DISCOVERING STATISTICS USING R

Exploring assumptions

1

DISCOVERING STATISTICS USING R

Aims and Objectives

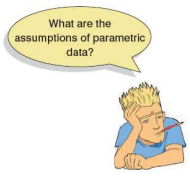
- Assumptions of parametric tests based on the normal distribution
- Understand the assumption of normality
  - Graphical displays
  - Skew/Kurtosis
  - Shapiro-Wilk test
- Understand homogeneity of variance
  - Levene's test
- Know how to correct problems in the data
  - Log, square root and reciprocal transformations
  - Pitfalls and alternatives
  - Nonparametric and robust tests

2

DISCOVERING STATISTICS USING R

Assumptions

- Parametric tests based on the normal distribution assume:
  - Response variable to be continuous
  - Normally distributed data
    - Sampling distribution
    - Residuals
  - Homogeneity of variance
  - Independent scores
    - Between subjects
    - Model error



3

Assessing normality

- We don't have access to the sampling distribution so we test the observed data instead
  - Central limit theorem:
    - if $N > 30$, the sampling distribution is normal anyway
- Graphical displays
  - Histogram
  - Q-Q plot
- Values of skew/kurtosis
  - 0 in a normal distribution
  - Convert to *z*-scores (by dividing value by SE)
- Shapiro-Wilk test
  - Tests if data differ from a normal distribution
  - Significant= non-normal data

4

Example

- A biologist was interested in the potential health effects of music festivals
- Measured the hygiene of 810 concert-goers over the three days of a festival
- Hygiene was measured using a standardized 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

5

Histogram and Q-Q Plot

- Draw plots of hygiene scores for day 1

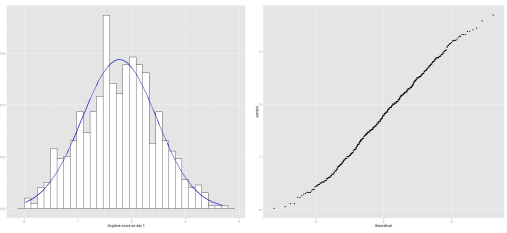
```
> hist.day1<- ggplot(dlf, aes(day1)) +
              geom_histogram(aes(y= ..density..),
                   col= "black", fill= "white") +
              stat_function(fun= dnorm, args=
                   list(mean(dlf$day1, na.rm= T),
                        sd(dlf$day1, na.rm = T)),
                   col= "blue", size= 1) +
              labs(x= "Hygiene score on day 1",
                   y= "Density")
> qqplot.day1<- qplot(sample= dlf$day1)

> hist.day1; qqplot.day1
```

6

DISCOVERING STATISTICS USING R

Histogram and Q-Q Plot



7

DISCOVERING STATISTICS USING R

Another example

- Performance on statistics exam
- Participants
  - *N*= 100 students
- Measures
  - Exam: first-year exam scores as a percentage
  - Computer: measure of computer literacy %
  - Lecture: percentage of lectures attended
  - Numeracy: a measure of numerical ability out of 15
  - Uni: whether the student attended Sussex University or Duncetown University

8

DISCOVERING STATISTICS USING R

Assessing skew and kurtosis

- Using '*by()*' and '*describe()*' from the 'psych' package

```
> by(rexam$exam, rexam$uni, describe)
```

- Using '*by()*' and '*stat.desc()*' from the 'pastecs' package

```
> by(rexam$exam, rexam$uni, stat.desc)
```

9

DISCOVERING STATISTICS USING R

Assessing skew and kurtosis

- If we want descriptive statistics for multiple variables, we can use '*cbind()*'

```
> by(cbind(rexam$exam, rexam$numeracy),
            rexam$uni, describe)
```

- We can also use '*describe()*' and '*stat.desc()*' with more than one variable at the same time using '*cbind()*'

```
> describe(cbind(dlf$day1, dlf$day2, dlf$day3))
> stat.desc(cbind(dlf$day1, dlf$day2, dlf$day3),
            basic= F, norm= T)
```

10

DISCOVERING STATISTICS USING R

Assessing skew and kurtosis

	day1	day2	day3
median	1.790	0.790	0.760
mean	1.793	0.961	0.977
SE.mean	0.033	0.044	0.064
CI.mean.0.95	0.065	0.087	0.127
var	0.892	0.520	0.504
std.dev	0.944	0.721	0.710
coef.var	0.527	0.750	0.727
skewness	8.833	1.083	1.008
skew.2SE	51.407	3.612	2.309
kurtosis	168.967	0.755	0.595
kurt.2SE	492.314	1.265	0.686
normtest.W	0.654	0.908	0.908
normtest.p	0.000	0.000	0.000

11

DISCOVERING STATISTICS USING R

Assessing normality with a statistical test

- Shapiro-Wilk test:

```
> shapiro.test(rexam$exam)

> shapiro.test(rexam$numeracy)
```

- Shapiro-Wilk test split by university, *e.g.*

```
> by(rexam$exam, rexam$uni, shapiro.test)
```

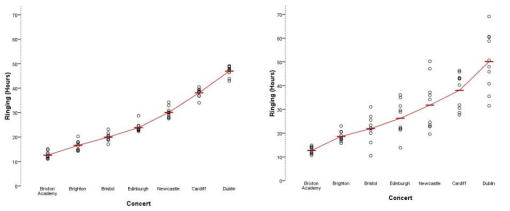
12

Assessing homogeneity of variance

- Graphs (Chapter 7)
- Levene's test
  - Tests if variances in different groups are the same
  - Significant= variances are not equal
- Variance ratio (or Hartley's F_{max})
  - With 2 or more groups
  - VR= largest variance/smallest variance
  - If VR< critical value in Figure 5.8: homogeneity can be assumed

13

Assessing homogeneity of variance



Homogeneous Not homogeneous

14

Assessing homogeneity of variance

- Use the '*leveneTest()*' function from the '*car*' package, *e.g.*:

```
> leveneTest(rexam$exam, rexam$uni)
```

- When sample size is large, small differences in group variances can already lead to significance
  - Interpret Levene's test together with the VR

15

Correcting data problems

- Deal with outliers
  - Remove the case
  - Transform the data
  - Change the score
- Deal with non-normality and heterogeneity of variance
  - Transform the data
  - Use nonparametric tests
  - Use robust tests
    - Bootstrap (Chapter 6 & 7)
    - Trimmed mean
    - M-estimator

→ "WRS" package
Not discussed/used in course

16

Correcting data problems

- 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, $1/(X_{Highest} - X_i + 1)$

17

Correcting data problems

- Log transformation

```
> dlf$logday1<- log(dlf$day1)
> dlf$logday1<- log(dlf$day1 + 1)
```

- Square root transformation

```
> dlf$sqrtday1<- sqrt(dlf$day1)
```

- Reciprocal transformation

```
> dlf$recday1<- 1/(dlf$day1 + 1)
```

18

DISCOVERING STATISTICS USING R

To transform … or not

- Transforming the data helps as often as it hinders the accuracy of *F*
  - The central limit theorem: sampling distribution will be normal in samples > 30 anyway
  - Transforming the data changes the hypothesis being tested
    - *E.g.* when using a log transformation and comparing means, you change from comparing arithmetic means to comparing geometric means
  - In small samples it is tricky to determine normality one way or another
  - The consequences for the statistical model of applying the 'wrong' transformation could be worse than the consequences of analysing the untransformed scores

(Games & Lucas 1966; Games 1984)

19

DISCOVERING STATISTICS USING R

To transform … or not

- When it all goes horribly wrong…
  - Look for nonparametric equivalent
  - Perform a robust analysis (*e.g.* based on bootstrapping)

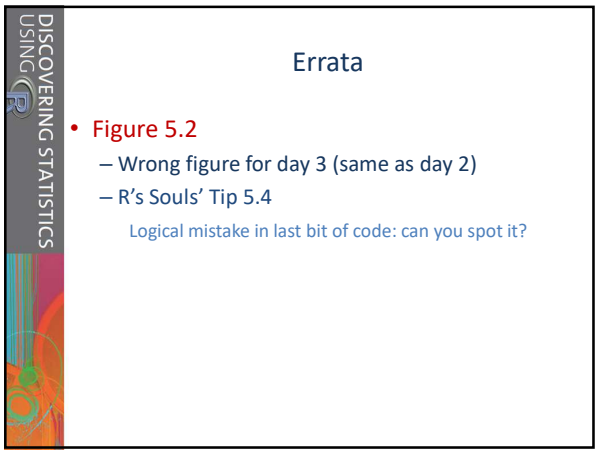
20

DISCOVERING STATISTICS USING R

Rest of morning and afternoon…

- Practical Chapter 5
  - Read § 5.1 - 5.3, "Cramming Sam's Tips" and "What have I discovered about statistics?"
  - Skip sections involving R Commander (Rcmdr)
  - Do all self-tests
  - Solve Smart Alex's tasks 1 & 2

21


DISCOVERING STATISTICS USING R
Errata
• Figure 5.2
– Wrong figure for day 3 (same as day 2)
– R's Souls' Tip 5.4
Logical mistake in last bit of code: can you spot it?

22