Global PaedSurg Research Training Fellowship



**Sessione 6: 25 Aprile 2019**

**Pulizia e Analisi dei Dati**

**Di Dr. Emily Smith and Tessa Concepcion**

# Pulizia dei Dati

* Pensate ai dati che avete raccolto
* Qual è la domanda della ricerca?
* Cosa state cercando di far dire ai vostri dati?
* Che test statistici sono i migliori per aiutarvi a rispondere alla domanda della ricerca? Questo cambia in base alla domanda della ricerca
* Contattate il team di ricercar/gli statistici per discutere come analizzare I dati

# Processo di analisi dei dati



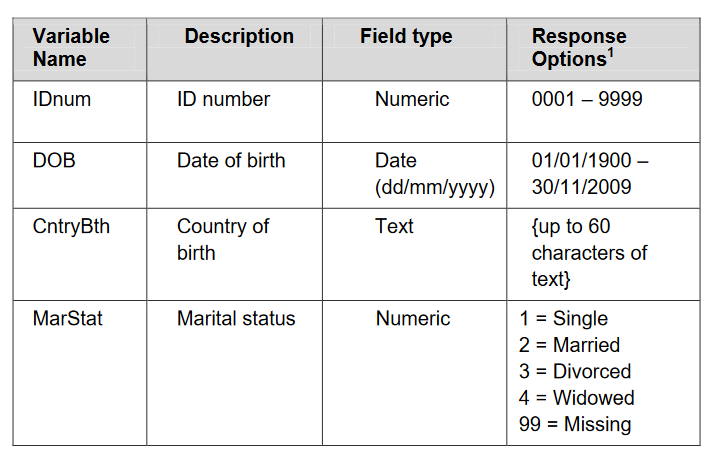
* Gestire i dati tramite Excel o REDCap
* Analizzare e interpretare un gran numero di dati utilizzando dei software statistici

# Step 1: Creare un piano di analisi

* Formulate un piano in accord con l’obiettivo e l’impostazione della vostra ricerca
* Raccogliete I dati
* Ora che avete raccolto I dati, considerate come stabilire la loro accuratezza e precision. Esempio: il processo di validazione di Global PaedSurg.

# Step 2: Gestire i dati

* Create una legenda dei dati
* La legenda deve includere almeno:
  + I nomi delle variabili
  + La descrizione delle variabili – cosa significano le variabili
  + Tipo di variabili
  + Le opzioni di risposta e il codice usato
* Alcune legend includono le colonne del questionario dove si può ritrovare la variabile. Microsoft Excel è una grande risorsa per le legend dei dati. Di seguito potete trovare un esempio.

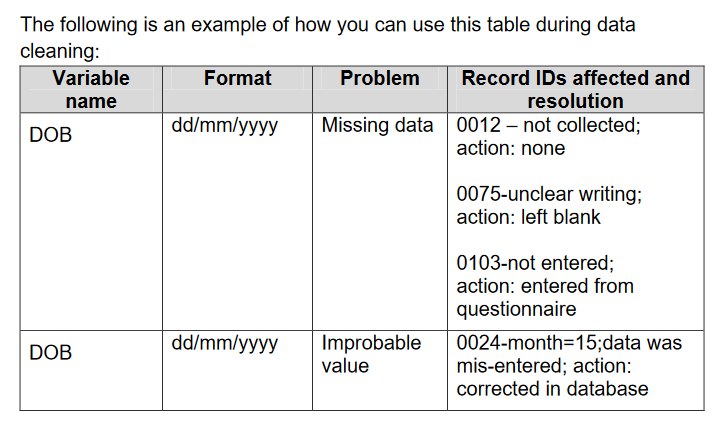


*NB: Assicuratevi di aggiungere una Colonna che descriva come codificare i dati mancanti.*

# Step 3: Pulizia dei dati

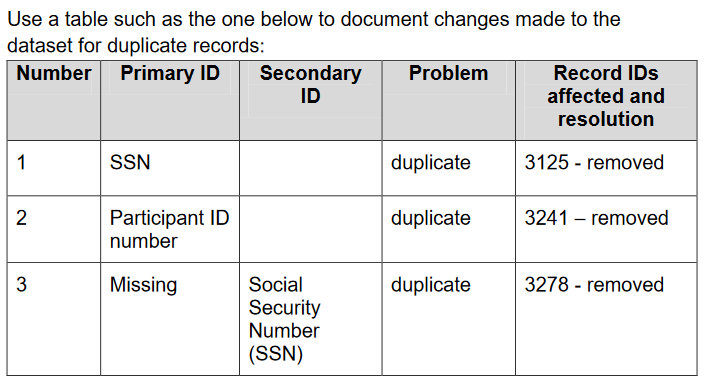
* Prima di iniziare salva una copia del dataset originale!!
* Pochi database sono privy di errori e dati mancanti
* Riguardare un database per identificare gli errori prima dell’analisi è importante
* Questo è un processo iterativo
* Documentate, documentate, documentate tutti i cambiamenti che fate! – annotate tutti i cambiamenti che fate per esempio
  + Cambi nel foglio di raccolta dati
  + Decisioni su come valutare un determinate campo
* Documentarlo ti permetterà di prendere decisioni coerenti e fornirà un document nel caso in cui ti facciano delle domande sulle tue analisi

Di seguito trovate un esempio di come usare una tabella per documentare i cambiamenti



Controllate che non ci siano dati doppi

* Identificate quanti records ci sono nel set di dati. Usate il vostro software statistico per controllare la conta.
* Determinare se il numero di records è uguale al numero di questionari/pazienti inclusi
* Se il numero di records è maggiore del numero di questionari/pazienti inclusi, eseguire un elenco di frequenze per cercare più record con le stesse informazioni identificative (come numero ID o nome). Se i dati sono anonimi, è comunque possibile valutare la presenza di duplicati: il software statistico può identificare voci di record uguali; con stesso peso, età gestazionale e età alla diagnosi.
* Se ci sono due records con lo stesso ID o stesso nome, selezionateli ed esaminateli per verificare se sono identici (record duplicato) o se il nome o l’ID sono stati inseriti in maniera non corretta.
* Usate una tabella come quella riportata qui sotto per documentare i cambiamenti effettuati sul set di dati per i record duplicate.



# Step 4: Identificare e correggere i dati mancanti, codificati in maniera errata o al di fuori del range di valori.

* Pochi dataset sono completi o accurati al 100%
* Normalmente ci sono alcuni dati strani o mancanti
* A volte questo succede in maniera causale altre in pattern

## Tipi di dati mancanti

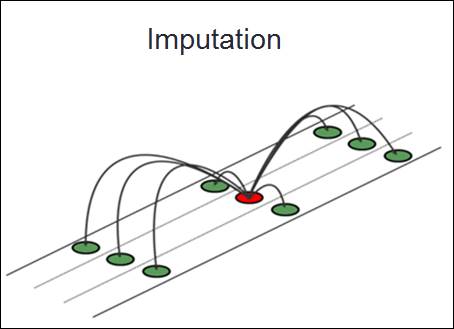
* MCAR (Mancanti in maniera totalmente causale): dati mandancti che sono indipendenti dalla variabile e si verificano in maniera casuale
* MAR (Mancanti in maniera casuale): La loro mancanza è collegata a una particoalre variabile, ma non è correlata al valore della variabile del dato mancante (dimenticanza accidentale di una risposta in un questionario)
* MNAR (Mancanti non in maniera casuale): Mancanti per una ragione, per esempio alcuni individui potrebbere essere esclusi.

Il modo migliore per identifiare le variabili mancanti è di creare delle tabelle di frequenza che mostrano la distribuzione e il totale delle vostre variabili.

## Maneggiare i dati mancanti

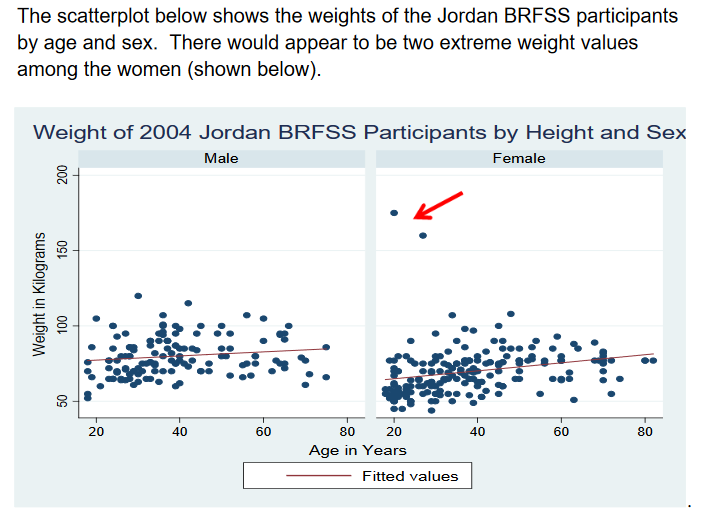
* *Analisi completa dei casi (analisi completa dei partecipanti)*
* *Usare l’intero* dataset così com’è (+/- rimuovere tutti coloro che hanno dati mancanti in una o più variabili)
* Può ridurre la precisione
* Senza bias in una vasta gamma di circostanze

### Imputing

Stima statistica di quali potrebbero essere le variabili mancanti basata su altri partecipanti con variabili simili

* Attribuire un valore di Default
* Inserire una media
* Regression imputation
* Attribuzioni multiple
* Ponderazione di probabilitù inversa

### Identificare valori anomali

Creare uno scatterplot che illustri i valori di una variabile sull’asse X e o valori di un’altra sull’asse Y.

# Analisi dei dati

* Dopo aver completato il cleaning dei dati.
* Ricorda che è necessario che tu abbia una copia del database originale con i dati grezzi oltre che il database con i dati controllati e *ripuliti*
* Il cleaned dataset varrà usato per le analisi

## Tipi di analisi statistiche

Statistiche descrittive

Descrivono un fenomeno, come *quanto/i?*

* Frequenze
* Misure di base

Possono essere presentati in una tabella che includa i numeri assoluti e le percentuali del totale che i numeri assoluti rappresentano. Usate i grafici a barre o a torta per le variabili discrete. Per la distribuzione, è possibile utilizzare grafici a scatole e baffi e istogrammi. Le variabili continue possono essere analizzate in categorie come l’età, usando diverse categorie con differenti range.

### Inferenza statistica

Giungere a delle conclusioni su un fenomeno, come provare o smentire delle teorie, associazioni tra fenomeni, se il campione ha una relazione con una popolazione più estesa, es la dieta e la salute; e determinare se i risultati sono significativi.

* Testare le ipotesi
* Intervallo di confidenza
* Testare la significatività
* Predizioni
* Correlazioni

## Analidi dei dati continui

### Correlazioni

Quando usarle?

* Quando vuoi conoscere l’associazione o le relazioni tra due variabili continue.
* C:\Users\ddavidov\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\N3L1XTC8\MC900014156[1].wmfEsempio: intake alimentare e peso; dosaggio del farmaco e pressione arteriosa; temperatura atmosferica e metabolismo ecc.

Cosa ci dicono?

* Se esiste una relazione lineare tra le due variabili e la forza di questa relazione

Come appaiono i risultati?

* Coefficiente di correlazione = Pearson’s *r (indice di correlazione di Pearson)*
* Range da -1 a +1

### T-tests

Cosa ci dice il t-test?

* Se c’è una differenza statisticamente significativa tra la media o i valori dei due gruppi (sia nello stesso gruppo di persone prima e dopo, sia in due gruppi diversi).

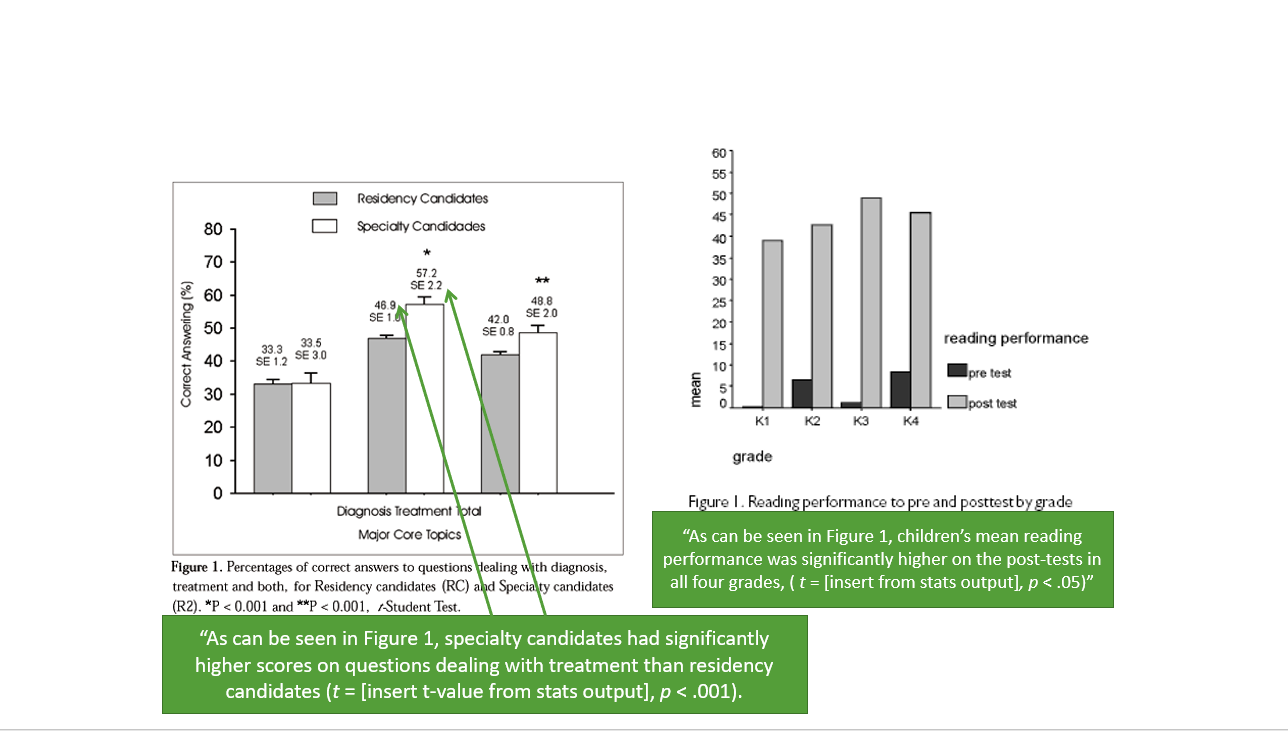
*C:\Users\ddavidov\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\F0382YT3\MC900014158[1].wmf*

Come appaiono i risultati?

* Students’ *t*

Come li interpreto?

* Guardando al corrispondente p-value
  + Se p < 0,05 la differnenza è statisticamente significativa
  + Se p > 0,05 non c’è differenza statisticamente significativa

Si possono riportare come illustrato di seguito

## Analisi delle variabili discrete (nominali)

### Chi-squared

Quando usarlo?

* Quando vuoi sapere se esiste un’associazione tra due variabili discrete (nominali) (es., tra esposizione e outcome)
  + Esempio: Fumo (si/no) e tumore ai polmoni (si/no)
  + Esempio: Obesità (si/no) e diabete (si/no)

Cosa ci dice il Chi-square?

* Se le frequenze osservate in ciascun gruppo sono statisticamente significative dalle frequenze attese (es. una differenza di proporzioni)

C:\Users\ddavidov\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\N3L1XTC8\MC900434529[1].wmf

Come appaiono i risultati?

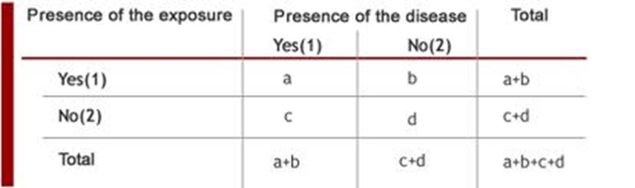
* Chi-square test statistics = *X*2

Come li interpreto?

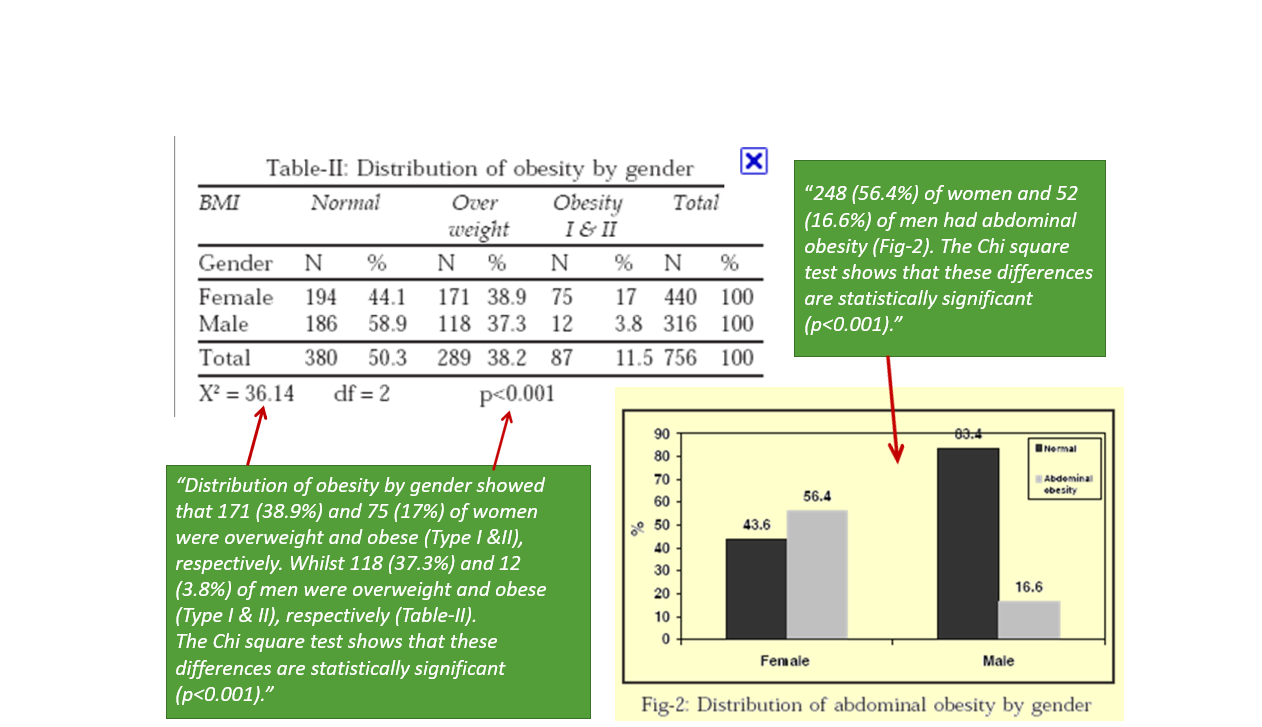
* Normalmente più è elevato il Chi-square più è versosimile che i risultati siano significativi, ma bisogna considerare il p-value per determinare la significatività

*Tip: Chi square richiede che ci sia 5 o di più in una tabella 2x2 e 5 o più in almeno 80% delle celle in una tabella più grande. Nessuna tabella può avere 0 come valore.*

Esempio di tabella di seguito:



Come riportarlo è descritto di seguito (P value):



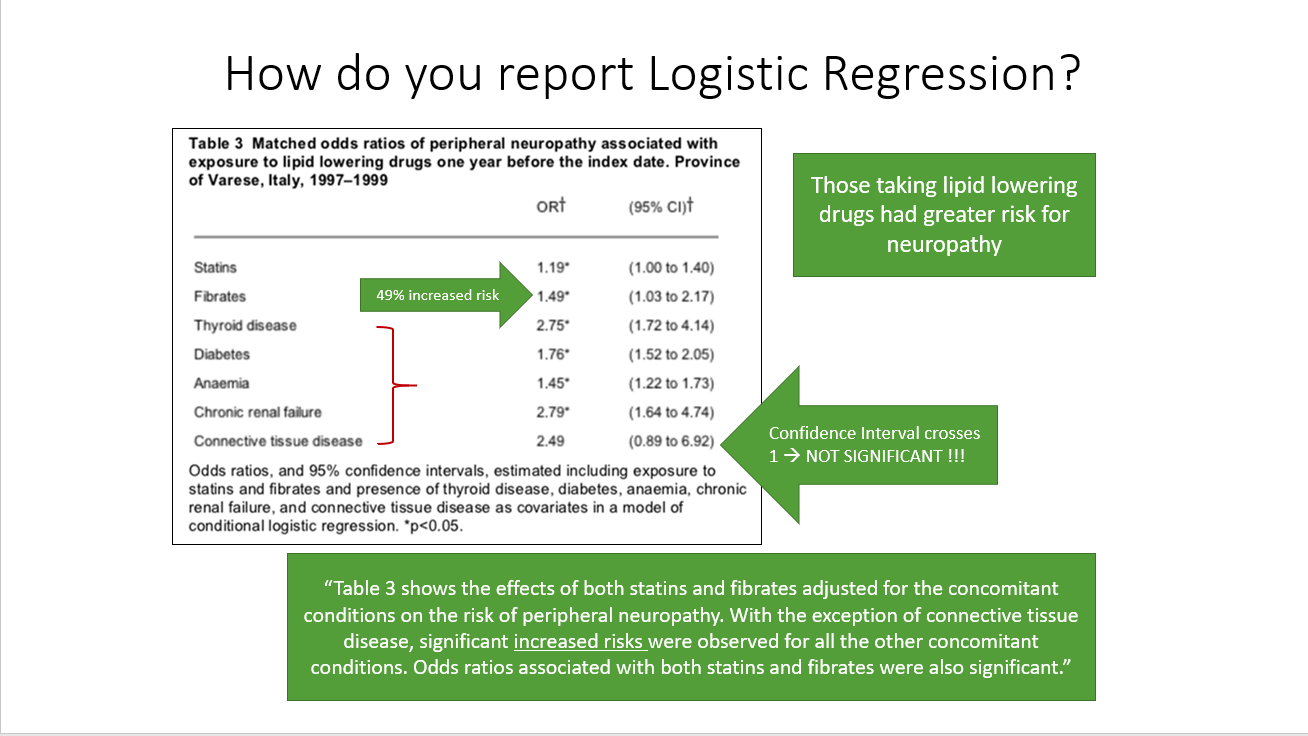
### Regressione logistica

Quando usarla?

* Quando si vuole misurare la forza e la direzione dell’associazione tra dye variabili, quando la variabile dipendente è discontinua (es. si/no)
* Quando si vuole predire la probabilità di un’outcome controllando i fattori confondenti

Come interpretare i risultati?

* La significatività può essere dedotta usando gli intervalli di confidenza:
* Se gli intervalli non comprendono 1 (es., 0.04 – 0.08 o 1.50 – 3.49), allora il risultato è significativo.
  + Se OR > 1 🡪 l’outcome è tante volte più probabile quanto indicato dal numero
  + La variabile indipendente potrebbe essere un FATTORE DI RISCHIO
  + 2.0 = due volte più probabile
* Se OR < 1 🡪 l’outcome è tante volte meno probabile quanto indicato dal numero
  + La variabile indipendente potrebbe essere un FATTORE PROTETTIVO
  + 0.50 = 50% meno probabile

Come riportare i risultati è mostrato di seguito

## Riassunto dei test statistici

|  |  |  |  |
| --- | --- | --- | --- |
| **Test** | **Tipo di dati necessari** | **Test statistico** | **Esempio** |
| **Correlazione** | 2 variabili continue | Pearson’s r | Pressione arteriosa e peso sono correlati? |
| **T-tests/ANOVA** | Media di variabili continue prese da due gruppi | Student’s *t* | I pazienti con peso normale (gruppo 1) hanno una pressione più bassa dei pazienti obesi (gruppo 2)? |
| **Chi-square** | 2 variabili discontinue | Chi-square *X*2 | I pazienti obesi (obesi vs non obesi) hanno più probabilità di avere un ictus? (ictus vs non ictus) |
| **Regressione logistica** | Variabile dicotomica come outcome | Odds Ratios (OR) & 95% intervalli di confidenza (CI) | L’obesità è un fattore di rishcio per l’ictus (ictus vs non ictus) quando corretta per altre variabili? |

# Bibliografia

*Essential Medical Statistics*. Kirkwood & Sterne, 2nd Edition. 2003

http://ocw.tufts.edu/Content/1/lecturenotes/193325

http://stattrek.com/AP-Statistics-1/Association.aspx?Tutorial=AP

http://udel.edu/~mcdonald/statcentral.html

*Background to Statistics for Non-Statisticians*. Powerpoint Lecture. Dr. Craig Jackson , Prof. Occupational Health Psychology , Faculty of Education, Law & Social Sciences, BCU. *ww.hcc.uce.ac.uk/****craigjackson****/Basic%20****Statistics****.ppt.*