Re-sampling

Re-sampling - Introduction

We have relied on idealized models of the origins of our data ($\varepsilon \sim N$) to make inferences

But, these models can be inadequate

Re-sampling techniques allow us to base the analysis of a study solely on the design of that study, rather than on a poorly-fitting model

Why Re-sampling?

- Fewer assumptions
  - Ex: re-sampling methods do not require that distributions be Normal or that sample sizes be large
- Generality: Re-sampling methods are remarkably similar for a wide range of statistics and do not require new formulas for every statistic
- Promote understanding: Bootstrap procedures build intuition by providing concrete analogies to theoretical concepts

Re-sampling

Collection of procedures to make statistical inferences without relying on parametric assumptions

- Bias
- Variance, measures of error
- Parameter estimation
- Hypothesis testing

Error Estimation

- Error estimation is concerned with establishing whether the results we have obtained on a particular experiment are representative of the truth or whether they are meaningless.
- Traditionally, error estimation was performed using the classical parametric (and sometimes, non-parametric).
- More recently, however, new tests have emerged for error estimation, based on re-sampling methods, that have the advantage of not making distributional assumptions the way parametric tests do. The tradeoff, though, is that such tests require high computational power.

Traditional Statistical Methods versus Resampling Methods

- Classical parametric tests compare observed statistics to theoretical sampling distributions.
- Re-sampling makes statistical inferences based upon repeated sampling within the same sample.
- Re-sampling methods stem from Monte Carlo simulations, but differ from them in that they are based upon some real data; Monte Carlo simulations, on the other hand, could be based on completely hypothetical data.

Error Estimation through Resampling Techniques in Machine Learning

- Error estimation through re-sampling techniques is concerned with finding the best way to utilize the available data to assess the quality of our algorithms.

- In other words, we want to make sure that our classifiers are tested on a variety of instances, within our sample, presenting different types of properties, so that we don't mistaken good performance on one type of instances as good performance across the entire domain.

Resampling

With replacement

Without replacement

Re-Sampling Approaches

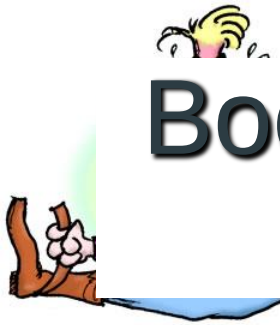
- Cross-validation
- Jackknife (Leave-one-out)
- Bootstrapping
- Randomization

Cross-Validation

- A sample is randomly divided into two or more subsets and test results are validated by comparing across sub-samples.
- The purpose of cross-validation is to find out whether the result is replicable or whether it is just a matter of random fluctuations.
- If the sample size is small, there is a chance that the results obtained are just artifacts of the sub-sample. In such cases, the jackknife procedure is preferred.

Jackknife

- In the Jackknife or Leave-One-Out approach, rather than splitting the data set into several subsamples, all but one sample is used for training and the testing is done on the remaining sample. This procedure is repeated for all the samples in the data set.
- The procedure is preferable to cross-validation when the distribution is widely dispersed or in the presence of extreme scores in the data set.
- The estimate produced by the Jackknife approach is less biased, in the two cases mentioned above than cross-validation.



Bootstrapping and Randomization: Main Ideas

- Bootstrapping makes the assumption that the sample is representative of the original distribution, and creates over a thousand bootstrapped samples by drawing, *with replacement,* from that pseudo-population.

- Randomization makes the same assumption, but, instead of drawing samples with replacement, it reorders (shuffles) the data systematically or randomly a thousand times or more. It calculates the appropriate test statistic on each reordering.

- Since shuffling the data amounts to sampling *without* replacement, the issue of replacement is one factor that differentiates the two approaches.

Bootsrapping

- One can understand the concept of bootstrapping by thinking of what can be done when not enough is known about the data.

- For example, let us assume that we don't know the standard error of the difference of medians. One solution consists of drawing many pairs of samples, calculating and recording, for each of these pairs of samples, the difference between the medians, and outputting the standard deviation of these differences in lieu of the standard error of the difference of medians.

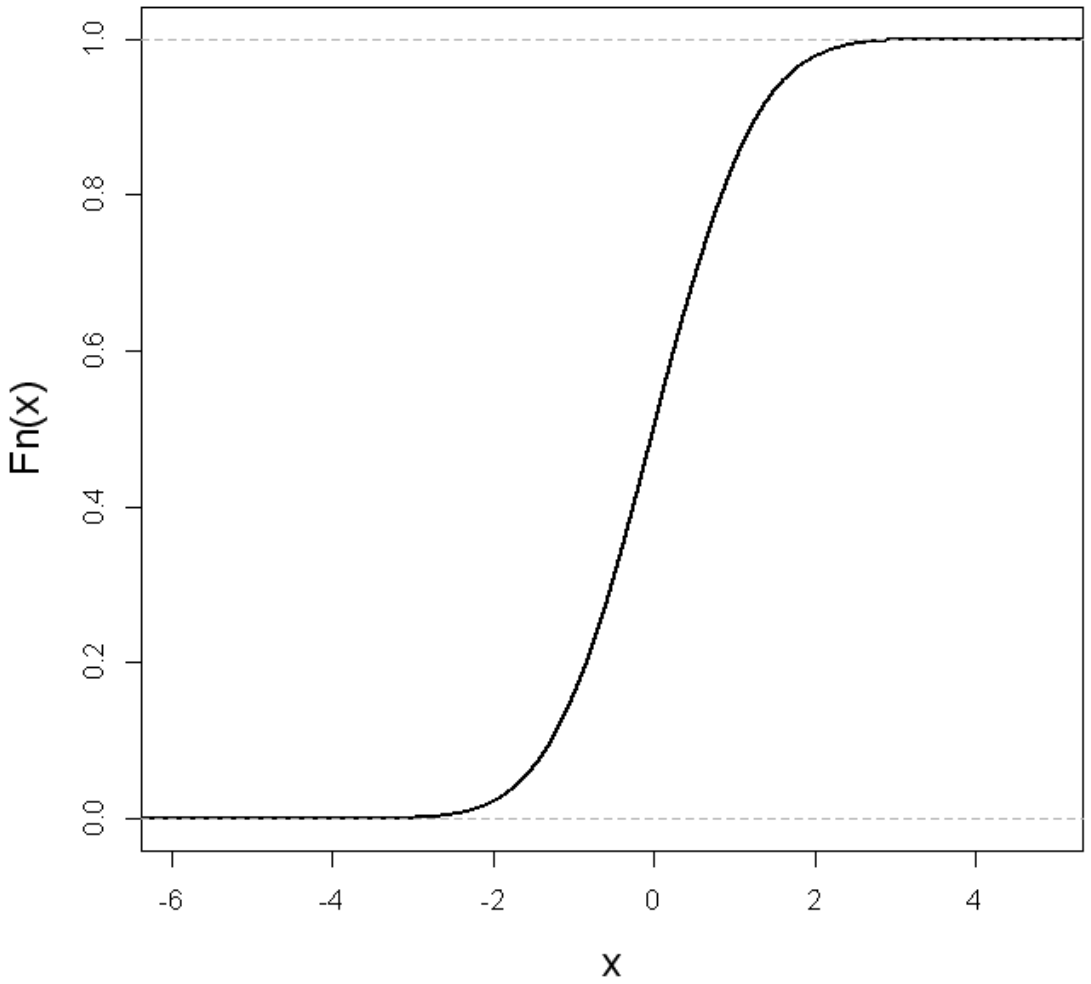
- In other words, bootstrapping consists of using an empirical, brute-force solution when no analytical solution is available.

- Bootstrapping is also very useful in cases where the sample is too small for techniques such as cross-validation or leave-one out to provide a good estimate, due to the large variance a small sample will cause in such procedures.

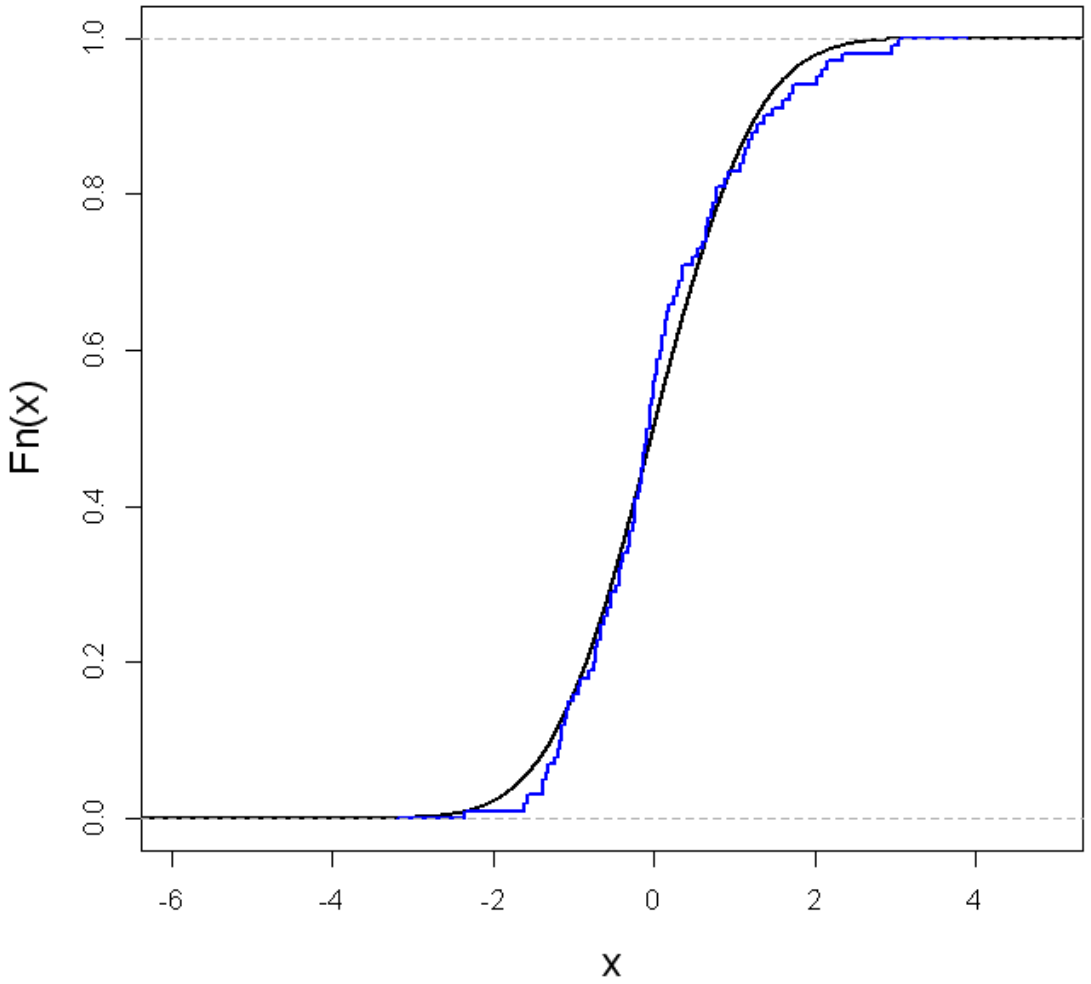
Resampling Bootstrap

- Hypothesis testing, parameter estimation, assigning measures of accuracy to sample estimates e.g.: se, CI
- Useful when:
  - formulas for parameter estimates are based on assumptions that are not met
  - computational formulas only valid for large samples
  - computational formulas do not exist
- Assume that sample is representative of population
- Approximate the distribution of the population by repeatedly resampling (with replacement) from the sample

Resampling Bootstrap



Resampling Bootstrap



Resampling Bootstrap

Non-parametric bootstrap

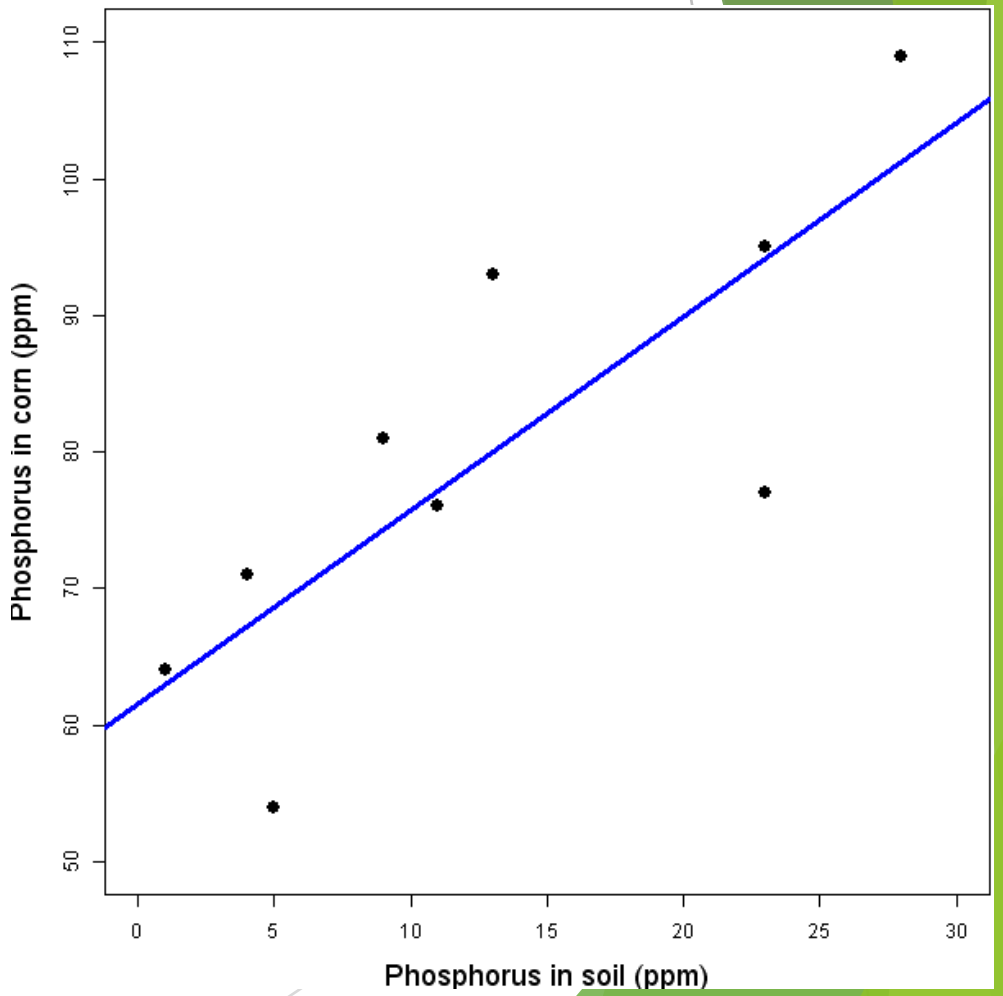
resample observation from original samples

Parametric bootstrap

fit a particular model to the data and then use this model to produce bootstrap samples

Confidence intervals Parametric Bootstrap

```
> preg
   psoil pcorn
1      1    64
2      4    71
3      5    54
4      9    81
5     13    93
6     11    76
7     23    77
8     23    95
9     28   109
> 
```



Confidence intervals Parametric Bootstrap

```
> preg
   psoil pcorn   fits    res
1      1    64  63.00   1.00
2      4    71  67.25   3.75
3      5    54  68.66 -14.66
4      9    81  74.33   6.67
5     13    93  80.00  13.00
6     11    76  77.17  -1.17
7     23    77  94.17 -17.17
8     23    95  94.17   0.83
9     28   109 101.25   7.75
>
```

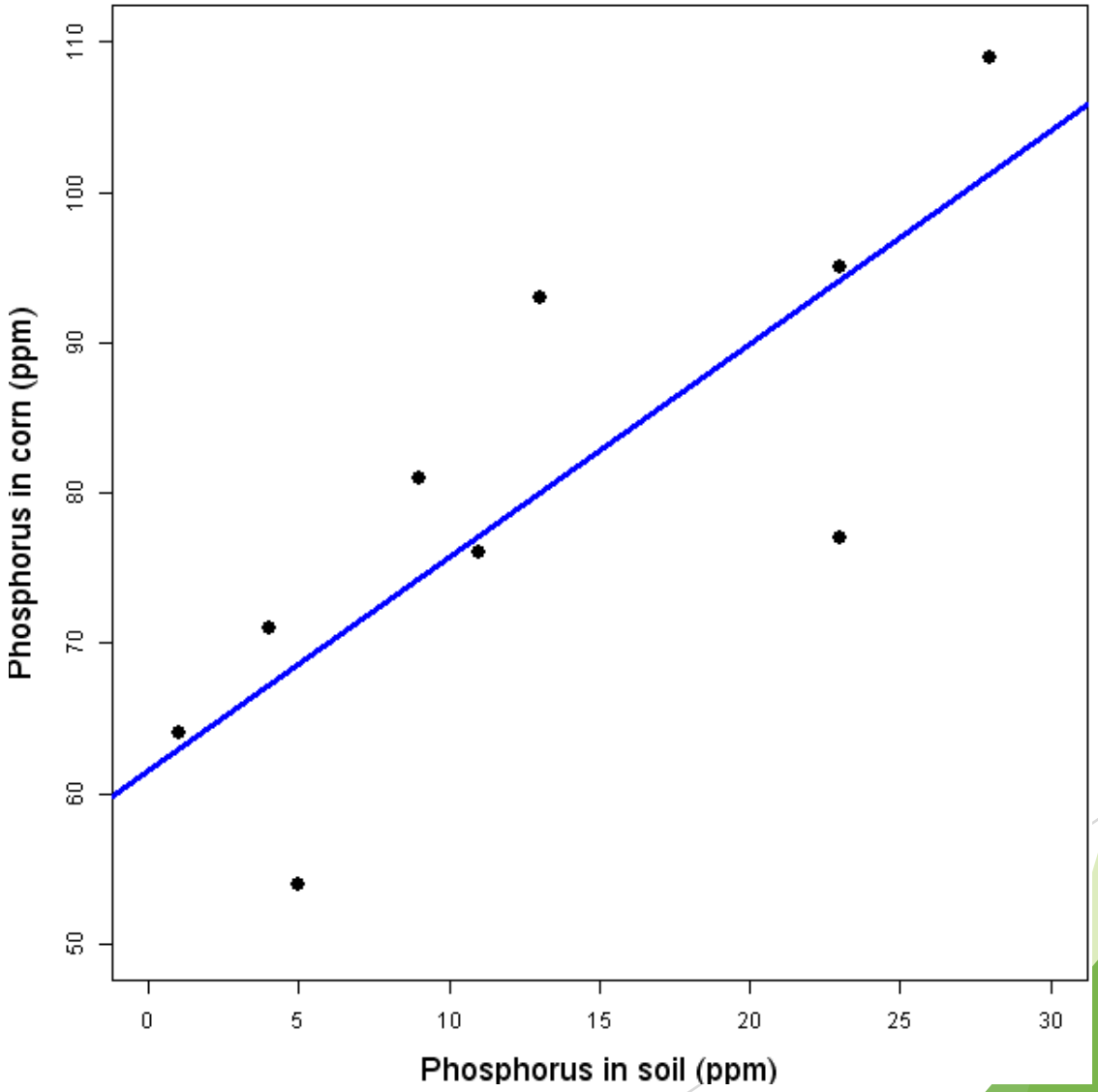
Confidence intervals Parametric Bootstrap

```
> preg
   psoil pcorn   fits     res resboot
1      1    64  63.00    1.00    6.67
2      4    71  67.25    3.75    7.75
3      5    54  68.66  -14.66    1.00
4      9    81  74.33    6.67    7.75
5     13    93  80.00   13.00    3.75
6     11    76  77.17   -1.17    6.67
7     23    77  94.17  -17.17  -17.17
8     23    95  94.17    0.83    0.83
9     28   109 101.25    7.75  -14.66
>
```

Confidence intervals Parametric Bootstrap

```
> preg
   psoil pcorn   fits     res resboot pcornboot
1      1    64  63.00    1.00    6.67     69.67
2      4    71  67.25    3.75    7.75     75.00
3      5    54  68.66  -14.66    1.00     69.66
4      9    81  74.33    6.67    7.75     82.08
5     13    93  80.00   13.00    3.75     83.75
6     11    76  77.17   -1.17    6.67     83.84
7     23    77  94.17  -17.17  -17.17     77.00
8     23    95  94.17    0.83    0.83     95.00
9     28   109 101.25    7.75  -14.66     86.59
>
```

Confidence intervals Parametric Bootstrap

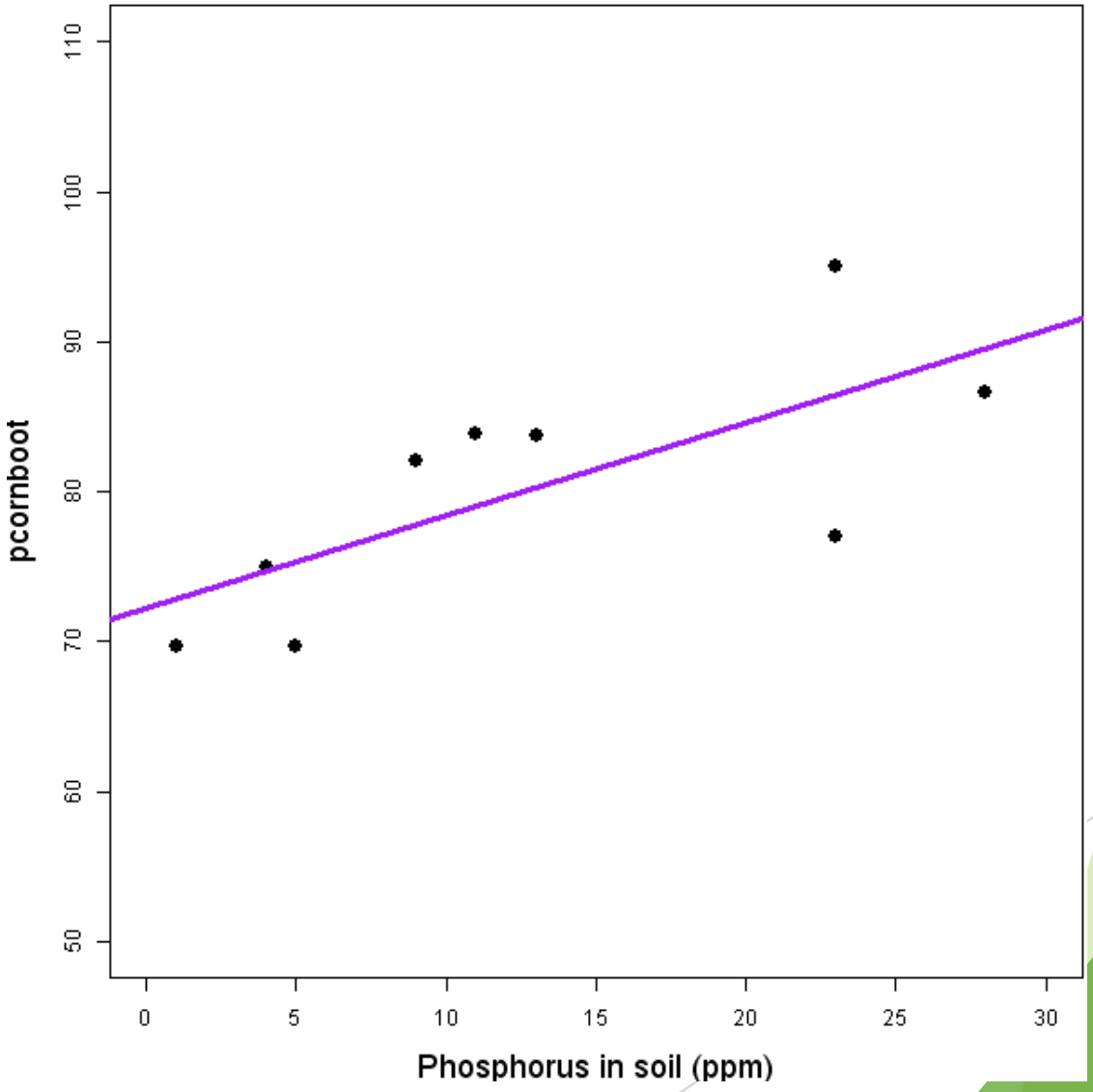


Params

β

α

Confidence intervals Parametric Bootstrap

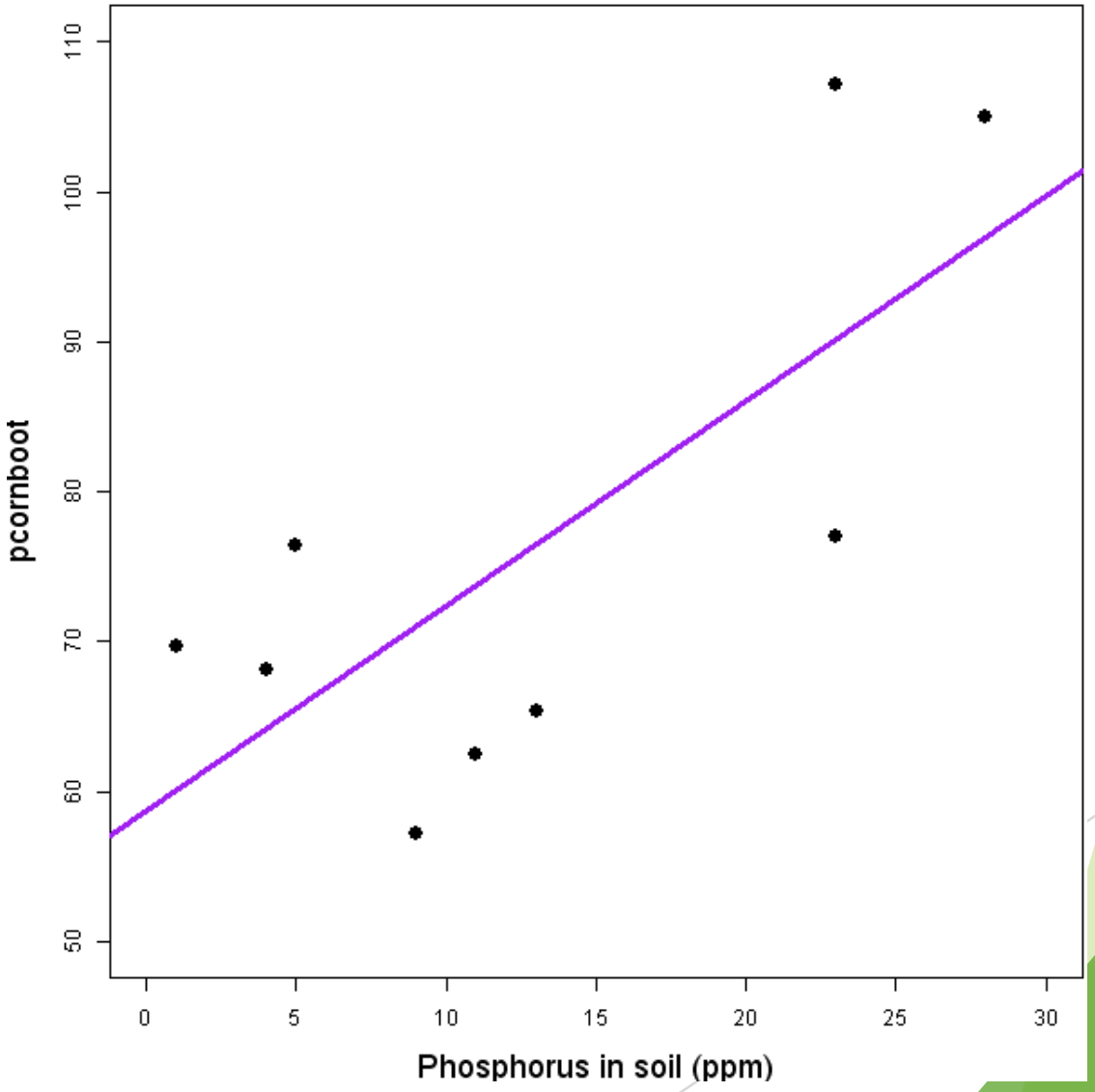


Params

β^*_1

α^*_1

Confidence intervals Parametric Bootstrap

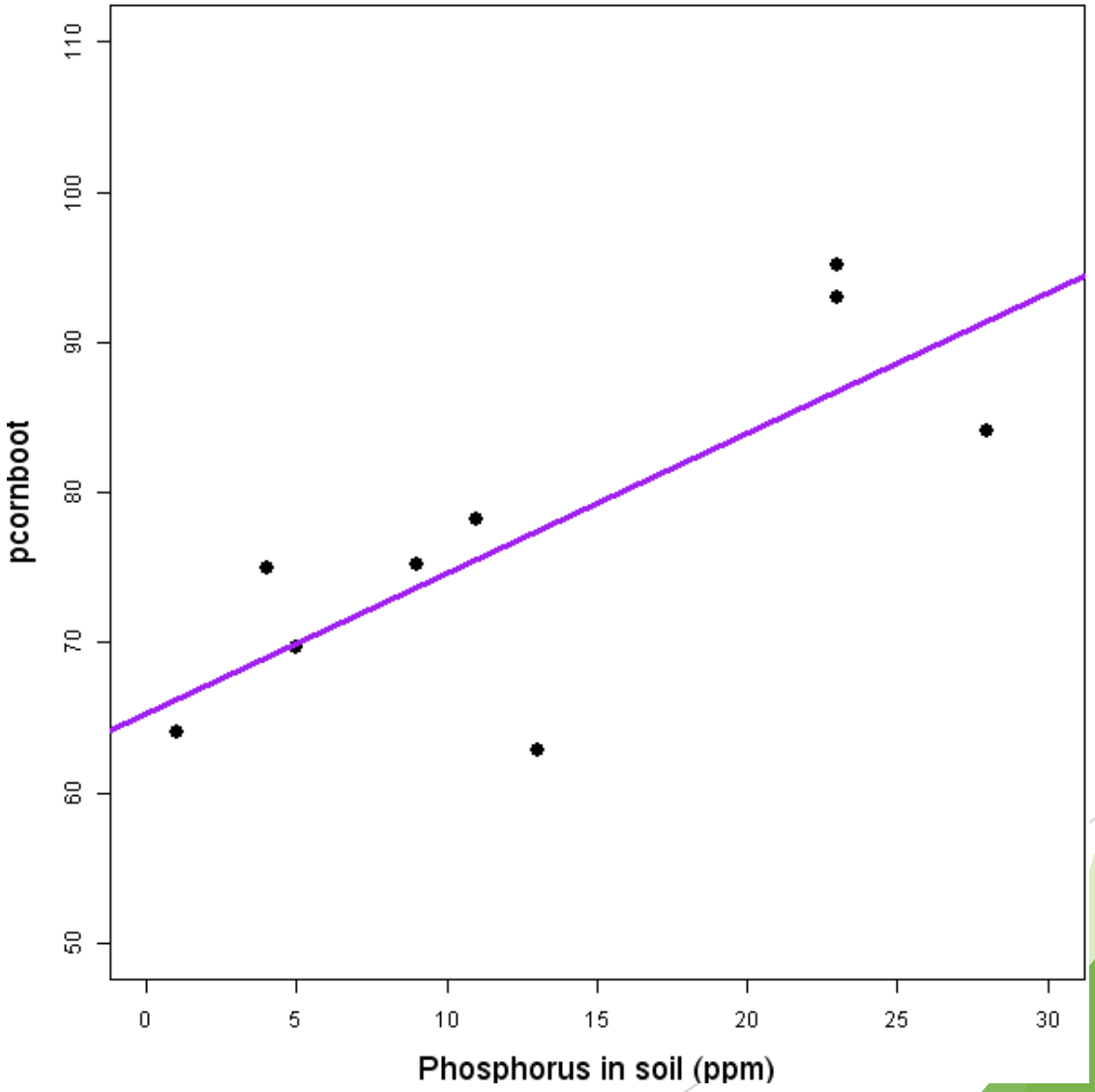


Params

β^*_2

α^*_2

Confidence intervals Parametric Bootstrap



Params

β^*_{nboot}

α^*_{nboot}

Bootstrap Caveat

Independence

Incomplete data

Outliers

Cases where small perturbations to the data-generating process produce big swings in the sampling distribution

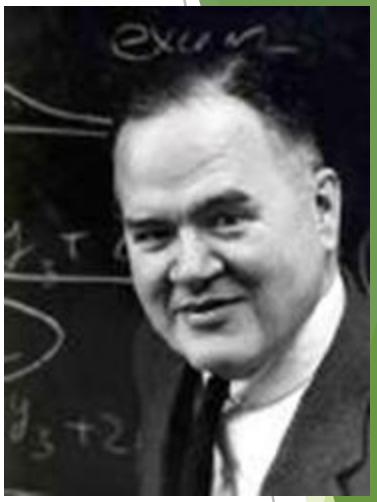
Resampling Jackknife

NOTES ON BIAS IN ESTIMATION

By M. H. QUENOUILLE

Research Techniques Unit, London School of Economics and Political Science

Quenouille 1956



Tukey 1958

Estimate bias and variance of a statistic

Concept: Leave one observation out and recompute statistic

Cross-validation Jackknife

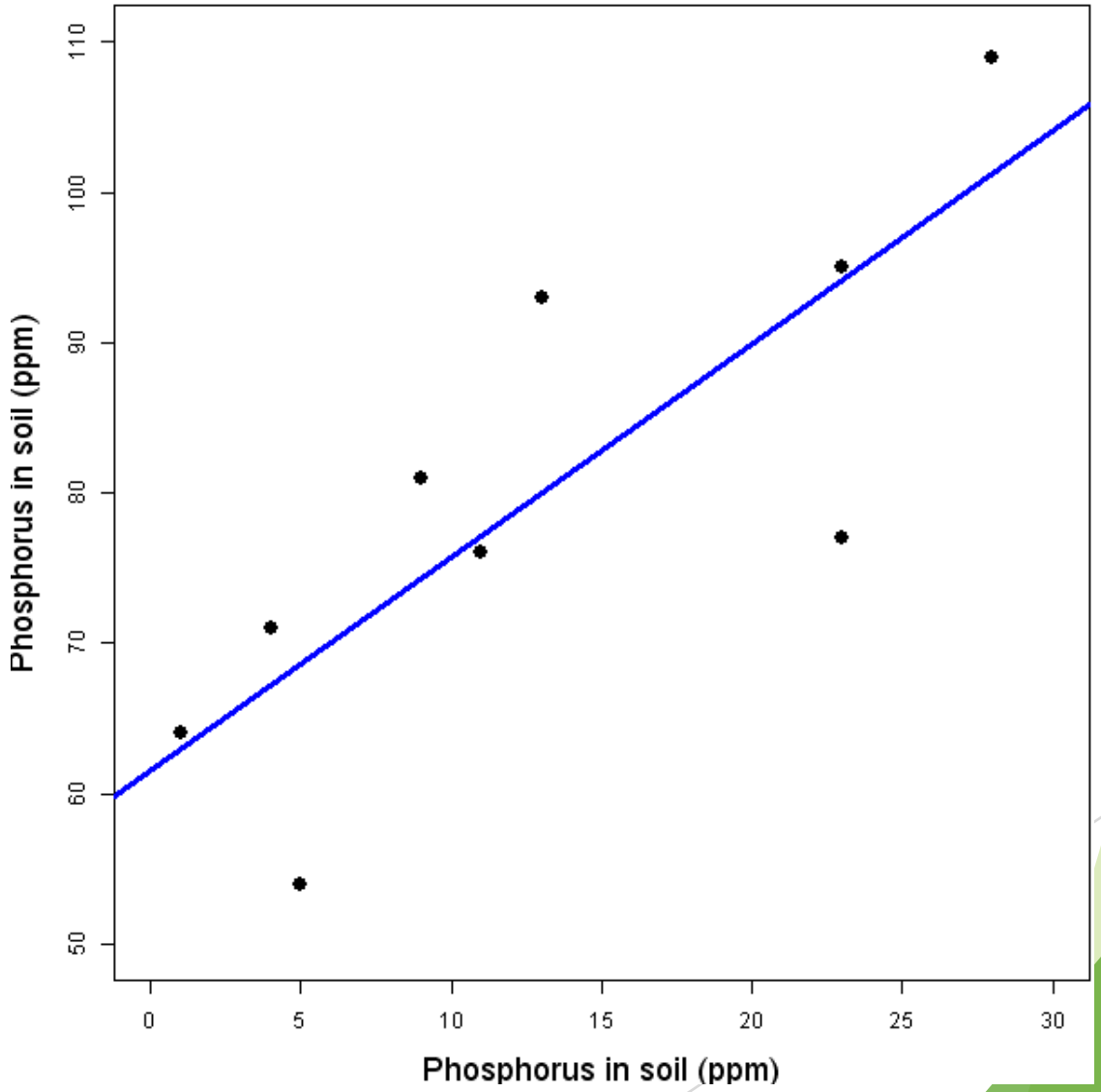
Assess the performance of the model

How accurately will the model predict a new observation?

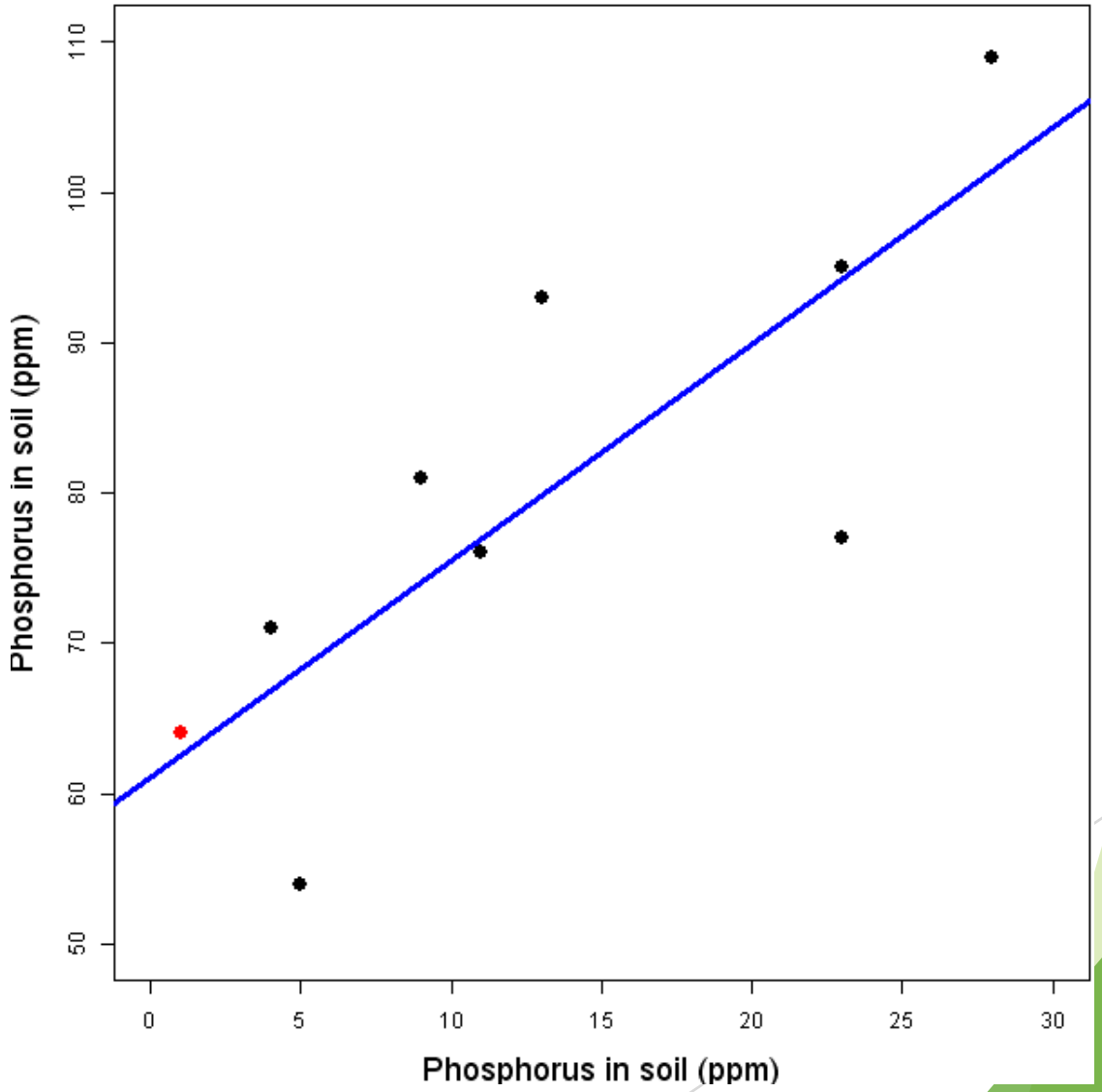
Cross-validation Jackknife

```
> preg
  psoil pcorn
1     1    64
2     4    71
3     5    54
4     9    81
5    13    93
6    11    76
7    23    77
8    23    95
9    28   109
> 
```

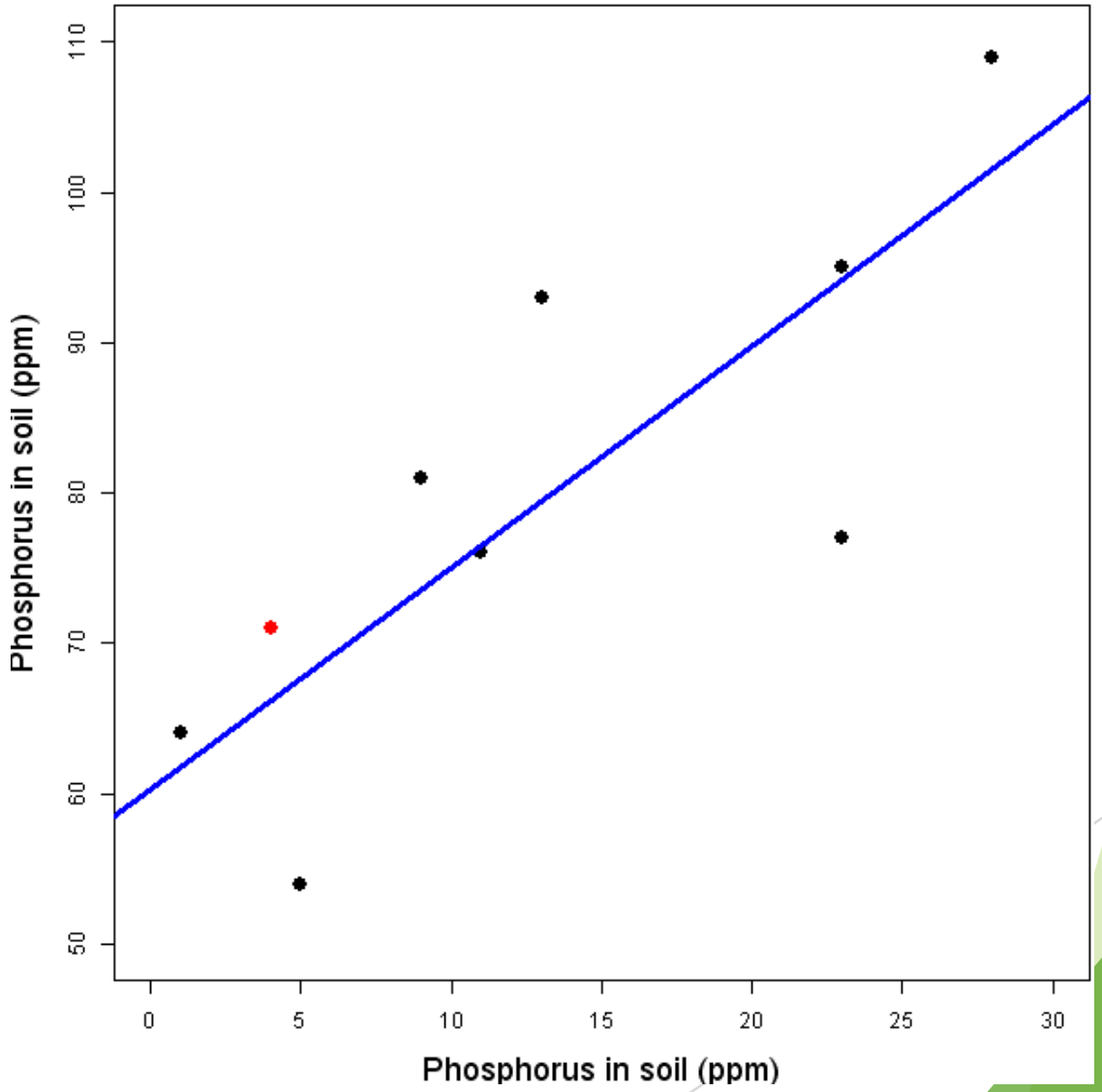
Cross-validation Jackknife



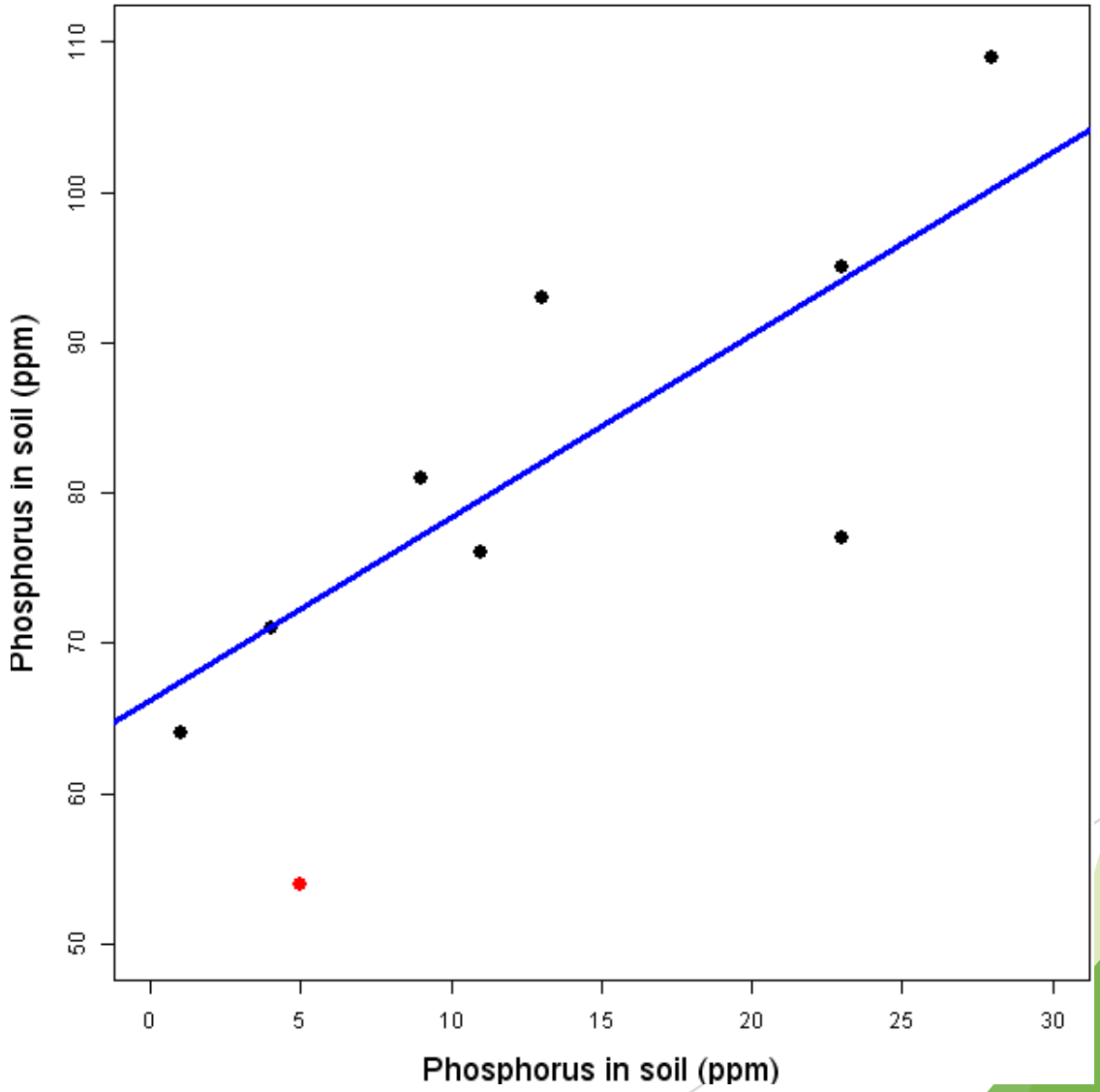
Cross-validation Jackknife



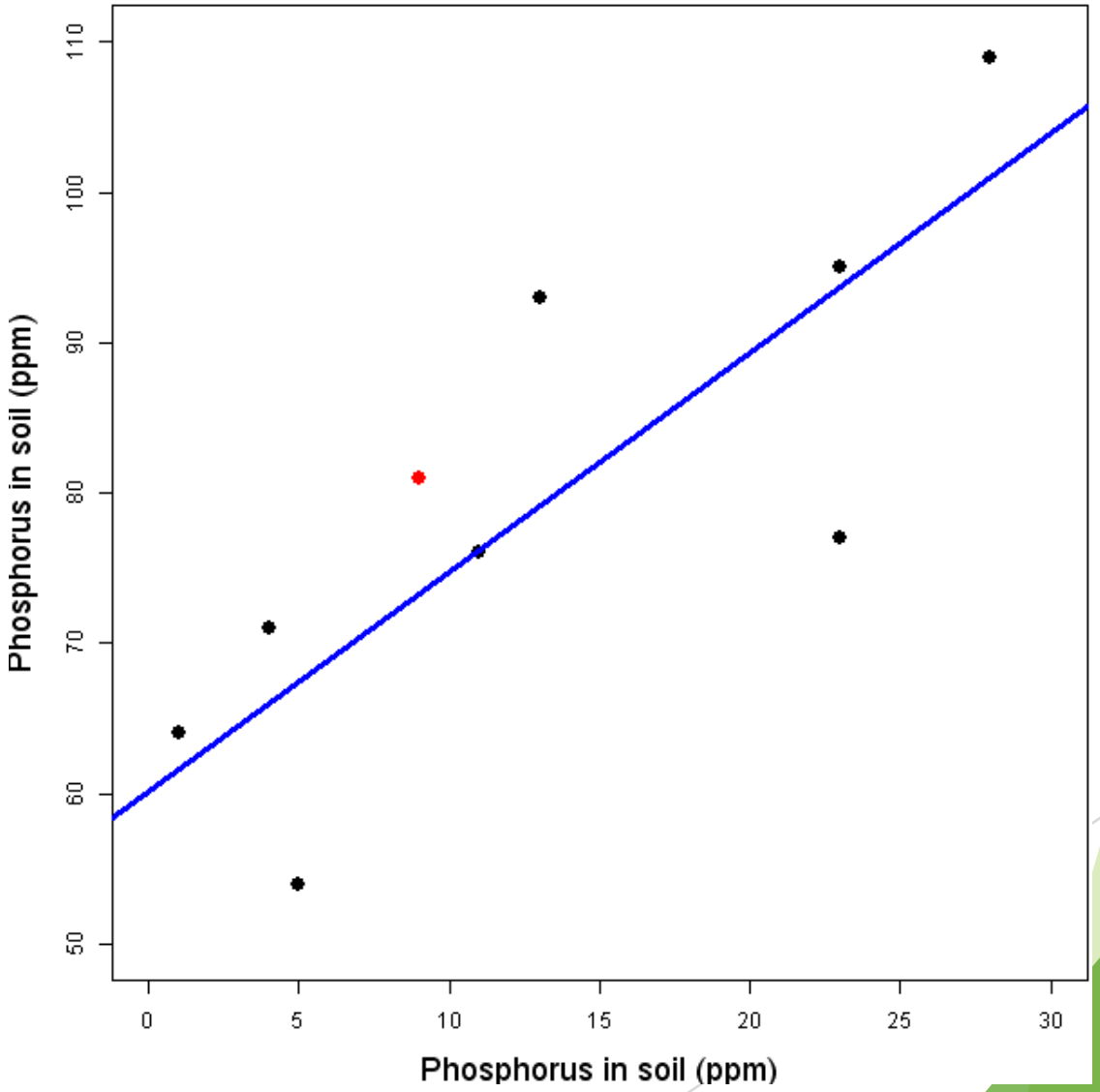
Cross-validation Jackknife



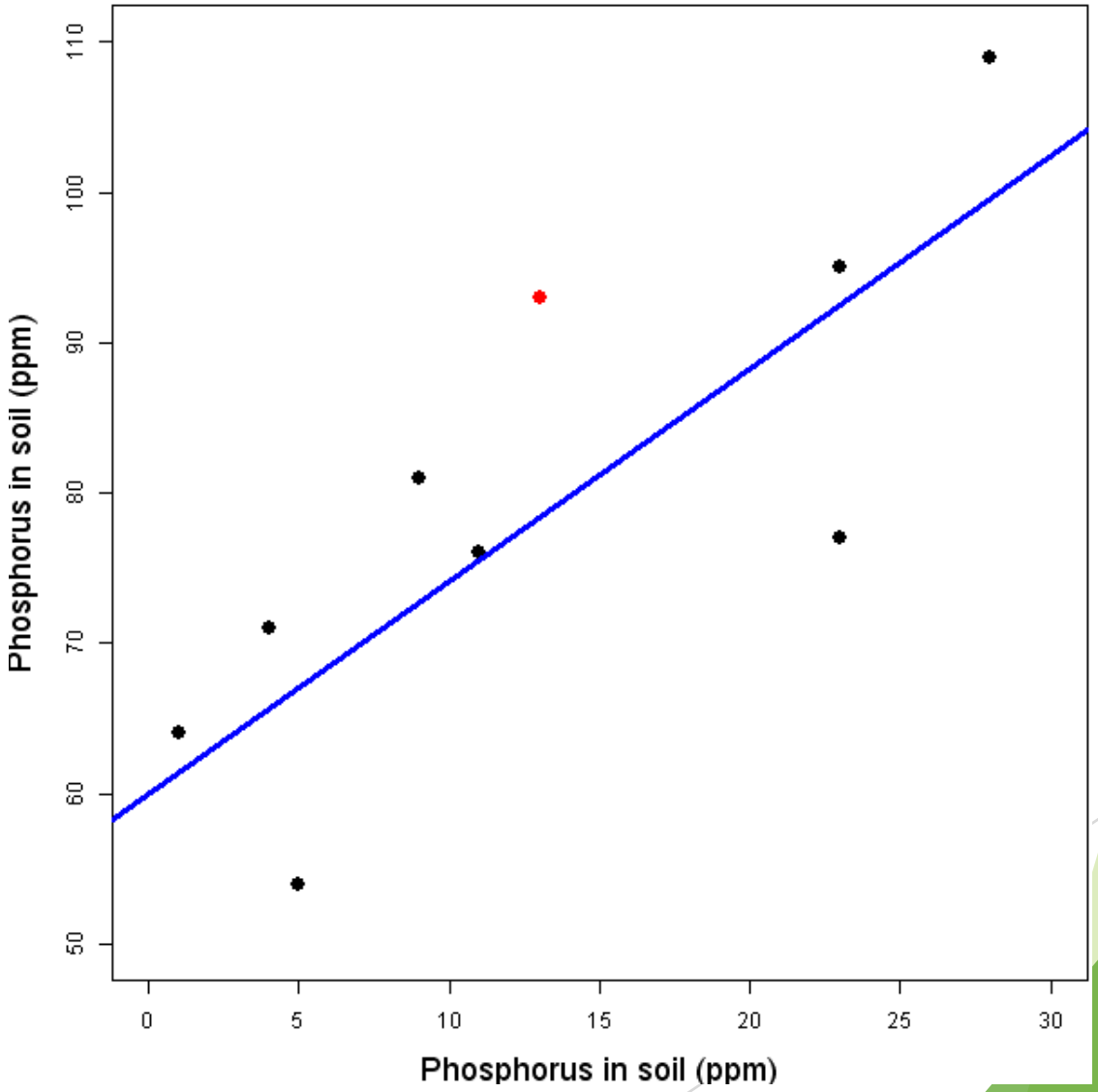
Cross-validation Jackknife



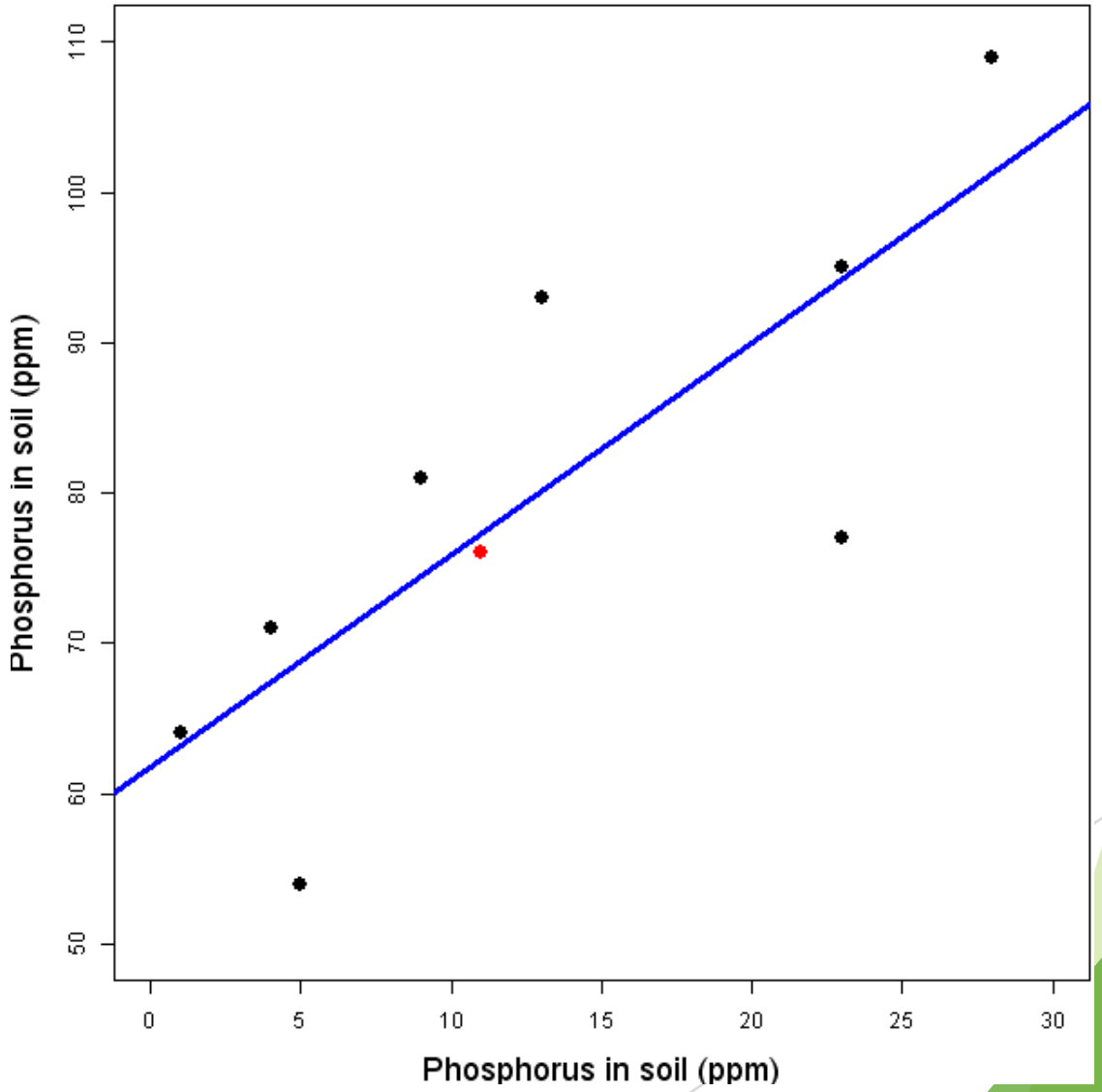
Cross-validation Jackknife



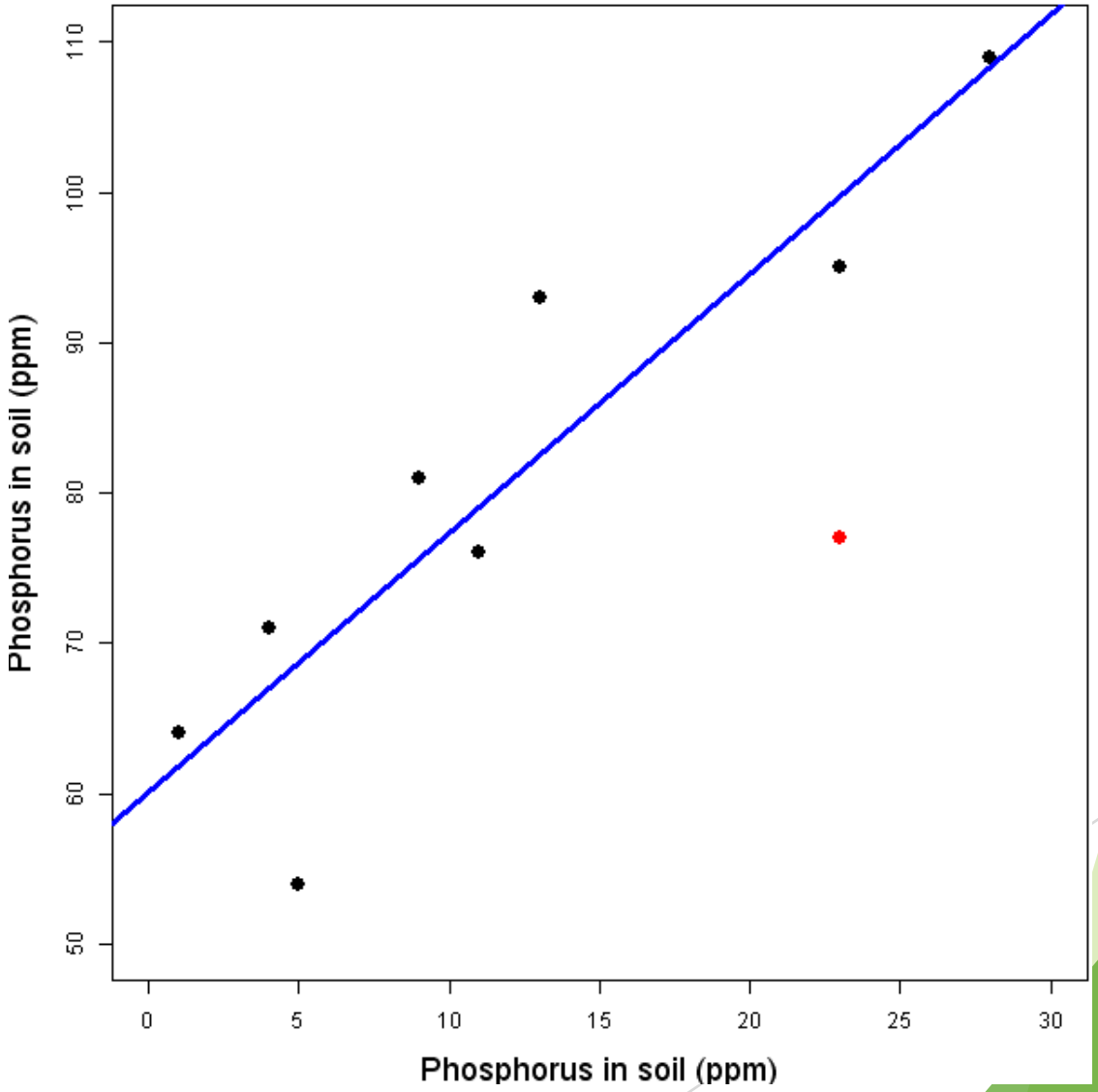
Cross-validation Jackknife



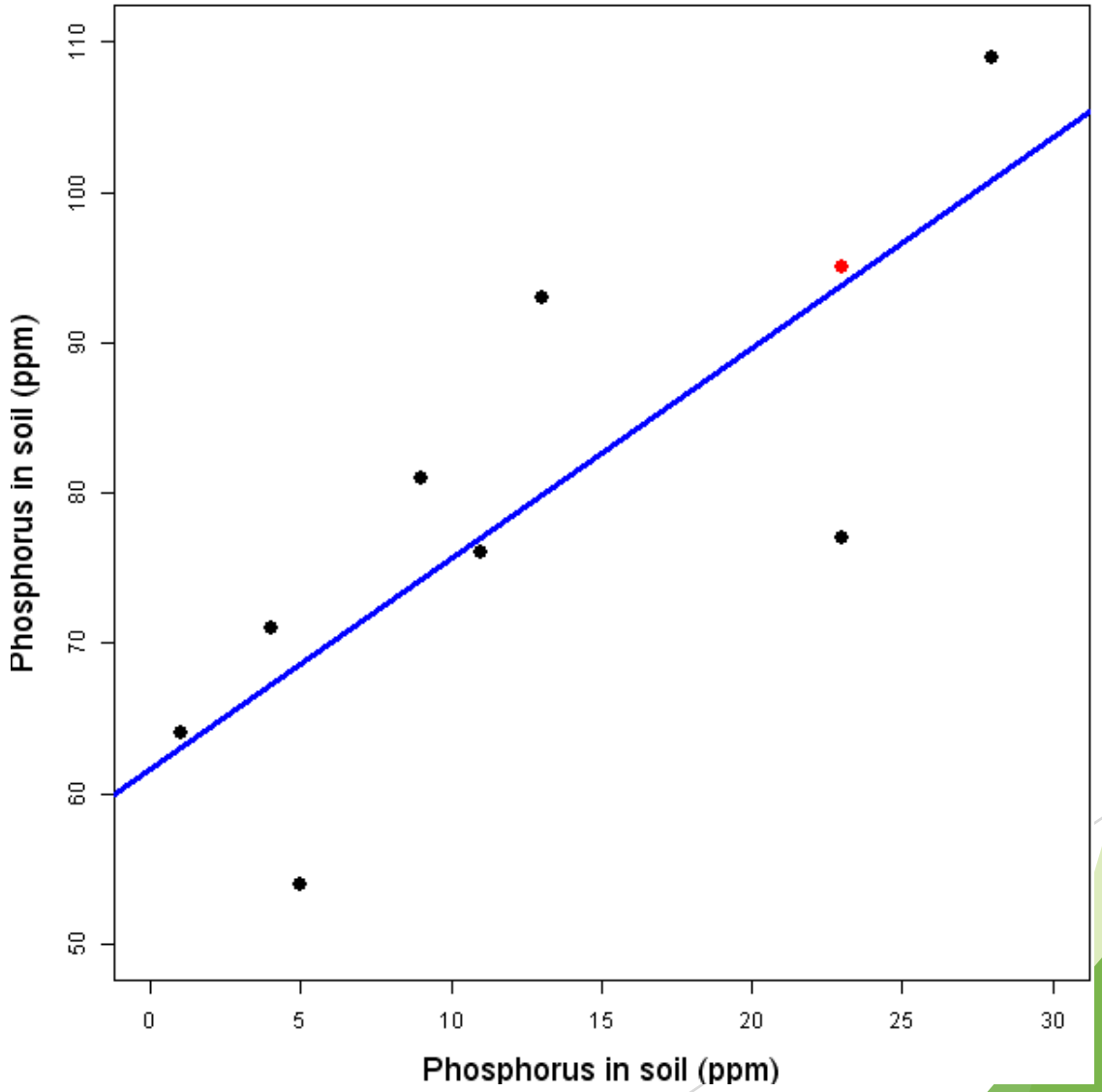
Cross-validation Jackknife



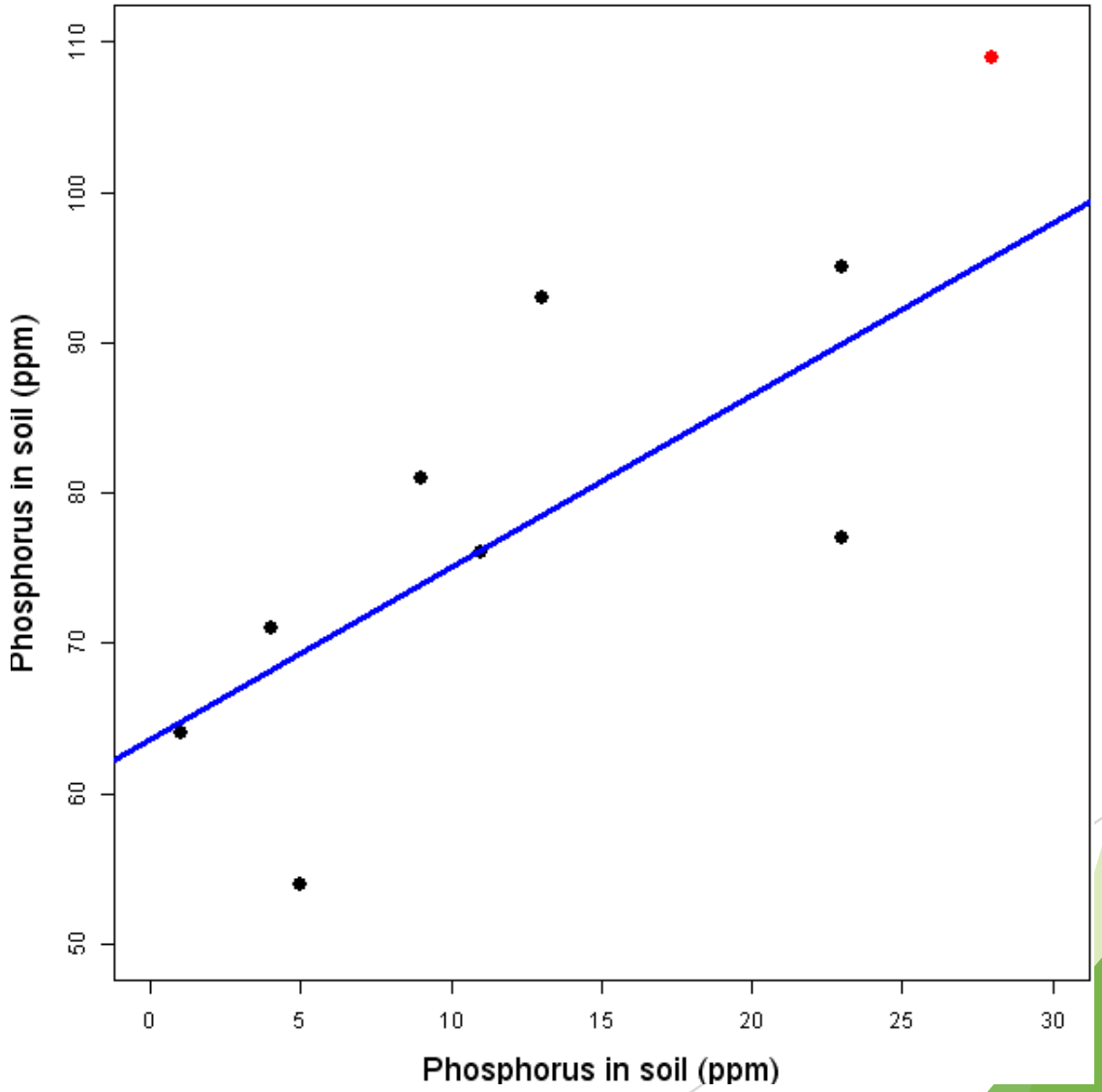
Cross-validation Jackknife



Cross-validation Jackknife



Cross-validation Jackknife



Cross-validation Jackknife

```
> preg
  psoil pcorn jackpred
1     1    64    62.55
2     4    71    66.18
3     5    54    72.29
4     9    81    73.31
5    13    93    78.38
6    11    76    77.32
7    23    77    99.81
8    23    95    93.90
9    28   109    95.70
>
```

Cross-validation Jackknife

```
> preg
  psoil pcorn jackpred percerror
1     1    64    62.55      2.27
2     4    71    66.18      6.79
3     5    54    72.29     33.87
4     9    81    73.31      9.49
5    13    93    78.38     15.72
6    11    76    77.32      1.74
7    23    77    99.81     29.62
8    23    95    93.90      1.16
9    28   109    95.70     12.20
>
```

Cross-validation Jackknife

```
> preg
  psoil pcorn jackpred percerror
1     1    64    62.55      2.27
2     4    71    66.18      6.79
3     5    54    72.29     33.87
4     9    81    73.31      9.49
5    13    93    78.38     15.72
6    11    76    77.32      1.74
7    23    77    99.81     29.62
8    23    95    93.90      1.16
9    28   109    95.70     12.20
> 
```

```
> summary(preg$percerror)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.162   2.262   9.493  12.540  15.730  33.870 
> 
```

Jackknife Bootstrap Differences

Both estimate variability of a statistic between subsamples

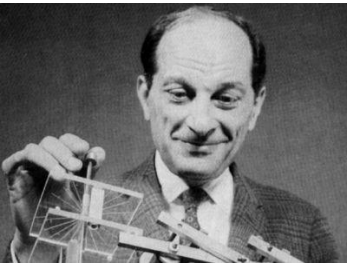
Jackknife provides estimate of the variance of an estimator

Bootstrap first estimates the distribution of the estimator. From this distribution, we can estimate the variance

Using the same data set:

- bootstrap results will always be different (slightly)
- jackknife results will always be identical

Resampling Monte Carlo



Stanisław Ulam

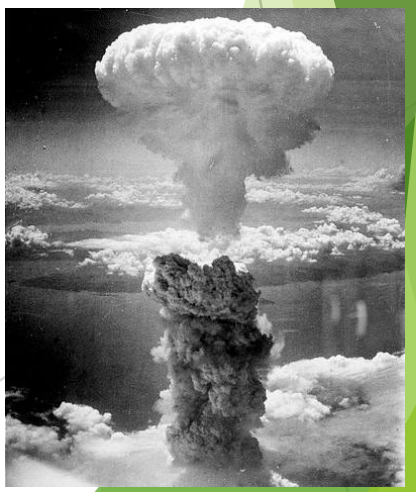
John von Neumann

Los Alamos NATIONAL LABORATORY EST. 1943

MANHATTAN PROJECT BOMB

Mid 1940s

"The first thoughts and attempts I made to practice [the Monte Carlo Method] were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires. The question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays."



Resampling Monte Carlo

Monte Carlo methodS:

not just one
no clear consensus on how they should be defined

Commonality:

repeated sampling from populations with known characteristics,

i.e. we assume a distribution and create random samples that follow that distribution, then compare our estimated statistic to the distribution of outcomes

Comparison

In principle both the parametric and the non-parametric bootstrap are special cases of Monte Carlo simulations used for a very specific purpose: estimate some characteristics of the sampling distribution.

The idea behind the bootstrap is that the sample is an estimate of the population, so an estimate of the sampling distribution can be obtained by drawing many samples (with replacement) from the observed sample, compute the statistic in each new sample.

Monte Carlo simulations are more general: basically it refers to repeatedly creating random data in some way, do something to that random data, and collect some results.

This strategy could be used to estimate some quantity, like in the bootstrap, but also to theoretically investigate some general characteristic of an estimator which is hard to derive analytically.

In practice it would be pretty safe to presume that whenever someone speaks of a Monte Carlo simulation they are talking about a theoretical investigation, e.g. creating random data with no empirical content what so ever to investigate whether an estimator can recover known characteristics of this random `data', while the (parametric) bootstrap refers to an emprical estimation. The fact that the parametric bootstrap implies a model should not worry you: any empirical estimate is based on a model.

Reasons for Supporting resampling

- Do not make assumptions about the sample and the population.
- Conceptually clean and simple.
- Useful when sample sizes are small and the distributional assumptions made by classical techniques cannot be made.
- Some people argue that re-sampling techniques will work even if the data sample is not random. Others remain skeptical, however, since non-random samples may not be representative of the population.

Reasons for Supporting resampling

- Even if a data set meets parametric assumptions, if that set is small, the power of the conclusions, in classical statistics will be low. Re-sampling techniques should suffer less from this.

- If the data set is too large, any null hypothesis can be supported using classical techniques. Cross-validation can help relieve this problem.

- Classical procedures do not inform researchers of how likely the results are to be replicated. Cross-validation and Bootstrapping can be seen as internal replications (external replication is still necessary for confirmation purposes, but internal replication is useful to establish as well).

Reasons for Supporting resampling

- Re-sampling techniques are not devoid of assumptions. The hidden assumption is that the same numbers are used over and over to get an answer that cannot be obtained in any other way.
- Because re-sampling techniques are based on a single sample, the conclusions do not generalize beyond that particular sample.
- Confidence intervals obtained by simple bootstrapping are always biased.
- If the collected data is biased, then re-sampling technique could repeat and magnify that bias.
- If researchers do not conduct enough experimental trials, then the accuracy of re-sampling estimates may be lower than those obtained by conventional parametric techniques.

Discussion

Was the bootstrap example showed parametric or non-parametric?

Could you think an example of the other case?

So, what's the difference between a bootstrap and a Monte Carlo?