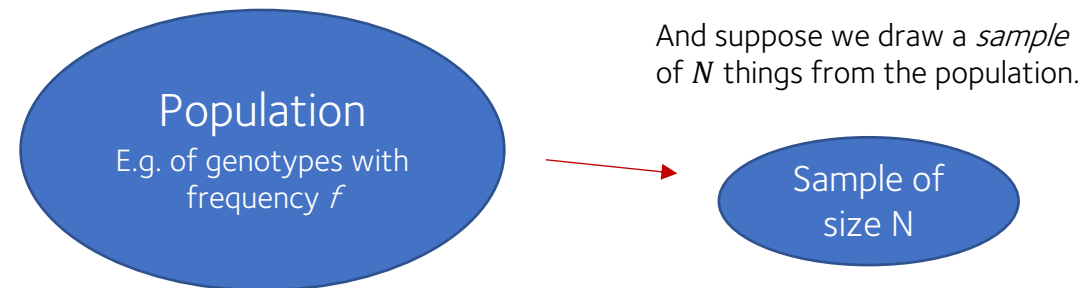


# Sampling and asymptotics cheatsheet

Gavin Band, [WHG GMS Programme](#) 2021

Suppose we have a big bag of things – for example, a population of people, with a particular genotype  $G$  that occurs at frequency  $f$ :



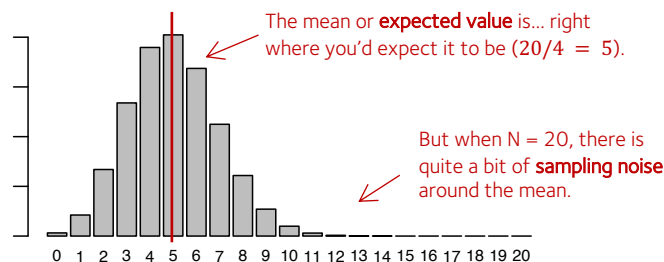
The number of  $G$  genotypes in our sample depends on what sample we drew (i.e. it is a “random variable”) – it would vary from one sample to the next. How does it vary?

It turns out that – if we made sure not to sample the same person twice (“**sampling without replacement**”), then the number has a “**hypergeometric distribution**”. This is actually a bit annoying because that distribution is a bit tricky to work with – it depends on knowing the full population size, and it makes the samples not independent of each other. However, *if our population is very large and the sample is much smaller* then we will never sample the same thing twice anyway. We might as well imagine we are **sampling with replacement** instead – allowing us to model using the much simpler **binomial distribution**.

number of  $G$  genotypes in the sample  $\sim \text{binomial}(N, f)$  ← Only depends on the sample size and the true frequency  $f$

This is the situation we’re often in in genetics – we have a small sample from a large population, and we would like to make statements about the population by looking at the sample.

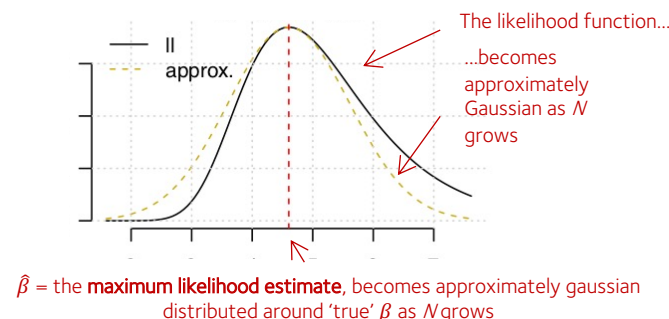
For example, here is the distribution of a sample of size 20 when the population frequency  $f = 25\%$ .



The central limit theorem implies that many likelihood functions ‘become gaussian’ as the amount of data grows. Specifically:

1. the likelihood function will approximate a Gaussian density (up to a constant) as  $N \rightarrow \infty$ .

And 2. the location of the likelihood function itself will become approximately Gaussian around the ‘true’ value as  $N \rightarrow \infty$ .



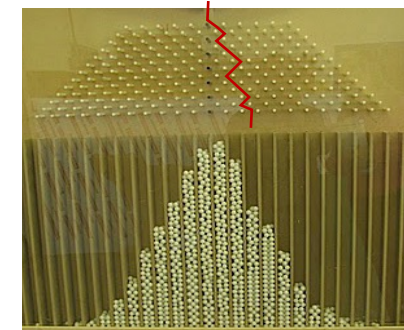
This is referred to as ‘asymptotic local normality’ and ‘Le Cam’ theory. For them to work the likelihood must be smooth, the ‘true  $\beta$ ’ should be in the interior of parameter space, and the data should have some level of independence.

Here is another way to think of sampling – via a [Galton board](#):

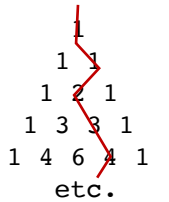
We drop marbles in at the top

At each level the ball ‘samples’ either a left or a right.

After a while the marbles draw a binomial distribution:

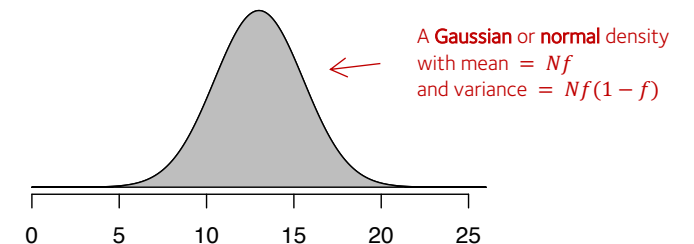


Number of possible routes is given by Pascal's triangle:



The number of possible routes ending up in each box is given by **Pascal's triangle** – in other words, they are the **binomial coefficients**. Hence the name!

The most crucial fact in all of statistics is that, as we let the number of samples get large, the distribution above becomes approximately **Gaussian**.



This famous fact is known as the **central limit theorem**. An important consequence is the asymptotic theory of likelihoods explained below.

This is in fact how tools like `glm()` compute their output – here’s an example:

```
> fit = glm( Y ~ X, data = data, family = "binomial" )
> summary(fit)$coefficient
      Estimate Std. Error  z value    Pr(>|z|)
(Intercept) -1.1155710  0.2409409 -4.630061 3.655588e-06
X            -0.2491656  0.2115801 -1.177642 2.389394e-01
```

$\hat{\beta}$  (the maximum likelihood estimate or MLE)      se (the standard error of the MLE, estimated from the likelihood curvature)       $z = \beta/\text{se}$  (the z-score)       $P = \text{pnorm}(-|z|) \times 2$  (the P-value)

Interpretation: our estimate of  $\beta$  is -0.25, but since  $P = 0.23$  this is consistent with a true effect of zero and a nonzero observed effect due to random sampling of the outcome variable  $Y$ .

Moreover 3. these gaussians have the same variance!

Likelihood function:

$$P(\text{data}|\beta = x) \propto N\left(x; \hat{\beta}, \frac{I}{N}\right)$$

MLE estimate distribution:

$$\hat{\beta} \sim N\left(\text{true } \beta, \frac{I}{N}\right)$$

These have the same variance, scaling like  $1/N$ . (Here  $I$  is some value that doesn't depend on the data or parameters.)

**Conclusion 1:** often we only need to report the maximum likelihood estimate and its std. error.

**Conclusion 2:** the standard error can be read off from the likelihood function. (This is indeed how standard errors and P-values in functions like `lm()` and `glm()` are often computed.)