March 2

- Lecture
  - Boxplots & Lineplots (from Chapter 4)
  - Chapter 5: Assumptions
- Practical
  - Will upload this afternoon
  - If concerned about time tomorrow, take a look today and download any packages
- Homework
  - Read Chapter 6
  - Homework assignment upload this afternoon
    - Due at 10 before next lecture

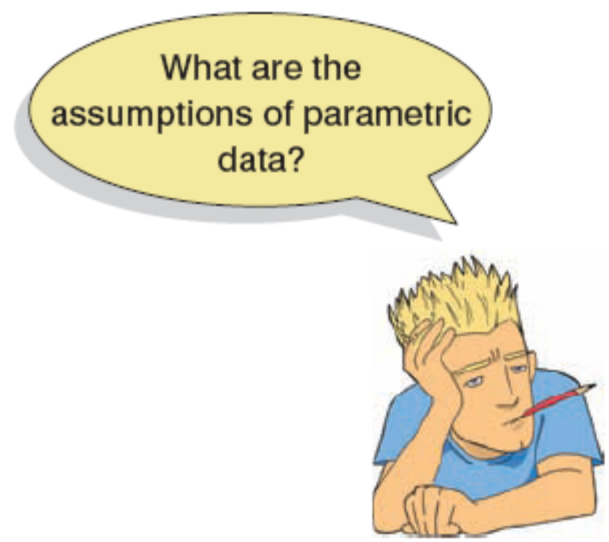
Exploring Assumptions

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

Assumptions

- Parametric tests based on the normal distribution assume:
  - Normally distributed
    - Sampling distribution
    - Errors
  - Homogeneity of variance
  - Interval or ratio level data
  - Independent data points



Assessing 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 (generally) normal anyway
- Graphical displays
  - Histogram
  - Q-Q plot (quantile-quantile plot)
- Values of skew/kurtosis
  - 0 in a normal distribution
- Shapiro-Wilk test
  - Tests if data differ from a normal distribution
  - Significant = non-normal data
  - Non-significant = normal data

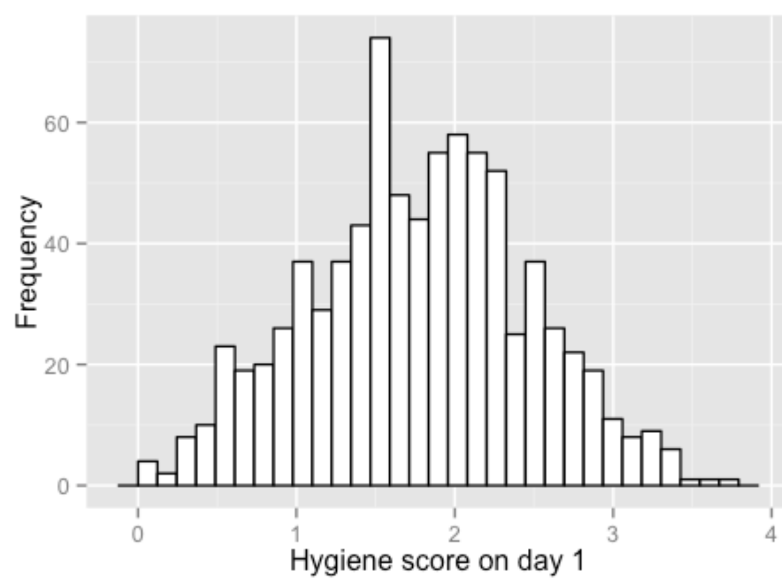
Normality Example

- A biologist was worried about the potential health effects of music festivals.
- Data from Download Music Festival
- Measured the hygiene of 810 concert-goers over the three days of the festival.
- Hygiene was measured using a standardized technique :
  - Score ranged from 0 to 4
    - 0 = you smell terrible
    - 4 = you smell lovely

Basic Histogram (with counts)

- To draw a histogram (for day 1 of the festival)

```
hist.day1 <- ggplot(dlf, aes(day1)) +
geom_histogram(colour="black", fill="white") +
labs(x="Hygiene score on day 1", y="Frequency") +
theme(legend.position="none"); hist.day1
```



Histograms (w/ normal distribution)

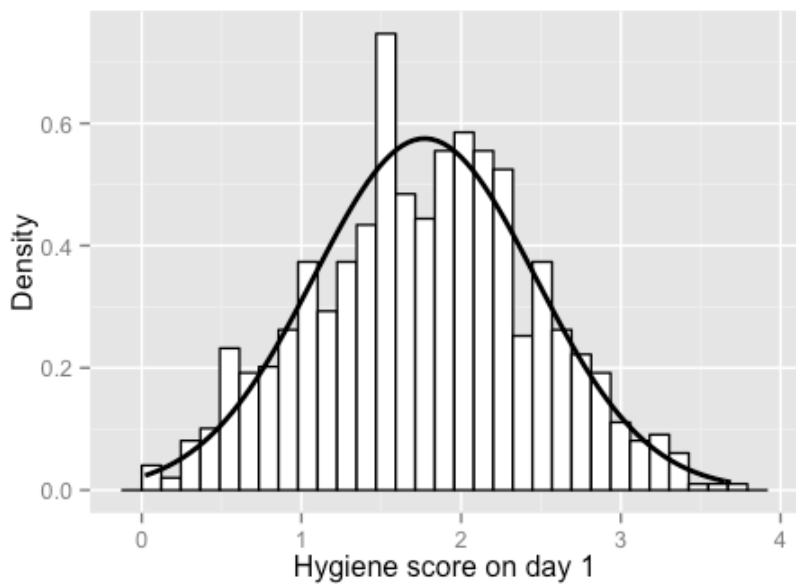
- To draw a histogram (for day 1 of the festival)

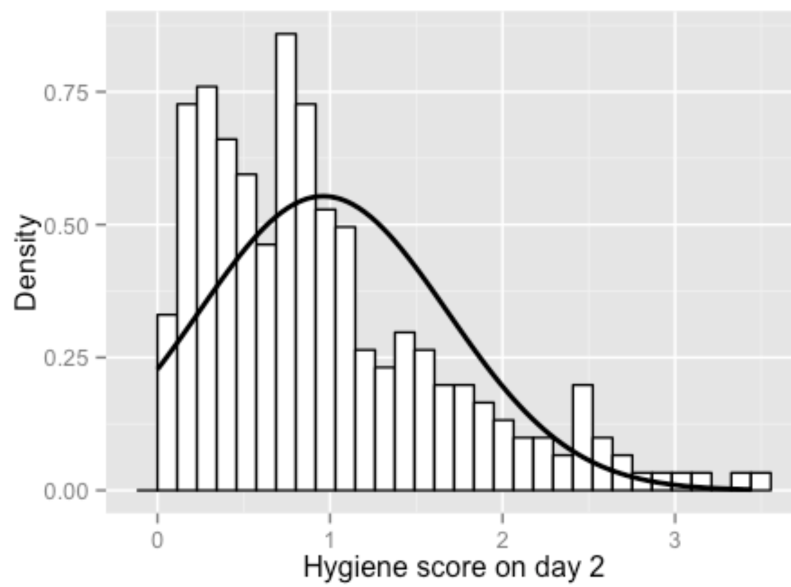
```
hist.day1 <- ggplot(dlf, aes(day1)) +
geom_histogram(aes(y = ..density..), colour="black",
fill="white") + labs(x="Hygiene score on day 1",
y="Density") + theme(legend.position="none"); hist.day1
```

- To superimpose a normal curve

```
mean_and_sd <- list(mean=mean(dlf$day1,
na.rm=TRUE), sd=sd(dlf$day1, na.rm=TRUE))
hist.day1 + stat_function(fun=dnorm,
args=mean_and_sd, colour=“black”, size=1)
```

Histograms (w/ normal distribution)





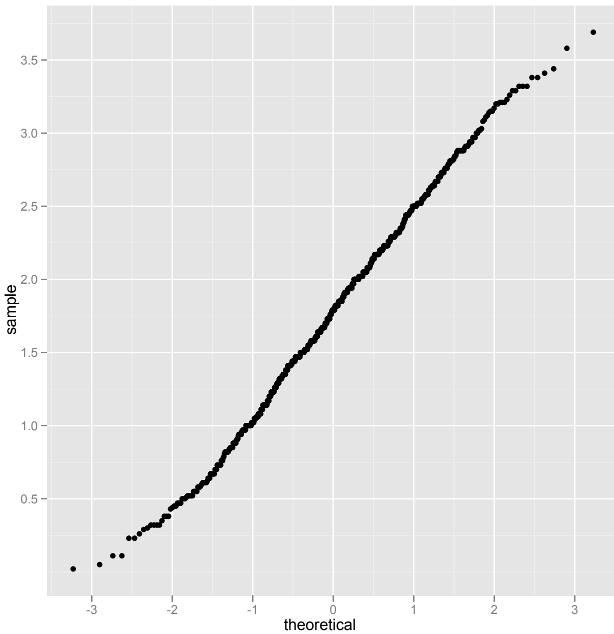
The Q-Q Plot

- To draw a Q-Q plot of the hygiene scores for day 1 of the music festival:

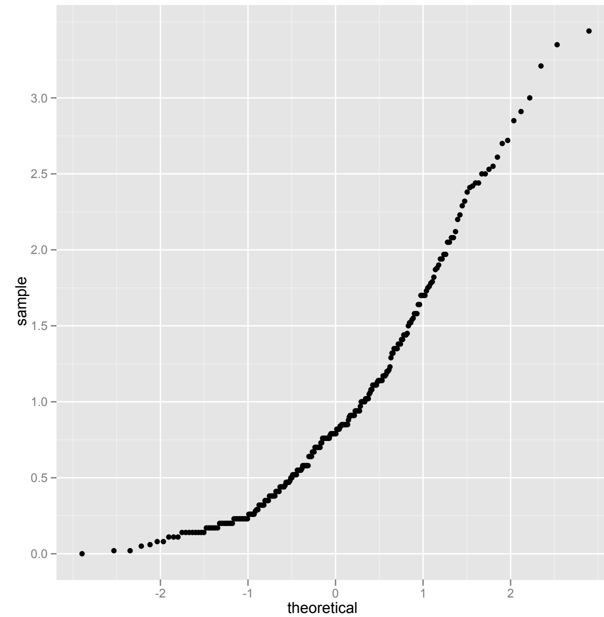
```
qqplot.day1 <- qplot(sample = dlf$day1, stat="qq")
qqplot.day1
```

The Q-Q Plot

Hygiene Scores: Day 1



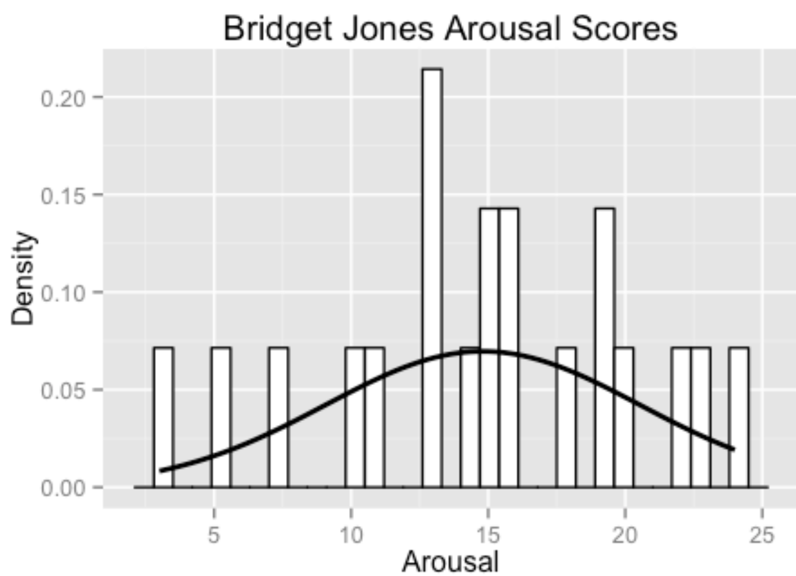
Hygiene Scores: Day 2

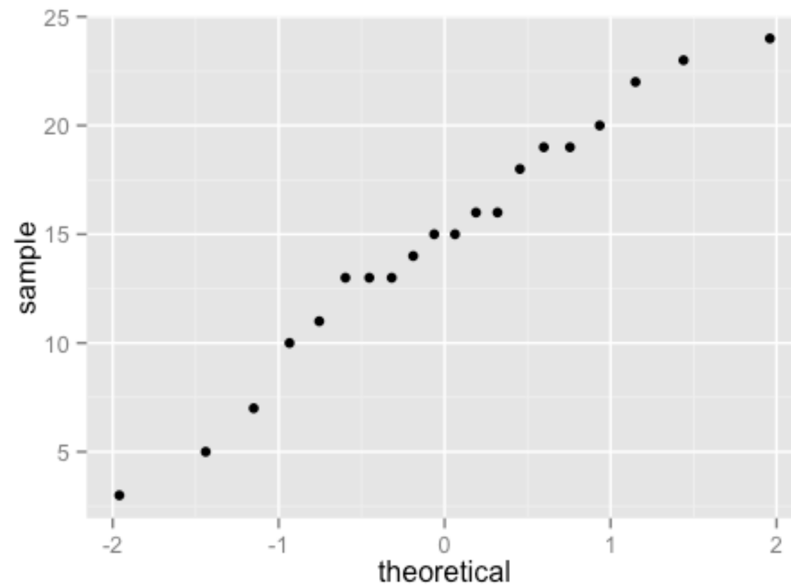


Normal

Not Normal

What does normal look like with small sample sizes?





Assessing Skew and Kurtosis

- Using stat.desc() (from pastecs package)

  stat.desc(dlf$day1, basic=FALSE, norm=TRUE)

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

  stat.desc(cbind(dlf$day1, dlf$day2, dlf$day3), basic=FALSE, norm=TRUE)

Assessing Skew and Kurtosis

	day1	day2	day3
median	1.790000000	7.900000e-01	7.600000e-01
mean	1.770828183	9.609091e-01	9.765041e-01
SE.mean	0.024396670	4.436095e-02	6.404352e-02
CI.mean.0.95	0.047888328	8.734781e-02	1.267805e-01
var	0.481514784	5.195239e-01	5.044934e-01
std.dev	0.693912663	7.207801e-01	7.102770e-01
coef.var	0.391857702	7.501022e-01	7.273672e-01
skewness	-0.003155393	1.082811e+00	1.007813e+00
skew.2SE	-0.018353763	3.611574e+00	2.309035e+00
kurtosis	-0.423991408	7.554615e-01	5.945454e-01
kurt.2SE	-1.234611514	1.264508e+00	6.862946e-01
normtest.W	0.995907247	9.083185e-01	9.077513e-01
normtest.p	0.031846386	1.281495e-11	3.804334e-07

Shapiro-Wilk Test

- Compares your sample to a normal distribution with same mean and SD as your sample
  - significant: your sample is not normal
  - non-significant: your sample is normal
- Beware large sample sizes!

Shapiro-Wilk Test

- shapiro.test(dlf$day1)

```
Shapiro-Wilk normality test

data:  dlf$day1
W = 0.9959, p-value = 0.03198
```

- Reporting
  - "According to a Shapiro-Wilk test, the hygiene scores on day 1, W=0.9959 and p-value=0.03198, were significantly non-normal"

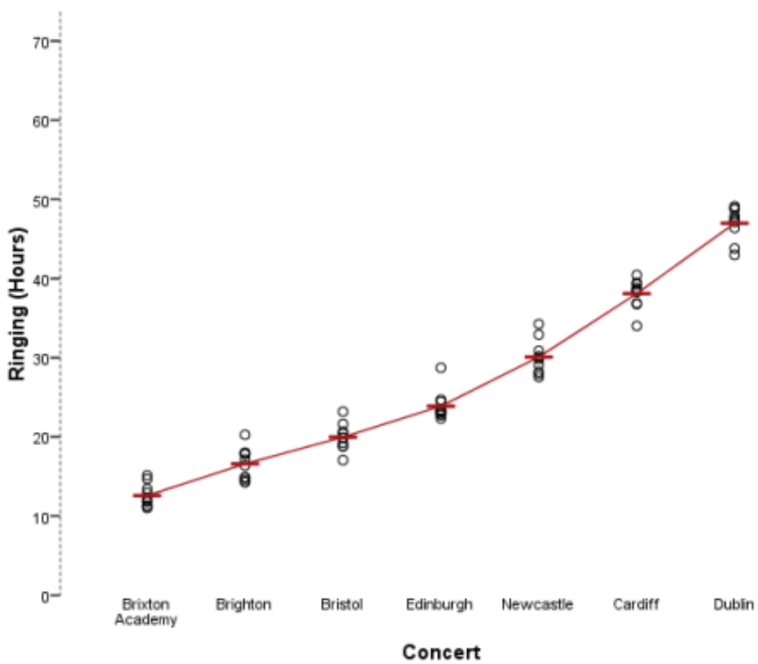
Summary for Day 1 of Festival

- Shapiro-Wilk: significantly not normal
- Skew: 0
- Kurtosis: not 0
- Q-Q Plot & histogram: appear normal
- Sample size: >30

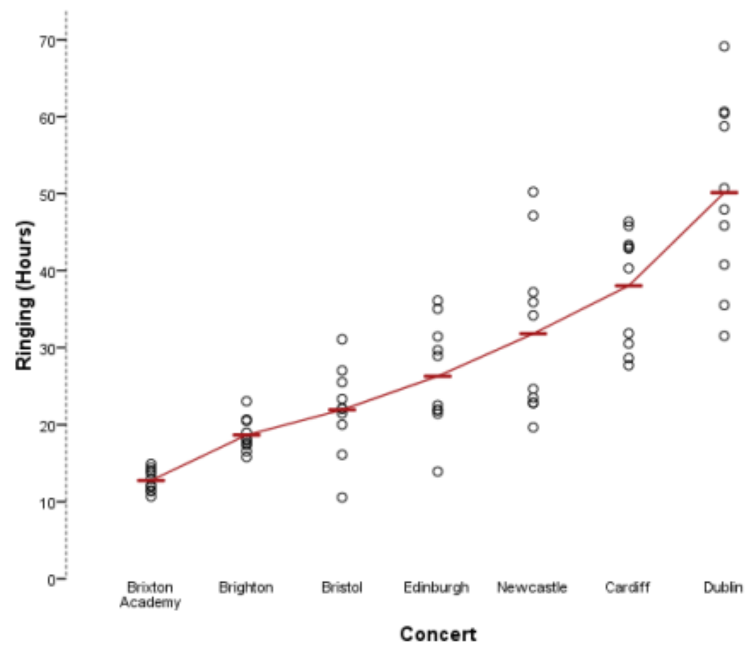
Assessing Homogeneity of Variance

- Graphs
- Levene's test
  - Tests if variances in different groups are the same.
  - Significant = variances not equal
  - Non-significant = variances are equal
- Variance ratio
  - With 2 or more groups
  - VR = largest variance/smallest variance
  - If VR < 2, homogeneity can be assumed.

Homogeneity of Variance



Homogeneous



Heterogeneous

Assessing Homogeneity of Variance with **R**

- Use the *leveneTest()* function from the *car* package:

  leveneTest(outcome variable, group, center = median/mean)

  – default is median

- Levene's test for exam scores from 2 different universities

  leveneTest(rexam$exam, rexam$uni)

Output for Levene's Test

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

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  2.0886 0.1516
      98
```

- Reporting
  - "For the scores on the exam, the variances were similar for the two universities, $F(1,98) = 2.09$, $p=0.152$."

Dealing with outliers

- Z-score of +/- 3.29 cuts off 99.9% of the data
  - any datapoints with z-scores w/ a greater absolute value than this are considered outliers (extreme values)
  - can bias mean and inflate standard deviation
- What to do?
  - remove to point (only if you don't actually think it is from the population)
  - change to next highest score +/- 1 unit
  - the mean +/- 2 standard deviation