1

Aims

- Assumptions of parametric tests based on the normal distribution
- Understand the assumption of normality
  - *Graphical displays*
  - *Skew*
  - *Kurtosis*
  - *Normality tests*
- Understand Homogeneity of Variance
  - *Levene's Test*
- Know how to correct problems in the data
  - *Log, Square Root and Reciprocal Transformations*
  - *Pitfalls and alternatives*
  - *Robust tests*

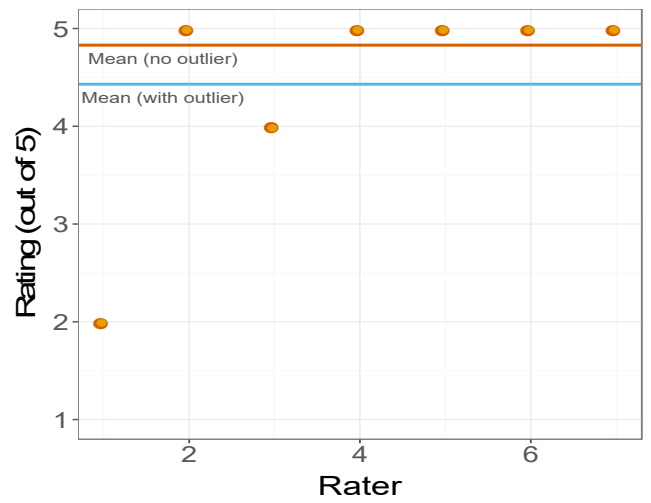
2

Assumptions

- Parametric tests based on the normal distribution assume:
  - *Additivity and linearity*
  - *Normality*
  - *Homogeneity of Variance*
  - *Independence*

3

Outliers can bias a parameter estimate



4

...and the error associated with that estimate

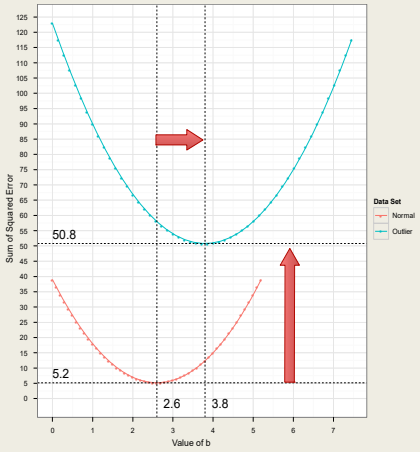


Figure 5.3: The effect of an outlier on a parameter estimate (the mean) and it's associated estimate of error (the sum of squared errors)

5

Additivity and linearity

- The outcome variable is, in reality, linearly related to any predictors.
- If you have several predictors then their combined effect is best described by adding their effects together.
- If this assumption is not met then your model is invalid.

6

Normal Distribution

- The normal distribution is relevant to:
  - *Parameters*
  - *Confidence intervals around a parameter*
  - *Null hypothesis significance testing*
- This assumption tends to get incorrectly translated as 'your data need to be normally distributed'.

7

When does the assumption of normality matter?

- In small samples.
  - *The central limit theorem allows us to forget about this assumption in larger samples.*
- In practical terms, as long as your sample is fairly large, outliers are a much more pressing concern than normality.

8

Homoscedasticity/ Homogeneity of Variance

- When testing several groups of participants, samples should come from populations with the same variance.
- In correlational designs, the variance of the outcome variable should be stable at all levels of the predictor variable.

9

Homoscedasticity/ Homogeneity of Variance

- Can affect the two main things that we might do when we fit models to data:
  - *Parameters*
  - *Null Hypothesis significance testing*

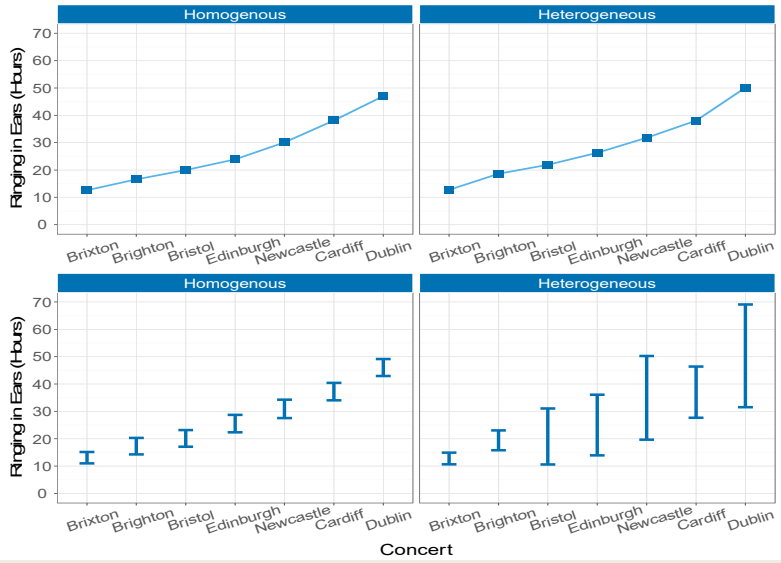
10

Assessing homoscedasticity/ homogeneity of variance

- Graphs (see lectures on regression)
- Levene's Tests
  - *Tests if variances in different groups are the same.*
  - *Field Says: Don't use:*
    - In large samples, small differences will be significant.
    - In small samples, big differences won't be significant
- Solutions
  - *Robust tests (Welch's t, Welch's F)*
  - *Adjusted standard errors*

11

Homogeneity of variance



12

Independence

- The errors in your model should not be related to each other.
- If this assumption is violated:
  - *Confidence intervals and significance tests will be invalid.*
  - *You should apply the techniques covered in Chapter 21.*

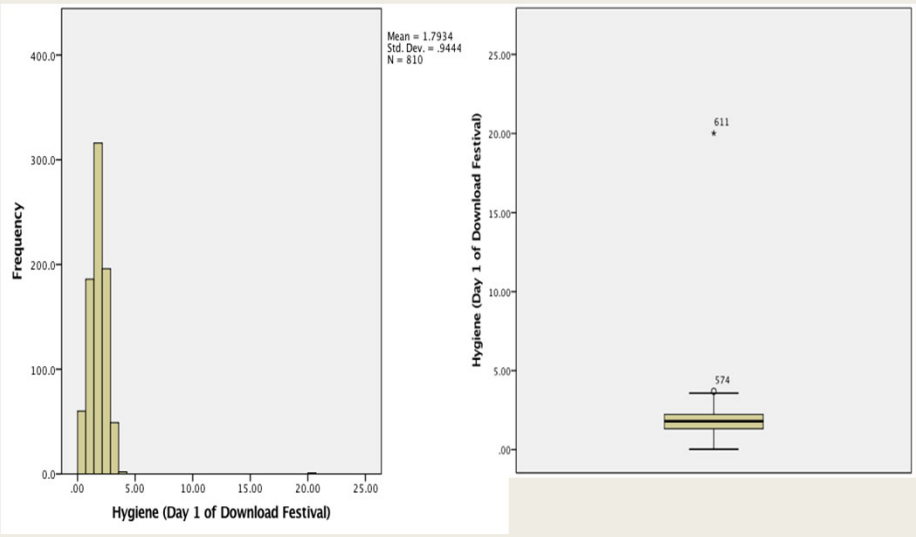
13

Spotting outliers example

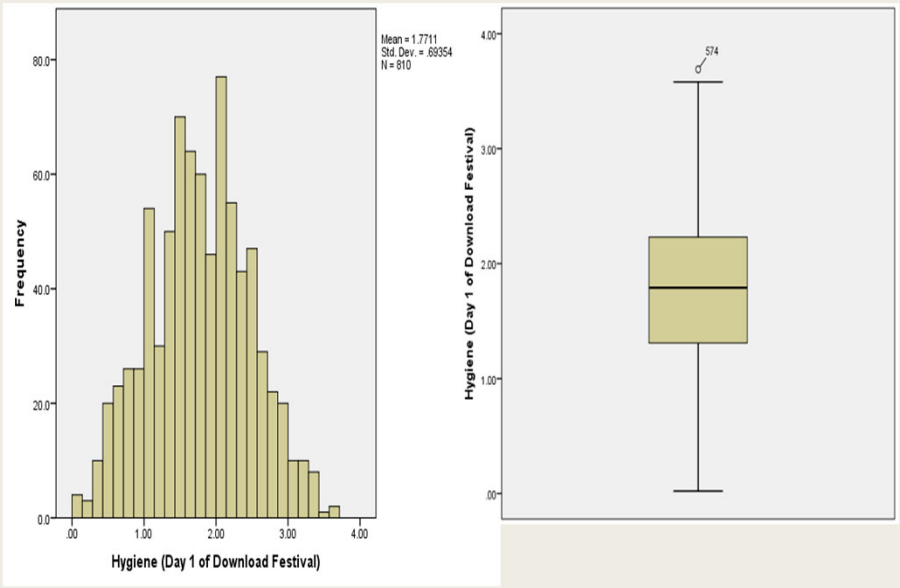
- A biologist was worried about the potential health effects of music festivals.
- Download Music Festival
- Measured the hygiene of 810 concert-goers over the three days of the festival.
- Hygiene was measured using a standardised technique:
  - *Score ranged from 0 to 4*
    - 0 = you smell like a corpse rotting up a skunk's arse
    - 4 = you smell of sweet roses on a fresh spring day

14

Spotting outliers with graphs



15



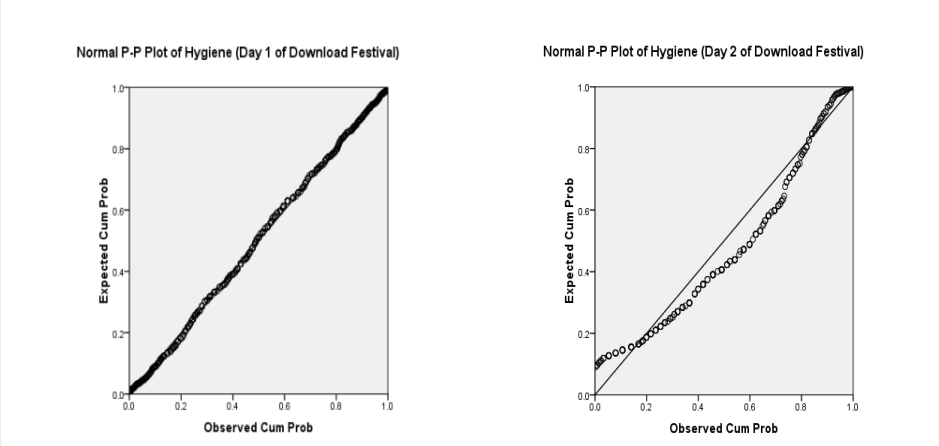
16

Spotting normality

- We don't have access to the sampling distribution so we usually test the observed data
- Central Limit Theorem
  - *If N > 30, the sampling distribution is normal anyway (arguably)*
- Graphical displays
  - *P-P Plot (or Q-Q plot)*
  - *Histogram*
- Values of skew/kurtosis
  - *0 in a normal distribution*
- Kolmogorov-Smirnov Test
  - *Tests if data differ from a normal distribution*

17

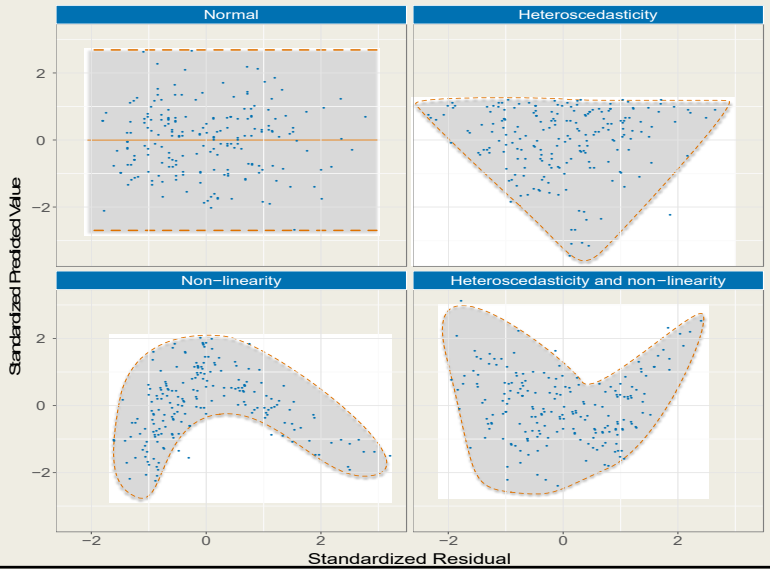
The P-P plot



Normal Not normal

18

Spotting problems with linearity or homoscedasticity



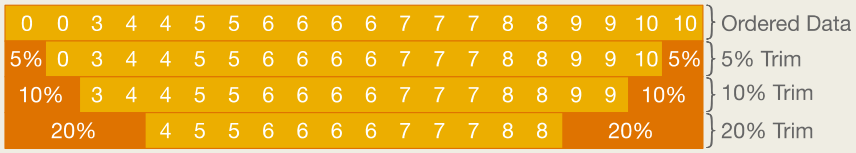
19

Reducing bias

- Analyse with robust methods:
  - *Bootstrapping*
  - *Methods based on medians and trims.*
- Trim the data:
  - *Delete a certain amount of scores from the extremes.*
- Windsorizing/Truncating:
  - *Substitute outliers with the highest value that isn't an outlier*
- Transform the data:
  - *By applying a mathematical function to scores.*

20

Trimming the data



21

Transforming data

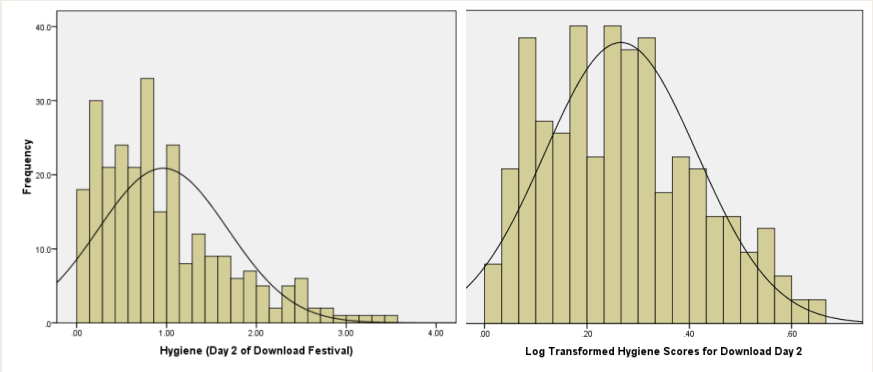
- Log transformation ($\log(x_i)$)
  - *Reduce positive skew.*
- Square root transformation ($\sqrt{x_i}$):
  - *Also reduces positive skew. Can also be useful for stabilizing variance.*
- Reciprocal transformation ($1/ x_i$):
  - *Dividing 1 by each score also reduces the impact of large scores. This transformation reverses the scores, you can avoid this by reversing the scores before the transformation, $1/(x_{highest} - x_i)$.*

22

Log transformation

Before

After

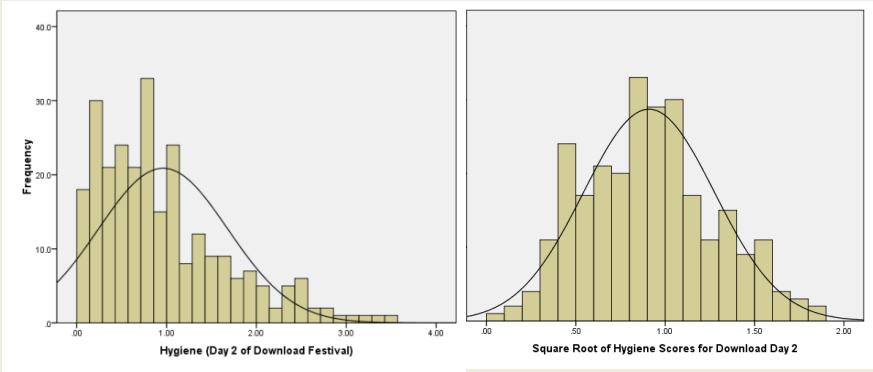


23

Square root Transformation

Before

After

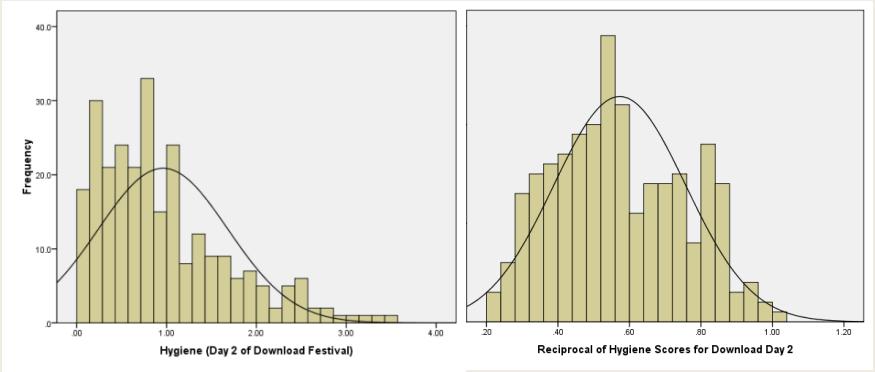


24

Reciprocal transformation

Before

After



25