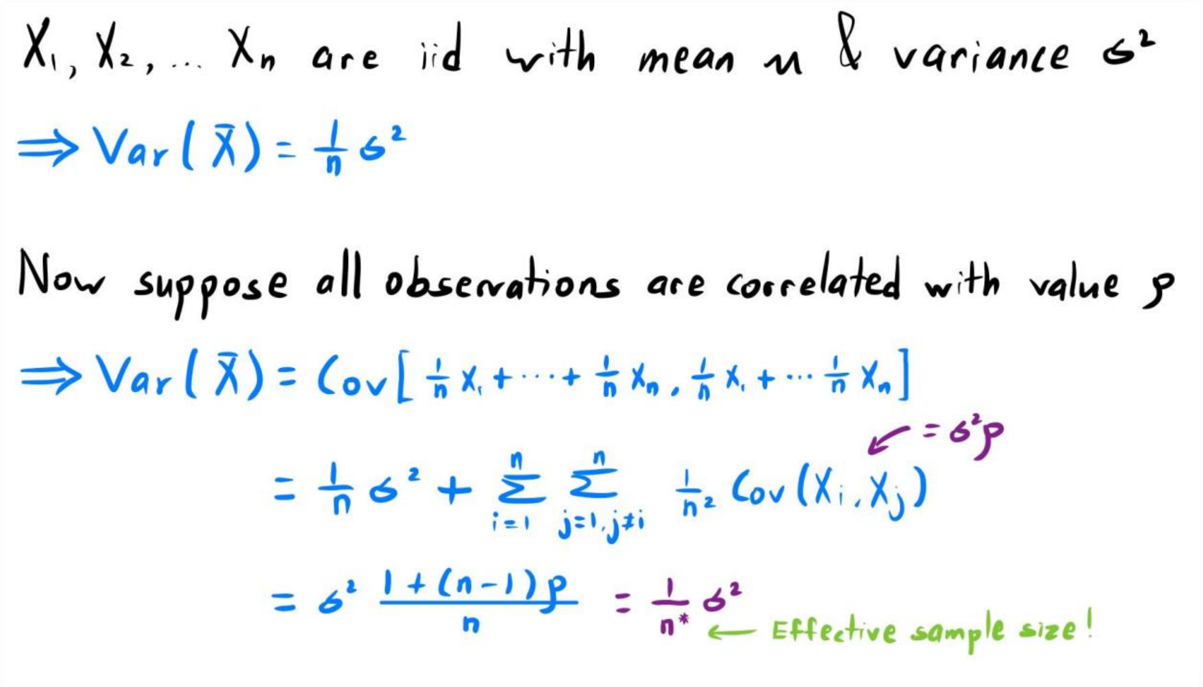
Effective Sample Size, Variance and Belief

A quick search of effective sample size exposes many resources which either define it mathematically or in a verbose manner. We state both approaches here and start with an independent and identically distributed (iid) sample where the true mean and variance are known. The effective sample size n* is defined as follows.



$X_1, X_2, \dots X_n$ are iid with mean μ & variance σ^2

$$\Rightarrow \mathrm{Var}(\bar{X}) = \frac{1}{n}\sigma^2$$

Now suppose all observations are correlated with value ρ

$$\Rightarrow \mathrm{Var}(\bar{X}) = \mathrm{Cov}\left[\frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n, \frac{1}{n}X_1 + \cdots \frac{1}{n}X_n\right]$$

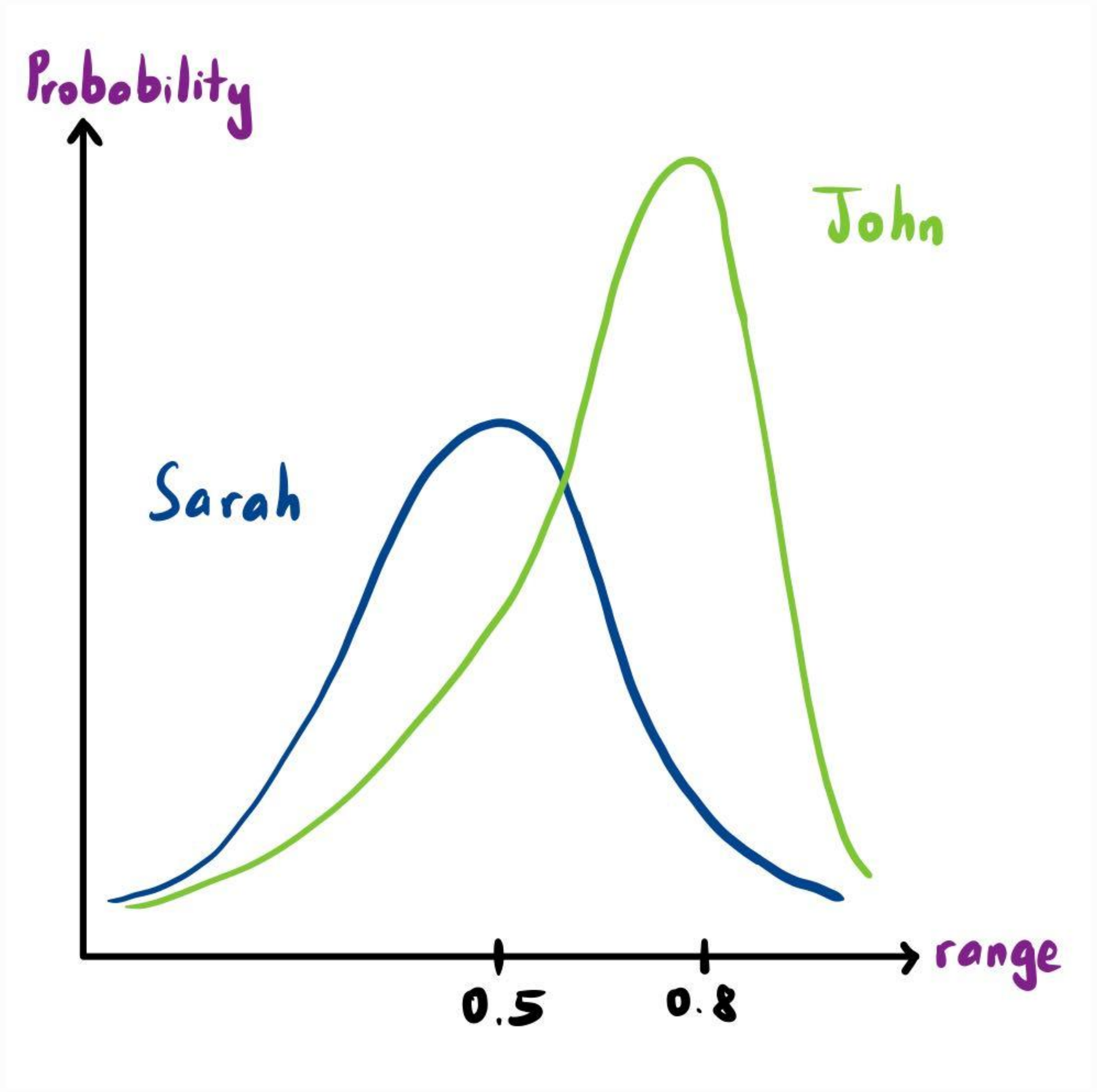
$$\swarrow = \sigma^2\rho$$

$$= \frac{1}{n}\sigma^2 + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{1}{n^2}\mathrm{Cov}(X_i, X_j)$$

$$= \sigma^2 \frac{1+(n-1)\rho}{n} = \frac{1}{n^*}\sigma^2 \leftarrow \text{Effective sample size!}$$

Now that is all well and good, but why did statisticians bother defining n* instead of leaving the coefficient of sigma squared as is? Turns out, **n* represents the size of an iid sample that contains the same amount of precision as our dataset!** Notice here, and in many other web resources, the use of "precision" instead of "information". To explain why the word precision is used, we need to dive into the nuances of variance and discover what it really represents.

Suppose you are asked how likely it will rain tomorrow. If you live in the UK, perhaps your answer would be 80% as you think it will rain more likely than not, but how would we explain this thought process mathematically? Human guesses can be modelled with the Beta distribution due to three characteristics. First, it is defined on the interval [0, 1] which represents the range of probability. Second, it can be skewed right or left depending on whether an event is likely or unlikely. Third, the variance can be increased or reduced depending on the strength of belief. To illustrate how variance ties with belief, consider the following example of Sarah and John. Sarah thinks the probability it will rain follows a Beta distribution with mean 0.5 and variance 0.5. John also assumes a Beta distribution but with mean 0.8 and variance 0.1. Anyone with an idea of variance can tell you that John is more **precise** (confident) about his guess than Sarah. The two distributions are illustrated to make this even clearer.



Thus, we can conclude that the word precise is used because n* represents the size of an iid sample that has the same variance as our dataset. That covers the basics of this intriguing topic, but I encourage you to think about correlation structures and their effect on the value of n*. For example, when correlation is constant and negative, the effective sample size may be larger than the original sample size! It is amazing that this is possible as intuitively, one might assume any correlation among our observations will lead to lower precision since the existence of correlation implies a linear pattern. However, this is not the case and most of the time correlation is not even constant!