Dipartimento di Biologia
e Biotecnologie
Charles Darwin

Sapienza
Università di Roma

Statistics for Biologists

16th -17th June 2025

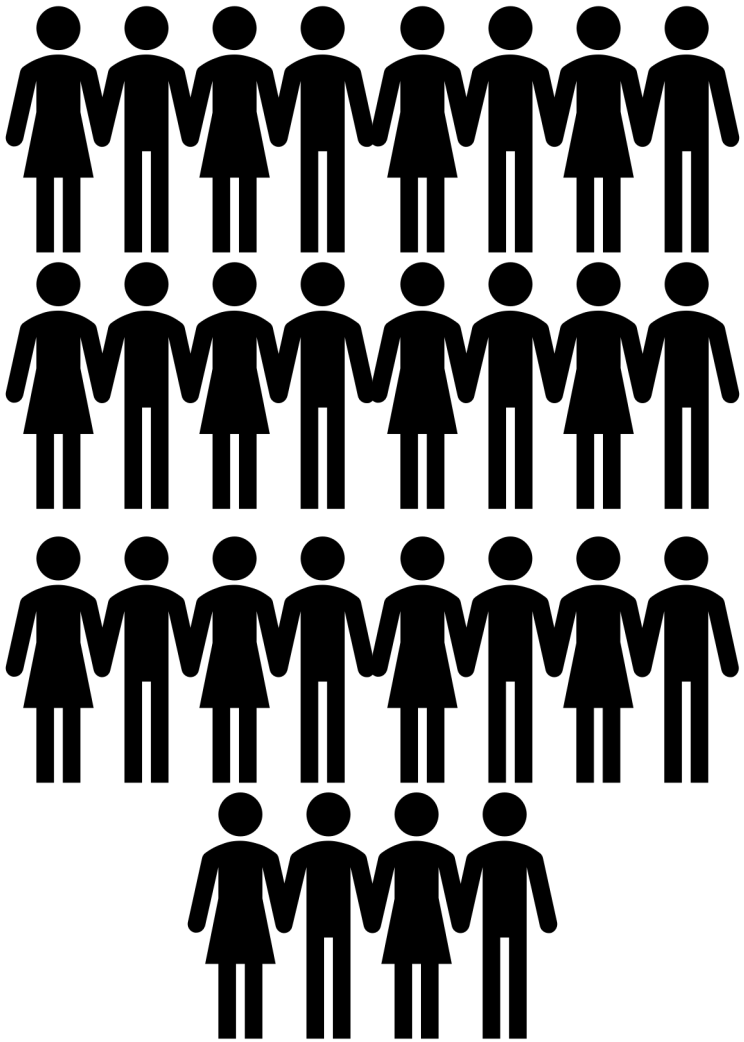
Città universitaria
Ed. Fisiologia generale e Antropologia CU026
E01PS1L101 Aula I multimediale

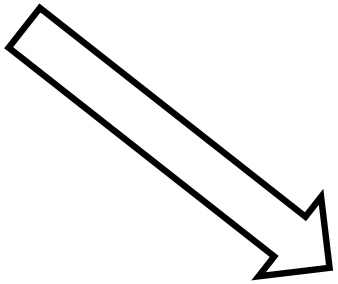
Dr. Mario Fordellone, PhD
Mail: mario.fordellone@unicampania.it
Dipartimento di Medicina Preventiva
Università degli Studi della Campania
"Luigi Vanvitelli"

Unità di
Statistica Medica

Statistical inference

Fundamental ingredients

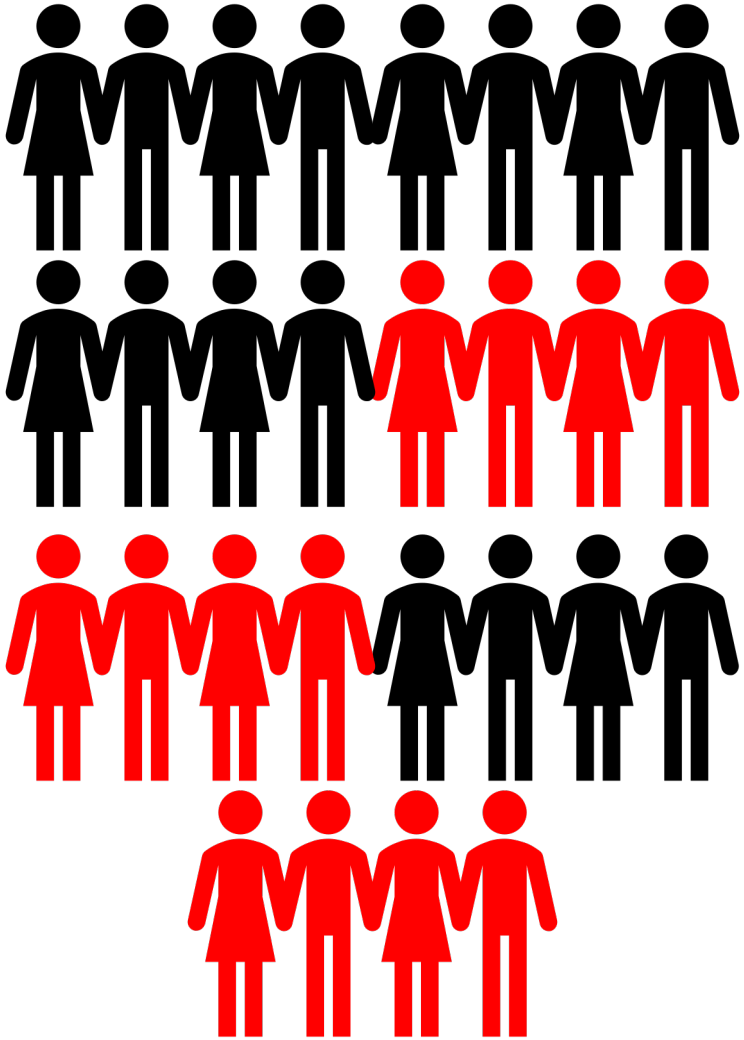




Population

Statistical inference

Fundamental ingredients



Sample

Statistical inference

Fundamental ingredients

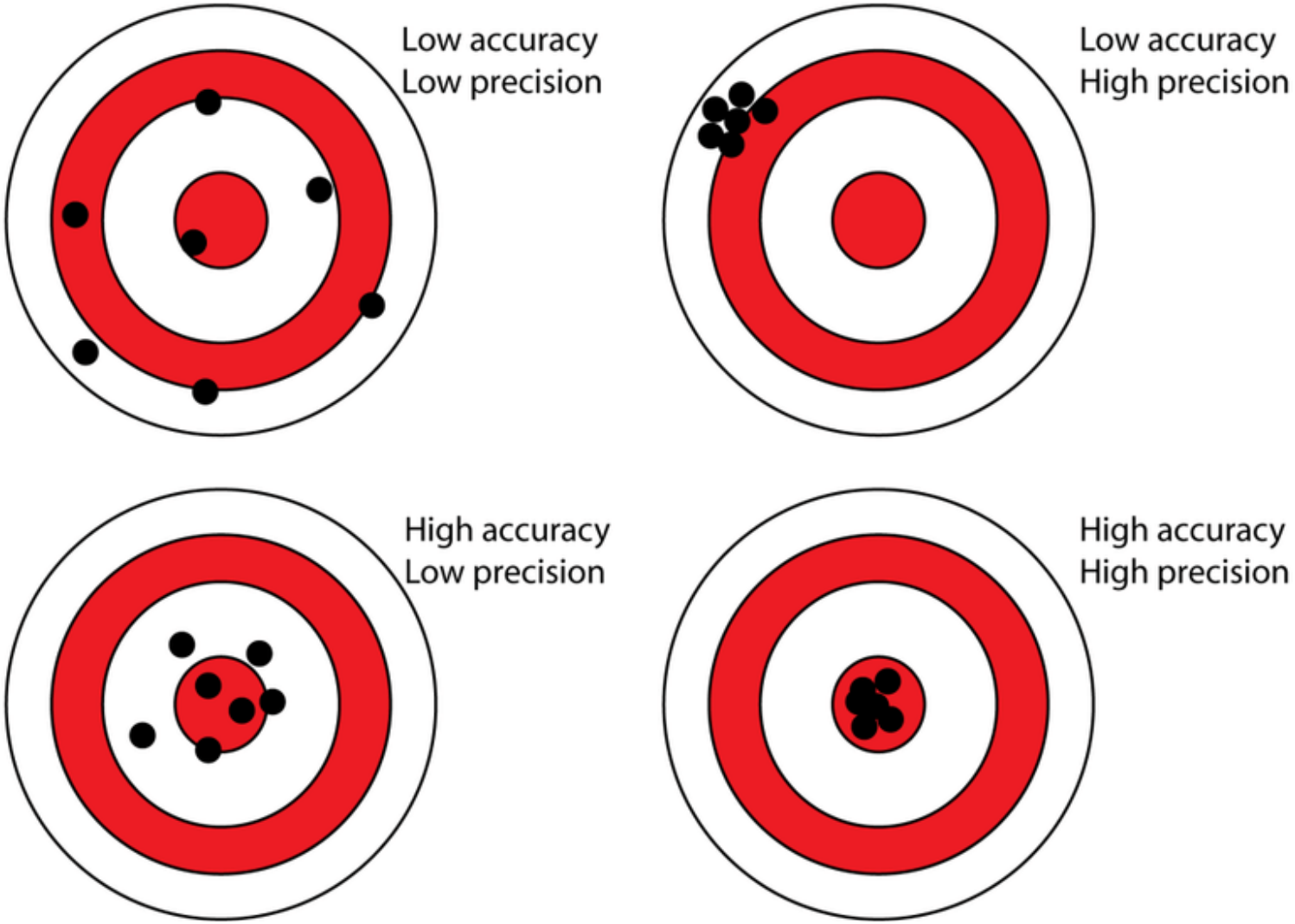
- Point estimation: find a number value (estimate) for the unknown population parameter based on the sample information.

- Interval estimation: find an interval of values (estimate) for the unknown population parameter based on the sample information.

- Hypothesis testing: make a statement about the population based on two complementary hypothesis.

Point estimation

Properties



Point estimation

Properties

The optimal estimator is a random variable with expected value E[x] equal to the parameter to be estimated (unbiased estimator) and variance inversely related to the sample size with the minimum value (minimum-variance estimator).

A good estimator minimizes the mean squared error (MSE):

$$MSE(t) = E[t - \theta]^2 + Var[t]$$

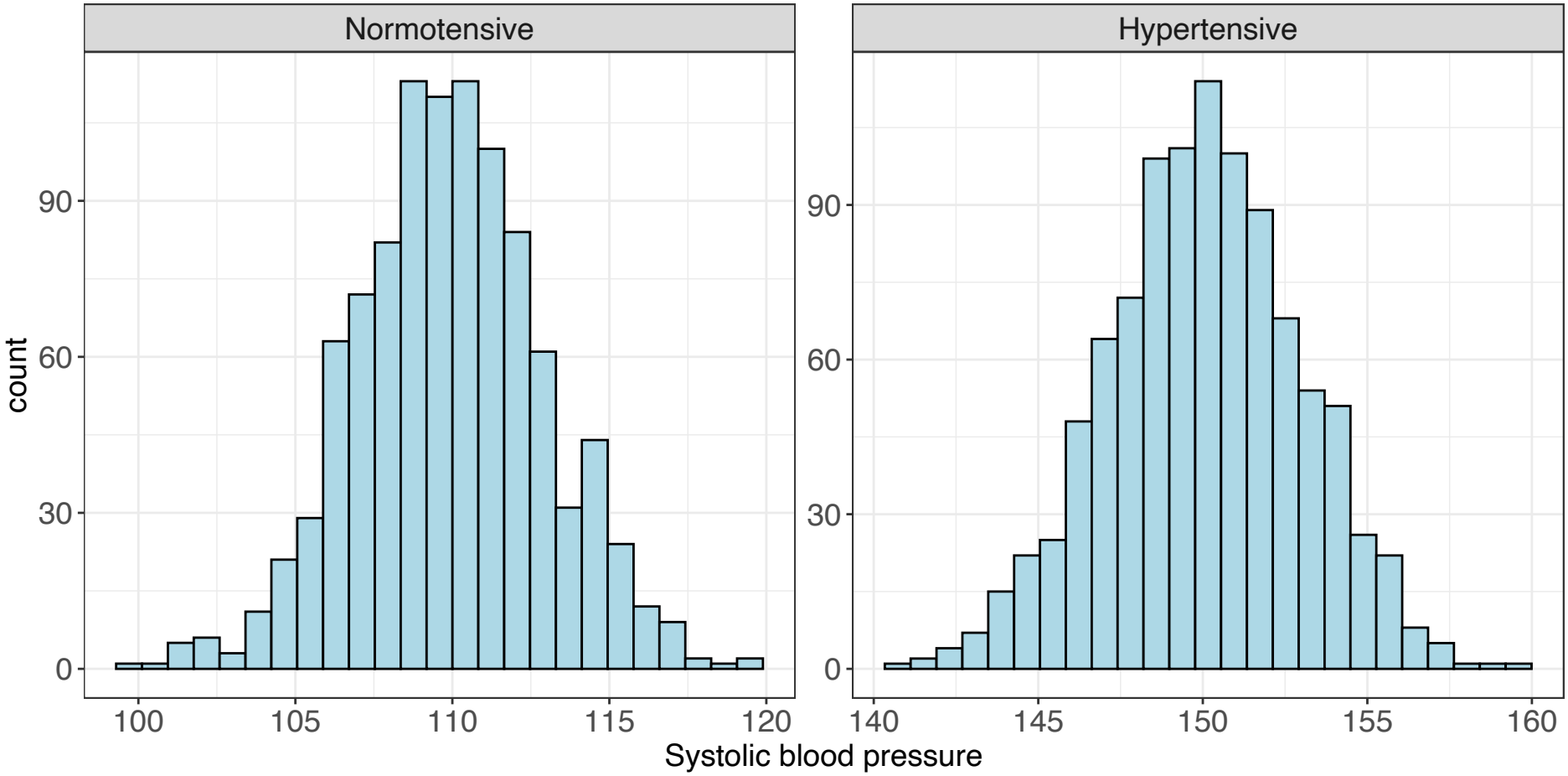
Point estimation

What are the estimators?

POPULATION PARAMETER	ESTIMATOR	MOMENTS
$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$E[\bar{X}] = \mu$ $Var[\bar{X}] = \frac{S^2}{n}$
$\pi = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$E[\hat{\pi}] = \pi$ $Var[\hat{\pi}] = \frac{\pi(1-\pi)}{n}$
$\mu_1 - \mu_2$ (equal variance, non-paired samples)	$\bar{X}_1 - \bar{X}_2$	$E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$ $Var[\bar{X}_1 - \bar{X}_2] = \frac{S^2}{n_1} + \frac{S^2}{n_2}$
$\mu_1 - \mu_2$ (unequal variance, non-paired samples)	$\bar{X}_1 - \bar{X}_2$	$E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$ $Var[\bar{X}_1 - \bar{X}_2] = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$
$\mu_{PRE} - \mu_{POST}$ (paired samples)	$\bar{d} = \bar{X}_{PRE} - \bar{X}_{POST}$	$E[\bar{d}] = \mu_{PRE} - \mu_{POST}$ $Var[\bar{d}] = \frac{S_d^2}{n}$
$\pi_1 - \pi_2$	$\hat{\pi}_1 - \hat{\pi}_2$	$E[\hat{\pi}_1 - \hat{\pi}_2] = \pi_1 - \pi_2$ $Var[\hat{\pi}_1 - \hat{\pi}_2] = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$

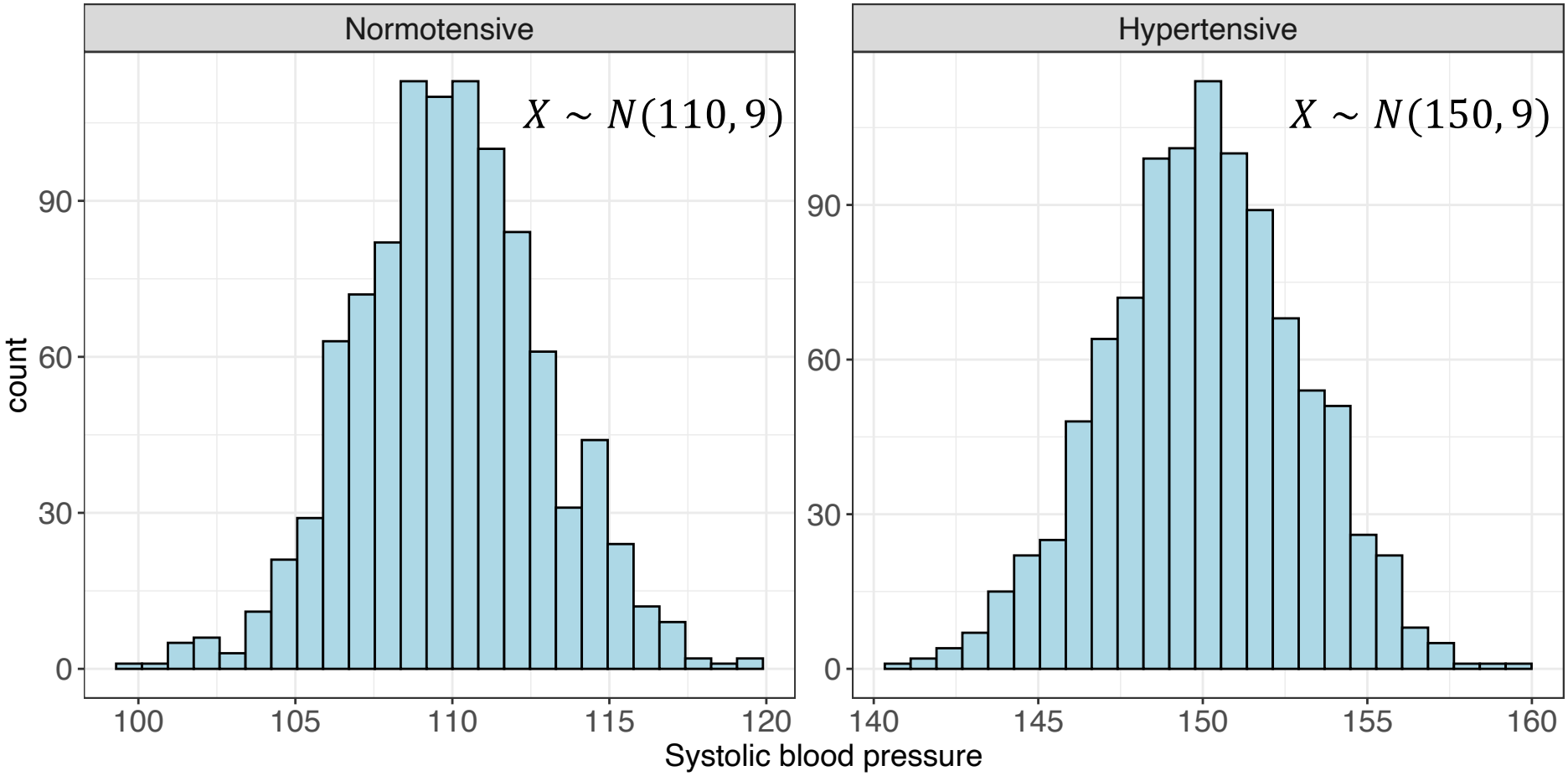
Point estimation

Sample mean estimator



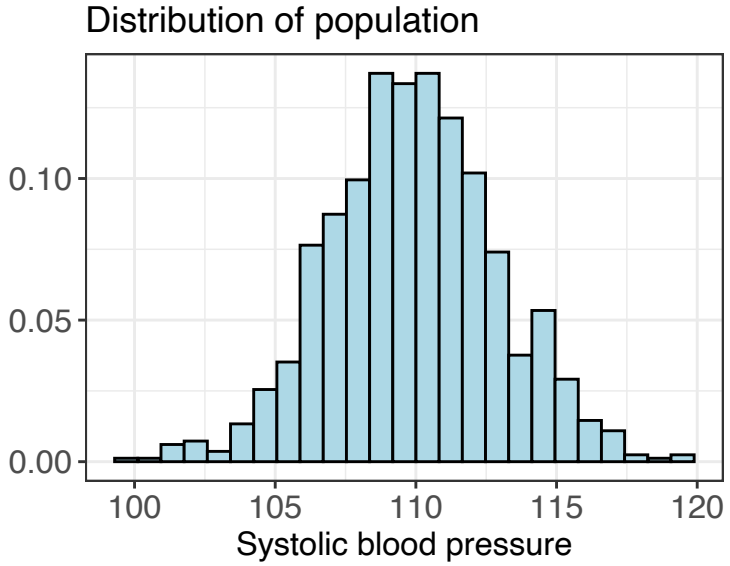
Point estimation

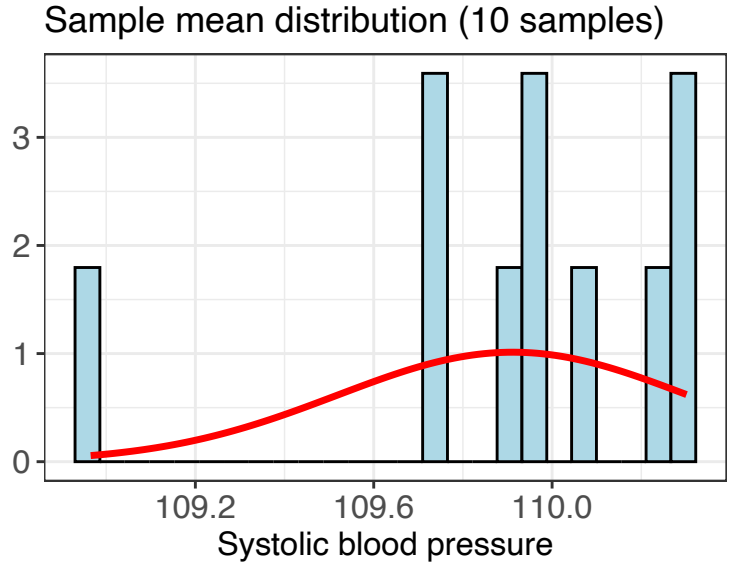
Sample mean estimator

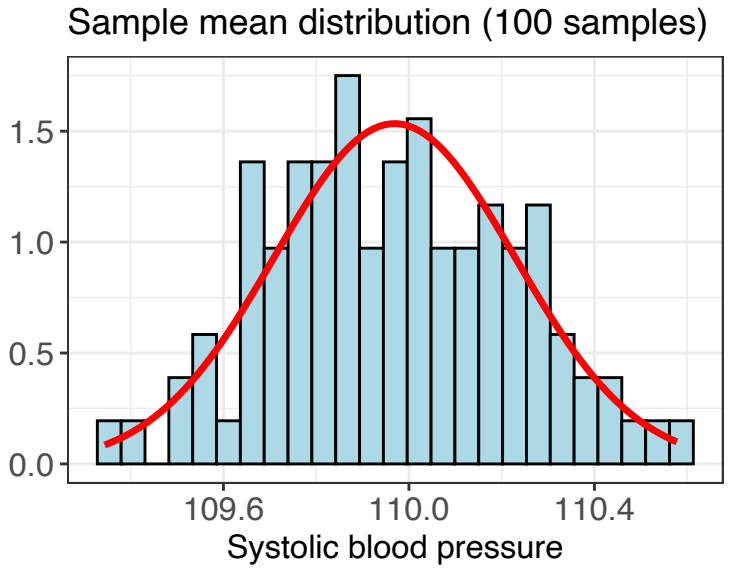


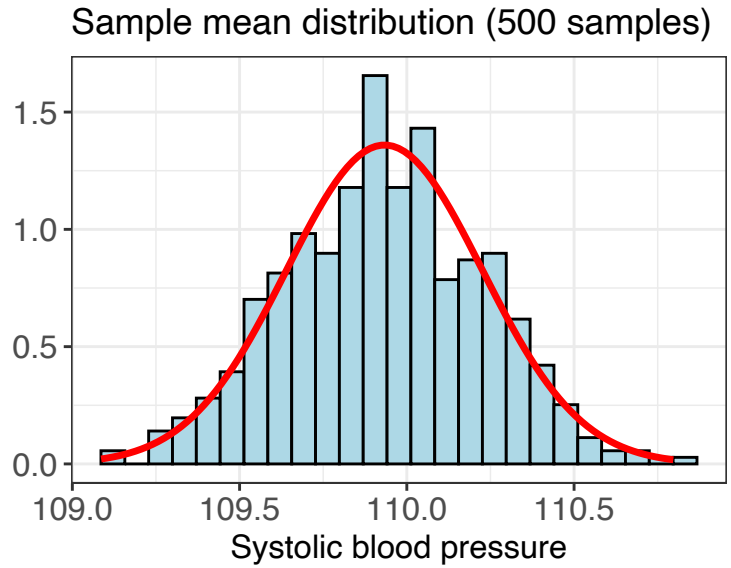
Point estimation

Sample mean distribution



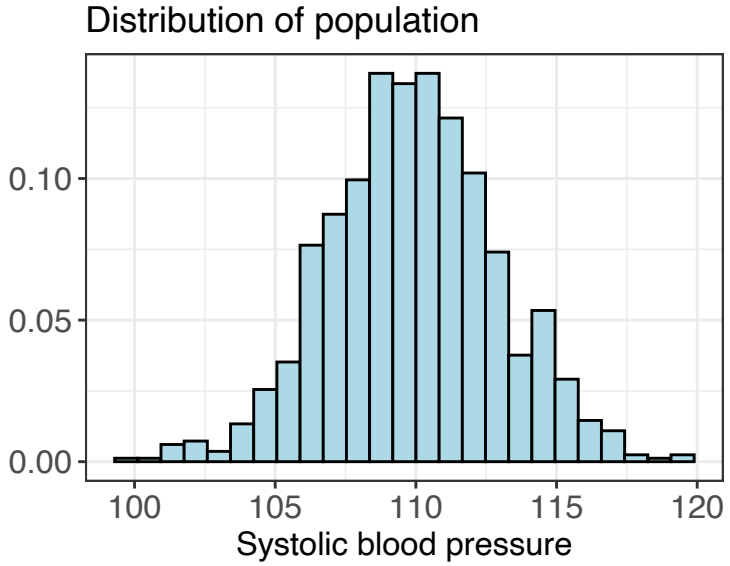


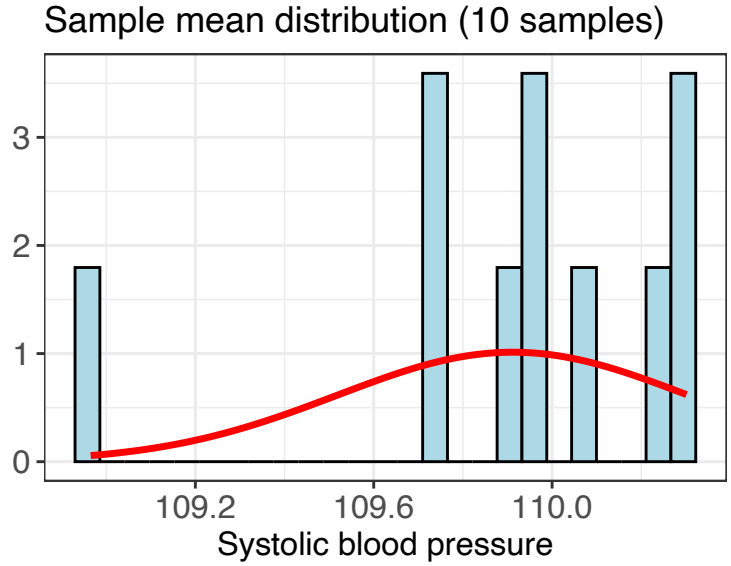


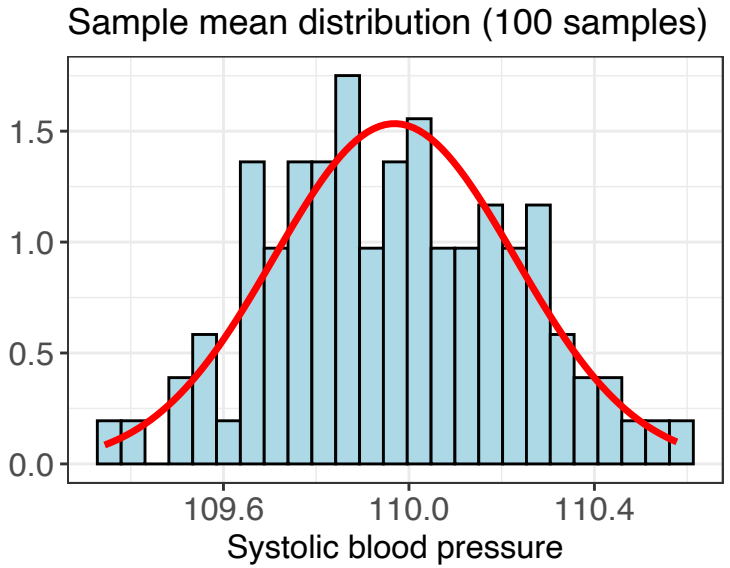


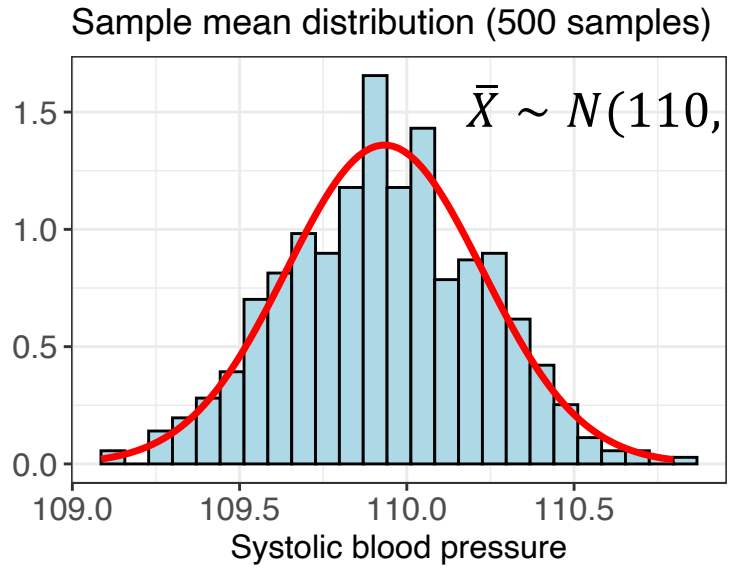
Point estimation

Sample mean distribution



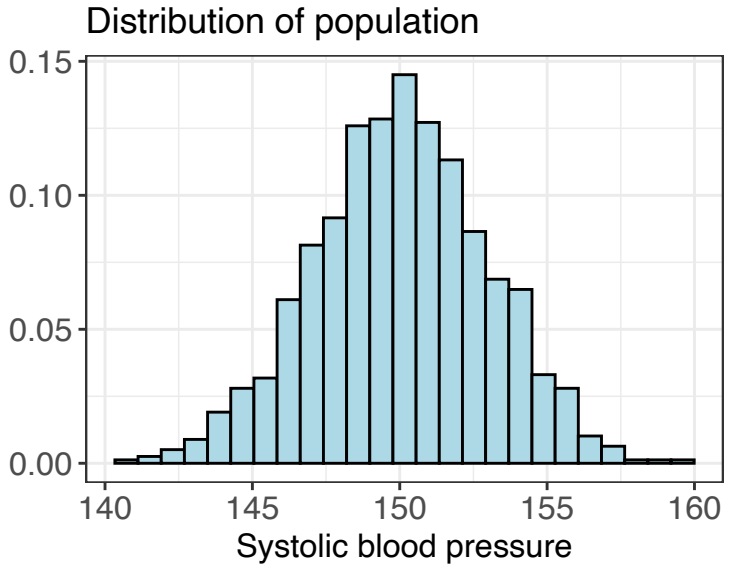


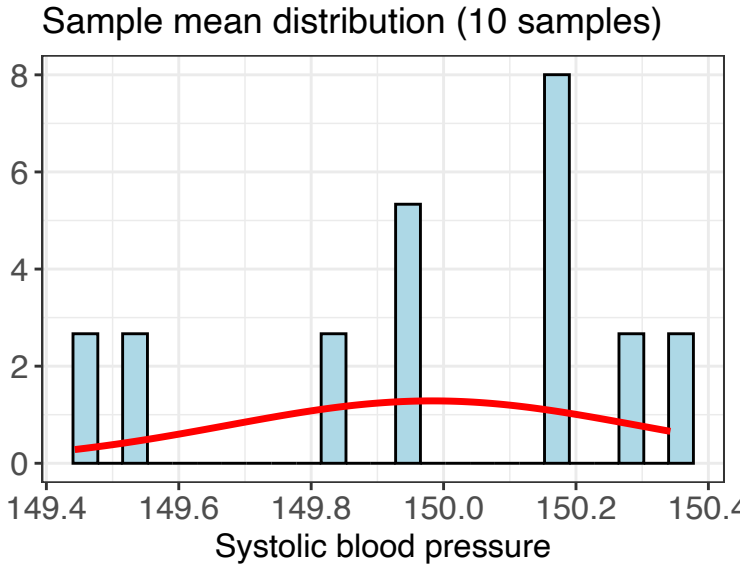


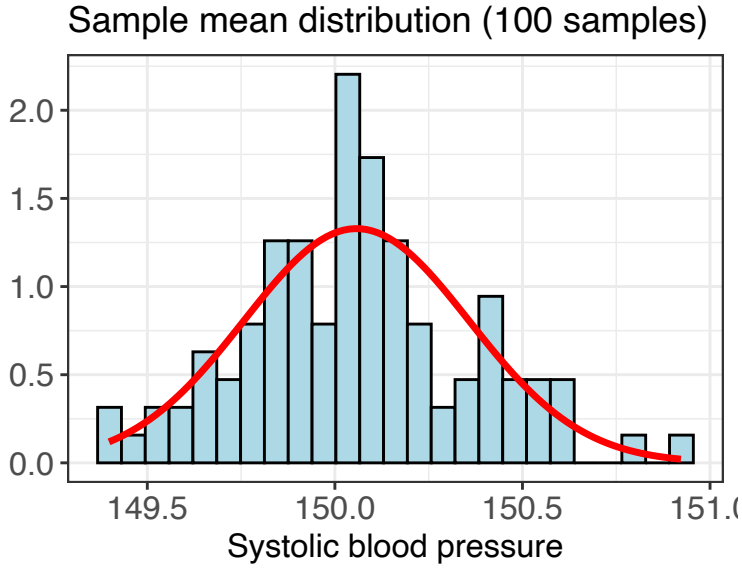


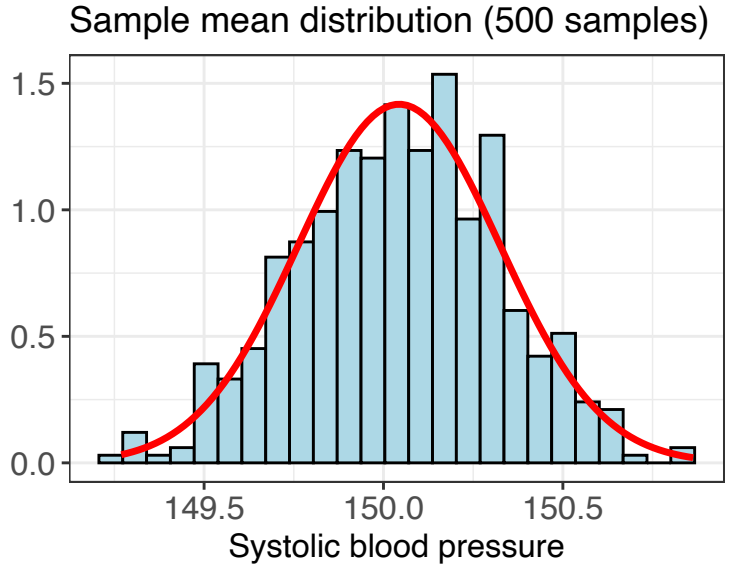
Point estimation

Sample mean distribution



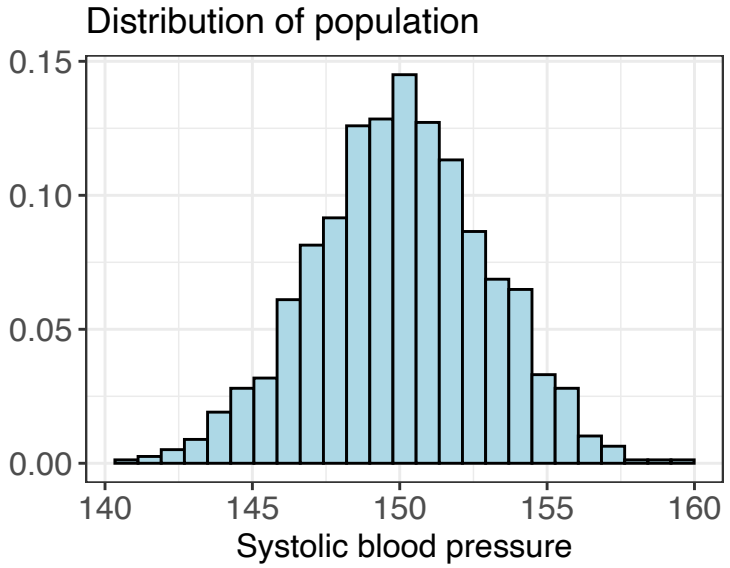


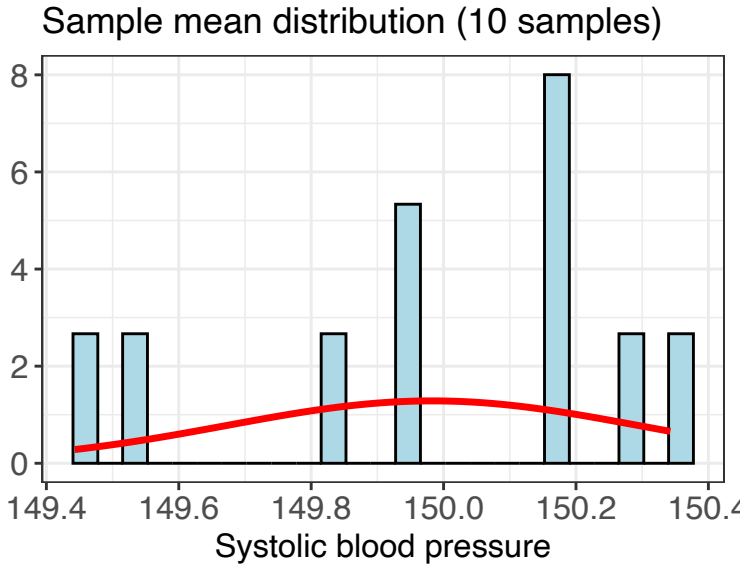


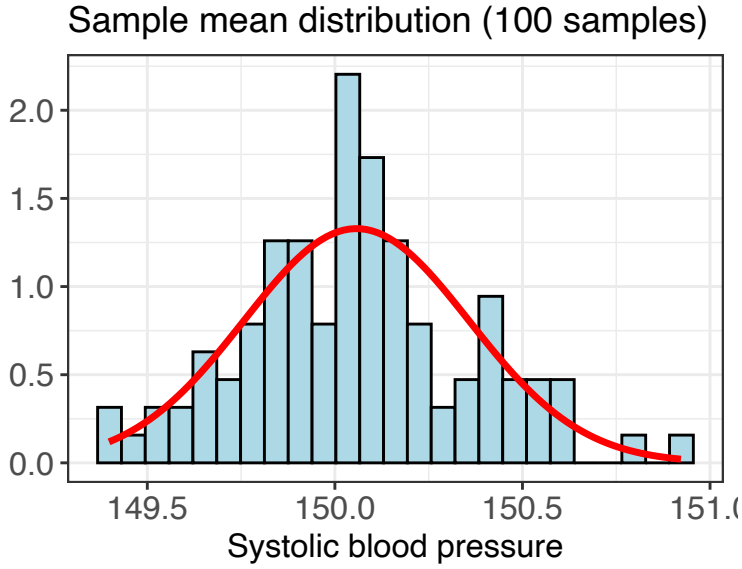


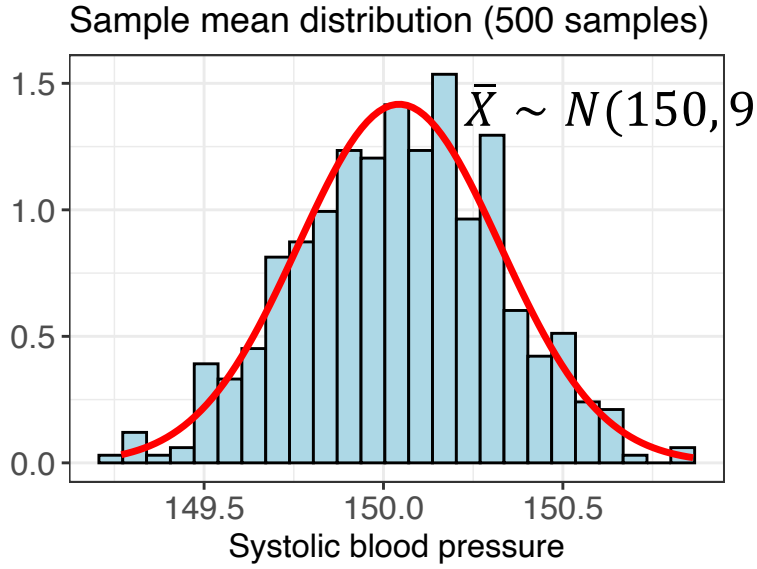
Point estimation

Sample mean distribution







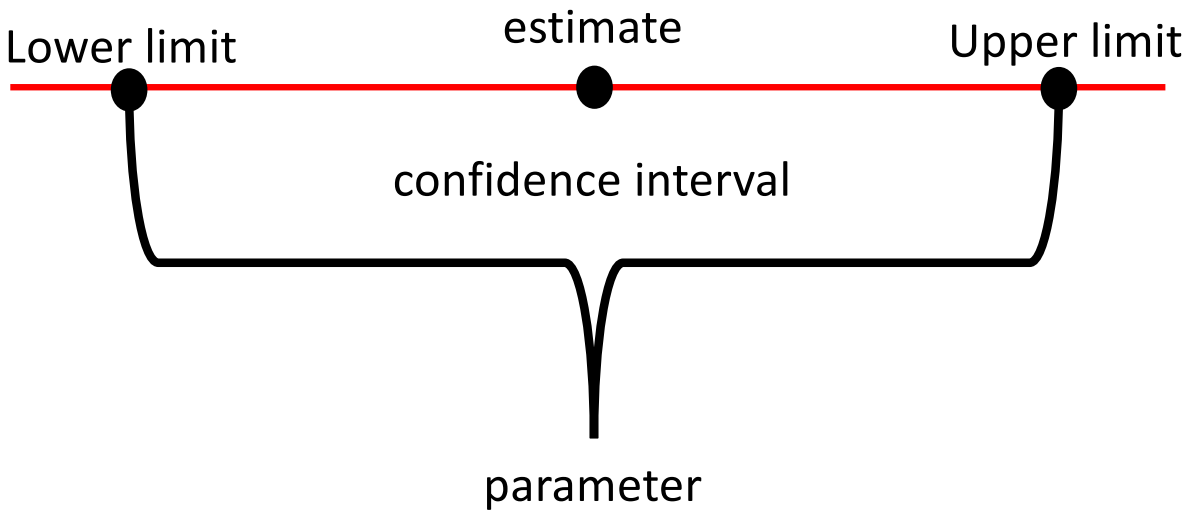


$\bar{X} \sim N(150, 9/100)$

Interval estimation

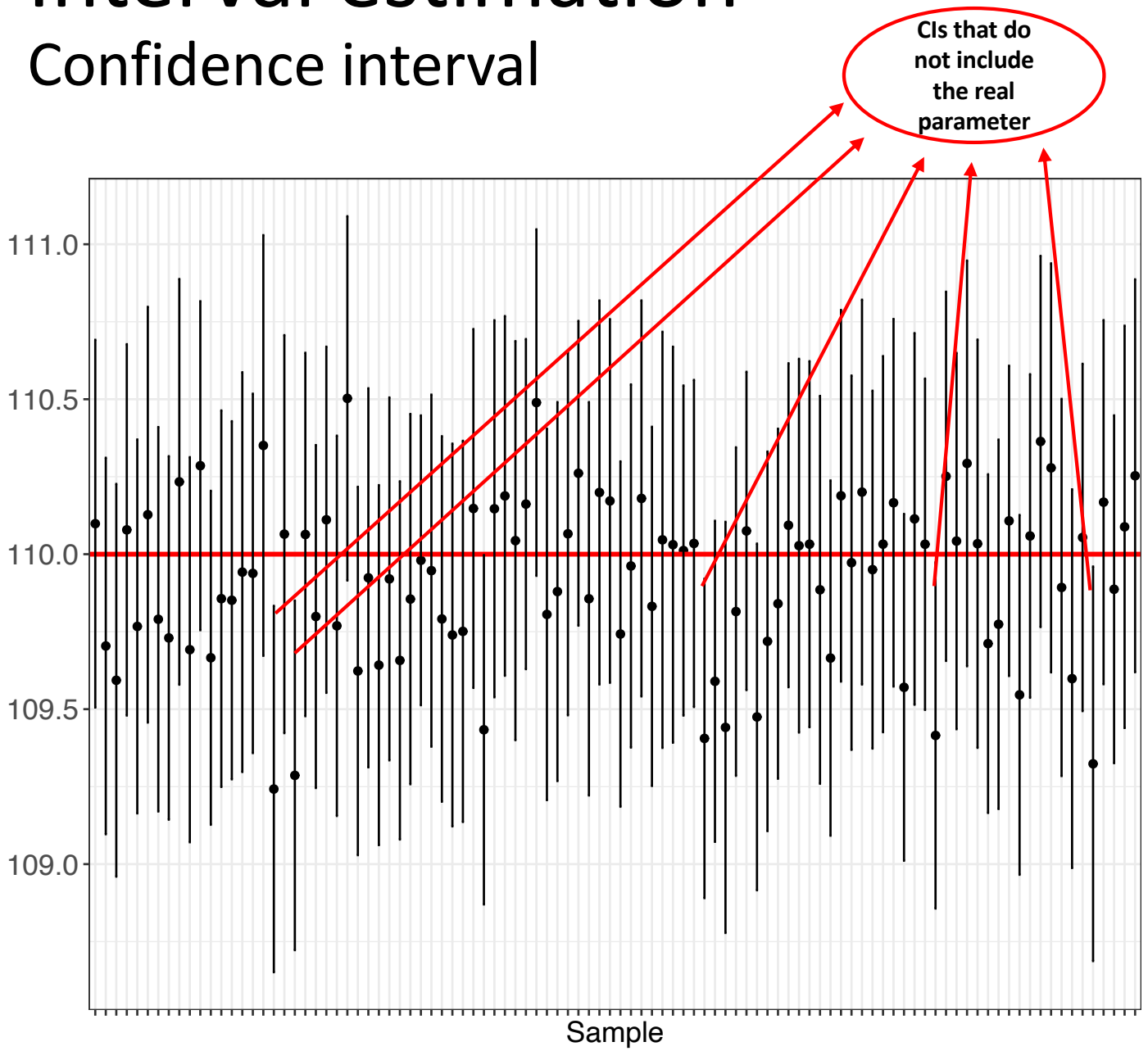
Confidence interval

Through the construction of a Confidence Interval, an interval of values is identified within which the parameter of interest of the population falls, with a certain degree of confidence.



Interval estimation

Confidence interval

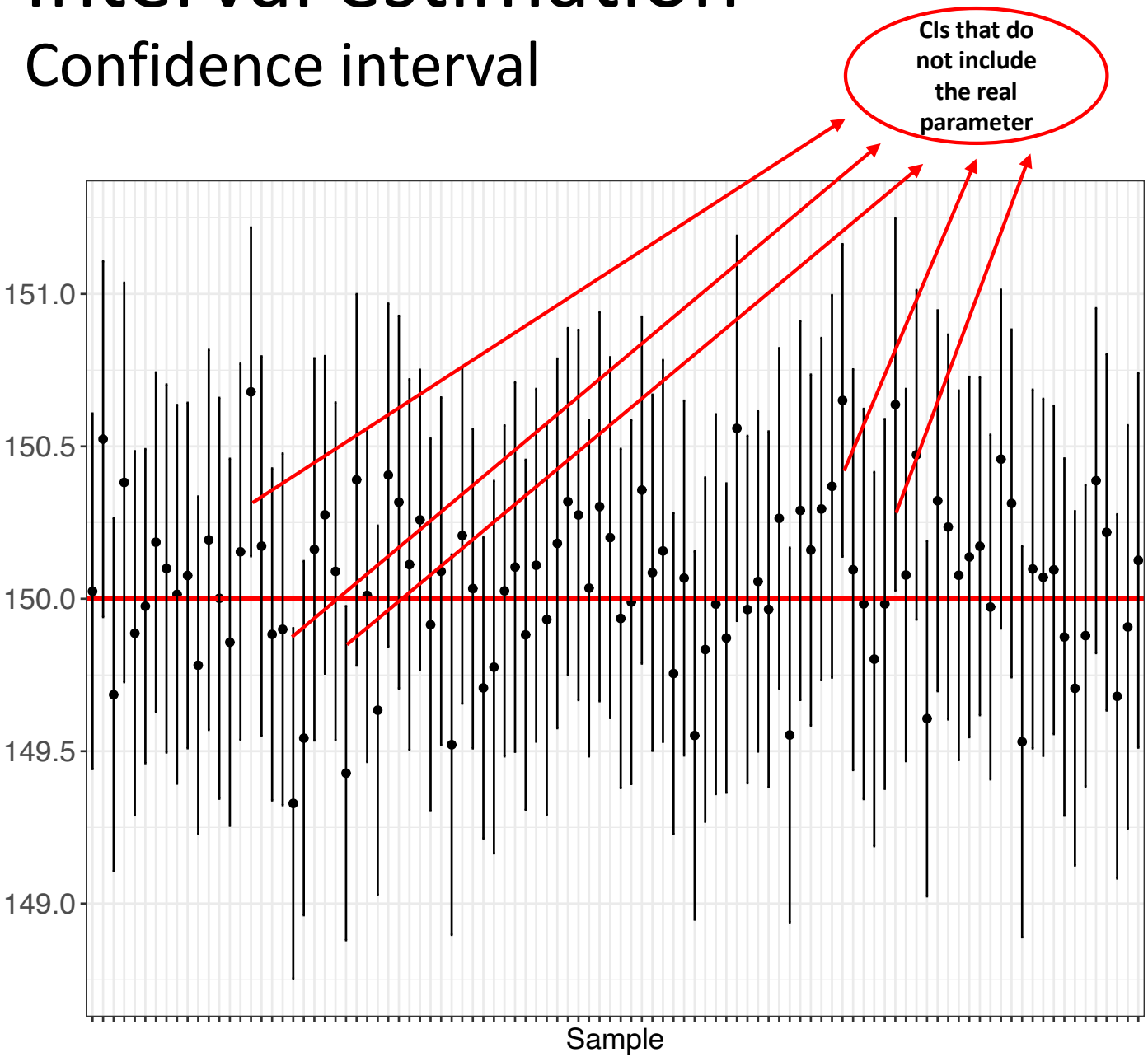


100 samples with size equal to 100.

5 confidence intervals at 95% on 100 do not include the real parameter of normotensive population (i.e., 110 mm/Hg)

Interval estimation

Confidence interval



100 samples with size equal to 100.

5 confidence intervals at 95% on 100 do not include the real parameter of hypertensive population (i.e., 150 mm/Hg)

Interval estimation

Confidence interval

From the asymptotic properties of the sample mean estimator, we know that:

$$\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

Then, the confidence interval at $1-\alpha$ *confidence degree is*

$$P\left(-Z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq +Z_{\alpha/2}\right) = 1-\alpha$$

Interval estimation

Confidence interval

From the asymptotic properties of the sample mean estimator, we know that:

$$\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

Then, the confidence interval at $1-\alpha$ *confidence degree is*

$$\text{C.I. } (1-\alpha) = \left[\bar{X} \pm Z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right]$$

Interval estimation

Confidence interval

From the asymptotic properties of the sample proportion estimator, we know that:

$$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$$

Then, the confidence interval at $1 - \alpha$ *confidence degree is*

$$P\left(-Z_{\alpha/2} \leq \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq +Z_{\alpha/2}\right) = 1 - \alpha$$

Interval estimation

Confidence interval

From the asymptotic properties of the sample proportion estimator, we know that:

$$\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$$

Then, the confidence interval at $1 - \alpha$ *confidence degree is*

$$\text{C.I. } (1-\alpha) = \left[\hat{\pi} \pm Z_{\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}}\right]$$

Interval estimation

Confidence interval

From the asymptotic properties of the sample difference of two mean (equal variance) estimator, we know that:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \sim N(0,1)$$

Then, the confidence interval at $1 - \alpha$ *confidence degree is*

$$P\left(-Z_{\alpha/2} \leq \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \leq +Z_{\alpha/2}\right) = 1 - \alpha$$

Interval estimation

Confidence interval

From the asymptotic properties of the sample difference of two mean (equal variance) estimator, we know that:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \sim N(0,1)$$

Then, the confidence interval at $1 - \alpha$ *confidence degree is*

$$\text{C.I. } (1-\alpha) = \left[(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \right]$$

Interval estimation

Confidence interval

From the asymptotic properties of the sample difference of two mean (not equal variance) estimator, we know that:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Then, the confidence interval at $1 - \alpha$ *confidence degree is*

$$P\left(-Z_{\alpha/2} \leq \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq +Z_{\alpha/2}\right) = 1 - \alpha$$

Interval estimation

Confidence interval

From the asymptotic properties of the sample difference of two mean (not equal variance) estimator, we know that:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Then, the confidence interval at $1 - \alpha$ *confidence degree is*

$$\text{C.I. } (1-\alpha) = \left[(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

Interval estimation

Confidence interval

From the asymptotic properties of the sample difference of two proportions estimator, we know that:

$$\frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$$

Then, the confidence interval at $1-\alpha$ *confidence degree is*

$$P\left(-Z_{\alpha/2} \leq \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \leq +Z_{\alpha/2}\right) = 1-\alpha$$

Interval estimation

Confidence interval

From the asymptotic properties of the sample difference of two proportions estimator, we know that:

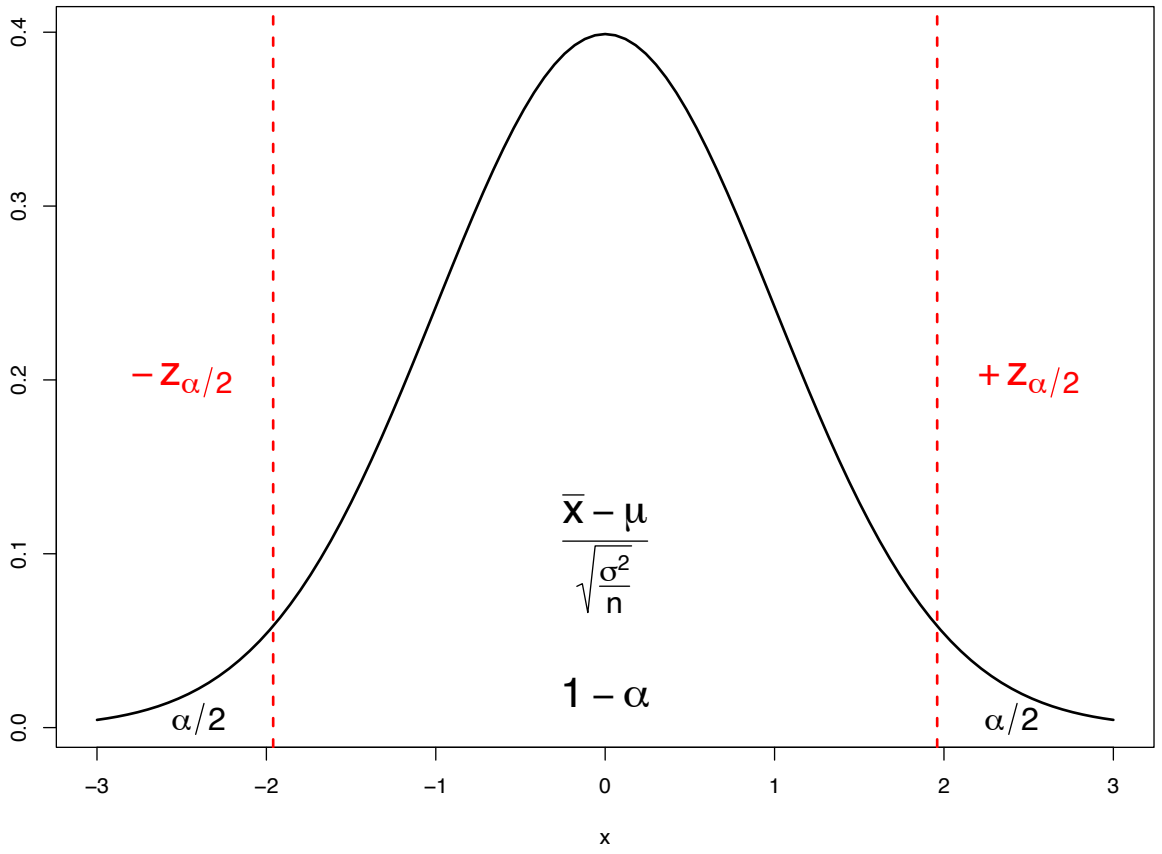
$$\frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$$

Then, the confidence interval at $1-\alpha$ confidence degree is

$$\text{C.I. } (1-\alpha) = \left[(\hat{\pi}_1 - \hat{\pi}_2) \pm Z_{\alpha/2} \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}} \right]$$

Interval estimation

Confidence interval



Interval estimation

Confidence interval

Notes:

Usually, the variance of population σ^2 is unknown. However, it can be estimated with the following correct estimator:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

In this case the observed statistic is not Normal distributed:

$$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)$$

Point and interval estimation

Three samples with size equal to 10, 25, and 50

NORMOTENSIVE POPULATION

Point and interval estimation

Three samples with size equal to 10, 25, and 50

NORMOTENSIVE POPULATION

SAMPLE 1

SAMPLE 2

SAMPLE 3

$n = 10$

$\bar{X} = 108.65$ mm/Hg

$SE(\bar{X}) = 2.69/\sqrt{10}$

$CI_{95\%} = [106.73, 110.58]$

Point and interval estimation

Three samples with size equal to 10, 25, and 50

NORMOTENSIVE POPULATION

SAMPLE 1

$n = 10$
$\bar{X} = 108.65$ mm/Hg
$SE(\bar{X}) = 2.69/\sqrt{10}$
$CI_{95\%} = [106.73, 110.58]$

SAMPLE 2

$n = 25$
$\bar{X} = 109.14$ mm/Hg
$SE(\bar{X}) = 3.79/\sqrt{25}$
$CI_{95\%} = [107.58, 110.71]$

SAMPLE 3

Point and interval estimation

Three samples with size equal to 10, 25, and 50

NORMOTENSIVE POPULATION

SAMPLE 1

$n = 10$
$\bar{X} = 108.65$ mm/Hg
$SE(\bar{X}) = 2.69/\sqrt{10}$
$CI_{95\%} = [106.73, 110.58]$

SAMPLE 2

$n = 25$
$\bar{X} = 109.14$ mm/Hg
$SE(\bar{X}) = 3.79/\sqrt{25}$
$CI_{95\%} = [107.58, 110.71]$

SAMPLE 3

$n = 50$
$\bar{X} = 110.48$ mm/Hg
$SE(\bar{X}) = 2.56/\sqrt{50}$
$CI_{95\%} = [109.75, 111.20]$

Point and interval estimation

Three samples with size equal to 10, 25, and 50

Point and interval estimation

Three samples with size equal to 10, 25, and 50

HYPERTENSIVE POPULATION

SAMPLE 1

SAMPLE 2

SAMPLE 3

$n = 10$

$\bar{X} = 150.01$ mm/Hg

$SE(\bar{X}) = 2.45/\sqrt{10}$

$CI_{95\%} = [148.25, 151.76]$

Point and interval estimation

Three samples with size equal to 10, 25, and 50

HYPERTENSIVE POPULATION

SAMPLE 1

$n = 10$
$\bar{X} = 150.01$ mm/Hg
$SE(\bar{X}) = 2.45/\sqrt{10}$
$CI_{95\%} = [148.25, 151.76]$

SAMPLE 2

$n = 25$
$\bar{X} = 150.36$ mm/Hg
$SE(\bar{X}) = 3.18/\sqrt{25}$
$CI_{95\%} = [149.05, 151.68]$

SAMPLE 3

Point and interval estimation

Three samples with size equal to 10, 25, and 50

HYPERTENSIVE POPULATION

SAMPLE 1

$n = 10$
$\bar{X} = 150.01$ mm/Hg
$SE(\bar{X}) = 2.45/\sqrt{10}$
$CI_{95\%} = [148.25, 151.76]$

SAMPLE 2

$n = 25$
$\bar{X} = 150.36$ mm/Hg
$SE(\bar{X}) = 3.18/\sqrt{25}$
$CI_{95\%} = [149.05, 151.68]$

SAMPLE 3

$n = 50$
$\bar{X} = 150.03$ mm/Hg
$SE(\bar{X}) = 3.01/\sqrt{50}$
$CI_{95\%} = [149.17, 150.89]$

Hypothesis testing

Theoretical framework

Decisional scheme:

		Conclusion about null hypothesis from statistical test	
		Accept Null	Reject Null
Truth about null hypothesis in population	True	Correct	Type I error Observe difference when none exists
	False	Type II error Fail to observe difference when one exists	Correct

Hypothesis testing
Theoretical framework

Sample size:

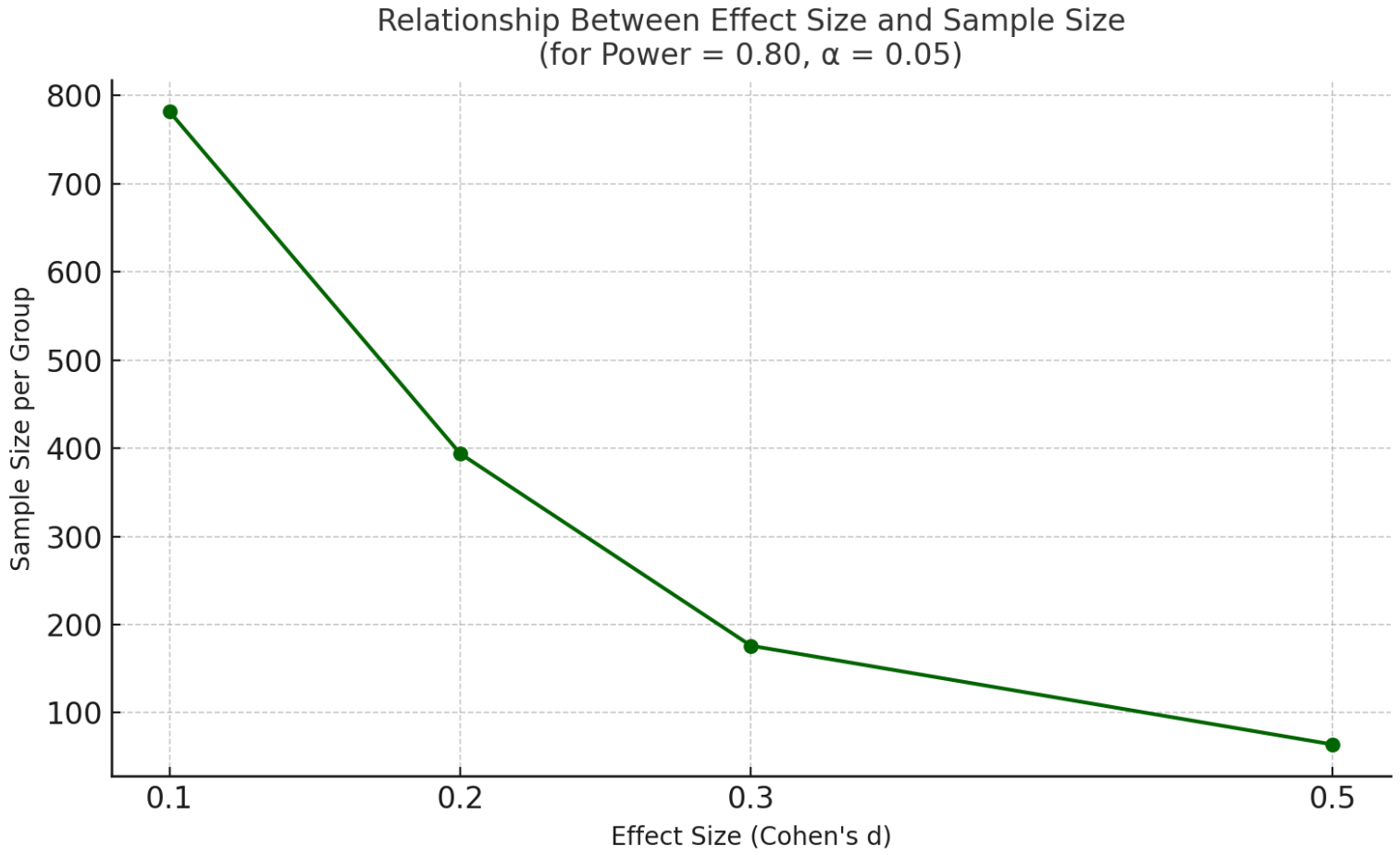
The sample size is a fundamental parameter in the design of a statistical study. It directly influences the statistical power, which is the probability of detecting a true effect when it exists, thereby minimizing the risk of a Type II error (false negative).

Power is commonly set at 0.80, indicating a 20% chance of failing to detect a real effect. Another key factor is the effect size, which quantifies the magnitude of the expected difference or association.

Small effect sizes (e.g., Cohen's d = 0.2 or r = 0.1) require substantially larger sample sizes compared to medium or large effects.

Hypothesis testing

Theoretical framework



Hypothesis testing
Theoretical framework

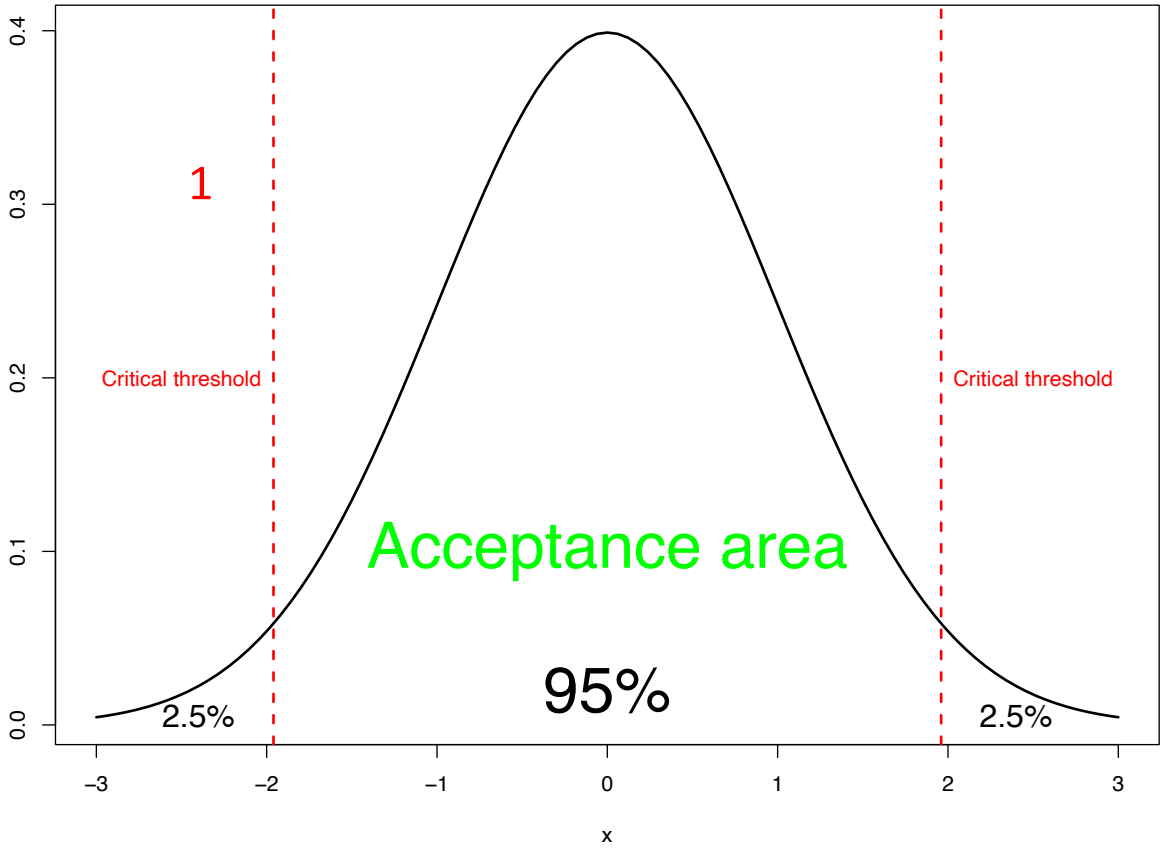
P-value:

In statistical inference, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct.

The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

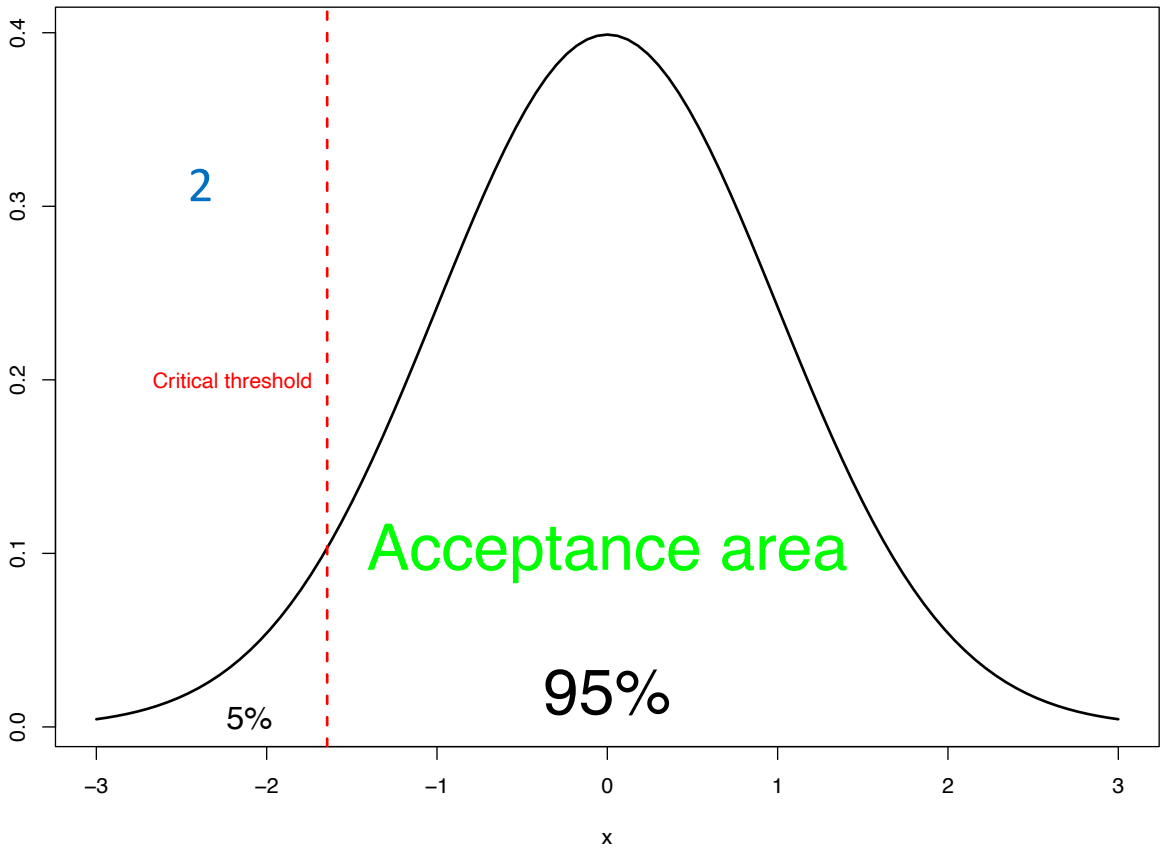
Hypothesis testing

Theoretical framework



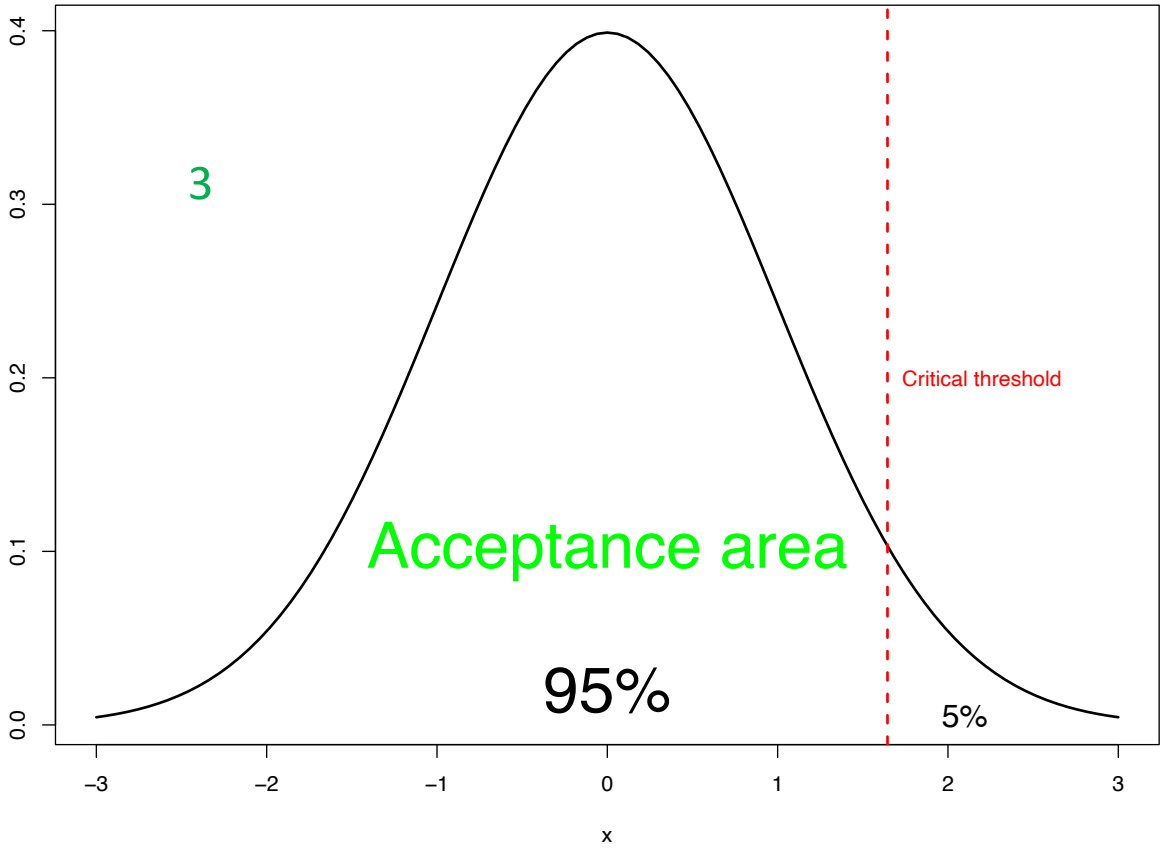
Hypothesis testing

Theoretical framework



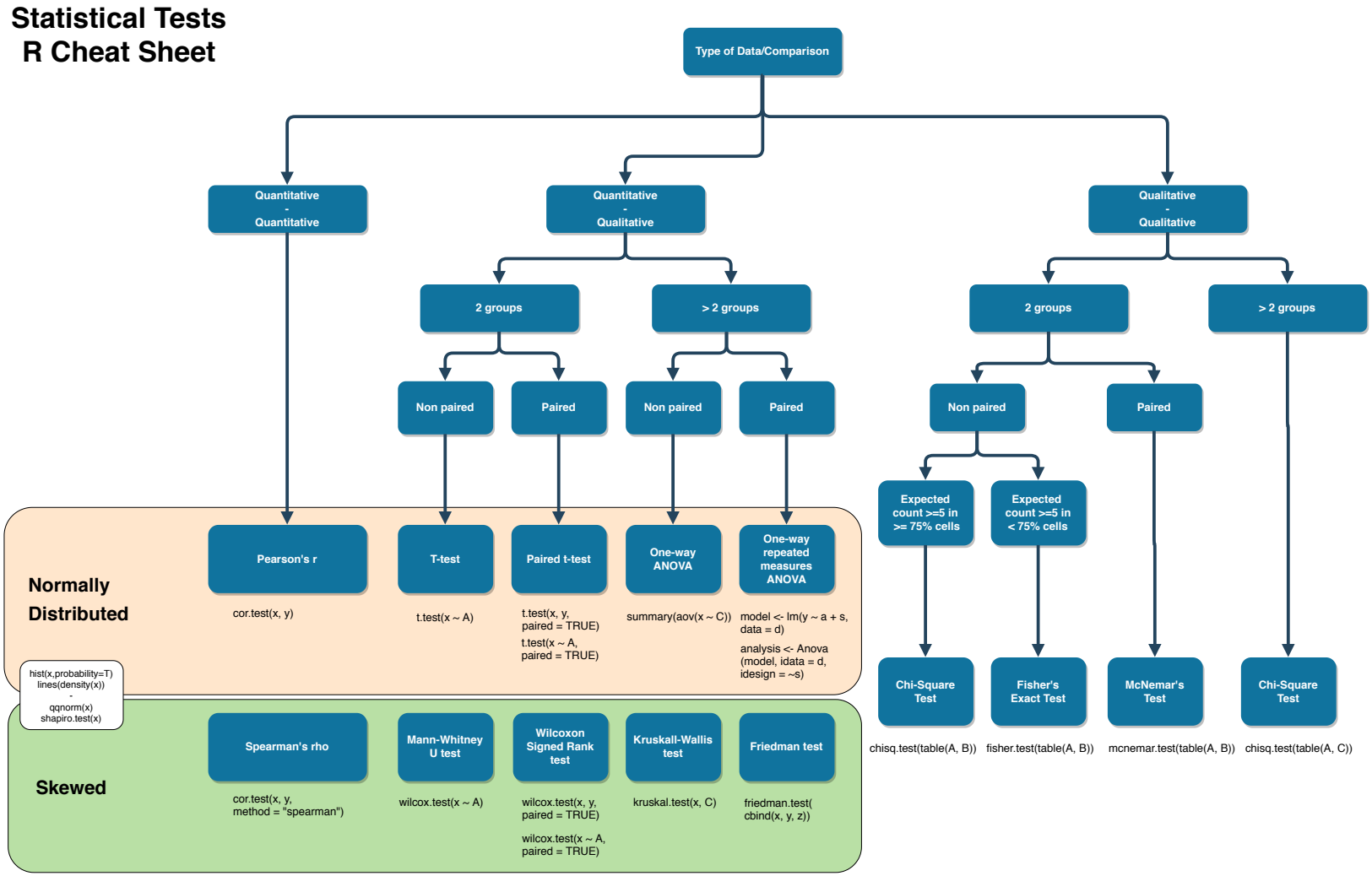
Hypothesis testing

Theoretical framework



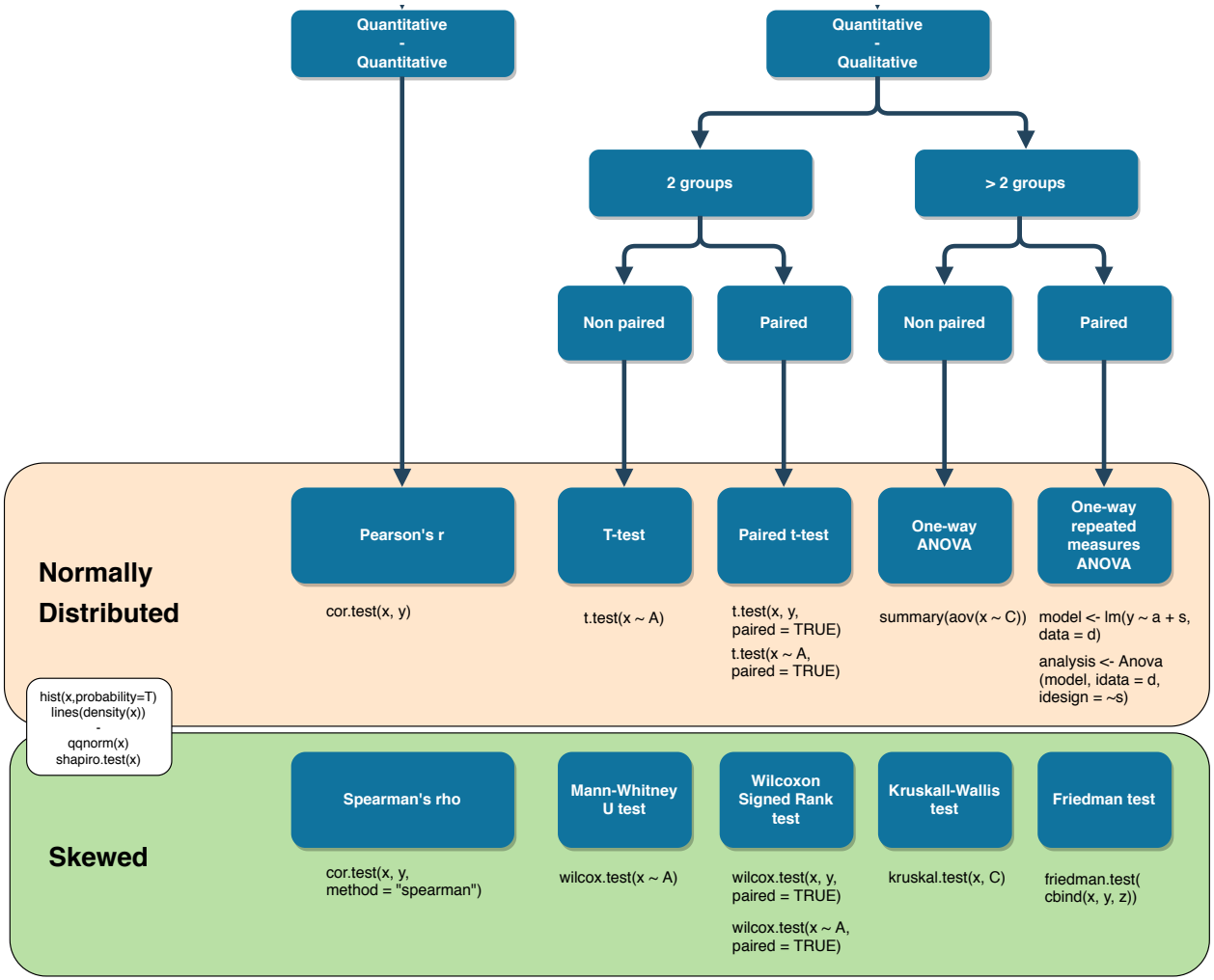
Hypothesis testing

Group comparison



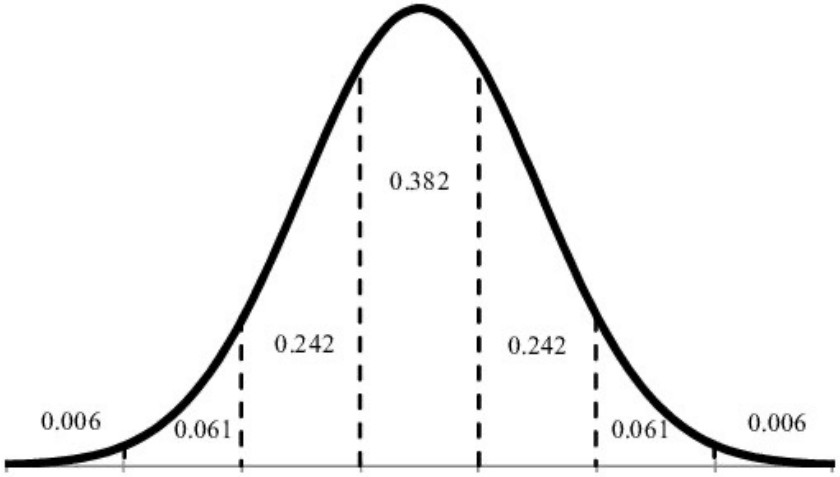
Hypothesis testing

Group comparison

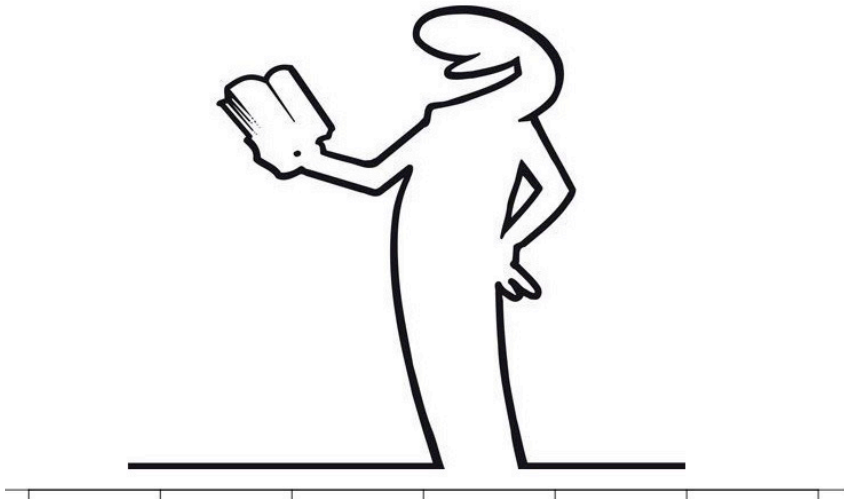


Hypothesis testing

Parametric or non-parametric?



NORMAL DISTRIBUTION



ABNORMAL DISTRIBUTION

Hypothesis testing

Parametric or non-parametric?

Shapiro-Wilk normality test:

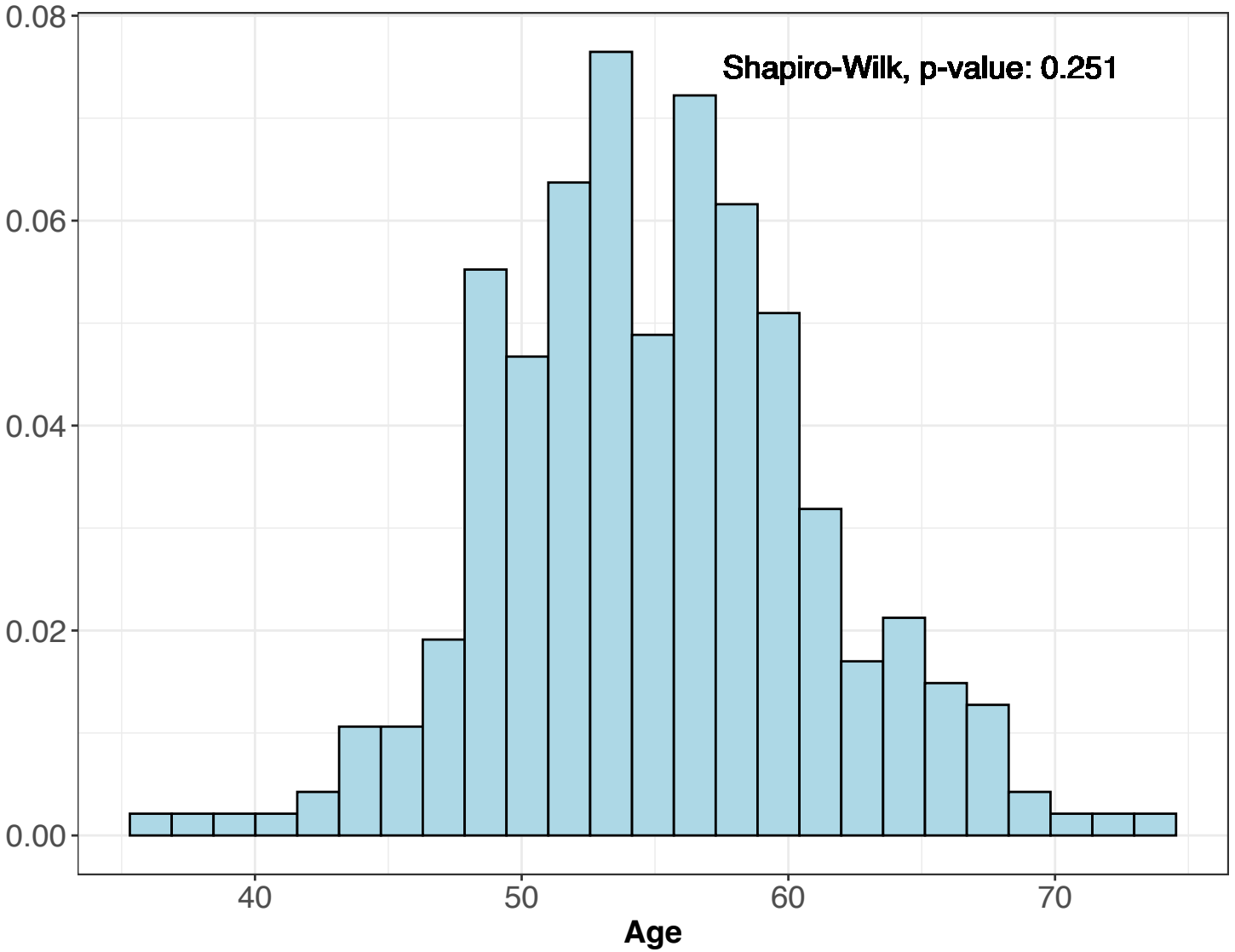
The Shapiro–Wilk test is a test of normality in frequentist statistics. It was published in 1965 by Samuel Sanford Shapiro and Martin Wilk.

The null-hypothesis of this test is that the population is normally distributed. Thus, if the p-value is lesser than the chosen alpha level (e.g., 0.05), then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed.

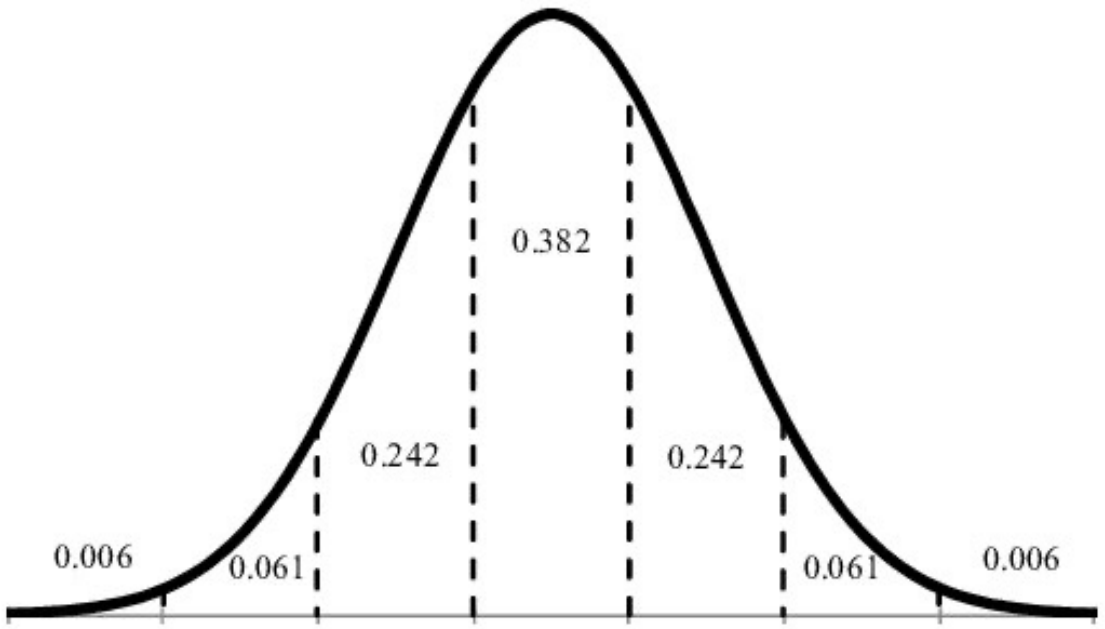
On the other hand, if the p-value is greater than the chosen alpha level, then the null hypothesis (that the data came from a normally distributed population) can not be rejected.

Hypothesis testing

Parametric or non-parametric?



Parametric Hypothesis testing



Hypothesis testing

Parametric: quantitative vs quantitative

Hypothesis testing

Parametric: quantitative vs quantitative

Pearson correlation test:

The Pearson correlation test is a parametric statistical test that measures the strength and direction of the linear relationship between two continuous variables.

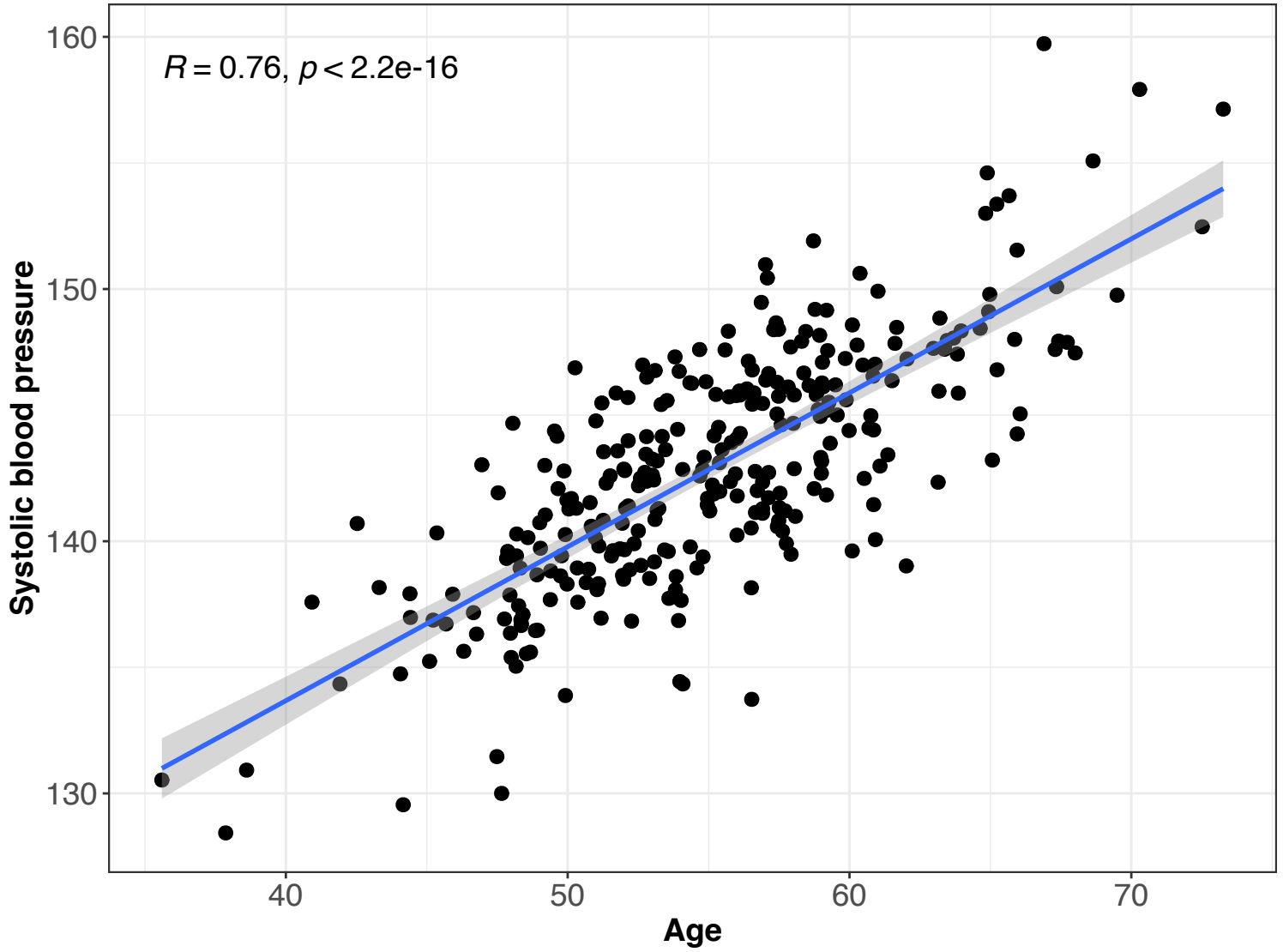
A value between -1 and 1 that quantifies the linear correlation:

- ✓ r = 1: Perfect positive linear relationship.
- ✓ r = -1: Perfect negative linear relationship.
- ✓ r = 0: No linear relationship (null hypothesis H_0)

A significant result (typically $p < 0.05$) leads to rejecting the null hypothesis, indicating that a linear relationship exists between the variables.

Hypothesis testing

Parametric: quantitative vs quantitative



Hypothesis testing

Parametric: quantitative vs qualitative (2 groups)

Hypothesis testing

Parametric: quantitative vs qualitative (2 groups)

T-Student test (independent groups):

The independent two-sample t-test compares the means of two independent groups to determine if there is a statistically significant difference between them.

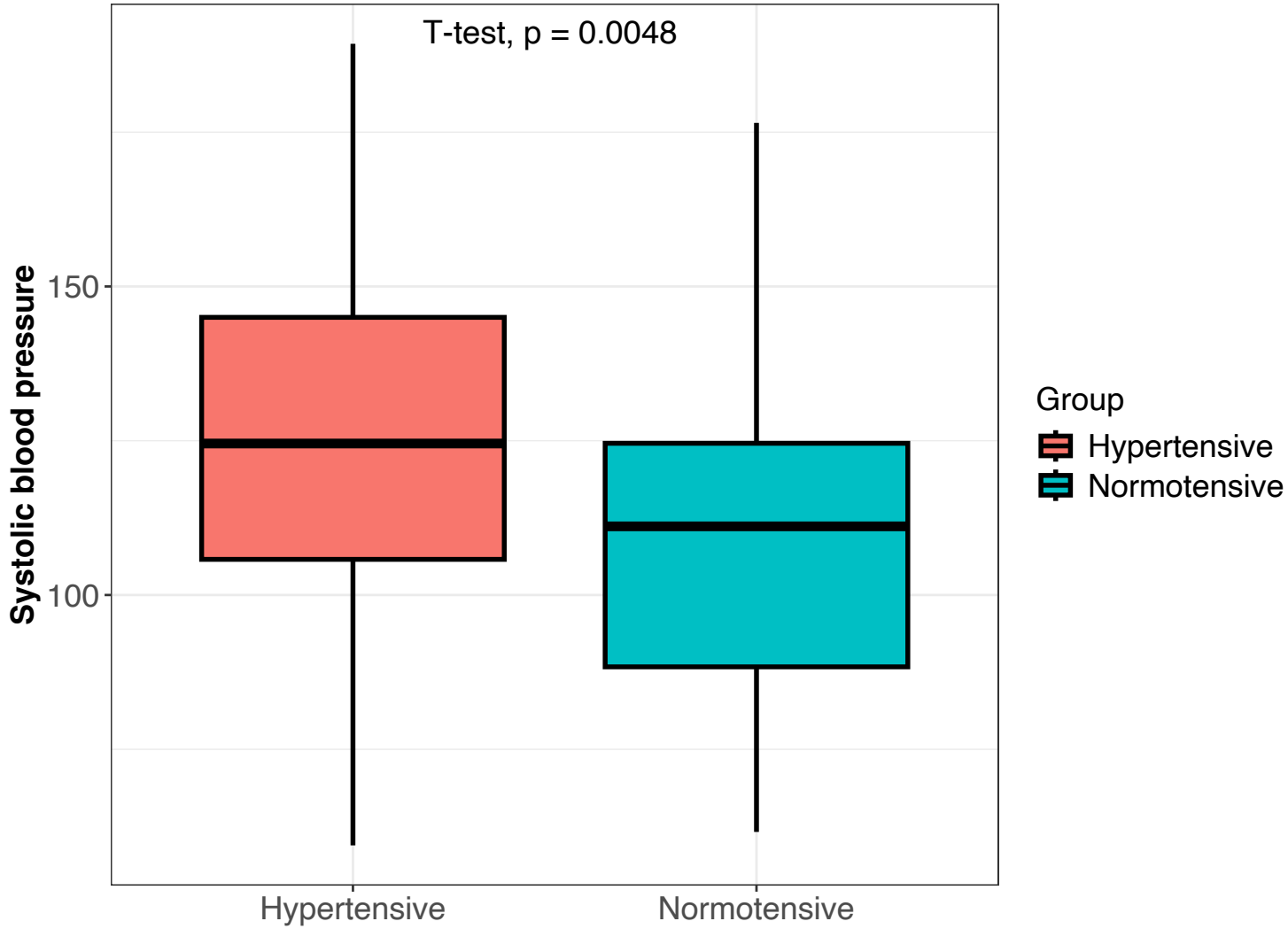
The null hypothesis H_0 is that the means of the two groups are equal. The t-test statistic is calculated as:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{\alpha;\, n_1+n_2-2}$$

A significant result (typically $p < 0.05$) suggests that there is a significant difference between the means of the two groups.

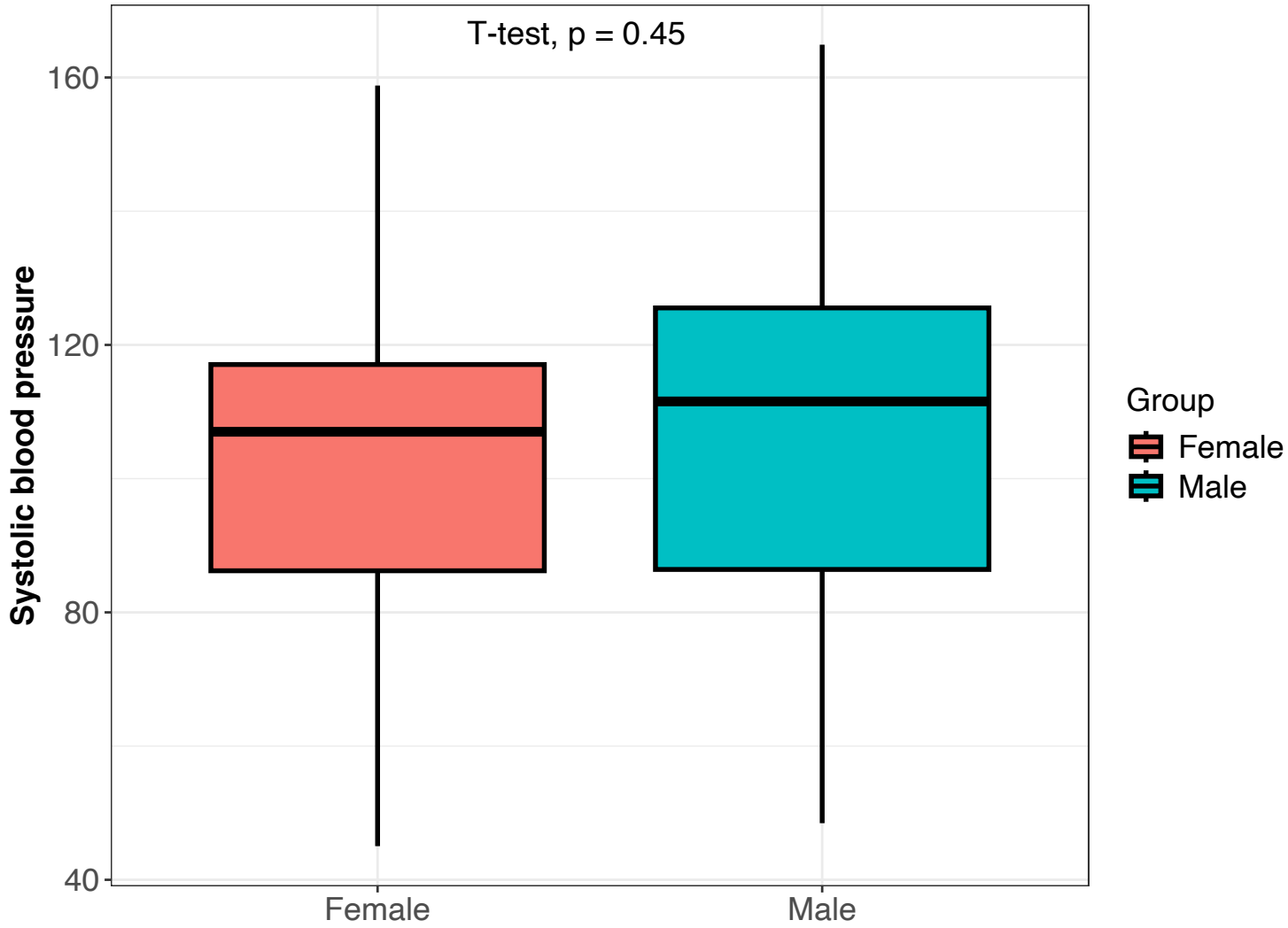
Hypothesis testing

Parametric: quantitative vs qualitative (2 groups)



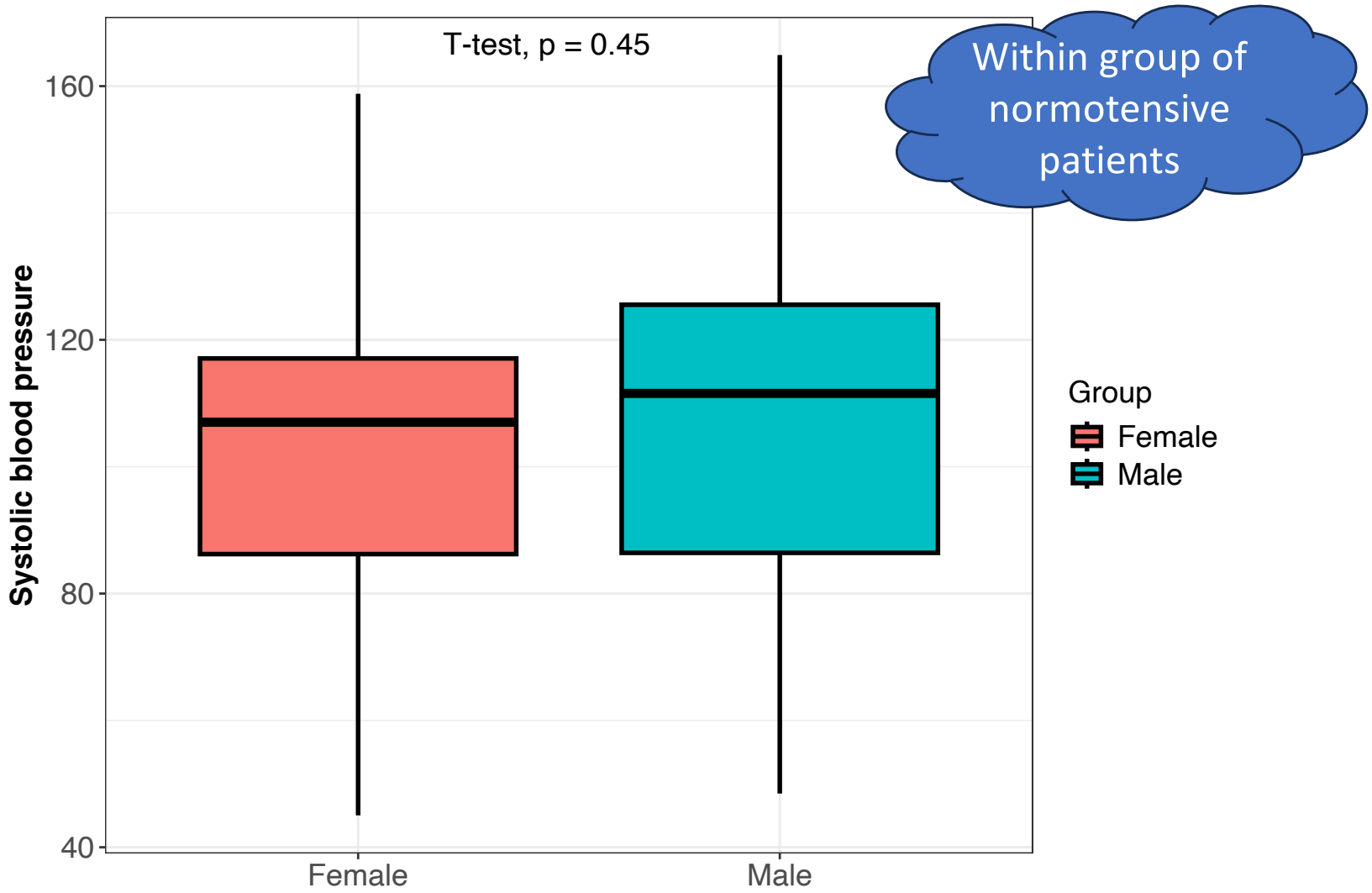
Hypothesis testing

Parametric: quantitative vs qualitative (2 groups)



Hypothesis testing

Parametric: quantitative vs qualitative (2 groups)



Hypothesis testing

Parametric: quantitative vs qualitative (2 groups)

T-Student test (paired groups):

The paired t-test compares the means of two related groups to determine if there is a statistically significant difference between them. This test is often used when the same subjects are used in both conditions.

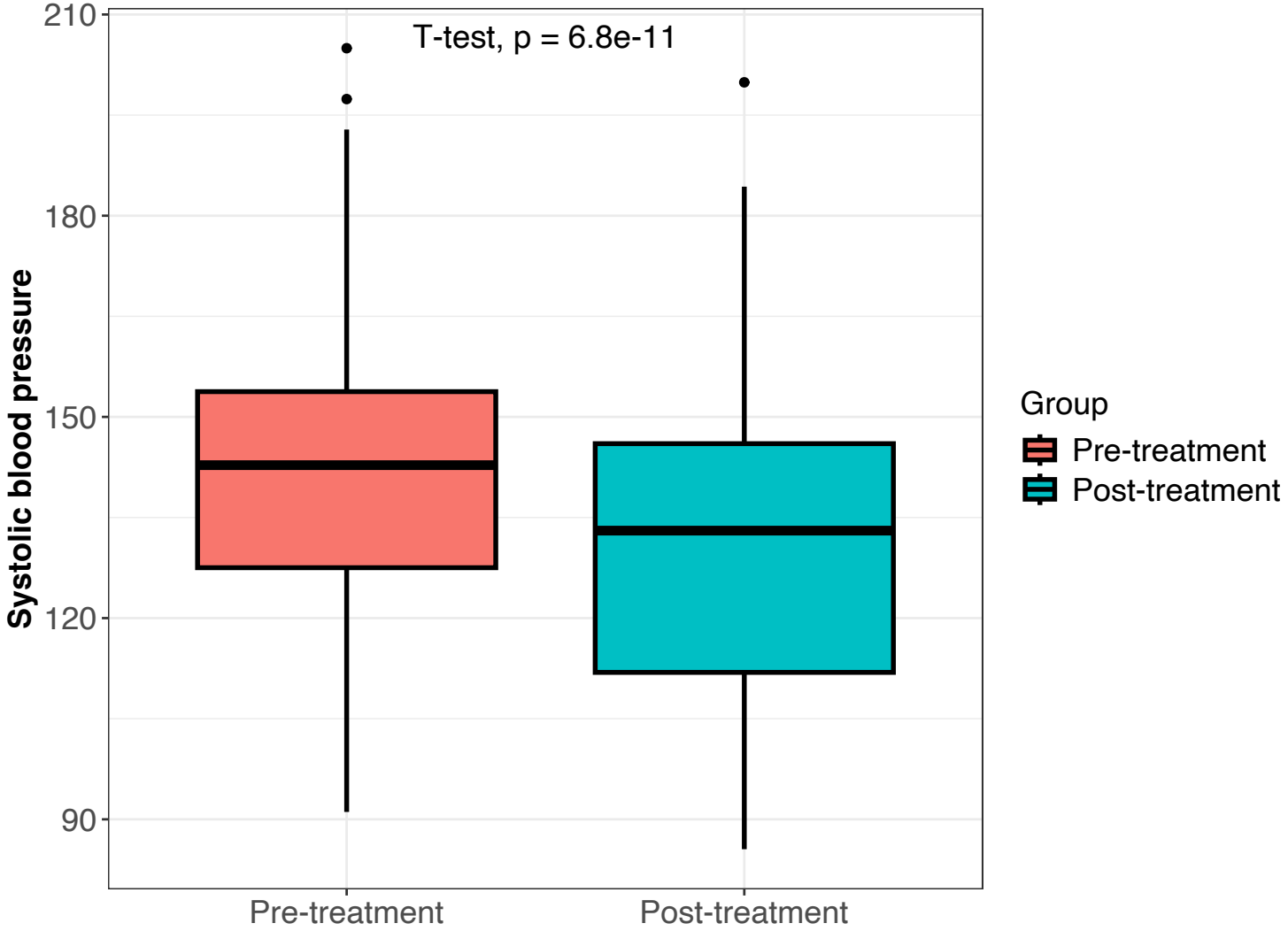
The null hypothesis H_0 is that the mean difference between the paired observations is zero. The t-test statistic is calculated as:

$$\frac{\bar{d}}{S_d / \sqrt{n}} \sim t_{\alpha;n-1}$$

A significant result (typically $p < 0.05$) suggests that there is a significant difference between the paired observations.

Hypothesis testing

Parametric: quantitative vs qualitative (2 groups)



Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)

Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)

Analysis of Variance (ANOVA) + Tukey's HSD post-hoc test:

One-way analysis of variance (abbreviated one-way ANOVA) is a technique that can be used to compare whether two o more samples means are significantly different or not.

Typically, however, the one-way ANOVA is used to test for differences among at least three groups, since the two-group case can be covered by a t-test. When there are only two means to compare, the t-test and the ANOVA e test are equivalent.

Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)

Analysis of Variance (ANOVA) + Tukey's HSD post-hoc test:

One-way analysis of variance (abbreviated one-way ANOVA) is a technique that can be used to compare whether two o more samples means are significantly different or not.

Typically, however, the one-way ANOVA is used to test for differences among at least three groups, since the two-group case can be covered by a t-test. When there are only two means to compare, the t-test and the ANOVA e test are equivalent.

Hypotheses of ANOVA test can be formalized as follow:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: \quad otherwise \end{cases}$$

Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)

Analysis of Variance (ANOVA) + Tukey's HSD post-hoc test:

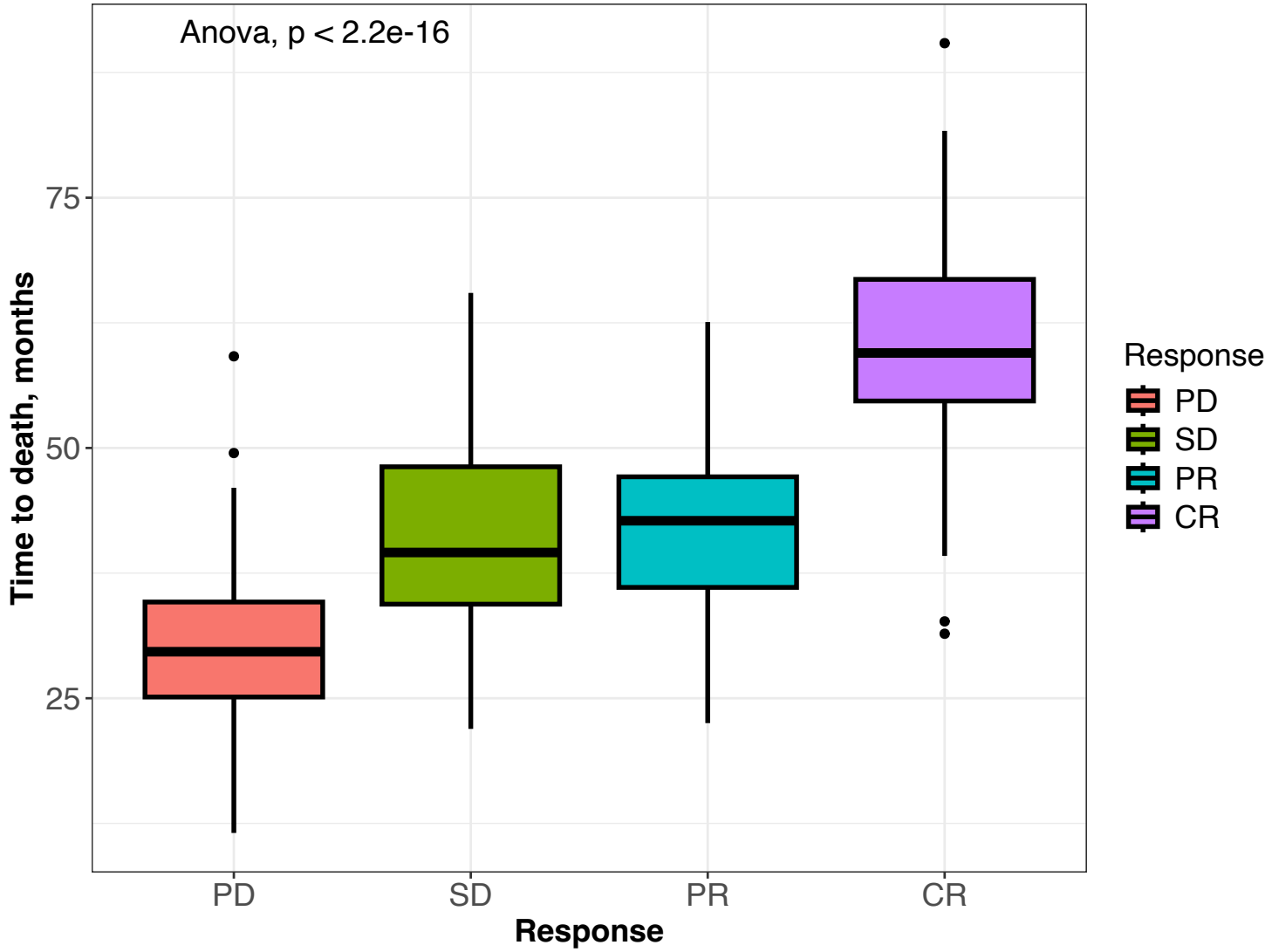
The Tukey Test (or Tukey procedure), also called Tukey's Honest Significant Difference test, is a post-hoc test based on the studentized range distribution.

An ANOVA test can tell you if your results are significant overall, but it won't tell you exactly where those differences lie. After you have run an ANOVA and found significant results, then you can run Tukey's HSD to find out which specific groups' means (compared with each other) are different.

The test compares all possible pairs of means.

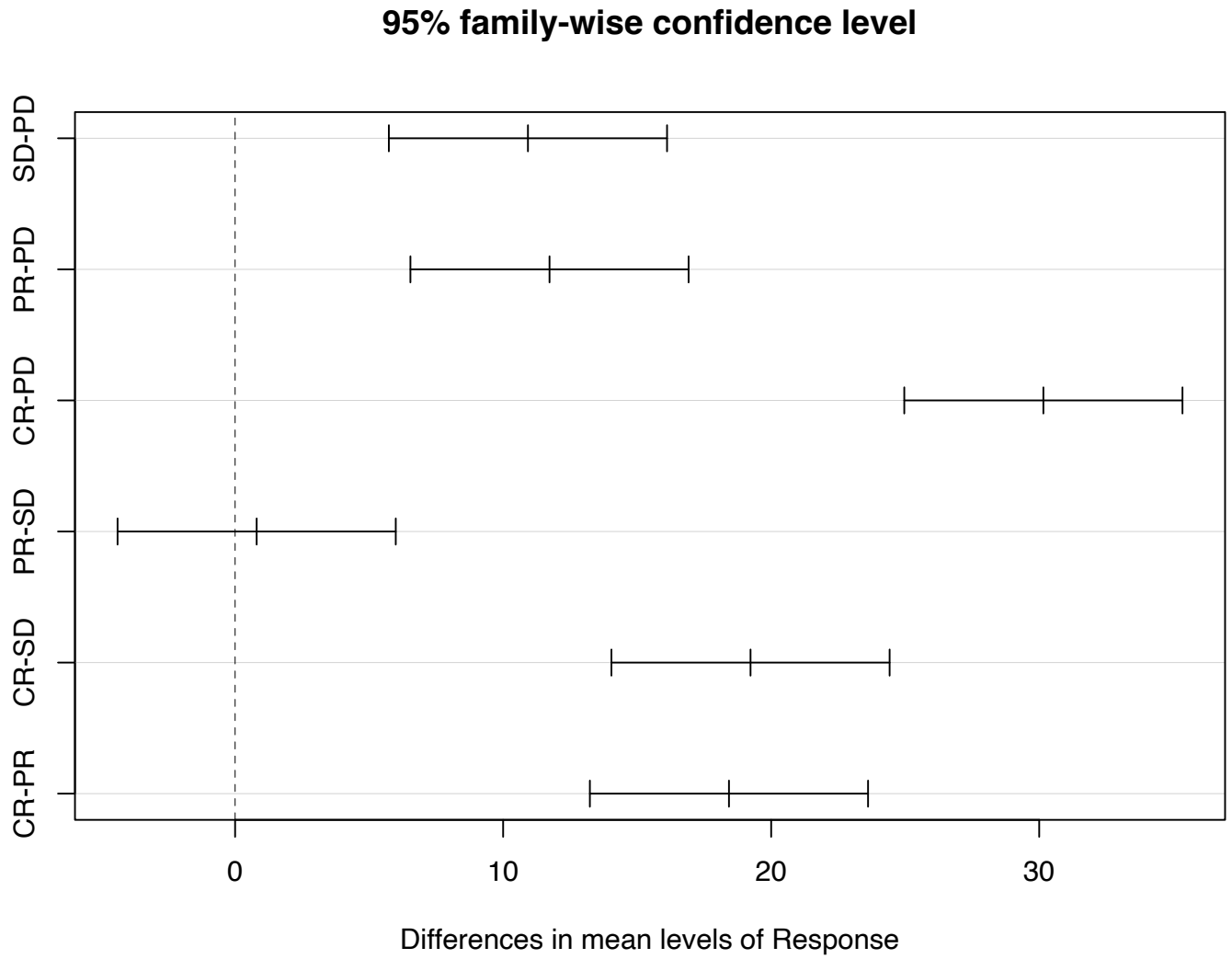
Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)



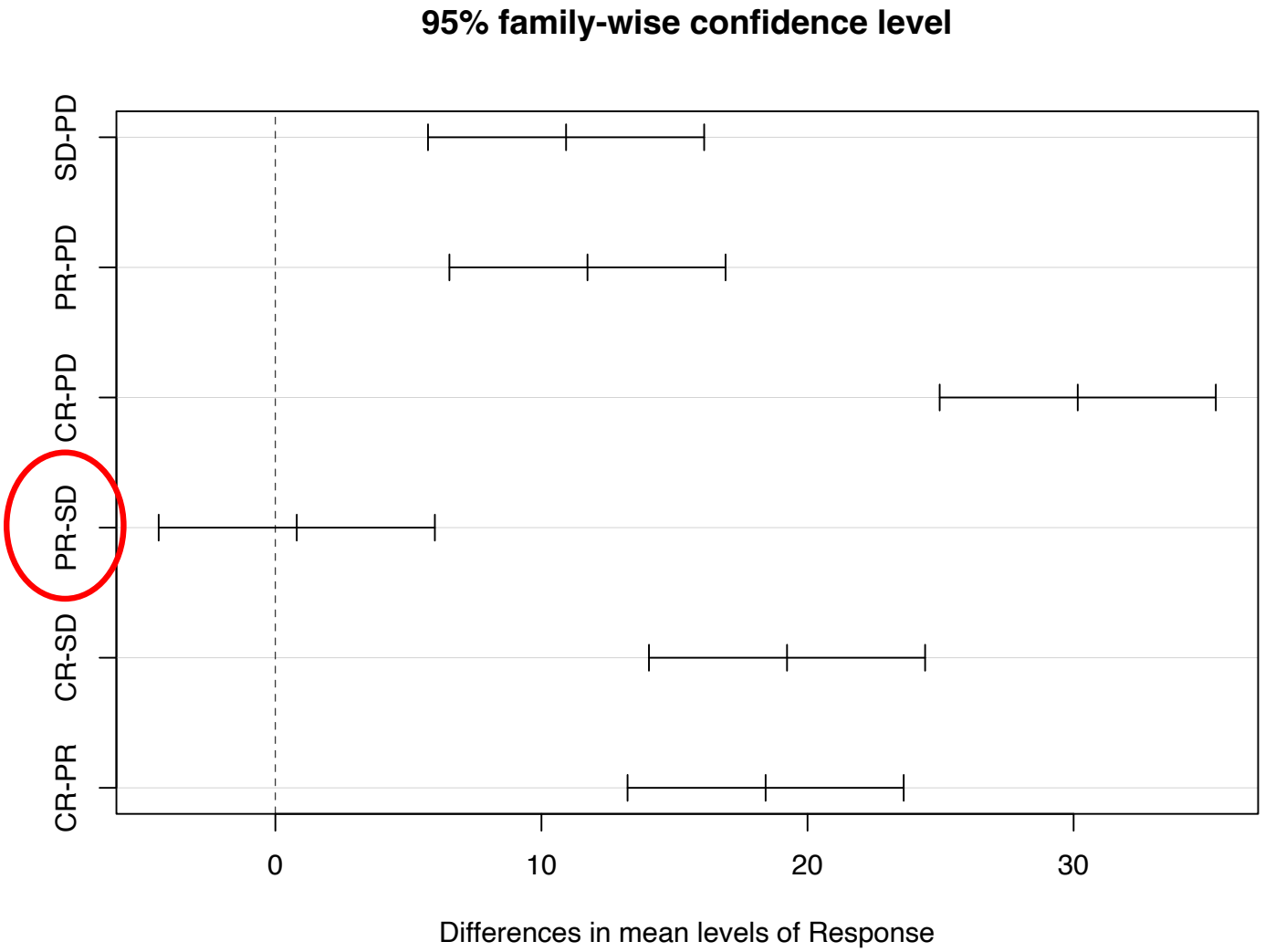
Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)



Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)



Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)

Pairwise t-test:

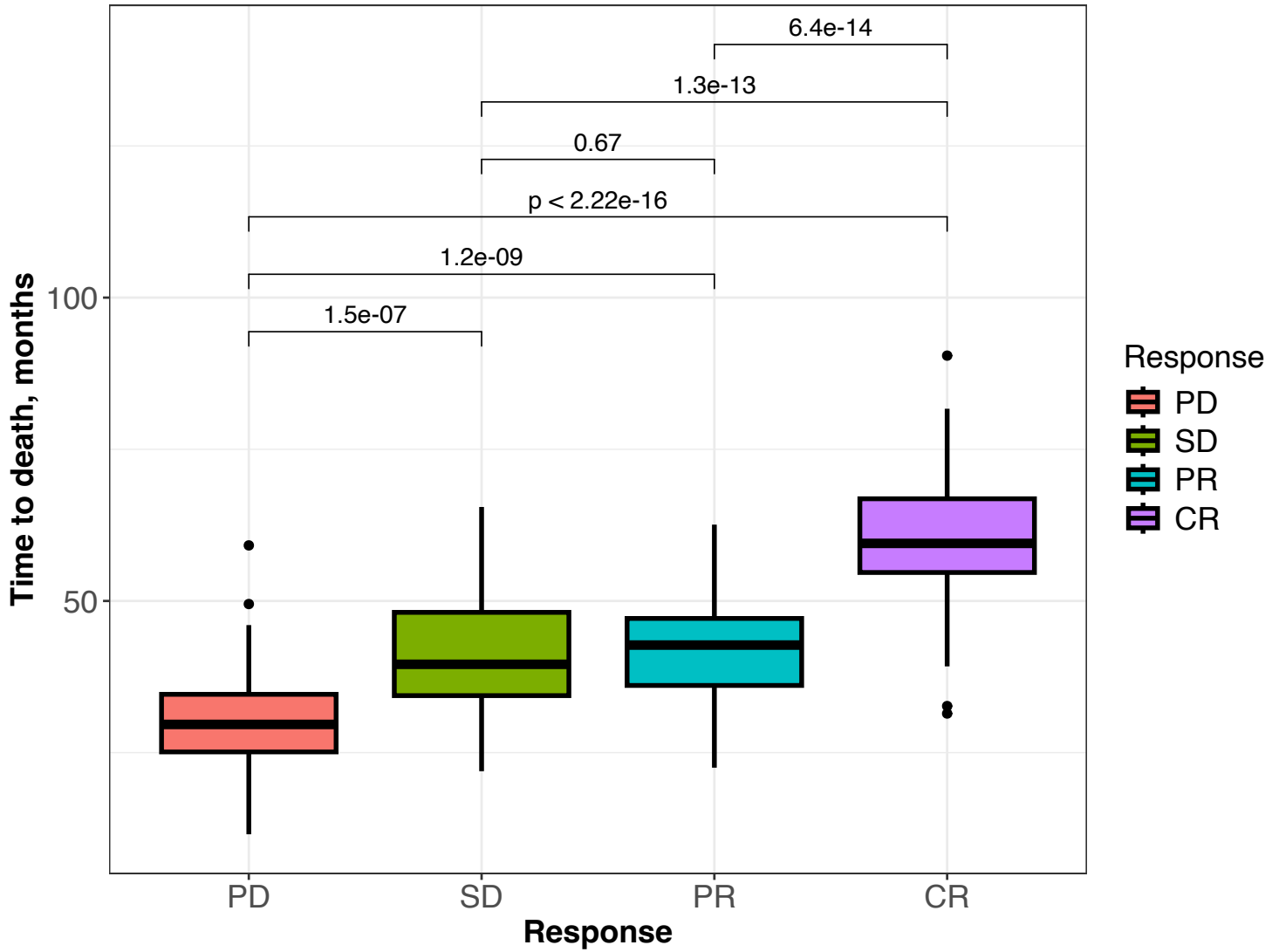
The pairwise t-test is used to compare the means of pairs of groups to determine if there are statistically significant differences between them. This test is often used after or as alternative of the ANOVA to identify which specific groups differ from each other.

When performing multiple pairwise comparisons, the risk of Type I error (false positive) increases. Common methods to adjust for this include:

- ✓ Bonferroni adjustment: adjust the significance level by dividing it by the number of comparisons.
- ✓ Holm adjustment: a stepwise method that is less conservative than Bonferroni.

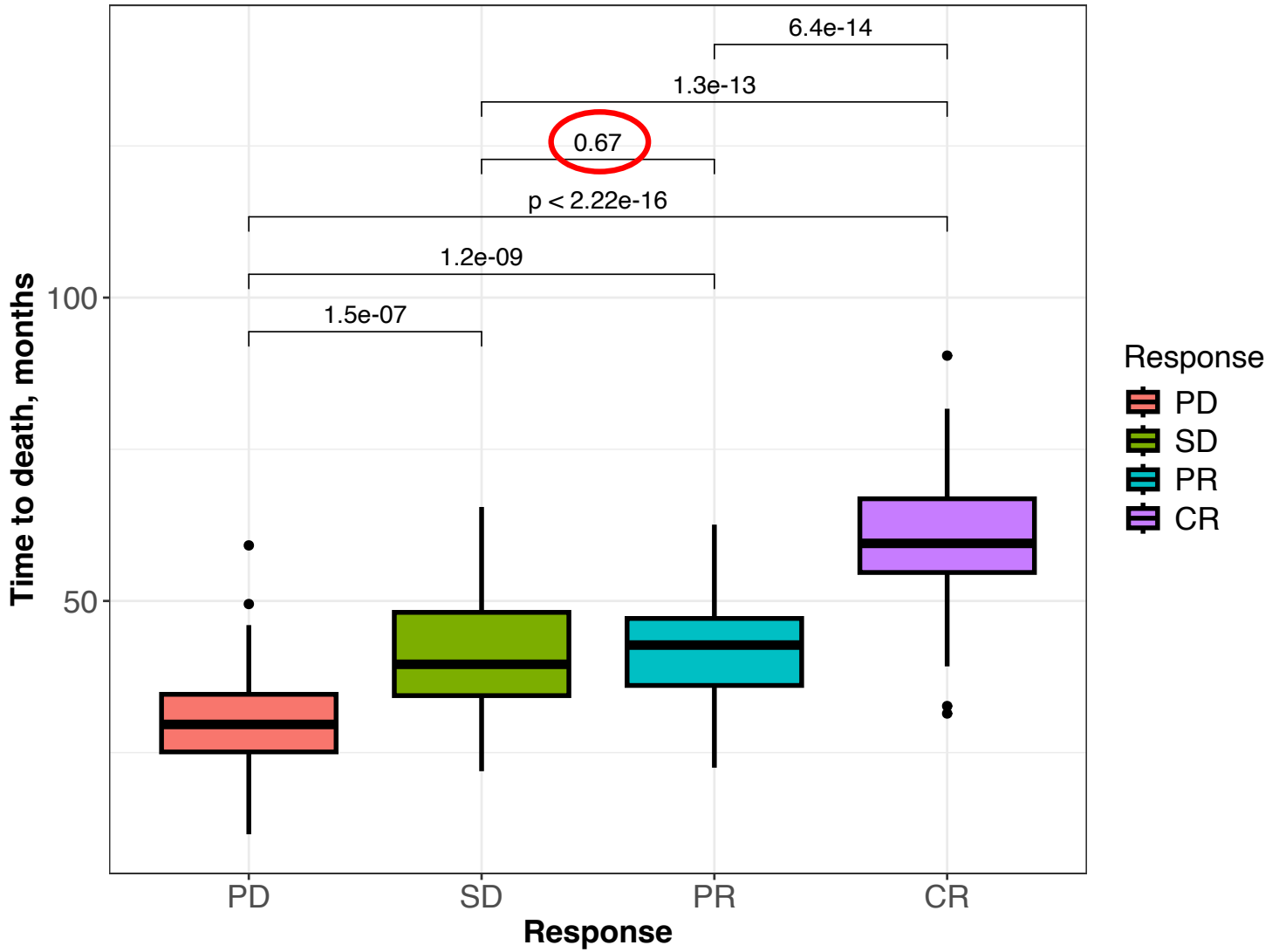
Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)



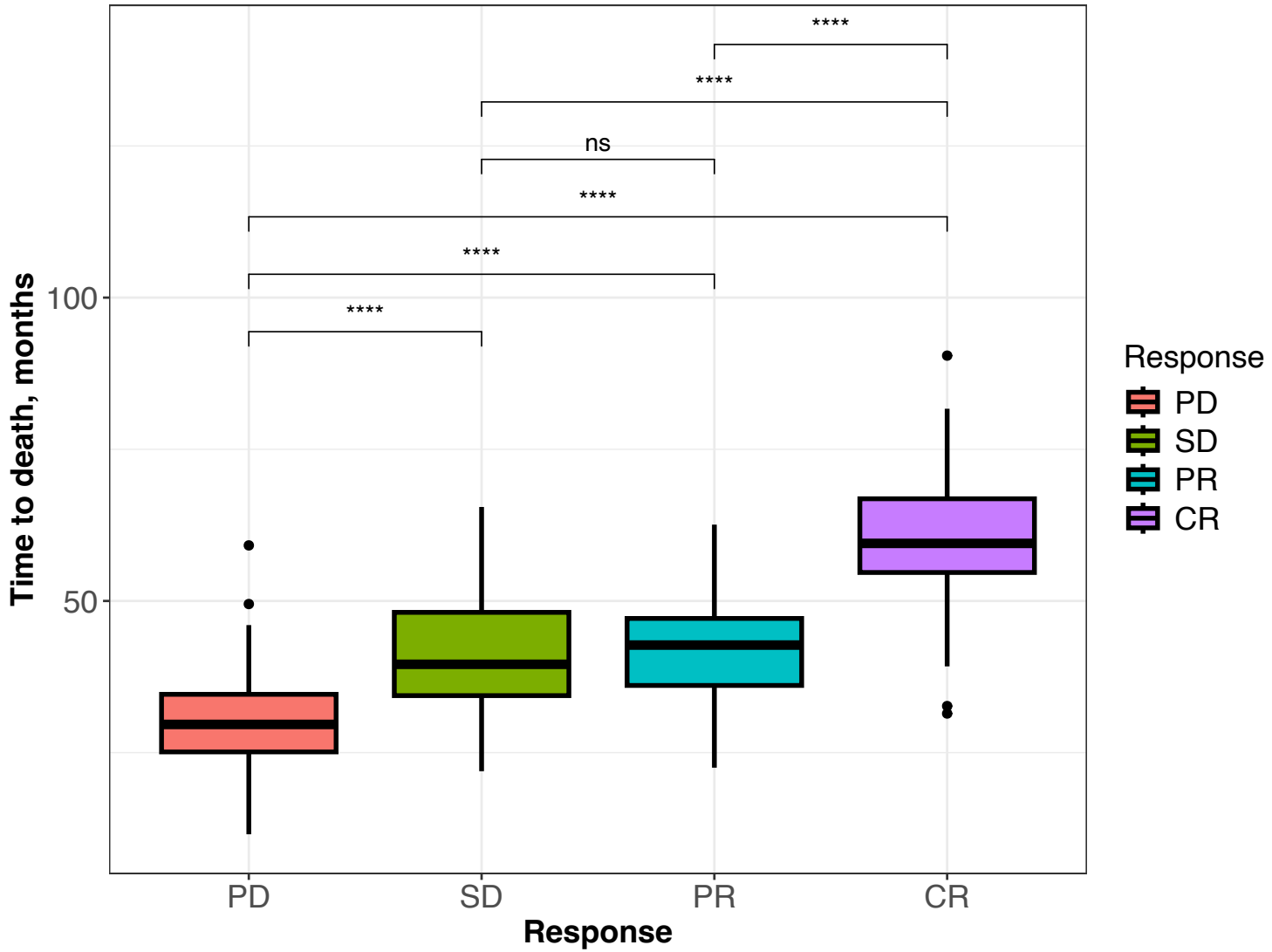
Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)



Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)



Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)

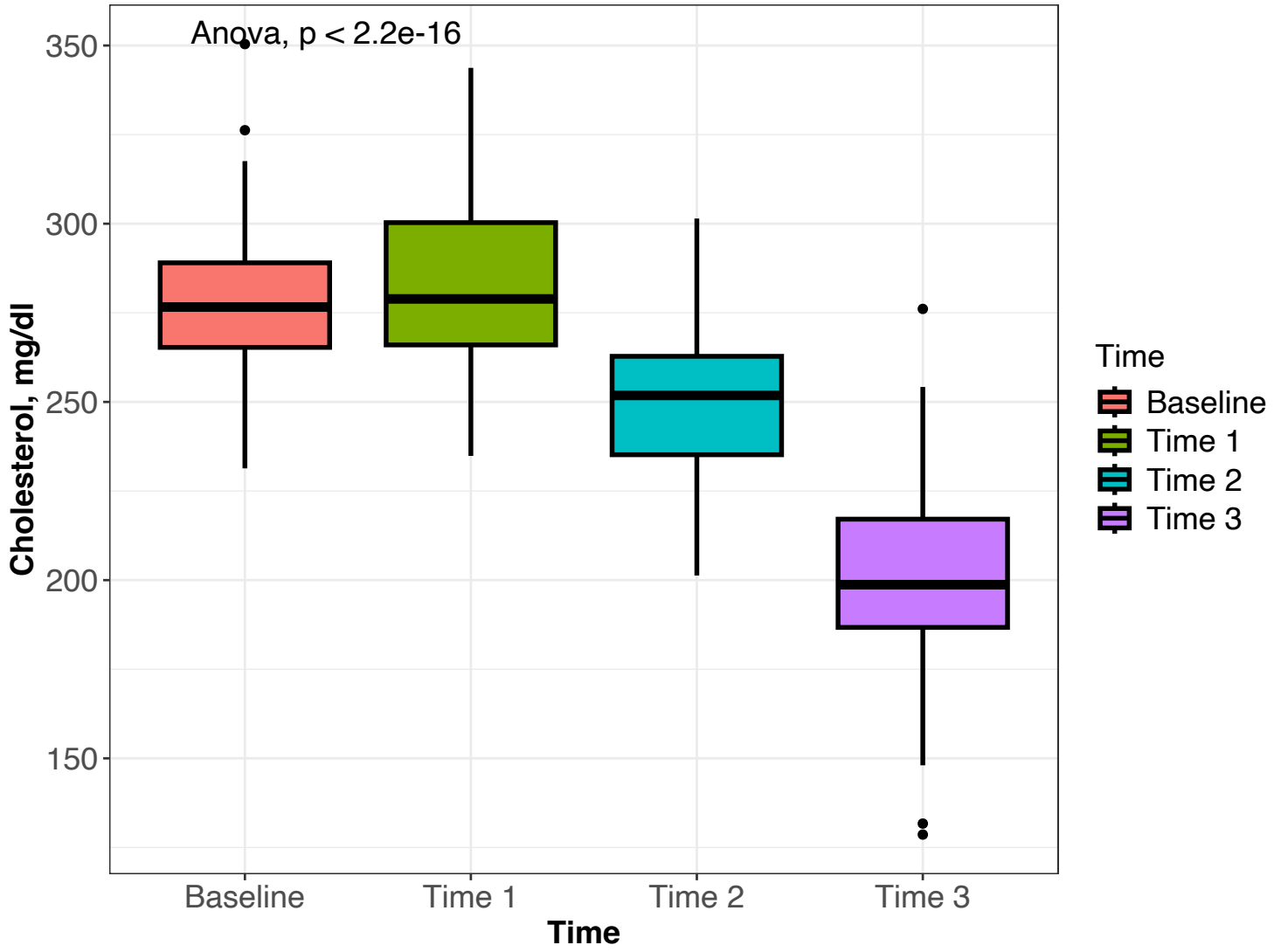
ANOVA for repeated measures + Tukey's HSD post-hoc test:

The repeated measures ANOVA compares the means of three or more related groups to determine if there is a statistically significant difference among them. It is used when the same subjects are measured multiple times under different conditions or over different time points.

An ANOVA test can tell you if your results are significant overall, but it won't tell you exactly where those differences lie. After you have run an ANOVA and found significant results, then you can run Tukey's HSD to find out which specific groups' means (compared with each other) are different.

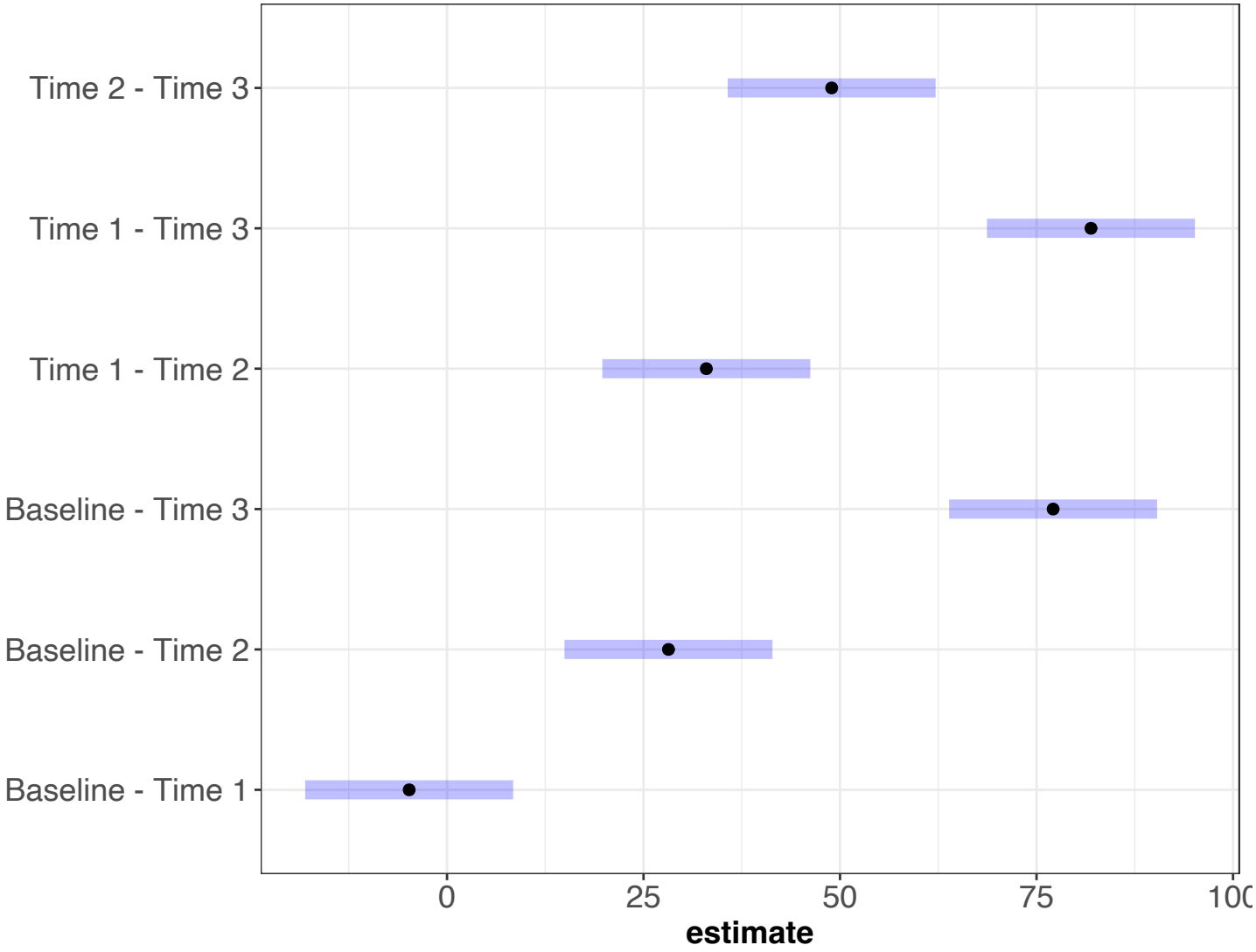
Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)



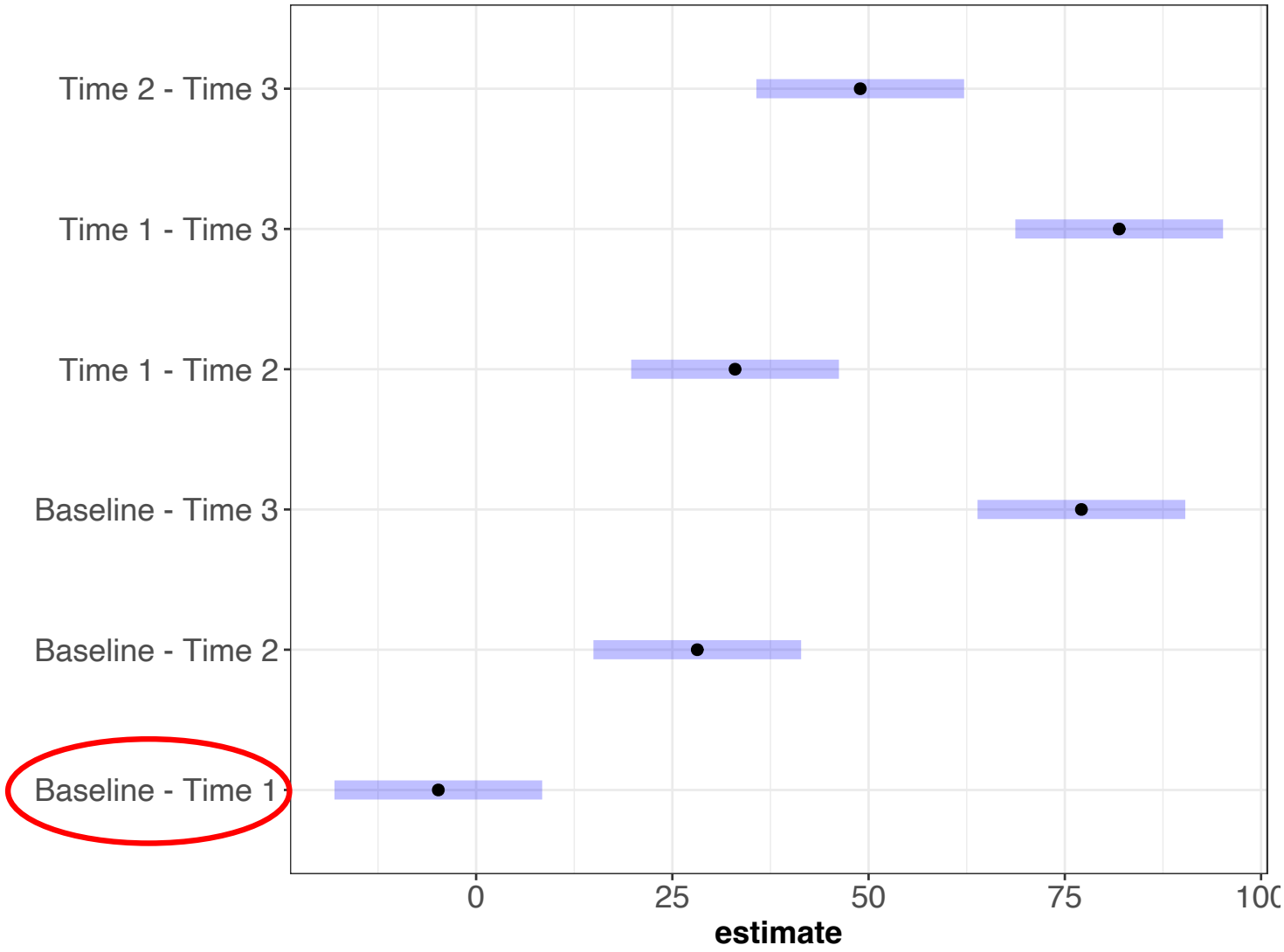
Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)



Hypothesis testing

Parametric: quantitative vs qualitative (> 2 groups)



non-parametric Hypothesis testing

Hypothesis testing

non-parametric

Example with VETERAN dataset:

The Veterans' Administration Lung Cancer study dataset is available in the 'survival' R package. It is a randomized trial of two treatment regimens for lung cancer. This is a dataset composed by 137 observations and 8 variables:

1. **trt:** 1=standard 2=test
2. **celltype:** 1=squamous, 2=smallcell, 3=adeno, 4=large
3. **time:** survival time in days
4. **status:** censoring status
5. **karno:** Karnofsky performance score (100=good)
6. **diagtime:** months from diagnosis to randomisation
7. **age:** in years
8. **prior:** prior therapy 0=no, 10=yes

Hypothesis testing

non-parametric

Example with VETERAN dataset:

The Veterans' Administration Lung Cancer study dataset is available in the 'survival' R package. It is a randomized trial of two treatment regimens for lung cancer. This is a dataset composed by 137 observations and 8 variables:

1. **trt:** 1=standard 2=test
2. **celltype:** 1=squamous, 2=smallcell, 3=adeno, 4=large
3. **time:** survival time in days
4. **status:** censoring status
5. **karno:** Karnofsky performance score (100=good)
6. **diagtime:** months from diagnosis to randomisation
7. **age:** in years
8. **prior:** prior therapy 0=no, 10=yes

Hypothesis testing

non-parametric

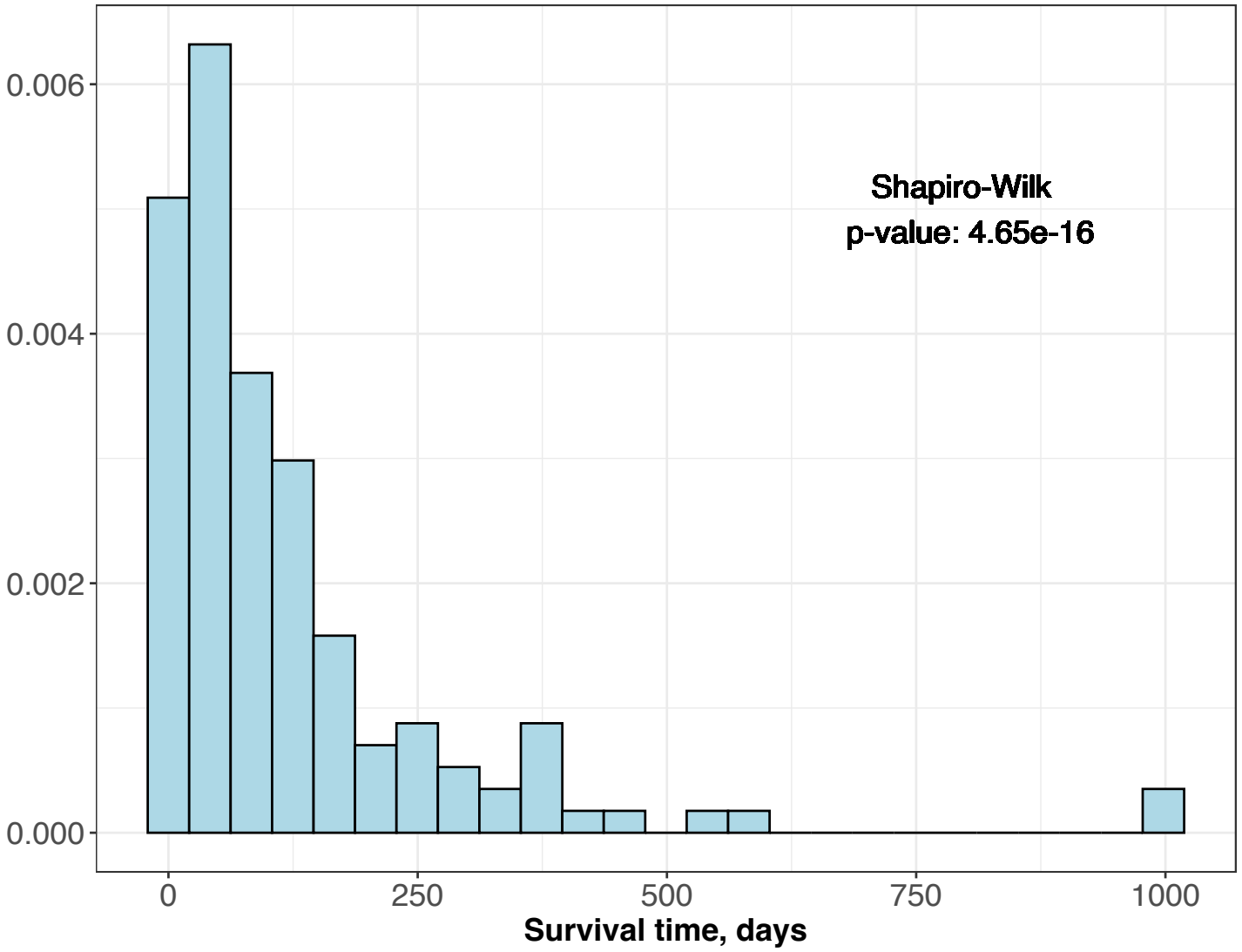
```
> skim(data = data)
── Data Summary ────────────────────────
                           Values
Name                       data
Number of rows             137
Number of columns          8
_______________________
Column type frequency:
  factor                   4
  numeric                  4
_______________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 trt                   0             1 FALSE          2 Sta: 69, Exp: 68
2 celltype              0             1 FALSE          4 sma: 48, squ: 35, ade: 27, lar: 27
3 status                0             1 FALSE          2 Dea: 128, Ali: 9
4 prior                 0             1 FALSE          2 No: 97, Yes: 40

── Variable type: numeric ─────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean    sd p0 p25 p50 p75 p100 hist
1 time                  0             1 122.   158.   1  25  80 144  999 ▇▁▁▁
2 karno                 0             1  58.6   20.0 10  40  60  75   99 ▂▅▇▃
3 diagtime              0             1   8.77  10.6  1   3   5  11   87 ▇▁▁▁
4 age                   0             1  58.3   10.5 34  51  62  66   81 ▂▅▇▃
```

Hypothesis testing

non-parametric



Hypothesis testing

non-parametric: quantitative vs quantitative

Hypothesis testing

non-parametric: quantitative vs quantitative

Spearman's rank correlation test:

The Spearman's rank correlation test is a non-parametric statistical test that measures the strength and direction of the linear relationship between two continuous variables.

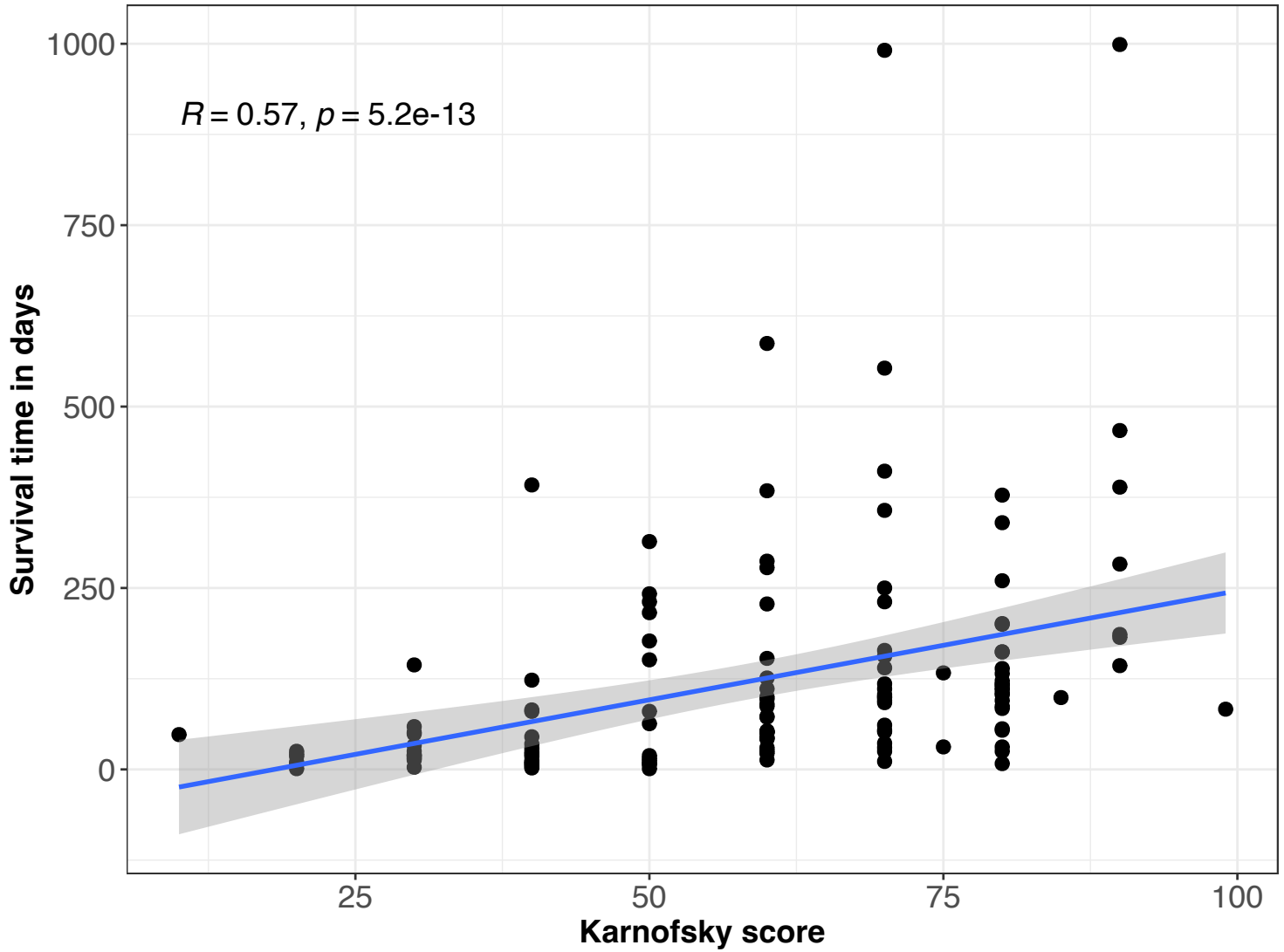
A value between -1 and 1 that quantifies the linear correlation:

- ✓ r = 1: Perfect positive linear relationship.
- ✓ r = -1: Perfect negative linear relationship.
- ✓ r = 0: No linear relationship (null hypothesis H_0)

A significant result (typically $p < 0.05$) leads to rejecting the null hypothesis, indicating that a linear relationship exists between the variables.

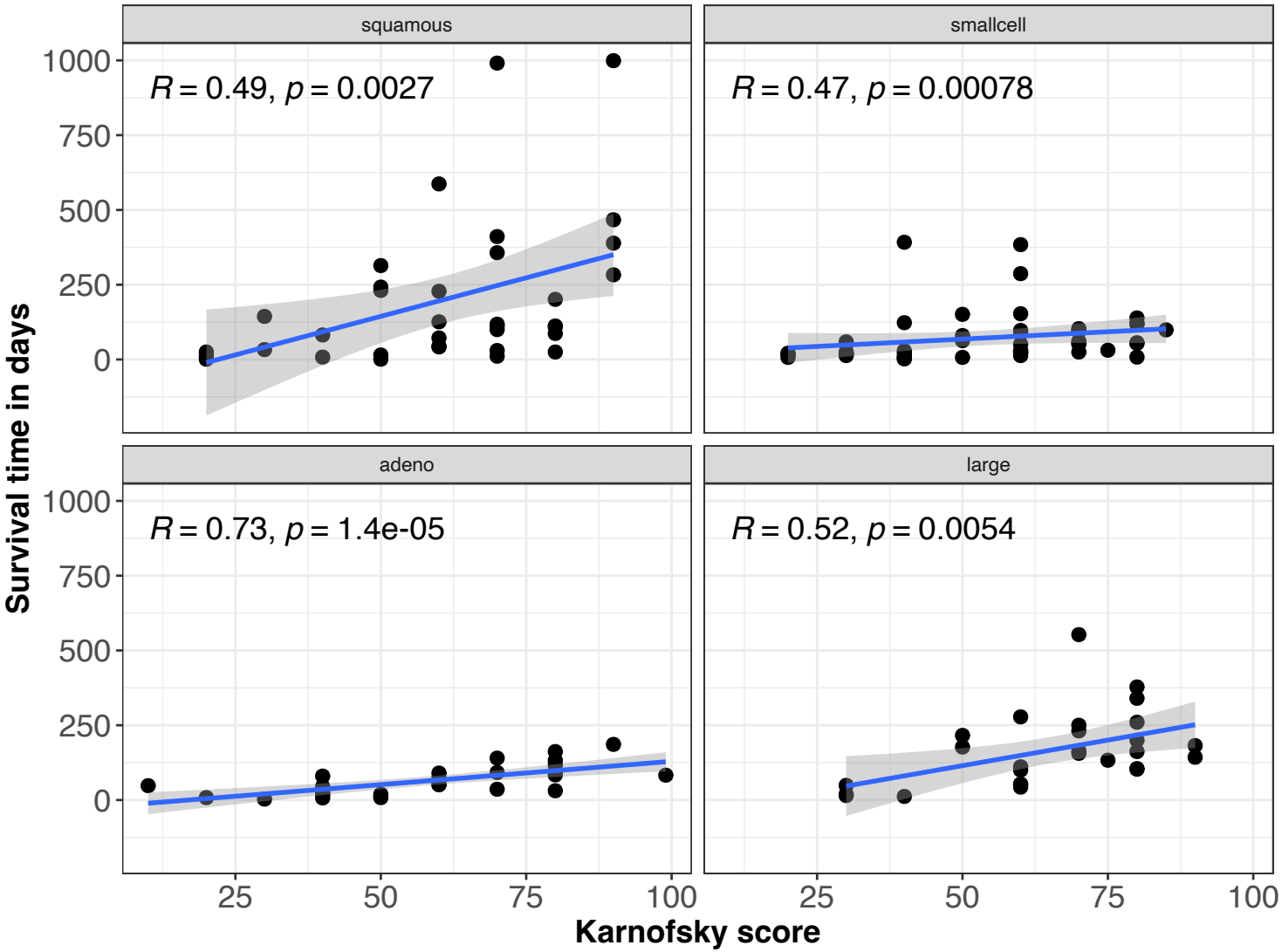
Hypothesis testing

non-parametric: quantitative vs quantitative



Hypothesis testing

non-parametric: quantitative vs quantitative



Hypothesis testing

non-parametric: quantitative vs qualitative (2 groups)

Hypothesis testing

non-parametric: quantitative vs qualitative (2 groups)

Wilcoxon test:

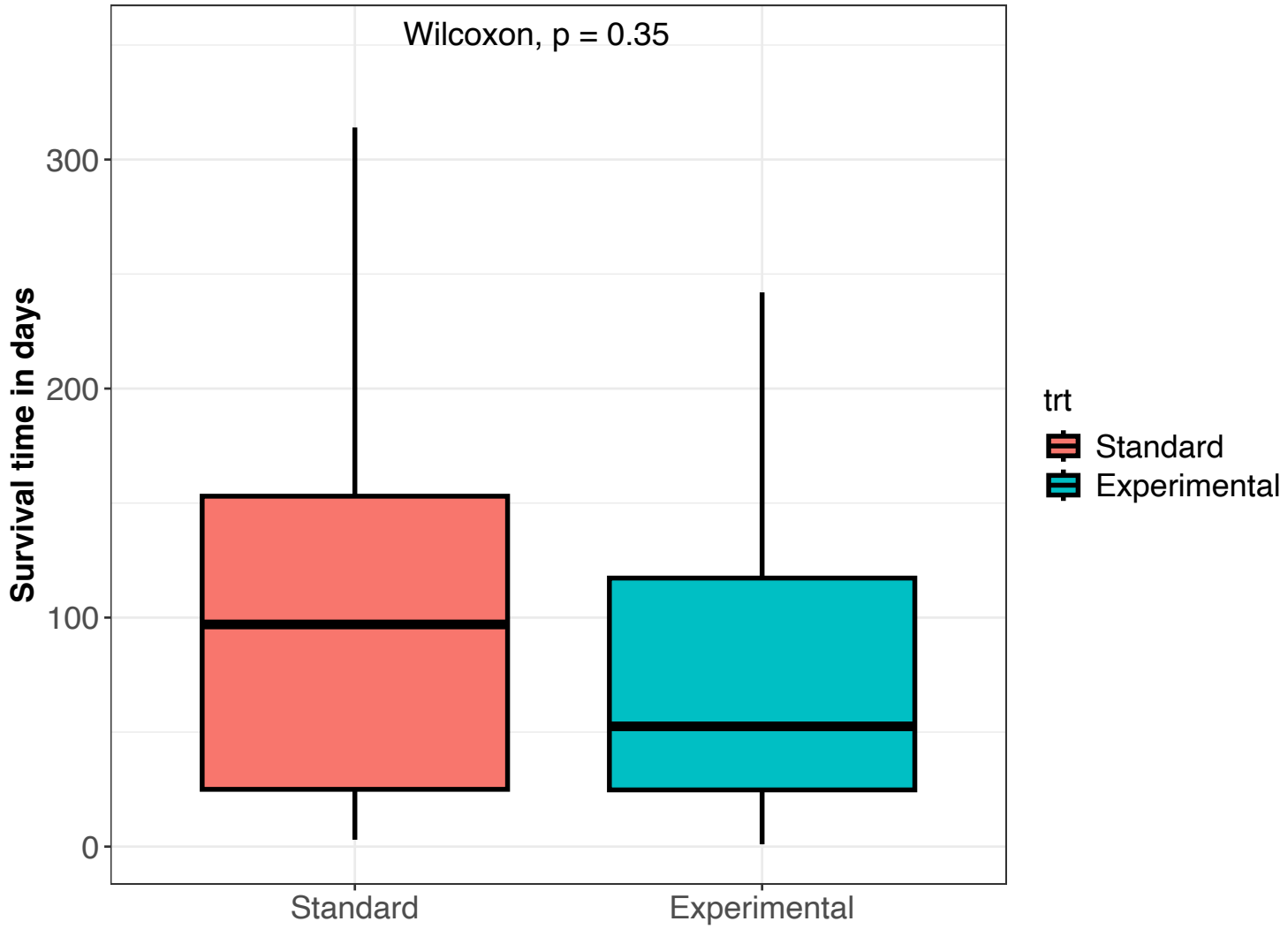
The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e., it is a paired difference test).

It can be used as an alternative to the paired Student's t-test when the distribution of the difference between two samples' means cannot be assumed to be normally distributed.

A Wilcoxon signed-rank test is a non-parametric test that can be used to determine whether two dependent samples were selected from populations having the same distribution.

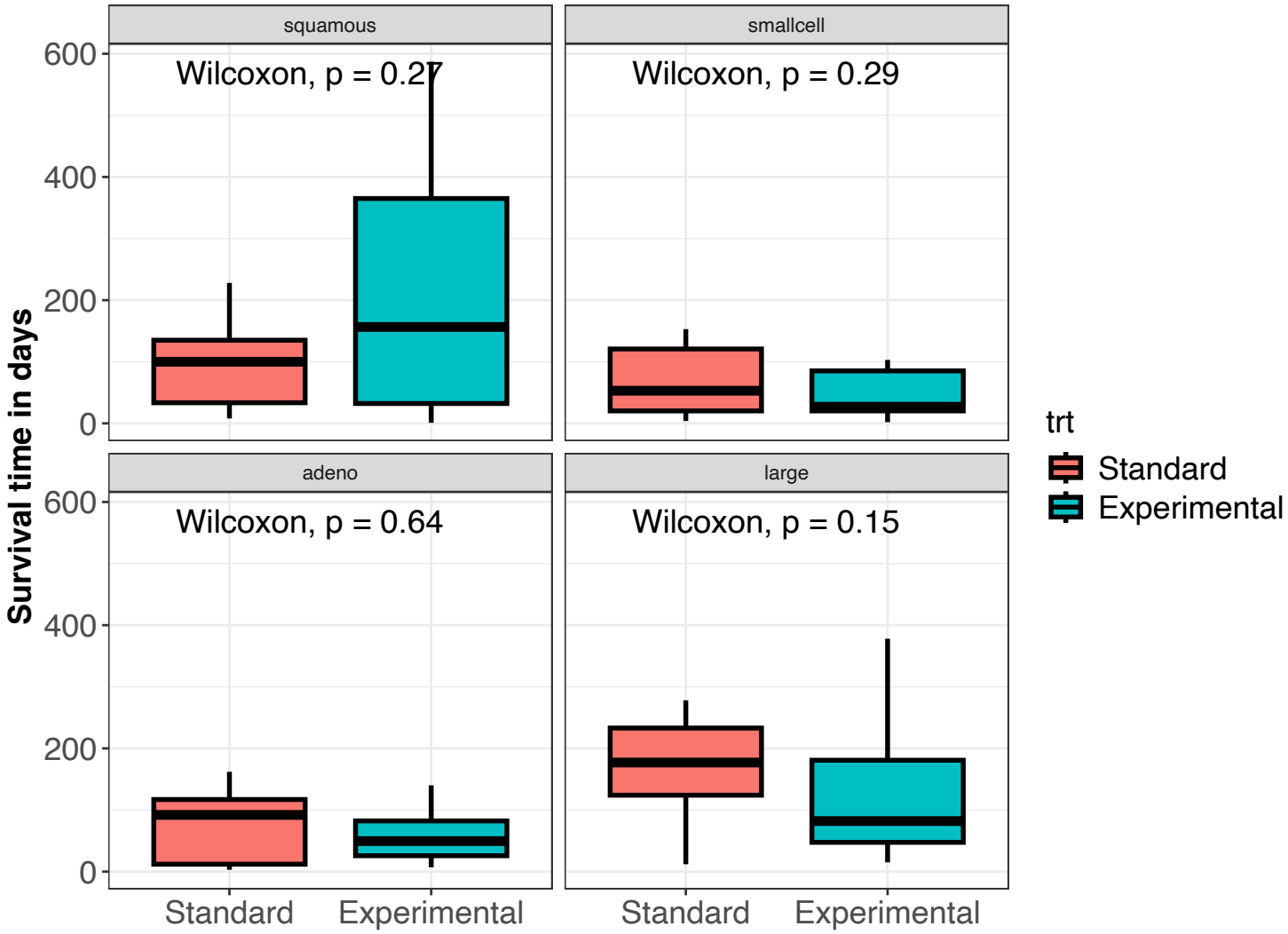
Hypothesis testing

non-parametric: quantitative vs qualitative (2 groups)



Hypothesis testing

non-parametric: quantitative vs qualitative (2 groups)



Hypothesis testing

non-parametric: quantitative vs qualitative (> 2 groups)

Hypothesis testing

non-parametric: quantitative vs qualitative (> 2 groups)

Kruskal-Wallis test + Dunn's post-hoc test:

The Kruskal-Wallis test is a non-parametric method for testing whether samples originate from the same distribution.

It is used for comparing two or more independent samples of equal or different sample sizes, and extends the Mann-Whitney U test, which is used for comparing only two groups.

The parametric equivalent of the Kruskal-Wallis test is the one-way analysis of variance (ANOVA).

A significant Kruskal-Wallis test indicates that at least one sample stochastically dominates one other sample. The test does not identify where this stochastic dominance occurs or for how many pairs of groups stochastic dominance obtains.

Hypothesis testing

non-parametric: quantitative vs qualitative (> 2 groups)

Kruskal-Wallis test + Dunn's post-hoc test:

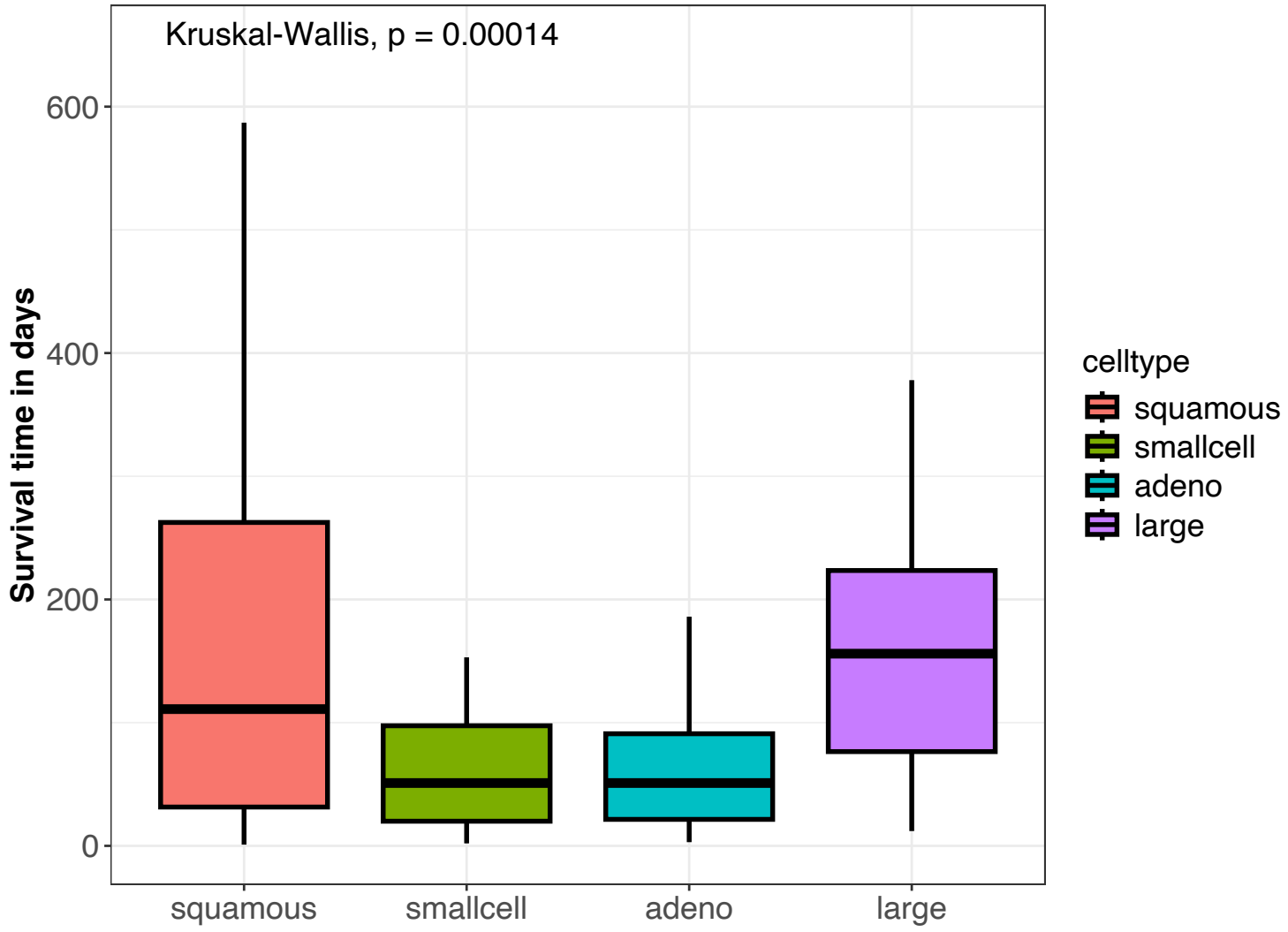
For analyzing the specific sample pairs for stochastic dominance, Dunn's test can be used.

It is the appropriate non-parametric pairwise multiple-comparison procedure when a Kruskal-Wallis test is rejected.

After you have run a Kruskal-Wallis and found significant results, then you can run Dunn's test to find out which specific groups are different.

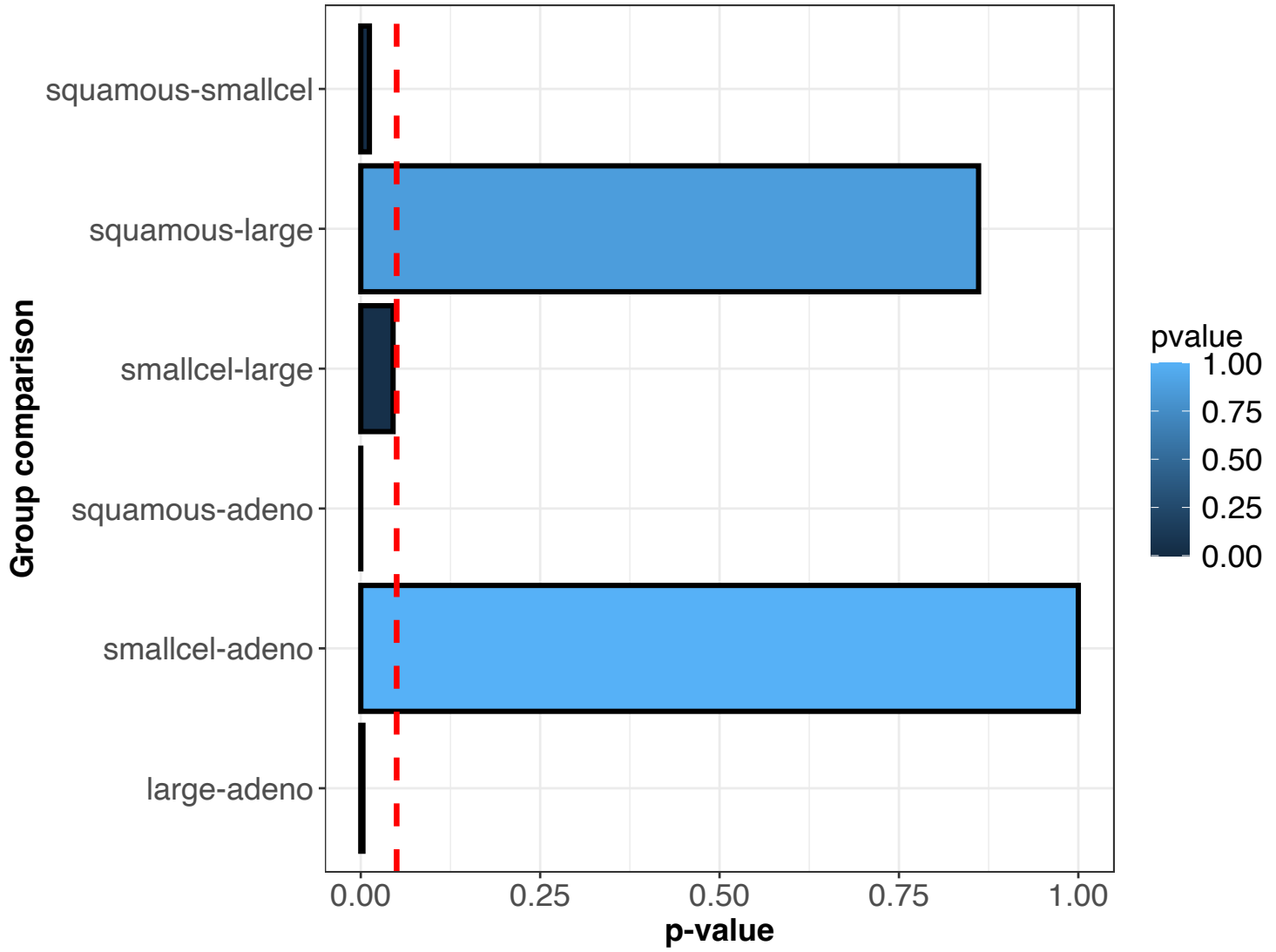
Hypothesis testing

non-parametric: quantitative vs qualitative (> 2 groups)



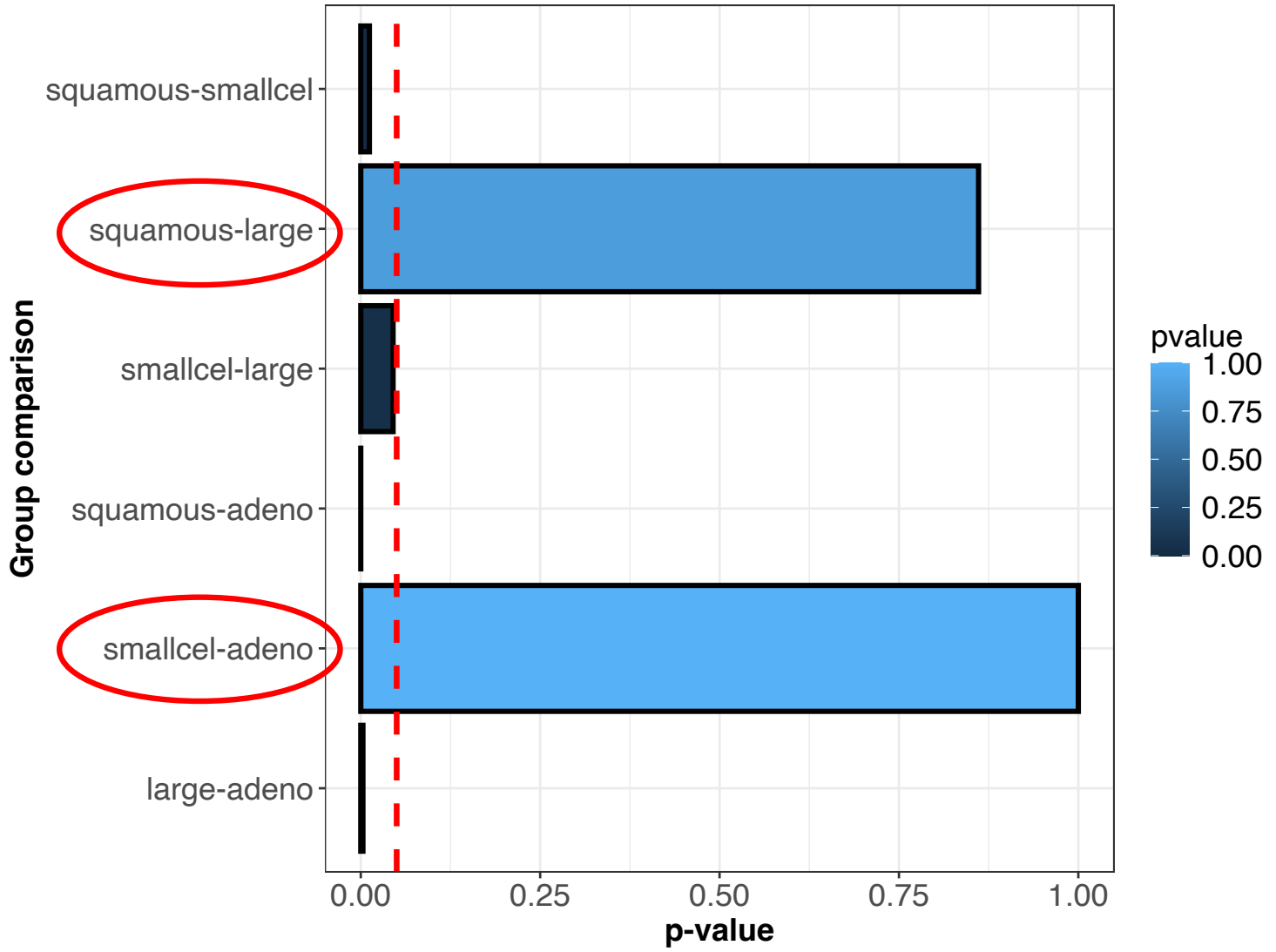
Hypothesis testing

non-parametric: quantitative vs qualitative (> 2 groups)



Hypothesis testing

non-parametric: quantitative vs qualitative (> 2 groups)



Survival analysis

Kaplan-Meier estimator

Survival analysis

Kaplan-Meier estimator

Survival function:

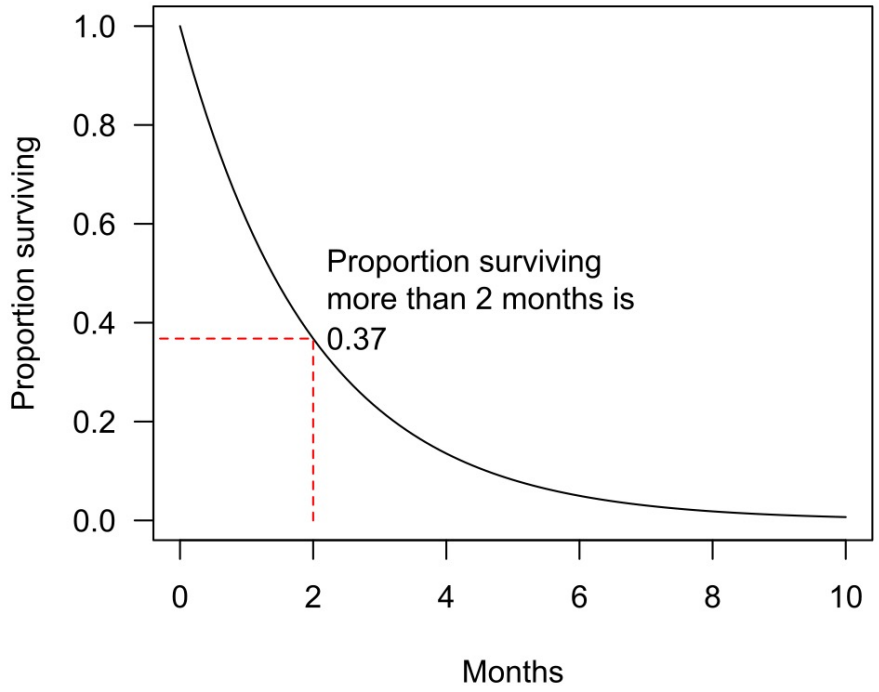
The survival function is a function that gives the probability that a patient, device, or other object of interest will survive beyond any specified time.

Survival analysis

Kaplan-Meier estimator

Survival function:

The survival function is a function that gives the probability that a patient, device, or other object of interest will survive beyond any specified time.



Survival analysis

Kaplan-Meier estimator

The Kaplan-Meier estimator is a non-parametric statistic used to estimate the survival function from lifetime data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment.

In other fields, Kaplan-Meier estimators may be used to measure the length of time people remain unemployed after a job loss, the time-to-failure of machine parts, or how long fleshy fruits remain on plants before they are removed by frugivores.

The estimator is named after Edward L. Kaplan and Paul Meier, who each submitted similar manuscripts to the Journal of the American Statistical Association.

Survival analysis

Kaplan-Meier estimator

To compare the survival distributions of two or more groups, the log-rank test is used. It is a non-parametric statistical test commonly used in survival analysis to test the null hypothesis that there is no difference in survival between the groups.

The log-rank test compares the observed number of events (e.g., deaths) to the expected number of events under the null hypothesis, at each observed event time.

The null hypothesis is that there is no difference in survival between the groups, and then A significant result (typically $p < 0.05$) indicates that there is a difference in the survival distributions between the groups.

Survival analysis

Kaplan-Meier estimator

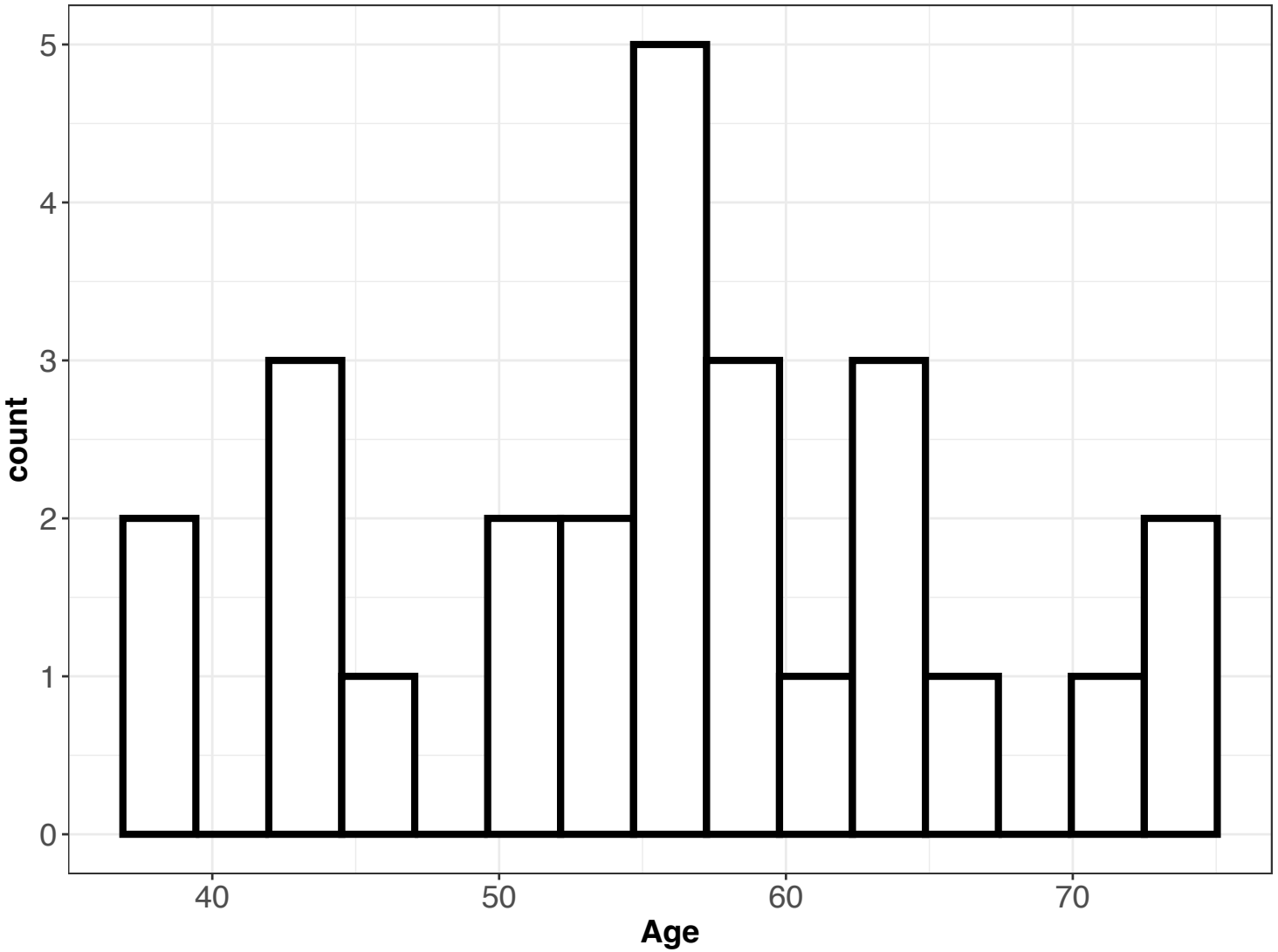
Example with Ovarian dataset:

The Ovarian dataset is available in the 'survival' R package. It is a randomised trial comparing two treatments for ovarian cancer. This is a dataset composed by 26 observations and 6 variables:

1. futime: survival or censoring time
2. fustat: censoring status
3. age: in years
4. resid.ds: residual disease present (1=no,2=yes)
5. rx: treatment group
6. ecog.ps: ECOG performance status (1 is better, see reference)

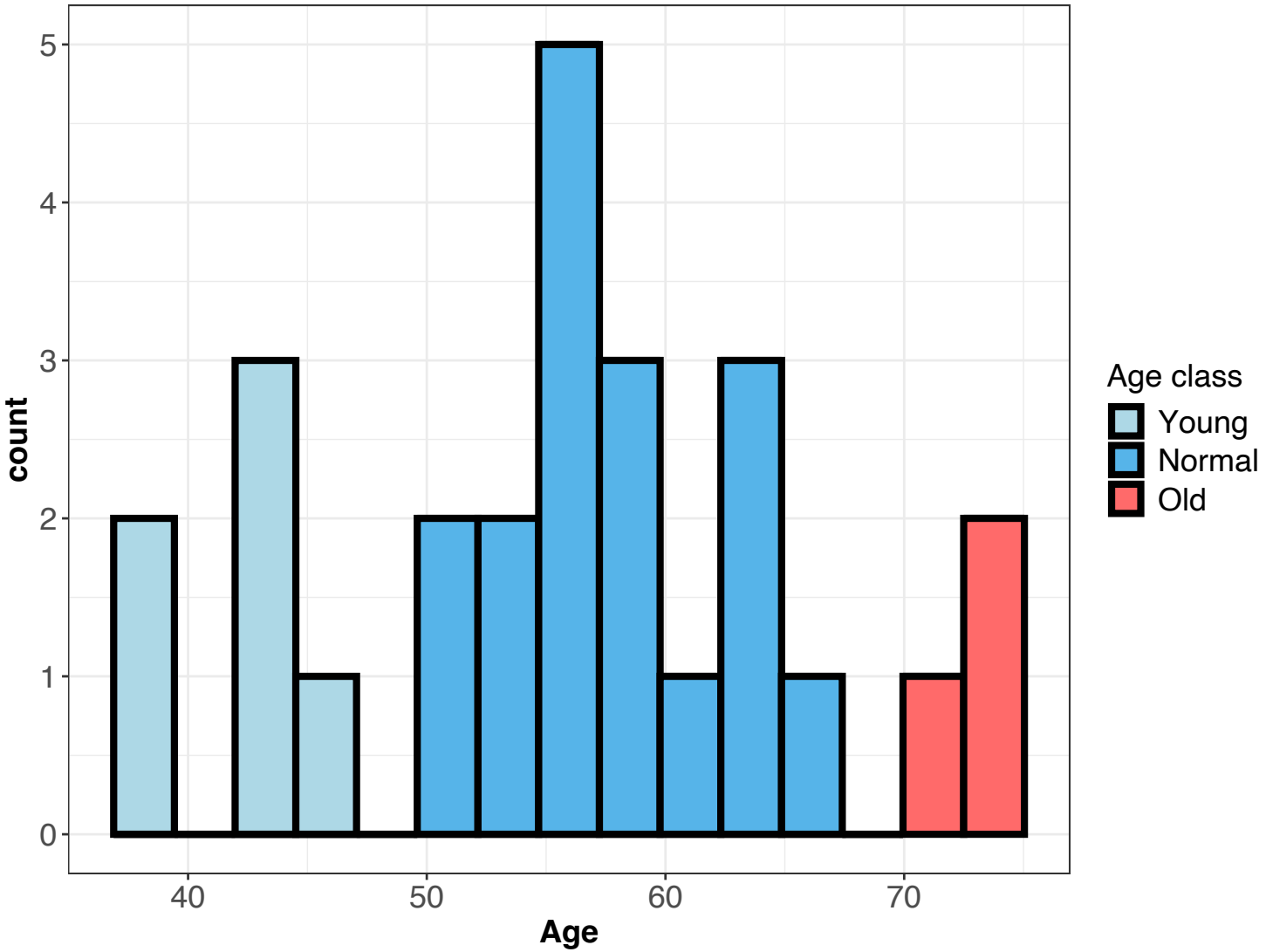
Survival analysis

Kaplan-Meier estimator



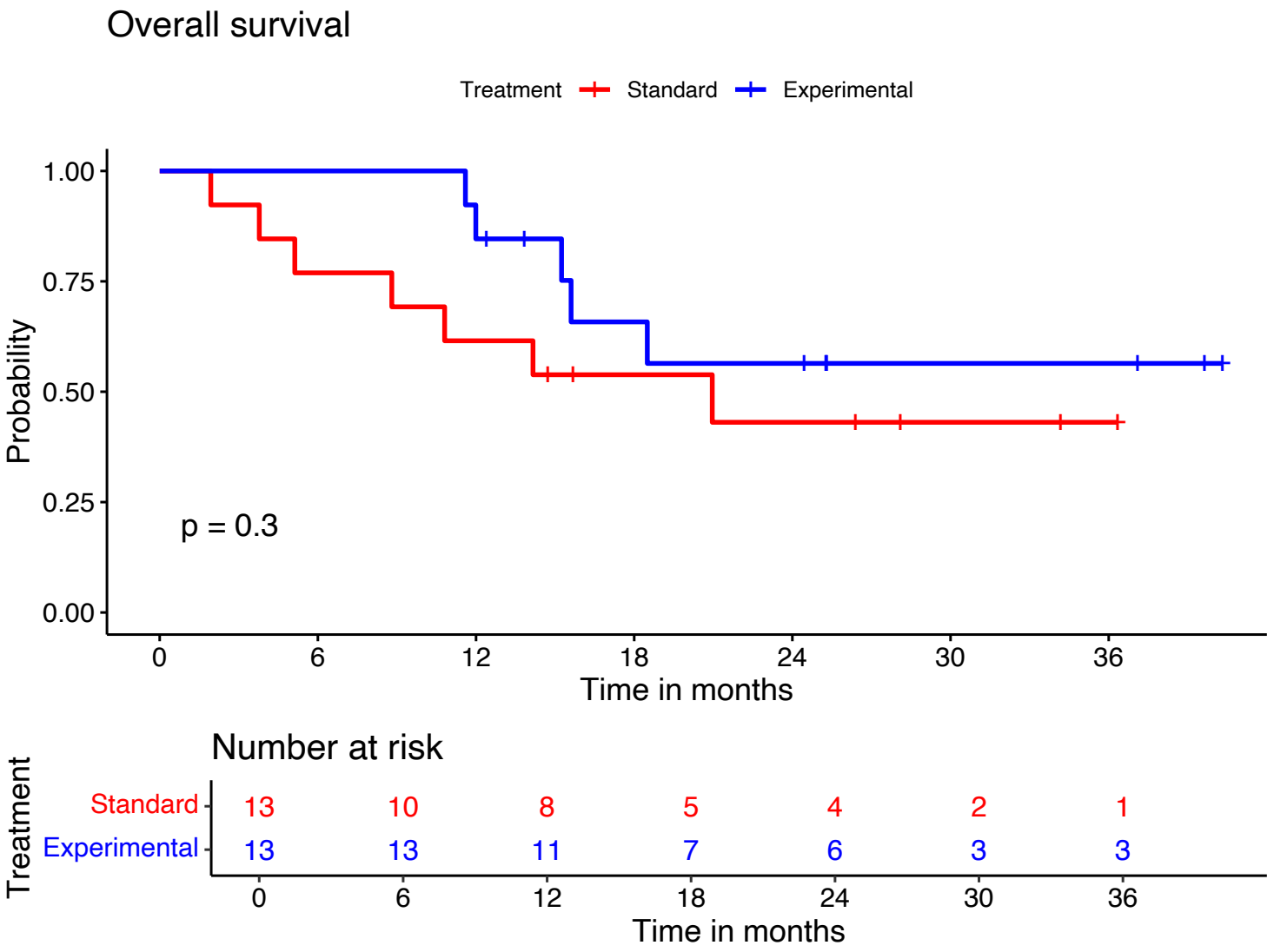
Survival analysis

Kaplan-Meier estimator



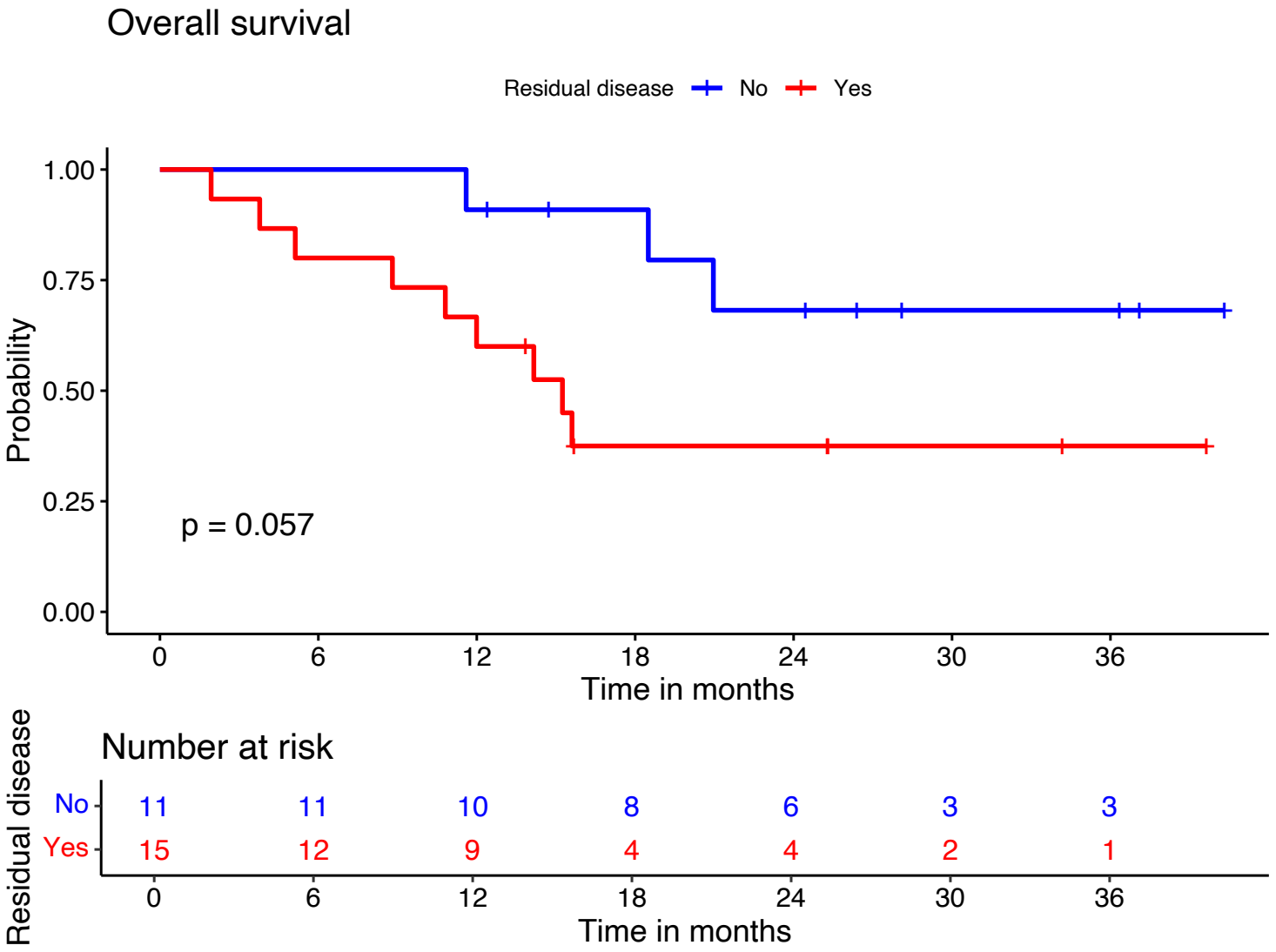
Survival analysis

Kaplan-Meier estimator



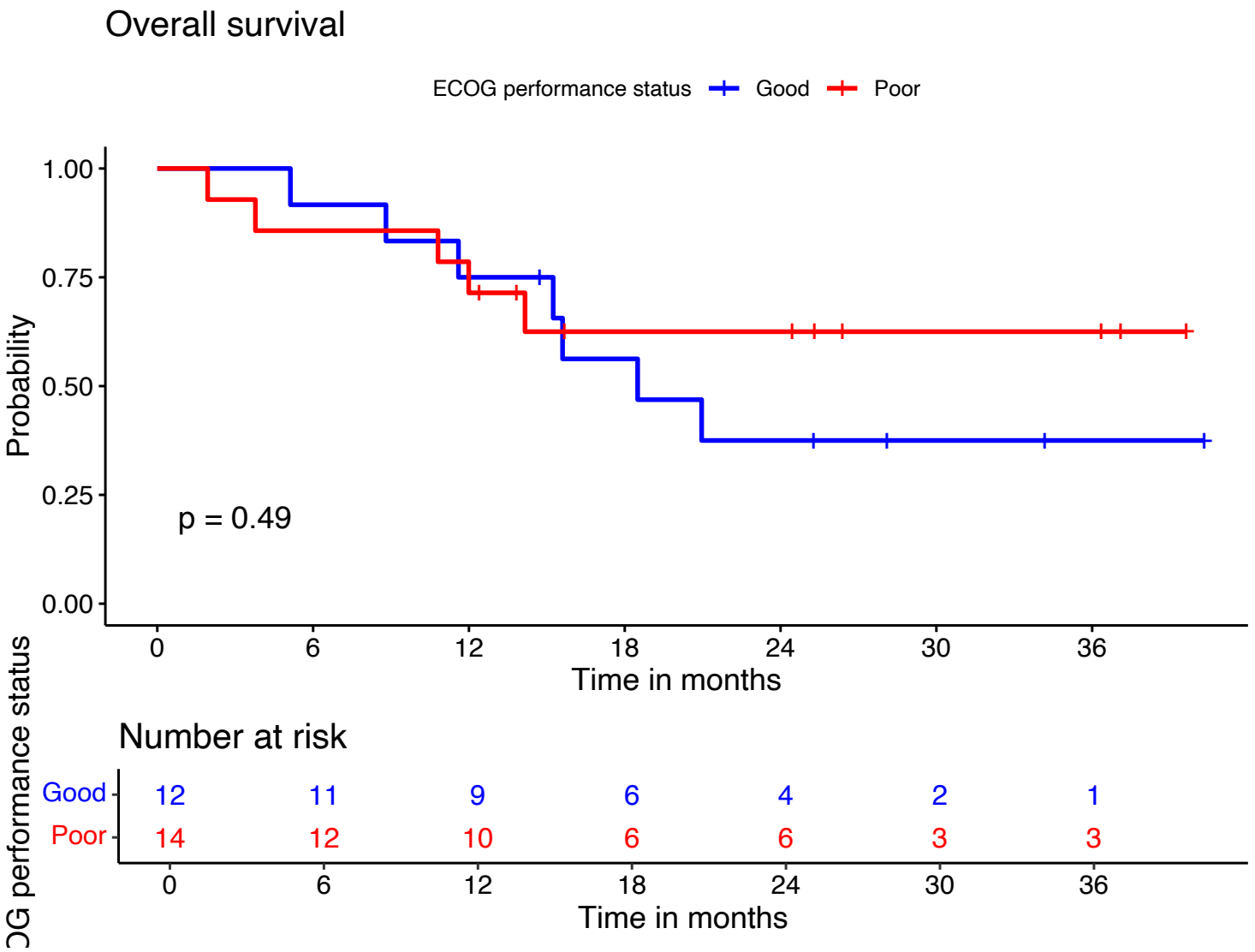
Survival analysis

Kaplan-Meier estimator



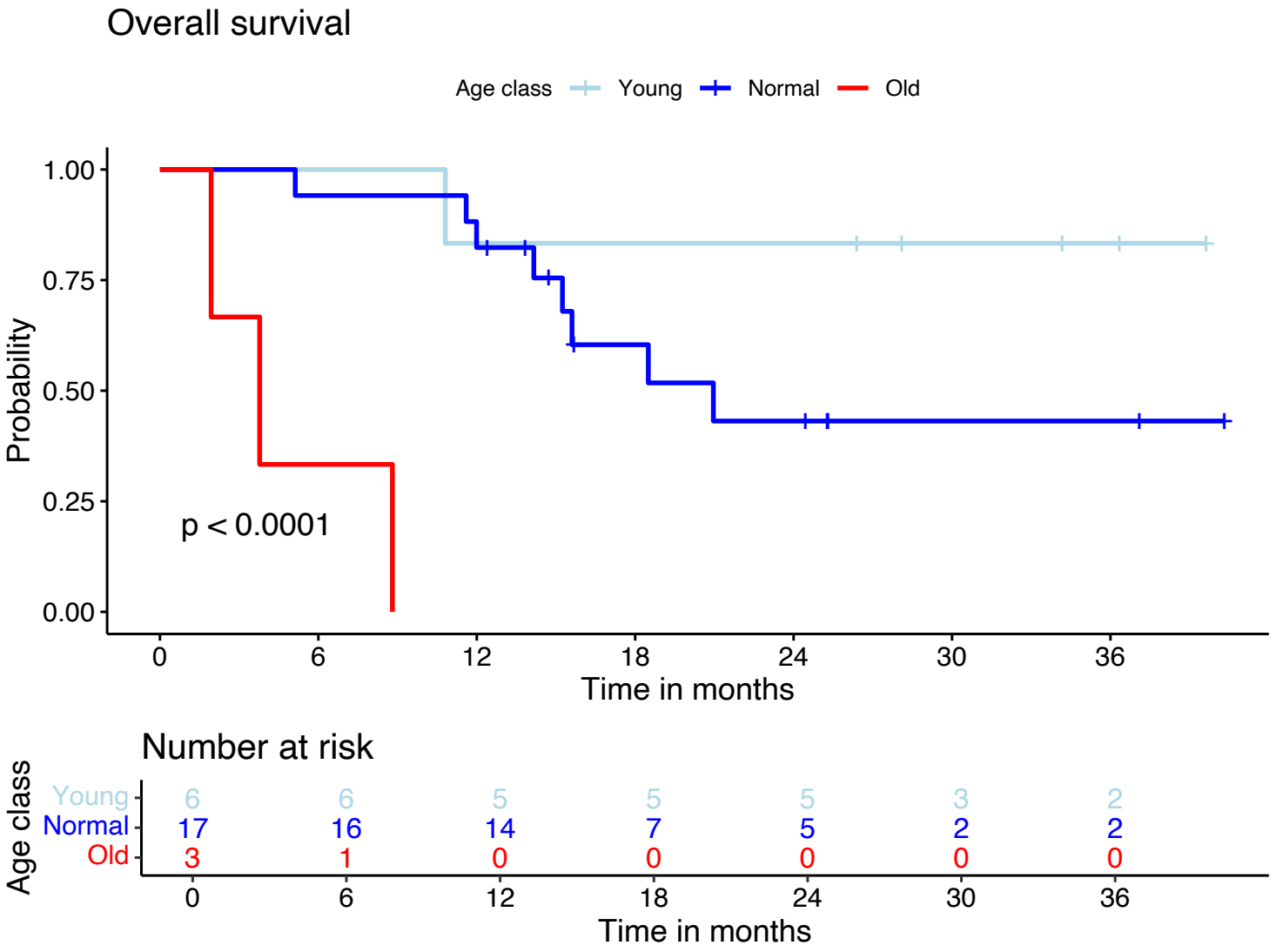
Survival analysis

Kaplan-Meier estimator



Survival analysis

Kaplan-Meier estimator




Thank You!