

Stat 101L: Lecture 15

Re-expressing Data

- *Chapter 6 – Normal Model
 - What if data do not follow a Normal model?
- *Chapters 8 & 9 – Linear Model
 - What if a relationship between two variables is not linear?

1

Re-expressing Data

- *Re-expression is another name for changing the scale of (transforming) the data.
- *Usually we re-express the response variable, Y.

2

Goals of Re-expression

- *Goal 1 – Make the distribution of the re-expressed data more symmetric.
- *Goal 2 – Make the spread of the re-expressed data more similar across groups.

3

Stat 101L: Lecture 15

Goals of Re-expression

- *Goal 3 – Make the form of a scatter plot more linear.
- *Goal 4 – Make the scatter in the scatter plot more even across all values of the explanatory variable.

4

Ladder of Powers

- *Power: 2
- *Re-expression: y^2
- *Comment: Use on left skewed data.

5

Ladder of Powers

- *Power: 1
- *Re-expression: y
- *Comment: No re-expression.
Do not re-express the data if they are already well behaved.

6

Stat 101L: Lecture 15

Ladder of Powers

- *Power: $\frac{1}{2}$
- *Re-expression: \sqrt{y}
- *Comment: Use on count data or when scatter in a scatter plot tends to increase as the explanatory variable increases.

7

Ladder of Powers

- *Power: "0"
- *Re-expression: $\log(y)$
- *Comments: Not really the "0" power. Use on right skewed data. Measurements cannot be negative or zero.

8

Ladder of Powers

- *Power: $-\frac{1}{2}, -1$
- *Re-expression: $\frac{1}{\sqrt{y}}, \frac{1}{y}$
- *Comments: Use on right skewed data. Measurements cannot be negative or zero. Use on ratios.

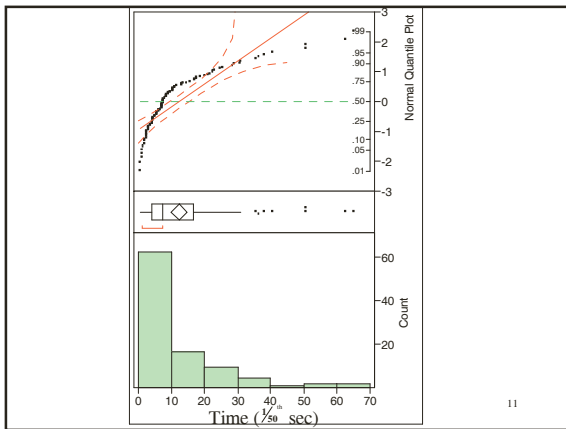
9

Stat 101L: Lecture 15

Goal 1 - Symmetry

- * Data are obtained on the time between nerve pulses along a nerve fiber.
- * Time is rounded to the nearest half unit where a unit is $\frac{1}{50}$ th of a second.
-30.5 represents $30.5/50 = 0.61$ sec

10

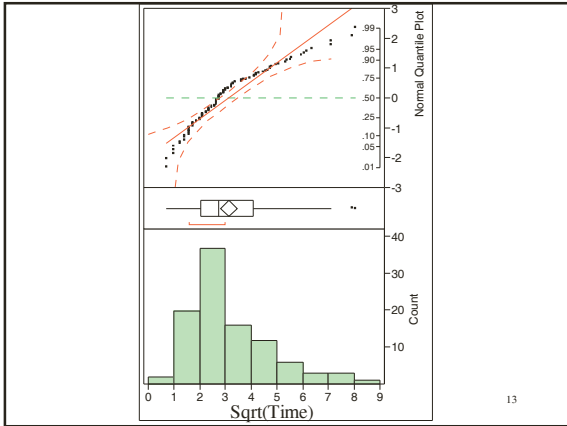


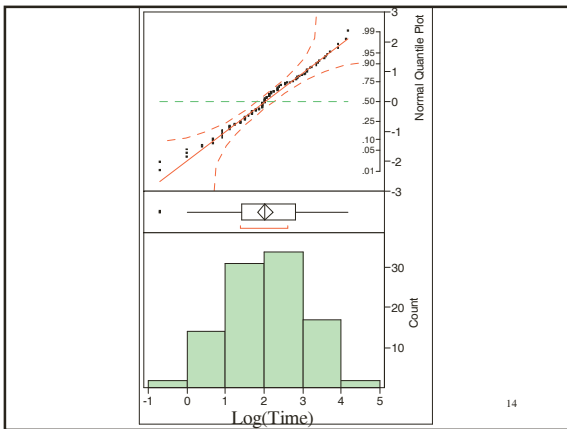
Time – Nerve Pulses

- * Distribution is skewed right.
- * Sample mean (12.305) is much larger than the sample median (7.5).
- * Many potential outliers.
- * Data not from a Normal model.

12

Stat 101L: Lecture 15





Summary

- * Time – Highly skewed to the right.
- * Sqrt(Time) – Still skewed right.
- * Log(Time) – Fairly symmetric and mounded in the middle.
 - Could have come from a Normal model.

15
