Statistics in Biological Science

Dec 6-8

Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

Reading:

- Pollard et al. 2019 Empowering statistical methods for cellular and molecular biologists. Molecular Biology of the Cell 30: 1359-1368
- Broman and Woo, 2018. Data organization in spreadsheets. The American Statistician 71:1 2-10 DOI: 10.1080/00031305.2017.1375989

Software:

- R: https://www.r-project.org
- Rstudio: https://www.rstudio.com

Resources:

- *The Analysis of Biological Data* Whitlock and Schluter 2009
- Canvas folder

Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

Question Experiment Data Collection Statistical Analysis

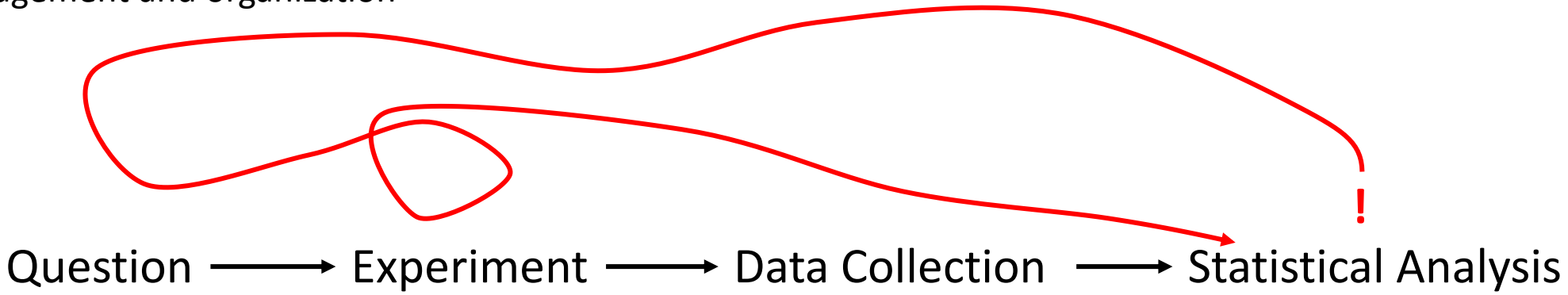
Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

Question ⟶ Experiment ⟶ Data Collection ⟶ Statistical Analysis

Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization



Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

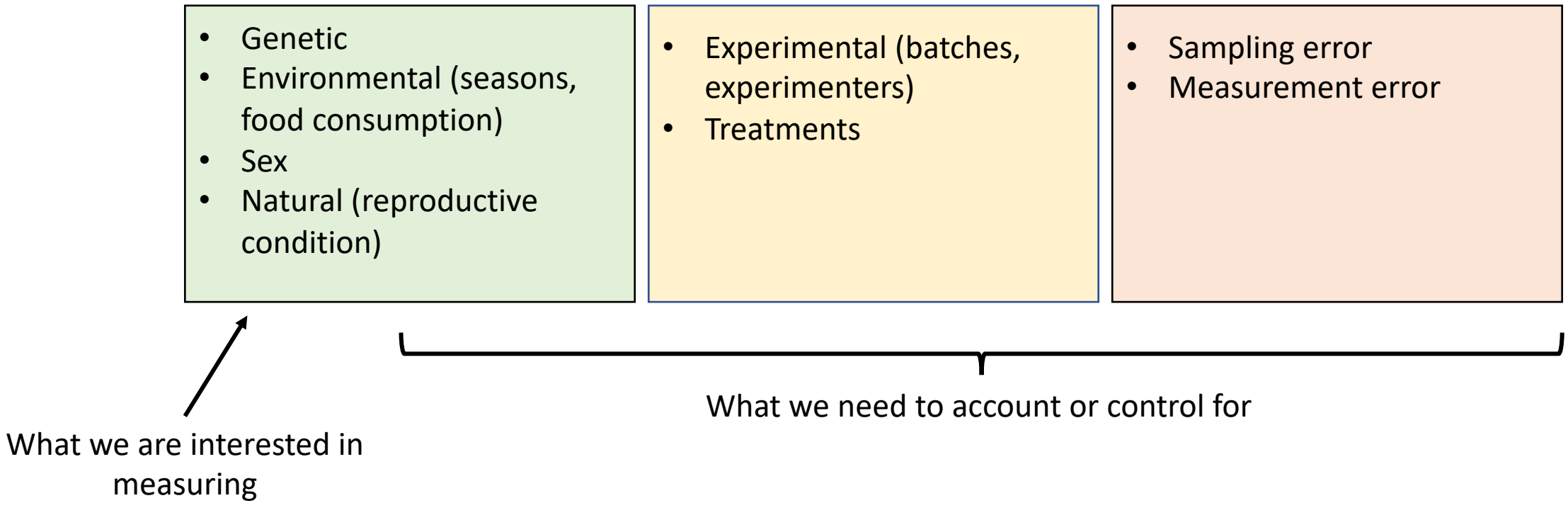
Sources of Variation

- Genetic
- Environmental (seasons, food consumption)
- Sex
- Natural (reproductive condition)

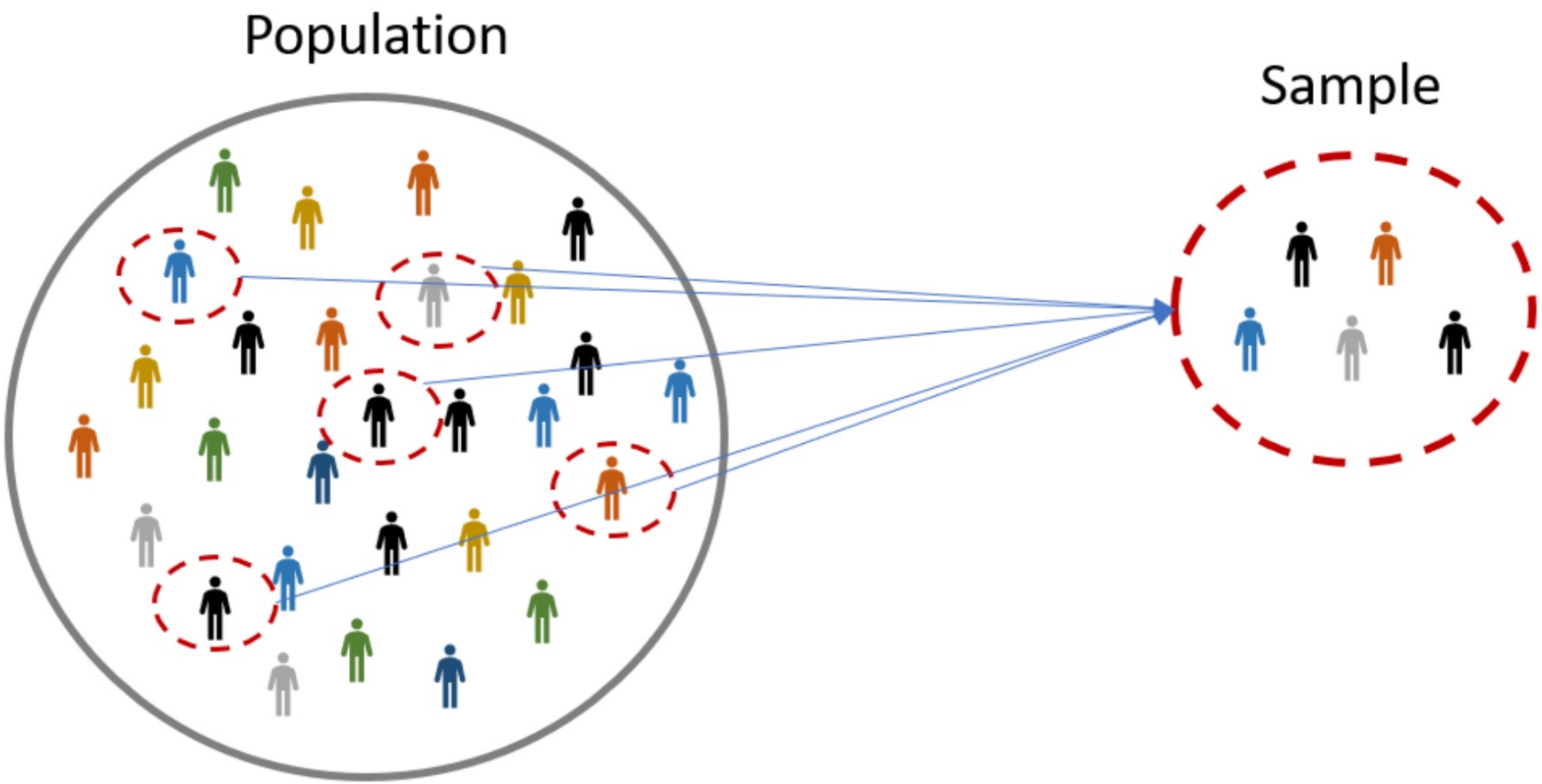
- Experimental (batches, experimenters)
- Treatments

- Sampling error
- Measurement error

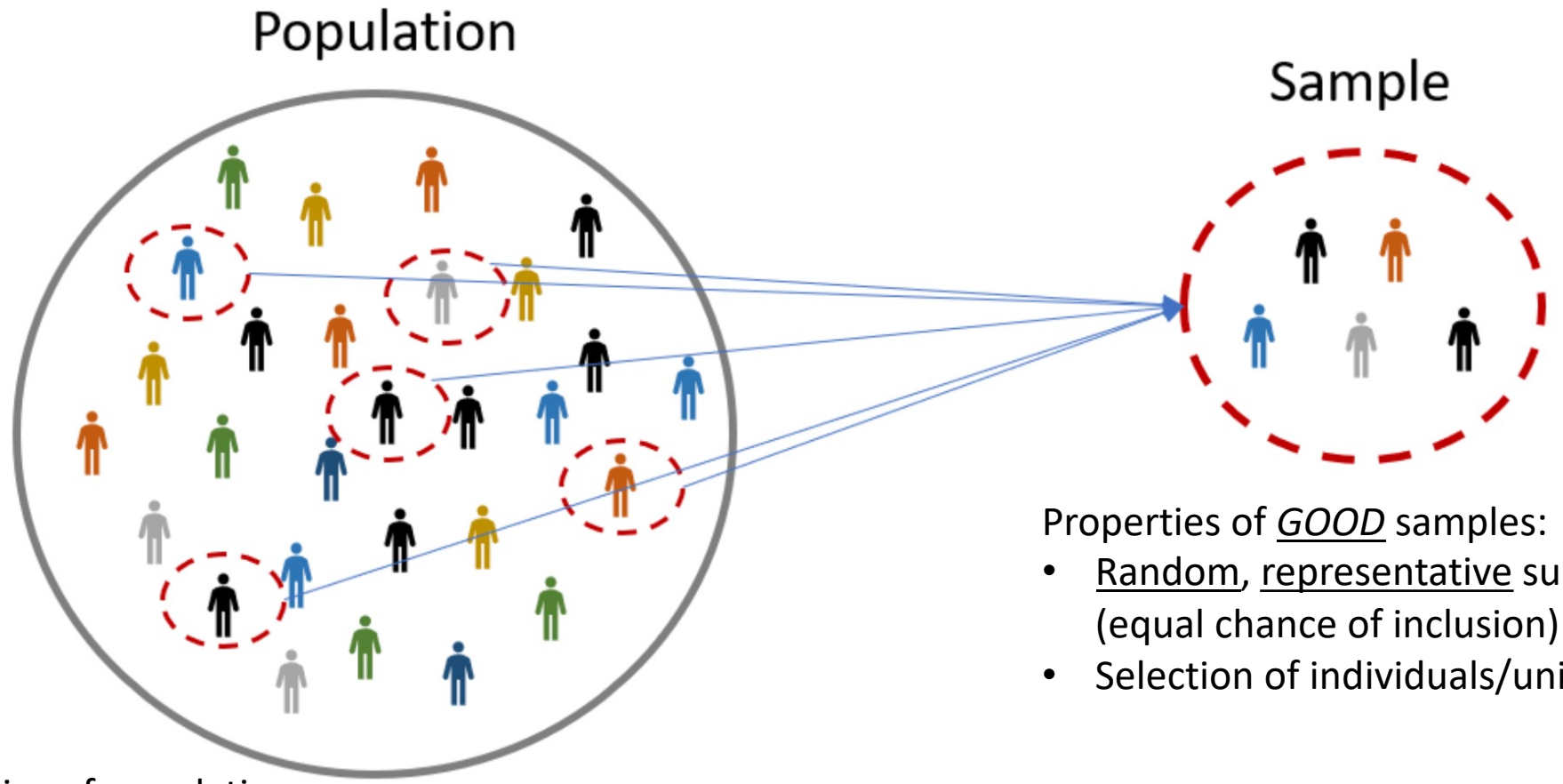
Sources of Variation



Sampling Error: Deviation of the sample mean from the population mean



Sampling Error: Deviation of the sample mean from the population mean



Properties of populations:

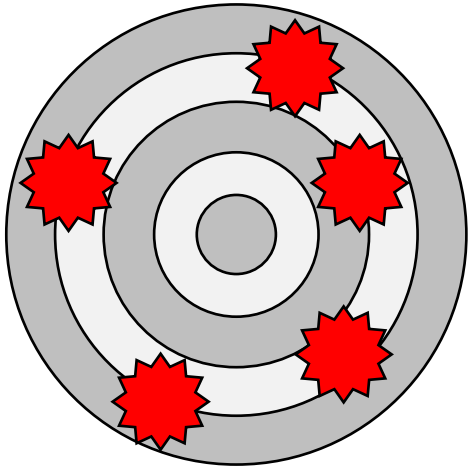
- All the individuals/units of interest
- Assumed to be infinitely large

Properties of *GOOD* samples:

- Random, representative subset of population (equal chance of inclusion)
- Selection of individuals/units is independent

Sampling Error: Deviation of the sample mean from the population mean is exacerbated by **Measurement Error**

Precision, Accuracy, Repeatability, and Bias:



Not Accurate ← Away from bullseye

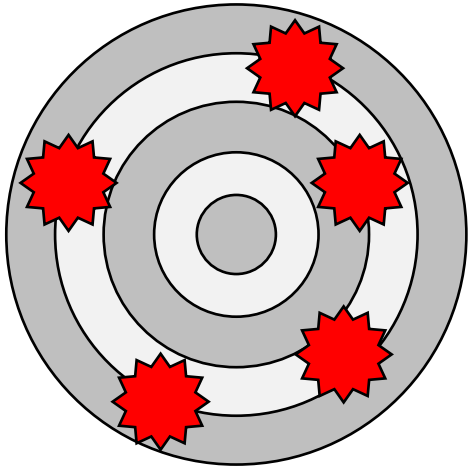
Not Precise ← Scattered and large

Not Repeatable ← Scattered

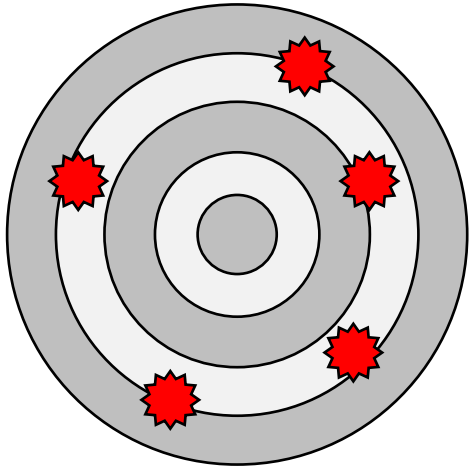
Not Biased ← Distribution from target

Sampling Error: Deviation of the sample mean from the population mean is exacerbated by **Measurement Error**

Precision, Accuracy, Repeatability, and Bias:



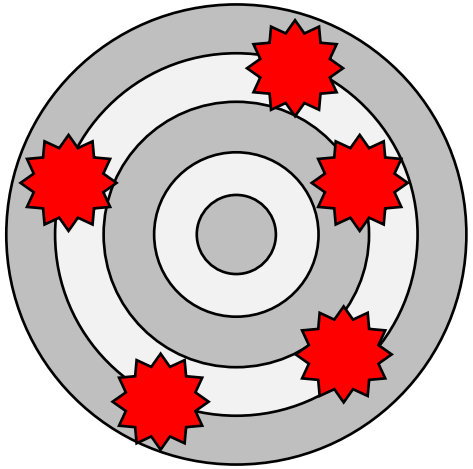
Not Accurate
Not Precise
Not Repeatable
Not Biased



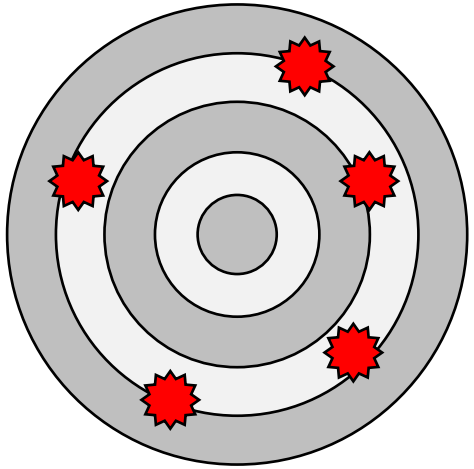
Not Accurate
More Precise
Not Repeatable
Not Biased

Sampling Error: Deviation of the sample mean from the population mean is exacerbated by **Measurement Error**

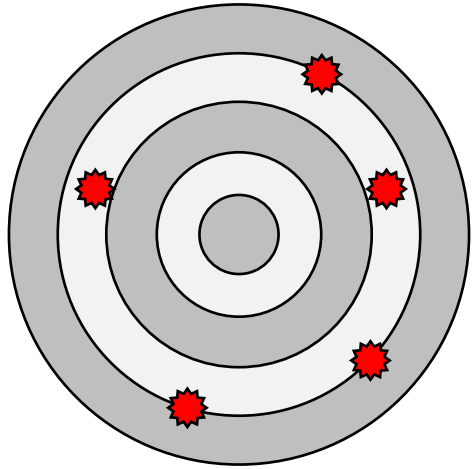
Precision, Accuracy, Repeatability, and Bias:



Not Accurate
Not Precise
Not Repeatable
Not Biased



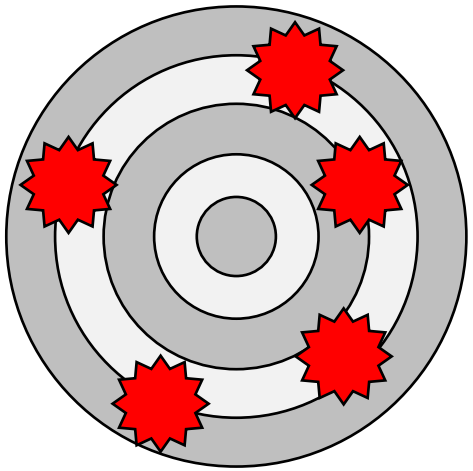
Not Accurate
More Precise
Not Repeatable
Not Biased



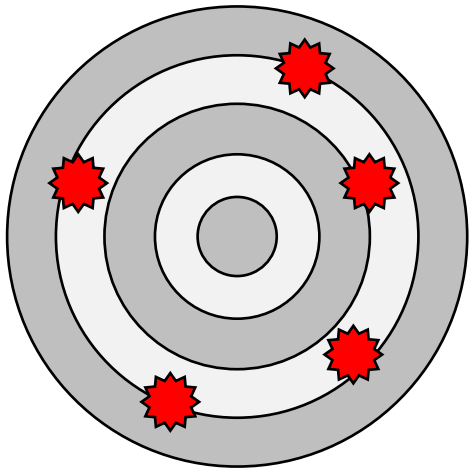
Not Accurate
Even More Precise
Not Repeatable
Not Biased

Sampling Error: Deviation of the sample mean from the population mean is exacerbated by **Measurement Error**

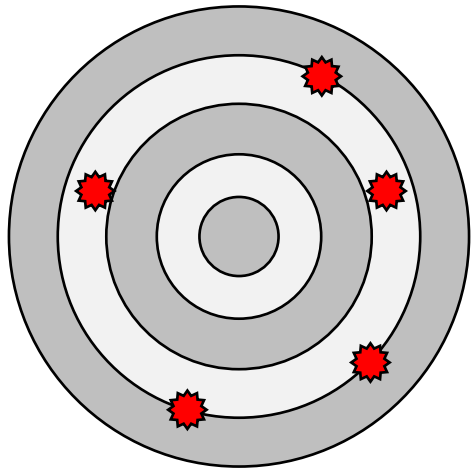
Precision, Accuracy, Repeatability, and Bias:



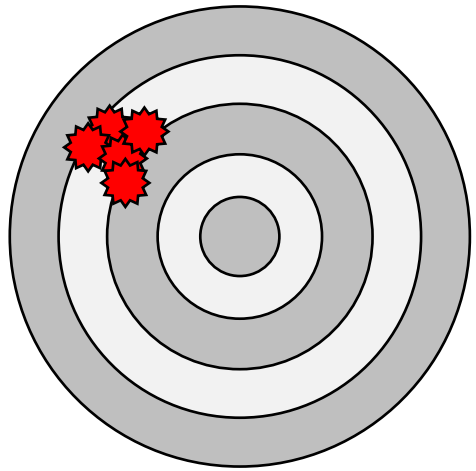
Not Accurate
Not Precise
Not Repeatable
Not Biased



Not Accurate
More Precise
Not Repeatable
Not Biased



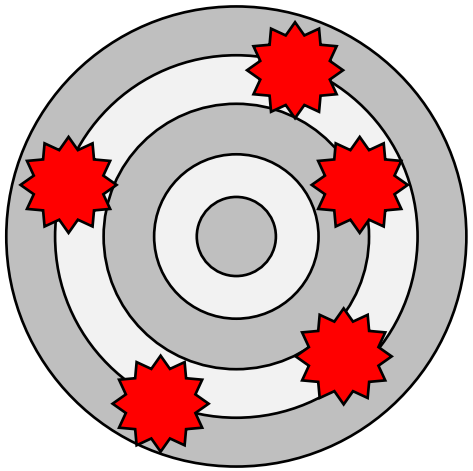
Not Accurate
Even More Precise
Not Repeatable
Not Biased



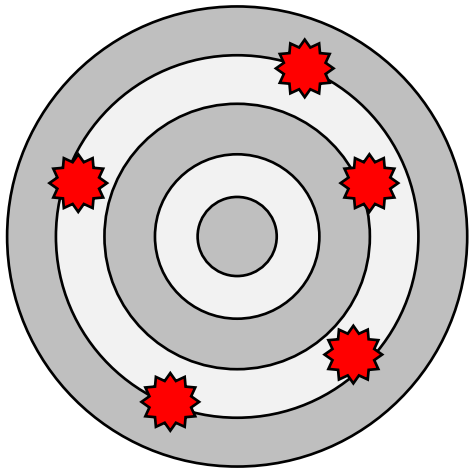
Not Accurate
More Precise
More Repeatable
Biased

Sampling Error: Deviation of the sample mean from the population mean is exacerbated by **Measurement Error**

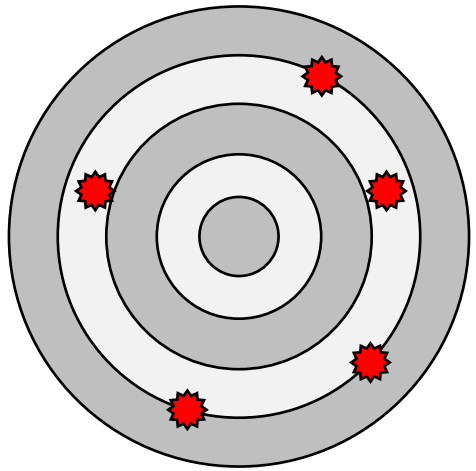
Precision, Accuracy, Repeatability, and Bias:



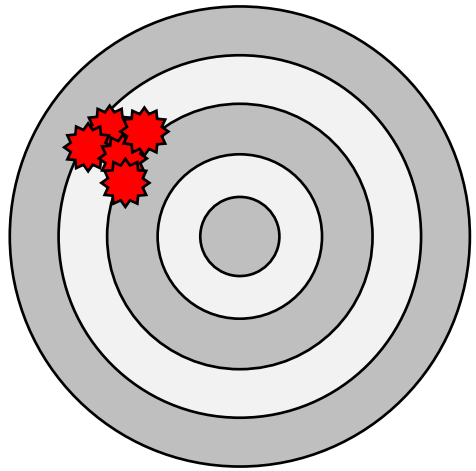
Not Accurate
Not Precise
Not Repeatable
Not Biased



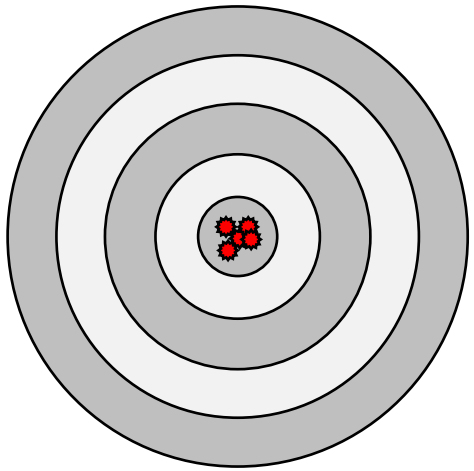
Not Accurate
More Precise
Not Repeatable
Not Biased



Not Accurate
Even More Precise
Not Repeatable
Not Biased



Not Accurate
More Precise
More Repeatable
Biased



Most Accurate
Most Precise
Most Repeatable
Not Biased

What do we know so far?

- Need representative, random, and independent samples
- Need to *maximize* precision, repeatability, and accuracy and *minimize* bias

What do we know so far?

- Need representative, random, and independent samples
- Need to *maximize* precision, repeatability, and accuracy and *minimize* bias

How do we know when a sample is “large enough”?

Influenced by:

- Properties of the population of interest = the shape of the distribution of observations in the population

What do we know so far?

- Need representative, random, and independent samples
- Need to *maximize* precision, repeatability, and accuracy and *minimize* bias

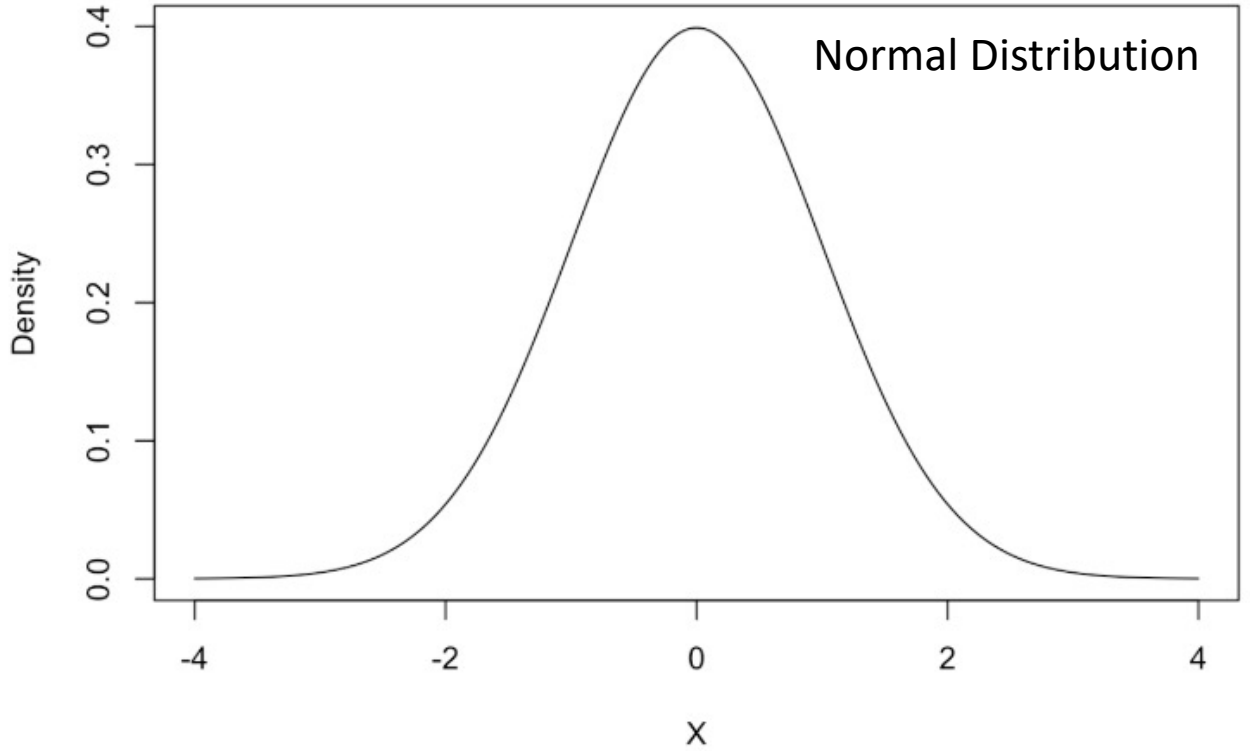
How do we know when a sample is "large enough"?

Influenced by:

- Properties of the population of interest = the shape of the distribution of observations in the population

Properties of a Normal Distribution:

- Continuous distribution, meaning probability is measure by the area under the curve rather than height of the curve
- Symmetric around the mean
- Single mode
- Probability density is highest exactly at the mean
- Mean, median, and mode are all equal
- About 2/3 of observations fall within 1 standard deviation (SD) of the mean; 95% fall within 2 SD of the mean



What do we know so far?

- Need representative, random, and independent samples
- Need to *maximize* precision, repeatability, and accuracy and *minimize* bias

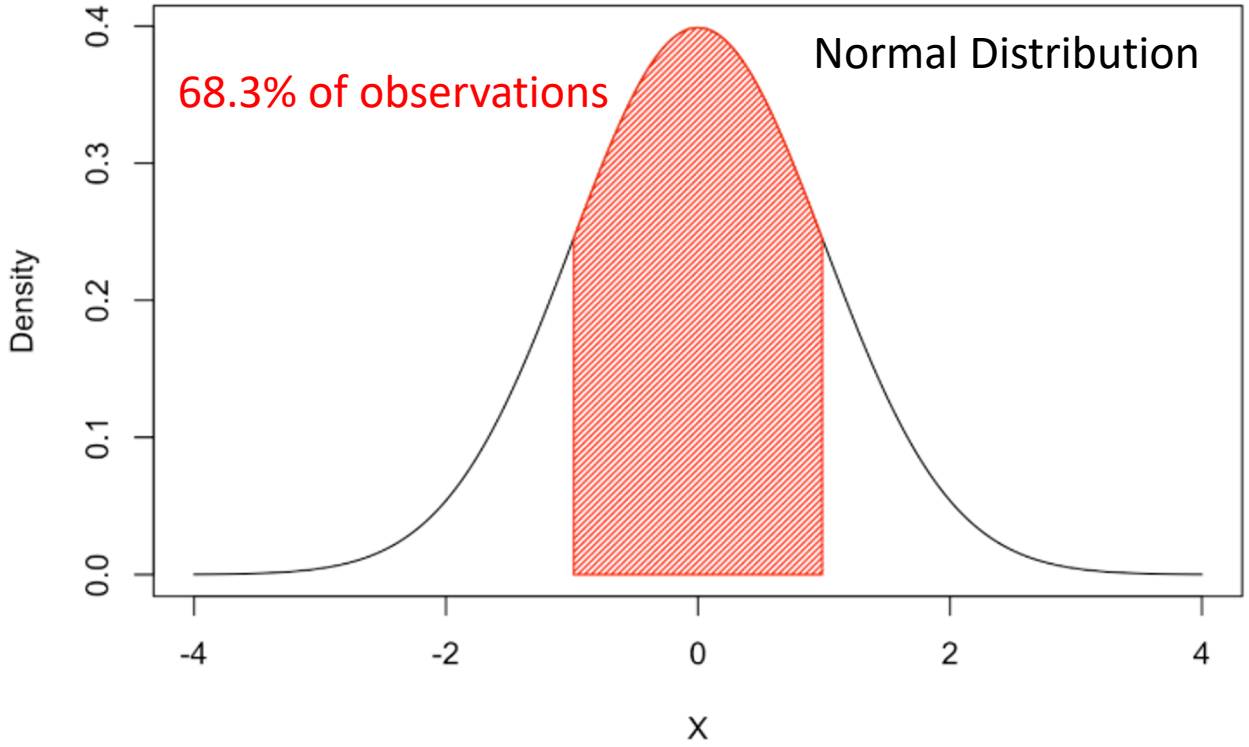
How do we know when a sample is "large enough"?

Influenced by:

- Properties of the population of interest = the shape of the distribution of observations in the population

Properties of a Normal Distribution:

- Continuous distribution, meaning probability is measure by the area under the curve rather than height of the curve
- Symmetric around the mean
- Single mode
- Probability density is highest exactly at the mean
- Mean, median, and mode are all equal
- About 2/3 of observations fall within 1 standard deviation (SD) of the mean; 95% fall within 2 SD of the mean



What do we know so far?

- Need representative, random, and independent samples
- Need to *maximize* precision, repeatability, and accuracy and *minimize* bias

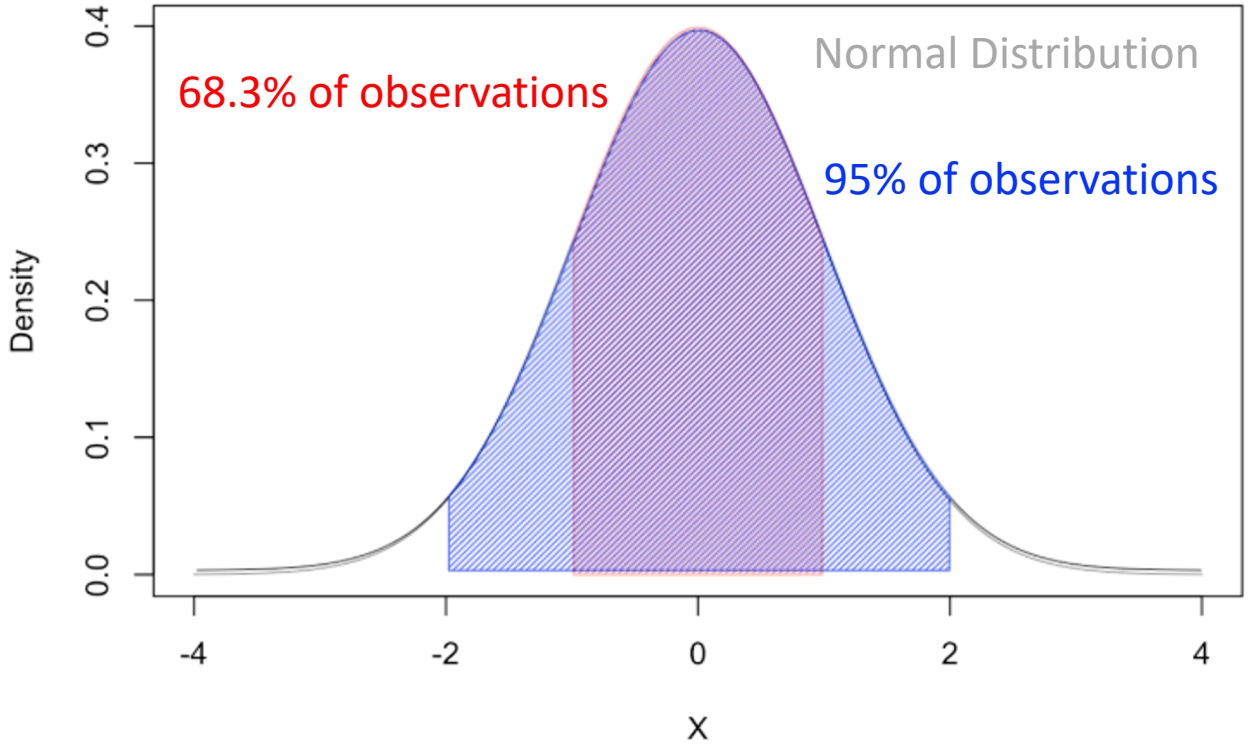
How do we know when a sample is "large enough"?

Influenced by:

- Properties of the population of interest = the shape of the distribution of observations in the population

Properties of a Normal Distribution:

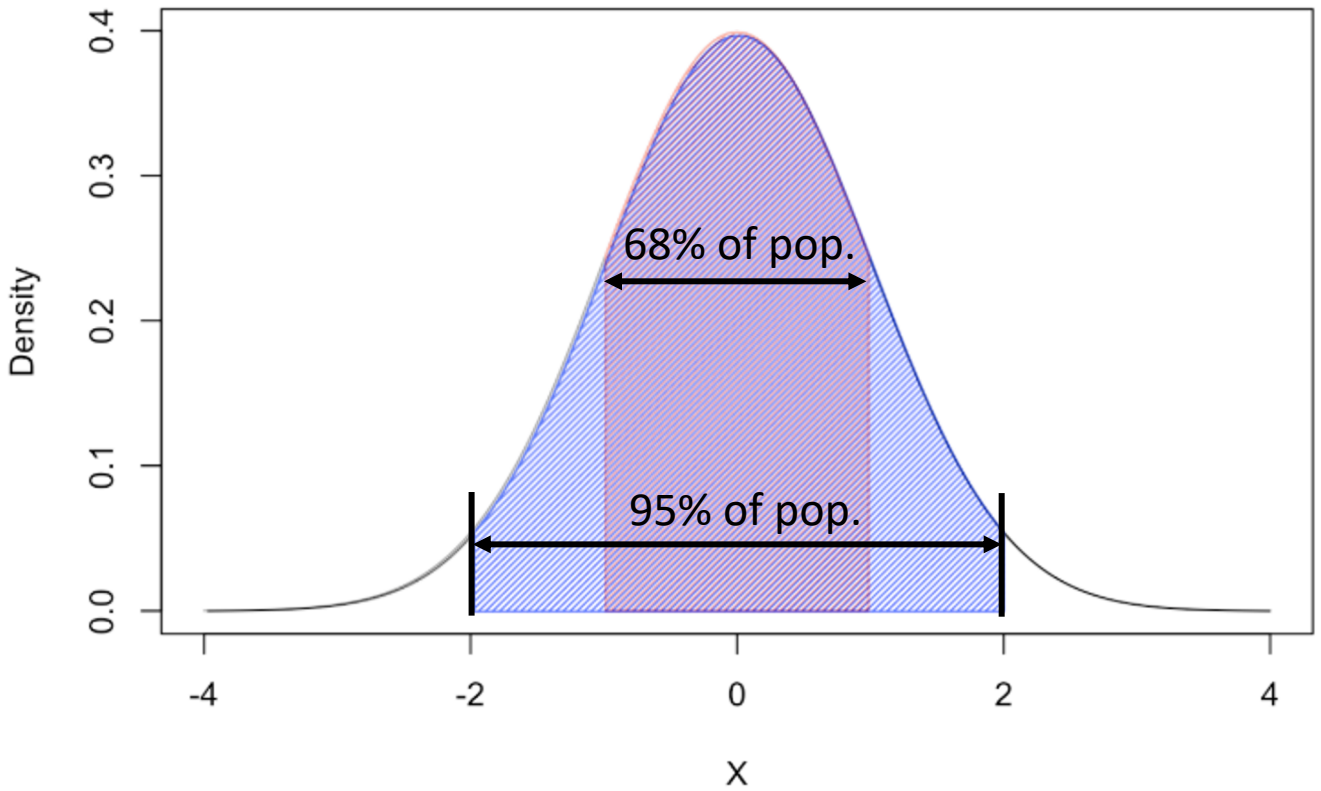
- Continuous distribution, meaning probability is measure by the area under the curve rather than height of the curve
- Symmetric around the mean
- Single mode
- Probability density is highest exactly at the mean
- Mean, median, and mode are all equal
- About 2/3 of observations fall within 1 standard deviation (SD) of the mean; 95% fall within 2 SD of the mean



The Empirical Rule:

When data are normally distributed, we can assume

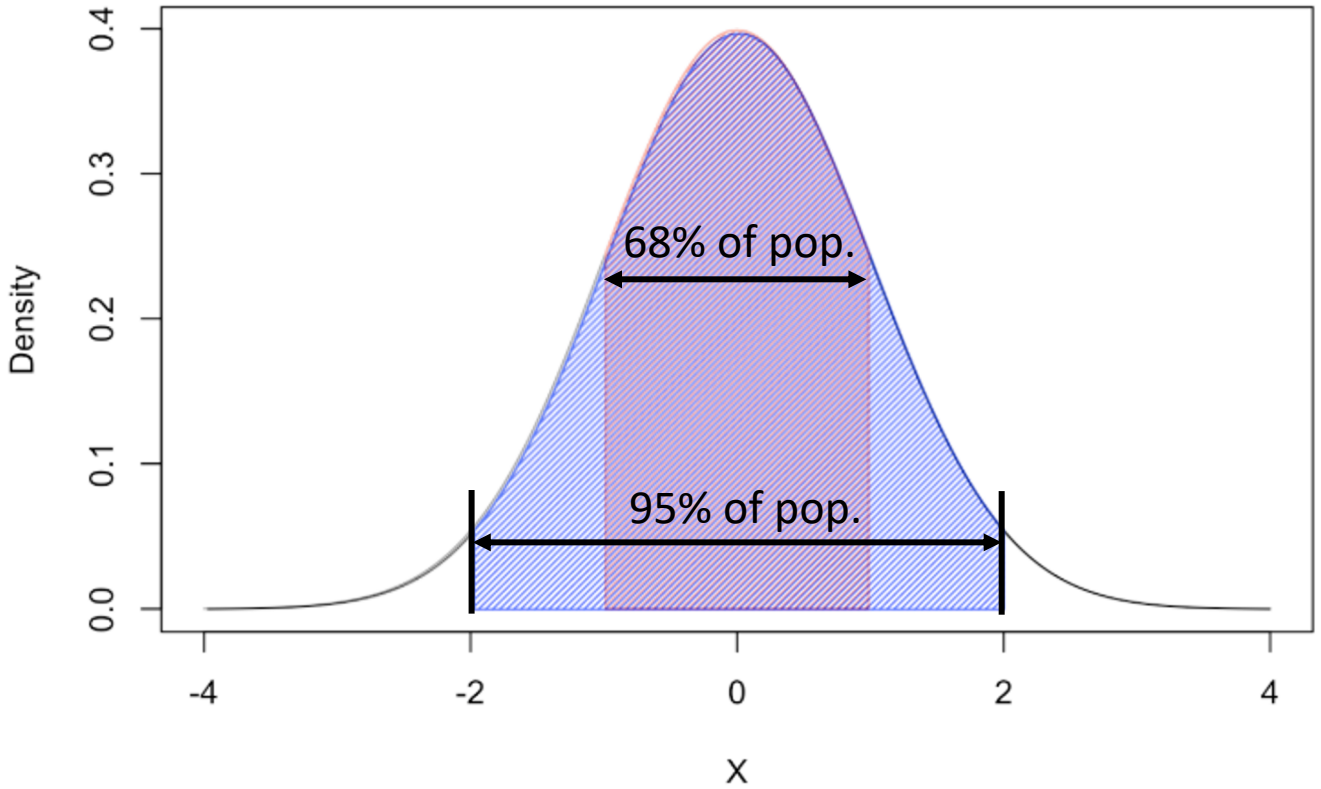
- ~68% of the population lies within +/- 1 SD of the mean
- ~95% of the population lies within +/- 2 SD of the mean
- This property is the basis for all parametric statistical tests (t tests, ANOVA, linear regression, etc.)



The Empirical Rule:

When data are normally distributed, we can assume

- ~68% of the population lies within +/- 1 SD of the mean
- ~95% of the population lies within +/- 2 SD of the mean
- This property is the basis for all parametric statistical tests (t tests, ANOVA, linear regression, etc.)

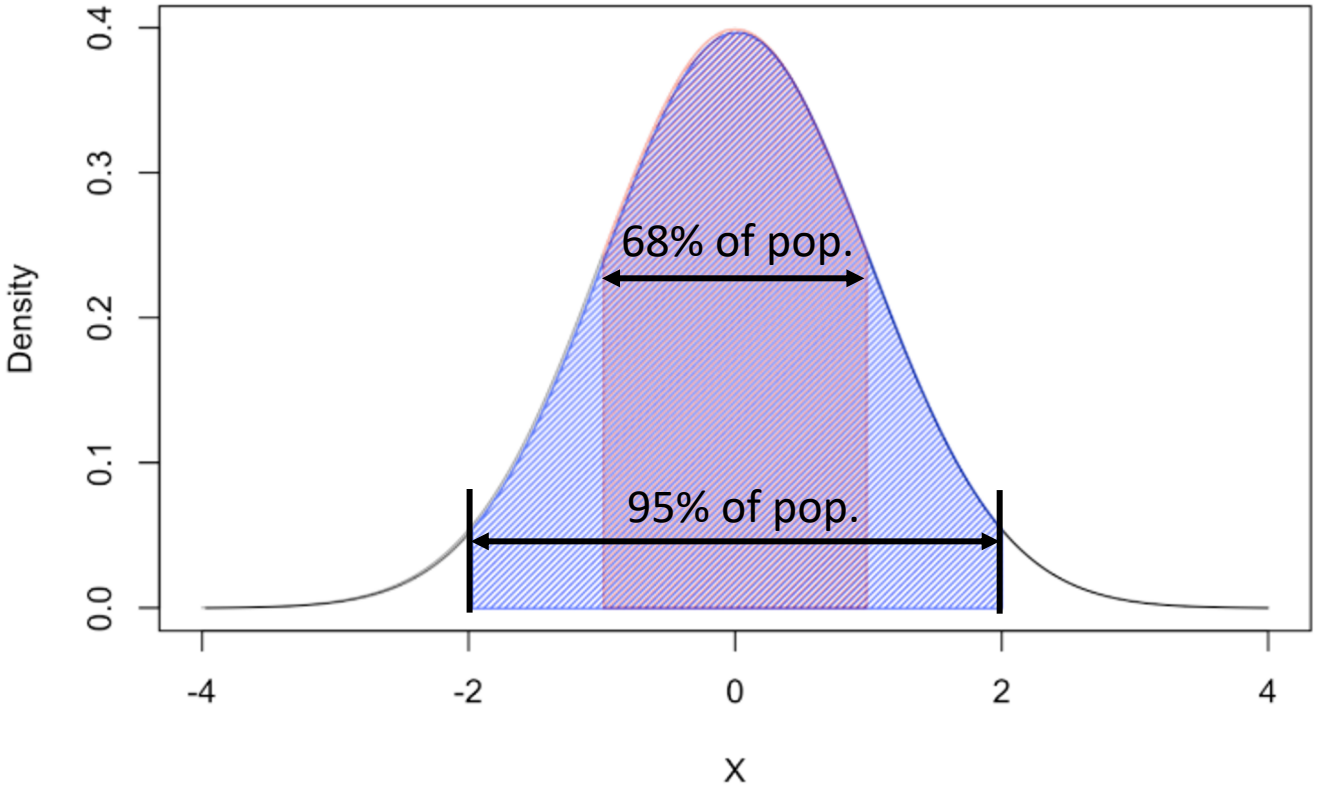


Why is p value < 0.05 significant?

The Empirical Rule:

When data are normally distributed, we can assume

- ~68% of the population lies within +/- 1 SD of the mean
- ~95% of the population lies within +/- 2 SD of the mean
- This property is the basis for all parametric statistical tests (t tests, ANOVA, linear regression, etc.)

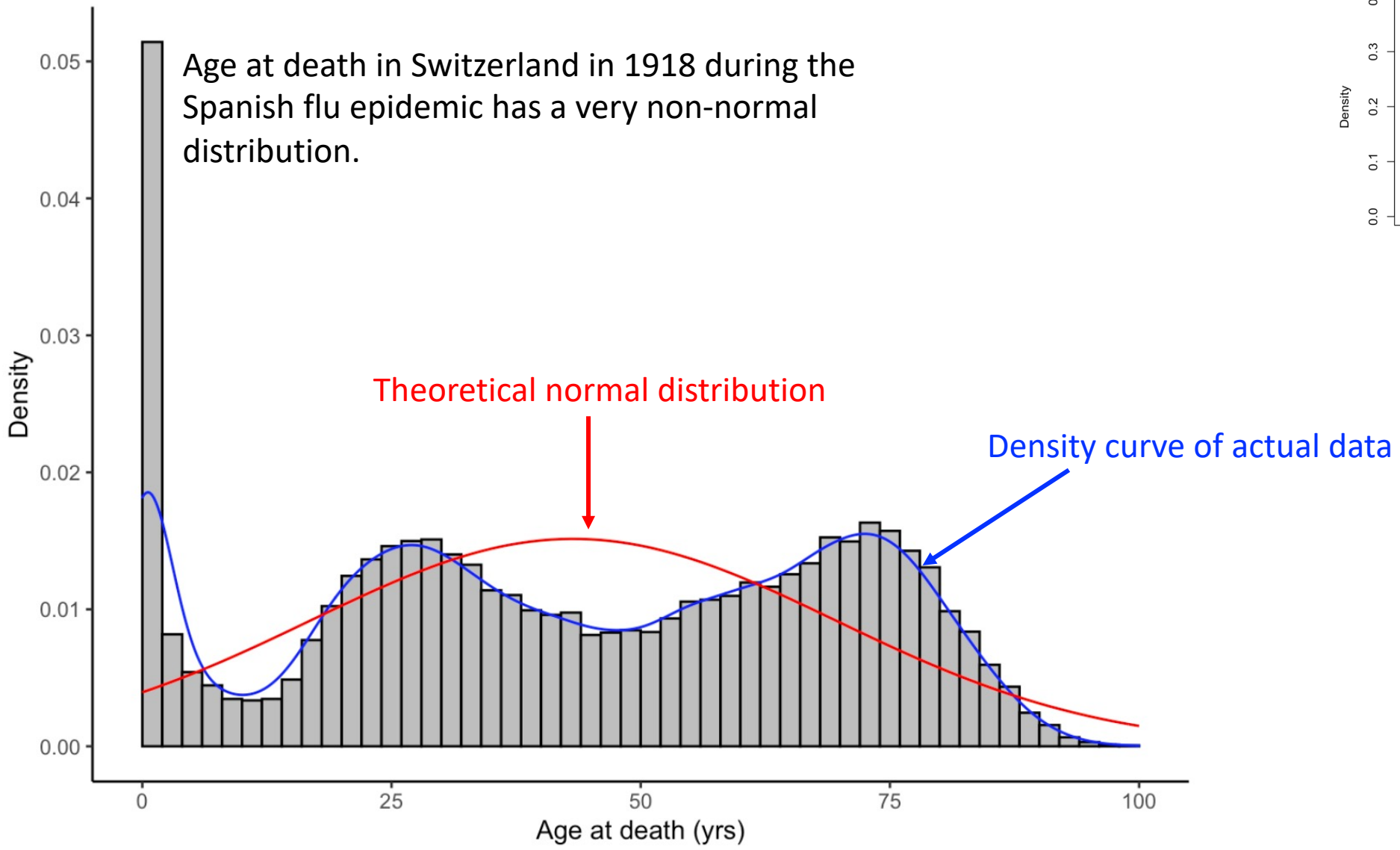


Why is p value < 0.05 significant?

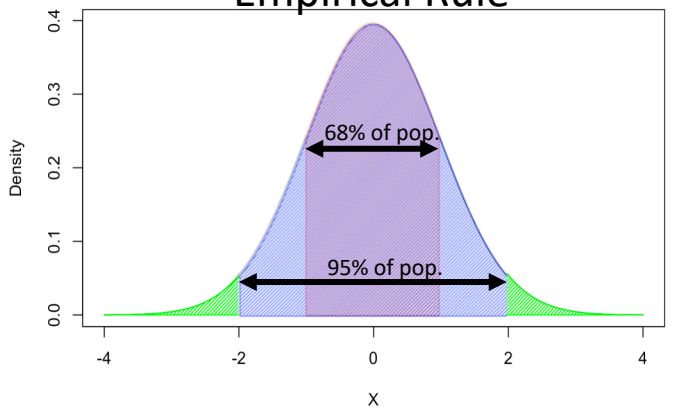
- Tied to Type I Error Rate alpha (α):
  - If α = 0.05 (5%), we accept a 5% chance that differences we detect will be incorrect
  - The p value is the observed probability that our conclusions are incorrect.
  - P = 0.01, there is an observed probability of 1% that conclusions are incorrect.

How do we know when a sample is “large enough”?

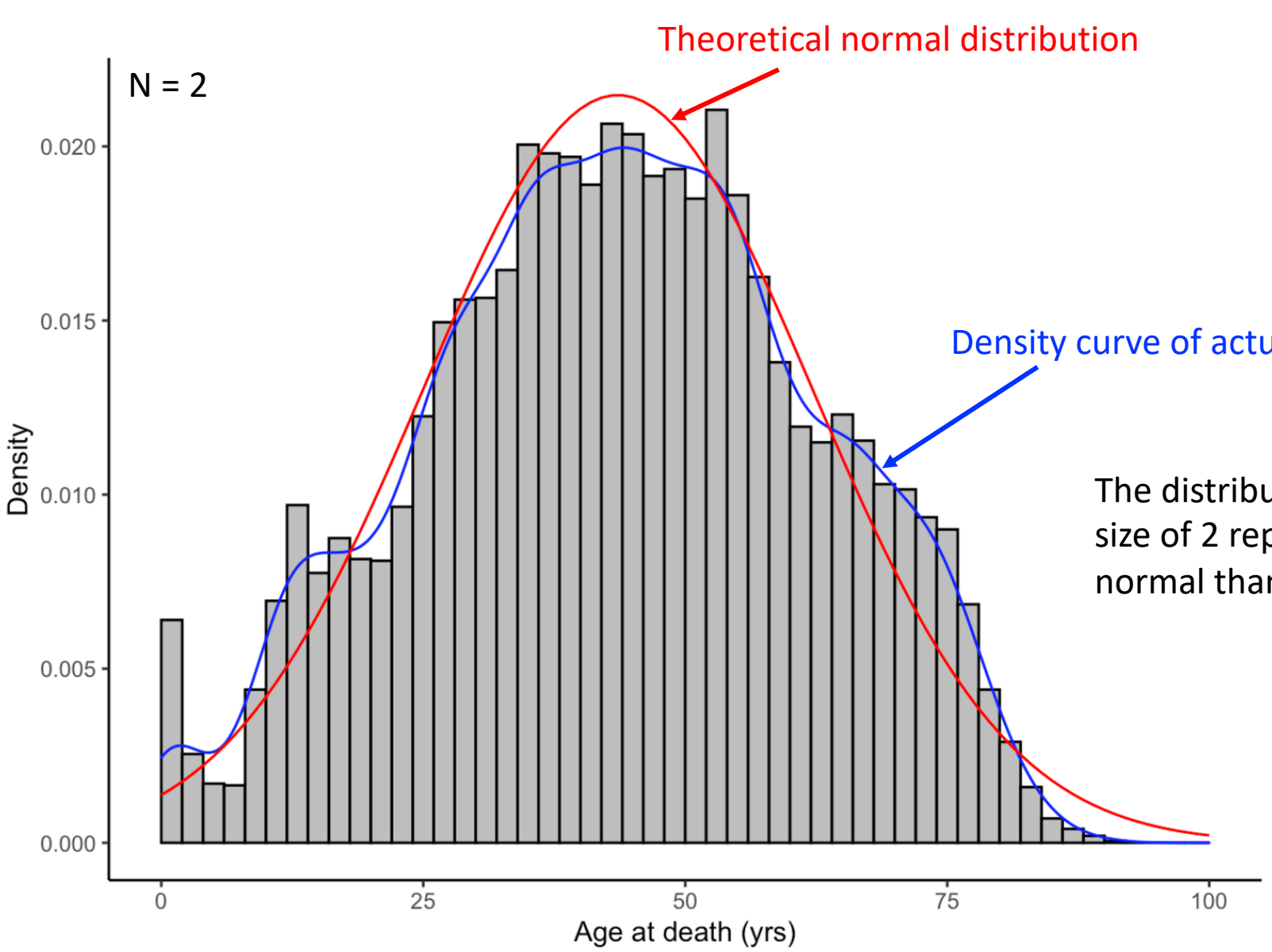
Age at death in Switzerland in 1918 during the Spanish flu epidemic has a very non-normal distribution.

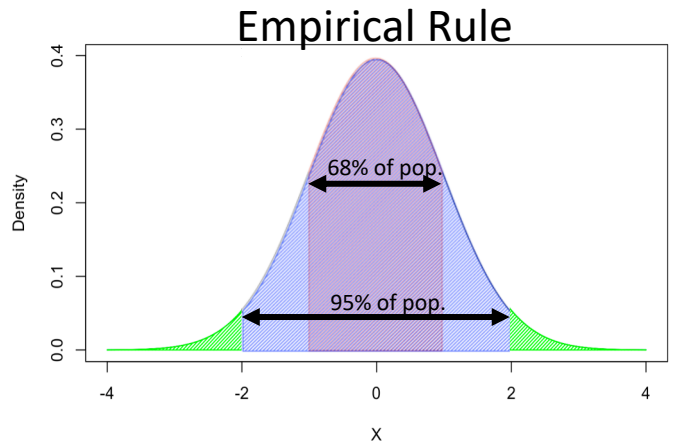


Empirical Rule



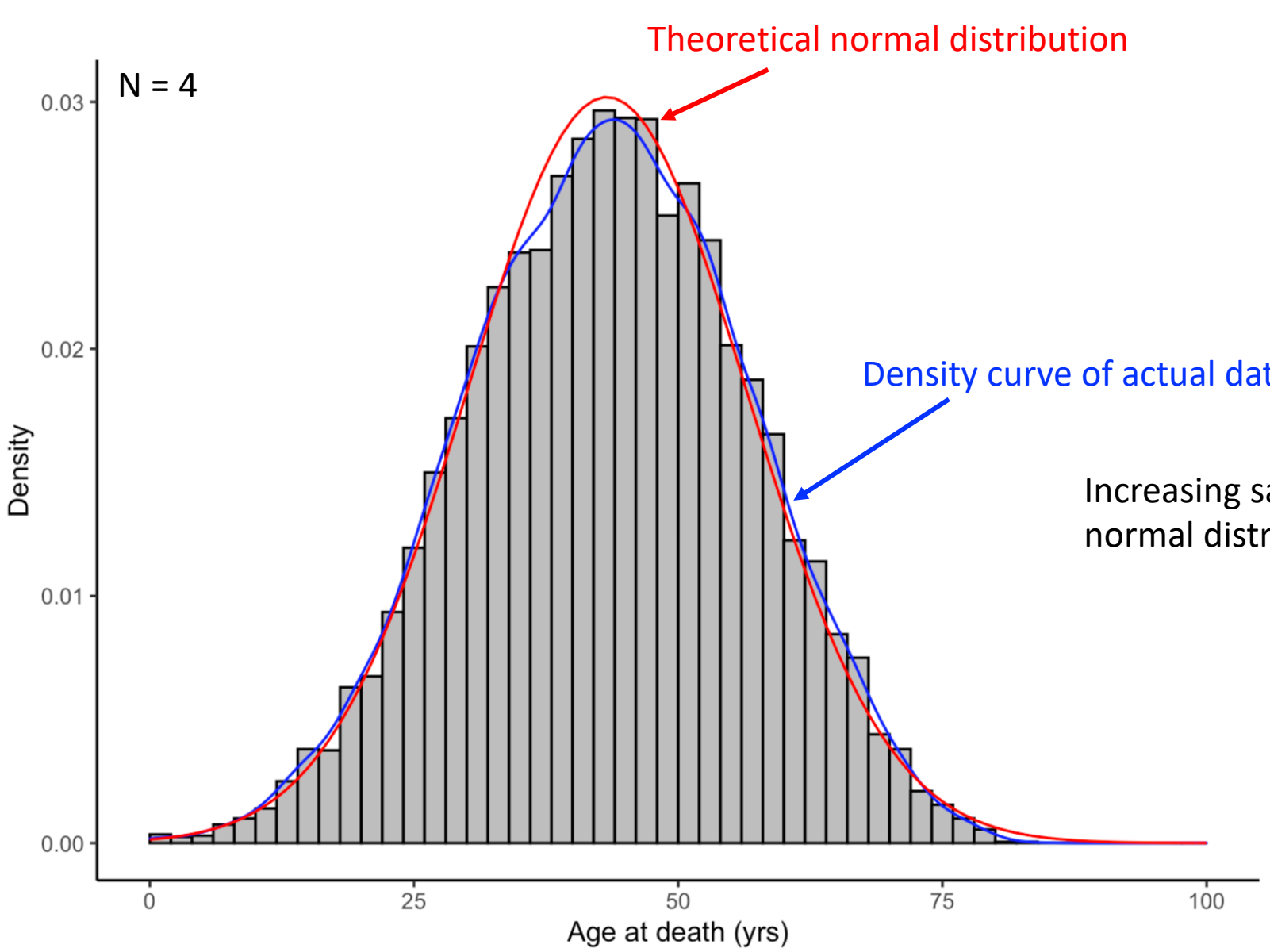
How do we know when a sample is “large enough”?

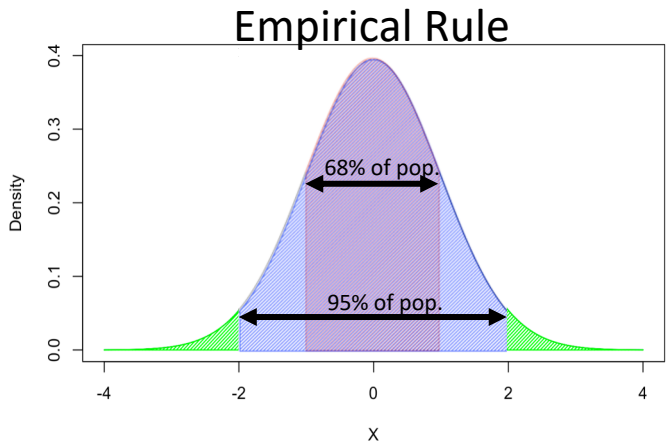




The distribution of means from a small sample size of 2 repeated 10,000 times is much closer to normal than the starting distribution.

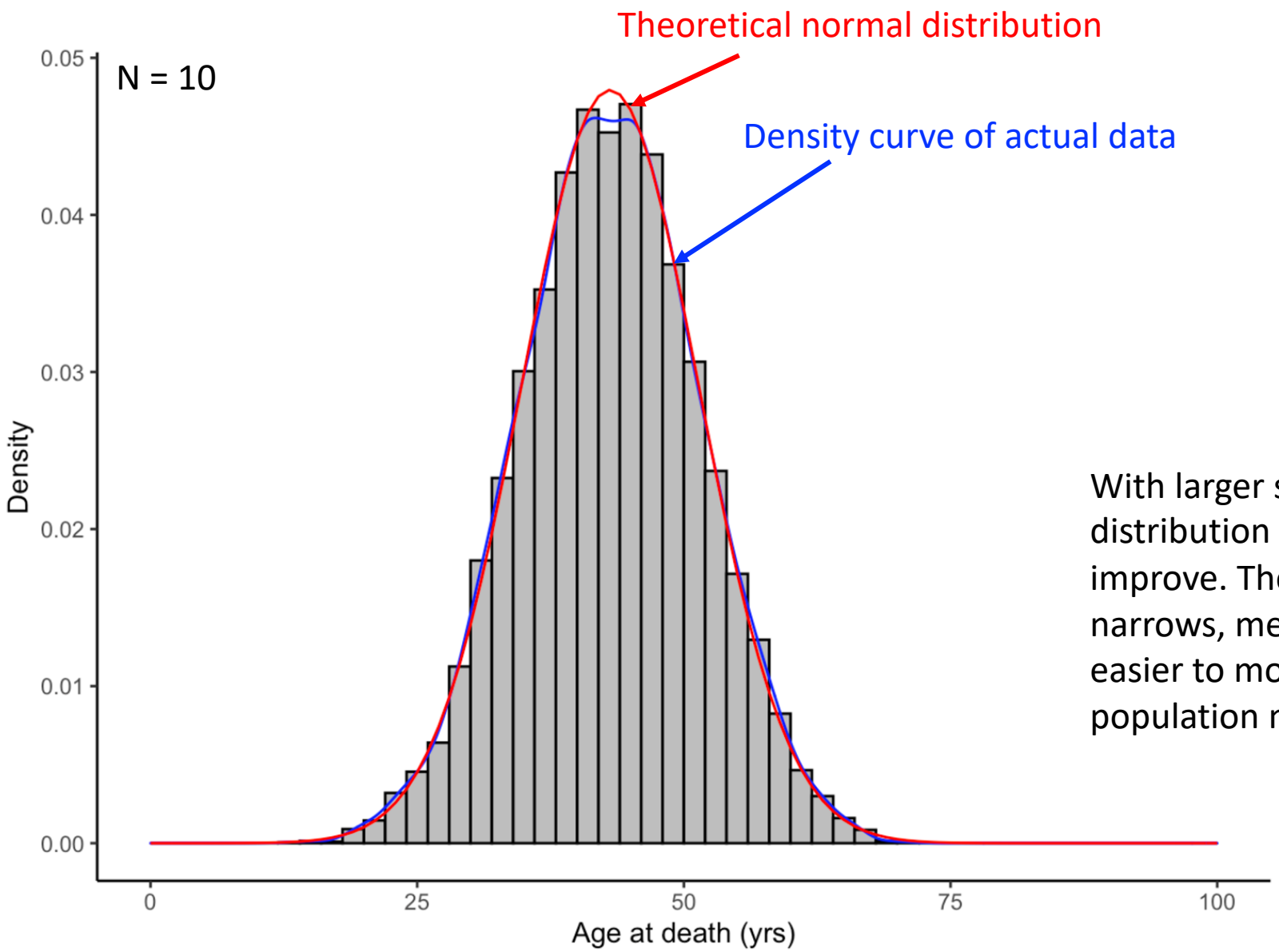
How do we know when a sample is “large enough”?

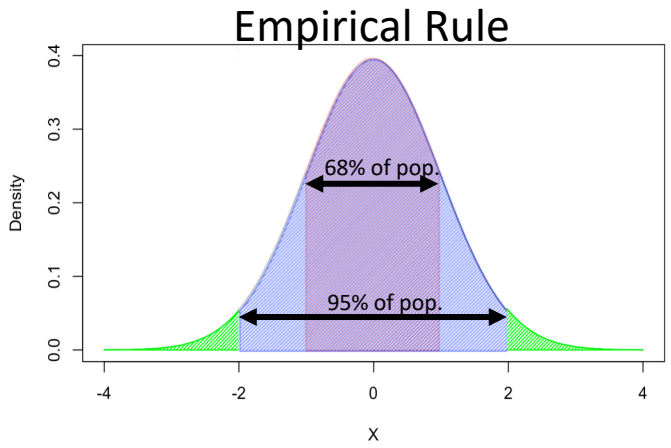




Increasing sample size to 4 approximates a normal distribution

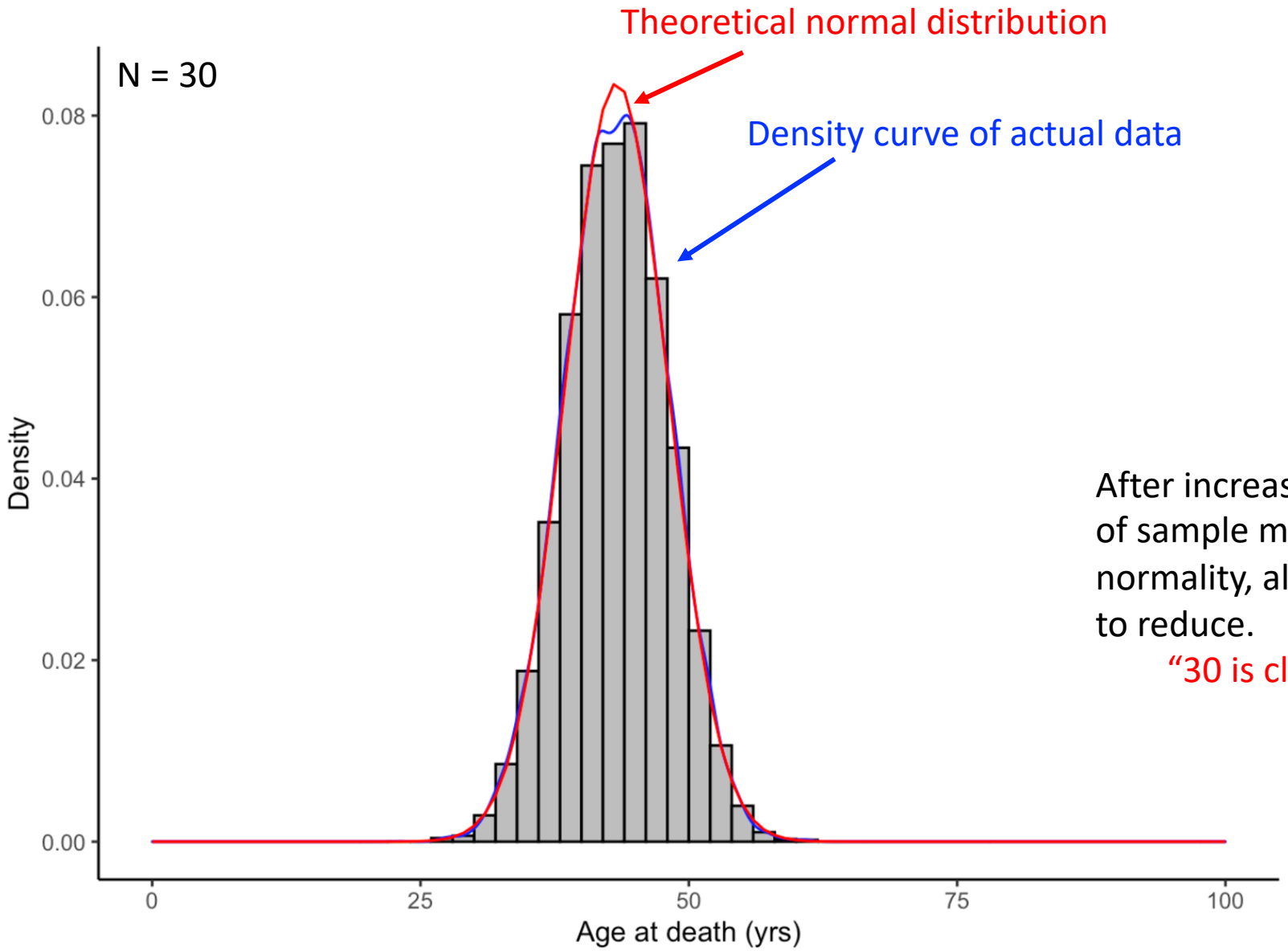
How do we know when a sample is “large enough”?

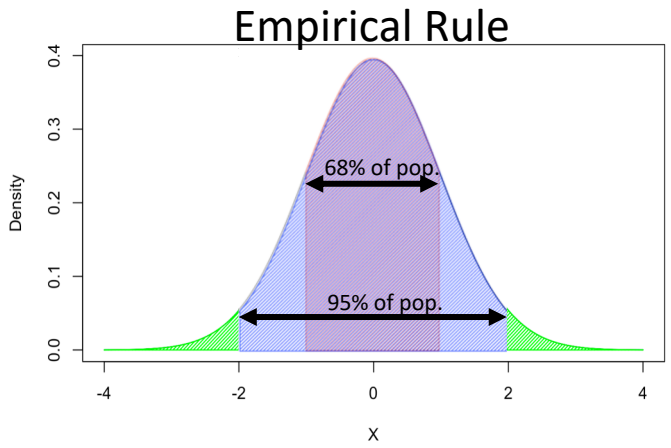




With larger sample sizes (n = 10), the normal distribution of the sample estimates continues to improve. The spread of the distribution also narrows, meaning larger sample sizes make it easier to more accurately estimate the true population mean.

How do we know when a sample is “large enough”?





After increasing sample size to 30, the distribution of sample means doesn’t improve much regarding normality, although the spread of data continues to reduce.

“30 is closer to infinity than 0 in statistics”

Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

What do we know so far?

- Need representative, random, and independent samples
- Need to *maximize* precision, repeatability, and accuracy and *minimize* bias
- The Empirical Rule and properties of the Normal Distribution form the basis of **ALL parametric statistical tests**
  - Our statistical tests are most powerful when data is normally distributed
- Increasing sample size in experiments can compensate for non-normal data with caveat that it is possible to reach a point of diminishing returns

Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

1_CentralLimitTheoremCode.R
Contains code for generating all figures up to this point (need R and Rstudio to run)

What do we know so far?

- Need representative, random, and independent samples
- Need to *maximize* precision, repeatability, and accuracy and *minimize* bias
- The Empirical Rule and properties of the Normal Distribution form the basis of **ALL parametric statistical tests**
  - Our statistical tests are most powerful when data is normally distributed
- Increasing sample size in experiments can compensate for non-normal data with caveat that it is possible to reach a point of diminishing returns

Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

Objectives:

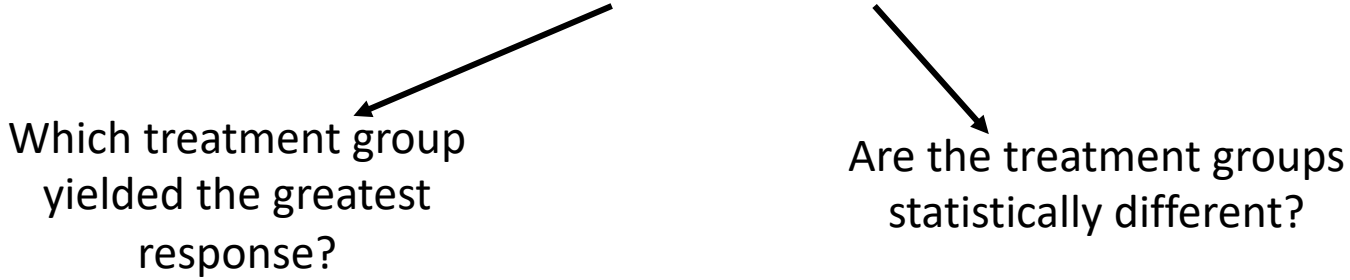
- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

Descriptive statistics: Provide estimates of the location and spread of the data distribution

Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

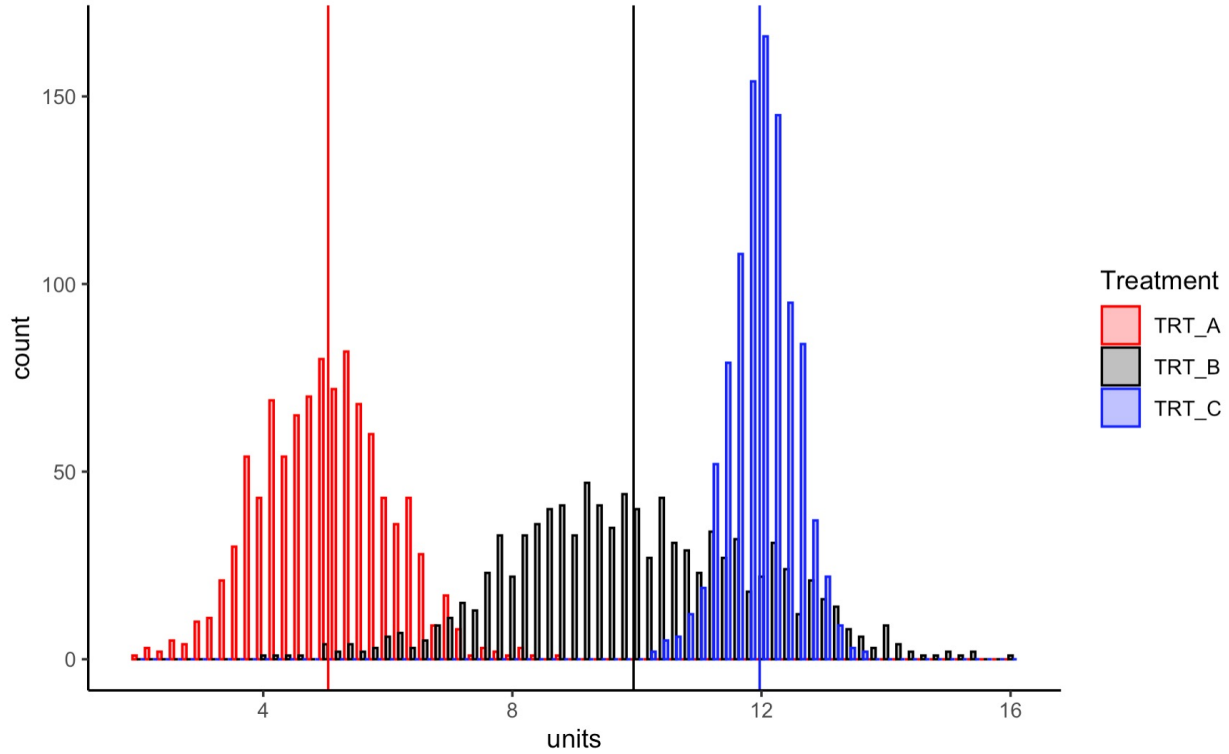
Descriptive statistics: Provide estimates of the location and spread of the data distribution



Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

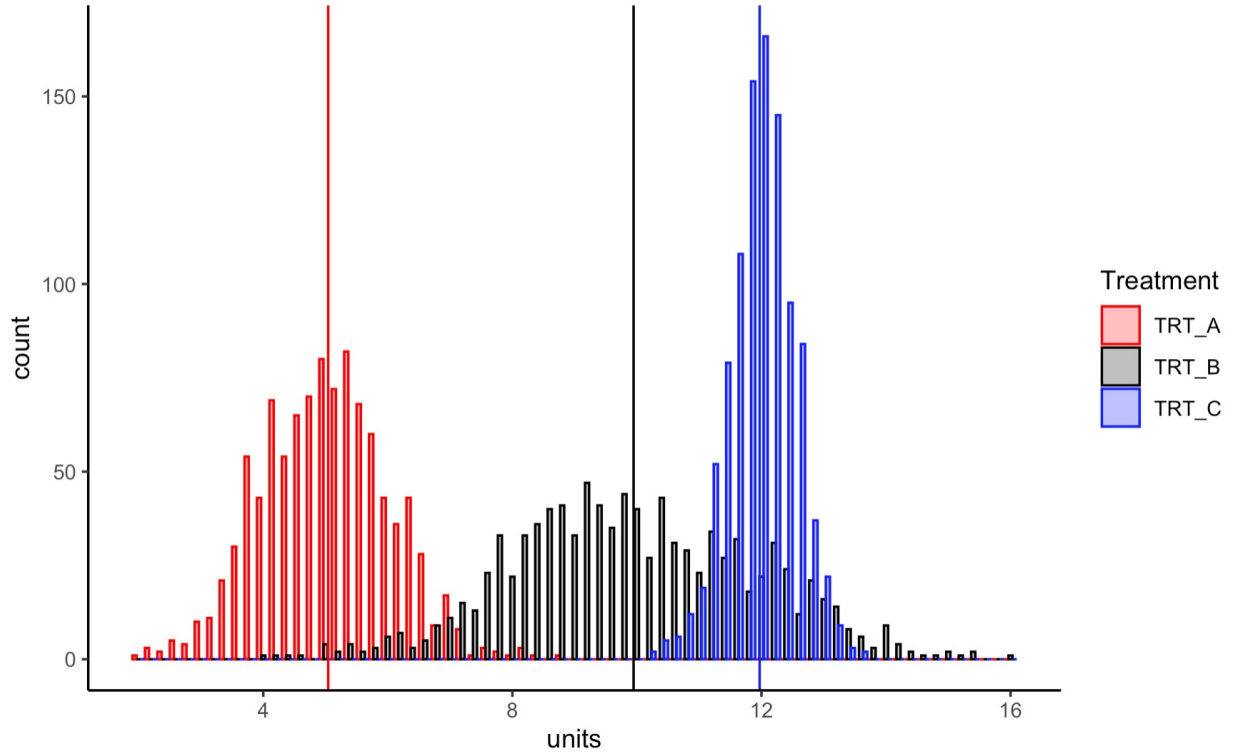
Descriptive statistics: Provide estimates of the location and spread of the data distribution



Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

Descriptive statistics: Provide estimates of the location and spread of the data distribution



Did the treatment influence the response variable?

- Are the means different?
  - The mean itself
  - Variation around the mean

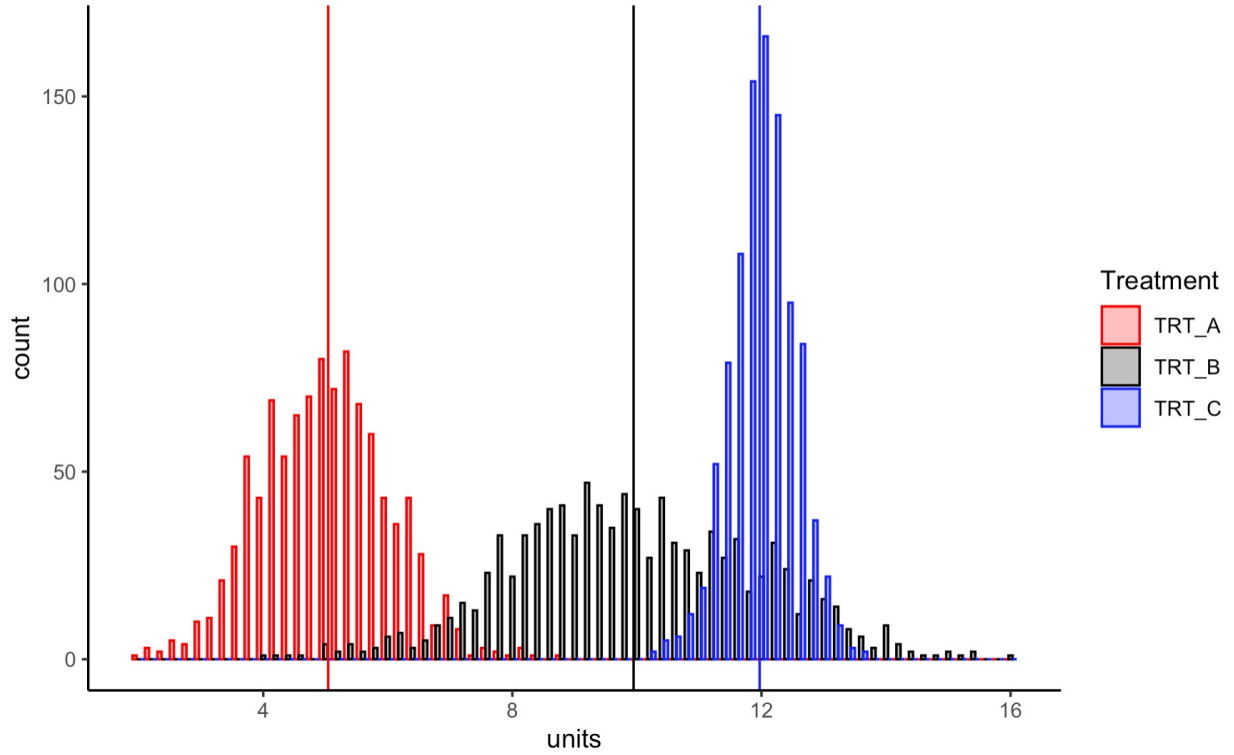
Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

Model Assumptions:

- Independent samples
- Data are normally distributed
- Variation in each sample is appox. equal
- No extreme outliers

Descriptive statistics: Provide estimates of the location and spread of the data distribution



Did the treatment influence the response variable?

- Are the means different?
  - The mean itself
  - Variation around the mean

Functions for calculating descriptive statistics in R:

- Maximum: `max()`
- Minimum: `min()`
- Sample size: `n()`
- Mean: `mean()`
- Median: `median()`
- Standard deviation: `sd()`
- Variance: `var()`
- Quantiles: `quantile()`

2_DescriptiveStatisticsCode.R
Contains code for describing and plotting data in following examples (need R and Rstudio to run)

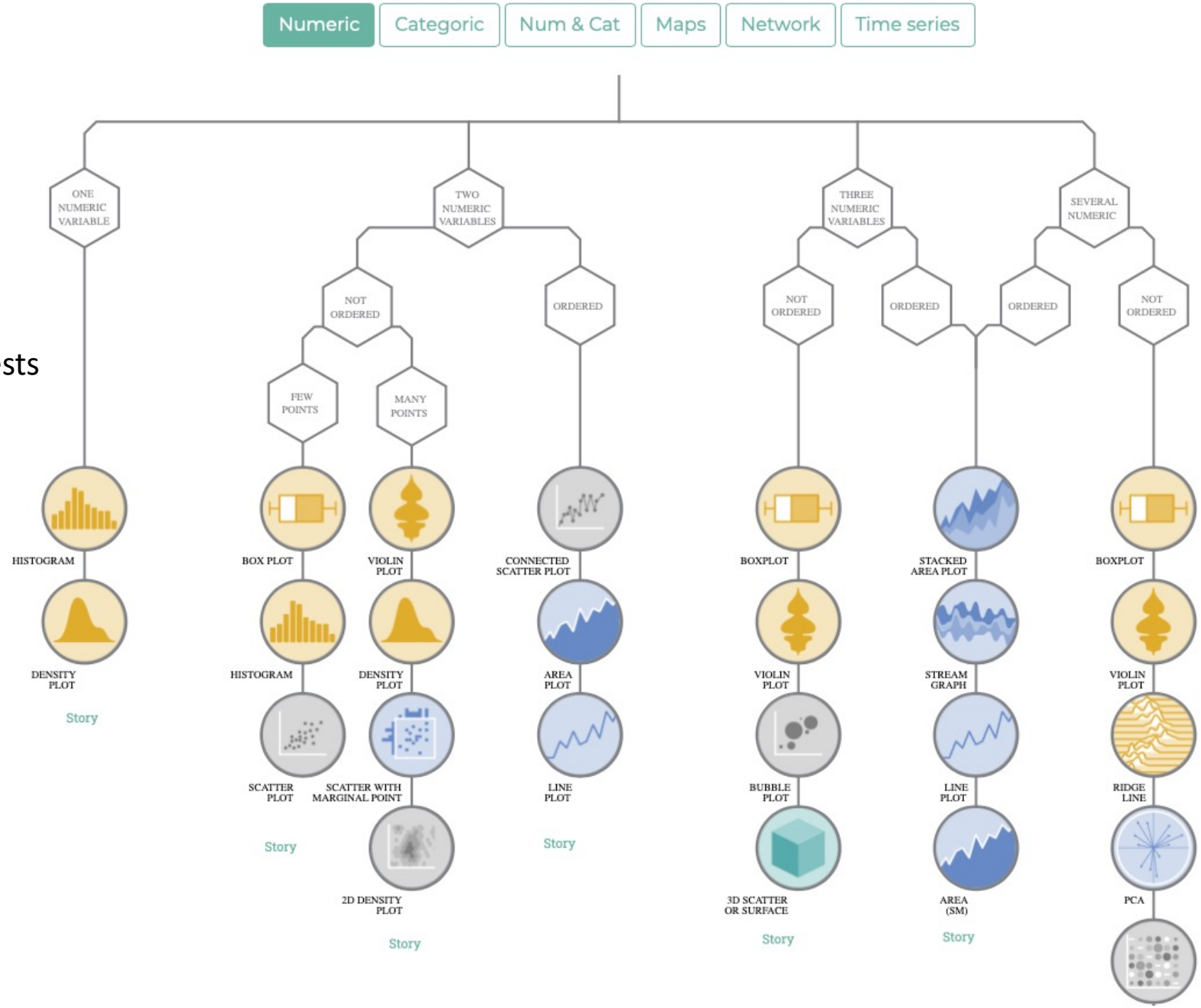
Functions for checking model assumptions in R:

- Normality: `shapiro.test()`
- Homogeneity of variance: `leveneTest()`

https://www.data-to-viz.com

Plot Raw Data:

- Check data entry
- Check data meet assumptions for tests
  - Variance
  - Normality
  - Outliers
- Check for unusual trends



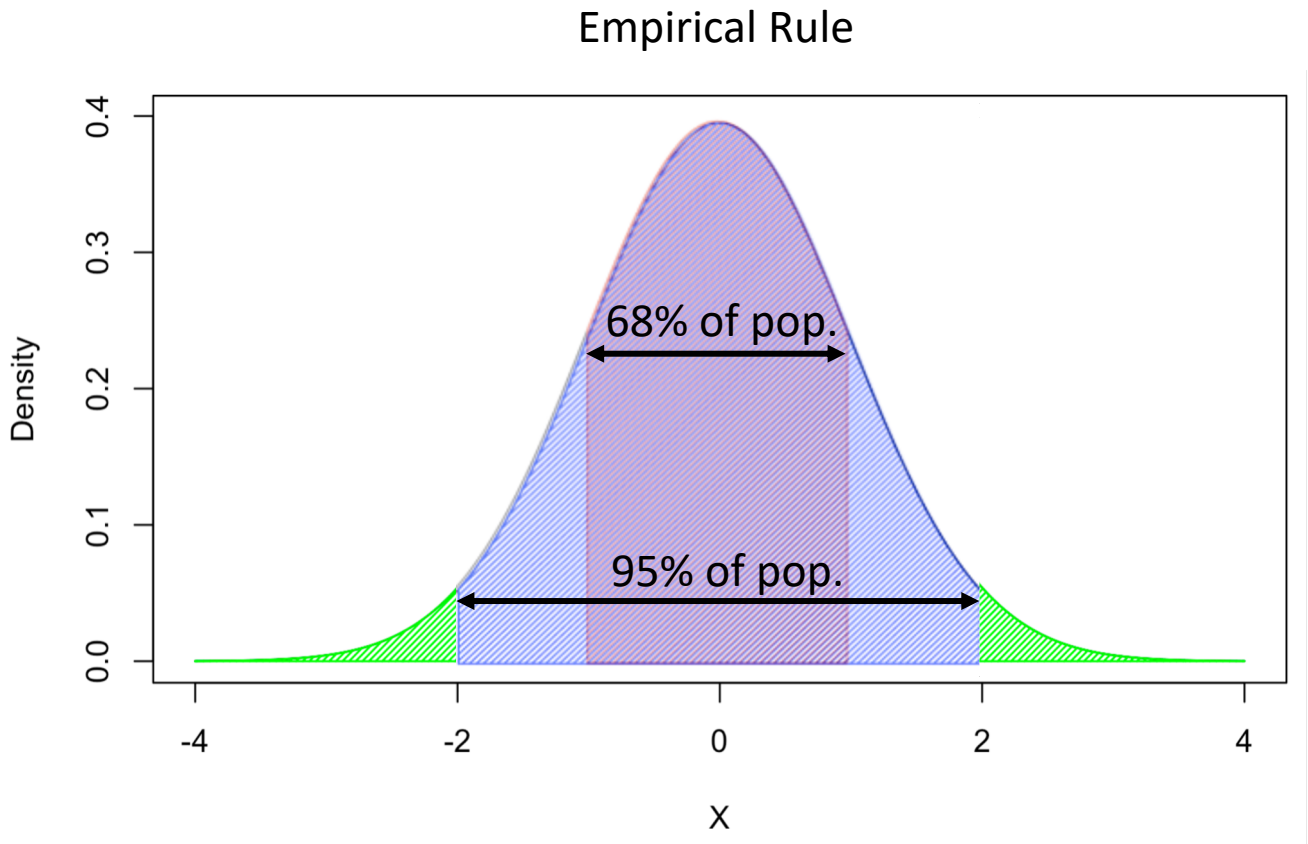
Dataset: Mutant versus Wildtype cells

- Hypothesis—the two cell types are different.

Model Assumptions:

- Independent samples
- Data are normally distributed
- Equal variance
- No extreme outliers

Are data normally distributed?



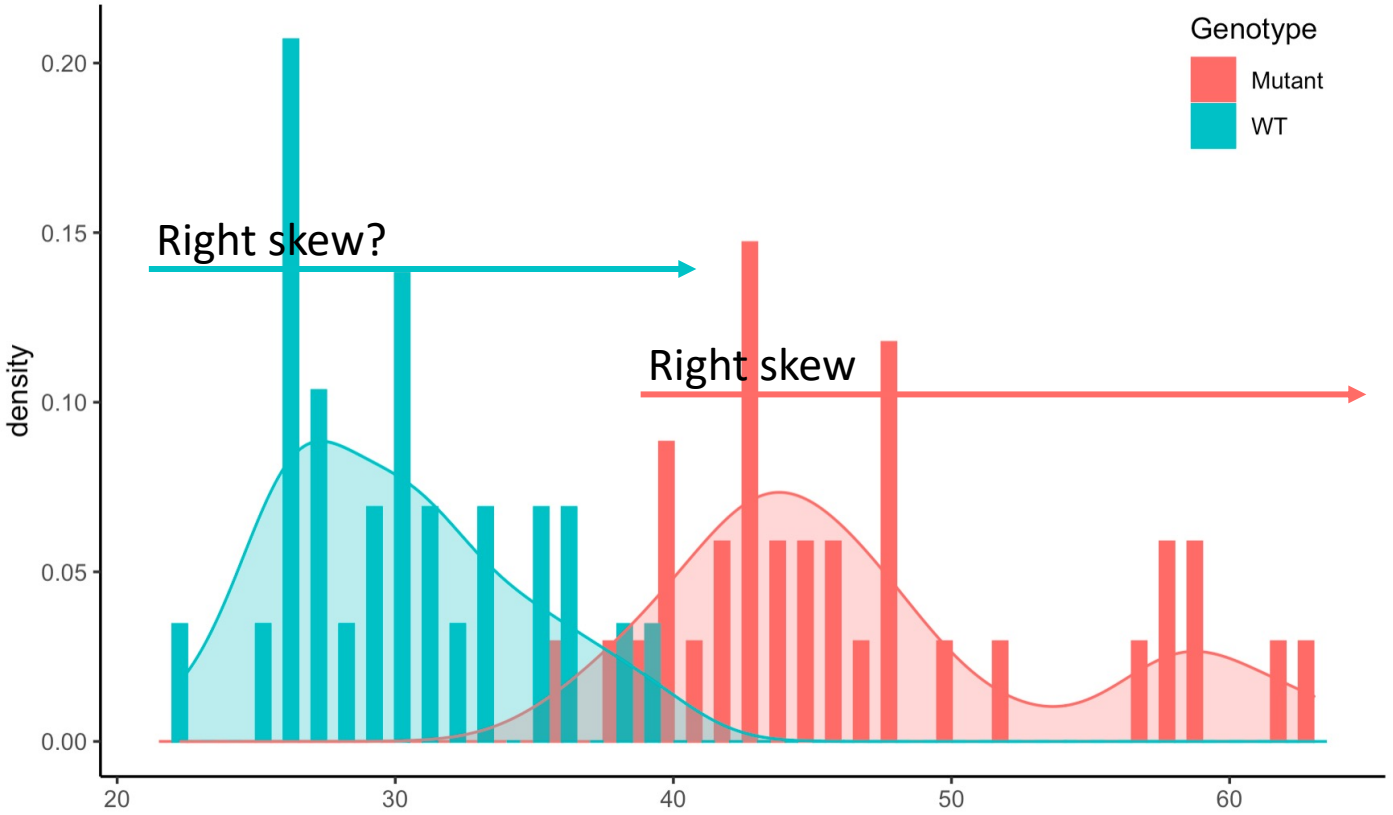
Dataset: Mutant versus Wildtype cells

- Hypothesis—the two cell types are different.

Model Assumptions:

- Independent samples
- Data are normally distributed
- Equal variance
- No extreme outliers

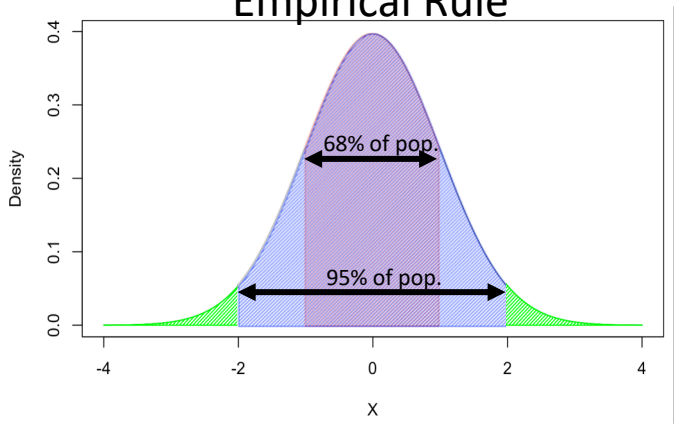
Are data normally distributed?



Shapiro Wilk's Test for normality

- WT: W = 0.95, p = 0.23 (Normal)
- Mutant: W = 0.90, p = 0.004 (Not Normal)

Empirical Rule



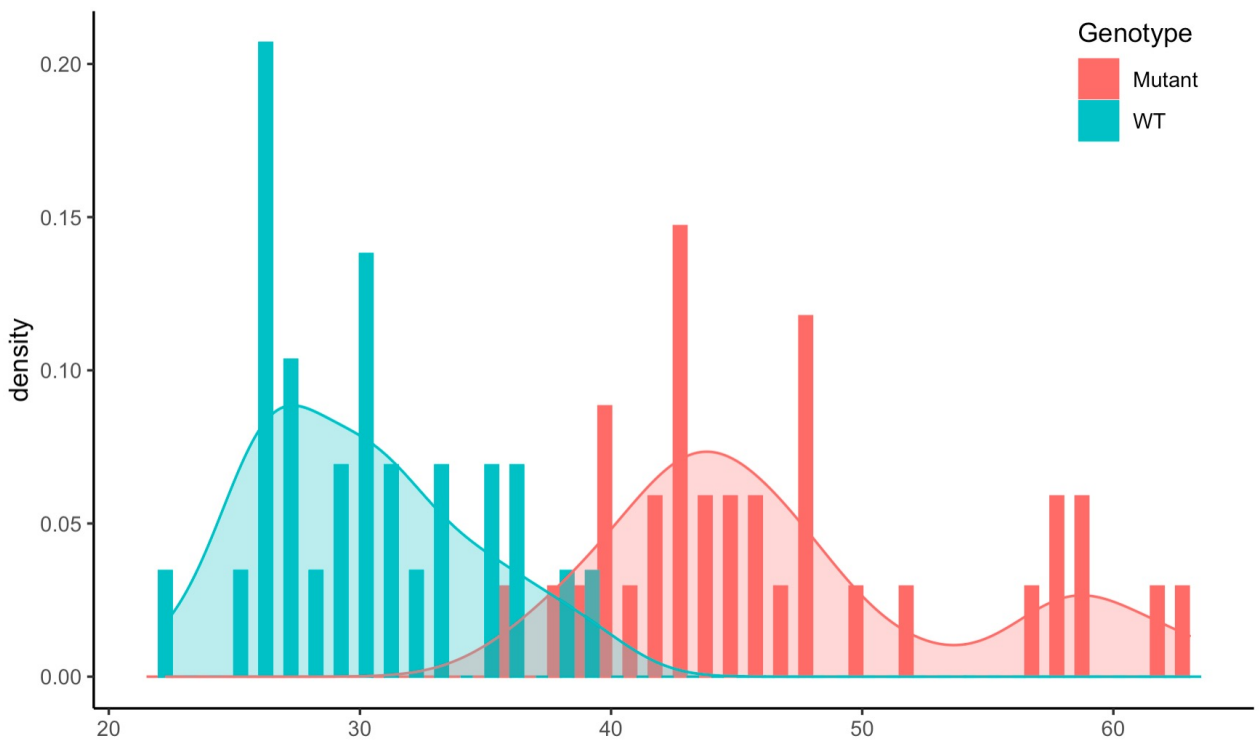
Model Assumptions:
- Independent samples
- Data are normally distributed
- Equal variance
- No extreme outliers

Key non-parametric tests

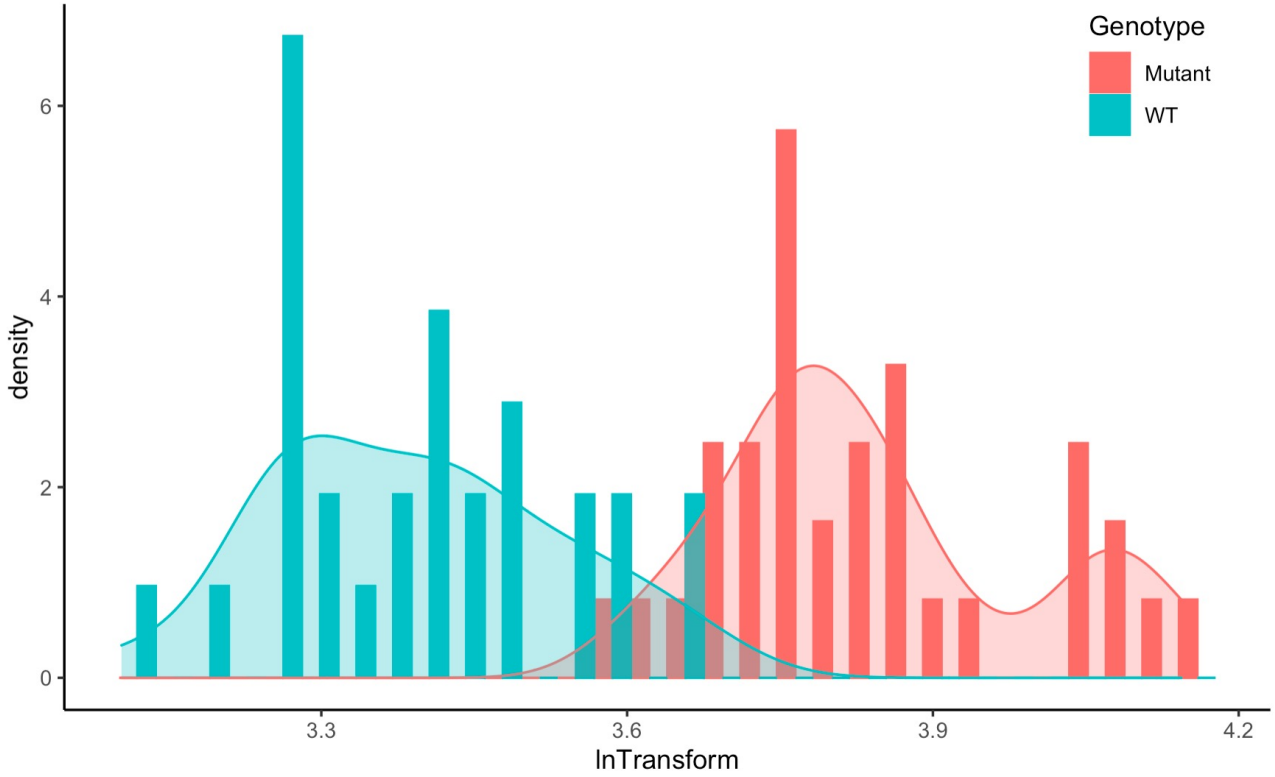
Parametric test	What to check for normality	Non-parametric test
Independent t-test	Dependent variable	Mann-Whitney test
Paired t-test	Paired differences	Wilcoxon signed rank test
One-way ANOVA	Residuals/dependent variable	Kruskal-Wallis test
Repeated measures ANOVA	Residuals at each time point	Friedman test
Pearson's correlation coefficient	Both variables should be normally distributed	Spearman's correlation coefficient
Simple linear regression	Residuals	N/A

Data transformation *can help* meet parametric model assumptions

Shapiro Wilk's Test for normality
- WT: W = 0.95, p = 0.23 (Normal)
- Mutant: W = 0.90, p = 0.004 (Not Normal)



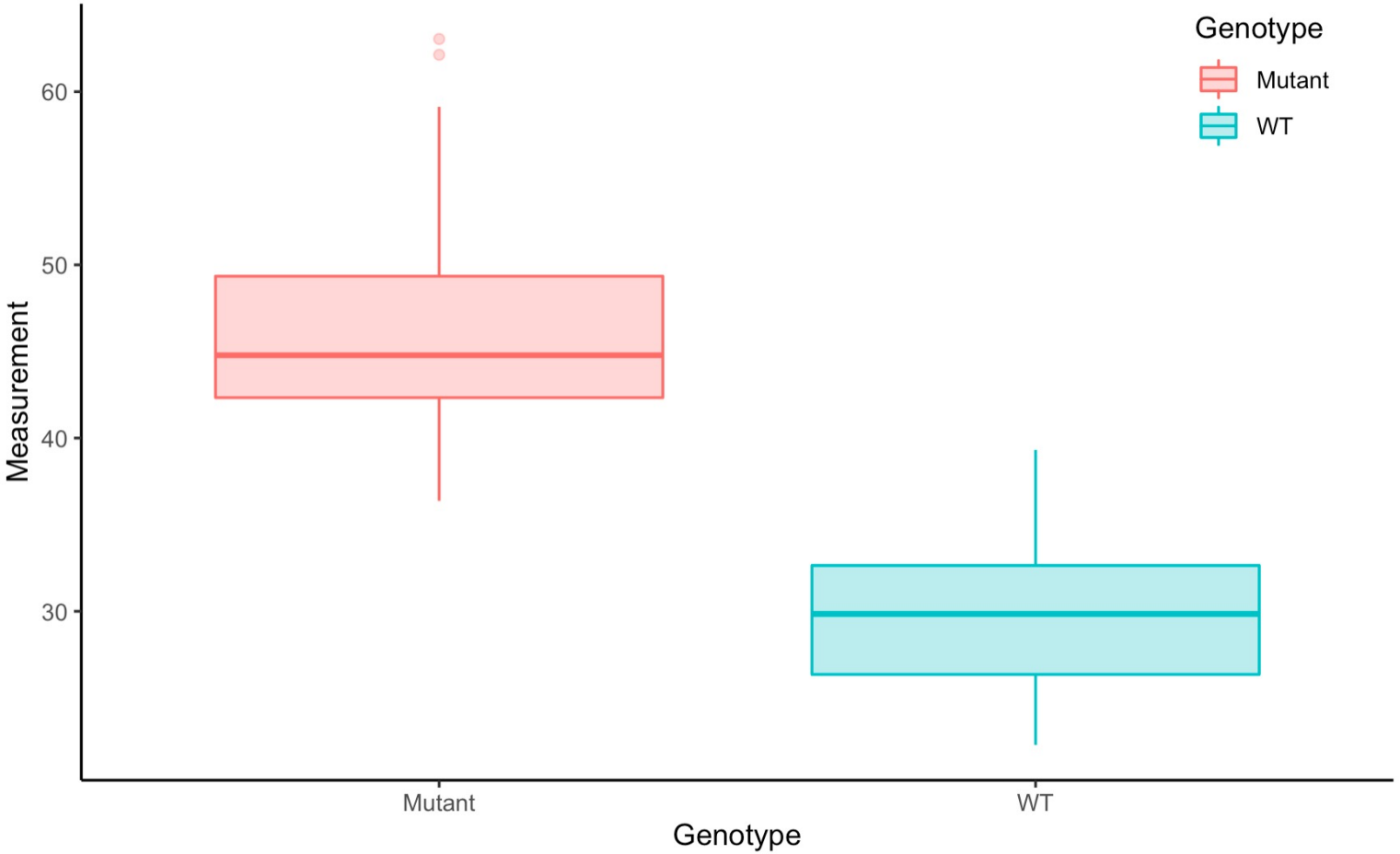
Shapiro Wilk's Test for normality
- WT: W = 0.97, p = 0.5 (Normal, improved)
- Mutant: W = 0.92, p = 0.03 (*Less* Not Normal)



Dataset: Mutant versus Wildtype cells

- Hypothesis—the two cell types are different.

Is variation in each sample approximately equal?



Model Assumptions:

- Independent samples
- Data are normally distributed—NO!
- Equal variance—Not really…
- No extreme outliers — possibly?

Genotype	N	Mean	SD	SE	Var
WT	29	29.98	4.25	0.79	18.1
Mutant	34	47.06	7.30	1.25	53.4

Mutant measurements are 3x more variable

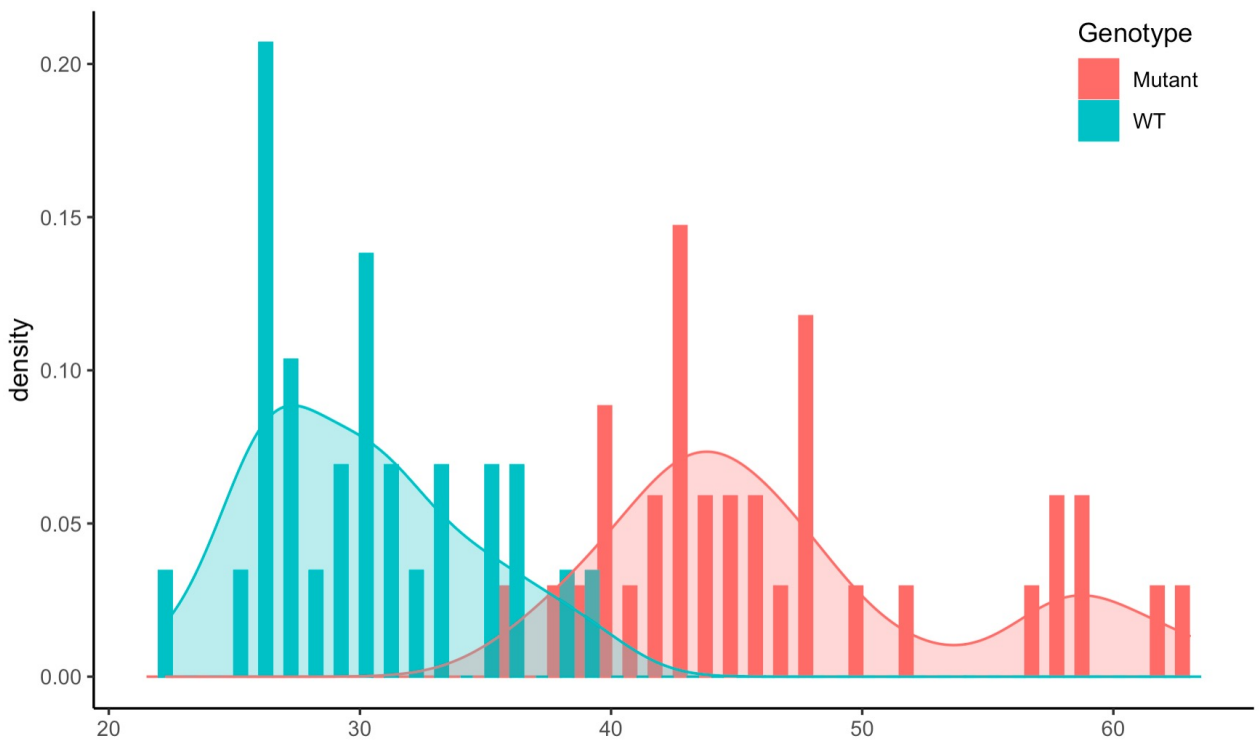
Levene's Test for homogeneity of variance

- F = 3.7, P = 0.059

Data transformation *can help* meet parametric model assumptions

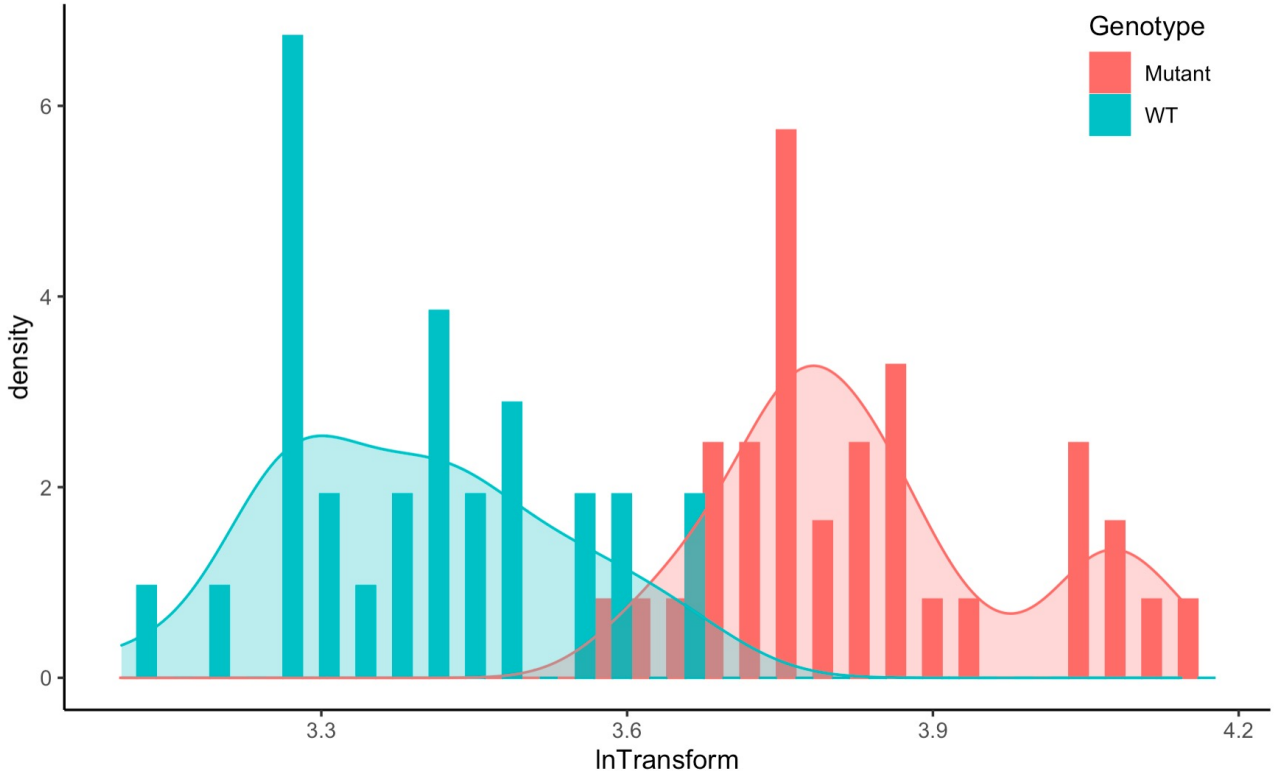
Shapiro Wilk's Test for normality
- WT: W = 0.95, p = 0.23 (Normal)
- Mutant: W = 0.90, p = 0.004 (Not Normal)

Levene's Test for equal variance
- F = 3.7, p = 0.059



Shapiro Wilk's Test for normality
- WT: W = 0.97, p = 0.5 (Normal, improved)
- Mutant: W = 0.92, p = 0.03 (*Less* Not Normal)

Levene's Test for equal variance
- F = 0.001, p = 0.98

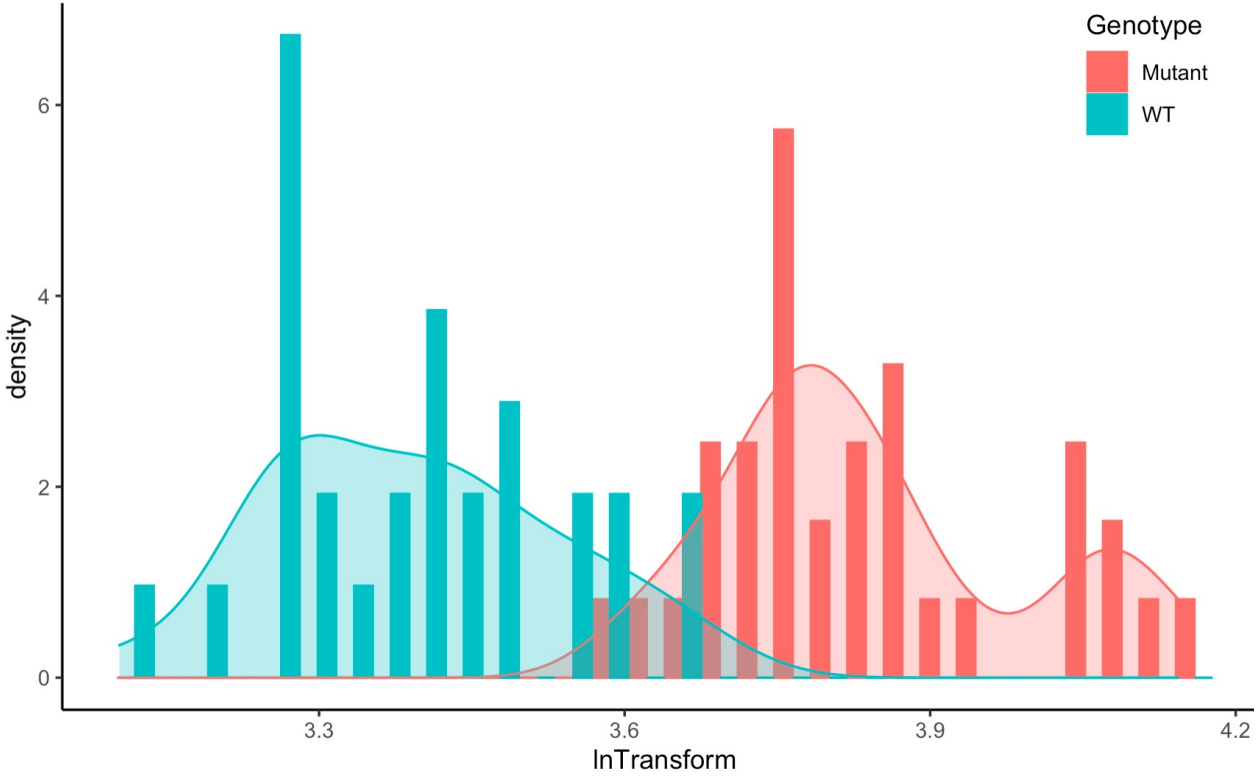


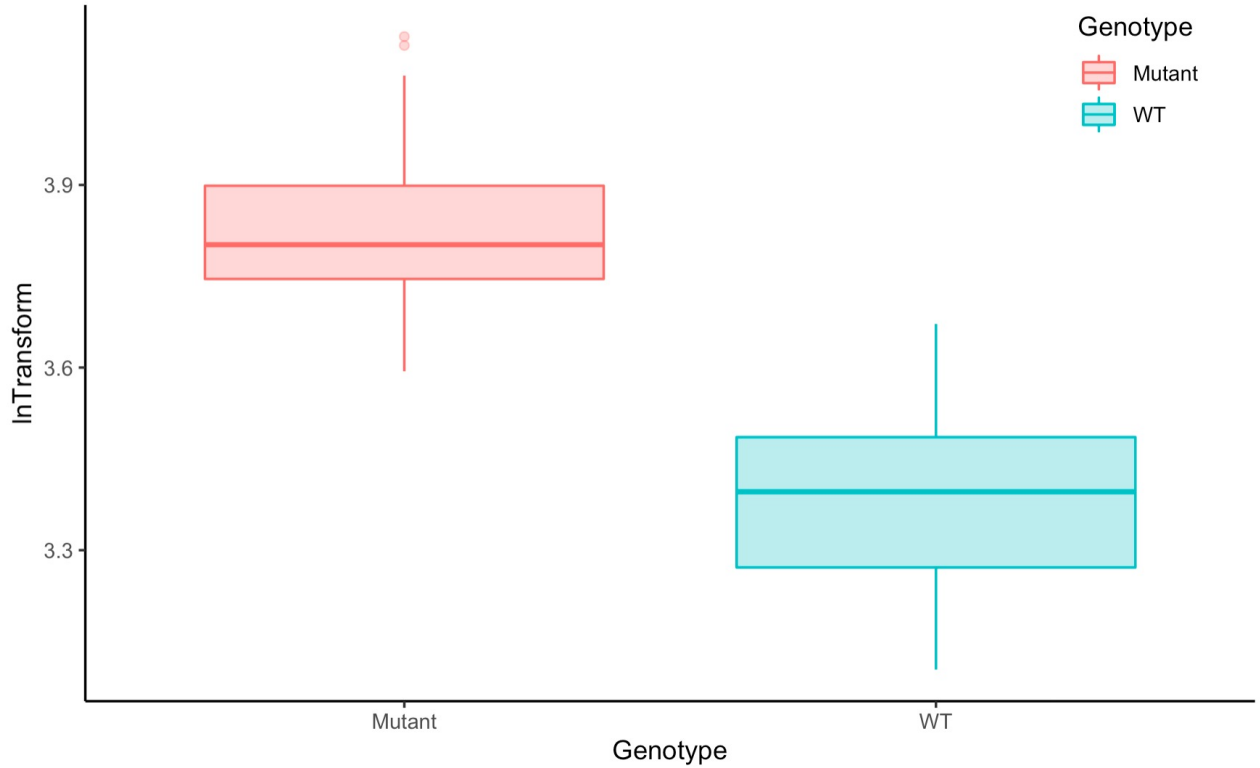
Dataset: Mutant versus Wildtype cells

- Hypothesis—the two cell types are different.

Did the treatment influence the response variable?

- Are the means different?
  - The mean itself
  - Variation around the mean

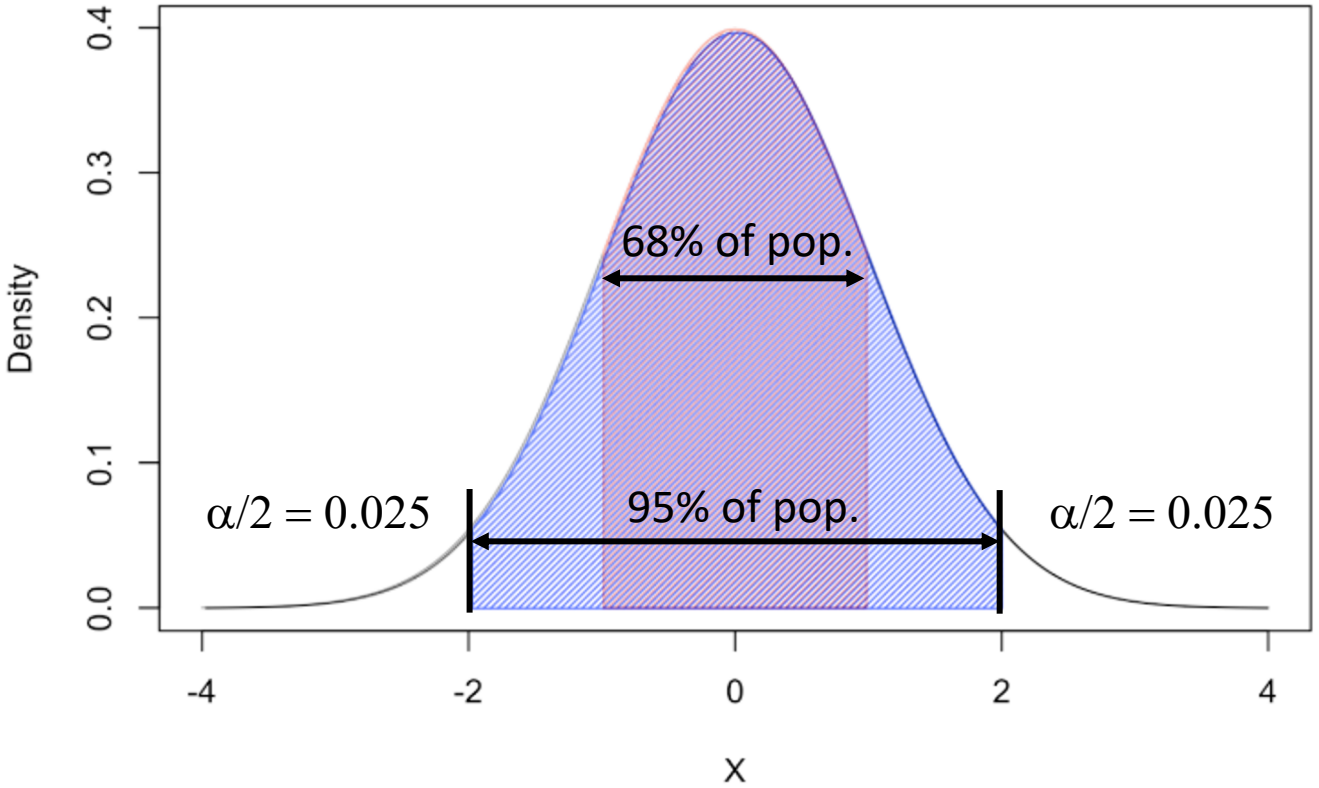




The Empirical Rule:

When data are normally distributed, we can assume

- ~68% of the population lies within +/- 1 SD of the mean
- ~95% of the population lies within +/- 2 SD of the mean
- This property is the basis for all parametric statistical tests (t tests, ANOVA, linear regression, etc.)



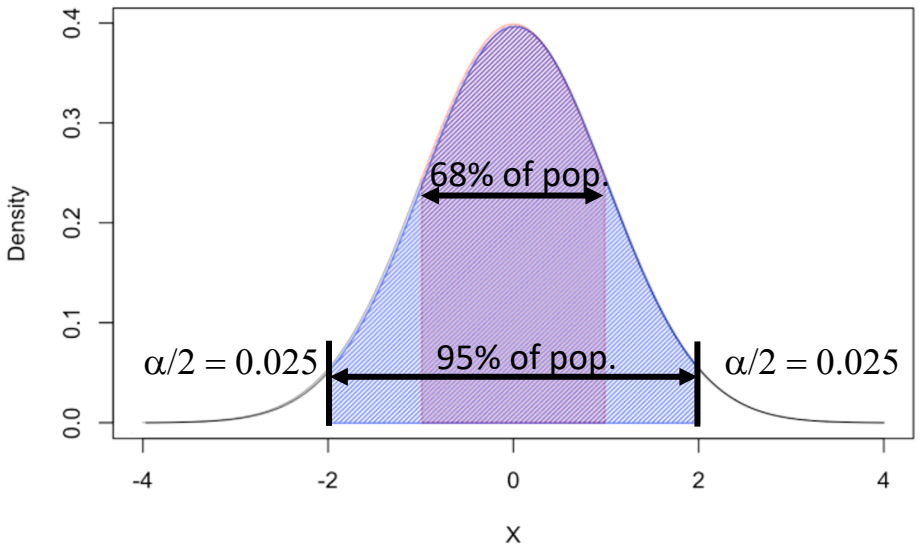
Why is p value < 0.05 significant?

- Tied to Type I Error Rate alpha (α):
  - If α = 0.05 (5%), we accept a 5% chance that differences we detect will be incorrect
  - The p value is the observed probability that our conclusions are incorrect.
  - P = 0.01, there is an observed probability of 1% that conclusions are incorrect.

Calculate by hand:

# WT:
WT.n = 29
WT.sd = 0.14
WT.se = WT.sd / sqrt(WT.n) = 0.026
WT.df = WT.n – 1 = 28

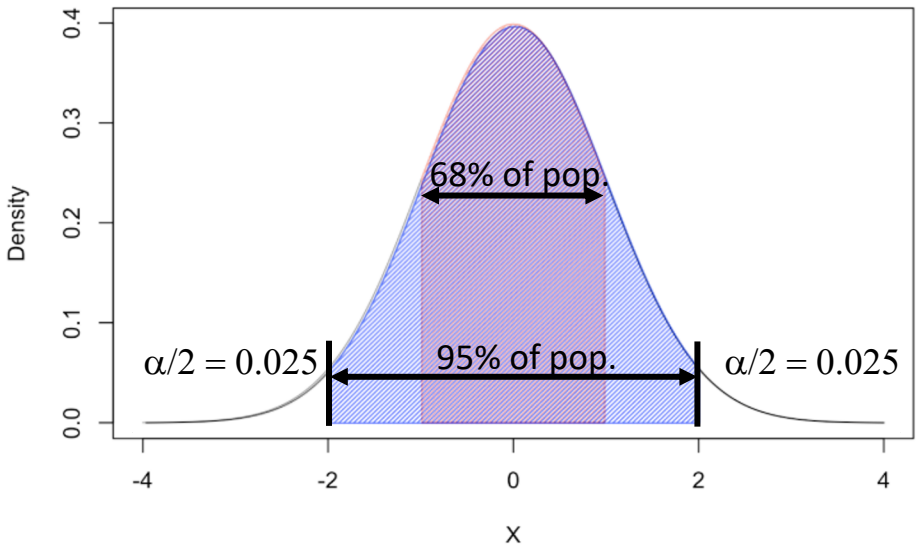
# Mutant:
mutant.n = 34
mutant.sd = 0.15
mutant.se = mutant.sd / sqrt(mutant.n) = 0.026
mutant.df = mutant.n - 1 = 33



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

Calculate by hand:

# WT:
WT.n = 29
WT.sd = 0.14
WT.se = WT.sd / sqrt(WT.n) = 0.026
WT.df = WT.n – 1 = 28
CI value = 2.04841 * 0.026 = 0.053
95% CI = mean +/- CI value



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752		31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986		6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363		4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847		3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048		3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180		3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579		2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548		2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113		2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461		2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885		2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288		2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933		2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310		2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050		2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884		2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607		2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064		2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133		2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718		2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743		2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144		2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872		2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882		2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141		2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618		2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288		2.47266	2.77068	3.6896
28	2.04841							
29	0.255684	0.683044	1.311434	1.699127		2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261		2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854		2.32635	2.57583	3.2905
CI	———	———	80%	90%		98%	99%	99.9%

Calculate by hand:

# WT:
WT.n = 29
WT.sd = 0.14
WT.se = WT.sd / sqrt(WT.n) = 0.026
WT.df = WT.n – 1 = 28
CI value = 2.04841 * 0.026 = 0.053
95% CI = mean +/- CI value

# Mutant:
mutant.n = 34
mutant.sd = 0.15
mutant.se = mutant.sd / sqrt(mutant.n) = 0.026
mutant.df <- mutant.n - 1 = 33
CI value = 2.0345 * 0.026 = 0.052
95% CI = mean +/- CI value

Using R to calculate statistics eliminates needing to look up a table and provides values that may not be listed – See 2_DescriptiveStatistics.Rmd for code.

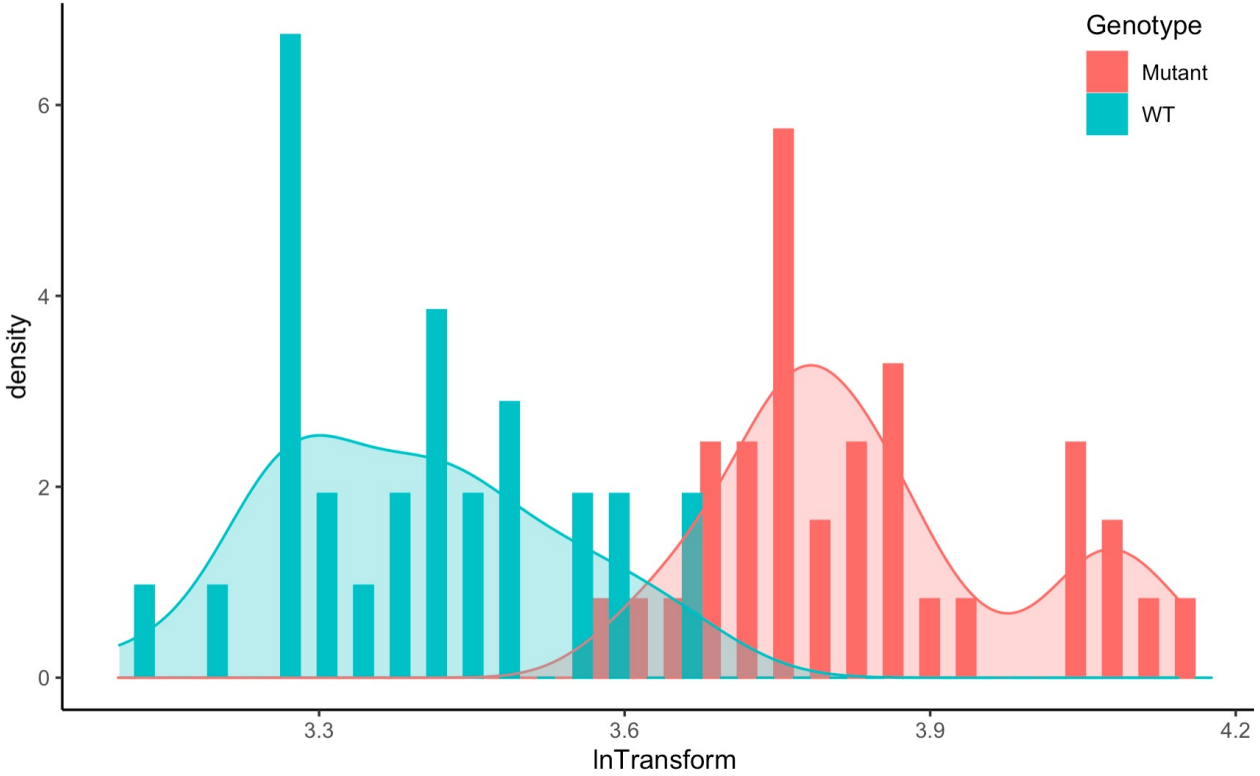
df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

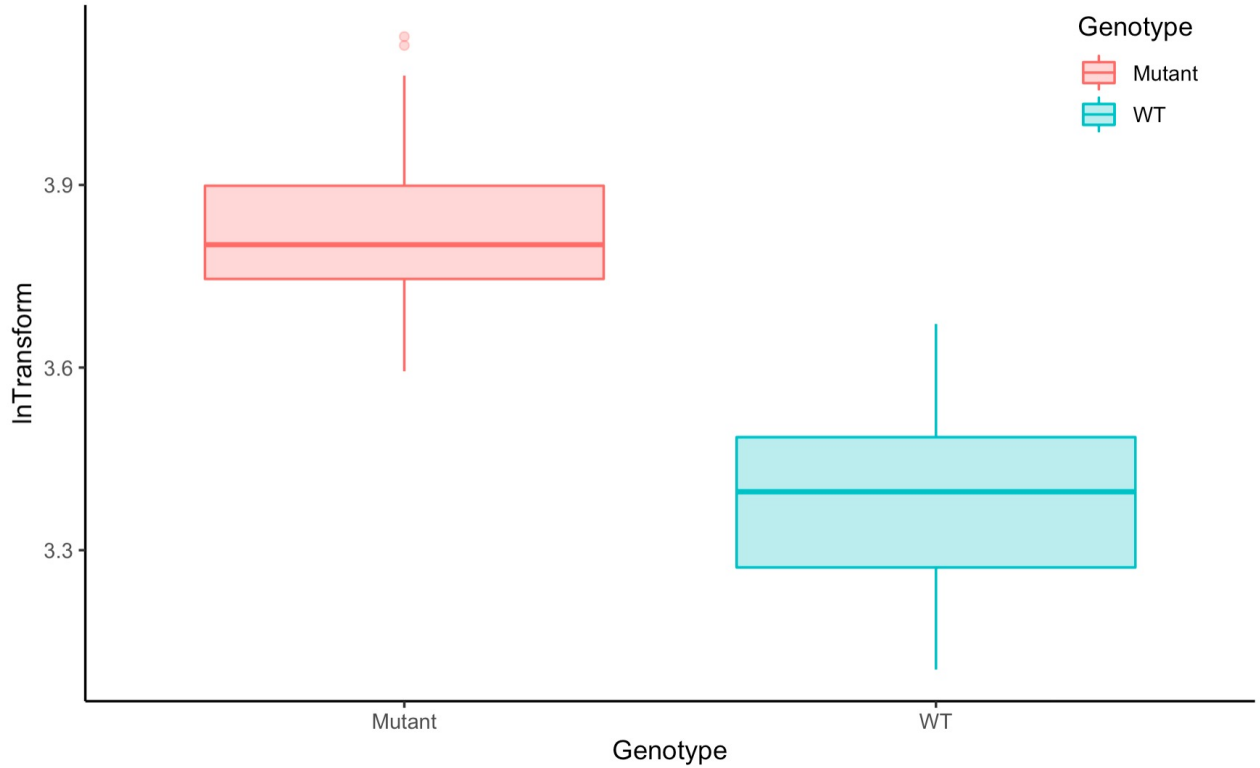
Dataset: Mutant versus Wildtype cells

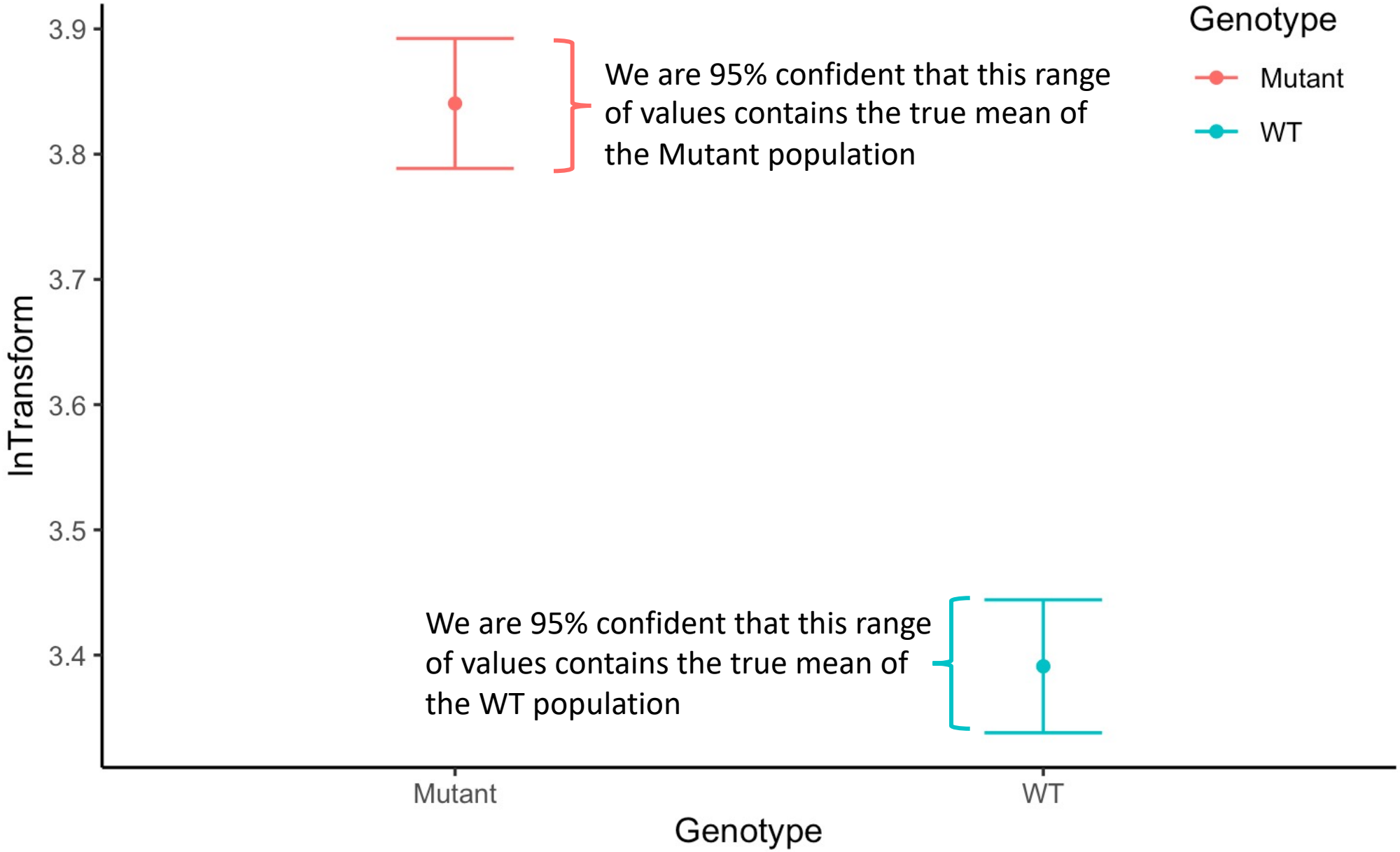
- Hypothesis—the two cell types are different.

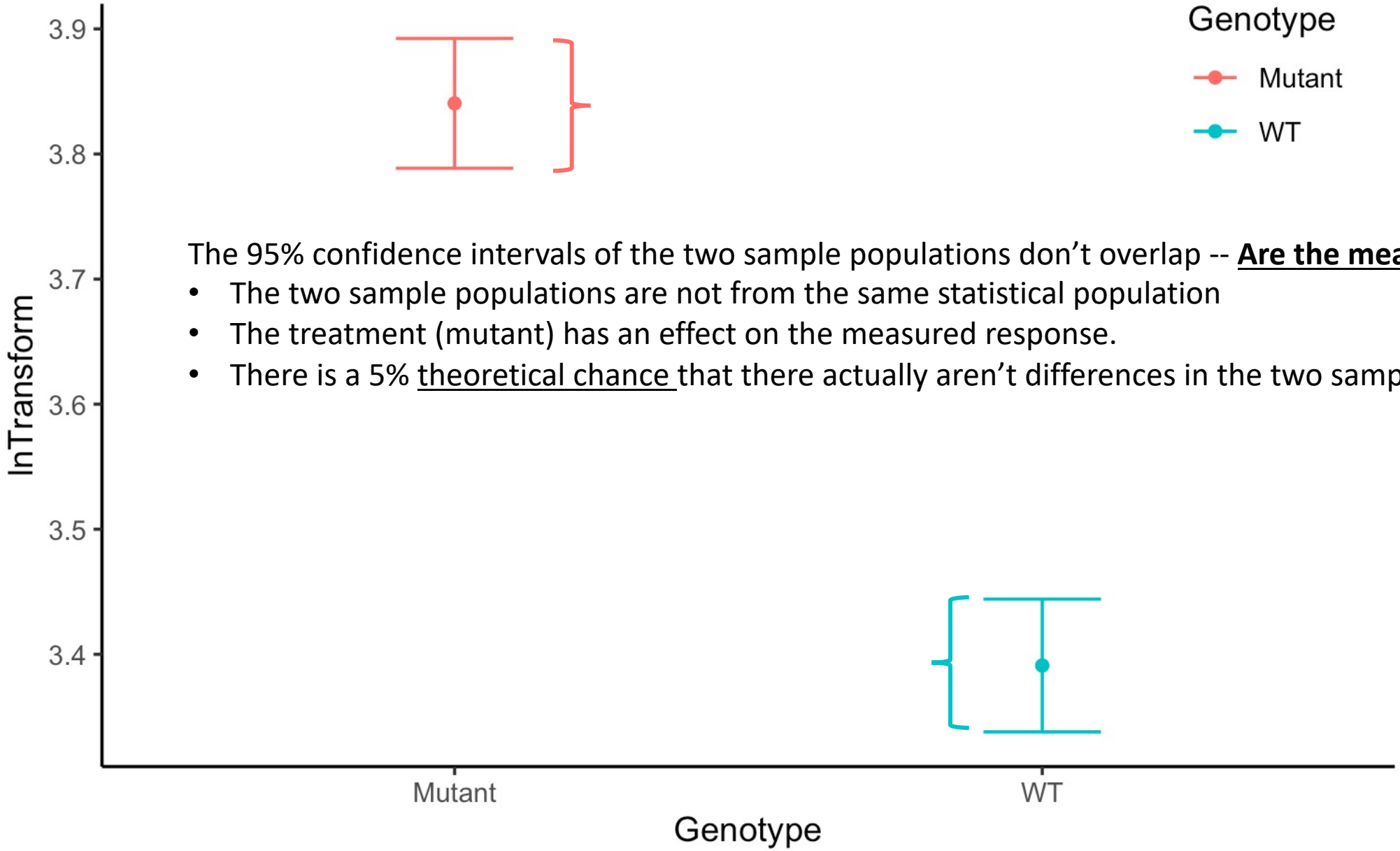
Did the treatment influence the response variable?

- Are the means different?
  - The mean itself
  - Variation around the mean



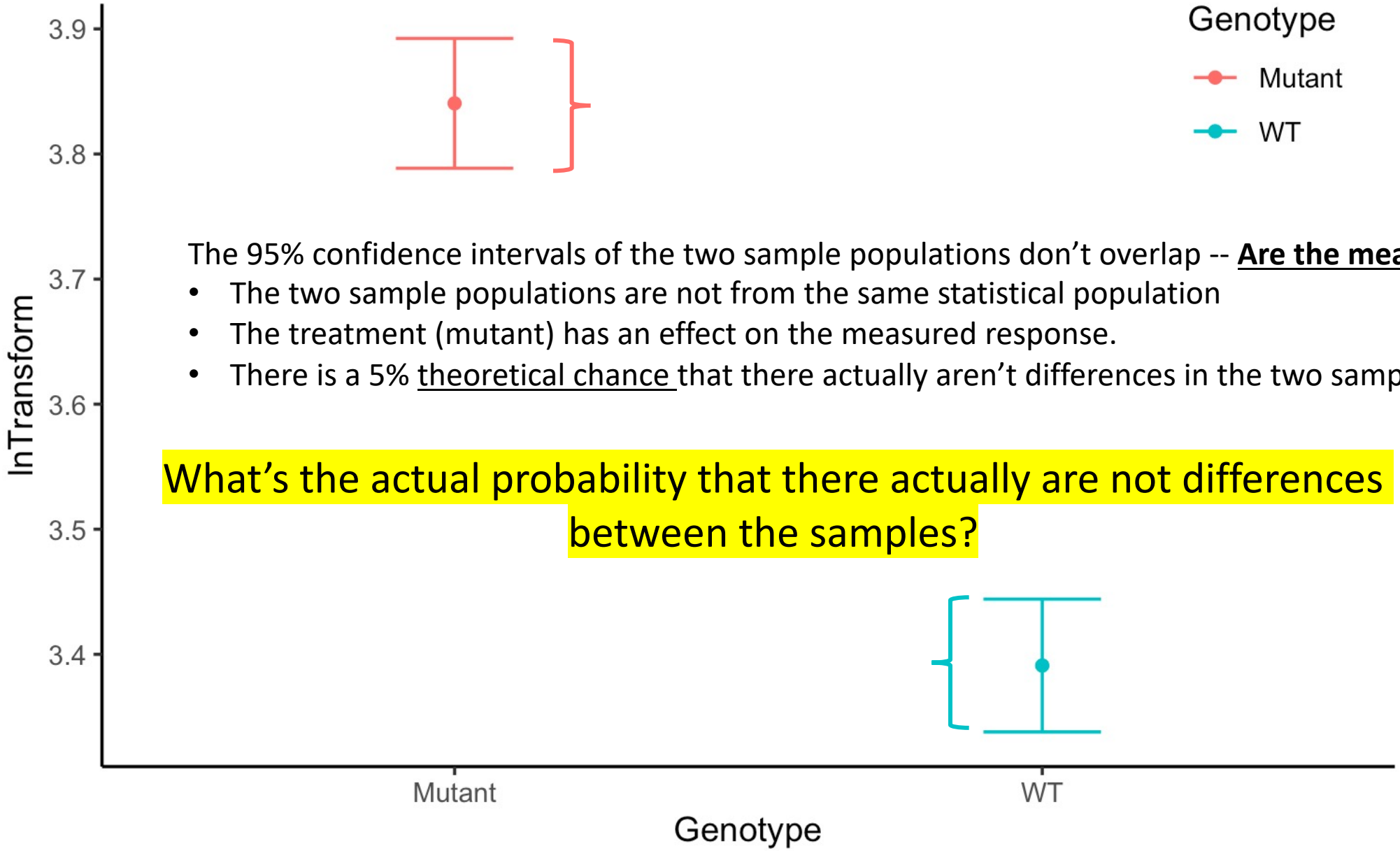



3.9
3.8
3.7
3.6
3.5
3.4
lnTransform
We are 95% confident that this range of values contains the true mean of the Mutant population
We are 95% confident that this range of values contains the true mean of the WT population
Genotype
Mutant
WT
Mutant
WT
Genotype



The 95% confidence intervals of the two sample populations don't overlap -- **Are the means different?**

- The two sample populations are not from the same statistical population
- The treatment (mutant) has an effect on the measured response.
- There is a 5% theoretical chance that there actually aren't differences in the two samples



The 95% confidence intervals of the two sample populations don't overlap -- **Are the means different?**

- The two sample populations are not from the same statistical population
- The treatment (mutant) has an effect on the measured response.
- There is a 5% theoretical chance that there actually aren't differences in the two samples

What's the actual probability that there actually are not differences between the samples?

Objectives:

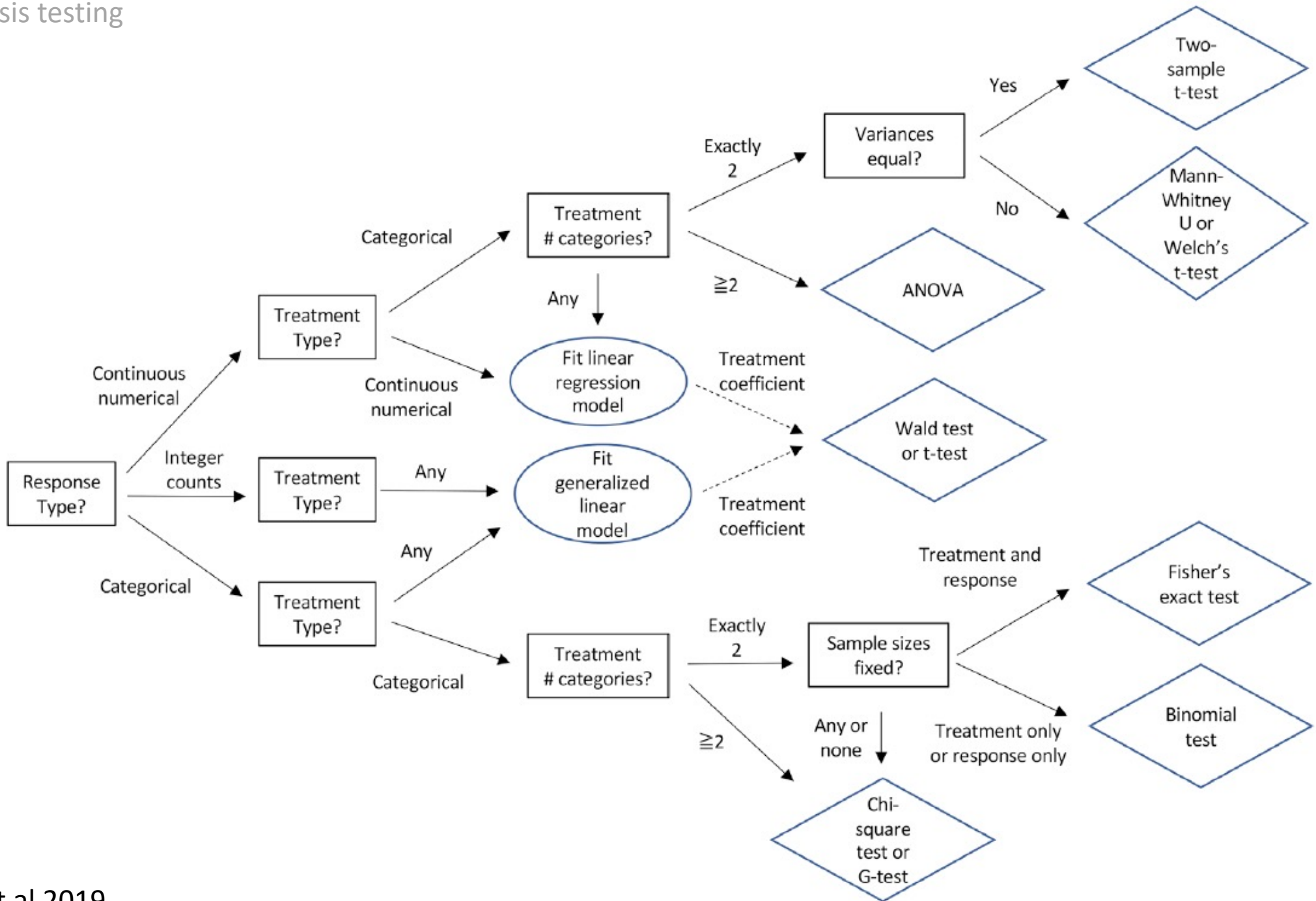
- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

3_HypothesisTestingCode.R
Contains code for describing and plotting data in following examples (need R and Rstudio to run)

What do we know so far?

- Need representative, random, and independent samples
- Need to *maximize* precision, repeatability, and accuracy and *minimize* bias
- The Empirical Rule and properties of the Normal Distribution form the basis of **ALL parametric statistical tests**
  - Our statistical tests are most powerful when data is normally distributed
- Increasing sample size in experiments can compensate for non-normal data with caveat that it is possible to reach a point of diminishing returns
- Check and calculate descriptive statistics for raw data to look for unexpected data points, check model assumptions, and to make early inferences

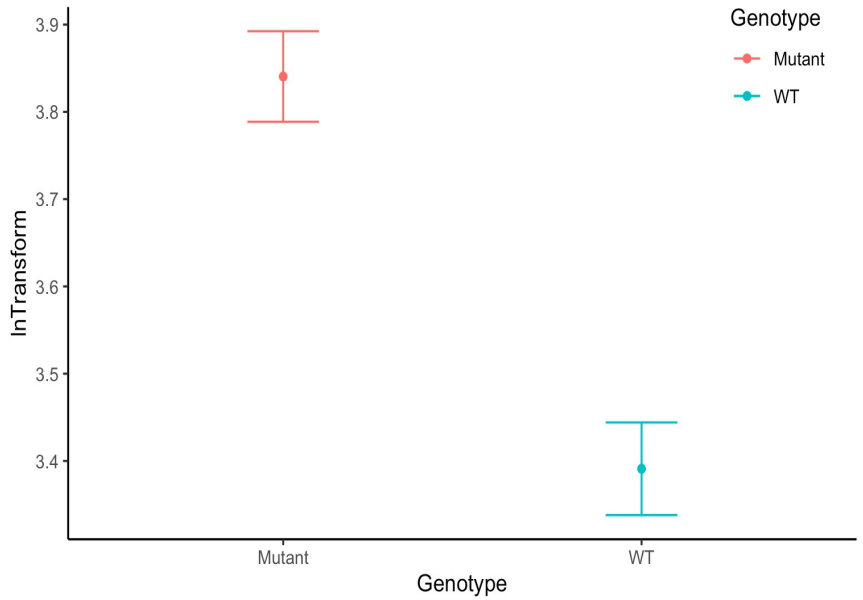
Hypothesis testing



Pollard et al 2019

Experiment Design:

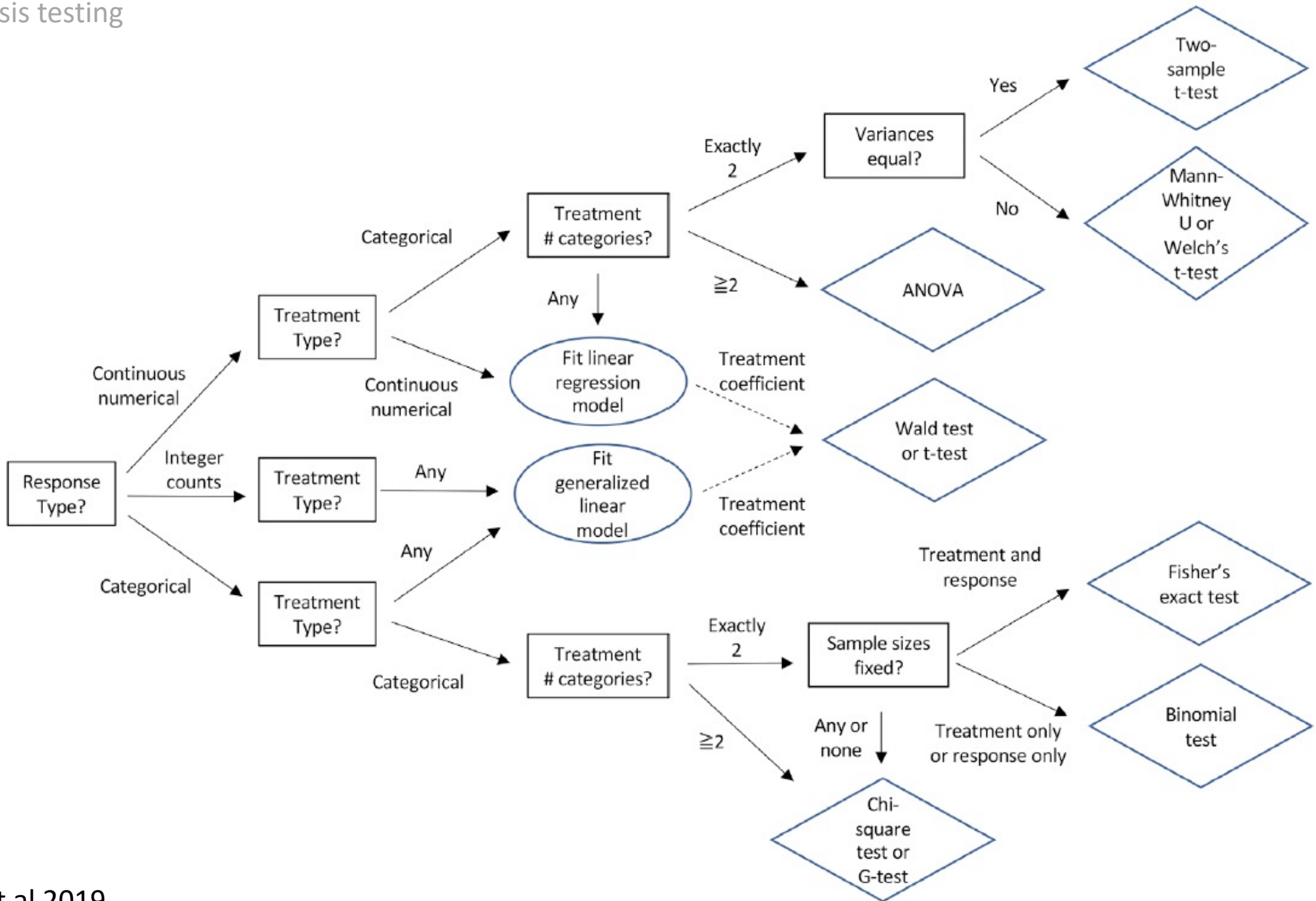
- Mutant vs WT cells
  - 2 treatment categories
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - ?
- Statistical Test:
  - ?



Are the treatment categories different?

What's the actual probability that there actually are not differences between the samples?

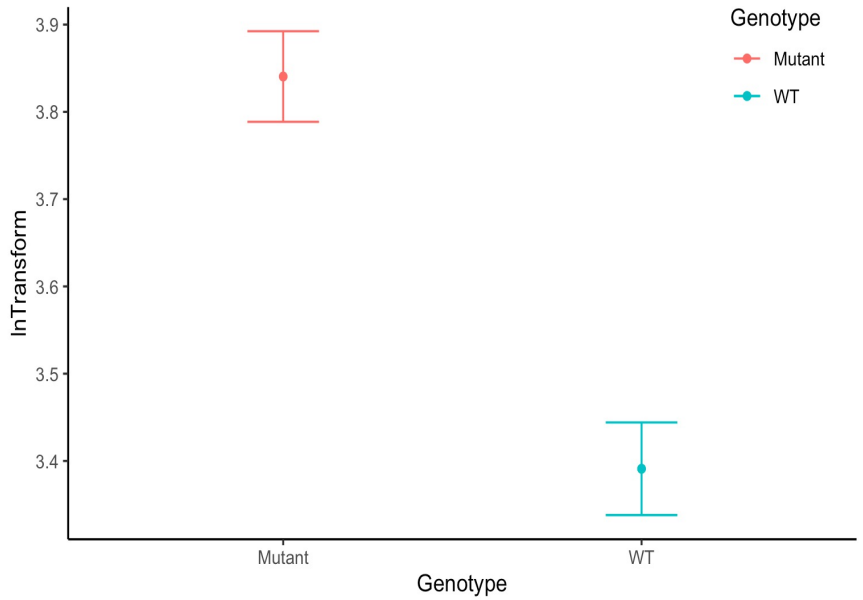
Hypothesis testing



Pollard et al 2019

Experiment Design:

- Mutant vs WT cells
  - 2 treatment categories
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - Students T Test

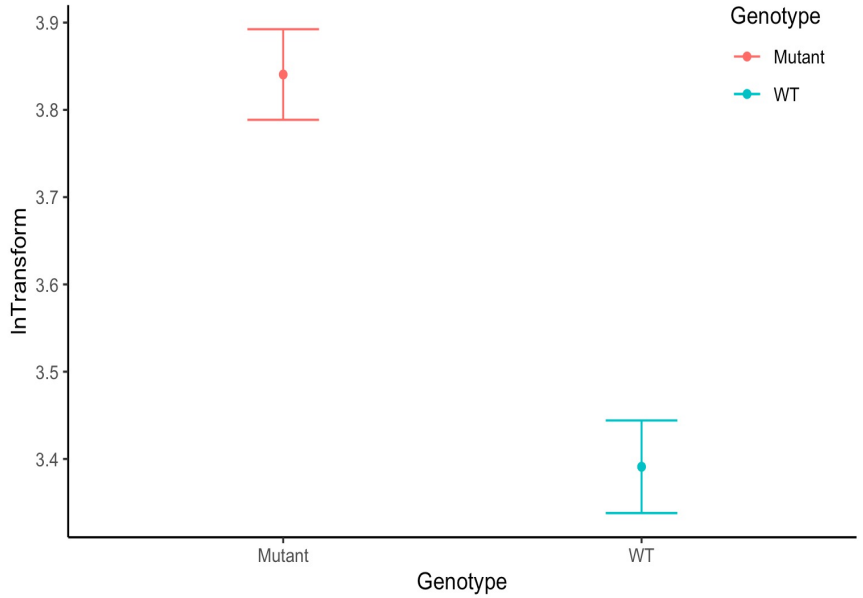


Are the treatment categories different?

What's the actual probability that there actually are not differences between the samples?

Experiment Design:

- Mutant vs WT cells
  - 2 treatment categories
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - Students T Test



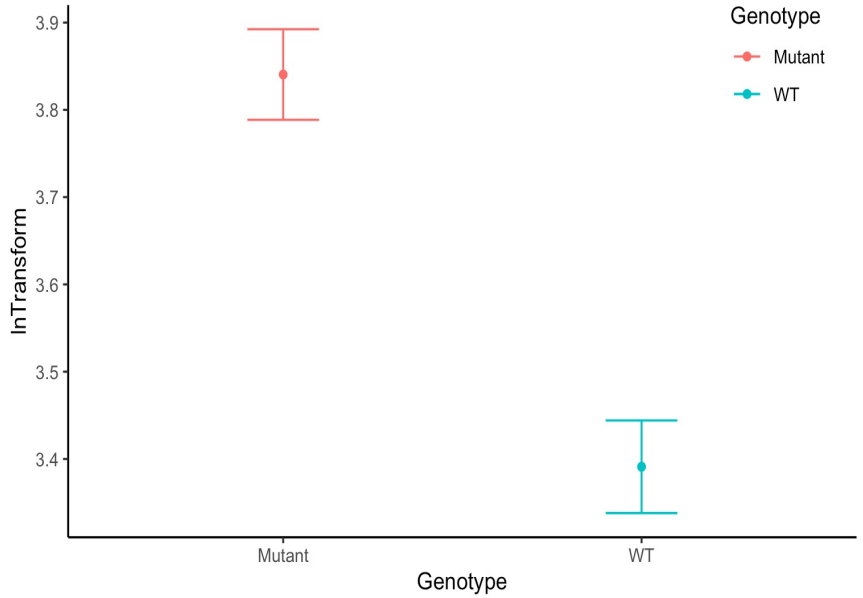
Are the treatment categories different?

What's the actual probability that there actually are not differences between the samples?

```
                    Welch Two Sample t-test
data: lnTransform by Genotype
t = 12.362, df = 60.417, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
       0.3767029 0.5221221
sample estimates:
mean in group Mutant       mean in group WT
            3.840482                3.391070
```

Experiment Design:

- Mutant vs WT cells
  - 2 treatment categories
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - Students T Test



Are the treatment categories different?

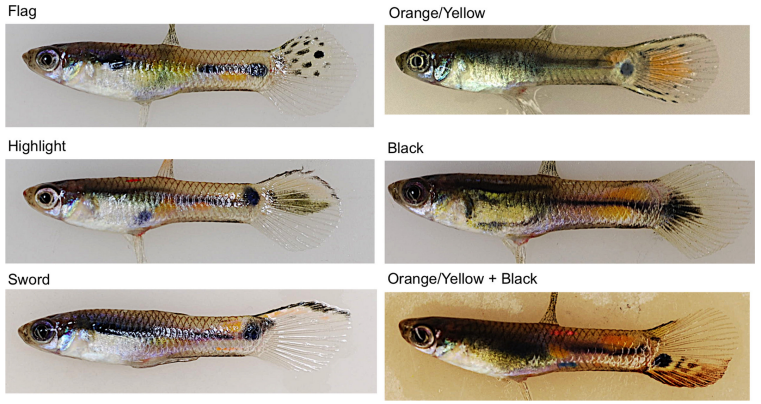
What's the actual probability that there actually are not differences between the samples?

There is a 0.00000000000022% chance we are incorrect if we conclude that the treatments are different.

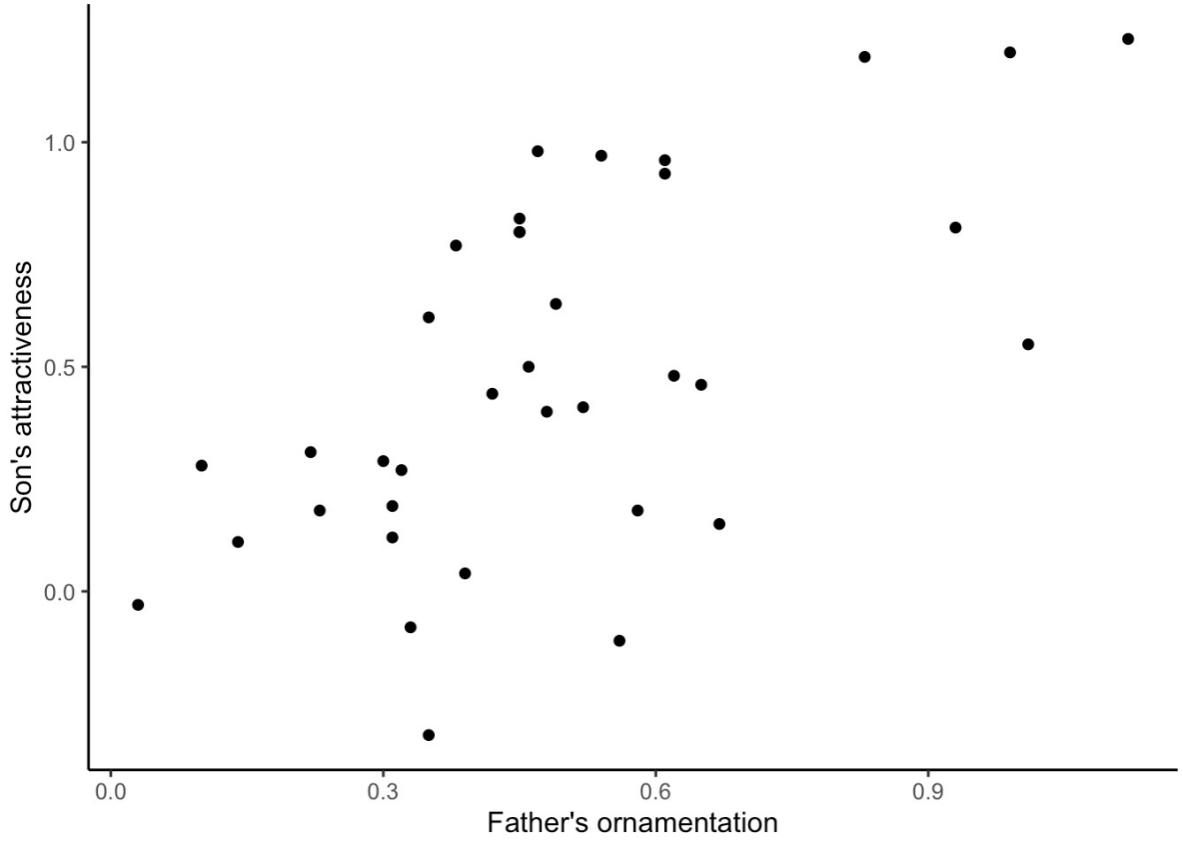
```
                Welch Two Sample t-test
data: lnTransform by Genotype
t = 12.362, df = 60.417, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
      0.3767029 0.5221221
sample estimates:
mean in group Mutant       mean in group WT
            3.840482               3.391070
```

Experiment Design:

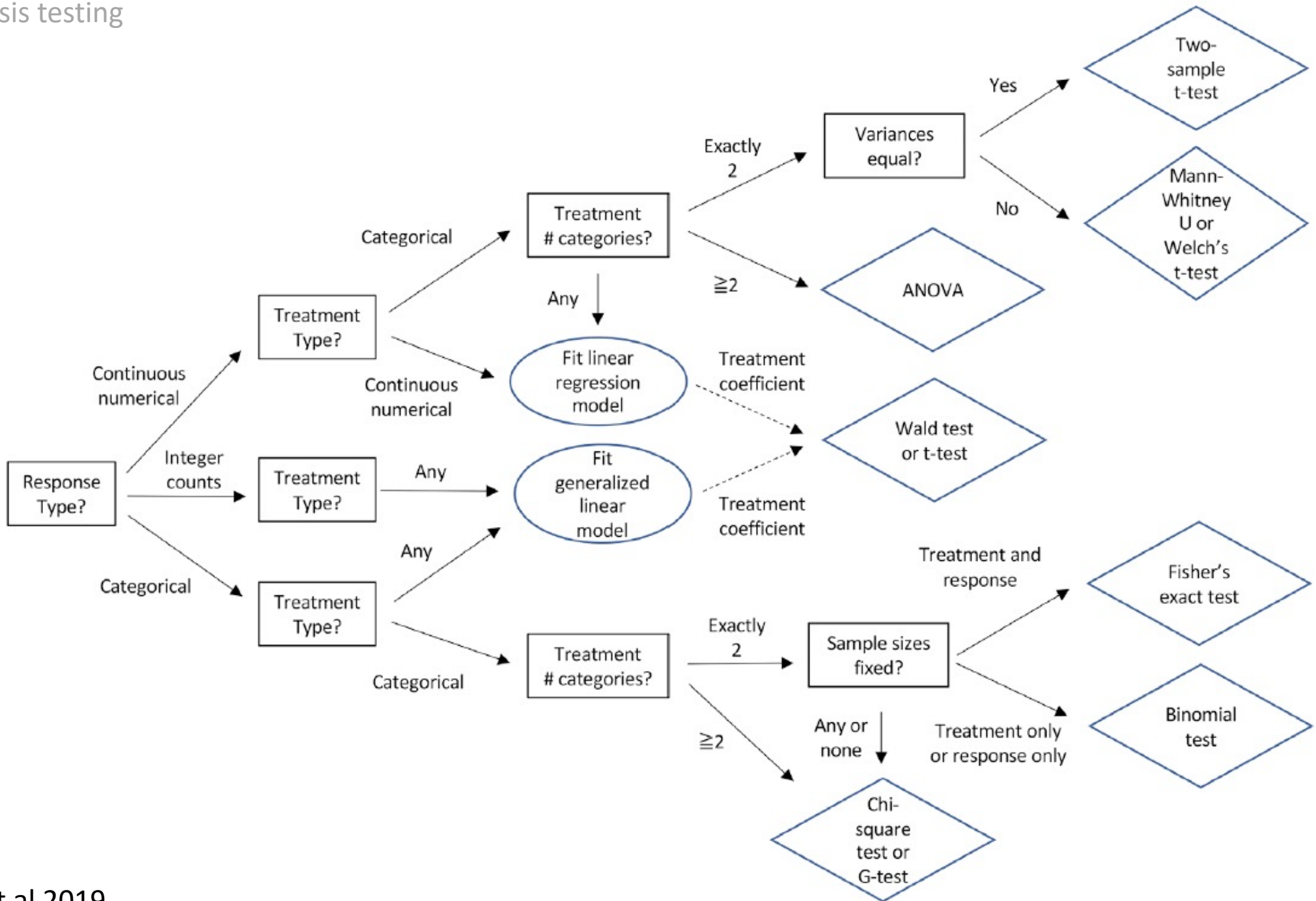
- Measuring son attractiveness
  - 1 treatment
  - Continuous
- Response variable
  - Continuous
- Meets model assumptions?
  - ?
- Statistical Test:
  - ?



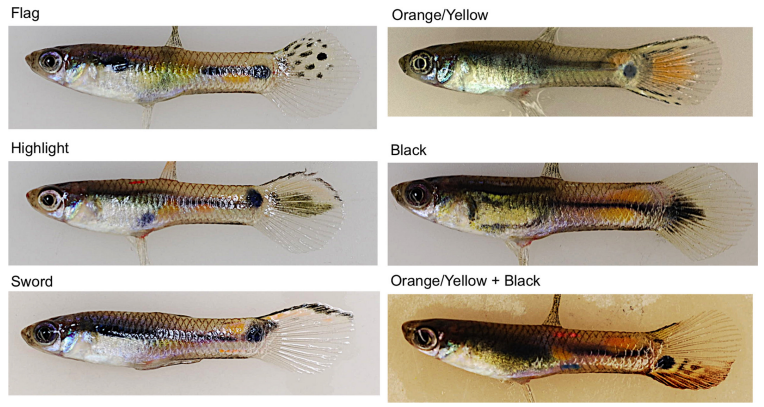
Are son attractiveness and father ornamentation correlated?



Hypothesis testing



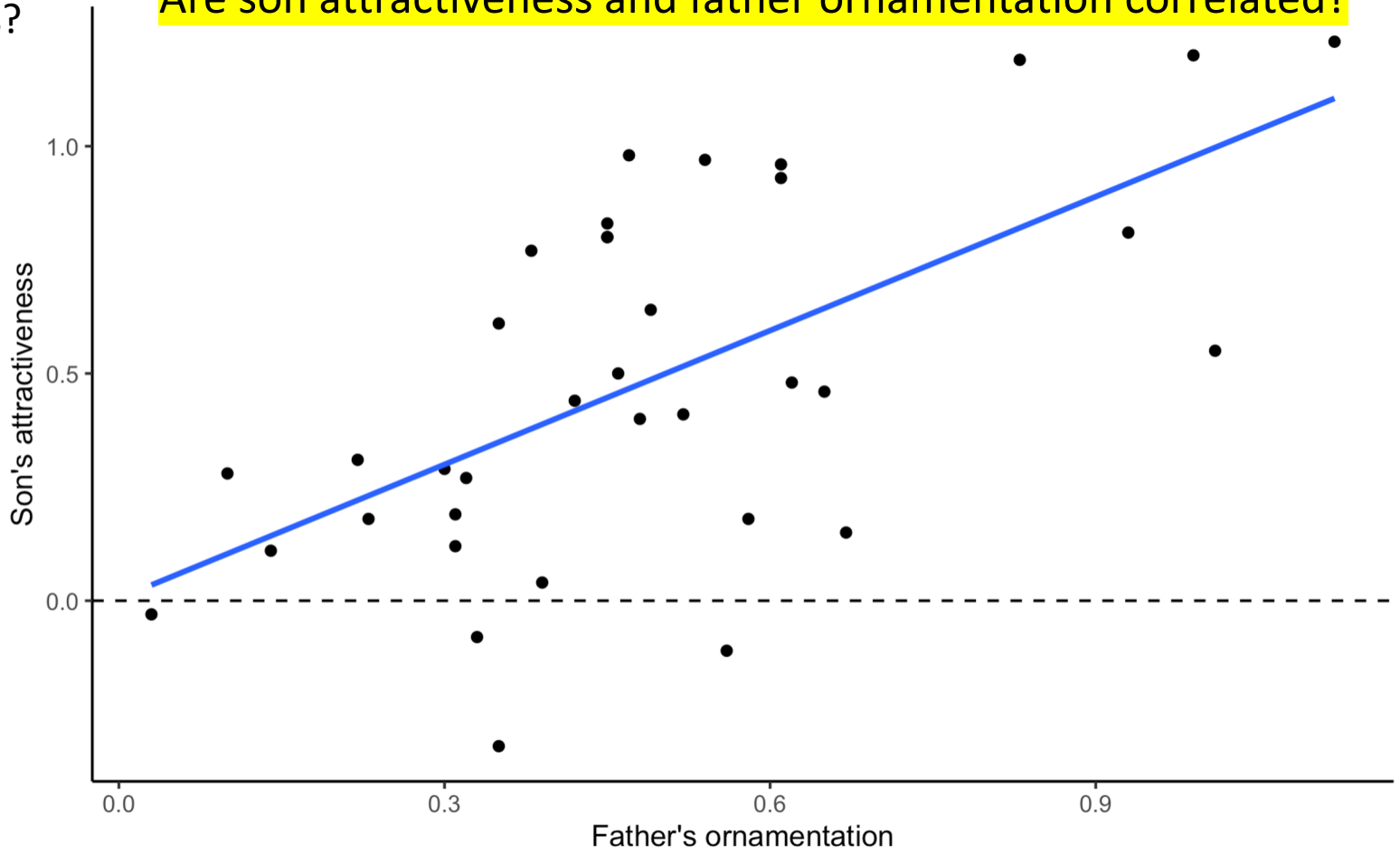
Pollard et al 2019



Experiment Design:

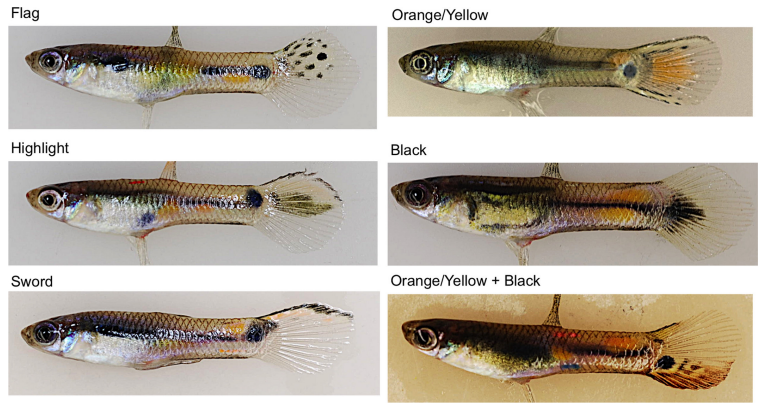
- Measuring son attractiveness
  - 1 treatment
  - Continuous
- Response variable
  - Continuous
- Meets model assumptions?
  - ?
- Statistical Test:
  - Linear regression

Are son attractiveness and father ornamentation correlated?

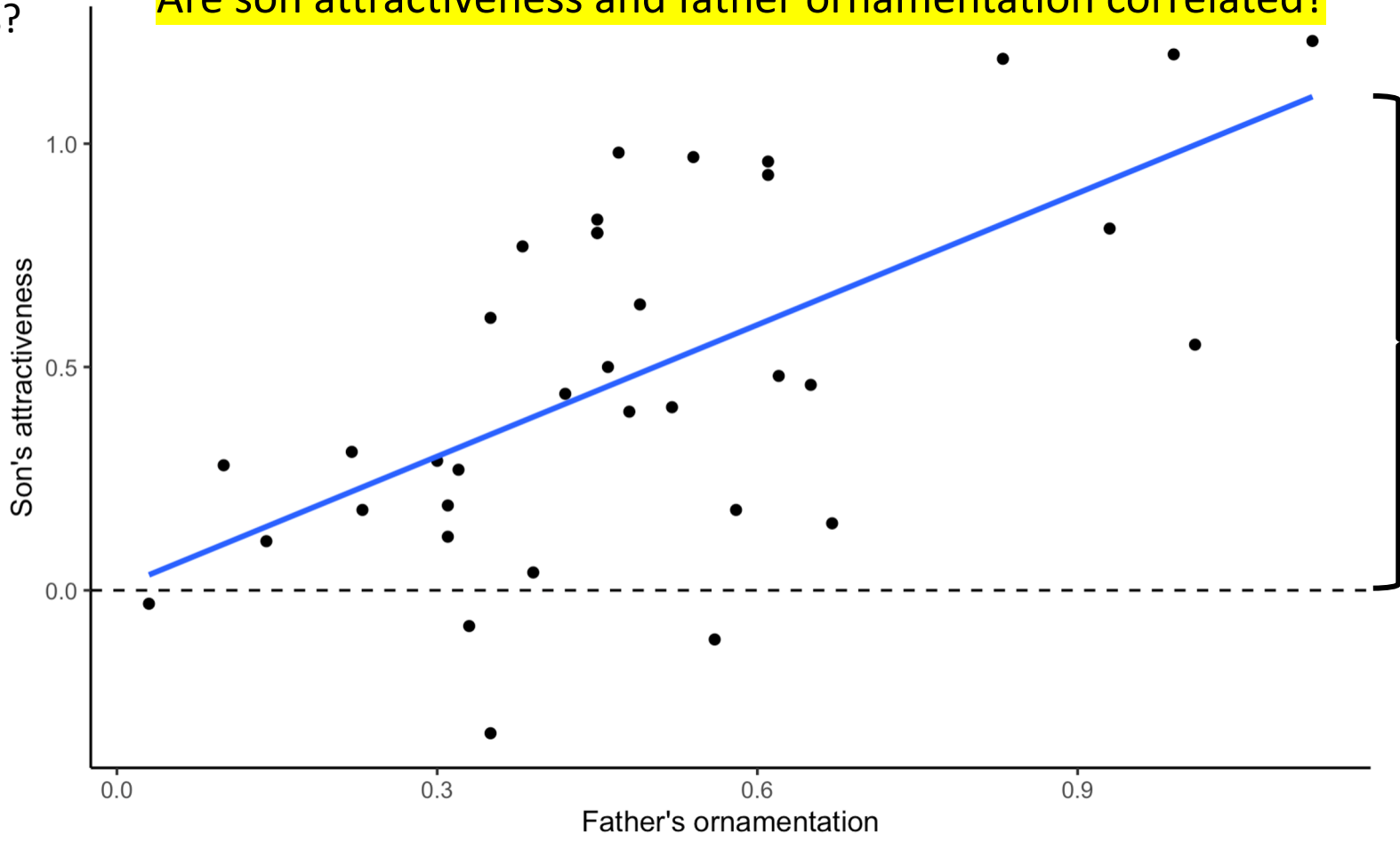


Experiment Design:

- Measuring son attractiveness
  - 1 treatment
  - Continuous
- Response variable
  - Continuous
- Meets model assumptions?
  - ?
- Statistical Test:
  - Linear regression



Are son attractiveness and father ornamentation correlated?



Is the correlation significantly different from 0?

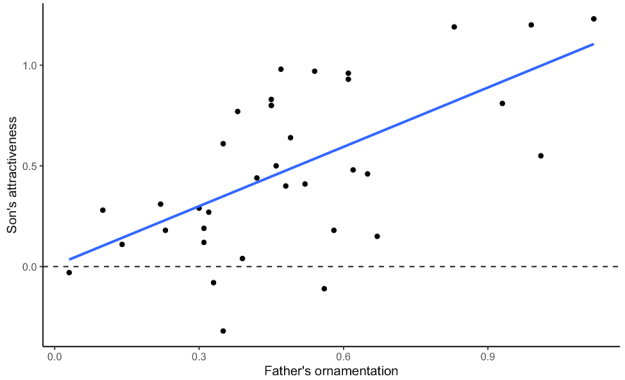
Is the slope of the regression significantly different from 0?

Experiment Design:

- Measuring son attractiveness
  - 2 treatment comparison
  - Continuous
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - Linear regression

Correlation coefficient: r = 0.61, 61%

Coefficient of determination: r^2 = 0.38, 38%



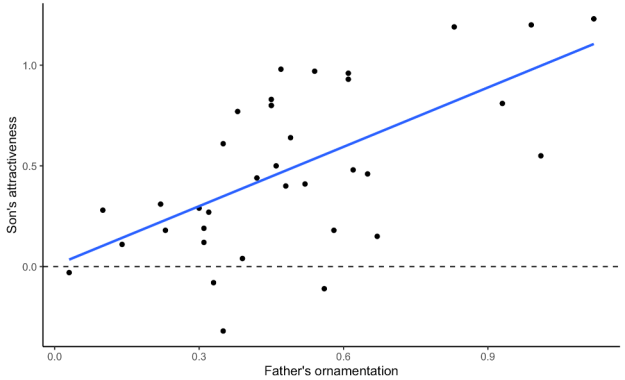
38% of variation in son attractiveness is explained by father ornamentation

Experiment Design:

- Measuring son attractiveness
  - 2 treatment comparison
  - Continuous
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - Linear regression

Correlation coefficient: r = 0.61, 61%
Coefficient of determination: r^2 = 0.38, 38%

Is the correlation significantly different from 0?

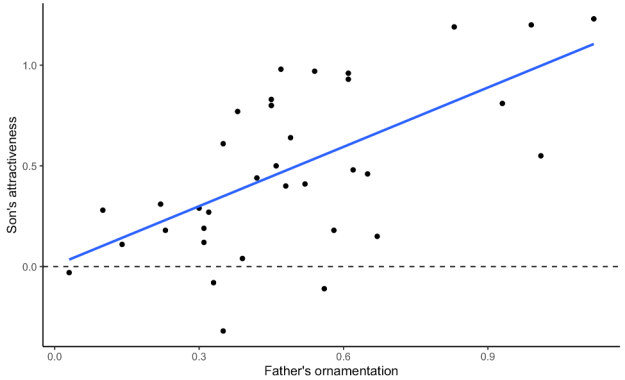


Experiment Design:

- Measuring son attractiveness
  - 2 treatment comparison
  - Continuous
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - Linear regression

Correlation coefficient: r = 0.61, 61%
Coefficient of determination: r^2 = 0.38, 38%

Is the correlation significantly different from 0?




```
                Pearson's product-moment correlation

data:  guppyData$fatherOrnamentation and guppyData$sonAttractiveness
t = 4.5371, df = 34, p-value = 6.784e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3577455 0.7843860
sample estimates:
      cor
0.6141043
```

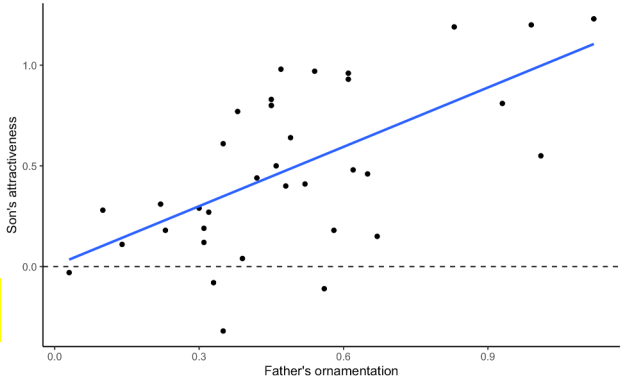
Experiment Design:

- Measuring son attractiveness
  - 2 treatment comparison
  - Continuous
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - Linear regression

Correlation coefficient: r = 0.61, 61%

Coefficient of determination: r^2 = 0.38, 38%

Is the regression significant (is the slope different from 0)?

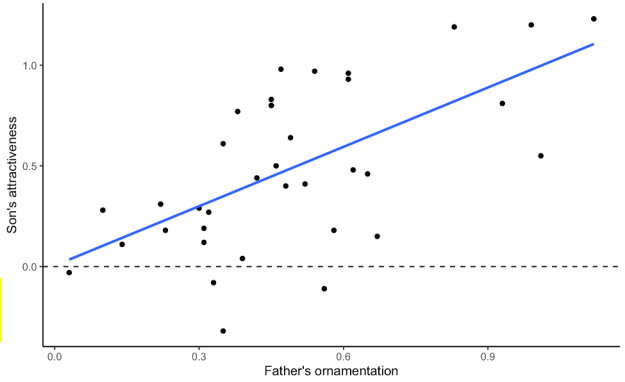


Experiment Design:

- Measuring son attractiveness
  - 2 treatment comparison
  - Continuous
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - Linear regression

Correlation coefficient: r = 0.61, 61%

Coefficient of determination: r^2 = 0.38, 38%



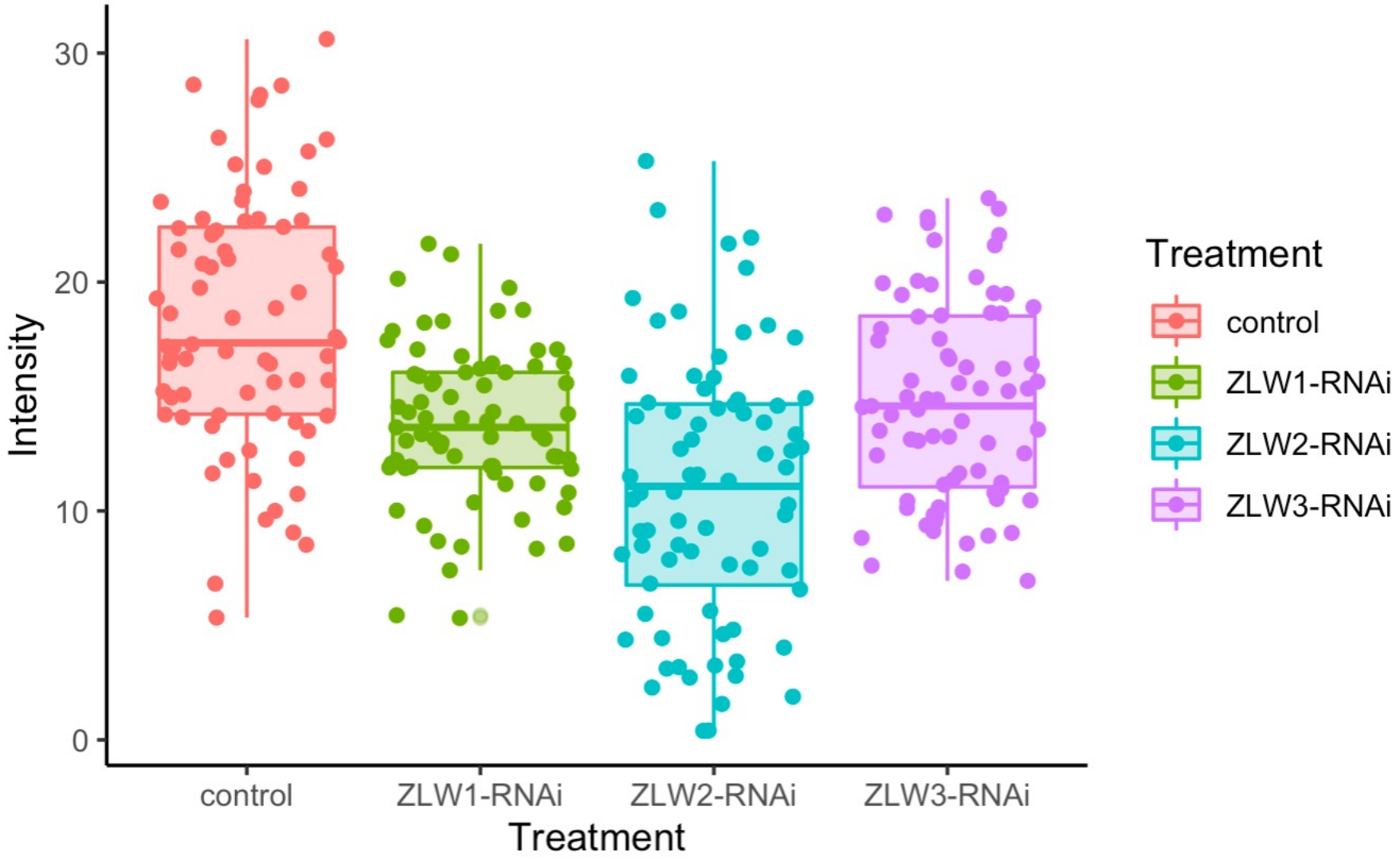
Is the regression significant (is the slope different from 0)?

```
                                   Call:
lm(formula = sonAttractiveness ~ fatherOrnamentation, data = guppyData)

                                Residuals:
                     Min       1Q   Median       3Q      Max
                -0.66888 -0.14647 -0.02119  0.27727  0.51324

                                Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
                (Intercept)  0.005084   0.118988   0.043    0.966
fatherOrnamentation  0.982285   0.216499   4.537 6.78e-05 ***
                                   ---
                Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                Residual standard error: 0.3212 on 34 degrees of freedom
                Multiple R-squared:  0.3771,  Adjusted R-squared:  0.3588
                F-statistic: 20.59 on 1 and 34 DF,  p-value: 6.784e-05
```

There is a 0.000678% chance the slope is not different from 0

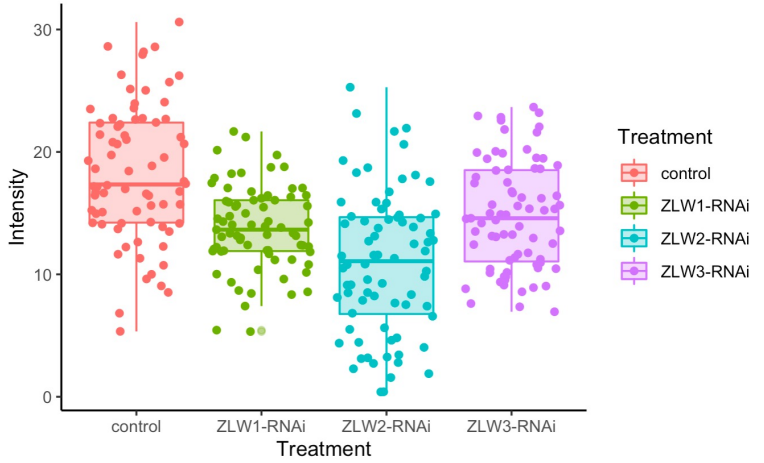
Experiment Design: RNAi

- Measuring fluorescence intensity of a reporter
  - Treatment with 4 groups
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - ?
- Statistical Test:
  - ?



Experiment Design: RNAi

- Measuring fluorescence intensity of a reporter
  - Treatment with 4 groups
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - ?
- Statistical Test:
  - ?

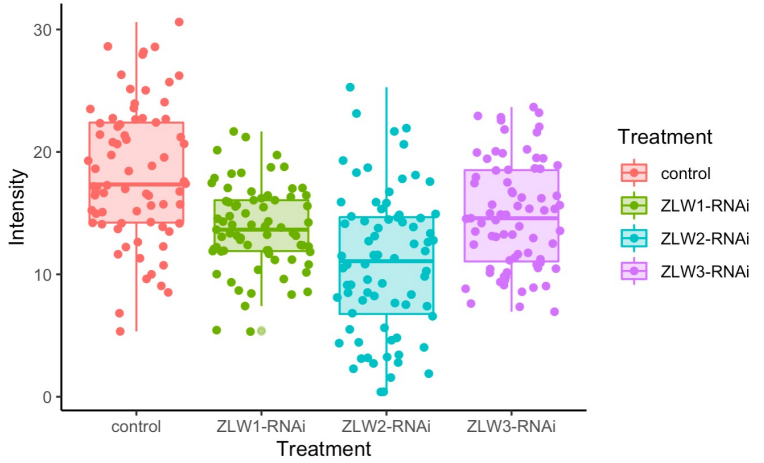


Multiple T Tests?

- Control vs ZLW1-RNAi ← 5% Type I error rate
- Control vs ZLW2-RNAi ← 5% Type I error rate
- Control vs ZLW3-RNAi ← 5% Type I error rate

Experiment Design: RNAi

- Measuring fluorescence intensity of a reporter
  - Treatment with 4 groups
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - ?
- Statistical Test:
  - ?



Multiple T Tests?

- Control vs ZLW1-RNAi ← 5% Type I error rate
- Control vs ZLW2-RNAi ← 5% Type I error rate
- Control vs ZLW3-RNAi ← 5% Type I error rate

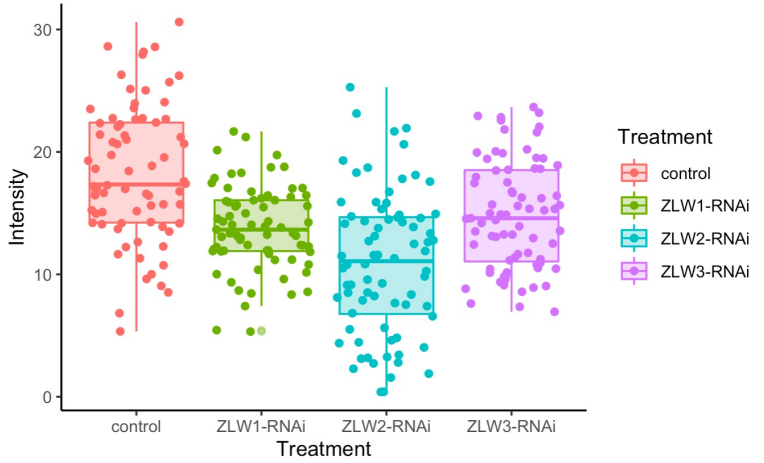
$$1 - (1 - \alpha)^N = 1 - (1 - 0.05)^3 = 14.3\%$$ Experiment-wide Type I error rate

Number of tests

Probability of incorrectly concluding treatment has an effect is much higher

Experiment Design: RNAi

- Measuring fluorescence intensity of a reporter
  - Treatment with 4 groups
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - ?
- Statistical Test:
  - ANOVA



Multiple T Tests

- Control vs
- Control vs -RN
- Control vs -RNAi

e I error rate

5% error rate

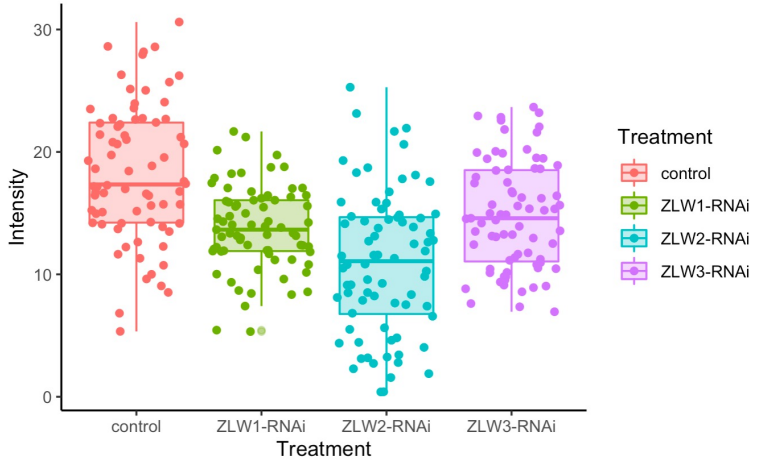
5% error rate

$1-(1-\alpha)^{N}$ $(1-0.05)$ periment-wide Type I error rate

1 ANOVA (analysis of variance) and post hoc comparisons instead of multiple t tests.

Experiment Design: RNAi

- Measuring fluorescence intensity of a reporter
  - Treatment with 4 groups
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - ANOVA

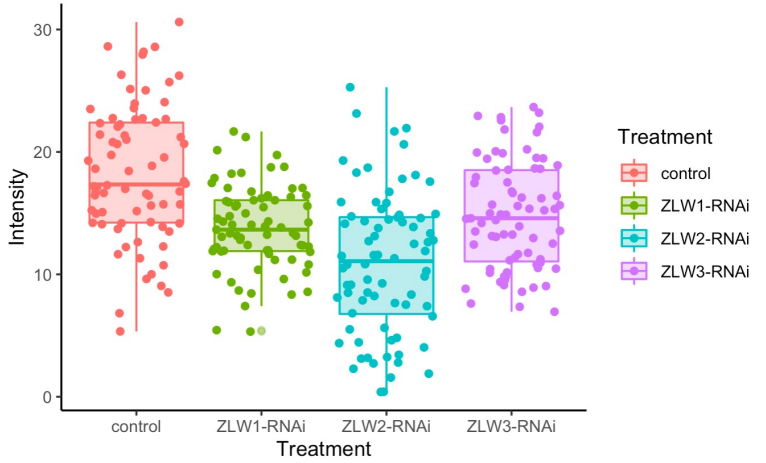


Do any of the strains influence fluorescence intensity?

Which strains influence fluorescence intensity?

Experiment Design: RNAi

- Measuring fluorescence intensity of a reporter
  - Treatment with 4 groups
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - ANOVA



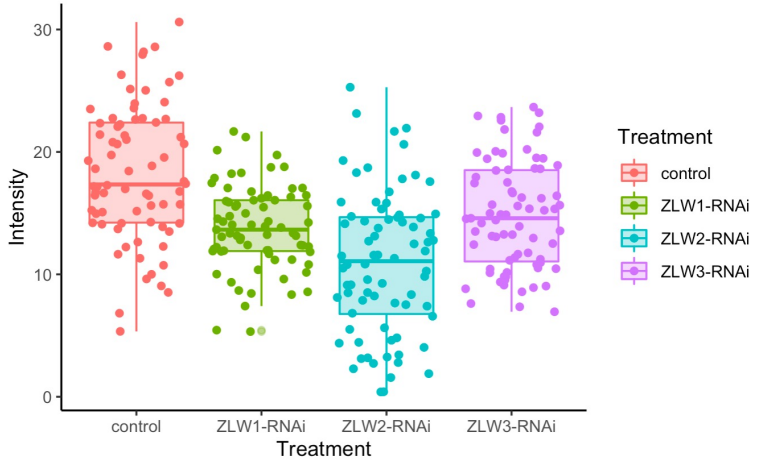
Do any of the strains influence fluorescence intensity?

Which strains influence fluorescence intensity?

Source	DF	SS	MS	F value	P value
Treatment	3	2074	691.4	28.68	2.7e-16
Residuals	294	7087	24.1		

Experiment Design: RNAi

- Measuring fluorescence intensity of a reporter
  - Treatment with 4 groups
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - ANOVA

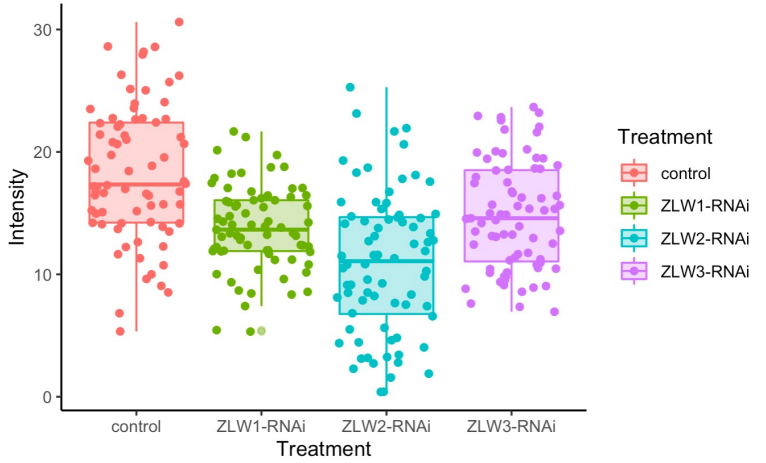


Do any of the strains influence fluorescence intensity?

Which strains influence fluorescence intensity?

Experiment Design: RNAi

- Measuring fluorescence intensity of a reporter
  - Treatment with 4 groups
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - ANOVA



Do any of the strains influence fluorescence intensity?

Which strains influence fluorescence intensity?

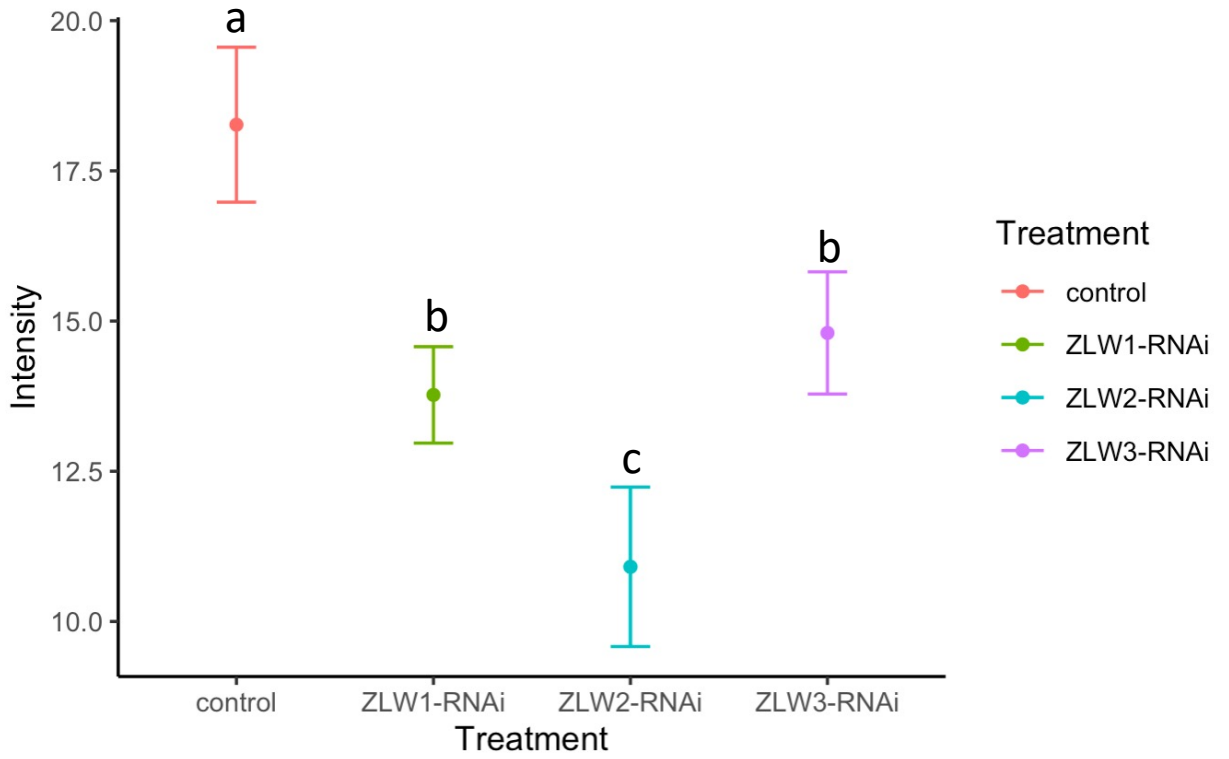
```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Intensity ~ Treatment, data = data5)

$Treatment
                             diff       lwr        upr     p adj
ZLW1-RNAi-control      -4.497413 -6.590101 -2.4047251 0.0000004
ZLW2-RNAi-control      -7.358565 -9.430357 -5.2867738 0.0000000
ZLW3-RNAi-control      -3.466019 -5.544614 -1.3874250 0.0001315
ZLW2-RNAi-ZLW1-RNAi    -2.861152 -4.940121 -0.7821835 0.0024652
ZLW3-RNAi-ZLW1-RNAi     1.031394 -1.054355  3.1171419 0.5778909
ZLW3-RNAi-ZLW2-RNAi     3.892546  1.827764  5.9573277 0.0000108
```

← Type I error rate remains at 5%

Experiment Design: RNAi

- Measuring fluorescence intensity of a reporter
  - Treatment with 4 groups
  - Categorical
- Response variable
  - Continuous
- Meets model assumptions?
  - Yes
- Statistical Test:
  - ANOVA

```
                Tukey multiple comparisons of means
                  95% family-wise confidence level

Fit: aov(formula = Intensity ~ Treatment, data = data5)

$Treatment
                                diff       lwr        upr     p adj
ZLW1-RNAi-control       -4.497413 -6.590101 -2.4047251 0.0000004
ZLW2-RNAi-control       -7.358565 -9.430357 -5.2867738 0.0000000
ZLW3-RNAi-control       -3.466019 -5.544614 -1.3874250 0.0001315
ZLW2-RNAi-ZLW1-RNAi     -2.861152 -4.940121 -0.7821835 0.0024652
ZLW3-RNAi-ZLW1-RNAi      1.031394 -1.054355  3.1171419 0.5778909
ZLW3-RNAi-ZLW2-RNAi      3.892546  1.827764  5.9573277 0.0000108
```

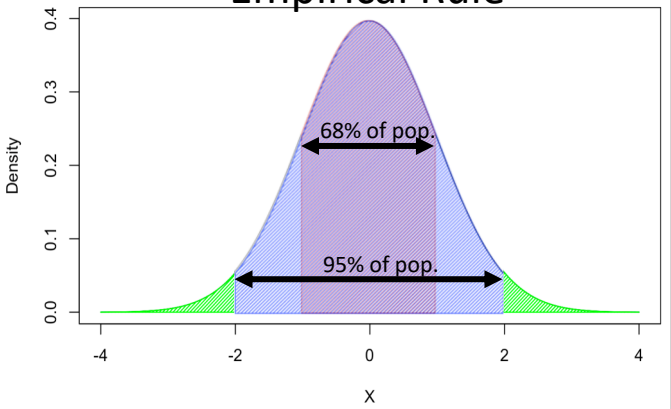


The treatment strain influenced fluorescence intensity ($F_{3,294}$ = 28.68, p << 0.005). All of the RNAi strains had lower intensity compared to the control strain (Tukey HSD test: p < 0.05, experiment-wide alpha = 0.05); however, fluorescence intensity was not significantly different between the ZLW1-RNAi and ZLW3-RNAi strains (adjusted p = 0.57).

Experiment Design:

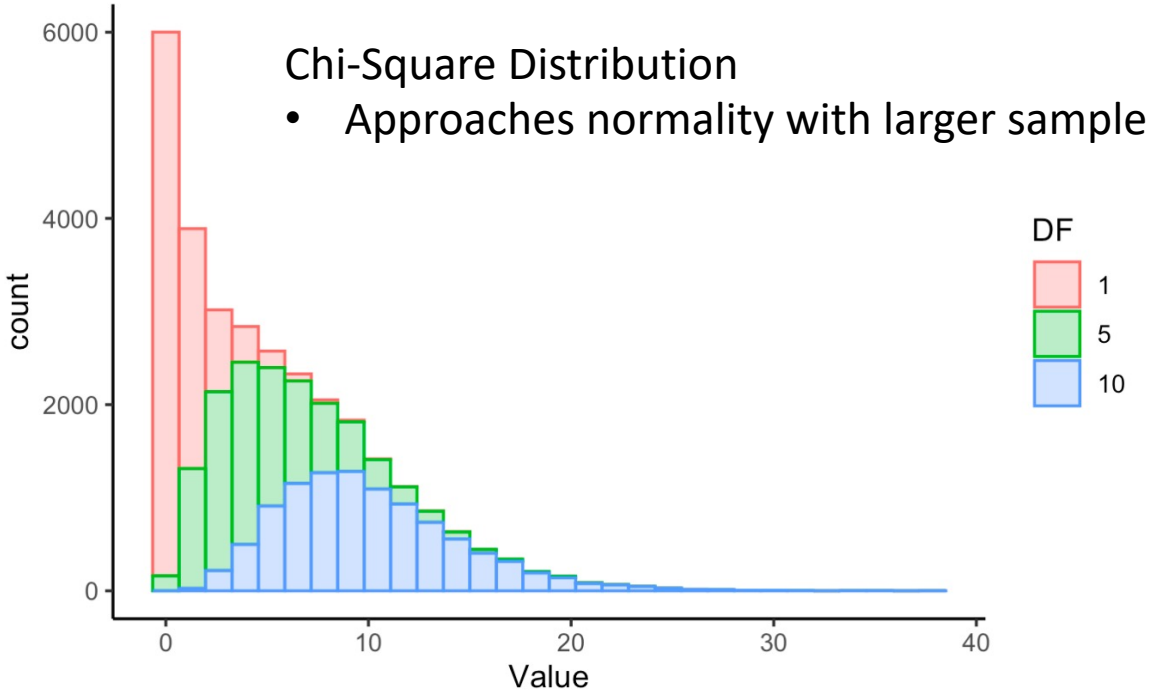
- Measuring expression of Sox2
  - Treatment with 3 groups
  - Categorical
- Response variable
  - Categorical
- Meets model assumptions?
  - Data are independent
  - <20% of expected values are < 5
- Statistical Test:
  - Chi-Square test
    - Observed frequency
    - Expected frequency

Empirical Rule



Chi-Square Distribution

- Approaches normality with larger sample sizes…

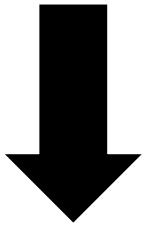


Experiment Design:

- Measuring expression of Sox2
  - Treatment with 3 groups
  - Categorical
- Response variable
  - Categorical
- Meets model assumptions?
  - Data are independent
  - <20% of expected values are < 5
- Statistical Test:
  - Chi-Square test
    - Observed frequency
    - Expected frequency

Raw Data:

Treatment	**Phenotype**
Control-MO	Decreased
JmjD2A-sbMO+tbMO	WT
JmjD2A-sbMO+pCI-JmjD2A	



Convert to count data for each category

Observed Frequency:

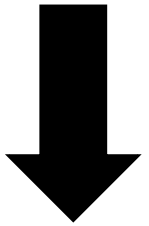
Treatment	Phenotype	
	Decreased	WT
Control-MO	2	20
JmjD2A-sbMO+pCI-JmjD2A	6	9
JmjD2A-sbMO+tbMO	34	15

Experiment Design:

- Measuring expression of Sox2
  - Treatment with 3 groups
  - Categorical
- Response variable
  - Categorical
- Meets model assumptions?
  - Data are independent
  - <20% of expected values are < 5
- Statistical Test:
  - Chi-Square test
    - Observed frequency
    - Expected frequency

Counts Table:

Treatment	Phenotype: Decreased	Phenotype: WT
Control-MO	2	20
JmjD2A-sbMO+pCI-JmjD2A	6	9
JmjD2A-sbMO+tbMO	34	15



Calculate expected values for each category

Expected Frequency:

Treatment	Phenotype: Decreased	Phenotype: WT
Control-MO	10.744186	11.255814
JmjD2A-sbMO+pCI-JmjD2A	7.325581	7.674419
JmjD2A-sbMO+tbMO	23.930233	25.069767

Experiment Design:

- Measuring expression of Sox2
  - Treatment with 3 groups
  - Categorical
- Response variable
  - Categorical
- Meets model assumptions?
  - Data are independent
  - <20% of expected values are < 5
- Statistical Test:
  - Chi-Square test
    - Observed frequency
    - Expected frequency

Counts Table:

Treatment	Phenotype: Decreased	Phenotype: WT
Control-MO	2	20
JmjD2A-sbMO+pCI-JmjD2A	6	9
JmjD2A-sbMO+tbMO	34	15

Calculate expected values for each category

Expected Frequency:

Treatment	Phenotype: Decreased	Phenotype: WT
Control-MO	10.744186	11.255814
JmjD2A-sbMO+pCI-JmjD2A	7.325581	7.674419
JmjD2A-sbMO+tbMO	23.930233	25.069767

$$X^2 = \sum \frac{(observed - expected)^2}{expected}$$

Experiment Design:

- Measuring expression of Sox2
  - Treatment with 3 groups
  - Categorical
- Response variable
  - Categorical
- Meets model assumptions?
  - Data are independent
  - <20% of expected values are < 5
- Statistical Test:
  - Chi-Square test
    - Observed frequency
    - Expected frequency

Counts Table:

Treatment	Phenotype: Decreased	Phenotype: WT
Control-MO	2	20
JmjD2A-sbMO+pCI-JmjD2A	6	9
JmjD2A-sbMO+tbMO	34	15

Calculate expected values for each category

Treatment	Phenotype: Decreased	Phenotype: WT
Control-MO	10.744186	11.255814
JmjD2A-sbMO+pCI-JmjD2A	7.325581	7.674419
JmjD2A-sbMO+tbMO	23.930233	25.069767

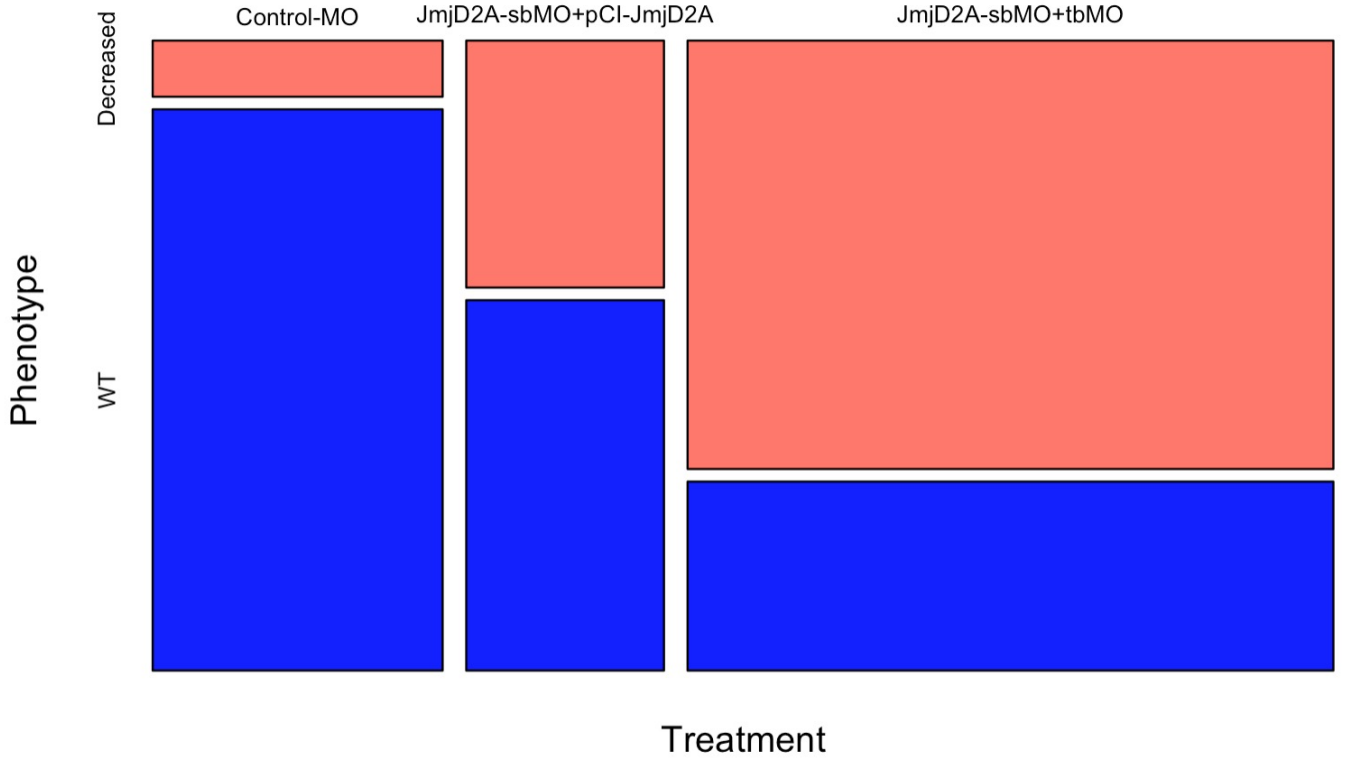
Reject the null hypothesis that treatment does not impact Sox2 expression

```
        Pearson's Chi-squared test

data:  data4.table
X-squared = 22.66, df = 2, p-value = 1.201e-05
```

Experiment Design: RNAi

- Measuring expression of Sox2
  - Treatment with 3 groups
  - Categorical
- Response variable
  - Categorical
- Meets model assumptions?
  - Data are independent
  - <20% of expected values are < 5
- Statistical Test:
  - Chi-Square test
    - Observed frequency
    - Expected frequency

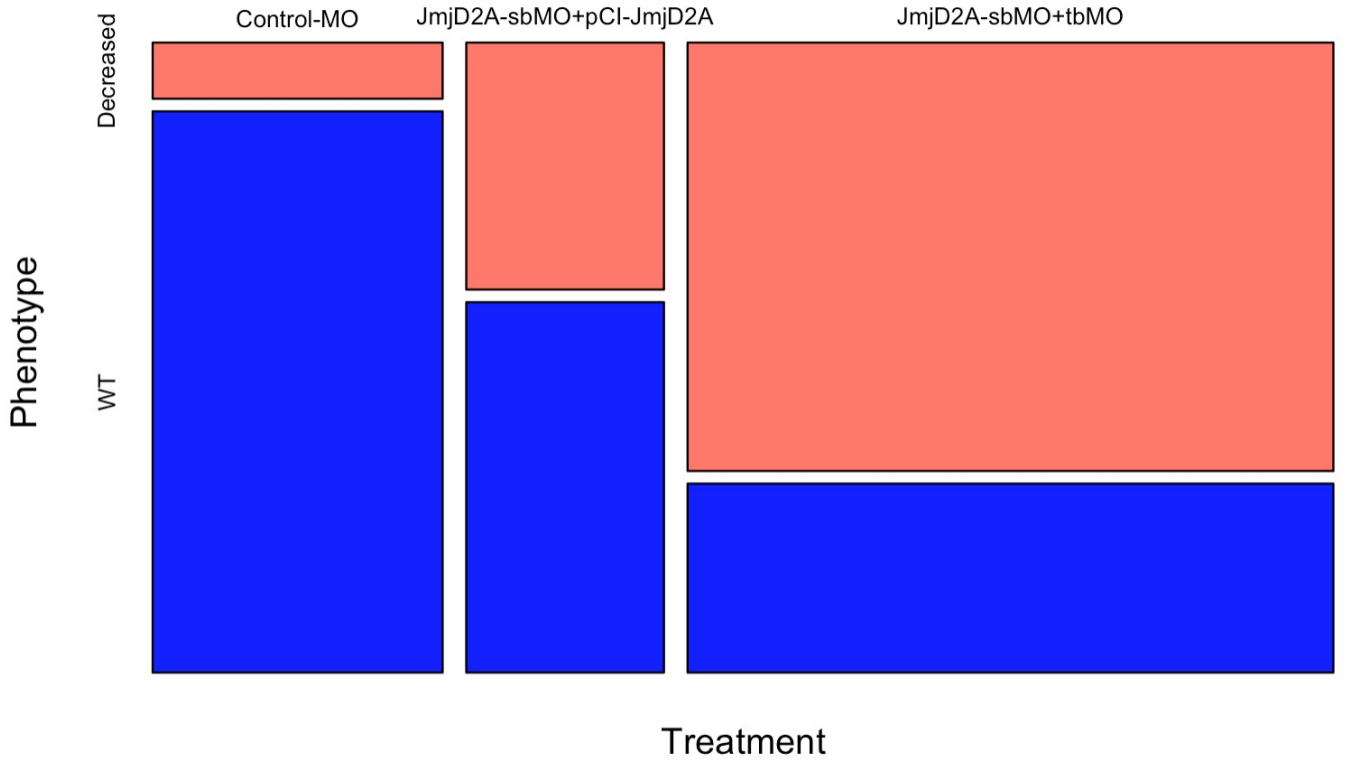
```
        Pearson's Chi-squared test

data:  data4.table
X-squared = 22.66, df = 2, p-value = 1.201e-05
```



Experiment Design: RNAi

- Measuring expression of Sox2
  - Treatment with 3 groups
  - Categorical
- Response variable
  - Categorical
- Meets model assumptions?
  - Data are independent
  - <20% of expected values are < 5
- Statistical Test:
  - Chi-Square test
    - Observed frequency
    - Expected frequency

```
        Pearson's Chi-squared test

data:  data4.table
X-squared = 22.66, df = 2, p-value = 1.201e-05
```

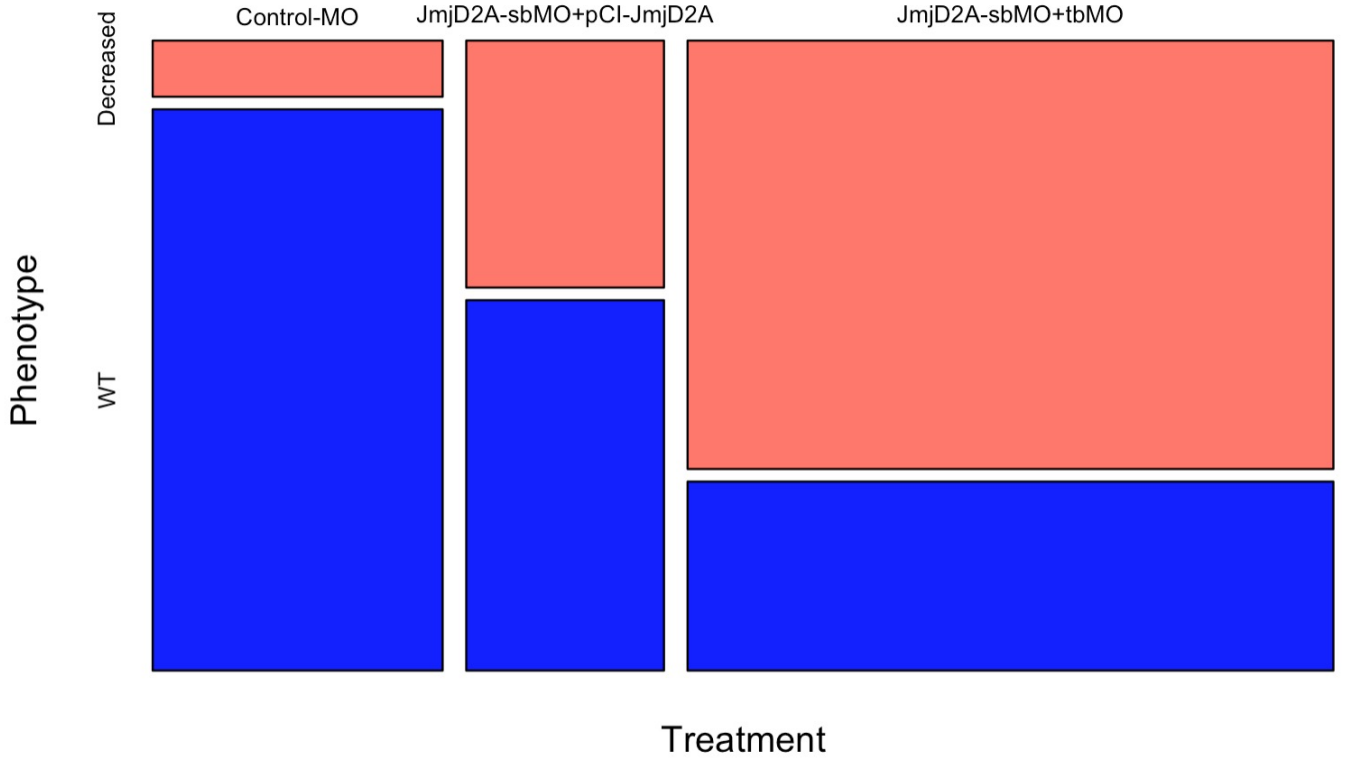


For which treatments does the response vary?

Experiment Design: RNAi

- Measuring expression of Sox2
  - Treatment with 3 groups
  - Categorical
- Response variable
  - Categorical
- Meets model assumptions?
  - Data are independent
  - <20% of expected values are < 5
- Statistical Test:
  - Chi-Square test
    - Observed frequency
    - Expected frequency

```
        Pearson's Chi-squared test

data:  data4.table
X-squared = 22.66, df = 2, p-value = 1.201e-05
```



Alternative test:
Fisher's exact test—no assumptions for minimum expected values

Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

Objectives:

- Basic statistics principles
- Describing data
- Hypothesis testing
- Analytical tools
- Data management and organization

Why does it matter how you record your data?

- Statistical tools require data to be stored in specific file types and with specific formats
- Projects take a long time. If you don't manage data well, it's possible to forget important components of the experiment

Organizing Data in Spreadsheets:

- Be consistent
- Choose good names for things
- Format dates YYYY-MM-DD or YYYYMMDD
- Don't include empty cells
- Put only one thing in a cell
- Don't do calculations in raw data files
- Maintain a data dictionary
- Don't use color, font formatting, highlighting as data
- Make backups
- Save data as plain text files

Common Messy Data Problems:

- Column headers are values instead of variable names
  - Fix by melting (converting format from wide to long)
- Multiple variables in one column
- Variables stored in rows and columns

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—

id	date	element	value
MX17004	2010-01-30	tmax	27.8
MX17004	2010-01-30	tmin	14.5
MX17004	2010-02-02	tmax	27.3
MX17004	2010-02-02	tmin	14.4
MX17004	2010-02-03	tmax	24.1
MX17004	2010-02-03	tmin	14.4
MX17004	2010-02-11	tmax	29.7
MX17004	2010-02-11	tmin	13.4
MX17004	2010-02-23	tmax	29.9
MX17004	2010-02-23	tmin	10.7

(a) Molten data

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

(b) Tidy data