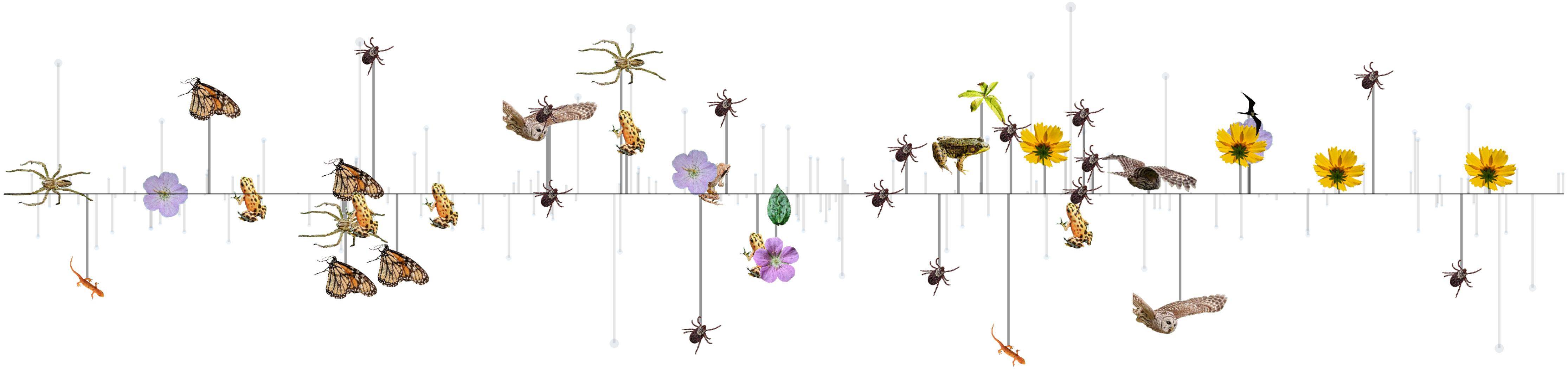
Intro to Quantitative Ecology
UMass Amherst – Michael France Nelson

Deck 6: Inferential Statistics and Differences Between Two Groups



If you’re feeling stuck, come to office hours!

Schedule
Mike • Mondays 3:00 – 4:30 • By appt. **Ana** • Fridays 4:00 – 5:00 • By appt.

Plot types for differences

What types of graphs do we know for plotting differences?

Good choices for plotting differences among groups are:

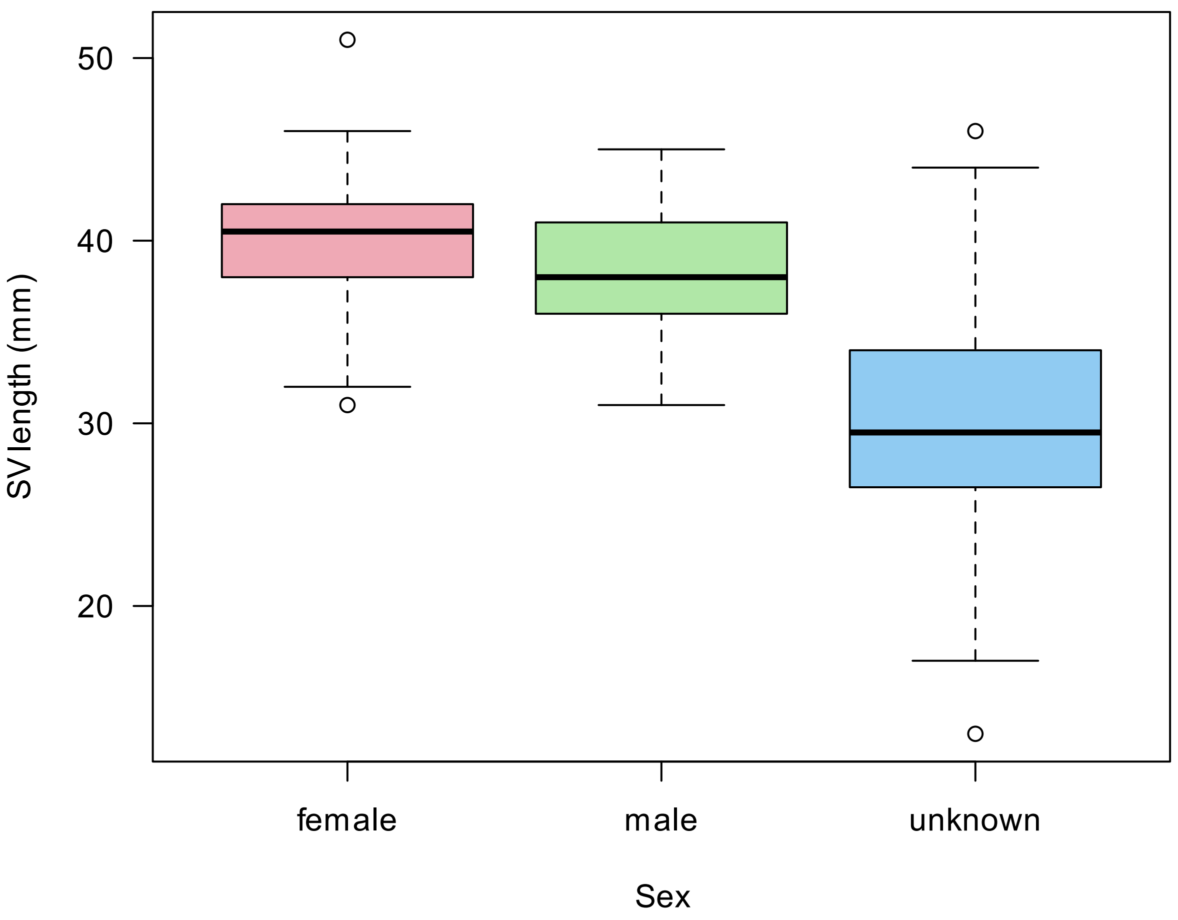
- **Conditional Boxplots**
- Barplots*
- Multi-series density plots or histograms*
- We'll focus on boxplots in this course. They illustrate differences better than barplots and they are easier to code than multi-series histograms/density plots.

Beyond graphs, Towards statistics

- Graphs are powerful tools that provide insight and understanding of the patterns and relationships in the data.
- Graphs alone don't give us the complete answer. We need to **quantify** the relationships we see in our plots.
- What other tools do we have to **support** our conclusions?

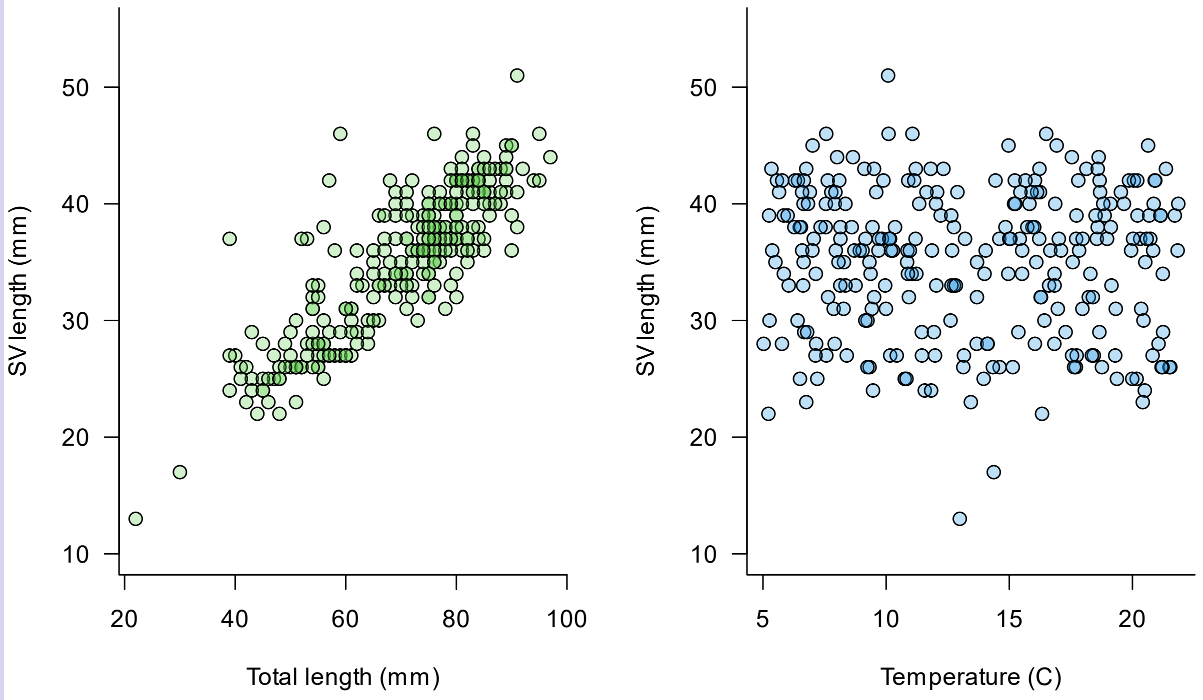
Beyond graphs, Towards statistics

- How can we **quantify** our evidence for relationships?
- Are differences between groups *significant*?
- Are differences between groups *meaningful*?



Beyond graphs, Towards statistics

- How can we **quantify** our evidence for relationships?
- Are associations between 2 variables *significant*?
- Are associations between 2 variables *meaningful*?



Beyond graphs, Towards statistics

- Statistics is the tool we use to formally answer these questions!
- Differences *are/are not* significant?
- Associations *are/are not* significant?

Wait a second… what do we mean when we say significant?

Hypotheses

Frequentist Statistics: Hypothesis Testing

- Frequentist paradigm ideas:
  - Population is infinite, parameters are unknowable
  - Samples are finite, we use sample statistics to make educated guesses about population parameters.
- Hypothesis testing is the heart of frequentist statistics.
  - Quantification of the evidence that something ‘interesting’ is happening in our data.
- Frequentist statistical hypotheses
  - Null hypothesis: There is nothing interesting happening: no associations, differences, or correlations
  - Alternative hypothesis: There is an association, difference, or correlation. This is what we think is actually happening!

Frequentist Statistics: Hypothesis Testing

- We can quantify **relationships** among variables in our data (associations, correlations, differences).
- We can characterize the **uncertainty** or noise in our data.
- Hypothesis testing is an objective way to **quantify the evidence** for any patterns we observe.
- Much more later!

Let’s examine some plots to gain intuition:

Lake Trout

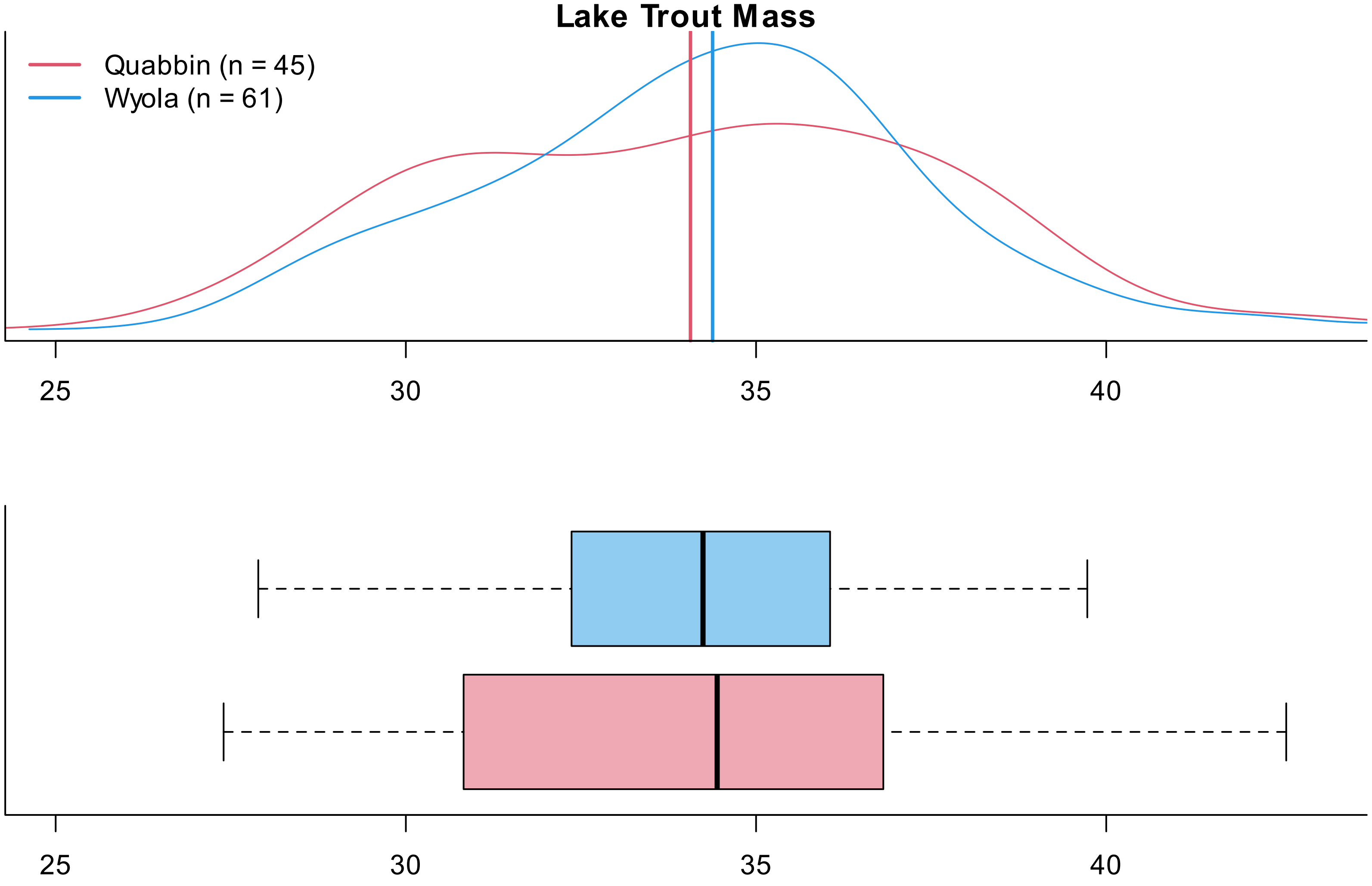
- Scenario: We want to know whether the size of lake trout (*Salvelinus namaycush*) are larger in some Massachusetts lakes than others.
- We have collected trout data for from Wyola Lake and the Quabbin Reservoir in Western Mass.

Lake Trout: *Salvelinus namaycush*

Fish Data: Scenario 1

Ask yourself…

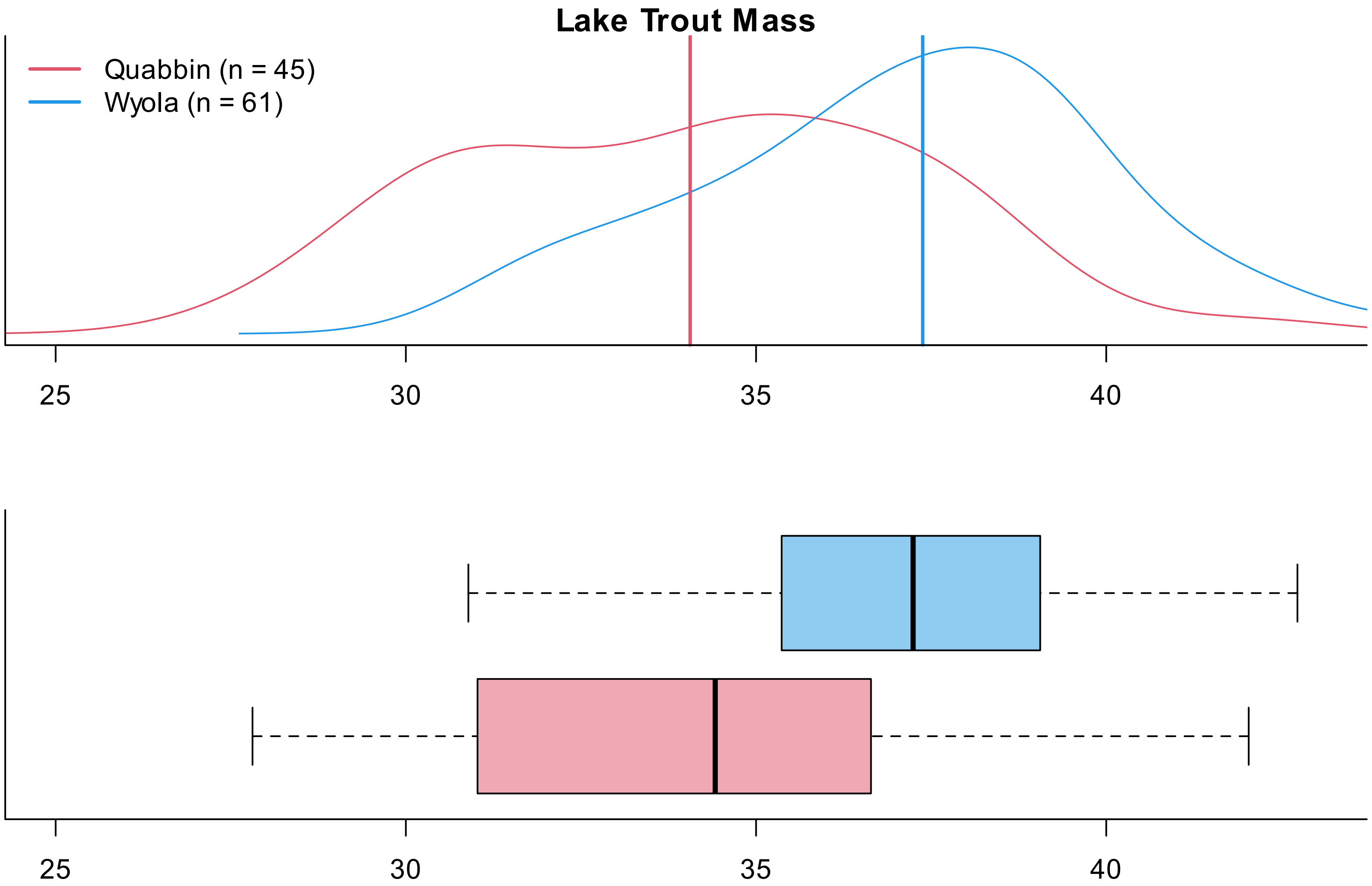
- Are differences between lakes *significant*?
- Are differences between lakes *meaningful*?



Fish Data: Scenario 2

Ask yourself…

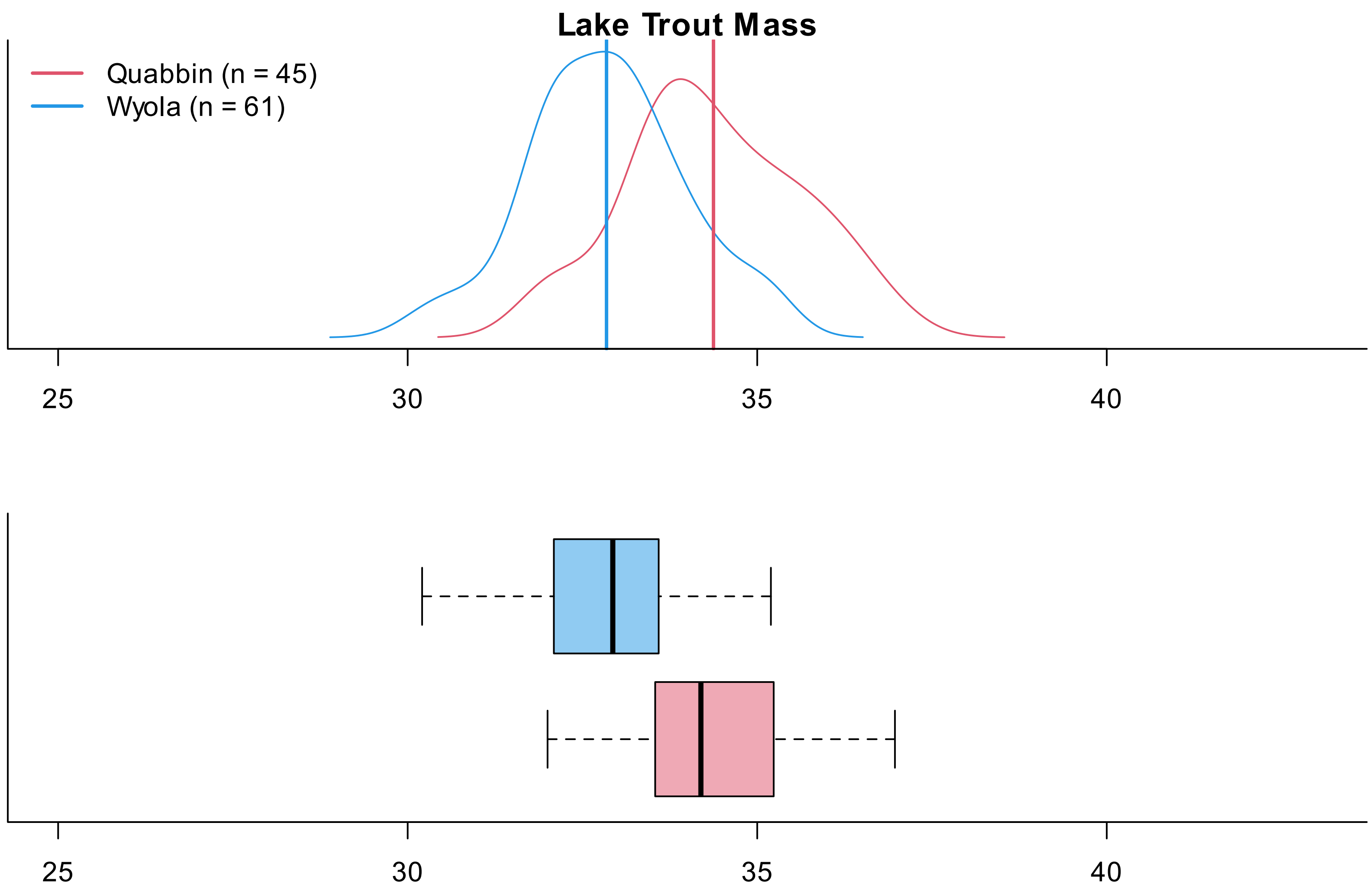
- Are differences between lakes *significant*?
- Are differences between lakes *meaningful*?



Fish Data: Scenario 3

Ask yourself…

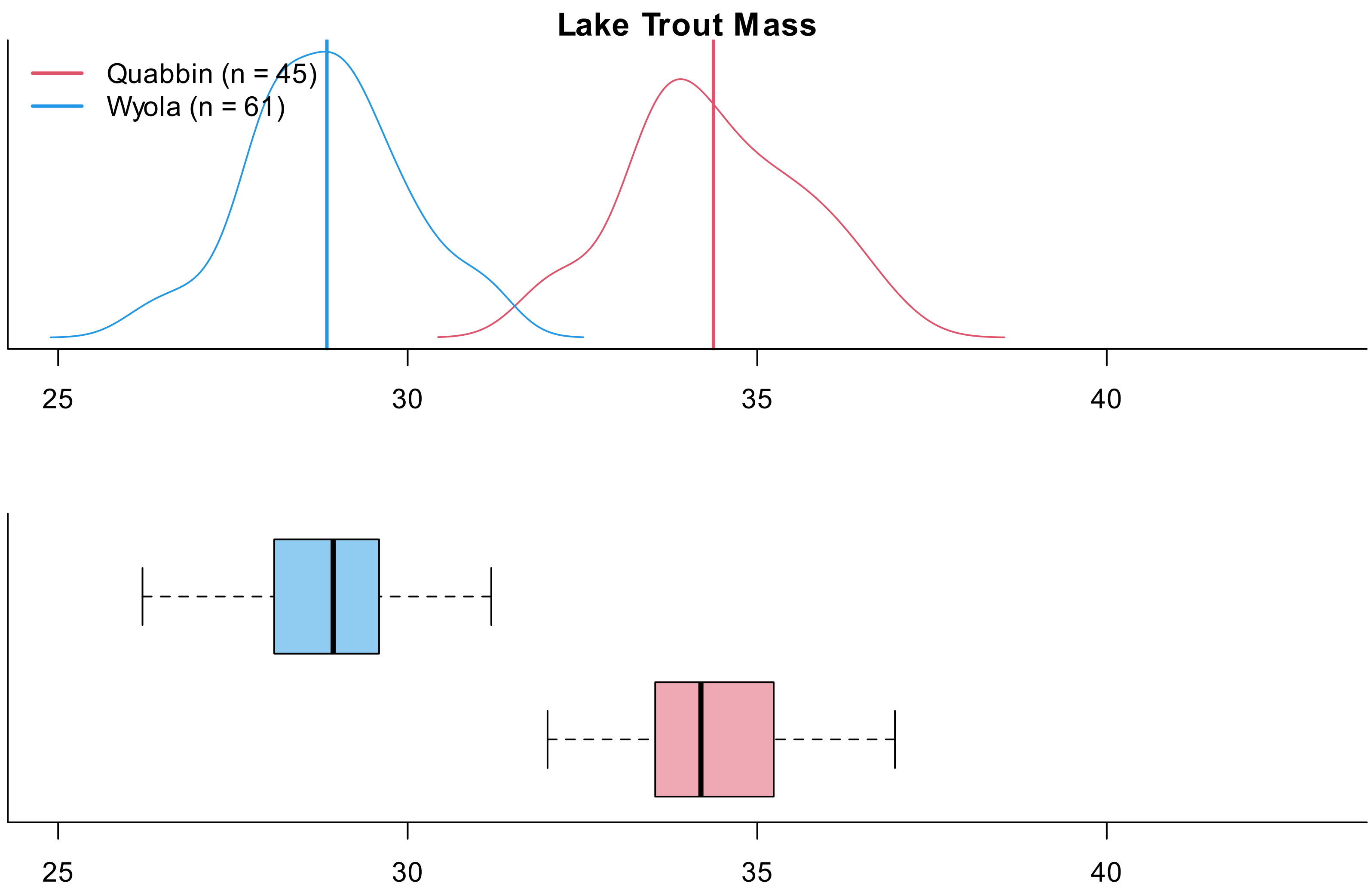
- Are differences between lakes *significant*?
- Are differences between lakes *meaningful*?



Fish Data: Scenario 4

Ask yourself…

- Are differences between lakes *significant*?
- Are differences between lakes *meaningful*?



Tests for differences

Often, we want to know if the *means* or *medians* of two or more groups are different.

- Are the differences *statistically significant*?

To determine the significance of differences between **two groups**, we need a statistical test:

- *t-test*
- *U-test*

Tests for differences: intuition

To build intuition about testing for differences between two groups, let’s consider:

- What information would we need to know?
- What kinds of evidence would support our conclusion?
- How do we define *different*?

Tests for differences: intuition

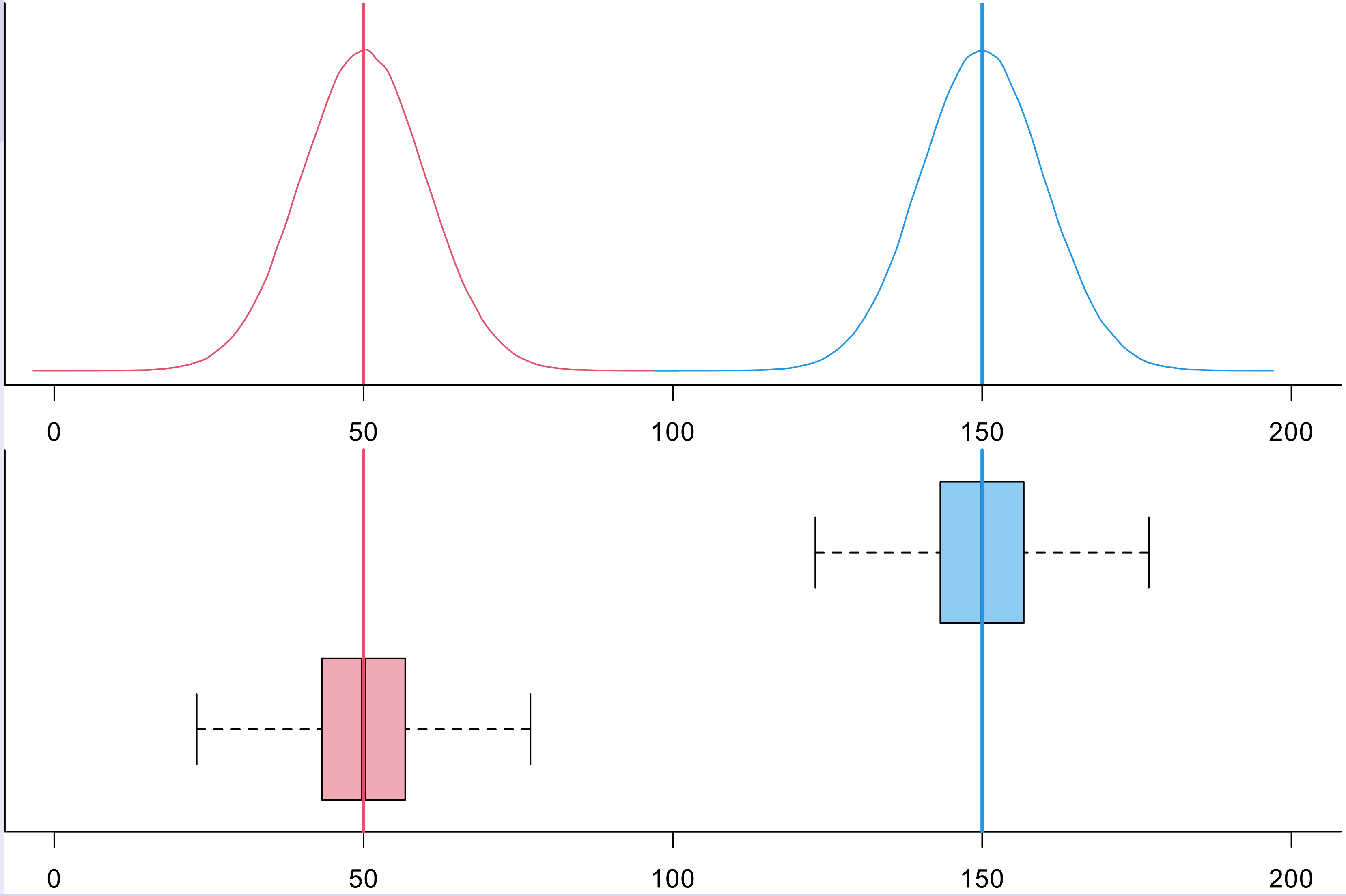
To build intuition about testing for differences between two groups, let's consider:

- What information would we need to know?
  - Center and spread of each group?
  - Difference in means?
- What kinds of evidence would support our conclusion?
  - Large difference in means?
- How do we define *different*?

Intuition: Large Difference In Means

Are the differences significant?

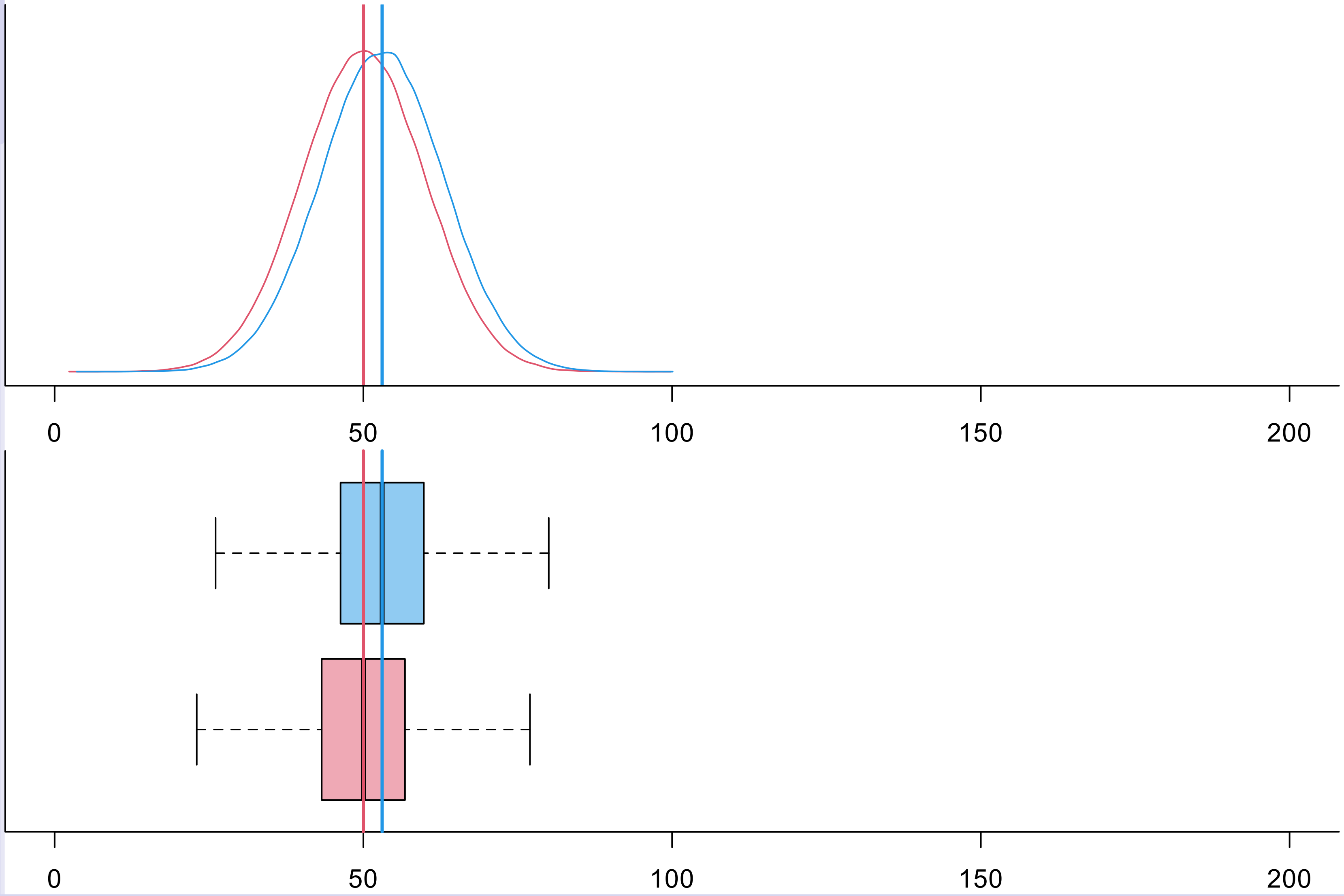
- Large difference in means.
- Very little overlap in the distributions.
- You observed an individual that weighed 50 grams, which group does it belong to?
- Difference is probably significant!



Intuition: Small Difference In Means

Are the differences significant?

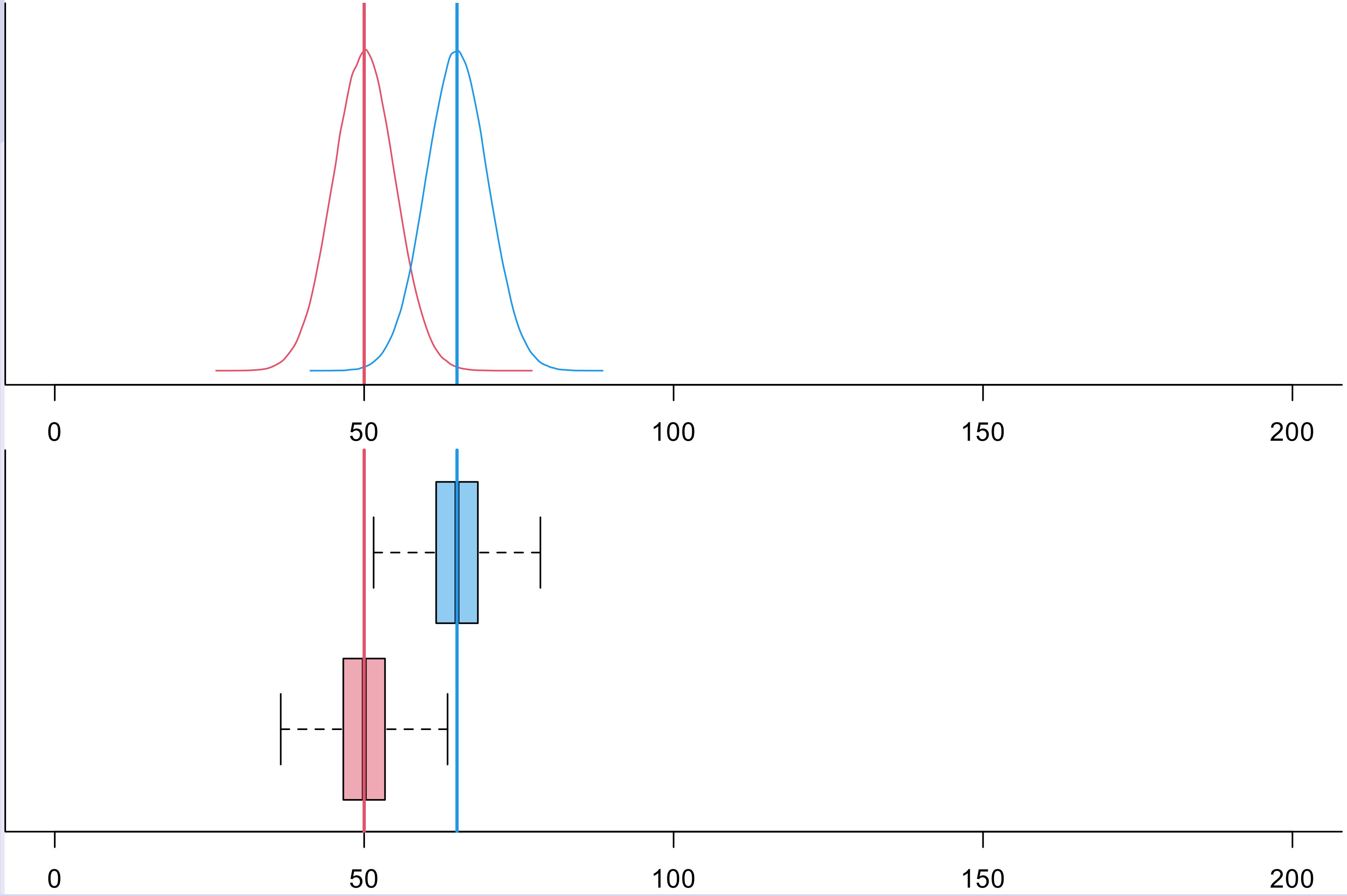
- Small difference in means.
- Distributions are mostly overlapping.
- You observed an individual that weighed 50 grams, which group does it belong to?
- Difference is probably not significant!



Intuition: Small Difference In Means

Are the differences significant?

- Small-ish difference in means, but distributions are narrow.
- Distributions are slightly overlapping.
- You observed an individual that weighed 50 grams, which group does it belong to?
- Difference is probably significant!



Differences: t-test

Purpose:

- compare the means of two samples (say a and b)

Assumptions:

- both samples normally distributed
- both samples have equal variances

Differences: t-test

Purpose:

- compare the means of two samples (say a and b)

Assumptions:

- both samples normally distributed
- both samples have equal variances (we can work around this one)

$$t = \frac{|\bar{x}_a - \bar{x}_b|}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

- t :the t-statistic
- $\bar{x}$:sample mean
- s: sample standard deviation
- n: sample size

Differences: t-test

Purpose:

- compare the means of two samples (say a and b)

Assumptions:

- both samples normally distributed
- both samples have equal variances

$$t = \frac{|\bar{x}_a - \bar{x}_b|}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

- if $|\bar{x}_a - \bar{x}_b|$ is large, then t is large
- if $\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$ is large, then t is small

Differences: t-test

Purpose:

- compare the means of two samples (say a and b)

Assumptions:

- both samples normally distributed
- both samples have equal variances

$$t = \frac{|\bar{x}_a - \bar{x}_b|}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

- if $|\bar{x}_a - \bar{x}_b|$ is large, then t is
- if $\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$ is large, then t is

What if there are more than 2 groups???

- T-tests can only handle one or two groups.
- Analysis of Variance (ANOVA) works for 3 or more groups.
  - Conceptually similar to t-test (it tests for differences among groups).
  - How do you know which group or groups are different? This question is more complicated with 3 groups!
  - P-value interpretation is similar.

Multi-Panel Plots and Formula Notation

But first, some background!

R's Formula Notation

- Formula notation: a powerful syntax that reflects our model, or plot, structure!
- The syntax is similar to how we express mathematical functions:
  - Response variable on the left.
  - Explanatory variables on the right.
- The tilde character: ~
  - Like the equals sign in a function.
  - Symbolizes that we propose a relationship between the variable on the left (the response variable) and the variables on the right (the predictors).

Formulas in R

```
penguins$flipper_length_mm ~ penguins$body_mass_g
```

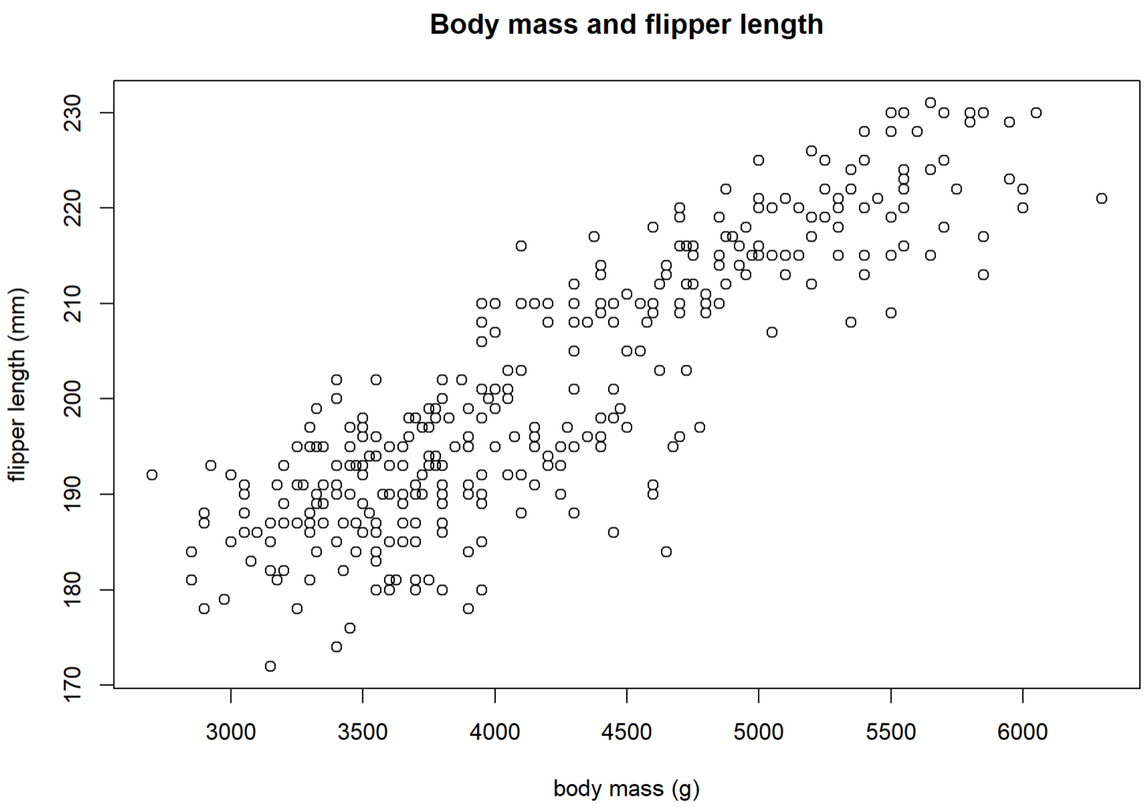
Mathematical Formulas	Formulas in R
Recall the equation of a line: y = mx + b • We can think about this in terms of the penguins: • Flipper length = m*(body mass) + (a constant)	We could write the flipper/body mass relationship using the formula notation in R using the tilde symbol. Read the tilde as: • “y explained by x” • “y in terms of x”

Formula example

```
plot(
  penguins$flipper_length_mm ~ penguins$body_mass_g,
  main = "Body mass and flipper length",
  xlab = "body mass (g)",
  ylab = "flipper length (mm)"
)
```

Alternate syntax using the data argument:

```
plot(
  flipper_length_mm ~ body_mass_g,
  data = penguins,
  main = "Body mass and flipper length",
  xlab = "body mass (g)",
  ylab = "flipper length (mm)"
)
```



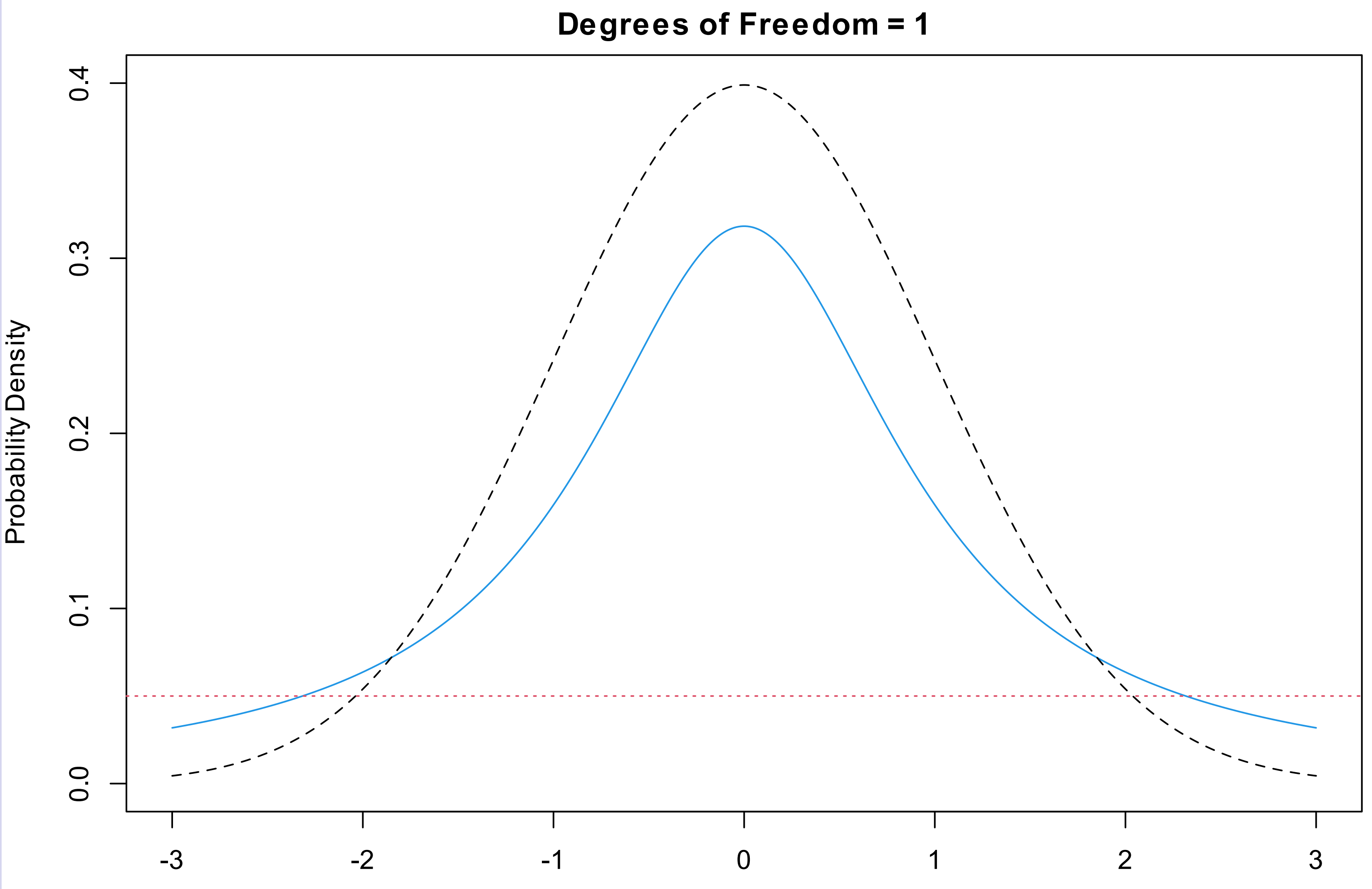
Multi-Panel Plotting

- Note the use of the par() function. See the example in the assignment walkthrough.

Understanding the *t-distribution*:

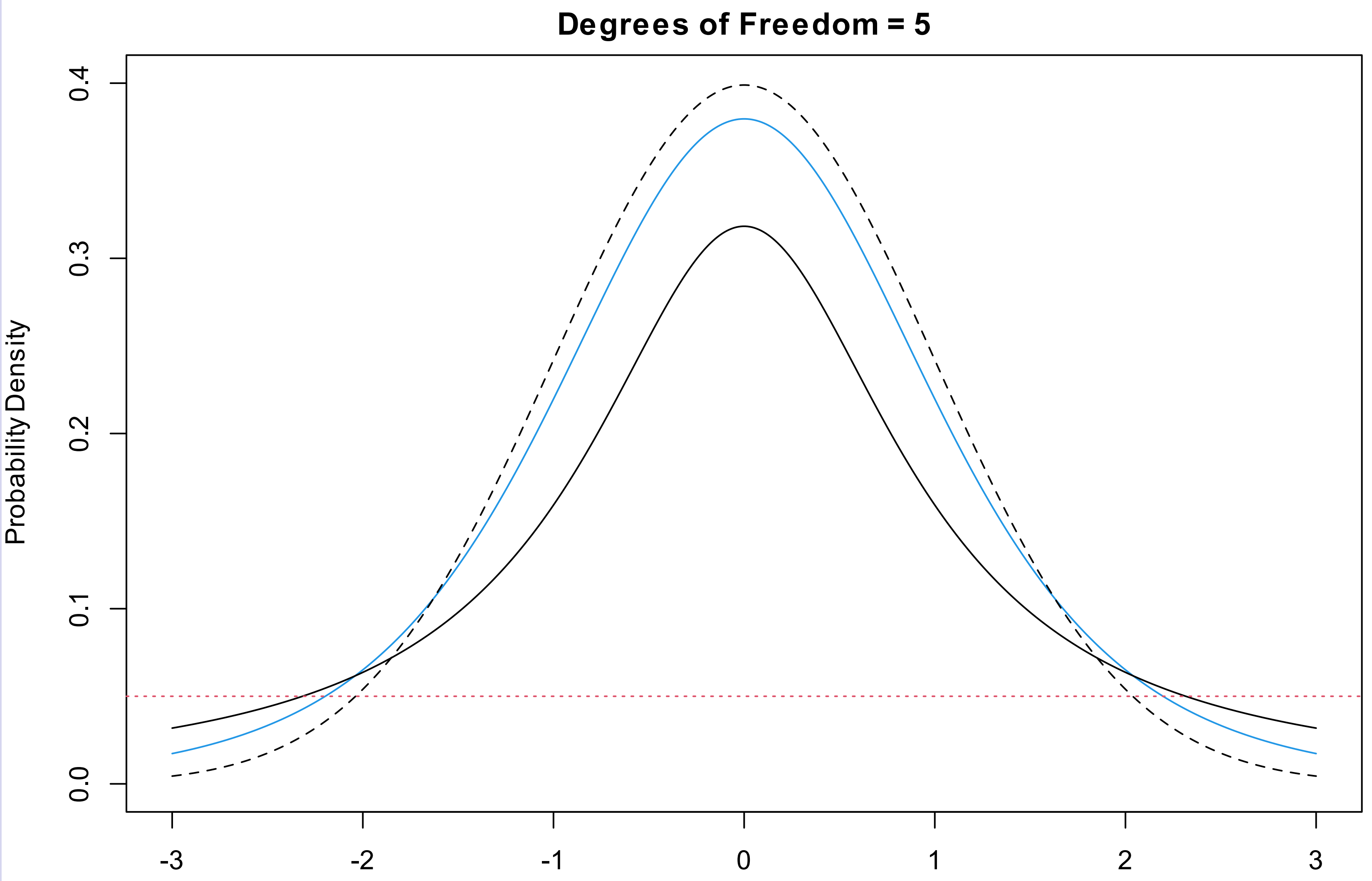
The t-distribution has fatter tails than the normal distribution.

The t-distribution has more uncertainty because we're estimating from data in a sample (rather than a population).



Understanding the *t-distribution*:

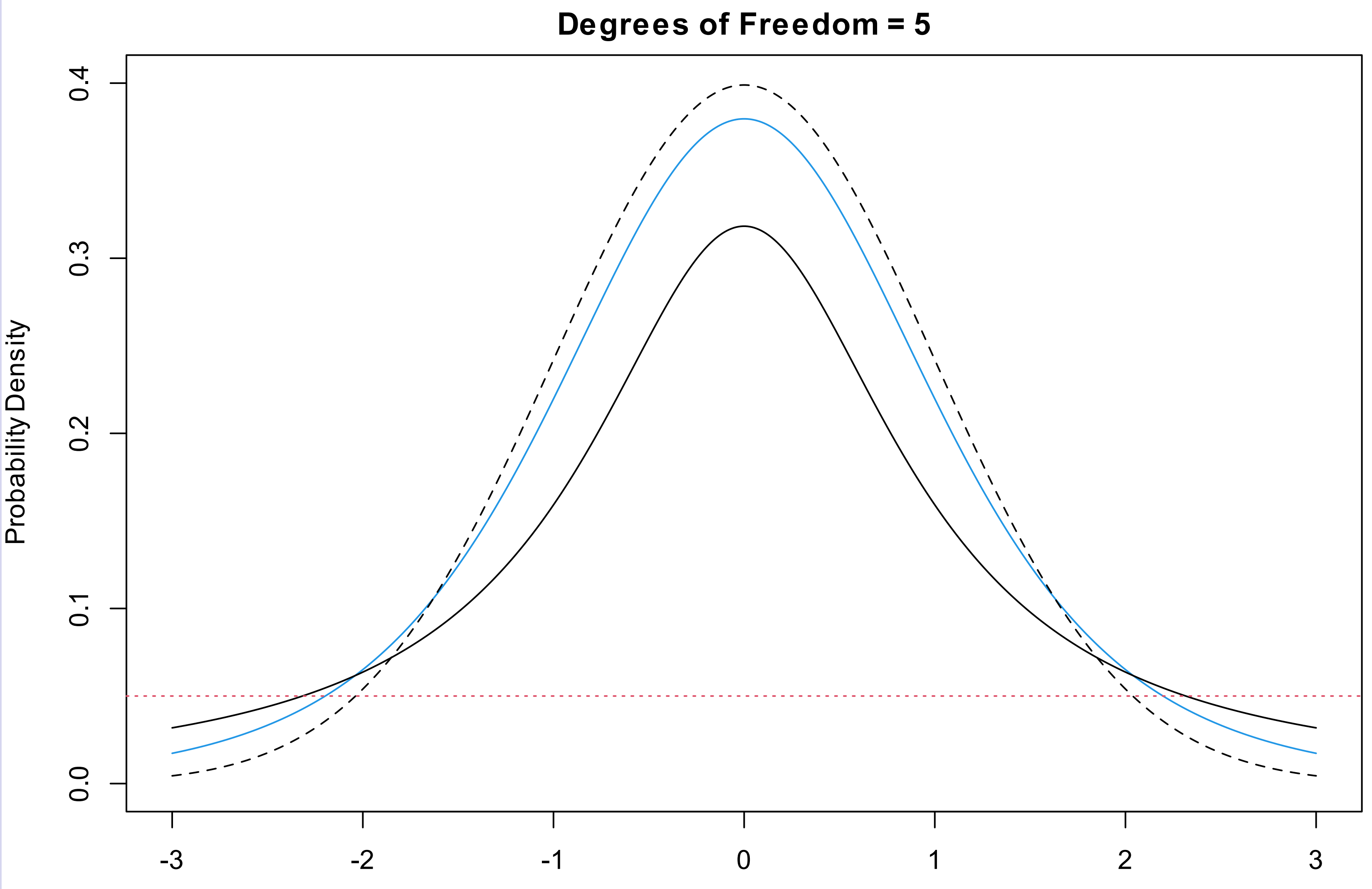
The t-distribution gets closer to the normal distribution with more observations.



Understanding the *t-distribution*:

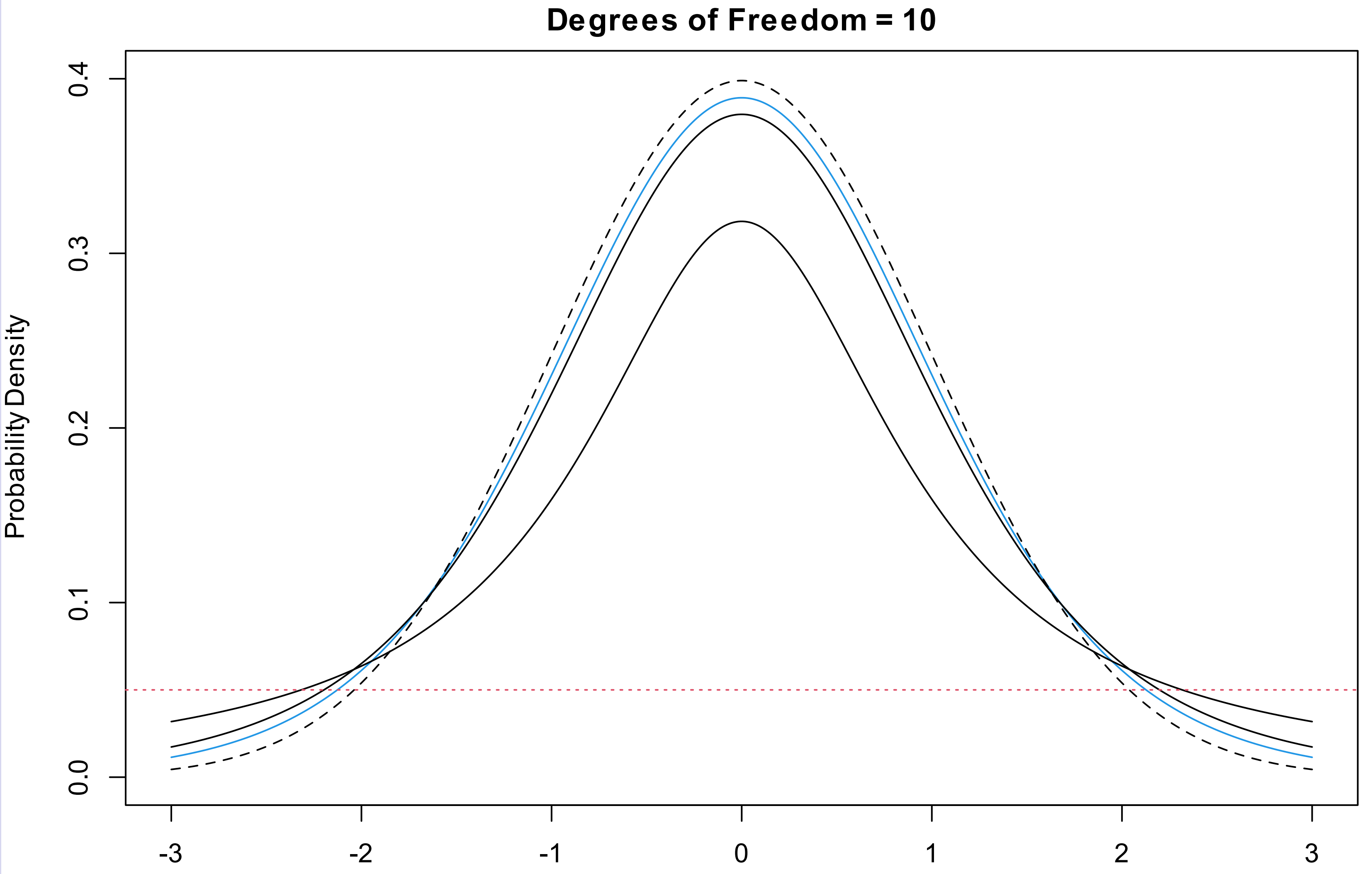
The t-distribution gets closer to the normal distribution with more observations.

This should make intuitive sense since larger samples are more likely to be representative of the population!



Understanding the *t-distribution*:

The t-distribution is nearly identical to the normal distribution with 30 df.



Differences: t-test

Understanding the *t-distribution*:

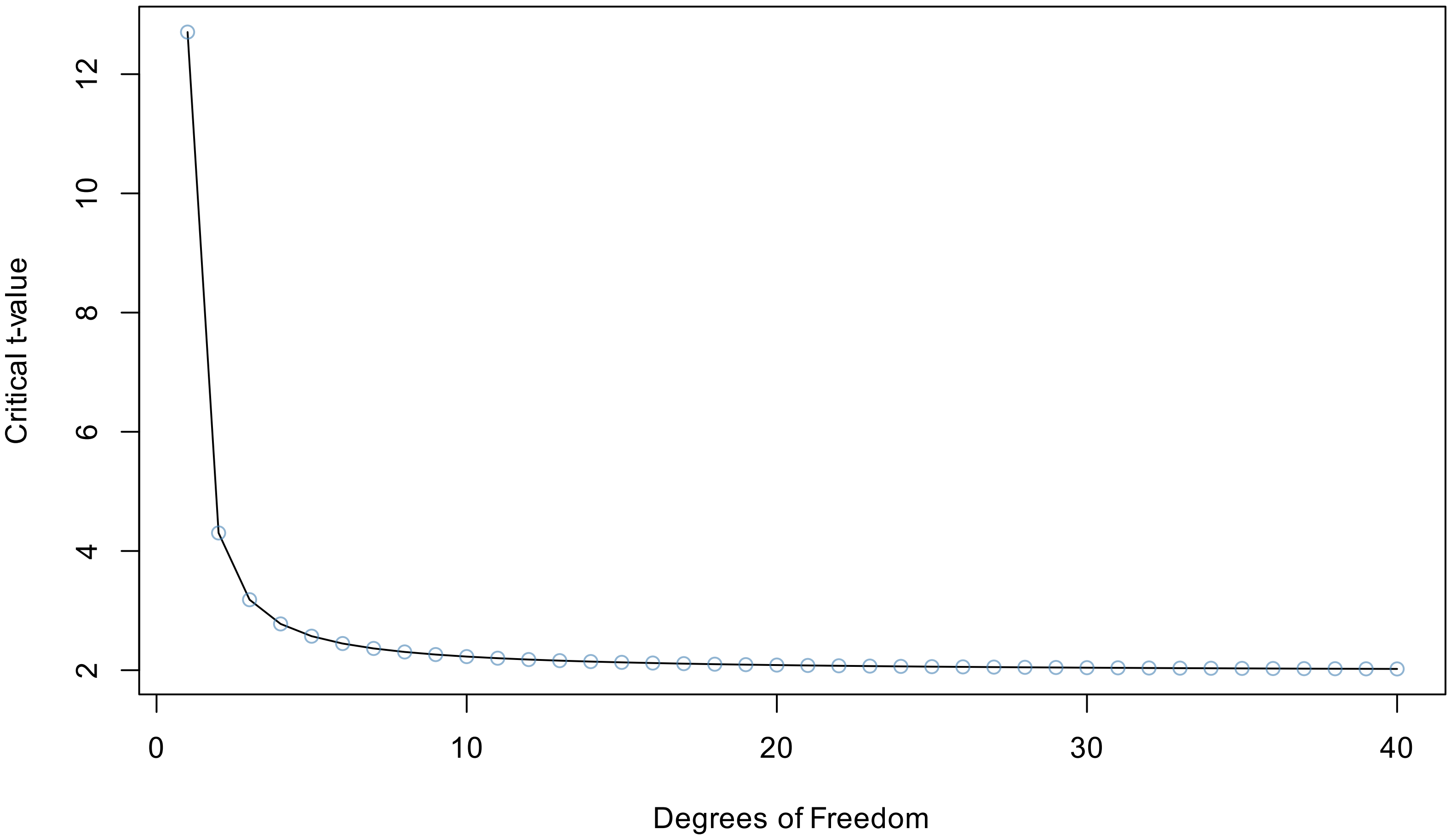
- whether a difference is significant depends on:
- the *t-statistic*
- degrees-of-freedom ($n_a - 1 + n_b - 1$)
- larger *t-statistics* more likely to be significant.

Understanding the *t-distribution*:

Large sample sizes have smaller critical t-values.

- Degrees of freedom is related to the sample size.
- But note the diminishing returns as you increase n.

95% Critical t-value for different degrees of freedom



Understanding the *p-value*:

p-value is the probability of observing a *t-statistic* as extreme as we did by chance *if the null hypothesis were true*.

- if *p-value* is lower than significance level (e.g. 5%):
  - difference is significant
  - reject the null hypothesis
- we don't *accept* the alternative hypothesis
  - But we can say we have good evidence in favor of the alternative over the null.
- P-values are always interpreted in reference to the null hypothesis

Differences: t-test

Which *t-test*?

- standard *t-test*
- compare two independent samples
- both normally distributed
- equal (similar) variances
- samples sizes can be the same or not

Differences: t-test Formula Revisited

Components of the t-test

$$t = \frac{|\bar{x}_a - \bar{x}_b|}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

- t :the t-statistic
- $\bar{x_a}$:sample mean, group a
- s_a^2: sample variance deviation, group a
- n_a: sample size, group a

T-test intuition

Numerator	Denominator
• Numerator is the difference in means • If numerator is large, t will be large	• Denominator is related to the sample dispersion • If groups have lots of dispersion, i.e. the numbers are very variable, the denominator will be large. • If denominator is large, t will be small.

P-value intuition

The p-value can be hard to understand, we'll revisit its meaning many times.

Some ways to think about the p-value:

- A low p-value (0.05 or lower) means we have observed something significant, or perhaps interesting.
- If you observe an unlabeled individual, how hard is it to decide which group it goes in? A low p-value means that it is easy to guess.
- How much overlap is there between the distributions of the group? Minimal overlap corresponds to low p-values.
- False positive rate: What is the probability that the pattern we observed is due to chance (i.e. sampling error)?
- Null hypothesis: A low p-value means we can reject the null hypothesis.

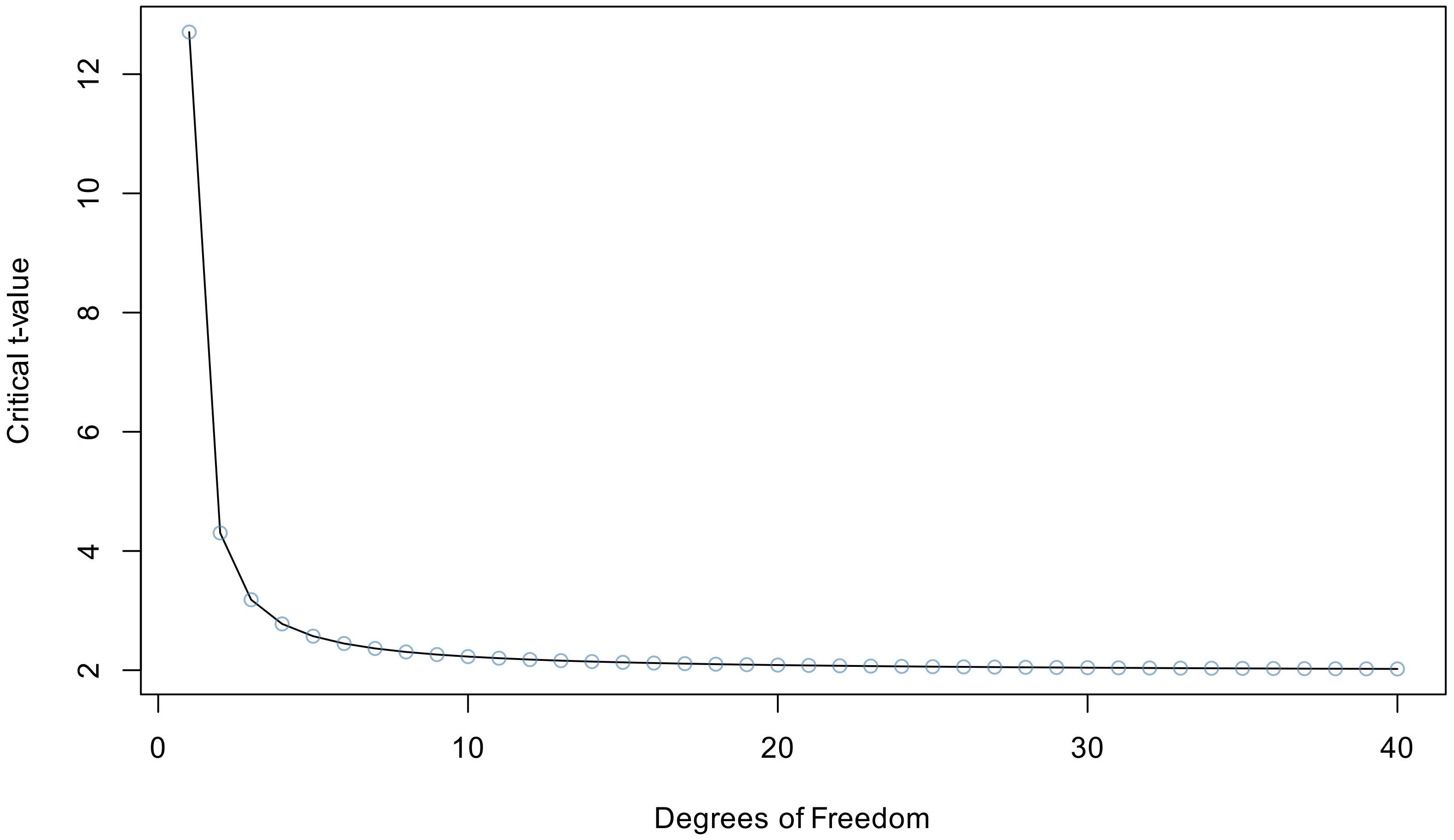
*Note, all of these (except the last two) are not technical definitions, they are meant to build intuition.

Understanding the *t-distribution*:

Large sample sizes have smaller critical t-values.

- Degrees of freedom is related to the sample size.
- But note the diminishing returns as you increase n.
- T-tests on the following slides used 30 d.f.
  - Requires a t-value of around 2 to be significant

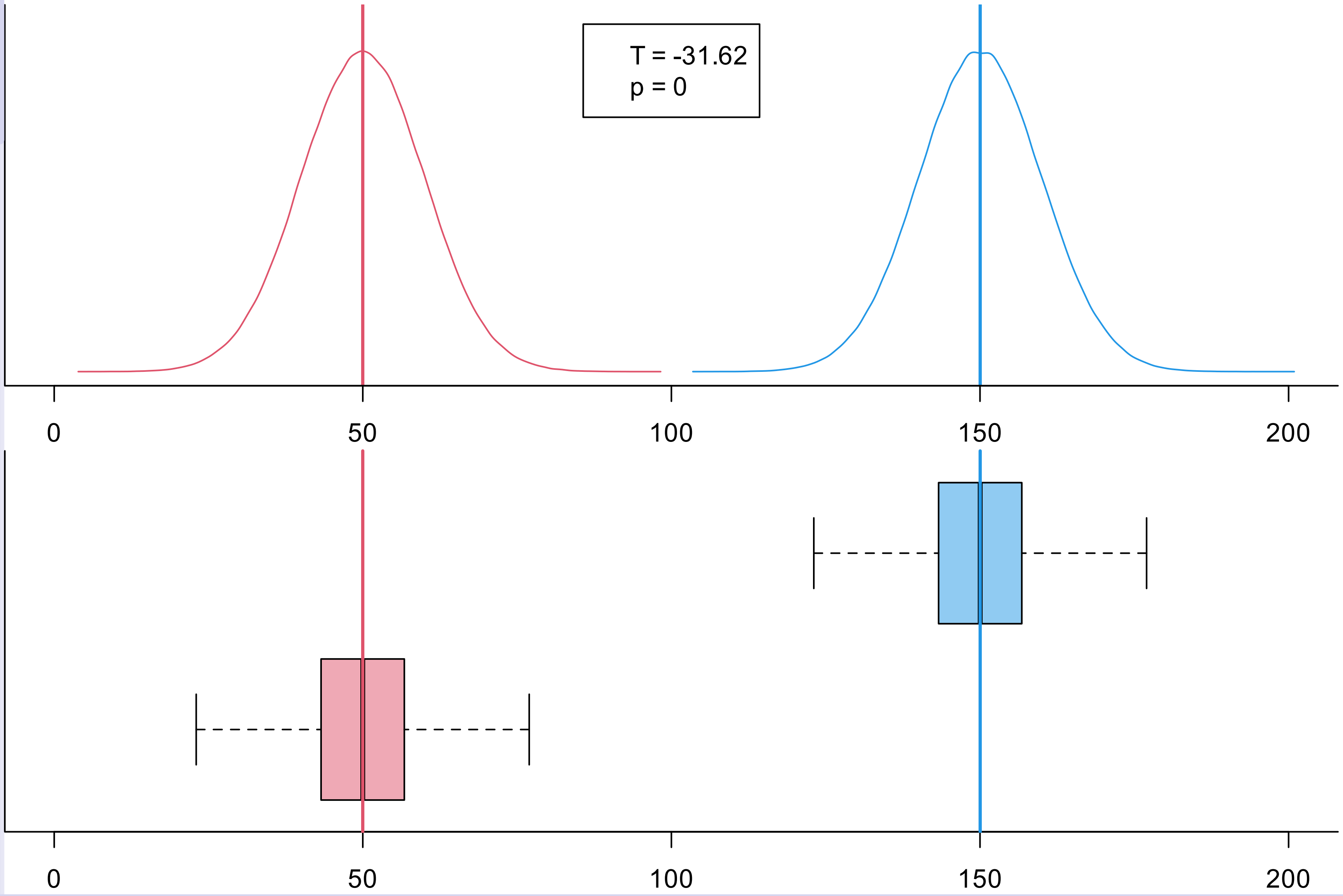
95% Critical t-value for different degrees of freedom



Intuition: Large Difference In Means

Are the differences significant?

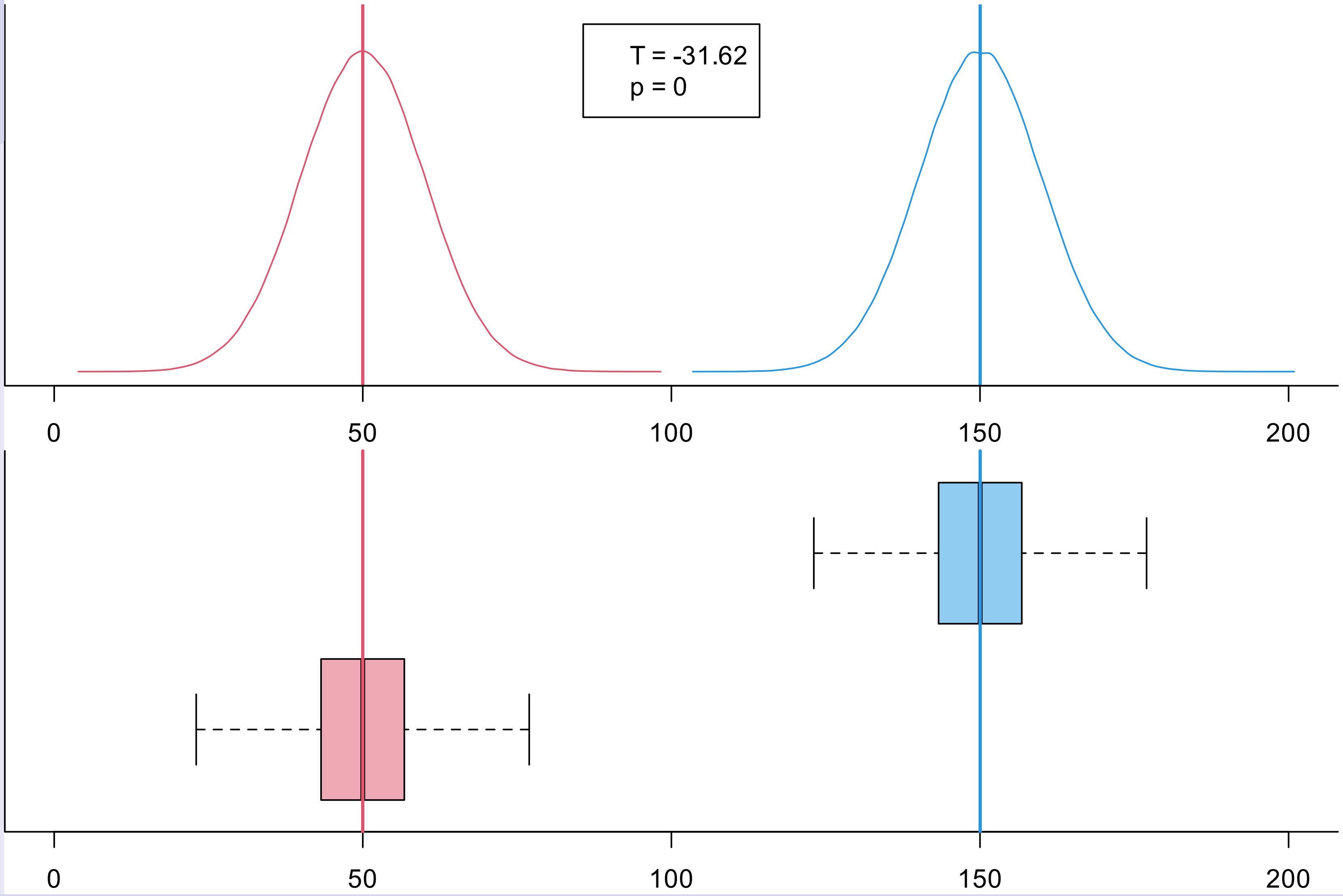
- Very large numerator (difference in means)
- Moderate denominator (sample variance)
- Large t-value



Intuition: Large Difference In Means

Are the differences significant?

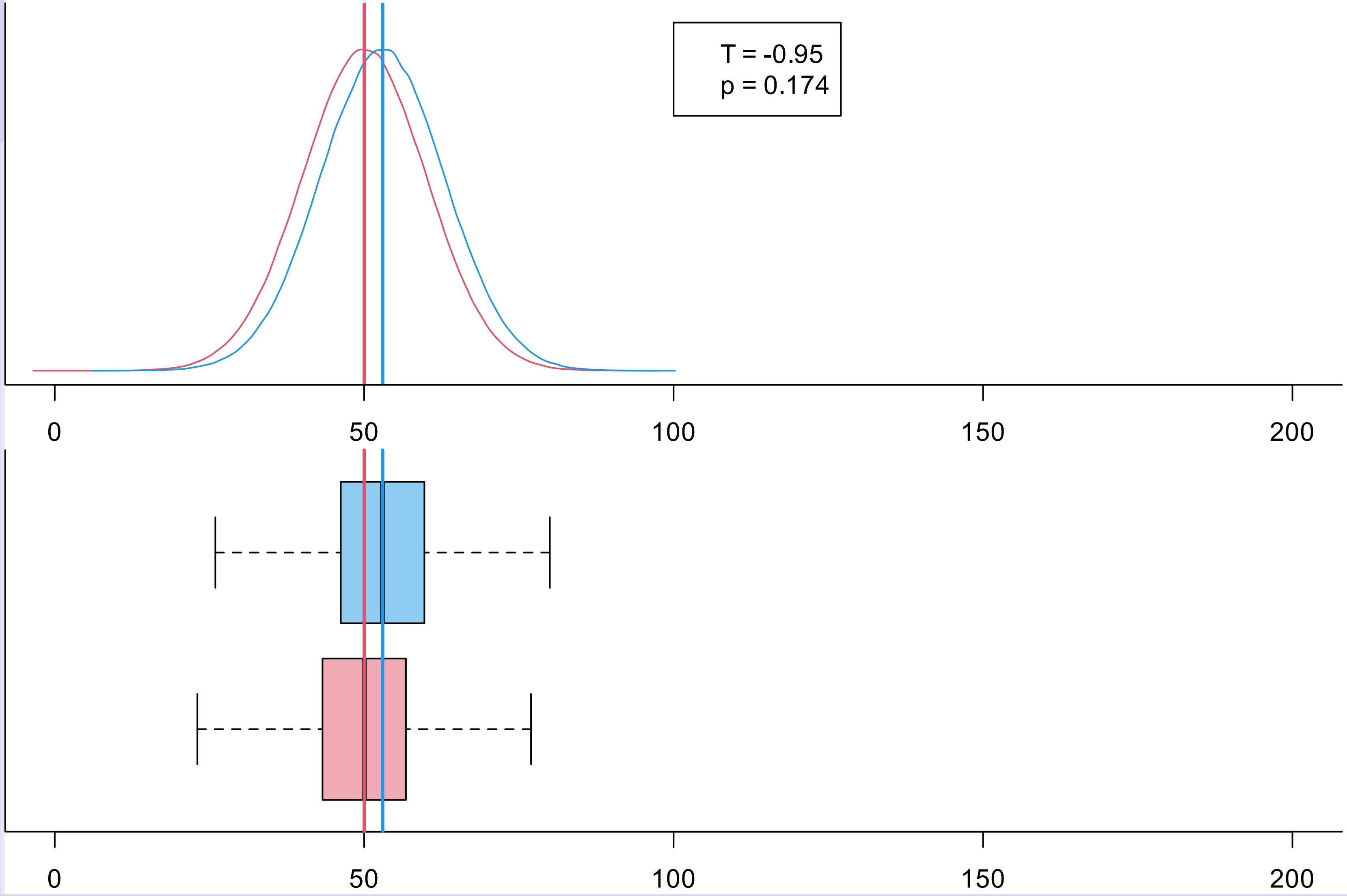
- Large difference in means
  - Big separation between group distributions
- Dispersion is moderate
- T is large (and negative)
- Very low p-value



Intuition: Small Difference In Means

Are the differences significant?

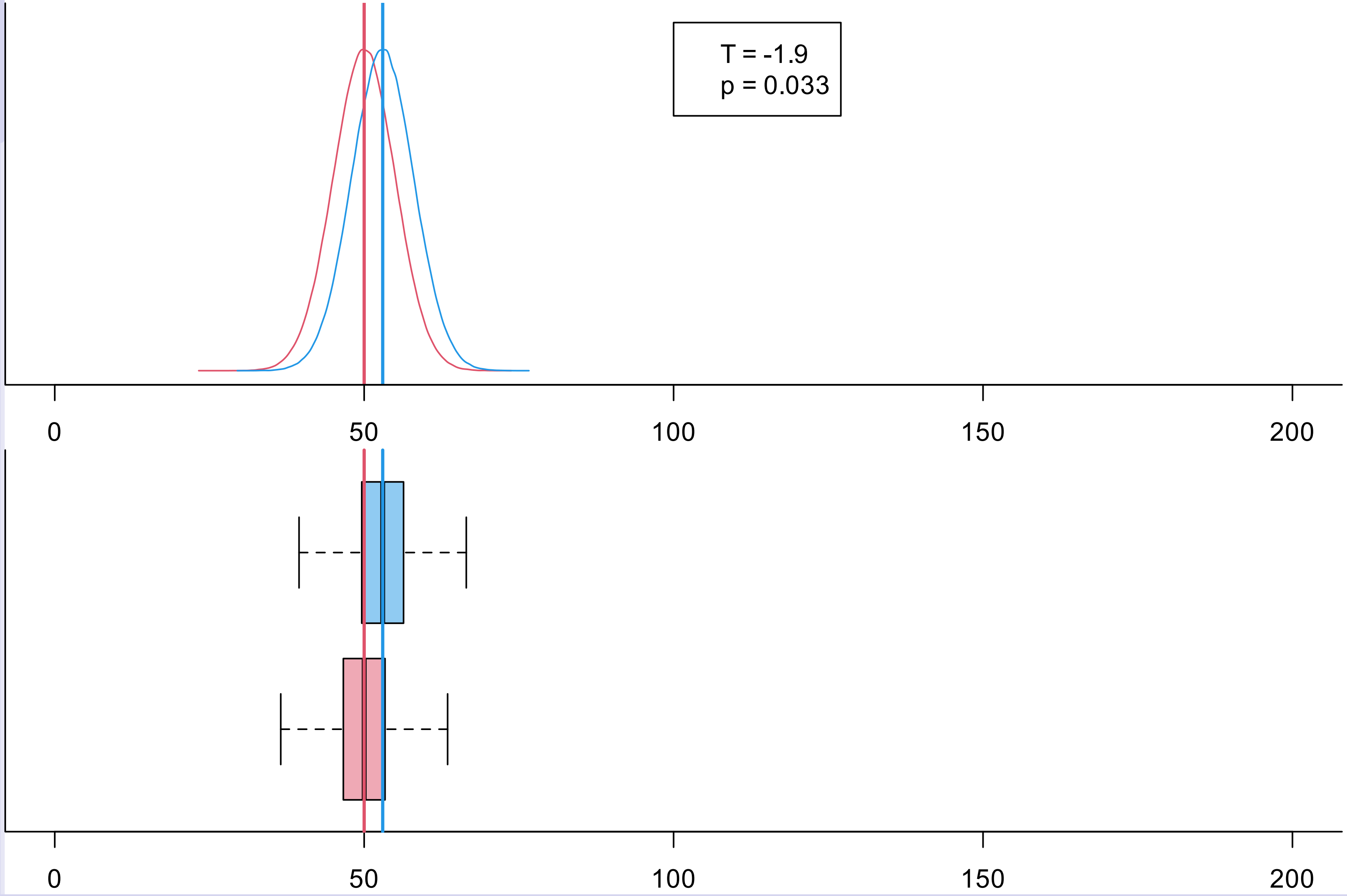
- Small difference in means.
- Dispersion is moderate
- Lots of overlap in the distributions
- T is small (and negative)
- P-value of 0.17 is not significant.
- What is the null hypothesis?



Intuition: Small Difference In Means

Are the differences significant?

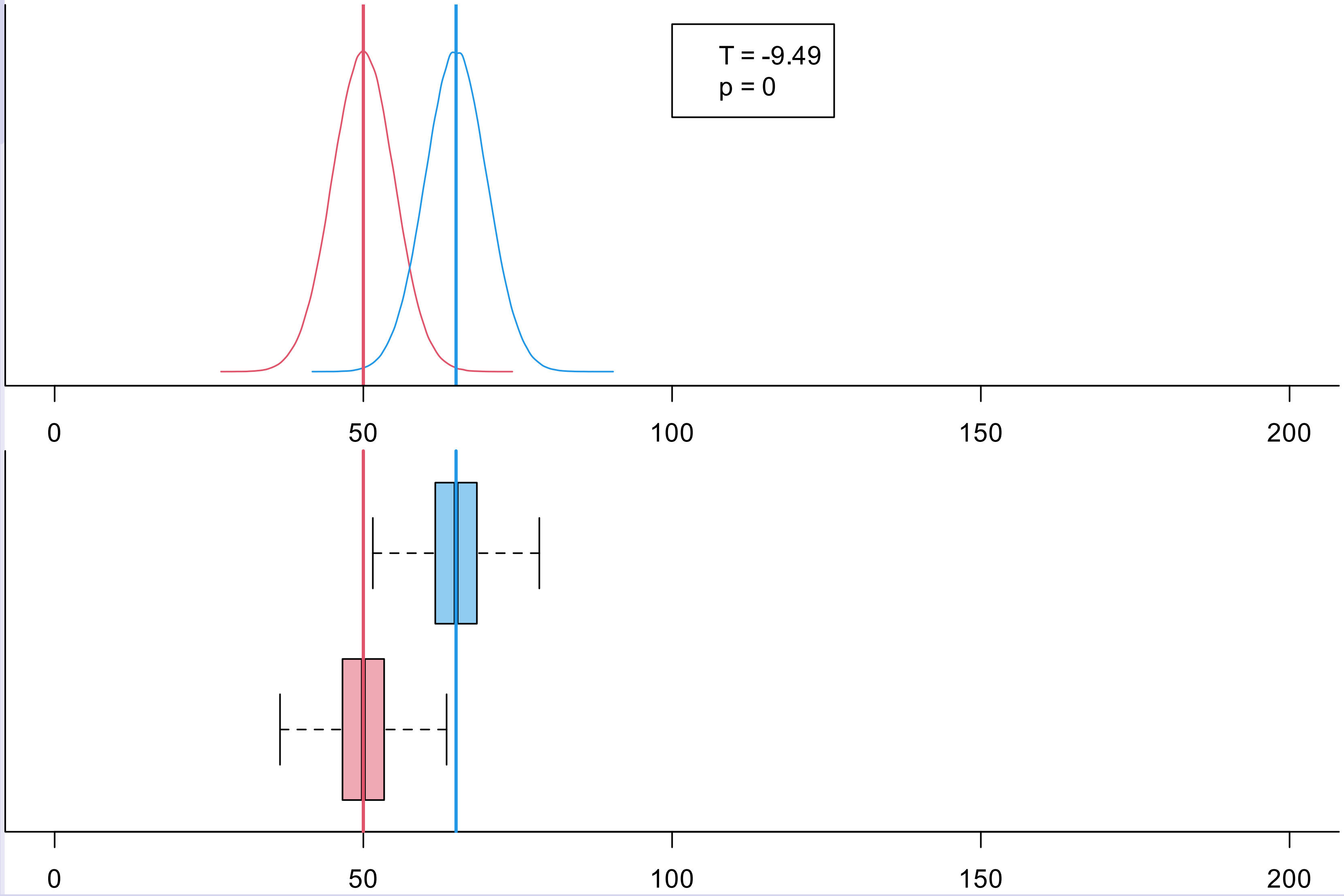
- Small difference in means.
- Dispersion is small
- High of overlap in the distributions
- T is small (and negative)
- P-value of 0.03 is significant.
- What is the null hypothesis?



Intuition: Small Difference In Means

Are the differences significant?

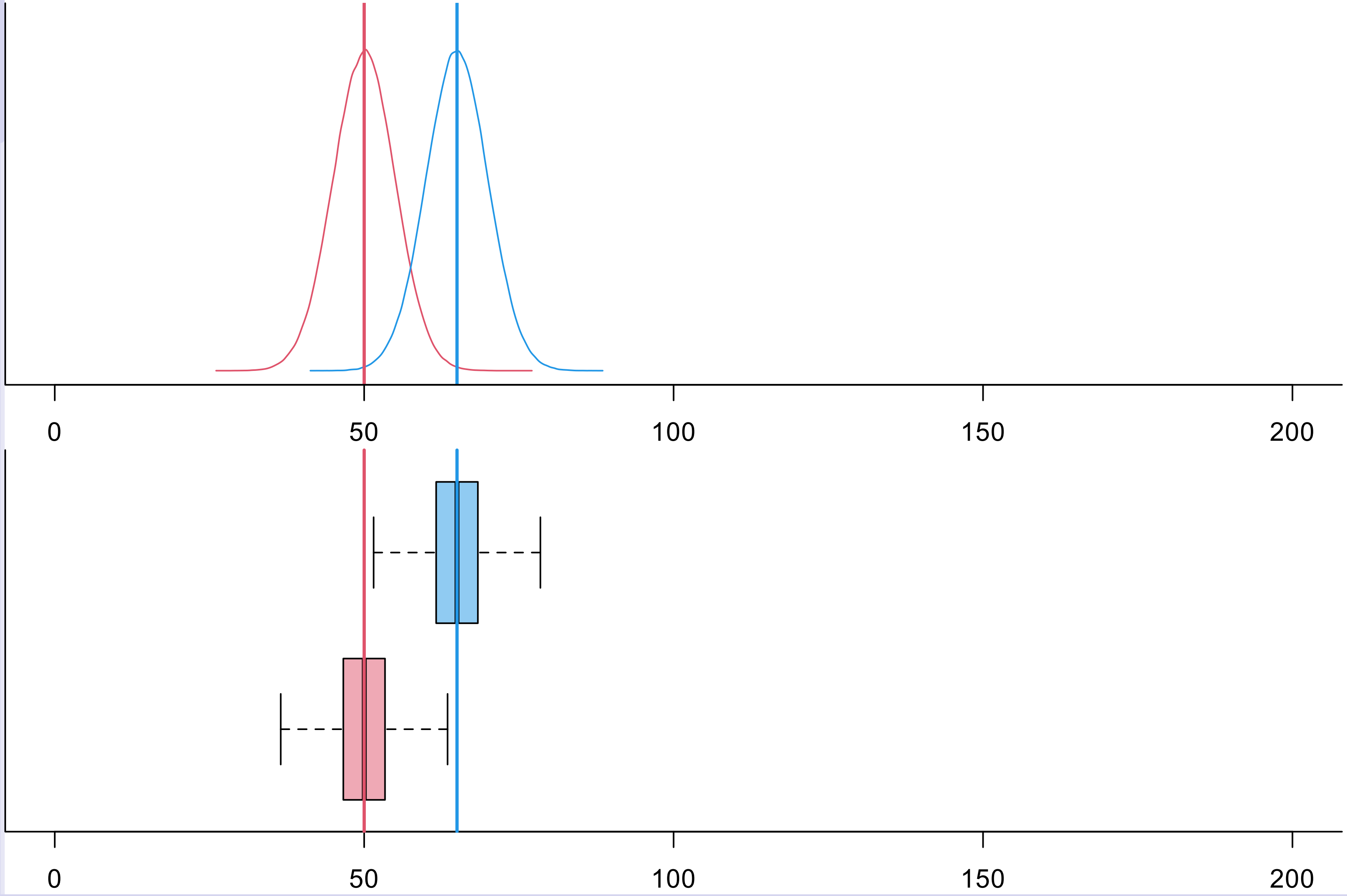
- Moderate difference in means.
- Dispersion is small
- Small of overlap in the distributions
- T is small (and negative)
- Very low p-value is highly significant.
- What is the null hypothesis?



Intuition: Small Difference In Means

Are the differences significant?

- Small-ish difference in means, but distributions are narrow.
- Distributions are slightly overlapping.
- You observed an individual that weighed 50 grams, which group does it belong to?
- Difference is probably significant!



Differences: paired t-test

Sometimes samples are meaningfully paired

- compare pairs of samples: test is on the ***mean difference***, not the mean values
- e.g., before-after
- e.g., north-south
- e.g., left-right
- Both samples on the same physical structure (like a target)

- Assumptions (as usual!)
  - Both normally distributed
  - Equal (similar) variances
  - Samples sizes *must* be the same. Why?

Differences: paired t-test

Which *t-test*?

- paired *t-test*
- compare pairs of samples
- both normally distributed
- equal (similar) variances
- samples sizes are ______ ?

$$t = \frac{\bar{D}}{\sqrt{\frac{s_D^2}{n}}}$$

- t: the t-statistic
- $\bar{D}$: mean of the *differences*
- s: standard deviation of the *differences*
- n: number of *paired* samples

Differences: When might the t-test be inappropriate?

Those pesky assumptions!

T-test Assumptions

- The t-test is a parametric statistical test.
- The theoretical justification of the t-test relies on certain assumptions:
  - Measurements are [approximately] normally distributed within groups.
  - The dispersion is approximately equal in the groups.
  - All measurements are independent.
  - Others, but these are the ones we'll focus on for now!
- We can test these assumptions: we'll focus on the normality assumption.

Differences: U-test

- compare two samples
- one or both *not* normally distributed
- based on *median*, *range*, and *ranks*
- rank all values as one sample, calculate group rank sums R
- calculate a *U*-value, a measure of overlap

Differences: U-test

- compare two samples
- both or differences *not* normally distributed
- based on *median*, *range*, and *ranks*
- rank all values as one sample, calculate group rank sums R
- calculate a *U*-value, a measure of overlap

$$U_a = n_a \times n_b + \frac{n_a(n_a + 1)}{2} - R_a$$

$$U_b = n_b \times n_a + \frac{n_b(n_b + 1)}{2} - R_b$$

- n_a: number of samples in sample a
- n_b: number of samples in sample b
- R_a: sum of the ranks of values in a
- R_b: sum of the ranks of values in b

Differences: U-test

- compare two samples
- both or differences *not* normally distributed
- based on *median*, *range*, and *ranks*
- rank all values as one sample, calculate group rank sums R
- calculate a *U*-value, a measure of overlap

$$U_a = n_a \times n_b + \frac{n_a(n_a + 1)}{2} - R_a$$

$$U_b = n_b \times n_a + \frac{n_b(n_b + 1)}{2} - R_b$$

- smallest is used to find the p-value
- unlike the t-statistic, lower U-values are more likely to be significant

Differences: Wilcoxon matched-pairs test

- both or differences *not* normally distributed
- based on ranked *differences*
- first calculate the differences
- second rank the differences
- 0's not ranked
- sum and compare +ve and -ve ranks

$$W^+ = \sum R^+$$
$$W^- = \sum R^-$$

- W^+: the Wilcoxon test statistic for positive differences
- W^+: the Wilcoxon test statistic for negative differences
- R^+: the sum of the ranks of positive differences
- R^+: the sum of the ranks of negative differences

Differences: Wilcoxon matched-pairs test

- pairs or differences *not* normally distributed
- based on ranked *differences*
- first calculate the differences
- second rank the differences
- sum and compare +ve and -ve ranks

$$W^+ = \sum R^+$$
$$W^- = \sum R^-$$

- smallest is used to find the p-value
- lower W-values are more likely to be significant

Two-sample tests in R

It's easy to conduct a t-test in R, but first we need to specify our hypotheses and check that our data meet the required assumptions.

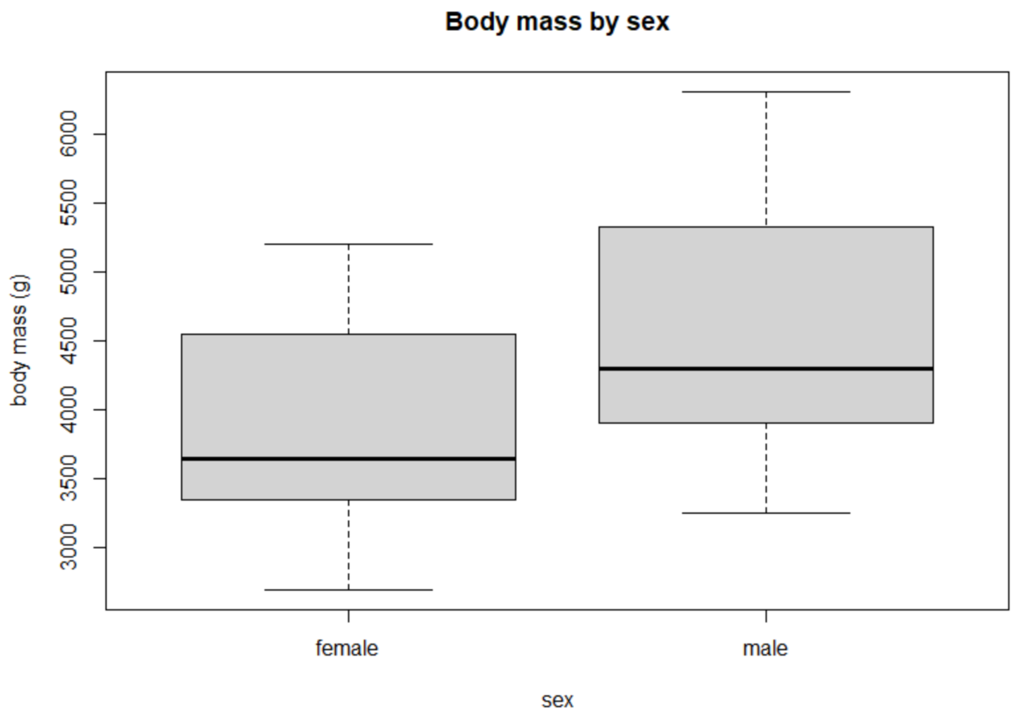
Two-sample test in R procedure:

1. Create a conditional boxplot to explore your data
   1. Using the formula notation is the easiest way.
2. Check the assumption of normality with shapiro.test()
   1. The null hypothesis is that the data are normal.
3. Decide which test to use
4. Conduct the test with t.test() or wilcox.test() and interpret the results.
   1. The null hypothesis is that there is no difference between groups.

Conditional Boxplot

Syntax	Plot
• Note the formula notation:	• What do you notice?

```
boxplot(
  body_mass_g ~ sex,
  data = penguins,
  main = "Body mass by sex",
  ylab = "body mass (g)"
)
```



Check Assumptions: Shapiro test in R

Syntax	Results

Syntax

```r
dat_male = subset(
  penguins,
  sex == "male")
dat_female = subset(
  penguins,
  sex == "female")
# Shapiro test on male penguins
shapiro.test(
  dat_male$body_mass_g
)
# Shapiro test on male penguins
shapiro.test(
  dat_female$body_mass_g
)
```

Results

```
> shapiro.test(
+   dat_male$body_mass_g
+ )

        Shapiro-Wilk normality test

data:  dat_male$body_mass_g
W = 0.92504, p-value = 1.227e-07

> shapiro.test(
+   dat_female$body_mass_g
+ )

        Shapiro-Wilk normality test

data:  dat_female$body_mass_g
W = 0.91931, p-value = 6.155e-08
```

Which test?

- What’s the Shapiro test null hypothesis?
- Do we have evidence for or against normality?
- Can we use a t-test?

Which test?

- What's the Shapiro test null hypothesis?
  - That the data are normal
- Do we have evidence for or against normality?
  - Low p-value is evidence that data are non-normal
- Can we use a t-test?
  - No, our data are too non-normal
  - We can use a U-test
    - The syntax is nearly identical to that of the t.test() function.

The U-Test

Syntax	Results
• Note the formula notation:	

```
wilcox.test(
  body_mass_g ~ sex,
  data = penguins
)
```

```
        Wilcoxon rank sum test with
        continuity correction

data:  body_mass_g by sex
W = 6874.5, p-value = 1.813e-15
alternative hypothesis: true location shift
 is not equal to 0
```

What's the p-value?

Is there evidence for a true difference?