

Inferential statistics

Inferential Statistics (Mod 3/4 highlights)

Goals

- Code and interpret results of t-tests: comparing **two** groups on some numerical measure
- Code and interpret results of one-way ANOVA: comparing **three or more groups** on some numerical measure
- Code and interpret results of linear regression: is there a **relationship between two numeric variables**?

In the code-a-longs, typically we would incorporate all the major concepts for a given example:

- Load and filter data
- Generate summary statistics
- Create one or more data visualizations
- Calculate and interpret statistics

In the interest of time, the code for summary statistics and data visualizations is provided.

t-tests

A two sample t-test is a way of evaluating if the means of **two** populations are different, given our samples of those populations.

A t-test relies on the calculate a t-score. This quantity depends on our sample mean, our sample standard deviation, and the size of our sample.

The formula of the t-score for a two sample t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The primary output of a t-test is a p-value. A p value represents the probability that the difference between our sample means would have occurred by chance.

We can use this p-value to assess a statistical hypothesis.

Statistical hypotheses are formulated as a null and an alternative hypotheses:

H_0 (null hypothesis) - There is **no difference** in the means of the populations we sampled from.

H_a (alternative hypothesis) - The means of the populations we sampled from **are different**.

Again, our p-value is the decimal probability that our data occurred by chance. For instance, a p-value of 0.05 would mean there is a 5% probability that the null hypothesis is true, given our observations.

Scenario: Fishing and leopard seals

As you are well aware by now, our main source of food (fish) has been compromised. Rather than starve or leave, we decide to source our fish from waters of Antarctica.

The problem is, the places we'd fish are also foraging grounds for leopard seals. To minimize the impact of our fishing on the seal population, we'd like to know where and when the presence of fish/seals are greater.

Luckily, we've been collecting relevant data for awhile. We have: radio tags on seals, and the number of seals at a given location. We also net traps to count humped rock cod, our shared food source.

Data exploration

Let's take a look at our new data. Because we have two different collection schemes, our data are separated in to two data frames (and files).

```
library(tidyverse)
```

```
— Attaching core tidyverse packages — tidyverse 2.0.0 —
✓ dplyr      1.1.4      ✓ readr      2.1.5
✓ forcats    1.0.0      ✓ stringr    1.5.1
✓ ggplot2    3.5.1      ✓ tibble     3.2.1
✓ lubridate  1.9.3      ✓ tidyr      1.3.1
✓ purrr      1.0.2

— Conflicts — tidyverse_conflicts() —
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
#seal data
```

```
seal_data = read_csv("antarctic-seals.csv")
```

```
Rows: 640 Columns: 5
```

```
— Column specification —
Delimiter: ","
chr  (2): time, bay
dbl  (2): area, num_seals
date (1): date
```

```
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#fish data
```

```
fish_data = read_csv("antarctic-fish.csv")
```

Rows: 640 Columns: 5

— Column specification —

Delimiter: ","

chr (2): time, bay

dbl (2): net, num_fish

date (1): date

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show_col_types = FALSE` to quiet this message.

- We'd like to know if seal and fish counts are different during times of day observed.
- Based on our goals, what quantities do we want to compare?
- Create null and alternative hypotheses to evaluate our data

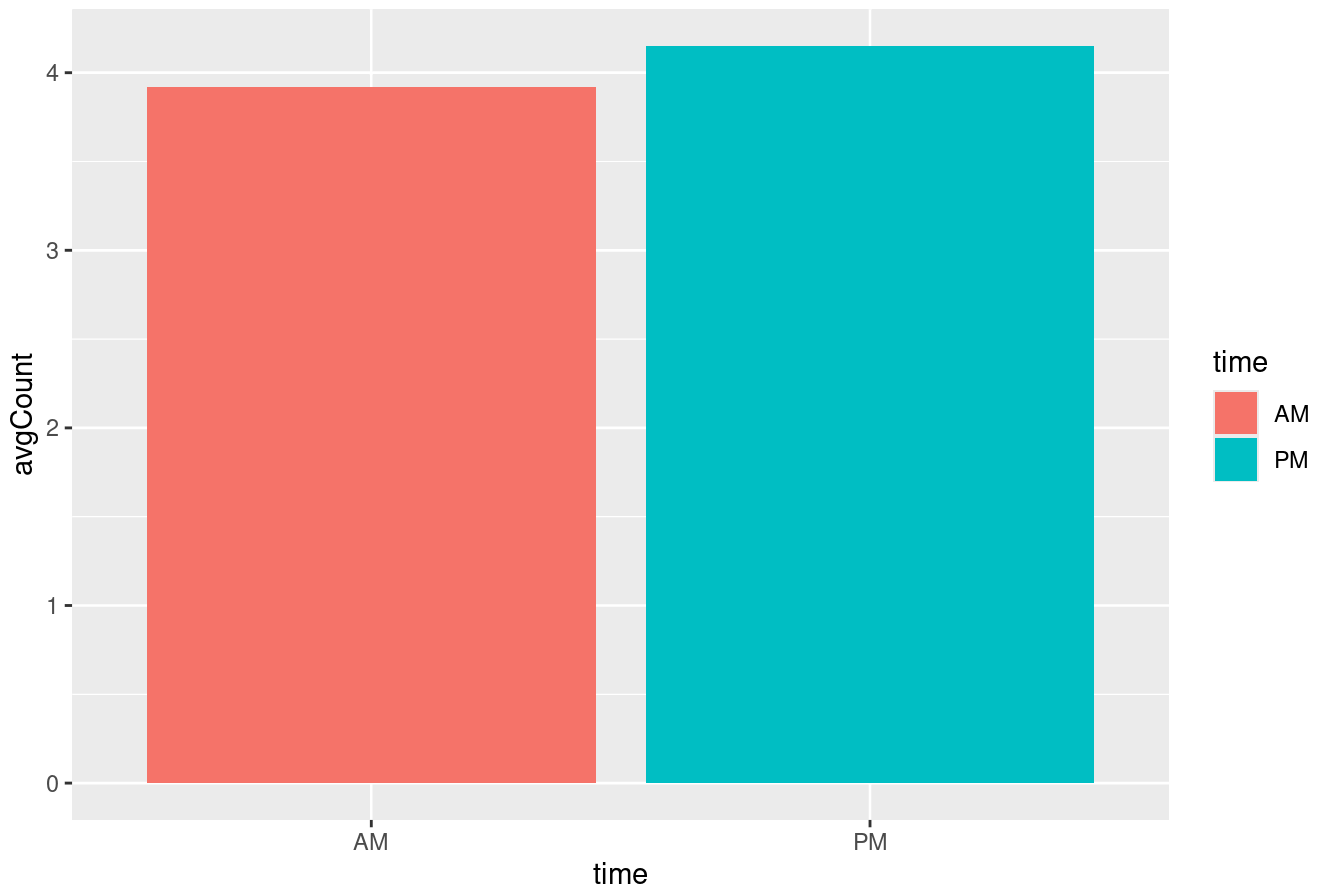
H0: no difference between the mean count of fish between the times of day

Ha: there *is* a difference

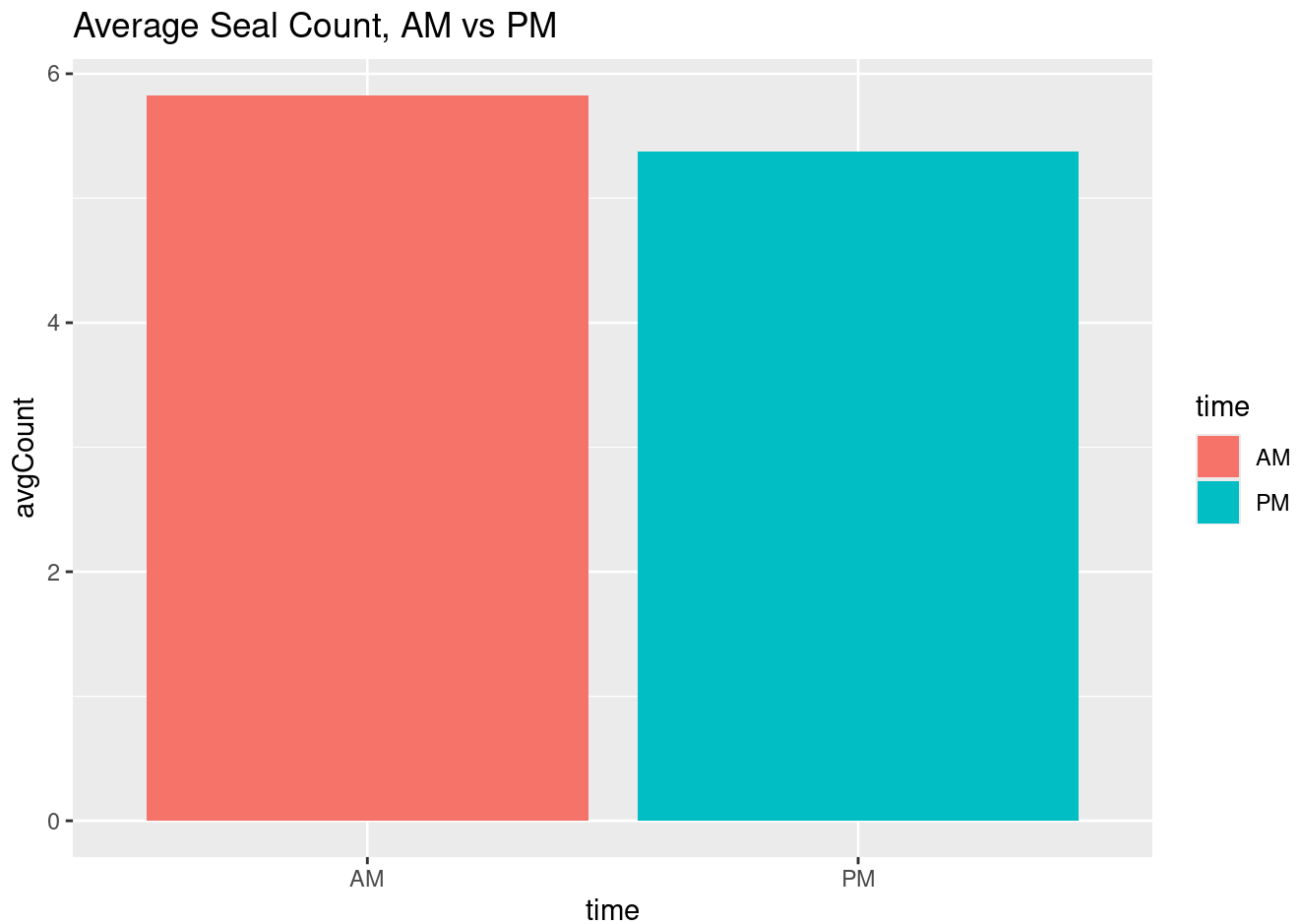
Transform and visualize the data

```
#fish
fish_data |>
  group_by(time) |>
  summarize(avgCount=mean(num_fish)) |>
  ggplot(mapping=aes(y=avgCount, x=time, fill=time)) +
  geom_bar(stat="identity")+
  labs(title="Average Fish Count, AM vs PM")
```

Average Fish Count, AM vs PM



```
#seals
seal_data |>
  group_by(time) |>
  summarize(avgCount=mean(num_seals)) |>
  ggplot(mapping=aes(y=avgCount, x=time, fill=time)) +
  geom_bar(stat="identity")+
  labs(title="Average Seal Count, AM vs PM")
```



Performing our tests

- Perform a t-test to evaluate our hypothesis
- Interpret the results using the p-value

```
#fish  
t.test(num_fish ~ time, data=fish_data)
```

Welch Two Sample t-test

```
data: num_fish by time  
t = -1.5665, df = 635.07, p-value = 0.1177  
alternative hypothesis: true difference in means between group AM and group PM is not  
equal to 0  
95 percent confidence interval:  
-0.52114501 0.05864501  
sample estimates:  
mean in group AM mean in group PM  
3.91875 4.15000
```

```
#seals  
t.test(num_seals ~ time, data = seal_data)
```

Welch Two Sample t-test

```
data: num_seals by time  
t = 2.6898, df = 635.23, p-value = 0.007338  
alternative hypothesis: true difference in means between group AM and group PM is not  
equal to 0  
95 percent confidence interval:  
 0.121469 0.778531  
sample estimates:  
mean in group AM mean in group PM  
      5.825      5.375
```

ANOVA: ANalysis Of VAriance

When can I use an ANOVA? Why would I?

- Independent variable is categorical and the response is numerical
- Goal: to compare means among groups

Assumptions of ANOVA

- Data are "normally distributed" => look at the histogram
- Data are "equally varied" => standard deviations reasonably similar
- Samples are independent of one another

The null and alternative hypotheses

H_0 (null hypothesis) - The means of the populations we sampled from **are all equal**:

$$\mu_1 = \mu_2 = \dots = \mu_i$$

H_a (alternative hypothesis) - The means of the populations we sampled from **are not all equal**

Updated scenario: more bays

We have figured out the best option for minimizing our impact on leopard seals while keeping ourselves fed between two bays: Wilhelmina and Marguerite. But there are more bays! And ideally we would use two or more bays to spread out our fishing efforts among multiple humped rock cod populations.

Our team has collected similar data, as we had for Wilhelmina and Marguerite, on four more bays: Emperor, Hope, Sil

We are going to examine the fish populations in class, and you will work with the leopard seals for your homework.

Read in the data

```
fishManyBays <- read_csv("antarctic_fish_many_bays.csv")
```

Rows: 1920 Columns: 5

— Column specification —

Delimiter: ","

chr (2): time, bay

dbl (2): net, num_fish

date (1): date

i Use ``spec()`` to retrieve the full column specification for this data.

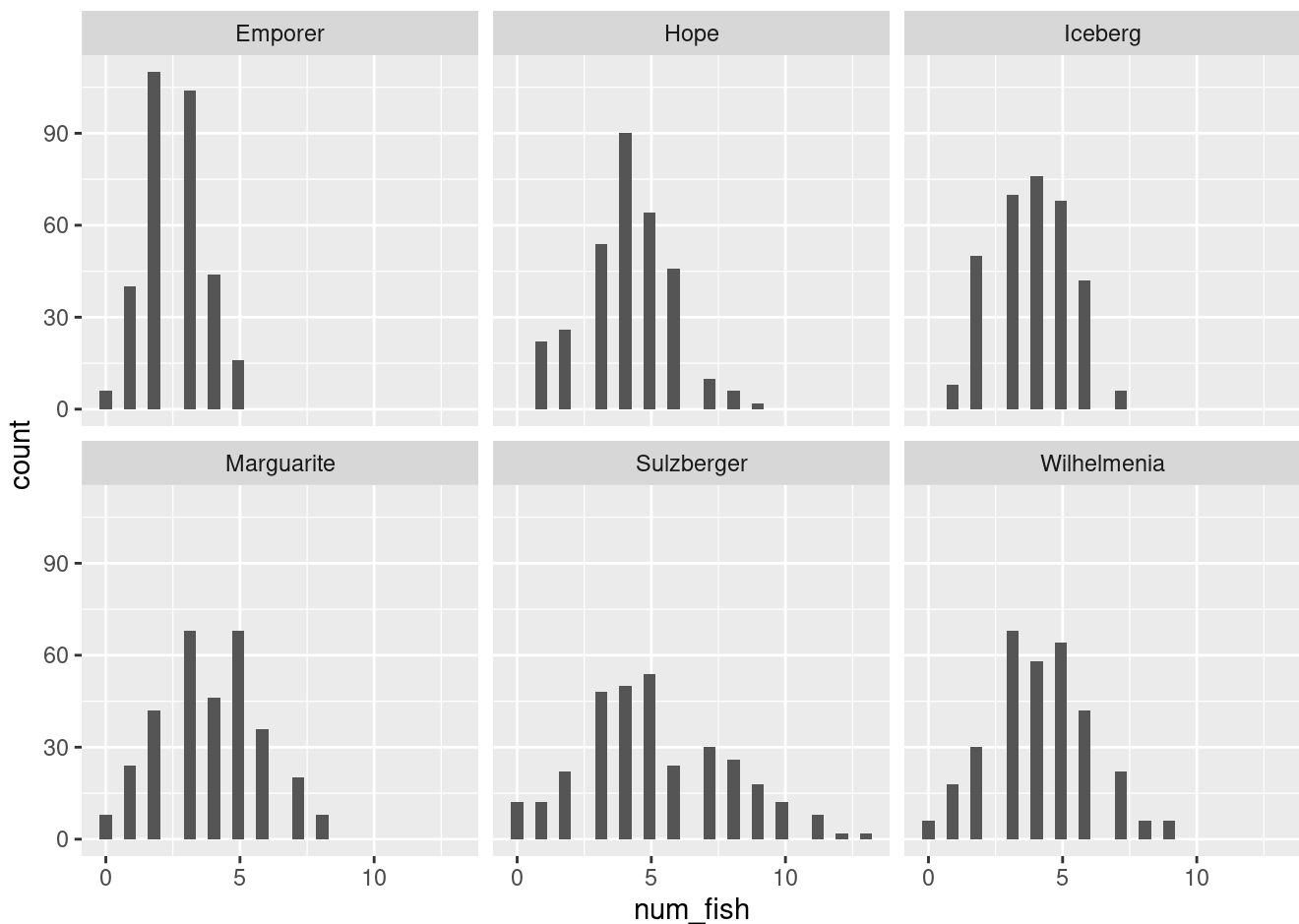
i Specify the column types or set ``show_col_types = FALSE`` to quiet this message.

It's always a good idea to visualize your data first. This gives you some perspective on the distribution of the data. What type of data viz is best for viewing the distribution of one variable?

4.

```
ggplot(data = fishManyBays, aes(x = num_fish)) +  
  geom_histogram() +  
  facet_wrap(~ bay)
```

``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.



5. Now let's calculate some summary statistics. What do you notice?

```
fishSummary <- fishManyBays |>
  group_by(bay) |>
  summarize(meanFish = mean(num_fish, na.rm=TRUE),
            standDevFish = sd(num_fish, na.rm=TRUE),
            sampleSize = n()) |>
  arrange(desc(meanFish))

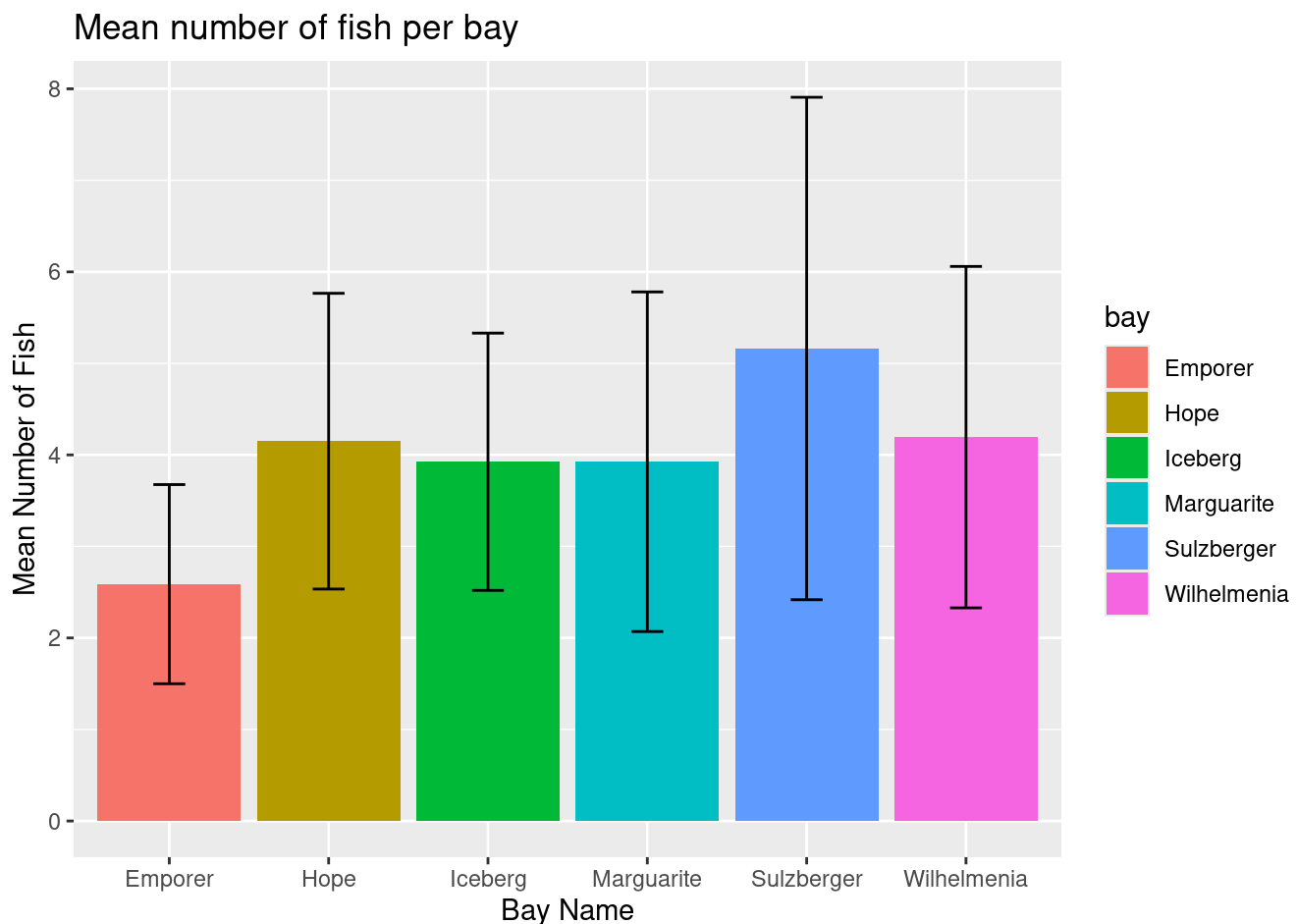
fishSummary
```

A tibble: 6 × 4

	bay	meanFish	standDevFish	sampleSize
	<chr>	<dbl>	<dbl>	<int>
1	Sulzberger	5.16	2.74	320
2	Wilhelmenia	4.19	1.87	320
3	Hope	4.15	1.62	320
4	Iceberg	3.92	1.41	320
5	Margarite	3.92	1.86	320
6	Emporer	2.59	1.09	320

6. Let's create a bar graph to compare the summary stats between the groups. Does it seem like the groups are different?

```
ggplot(data = fishSummary, mapping=aes(x=bay, y=meanFish, fill = bay)) +
  geom_bar(stat = "identity") +
  geom_errorbar(mapping=aes(ymin = meanFish-standDevFish,
                           ymax = meanFish + standDevFish), width = 0.2) +
  labs(x="Bay Name",
       y="Mean Number of Fish",
       title="Mean number of fish per bay")
```



7. Finally, let's code for the ANOVA. The syntax is dependent variable ~ independent variable

```
fishModel <- aov(data = fishManyBays, num_fish ~ bay)
summary(fishModel)
```

```
      Df Sum Sq Mean Sq F value Pr(>F)
bay      5   1094   218.71   64.88 <2e-16 ***
Residuals 1914   6452     3.37
```

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

ANOVAs are incredibly useful to tell you if there is a difference in the means of any of the groups. However, they do not tell you which means differ from another. To do that, you need to use a class of tests called Post Hoc Tests. Post hoc tests take into account the problem of running multiple pairwise comparisons, which is the increasing chance of error rates. The most common is Tukey's HSD, but there are others depending on the specifics of your data set. You don't need to worry about understanding Tukey's test, but here I am going to show you how it works and an overview of the interpretation of it.

```
TukeyHSD(fishModel)
```

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

```
Fit: aov(formula = num_fish ~ bay, data = fishManyBays)
```

\$bay

	diff	lwr	upr	p adj
Hope-Emporer	1.562500e+00	1.1484366	1.9765634	0.0000000
Iceberg-Emporer	1.337500e+00	0.9234366	1.7515634	0.0000000
Margarite-Emporer	1.337500e+00	0.9234366	1.7515634	0.0000000
Sulzberger-Emporer	2.575000e+00	2.1609366	2.9890634	0.0000000
Wilhelmenia-Emporer	1.606250e+00	1.1921866	2.0203134	0.0000000
Iceberg-Hope	-2.250000e-01	-0.6390634	0.1890634	0.6316815
Margarite-Hope	-2.250000e-01	-0.6390634	0.1890634	0.6316815
Sulzberger-Hope	1.012500e+00	0.5984366	1.4265634	0.0000000
Wilhelmenia-Hope	4.375000e-02	-0.3703134	0.4578134	0.9996682
Margarite-Iceberg	4.440892e-16	-0.4140634	0.4140634	1.0000000
Sulzberger-Iceberg	1.237500e+00	0.8234366	1.6515634	0.0000000
Wilhelmenia-Iceberg	2.687500e-01	-0.1453134	0.6828134	0.4328039
Sulzberger-Margarite	1.237500e+00	0.8234366	1.6515634	0.0000000
Wilhelmenia-Margarite	2.687500e-01	-0.1453134	0.6828134	0.4328039
Wilhelmenia-Sulzberger	-9.687500e-01	-1.3828134	-0.5546866	0.0000000

Linear Regression

Latest scenario: we want to build a road to access the new fishing sites on a path that minimizes impact on hair grass.

There are many environmental conditions that may be associated with hair grass density. For today's code along, we are going to focus on two: soil pH and nitrogen content.

Let's look at nitrogen content first.

We always should start with a data visualization and some descriptive statistics.

```
# load in the tidyverse
#library("tidyverse")
# load in the data
hairgrass <- read_csv("hairgrass_data.csv")
```

Rows: 480 Columns: 12

— Column specification —

Delimiter: ","

dbl (11): location_ID, soil_pH, p_content, percent_soil_rock, max_windspeed...

date (1): date

i Use `spec()` to retrieve the full column specification for this data.

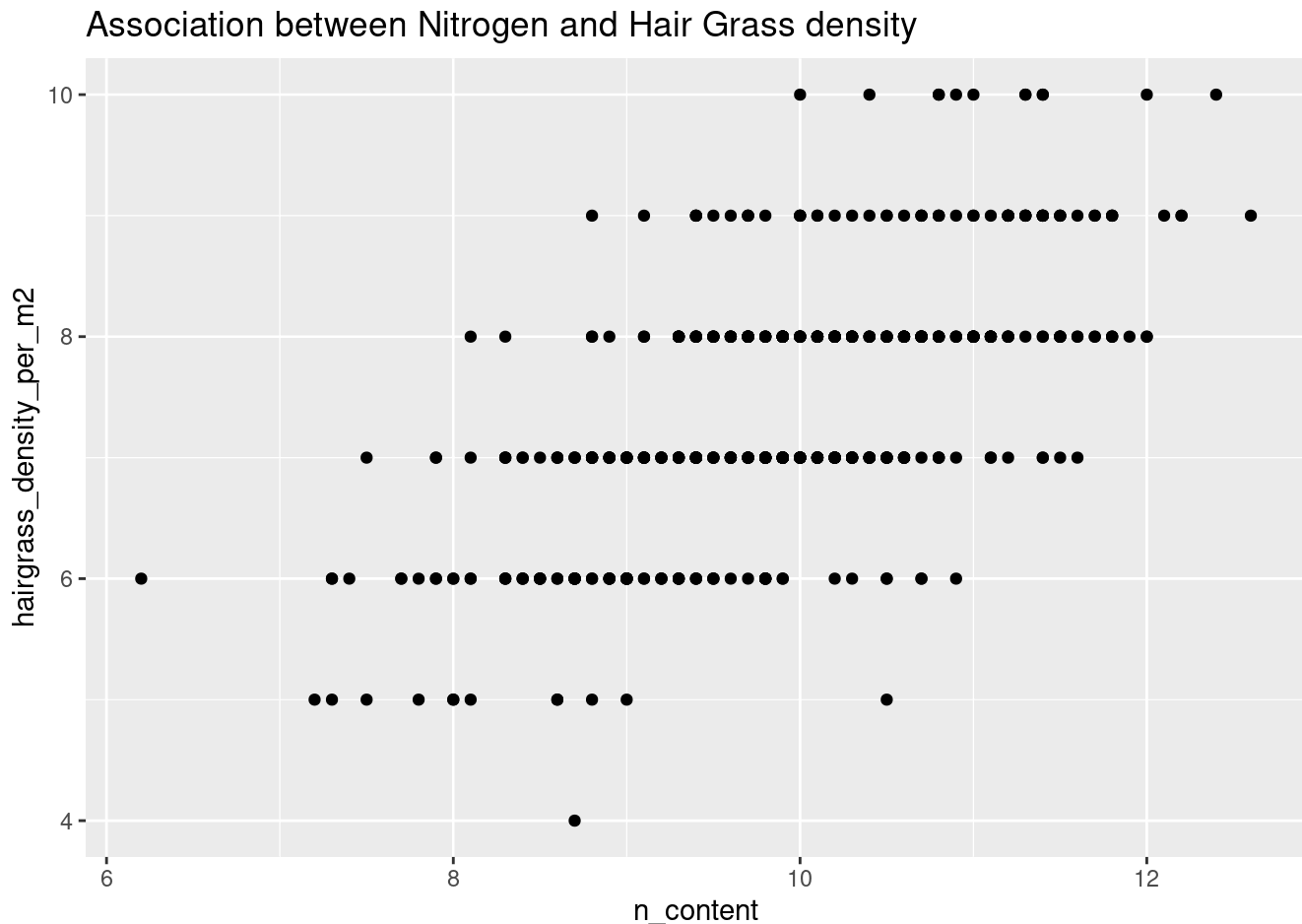
i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```
# What are our independent and dependent variables?
# What kind of variables are they?
# What kind of viz should we do?
```

```
hairgrass |>
  summarize(maxNitrogen=max(n_content),
            minNitrogen=min(n_content),
            avgNitrogen=mean(n_content),
            sdNitrogen=sd(n_content))
```

```
# A tibble: 1 × 4
  maxNitrogen minNitrogen avgNitrogen sdNitrogen
    <dbl>      <dbl>      <dbl>      <dbl>
1    12.6        6.2      9.93      1.02
```

```
hairgrass |>
  ggplot(mapping=aes(x = n_content, y = hairgrass_density_per_m2)) +
  geom_point()+
  # geom_jitter() #as an alternative
  labs(title="Association between Nitrogen and Hair Grass density")
```



do you see a pattern? Do you think these data are correlated? What do you think

Now let's actually calculate the correlation coefficient, r . As a reminder, the correlation coefficient is a number between -1 and 1 that looks at the strength and direction of the relationship between two

numeric variables. The greater the magnitude of the correlation coefficient, the stronger the correlation (All the points fall exactly on the line of best fit if $r = 1$ or -1).

```
r = cor(hairgrass$hairgrass_density_per_m2, hairgrass$n_content)
r
```

```
[1] 0.6326895
```

```
# What do we expect based on this correlation coefficient?
```

We often think about the correlation in terms of r-squared. All we have to do is square the value we calculated above. How do we interpret r-squared for this relationship?

```
r^2
```

```
[1] 0.400296
```

```
# Means that 40% of the variation in hair grass density can be explained by the v
```

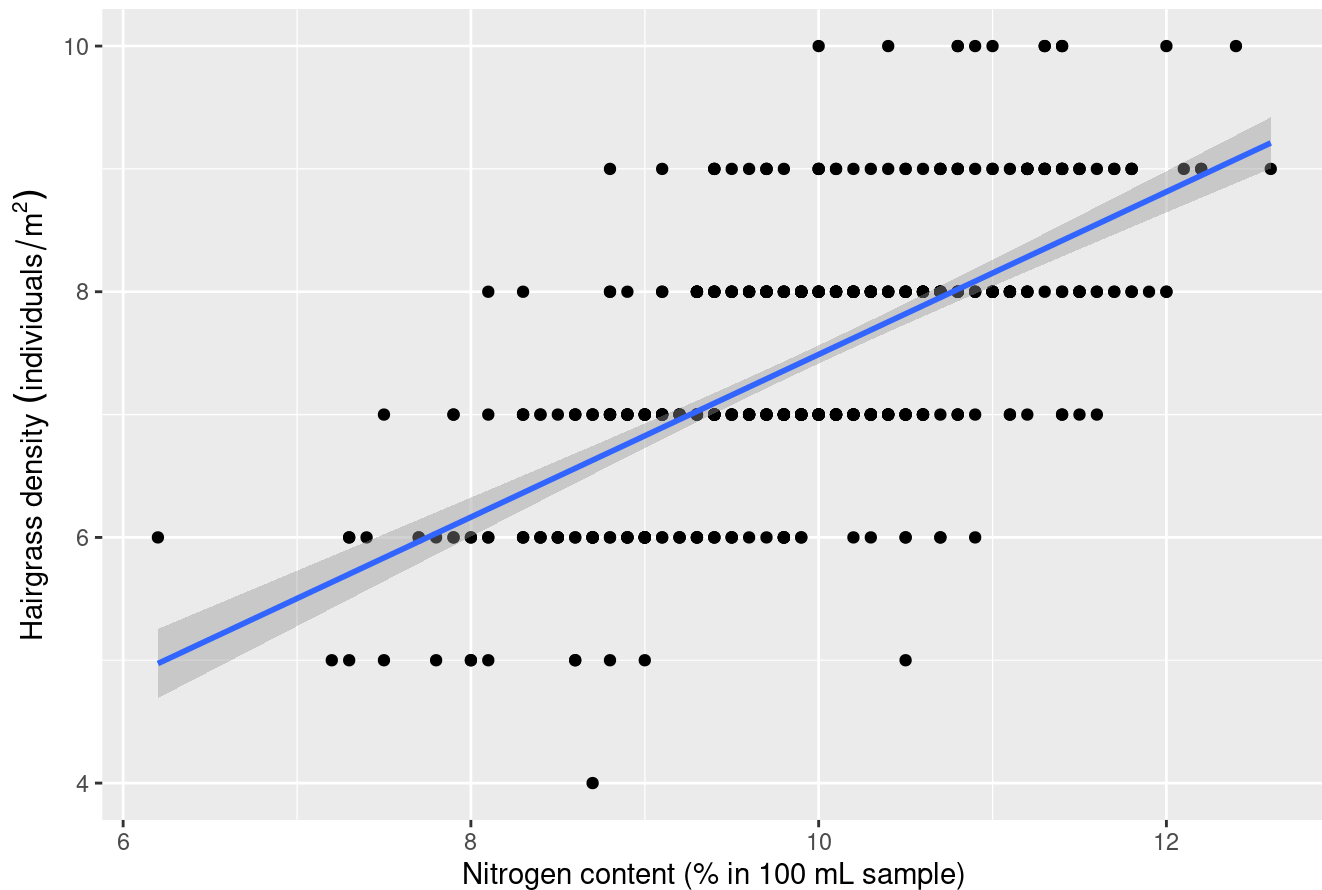
Adding our line of best fit to the data

```
# is that what we expected based on that correlation coefficient?

hairgrass |>
  ggplot(mapping=aes(y = hairgrass_density_per_m2, x = n_content)) +
  geom_point() +
  geom_smooth(method = "lm") + # this is new - adds line of best fit
  labs(title="Nitrogen content and hairgrass density",
       x="Nitrogen content (% in 100 mL sample)",
       y=bquote('Hairgrass density ' (individuals / m^2))) # bquote for math notation
```

`geom_smooth()` using formula = 'y ~ x'

Nitrogen content and hairgrass density



If we want to add statistical rigor, we need to use regression analysis. A regression analysis approximates the relationship between a dependent variable and one or more independent variables and evaluates the strength of that relationship (giving us a p-value).

We will use linear regressions in this unit. This simply means that the model will take the form of $y = mx + b$, where:

- **y** is the dependent variable
- **x** is the independent variable
- **m** is the slope
- **b** is the y-intercept.

What would the model for our question about nitrogen content be? (it's okay that we haven't yet calculated the values)

```
# hair grass density = m * n_content + b
```

What is the null hypothesis? What is the alternative hypothesis?

```
# null: There is no relationship between hairgrass density and n_content
```

```
# alt: There is a relationship between hairgrass density and n_content
```

R can actually calculate what this model would be for us. The formula for the line of best fit ($y = mx + b$) aims to minimize the distance between each observation (point) and the line. What is the model?

```
summary(lm(formula=hairgrass_density_per_m2 ~ n_content, data = hairgrass))
```

Call:

```
lm(formula = hairgrass_density_per_m2 ~ n_content, data = hairgrass)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.82079	-0.55590	-0.02612	0.57654	2.51032

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.86739	0.37000	2.344	0.0195 *
n_content	0.66223	0.03707	17.862	<2e-16 ***

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

Residual standard error: 0.8294 on 478 degrees of freedom

Multiple R-squared: 0.4003, Adjusted R-squared: 0.399

F-statistic: 319.1 on 1 and 478 DF, p-value: < 2.2e-16

```
# model: hairgrass density = 0.87 + 0.66 *n_content
```

So what can we conclude about soil pH and hair grass density?

```
# stats interpretation
# Because the p-value associated with the F statistic (319) was 2.2x10(-16), we r
# interpretation in light of scenario: we should pay attention to n content as we
```

Moving on to soil pH

Data visualization, with the line of best fit, and summary statistics for soil pH

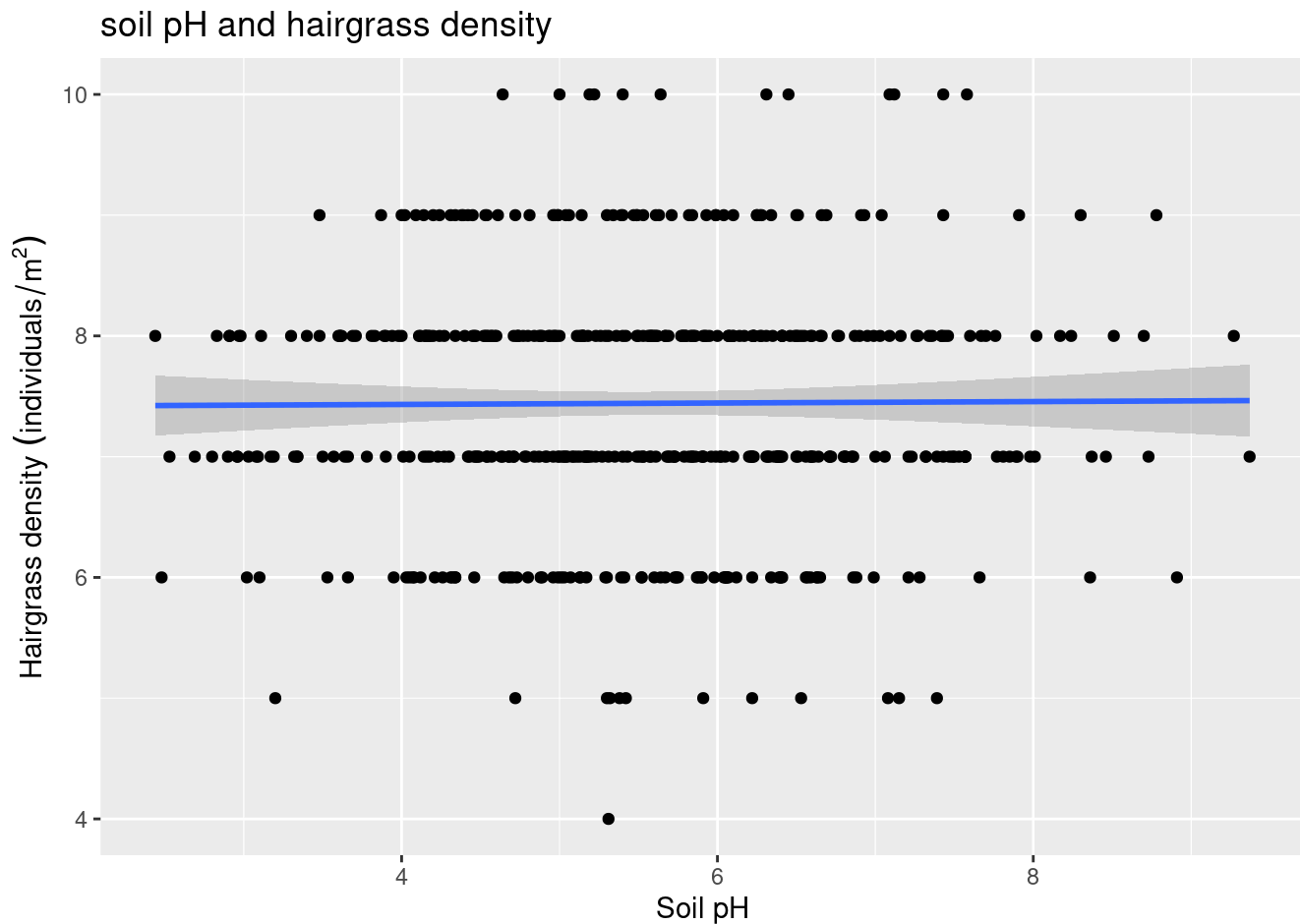
```
# look at summary statistics
hairgrass |>
  summarize(max(soil_pH), min(soil_pH), mean(soil_pH), sd(soil_pH))
```

A tibble: 1 × 4

	`max(soil_pH)`	`min(soil_pH)`	`mean(soil_pH)`	`sd(soil_pH)`
	<dbl>	<dbl>	<dbl>	<dbl>
1	9.37	2.44	5.55	1.30

```
# plot
hairgrass |>
  ggplot(mapping=aes(x = soil_pH, y = hairgrass_density_per_m2 )) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title="soil pH and hairgrass density",
       x="Soil pH",
       y=bquote('Hairgrass density ' (individuals / m2)))
```

`geom_smooth()` using formula = 'y ~ x'



What is the correlation coefficient?

```
cor(hairgrass$hairgrass_density_per_m2, hairgrass$soil_pH)
```

```
[1] 0.007200444
```

What is the model for our question about soil pH, without values?

```
# hairgrass density = a * soil_pH + b
```

Create the model in R and calculate the values for a and b.

```
summary(lm(hairgrass_density_per_m2 ~ soil_pH, data = hairgrass))
```

Call:

```
lm(formula = hairgrass_density_per_m2 ~ soil_pH, data = hairgrass)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.4403	-0.4491	-0.4271	0.5631	2.5637

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.408859	0.214051	34.613	<2e-16 ***
soil_pH	0.005915	0.037570	0.157	0.875

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

Residual standard error: 1.071 on 478 degrees of freedom

Multiple R-squared: 5.185e-05, Adjusted R-squared: -0.00204

F-statistic: 0.02478 on 1 and 478 DF, p-value: 0.875

```
# model: hairgrass density = 7.4 + 0.006 * soil pH
```

At alpha = 0.05, what do we conclude about the relationship between soil pH and hairgrass density and why?

```
# stats interpretation
```

```
# Because the p-value associated with the F statistic was 0.875, we accept the null hypothesis
```

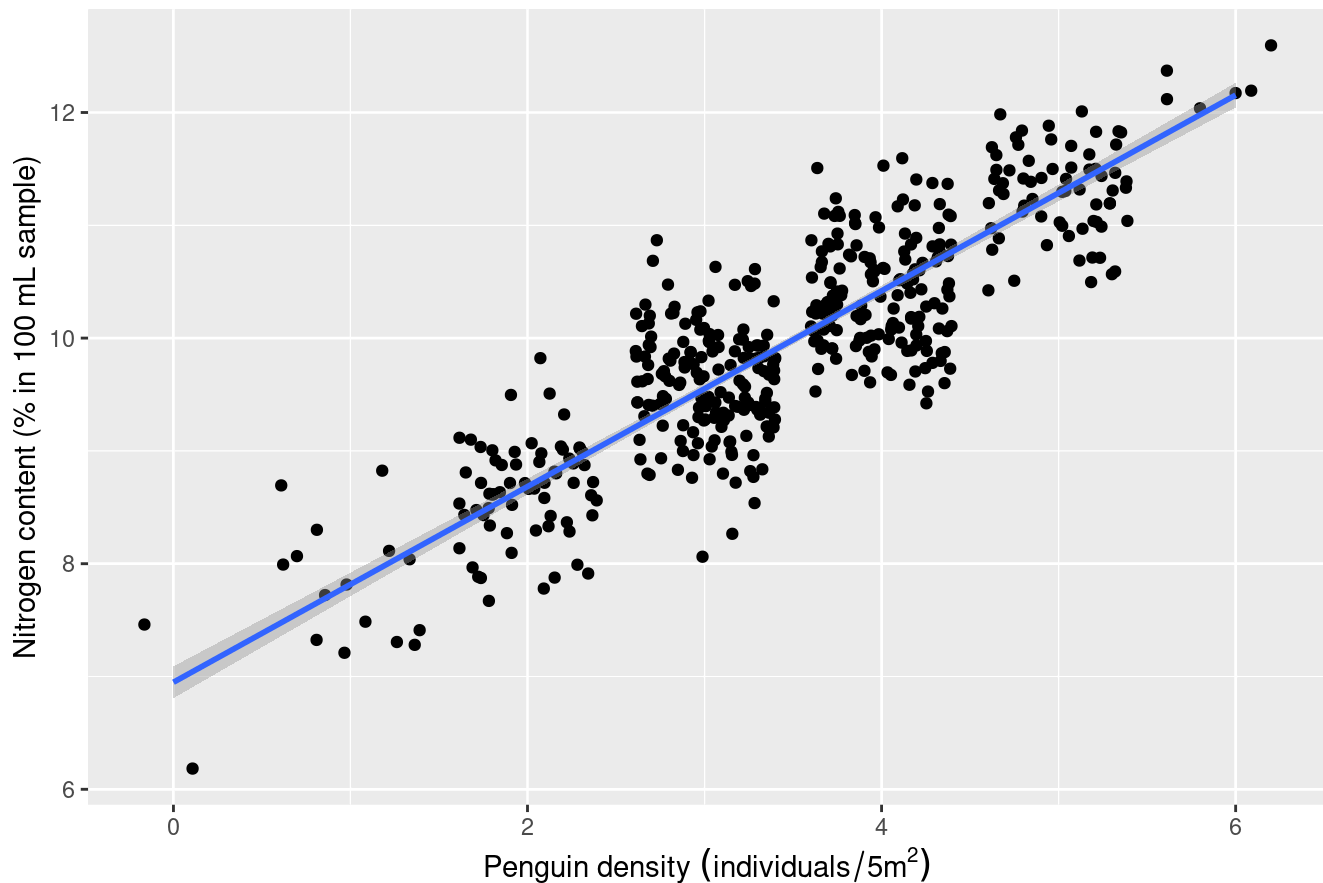
What does this mean for the road we are building?

```
# interpretation in light of scenario: we shouldn't worry about soil pH as we think it's unrelated
```

```
hairgrass |>
  ggplot(aes(x= penguin_density_per_5m2_within_100m, y = n_content)) +
  geom_jitter() +
  geom_smooth(method = "lm") +
  xlab(bquote('Penguin density ' (individuals /5 *m^2))) +
  ylab("Nitrogen content (% in 100 mL sample)") +
  ggtitle("Penguin density is related to nitrogen content")
```

`geom_smooth()` using formula = 'y ~ x'

Penguin density is related to nitrogen content



```
summary(lm(n_content ~ penguin_density_per_5m2_within_100m, data = hairgrass))
```

Call:

```
lm(formula = n_content ~ penguin_density_per_5m2_within_100m,  
    data = hairgrass)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.45157	-0.31911	-0.00157	0.31401	1.34843

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.94895	0.07196	96.56	<2e-16 ***
penguin_density_per_5m2_within_100m	0.86754	0.02004	43.28	<2e-16 ***

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

Residual standard error: 0.4614 on 478 degrees of freedom

Multiple R-squared: 0.7967, Adjusted R-squared: 0.7963

F-statistic: 1873 on 1 and 478 DF, p-value: < 2.2e-16

