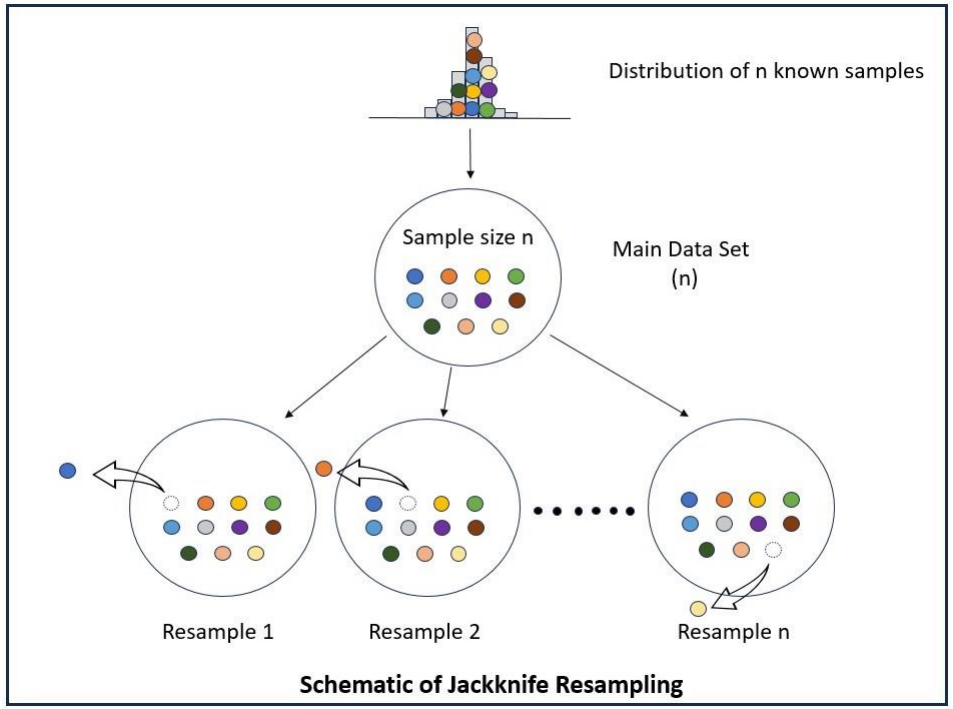
Re-sampling methods

- Jackknife Method
- Bootstrap Method (more general)

How to quantify uncertainties for very complex models or data?

Jackknife Method



https://en.wikipedia.org/wiki/Jackknife_resampling

Simple method that can be used to estimate linear regression parameters and their uncertainties. It is an effective resampling method to estimate and reduce the bias of parameter estimates.

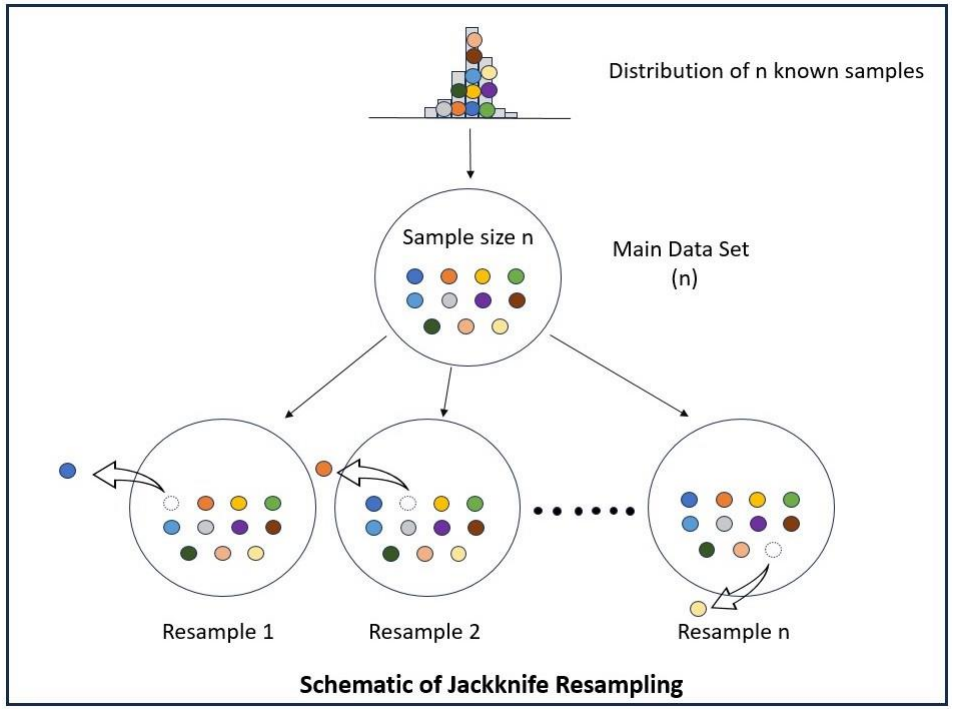
Procedure:

- Generate a re-sampled dataset Z_j, obtained by deleting the *j*th element from the dataset. Each of the *N* re-sampled dataset has therefore dimension *N-1*.

- Each dataset Z_j is used to estimate the parameters of interest in the same way as in the original dataset Z. For example, if the method is used to estimate the two parameters of the linear regression, the re-sampled dataset Z_j yields the best-fit values $\hat{\theta}_j \equiv \{a_j, b_j\}$ (intercept and slope).

- The parameters of interest are also calculated for the full-dimensional dataset Z. The best-fit parameters from Z yields $\hat{\theta} \equiv \{a, b\}$.

Jackknife Method



https://en.wikipedia.org/wiki/Jackknife_resampling

- For each dataset Z_j, we can define and obtain the *pseudo-values* $\theta_j^* = N\hat{\theta} - (N-1)\hat{\theta}_j$.

- The jackknife estimator of the parameters of interest and their uncertainties are given by the following equations:

$$\bar{\theta}^* = \frac{1}{N}\sum_{j=1}^{N}\theta_j^*$$

$$\sigma_{\theta^*}^2 = \frac{1}{N(N-1)}\sum_{j=1}^{N}\left(\theta_j^* - \bar{\theta}^*\right)^2$$

Limitations:

JM can fail if the estimator is not smooth, i.e., the parameters being estimated change abruptly or shows discontinuities when one of the observations is left out (e.g., median can vary significantly when one observation is removed).

JM is sensitive to data with outliers or heavily skewed.

Jackknife Method

It is possible to show that the jackknife method is usually an unbiased estimator for the linear regression in the asymptotic limit of sufficiently large number of samples, although the variance of the estimator may be larger than that of the least-square estimate.

Wu, C. F. J.: Jackknife, bootstrap and other resampling methods in regression analysis. Ann. Stat. 14, 1261-1295 (1986).

Limitations:

- Observations are assumed to be “iid” (independent and identically distributed). Jackknife is not ideal for time series.

- Works ok for statistics which are linear functions of the parameters or the data, and whose distribution is continuous (or “smooth enough”).

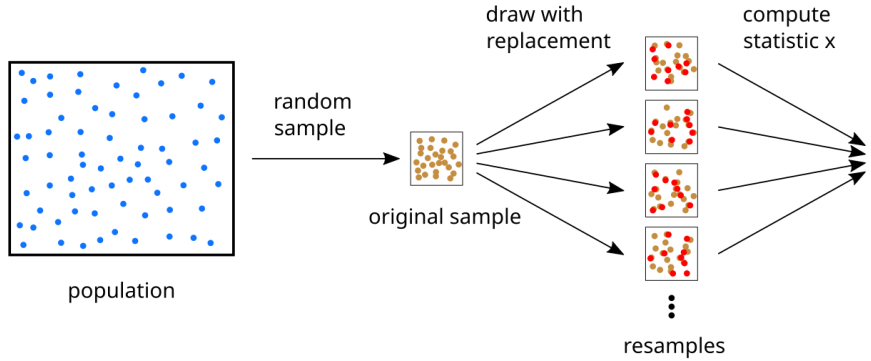
Abdi, H. and Williams, L.J. (2010) Jackknife. In: Salkind, N. and Frey, B., Eds., Encyclopedia of Research Design, Sage, Thousand Oaks.

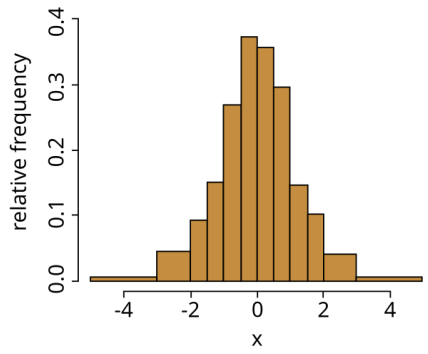
Bootstrap Method

Bootstrapping is any test or metric that uses **random sampling with replacement** (e.g., mimicking the sampling process) and falls under the broader class of resampling methods. Bootstrapping assigns measures of accuracy (bias, variance, confidence intervals, prediction error, etc.) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods.

Procedure:

- Draw at random *N* data points from the original dataset Z, **with replacement (allowing duplicates)**, to form a synthetic bootstrap dataset Z_i. The new dataset, therefore, has the same dimension as the original dataset, but a few of the original points may be repeated and a few are missing. For example, suppose a study collects five data points and creates four bootstrap samples:





Original	Bootstrap1	Bootstrap2	Bootstrap3	Bootstrap4
1	1	2	1	1
2	1	3	2	1
3	3	3	3	1
4	3	3	5	4
5	5	4	5	5

https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

Bootstrap Method

- From the bootstrap dataset Z_i, calculate the parameters of interest θ_i (it can be mean, median, parameters for linear regression, etc.). For example, for the linear regression case, $\theta_i = \{a_i, b_i\}$ which can be calculated using maximum likelihood method or least-square.

- The two steps before are repeated a large number of times, say $i = 1, \dots, N_{boot}$.

- At the end of the process, the parameter values θ_i are used to approximate the sampling distribution of the parameters, and therefore obtain an estimate of the best-fit values and confidence intervals. We will be able to visualize the histograms of $\hat{\theta} = \{\hat{a}, \hat{b}\}$ and make inferences.

$$\bar{\theta} = \frac{1}{N_{boot}} \sum_{i=1}^{N_{boot}} \theta_i$$

$$s_\theta^2 = \frac{1}{N_{boot}} \sum_{i=1}^{N_{boot}} (\theta_i - \bar{\theta})^2$$

Such re-sampling methods can be used even when the errors on the data points are not available!