

Practical Statistics (in R)

Josh Sumner

Outline

- **Scientific process**
- **Statistics workflow**
- **Example scenarios**

The Scientific Process

- 1) Theorize question or problem
- 2) Develop Hypotheses
- 3) Design Experiment
- 4) Record observations
- 5) Analyze observations

The Scientific Process

- 1) Theorize question or problem
- 2) Develop Hypotheses
- 3) Design Experiment
- 4) Record observations
- 5) Analyze observations

- Not all designs are created equally



- poor design leads to impossible inferences

The Scientific Process

- 1) Theorize question or problem
- 2) Develop Hypotheses
- 3) Design Experiment
- 4) Record observations
- 5) Analyze observations

- Not all designs are created equally

- poor design leads to impossible inferences

- Not all observations are created equally

- poor statistical methods leads to false-positives and false-negatives

- **continuous > ordinal >= nominal**

The Scientific Process

- 1) Theorize question or problem
- 2) Develop Hypotheses
- 3) Design Experiment
- 4) Record observations
- 5) Analyze observations

- Not all designs are created equally

- poor design leads to impossible inferences

- Not all observations are created equally

- poor statistical methods leads to false-positives and false-negatives
- **continuous** > **ordinal** >= **nominal**

- Not all methods are created equally

- many methods can only test a handful of hypotheses and use p-values

The Scientific Process

- 1) Theorize question or problem
- 2) Develop Hypotheses
- 3) Design Experiment
- 4) Record observations
- 5) Analyze observations

- Before starting you should consider

What outcomes are possible?

Are the questions you're interested in answerable with the design?

What format will your data be in?

What test will you use and what hypotheses are possible in it?

Do you have the replication required?

The Scientific Process

- 1) Theorize question or problem
- 2) Develop Hypotheses
- 3) Design Experiment
- 4) Record observations
- 5) Analyze observations

The **Stats in RCR** Workshop gets into These steps some And talks about how To know when you're Off of the flow chart.

The **Troubleshooting In R** workshop gets into some coding problems in this step.

- Before starting you should consider

What outcomes are possible?

Are the questions you're interested in answerable with the design?

What format will your data be in?

What test will you use and what hypotheses are possible in it?

Do you have the replication required?

Focus of the **Stats In R** and **Stats in pcvr** Workshops

The **Power Analysis** Workshop focuses On these parts.

The Scientific Process

- 1) Theorize question or problem
- 2) Develop Hypotheses
- 3) Design Experiment
- 4) Record observations
- 5) Analyze observations

- Before starting you should consider

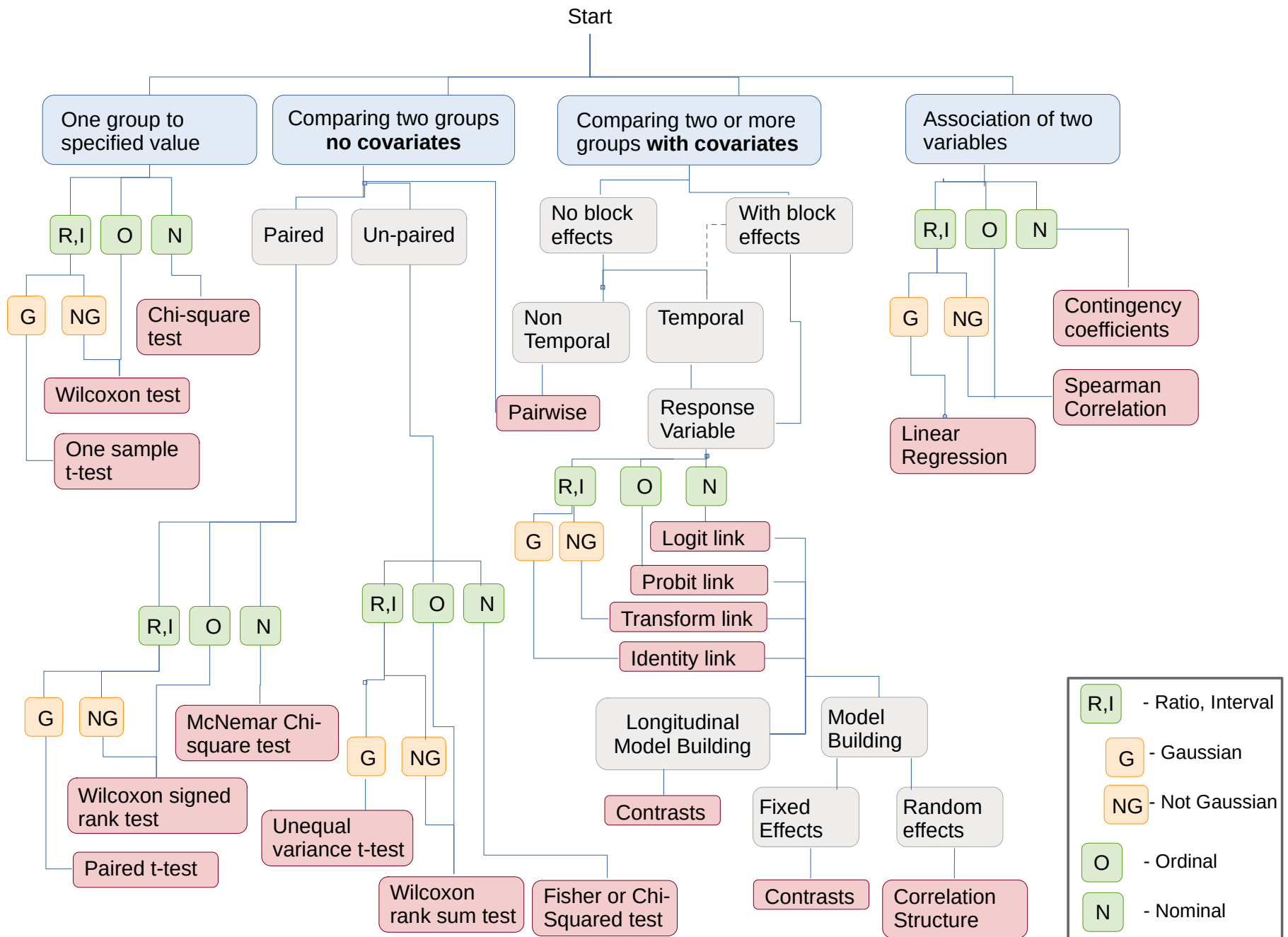
What outcomes are possible?

Are the questions you're interested in answerable with the design?

What format will your data be in?

What test will you use and what hypotheses are possible in it?

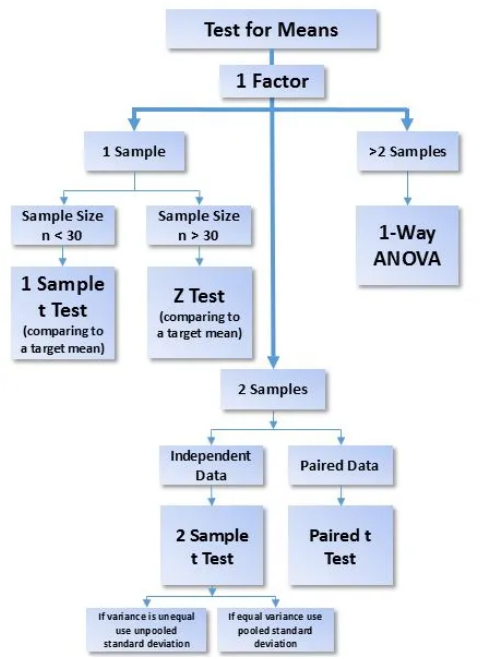
Do you have the replication required?



* If the outcome and main predictor are both two-level factors, Breslow-Day and Cochren-Mantel-Hanzel tests are better

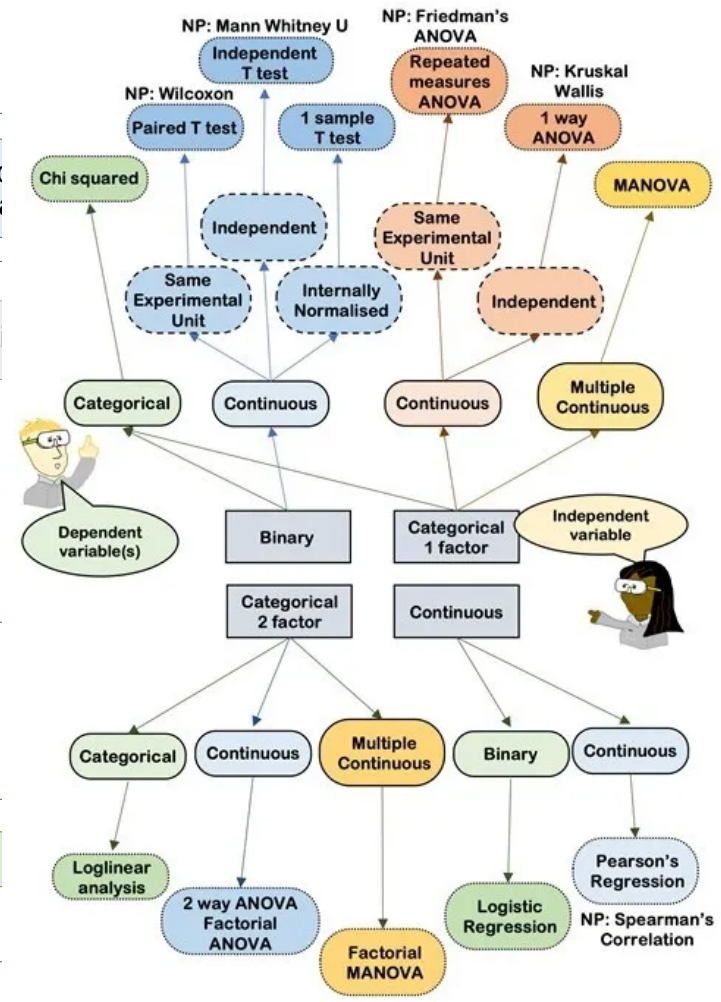
Parametric Test Flowchart

www.six-sigma-material.com



Comparing two
no covari

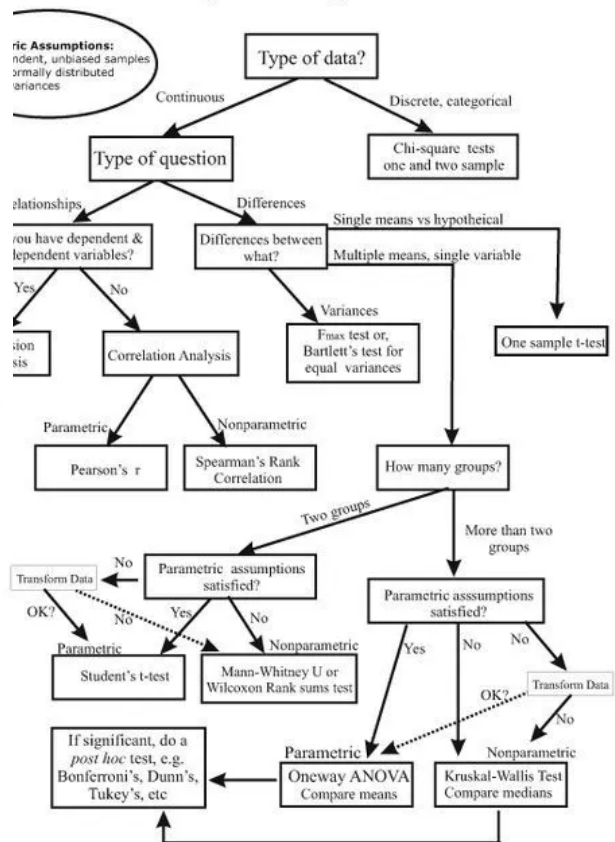
red Un-pa



association of two
variables

I O N

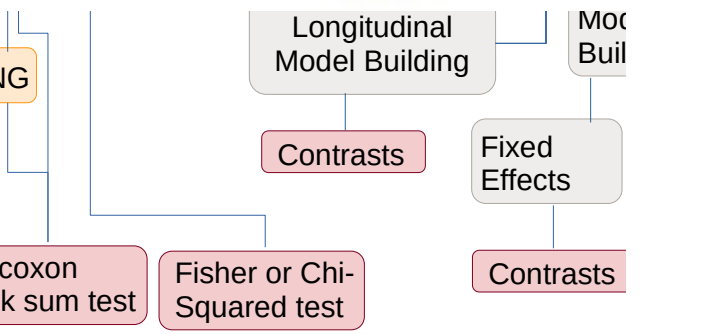
Chart for Selecting Commonly Used Statistical Tests



R, I O N

G NG

unequal variance t-test



Longitudinal
Model Building

Contrasts

Fixed Effects

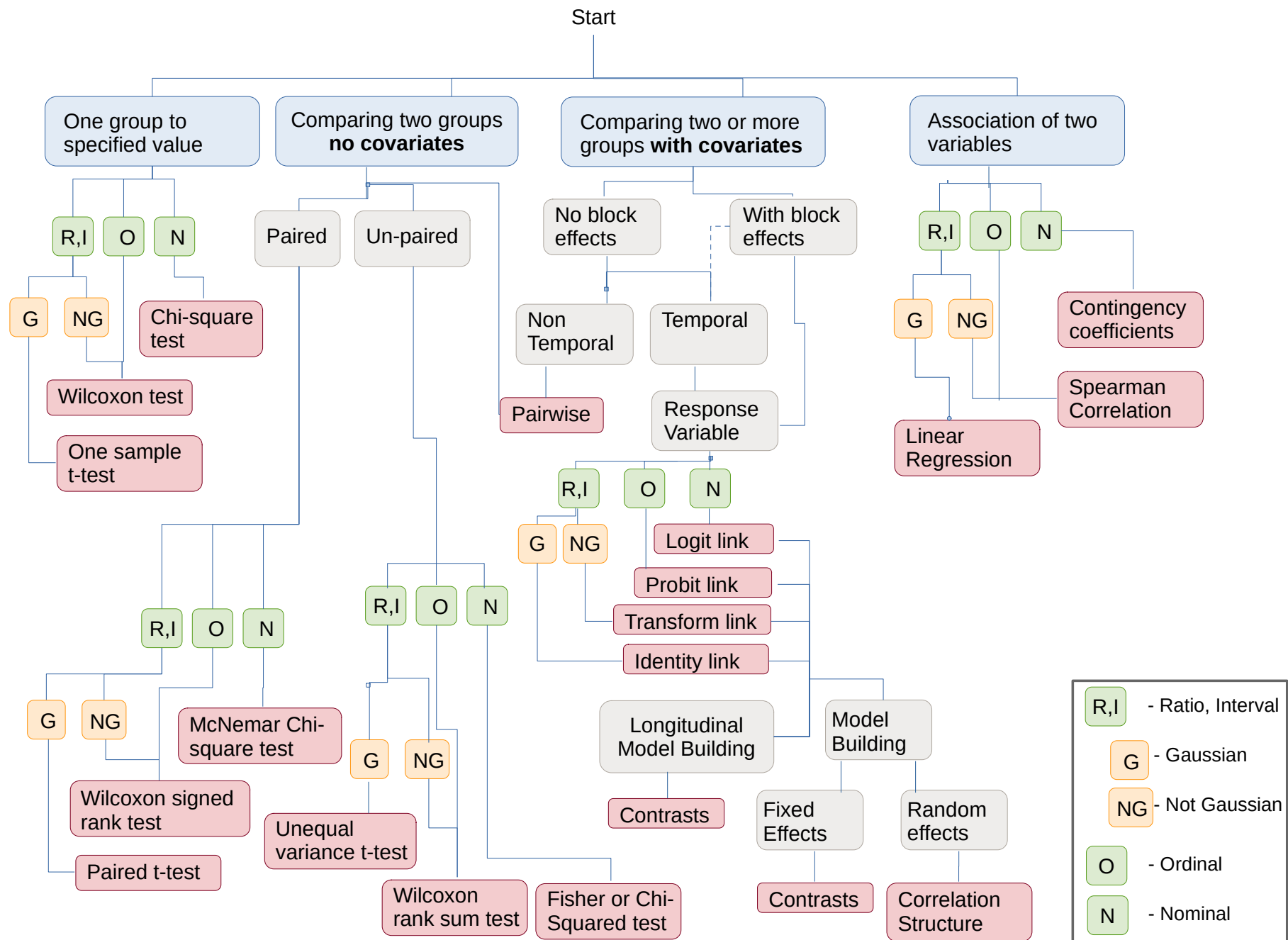
Contrasts

Wilcoxon
rank sum test

Fisher or Chi-
Squared test

are both two-level factors, Breslow-Day and Cochren-Mantel-Hanzel tes

- [illegible]



* If the outcome and main predictor are both two-level factors, Breslow-Day and Cochren-Mantel-Hanzel tests are better

R,I

- **Ratio, Interval**

- * Continuous scale measurements
- * Ex: 1.3, 3.45, 2.98, ...
 - * Plant height, CFUs, q-PCR, Watersoaking area, etc

O

- **Ordinal**

- * Ordered discrete scale measurements
- * Ex: 19, 13, 18, ...
 - * Severity scores, hull vertices, number of leaves, etc

N

- **Nominal**

- * Non-ordered discrete scale measurements
- * Ex: Red, Yellow, Cyan, ...
 - * Diseased vs not-diseased, dead vs alive, punnett squares, etc

R,I

- **Ratio, Interval**
 - Continuous scale measurements
 - Ex: 1.3, 3.45, 2.98, ...
 - Plant height, CFUs, q-PCR, Watersoaking area, etc

R,I

- **Ratio, Interval**

- Continuous scale measurements
- Ex: 1.3, 3.45, 2.98, ...
 - Plant height, CFUs, q-PCR, Watersoaking area, etc

NG

- **Not Gaussian distributed**

- Usually means non-parametric test
- If it follows a different distribution,
likelihood ratio test or other methods using
that distribution

G

- **Gaussian distributed**

- Parametric testing

R,I

- **Ratio, Interval**

- Continuous scale measurements
- Ex: 1.3, 3.45, 2.98, ...
 - Plant height, CFUs, q-PCR, Watersoaking area, etc

NG

- **Not Gaussian distributed**

- Usually means non-parametric test
- If it follows a different distribution,
likelihood ratio test or other methods using
that distribution

G

- **Gaussian distributed**
- Parametric testing

- Is data gaussian?

- 1) Visualize data
 - QQ plots
- 2) shapiro.test()
- 3) ks.test()

```
> shapiro.test(dat$Values[dat$Group == "Gaussian"])
```

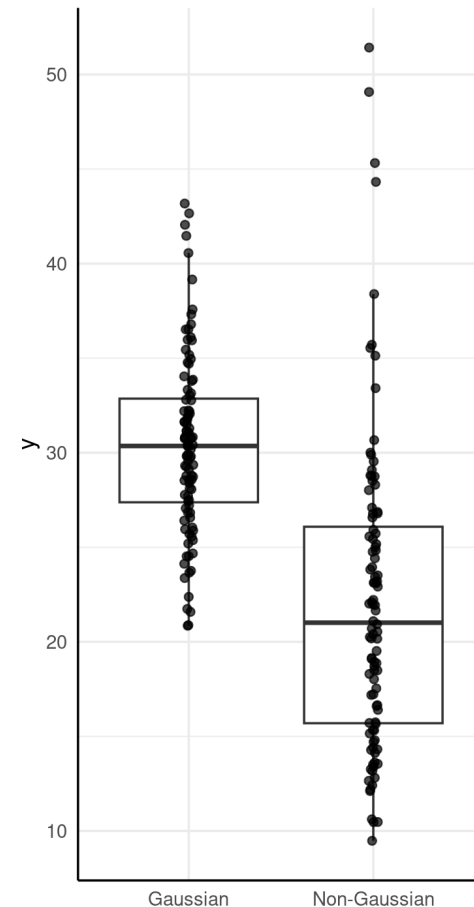
Shapiro-Wilk normality test

```
data: dat$Values[dat$Group == "Gaussian"]  
W = 0.98484, p-value = 0.3094
```

```
> shapiro.test(dat$Values[dat$Group == "Not Gaussian"])
```

Shapiro-Wilk normality test

```
data: dat$Values[dat$Group == "Not Gaussian"]  
W = 0.86899, p-value = 6.384e-08
```



R,I

- Ratio, Interval

- Continuous scale measurements
- Ex: 1.3, 3.45, 2.98, ...
 - Plant height, CFUs, q-PCR, Watersoaking area, etc

NG

- Not Gaussian distributed

- Usually means non-parametric test
- If it follows a different distribution,
likelihood ratio test or other methods using
that distribution

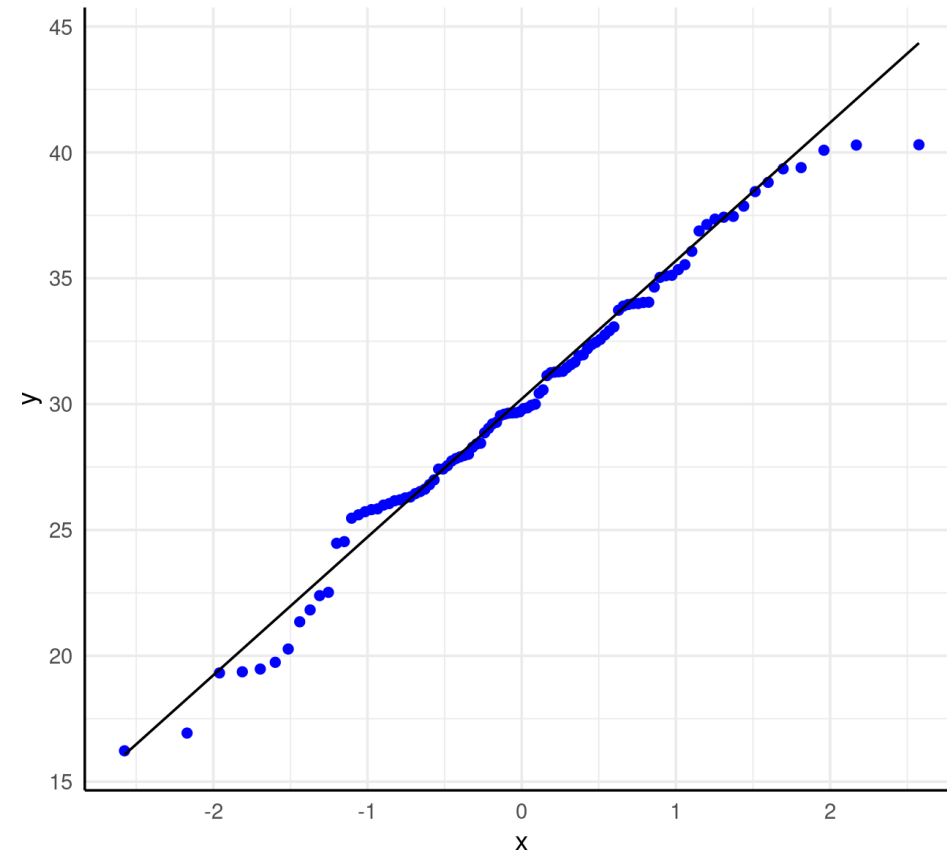
G

- Gaussian distributed
- Parametric testing

- Is data gaussian?

- 1) Visualize data
 - QQ plots
- 2) shapiro.test()
- 3) ks.test()

Gaussian Sample QQ plot



R,I

- Ratio, Interval

- Continuous scale measurements
- Ex: 1.3, 3.45, 2.98, ...
 - Plant height, CFUs, q-PCR, Watersoaking area, etc

NG

- Not Gaussian distributed

- Usually means non-parametric test
- If it follows a different distribution,
likelihood ratio test or other methods using
that distribution

G

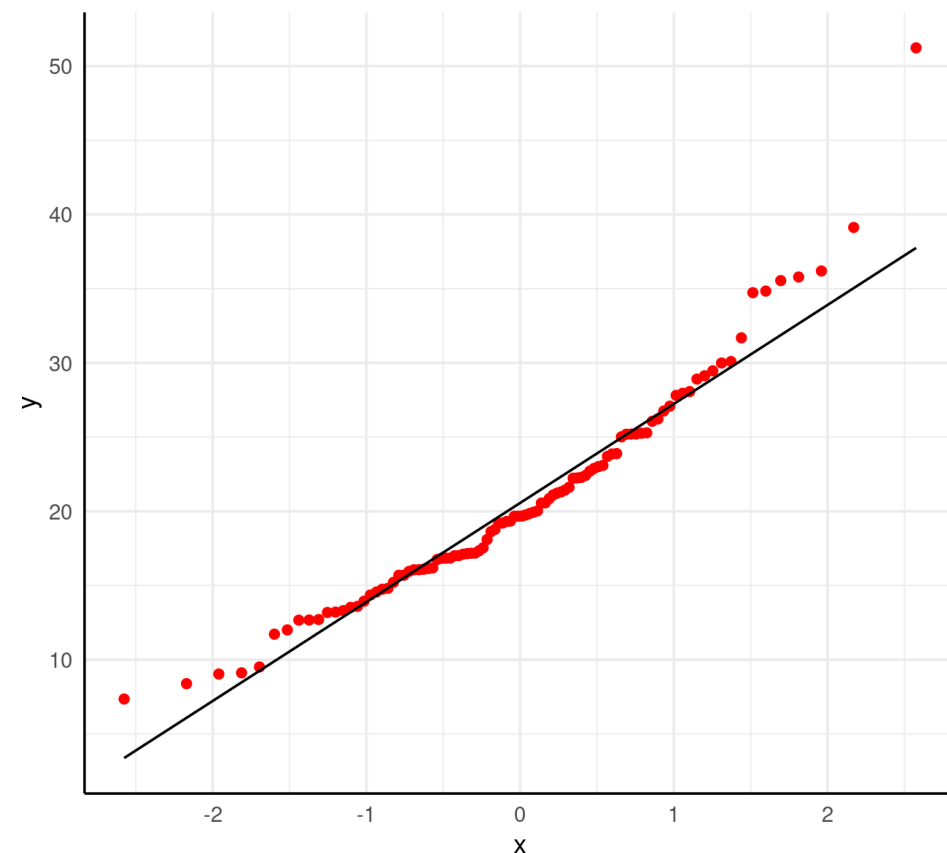
- Gaussian distributed

- Parametric testing

- Is data gaussian?

- 1) Visualize data
 - QQ plots
- 2) shapiro.test()
- 3) ks.test()

Non-Gaussian Sample QQ plot



R,I

- Ratio, Interval

- Continuous scale measurements
- Ex: 1.3, 3.45, 2.98, ...
 - Plant height, CFUs, q-PCR, Watersoaking area, etc

NG

- Not Gaussian distributed

- Usually means non-parametric test
- If it follows a different distribution,
likelihood ratio test or other methods using
that distribution

G

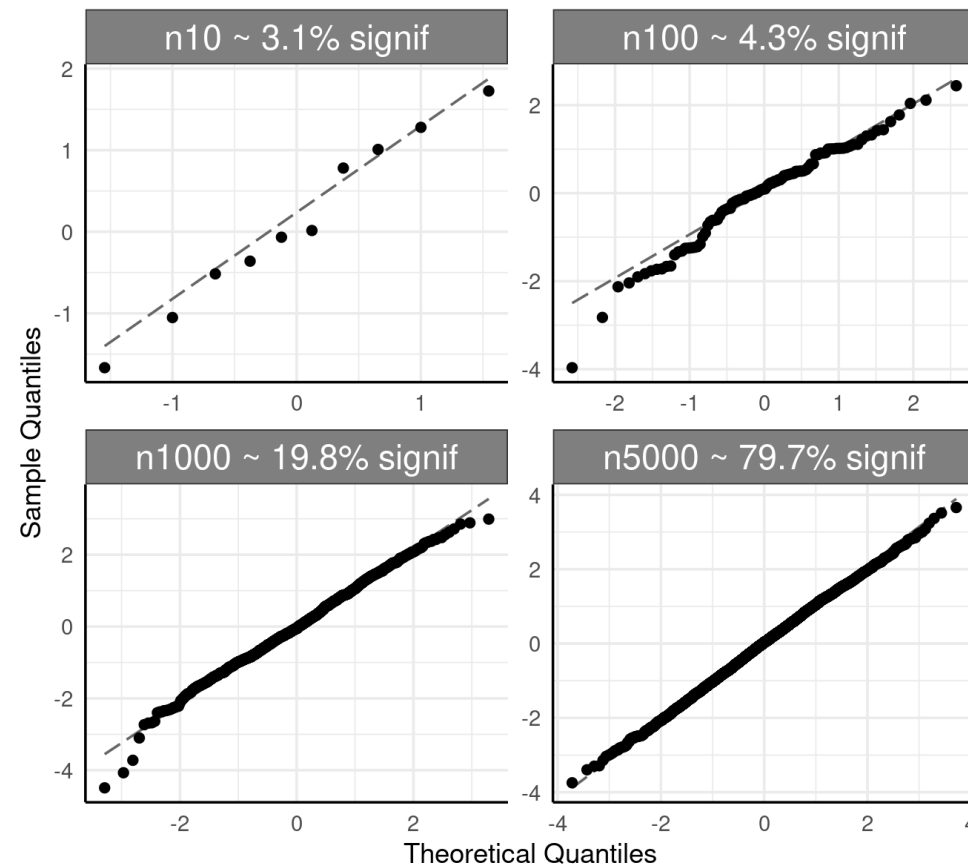
- Gaussian distributed
- Parametric testing

- Is data gaussian?

- 1) Visualize data
 - QQ plots
- 2) shapiro.test()
- 3) ks.test()

Broadly,
visualizations are
more useful than
tests of normality.

Percentage of Significant Shapiro Tests



R,I

- Ratio, Interval

- Continuous scale measurements

- Ex: 1.3, 3.45, 2.98, ...

- Plant height, CFUs, q-PCR, Watersoaking area, etc

NG

- Not Gaussian distributed

- Usually means non-parametric test

- If it follows a different distribution,
likelihood ratio test or other methods using
that distribution

G

- Gaussian distributed

- Parametric testing

- Is data gaussian?

1) Visualize data

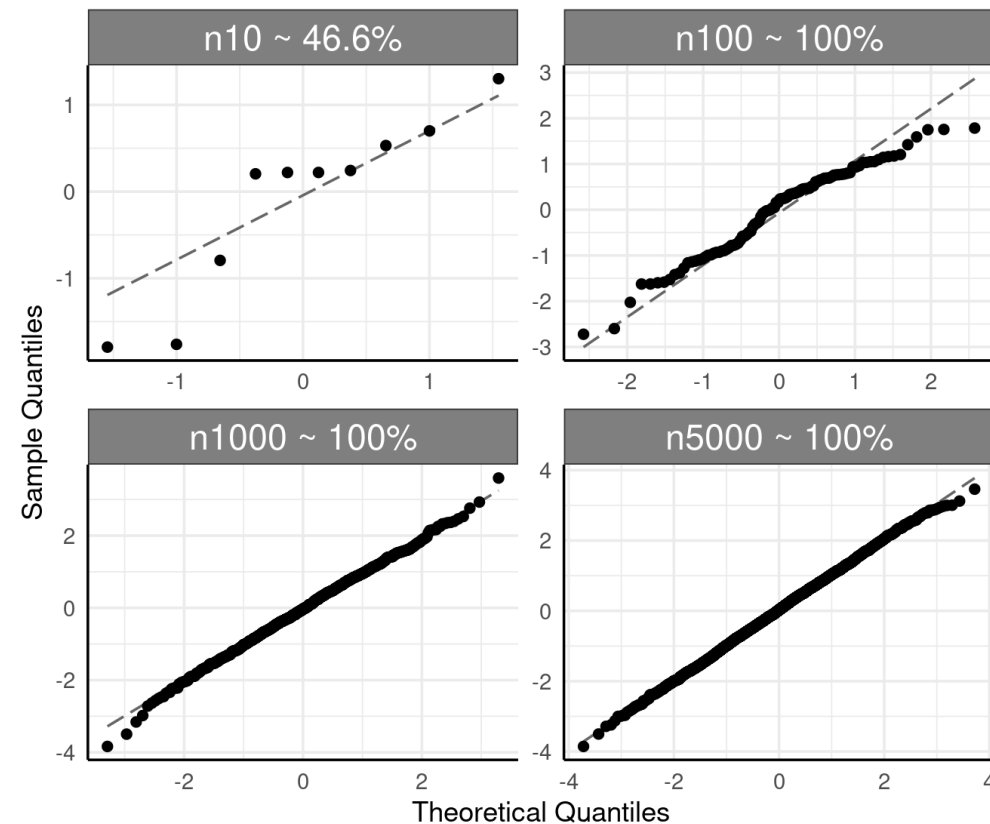
- QQ plots

2) shapiro.test()

3) ks.test()

Broadly,
visualizations are
more useful than
tests of normality.

Percentage of Significant KS Tests



R,I

- Ratio, Interval

- Continuous scale measurements
- Ex: 1.3, 3.45, 2.98, ...
 - Plant height, CFUs, q-PCR, Watersoaking area, etc

NG

- Not Gaussian distributed

- Usually means non-parametric test
- If it follows a different distribution,
likelihood ratio test or other methods using
that distribution

G

- Gaussian distributed

- Parametric testing

O

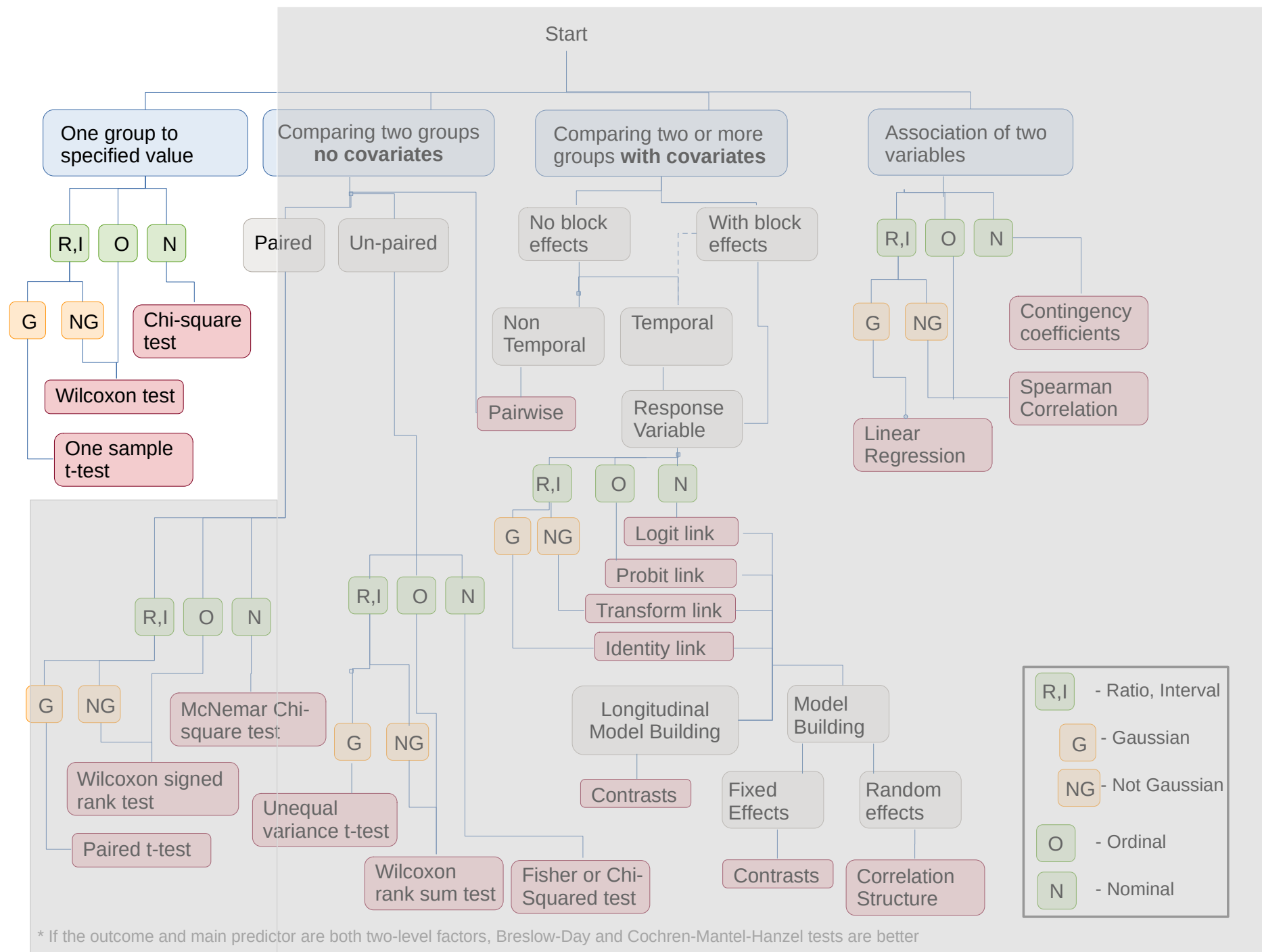
- Ordinal

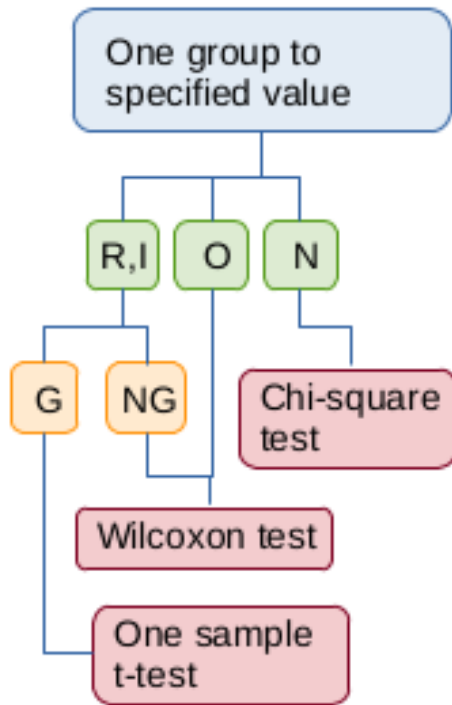
- Ordered discrete scale measurements
- Ex: 19, 13, 18, ...
 - Severity scores, hull vertices, number of leaves, etc

N

- Nominal

- Non-ordered discrete scale measurements
- Ex: Red, Yellow, Cyan, ...
 - Diseased vs not-diseased, dead vs alive, punnett squares, etc

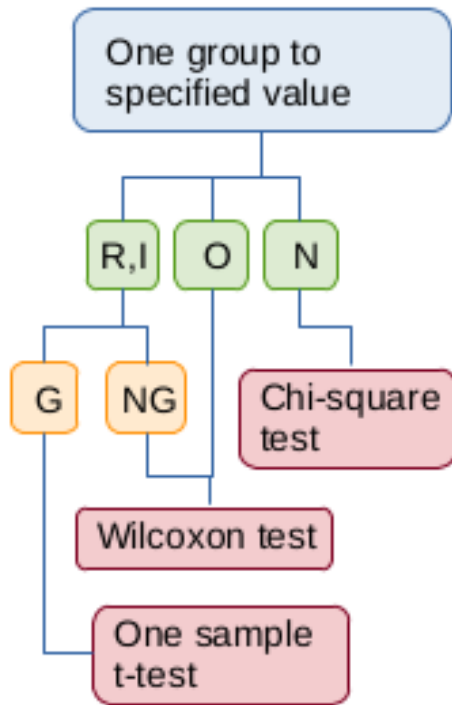




Scenario 1

You read a paper that says the average root length in arabidopsis 10 days after planting is 5.8cm. Your own work has found a gene complex that is associated to root length and you knock out one of the components to test it's effect in this phenotype. You do 20 reps and compare to this reported value.

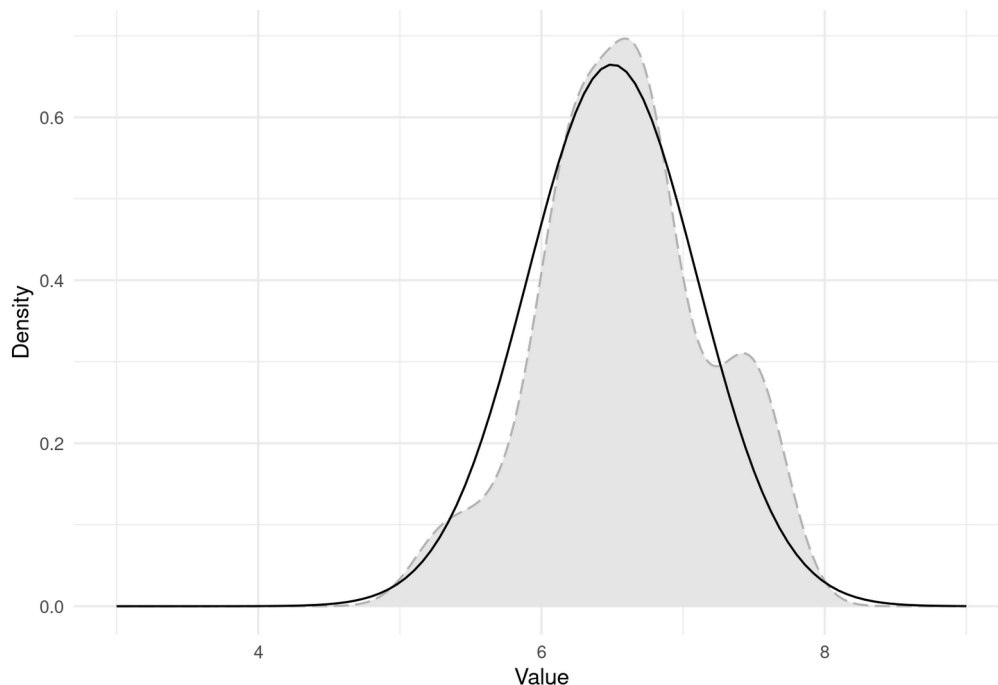
- What are the steps that lead to the right test?



Scenario 1

You read a paper that says the average root length in arabidopsis 10 days after planting is 5.8cm. Your own work has found a gene complex that is associated to root length and you knock out one of the components to test it's effect in this phenotype. You do 20 reps and compare to this reported value.

- What are the steps that lead to the right test?

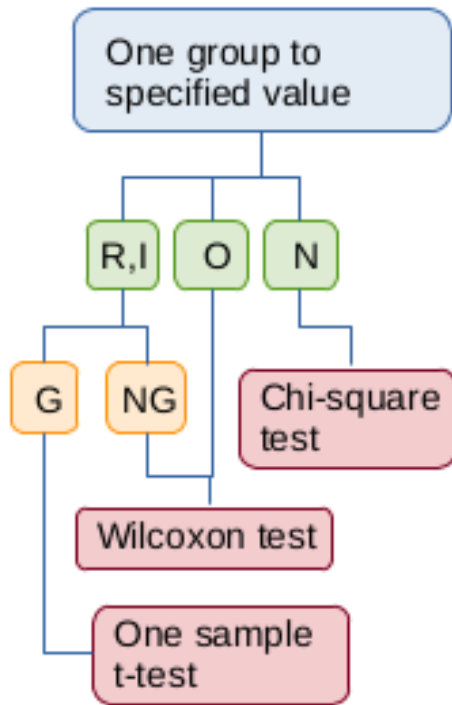


```
> shapiro.test(s1)
```

Shapiro-Wilk normality test

data: s1

W = 0.96858, p-value = 0.7247



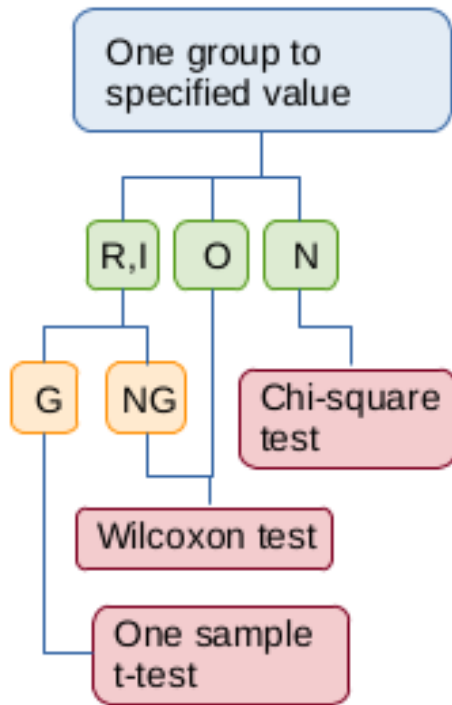
Scenario 1

You read a paper that says the average root length in arabidopsis 10 days after planting is 5.8cm. Your own work has found a gene complex that is associated to root length and you knock out one of the components to test it's effect in this phenotype. You do 20 reps and compare to this reported value.

- What are the steps that lead to the right test?

Answer

Root length is measured continuously so we have R,I data. Our data looks gaussian and the shapiro test agrees, so we use a one sample T test.



Scenario 1

You read a paper that says the average root length in arabidopsis 10 days after planting is 5.8cm. Your own work has found a gene complex that is associated to root length and you knock out one of the components to test it's effect in this phenotype. You do 20 reps and compare to this reported value.

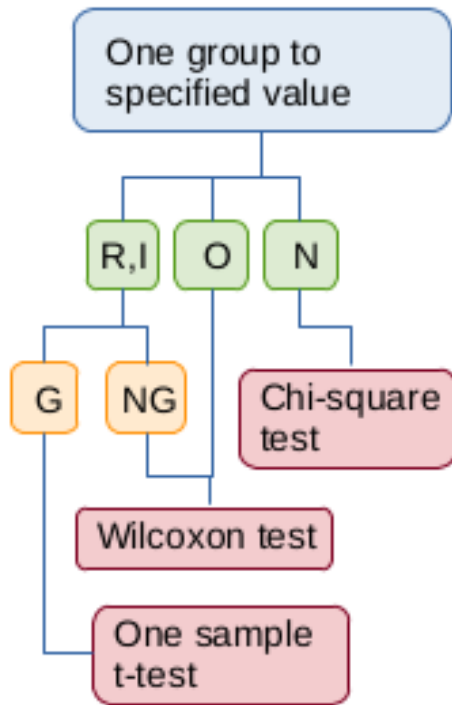
- What are the steps that lead to the right test?

R

```
> t.test(s1, mu=5.8)
```

One Sample t-test

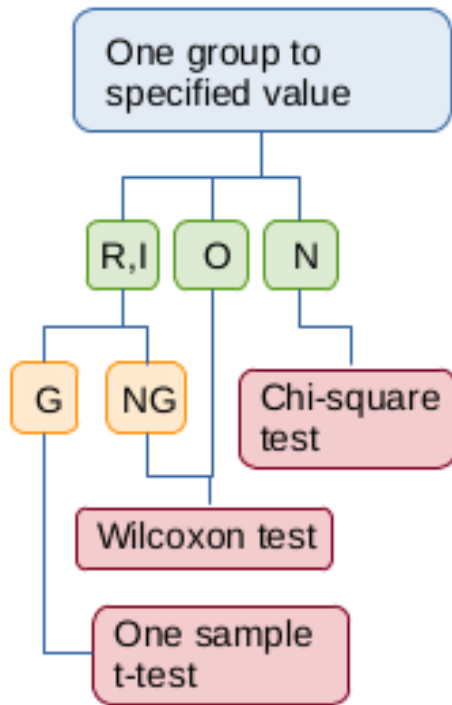
```
data: s1
t = 6.0153, df = 19, p-value = 8.693e-06
alternative hypothesis: true mean is not equal to 5.8
95 percent confidence interval:
 6.311841 6.858107
sample estimates:
mean of x
 6.584974
```



Scenario 2

- You read a paper that found a SNP marker in *Gossypium arboreum* (diploid) that effectively doubles the amount of yield when homozygous recessive but it's unclear if the population of cotton in a particular field is under Hardy-Weinberg equilibrium for this allele. Because you're so excited about this, you go to the field and genotype 1000 plants to test if they are in HWE.

-What are the steps that lead to the right test?



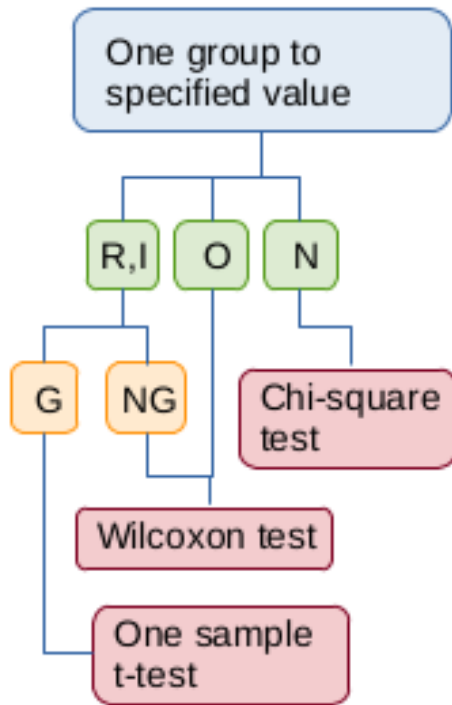
Scenario 2

- You read a paper that found a SNP marker in *Gossypium arboreum* (diploid) that effectively doubles the amount of yield when homozygous recessive but it's unclear if the population of cotton in a particular field is under Hardy-Weinberg equilibrium for this allele. Because you're so excited about this, you go to the field and genotype 1000 plants to test if they are in HWE.

-What are the steps that lead to the right test?

Answer

Genotypes fall into aa, Aa, and AA categories which are typically thought of as unordered so you follow the **N** path. Hardy-Weinberg equilibrium means the genotypes are in a ratio of 1 : 2 : 1 and you use these as the expected ratios for the correct test which is a **Chi-Square Test**.



Scenario 2

- You read a paper that found a SNP marker in *Gossypium arboreum* (diploid) that effectively doubles the amount of yield when homozygous recessive but it's unclear if the population of cotton in a particular field is under Hardy-Weinberg equilibrium for this allele. Because you're so excited about this, you go to the field and genotype 1000 plants to test if they are in HWE.

-What are the steps that lead to the right test?

R

```
> head(s2)
```

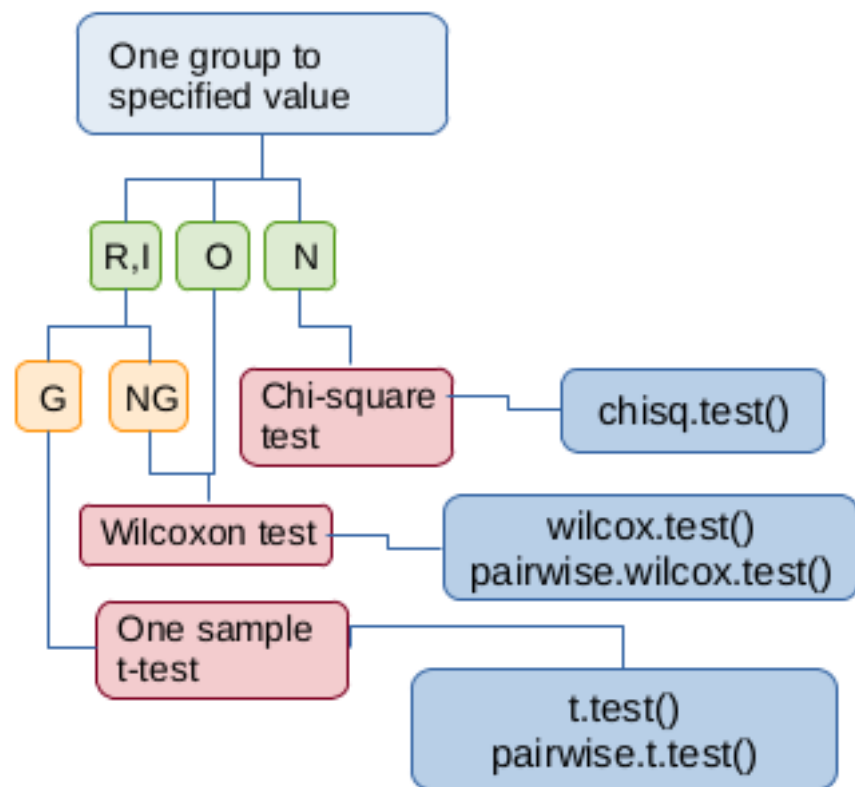
```
[1] "Aa" "Aa" "Aa" "Aa" "Aa" "aa"
```

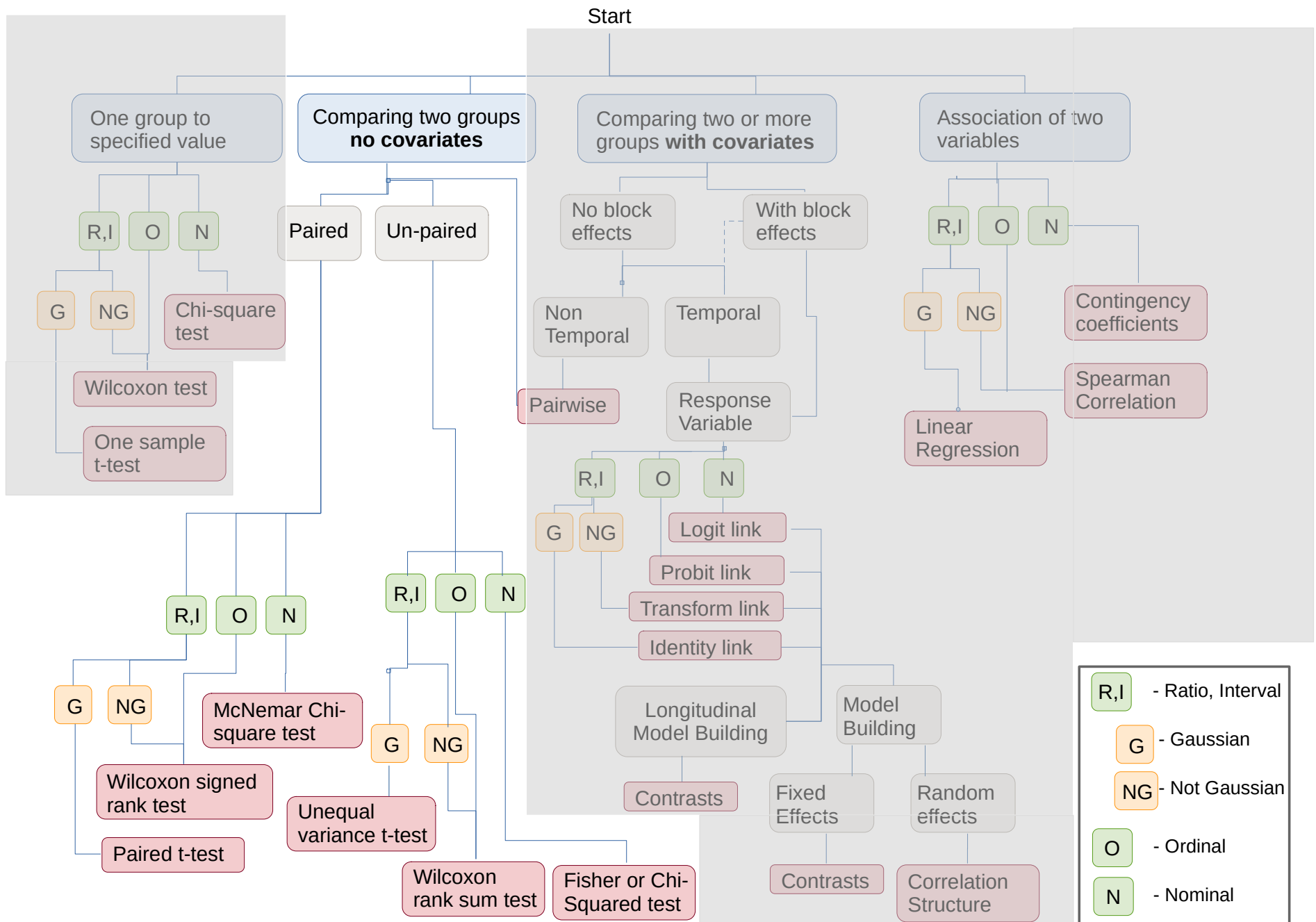
```
> chisq.test(table(s2), p = c(0.25, 0.5, 0.25) )
```

Chi-squared test for given probabilities

```
data: table(s2)
```

```
X-squared = 2.822, df = 2, p-value = 0.2439
```





* If the outcome and main predictor are both two-level factors, Breslow-Day and Cochren-Mantel-Hanzel tests are better

Paired

- Sometimes called “**before and after**” data.
 - This is for when you measure every replicate before an intervention and then again after. This accounts for variance between individuals, which are assumed to be independent.

Un-paired

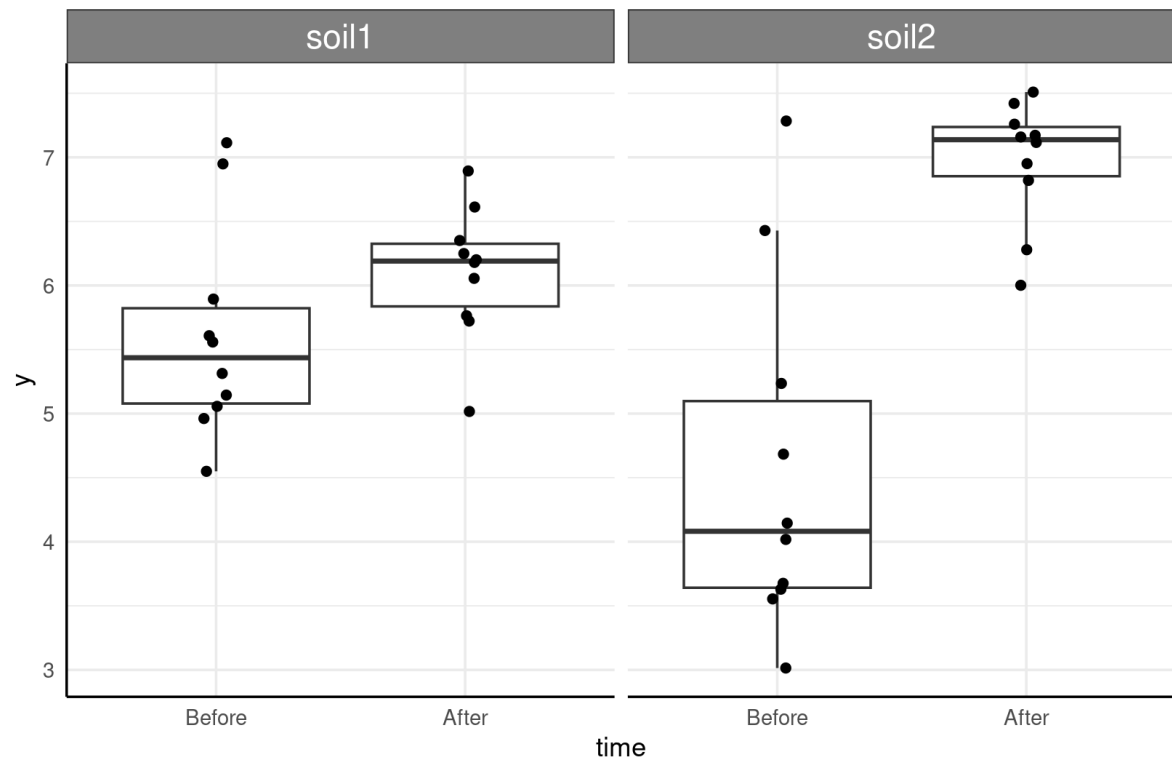
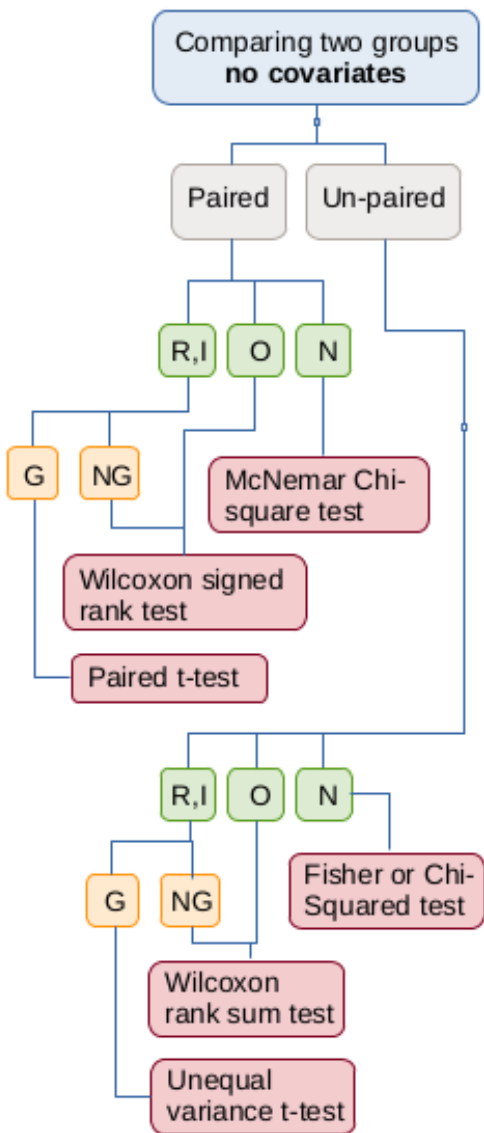
- This is the most common data that is collected.
 - Individual replicates are assumed to be independent.

When more than two timepoints are used the data is considered longitudinal which introduces another type of correlation structure.

Scenario 3

You'd like to compare the soil moisture readings of two soil types each having 10 reps both before and after adding 200ml of water and mixing thoroughly. You'd like to test if the moisture reading is significantly different for either soil after the water is added.

- What are the steps that lead to the right test?



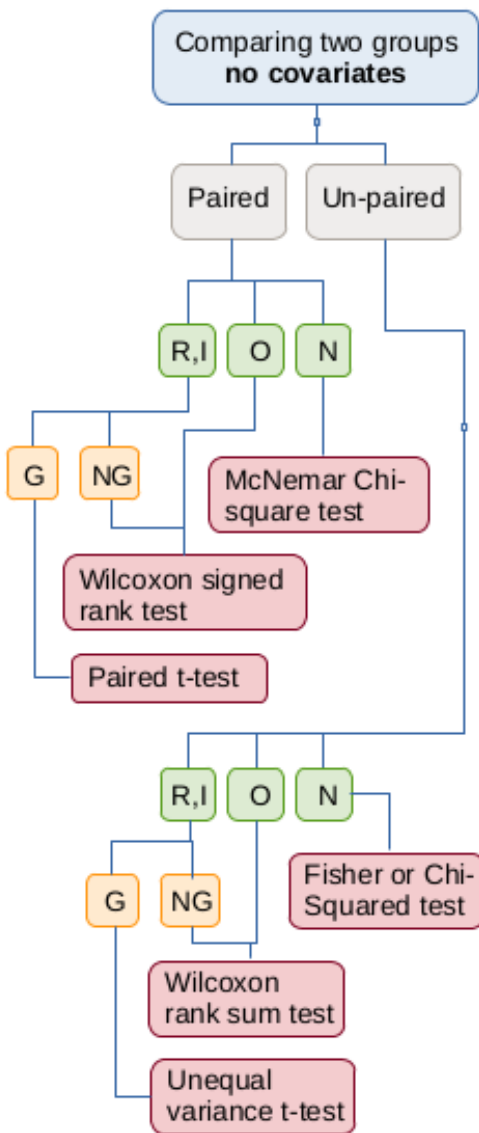
Scenario 3

You'd like to compare the soil moisture readings of two soil types each having 10 reps both before and after adding 200ml of water and mixing thoroughly. You'd like to test if the moisture reading is significantly different for either soil after the water is added.

- What are the steps that lead to the right test?

Answer

First we need to recognize that this is paired data since we measured every rep before and after we added water so we'll follow the **Paired** path. Next we determine that the measurement we're taking falls in the continuous scales so we follow the **R,I** path. Then we use boxplots to discover that the observations are **not Gaussian** which leads us to a **Wilcoxon signed rank test**.



Scenario 3

You'd like to compare the soil moisture readings of two soil types each having 10 reps both before and after adding 200ml of water and mixing thoroughly. You'd like to test if the moisture reading is significantly different for either soil after the water is added.

- What are the steps that lead to the right test?

R

```
> wilcox.test(x=s1_1, y = s1_2, paired=TRUE)
```

Wilcoxon signed rank exact test

data: s1_1 and s1_2

V = 9, p-value = 0.06445

alternative hypothesis: true location shift is not equal to 0

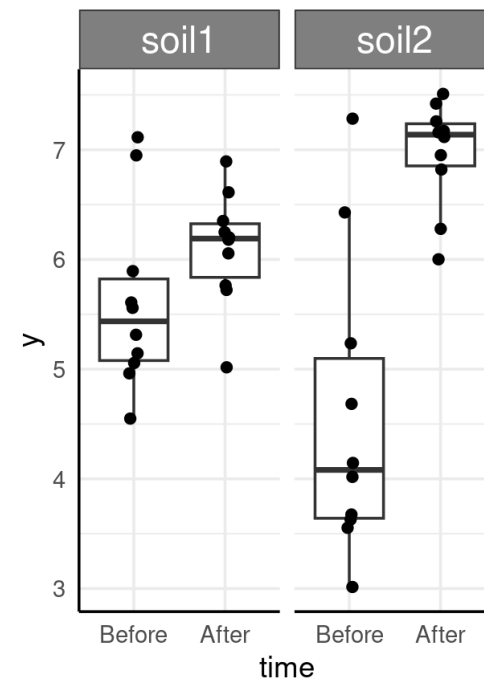
```
> wilcox.test(x=s2_1, y = s2_2, paired=TRUE)
```

Wilcoxon signed rank exact test

data: s2_1 and s2_2

V = 1, p-value = 0.003906

alternative hypothesis: true location shift is not equal to 0



Why not transform the data?

Transformation should generally not be the first thing you try with non-gaussian data, non-parametric testing will make your life easier and your results better.

Keeping data in the original parameter space make for easier interpretation. Once you use a transformation it is very difficult to think about effect sizes and variance at the original scale.

Why not transform the data?

Transformation should generally not be the first thing you try with non-gaussian data, non-parametric testing will make your life easier and your results better.

Keeping data in the original parameter space make for easier interpretation. Once you use a transformation it is very difficult to think about effect sizes and variance at the original scale.

If you still have to transform the data

Non-parametric tests are less powerful so you may sometimes need to transform your data. If you do then only use a **log**, **exponential**, **square-root**, or **square** transform depending on how your data look.

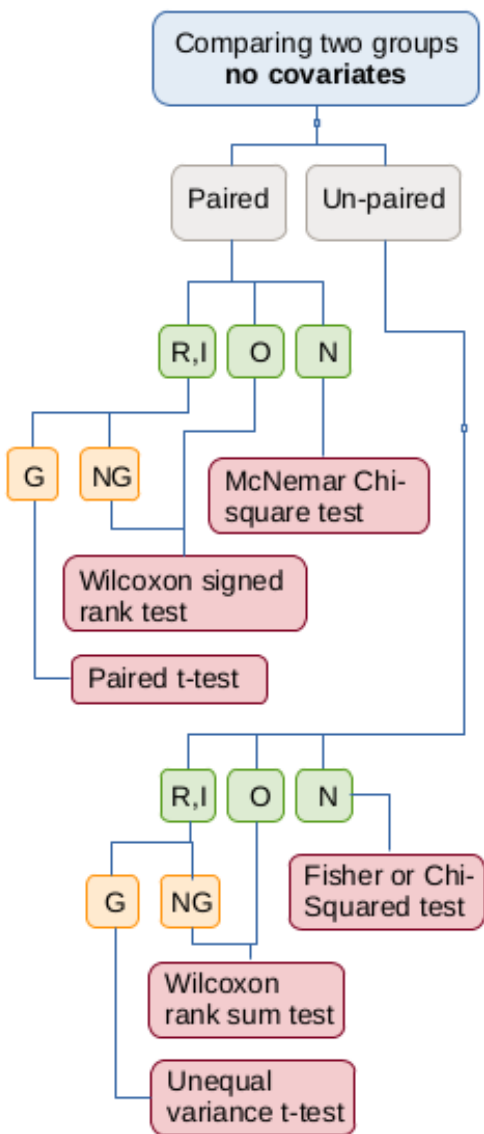
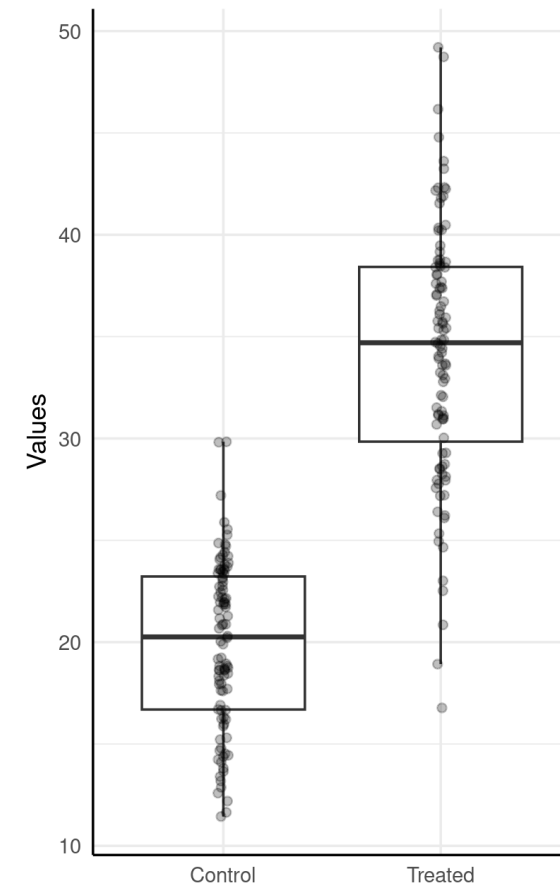
These transformations are most mathematically sound because of the relationships between the gaussian and other distributions. That will make for better and more interpretable results than the more outlandish transformations.



Scenario 4

- You're measuring plant height in a control (low N) and treated with a nitrogen fixing bacteria in 100 plants per condition. You have let the plants grow for two weeks and measure the height on the last day.

- What are the steps that lead to the right test?



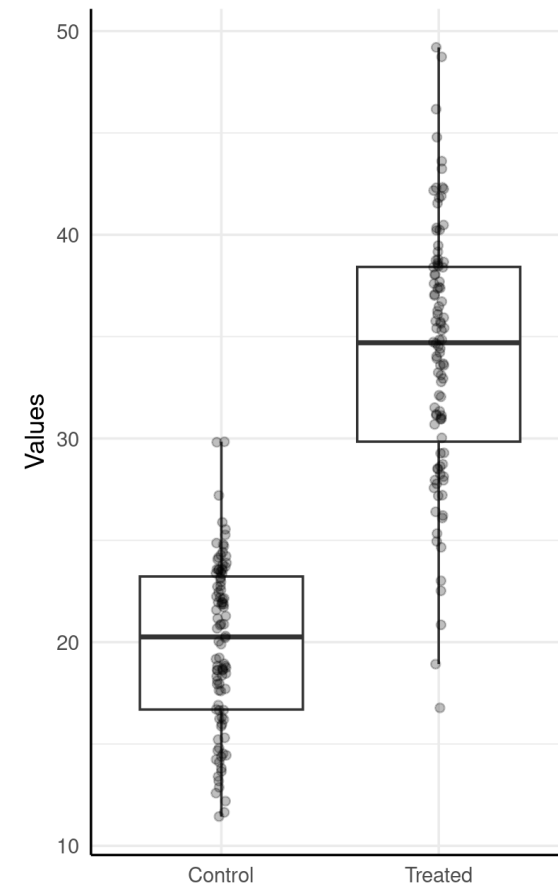
Scenario 4

- You're measuring plant height in a control (low N) and treated with a nitrogen fixing bacteria in 100 plants per condition. You have let the plants grow for two weeks and measure the height on the last day.

- What are the steps that lead to the right test?

Answer

These data do not compare times so they are **unpaired**. Plant height is measured continuously so we have **R,I** data. Our data looks very **gaussian** so we pick the **Unequal Variance T Test**.



Scenario 4

- You're measuring plant height in a control (low N) and treated with a nitrogen fixing bacteria in 100 plants per condition. You have let the plants grow for two weeks and measure the height on the last day.

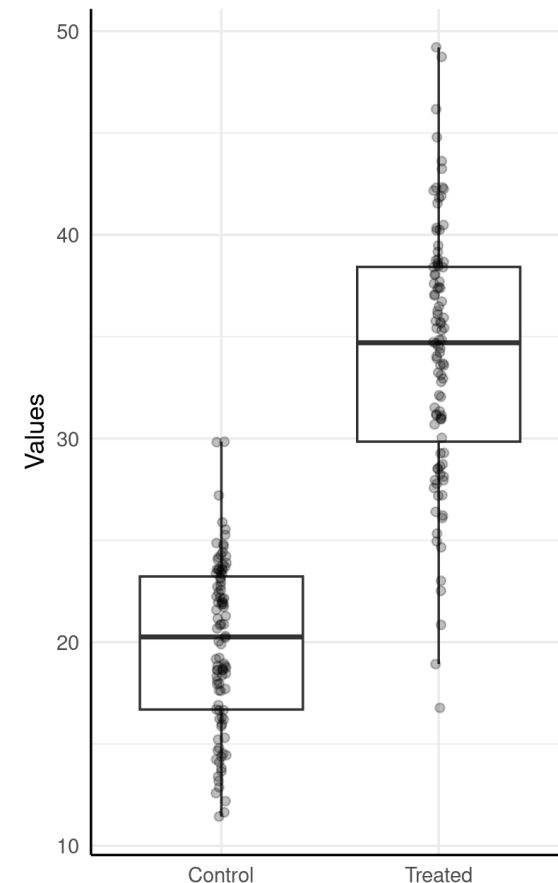
- What are the steps that lead to the right test?

R

```
> t.test(control, treated)
```

Welch Two Sample t-test

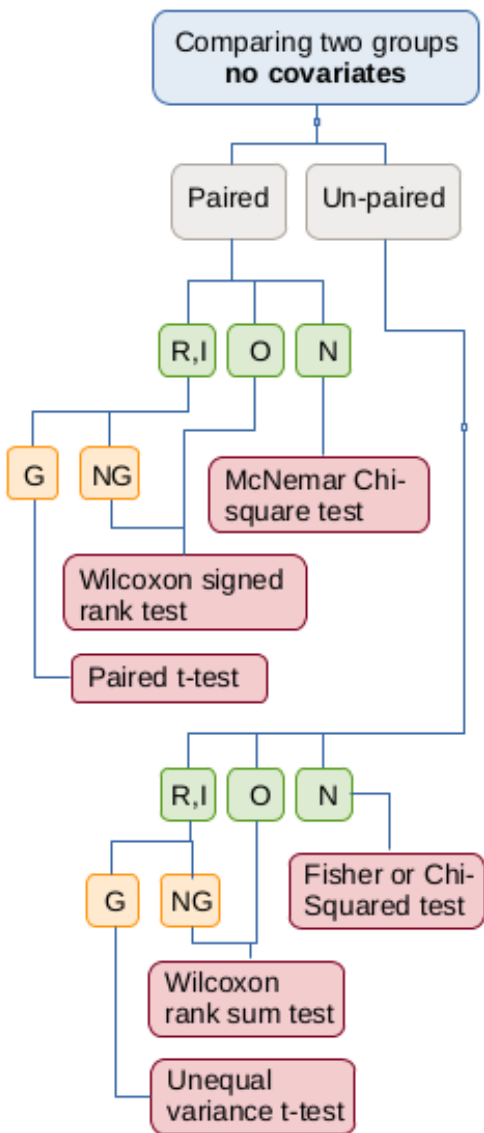
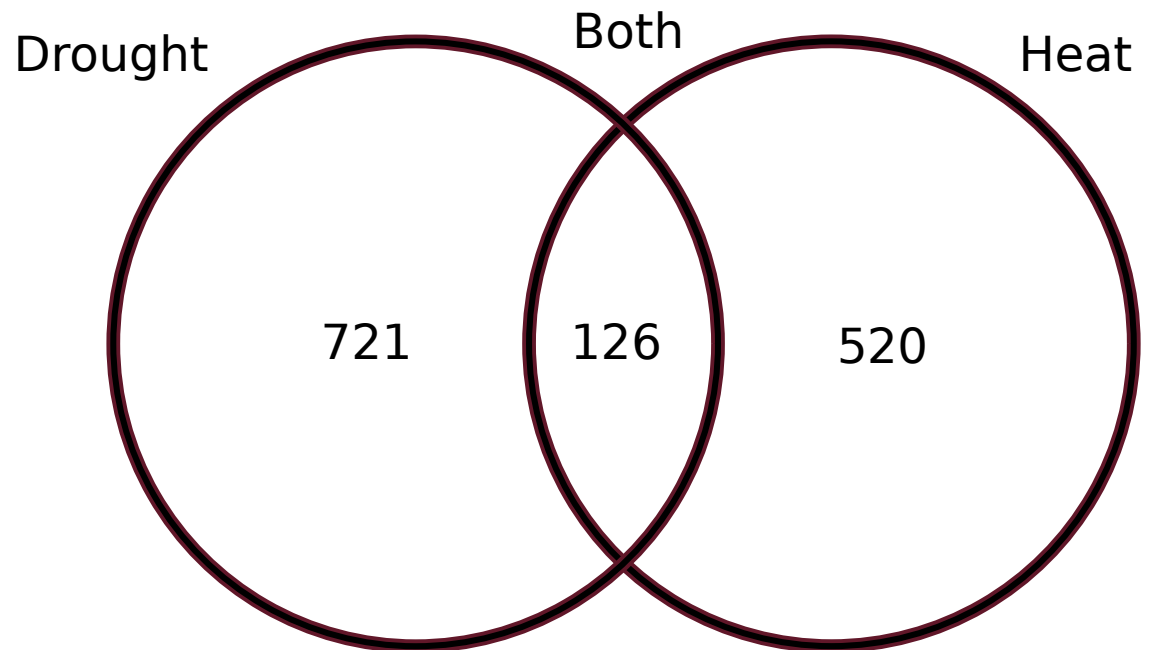
```
data: control and treated
t = -20.412, df = 166.78, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.34651 -12.63968
sample estimates:
mean of x mean of y
 20.36162  34.35472
```



Scenario 5

You do a transcriptomics experiment to identify upregulated genes in both heat (35C, 80% WC) and drought (30C, 50% WC) treatments relative to a common control (30C, 80% WC). For each of the two treatments you have upregulated genes, some of which overlap and you want to test if the overlap is due to random chance.

- What are the steps to test this hypothesis?



Scenario 5

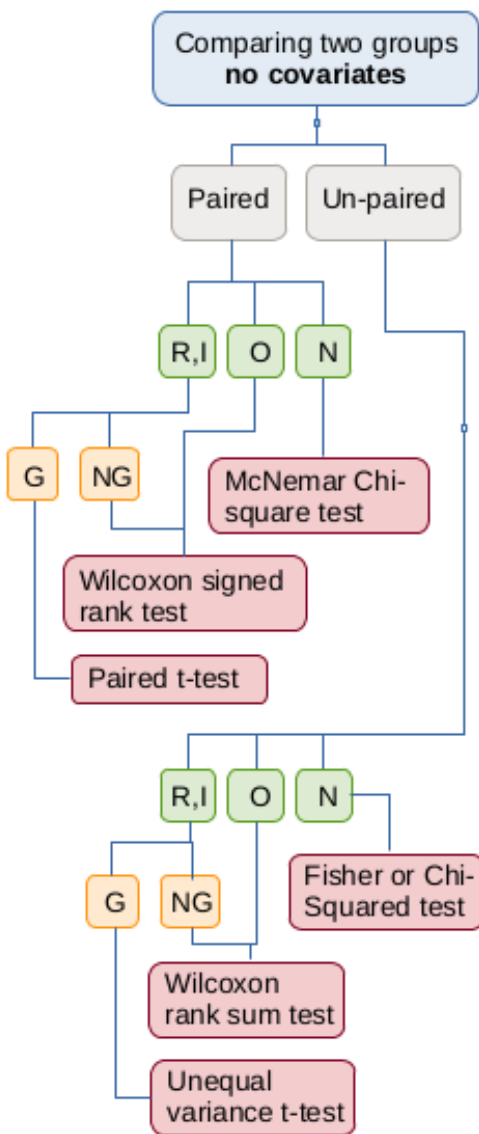
You do a transcriptomics experiment to identify upregulated genes in both heat (35C, 80% WC) and drought (30C, 50% WC) treatments relative to a common control (30C, 80% WC). For each of the two treatments you have upregulated genes, some of which overlap and you want to test if the overlap is due to random chance.

- What are the steps to test this hypothesis?

Answer

* These data are unpaired and the counts fall into three distinct, unordered categories so we follow the **N** path to the **Chi-Square Test**.

If all the counts are above 5, then Chi-square and Fisher produce the same results. If any counts are less than 5, use Fisher.



Scenario 5

You do a transcriptomics experiment to identify upregulated genes in both heat (35C, 80% WC) and drought (30C, 50% WC) treatments relative to a common control (30C, 80% WC). For each of the two treatments you have upregulated genes, some of which overlap and you want to test if the overlap is due to random chance.

- What are the steps to test this hypothesis?

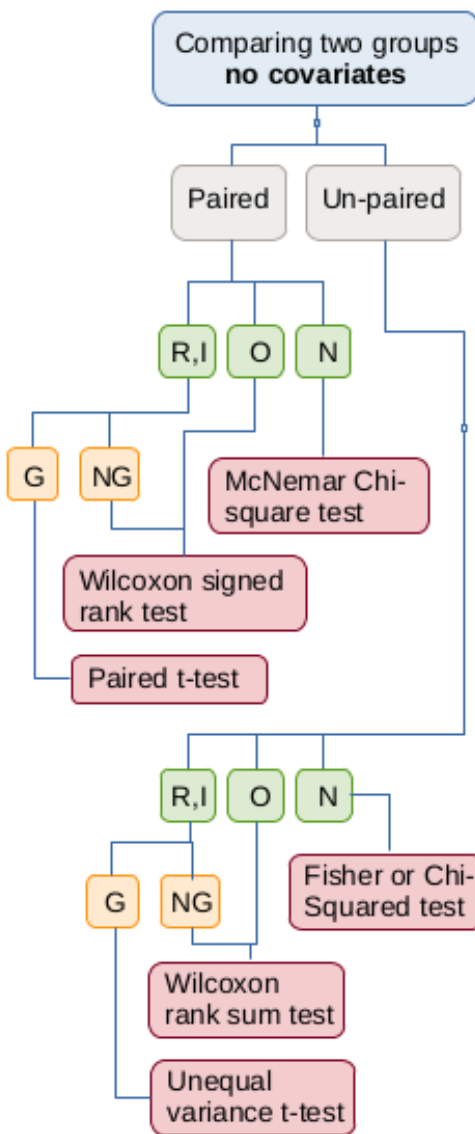
R

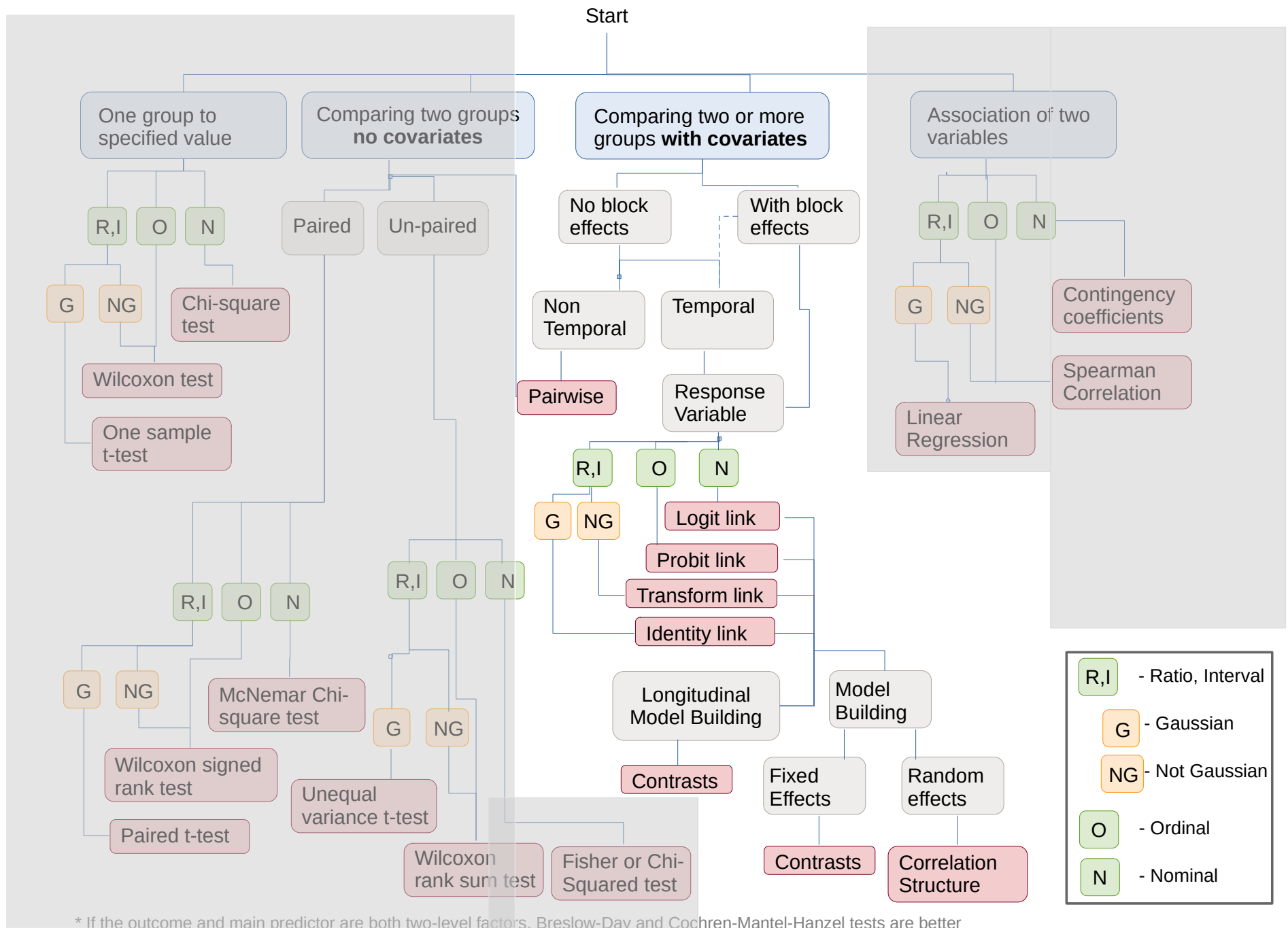
```
> obs <- c(721, 126, 520)
> chisq.test(obs)
```

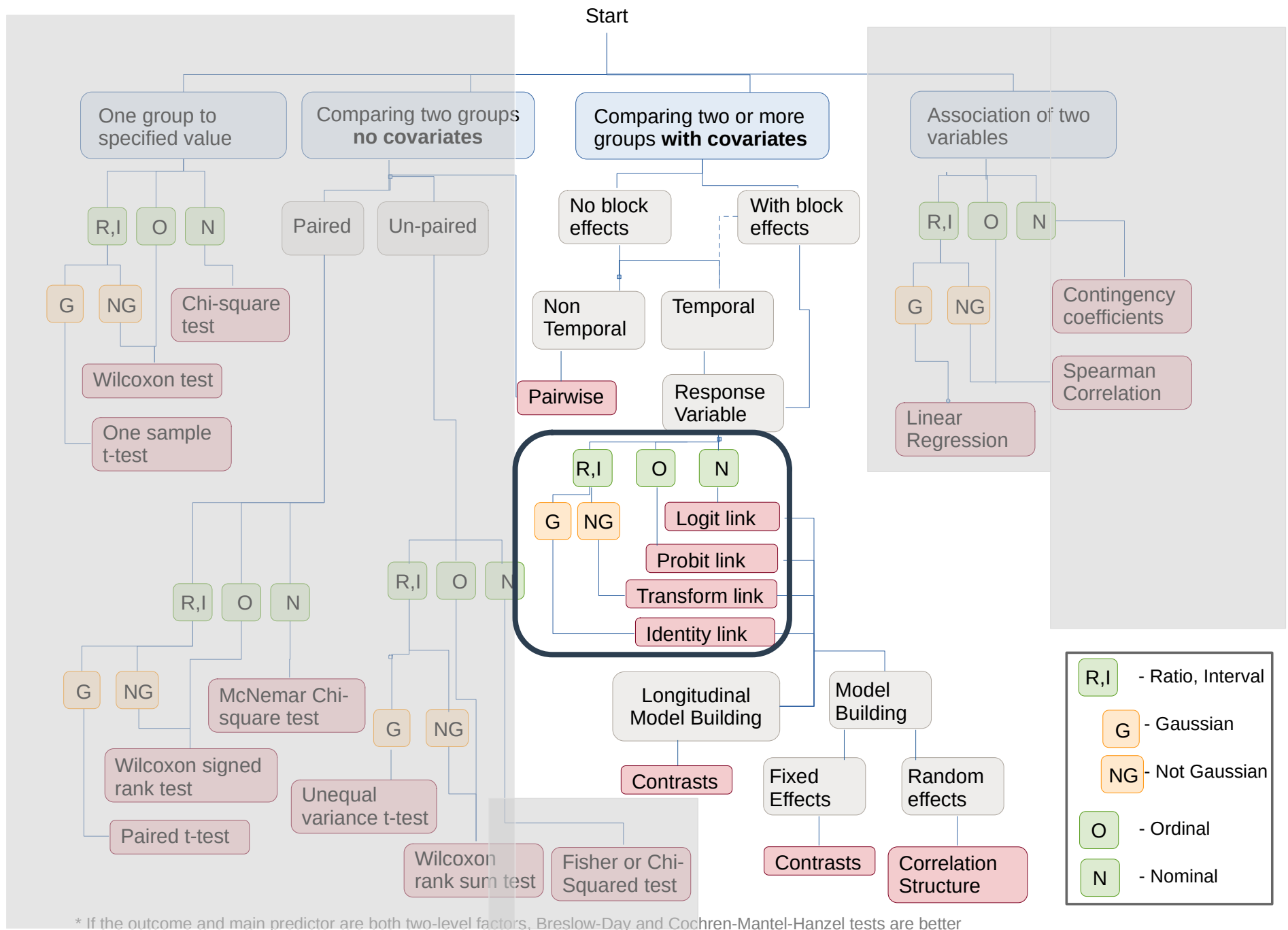
Chi-squared test for given probabilities

```
data: obs
```

```
X-squared = 402.09, df = 2, p-value < 2.2e-16
```







No block
effects

With block
effects

Temporal

Non-
Temporal

Fixed Effects

Random
Effects

Link Function

No block effects

* Cofactor DOES NOT effect the components of the other design parameters equally

With block effects

* Cofactor DOES effect the components of the other design parameters equally

Temporal

Non-Temporal

Fixed Effects

Random Effects

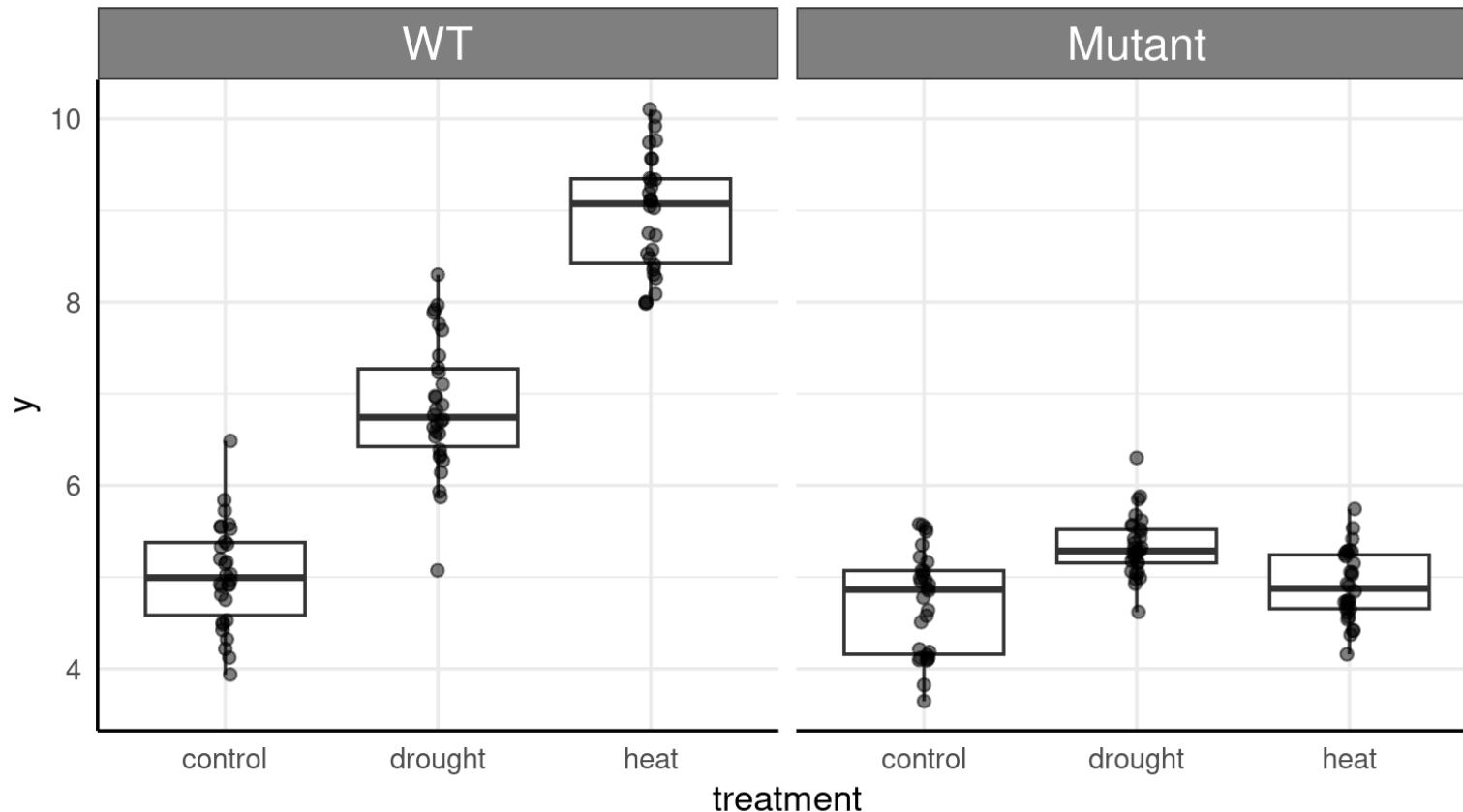
Link Function

No block effects

* Cofactor DOES NOT effect the components of the other design parameters equally

With block effects

* Cofactor DOES effect the components of the other design parameters equally



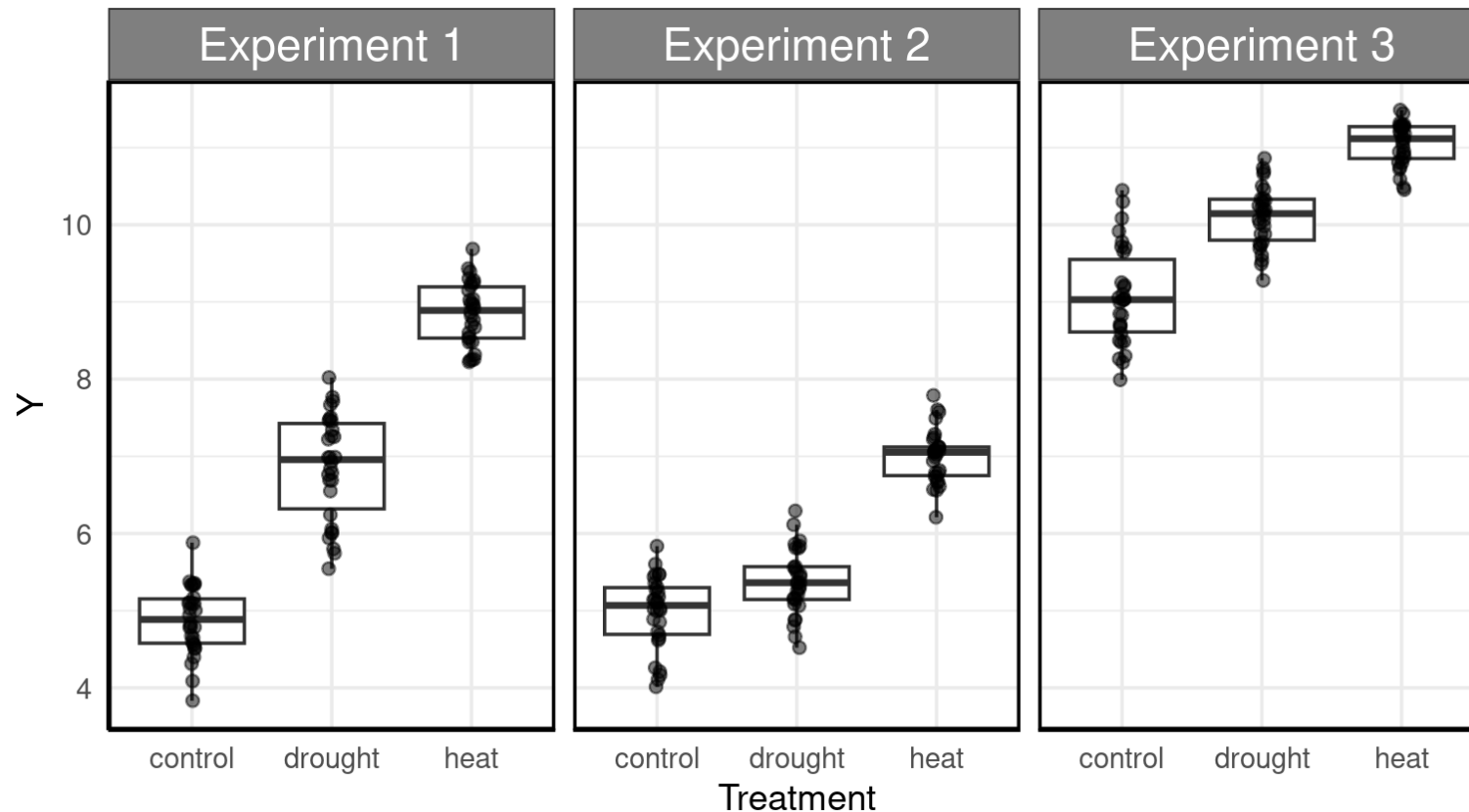
This is an interaction effect, not a blocking effect. Our controls are at the same scale but genotypes are responding differently.

No block effects

* Cofactor DOES NOT effect the components of the other design parameters equally

With block effects

* Cofactor DOES effect the components of the other design parameters equally



Looking at experiment 3 there is a clear blocking effect. Something was different but we can still use the data.

No block effects

* Cofactor DOES NOT effect the components of the other design parameters equally

With block effects

* Cofactor DOES effect the components of the other design parameters equally

Temporal

* Individuals' data is collected over time (>2 timepoints)

Non-Temporal

* Individuals are not measured over time

Fixed Effects

Random Effects

Link Function

Temporal

* Individuals' data is collected over time (>2 timepoints)

Non-Temporal

* Individuals are not measured over time

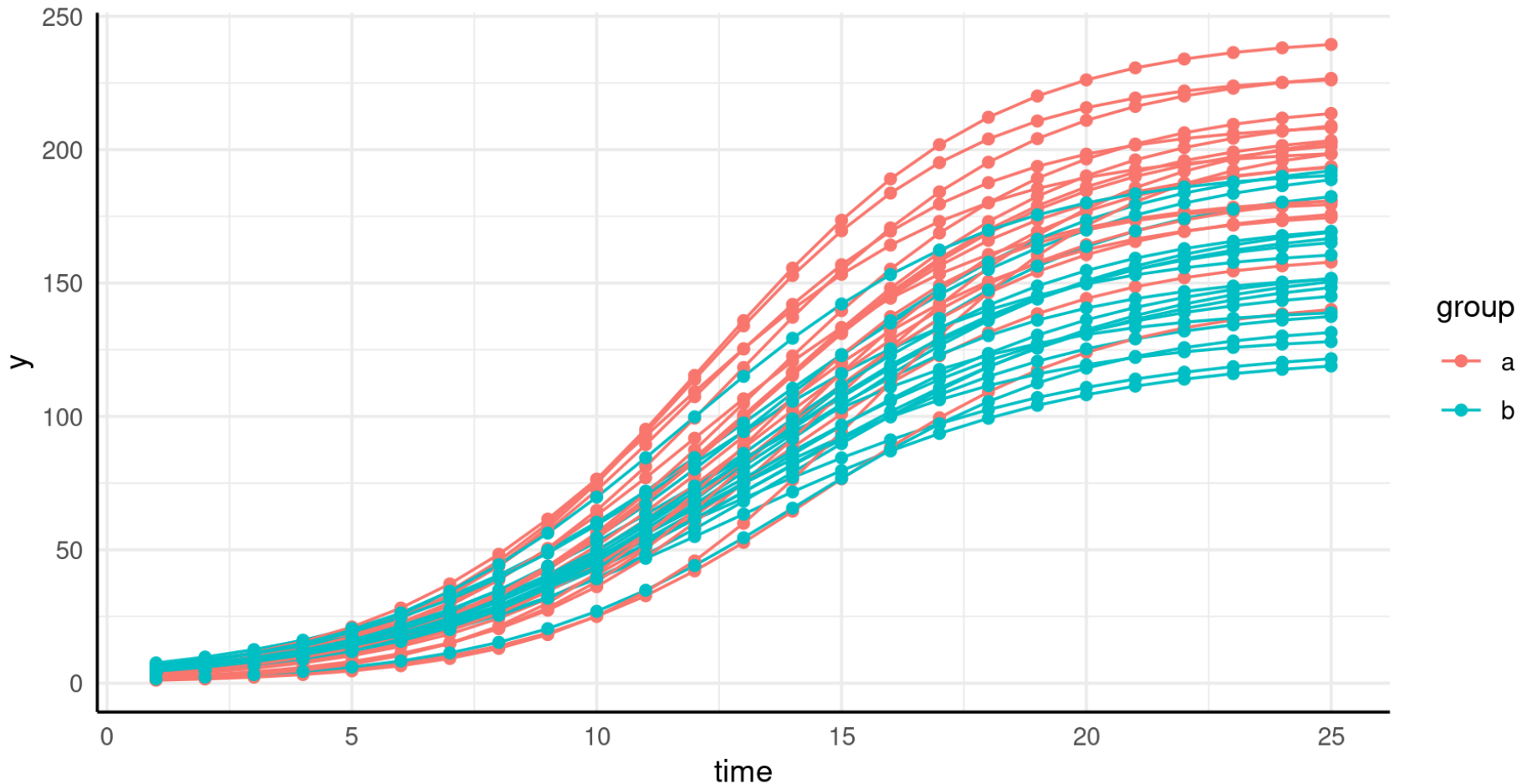
Temporal

* Individuals' data is collected over time (>2 timepoints). Note if you can use `ggplot2::geom_line` that's a hint that your data are likely longitudinal.

Non-Temporal

* Individuals are not measured over time

Longitudinal Data



Temporal

* Individuals' data is collected over time (>2 timepoints)

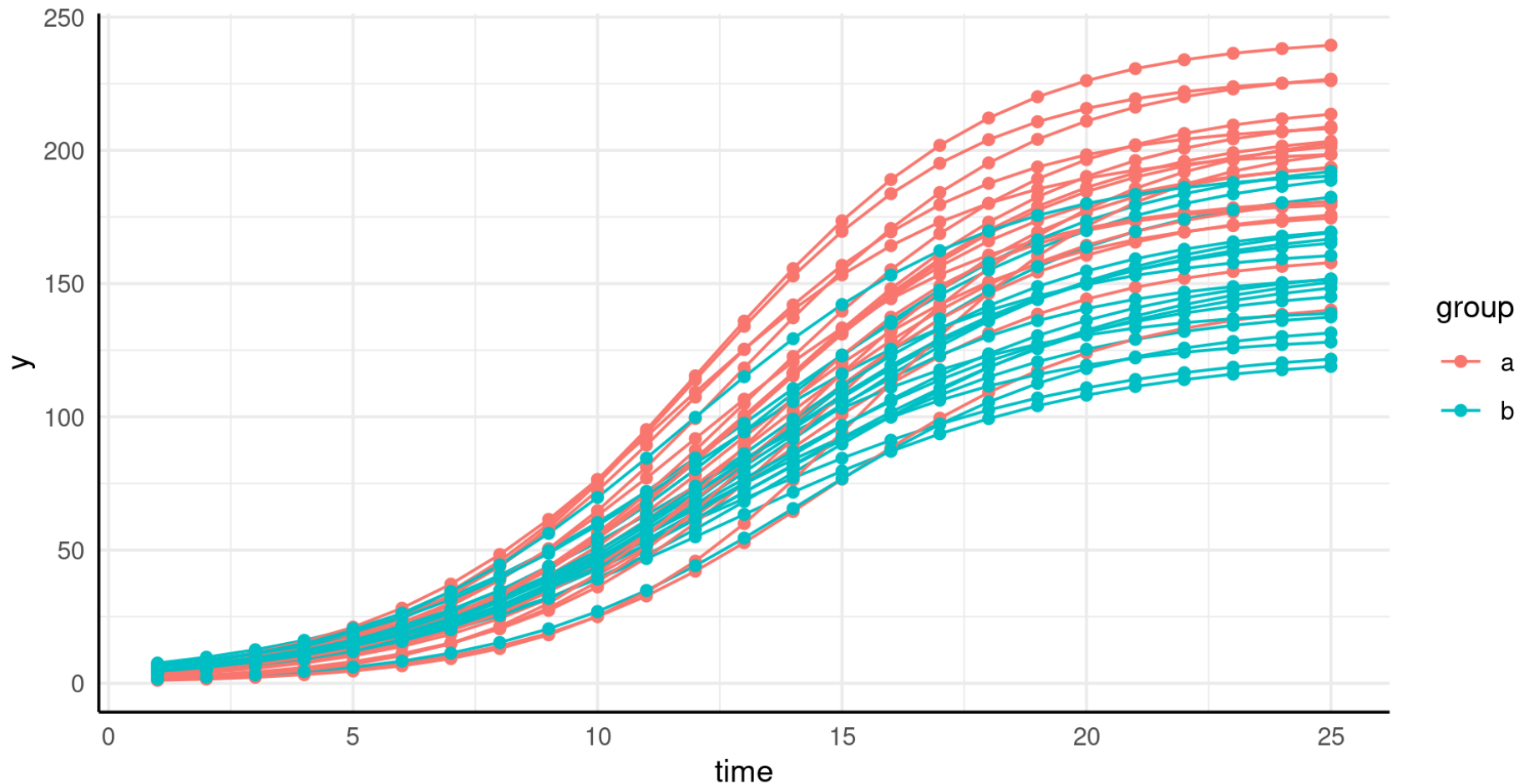
Non-Temporal

* Individuals are not measured over time

Challenges

- Autocorrelation
- Non-linearity
- Heteroskedasticity

Longitudinal Data



Temporal

* Individuals' data is collected over time (>2 timepoints)

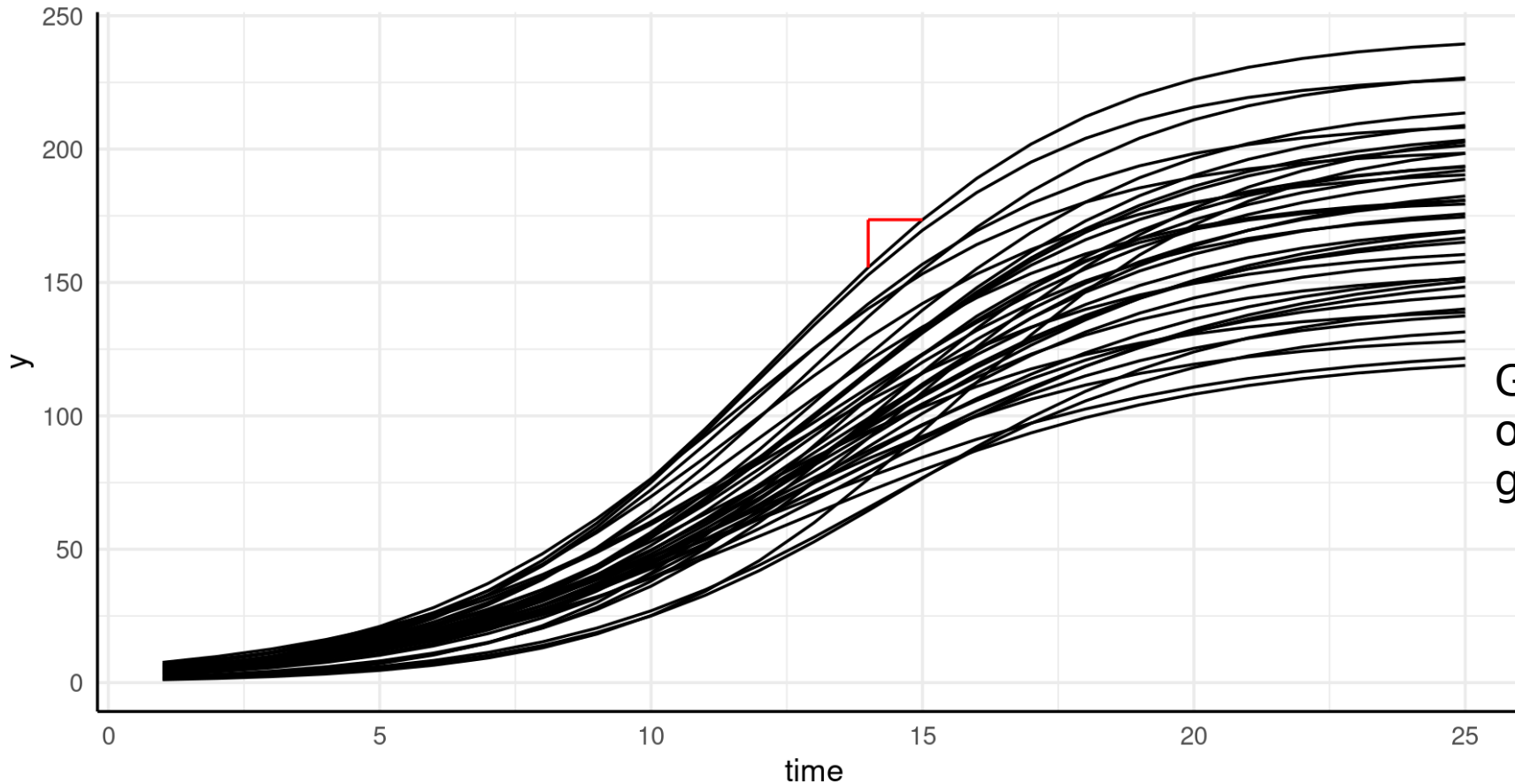
Non-Temporal

* Individuals are not measured over time

Challenges

- **Autocorrelation**
- Non-linearity
- Heteroskedasticity

Autocorrelation



Generally AR1 or ARMA1 are good options

Temporal

* Individuals' data is collected over time (>2 timepoints)

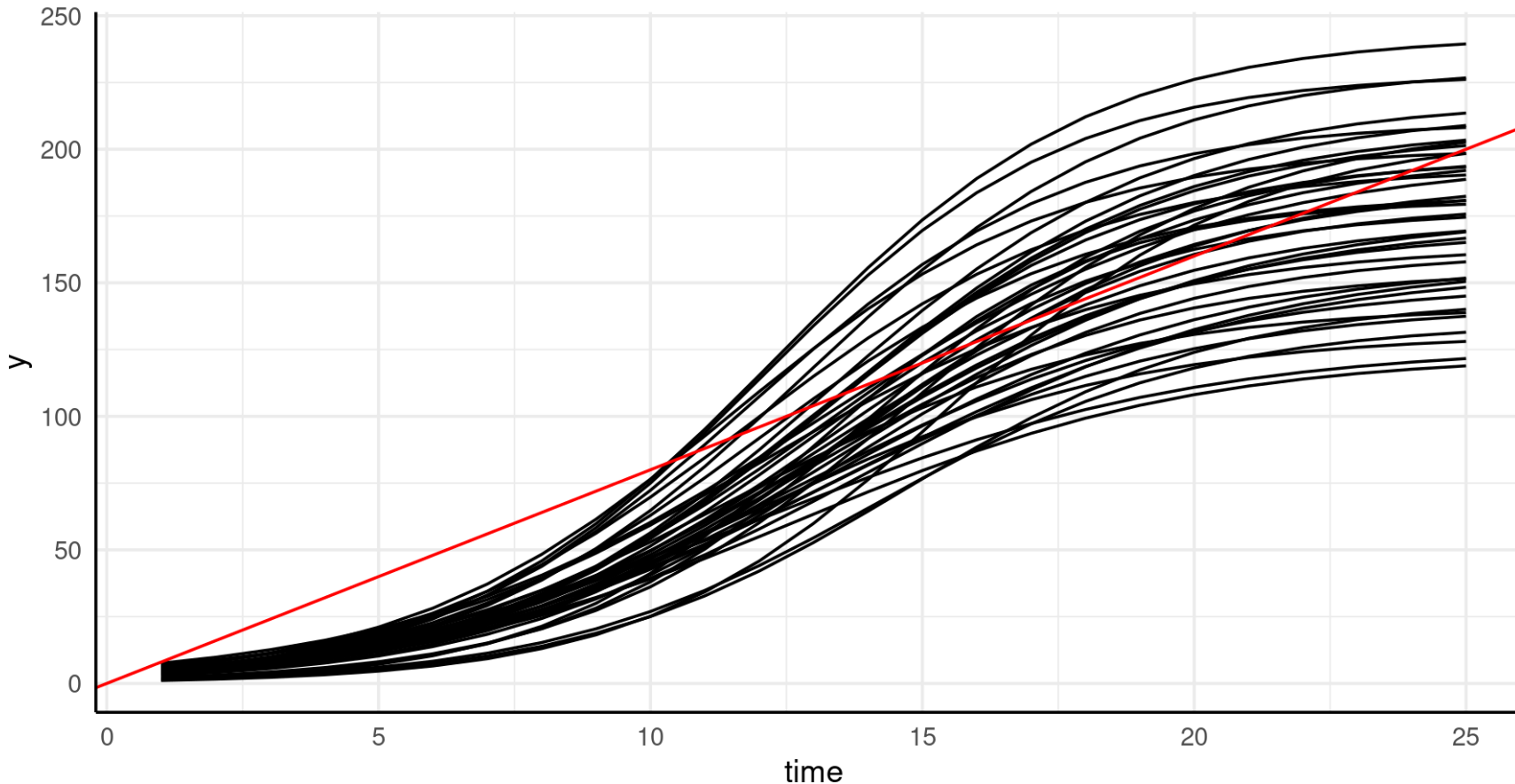
Non-Temporal

* Individuals are not measured over time

Challenges

- Autocorrelation
- **Non-linearity**
- Heteroskedasticity

Non-Linearity



Temporal

* Individuals' data is collected over time (>2 timepoints)

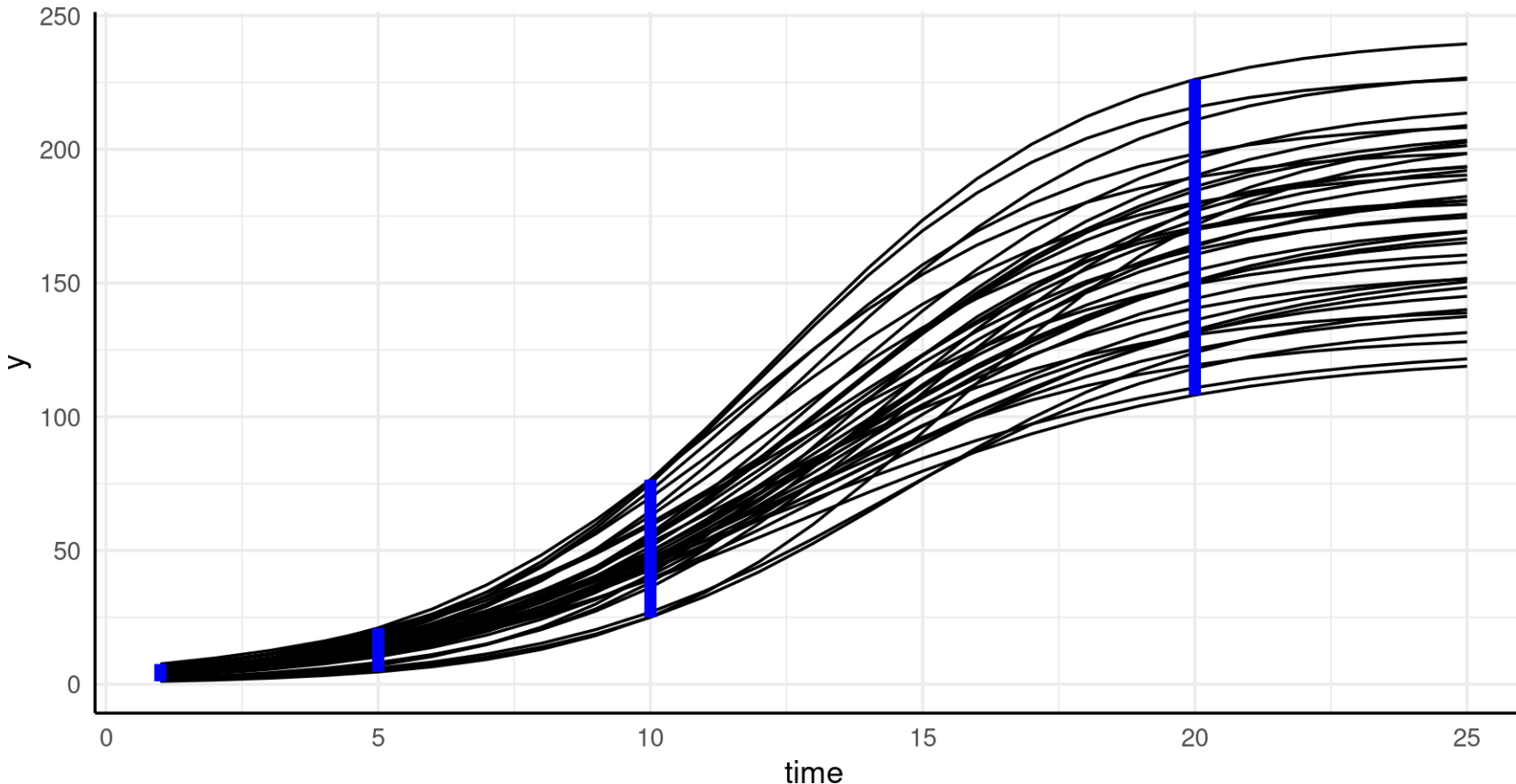
Non-Temporal

* Individuals are not measured over time

Challenges

- Autocorrelation
- Non-linearity
- **Heteroskedasticity**

Heteroskedasticity

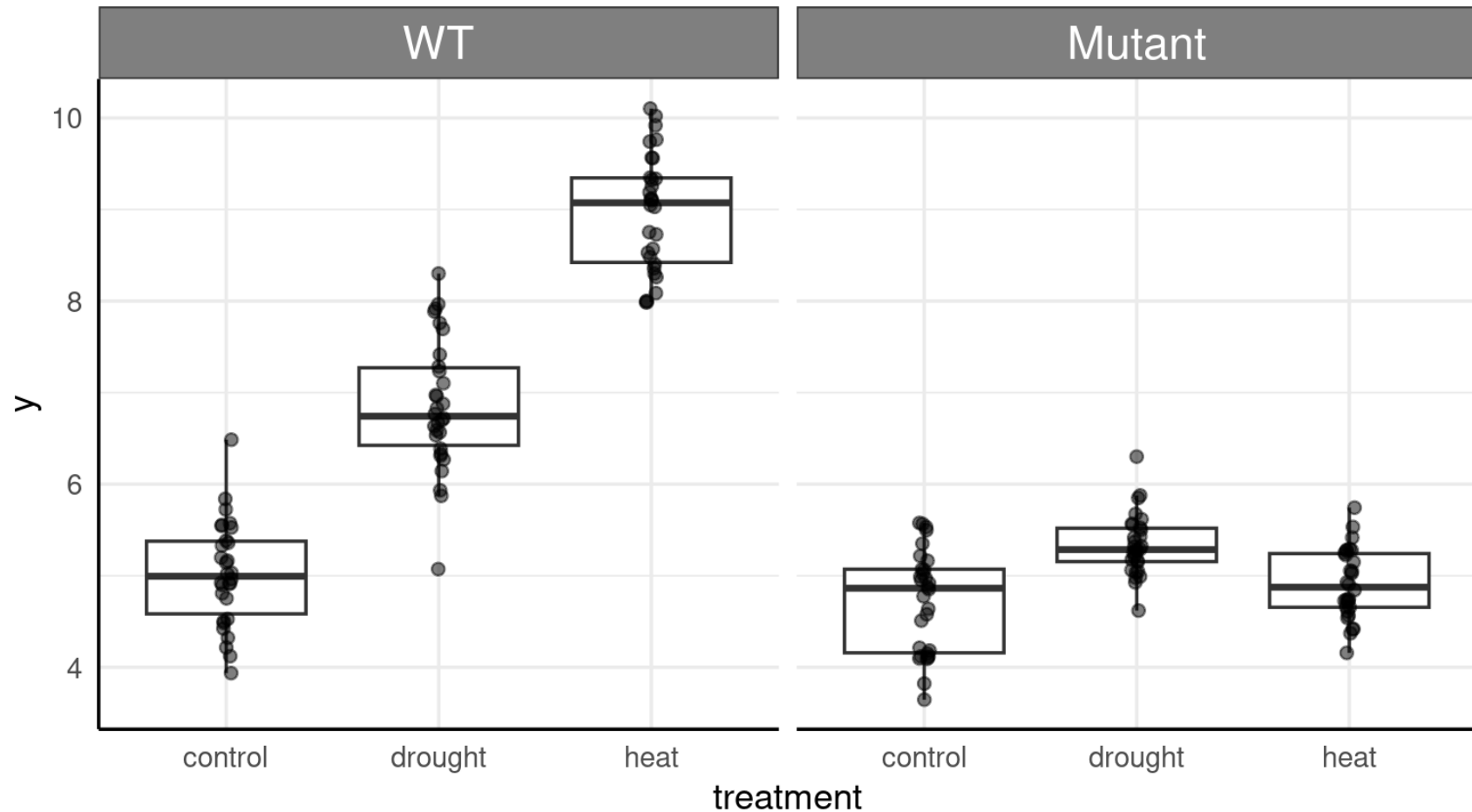


Temporal

* Individuals' data is collected over time (>2 timepoints)

Non-Temporal

* Individuals are not measured over time



No block effects

* Cofactor DOES NOT effect the components of the other design parameters equally

With block effects

* Cofactor DOES effect the components of the other design parameters equally

Temporal

* Individuals' data is collected over time (>2 timepoints)

Non-Temporal

* Individuals are not measured over time

Fixed Effects

* Effects from your experimental design that you intend to compare and with **fixed** levels (treatments, genotypes, etc)

Random Effects

* Effects from your experimental design that add noise across a **random** sample of levels (growth chambers, experiment #)

Link Function

Fixed Effects

* Effects from your experimental design that you intend to compare and with **fixed** levels (treatments, genotypes, etc)

Random Effects

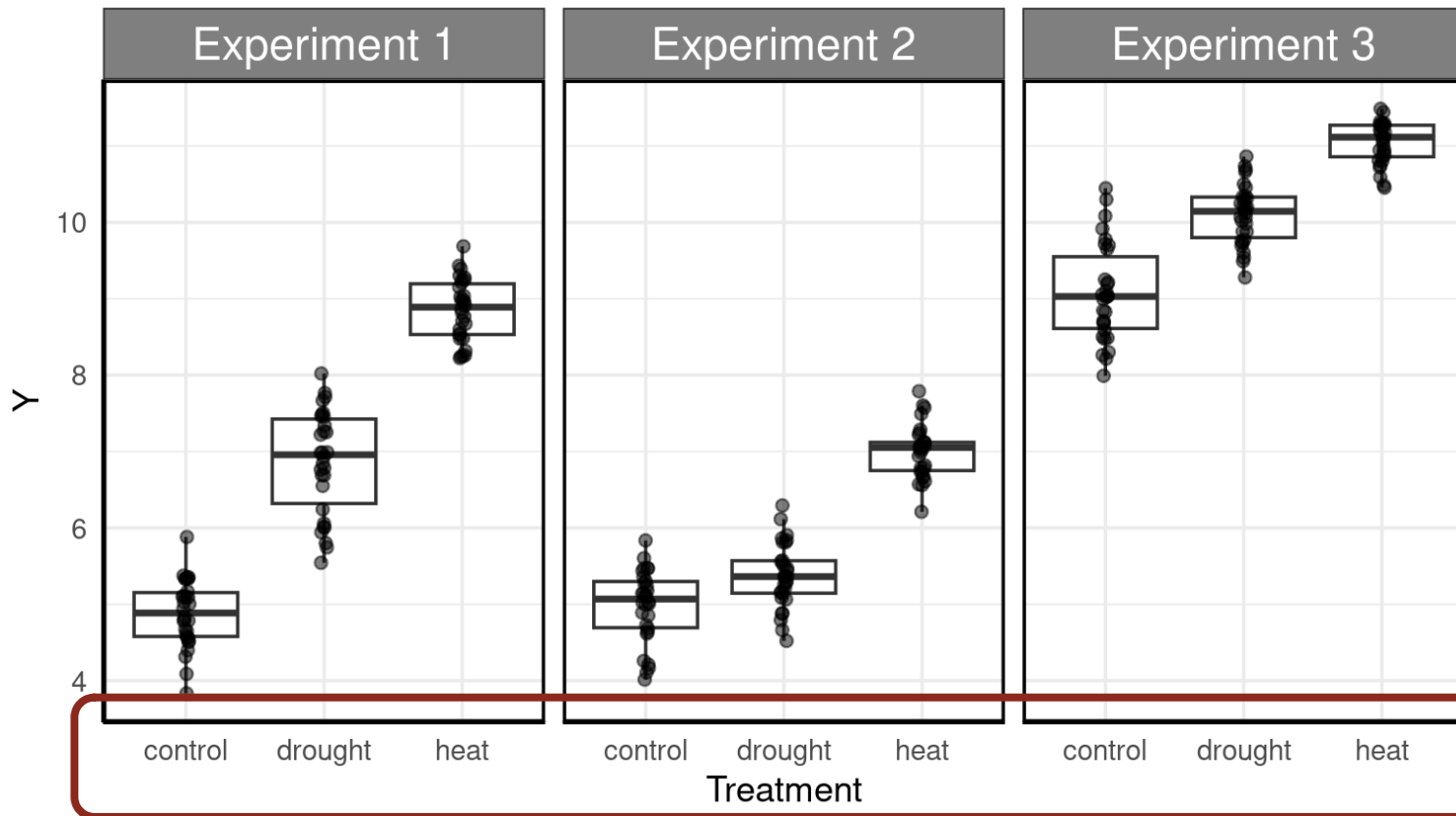
* Effects from your experimental design that add noise across a **random** sample of levels (growth chambers, experiment #)

Fixed Effects

* Effects from your experimental design that you intend to compare and with **fixed** levels (treatments, genotypes, etc)

Random Effects

* Effects from your experimental design that add noise across a **random** sample of levels (growth chambers, experiment #)



* We included 3 treatments and we want to be able to compare between them. These are **fixed** treatment groups. There are other treatments we could have included, but we are not claiming that we can generalize to those (here, cold stress)

Fixed Effects

* Effects from your experimental design that you intend to compare and with **fixed** levels (treatments, genotypes, etc)

Random Effects

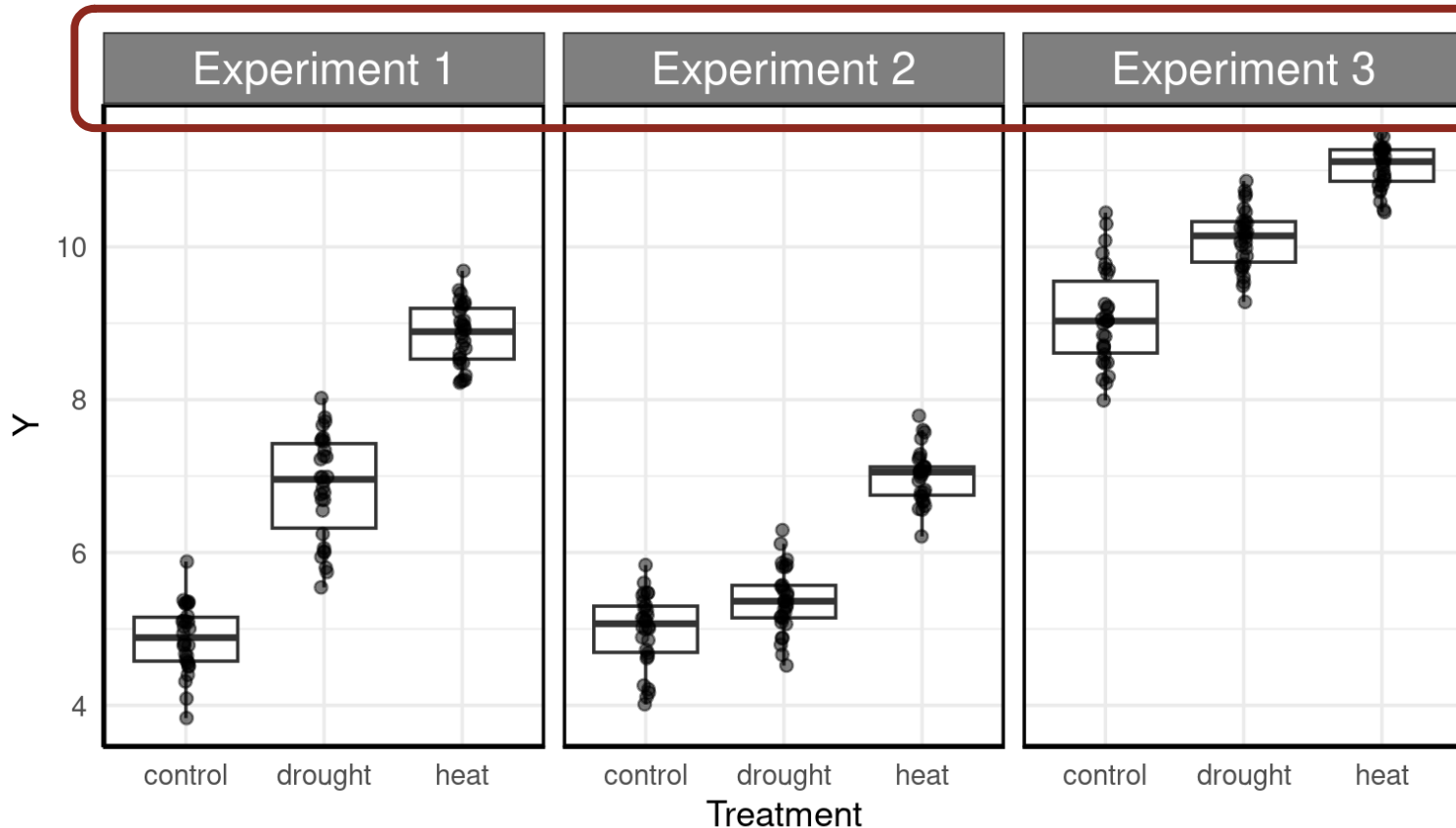
* Effects from your experimental design that add noise across a **random** sample of levels (growth chambers, experiment #)

Fixed Effects

* Effects from your experimental design that you intend to compare and with **fixed** levels (treatments, genotypes, etc)

Random Effects

* Effects from your experimental design that add noise across a **random** sample of levels (growth chambers, experiment #)



* We ran this experiment 3 times. We don't really want to compare those runs against each other and we might run the experiment 4 or more times still, so experiment number is a **random sample** of the possible times we could have done this.

No block effects

- * Cofactor DOES NOT effect the components of the other design parameters equally

With block effects

- * Cofactor DOES effect the components of the other design parameters equally

Temporal

- * Individuals' data is collected over time (>2 timepoints)

Non-Temporal

- * Individuals are not measured over time

Fixed Effects

- * Effects from your experimental design that you intend to compare and with **fixed** levels (treatments, genotypes, etc)

Random Effects

- * Effects from your experimental design that add noise across a **random** sample of levels (growth chambers, experiment #)

Link Function

- * A function applied to your response variable for computational simplicity/interpretability.

Link Function

* A function applied to your response variable for computational simplicity/interpretability.

Link Function

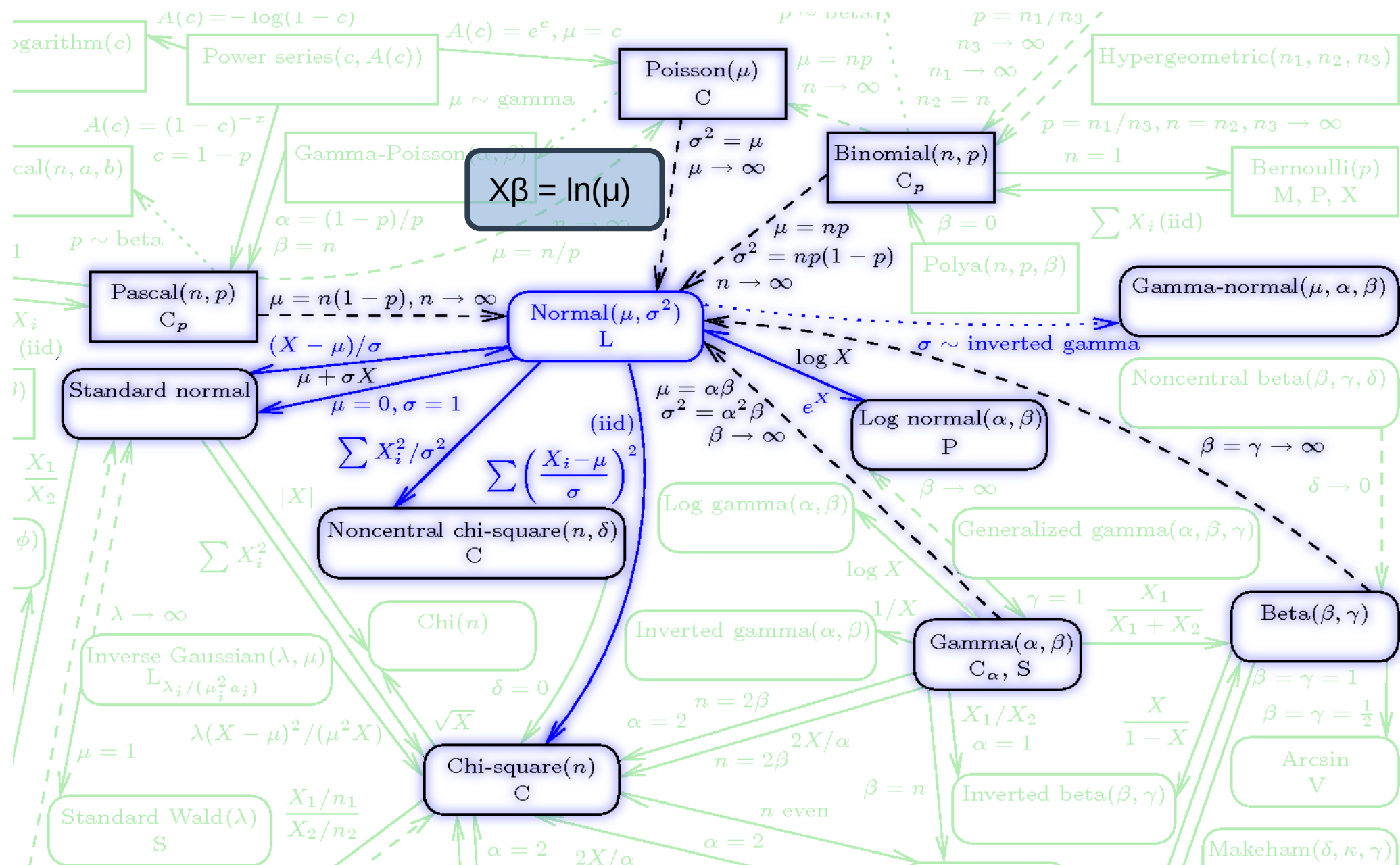
* A function applied to your response variable for computational simplicity/interpretability.

How does this work?

* These functions relate a linear predictor to the mean of the response variable. These generally come from mathematical statistics, where we can find the canonical link between distributions.

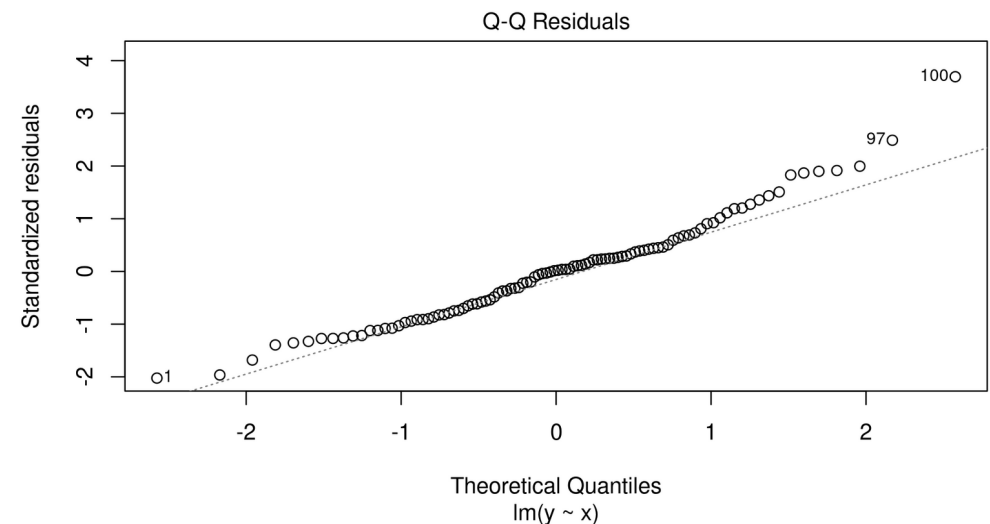
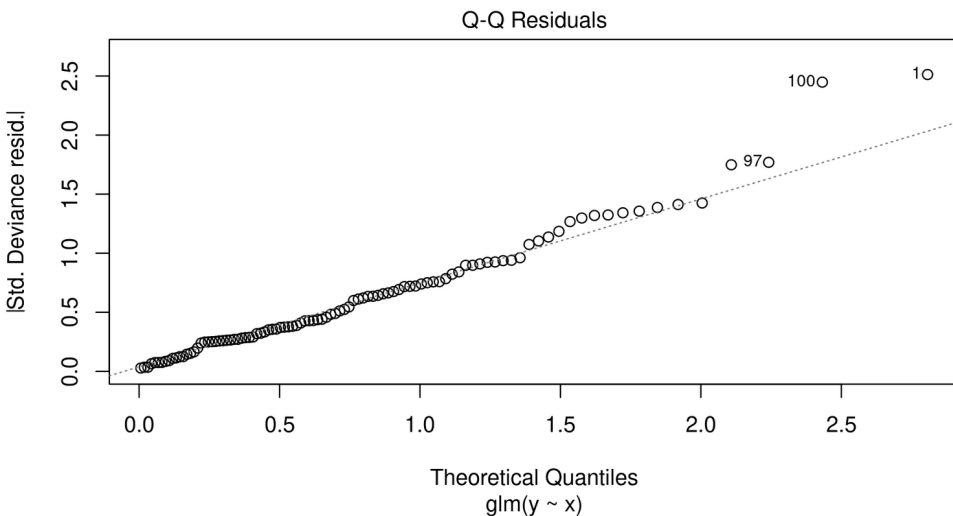
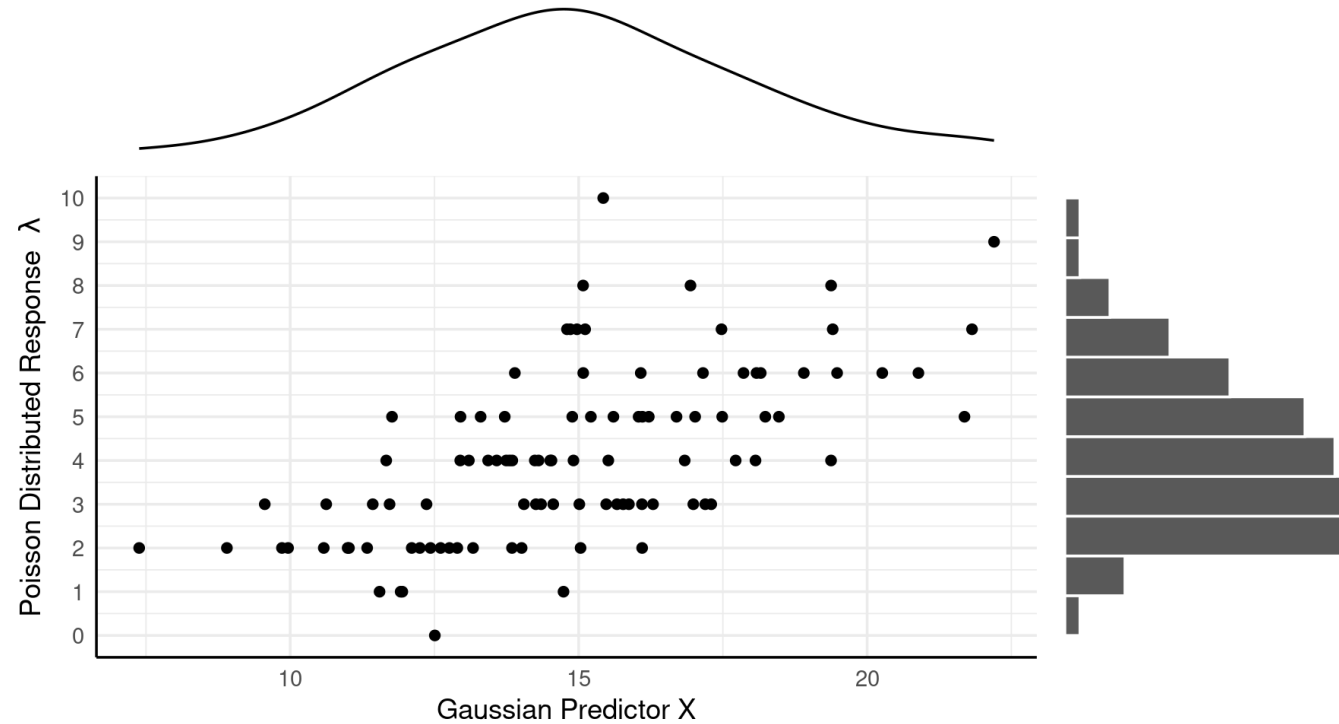


Oh cool, mathematical statistics again!

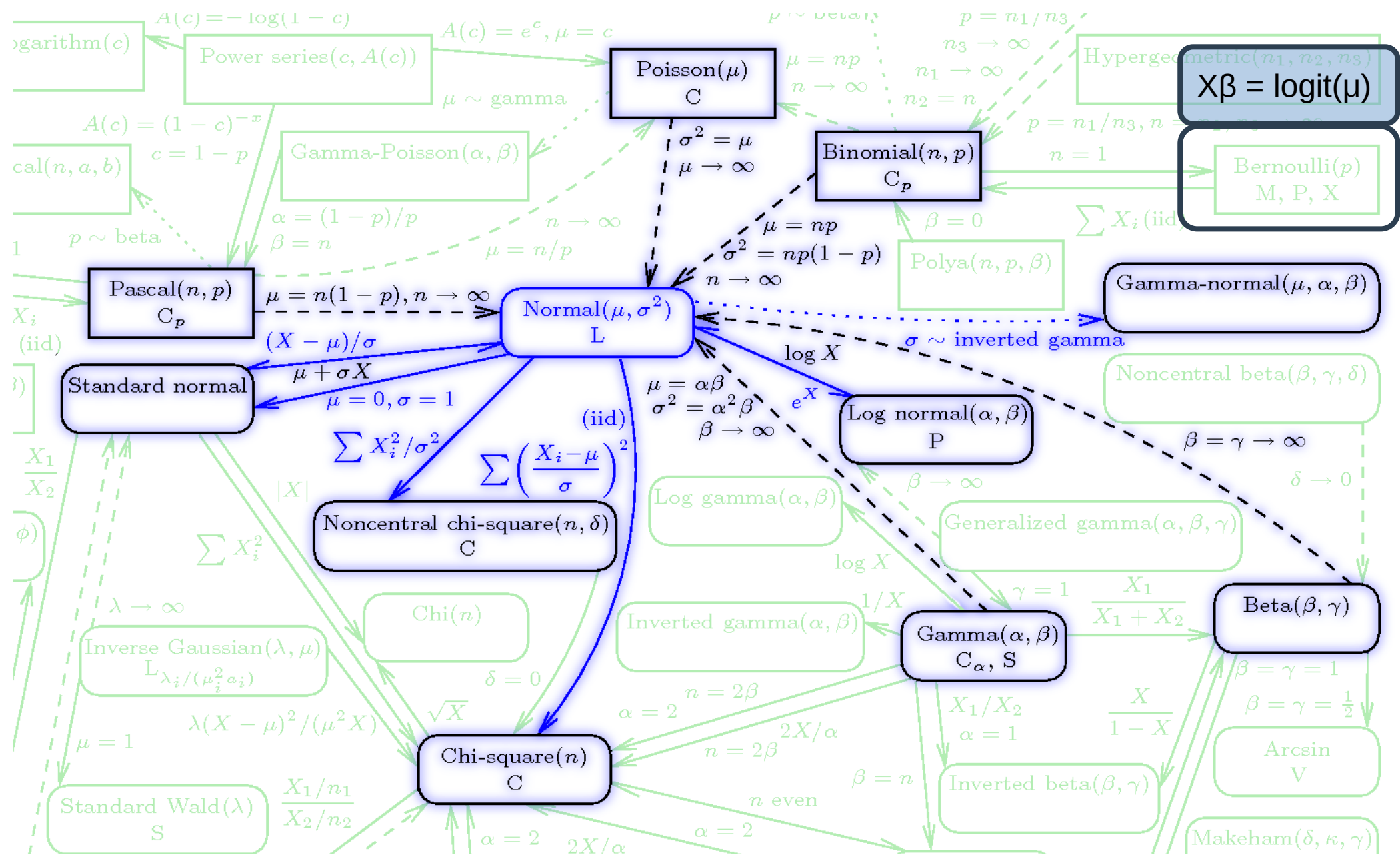


Link Function

* A function applied to your response variable for computational simplicity/interpretability.

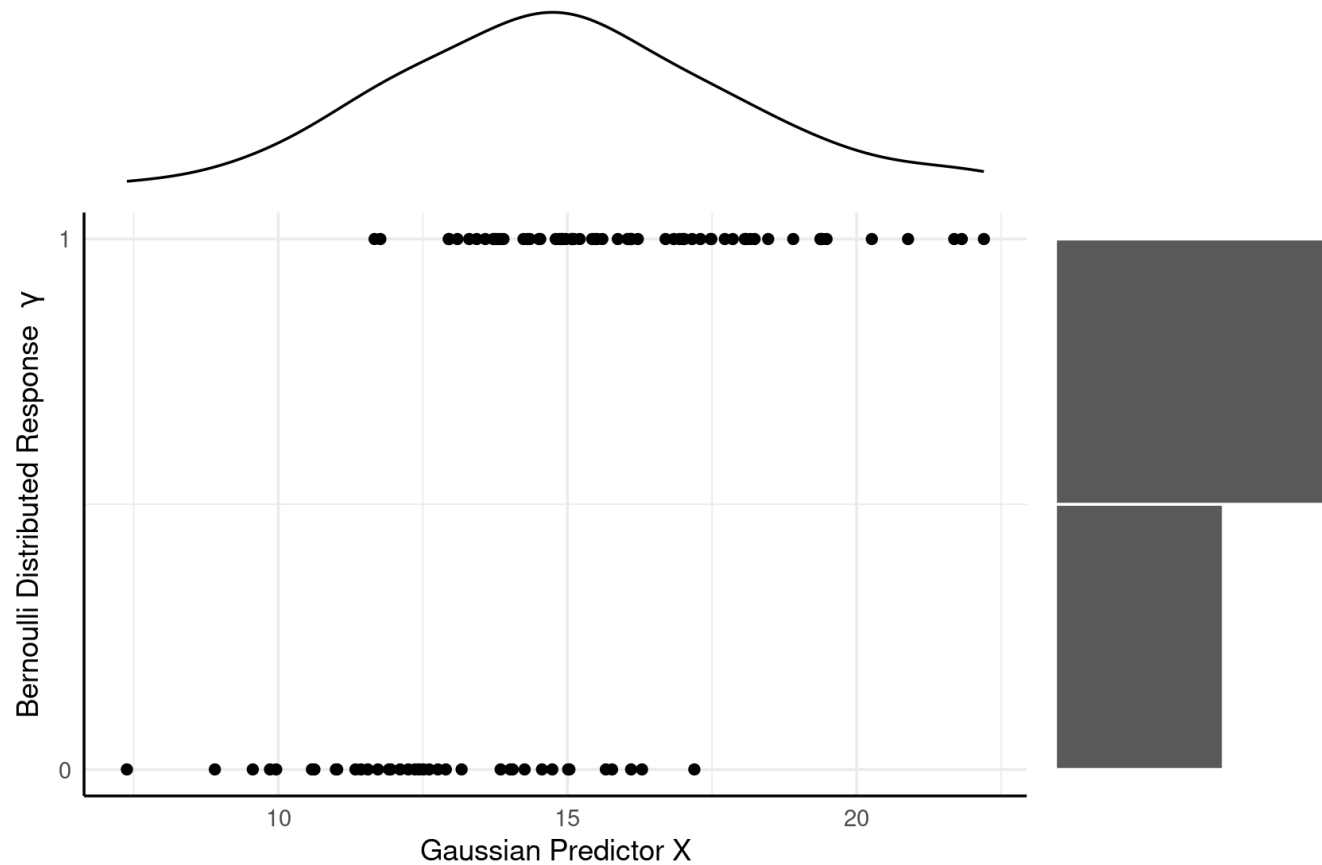


One More Time!



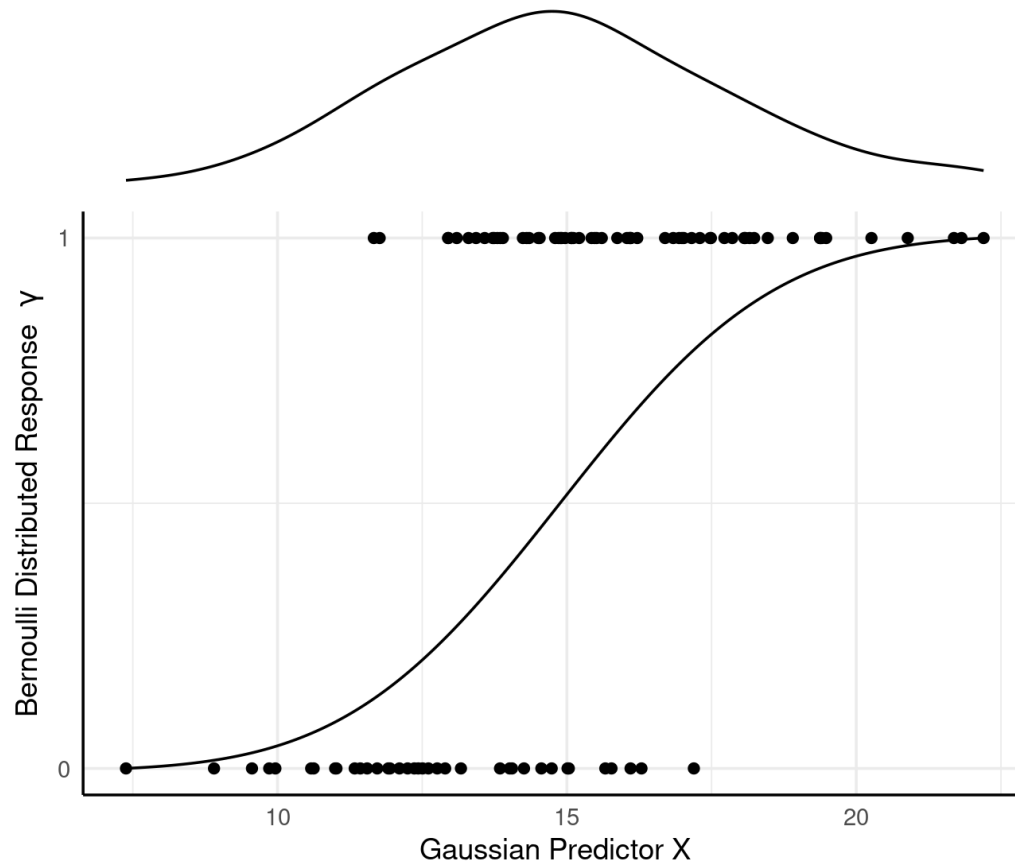
Link Function

* A function applied to your response variable for computational simplicity/interpretability.



Link Function

* A function applied to your response variable for computational simplicity/interpretability.



Link Function

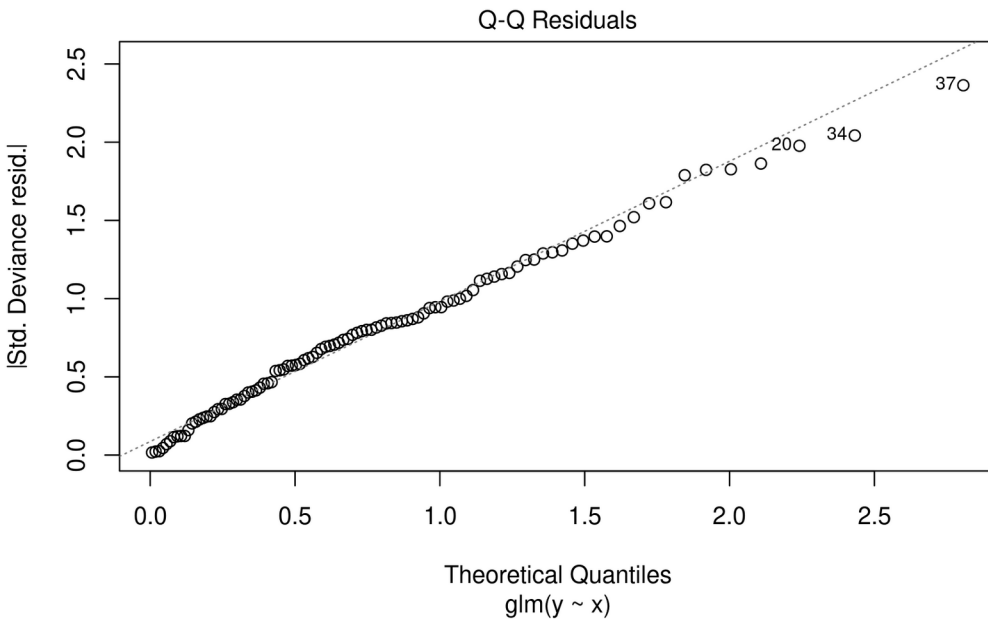
* A function applied to your response variable for computational simplicity/interpretability.

```
> m <- glm( y ~ x, family=binomial(link="probit"), data = df)
> m
```

Call: `glm(formula = y ~ x, family = binomial(link = "probit"), data = df)`

Coefficients:

(Intercept)	x
-5.8072	0.4257



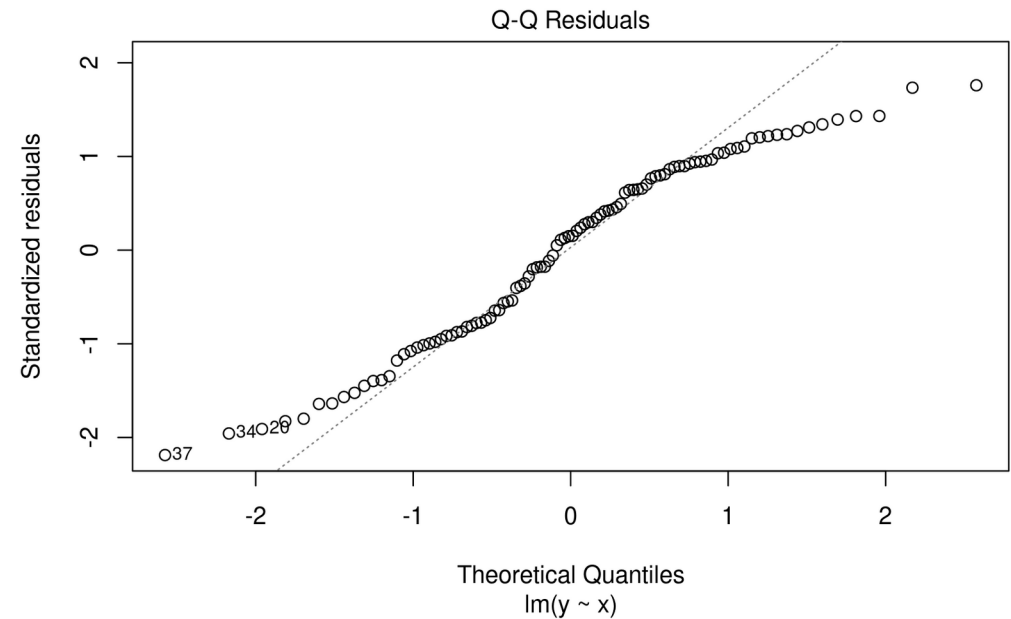
```
> m2 <- lm( y ~ x, data = df)
> m2
```

Call:

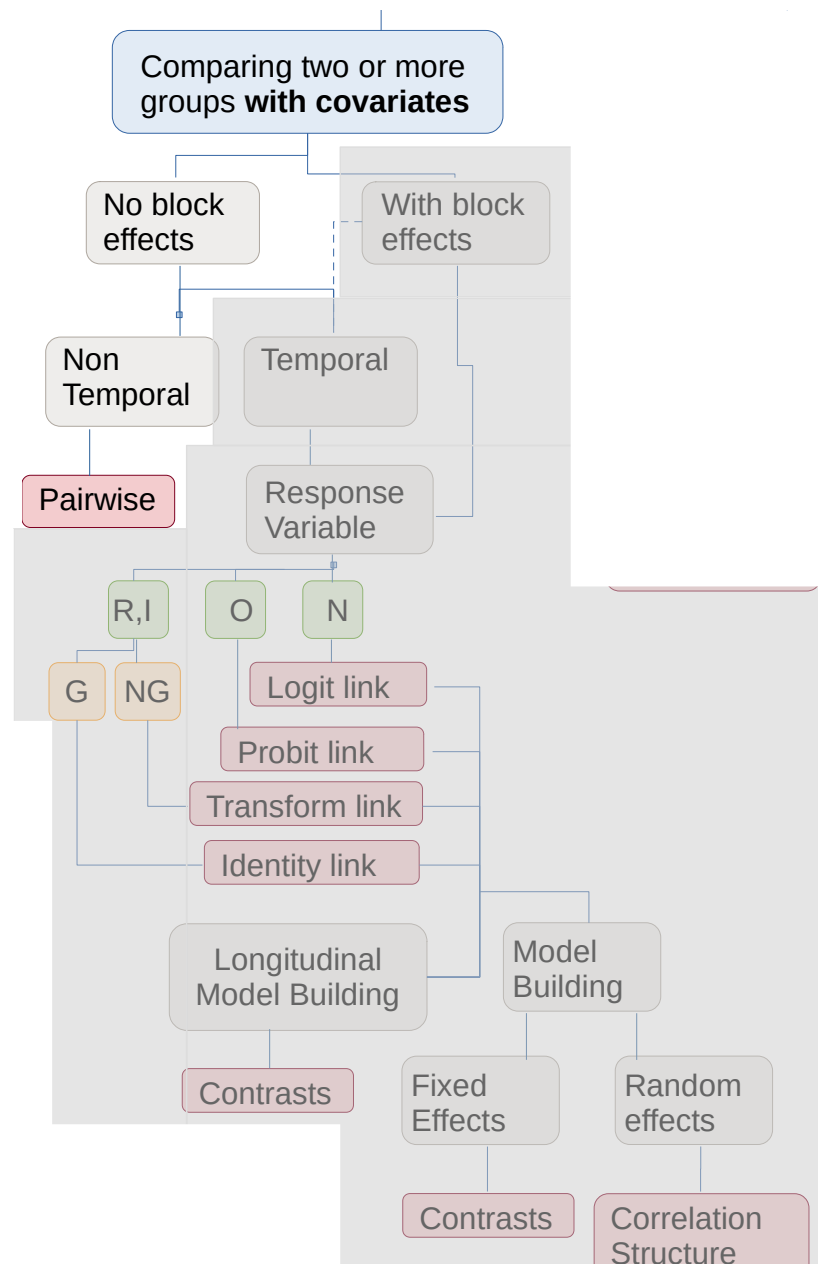
`lm(formula = y ~ x, data = df)`

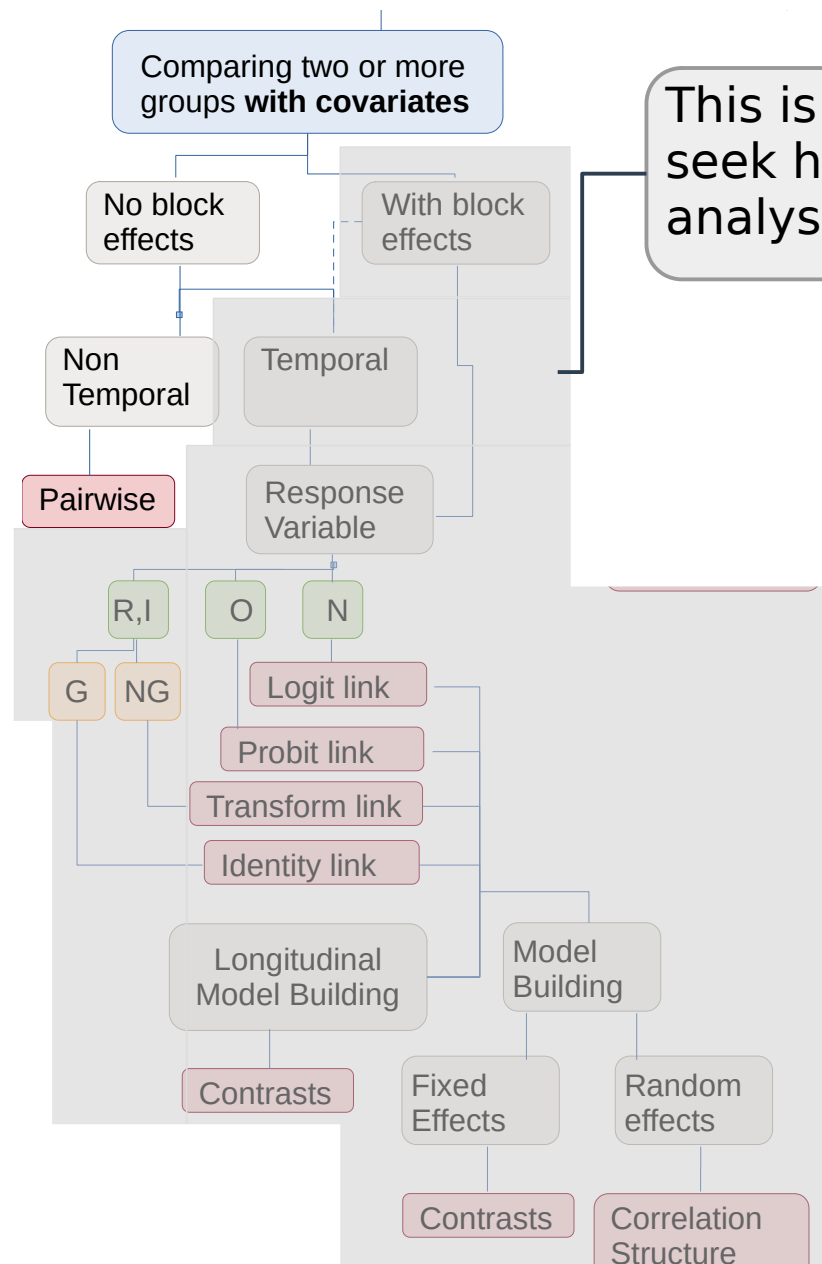
Coefficients:

(Intercept)	x
-0.82709	0.09784

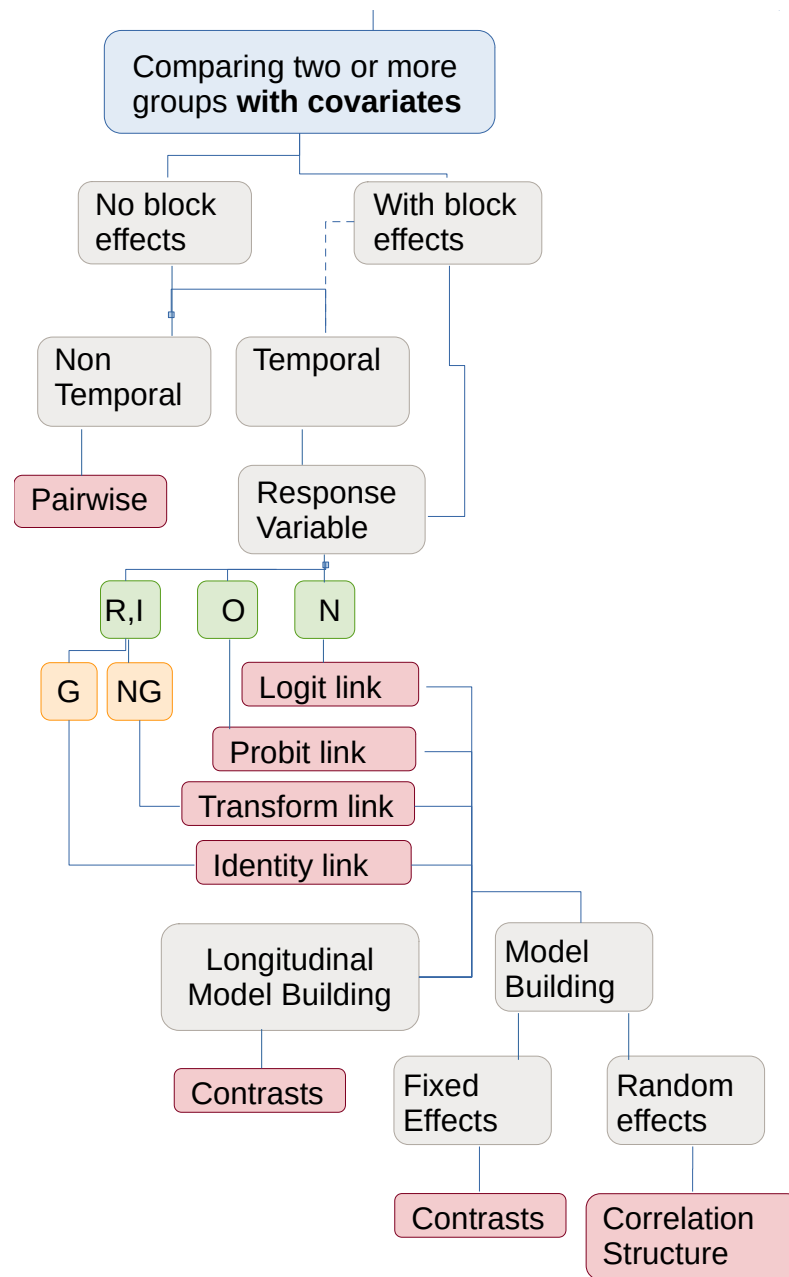


Two group comparisons between all groups of interest





This is a good place to seek help with your analysis.



Scenario 6

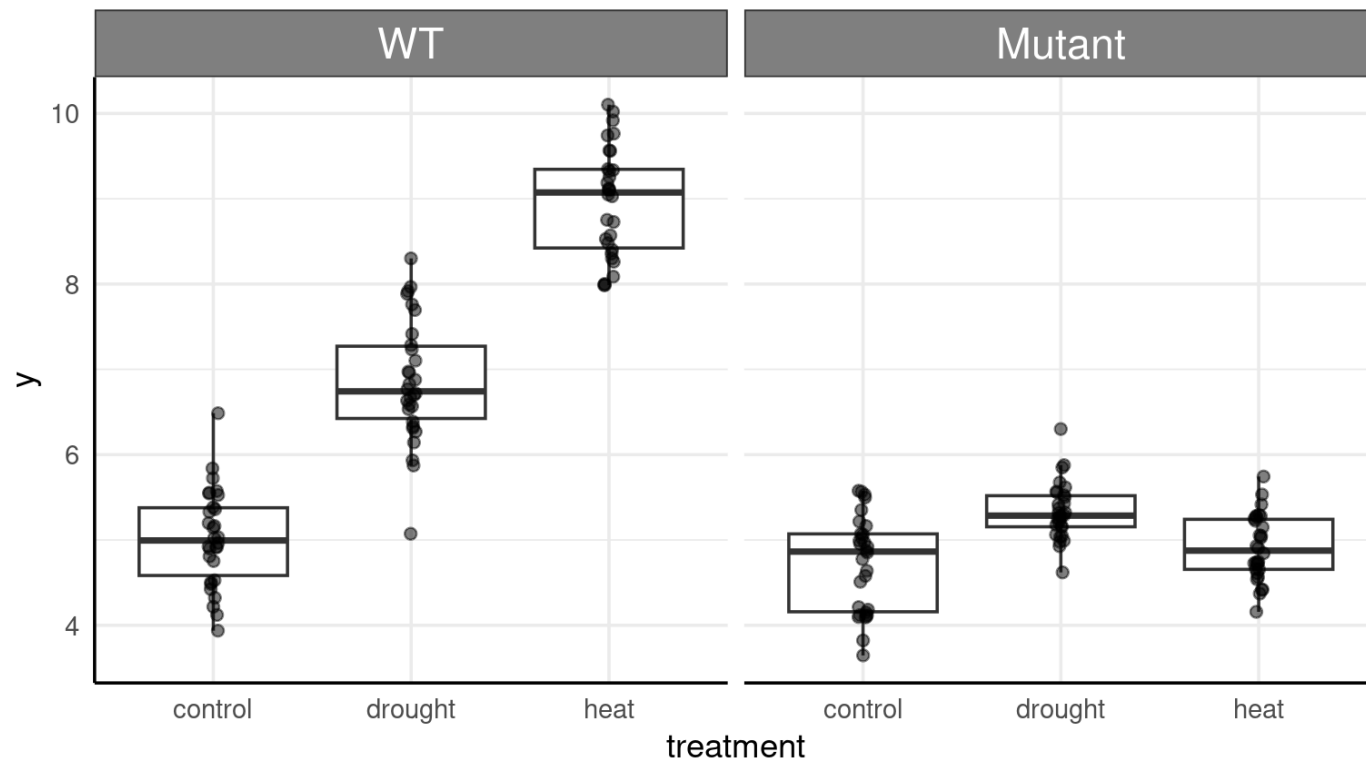
Comparing two or more groups **with covariates**

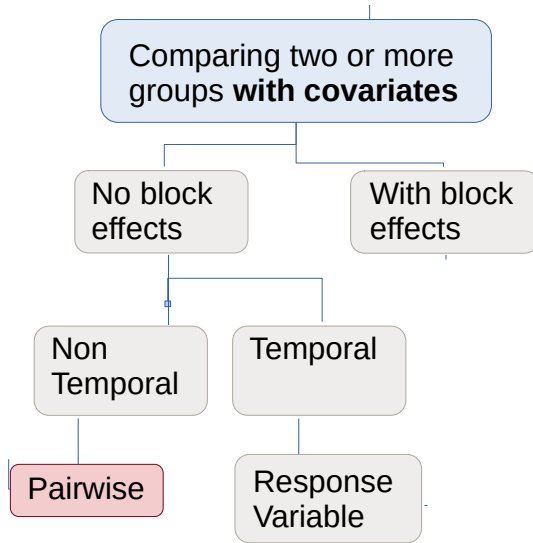
No block effects

With block effects

You want to increase gene expression of your favorite gene so you create a knock-in mutant and compare to WT. You grow the plants in control conditions, drought, and heat. When you gather the expression values and plot them this is what you see.

- What steps lead you do the correct test to say you have induced gene expression?

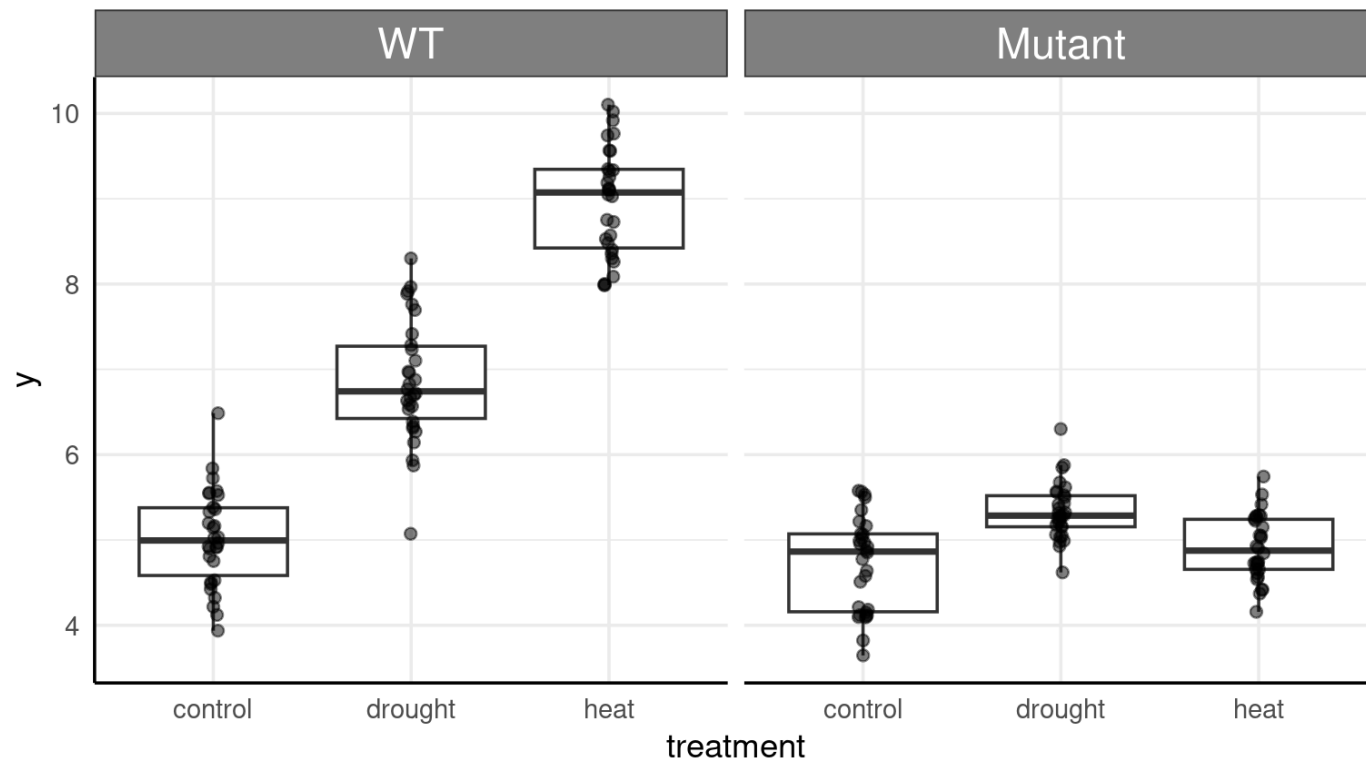




Scenario 6

You want to increase gene expression of your favorite gene so you create a knock-in mutant and compare to WT. You grow the plants in control conditions, drought, and heat. When you gather the expression values and plot them this is what you see.

- What steps lead you do the correct test to say you have induced gene expression?



Scenario 6

You want to increase gene expression of your favorite gene so you create a knock-in mutant and compare to WT. You grow the plants in control conditions, drought, and heat. When you gather the expression values and plot them this is what you see.

- What steps lead you do the correct test to say you have induced gene expression?

Answer

These data do not show a block effect and are not longitudinal, since the controls are at the same scale this is an interaction effect, so we use pairwise comparisons.

Scenario 6

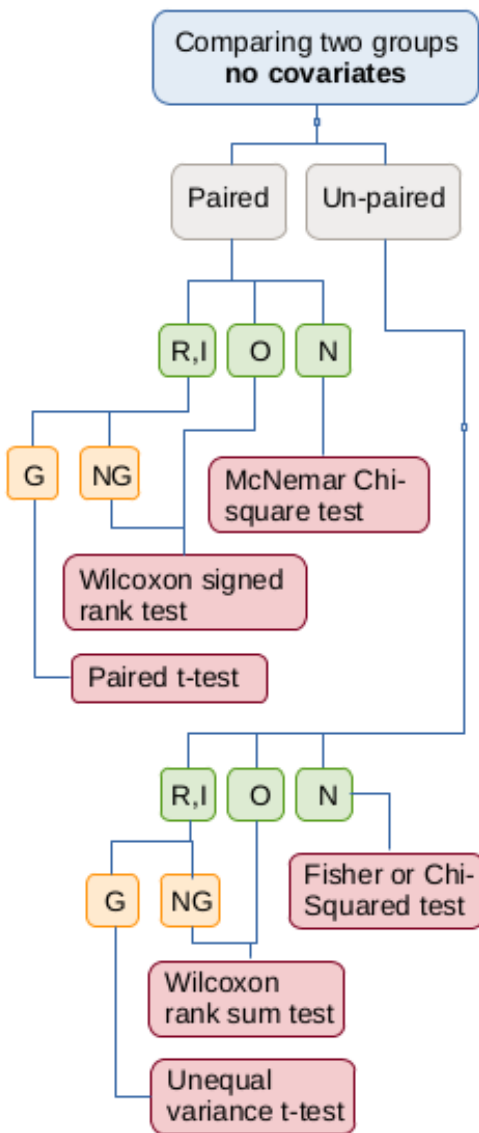
You want to increase gene expression of your favorite gene so you create a knock-in mutant and compare to WT. You grow the plants in control conditions, drought, and heat. When you gather the expression values and plot them this is what you see.

- What steps lead you do the correct test to say you have induced gene expression?

Answer

These data do not show a block effect, since the controls are at the same scale this is an interaction effect, so we use pairwise comparisons.

Now we restart in the “no covariates” option since we will pick which covariate combinations to test. Here we have unpaired data measured as R,I and that appears Gaussian, so we will use several unequal variance T tests.



Scenario 6

You want to increase gene expression of your favorite gene so you create a knock-in mutant and compare to WT. You grow the plants in control conditions, drought, and heat. When you gather the expression values and plot them this is what you see.

- What steps lead you do the correct test to say you have induced gene expression?

R

```
> # selected comparisons
> s1 <- int_df[int_df$treatment=="heat" & int_df$genotype=="WT", "y"]
> s2 <- int_df[int_df$treatment=="heat" & int_df$genotype=="Mutant", "y"]
> t.test(s1,s2)
```

Welch Two Sample t-test

data: s1 and s2

t = 29.899, df = 47.305, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

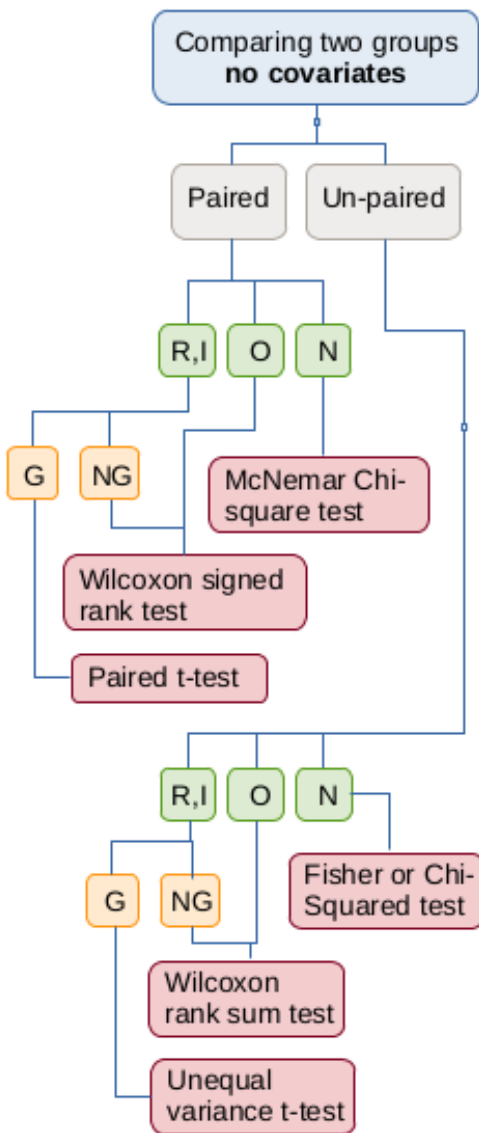
95 percent confidence interval:

3.783558 4.329343

sample estimates:

mean of x mean of y

8.966105 4.909654



Scenario 6

You want to increase gene expression of your favorite gene so you create a knock-in mutant and compare to WT. You grow the plants in control conditions, drought, and heat. When you gather the expression values and plot them this is what you see.

- What steps lead you do the correct test to say you have induced gene expression?

R

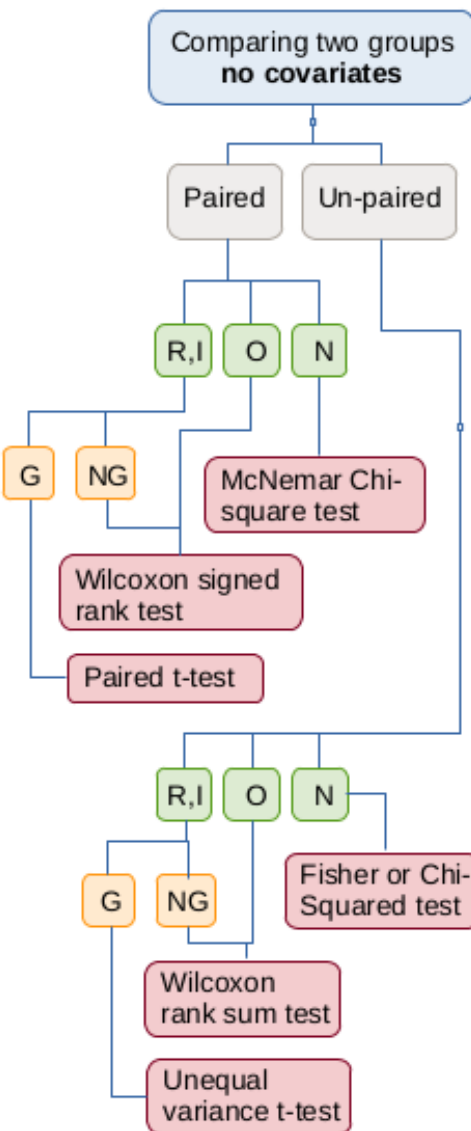
```
> pairwise.t.test(int_df$y, interaction(int_df$treatment, int_df$genotype))
```

Pairwise comparisons using t tests with pooled SD

data: int_df\$y and interaction(int_df\$treatment, int_df\$genotype)

	control.WT	drought.WT	heat.WT	control.Mutant	drought.Mutant
drought.WT	< 2e-16	-	-	-	-
heat.WT	< 2e-16	< 2e-16	-	-	-
control.Mutant	0.11061	< 2e-16	< 2e-16	-	-
drought.Mutant	0.11013	< 2e-16	< 2e-16	0.00015	-
heat.Mutant	0.43699	< 2e-16	< 2e-16	0.37429	0.01540

P value adjustment method: holm



Scenario 6

You want to increase gene expression of your favorite gene so you create a knock-in mutant and compare to WT. You grow the plants in control conditions, drought, and heat. When you gather the expression values and plot them this is what you see.

- What steps lead you do the correct test to say you have induced gene expression?

R

```
> pairwise.t.test(int_df$y, interaction(int_df$treatment, int_df$genotype))
```

Pairwise comparisons using t tests with pooled SD

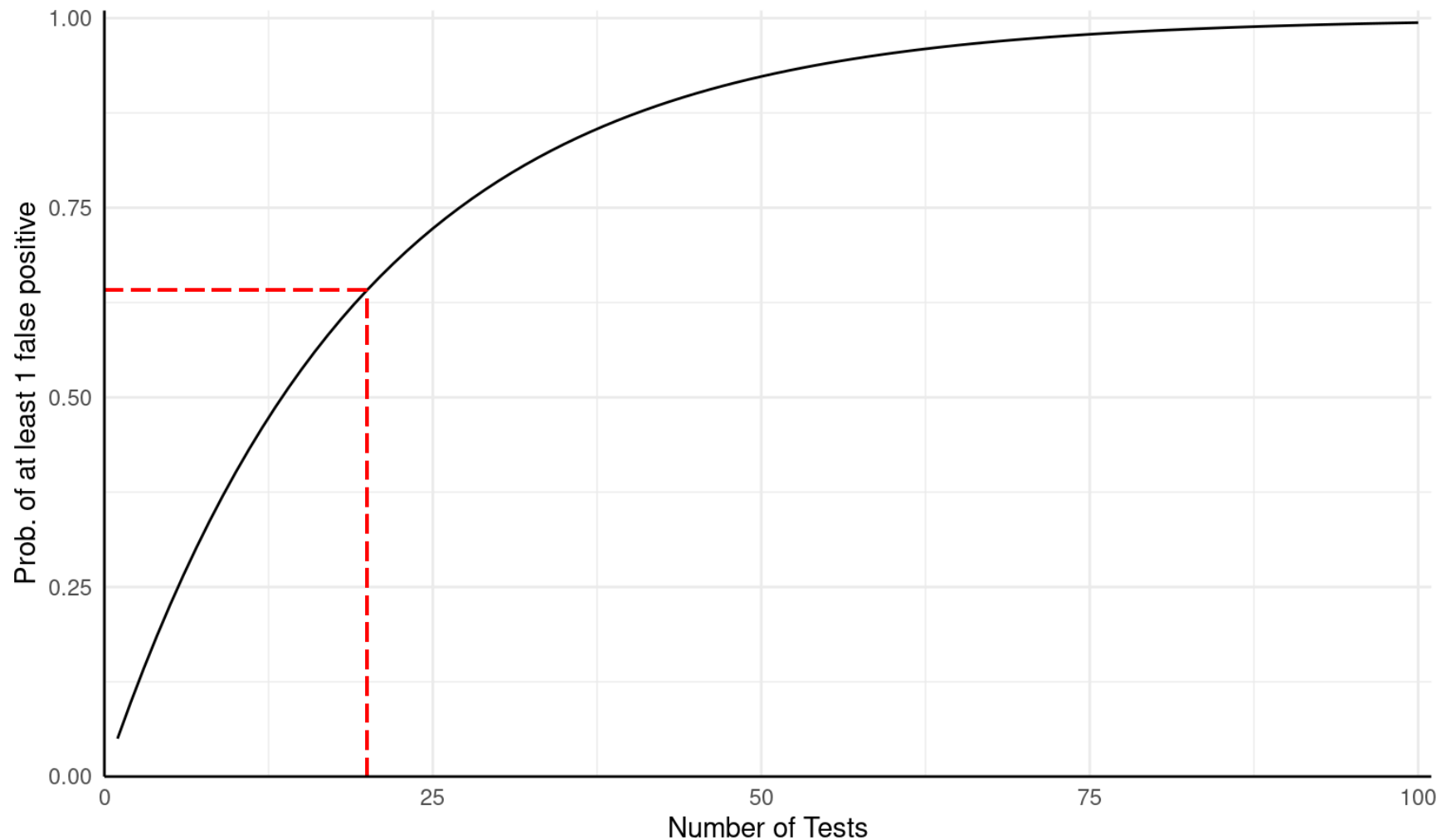
data: int_df\$y and interaction(int_df\$treatment, int_df\$genotype)

	control.WT	drought.WT	heat.WT	control.Mutant	drought.Mutant
drought.WT	< 2e-16	-	-	-	-
heat.WT	< 2e-16	< 2e-16	-	-	-
control.Mutant	0.11061	< 2e-16	< 2e-16	-	-
drought.Mutant	0.11013	< 2e-16	< 2e-16	0.00015	-
heat.Mutant	0.43699	< 2e-16	< 2e-16	0.37429	0.01540

P value adjustment method:holm

Multiple Test Corrections

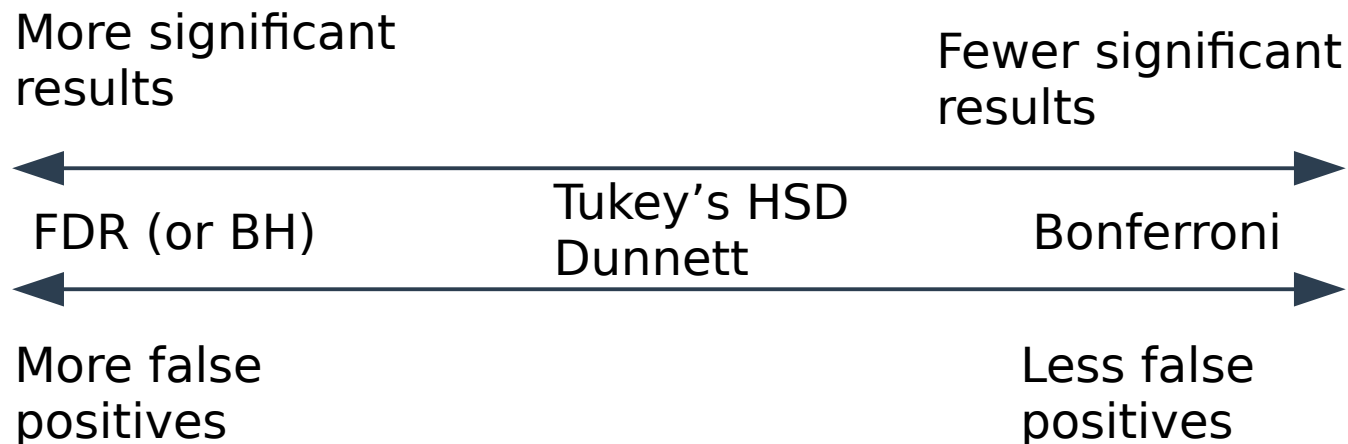
We generally apply multiple testing corrections if we do 20 or more tests under the heuristic that $\alpha = 0.05$ and $1/\alpha = 20$.



Multiple Test Corrections

We generally apply multiple testing corrections if we do 20 or more tests under the heuristic that $\alpha = 0.05$ and $1/\alpha = 20$.

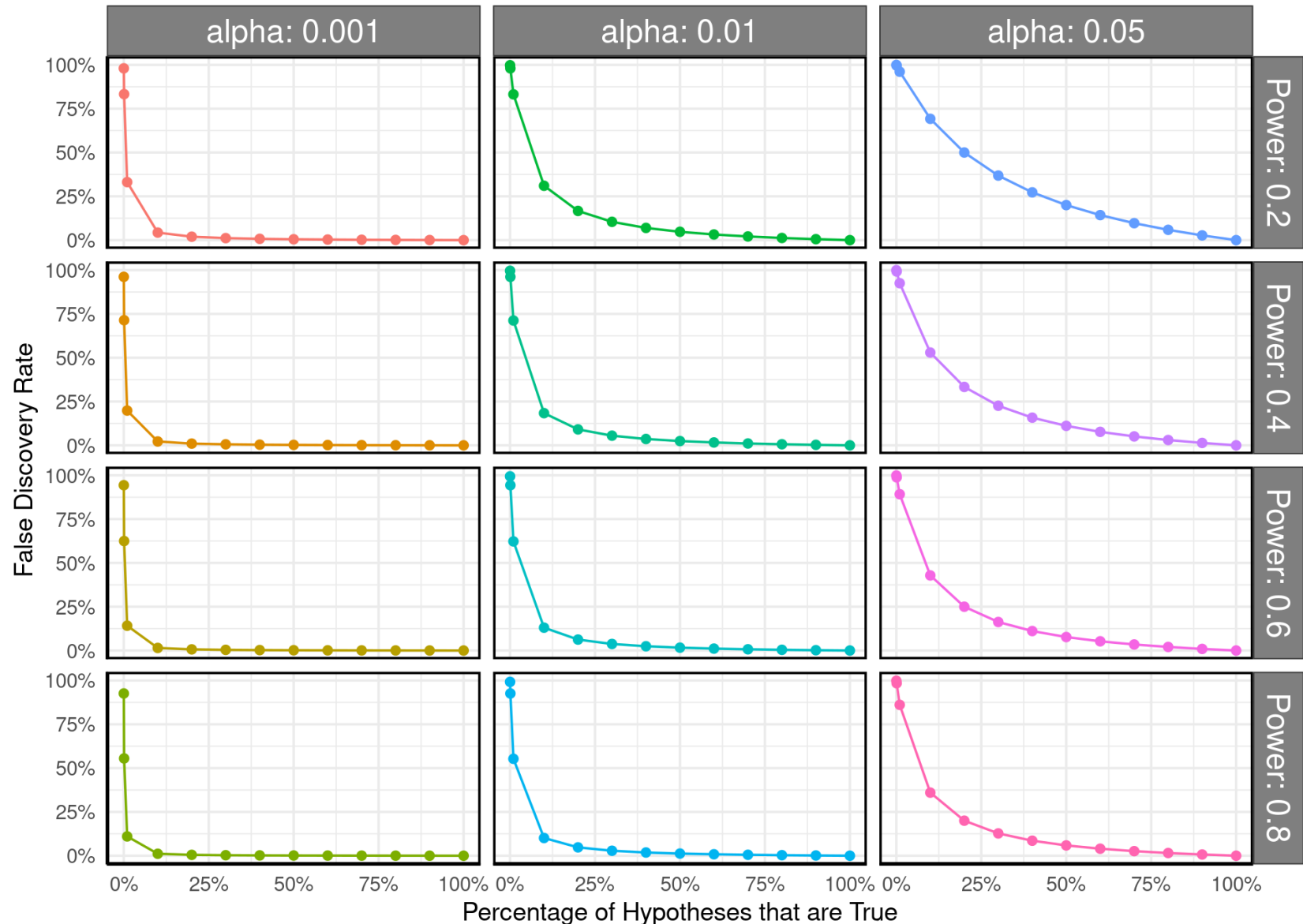
Correction methods fall on a spectrum



Multiple Test Corrections

This all assumes a constant and relatively high percentage of hypotheses that are True.

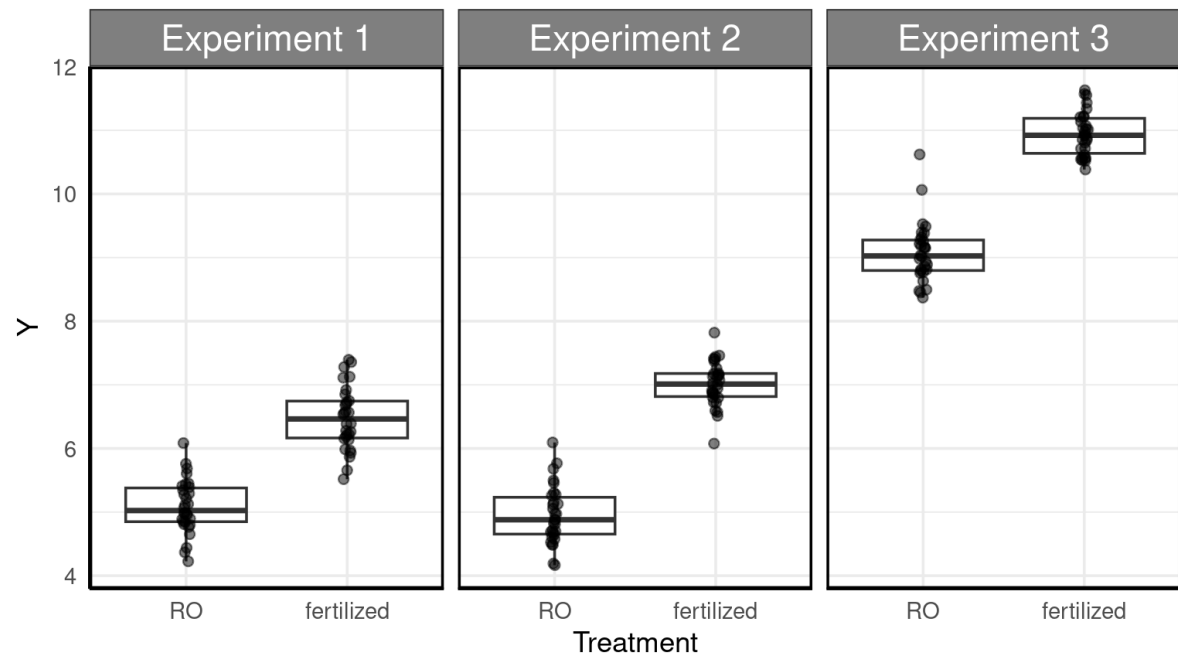
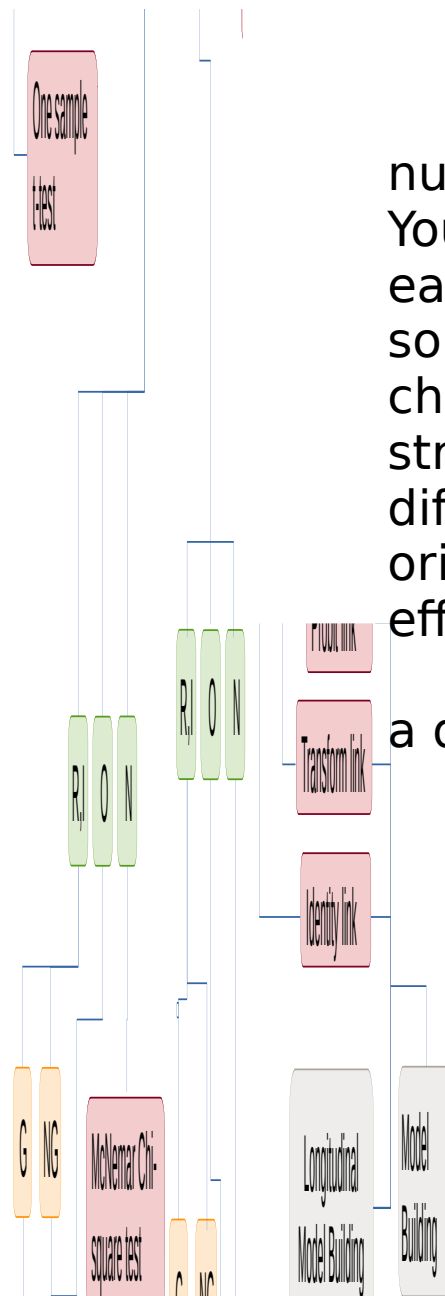
Bayesian methods can address that assumption.



Scenario 7

You're trying to determine if it's necessary to add nutrients to a metromix 360 and turfcase soil blend. You conduct your first experiment with 10 reps in each condition. You see a very minor treatment effect so you do it again but couldn't get the same growth chamber. In the second experiment you see a stronger effect but all the plants grew larger in the different chamber. In a third experiment you get the original chamber again and see a better treatment effect.

- What steps lead you do the correct test to make a decision on fertilizer?



Scenario 7

You're trying to determine if it's necessary to add nutrients to a metromix 360 and turface soil blend. You conduct your first experiment with 10 reps in each condition. You see a very minor treatment effect so you do it again but couldn't get the same growth chamber. In the second experiment you see a stronger effect but all the plants grew larger in the different chamber. In a third experiment you get the original chamber again and see a better treatment effect.

- What steps lead you do the correct test to make a decision on fertilizer?

Answer

These data do show a block effect so we need to account for that.

Here our data is **R,I** and appears **Gaussian**, so we will use an **identity link** in our model



Scenario 7

You're trying to determine if it's necessary to add nutrients to a metromix 360 and turfcase soil blend. You conduct your first experiment with 10 reps in each condition. You see a very minor treatment effect so you do it again but couldn't get the same growth chamber. In the second experiment you see a stronger effect but all the plants grew larger in the different chamber. In a third experiment you get the original chamber again and see a better treatment effect.

- What steps lead you do the correct test to make a decision on fertilizer?

R

```
> lme4::lmer(y ~ fertilizer + 1|experiment, data = blk_df)
```

Linear mixed model fit by REML ['lmerMod']

Formula: y ~ fertilizer + 1 | experiment

Data: blk_df

REML criterion at convergence: 238.2419

Random effects:

Groups	Name	Std.Dev.	Corr
experiment	(Intercept)	4.2020	
	fertilizer.L	1.2655	0.84

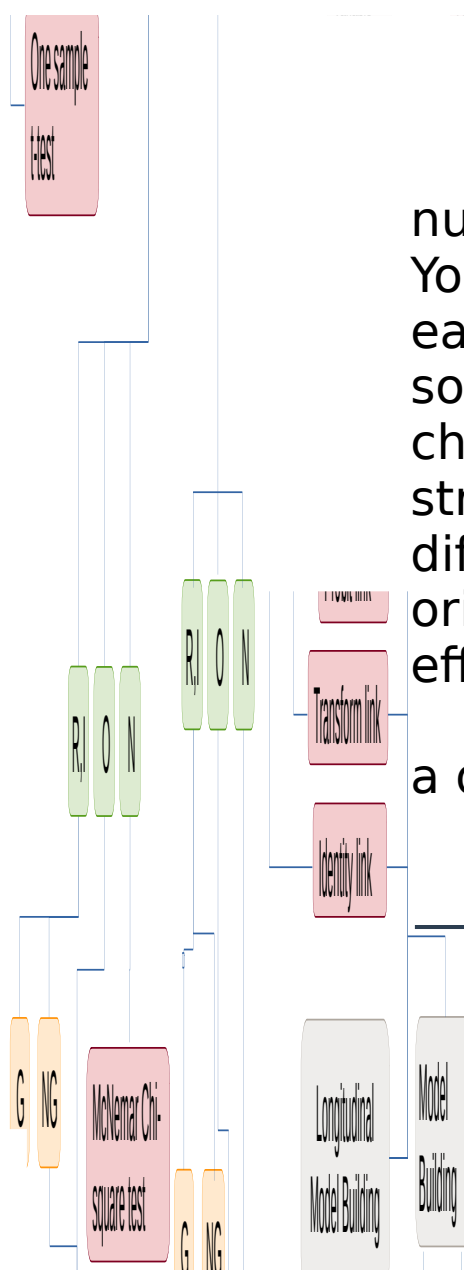
Residual	0.4247
----------	--------

Number of obs: 180, groups: experiment, 3

Fixed Effects:

(Intercept)

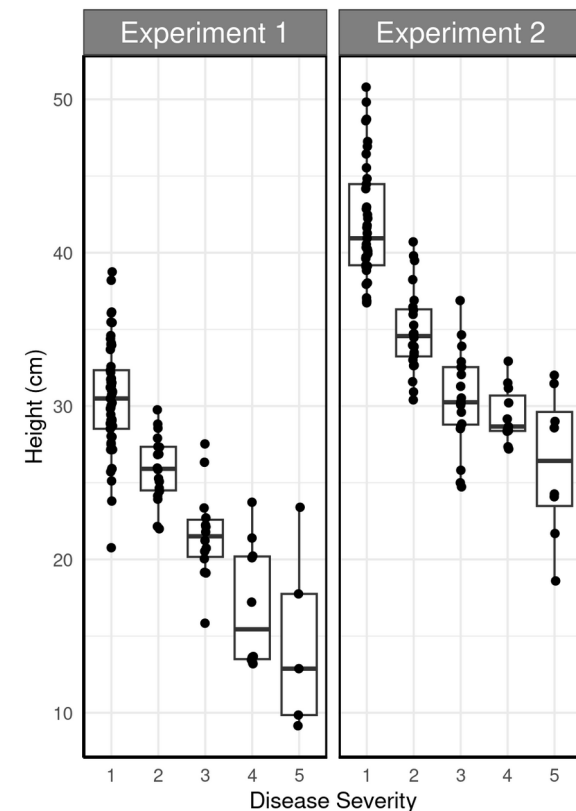
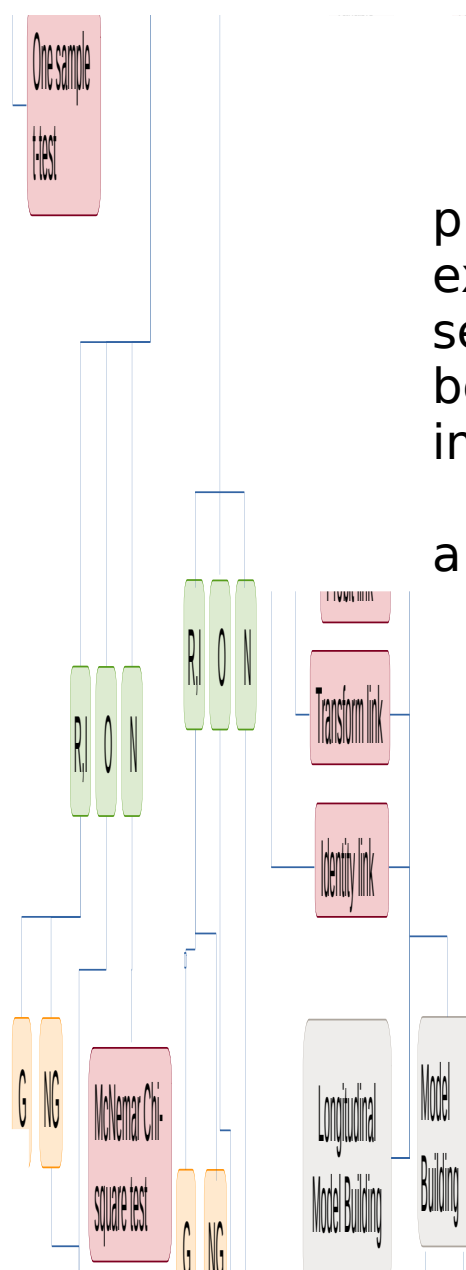
3.777



Scenario 8

You have run two experiments each with 50 plants infected with a virus. At the end of the experiment you label each plant with a disease severity score from 1-5 with 1 being healthiest and 5 being most diseased. You want to know if height is impaired by more severe disease scores.

- What steps lead you to the correct test to make a decision about disease severity scores and height?



Scenario 8

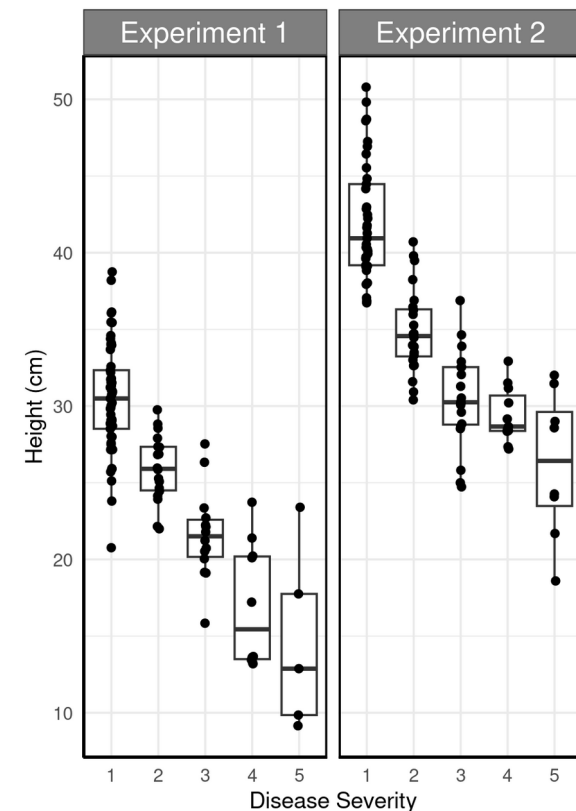
You have run two experiments each with 50 plants infected with a virus. At the end of the experiment you label each plant with a disease severity score from 1-5 with 1 being healthiest and 5 being most diseased. You want to know if height is impaired by more severe disease scores.

- What steps lead you to the correct test to make a decision about disease severity scores and height?

Answer

These data do show a block effect so we need to account for that.

Here our data is **O**, so we will use an **Probit link** in our model where we will compare our fixed effects.



Scenario 8

You have run two experiments each with 50 plants infected with a virus. At the end of the experiment you label each plant with a disease severity score from 1-5 with 1 being healthiest and 5 being most diseased. You want to know if height is impaired by more severe disease scores.

- What steps lead you to the correct test to make a decision about disease severity scores and height?



```
> m1<-glmmTMB(disease ~ height + (1|ex), data=df, family = binomial(link="probit"),
+             control = glmmTMBControl(optimizer=optim, optArgs=list(method="BFGS")) )
> fixef(m1)
```

Conditional model:

(Intercept)	height
12.5957	-0.3923

```
> summary(m1)$coefficients$cond
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	12.595671	2.33278905	5.399404	6.686264e-08
height	-0.392347	0.05374494	-7.300166	2.874125e-13

```
>
```

```
> m2 <- lme4::glmer(disease ~ height + (1|ex), data=df, family = binomial(link="probit"))
```

```
> fixef(m2)
```

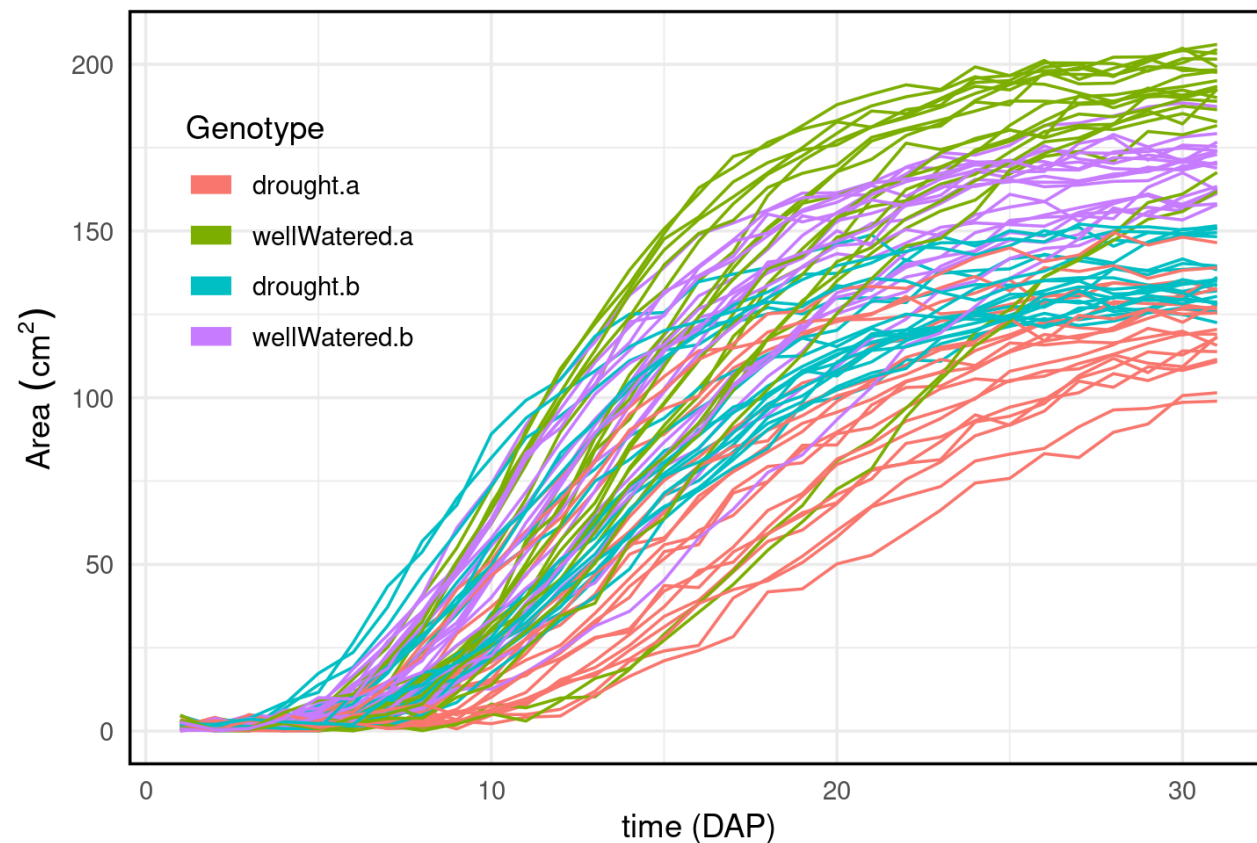
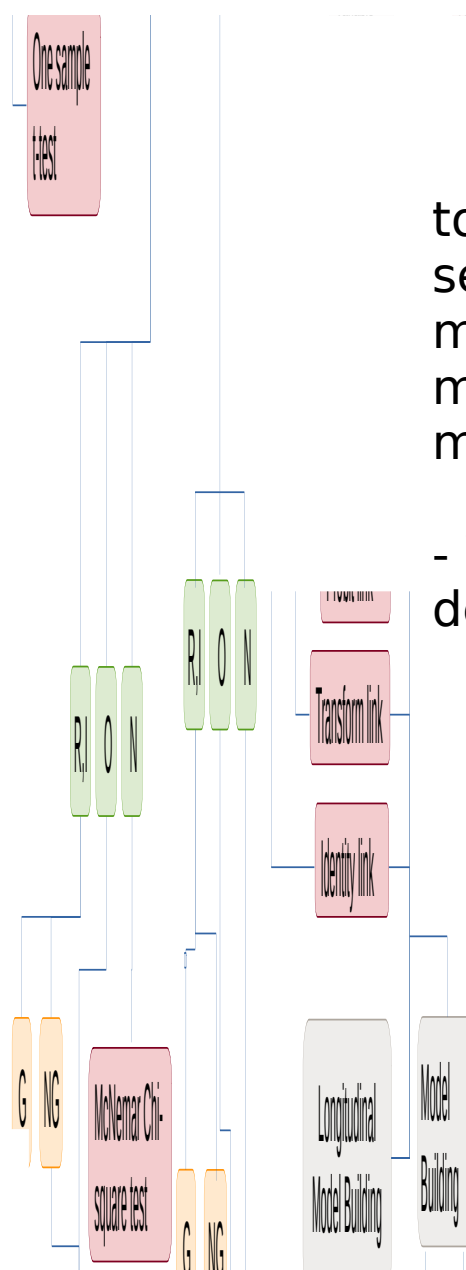
(Intercept)	height
12.6386559	-0.3936969

```
> #multcomp::glht()
```

Scenario 9

You are interested in testing the drought tolerance between two maize genotypes (a,b) so you set up an experiment where photos are taken of 20 maize plants per condition as they grow over one month. You analyze the images and extract area measurements from each.

- What steps lead you to the correct test to make a decision about drought tolerance?



Scenario 9

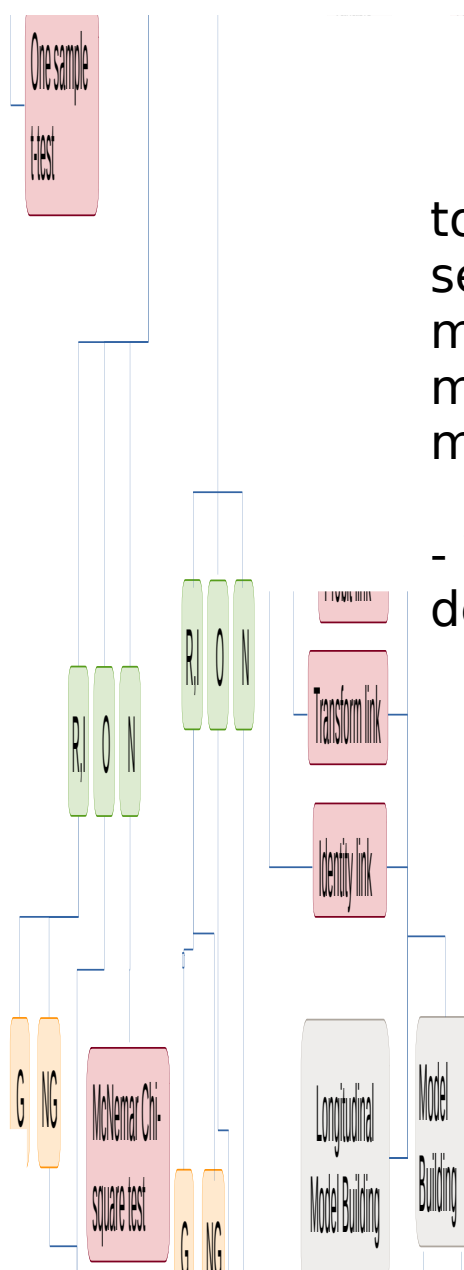
You are interested in testing the drought tolerance between two maize genotypes (a,b) so you set up an experiment where photos are taken of 20 maize plants per condition as they grow over one month. You analyze the images and extract area measurements from each.

- What steps lead you to the correct test to make a decision about drought tolerance?

Answer

This experiment was only conducted once and we do not see obvious confounding, so we have No block effects, but we collected many timepoints so we do have longitudinal data and take the Temporal path to model building.

Since our data is R,I and looks Gaussian we use the identity link to build our model before testing parameters using Contrasts.



Scenario 9

You are interested in testing the drought tolerance between two maize genotypes (a,b) so you set up an experiment where photos are taken of 20 maize plants per condition as they grow over one month. You analyze the images and extract area measurements from each.

- What steps lead you to the correct test to make a decision about drought tolerance?

R

```
> prior1 <- prior(lognormal(log(130), .25), nlpar = "A", lb=0) +
+   prior(lognormal(log(12), .25), nlpar = "B", lb=0) +
+   prior(lognormal(log(1.2), .25), nlpar = "C", lb=0) +
+   prior(lognormal(log(20), .25), nlpar = "subA", lb=0) +
+   prior(lognormal(log(12), .25), nlpar = "subB", lb=0) +
+   prior(lognormal(log(1.2), .25), nlpar = "subC", lb=0) +
+   prior(gamma(2,0.1), class="nu")
>
> fit9 <- brm(bf(y ~ A*exp(-B*exp(-C*time)),
+   nlf(sigma~ subA/(1+exp((subB-time)/subC))),
+   A+B+C+subA+subB+subC ~ 0+genotype:treatment,
+   autocor = ~arma(~time|sample:treatment:genotype,1,1), nl = TRUE),
+   family = student, prior = prior1, data = df, iter = 2000,
+   cores = 4, chains = 4, backend = "cmdstanr", silent=0,
+   control = list(adapt_delta = 0.999, max_treedepth = 20),
+   init = function(){list(b_A=rgamma(4,1), b_B=rgamma(4,1), b_C=rgamma(4,1),
+     b_subA=rgamma(4,1), b_subB=rgamma(4,1), b_subC=rgamma(4,1))})
```


Scenario 9

You are interested in testing the drought tolerance between two maize genotypes (a,b) so you set up an experiment where photos are taken of 20 maize plants per condition as they grow over one month. You analyze the images and extract area measurements from each.

- What steps lead you to the correct test to make a decision about drought tolerance?

R

```
> hypothesis(fit9, "1.2*(A_genotypea:treatmentdrought / A_genotypea:treatmentwellWatered)
< (A_genotypeb:treatmentdrought / A_genotypeb:treatmentwellWatered)")
```

Hypothesis Tests for class b:

	Hypothesis	Estimate	Est.Error	CI.Lower	CI.Upper	Evid.Ratio	Post.Prob
1	(1.2*(A_genotypea... < 0	-0.04	0.02	-0.07	0	20.51	0.95
	Star						
1	*						

'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.

Scenario 9

You are interested in testing the drought tolerance between two maize genotypes (a,b) so you set up an experiment where photos are taken of 20 maize plants per condition as they grow over one month. You analyze the images and extract area measurements from each.

- What steps lead you to the correct test to make a decision about drought tolerance?

R using **pcvr**

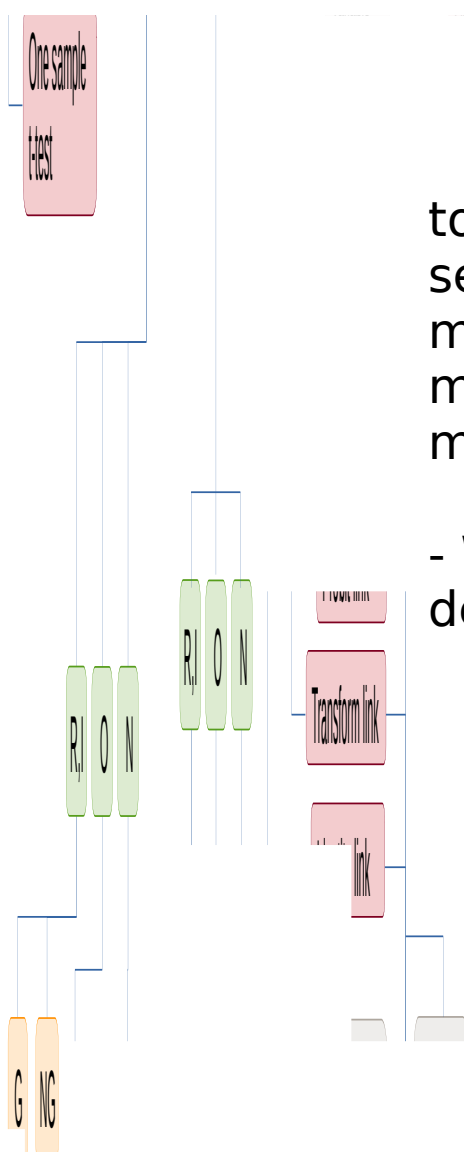
```
# devtools::install_github("danforthcenter/pcvr")
library(pcvr)
df$grouping <- interaction(df$treatment, df$genotype)
ss <- growthSS(model="gompertz", form=y~time|sample/grouping,
               df=df, sigma="logistic", type="brms",
               start = list("A"=130, "B"=12, "C"=1,
                           "subA"=20, "subB"=12, "subC"=1))
fit9_pcvr <- fitGrowth(ss, cores = 4, chains=4,
                      control = list(adapt_delta = 0.999, max_treedepth = 20))
```

Scenario 9

You are interested in testing the drought tolerance between two maize genotypes (a,b) so you set up an experiment where photos are taken of 20 maize plants per condition as they grow over one month. You analyze the images and extract area measurements from each.

- What steps lead you to the correct test to make a decision about drought tolerance?

R using **pcvr**



```
hypothesis(fit9_pcvr, paste0("1.2*(A_groupingdrought.a / A_groupingwellWatered.a) ",  
                             "< (A_groupingdrought.b / A_groupingwellWatered.b)"))
```

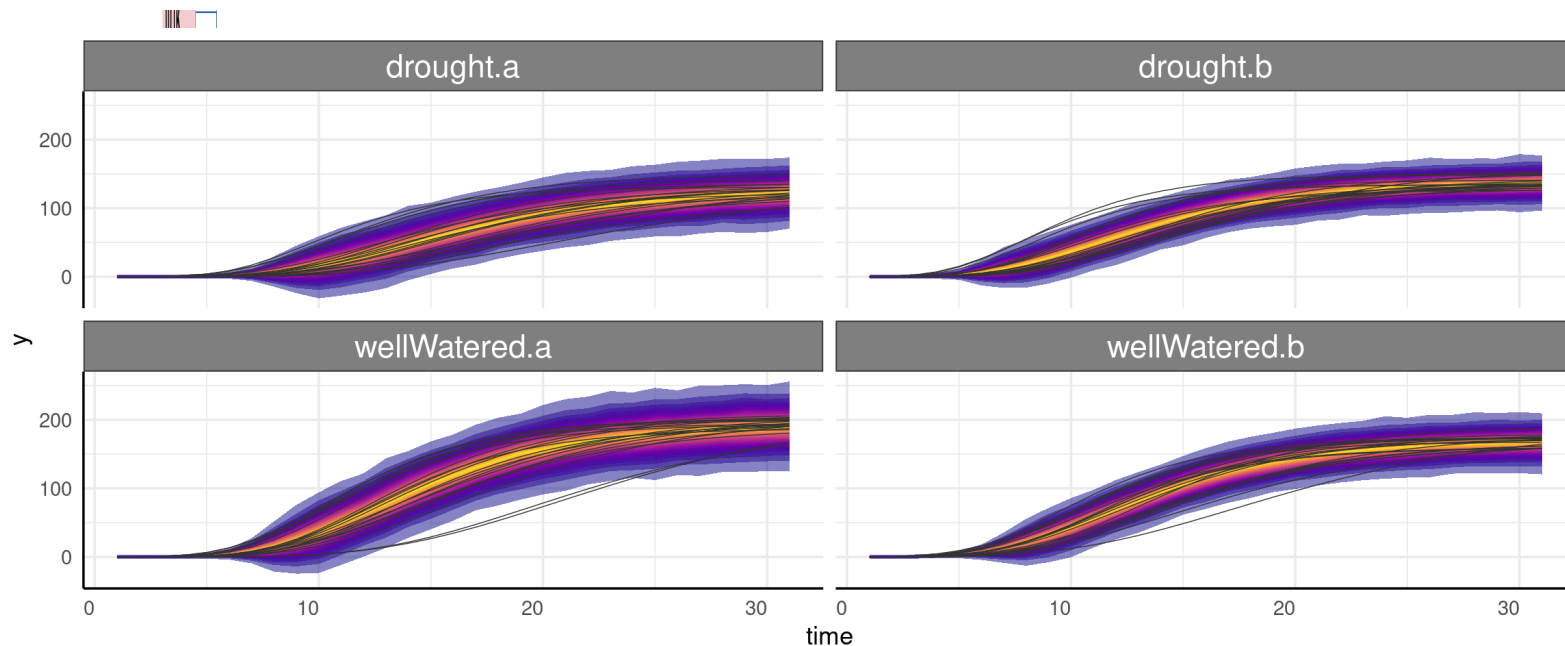
Scenario 9

You are interested in testing the drought tolerance between two maize genotypes (a,b) so you set up an experiment where photos are taken of 20 maize plants per condition as they grow over one month. You analyze the images and extract area measurements from each.

- What steps lead you to the correct test to make a decision about drought tolerance?

R using **pcvr**

```
growthPlot(fit9_pcvr, form=ss$pcvrForm, df=ss$df)
```



Scenario 9

You are interested in testing the drought tolerance between two maize genotypes (a,b) so you set up an experiment where photos are taken of 20 maize plants per condition as they grow over one month. You analyze the images and extract area measurements from each.

- What steps lead you to the correct test to make a decision about drought tolerance?

R using **pcvr**

growthSS {pcvr}

R Documentati

```
?pcvr::growthSS  
?pcvr::fitGrowth
```

Ease of use growth model helper function for 6 model parameterizations

Description

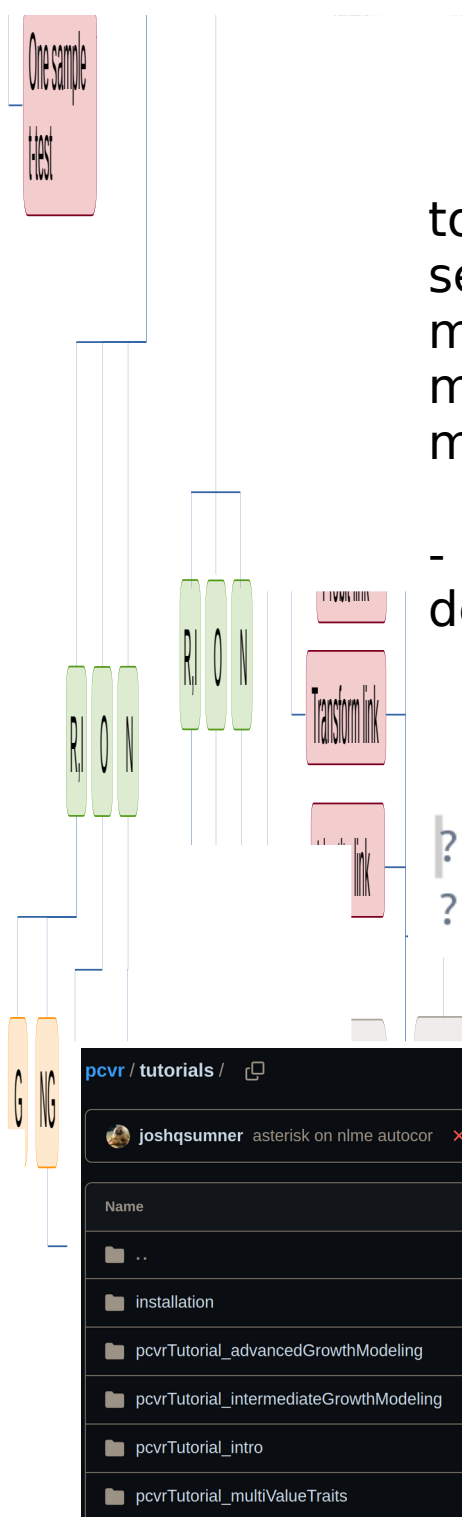
Ease of use growth model helper function for 6 model parameterizations

Usage

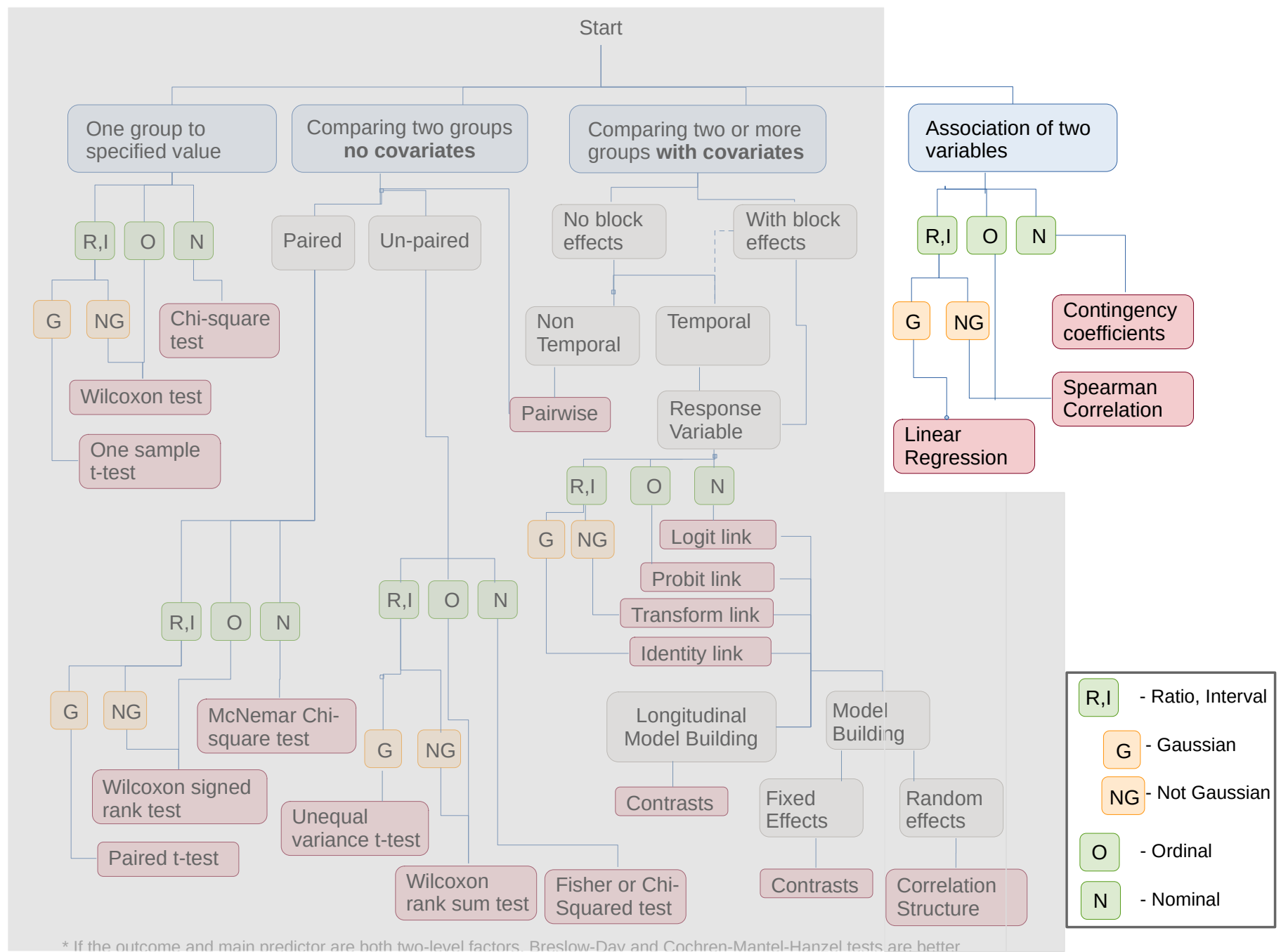
```
growthSS(  
  model,  
  form,  
  sigma = NULL,  
  df,  
  start = NULL,  
  pars = NULL,  
  type = "brms",  
  tau = 0.5  
)
```

Arguments

The name of a model as a character string. Supported options are c("logistic", "gompertz", "monomolecular", "exponential", "linear", "power law", "double logistic", "double gompertz", "gam", "int"), with "int" representing an intercept only model which is only used in brms (and is expected to only be used in threshold models or

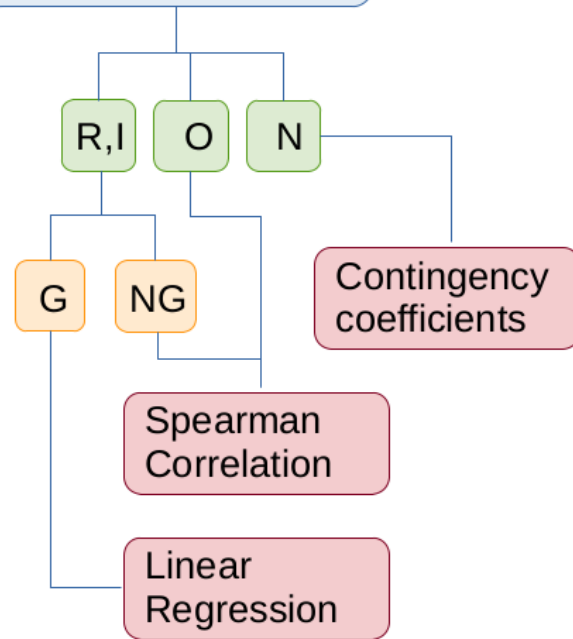


pcvr / tutorials /	
joshqsumner asterisk on nlme autocor	
Name	Last commit message
..	
installation	removing caches
pcvrTutorial_advancedGrowthModeling	longitudinal vignette updates
pcvrTutorial_intermediateGrowthModeling	asterisk on nlme autocor
pcvrTutorial_intro	oops
pcvrTutorial_multiValueTraits	mvt tutorial



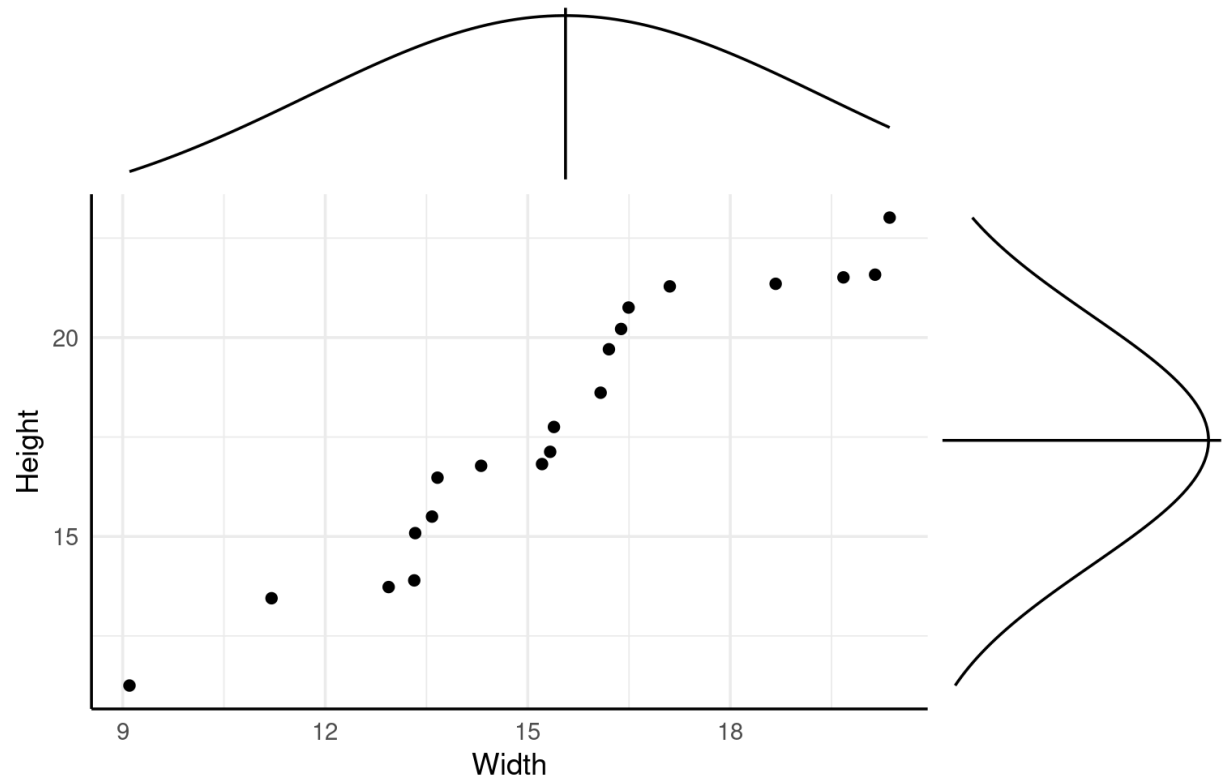
Scenario 10

Association of two variables



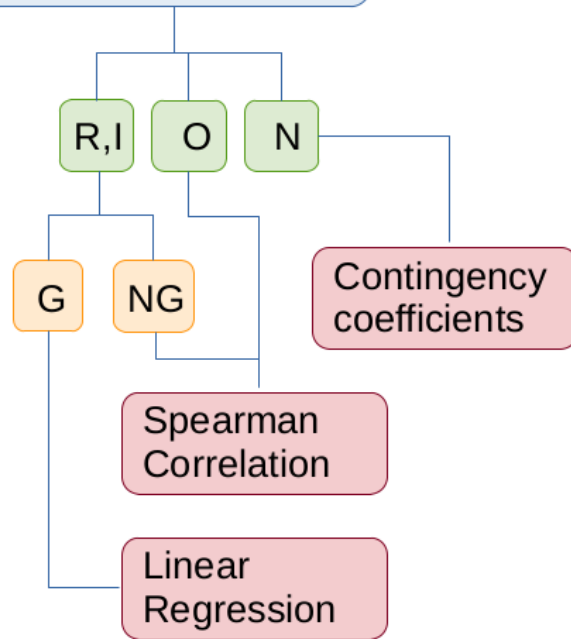
You're curious about general sorghum shape and decide to look at width vs height. You sow 20 plants and let them grow in the exact same conditions and after 2 weeks, you cut the shoot at soil level, lay them down and manually measure maximum width and height of the plant. You see a trend like below.

- What path should you follow to conclude they are associated?



Scenario 10

Association of two variables

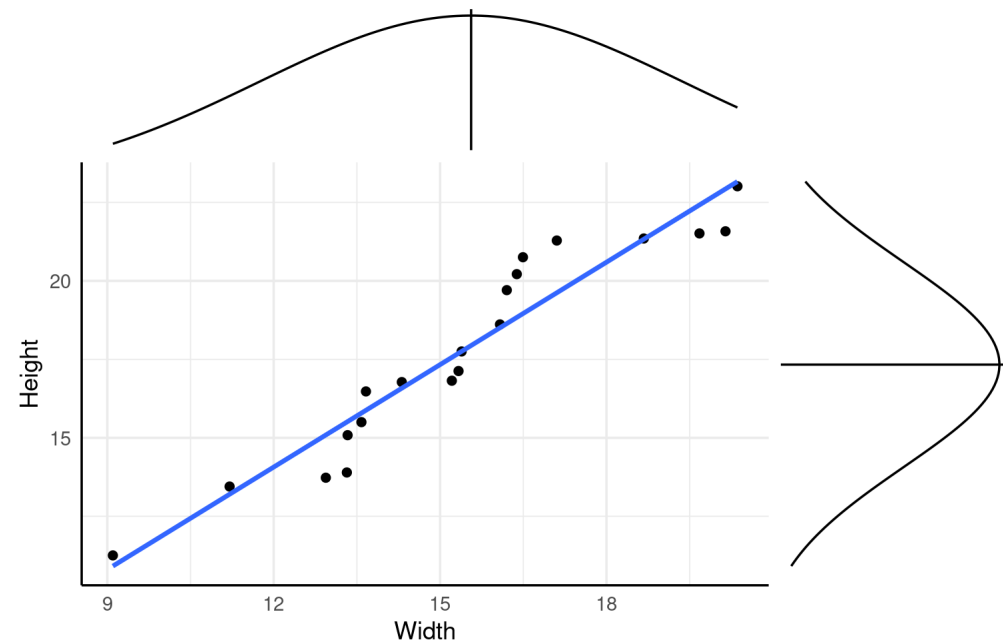


You're curious about general sorghum shape and decide to look at width vs height. You sow 20 plants and let them grow in the exact same conditions and after 2 weeks, you cut the shoot at soil level, lay them down and manually measure maximum width and height of the plant. You see a trend like below.

- What path should you follow to conclude they are associated?

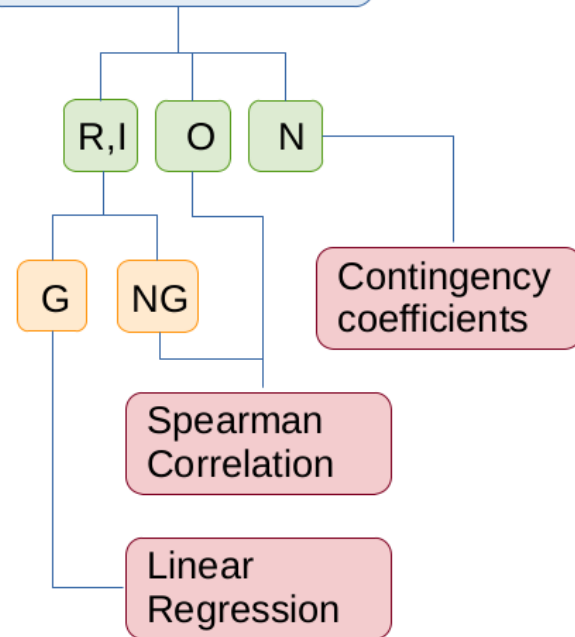
Answer

Both measures here are continuous so we can use the R,I path. The data looks Gaussian, so we will use linear regression.



Scenario 10

Association of two variables



You're curious about general sorghum shape and decide to look at width vs height. You sow 20 plants and let them grow in the exact same conditions and after 2 weeks, you cut the shoot at soil level, lay them down and manually measure maximum width and height of the plant. You see a trend like below.

- What path should you follow to conclude they are associated?

R

```
> summary(lm(x ~ y , df))
```

Call:

```
lm(formula = x ~ y, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.41925	-0.69753	0.04247	0.56226	1.53808

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.46582	1.09471	0.426	0.676
y	0.84063	0.06053	13.889	4.64e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

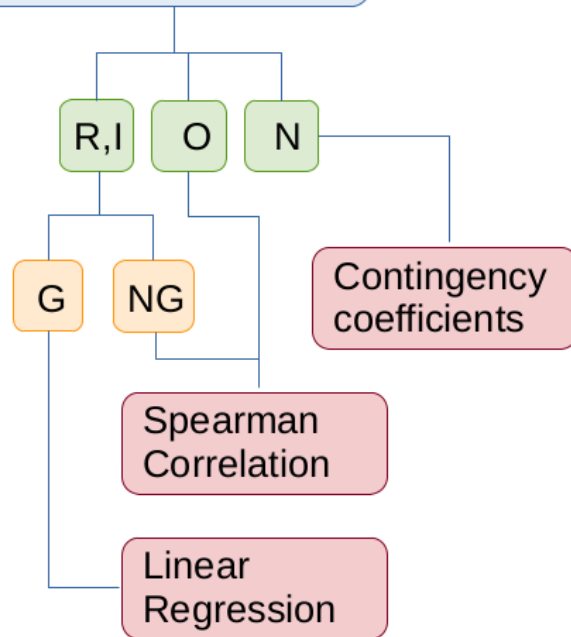
Residual standard error: 0.8758 on 18 degrees of freedom

Multiple R-squared: 0.9147, Adjusted R-squared: 0.9099

F-statistic: 192.9 on 1 and 18 DF, p-value: 4.638e-11

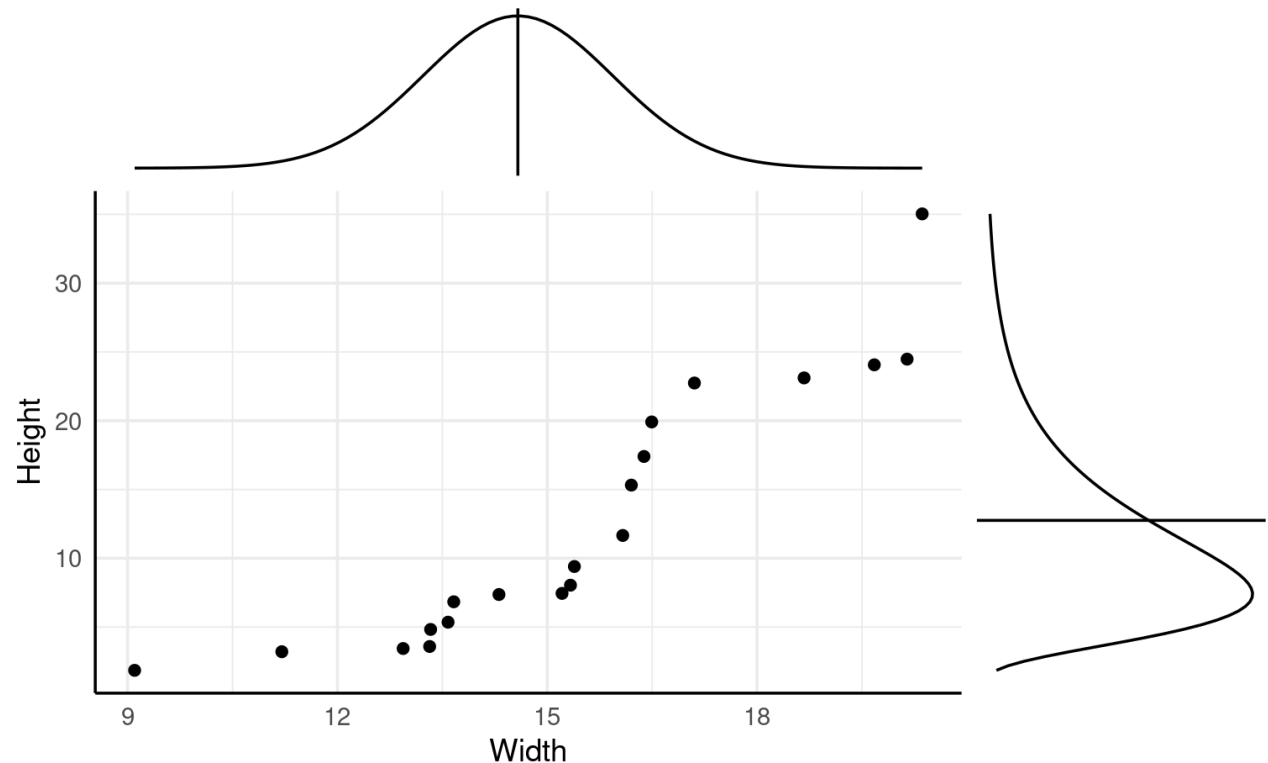
Scenario 10-2

Association of two variables



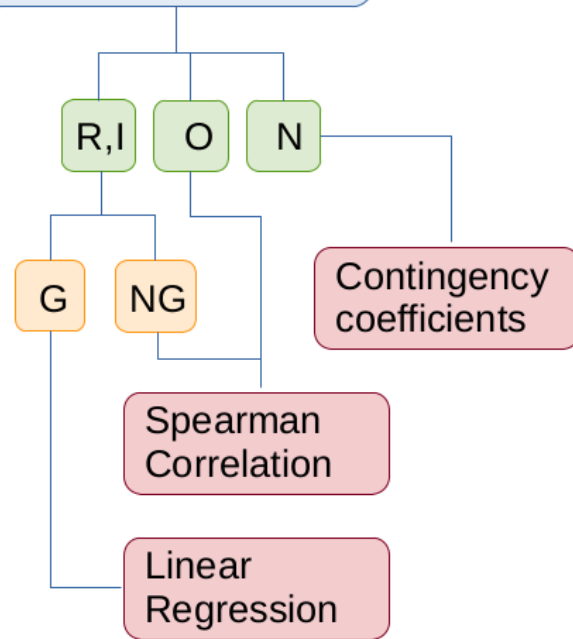
You're curious about general sorghum shape and decide to look at width vs height. You sow 20 plants and let them grow in the exact same conditions and after 2 weeks, you cut the shoot at soil level, lay them down and manually measure maximum width and height of the plant. You see a trend like below.

- What path should you follow to conclude they are associated?



Scenario 10-2

Association of two variables

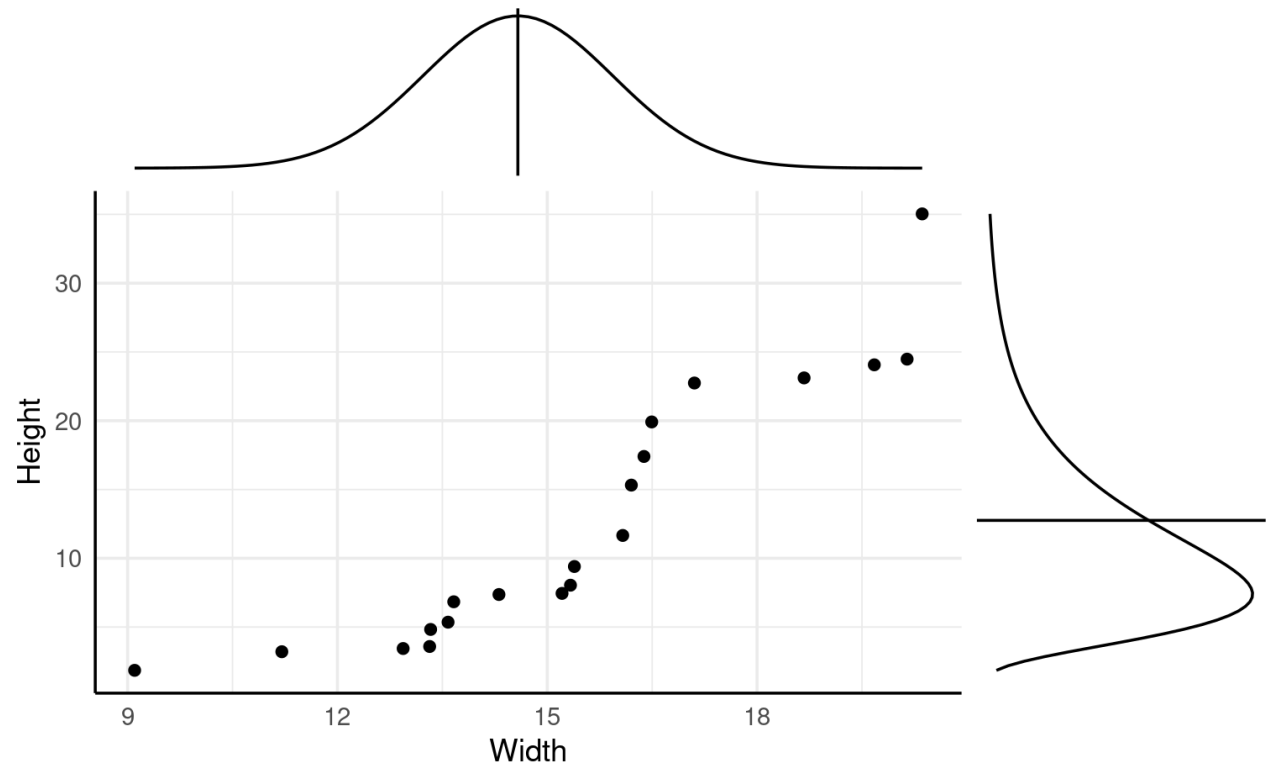


You're curious about general sorghum shape and decide to look at width vs height. You sow 20 plants and let them grow in the exact same conditions and after 2 weeks, you cut the shoot at soil level, lay them down and manually measure maximum width and height of the plant. You see a trend like below.

- What path should you follow to conclude they are associated?

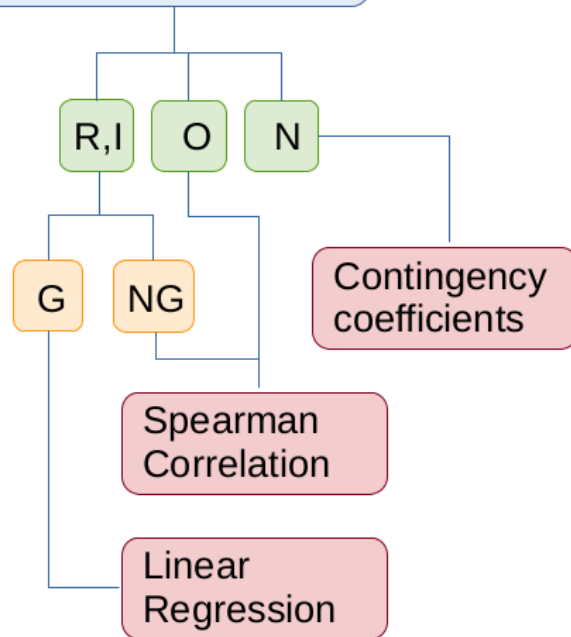
Answer

Both measures here are continuous so we can use the R,I path. The data is non-gaussian, so we will use Spearman Correlation.



Scenario 10-2

Association of two variables



You're curious about general sorghum shape and decide to look at width vs height. You sow 20 plants and let them grow in the exact same conditions and after 2 weeks, you cut the shoot at soil level, lay them down and manually measure maximum width and height of the plant. You see a trend like below.

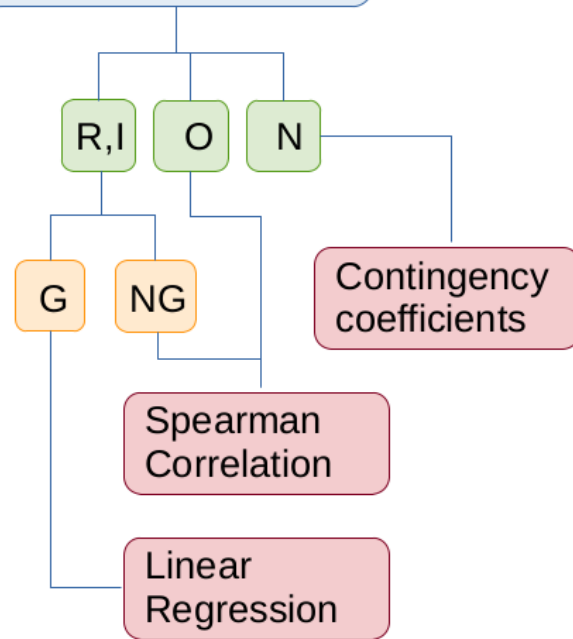
- What path should you follow to conclude they are associated?

R

```
> cor(df$x, df$y, method="spearman")  
[1] 1
```

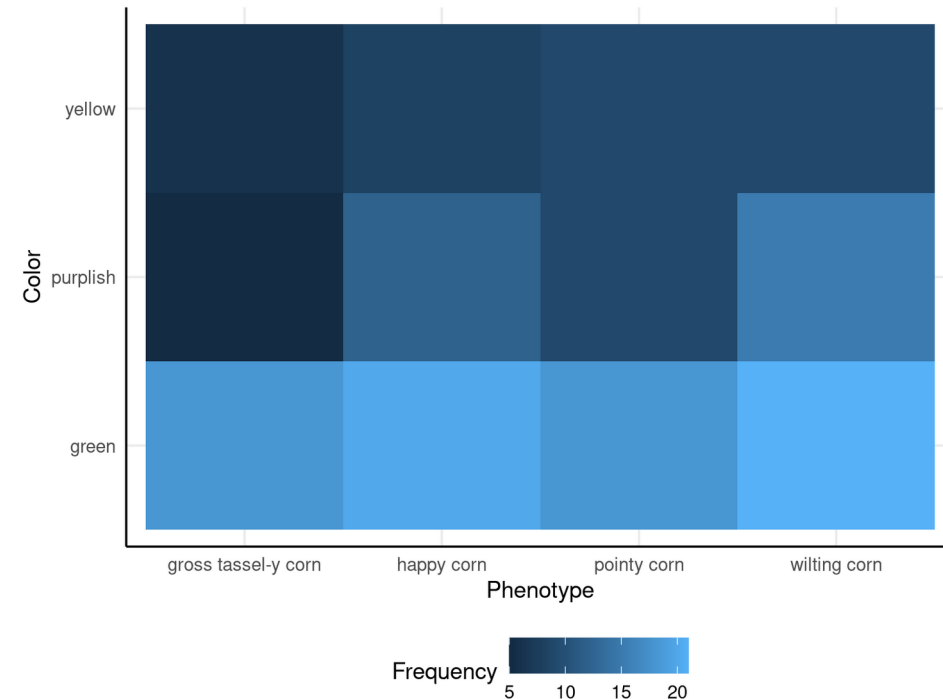
Scenario 11

Association of two variables



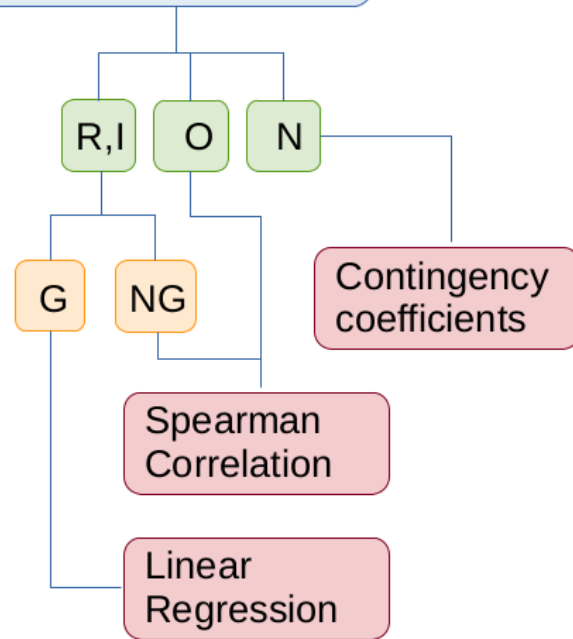
You are a statistician at a plant science center and you've been given a image dataset with no clear hypothesis and the question “are these groups different”. Without being given more clear instructions you categorize corn phenotypes as “pointy corn”, “wilting corn”, “happy corn” and “gross tassel-y corn” and colors as “green”, “yellow”, and “purplish”.

- What path should you follow to conclude they are associated?



Scenario 11

Association of two variables

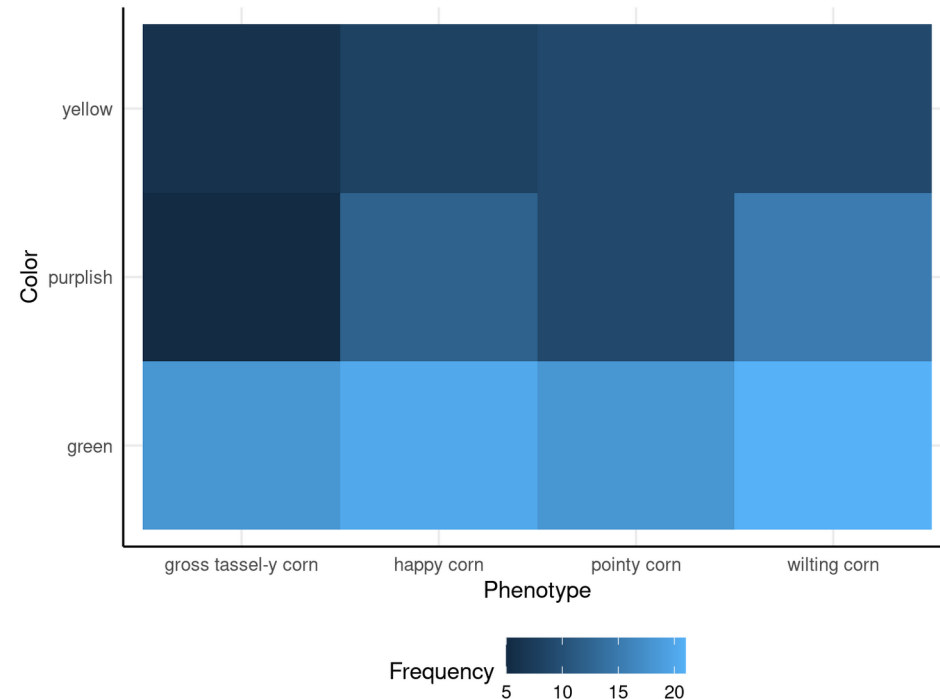


You are a statistician at a plant science center and you've been given a image dataset with no clear hypothesis and the question “are these groups different”. Without being given more clear instructions you categorize corn phenotypes as “pointy corn”, “wilting corn”, “happy corn” and “gross tassel-y corn” and colors as “green”, “yellow”, and “purplish”.

- What path should you follow to conclude they are associated?

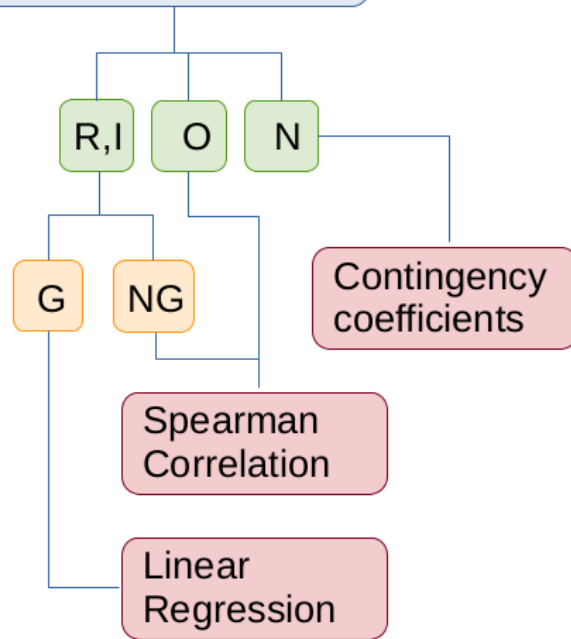
Answer

He have categorical unordered data so we have to use the **N** path, so we will use Contingency coefficients.



Scenario 11

Association of two variables

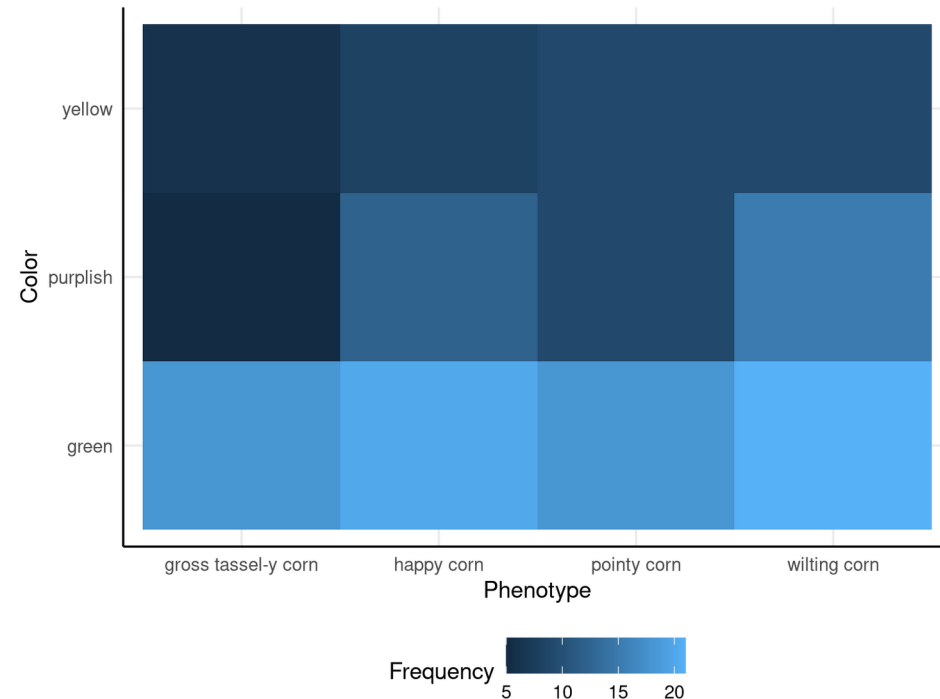


R

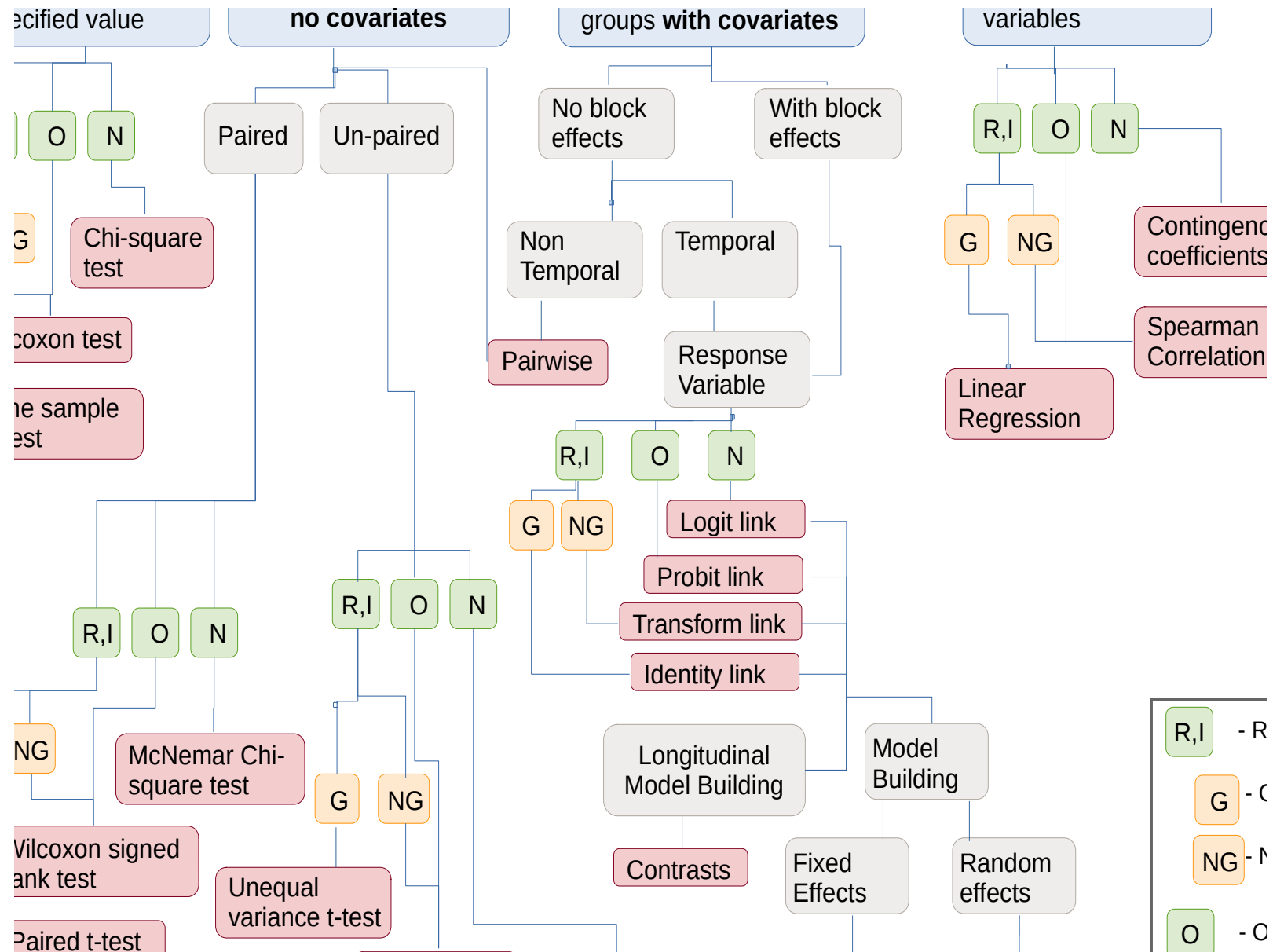
```
> chi <- as.numeric(chisq.test(ps, cs)$statistic)
> sqrt(chi/(chi+length(cs))) # in [0,1], 0 being independence
[1] 0.140501
```

You are a statistician at a plant science center and you've been given a image dataset with no clear hypothesis and the question “are these groups different”. Without being given more clear instructions you categorize corn phenotypes as “pointy corn”, “wilting corn”, “happy corn” and “gross tassel-y corn” and colors as “green”, “yellow”, and “purplish”.

- What path should you follow to conclude they are associated?



Conclusion



Conclusion

- Tomorrow we will expand on this some talking about pcvr and:
 - Intro to Bayesian statistics
 - More about longitudinal modeling
 - Non-gaussian distributions