Biostatistics

林　建　甫
C.F. Jeff Lin, MD. PhD.

台北大學統計系助理教授
台北榮民總醫院生物統計顧問
美國密西根大學生物統計博士

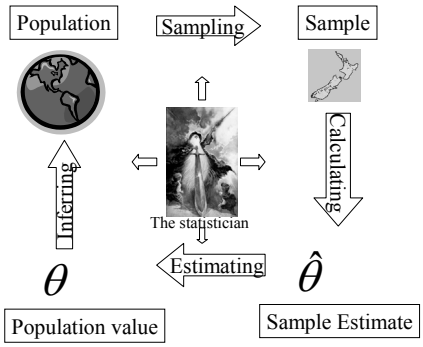
Some Aims of Course

- Learn more about how statistics can help solve real biomedical problems.
  - Including how to figure out what the problem really is!
  - What the statistical result means!
- Learn more about communicating the results of statistical manipulations to the 'client',
- Gain and improve specific skills in statistical analysis eg. regression, analysis of variance, multivariate analysis,

Some Aims of Course

- But **NOT** to learn mathematical theorems which may be behind statistical procedures
- Note the emphasis is on practical statistical inference.
- **R** will be used to perform all statistical calculations. We will also use Excel for data management.
- **R –** http://lib.stat.cmu.edu/R/CRAN

The Statistical Process



All about R

- Download is about 20+MB
- Quit moaning about it! It won't change my mind ☺
- Keep notes as you work

Exploration and Presentation of Univariate Data

- Single variable. (there may be multiple samples of the same variable)
- E.g. Reported cancer death for 2003 in Taiwan
  - We could group these into regions.
  - We will look at responses grouped into males and females.

Statistical Inference About a Relationship

- We have observations on two or more variables for a number of items.
- We believe there is an underlying (linear) relationship between the variables, but we don't know the parameters.
- Assuming a ***random*** error structure on the observations.
- We make inferences about the parameters of the relationship.
- Whether or not the conclusions extend to the population of items depends how the ones we measured were selected.

Inference on a Relationship

- Observational study--which items go into the treatment group(s) and the control group are not determined using randomisation.
- Randomised Experiment--we use randomisation to determine which items go into treatment groups and which go into control group.
- We can infer the relationship is a **causal** one only when the data comes from a randomised experiment.

Univariate Exploratory Data Analysis (EDA)

- Univariate means the data is comes from measurement of one variable
- Sometimes we have additional information to divide the data into groups
- Analysis and depiction depends on the type of data

Different Types of Data: Statistics

- Discrete data
- Continuous data

Different Types of Data: Statistics

- Discrete data
  - Count data
  - Categorical data
  - Ordinal data
- For example:
  1. Number of road accidents due to speed in a year
  2. Ethnicity question on the Taiwan Census
  3. Socio-economic class (low/med/high)

Different Types of Data: Statistics

- Continuous data
  - Measurements

For example:

1. Income in dollars
2. Amount of poison (in g/L) required to kill subject
3. Age, WT, HT, BP, Cholesterol, Blood sugar

Different Types of Data: Statistics

- Sometimes discrete data can be treated as continuous data, especially when there are many possible outcomes.
- For example – smashing a window and counting the number of glass fragments on 40cmx40cm sheet of paper around the breaker (2-4000)

Plotting Data

- Why plot data?
  - Often the best way to understand the data
- What do we look for when we plot data?
  - Depends on what kind of plot we have.
- How do we select the appropriate plot?
  - Depends on the data

Plotting Discrete Data

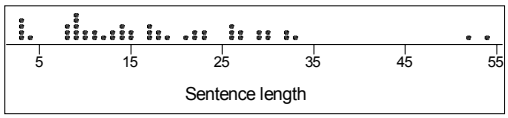
- For count data we have 4 options
  1. Dotplots
  2. Stem and leaf plots
  3. Histograms
  4. Boxplots

- In the following examples we have data on sentence lengths (in words) from 8 books

1. Eye of the Dragon
2. The Tommy- knockers
3. The Shining
4. The Stand
5. The Dark Half
6. Four Past Midnight
7. Rising Sun
8. Disclosure

- The first six are written by Stephen King and the last two by Michael Crichton

Dotplot



From the dotplot we can see:
- The location of the data
- The spread of the data
- Extremes in the data (outliers)
- Individual values

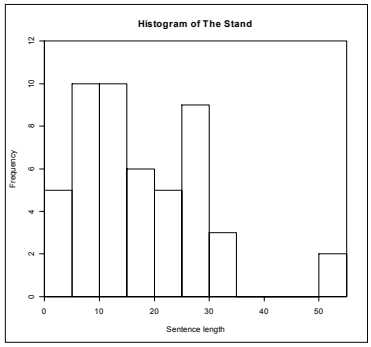
Stem and Left Plot

```
Stem-and-leaf of The Stand  N  = 50
The decimal point is 1 digit(s) to
   the right of the |

  0 | 33334
  0 | 88899999
  1 | 0011233444
  1 | 55777889
  2 | 12233
  2 | 6667799
  3 | 00223
  3 |
  4 |
  4 |
  5 | 24
```

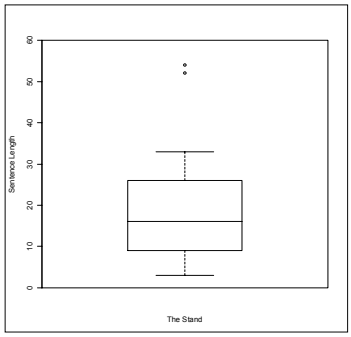
We can see:
- Location
- Spread
- Order statistics
- Individual values

Histogram



We can see:
- Location
- Spread
- Approximate mode
- Outliers

Boxplot



We can see:
- Location
- Spread
- Median
- IQR
- Outliers

Which Plot?

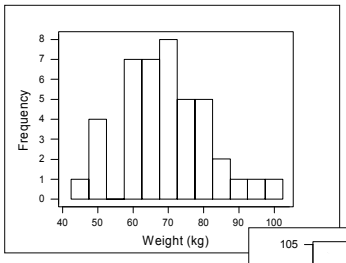
- It should be obvious that each of the plots we've seen has strengths and weaknesses
- Dotplots
  - Good for small data sets (n < 50)
  - and data that is not over dispersed
- Stem and leaf plots
  - Good for very small data sets (n < 20).
  - and data that is not over dispersed
  - Useful for ordering data
  - Not useful for any real problem.
  - Sensitive to choice of stems

Which Plot?

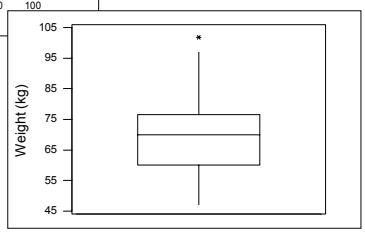
- Histograms
  - Good for larger data sets (> 20)
  - Often the best way to summarise data
  - Sensitive to choice of "class intervals"
  - Doesn't highlight individual values
- Boxplots
  - Good for larger data sets (> 15)
  - Often the best way to summarise grouped data
  - Doesn't highlight individual values
  - Does highlight potential outliers (3SD)

Plotting Continuous (cts.) Data

- The choice of plots is almost the same
  - Dotplots are not much use for cts. data
  - Similarly steam and leaf plots
  - Histograms and boxplots display continuous data well
  - Histograms are even more sensitive for cts. data
    - i.e. no "natural" class breaks



Histogram shows:
- Shape of data well
- Spread and mode



Boxplot shows:
- Spread
- Median (IQR)
- Outlier?

Which Plot?

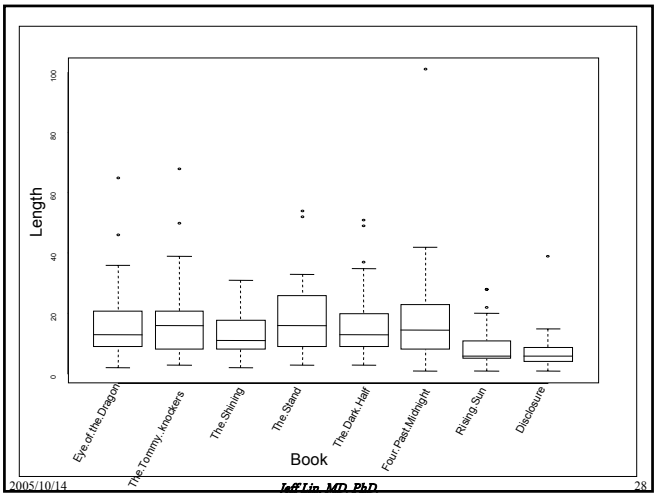
- Think about the following:
  - What am I trying to show?
  - Which features of the data are important?
    - Shape?
    - Location/Spread?
    - Outliers?
  - Is the plot easy to understand?
    - Chart junk!
    - Clutter!

Grouped Data

- Some ancillary (additional) information on grouping
- We want to compare between groups. Why?
  - To highlight or look for differences between groups
- Points to remember
  - Compare apples with apples!

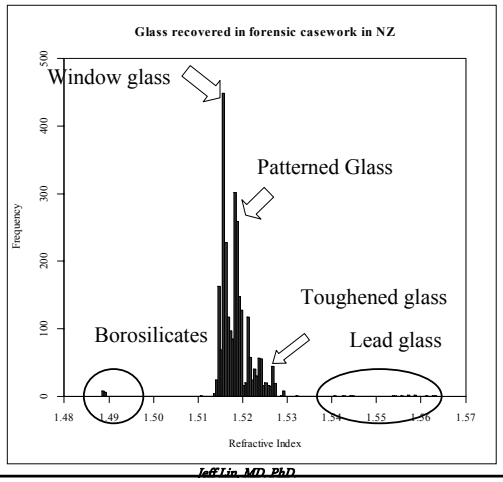
Plotting Grouped Data

- Two groups
  - Back to back stem and leaf plots
  - Parallel or back to back dot plots
- Two or more groups
  - Histogram matrix
  - Boxplots
- Boxplots are often the best way to compare groups



Data Description and Transformation

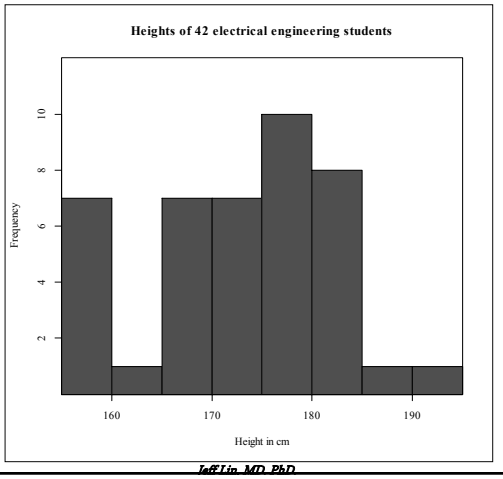
- What do we look for when we describe univariate data?
  - Peaks/Bumps
  - Location
  - Spread
  - Outliers

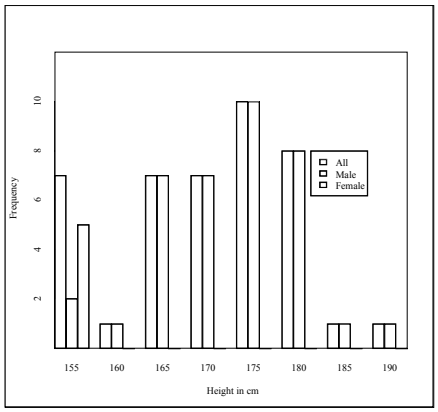


Data Description

- Peaks/Bumps
  - Unimodal
  - Bimodal
  - Multimodal

Why is there more than one mode?
- Contamination
- Mixing of more than one process?



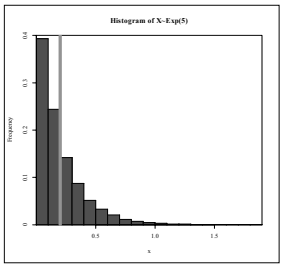


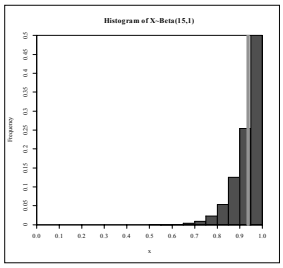
Outliers

- Outliers are:
  - Abnormal data points – "different from the bulk of the data".
- They may be:
  - Data entry errors
  - Measurement errors
  - Just really different pieces of data
- Outliers should always be investigated!

Symmetry and Skewness

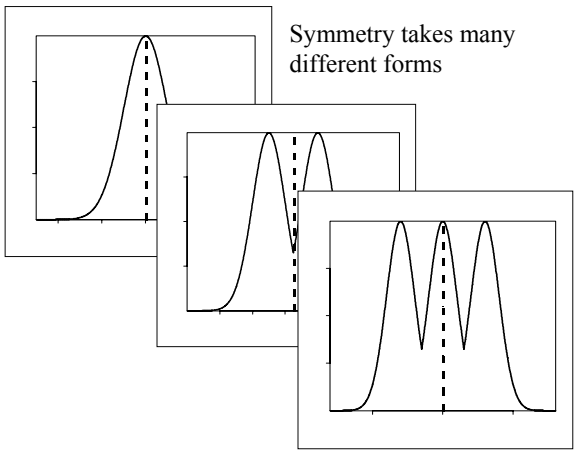
- Data may also be described in terms of symmetry or skewness
- If the extreme values of the data set lie to the left of the mean or median then the data is said to be left skewed.
- Conversely, if the extreme values of the data set lie to the right of the mean or median then the data is said to be right skewed.





- **Right** Skewed/Postively Skewed
- Long tail to the **right**
- Bulk of the data to the **left** of the mean $\Pr(X > \bar{x}) \approx 0.36$
- Mean = 0.20

- **Left** skewed/Negatively Skewed
- Long tail to the **left**
- Bulk of the data to the **right** of the mean $\Pr(X > \bar{x}) \approx 0.61$
- Mean = 0.93



Summary Statistics

- Measures of location
  - mean
  - median
- Measures of spread
  - variance/std. deviation
  - range
  - interquartile range (IQR)

Order Statistics and Quantiles

If our data are $x_1, x_2, \ldots, x_n$ then the order statistics are simple the data sorted into ascending order and denoted $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ and $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$

If α is of the form $\alpha_i = \frac{i}{n+1}$ for some integer i, define the corresponding quantile to be the ith order statistic, i.e. $q_{\alpha_i} = x_{(i)}$

- The α quantile, q_α, is chosen such that approximately $100\alpha\%$ of the data is less than q_α.
- For example if $\alpha = 0.1$ then approximately 10% of the data has a value less than the 0.1 quantile $q_{0.1}$

Example

- Let x = {9, 6, 10, 5, 3, 1, 5, 5, 10, 6}
- Then the order statistics are {1, 3, 5, 5, 5, 6, 6, 9, 10, 10}
- The first order statistic $x_{(1)} = 1$, is the $\alpha = \frac{1}{10+1} \approx 0.09$ quantile
- That is, $x_{(1)} = q_{0.09}$
- The last order statistic is $x_{(10)} = 10$, is 0.91 quantile
- Why isn't $x_{(10)}$ the 1.0 quantile?
- Imagine that we're finding quantiles for the normal distribution. What is the point x such that $\Pr(X < x) = 1$?
- The answer is $+\infty$ (positive infinity). This will cause us problems later on.

Quantiles

- The,0.25 quantile is the lower quartile
- The 0.5 quantile is the median
- The 0.75 quantile is the upper quartile

If you want a quantile from the data that is not represented by any α_i (e.g. if $n = 10$, and $\alpha = 0.1$), then it can be interpolated from the data by taking a weighted sum of the two surrounding quantiles

The Symmetry Statistic

- The symmetry statistic provides a numerical summary of symmetry.
- It is defined as:

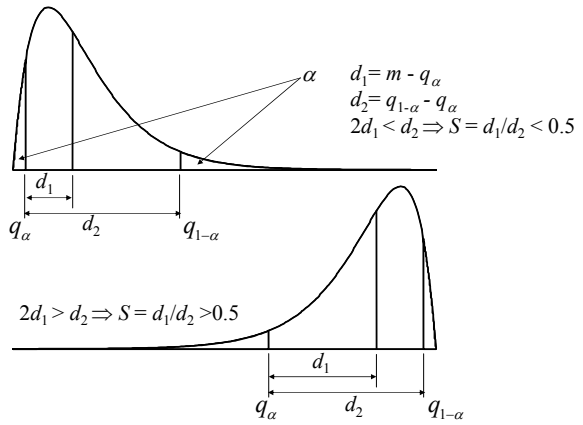
$$S = \frac{q_{0.5} - q_\alpha}{q_{(1-\alpha)} - q_\alpha}$$

"The ratio of the distance from the median to the lower quantile to the interquantile difference"

α is usually chosen to be 0.1 or 0.05. We will use 0.1

The Symmetry Statistic

- If S is less than 0.5 then the data is right skewed
  - The α quantile is closer to the median than the 1 - α quantile
- If S is greater than 0.5 then the data is left skewed
  - The 1 - α quantile is closer to the median than the α quantile



Transforming Data

- Sometimes we want to make data more symmetric. Why?
  - Makes it easier to see what is going on
  - Meet the assumptions of some parametric test
  - Makes the data better behaved
- How do we transform the data?
  - Can apply any mathematical function but typically we use power transformations

Power Transformations

- Note these only work for positive data
- Can raise data to an positive integer or fractional power p, for example p=0.5 (square root) or p = 3 (cubed data).
- Or raise data to a negative power – remember $x^{-p} = 1/x^p$.
- When p = 0, we use the logarithm. Why?

Choosing an Optimal Transformation

- In general:
  1. Increasing power (>1) will correct left skew
  2. Decreasing power (<1) will correct right skew
  3. Always choose a transformation you can explain (rounding is a good start)
  4. A log transformation is often all you need
- Symmetry power plot
  - Plot power (p) versus symmetry statistic S

Optimal Transformation

- Note: for $p < 0$ we must transform the data x to $-x^p$ to preserve the order of the data.
- Box-Cox transformations

$$x^{(p)} = \begin{cases} \dfrac{x^p - 1}{p}, & p \neq 0 \\ \log x, & p = 0 \end{cases}$$

–Box-Cox transformations try to make the data more normal

– Box-Cox method assumes data raised to some power p, comes from a normal distribution with mean μ and std. dev. σ

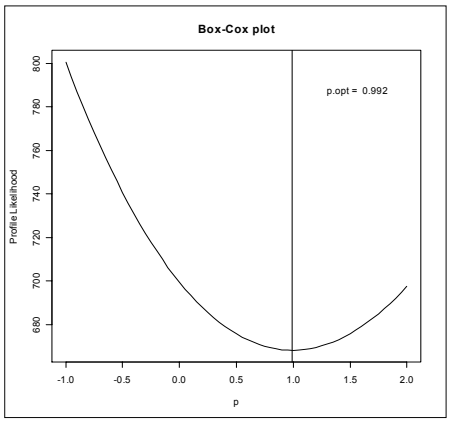
–To find p, μ and σ we use maximum likelihood estimation (mle)

Box-Cox Transformations

Minimise:

$$n \log s^{(p)} - (p-1)\sum_{i=1}^{n} \log x_i$$

This is a "profile likelihood" – which you might learn about in the third or fourth year. **YOU DO NOT NEED TO REMEMBER THIS OR EVEN KNOW HOW IT WORKS!**
Typically this is minimised using a graphical method or by some sort of numerical optimisation routing. R can do both. The routine I have supplied is available on the web page and from the server



Transformations - Summary

- **Only works on positive data**
- Box-Cox usually works well
- Make sure you can explain your transformation
- Taking the log often is the easiest thing to do
- Don't transform if you don't need to!

Symmetry and Skewness Measures

- We've seen the symmetry statistic S.
  - Robust to outliers
  - Crude symmetry statistic can be obtained using the quartiles. These can be gained in R by typing: `quantile(x)`
  - For the standard symmetry statistic, there is a function called s in the file called symmetry.r on the web and in the course folder. We type `s(x)`

Skewness Statistic

- Traditional measure of skewness is defined as:

$$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

where s is the sample standard deviation. This is sometimes called the 3rd central moment (the variance is the second)

– If this statistic is positive, then the data are right skewed and

– if it is negative then they are left skewed.

Quantile Plots

- Box-Cox transformations try to "transform to normality"
- Is there any way we can find out whether the data behave as though they came from a normal distribution?
- Yes – a normal quantile plot or a norplot.

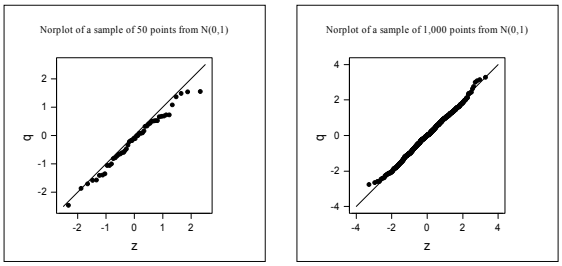
Norplots

- Define the quantiles of a standard normal distribution $N(0,1)$

$$z_\alpha = z \text{ s.t. } \Pr(Z < z) = \alpha$$

If the data come from a standard normal distribution then the empirical quantiles will be in approximately the same position as the theoretical quantiles

Example of a norplot

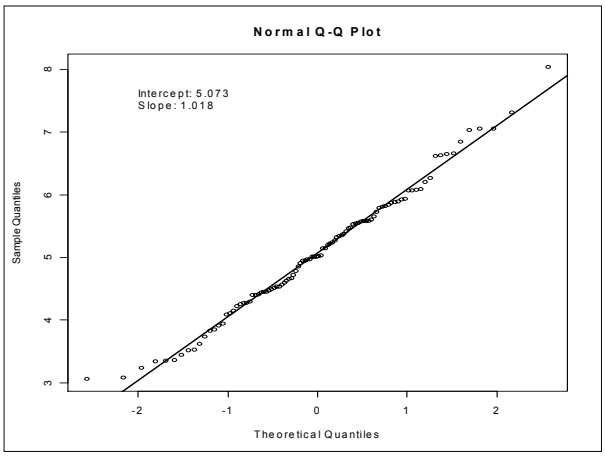


Norplots

- What happens if the data don't come from a standard normal?
- Remember that if $Y \sim N(\mu,\sigma)$ and $Z \sim N(0,1)$ then

$$Y = \sigma Z + \mu$$

- Therefore if the data are normal, but not standard normal, the data will have slope σ and intercept μ



Quantile Plots

- It is relatively easy to construct quantile plots for any of the standard distributions.
- However, the parameters are often unknown and so most people don't bother for anything other than a uniform distribution
- Sometimes, however, we wish to see if two sets of data are similarly distributed

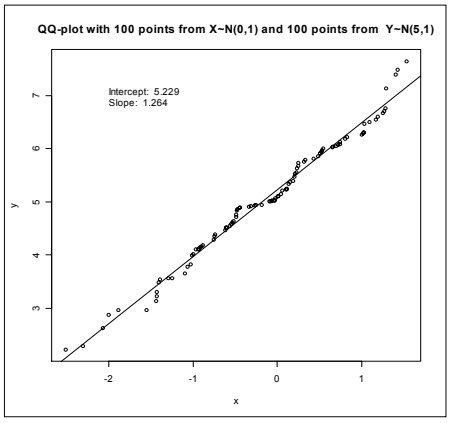
QQ-Plots

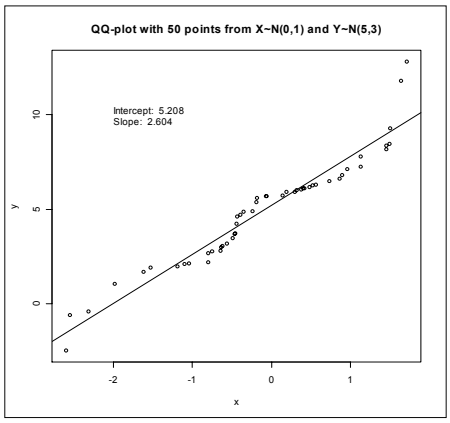
- QQ or quantile-quantile plots compare the quantiles of two sets of data
- If the data sets are similarly distributed then then quantiles will be similar and hence the points will cluster around the 45° line
- If the data sets have the same spread but differing means the relationship will be linear but the intercept will be different from zero
- If the data sets have the differing means and spreads the relationship will be linear but the slope will be different from 1 and the intercept non-zero

Constructing a QQ-plot

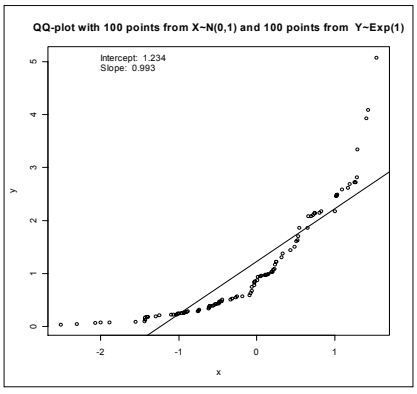
- Let X represent data set 1, and Y data set 2.
- If X and Y are of equal length then we can plot the order statistics of X vs. the order statistics of Y
- If $n_x > n_Y$ then we need to interpolate the quantiles from Y with

$$\alpha_i = \frac{i}{n_X + 1}$$





QQ-Plot – Different Distributions



Some Points about QQ-plots

- Easier to interpret if aspect ratio is 1:1
- Good for detect mean and variance/std. dev. shifts
- Not really convincing evidence against normality (need a lot of data)
- Heavy curvature usually means the data sets are from different distributions

A Small Case Study

- Description: Rainfall from Cloud-Seeding. The rainfall in acre-feet from 52 clouds, 26 of which were chosen at random and seeded with silver nitrate.
- Reference: Chambers, Cleveland, Kleiner, and Tukey. (1983). Graphical Methods for Data Analysis. Wadsworth International Group, Belmont, CA, 351.
- Original Source: Simpson, Alsen, and Eden.(1975). A Bayesian analysis of a multiplicative treatment effect in weather modification. Technometrics 17, 161-166.

Number of cases: 26

Variable Names:

1. **Unseeded**: Amount of rainfall from unseeded clouds (in acre-feet)
2. **Seeded**: Amount of rainfall from seeded clouds with silver nitrate (in acre-feet)

Unseeded	Seeded	Unseeded	Seeded
1202.6	2745.6	41.1	200.7
830.1	1697.8	36.6	198.6
372.4	1656	29	129.6
345.5	978	28.6	119
321.2	703.4	26.3	118.3
244.3	489.1	26.1	115.3
163	430	24.4	92.4
147.8	334.1	21.7	40.6
95	302.8	17.3	32.7
87	274.7	11.5	31.4
81.2	274.7	4.9	17.5
68.5	255	4.9	7.7
47.3	242.5	1	4.1

Step 1: Visually Inspect the Data

- It is easier to look at the order statistics ⇒ sort the data. First let's make the data easier to work with

```
clouds<-data.frame(seed=Seeded, u.seed=Unseeded)
attach(clouds)
```

Now sort the data. Type:

```
seed<-sort(seed)
u.seed<-sort(u.seed)
```

Type `clouds` to display the data

Visual Inspection of the Data

- There are some very large values
- There are some very small values
- The data does not increase linearly
- The data sets look as though they increase at the same rate but there is a shift in mean

Plot the Data

- A histogram of each data set would be helpful
  - Type:

```
par(mfrow=c(2,1))     # this sets the
graph screen   #to 2 rows 1 column
hist(u.seed)
hist(seed)
```


Interpret the Histograms

- Note the histograms don't have the same scales so it makes them difficult to compare
- Each histogram has a long tail to the right so the data sets are right skewed.
- Confirm this by calculating a symmetry statistic for each.
- Type:

```
s(seed)
s(u.seed)
```

Note the 0.1 – this means we're using the 0.1 quantile and the 0.9 quantile to calculate the symmetry statistic

Get Summary Statistics

- Type:
  ```
  summary(seed)
  summary(u.seed)
  ```
- This gives you a five number summary (Min, LQ, Med, UQ, Max) plus the mean
- To get the standard deviations type
  ```
  sd(seed)
  sd(u.seed)
  ```
- To get the skewness statistics type
  ```
  skewness(seed)
  skewness(u.seed)
  ```


Descriptive Statistics

- Note that the skewness statistic confirms our histogram interpretation and our symmetry statistic.
- We can plot the mean and median on each histogram:
  ```
  hist(seed)
  abline(v=median(seed),col="red",lwd=2)
  abline(v=mean(seed),col="blue",lwd=2)
  hist(u.seed)
  abline(v=median(u.seed),col="red",lwd=2)
  abline(v=mean(u.seed),col="blue",lwd=2)
  ```
  Notice how the median lines up with the centre of the data much better than the mean.
- If we want to compare groups it is going to be easier if we transform the data

Transforming the Data

- Choose a sensible transformation either using the `sympowerplot` function, e.g.
  ```
  sympowerplot(u.seed)
  sympowerplot(seed)
  ```
- Or by using the Box-Cox method.
  ```
  bcp(u.seed)
  bcp(seed)
  ```


Transforming the Data

- Box-Cox recommends $p = 0.047$ and $p = 0.118$ – therefore a log transformation might work well. If you're willing to wait awhile try
  - `bcp.bounds(seed)` and `bcp.bounds(u.seed)`
  - These will give approximate 95% confidence intervals on the powers and they both include 0
- Symmetry power plots recommend $p = -0.1$ and $p = 0.16$ for unseeded and seeded respectively.
- Log is easiest to explain, so try that.
- Look at log transformed histograms
  - Still lumpy, but skewness has been reduced

Compare the Data

- Boxplot the transformed data
  - Join the data into one vector
    ```
    rainfall<-c(log.u.seed, log.seed)
    ```
  - Use the `rep` command to make a vector of labels corresponding to the unseeded and seeded data
    ```
    gp<-rep(c("Unseeded","Seeded"),c(26,26))
    boxplot(split(rainfall,gp))
    ```
- QQ-Plot the data
  ```
  qqplot(log.u.seed,log.seed)
  qq.plot(log.u.seed,log.seed)
  ```


Compare the Data

- QQ-plot shows linear trend ⇒ logged data are from similar distributions
- Slope is approximately equal to 1 ⇒ similar spread
- Is the difference in the means significant ?
- Two-sample *t*-test

Bivariate Data

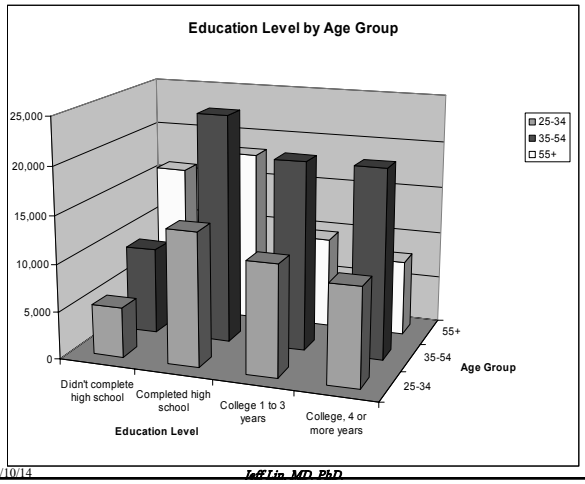
Bivariate Data

- Bivariate data is simply data where we have measurements on two variables.
- The variables can be either discrete or continuous, leading to three possible combinations: discrete/discrete, continuous/ discrete, continuous/continuous
- The type of analysis depends on the combination you have.
- Note if one of your variables is in continuous time units (such as years or seconds) then time series analysis is usually the appropriate treatment. We will not cover this subject

Displaying Discrete/Discrete Data

- Some texts recommend a three dimensional barplot

	Age Group			
Education	25-34	35-54	55+	Total
Didn't complete high school	5,325	9,152	16,035	30,512
Completed high school	14,061	24,070	18,320	56,451
College 1 to 3 years	11,659	19,926	9,662	41,247
College, 4 or more years	10,342	19,878	8,005	38,225
Total	41,387	73,026	52,022	166,435



3D Perspective Bar Charts

- Difficult to read
- Increase in categories leads to increase in complexity
- Increase in categories leads to more obscuration of data/features
- Perspective distorts the information
- **Don't do it!**

Displaying Discrete/Discrete Data

- **Two-way tables** are usually the best
- Think about what your aims are:
  - If the data are for someone else to use in an analysis or for reference in an appendix, then apart from organising the data in a legible fashion don't do anything to it.
  - If the data are for presentation or summary then apply Ehrenberg's rules

Displaying Tabular Data – Ehrenberg's Rules

1. Round drastically "2 busy digits"
2. Arrange the numbers so that comparisons are made column-wise and not row-wise
3. Order the columns by size or use ordinal information
4. Use row and column averages or sums as a focus
5. Use white space well
6. Provide verbal summaries

	Age Group			
Education	25-34	35-54	55+	Total
Didn't complete high school	5,325	9,152	16,035	30,512
Completed high school	14,061	24,070	18,320	56,451
College 1 to 3 years	11,659	19,926	9,662	41,247
College, 4 or more years	10,342	19,878	8,005	38,225
Total	41,387	73,026	52,022	166,435

	Age Group			
Education	25-34	35-54	55+	Total
Didn't complete high school	5,300	9,100	16,000	30,400
Completed high school	14,000	24,000	18,300	56,300
College 1 to 3 years	11,700	20,000	9,700	41,400
College, 4 or more years	10,300	20,000	8,000	38,300
Total	41,300	73,100	52,000	166,400

	Didn't complete high school	Completed high school	College 1 to 3 years	College, 4 or more years
25-34	17%	25%	28%	27%
35-54	30%	43%	48%	52%
55+	53%	33%	23%	21%

Statistical Analysis of Discrete Data

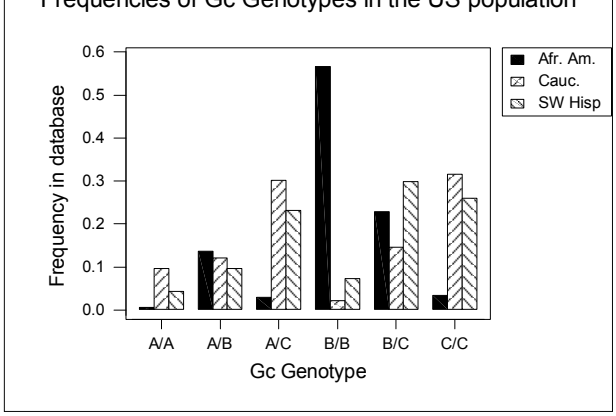
- There are two possible situations when dealing with bivariate discrete data
  1. Single sample cross-classified by two variables
     - E.g. 221 students classified by age group and birth order
  2. Multiple samples classified by a single variable
     - Genotype of individuals in different ethnic DNA databases
- As you would expect the methods of analysis are quite different

Multiple Samples Classified By a Single Categorical Variable

- If the samples are independent (and they usually are) then the category proportions can be compared across samples
- For example, in the US the FBI has DNA databases for African Americans, Caucasians and South Western Hispanics (Florida keeps it's own database for South Eastern Hispanics)

Analysing Multiple Sample Discrete Data

- Say we wished to compare genotype frequencies at a particular DNA locus called Gc
- Gc has 3 alleles A, B, and C. We have one allele from our mother and one from our father to make a genotype, so there are six possible genotype frequencies: A/A, A/B, A/C, B/B, B/C, and C/C
- How do we go about comparing the proportions of each genotype?
- We can plot them – always a good idea



Analysing Multiple Sample Discrete Data

- How about a statistical test?
- It boils down to the question we want to answer.
- If that question is: "Is there a difference in genotype probabilities between the races" then we can answer it. How?
- A chi-square analysis.
- Note, it turns out that this is the same analysis that we apply to discrete data from a single sample cross classified by two (or more) categorical variables, so we will assume we are talking about either case from now on.

EDA for Bivariate Data

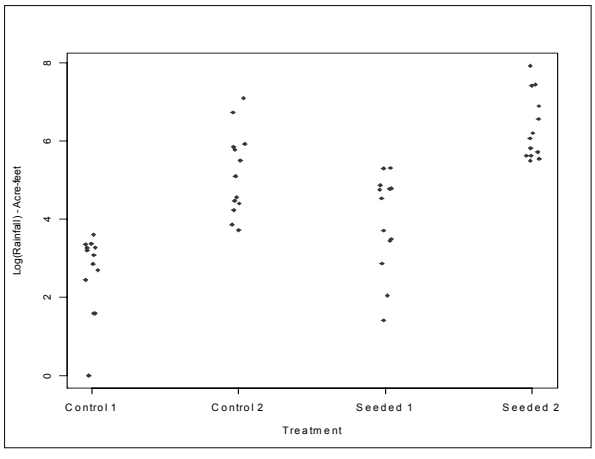
- Discrete/Continuous variables
  - Grouped data – combination of a categorical variable and a cts. measurement
    - Root weight and genotype (phenotype and genotype)
  - Continuous measurement/discrete measurement
    - Weight of cat and number of litters produced

Displaying Grouped Data

- For 2 groups
  - Back to back stem and leaf plots
  - Back to back histograms
  - Parallel or back to back dot plots
- For 2 or more groups
  - Boxplots
  - Histogram matrices
    - Jitter plots

Jitter Plots

- Plot continuous variable by group, but…
  - Add a small amount of random noise to the grouping variable so that all the data is visible.
  - The addition of random noise is called jittering
  - Jittering can be added to either the ***x*** or ***y*** variable
  - Usually jittering is done by adding a small amount of uniformly distributed noise
  - The range of the uniform should be relative to the size of spacings between the groups – i.e. don't obscure the information



Analysing Grouped Data

- 2 groups
  - Paired data (before and after data say)
    - Paired *t*-test
    - Sign test (non-parametric)
  - 2 independent groups
    - 2 sample *t*-test
    - Wilcoxon rank sum test (non-parametric)

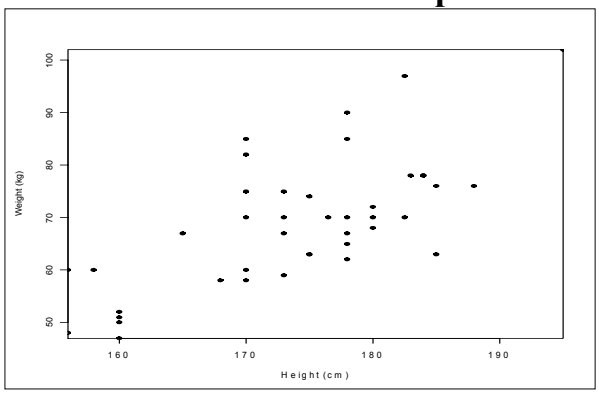
Analysing Grouped Data

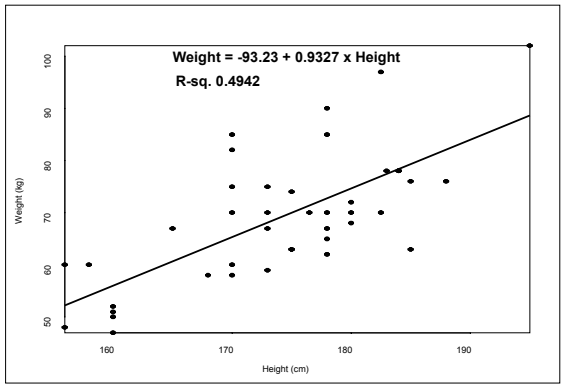
- *k* independent groups
  - One-way analysis of variance (ANOVA)
  - Kruskal-Wallis test (non-parametric)
- Plotting options for *k* groups
  - LSD plots
  - Tukey's HSD

Displaying Continuous Bivariate Data

- Scatterplot
  - Looking for a relationship between the two variables
    - This could be a **linear** relationship - the data follows a straight line
    - The relationship could be **non-linear** – the data follows a curve
  - It is rare that any relationship is perfect

A Linear Relationship?





Non-linear Relationships

- Sometimes non-linear relationships can be described by simple mathematical functions
  - Log/Exponential

$$y = \exp(\beta_0 + \beta_1 x) \text{ or } \log(y) = \beta_0 + \beta_1 x$$

  - Polynomial regression

$$y = \beta_0 + \beta_1 x^2$$

  - Trigonometric functions (uncommon)

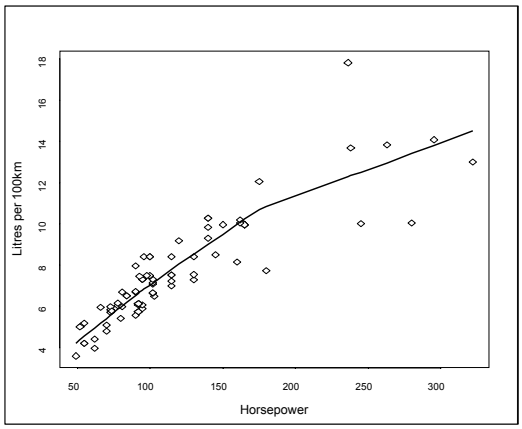
$$y = \beta_0 \sin(x)$$

More Advanced Problems

- Sometimes no mathematical function will adequately describe the relationship that exists in the data
- Solutions to this problem are harder
  - lowess (locally weighted least squares)
  - basis splines (beyond the scope of this course)

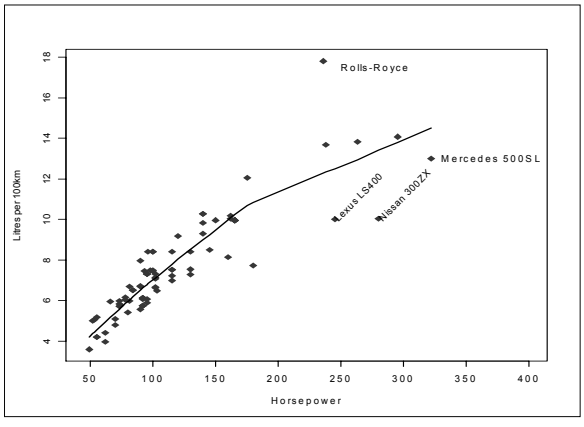
Lowess Scatterplot Smoothing

- We will discuss the methodology behind lowess smoothing when we learn about regression
- Why smooth?
  - Why do we plot data?
  - To describe a relationship
  - If data fail to conform to standard models do we give up?



Useful things to add to a scatterplot

- A fitted line
- Identifiers on extreme or outlying points
- Correlation or R^2 (only good for linear data)
- Identify groups or clusters of data



Thanks !