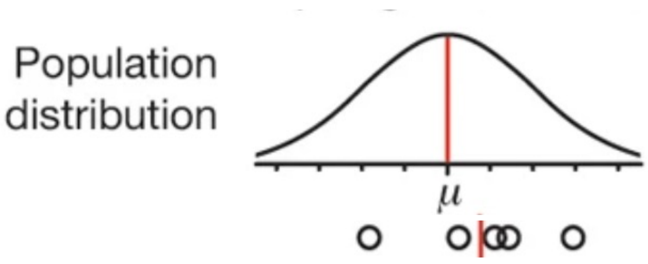
Bootstrapping

MCDB 170/270

Statistical testing

T-test

1. We sample from a population



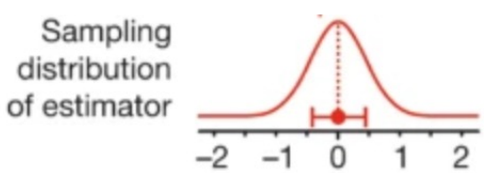
Gaussian distribution

2. We define a test statistics, such as a mean

$$\bar{X} = \frac{\sum_i x_i}{n}$$

Mean

3. Then we want to test this statistics for something (e.g. Is it significantly different from zero?) → sampling distribution



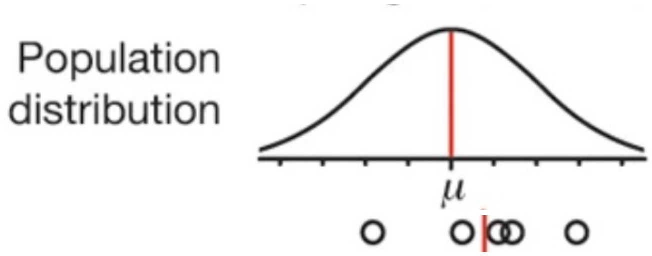
Student's t-distribution

Statistical testing

I.E., What if we don't have a pre-formulated test?

1. We sample from a population

What if we don't know the population distribution?

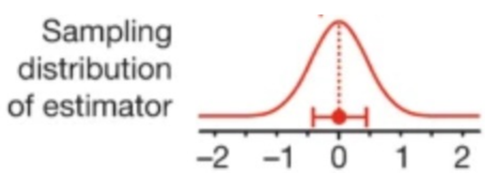


2. We define a test statistics, such as a mean

$$\bar{X} = \frac{\sum_i x_i}{n}$$

What if our statistics is not trivial (i.e. not a simple mean)?

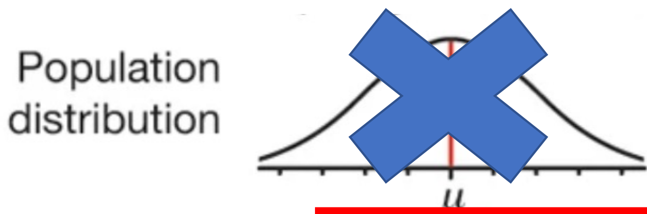
3. Then we want to test this statistics for something (e.g. Is it significantly different from zero?) → sampling distribution

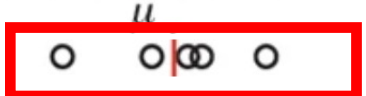


What if we don't know the sampling distribution?

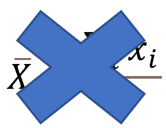
Bootstrapping (Efron 1978)

1. We sample from a population



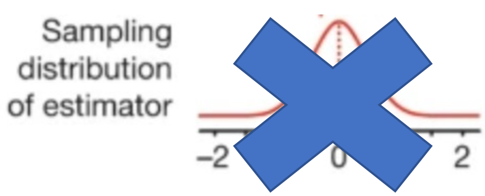


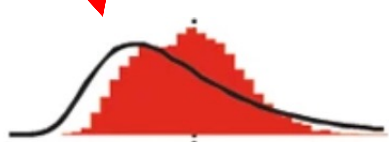
2. We define a test statistics, such as a mean



$$\boxed{S = f(x_i)}$$

3. Then we want to test this statistics for something (e.g. Is it significantly different from zero?) → sampling distribution



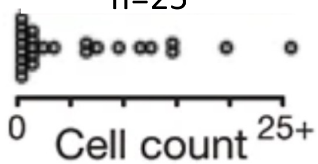


Example: Cell counting (see the nature article for description)

https://doi.org/10.1038/nmeth.3414

1. Samples

n=25



2. Statistics

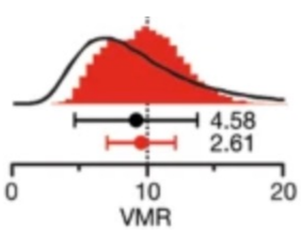
$$VMR = \frac{var(x)}{mean(x)}$$

3. Test if VMR > 1

Randomly sample ‘n’ times (n=25)
with replacement

Calculate VMR:
This is a point in the
estimated sampling distribution

Repeat 5,000 times



Bootstrapping

The most important assumption is that **our data well represent the true population distribution**.

→ To this end, we need as many data as possible.

→ 5 is never enough. 10 is OK. 15 is good. 20 is great. 50 is even better. But more than 50 may not be necessary.

The 'shape' (dimension, size, etc) of each sampling must be identical to the original data.

In general, less sensitive than standard tests because bootstrapping uses less assumptions (i.e., less information).

What about parametric bootstrapping?

- Nonparametric bootstrapping: Basically, the data solely represent the population distribution.
- Parametric bootstrapping: We know the population distribution and use the data to estimate the parameters of the population distribution. Then we sample from this estimated distribution. It is generally more sensitive compared to the nonparametric bootstrapping because of the additional information we assume. (We can implement the t-test using parametric bootstrapping by estimating the mean and variance of the gaussian distribution using the data.)

Cf) Bootstrap methods can be considered as a special case of another widely used statistical technique called "Monta Carlo" simulation.