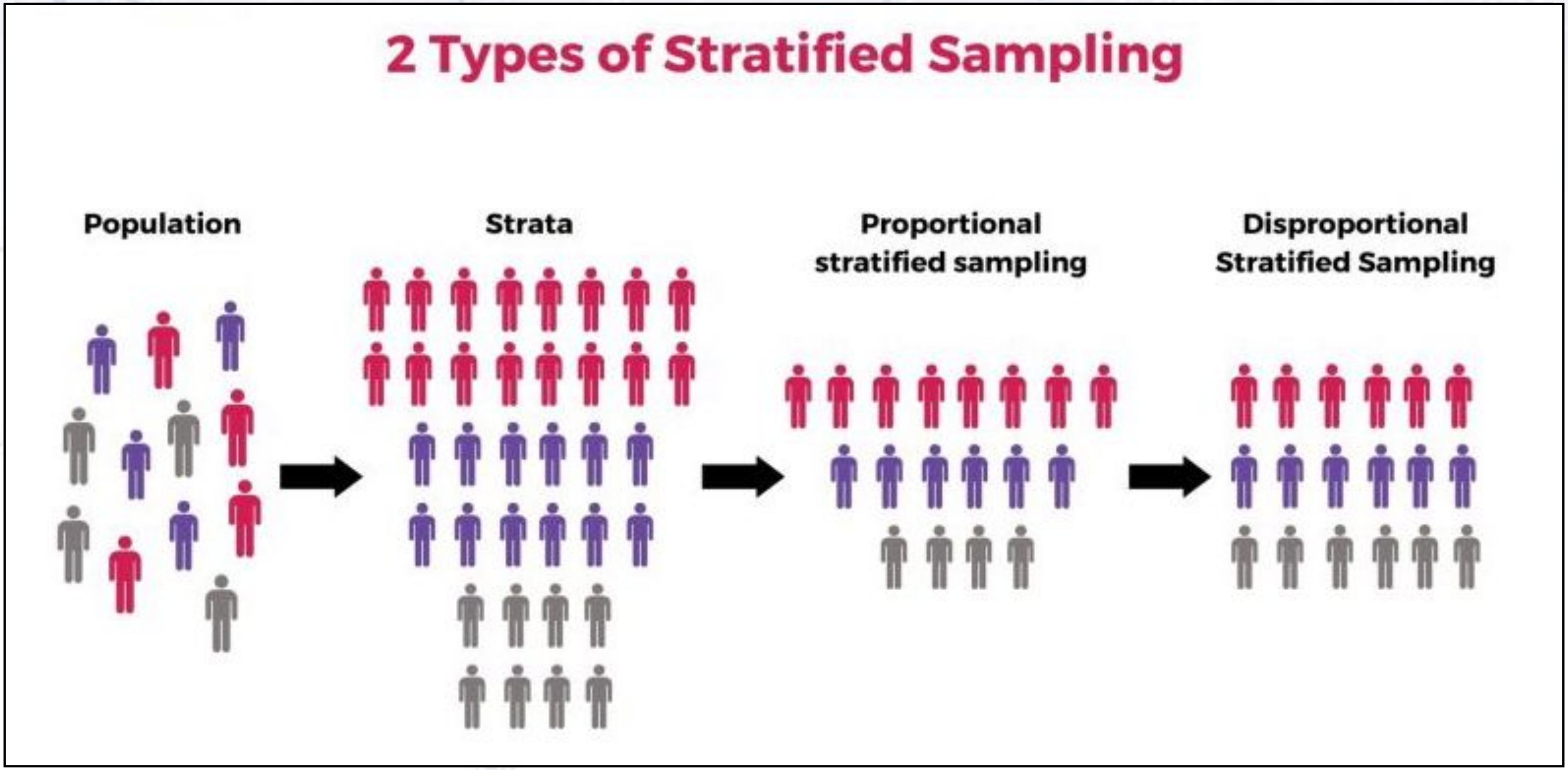
In stratified sampling, you try to get the best out of the small sample you make your estimates from. The optimal way is to divide the sample (into strata), and sample equally in each of these.



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

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.


2 Types of Stratified Sampling
Population
Strata
Proportional stratified sampling
Disproportional Stratified Sampling

Example of Stratification

Imagine that we wanted to estimate the average income of a large population, where it is known (*and this part is required knowledge!*) that:

- 90% has a very standard income (μ = 400k kr, σ = 10k kr).
- 10% has a highly variable income (μ = 500k kr, σ = 100k kr).

As a single large population, they overall have μ = 410k kr and σ = 52.9k kr.

$$V = f_1 V_1 + f_2 V_2 + f_1 f_2 (\mu_1 - \mu_2)^2$$

(Barlow p.94)

Now, we can take two strategies:

- **Normal sampling**: Simply sample 100 random persons. **Result: σ = ± 5.29k kr.**
- **Stratified sampling**: Use the above information and stratify the 100 samplings. This can be done in two ways:

  a: **Proportional**: Proportional to the known strata sizes.
  This approach "kills" the f_1 f_2 $\Delta\mu^2$ term! **Result: σ = ± 4.24k kr.**

  b: **Disproportionally**: Proportional to the fraction times standard deviation.
  This approach further uses knowledge of the Std.
  Now we sample 47/53 (most from small group). **Result: σ = ± 3.61k kr.**

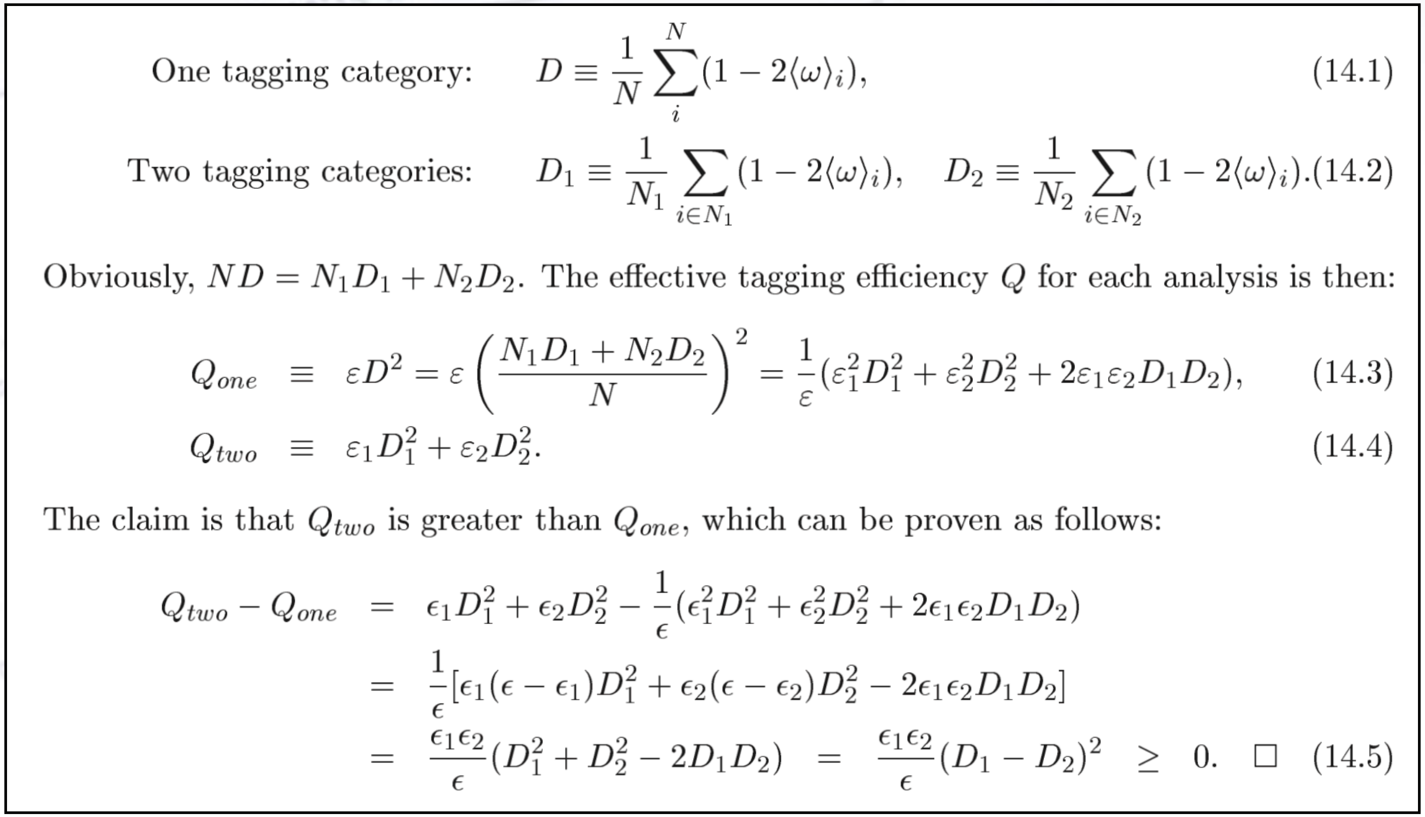
$$V = f_1^2 V_1 / m_1 + f_2^2 V_2 / m_2$$

(Barlow p.95)

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 data separately - getting the full power of both (but especially the good) - and combine these afterwards.



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

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

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

$$\begin{aligned} Q_{one} &\equiv \varepsilon D^2 = \varepsilon\left(\frac{N_1D_1+N_2D_2}{N}\right)^2 = \frac{1}{\varepsilon}(\varepsilon_1^2D_1^2 + \varepsilon_2^2D_2^2 + 2\varepsilon_1\varepsilon_2D_1D_2), && (14.3)\\ Q_{two} &\equiv \varepsilon_1D_1^2 + \varepsilon_2D_2^2. && (14.4) \end{aligned}$$

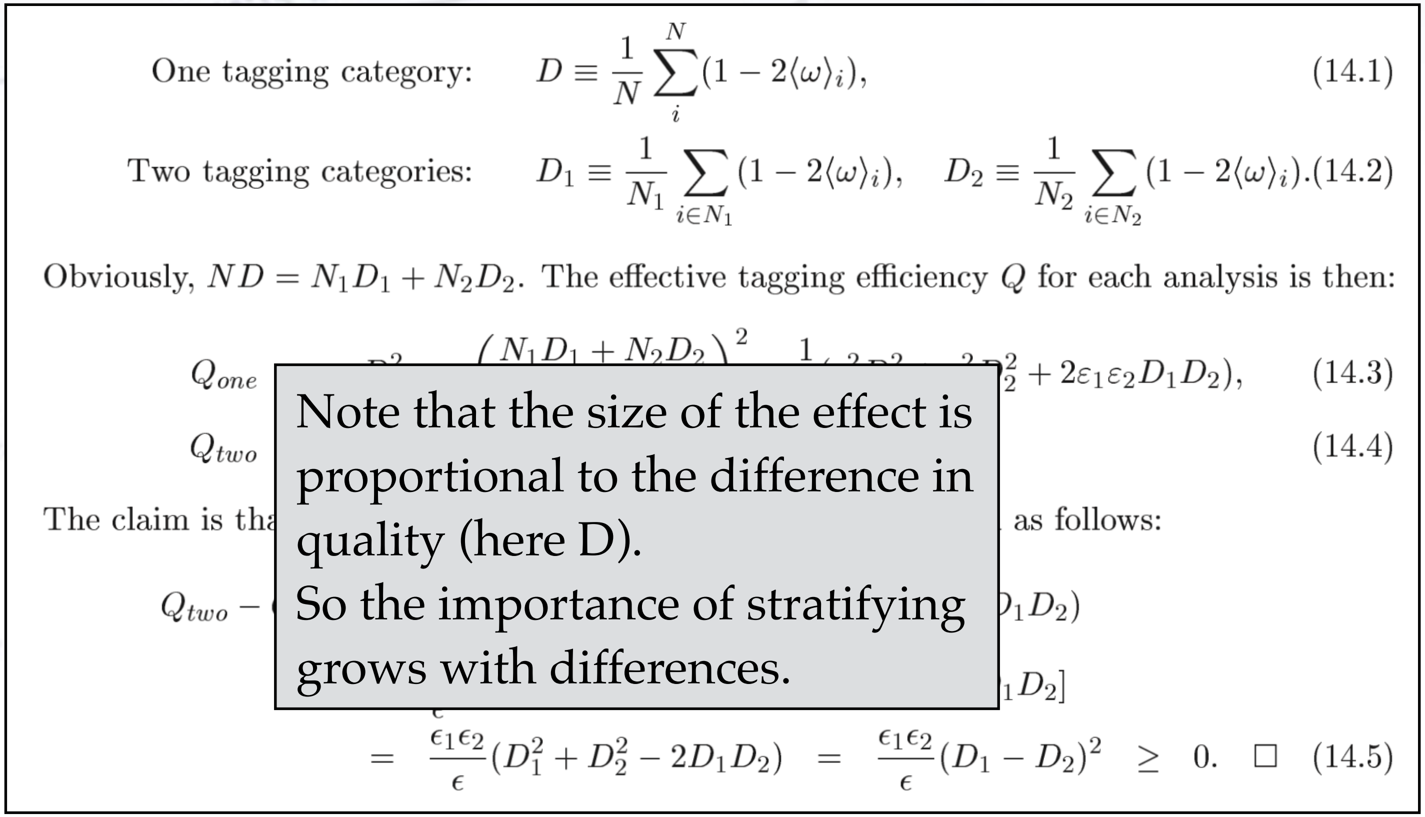
The claim is that Q_{two} is greater than Q_{one}, which can be proven as follows:

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

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 data separately - getting the full power of both (but especially the good) - and combine these afterwards.



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

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

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

Q_{one} … $\left(N_1D_1 + N_2D_2\right)^2$ … $D_2^2 + 2\varepsilon_1\varepsilon_2 D_1 D_2),$ (14.3)

Q_{two} … (14.4)

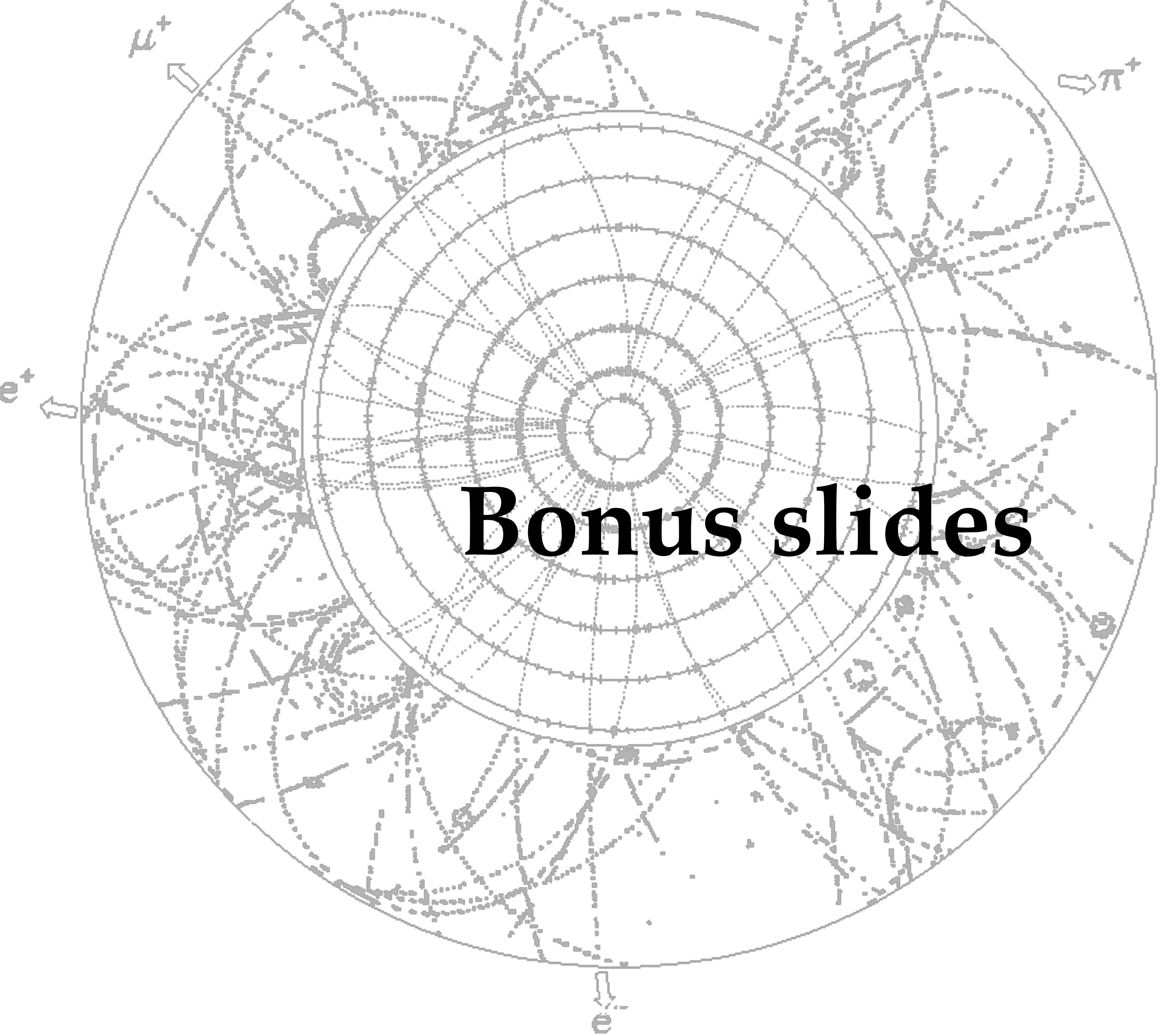
The claim is tha… as follows:

$Q_{two} -$ … $D_1D_2)$

… $D_1D_2]$

$$= \frac{\epsilon_1\epsilon_2}{\epsilon}(D_1^2 + D_2^2 - 2D_1D_2) = \frac{\epsilon_1\epsilon_2}{\epsilon}(D_1 - D_2)^2 \geq 0. \quad \square \qquad (14.5)$$

Note that the size of the effect is proportional to the difference in quality (here D).
So the importance of stratifying grows with differences.



Bonus slides