Chaves e Guias de uso para testes e gráficos básicos

Testes e gráficos são ferramentas de análise complementares no estabelecimento e descrição de relações. Uma tendência pode parecer significativa em um gráfico, mas o teste mostra que na realidade ela não é. Um teste pode dar um resultado significativo, mas um gráfico pode mostrar que as premissas não são válidas. Além disto, há uma forte relação entre tipos de testes e tipos de gráficos, tanto que **as chaves para testes servem também como chaves para gráficos**. Apenas por uma questão de organização, apresentaremos separadamente as explicações para o uso dos testes e para o uso de gráficos, mas o uso deve ser em conjunto.

Antes de usar qualquer teste, lembre-se da premissa fundamental de "independência entre unidades amostrais". Problemas na estrutura dos dados como "pseudorreplicações" invalidam análises. Se você não tem uma idéia clara do que é independência, leia o capítulo 5. Outras premissas serão apresentadas para os respectivos testes ao correr da chave. Abaixo das tabelas há comentários e o caminho para cada teste no Mystat12, mas provavelmente todos estes testes podem ser resolvidos em sites que os disponibilizam online. Os conceitos para correr estas chaves estão no capítulo 3.

Tabela 1: Testes com duas variáveis com unidades amostrais simples

Tabela 1. São os testes mais utilizados, pois são os mais básicos para avaliar relações entre: a) uma Variável Independente (VI) e uma Variável Dependente (VD) ou b) duas dependentes (V1 e V2) de uma terceira externa (correlação). Para simplificar, vamos nos referir apenas à situação "a", destacando diferenças quando forem relevantes. Nos testes desta tabela, em contraste com os testes da tabela 2, as unidades amostrais são simples (i.e. não são pares ou blocos divididos em subunidades).

V. dep→ / V. ind.↓	Binário (Categ. de 2)	Categórico	Ordinal ou Quantitativo Condição NP*	Quantitativo Condição P*
Binário (cat. de 2)	Teste de 2 prop., T. Exato de Fisher ou T.C. (a) Graf: Seção III	Tabela de Contingência (TC) (b) Graf: Seção III	Mann- Whitney; Cochran's TT (c) Graf: Seção IV	Teste t (de 2 grupos) (d) Graf: Seção IV
Categórico	Tabela de Contingência (b) Graf: Seção III	Tabela de Contingência (b) Graf: Seção III	Kruskal-Wallis (e) Graf: Seção IV	Análise de Variância (f) Graf: Seção IV
Ordinal	Mann- Whitney; Cochran's TT (c) Graf: Seção VI	Dicotomizar VI ou VD e usar teste apropriado (Max. Balanço) (g)	Correlação de Postos/ RNL(h) Graf: Seção V	Correlação de Postos/ RNL(h) Graf: Seção V
Quantitativo	Regressão Logística (i) Graf: Seção VI	Dicotomizar VI ou VD (Max. Balanço) (g)	Correlação de Postos/ RNL(h) Graf: Seção V	Pearson/ Regressão Linear/ RNL (j) Graf: Seção V

***OBS- V. D. Quantitativa Condição Não Paramétrica:** a) número de níveis na variável dependente entre 3 e 7, **ou** b) forte desvio da normalidade na distribuição da variável dependente (VD) para algum nível da variável independente (VI), **ou** c) forte desvio de homogeneidade de variâncias de VD em cada nível de VI (exceto Teste T). Considere a possibilidade de transformar dados quantitativos (seção 3.4) antes de utilizar a estatística não- paramétrica desta coluna. **V. D. Quantitativa Condição Paramétrica:** a) número de níveis maior que 10 **e** b) normalidade na distribuição da VD em todos os níveis da VI **e** c) Homogeneidade de Variâncias de VD em cada nível de VI ("c" é desnecessário em VI Binária). **Condições intermediárias:** Em situações intermediárias, dependerá do pesquisador assumir uma postura mais conservadora (escolhendo condição NP) ou ousada (escolhendo condição P).

a) Quando há uma variável independente e uma dependente, havendo tamanho amostral suficiente, o teste mais objetivo (logo o mais poderoso) é o de duas proporções. No Mystat12: [Analyze/ Hypotesis testing/ Proportions/ Equality of two proportions/ opção aggregate/ entre valores, escolha "not equal" para unicaudal e "larger than" (maior que) ou "smaller than" (menor que) para unicaudal conforme a hipótese]. Para usar este teste, é necessários que: [(n1 p1) > 5 **e** n1(1 − p1) > 5 **e** (n2 p2) > 5 **e** n2(1 − p2) > 5], onde n é o tamanho amostral de cada amostra e p é a proporção de cada amostra. Por exemplo, verificar se há uma relação entre o sexo do estudante (VI) e se ele sabe nadar (VD) com 20 meninos que sabem e 10 que não sabem (n1=30; p1=0,5) e nove meninas que sabem e 22 que não sabem (n2=29; p2= 0,31). Primeiro verificamos que os tamanhos amostrais e as proporções permitem o uso do teste (do contrário passaríamos a utilizar o Teste Exato de Fisher - ver próximo parágrafo). Em "sample1" entre em "number of trials" o total do primeiro nível da variável independente (30 meninos) e em "number of Successes" entre o número de ocorrências do primeiro nível da dependente (20 meninos) e em "sample2" entre o total do segundo nível da variável independente (31 meninas) e entre em "number of Successes" o número de ocorrências do primeiro nível da dependente (nove meninas). Se a hipótese alternativa fosse que os meninos nadam melhor, a escolha do "alternative type" seria "larger than". Geralmente, mais de uma probabilidade é apresentada, pois o Mystat12 usa abordagens alternativas (testes Binomiais Exatos ou testes de aproximação normal). Com tamanhos amostrais pequenos (N<30), o Mystat12 fará o teste Binomial Exato que é o melhor neste caso, e uma ou duas aproximações normais que podem ser desprezadas. Em tamanhos amostrais maiores (N≥30), o Mystat12 mostra uma ou duas formas de aproximação normal. Para simplificar, recomendo que você faça o teste conforme indicado acima e simplesmente escolha o valor de P mais conservador (maior P) que for apresentado.

Se o teste for para avaliar uma relação sem variáveis independente e dependente (VI e VD), mas com variáveis dependentes (V1 e V2) de uma terceira oculta, estaria errado se pode utilizar o teste de duas proporções e a melhor escolha é o Teste Exato de Fisher. No Mystat12: [Analyze/ Tables/ Two way/ Measures/ Fisher Exact Test].Por exemplo, o teste de associação de espécies: considerando várias unidades amostrais, V1= presença da espécie 1 [sim ou não] e V2= presença da espécie 2 [sim ou não]) (a variável oculta seria um fator ou um conjunto de fatores ambientais que afetam as duas espécies da mesma forma ou de forma oposta). O programa calcula a probabilidade do teste bicaudal (PTB). Para se obter a probabilidade do teste unicaudal (PTU), divide-se a PTB por 2 se a tendência for em direção de rejeição de H_0, do contrário use PTU= 1-(PTB/2). Tabelas de Contingência (TC) também poderiam ser utilizadas, mas são menos exatas, especialmente para tamanhos amostrais pequenos. Os gráficos para estes testes estão representados na seção III da apostila de gráficos.

b) No Mystat12: [Analyze/ Tables/ Two way/ uma variável vai em "row variable" e a outra em "column variable"]. Este teste também tem algumas limitações: a) quanto mais células na Tabela de Contingência, mais fraco é o teste. b) Se a freqüência em alguma célula for inferior a cinco, o teste é considerado suspeito pelo programa MYSTAT. Para obter mais poder e para evitar freqüências baixas nas células é recomendada a redução no número de níveis ao mínimo necessário por exclusão ou por fusão de categorias na VI e/ou na VD (se chegar a 2x2 mudar para opção de testes "a" descritos acima). Os gráficos para esta situação estão representados na seção III da apostila de gráficos.

c) Pode utilizar o Teste T se a única premissa de "Condição P" da Tabela 1 violada for a da homogeneidade de variâncias, do contrário terá de utilizar um não paramétrico. O teste Mann Whitney (MW) e o "Cochrans Test for Trend" (CTT) podem ser utilizados com variável independente binária e dependente ordinal ou quantitativa ou com variável independente ordinal ou quantitativa e dependente binária.

Quando o número de níveis na variável dependente for superior a 10, o teste mais recomendado é o Mann-Whitney (MW). No MYSTAT12: [Analyse/ Non Parametric tests/ Kruskal/ entre a variável binária em "grouping variable" e a ordinal ou quantitativa em "selected variable"], (a opção é realmente chamada "Kruskal", mas o programa detectará automaticamente que a "grouping variable" é binária e realizará o teste MW). Quando você está em dúvida se vai utilizar o Teste t ou o MW devido à premissa de normalidade, realize primeiro o Teste t, pois quando Mystat12 realiza este teste, ele mostra automaticamente um gráfico que ajuda a verificar qualitativamente a normalidade. Embora o MW seja mais robusto que um teste paramétrico, ele não é totalmente "distribution free". Em especial, é necessário cuidado com comparações com excesso de valores zero. Se o número de zeros for superior a 25% dos dados é melhor evitar o MW. A alternativa mais simples é transformar os dados da variável dependente em binários (0/1; presença/ausência) e utilizar um teste de duas proporções ou teste exato de Fisher (ver alternativa "a" acima). Outro problema ocorre quando o número de níveis da variável dependente for muito baixo.

Quando o número de níveis na variável dependente estiver entre 3 e 6, o teste mais recomendado é o CTT que no MYSTAT está em Analyze/ Tables/ Two-Way/ entra VI e y/ Marca aba Measures/ marca Cochrans test for Linear Trend. Em situações intermediárias em número de níveis, o CTT será mais conservador e o MW será mais ousado, você decide. Os gráficos para esta situação estão representados na seção IV da apostila de gráficos.

d) O Teste t no Mystat12 está em [Analyze, Hypothesis testing, mean, two sample t test]. Ao realizar o teste, um gráfico é mostrado, verifique se o número de níveis e a normalidade são apropriados para um teste paramétrico. Se não houver normalidade, considere a possibilidade de transformar os dados (seção 3.4) e repita o teste com a nova variável antes de passar para um teste não paramétrico. Utilize sempre a probabilidade da opção variâncias separadas. O Teste t admite hipóteses unicaudais (opção "alternative types"). Os gráficos para esta situação estão representados na seção IV da apostila de gráficos.

e) Para o teste Kruskal Wallis (KW) no Mystat12: [Analyze, Non Parametric tests, Kruskal] informe a variável dependente em "Selected variable", a variável independente em "grouping variable".] Detectada uma diferença estatisticamente significativa, pode se utilizar múltiplos testes Mann Whitney para o contraste (teste das diferenças entre níveis) de forma semelhante ao que se faz com o Teste Tukey em ANOVA. Embora o KW seja mais robusto que a ANOVA, ele não é totalmente "distribution free". Em especial, é necessário cuidado com comparações com excesso de valores zero. Se o número de zeros for superior a 25% dos dados é melhor evitar o KW. A alternativa mais simples é transformar os dados da variável dependente em binários (0/1; presença/ausência) e utilizar uma tabela de contingência. Também é necessária cautela quando o número de níveis na variável dependente for abaixo de 10. Uma alternativa neste caso seria realizar múltiplos CTT e corrigir o α pelo número de testes. Outra alternativa seria realizar um teste de permuta. Os gráficos para esta situação estão representados na seção IV da apostila de gráficos.

f) Análise de Variância ou ANOVA de uma via é a versão mais simples de um teste que possui muitas ramificações para uma diversidade de aplicações. No Mystat12: [Analyze, Analisys of Variance, Estimate Model]. Como outros testes com variável independente categórica, quanto maior o número de níveis mais fraco será o seu poder. Após a ANOVA, geralmente há o interesse de se

determinar quais as diferenças entre grupos (níveis da variável independente) que são significativas. Este teste chama-se contraste e é realizado pelo teste Tukey ou equivalente, que não estão disponíveis no Mystat12. Entretanto, há vários sites na internet que disponibilizam testes online, inclusive ANOVA com contrastes (e.g. http://faculty.vassar.edu/lowry//anova1u.html ou procure "ANOVA ONLINE" na internet). Os gráficos para esta situação estão representados na seção IV da apostila de gráficos.

g) Existem técnicas avançadas que permitem realizar testes nestas condições, mas como estamos nos restringindo às técnicas básicas, nossa opção é dicotomizar uma ou ambas variáveis e usar um teste apropriado considerado as escalas das novas variáveis. A escolha depende do caso. Se a variável dependente categórica puder ser reduzida a duas categorias, teríamos Mann Whitney para VI ordinal e Regressão Logística para VI Quantitativa. Se não puder, então a variável VI poderia ser dicotomizada, o que resultaria em Tabela de contingência. Dar preferência à fusão que leve à menor diferença de número de casos entre os níveis da variável independente (melhor balanço). Obs. Os gráficos para esta situação estão representados na seção VI da apostila de gráficos.

h) Se o objetivo for apenas testar a relação, pode se utilizar dois testes de Correlação de Postos: Spearman ou Kendall, no Mystat12: Analyze/ Tables/ entre VI e Y/ na aba "measures" marque Spearman e Kendall. São testes muito semelhantes, e recomendo que ambos sejam realizados e a escolha do resultado seja pelo mais conservador dos dois (maior valor de P).

Se for importante descrever a relação, então temos duas opções, um modelo *a priori* caso haja uma expectativa sobre o formato da relação (e.g. um modelo logístico), ou a partir de um modelo *a posteriori*, que pode se basear na forma dos dados com uma curva com "Smooth=LOWESS" no gráfico Scatterplot. A partir de um modelo matemático (e.g. Y= a+b*X+c*X^2), pode se determinar os coeficientes pela função NONLIN do Mystat12: Analyze/ Regression/ Nonlinear/ Loss e entra o modelo trocando VD e VI pelos nomes das variáveis (a menos que tenha muita segurança, é bom fazer isto junto a um estatístico nas primeiras vezes). Os gráficos para esta situação estão representados na seção VI da apostila de gráficos.

i) A regressão logística no Mystat12 exige VD numérico. Se esta variável estiver na planilha como categórica ("string"), como sexo$= "m" ou "f", então deve se criar uma variável binária numérica correspondente, (e.g. if sexo$="m" then let M1F2= 1- e o mesmo para fêmeas). [Analyze, Regression, Logit]. Os gráficos para esta situação estão representados na seção VI da apostila de gráficos.

j) Regressões e "correlações" retilineares*. A regressão retilinear [Analyze, Regression, Least Squares] e a correlação de Pearson [Analyze, Correlation, Simple, Pearson, Option Probabilities] apresentam o mesmo resultado (P calculado), então por que dois nomes? O nome do teste para verificar uma relação retilinear entre duas variáveis contínuas é uma questão que gera confusão. Muitos livros dividem Correlação e Regressão em dois capítulos e os autores dizem que a primeira refere-se a um estudo de associação e a segunda ao estudo de causalidade. Entretanto, a questão da causalidade está na "jurisdição" do desenho amostral (validação interna), não da análise numérica de dados (validação dados-> conclusão). Sokal & Rohlf (1988: pag. 564) explicam a questão mais profundamente. A medida de Correlação de Pearson descreve o quanto é forte a associação entre duas variáveis (seja devido a uma relação causal entre as duas ou devido a uma terceira). A regressão retilinear é um calculo de coeficientes para passar uma reta. Este cálculo da reta pode ser de dois tipos: 1) se tivermos uma variável independente com valores fixos e exatos, como normalmente ocorre em um experimento, então verificamos se a Regressão Retilinear (reta) simples ou Modelo I calculada pelo método dos mínimos quadrados é significativa e se os resíduos estão distribuídos de forma apropriada. Se estiver, verificamos P e acabou (não represente uma reta em um gráfico se P>α). Se os resíduos não estiverem apropriados (seção 3.3), conforme a situação, transformamos os dados (para obter normalidade e homocedasticidade) ou utilizamos uma regressão não retilinear (curva). Uma

regressão curvilinear pode ser obtida da forma descrita no item g. Podemos comparar estatisticamente se a relação curvilinear é significativa melhora significativamente o modelo em relação a uma regressão retilinear simples. Para isto, crie a variável XQuad (XQuad=X^2) e entre ela em regressão retilinear de mínimos quadrados com a fórmula do item g. 2). Se tivermos uma variável independente com valores aleatórios e/ou inexatos, então precisamos de uma Regressão Modelo II para determinar coeficientes mais apropriados de uma relação retilinear (reta). Há diferentes modelos, conforme o caso (o assunto é complexo e polêmico). Um modelo flexível é o "Reduced Major Axis Regression" que dá os coeficientes em Loss com a fórmula (Y-(a+b*X))^2/ABS(b). Os gráficos para esta situação estão representados na seção V da apostila de gráficos. *Considerando que uma curva é uma linha, seria mais apropriado utilizar o termo "curvilinear" para as regressões chamadas "não lineares" e o termo retilinear para as regressões chamadas de "lineares".

Tabela 2: Testes pareados ou com blocos

Estes testes são utilizados para verificar a relação entre duas variáveis, mas de uma forma "indireta". As variáveis dependentes (VD) e independentes (VI) da relação não são colunas na planilha EPR e por isto são denominadas "implícitas". Ao invés delas, utilizamos variáveis "medidas repetidas" (VMR) para "fatores intra-objeto" na analise da relação implícita. Este formato é denominado estrutura EPR longitudinal. Por exemplo, para analisar se há uma relação entre o número de baratas silvestres de serapilheira por m^2 (VD) e o período (Dia X Noite- VI), foram feitas medidas em 10 locais diferentes uma vez de dia e uma vez à noite. A Entidade (ou "Unidade Amostral" ou "Objeto") é o local, o fator intra-objeto é período, a VMR1 é número de baratas de dia e a VMR2 é número de baratas à noite. Cada m^2 é uma subunidade amostral. Para explicação da terminologia, lógica e aplicações ver seção 4.2.

Diferença entre valores quantitativos pareados com distribuição que pode ser considerada normal.	teste t pareado (k) Graf: Seção VII
Diferença entre dados ordinais pareados **ou** entre dados quantitativos pareados com distribuição das diferenças sem normalidade. Número de empates ("ties") inferior a 25% do N.	teste Wilkoxon Pareado (l) Graf: Seção VII
Diferença entre dados binários (+, -) pareados **ou** diferença entre dados ordinais ou quantitativos pareados com número de empates superior a 25% do N.	"Sign test" (m) Graf: Seção VII
Unidades amostrais com mais que duas medidas repetidas (paramétrico).	Anova de Medidas Repetidas(n) Graf: Seção VII
Unidades amostrais com mais que duas medidas repetidas (não paramétrico).	Friedman (o) Graf: Seção VII

k) Lembre que os dados precisam entrar em uma planilha em que a entidade é o par e cada variável de medida repetida é um nível do fator intra-objeto (no exemplo acima VMR1 é o número de baratas de noite e VMR2 é o número de baratas de dia). Para saber se há normalidade na diferença entre estes dois valores é necessário se calcular esta diferença da seguinte forma: DATA/ Transform/ Let/ Dif=VMR1-VMR2. Este procedimento cria a coluna das diferenças. A análise da normalidade pode ser por premissa, qualitativa ou quantitativa. Não assuma a normalidade por premissa se não tiver certeza que as diferenças são normais em situações semelhantes. A qualitativa precisa de um número de pares (N) >10 e é por um histograma: Graph/ Histogram/ entra Dif em "X-variable", que deve ter um padrão pelo menos grosseiramente normal. A quantitativa é feita com um teste, mas este tipo de teste só é confiável se N>30. Para verificar quantitativamente a normalidade da diferença use: Analyse/ Fitting distribution/ "Selected distribution"= Normal/ Entra variável Dif em "X-variable". Se o teste de normalidade apresentar P<0,05, então não há normalidade, utilize o teste Wilkoxon. Finalmente, se aceitar a normalidade, o teste pareado é feito no Mystat12: Analyze/ Hip Test/ Mean/ Paired T test/escolher opção de uma ou duas caudas (se utilizar os dados das duas variáveis) ou em

Analyze/ Hip test./ mean/ One sample T test/escolher opção de uma ou duas caudas (se utilizar as diferenças). O teste para duas caudas (opção "not equal") verifica se as diferenças são significativamente diferentes de zero. O teste para uma cauda verifica se as diferenças são maiores que zero (opção "greater than") ou se são menores que zero (opção "less than"). Os gráficos para esta situação estão representados na seção VII da apostila de gráficos.

l) Se a análise de normalidade das diferenças (ver parágrafo anterior) levar à conclusão que elas não podem ser consideradas normais, utilizamos o teste não paramétrico Wilkoxon (desde que não haja excesso de empates- ver abaixo). O Wilcoxon no Mystat12 está em: Analyze/ Non Parametric tests/ Wilkoxon/escolha a opção de caudas. O teste ranqueia os valores independentemente das colunas e verifica para duas caudas (opção "not equal") se as diferenças das posições são significativamente diferentes de zero ou para uma cauda se são maiores que zero (opção "greater than") ou se são menores que zero (opção "less than"). Este teste não é apropriado caso haja uma grande proporção de empates (>25%) entre os valores de cada entidade (unidade amostral, objeto), normalmente por excesso de valores nulos ou por número de níveis muito baixo na variável dependente implícita. Neste caso é melhor se utilizar o "sign test" (próximo teste). Os gráficos para esta situação estão representados na seção VII da apostila de gráficos.

m) Este teste é aplicável para desenhos pareados com variáveis binárias (presença/ausência; menor/maior) ou quando há excesso de empates entre os valores (ver parágrafo anterior). Lembre-se que a unidade é o par. Os valores das variáveis binárias devem ser 0 ou 1 que significam presença/ ausência ou maior/menor dentro de cada par. No caso de empate, coloque 0 e 0 para as duas variáveis da entidade. No caso de variáveis quantitativas, o próprio programa transformará os valores em 0 ou 1 para menor/maior. No Mystat: Analyze/ Non Param Test/ Sign. Se você ainda não tiver os dados entrados na planilha, o mais fácil é contar os sinais + e – das diferenças e ir para o teste Binomial (= teste de 1 proporção) Mystat12: Analyze/ Hyp Test./ Propor/ Simple Prop. e entrar opção "aggregate"; N em "number of trials"; o número de positivos em successes; Proportion=0.5; e a alternativa se será unicaudal ou bicaudal. Os gráficos para esta situação estão representados na seção III da apostila de gráficos, mas sem representação da relação pareada.)

n) A Análise de Variância de Medidas Repetidas (RM Anova) é semelhante ao teste t pareado, mas ao invés de duas medidas por entidade temos três ou mais. Por ser mais complexo, é importante uma consulta à seção 4.2 para entender bem a terminologia, lógica, premissas e aplicações. Lembre que os dados devem estar na forma "longitudinal", isto é, cada entidade medida é uma linha e as diferentes medidas dela estão em colunas. No MYSTAT12 Analyse/ Analysis of Variance/ entrar todas as variáveis de medidas repetidas na variável dependente/ Na aba Repeated Measures marcar "Perform Repeated Measures analysis" e o número de níveis em Level. Os gráficos para esta situação estão representados na seção VII da apostila de gráficos.

o) Quando as premissas de testes paramétricos não permitem uma RM Anova, a opção não paramétrica é o teste Friedman. Há duas estruturas EPR que permitem o teste Friedman, na forma longitudinal (como para RM Anova) e em um formato com VI e VD explícitas. No formato longitudinal, o caminho no MYSTAT12 é Analyse/ Non Parametric tests/ Friedman / entrar todas as variáveis de medidas repetidas em "Selected Variables". Outra forma é utilizando-se colunas com a Variável Independente, a Variável Dependente e uma variável identificando os blocos. No exemplo na legenda da tabela acima seriam Período, Número de Baratas e Local. Neste formato, entre em Analyse/ Non Parametric tests/ Friedman / VD em "Selected variables", VI em Grouping Variable e a variável dos blocos em Blocking Variable"). Uma alternativa ao Friedman é o teste "Quade" [Analyze, Non Parametric tests, Quade] veja o "Help" do Mystat12 para mais informações. Os gráficos para esta situação estão representados na seção VII da apostila de gráficos.

Tabela 3: Testes com uma variável

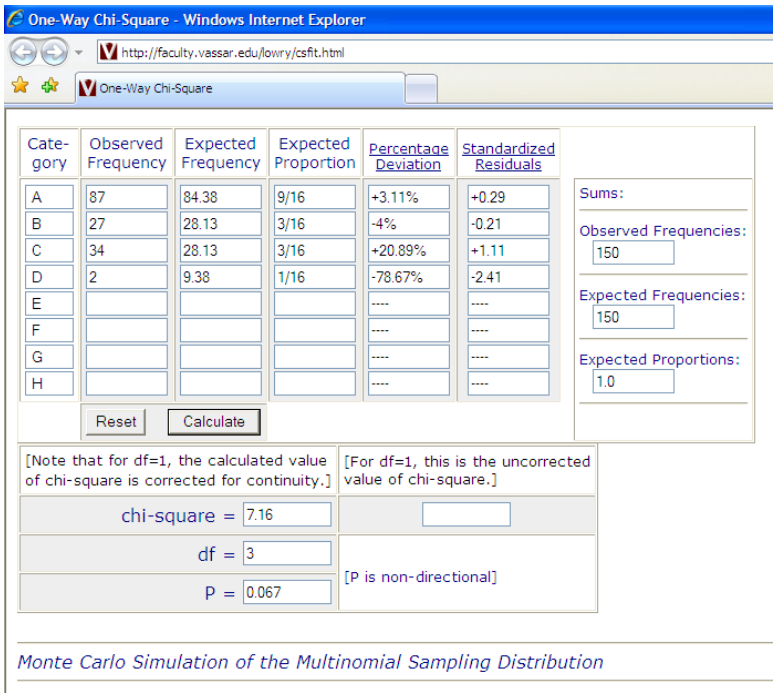
Tabela 3- Estes testes são testes chamados de testes de aderência ("goodness of fit") porque verificam como uma variável se ajusta a uma condição pré definida. Exemplos: a razão sexual está dentro do esperado (50%)? As freqüências observadas de tons de vermelho em rosas estão dentro da razão esperada de alelos pela segunda a lei de Mendel (9:3:3:1)? Este crânio fóssil único é significativamente maior que as medidas anteriores de vários crânios de outra localidade? Estas medidas de mercúrio estão significativamente maiores do que a média recomendada pelo governo? A distribuição de freqüências de tamanhos de peixes mudou este ano comparada com as distribuições de tamanhos de 30-50 anos atrás? Estes testes são geralmente denominados "testes de uma amostra", mas a denominação "testes com uma única variável" é mais apropriada na abordagem EPR (Entidade-Propriedade-Relação) adotada neste curso. (OBS- estes testes normalmente não "pedem" gráficos, mas, se necessário, algumas destas situações podem ser representadas conforme as seções I e II da apostila de gráficos.

Uma amostra com medidas **Binárias** têm freqüências compatíveis com freqüências teóricas esperadas? (Aderência de proporções).	Binomial exato; Teste z aprox. (p) Graf: Seção I
Uma amostra com medidas **Categóricas** são compatíveis com freqüências teóricas categóricas esperadas? (Aderência de proporções).	Qui2 ou teste g (q) Graf: Seção I
Um valor **quantitativo** é compatível com uma população de valores uma com distribuição normal com média e desvio padrão conhecidos? (Aderência de valor a uma média de pop. com distribuição normal- DN)	Teste z para um valor (r) (Sem gráfico)
Uma amostra com valores **quantitativos** é compatível com uma distribuição normal com média e desvio padrão conhecidos? (Aderência de média em pop. com DN)	Teste z para uma amostra (r) Graf: Seção II
Uma amostra com valores **quantitativos** é compatível com uma população de valores com distribuição normal com média e desvio padrão desconhecidos? (Aderência de média em pop. DN)	Teste t para uma amostra (s) Graf: Seção II
A distribuição observada de uma variável com medidas **ordinais ou quantitativas** é compatível com freqüências teóricas esperadas? (Aderência de valores a modelos)	Kolmogorov Smirnov (KS) (t) Graf: Seção II
A distribuição observada de uma variável com medidas **ordinais ou quantitativas** é compatível com uma curva normal? (Aderência de valores à normalidade)	Shapiro-Wilkes; KS(u) Graf: Seção II

p) Variável binária. Verifique a proporção esperada, o número de casos e o número de "sucessos". "Sucesso" é uma das possibilidades em questão que foi escolhida para ter a proporção analisada, por exemplo, "número 6" ao se jogar 1 dado de 6 faces, ou "macho", para se ver se a proporção de machos difere de 50%. Entre os valores em Mystat12 [Analyze/ Hip. Test/ Prop./ Single Proportion/Aggredate; escolher "Alternative Type" para bicaudal ou unicaudal]. A condição de tamanho amostral mínimo para o teste é [n .p0 > 10 e n (1 − p0) > 10], onde n é o tamanho amostral e p0 é a proporção teórica. Geralmente, mais de uma probabilidade é apresentada, pois o Mystat12 usa abordagens alternativas (testes Binomiais Exatos e testes de aproximação normal). Com tamanhos amostrais pequenos (N<30), o Mystat12 fará o teste Binomial Exato que é o melhor neste caso, e uma ou duas aproximações normais que devem ser desprezadas. Em tamanhos amostrais maiores (N≥30), o Mystat12 mostra uma ou duas formas de aproximação normal. Para simplificar, recomendo que você simplesmente escolha a abordagem que tiver o valor de P mais conservador (maior P). Por exemplo, uma pessoa disse que tem uma técnica para distinguir machos de fêmeas de pintinhos com um mês, o que é importante para granjas. De 50 pintinhos ele acertou 39 com a técnica, mas esta proporção de acertos é significativamente maior que 50% (acaso)? Entre os dados: "Number of trials"=50; "Number of successes"=39; "Proportion"=0.5; Alternative type: "Greater than". Explicando: é proporção

simples (“single proportion”) porque estamos comparando uma proporção obtida com uma esperada; o número de casos total (“number of cases”) é 50, o número de casos favoráveis (number of successes) é 39 e a chance unitária de sucesso (“chance de acertar o sexo chutando”=0.5- “Proportion”). A pergunta é unicaudal porque você quer saber se ele acerta “mais que 50%” e não “diferente de 50%” (alternative type). Entretanto, se você precisar de uma taxa de acerto igual ou superior a 75%, para valer a pena descartar os filhotes machos, mude “Proportion” para 0.75.

q) O teste de Qui2 de uma via com três ou mais categorias no Mystat12 pode ser feito apenas se a hipótese nula tiver proporções homogêneas (e.g. 25% em cada uma de 4 categorias): Analyze/ One way freq. tab. Se as proporções não forem homogêneas (e.g. a proporção 9:3:3:1 da 2ª lei de Mendel), o mais prático é se fazer o teste online disponível em alguns sites (e.g. http://faculty.vassar.edu/lowry/csfit.html- Acessado em Dezembro de 2010). Neste site entre proporções esperadas em Expected proportions assim: 9/16; 3/16, etc. e as observadas em “Observed Frequency” e depois pressione “calculate”. Se não estiver online, use o arquivo que está no pacote estatístico em “Programas/ Testes no Excel/ A1_Qui2_Prop_ hetero.xls”. Modifique o exemplo com os seus dados. O teste G é uma alternativa ao Qui2 recomendada em alguns livros, mas não é muito diferente, de forma que não a abordaremos. Estes testes não admitem hipóteses unicaudais.



r) Quando estamos querendo verificar se um valor está significativamente diferente do esperado para uma média e um desvio padrão pré-determinados ou “conhecidos”*, então utilizamos o teste Z. No Mystat12 entre em Utilities/ Probability Calculator/ Continous/ entre a média em “Location or mean”, o desvio padrão da população em “Scale or SD “e o valor que será testado em “Input Value” Por exemplo, se a distribuição de tamanho de ratos de laboratório é bem conhecida (média e desvio padrão definidos) e você quer saber se o peso de um determinado rato difere deste valor, então você pode utilizar o teste Z para um valor.

Quando estamos querendo comparar uma amostra com valores pré-determinados de média e desvio padrão, utilizamos o teste Z para uma amostra no Mystat12 [Analyse/ Hip. Test./ Mean/ One Sample Z test./ entrar a variável e os valores de média e desvio padrão pré determinados]. Por exemplo, se você recebe um lote de ratos e quer saber se eles estão dentro da média e desvio padrão conhecidos para ratos de laboratório, então você utiliza o teste Z para uma amostra.

*OBS1- É questionável o que podemos chamar de “média e desvio padrão conhecidos”; quase sempre estes valores foram determinados a partir de amostras. Levine sugere que quando o tamanho amostral para estes valores for superior a 200, pode se usar o teste Z, do contrário é melhor utilizar um teste t.

s) Quando estamos querendo verificar se uma amostra está significativamente diferente de uma média pré-definida, mas não temos um valor pré-definido para o desvio padrão ou quando queremos verificar se um valor está significativamente diferente dos valores de uma amostra considerada referência utilizamos o Teste t para uma amostra. No Mystat: Analyse/ Hip. Test./ Mean/ One Sample

t test./ entrar a variável e o valor de média a ser comparada]. Note que nos dois casos o desvio padrão é estimado pela amostra. Por exemplo, há padrões de valor máximo de mercúrio, mínimo de Oxigênio e ideais de pH para lagos em uma legislação estadual. Entretanto, você fez medidas distribuídas em uma área protegida (30 amostras) e percebeu que o valor das medidas era significativamente maior de mercúrio, menor de Oxigênio e diferente do ideal de pH. Com base nisto, você contesta a norma para cada variável, pois os valores de referência deveriam refletir os valores locais em ambientes protegidos. Note que o teste será unicaudal para mercúrio e oxigênio e será bicaudal para pH.

t) e u) O teste de Kolmogorov Smirnov (KS) para uma única variável depende de uma distribuição esperada para ser comparada com a obtida. O Mystat12 oferece diversas distribuições teóricas para esta comparação: Uniforme, Uniforme Discreta, Poison, Lognormal e Normal. Portanto, o KS pode ser utilizado como teste de normalidade, mas não é o único Shapiro-Wilkes e Anderson Darling são outros testes mais usados. No Mystat12 [Analyse/ Fitting distributions/ KS]. Quando se realiza testes, como em ANOVA, há opções para testar-se a normalidade.

Tabela 4: Testes com 1 Variável dependente e 2 variáveis independentes

Tabela 4. A inclusão de duas variáveis em um estudo experimental possibilita a avaliação de interações entre variáveis independentes. Se o estudo não for experimental, então a inclusão de uma variável pode servir principalmente para um controle. Como vimos acima, a análise de uma única variável independente já envolve diversas premissas. A situação se complica com a segunda variável dependente. Continua valendo a normalidade e a homogeneidade de variâncias. Adicionalmente, é necessário que haja independência entre as variáveis independentes. É fácil realizar estes testes, e ai mora o perigo. Um treino nestas técnicas está além do escopo deste curso; mas passamos as os caminhos no MYSTATpara quem quiser começar a aprender a usá-las.

Var. Dependente	V. Ind. 1	V. Ind. 2	Teste
Categórico	Categórico	Categórico	Tabela de Contingência de 3 vias (v)
Binário	Quantitativo	Quantitativo	Regressão logística múltipla (w)
Quantitativo	Categórico	Categórico	Análise de Variância de 2+ Vias (x)
Quantitativo	Quantitativo	Binário	ANCOVA (y)
Quantitativo	Quantitativo	Quantitativo	Regressão Múltipla (z)

V) A tabela de contingência de três vias no Mystat12 está em "Analyze/ Tables/ Multiway"

W) A regressão múltipla está em "Analyze/ Regression/ Logit/ Model"

X) A ANOVA de duas vias está em "Analyze/ ANOVA/ com X1 e X2 em factor

Y) A ANCOVA de duas vias está em "Analyze/ ANOVA/"com a X1 quantitativo em covariate e X2 categórico em factor.

Z) A ANOVA de duas vias está em "Analyze/ ANOVA/ com X1 e X2 em covariate.

O Uso de Gráficos

Gráficos são parte essencial da análise de dados e da comunicação científica. A aplicação de testes estatísticos deve ser feita juntamente com análises gráficas para se verificar se premissas não foram quebradas, e os gráficos permitem se apreciar a forma de uma relação melhor que números em um texto. Os gráficos prendem a atenção dos leitores e podem transmitir eficientemente idéias complexas. Entretanto, um número elevado de gráficos pode dispersar a atenção do leitor e o número permitido normalmente é limitado na hora de publicar, de forma que devem ser escolhidos só os mais relevantes e devem ser preparados com um cuidado especial. Gráficos mal feitos ou desnecessários desvalorizam muito uma publicação. Portanto, o domínio dos gráficos básicos é pré requisito para a autonomia na pesquisa e divulgação de suas descobertas.

Seção I- Gráficos com uma variável única categórica

Gráficos com uma variável nominal podem ser construídos com uma variável de um arquivo EPR (Entidade-Propriedade-Relação) ou com um Arquivo Síntese de EPR com uma coluna com os níveis de X e outra com a freqüência fornecida. Por exemplo, se nós temos um arquivo no qual a entidade é "pessoa" e há uma coluna "cidade de origem", nós podemos criar um gráfico com o eixo X com a lista das cidades deste arquivo e **o computador contará automaticamente** quantas há em cada cidade e construirá o gráfico da figura 1. Esta é a abordagem mais prática se já tivermos a planilha EPR montada. Entretanto, se tivermos os dados de freqüência de cada cidade, montar a planilha EPR pode ser trabalhoso (imagine entrar 100 linhas com o valor "Manaus" para 100 pessoas). Neste caso é mais prático criamos uma tabela síntese com cada nível de X em uma linha (e.g. o nome da cidade) e uma coluna com a quantidade de pessoas de cada cidade (100 para Manaus; 34 para Parintins, etc) este valor de freqüência é **inserido por nós na planilha**.

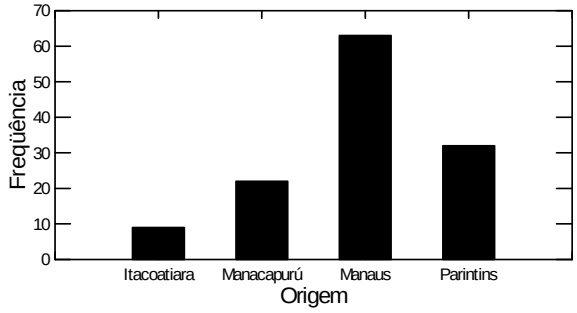


Figura 1- Gráfico de Barras Simples. Construído a partir de um arquivo com entidade= pessoa e propriedade= cidade, ou a partir de um arquivo Síntese de EPR com uma coluna com os nomes das cidades e outra coluna com os valores de freqüência.

Um arquivo síntese de EPR pode ser um novo arquivo EPR, entretanto, a entidade será outra. Neste exemplo, o arquivo EPR básico tinha pessoa como entidade e o arquivo síntese tem "local" como entidade e duas propriedades: "cidade de origem" e "freqüência de pessoas com esta cidade de origem". Embora seja possível se considerar esta freqüência como propriedade do local, nem sempre as planilhas síntese de EPR são arquivos cujas linhas são entidades com um sentido relevante e cada nível da "variável independente" aparece apenas uma vez, o que contrasta com as planilhas usadas para relações que tem replicatas de cada nível de "variáveis independentes" categóricas. É por isto que preferimos chamar planilhas nestas condições de planilhas síntese de EPR. A situação muda se outros dados são adicionados a cada entidade da planilha síntese e passa-se a utilizar cada linha da planilha como uma replicata em uma análise de relação entre "variáveis independentes" e "dependentes", incluindo variáveis de contagem.

OBS: Quando se trabalha com freqüências, deve-se apresentar o valor 0 (zero) no eixo Y do gráfico de barras. Estes dados também podem ser apresentados no formato de setores ("pizza"), mas este formato é considerado menos efetivo em geral.

MYSTAT: a) Barras simples com Arquivo EPR: Graph/ Bar Chart/ Variável→ Xvariable;/ colocar 0 (zero) no Ymin na aba Yaxis. b) Barras simples com arquivo Síntese de EPR: Graph/ Bar Chart/ Variável→ Xvariable;/ Freqüência em Yvariable/ colocar 0 (zero) no Ymin na aba Yaxis. c) Setores com Arquivo EPR: Graph/ Pie Chart/ Variável→ Xvariable;/ colocar 0 (zero) no Ymin na aba Yaxis. d) Setores com arquivo Síntese de EPR: Graph/ Pie Chart /Variável → Xvariable;/ Freqüência em Yvariable/ colocar 0 (zero) no Ymin na aba Yaxis.YSTAT: GRAPH/ Pie Chart/ X-> Xvariable)

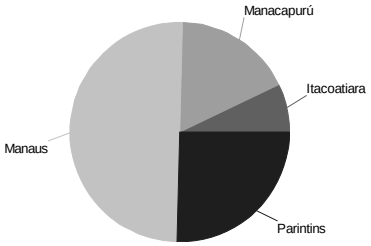


Figura 2. Mesmos dados em um Gráfico de Setores

Seção II- Gráficos com uma variável única quantitativa

Quando a variável é quantitativa, os dados são divididos em intervalos iguais e a freqüência é calculada para cada intervalo. O número de intervalos é definido pelo usuário em função do total de casos, normalmente entre 12 e 20, evitando muitos intervalos com freqüência de 1 (um) caso e evitando intervalos com valores "quebrados" (e. g. 4,256). Se o tamanho amostral não for muito pequeno, geralmente o número de intervalos está entre 7 e 15.

MYSTAT: Graph/ Histogram/ Variável→ Xvariable/ Options, Number of bars=7 (ou outro valor considerado apropriado.)

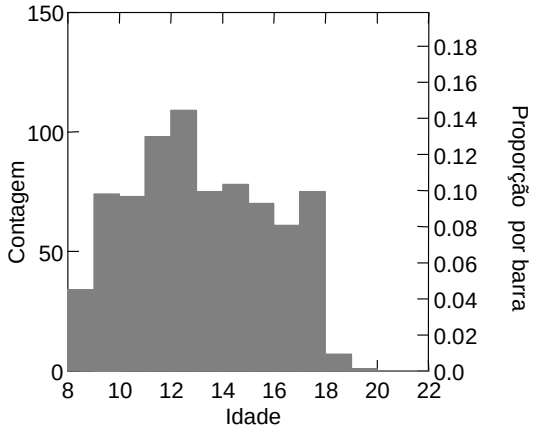


Figura 3- Histograma mostrando a quantidade de pessoas por intervalo de idade. O intervalo utilizado foi 1 ano, mas poderia ter sido outro.

Seção III- Gráficos com "variáveis dependentes" e "independentes" nominais

Nas seções anteriores, havia uma única variável em análise. A partir desta seção estamos lidando com análises de relações entre uma variável independente e uma dependente (ou duas dependentes de uma terceira). Esta é a única seção que lida com relações em que não utilizamos gráficos EPR, e vamos começar explicando o porquê. Os gráficos mais ricos em informação são os gráficos EPR nos quais os eixos X e Y são variáveis (colunas da planilha EPR) e cada ponto é uma entidade. Quando as "variáveis dependentes" e "independentes" são nominais (binárias, categóricas, ordinais ou quantitativas tratadas como nominais), o gráfico EPR seria na forma da figura 4.

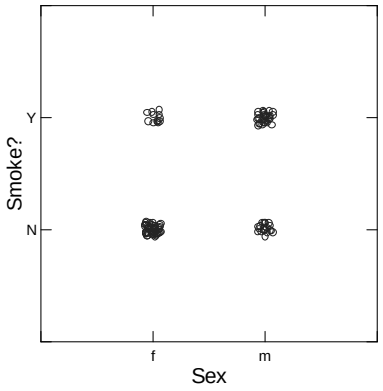


Figura 4- O que seria um gráfico EPR para variáveis nominais. Ao invés deste gráfico, utilizamos tabelas ou outro tipo de gráfico.

Obviamente é uma representação ruim devido à sobreposição dos pontos. Ao invés das nuvens de pontos poderíamos ter o número de dados, e então o gráfico se transforma em uma tabela (figura 5a). Esta tabela pode ser interpretada mais facilmente se forem utilizados os valores porcentuais por linha (se a "variável independente" tiver os níveis em linhas) ou por coluna (se a "variável independente" tiver os níveis como colunas) como na figura 5b. Estas tabelas podem ser construídas com a ferramenta tabela dinâmica do Excel ou com tabelas de contingência de 2 vias no MYSTAT. Se estiver com uma tabela síntese de EPR no MYSTAT, deve-se marcar a freqüência como frequency em [Data/ Case Weighting/ by frequency].

a)

Sex\ Smoke?	No	Yes
Females	50	12
Males	20	30

b)

Sex\ Smoke?	No	Yes
Females	78%	22%
Males	43%	57%

Figura 5- Tabelas que apresentam os dados da figura 4. Na primeira são apresentados os valores absolutos das contagens e na segunda a porcentagem dos valores por linha, pois a variável sexo é independente e está com seus níveis em linha

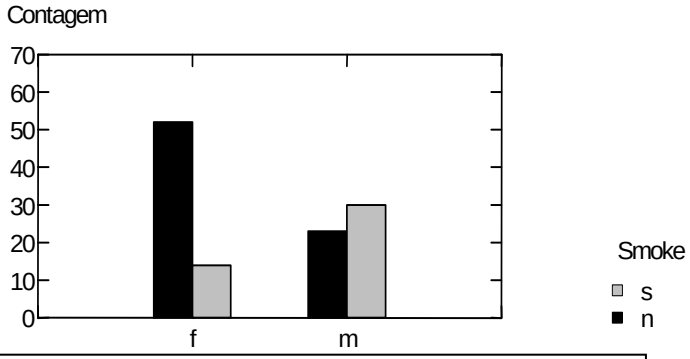


Figura 6- Gráfico de Barras Composto Horizontalmente.

Há uma alternativa gráfica para esta situação, a utilização de um gráfico de Barras Composto (fig. 6). Nesta situação, este gráfico preserva toda a informação (permite a reconstrução da planilha) e permite a apreciação da relação pelo contraste das proporções de colunas pretas e cinzas dentro de cada sexo. Da mesma forma que foi discutido na Seção I, este gráfico pode ser construído com uma tabela síntese de EPR. Neste caso, cada linha é uma combinação diferente dos níveis das "Variáveis Dependentes" e "Independentes" e há uma freqüência para cada combinação (e.g. L1- homem, fuma, 30, L2 homem, não fuma, 20, L3- mulher, fuma, 12, L4- mulher, não fuma, 50).

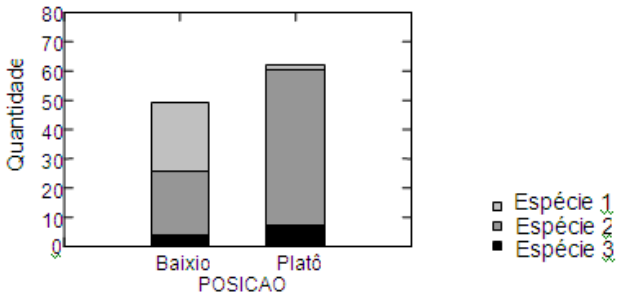


Figura 7- Gráfico de Barras Composto Verticalmente.

Outra opção é o gráfico de Barras Composto Vertical (BCV) no qual as barras em cada nível de X são sobrepostas. Para apresentar barras sobrepostas, é necessária a construção de uma planilha Síntese de EPR de outra forma. A primeira coluna tem em cada linha um nível da "variável independente" e as outras colunas são os níveis da "variável dependente". (e.g. Colunas= sexo/ fumam/ não fumam; L1- homem, mulher/ L2 30, 12/ L3 20, 50). O gráfico de Barras Composto Vertical é menos efetivo que o BCH quando as proporções de cada grupo totalizam 100%, mas pode ser superior quando os totais de cada grupo são diferentes, como na figura 7.

OBS- É importante considerar se a apresentação de um gráfico ou uma tabela se justifica nesta situação. No exemplo do estudo sobre o fumo, bastaria se informar que "57% dos homens e 22% das mulheres eram fumantes (n=50 e 62 respectivamente)". Normalmente não se justifica um gráfico ou uma tabela para uma informação que poderia ser apresentada em uma ou duas linhas, a menos que seja um dos resultados mais importantes de todo o estudo, para destacá-lo.

MYSTAT: Arquivo EPR: Gráfico de Barras Composto Horizontal (BCH) Graph/ Bar Chart / VI→ Xvariable/ VD→ Grouping Variable/ option Overlay Multiple Graphs. Arquivo Síntese de EPR: o mesmo, mas coloque **a freqüência** em Y variable. Para Barras sobrepostas (BCV), ver como construir síntese de EPR no texto. Graph/ Bar Chart/ VI→ X variable/ Variáveis de cada nível 1 da VD → Y Variable/ Opção Stackbars.

Seção IV- Gráficos com "variáveis dependentes" e "independentes" nominais

O melhor gráfico nesta situação normalmente é o Gráfico de Dispersão Categórico Normal ou "Dot Density" normal, pois, como vemos na figura abaixo, é o único que mostra toda a informação. Com base neste gráfico podemos ver o número de entidades em cada nível da "Variável Independente" (VI) (e se há ou não balanço), a média, a amplitude, a normalidade e se há homogeneidade de variâncias entre os níveis da VI, que são informações essenciais para uma avaliação estatística de dados. O gráfico de barras é mais comum em publicações talvez por desconhecimento, pois o "Dot density" está disponível em poucos programas aplicativos de estatística. Outra justificativa para os outros gráficos poderia ser "para se apresentar gráficos mais limpos". Entretanto, esta "limpeza" nos impede de distinguir entre situações ideais e situações problemáticas, pois podem estar escondidos "outliers", desbalanço, falta de normalidade, etc. Não se pode menosprezar a importância desta informação antes de optar por uma outra alternativa, e a justificativa deve ser pela qualidade da comunicação e não pela conveniência de se esconder uma situação fora do ideal.

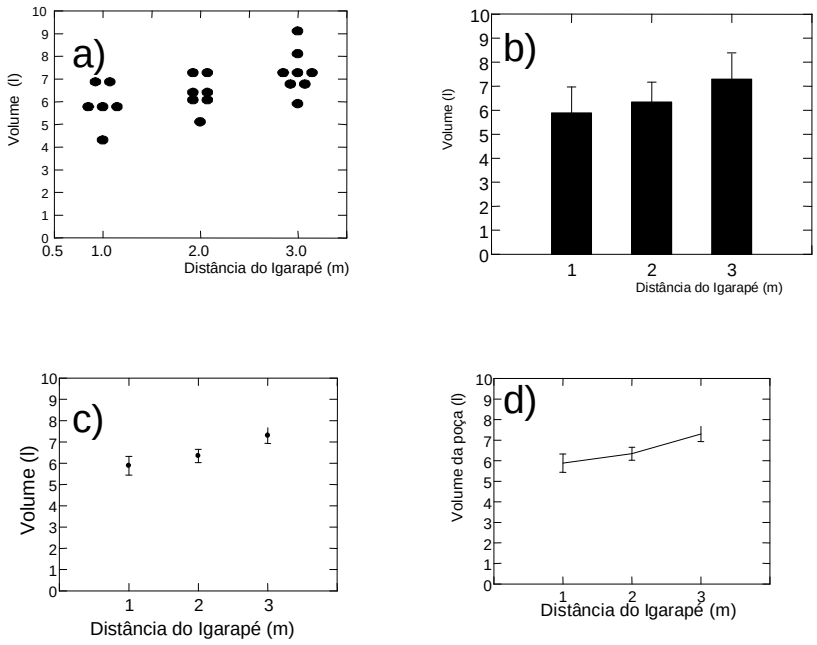


Figura 8- Quatro formas de se apresentar dados nominais (neste caso quantitativos tratados como categóricos): a) "Dot Density"; b) Barra simples V; c) Dot; d) Line. The error bars represent the standard error.

O gráfico de Barras Simples é preferível ao Dot Density se houver apenas um dado para cada nível de X, por exemplo, em um gráfico de precipitação mensal ao longo dos meses durante um ano, pois as barras são mais visíveis do que um ponto. Note que, diferente do que foi apresentado na seção I, o eixo Y representa uma "variável dependente". Nos casos em que há mais de um valor por nível da "variável independente", é apresentado um valor médio e podemos utilizar uma barra de erro padrão ou de desvio padrão. O desvio padrão é uma medida de variação importante para caracterizar uma população com distribuição normal, mas não ajuda a ajuda a sabermos se a média da amostra está próxima da média real sem o dado do tamanho da amostra. A barra de erro padrão deve ser preferida de forma geral, pois é um índice do intervalo de confiança da média, o que é útil para termos uma idéia se há diferença estatística entre os níveis de X, mesmo para populações que não tem distribuição normal.

O gráfico de “pontos médios” (ou “Dot”) é semelhante ao gráfico de barra, mas ao invés da barra usa um ponto apenas e pode ter barras de erro. Em situações com muitos níveis da “variável independente” e/ou subgrupos em cada nível que tornariam o gráfico muito complexo para um gráfico de barras ou um “Dot Density”, o gráfico “Dot” se justifica por reduzir a complexidade do conjunto para dar ênfase às diferenças entre determinados grupos.

MYSTAT: Dot Density: Graph/ Dot Density / VI→ X variable; VD→ Y variable/ em Type of Display o recomendado é o “Symmetrical Dot Density”; No **Barra Simples**: Graph/ Bar Chart / VI→ Xvariable; VD→ Y variable; No gráfico de Pontos Médios ou “DOT”: Graph/ Summary Charts/ Dot / VI→ X variable; VD→ Y variable; No gráfico de linha Graph/ Line Chart / VI→ X variable; VD→ Y variable. Nos três ultimos tipos de gráfico, pode-se incluir uma barra de erro padrão na aba error bar/ standard error. A sobreposição dos gráficos de “Dot density” e de linha pode dar bons resultados. Para isto deve se fazer cada um dos gráficos com as mesmas escalas e todas as opções definidas *a priori*. Depois entra-se em Graph/ Begin Overlay Mode/ Graph/ Dot Density/ OK/ Graph/ Line Chart/ OK/ End Overlay Mode.

O Gráfico de Linha tem uma linha que liga valores únicos ou médios de cada nível de X e é útil para destacar mudanças espaciais ou temporais. Deve-se evitar sua utilização para variáveis nominais em geral (binários, categóricos ou ordinais), pois não há continuidade entre categorias, mas em alguns casos isto se justifica (e.g. Seção VII). A princípio, é necessário que haja unidades equivalentes entre níveis de um X contínuo para usá-lo. Pode se ligar valores com de uma “variável independente” como mês, entretanto, os níveis de mês devem estar distanciados de forma apropriada, não se pode colocar os meses de janeiro, fevereiro e outubro eqüidistantes e ligá-los com uma linha porque fevereiro está próximo de janeiro e distante de outubro. Em um caso destes, podemos usar valores de dias para a posição do mês (janeiro=15, fevereiro=45...) e em [Data/Value labels] informar que 15= Janeiro, etc. de forma que cada mês fique em sua posição correta. Esta regra não se aplica obrigatoriamente quando se usa os outros gráficos nesta situação, mas a mesma diretriz pode ser seguida para transmitir uma informação de forma mais clara (fig. 10).

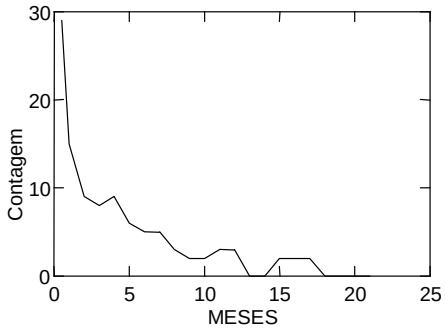
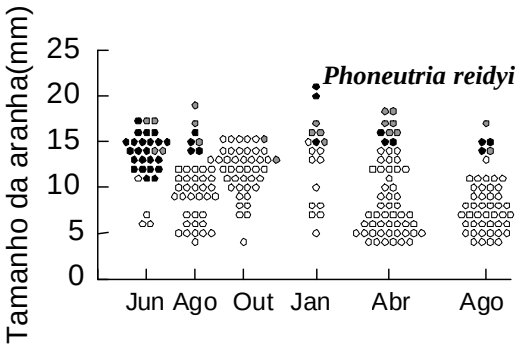


Figura 9- O gráfico de linha é uma boa opção para mostrar variações no tempo e no espaço, mas possui regras mais estritas que os anteriores.

Figura 10. Variação nos tamanhos de aranhas ao longo de 14 meses. O inclusão de distâncias informativas entre níveis categóricos da variável Independente (e.g. meses com distâncias proporcionais a diferenças de dias) é uma regra do gráfico de Linha ("Line") que pode ser aplicada ao "Dot Density", como neste exemplo, e aos outros gráficos desta seção. Para isto, os gráficos devem ser construídos com os valores em dias e depois os nomes dos meses podem ser ajustados em um processador de textos como o Word.



Outra alternativa para esta situação é o Box Plot (fig. 11). Este gráfico é recomendado para situações em que as distribuições dentro de cada nível da "variável independente" não seguem uma distribuição normal (em forma de sino). Neste gráfico a linha central em cada nível de X é a mediana e as outras linhas marcam os limites dos "quartis" (cada grupo de 25% dos dados mais próximos e mais distantes da mediana). É superior ao gráfico de Barras nesta situação, mas é inferior ao "Dot Density", pois não mostra qual a distribuição dos dados. Portanto, não é um gráfico recomendável.

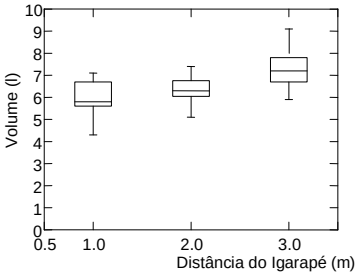


Figura 11- Box Plot para os mesmos dados da figura 8.

Seção V- Gráficos com "variáveis dependentes" e "independentes" quantitativas

O melhor gráfico nesta situação normalmente é o Gráfico de Dispersão Quantitativo ou "Scatterplot" (Fig. 10) que é um gráfico EPR. Apenas quando a sobreposição de pontos compromete a percepção da relação, o que normalmente ocorre quando o número de níveis da "variável independente" e/ou da "variável dependente" são muito pequenos que é recomendável a utilização do "Dot Density" com as variáveis numéricas tratadas como categóricas (fig 11).

MYSTAT: Diagrama de dispersão quantitativo: Graph/ Scatterplot/ VI→ X variable/ VD→Y Variable. Linhas de regressão linear e outros tipos de linhas de tendências podem ser escolhidas na aba Smooth. Uma linha de regressão apenas pode ser representada se a relação tiver sido comprovada estatisticamente.. Diagrama de dispersão nominal Graph/ Dot Density/ VI→ X variable/ VD→Y Variable/ Type of display=symmetrical. Se houver necessidade de se representar uma linha, isto pode ser feito com sobreposição de gráficos.

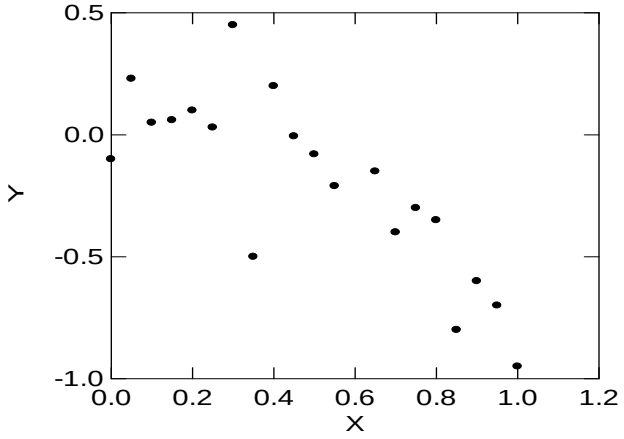


Figura 12- Diagrama de dispersão ou "Scatterplot"

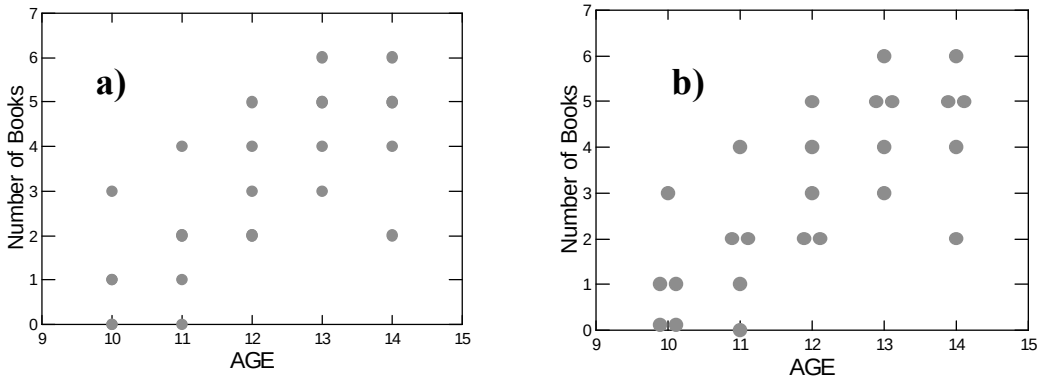


Figura 13- O "dot density" é melhor que o "scatterplot" se o número de níveis é baixo e a sobreposição de pontos comprometer a percepção da relação.

Seção VI- Gráficos com "variável dependentes" nominal e "independente" quantitativa

A apresentação de uma "variável dependente" nominal no eixo Y e uma "independente" quantitativa no eixo X é um pouco incomum, entretanto, é uma forma correta de se apresentar estes dados. Isto incomoda tanto que algumas pessoas chegam a inverter os eixos para deixar a variável nominal em X, mas isto está errado, pois a "variável independente" deve ficar no eixo X. Possivelmente seja pouco comum porque os testes utilizados nesta situação são a regressão logística (se VD for binária), que raramente é ensinada em cursos básicos de estatística, e outras análises realmente complexas se VD for categórica ou ordinal. As dificuldades de análise levaram os pesquisadores a utilizar abordagens alternativas (e.g. agrupamento de indivíduos em grupos para se trabalhar com taxas), o que levou a gráficos com VD quantitaviva. Entretanto, pelo menos no caso de VD binária, não é mais necessário se partir para abordagens alternativas. O gráfico EPR que representa esta situação é o Diagrama de Dispersão Nominal Transposto (DDNT ou "Dot density" transposto- Fig 14a).

Há uma alternativa para o DDNT, especialmente aplicável para situações como esta do exemplo, porque os níveis de X são fixos (determinados pelo pesquisador) e há um balanço entre o número de casos em cada nível (5 casos por nível). Neste caso, um Gráfico de Barras Composto Horizontalmente tem o mesmo poder informativo que o gráfico EPR (ambos permitem a reconstituição da base de dados), mas leva alguma vantagem porque os leitores estão mais acostumados a interpretar gráficos deste tipo. Entretanto, se tivéssemos verificando a relação entre a sobrevivência de peixes de poça (só um por poça) e o tamanho de poças naturais, o eixo X teria valores não fixos (ou "aleatórios"-mas este não é um termo apropriado), e fica mais difícil se interpretar o gráfico de barras devido principalmente ao desbalanço e perde-se a precisão da medida pela necessidade de se estabelecer intervalos na variável quantitativa.

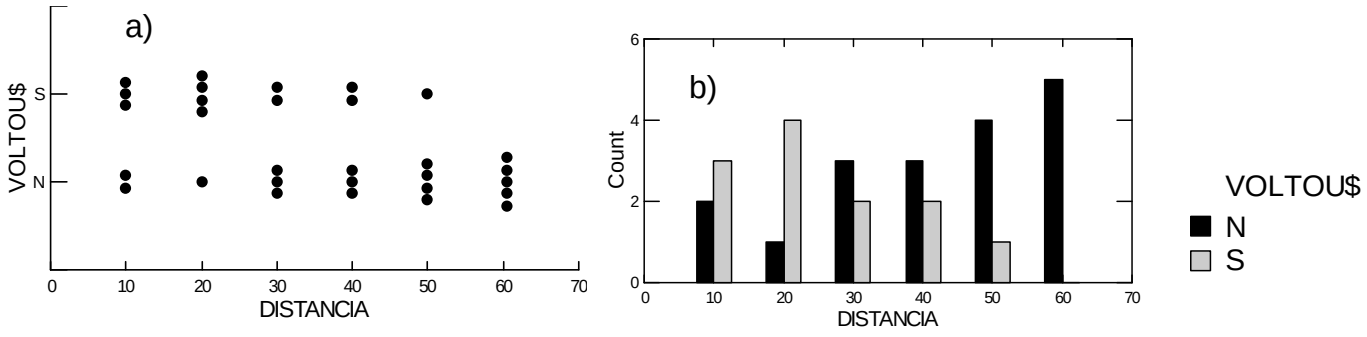


Figure 14- A distância a que formigas foram levadas de seu formigueiro e o sucesso de retorno. Duas formas de se apresentar dados com uma "variável independente" quantitativa e uma "variável dependente binária para os mesmos dados.

MYSTAT: DDNT ("dot density transposto") Graph/ Dot density/ VI → Y-Variable e VD → X variable (é invertido mesmo!)/ Display: Symmetrical/ Aba all axes: marcar transpose. Gráfico de barras- Se VI for "não fixa", deve-se dividir os dados da VI em intervalos iguais. A nova variável será VIb. Graph/ Bar Char/ VIb→X variable/ VD → "grouping variable/ Overlay multiple graphs. Se a variável independente for fixa, utilize-se diretamente VI. Graph/ Bar Char/ VI→X variable/ VD → "grouping variable/ Overlay multiple graphs.

Seção VII- Gráficos para representar resultados com desenhos em pares ou blocos para "variável dependente" quantitativa e "independente" nominal ou quantitativa

A apresentação gráficos de desenhos em pares ou blocos é um pouco complicada pois exige uma transposição da planilha original, ajustes antes da transposição e sobreposição de gráficos. Os passos são os seguintes: 1) Inicialmente temos uma planilha com entidades que são os blocos (ou pares), que é a planilha usadas na avaliação estatística destes dados. Esta planilha deve ser salva (e.g. Base_Original). 2) A variável que nomeia os blocos deve ser renomeada para LABEL$. 3) em seguida a planilha deve ser transposta [Data/ Reshape/ Transpose], selecionar as colunas que serão transpostas e marcar para salvar com outro nome (e.g. Base transposta). 4) Na planilha transposta a variável LABEL$ agora tem os nomes das antigas colunas. Estes nomes devem ser transformados para números na ordem que serão apresentados no eixo x, e em [View/ Variable/ Editor/ Value Labels] coloque a correspondência destes números para o que aparecerá no gráfico (e. g. 1=fraco; 2= médio e 3= forte). O resultado da planilha original para a transpota com a modificação está representado ao lado.

Startpage | Untitled.syo | Ar_cond_Ori.syz | Graph1 | Unt

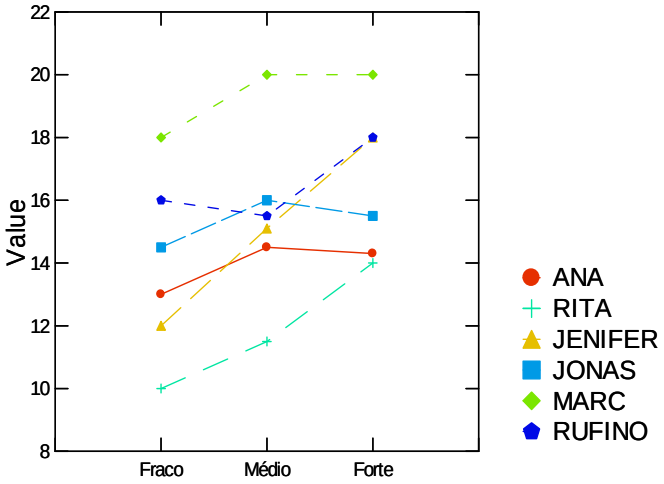
	AR_FRACO	AR_NORMAL	AR_FORTE	LABEL$
1	13.000	14.500	14.300	Ana
2	10.000	11.500	14.000	Rita
3	12.000	15.100	18.000	Jenifer
4	14.500	16.000	15.500	Jonas
5	18.000	20.000	20.000	Marc
6	16.000	15.500	18.000	Rufino
7				
8				

	ANA	RITA	JENIFER	JONAS	MARC	RUFINO	LABEL$
1	13.000	10.000	12.000	14.500	18.000	16.000	1
2	14.500	11.500	15.100	16.000	20.000	15.500	2
3	14.300	14.000	18.000	15.500	20.000	18.000	3
4							

5) faça o gráfico de linha [Graph/ Line Chart] entre Label$ em X variable e as variáveis que nomeiam os blocos em Y-variable. 6) faça o gráfico de pontos em [Graph/ Summary Charts/ Dot] da mesma forma e escolha os símbolos na aba Symbol. 7) Sobreponha os dois gráficos com [Graph/ Begin Overlay Mode/ refaça o gráfico Line e o gráfico de pontos/ End Overlay Mode].

Se a variável representada em X for quantitativa, cria-se uma variável com os valores que serão utilizados em X que substituirá a variável LABEL$, e o processo é o mesmo que foi descrito anteriormente.

Uma alternativa para este gráfico é um gráfico com símbolos ou números ao invés das linhas ligando os pontos que pode ser feito de uma forma mais simples. Monta-se uma planilha EPR com cada medida como entidade e as variáveis dependentes e independentes (no exemplo VI= intensidade do ar condicionado e VD= Nota na avaliação) e uma variável para os blocos. O gráfico pode ser montado por Dot density (seção IV) ou Diagrama de dispersão (seção V). O processo de construção do gráfico é bem mais simples e prático para uma avaliação preliminar, mas é menos recomendado para uma versão final por ser considerado menos efetivo para mostrar o efeito do tratamento.



Nível do ar condicionado