TUTORIAL 4
Introductory Statistics with *R*

Two-sample *t*-test
I. Data & Hypothesis

- `ozone ← read_csv("ozone.csv")`
- Question: Ozone level differs between east/west?
- Null Hypothesis (H_0): No difference ($\mu_1 = \mu_2$)

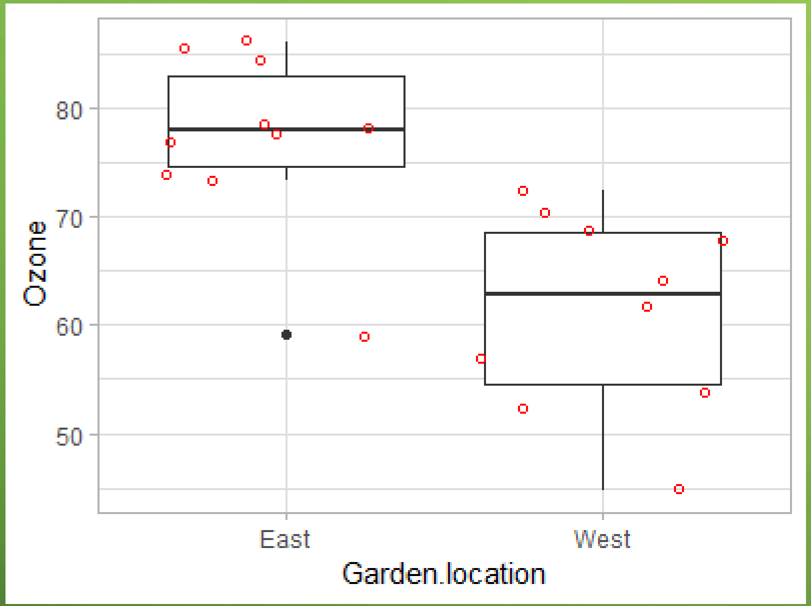
```
glimpse(ozone)

## Observations: 20
## Variables: 3
## $ Ozone           (dbl) 61.7, 64.0, 72.4, 56.8, 52.4, 4...
## $ Garden.location (fctr) West, West, West, West, West, ...
## $ Garden.ID       (fctr) G1, G2, G3, G4, G5, G6, G7, G8...
```

	Ozone	Garden.location	Garden.ID
	<dbl>	<chr>	<chr>
1	61.7	West	G1
2	64	West	G2
3	72.4	West	G3
4	56.8	West	G4
5	52.4	West	G5
6	44.8	West	G6
7	70.4	West	G7
8	67.6	West	G8
9	68.8	West	G9
10	53.7	West	G10
11	59.1	East	G11
12	78.5	East	G12
13	73.9	East	G13
14	86.1	East	G14
15	78	East	G15
16	84.4	East	G16
17	77.7	East	G17
18	76.8	East	G18
19	85.6	East	G19
20	73.3	East	G20

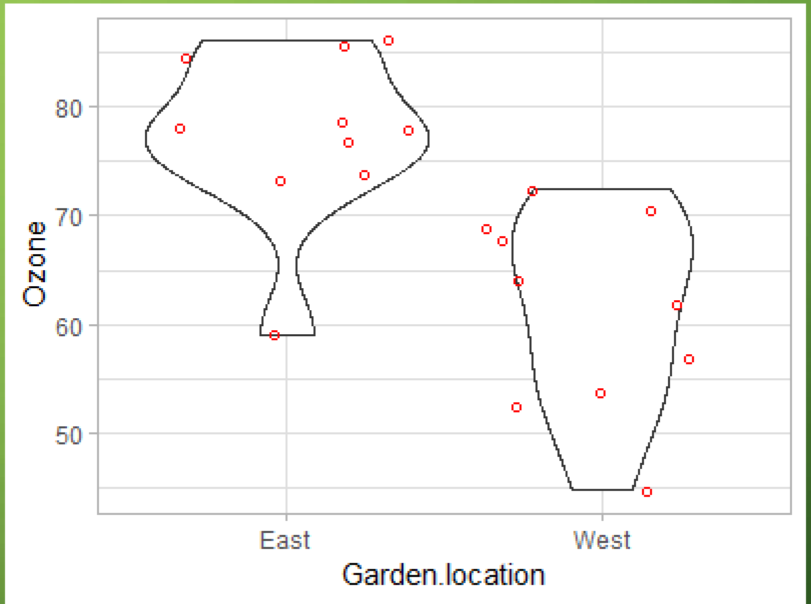
Two-sample *t*-test
II. Data Visualization

Boxplot




```
ozone %>%
     ggplot(data = ozone, aes(x =
Garden.location, y = Ozone)) +
     geom_boxplot() +
     geom_jitter(shape=1, color="red") +
     theme_bw()
```

Violin plot




```
Ozone %>%
     ggplot(data = ozone, aes(x = Garden.location, y
= Ozone)) +
     geom_violin() +
     geom_jitter(shape=1, color="red") +
     theme_bw()
```

Two-sample *t*-test

III. Run *t*-test

```r
# Do a t.test now....
t.test(Ozone ~ Garden.location, data = ozone)
```

```
##
##  Welch Two Sample t-test
##
## data:  Ozone by Garden.location
## t = 4.2363, df = 17.656, p-value = 0.0005159
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.094171 24.065829
## sample estimates:
## mean in group East mean in group West
##              77.34              61.26
```

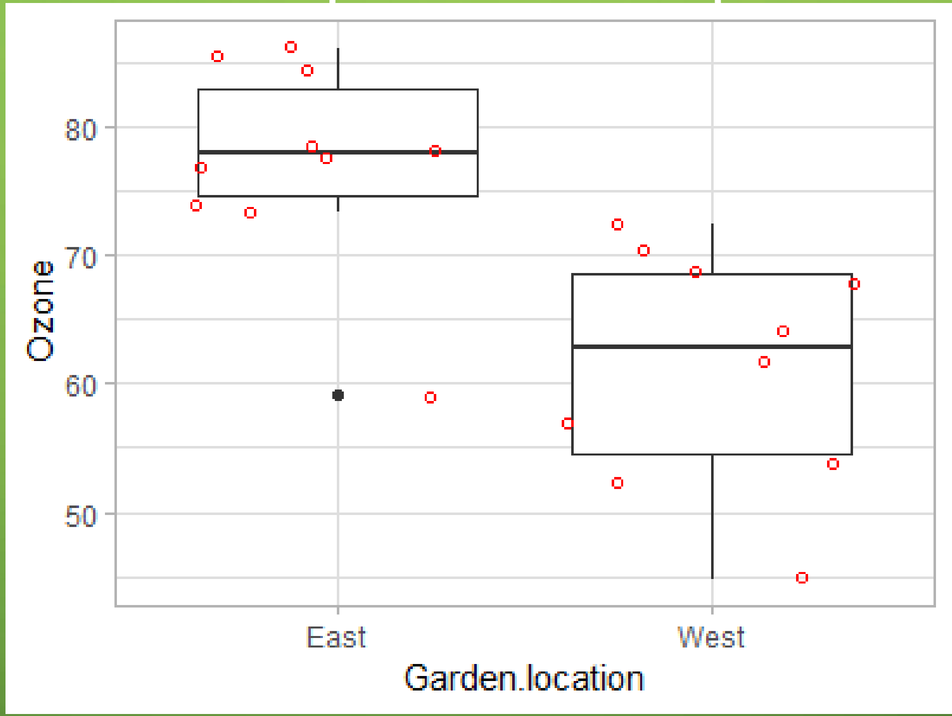
P value –

- Probability that observed difference is due to chance
- (more specifically) probability that t >= 4.2363 under null hypothesis (H_0)

T-test

IV. Re-plot & Conclude

$P = 5.2 \times 10^{-14}$



Conclusions:

- Statistical conclusion: The null hypothesis (same mean) is rejected at $p=5.2 \times 10^{-14}$
- Biological conclusion: The ozone level is significantly different between the east & west locations

Linear Regression

I. Data & Hypothesis

- Biological Question: Does soil moisture affect growth rate?
- Null Hypothesis (H_0): No correlation ($r=0$)

```
> plant_gr <- read_csv("plant.growth.rate.csv")
Parsed with column specification:
cols(
  soil.moisture.content = col_double(),
  plant.growth.rate = col_double()
)
> tbl_df(plant_gr)
# A tibble: 50 x 2
   soil.moisture.conte~ plant.growth.ra~
                  <dbl>            <dbl>
 1                0.470             21.3
 2                0.541             27.0
 3                1.70              39.0
 4                0.826             30.2
 5                0.857             37.1
 6                1.61              43.2
 7                0.250             22.7
 8                1.67              40.2
 9                1.46              46.9
10                0.473             28.8
# ... with 40 more rows
```

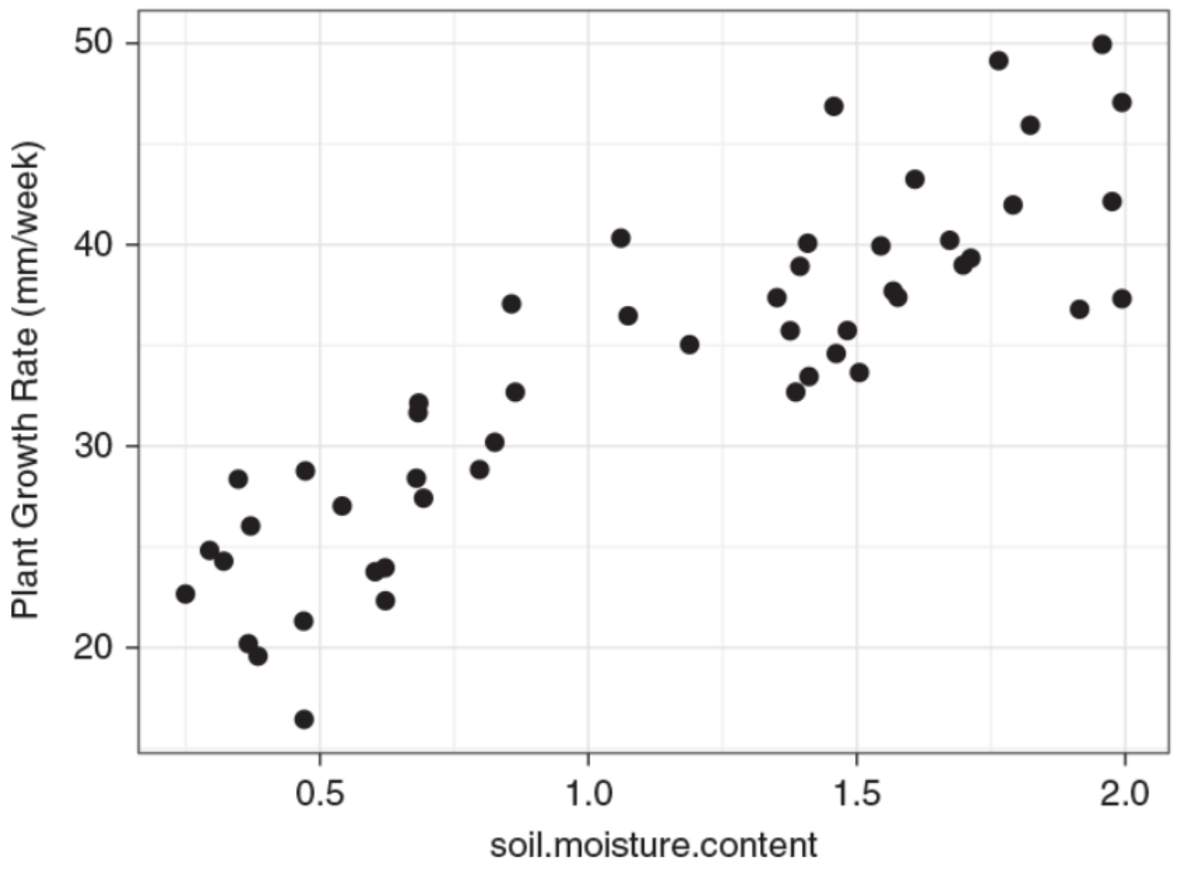
```
glimpse(plant_gr)

## Observations: 50
## Variables: 2
## $ soil.moisture.content (dbl) 0.4696876, 0.5413106, 1.6...
## $ plant.growth.rate     (dbl) 21.31695, 27.03072, 38.98...
```

Linear Regression

II. Visualization

```
ggplot(plant_gr,
       aes(x = soil.moisture.content, y = plant.growth.rate)) +
          geom_point() +
          ylab("Plant Growth Rate (mm/week)") +
          theme_bw()
```



Linear Regression

III. Run linear model

```r
model_pgr <- lm(plant.growth.rate ~ soil.moisture.content,
                data = plant_gr)
```

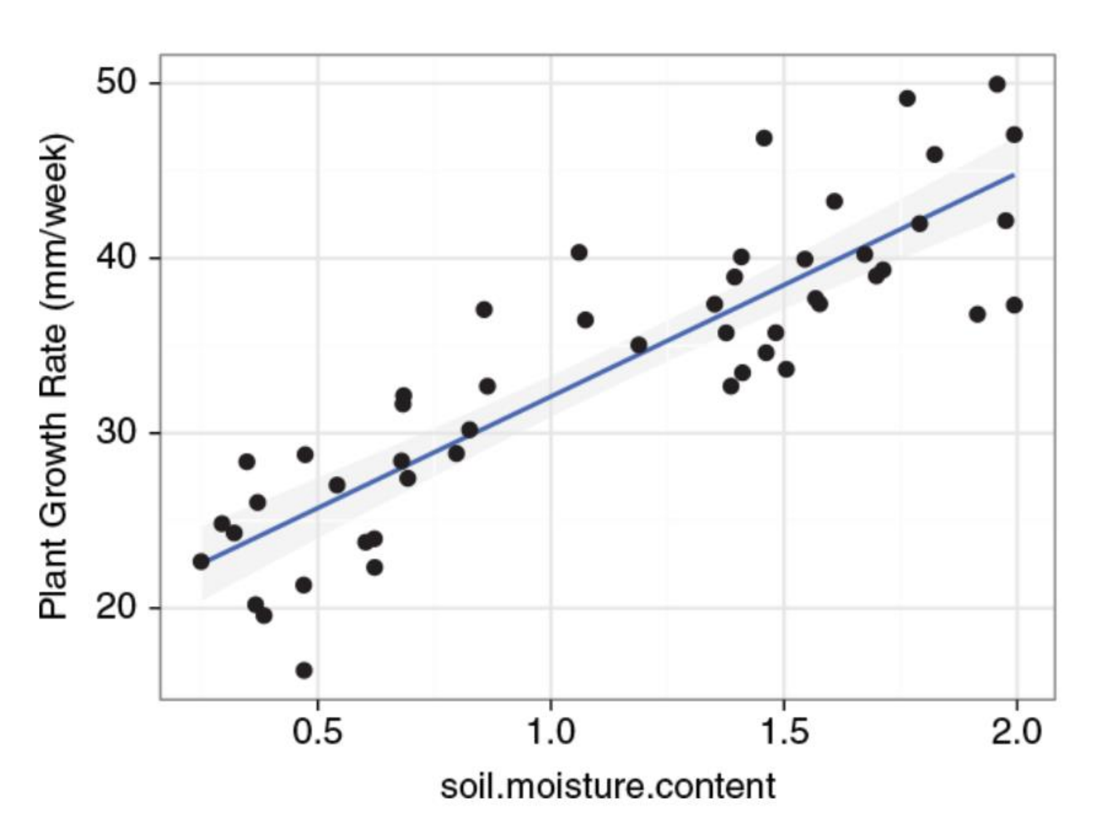
```r
summary(model_pgr)

##
## Call:
## lm(formula = plant.growth.rate ~ soil.moisture.content,
##    data = plant_gr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9089 -3.0747  0.2261  2.6567  8.9406
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             19.348      1.283   15.08   <2e-16
## soil.moisture.content   12.750      1.021   12.49   <2e-16
##
## (Intercept)           ***
## soil.moisture.content ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.019 on 48 degrees of freedom
## Multiple R-squared:  0.7648, Adjusted R-squared:  0.7599
## F-statistic: 156.1 on 1 and 48 DF,  p-value: < 2.2e-16
```

Conclusions:

- The null hypothesis (no correlation) is rejected at $p<2.2e\text{-}16$
- The plant growth rate is significantly correlated with soil moisture with $R^2=0.7599$

Linear Regression
IV. Re-plot (add regression line & confidence band)




```
ggplot(plant_gr, aes(x = soil.moisture.content,
          y = plant.growth.rate)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  ylab("Plant Growth Rate (mm/week)") +
  theme_bw()
```

Data Analysis Workflow

Hypothesis

- Define the biological question
- Design a null hypothesis (H_0)
- Define an alternative hypothesis (H_1)

Visualization

- Identify variables and variable types
- Show data distributions (transform if necessary)
- Show data relations

Modeling

- Run statistical tests (e.g., *t*-test, linear regression)
- Interpret test results (e.g., is null hypothesis rejected?)

Conclude

- Re-plot
- Draw statistical conclusions
- Draw biological conclusions

PRACTICE #4

- Does the "Sepal.Length" differ between the two species "virginica" & "vesicolor"? Perform a *t*-test and include all 4 steps
- How about the "Sepal Width"? Perform a *t*-test and include all 4 steps
- Are the "Sepal.Width" and "Sepal.Length" correlated in the species "setosa"? Show all 4 steps.
- How about in the other two species?
- Batch testing the above correlation in all 3 species at once
- Save all commands to a file "practice-4.R"