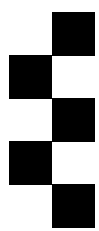
The Generalised Linear Model; Bootstrapping and Permutations

Philip Leifeld

GV903: Advanced Research Methods, Week 17



University of Essex

1. The Generalised Linear Model

The Exponential Family: One-Parameter Model

We have seen many distributions in this module: normal, binomial, Poisson etc.

The Exponential Family: One-Parameter Model

We have seen many distributions in this module: normal, binomial, Poisson etc.

A distribution or mass function is of the *exponential family* if it can be expressed in the following form:

$$f(x;\boldsymbol{\theta}) = \exp\left[\eta(\boldsymbol{\theta})^{\intercal} h(x) - A(\boldsymbol{\theta}) + c(x)\right]$$

The Exponential Family: One-Parameter Model

We have seen many distributions in this module: normal, binomial, Poisson etc.

A distribution or mass function is of the *exponential family* if it can be expressed in the following form:

$$f(x;\boldsymbol{\theta}) = \exp\left[\eta(\boldsymbol{\theta})^{\intercal} h(x) - A(\boldsymbol{\theta}) + c(x)\right]$$

This expression has different functions η, h, A, and c of the parameter of θ vectors and the data x.

The Exponential Family: One-Parameter Model

We have seen many distributions in this module: normal, binomial, Poisson etc.

A distribution or mass function is of the *exponential family* if it can be expressed in the following form:

$$f(x;\boldsymbol{\theta}) = \exp\left[\eta(\boldsymbol{\theta})^{\intercal} h(x) - A(\boldsymbol{\theta}) + c(x)\right]$$

This expression has different functions η, h, A, and c of the parameter of θ vectors and the data x.

Many of the distributions we have dealt with can be reformulated mathematically to conform to the exponential family form.

The Exponential Family: Two-Parameter Model

Some of the distributions we have dealt with can only be pressed into an extended version of the exponential family that has an additional *dispersion* parameter ϕ:

$$f(x;\theta,\phi) = \exp\left[\frac{x\theta - A(\theta)}{\phi} + c(x,\phi)\right]$$

The Exponential Family: Two-Parameter Model

Some of the distributions we have dealt with can only be pressed into an extended version of the exponential family that has an additional *dispersion* parameter ϕ:

$$f(x;\theta,\phi) = \exp\left[\frac{x\theta - A(\theta)}{\phi} + c(x,\phi)\right]$$

This is the notation in Ward and Ahlquist and in Pawitan (2001). Fox(2015) expresses it as:

$$p(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

The Exponential Family: Two-Parameter Model

Some of the distributions we have dealt with can only be pressed into an extended version of the exponential family that has an additional *dispersion* parameter ϕ:

$$f(x;\theta,\phi) = \exp\left[\frac{x\theta - A(\theta)}{\phi} + c(x,\phi)\right]$$

This is the notation in Ward and Ahlquist and in Pawitan (2001). Fox(2015) expresses it as:

$$p(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

This is called a two-parameter exponential-family model with *canonical* (or *location*) parameter θ and *dispersion* (or *scale*) parameter ϕ.

The Exponential Family: Two-Parameter Model

Some of the distributions we have dealt with can only be pressed into an extended version of the exponential family that has an additional *dispersion* parameter ϕ:

$$f(x;\theta,\phi) = \exp\left[\frac{x\theta - A(\theta)}{\phi} + c(x,\phi)\right]$$

This is the notation in Ward and Ahlquist and in Pawitan (2001). Fox(2015) expresses it as:

$$p(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

This is called a two-parameter exponential-family model with *canonical* (or *location*) parameter θ and *dispersion* (or *scale*) parameter ϕ.

Think of dispersion as the variance and location as the mean.

The Bernoulli Model as an Exponential Family

Ward and Ahlquist (2019)

Last week, we formulated the Bernoulli model for binary outcome data:

$$P(Y_i = y_i) = \theta_i^{y_i}(1 - \theta_i)^{1-y_i}$$

The Bernoulli Model as an Exponential Family

Ward and Ahlquist (2019)

Last week, we formulated the Bernoulli model for binary outcome data:

$$P(Y_i = y_i) = \theta_i^{y_i}(1-\theta_i)^{1-y_i}$$

We can show that this is an exponential-family model with $\phi = 1$:

$$\begin{aligned}
f_B(x;\theta) &= \theta^x(1-\theta)^{1-x} \\
&= \exp\left[\log(\theta^x(1-\theta)^{1-x})\right] \\
&= \exp\left[x\log\theta + (1-x)\log(1-\theta)\right] \\
&= \exp\left[x\log\frac{\theta}{1-\theta} + \log(1-\theta)\right] \\
&= \exp\left[x\text{logit}(\theta) + \log(1-\theta)\right].
\end{aligned}$$

The Bernoulli Model as an Exponential Family

Ward and Ahlquist (2019)

Last week, we formulated the Bernoulli model for binary outcome data:

$$P(Y_i = y_i) = \theta_i^{y_i}(1-\theta_i)^{1-y_i}$$

We can show that this is an exponential-family model with $\phi = 1$:

$$\begin{aligned} f_B(x;\theta) &= \theta^x(1-\theta)^{1-x} \\ &= \exp\left[\log(\theta^x(1-\theta)^{1-x})\right] \\ &= \exp\left[x\log\theta + (1-x)\log(1-\theta)\right] \\ &= \exp\left[x\log\frac{\theta}{1-\theta} + \log(1-\theta)\right] \\ &= \exp\left[x\text{logit}(\theta) + \log(1-\theta)\right]. \end{aligned}$$

Similar transformations can be applied to the normal distribution (for the linear model), the Poisson distribution for counts etc.

The Normal Distribution as an Exponential Family

Fox(2015): Applied Regression Analysis and Generalized Linear Models

Remember the normal distribution from Week 3:

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{(y-\mu)^2}{2\sigma^2}\right]$$

The Normal Distribution as an Exponential Family

Fox(2015): Applied Regression Analysis and Generalized Linear Models

Remember the normal distribution from Week 3:

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{(y-\mu)^2}{2\sigma^2}\right]$$

After some "heroic algebraic manipulation" of the normal distribution, we get:

$$p(y;\theta,\phi) = \exp\left\{\frac{y\theta - \theta^2/2}{\phi} - \tfrac{1}{2}\left[\frac{y^2}{\phi} + \log_e(2\pi\phi)\right]\right\}$$

The Normal Distribution as an Exponential Family

Fox(2015): Applied Regression Analysis and Generalized Linear Models

Remember the normal distribution from Week 3:

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{(y-\mu)^2}{2\sigma^2}\right]$$

After some "heroic algebraic manipulation" of the normal distribution, we get:

$$p(y;\theta,\phi) = \exp\left\{\frac{y\theta - \theta^2/2}{\phi} - \tfrac{1}{2}\left[\frac{y^2}{\phi} + \log_e(2\pi\phi)\right]\right\}$$

This is now in the form of the two-parameter exponential family.

Overview of Different Distributions

Fox(2015): Applied Regression Analysis and Generalized Linear Models

Table 15.9 Functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ for Constructing the Exponential Families

Family	$a(\phi)$	$b(\theta)$	$c(y, \phi)$
Gaussian	ϕ	$\theta^2/2$	$-\frac{1}{2}\left[y^2/\phi + \log_e(2\pi\phi)\right]$
Binomial	$1/n$	$\log_e(1+e^\theta)$	$\log_e\binom{n}{ny}$
Poisson	1	e^θ	$-\log_e y!$
Gamma	ϕ	$-\log_e(-\theta)$	$\phi^{-2}\log_e(y/\phi) - \log_e y - \log_e\Gamma(\phi^{-1})$
Inverse-Gaussian	ϕ	$-\sqrt{-2\theta}$	$-\frac{1}{2}\left[\log_e(\pi\phi y^3) + 1/(\phi y)\right]$

NOTE: In this table, n is the number of binomial observations, and $\Gamma(\cdot)$ is the gamma function.

Overview of Different Distributions

Fox(2015): Applied Regression Analysis and Generalized Linear Models

Table 15.9 Functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ for Constructing the Exponential Families

Family	$a(\phi)$	$b(\theta)$	$c(y, \phi)$
Gaussian	ϕ	$\theta^2/2$	$-\frac{1}{2}\left[y^2/\phi + \log_e(2\pi\phi)\right]$
Binomial	$1/n$	$\log_e(1+e^\theta)$	$\log_e\binom{n}{ny}$
Poisson	1	e^θ	$-\log_e y!$
Gamma	ϕ	$-\log_e(-\theta)$	$\phi^{-2}\log_e(y/\phi) - \log_e y - \log_e\Gamma(\phi^{-1})$
Inverse-Gaussian	ϕ	$-\sqrt{-2\theta}$	$-\frac{1}{2}\left[\log_e(\pi\phi y^3) + 1/(\phi y)\right]$

NOTE: In this table, n is the number of binomial observations, and $\Gamma(\cdot)$ is the gamma function.

This forms the *stochastic component* of a GLM.

Overview of Different Distributions

Fox(2015): Applied Regression Analysis and Generalized Linear Models

Table 15.9 Functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ for Constructing the Exponential Families

Family	$a(\phi)$	$b(\theta)$	$c(y, \phi)$
Gaussian	ϕ	$\theta^2/2$	$-\frac{1}{2}\left[y^2/\phi + \log_e(2\pi\phi)\right]$
Binomial	$1/n$	$\log_e(1+e^\theta)$	$\log_e\binom{n}{ny}$
Poisson	1	e^θ	$-\log_e y!$
Gamma	ϕ	$-\log_e(-\theta)$	$\phi^{-2}\log_e(y/\phi) - \log_e y - \log_e\Gamma(\phi^{-1})$
Inverse-Gaussian	ϕ	$-\sqrt{-2\theta}$	$-\frac{1}{2}\left[\log_e(\pi\phi y^3) + 1/(\phi y)\right]$

NOTE: In this table, n is the number of binomial observations, and $\Gamma(\cdot)$ is the gamma function.

This forms the *stochastic component* of a GLM.

The *systematic component* is the (additive) linear model part (called the *linear predictor*).

Components of the GLM

Fox(2015): Applied Regression Analysis and Generalized Linear Models

A generalized linear model (or GLM) consists of three components:

1. A random component, specifying the conditional distribution of the response variable, Y_i (for the ith of n independently sampled observations), given the values of the explanatory variables in the model. In the initial formulation of GLMs, the distribution of Y_i was a member of an exponential family, such as the Gaussian, binomial, Poisson, gamma, or inverse-Gaussian families of distributions.
2. A linear predictor—that is a linear function of regressors,

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

3. A smooth and invertible linearizing link function $g(\cdot)$, which transforms the expectation of the response variable, $\mu_i = E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

Link Functions

Fox(2015): Applied Regression Analysis and Generalized Linear Models

Table 15.1 Some Common Link Functions and Their Inverses

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\log_e \mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	η_i^2
Logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

NOTE: μ_i is the expected value of the response; η_i is the linear predictor; and $\Phi(\cdot)$ is the cumulative distribution function of the standard-normal distribution.

Why should we Care about GLM Theory?

Why should we Care about GLM Theory?

The GLM allows us to subsume most models we care about under a unified terminology and statistical framework.

Why should we Care about GLM Theory?

The GLM allows us to subsume most models we care about under a unified terminology and statistical framework.

Nelder and Wedderburn (1972), the original authors credited with the GLM, showed that Iteratively (re)Weighted Least Squares (IWLS), a simple optimisation algorithm, can be used to find the MLE for any model in the exponential family.

Why should we Care about GLM Theory?

The GLM allows us to subsume most models we care about under a unified terminology and statistical framework.

Nelder and Wedderburn (1972), the original authors credited with the GLM, showed that Iteratively (re)Weighted Least Squares (IWLS), a simple optimisation algorithm, can be used to find the MLE for any model in the exponential family.

Going forward, we will see that many models we deal with (e. g., for ordinal or count data) are GLMs with specific link functions and exponential-family stochastic components.

Why should we Care about GLM Theory?

The GLM allows us to subsume most models we care about under a unified terminology and statistical framework.

Nelder and Wedderburn (1972), the original authors credited with the GLM, showed that Iteratively (re)Weighted Least Squares (IWLS), a simple optimisation algorithm, can be used to find the MLE for any model in the exponential family.

Going forward, we will see that many models we deal with (e. g., for ordinal or count data) are GLMs with specific link functions and exponential-family stochastic components.

In `R`, many of them are conveniently implemented in a single function, `glm`, with argument `family`.
For example, `family = binomial(link = "logit")`.

2. Bootstrapping

Bootstrapping: Introduction

- Imagine a situation in which we have a sample but don't have access to the underlying population.

Bootstrapping: Introduction

- Imagine a situation in which we have a sample but don't have access to the underlying population.
- We must trust that the sample is a random selection from the population.

Bootstrapping: Introduction

- Imagine a situation in which we have a sample but don't have access to the underlying population.
- We must trust that the sample is a random selection from the population.
- We could calculate various statistics from the sample, such as the mean, median, standard deviation, and lots of other functions.

Bootstrapping: Introduction

- Imagine a situation in which we have a sample but don't have access to the underlying population.
- We must trust that the sample is a random selection from the population.
- We could calculate various statistics from the sample, such as the mean, median, standard deviation, and lots of other functions.
- But we also want to factor in the uncertainty around the respective statistic due to the sampling process.

Bootstrapping: Introduction

- Imagine a situation in which we have a sample but don't have access to the underlying population.
- We must trust that the sample is a random selection from the population.
- We could calculate various statistics from the sample, such as the mean, median, standard deviation, and lots of other functions.
- But we also want to factor in the uncertainty around the respective statistic due to the sampling process.
- For example, we want a confidence interval around the standard deviation.

Bootstrapping: Introduction

- Imagine a situation in which we have a sample but don't have access to the underlying population.
- We must trust that the sample is a random selection from the population.
- We could calculate various statistics from the sample, such as the mean, median, standard deviation, and lots of other functions.
- But we also want to factor in the uncertainty around the respective statistic due to the sampling process.
- For example, we want a confidence interval around the standard deviation.
- We cannot compute this because we have only that one sample.

Bootstrapping: Introduction

- Imagine a situation in which we have a sample but don't have access to the underlying population.
- We must trust that the sample is a random selection from the population.
- We could calculate various statistics from the sample, such as the mean, median, standard deviation, and lots of other functions.
- But we also want to factor in the uncertainty around the respective statistic due to the sampling process.
- For example, we want a confidence interval around the standard deviation.
- We cannot compute this because we have only that one sample.
- Bootstrapping is a way to do this anyway.

Bootstrapping: Introduction

- The term "bootstrapping" usually refers to a self-starting process that is supposed to proceed without external input.

Bootstrapping: Introduction

- The term "bootstrapping" usually refers to a self-starting process that is supposed to proceed without external input.
- It is a bit like black magic. It works, but is hard to wrap your head around why.

Bootstrapping: Introduction

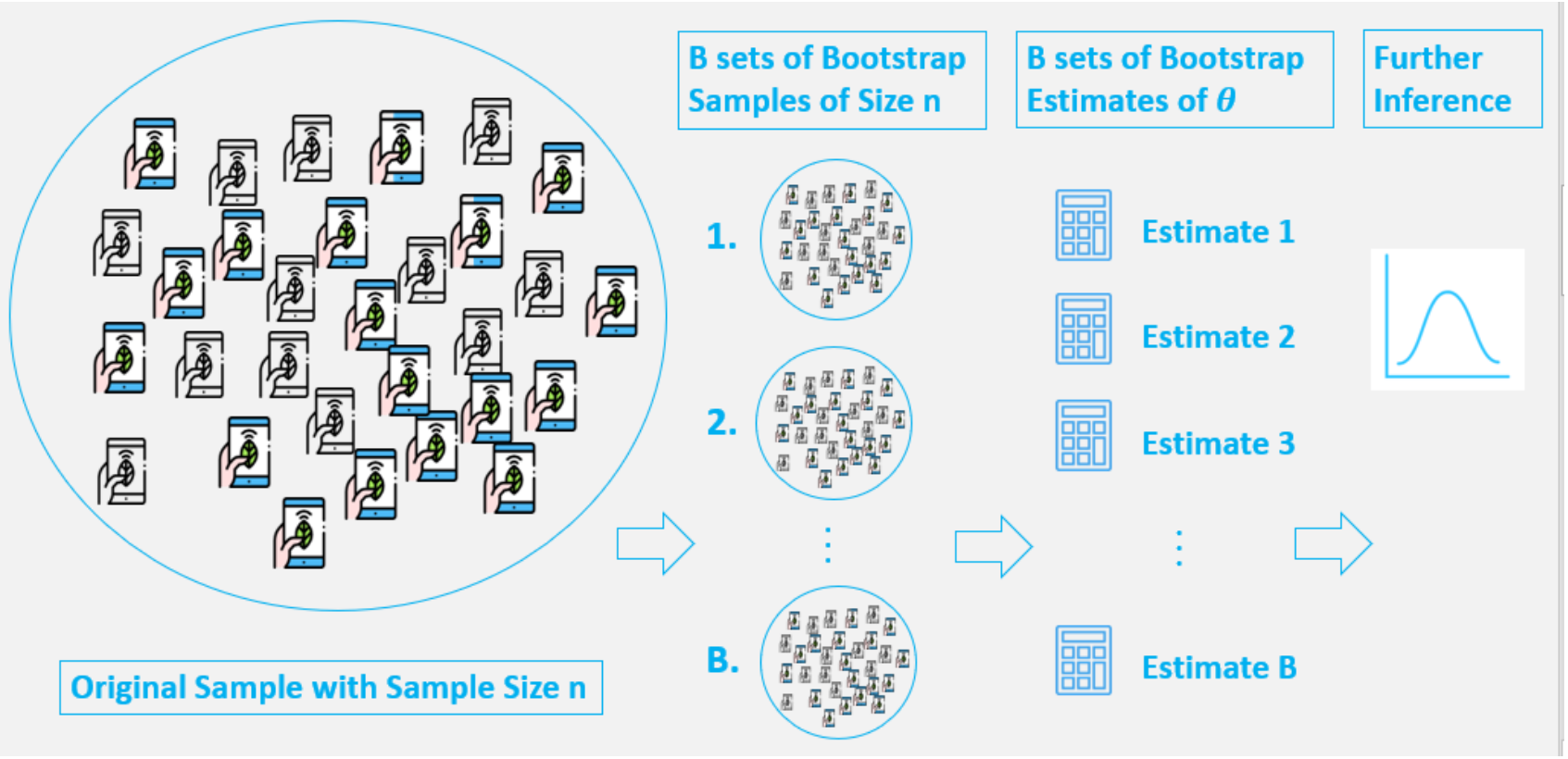
- The term "bootstrapping" usually refers to a self-starting process that is supposed to proceed without external input.
- It is a bit like black magic. It works, but is hard to wrap your head around why.
- The *bootstrapping* metaphor goes back to the 19^{th} century saying "to pull oneself over a fence by one's bootstraps", which is of course an impossible task.

Bootstrapping: Introduction

- The term "bootstrapping" usually refers to a self-starting process that is supposed to proceed without external input.
- It is a bit like black magic. It works, but is hard to wrap your head around why.
- The *bootstrapping* metaphor goes back to the 19^{th} century saying "to pull oneself over a fence by one's bootstraps", which is of course an impossible task.
- Applied to statistics: We sample with replacement from the sample itself.

Bootstrapping: Algorithm Overview

https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60



Bootstrapping: Formalisation

- Let's say we have a sample with $n = 10$ values:
  $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10$

Bootstrapping: Formalisation

- Let's say we have a sample with $n = 10$ values:
  $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10$
- We can re-sample n values with replacement from this:
  $b_1 = \{1, 3, 10, 3, 3, 6, 5, 1, 5\}$,
  $b_2 = \{9, 7, 10, 3, 2, 2, 1, 10, 10\}$ and so on...

Bootstrapping: Formalisation

- Let's say we have a sample with $n = 10$ values:
  $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10$
- We can re-sample n values with replacement from this:
  $b_1 = \{1, 3, 10, 3, 3, 6, 5, 1, 5\}$,
  $b_2 = \{9, 7, 10, 3, 2, 2, 1, 10, 10\}$ and so on...
- Note how some values appear multiple times while other don't appear in a specific bootstrapping sample due to the replacement.

Bootstrapping: Formalisation

- Let's say we have a sample with $n = 10$ values:
  $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10$
- We can re-sample n values with replacement from this:
  $b_1 = \{1, 3, 10, 3, 3, 6, 5, 1, 5\}$,
  $b_2 = \{9, 7, 10, 3, 2, 2, 1, 10, 10\}$ and so on...
- Note how some values appear multiple times while other don't appear in a specific bootstrapping sample due to the replacement.
- These bootstrapping samples form a set of samples
  $B = \{b_1, b_2, \ldots, b_K\}$.

Bootstrapping: Formalisation

- K should be at least 1,000 or more.

Bootstrapping: Formalisation

- K should be at least 1,000 or more.
- We can now pretend B was coming from the population and compute distributions of statistics rather than a single statistic.

Bootstrapping: Formalisation

- K should be at least 1,000 or more.
- We can now pretend B was coming from the population and compute distributions of statistics rather than a single statistic.
- For example, a distribution of standard deviations across B instead of a single standard deviation of S.

Bootstrapping: Formalisation

- K should be at least 1,000 or more.
- We can now pretend B was coming from the population and compute distributions of statistics rather than a single statistic.
- For example, a distribution of standard deviations across B instead of a single standard deviation of S.
- Then we can compute the standard error and confidence intervals based on this distribution.

Bootstrapping: Formalisation

- K should be at least 1,000 or more.
- We can now pretend B was coming from the population and compute distributions of statistics rather than a single statistic.
- For example, a distribution of standard deviations across B instead of a single standard deviation of S.
- Then we can compute the standard error and confidence intervals based on this distribution.
- Or use the 2.5 per cent and 97.5 quantiles of the distribution of B as the confidence interval if the distribution is not normal.

Bootstrapping Illustration in R

We will try this with $n = 20$ for illustration. With larger samples, it will be asymptotically unbiased.

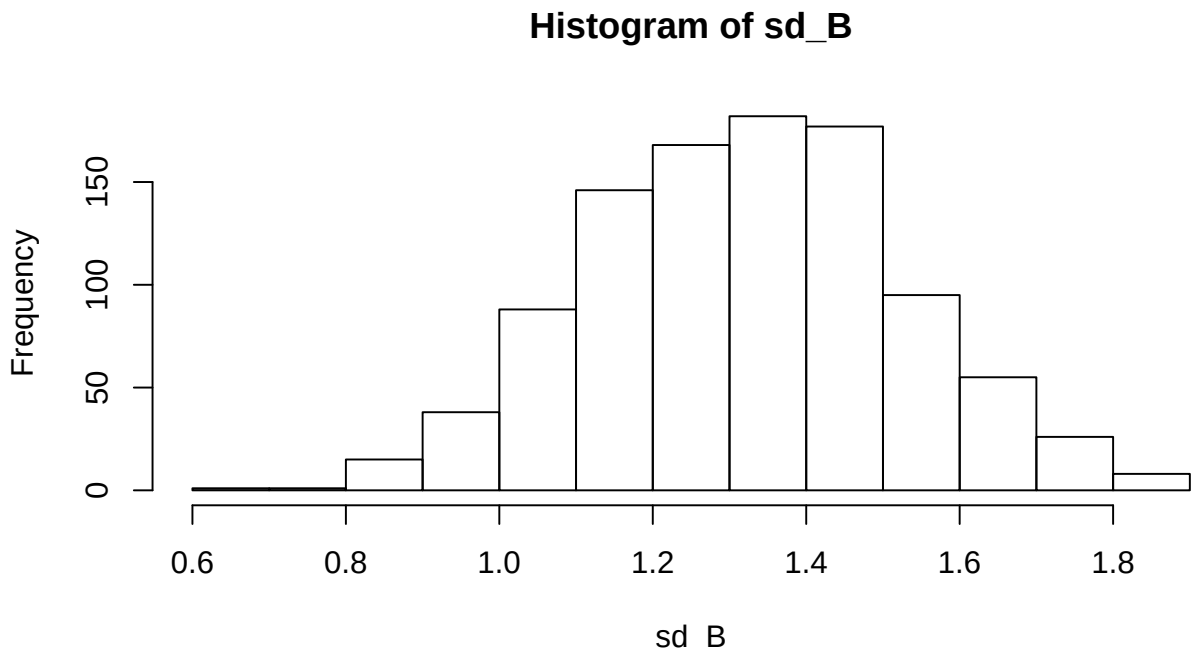
```
set.seed(123) # set a random seed for reproducibility
n <- 20
S <- rnorm(n = n, mean = 5.5, sd = 1.4) # generate some values
S
##  [1] 4.715334 5.177752 7.682192 5.598712 5.681003 7.901091
##  [7] 6.145283 3.728914 4.538406 4.876073 7.213715 6.003739
## [13] 6.061080 5.654956 4.721822 8.001678 6.196991 2.746736
## [19] 6.481898 4.838092

sd(S) # the empirical SD is slightly off due to small n
## [1] 1.361731

B <- matrix(0, nrow = 1000, ncol = n)
for (i in 1:1000) { # resample 1000 times
  B[i, ] <- sample(S, size = n, replace = TRUE)
}
```

Bootstrapping Illustration in R

```
B[1:2, ] # let's look at some of the resampled values
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
## [1,] 4.721822 4.876073 6.061080 6.145283 4.538406 4.538406
## [2,] 8.001678 4.838092 7.901091 7.213715 3.728914 6.145283
##          [,7]     [,8]     [,9]    [,10]    [,11]    [,12]
## [1,] 4.876073 6.145283 7.901091 5.177752 5.681003 3.728914
## [2,] 8.001678 6.196991 2.746736 6.196991 5.177752 5.598712
##         [,13]    [,14]    [,15]    [,16]    [,17]    [,18]
## [1,] 6.003739 6.061080 2.746736 4.715334 7.901091 4.721822
## [2,] 6.061080 5.681003 6.481898 4.838092 5.654956 7.682192
##         [,19]    [,20]
## [1,] 4.538406 4.721822
## [2,] 3.728914 8.001678

sd_B <- apply(B, 1, sd) # compute SD for each row
```

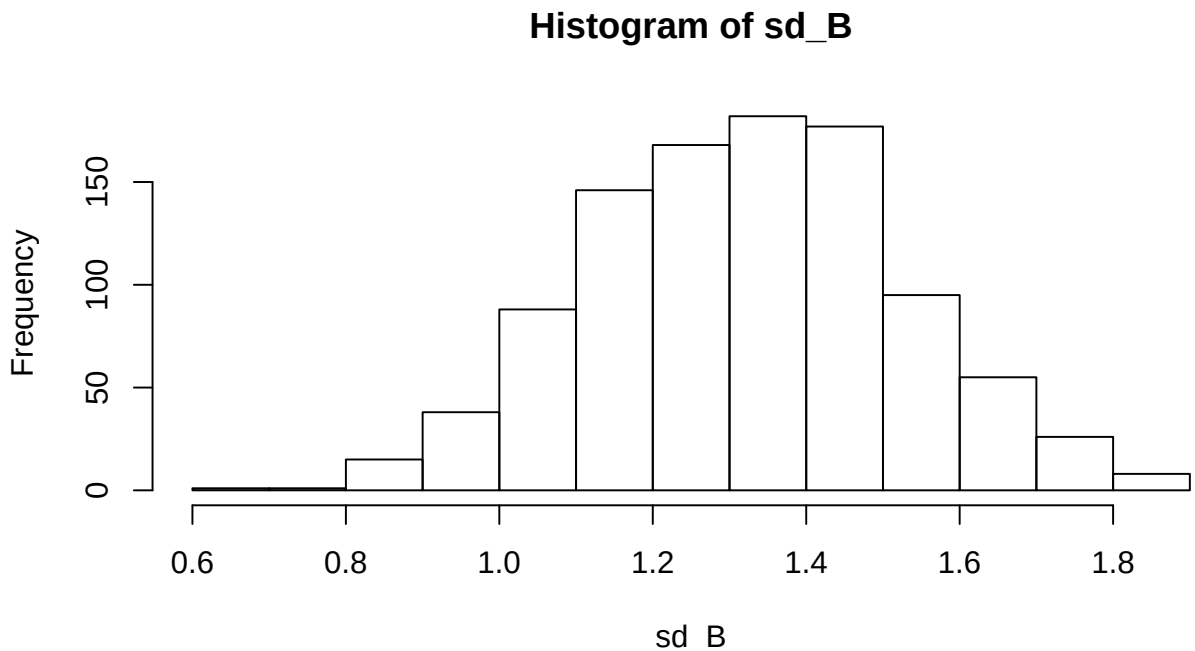
Bootstrapping Illustration in R

```
sd_hat <- mean(sd_B) # compute mean SD across samples
sd_hat # again, slightly off due to small n
## [1] 1.320427

hist(sd_B) # inspect the distribution of the SDs
```



Bootstrapping Illustration in R

```
sd_hat <- mean(sd_B) # compute mean SD across samples
sd_hat # again, slightly off due to small n
## [1] 1.320427

hist(sd_B) # inspect the distribution of the SDs
```



Here's the black magic: We have a sampling distribution of the SD and can now compute SE, CI etc.

Bootstrapping Illustration in R

The SE of any sample statistic is the SD of the sampling distribution for that statistic.

Bootstrapping Illustration in R

The SE of any sample statistic is the SD of the sampling distribution for that statistic.

```
SE <- sd(sd_B)
SE
## [1] 0.2022155

CI_lower <- sd_hat - qnorm(1 - (0.05 / 2)) * SE
CI_upper <- sd_hat + qnorm(1 - (0.05 / 2)) * SE
cat(sd_hat, " [", CI_lower, "; ", CI_upper, "]", sep = "")
## 1.320427 [0.9240922; 1.716762]
```

Bootstrapping Illustration in R

The SE of any sample statistic is the SD of the sampling distribution for that statistic.

```
SE <- sd(sd_B)
SE
## [1] 0.2022155

CI_lower <- sd_hat - qnorm(1 - (0.05 / 2)) * SE
CI_upper <- sd_hat + qnorm(1 - (0.05 / 2)) * SE
cat(sd_hat, " [", CI_lower, "; ", CI_upper, "]", sep = "")
## 1.320427 [0.9240922; 1.716762]
```

Or, if we cannot assume a normal distribution, we can just use the quantiles from the empirical sampling distribution:

```
quantile(sd_B, 0.025)
##      2.5%
## 0.9413218

quantile(sd_B, 0.975)
##    97.5%
## 1.722768
```

Why do we Need this?

Why do we Need this?

In small samples like here, estimating confidence intervals would lead to overconfidence.

Why do we Need this?

In small samples like here, estimating confidence intervals would lead to overconfidence. For comparison:

```
# lower CI bound, based on initial sample
sd(S) - qnorm(1 - (0.05 / 2)) * (sd(S) / sqrt(n))
## [1] 0.7649373

# upper CI bound, based on initial sample
sd(S) + qnorm(1 - (0.05 / 2)) * (sd(S) / sqrt(n))
## [1] 1.958526
```

Why do we Need this?

In small samples like here, estimating confidence intervals would lead to overconfidence. For comparison:

```
# lower CI bound, based on initial sample
sd(S) - qnorm(1 - (0.05 / 2)) * (sd(S) / sqrt(n))
## [1] 0.7649373

# upper CI bound, based on initial sample
sd(S) + qnorm(1 - (0.05 / 2)) * (sd(S) / sqrt(n))
## [1] 1.958526
```

Sometimes, the population distribution is unknown.

Why do we Need this?

In small samples like here, estimating confidence intervals would lead to overconfidence. For comparison:

```
# lower CI bound, based on initial sample
sd(S) - qnorm(1 - (0.05 / 2)) * (sd(S) / sqrt(n))
## [1] 0.7649373

# upper CI bound, based on initial sample
sd(S) + qnorm(1 - (0.05 / 2)) * (sd(S) / sqrt(n))
## [1] 1.958526
```

Sometimes, the population distribution is unknown. Example: if you do content analysis and count the number of occurrences of some word per speech, to compare two politicians' speech patterns. You want to report the counts along with confidence intervals. You can bootstrap the CIs without having to assume that the counts are generated by a Poisson process and looking up the variance of the Poisson distribution. Cool.

Case Study: Bootstrapping in the TERGM

- The temporal exponential random graph model (TERGM) is a model for panel network data.

Case Study: Bootstrapping in the TERGM

- The temporal exponential random graph model (TERGM) is a model for panel network data.
- For example, how do the friendships between students in a school class evolve over the weeks of the semester and why?

Case Study: Bootstrapping in the TERGM

- The temporal exponential random graph model (TERGM) is a model for panel network data.
- For example, how do the friendships between students in a school class evolve over the weeks of the semester and why?
- Estimation: MCMC-MLE. Computationally *very* expensive.

Case Study: Bootstrapping in the TERGM

- The temporal exponential random graph model (TERGM) is a model for panel network data.
- For example, how do the friendships between students in a school class evolve over the weeks of the semester and why?
- Estimation: MCMC-MLE. Computationally *very* expensive.
- With many observations (e. g., international conflict), the MLE cannot be computed in a hundred human lifetimes.

Case Study: Bootstrapping in the TERGM

- The temporal exponential random graph model (TERGM) is a model for panel network data.
- For example, how do the friendships between students in a school class evolve over the weeks of the semester and why?
- Estimation: MCMC-MLE. Computationally *very* expensive.
- With many observations (e. g., international conflict), the MLE cannot be computed in a hundred human lifetimes.
- A simpler estimation strategy, MPLE, is known to be fast but produces biased SEs.

Case Study: Bootstrapping in the TERGM

- The temporal exponential random graph model (TERGM) is a model for panel network data.
- For example, how do the friendships between students in a school class evolve over the weeks of the semester and why?
- Estimation: MCMC-MLE. Computationally *very* expensive.
- With many observations (e. g., international conflict), the MLE cannot be computed in a hundred human lifetimes.
- A simpler estimation strategy, MPLE, is known to be fast but produces biased SEs.
- Bootstrapping across time steps and recomputing the SEs has been shown to be unbiased.

Case Study: Bootstrapping in the TERGM

- The temporal exponential random graph model (TERGM) is a model for panel network data.
- For example, how do the friendships between students in a school class evolve over the weeks of the semester and why?
- Estimation: MCMC-MLE. Computationally *very* expensive.
- With many observations (e. g., international conflict), the MLE cannot be computed in a hundred human lifetimes.
- A simpler estimation strategy, MPLE, is known to be fast but produces biased SEs.
- Bootstrapping across time steps and recomputing the SEs has been shown to be unbiased.

Leifeld, Philip, Skyler J. Cranmer and Bruce A. Desmarais (2018): Temporal Exponential Random Graph Models with `btergm`: Estimation and Bootstrap Confidence Intervals. *Journal of Statistical Software* 83(6): 1–36.

3. Other Resampling Approaches

The Jackknife

- The jackknife is like bootstrapping.

The Jackknife

- The jackknife is like bootstrapping.
- But instead of resampling with replacement, we copy the existing sample n times and delete one (different) observation from each of these new samples.

The Jackknife

- The jackknife is like bootstrapping.
- But instead of resampling with replacement, we copy the existing sample n times and delete one (different) observation from each of these new samples.
- Then we can continue like with the bootstrap, based on the n reduced samples.

The Jackknife

- The jackknife is like bootstrapping.
- But instead of resampling with replacement, we copy the existing sample n times and delete one (different) observation from each of these new samples.
- Then we can continue like with the bootstrap, based on the n reduced samples.
- This is called a *leave-one-out* procedure.

The Jackknife

- The jackknife is like bootstrapping.
- But instead of resampling with replacement, we copy the existing sample n times and delete one (different) observation from each of these new samples.
- Then we can continue like with the bootstrap, based on the n reduced samples.
- This is called a *leave-one-out* procedure.
- Bootstrap is newer and more flexible.

Cross-Validation

- The leave-one-out procedure can also be applied for assessing predictive fit.

Cross-Validation

- The leave-one-out procedure can also be applied for assessing predictive fit.
- Leave one observation out; estimate model; predict left-out observation.

Cross-Validation

- The leave-one-out procedure can also be applied for assessing predictive fit.
- Leave one observation out; estimate model; predict left-out observation.
- Measure how well the prediction matches the left-out data.

Cross-Validation

- The leave-one-out procedure can also be applied for assessing predictive fit.
- Leave one observation out; estimate model; predict left-out observation.
- Measure how well the prediction matches the left-out data.
- Repeat for each single observation to generate a distribution of prediction errors.

Cross-Validation

- The leave-one-out procedure can also be applied for assessing predictive fit.
- Leave one observation out; estimate model; predict left-out observation.
- Measure how well the prediction matches the left-out data.
- Repeat for each single observation to generate a distribution of prediction errors.
- The mean of the distribution is a measure of model fit.

Cross-Validation

- The leave-one-out procedure can also be applied for assessing predictive fit.
- Leave one observation out; estimate model; predict left-out observation.
- Measure how well the prediction matches the left-out data.
- Repeat for each single observation to generate a distribution of prediction errors.
- The mean of the distribution is a measure of model fit.
- It can be used to compare different models.

Cross-Validation

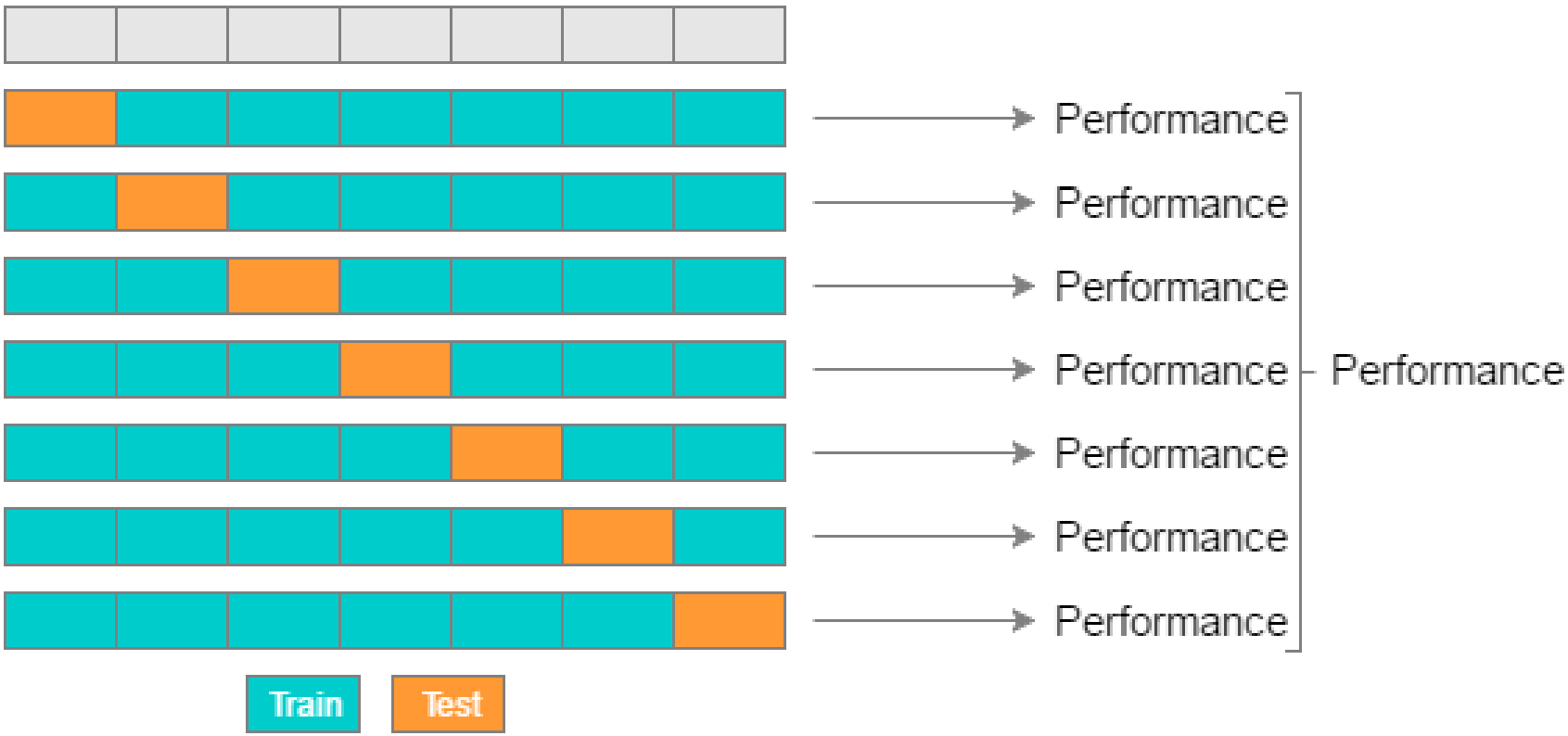
- The leave-one-out procedure can also be applied for assessing predictive fit.
- Leave one observation out; estimate model; predict left-out observation.
- Measure how well the prediction matches the left-out data.
- Repeat for each single observation to generate a distribution of prediction errors.
- The mean of the distribution is a measure of model fit.
- It can be used to compare different models.
- k-fold cross-validation: Divide the data into k parts and predict one left out segment based on a model of the remaining $k - 1$ segments; then assess distribution of prediction error.

Cross-Validation

- The leave-one-out procedure can also be applied for assessing predictive fit.
- Leave one observation out; estimate model; predict left-out observation.
- Measure how well the prediction matches the left-out data.
- Repeat for each single observation to generate a distribution of prediction errors.
- The mean of the distribution is a measure of model fit.
- It can be used to compare different models.
- k-fold cross-validation: Divide the data into k parts and predict one left out segment based on a model of the remaining $k - 1$ segments; then assess distribution of prediction error.
- Corrects for overfitting of the model to the data.

Illustration of k-fold Cross-Validation

`http://www.ebc.cat/2017/01/31/cross-validation-strategies/`



Performance can be assessed using AUC or mean squared prediction error.

Permutation Tests

Permutation Tests

- Imagine I run an unethical experiment: I split the students in this module randomly into two groups. The treatment group receives some wonder drug. I want to know whether that increases the capacity to understand the module materials. Exam marks serve as measurements.

Permutation Tests

- Imagine I run an unethical experiment: I split the students in this module randomly into two groups. The treatment group receives some wonder drug. I want to know whether that increases the capacity to understand the module materials. Exam marks serve as measurements.
- Let's say 10 random people get the drug. They receive average marks of 69 while the remaining students receive 57.

Permutation Tests

- Imagine I run an unethical experiment: I split the students in this module randomly into two groups. The treatment group receives some wonder drug. I want to know whether that increases the capacity to understand the module materials. Exam marks serve as measurements.
- Let's say 10 random people get the drug. They receive average marks of 69 while the remaining students receive 57.
- H_0: There is no difference. H_1: Positive effect on grades.

Permutation Tests

- Imagine I run an unethical experiment: I split the students in this module randomly into two groups. The treatment group receives some wonder drug. I want to know whether that increases the capacity to understand the module materials. Exam marks serve as measurements.
- Let's say 10 random people get the drug. They receive average marks of 69 while the remaining students receive 57.
- H_0: There is no difference. H_1: Positive effect on grades.
- I can now do a two-sample t-test or regression with a group dummy to check if there are any differences.

Permutation Tests

- Imagine I run an unethical experiment: I split the students in this module randomly into two groups. The treatment group receives some wonder drug. I want to know whether that increases the capacity to understand the module materials. Exam marks serve as measurements.
- Let's say 10 random people get the drug. They receive average marks of 69 while the remaining students receive 57.
- H_0: There is no difference. H_1: Positive effect on grades.
- I can now do a two-sample t-test or regression with a group dummy to check if there are any differences.
- OR I can do a permutation test – without any distributional assumptions!

Permutation Test – How does it Work?

Student name	Group	Mark
Student A	Treatment	43
Student B	Control	48
Student C	Treatment	97
Student D	Treatment	74
Student E	Control	65
...	...	...

Permutation Test – How does it Work?

Student name	Group	Mark
Student A	Treatment	43
Student B	Control	48
Student C	Treatment	97
Student D	Treatment	74
Student E	Control	65
...	...	...

1. We save the observed difference in group means.

Permutation Test – How does it Work?

Student name	Group	Mark
Student A	Treatment	43
Student B	Control	48
Student C	Treatment	97
Student D	Treatment	74
Student E	Control	65
...	...	...

1. We save the observed difference in group means.
2. We re-assign (reshuffle, permute) group labels randomly.

Permutation Test – How does it Work?

Student name	Group	Mark
Student A	Treatment	43
Student B	Control	48
Student C	Treatment	97
Student D	Treatment	74
Student E	Control	65
...	...	...

1. We save the observed difference in group means.
2. We re-assign (reshuffle, permute) group labels randomly.
3. We recompute the difference in group means again.

Permutation Test – How does it Work?

Student name	Group	Mark
Student A	Treatment	43
Student B	Control	48
Student C	Treatment	97
Student D	Treatment	74
Student E	Control	65
...	...	...

1. We save the observed difference in group means.
2. We re-assign (reshuffle, permute) group labels randomly.
3. We recompute the difference in group means again.
4. We repeat steps (2) and (3) many times. This yields a distribution of differences.

Permutation Test – How does it Work?

Student name	Group	Mark
Student A	Treatment	43
Student B	Control	48
Student C	Treatment	97
Student D	Treatment	74
Student E	Control	65
...	...	...

1. We save the observed difference in group means.
2. We re-assign (reshuffle, permute) group labels randomly.
3. We recompute the difference in group means again.
4. We repeat steps (2) and (3) many times. This yields a distribution of differences.
5. We check if our observed difference is in the tails of that distribution to check for significance.

Permutation Test in R

Let's first create some data for this experiment:

Permutation Test in R

Let's first create some data for this experiment:

```
set.seed(123)
dat <- data.frame(group = c(rep("t", 10),
                            rep("c", 10)),
                  mark = c(rnorm(10, 69, 10),
                           rnorm(10, 57, 10)))
head(dat)
##   group     mark
## 1     t 63.39524
## 2     t 66.69823
## 3     t 84.58708
## 4     t 69.70508
## 5     t 70.29288
## 6     t 86.15065
```

Permutation Test in R

Now let's compute the difference and the t test:

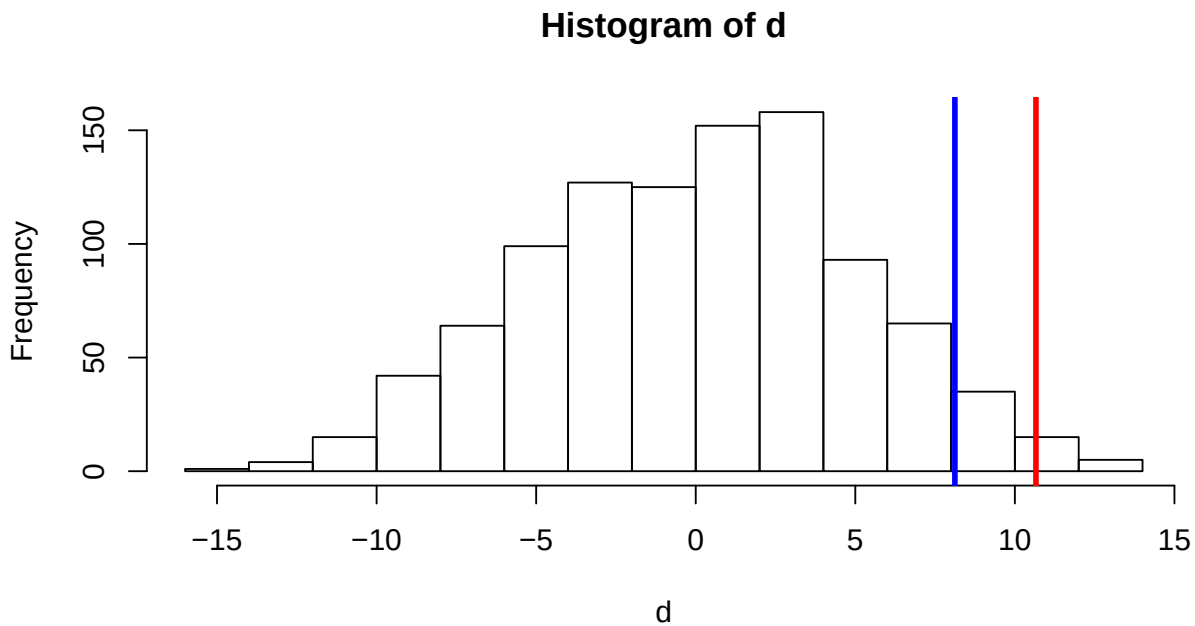
Permutation Test in R

Now let's compute the difference and the t test:

```
obs <- mean(dat$mark[dat$group == "t"]) -
    mean(dat$mark[dat$group == "c"])
obs
## [1] 10.66004

t.test(dat$mark[dat$group == "t"], dat$mark[dat$group == "c"])
##
##  Welch Two Sample t-test
##
## data:  dat$mark[dat$group == "t"] and dat$mark[dat$group == "c"]
## t = 2.3913, df = 17.872, p-value = 0.02801
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.289512 20.030562
## sample estimates:
## mean of x mean of y
##  69.74626  59.08622
```

Permutation Test in R

```r
d <- numeric(1000)
for (i in 1:1000) {
  dat2 <- dat
  dat2$group <- sample(dat2$group, replace = FALSE)
  d[i] <- mean(dat2$mark[dat2$group == "t"]) -
      mean(dat2$mark[dat2$group == "c"])
}
hist(d)
abline(v = quantile(d, 0.95), col = "blue", lwd = 3)
abline(v = obs, col = "red", lwd = 3)
```



Permutation Test in R

We can check what the 95 per cent quantile is (as the critical value) and compare the observed difference:

```
obs
## [1] 10.66004

quantile(d, 0.95)
##      95%
## 8.120266
```

Permutation Test in R

We can check what the 95 per cent quantile is (as the critical value) and compare the observed difference:

```
obs
## [1] 10.66004

quantile(d, 0.95)
##      95%
## 8.120266
```

The observed difference is clearly more extreme than the difference we would have observed if there was no difference in treatment and control group!

Permutation Test in R

We can check what the 95 per cent quantile is (as the critical value) and compare the observed difference:

```
obs
## [1] 10.66004

quantile(d, 0.95)
##      95%
## 8.120266
```

The observed difference is clearly more extreme than the difference we would have observed if there was no difference in treatment and control group!

We can also compute the p-value directly:

```
1 - ecdf(d)(obs) # create a cdf function for empirical values
## [1] 0.013
```

Permutation Tests: Conclusion

- Note how we did not need any distributional assumptions here.

Permutation Tests: Conclusion

- Note how we did not need any distributional assumptions here.
- We just computed the null distribution from the data.

Permutation Tests: Conclusion

- Note how we did not need any distributional assumptions here.
- We just computed the null distribution from the data.
- The same principle is applicable in many other situations, not just experiments.

Permutation Tests: Conclusion

- Note how we did not need any distributional assumptions here.
- We just computed the null distribution from the data.
- The same principle is applicable in many other situations, not just experiments.
- Example: Is a significant effect due to temporal variation or cross-sectional variation?

Permutation Tests: Conclusion

- Note how we did not need any distributional assumptions here.
- We just computed the null distribution from the data.
- The same principle is applicable in many other situations, not just experiments.
- Example: Is a significant effect due to temporal variation or cross-sectional variation?
- Just permute time and see if the effect is still there!

Permutation Tests: Conclusion

- Note how we did not need any distributional assumptions here.
- We just computed the null distribution from the data.
- The same principle is applicable in many other situations, not just experiments.
- Example: Is a significant effect due to temporal variation or cross-sectional variation?
- Just permute time and see if the effect is still there!
- Malang, Thomas, Laurence Brandenberger and Philip Leifeld (2019): Networks and Social Influence in European Legislative Politics. *British Journal of Political Science* 49(4): 1475–1498.

Permutation Tests: Conclusion

- Note how we did not need any distributional assumptions here.
- We just computed the null distribution from the data.
- The same principle is applicable in many other situations, not just experiments.
- Example: Is a significant effect due to temporal variation or cross-sectional variation?
- Just permute time and see if the effect is still there!
- Malang, Thomas, Laurence Brandenberger and Philip Leifeld (2019): Networks and Social Influence in European Legislative Politics. *British Journal of Political Science* 49(4): 1475–1498.
- But: If parametric (distribution-based) tests are feasible and plausible, they should be preferred!