Bolsas de Formação para Treinamento em Pesquisa Global PaedSurg



**Sessão 6: 25 Abril 2019**

**Limpeza e Análise de Dados**

**Por Dra. Emily Smith e Tessa Concepcion**

# Limpar o conjunto de dados

* Pense nos dados que você coletou.
* Qual é a pergunta da sua pesquisa?
* O que você está tentando fazer com que seus dados “digam”?
* Quais testes estatísticos ajudarão você a responder melhor a pergunta da sua pesquisa? Isso irá variar dependendo da a pergunta da sua pesquisa?
* Entre em contato com a equipe de pesquisa/ estatístico para discutir como analisar seus dados.

# Processo de Análise de Dados



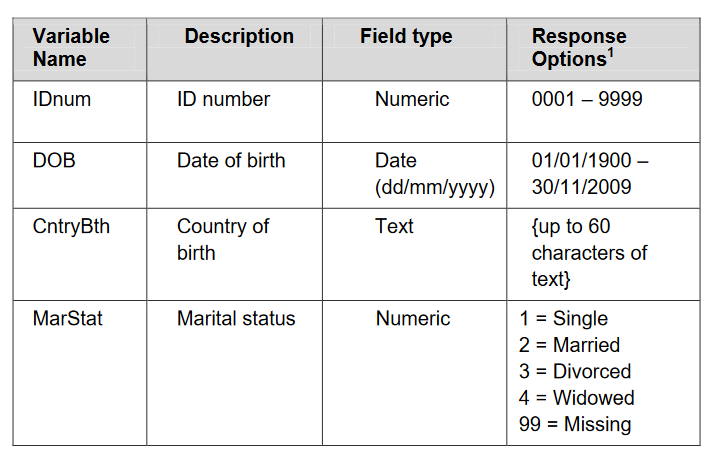
* Gerenciando dados – via Excel ou REDCap
* Analisar e Interpretar grandes conjuntos de dados usando software estatístico

# Etapa1 : Criando um plano de análise

* Formule seu plano de acordo com seu objetivo e configuração de pesquisa
* Colete os dados
* Agora que você que coletou os dados, considere como você avalia sua exatidão e precisão. Exemplo – processo de validação da Global PaedSurg.

# Etapa 2: Gerenciamento de Dados

* Crie um dicionário de dados.
* Um dicionário de dados deve incluir no mínimo:
  + Nomes das variáveis
  + Descrição das variáveis – o que a variável significa
  + Tipos de variáveis
  + Opções de resposta e quaisquer códigos usados
* Alguns dicionários de dados incluem a coluna do questionário onde a variável pode ser encontrada. O Microsoft Excel é um ótimo recurso para um dicionário de dados. Um exemplo está anexado abaixo

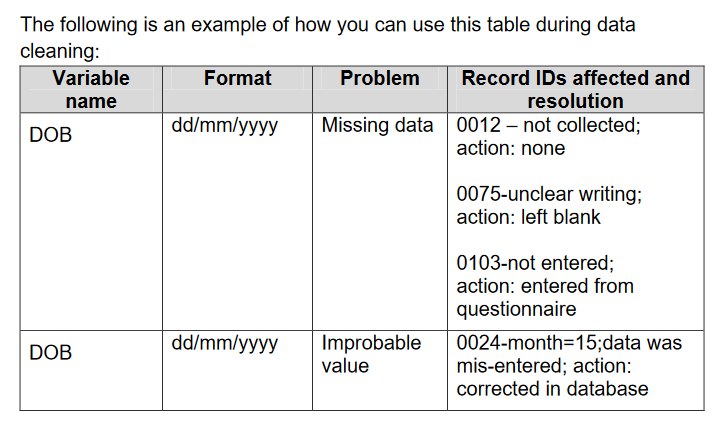


*NB: Certifique-se de adicionar uma coluna descrevendo como codificar os valores ausentes*

# Etapa 3: Limpeza de Dados

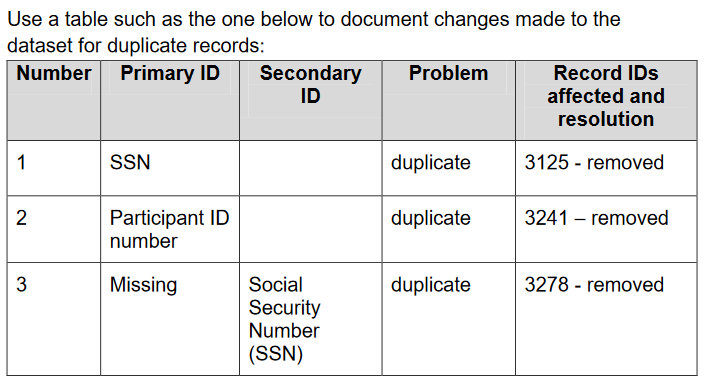
* Antes de começar, faça uma cópia do conjunto de dados original!!
* Poucas bases de dados estão livres de erros e valores faltantes
* Revisar o conjunto de dados para identificar erros antes da análise é importante
* Este é um processo iterativo
* Documente, documente, documente qualquer mudança que você fizer! – anote todas as alterações feitas como
  + Alterações no conjunto de dados
  + Decisões sobre como avaliar determinados campos
* A documentação assegurará que você tome decisões consistentes e fornecerá uma referência para aqueles que possam ter dúvidas sobre sua análise

A seguir, um exemplo de como usar essa tabela ao documentar qualquer alteração.



Verifique se há registros duplicados

* Identifique quantos registros estão no conjunto de dados. Use seu software estatístico para verificar a contagem de registros
* Determine se o número de registros corresponde ao número de questionários / inscrições
* Se os registros forem mais do que o número de questionários / inscrições, execute uma listagem de frequência para procurar vários registros com as mesmas informações de identificação (como número de ID ou nome). Se os dados forem anônimos, você ainda poderá avaliar as duplicatas - o software estatístico pode identificar as entradas de registro que são as mesmas, ou seja, o mesmo peso, idade gestacional e idade no momento do diagnóstico.
* Se houver dois registros com o mesmo número de ID ou nome, selecione os registros e examine-os para determinar se eles são idênticos (um registro duplicado) ou se um número de ID ou nome foi inserido incorretamente
* Use uma tabela como a abaixo para documentar as alterações feitas no conjunto de dados para registros duplicados



# Etapa 4: Detectando e corrigindo valores ausentes, incorretos ou fora do intervalo

* Poucos conjuntos de dados são 100% completos ou precisos
* Geralmente existem alguns valores estranhos ou ausentes
* Às vezes ocorre aleatoriamente ou em padrões

## Tipos de dados perdidos

* MCAR (ausente completa e aleatoriamente): os dados ausentes são independentes das variáveis e ocorrem aleatoriamente.
* MAR (ausente aleatoriamente): a ausência está relacionada a uma variável específica, mas não está relacionada ao valor da variável que possui dados faltantes (omitindo acidentalmente uma resposta em um questionário)
* MNAR (ausente não aleatoriamente): ausente por uma razão, por exemplo, alguns indivíduos podem ser excluídos

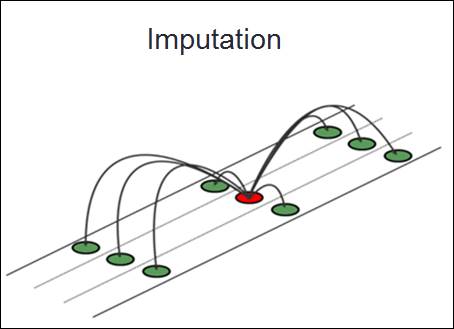
A melhor maneira de identificar as variáveis ausentes é executar as tabelas de frequência, elas mostram a distribuição da variável e os totais.

## Manipulando Dados Ausentes

### Análise completa do casos (análise completa do participante)

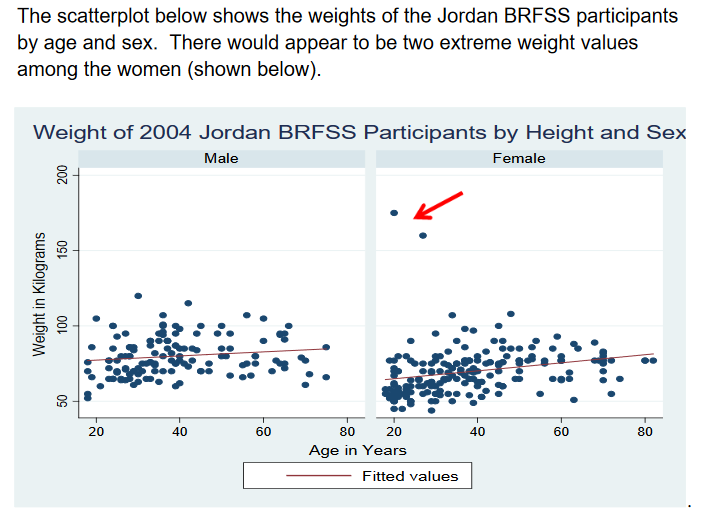
* Usa todo o conjunto de dados como é (+/- remove todos com dados ausentes em uma mais variáveis)
* Pode reduzir a precisão
* Imparcial em uma ampla variedade de circunstâncias

### Imputando

Estimativa estatística do que a variável ausente seria baseada em outros participantes com variáveis semelhantes.

* Valor padrão imputado
* Imputação média
* Imputação regressão
* Imputação múultipla
* Ponderação de probabilidade inversa

### Identificar outliers

Fazer um gráfico de dispersão ilustra o valor de uma variável no eixo X e o valor da outra no eixo Y.

# Análise de dados

* Agora que você limpou seus dados.
* Lembre-se de que você deve ter uma cópia dos dados brutos que não foram manipulados, bem como o conjunto de dados limpo.
* O conjunto de dados limpo será usado para análise

## Tipos de Análise Estatística

Estatística Descritiva

Descreve um fenômeno, como quantos?

* Frequências
* Medições básicas

Pode ser apresentado em uma tabela, incluindo o número bruto e a porcentagem do total que o número bruto representa. Uso de gráficos de barras e gráficos de pizza para dados categóricos. Para distribuição, diagrama de caixa e histogramas podem ser usados. Variáveis contínuas podem ser feitas em categorias como idade, usando várias categorias com diferentes intervalos.

### Estatística Inferencial

Infere sobre um fenômeno, como provar ou refutar teorias, associações entre fenômenos, se a amostra se relacionar a uma população maior, ou seja, dieta e saúde. E determine se as descobertas são significativas.

* Testando hipóteses
* Intervalos de confiança
* Teste de significância
* Predição
* Correlação

## Análise de Dados Contínuos

### Correlação

Quando usá-lo?

* Quando você quer saber sobre a associação ou relação entre duas variáveis contínuas.
* C:\Users\ddavidov\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\N3L1XTC8\MC900014156[1].wmfExemplo: ingestão de alimentos e peso; dosagem de droga e pressão sanguínea; temperatura do ar e taxa metabólica, etc.

O que isso diz a você?

* Se existe uma relação linear entre duas variáveis e quão forte é essa relação.

Como são os resultados?

* O coeficiente de correlação = r de Pearson
* Varia de -1 a +1

### Teste-T

O que um teste-t diz?

* Se houver diferença estatisticamente significante entre o escore médio ou o valor entre dois grupos (seja o mesmo grupo de pessoas antes e depois, ou dois grupos diferentes de pessoas).

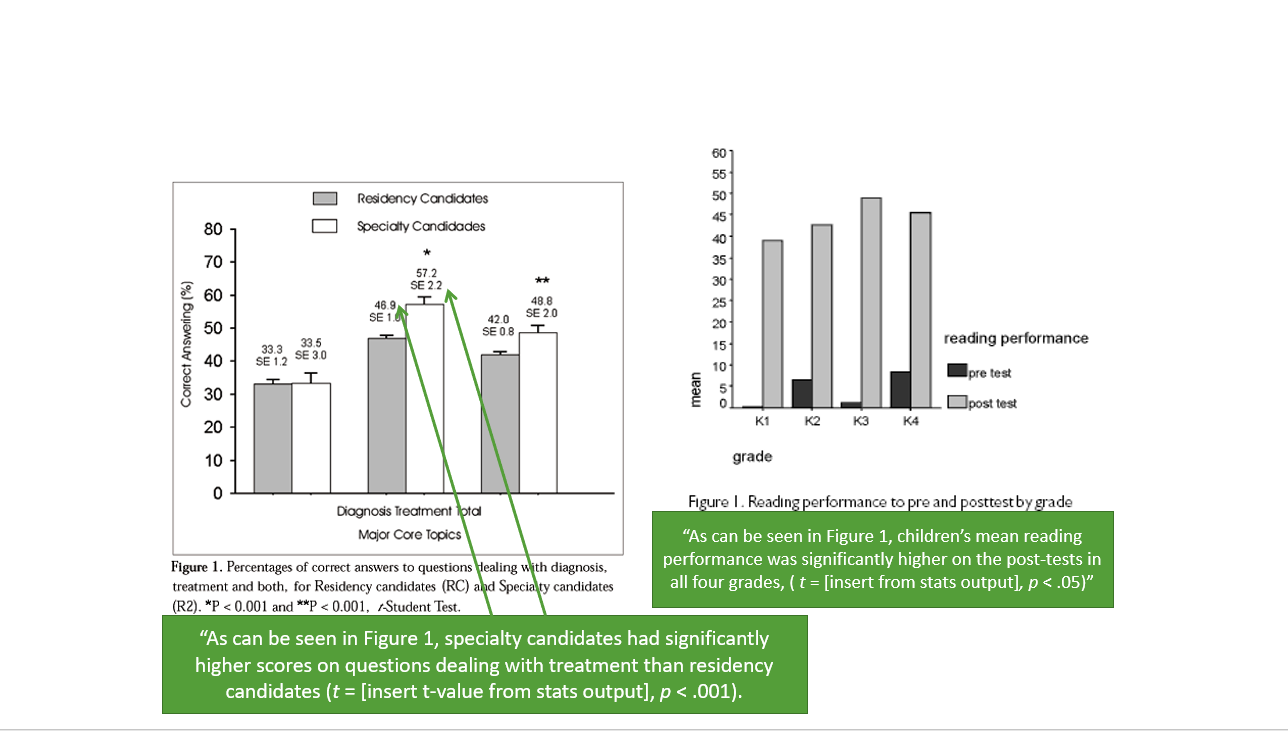
*C:\Users\ddavidov\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\F0382YT3\MC900014158[1].wmf*

Como é o resultado?

* T de Student

Como você interpreta isso?

* Olhando para o valor de p correspondente
  + Se p < 0,05 então significa que são significativamente diferentes um do outro
  + Se p > 0,05 então significa que são não significativamente diferentes um do outro

Como reportar é ilustrado no diagrama abaixo

## Análise de Dados Categóricos (Nominais)

### Qui-quadrado

Quando usá-lo?

* Quando você quer saber se há uma associação entre duas variáveis ​​categóricas (nominais) (ou seja, entre uma exposição e um resultado)
  + Exemplo: tabagismo (sim / não) e câncer de pulmão (sim / não)
  + Exemplo: obesidade (sim / não) e diabetes (sim / não)

O que um teste do qui-quadrado lhe diz?

* Se as frequências de ocorrência observadas em cada grupo são significativamente diferentes das frequências esperadas (ou seja, uma diferença de proporções)

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

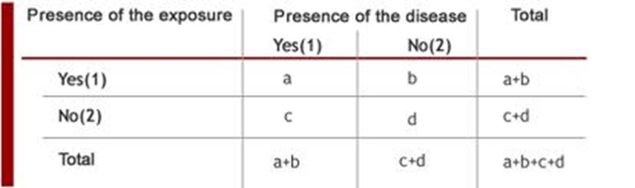
Como são os resultados?

* Estatísticas de teste do qui-quadrado = *X*2

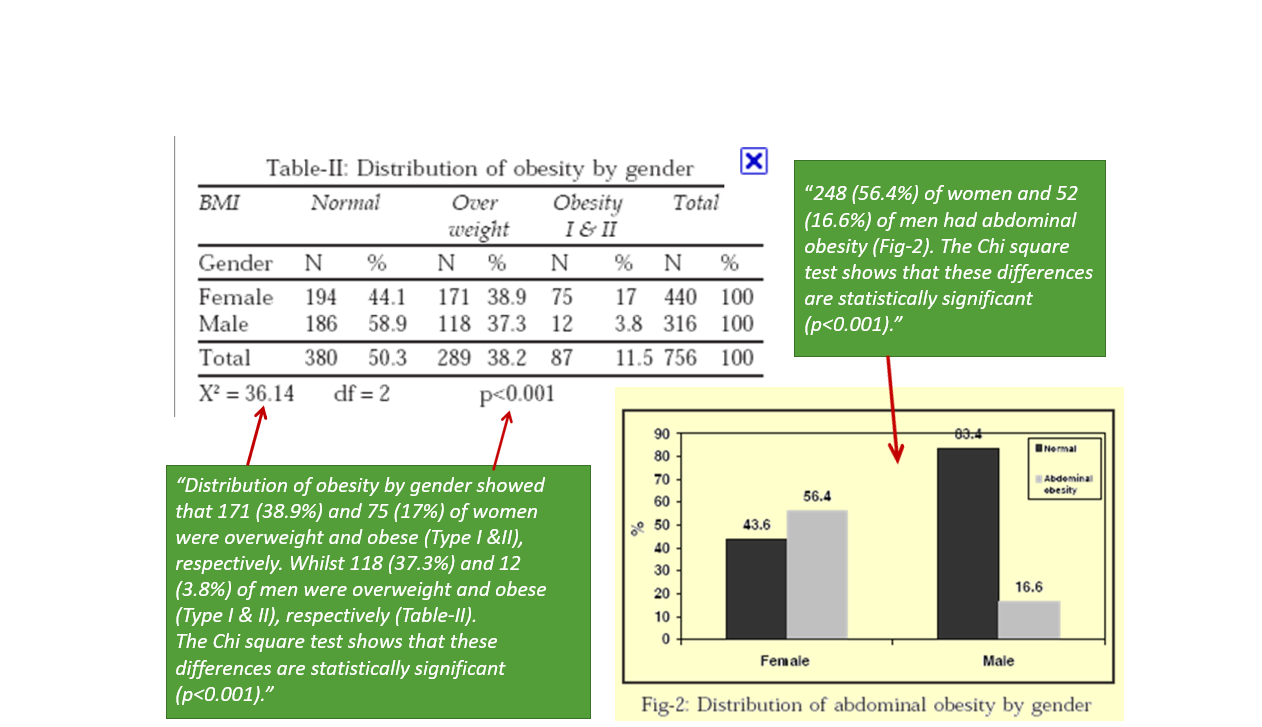
Como você interpreta isso?

* Geralmente, quanto maior a estatística do qui-quadrado, maior a probabilidade de o achado ser significativo, mas você deve observar o valor p correspondente para determinar a significância.

*Dica: O qui quadrado requer que haja 5 ou mais em cada célula de uma tabela 2x2 e 5 ou mais em 80% das células em tabelas maiores. Nenhuma célula pode ter uma contagem zero.*

Um exemplo está na tabela abaixo

Como reportar é ilustrado no diagrama abaixo (valor P):



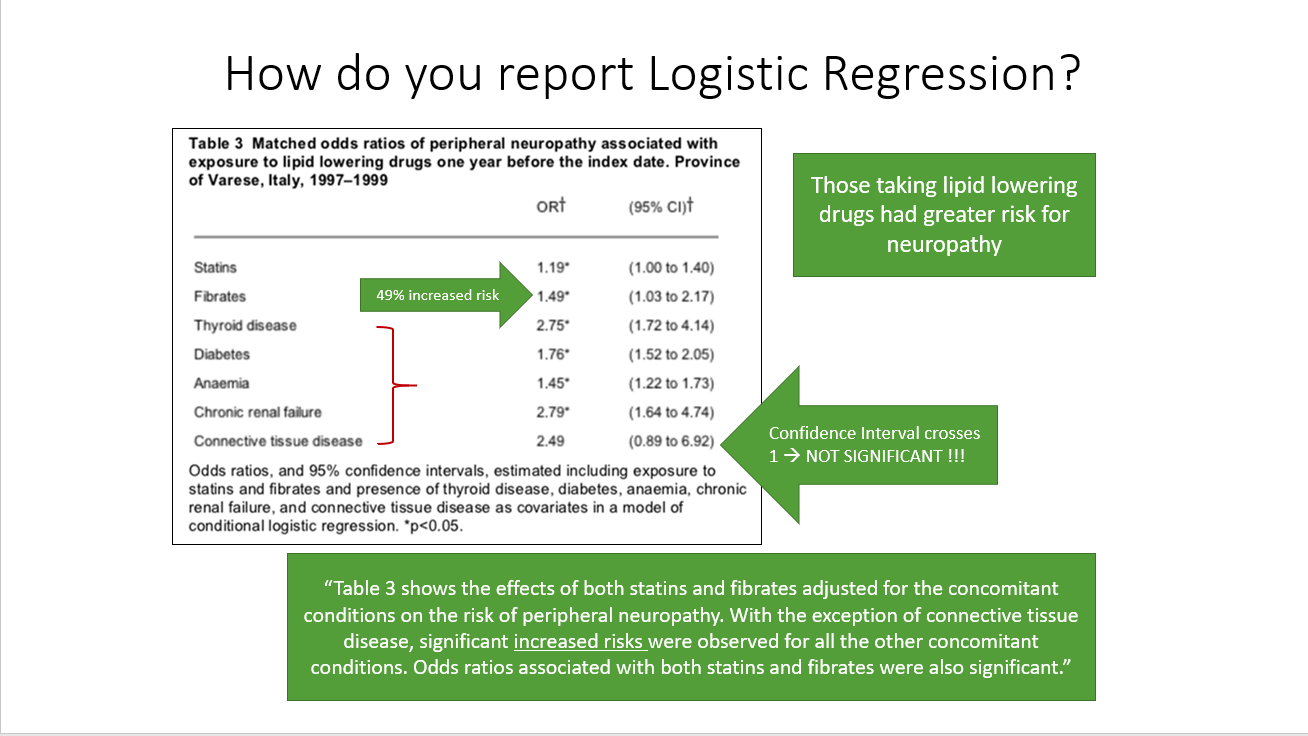
### Regressão Lógica

Quando usá-la?

* Quando você deseja medir a força e a direção da associação entre duas variáveis, em que a variável dependente ou de resultado é categórica (por exemplo, sim / não)
* Quando você deseja prever a probabilidade de um resultado enquanto controla os fatores de confusão.

Como você interpreta os resultados?

* A significância pode ser inferida usando-se os intervalos de confiança:
* Se o intervalo de confiança não ultrapassar 1 (por exemplo, 0,04 - 0,08 ou 1,50 - 3,49), então o resultado é significativo.
* Se OU> 1 🡪 O resultado é que muitas vezes é mais provável que ocorra
  + A variável independente pode ser um FATOR DE RISCO
  + 2.0 = duas vezes mais provável
* Se OU < 1 🡪 O resultado é que muitas vezes é provável que ocorra menos
  + A variável independente pode ser um FATOR PROTETOR
  + 0.50 = 50% MENOS PROBABILIDADE de experimentar o evento

Como relatar, é como ilustrado na imagem abaixo

## Resumo dos Testes Estatísticos

|  |  |  |  |
| --- | --- | --- | --- |
| **Teste Estatístico** | **Tipo de Dados Necessários** | **Estatística do teste** | **Exemplo** |
| **Correlação** | Duas variáveis contínuas | R de Pearson | A pressão arterial e o peso estão correlacionados? |
| **Teste-T/ ANOVA** | Uma variável contínua tirada de dois mais grupos | T de Student | Os pacientes com peso normal (grupo 1) têm pressão arterial menor que os obesos (grupo 2)? |
| **Qui-quadrado** | Duas variáveis categóricas | Qui-quadrado *X*2 | Indivíduos obesos (obesos versus não obesos) são significativamente mais propensos a ter um derrame (derrame vs. não derrame)? |
| **Regressão Lógica** | Uma variável dicotômica como o resultado | Odds Ratios (OR) & 95% Intervalo de Confiança (IC) | A obesidade prediz acidente vascular cerebral (derrame vs. não derrame) ao controlar outras variáveis? |

# Referências

*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.*