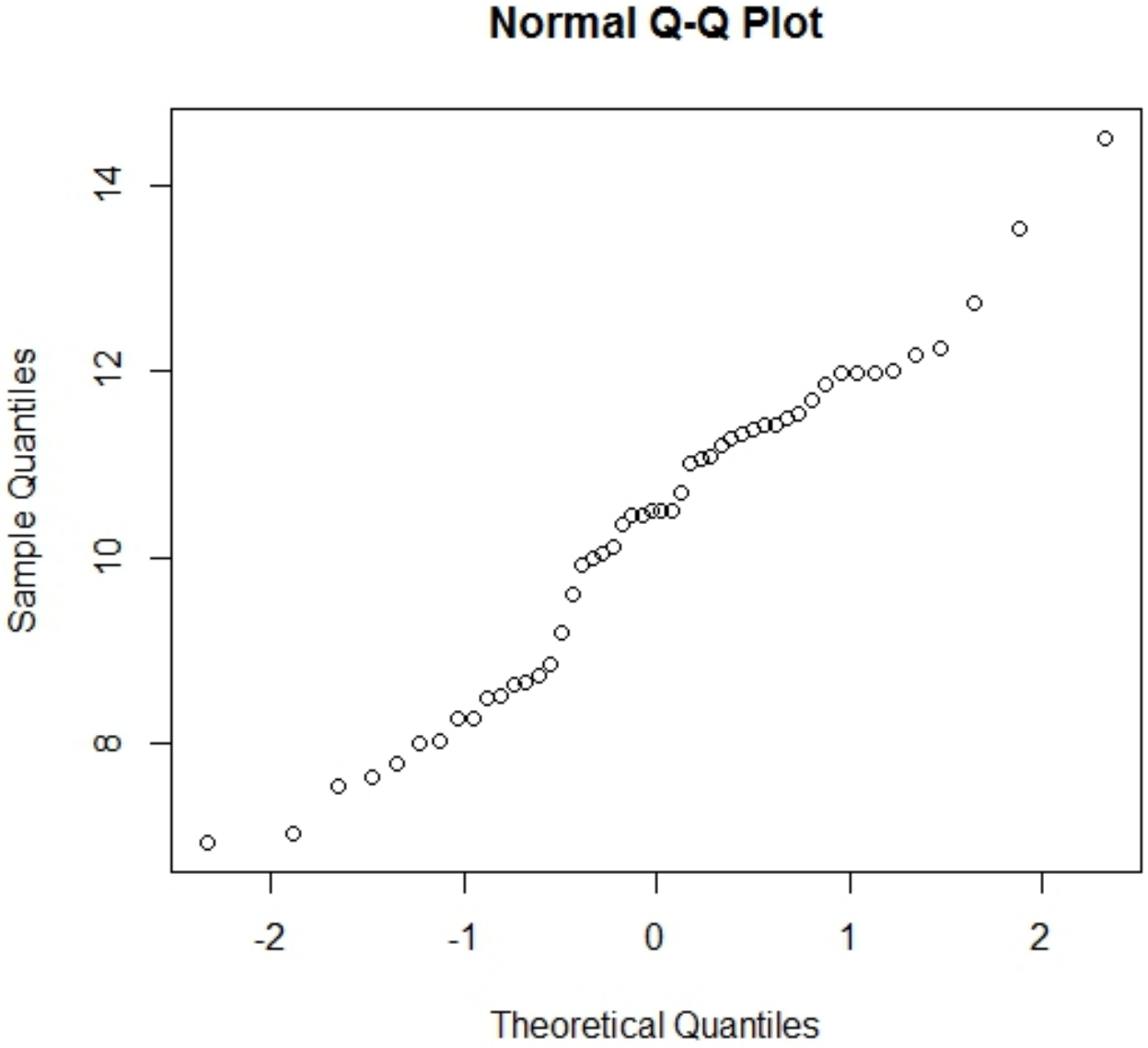
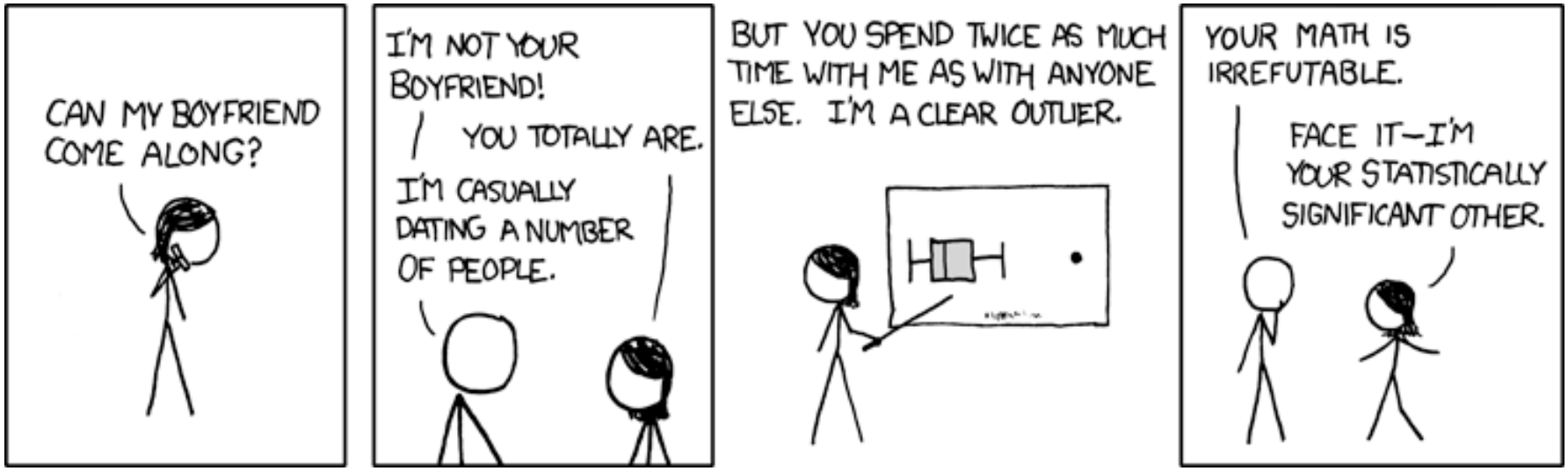
Week 4, February 3rd 2017


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
8
10
12
14
-2
-1
0
1
2

SINCE

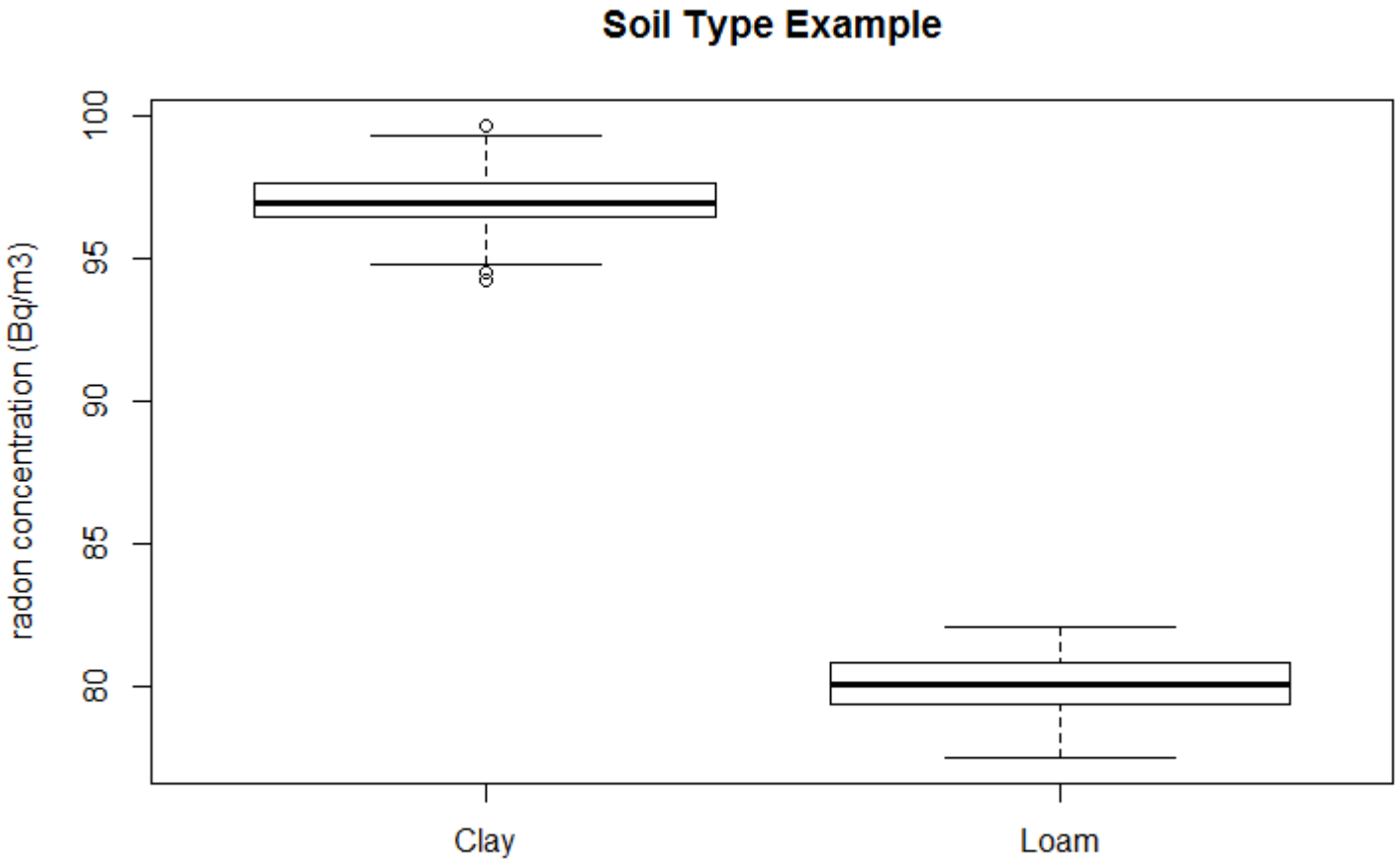
BIAS / BIASED / UNBIASED

SIGNIFICANT



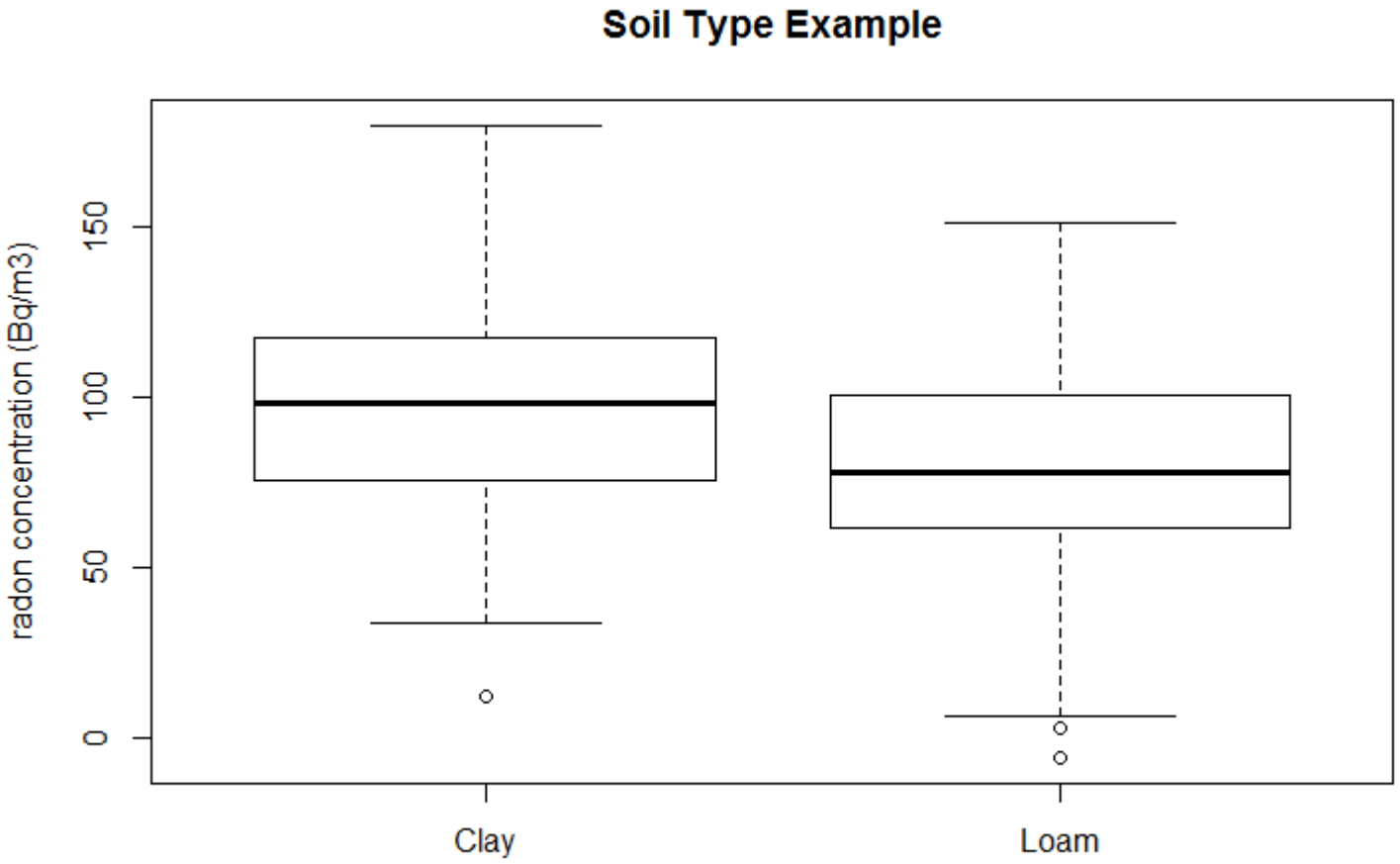
Statistical Hypothesis Testing

- We are evaluating the probability of observing the value of a statistic by chance alone
- The statistic depends on the test that we are running
- The probability of observing the statistic is expressed by the p-value
- We use this to evaluate whether our observation is real or random



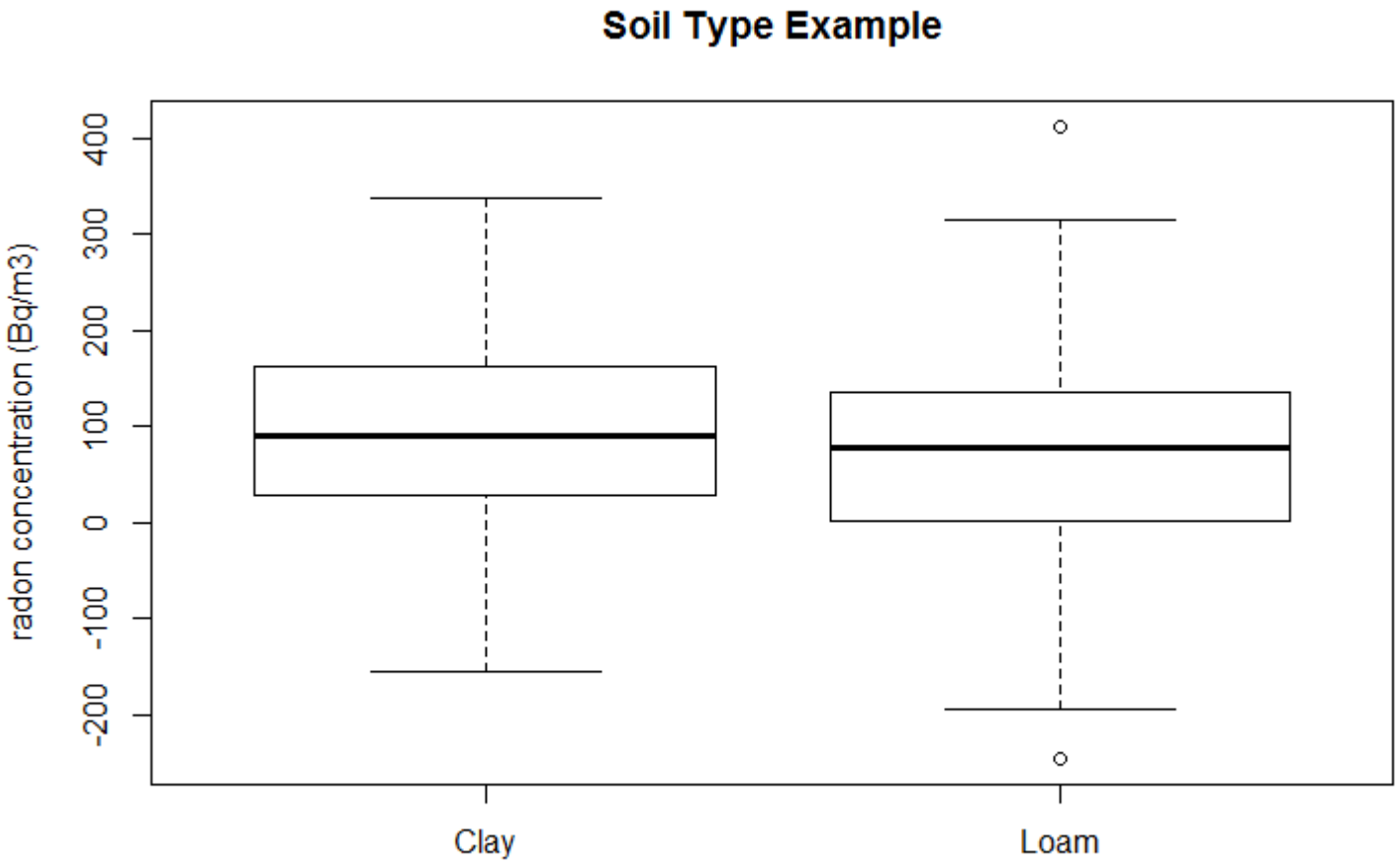
Statistical Hypothesis Testing

- We are evaluating the probability of observing the value of a statistic by chance alone
- The statistic depends on the test that we are running
- The probability of observing the statistic is expressed by the p-value
- We use this to evaluate whether our observation is real or random



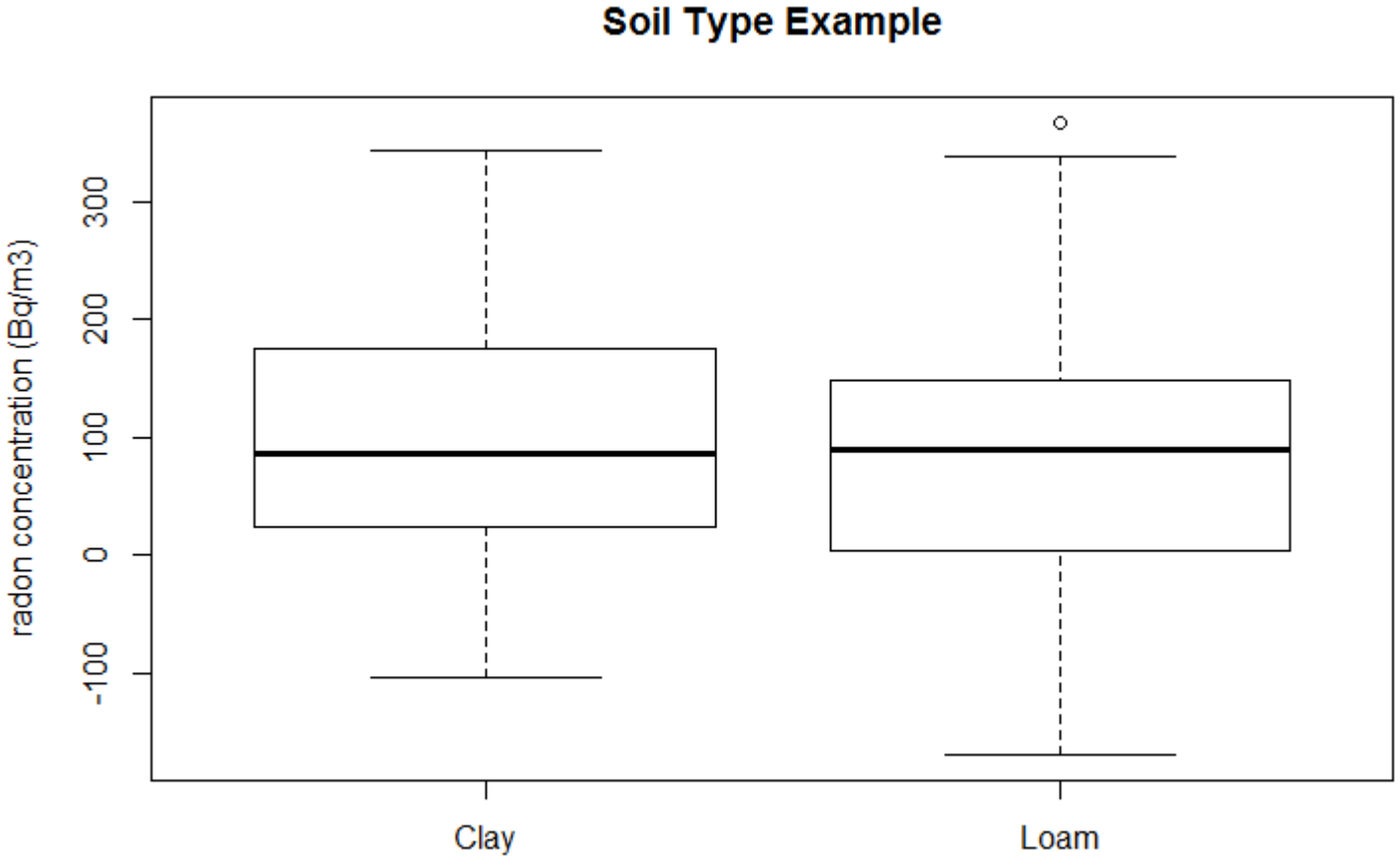
Statistical Hypothesis Testing

- We are evaluating the probability of observing the value of a statistic by chance alone
- The statistic depends on the test that we are running
- The probability of observing the statistic is expressed by the p-value
- We use this to evaluate whether our observation is real or random



Statistical Hypothesis Testing

- We are evaluating the probability of observing the value of a statistic by chance alone
- The statistic depends on the test that we are running
- The probability of observing the statistic is expressed by the p-value
- We use this to evaluate whether our observation is real or random



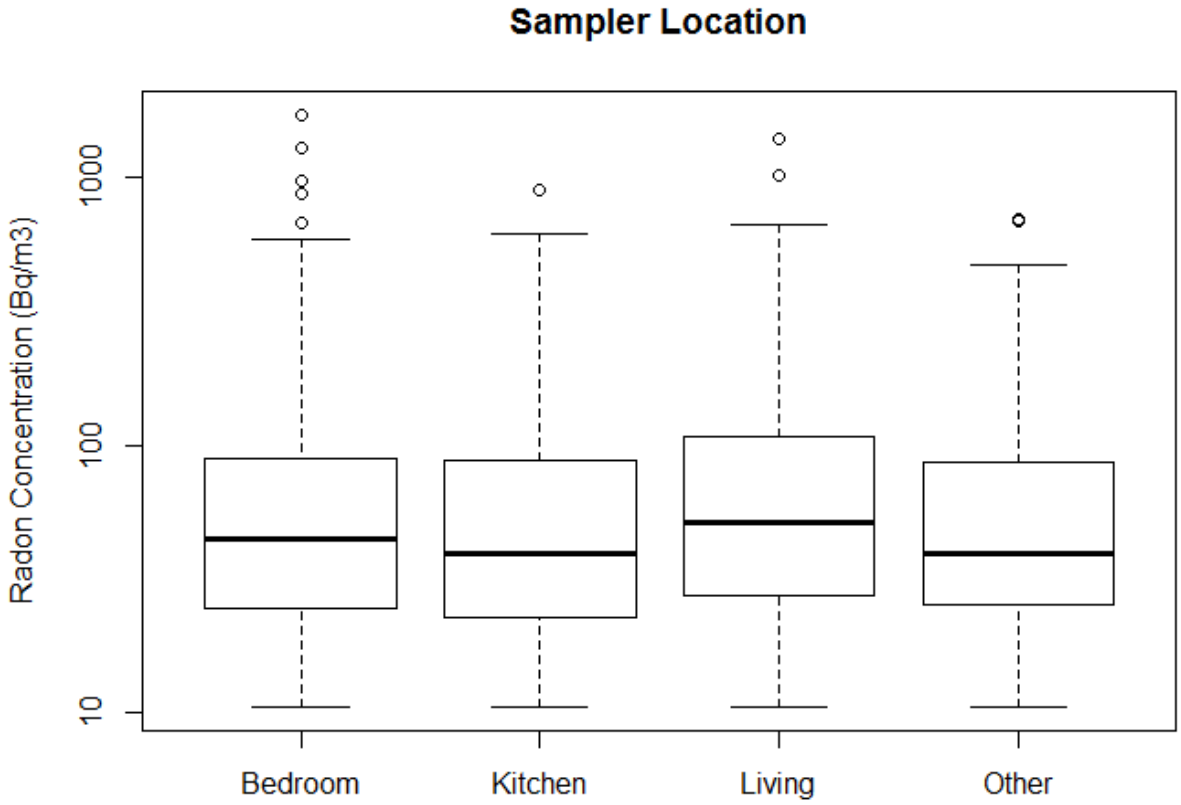
Assignment #3

Short report

- ***INTRODUCTION:*** *should provide background for why you are conducing this analysis, and should include your hypothesis about the relationship between radon and the variable you have chosen. You must include at least one citation to the peer-reviewed literature that supports the thinking behind your hypothesis.*
- ***METHODS:*** *describe the methods that you used to evaluate the association between radon and your chosen variables.*
- ***RESULTS:*** *describe the results of your analyses with the assistance of tables and figures, if necessary. Tables and figures should be properly labelled and referenced in the text. It is preferable that you structure your report as elegantly as possible. This means that you describe the result and refer to the table or figure in parentheses following that description. For example, I would like to see "The mean radon concentration for category 1 was XX.X Bq/m3 compared with XX.X Bq/m3 in category 2 (Table 1)" rather than this "Table 1 summarizes the mean radon concentrations in each category". Most good journals will not accept the latter, as it does not provide a flowing narrative for the reader because they must go look at the table to get the information necessary to interpret the rest of the paper.*
- ***DISCUSSION:*** *what did you find and what does it mean? Please end with a concluding statement about the relationship between the variables in your data.*

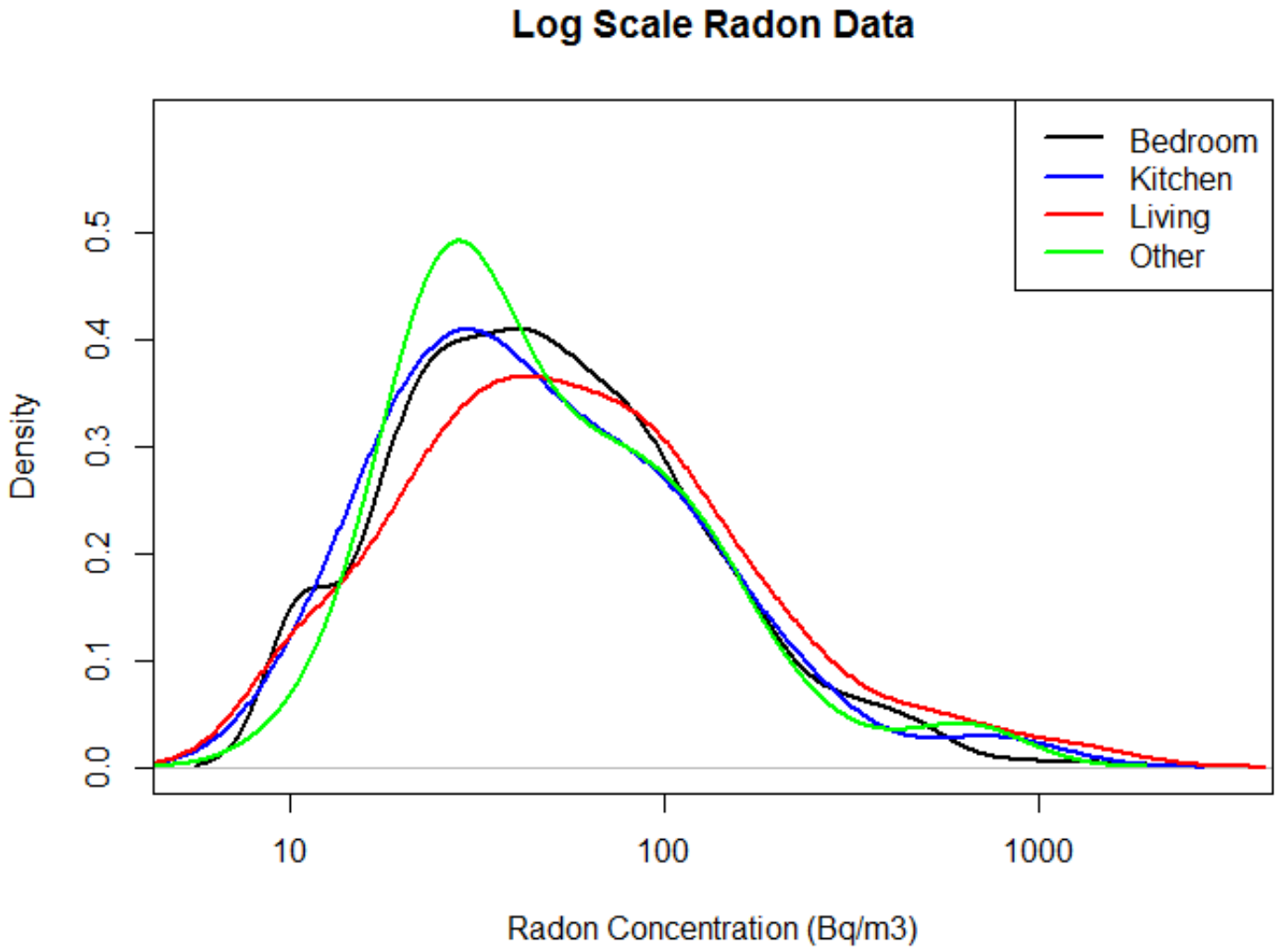
Categorical Variables

- What are they?
- Which variables in the radon dataset (as provided) are categorical?
- What hypotheses do we have about the association between these variables and radon concentrations?
- What other categorical variables would be nice to have in the dataset?
- How much data are we omitting due to missing information?



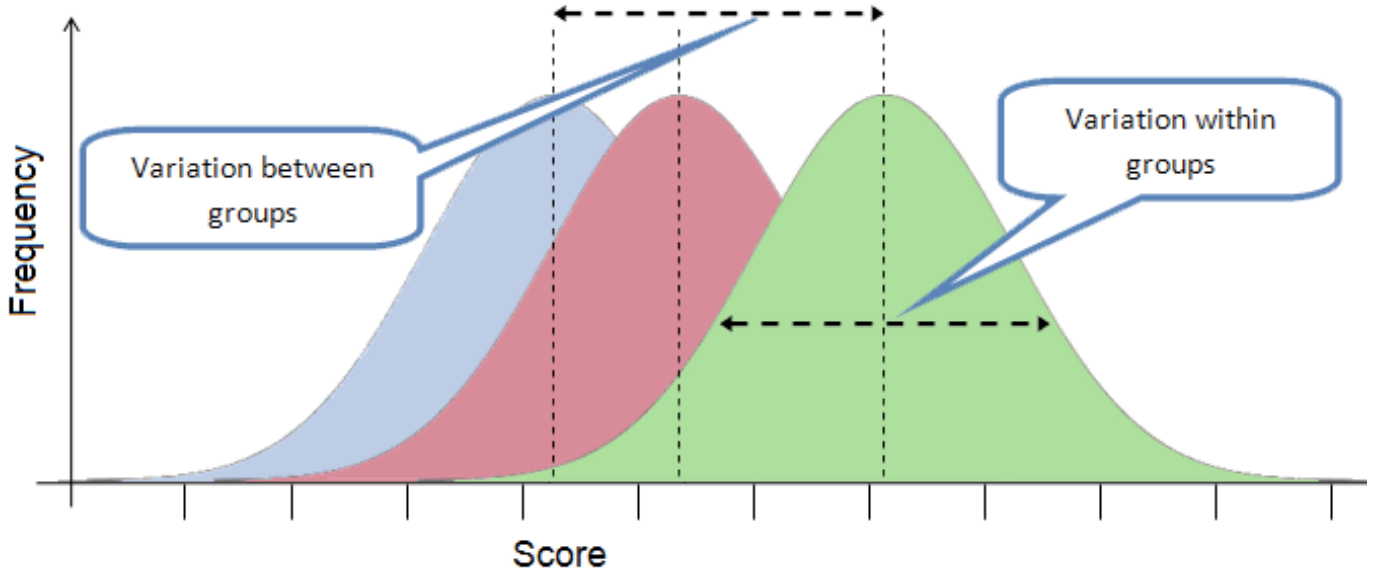
Density Plots

- Are we likely to see a statistically significant difference between these means?
- What is the reasoning behind your answer?



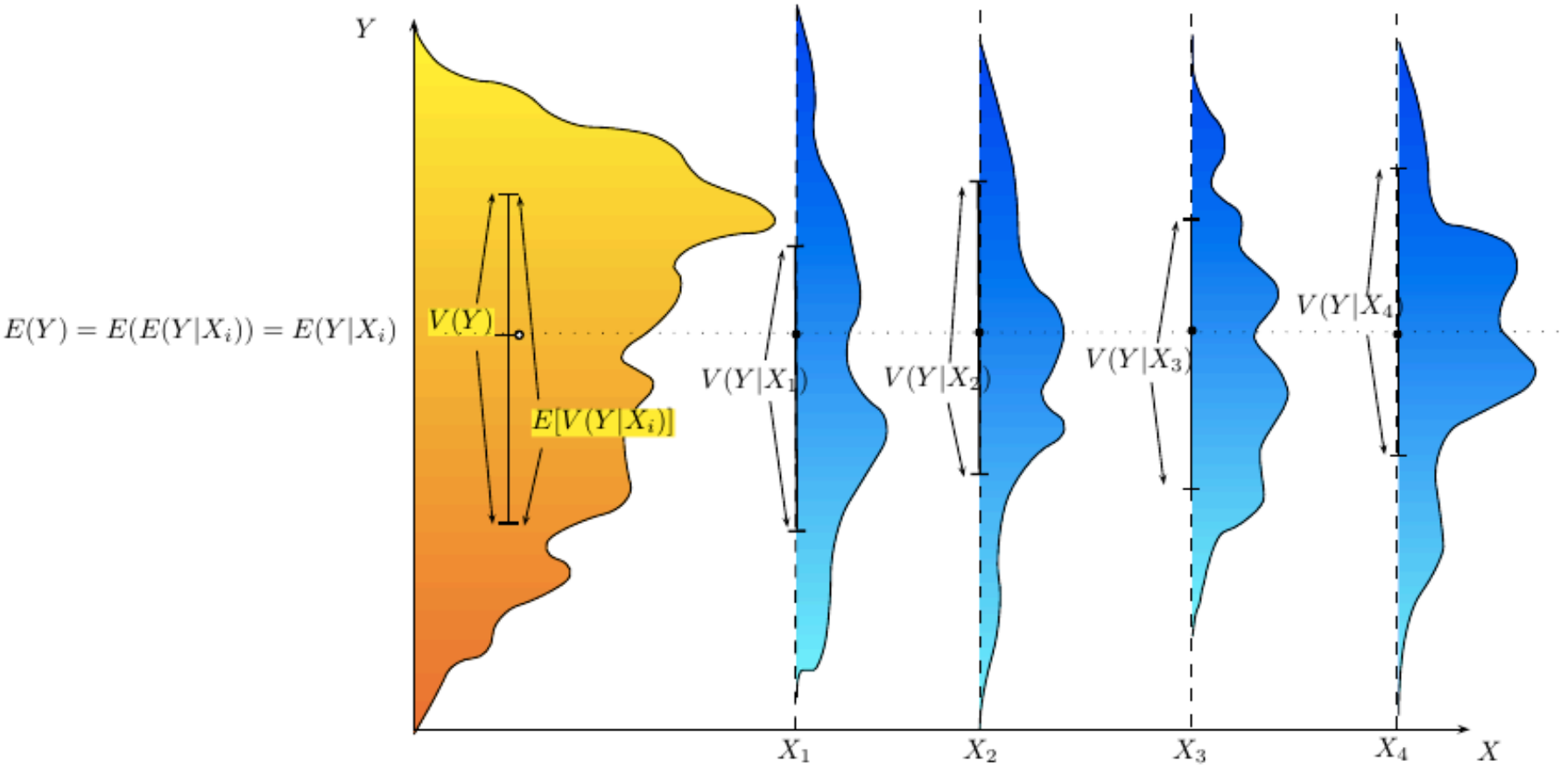
ANOVA on Multiple Means

- ANOVA = Analysis of Variance
- A one-way ANOVA is used when comparing the means of a continuous DEPENDENT variable across more than two groups of a categorical INDEPENDENT variable
- The t-test is a special case of ANOVA that is used when there are only two categories
- The one-way ANOVA separates variability into two components: BETWEEN groups and WITHIN groups
- Between groups is the sum of the square difference between each individual group mean and the GRAND MEAN
- Within groups is the sum of the square differences between each individual observation and the group mean



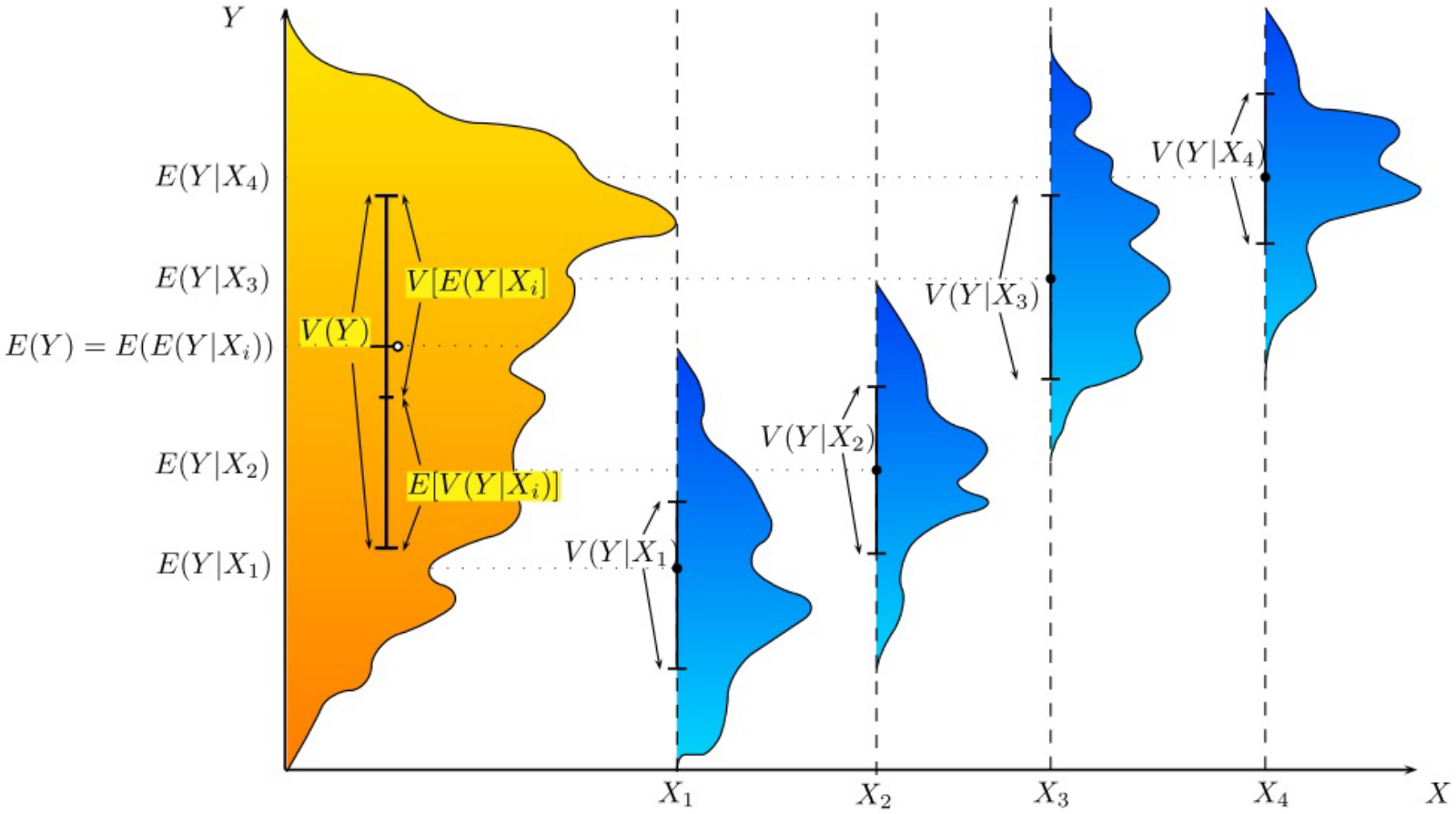
No Association

- Don’t worry about the equations, just look at the pictures (from Wikipedia)



Weak Association

- Don’t worry about the equations, just look at the pictures (from Wikipedia)



Sums of Squares

- ANOVA tests the association between a the categorical variable as a WHOLE and the continuous variable
- To generate the test statistic you must calculate the sum of squares (SS) and degrees of freedom (df) for the between group (SS_B, df_B) and within group (SS_W, df_W) portions of the variability

```
>   bedroom; n=869
[1] 80.13262
>   kitchen; n=68
[1] 82.94892
>   living; n=112
[1] 110.039
>   other; n=78
[1] 81.54504
>   grand; n=1127
[1] 83.37236
```

$SS_B = 869 * (83.37 - 80.13)^2 + 68 * (83.37 - 82.95)^2 + 112 * (83.37 - 110.04)^2 + 78 * (83.37 - 81.54)^2 =$

89038

$df_B =$ number of groups $- 1 = 3$

Sums of Squares

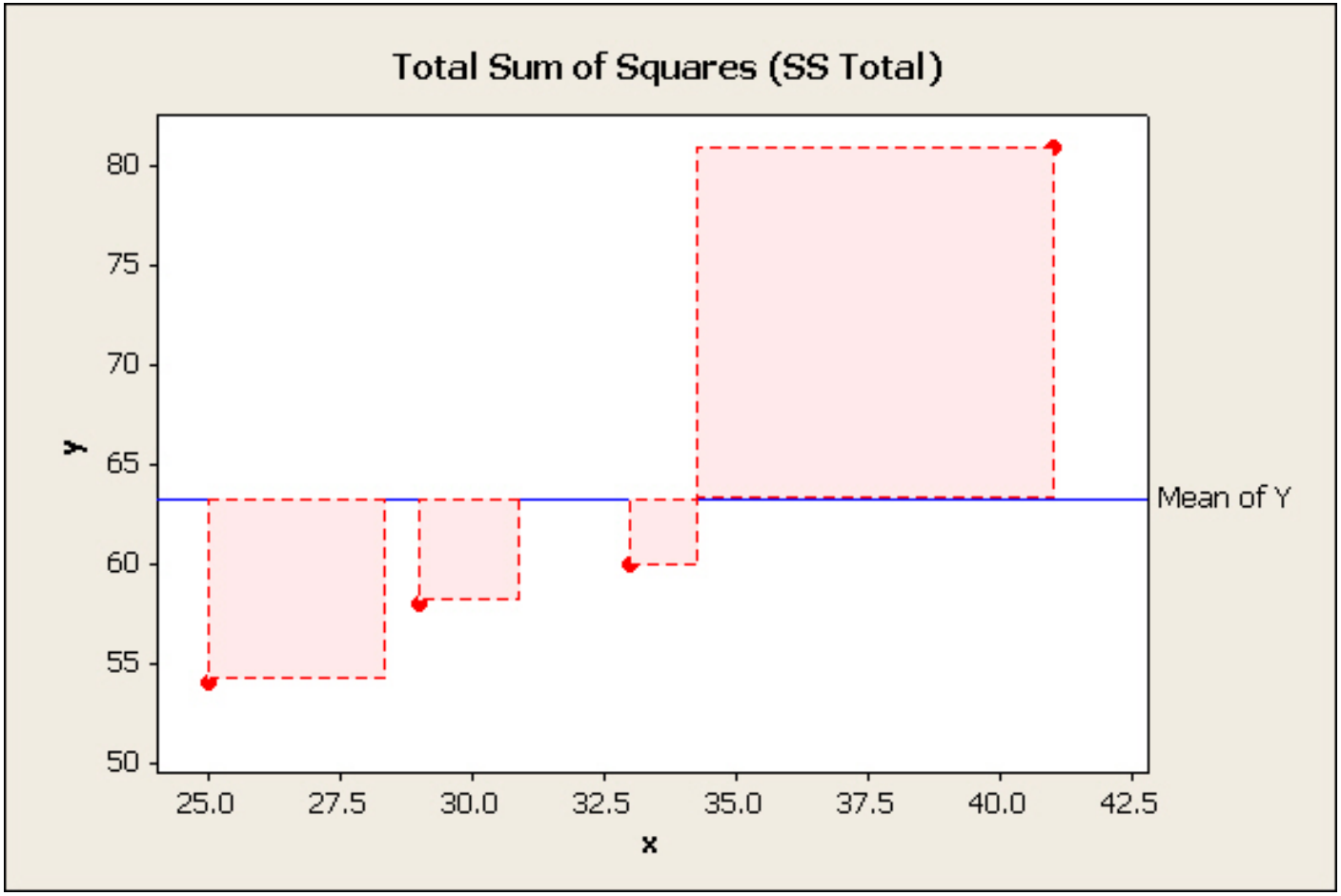
```
>   bedroom; n=869
[1] 80.13262
>   kitchen; n=68
[1] 82.94892
>   living; n=112
[1] 110.039
>   other; n=78
[1] 81.54504
>   grand; n=1127
[1] 83.37236
```

$SS_W = \Sigma(80.13 - x_{bedroom})^2 + \Sigma(82.95 - x_{kitchen})^2 + \Sigma(110.04 - x_{living})^2 + \Sigma(81.54 - x_{other})^2 =$

18467397

$df_W =$ n observations – n groups =
1127 (7 NA values) – 4 = 1123

Sums of Squares



The F Statistic

- The F statistic evaluates the mean SS per degree of freedom BETWEEN groups divided by the mean SS per degree of freedom WITHIN groups
- The H_0 is that the population means in each group are the same
- In other words: $\mu_{bedroom} = \mu_{kitchen} = \mu_{living} = \mu_{other}$
- Strictly speaking it assumes normality within groups, but it is robust with positively skewed data…you just have a higher chance of falsely rejecting the null hypothesis

$$F = \frac{SS_B/df_B}{SS_W/df_W} + \frac{89038/3}{18467397/1123} = 29679/16445 = 1.805$$

```
                  Df   Sum Sq Mean Sq F value Pr(>F)
radon$Location     3    89038   29679   1.805  0.145
Residuals       1123 18467397   16445
7 observations deleted due to missingness
```

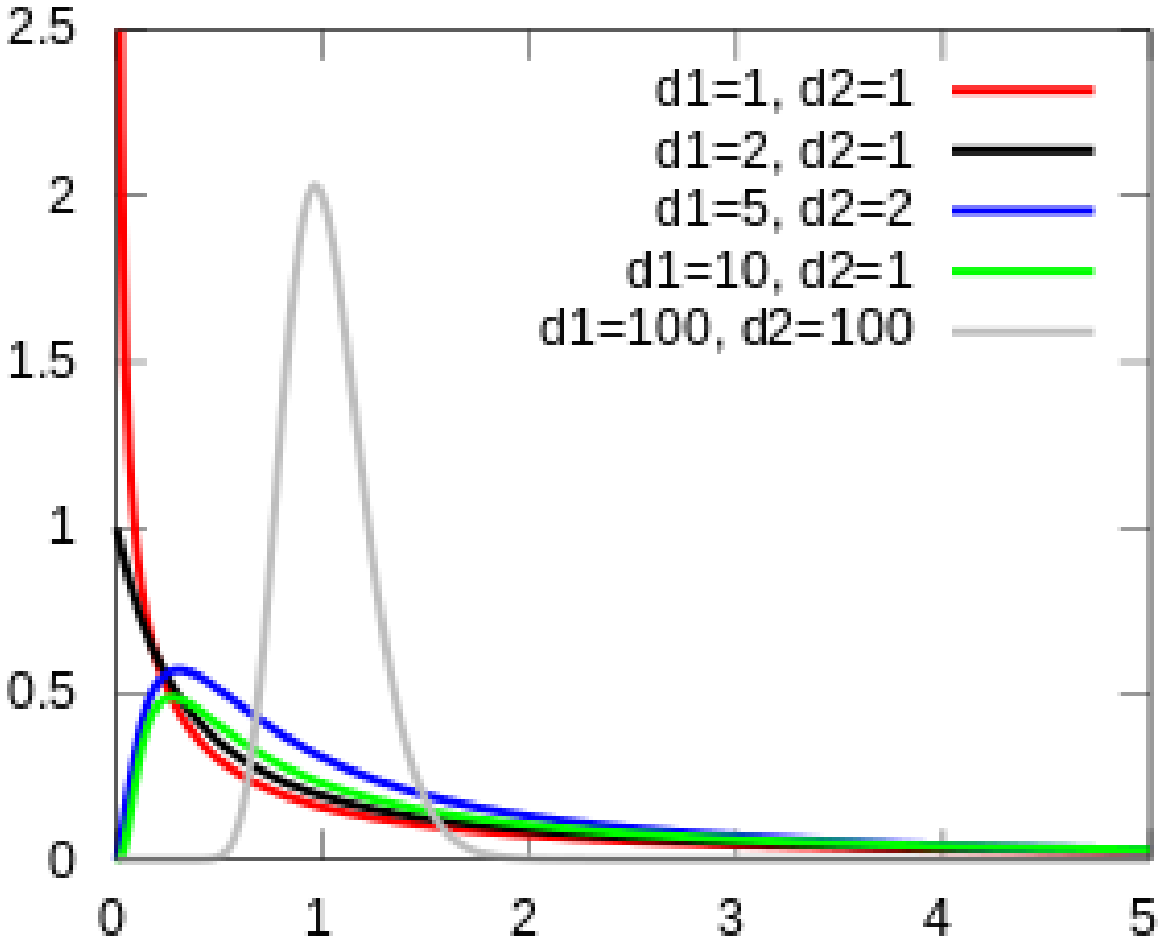
Untransformed data

```
                  Df Sum Sq Mean Sq F value Pr(>F)
radon$Location     3    2.7  0.9072   1.026   0.38
Residuals       1123  992.9  0.8841
7 observations deleted due to missingness
```

Log-transformed data

The F Distribution

- The shape of the F distribution depends on the df_B and df_W
- Rely on your software to give you the critical values



Tukey Tests

- ANOVA H_0 is that the means in each group are the same
- H_1 is that the population mean for AT LEAST ONE group is different
- ANOVA cannot tell you which mean(s) is/are different, you need to run a post-hoc (follow-up) test for that
- A Tukey test is similar to a series of pairwise t-tests, but it accounts for the type I error we would expect if doing a series of t-tests.
- i.e. if we did 20 pairwise t-tests we would expect to erroneously reject the null hypothesis in one case, whereas we do not expect that with the Tukey Honest Significant Difference (HSD) test

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

Fit: aov(formula = radon$MainRadon ~ radon$Location)

$`radon$Location`
                      diff         lwr      upr     p adj
Kitchen-Bedroom   2.816295 -38.730572 44.36316 0.9981150
Living-Bedroom   29.906360  -3.218072 63.03079 0.0934460
Other-Bedroom     1.412419 -37.586327 40.41116 0.9997101
Living-Kitchen   27.090065 -23.633016 77.81315 0.5158964
Other-Kitchen    -1.403877 -56.144219 53.33647 0.9998969
Other-Living    -28.493941 -77.151837 20.16395 0.4337747
```

Dummy Variables

- A categorical variable with four groups is going to be converted into THREE dummy variables with ONE reference category
- Each dummy variable is going to get a coefficient in the model
- Interpretation of the coefficients for categorical variables is a simple extension of the interpretation for dichotomous variables

Category	Value for DV1	Value for DV2	Value for DV3
Bedroom (reference)	0	0	0
Kitchen	1	0	0
Living	0	1	0
Other	0	0	1

$$Y = \beta_0 + \beta_{1DV1}X_{DV1} + \beta_{1DV2}X_{DV2} + \beta_{1DV3}X_{DV3}$$

- β_0 is still indicates the mean of the reference category
- β_{1DV1} is the the coefficient for the first dummy variable, so it indicates the effect of having the sampler in the KITCHEN ($X_{DV1} = 1$, $X_{DV2} = 0$, $X_{DV3} = 0$), compared with the BEDROOM
- β_{1DV2} is the the coefficient for the second dummy variable, so it indicates the effect of having the sampler in the LIVING area ($X_{DV1} = 0$, $X_{DV2} = 1$, $X_{DV3} = 0$), compared with the BEDROOM
- β_{1DV3} is the the coefficient for the third dummy variable, so it indicates the effect of having the sampler in an OTHER area ($X_{DV1} = 0$, $X_{DV2} = 0$, $X_{DV3} = 1$), compared with the BEDROOM
- In JMP you will need to set your variables to NOMINAL (rather than ORDINAL, even if they are ordinal) to interpret them this way
- For any ordinal variables it makes sense to use the LOWEST or HIGHEST value as the reference category

MainRadon = $\beta_0 + \beta_1$*location

- This is a report from R, but we can use it to get all of the information that you would find in any statistical software program
- What is the reference category?
- What is the mean in the reference category?
- What are the effects of the other categories?
- What are the confidence intervals around those effects?
- How much of the variation in MainRadon did we explain?

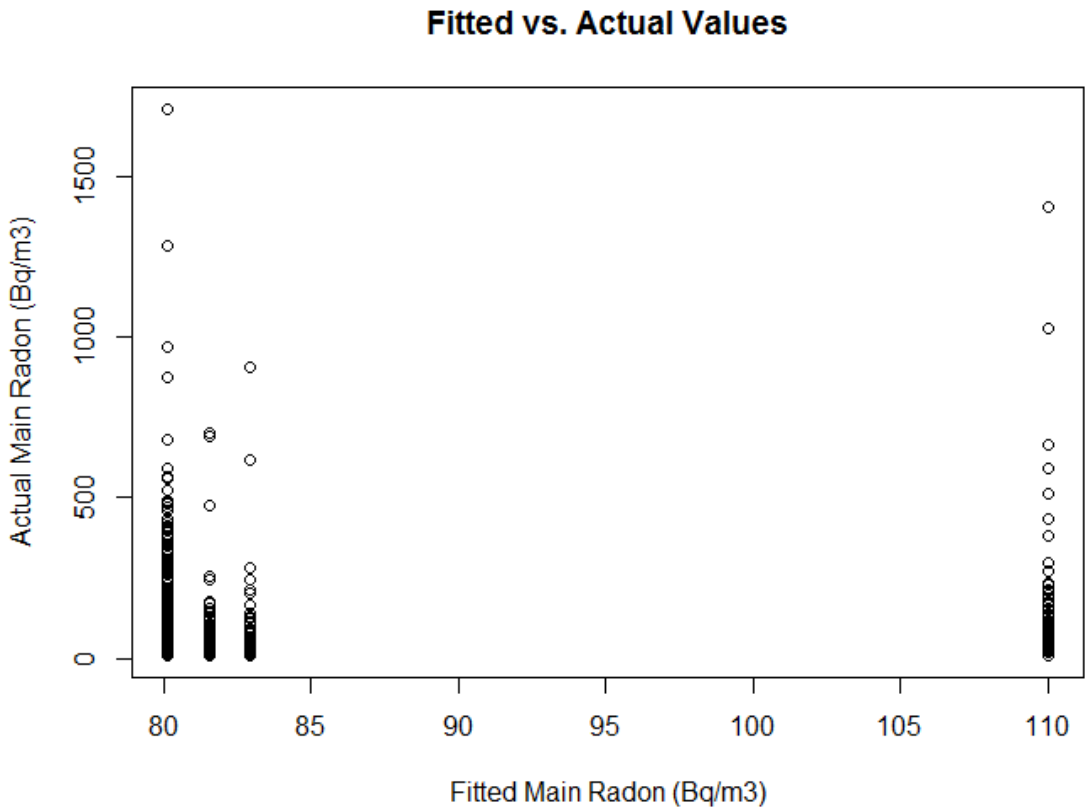
```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             80.133      4.350  18.421   <2e-16 ***
radon$LocationKitchen    2.816     16.148   0.174   0.8616
radon$LocationLiving    29.906     12.874   2.323   0.0204 *?
radon$LocationOther      1.412     15.158   0.093   0.9258
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 128.2 on 1123 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.004798,  Adjusted R-squared:  0.00214
F-statistic: 1.805 on 3 and 1123 DF,  p-value: 0.1445
```

Does this look familiar?!?

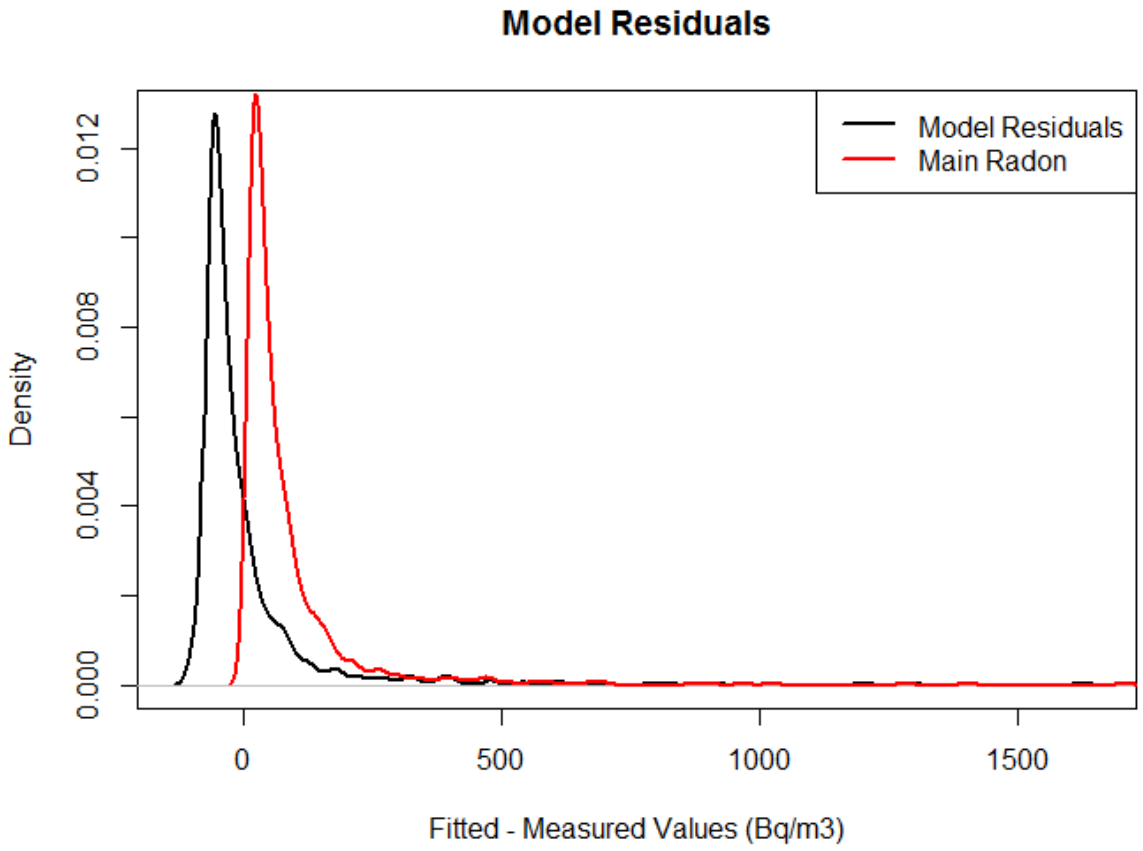
Fitted Values

- Which group indicates which category?
- Linear regression assumes that RESIDUAL values follow a normal distribution
- What are RESIDUAL values?
- Do you think these would follow a normal distribution?



Residual Values

- Arithmetic mean residual = 0 Bq/m^3 – is this surprising?
- Does it make sense that log-normally distributed data lead to violation of the assumption of normally distributed residuals?

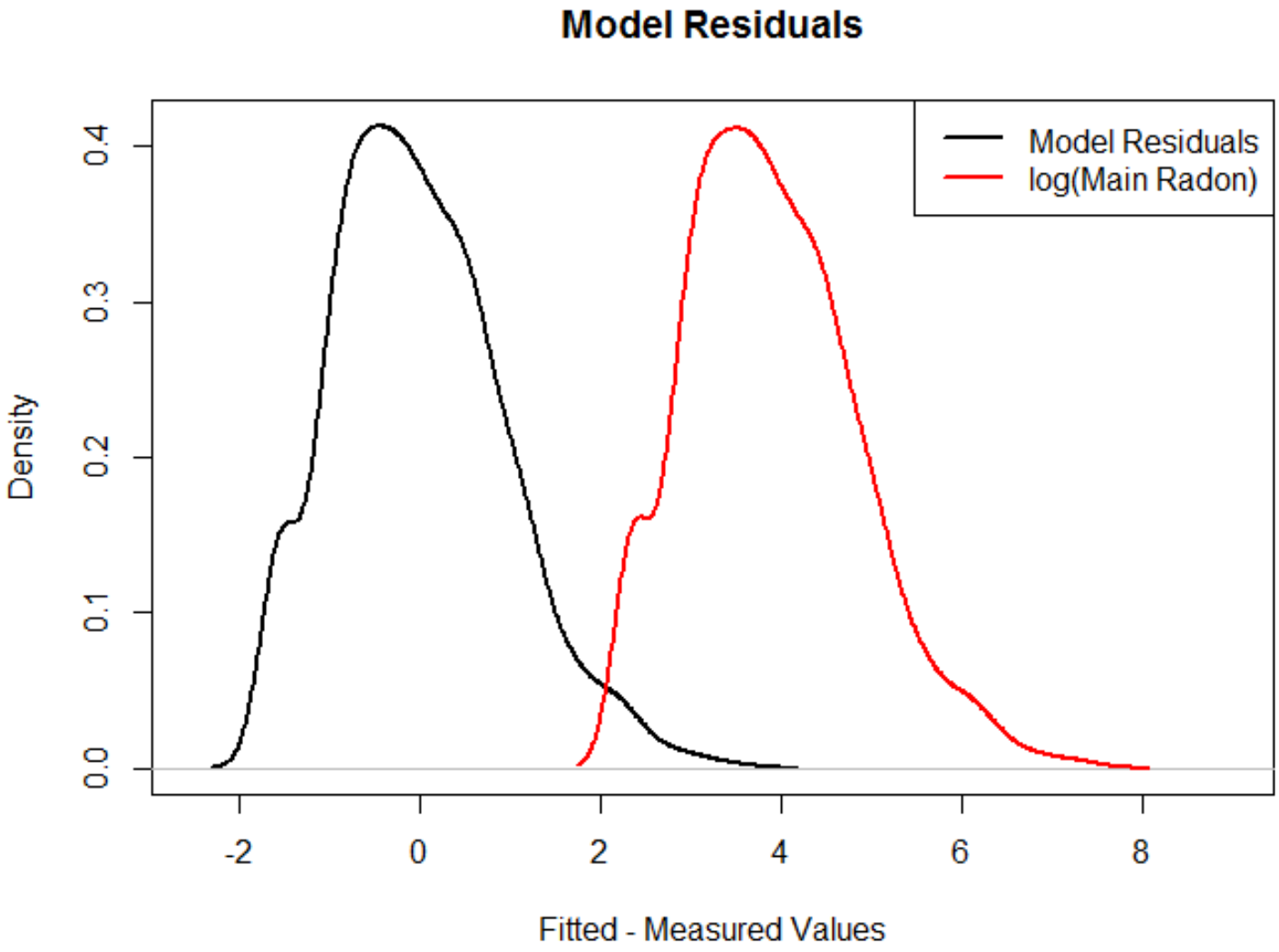


$\log(\text{MainRadon}) = \beta_0 + \beta_1 * \text{location}$

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.890090   0.031897 121.958   <2e-16 ***
radon$LocationKitchen -0.026812   0.118403  -0.226   0.8209
radon$LocationLiving   0.161388   0.094400   1.710   0.0876 .
radon$LocationOther    0.005386   0.111142   0.048   0.9614
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9403 on 1123 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.002734,  Adjusted R-squared:  6.958e-05
F-statistic: 1.026 on 3 and 1123 DF,  p-value: 0.3801
```

Residual Values



Next Week

- Assessing the relationship between two continuous variables
- Scatter plots to visualize
- Pearson's correlation
- Hypothesis generation
- Simple linear regression PART III
- Least squares regression
- Standard reporting
- Model diagnostics