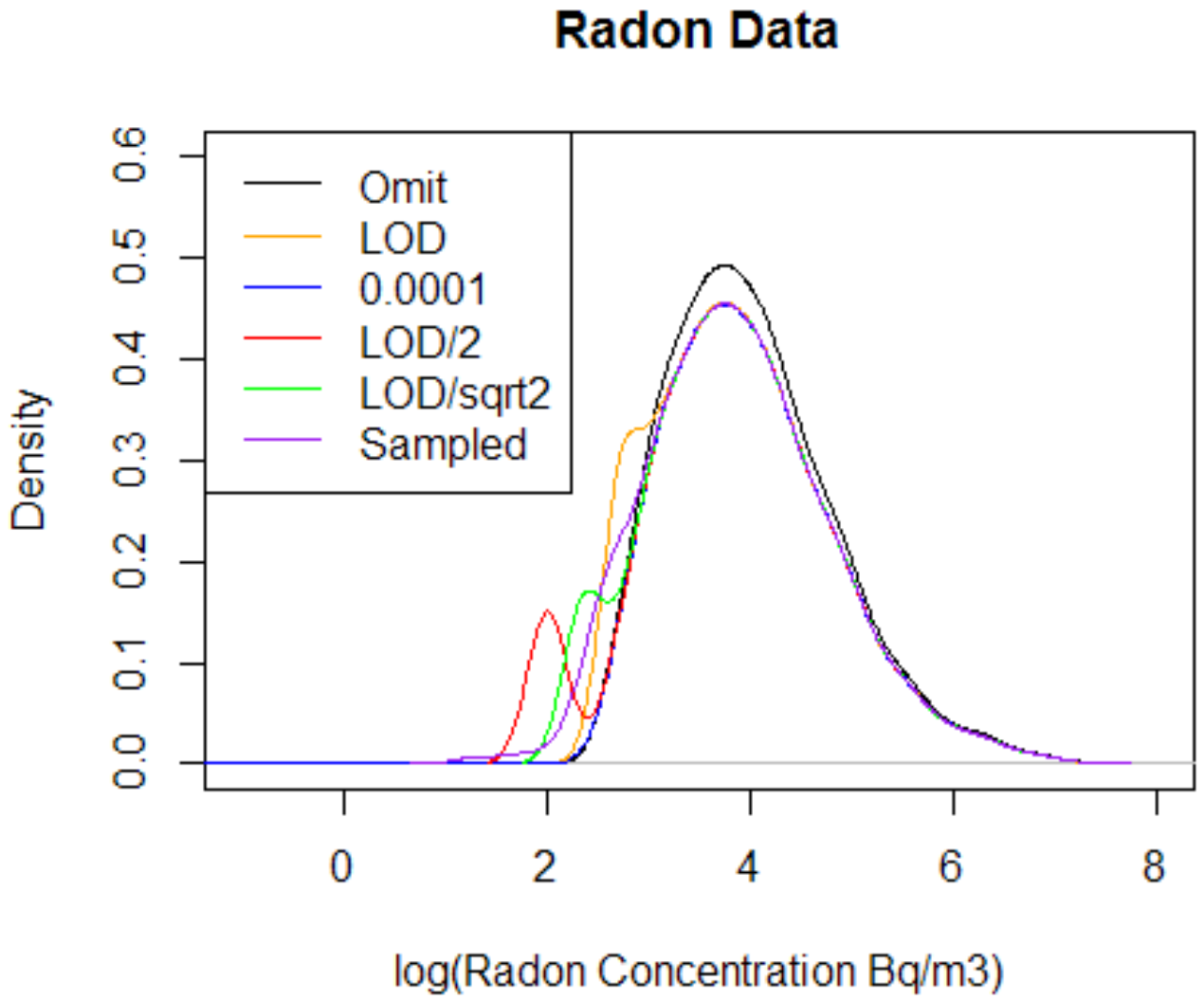
Week 3, January 27th 2017

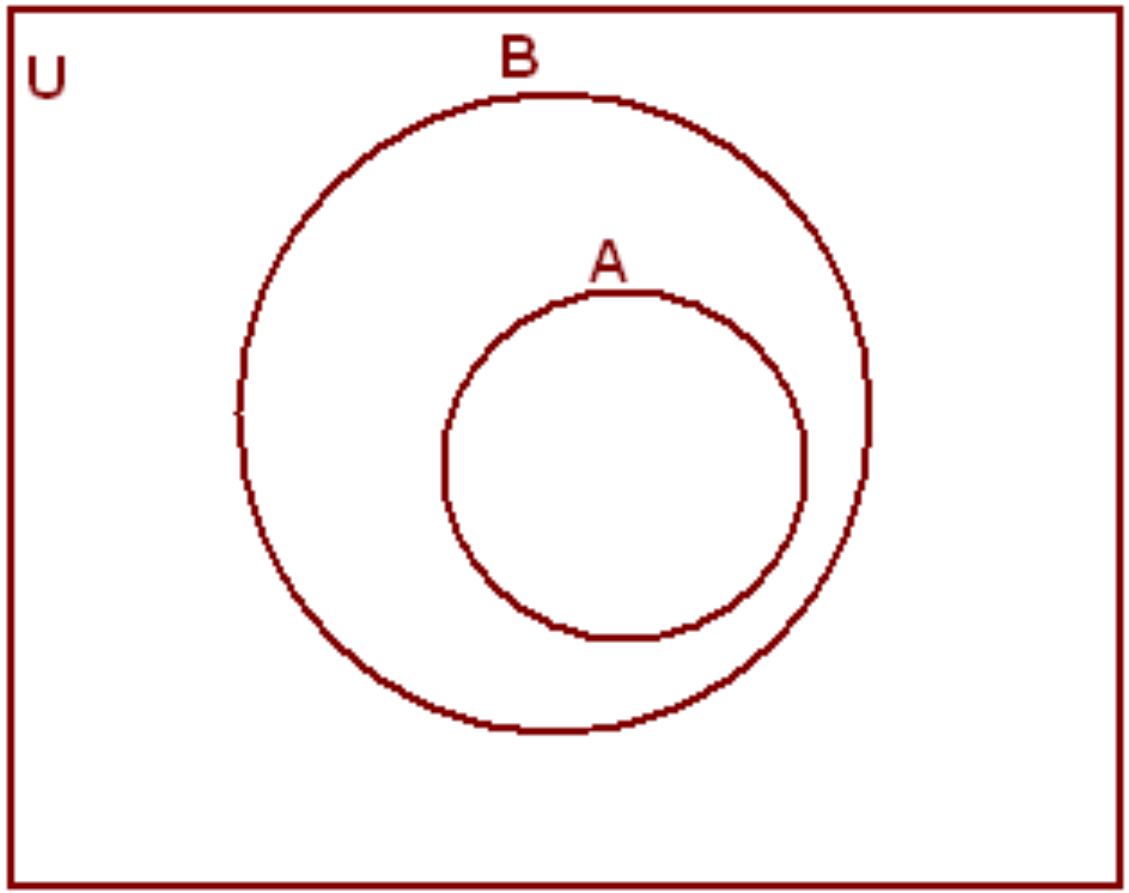
Assignment #1

- What did you feel confident about?
- What did you not feel confident about?



Assignment #2

- The most important thing Sarah considers when hiring new staff
- One of the most important things Sarah will consider when reading your assignments



A is subset of B

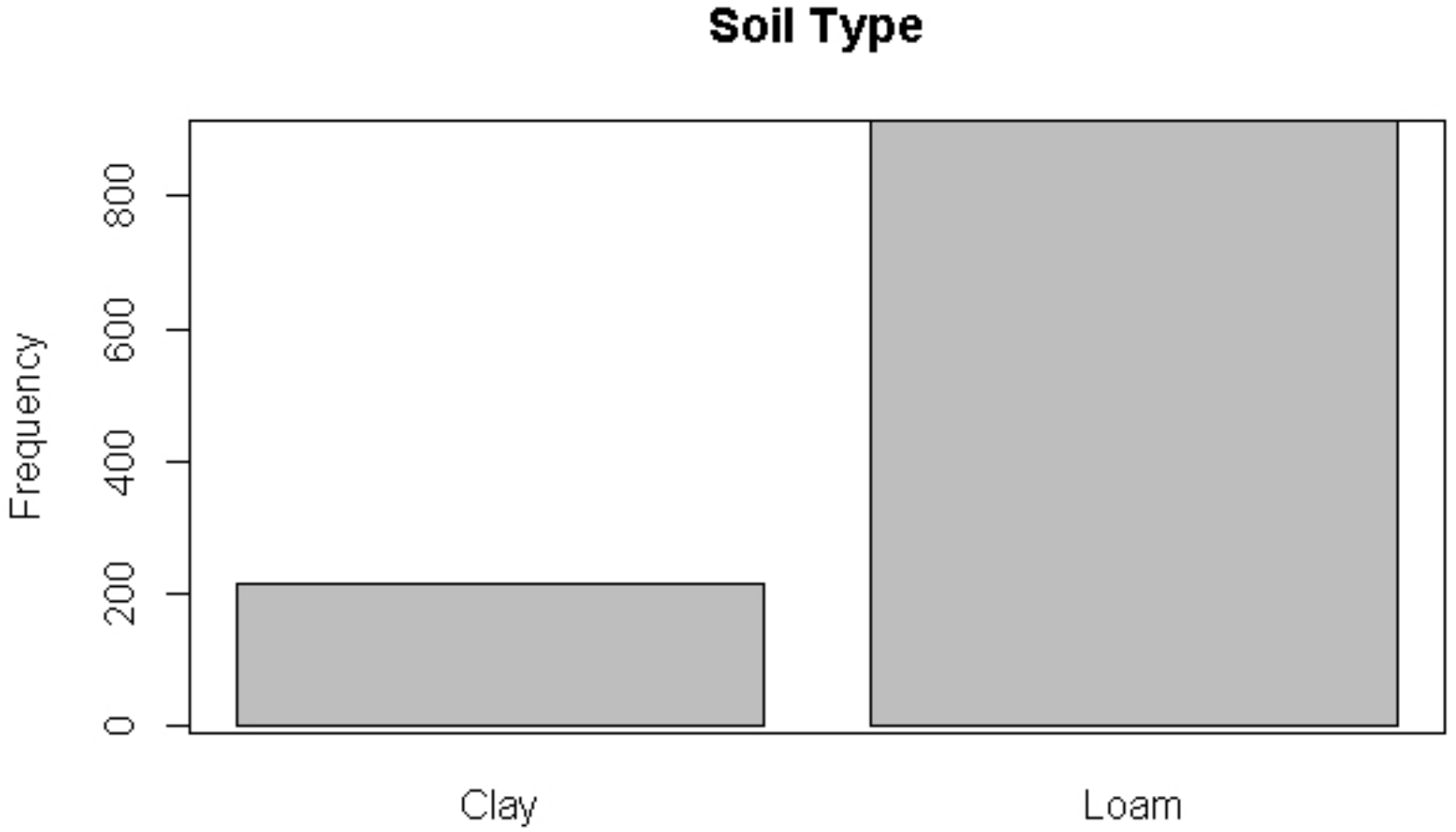
Assignment #2

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.*

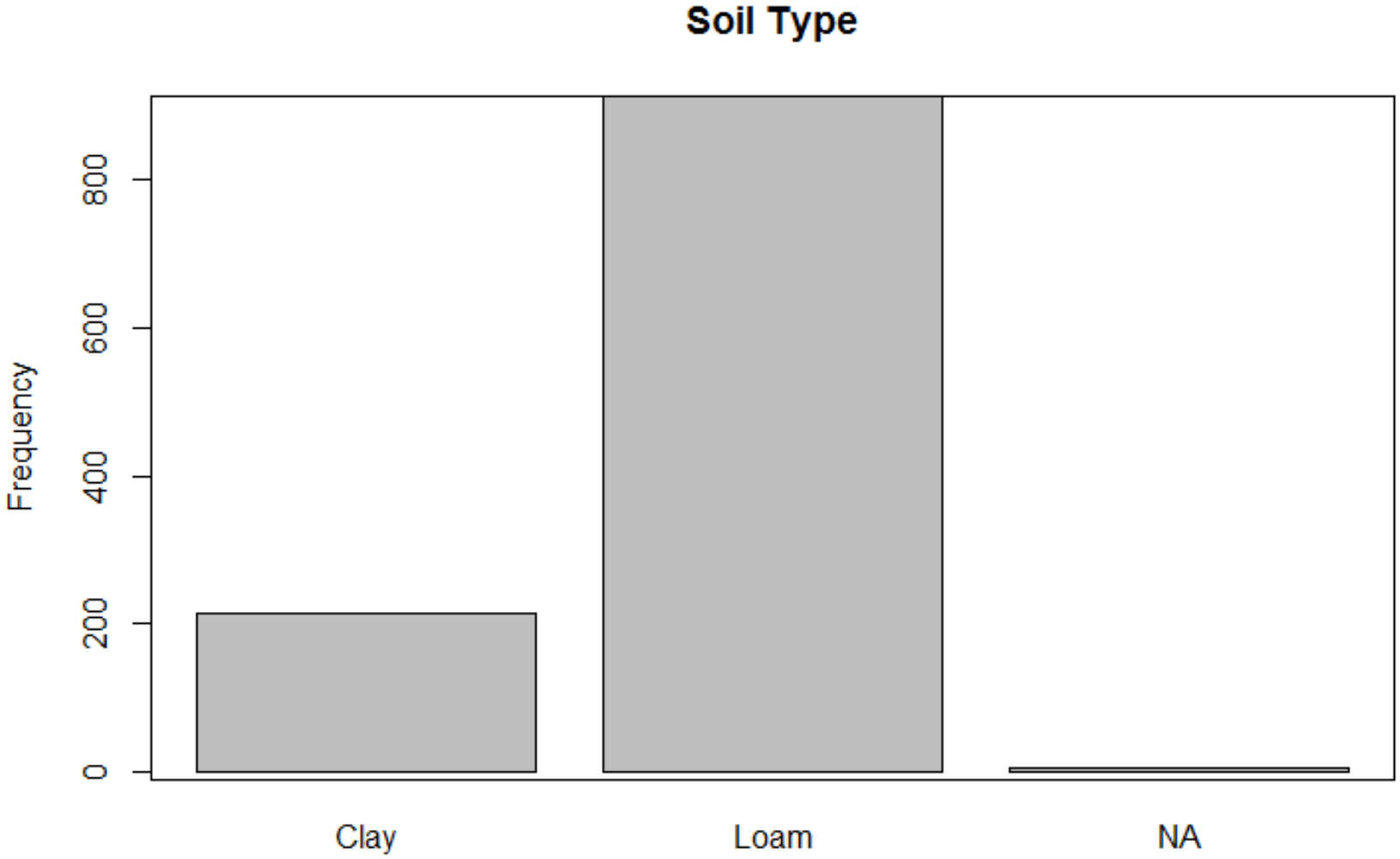
Dichotomous Variables

- What are they?
- Which variables in the radon dataset (as provided) are dichotomous?
- What hypotheses do we have about the association between these variables and radon concentrations?
- What other dichotomous variables would be nice to have in the dataset?



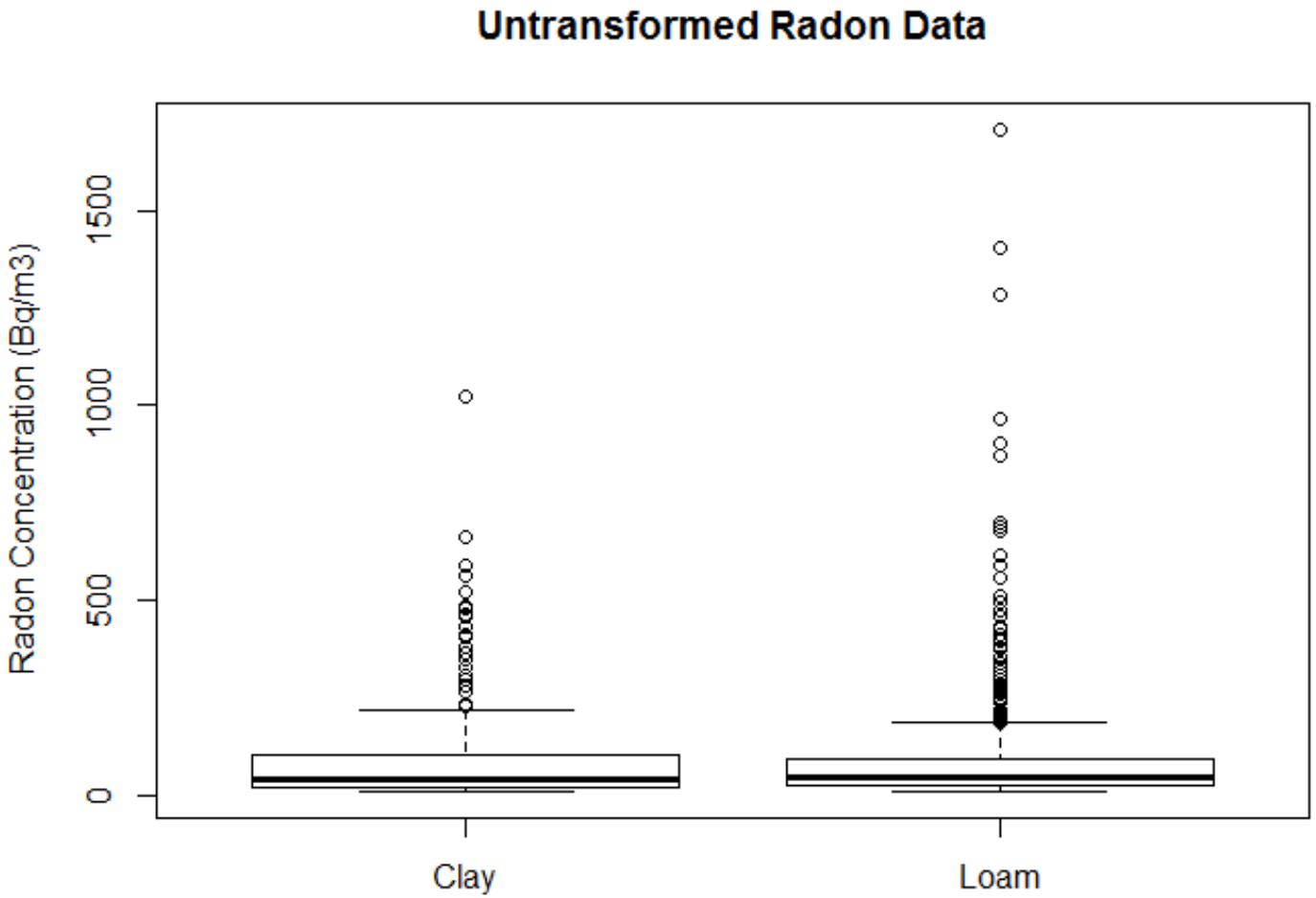
Missing Data

- What does this mean for us?!?
- How do we deal with them?
- ASIDE: What will Sarah do to you if you treat the word “data” as singular in your assignments?



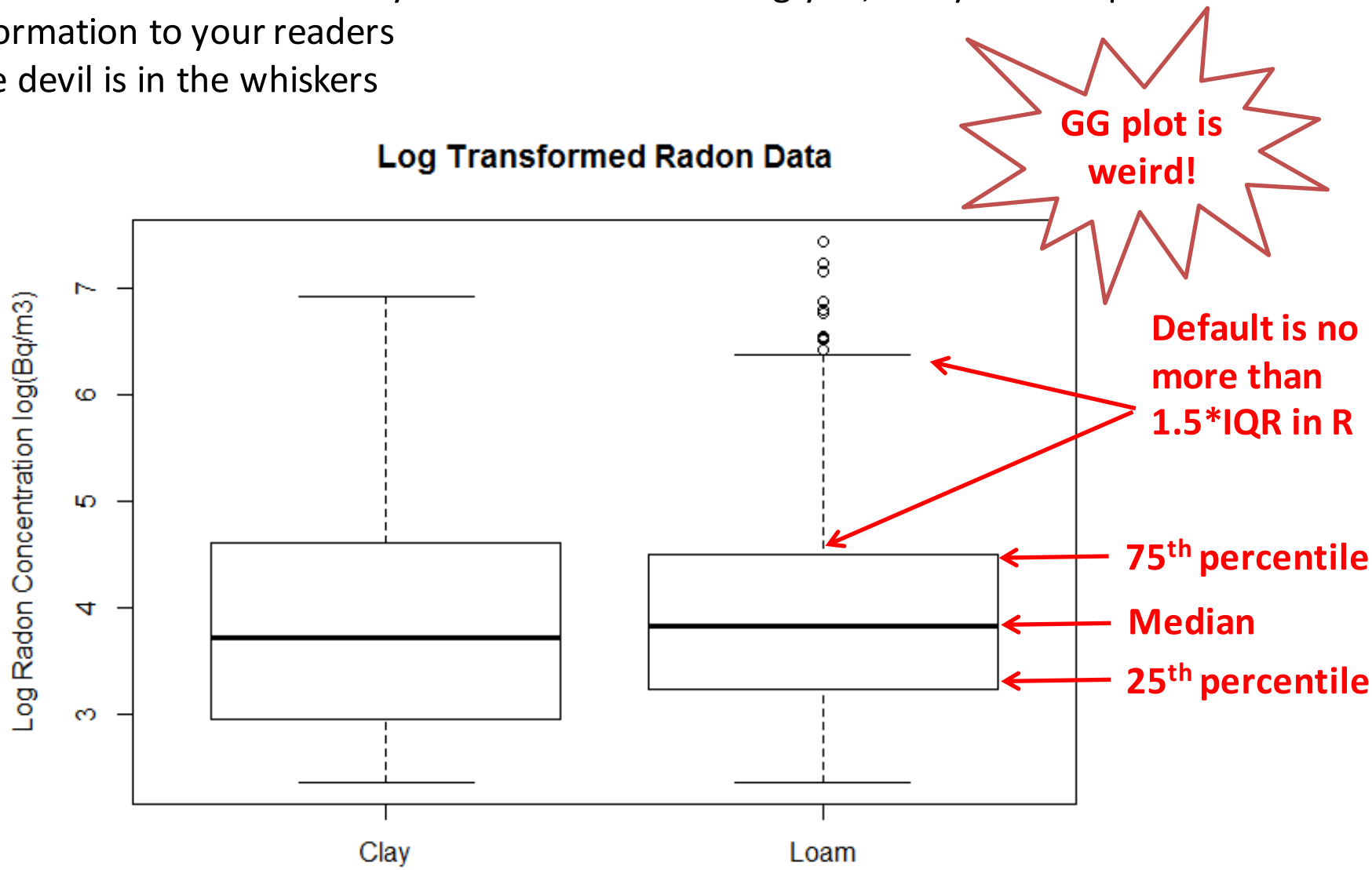
Box Plots

- Box plots are a nice way to visualize the relationship between a dichotomous variable and a continuous variable
- They also help us understand the distribution of a variable

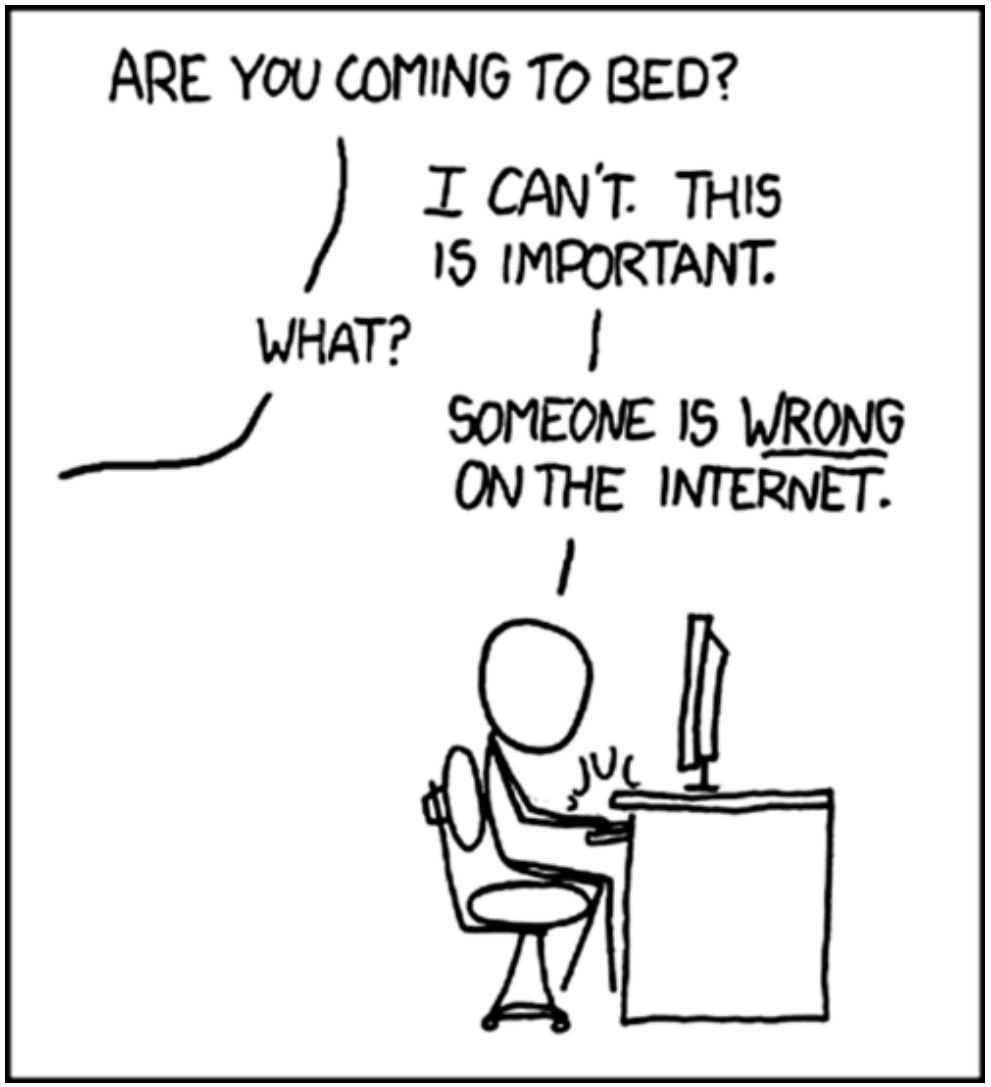


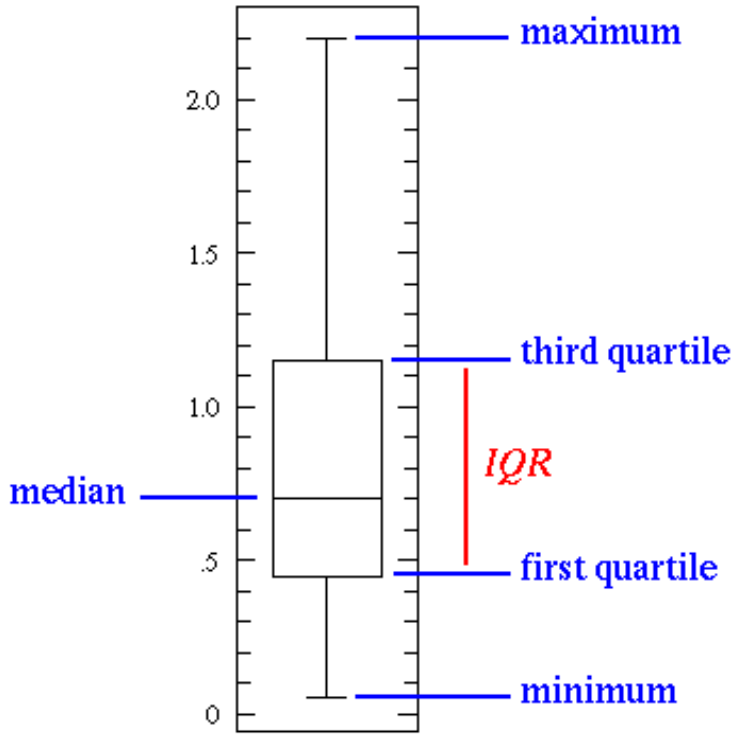
Box Plots

- Box plots are one of the LEAST STANDARDIZED visualization tools available in different statistical software platforms
- You must understand what your software is showing you, and you must provide that information to your readers
- The devil is in the whiskers



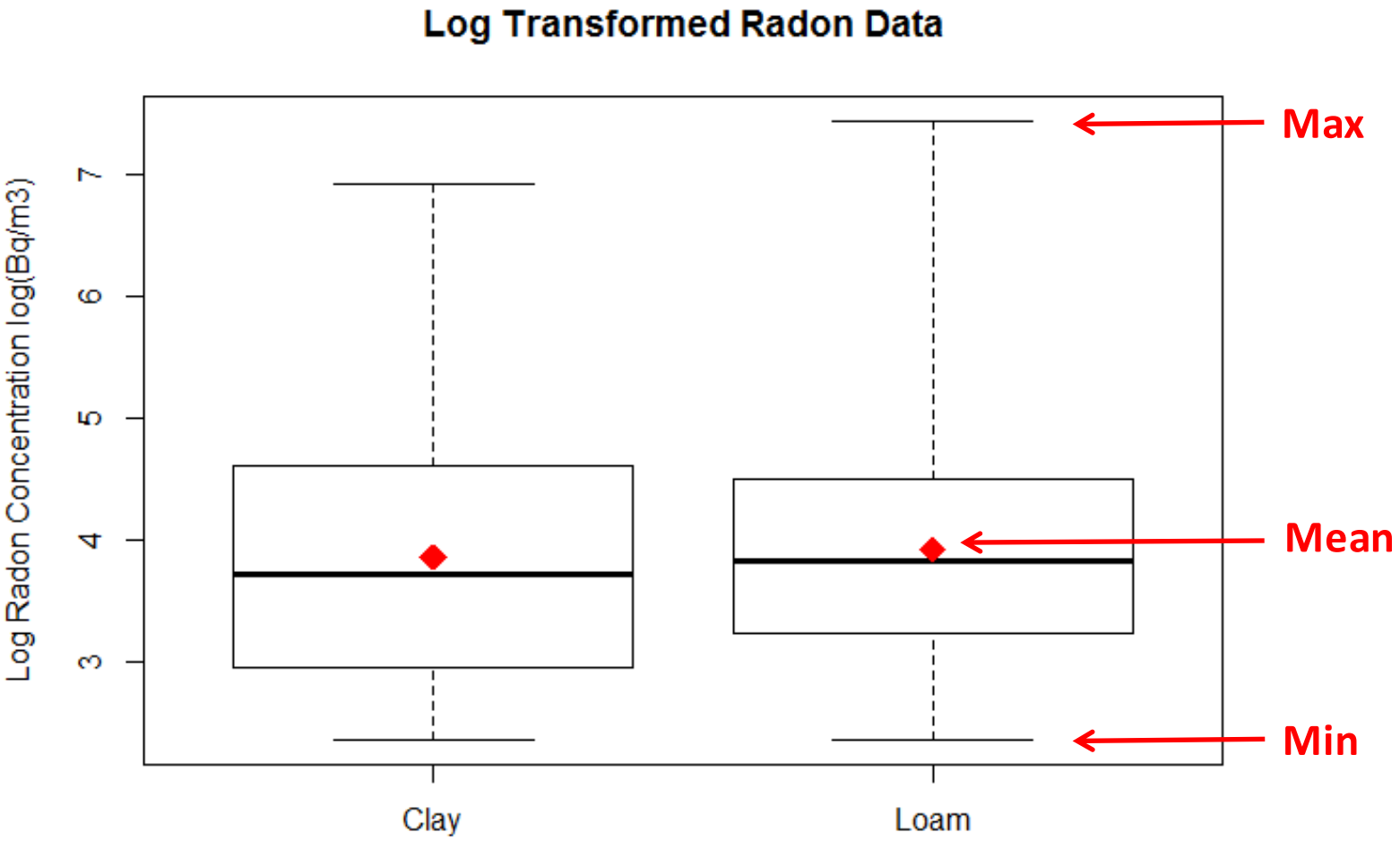
Someone is Wrong on the Internet





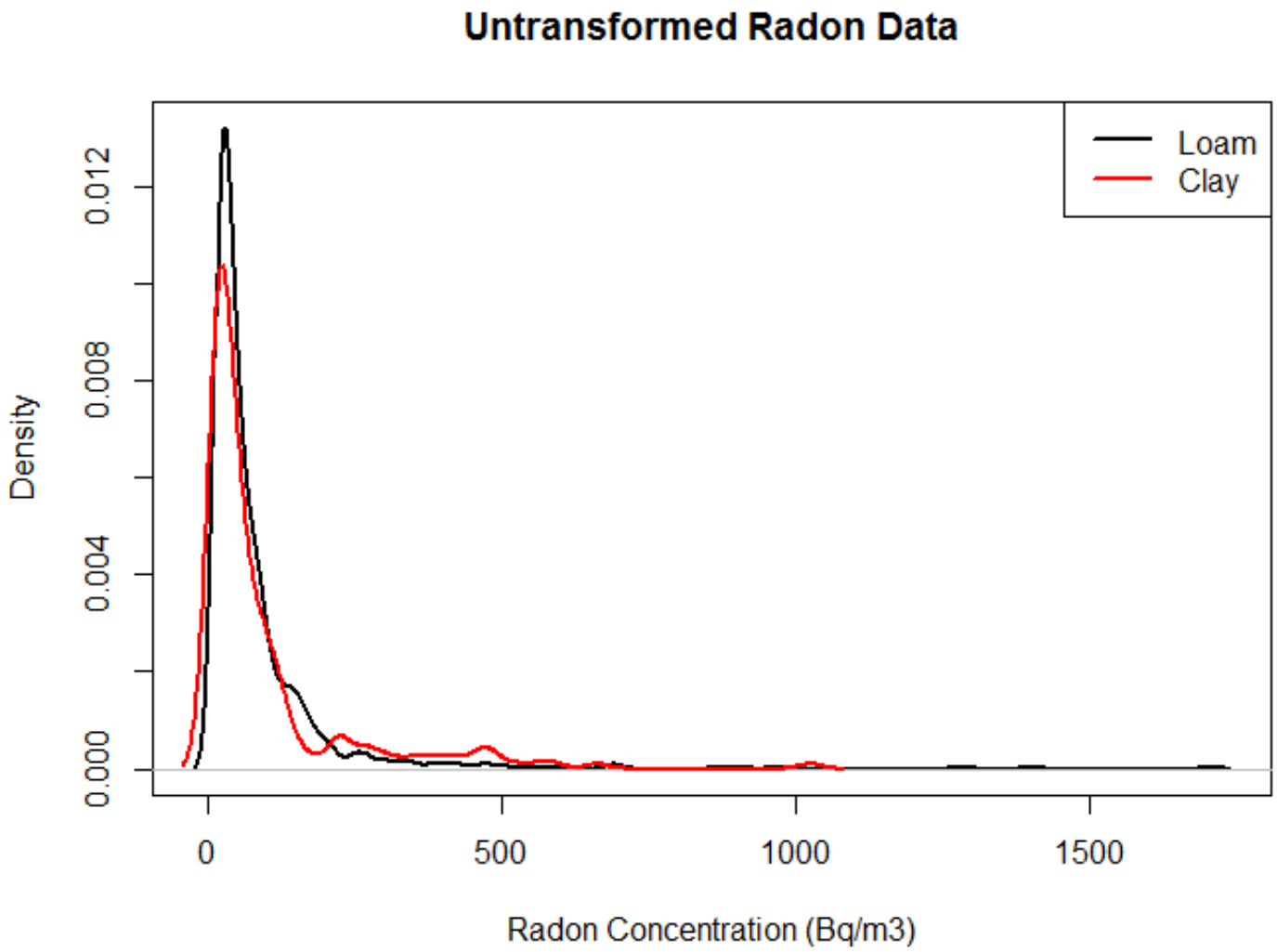
Box Plots

- Most statistical software gives you control over the parameters of your boxplots
- Deducer allows you to do some pretty crazy stuff, so make sure you understand what you're doing!



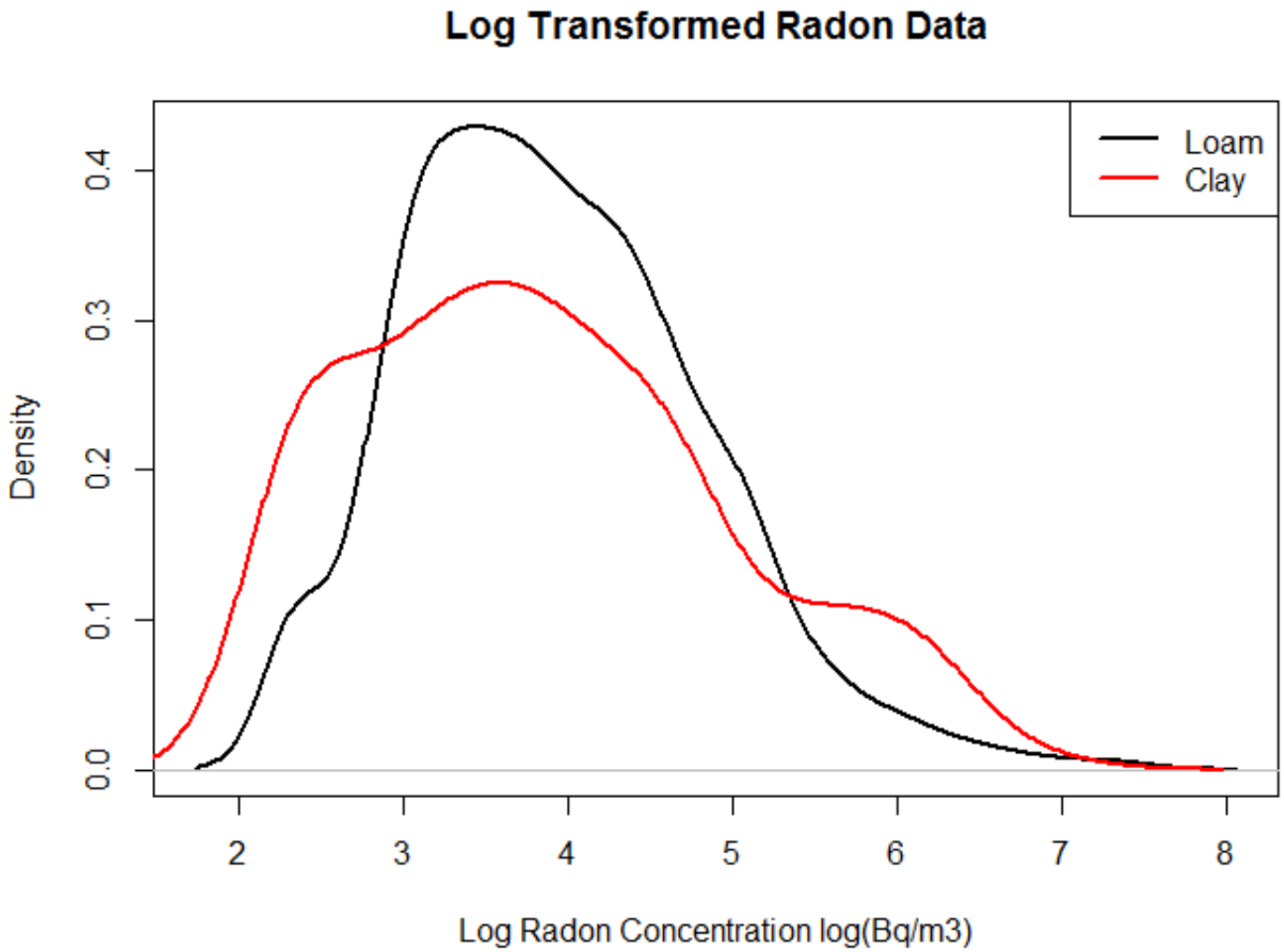
Density Plots

- Useful, but doesn't allow for the same visual comparison on medians and IQRS



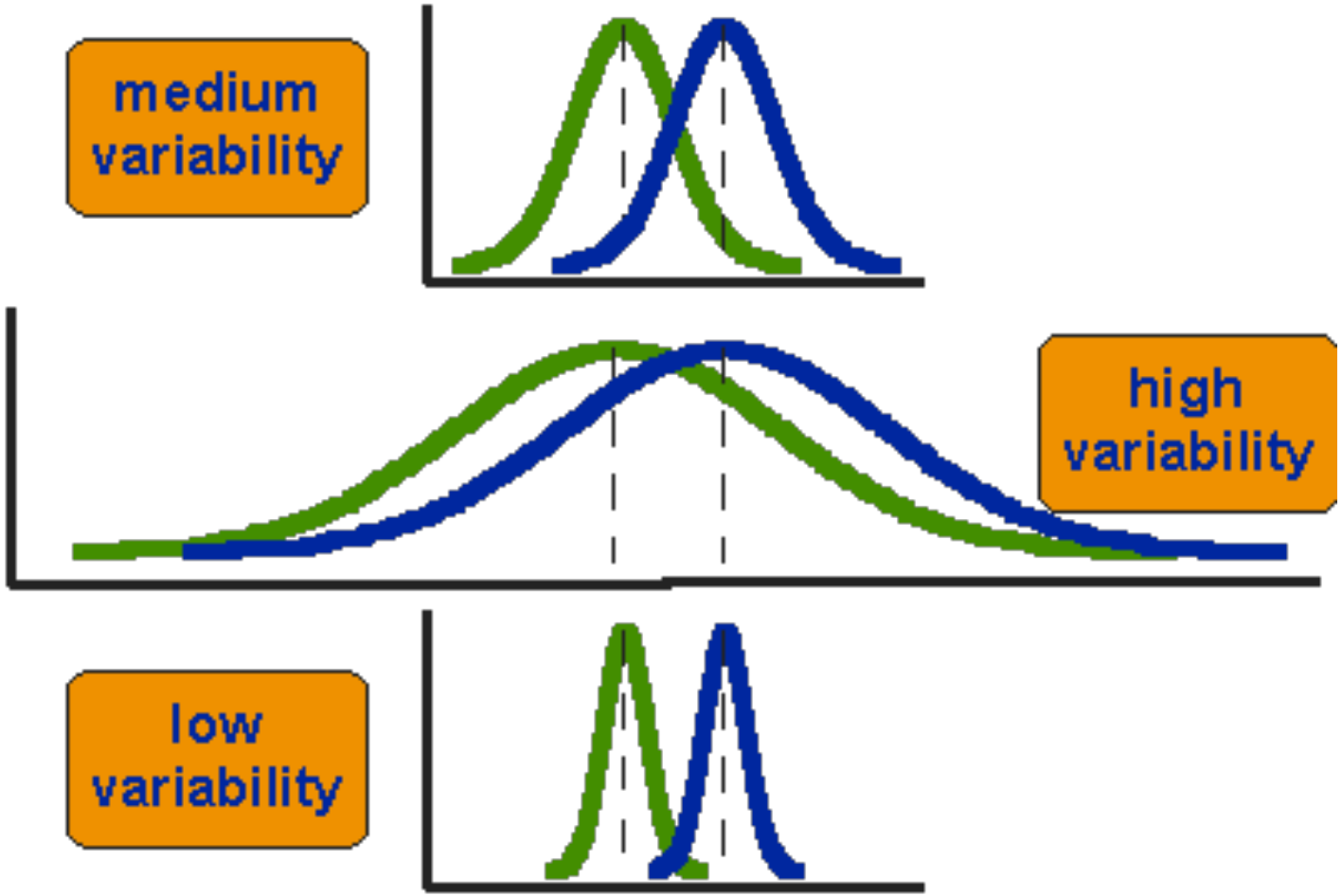
Density Plots

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



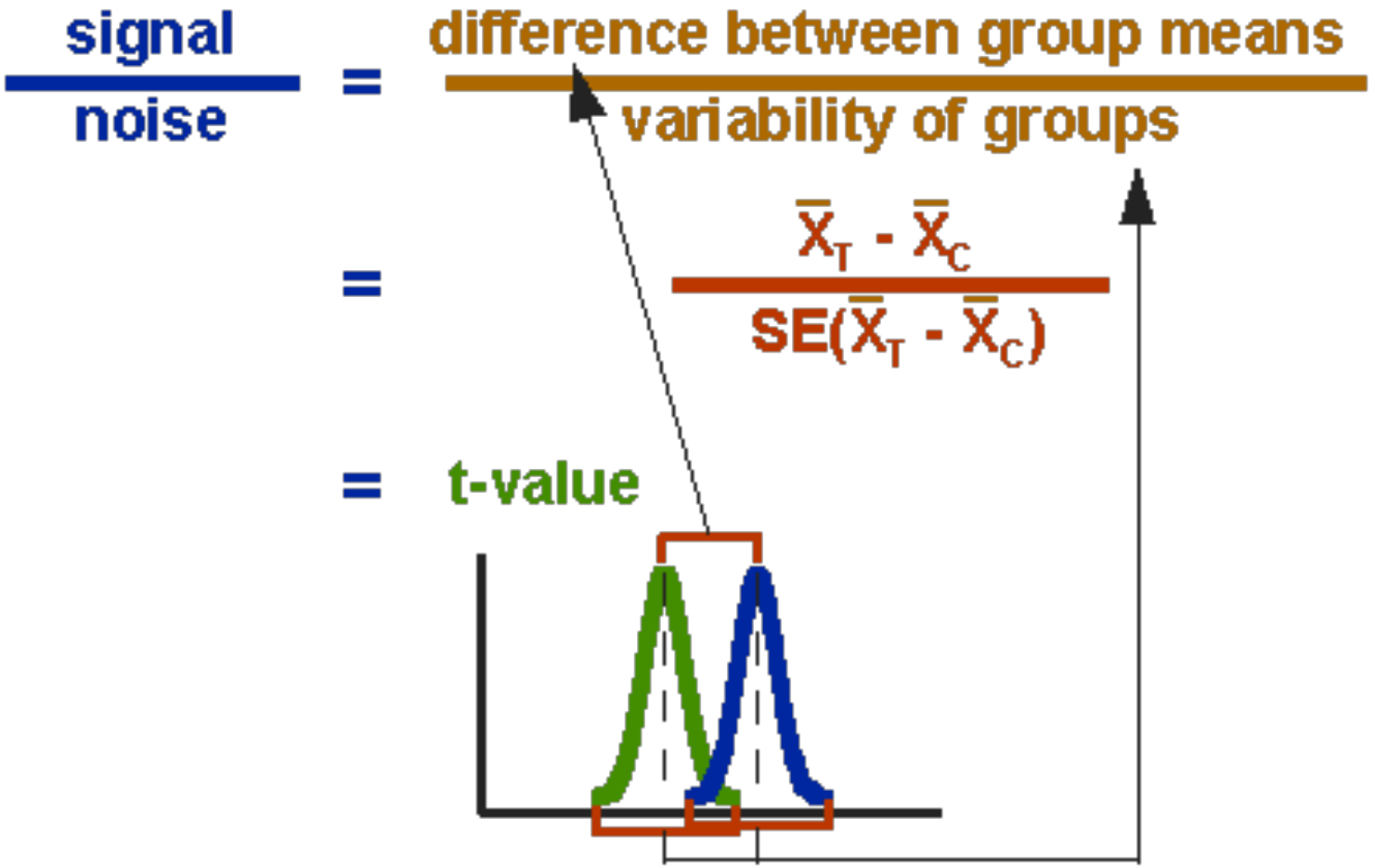
t-test of Two Means

- Means are the same in all cases
- Which do you think are significantly different?
- What are the features we need to assess?



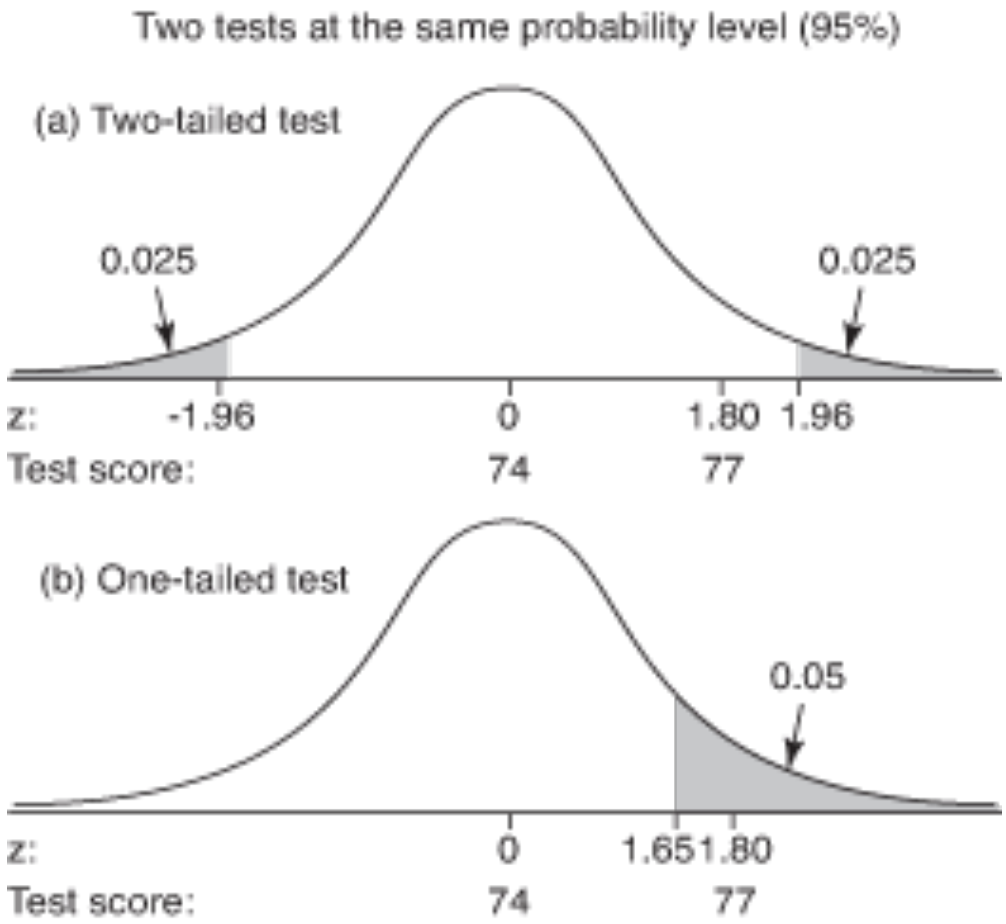
t-test of Two Means

- It comes down to the ratio of the difference between the groups and the variability within the groups
- This is otherwise known as the signal:noise ratio, which is fundamental to the field of statistics
- Why are larger sample sizes better from this perspective?



t-test of Two Means

- Test for a difference in the mean between two INDEPENDENT samples
- H_0: the difference between the means is 0
- H_1: the difference between the means is not 0
- Can be one-tailed (testing for a difference in one direction only) or two-tailed (testing for a difference in either direction).
- One tail more powerful, but only to be used if your hypothesis is in one direction – EXAMPLES?



The t Statistic

- The t statistic you calculate depends on a few different things
  - Are your samples paired or independent?
  - Are the sample sizes equal?
  - Do they have approximately equal variance?
- For unpaired samples with different sizes and approximately equal variance, the t statistic is calculated as follows
- Where s^2 is the sample VARIANCE, or the standard deviation squared
- Are the variances approximately equal in our clay and loam data?

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_{X_1X_2} = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}.$$

Welch's t Statistic

- For unpaired samples with different sizes and UNEQUAL variance, the t statistic is calculated as follows
- Where s^2 is the sample VARIANCE, or the standard deviation squared
- Which t statistic would we use for our data?
- Should we test the UNTRANSFORMED or LOG-TRANSFORMED data?
- What is the value of our t-statistic?

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

where

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The Student t Distribution

- The t statistic follows a t distribution
- This distribution is the standard normal distribution when the size of the sample approximates the size of the entire population (i.e. v = df = +∞)
- When the size of the sample is small the distribution is much wider
- We can use a t-table to look up the probability of observing our t statistic

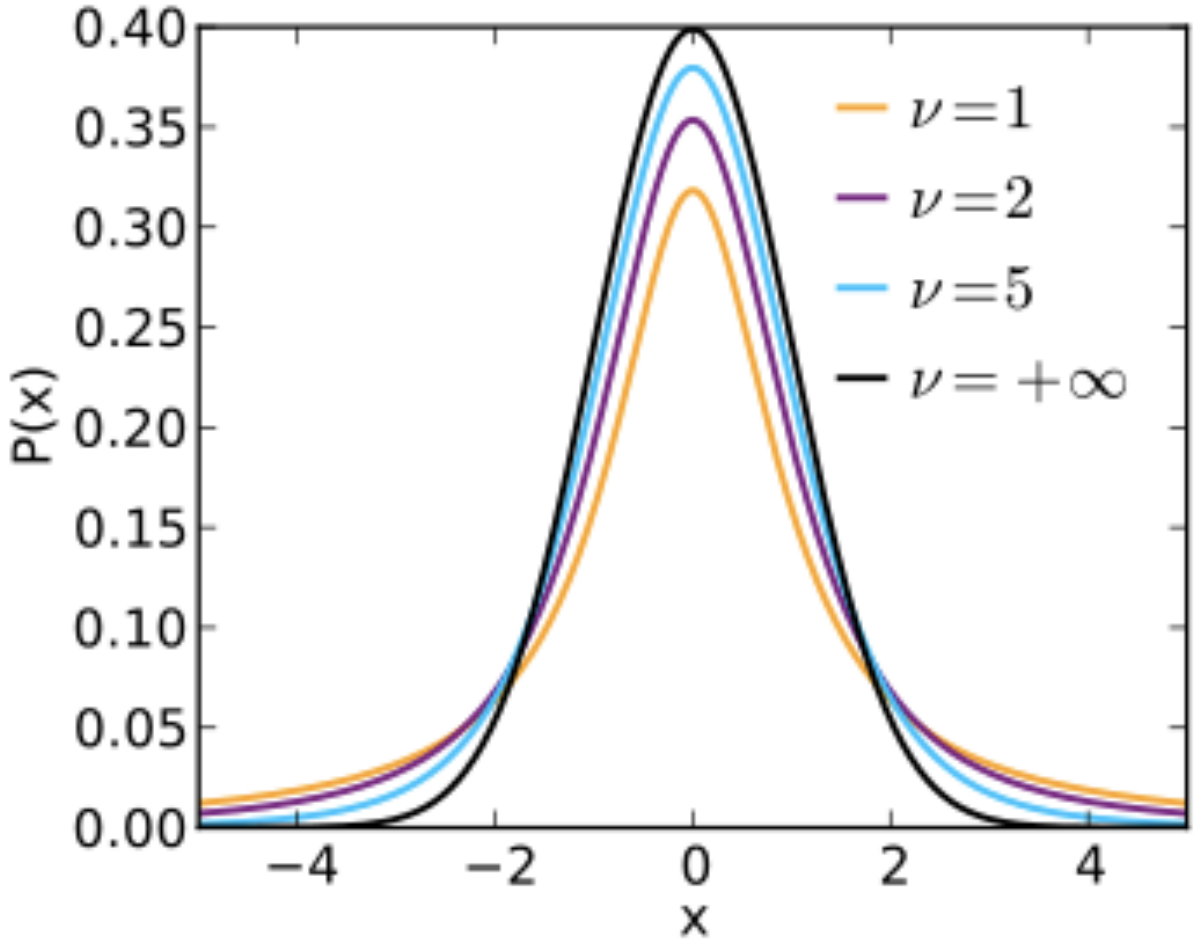


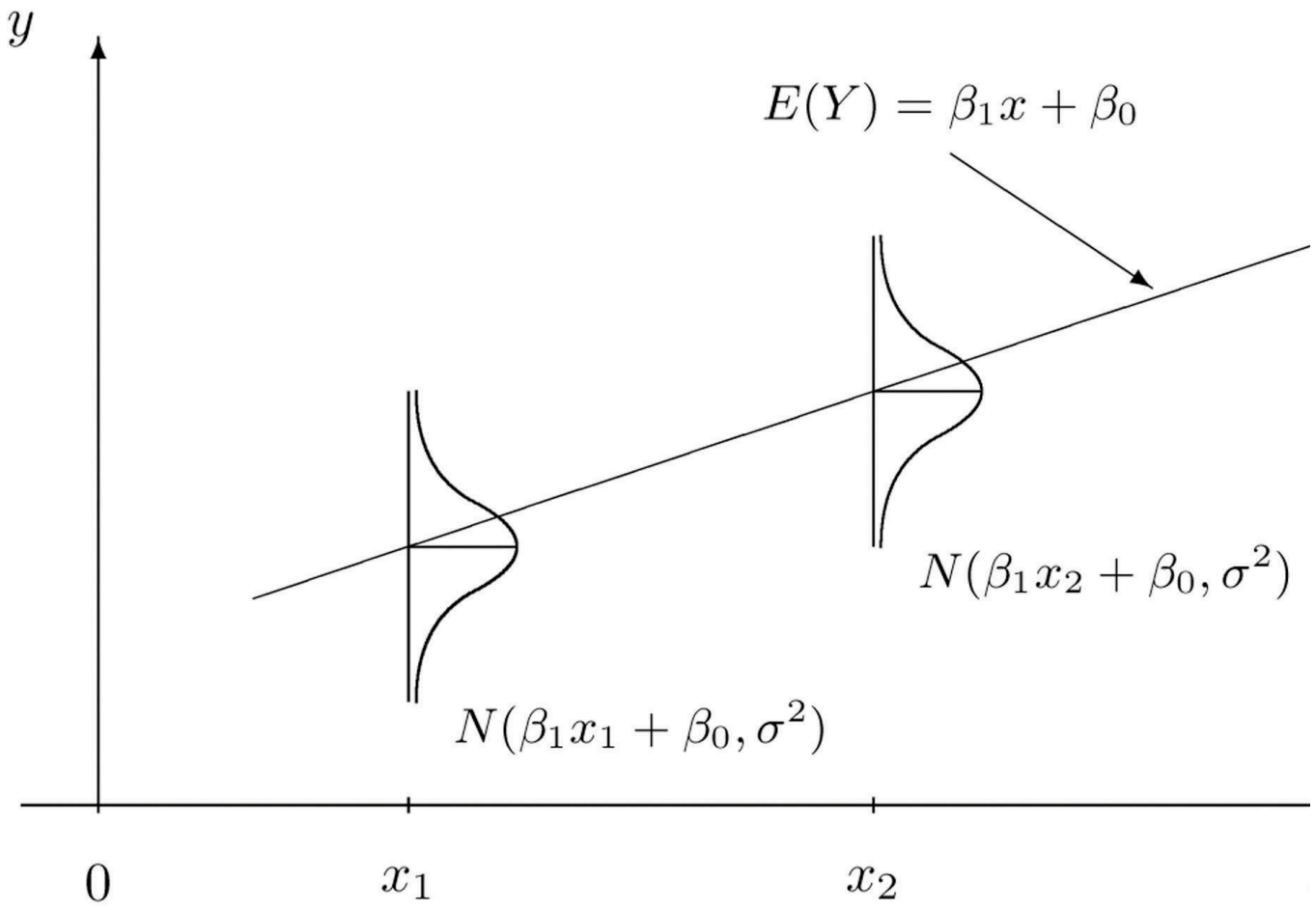
Table T Critical Values of the *t* Distribution

df	One-Tail = .4 Two-Tail = .8	.25 .5	.1 .2	.05 .1	.025 .05	.01 .02	.005 .01	.0025 .005	.001 .002	.0005 .001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Source: From *Biometrika Tables for Statisticians*, Vol. 1, Third Edition, edited by E. S. Pearson and H. O. Hartley, 1966, p. 146. Reprinted by permission of the Biometrika Trustees.

Simple Linear Regression

- Regression is a method we use to evaluate how MEAN VALUES of a DEPENDENT variable (i.e. the Y VALUES in your regression equation) are associated with values of one or more INDEPENDENT variables (i.e. the X VALUES in your regression equation)
- Linear regression assumes that the DEPENDENT variable is normally distributed
- In its simplest form your regression equation is $Y = \beta_0 + \beta_1 X$



$$Y = \beta_0 + \beta_1 X$$

- β_0 is a COEFFICENT called the INTERCEPT, and it indicates the mean of the DEPENDENT variable when the value of X is zero
- β_1 is the COEFFICIENT for X, and it indicates the EFFECT of X on Y
- When X is a dichotomous variable, one of the categories will be assigned a value of 0 and the other will be assigned a value of 1
- This conversion to 0 and 1 is called a DUMMY variable that stands in for the actual values
- The assignment will be ARBITRARY unless you specifically tell your statistical software which category should be 0
- The category that is 0 is called the REFERENCE category
- Thus, when X is a dichotomous variable, β_0 indicates the mean of the DEPEDENT variable for the REFERENCE category

MainRadon = $\beta_0 + \beta_1$*soil

- This is a report from R, which will be similar in Deducer. You will find similar information in the report from any statistical software program
- Let's write out this equation on the board

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       94.430      8.809  10.720   <2e-16 ***
radon$SoilLoam   -12.643      9.791  -1.291    0.197
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 129.2 on 1126 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.001478,  Adjusted R-squared:  0.0005917
F-statistic: 1.667 on 1 and 1126 DF,  p-value: 0.1969
```

MainRadon = $\beta_0 + \beta_1$*soil

- What does the significance of the intercept mean?

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       94.430      8.809  10.720   <2e-16 ***
radon$soilLoam   -12.643      9.791  -1.291    0.197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129.2 on 1126 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.001478,  Adjusted R-squared:  0.0005917
F-statistic: 1.667 on 1 and 1126 DF,  p-value: 0.1969
```

MainRadon = $\beta_0 + \beta_1$*soil

- Why was CLAY chosen as the reference category by the statistical software (the category with the DUMMY value of 0) and not LOAM?

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       94.430      8.809  10.720   <2e-16 ***
radon$SoilLoam   -12.643      9.791  -1.291    0.197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129.2 on 1126 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.001478,  Adjusted R-squared:  0.0005917
F-statistic: 1.667 on 1 and 1126 DF,  p-value: 0.1969
```

Loam soil is associated with a 12.6 Bq/m^3 DECREASE in radon concentrations COMPARED WITH clay soil ON AVERAGE.

Confidence Intervals

- What is the 95% confidence interval around the estimate for the effect of SOIL on RADON?

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      94.430      8.809  10.720   <2e-16 ***
radon$SoilLoam  -12.643      9.791  -1.291    0.197
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 129.2 on 1126 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.001478,  Adjusted R-squared:  0.0005917
F-statistic: 1.667 on 1 and 1126 DF,  p-value: 0.1969
```

LCI = β_1 - 1.96 * Standard Error of β_1 = -12.6 – 1.96*9.8 = -31.9

UCI = β_1 + 1.96 * Standard Error of β_1 = -12.6 + 1.96*9.8 = 6.6

Therefore the estimate for the effect of SOIL on RADON is a -12.6 [-31.9, 6.6] Bq/m^3 change for loam compared with clay.

R^2 = Coefficient of Determination

- A measure of how well your model fits your data
- Indicates the percent of the variability in your DEPENDENT variable that is described the variability in your INDEPENDENT variable(s)

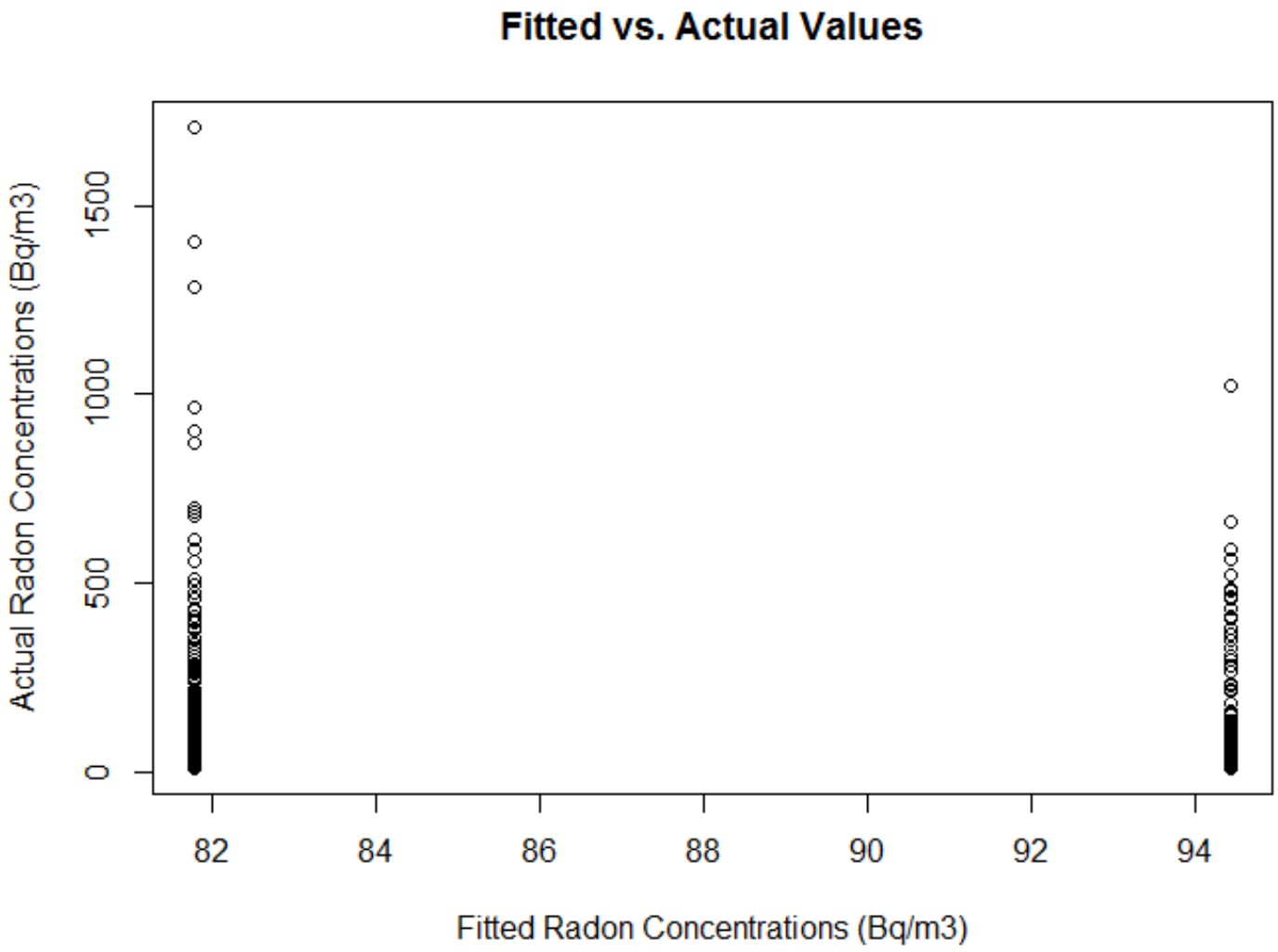
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       94.430      8.809  10.720   <2e-16 ***
radon$SoilLoam   -12.643      9.791  -1.291    0.197
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 129.2 on 1126 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared: 0.001478,  Adjusted R-squared:  0.0005917
F-statistic: 1.667 on 1 and 1126 DF,  p-value: 0.1969
```

A whopping 0.15%!!!

Fitted Values

- What radon concentration does the model estimate for a home built on clay soil?
- What radon concentration does the model estimate for a home build on loam soil?



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

- We can do it with the UNTRANSFORMED data, but this violates the normality of assumption of linear regression modelling
- We will talk more about the underlying assumptions in the coming weeks
- Better to use the LOG-TRANSFORMED data, but the math gets trickier!
- What does the INTERCEPT mean now?

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.85447    0.06440  59.853   <2e-16 ***
radon$SoilLoam  0.07002    0.07158   0.978    0.328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9443 on 1126 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.0008492, Adjusted R-squared:  -3.815e-05
F-statistic: 0.957 on 1 and 1126 DF,  p-value: 0.3282
```

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

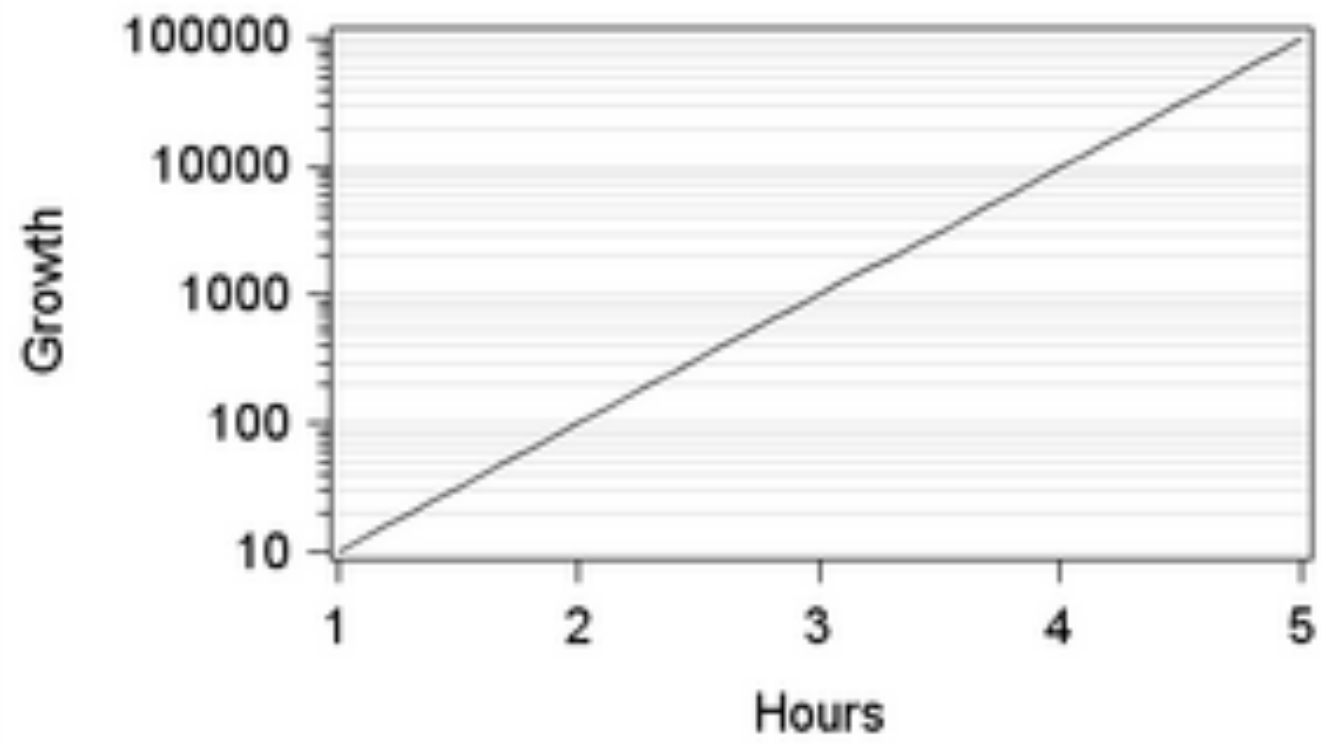
- How do we find the effect of SOIL on RADON concentrations now?

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.85447    0.06440  59.853   <2e-16 ***
radon$SoilLoam   0.07002    0.07158   0.978    0.328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9443 on 1126 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.0008492, Adjusted R-squared:  -3.815e-05
F-statistic: 0.957 on 1 and 1126 DF,  p-value: 0.3282
```

Loam soil is associated with a 0.07 $\log(Bq/m^3)$ INCREASE in the LOG radon concentrations COMPARED WITH clay soil.

Log Scales



- How many bacteria at 1 hour?
- How many at 2 hours?
- Are there twice as many at 2 hours than at 1 hour?
- Log scale = multiplicative framework

What does this really mean?

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.85447    0.06440  59.853   <2e-16 ***
radon$SoilLoam   0.07002    0.07158   0.978    0.328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9443 on 1126 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.0008492, Adjusted R-squared:  -3.815e-05
F-statistic: 0.957 on 1 and 1126 DF,  p-value: 0.3282
```

- Geometric mean of radon for clay soil is exp(3.85447) = 47.20 Bq/m^3
- Geometric mean of radon for loam soil is exp(3.85447+0.07002) = 50.63 Bq/m^3
- 50.63 / 47.20 = 1.07
- exp(0.07002) = 1.07

Loam soil is associated with a 1.07-fold difference in the geometric mean of radon concentrations COMPARED WITH clay soil. Can also be stated as a 7% increase.

Confidence Intervals

- What is the 95% confidence interval around the estimate for the effect of SOIL on RADON?

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.85447    0.06440  59.853   <2e-16 ***
radon$soilLoam   0.07002    0.07158   0.978    0.328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9443 on 1126 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.0008492, Adjusted R-squared:  -3.815e-05
F-statistic: 0.957 on 1 and 1126 DF,  p-value: 0.3282
```

LCI = $\exp(\beta_1 - 1.96 * \text{Standard Error of } \beta_1) =$
$\exp(0.07002 - 1.96*0.07158) = 0.93$

UCI = $\exp(\beta_1 + 1.96 * \text{Standard Error of } \beta_1) =$
$\exp(0.07002 + 1.96*0.07158) = 1.23$

Therefore the estimate for the effect of SOIL on the geometric mean of radon is a 1.07-fold difference [0.93, 1.23] for loam compared with clay. This can also be stated as a 7% increase [-7%, 23%].

Next Week

- Assessing the relationship between a categorical and continuous variable
- Box plots to visualize
- ANOVA to test for differences in the means
- Hypothesis generation
- More on DUMMY variables
- Simple linear regression PART II
- Standard reporting
- Model diagnostics