# Previsão e validade. Provas de diagnóstico

Para perceber como uma doença é transmitida e se desenvolve, e para providenciar cuidados de saúde eficazes é necessário distinguir na população quem tem a doença dos que não têm. Assim, a qualidade das provas de diagnóstico e rastreio é uma questão essencial. Temos de saber o quão bom é um teste a separar populações com e sem a doença em questão.

Mas antes de procedermos a esta questão, vamos abordar algumas características das medições.

## A. Planeando as medições: Precisão e validade

As medições descrevem os fenómenos em termos que podem ser analisados estatisticamente. A validade de um estudo depende de quão bem as variáveis desenhadas para o estudo representam o fenómeno de interesse.

**Escalas de medida:**

A classificação é importante porque alguns tipos de variáveis providenciam estatísticas mais informativas que outras, dando mais poder ao estudo e reduzindo o tamanho da amostra necessário.

1. Variáveis contínuas

As variáveis contínuas têm intervalos quantificados numa escala infinita de valores. O número de resultados possíveis de peso, por exemplo, só é limitado pela sensibilidade da máquina que é usada para o medir. As variáveis contínuas são ricas em informação.

Uma escala que tenha um número finito de intervalos (tal como o número de cigarros fumado por dia) é denominada **discreta**. As variáveis discretas ordenadas que têm um considerável número de valores possíveis assemelham-se às variáveis contínuas na análise estatística e são equivalentes para o propósito de desenho de medidas e análise.

2. Variáveis categóricas

Os fenómenos que não são apropriados para a quantificação podem ser muitas vezes medidos pela sua classificação em categorias.

Quanto ao número de variáveis:

- Dicotómicas: dois valores possíveis (ex: vivo/morto)
- Policotómicas: mais de dois valores possíveis. Estas podem ainda ser categorizadas quanto ao tipo de informação que contêm:
    - Variáveis nominais: categorias que não podem ser ordenadas (ex: tipo de sangue). Normalmente têm um carácter qualitativo e absoluto que as possibilita medir prontamente.
    - Variáveis ordinais: categorias que têm uma ordem (ex: dor leve, moderada, severa). Isto é uma vantagem relativamente às variáveis nominais, mas como as variáveis ordinais não especificam uma diferença uniforme ou numérica entre uma categoria e a próxima, a informação fornecida é menor do que nas variáveis discretas.

| Tipo de medida | Características da variável | Exemplo | Estatísticas descritivas | Informação fornecida |
|---|---|---|---|---|
| Categórica nominal | Categorias não ordenadas | Sexo, tipo de sangue | Contagem, proporções | Baixa |
| Categórica ordinal | Categorias ordenadas com intervalos que não são quantificáveis | Nível de dor | Para além das de cima: medianas | Intermédia |
| Contínua ou discreta ordenada | Espectro ordenado com intervalos quantificáveis | Peso, número de cigarros/dia | Para além das de cima: médias, variações padronizadas | Elevada |

**Como escolher uma medida?**

Uma boa regra geral é preferir as variáveis contínuas, porque a informação adicional que fornecem melhora a eficácia. Assim, o resultado final é um estudo com mais poder e/ou uma amostra mais pequena.

Há, no entanto, algumas excepções. Por exemplo, se um investigador apenas pretende estudar os determinantes do baixo peso ao nascimento, ele estará mais preocupado com os bebés cujo peso seja tão baixo que a sua saúde esteja comprometida do que com as diferenças observadas ao longo do largo espectro de pesos ao nascimento. Uma variável dicotómica como a proporção de bebés abaixo e acima dos 2500 g seria o ideal. Mas mesmo quando os dados categorizados são mais significativos continua a ser melhor recolher os dados como uma variável contínua. Assim, as opções analíticas, tais como mudar o *cut-off point* ficam em aberto.

Muitas características, particularmente os sintomas (ex: dor) ou aspectos do estilo de vida (ex: qualidade de vida), são difíceis de descrever com categorias e números. Mas estes fenómenos têm muitas vezes papéis importantes no diagnóstico e nas decisões de tratamento, e a tentativa de os medir é uma parte essencial da abordagem científica à descrição e análise. Isto é conseguido, por exemplo, com questionários estandardizados. O processo de classificação e medição, se feito correctamente, pode melhorar a objectividade do nosso conhecimento, reduzir os viés e providenciar um meio de comunicação.

## Resolução dos exercícios – primeira parte:

**1)**

### Precisão

A precisão (também chamada de fiabilidade ou consistência) está relacionada com a reprodutibilidade de um teste, ou seja, uma medida precisa é uma medida reprodutível que repetida nas mesmas condições obtém os mesmos resultados.

Uma balança pode medir o peso corporal com uma grande precisão, enquanto que uma entrevista para medir a qualidade de vida tem uma maior probabilidade de produzir valores diferentes de produzir valores diferentes de um observador para outro, sendo, portanto, menos precisa.

Idealmente, num estudo a única fonte de variabilidade existente deveria ser a variabilidade biológica intrínseca aos sujeitos em estudo, mas muitas vezes existe também variabilidade que está dependente da medição do observador ou do instrumento que mede.

A precisão tem uma influência muito importante no poder de um estudo. Quanto mais precisa uma medida, maior o poder estatístico para estimar valores médios e para testar hipóteses, mantendo o tamanho amostral constante.

A precisão é afectada por **erros aleatórios**. Quanto maior o erro, menos precisa é a medida. Existem três principais fontes de erros de precisão:

- Variabilidade do observador – devido ao observador. Ex: escolha de palavras numa entrevista, habilidade no uso de um instrumento mecânico.
- Variabilidade do sujeito – variabilidade biológica intrínseca nos objectos de estudo devido a, por exemplo, flutuações de humor ou tempo desde a última medicação.
- Variabilidade do instrumento – devido a factores ambientais variáveis, como a temperatura, ou a componentes mecânicos envelhecidos, entre outros.

A precisão é avaliada pela consistência de medidas repetidas:

- Reprodutibilidade intra-observador
- Reprodutibilidade inter-observador
- Reprodutibilidade intra-instrumental
- Reprodutibilidade inter-instrumental

Os métodos usados para quantificar a concordância dependem da escala das variáveis em estudo:

| | **Escala** |
|---|---|
| Grau de concordância | Categórica |
| Estatística Kappa | Categórica dicotómica |
| Estatística Kappa ponderada | Categórica ordinal |
| Limites de concordância (Bland & Altman) | Continua |
| Coeficiente de Correlação Intra-classe | Continua |

} Exercício 3

**Estratégias para aumentar a precisão:**

1. Estandardizar os métodos de medição – com definições operacionais (instruções específicas para fazer as medições)
2. Treinar e certificar os observadores – pode aumentar a reprodutibilidade de um procedimento especialmente na concordância inter-observador
3. Refinar os instrumentos – os instrumentos mecânicos e electrónicos podem ser desenhados e modificados para reduzir a variabilidade. Também as entrevistas e questionários podem ser escritos para aumentar a clareza e evitar potenciais ambiguidades.
4. Automatizar os instrumentos – variações na maneira como observadores humanos fazem as medições podem ser eliminadas com o uso de aparelhos automáticos e questionários *self-response.*
5. Repetição – o efeito dos erros aleatórios pode ser reduzido se repetirmos a medição e usarmos a média das duas leituras. Assim podemos aumentar muito a precisão, mas tem como limitações o custo adicional e as dificuldades práticas de repetir as medições.

Para cada medição no estudo, o investigador tem de decidir quão vigorosamente perseguir cada uma destas estratégias. Em geral, as primeiras duas estratégias (estandardização e treino) devem ser sempre ser usadas, e a quinta (repetição) é uma opção garantida para aumentar a precisão quando é possível e o custo é suportável.

## Validade[1]

A validade está relacionada com a exactidão de um estudo; é a capacidade que um teste tem de medir aquilo que queremos medir; diz respeito ao quão bem uma medida representa o fenómeno de interesse. Se um teste for muito válido sabemos que estamos a medir o real.

- Interna – estudo que dá resultados correctos para a população alvo daquele estudo.
- Externa – estudo que dá a possibilidade de se generalizar para a população geral.

[1] Alguns autores costumam chamar à validade "Exactidão", reservando o termo "validade" para uma forma de exactidão usada para variáveis abstractas e subjectivas (exemplo: dor e qualidade de vida), para as quais não há gold standart concreto. No entanto, nas aulas apenas foi referido o termo "Validade" em geral.

A validade é sobretudo influenciada por **erros sistemáticos**. Os três principais erros sistemáticos são:

- Erro do observador – distorção, consciente ou inconsciente, na percepção ou relato da medição pelo observador. Pode representar erros sistemáticos na maneira como um instrumento é operado, como, por exemplo, uma tendência para arredondar as medições de pressão arterial.
- Erro do sujeito – por exemplo, o viés de memória. Pacientes com cancro da mama que acreditam que a dieta rica em gordura é causa de cancro podem recordar exageradamente as quantidades de gordura consumidas quando mais novos.
- Erro do instrumento - por exemplo: má calibração, dando consistentemente resultados errados.

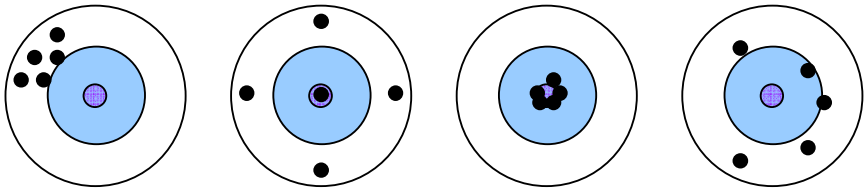
A **validade** de uma medida é melhor avaliada comparando-a com um ***"gold standard"***, que é uma técnica de referência que é considerada como válida, exacta. Para medições numa escala contínua, pode-se determinar a diferença média entre a medição sobre investigação e o gold standart. Para medições numa escala categórica, a comparação da validade da medição com o gold standart pode ser descrita em termos de sensibilidade (capacidade de identificar correctamente aqueles que têm a doença) e de especificidade (capacidade de identificar correctamente quem não tem a doença) – ver segunda parte da aula. Quando um gold standart não está disponível, o investigador deve utilizar outras medidas para avaliar a validade.

**Estratégias para aumentar a validade:**
(Inclui as primeiras 4 listadas anteriormente para a precisão e 3 adicionais)

1. Estandardizar os métodos de medição
2. Treinar e certificar os observadores
3. Refinar os instrumentos
4. Automatizar os instrumentos
5. Utilizar métodos não invasivos – é possível desenhar métodos de medição dos quais os sujeitos não estejam cientes, eliminando assim a possibilidade de eles enviesarem conscientemente a variável (por exemplo, medir a frequência respiratória enquanto medimos a frequência cardíaca, sem dizer nada ao paciente).
6. Ocultamento – esta é uma estratégia clássica que não assegura a validade geral das medições, mas pode eliminar viés diferenciais que afectem um grupo de estudo mais do que outro. Numa experiência com ocultamento duplo, nem o observador, nem o sujeito sabem se lhes foi atribuído o medicamento ou o placebo, assegurando que as medições do outcome não vão ter diferentes graus de validade nos dois grupos. Os estudos observacionais também usam o ocultamento para resguardar os valores das variáveis preditivas daqueles que vão medir os outcomes.
7. Calibrar os instrumentos

Tal como no caso da precisão, o investigador é que decide quais as estratégias a usar, tendo em consideração a importância da variável, a magnitude do impacto potencial que o grau antecipado de falta de validade terá nas conclusões do estudo, a possibilidade de realizar a estratégia e o seu custo. As primeiras duas estratégias devem ser sempre usadas, o ocultamento é essencial quando possível, e a calibração é necessária para qualquer instrumento que tenha o potencial de mudar ao longo do tempo.

**Nota: a precisão é diferente da validade e as duas não estão necessariamente ligadas. Um estudo pode ser muito preciso e não ter validade.**
Exemplo: Se o nível de colesterol for medido repetidamente usando padrões que foram inadvertidamente diluídos duas vezes, o resultado não iria representar a realidade mas poderia continuar a ser preciso (isto é, a dar sempre resultados semelhantes).



Muito preciso
Pouco válido

Pouco preciso
Muito válido

Muito preciso
Muito válido

Pouco preciso
Pouco válido

| | **Precisão** | **Validade** |
|---|---|---|
| **Definição** | Capacidade de uma variável apresentar quase o mesmo valor quando medida várias vezes | Capacidade de uma medida representar de facto o que pretende representar |
| **Melhor maneira de aumentar** | Comparação entre medições repetidas | Comparação com um gold standart |
| **Valor para o estudo** | Aumenta o poder para detectar efeitos | Aumenta a validade das conclusões |
| **Ameaçada por** | Erros aleatórios | Erros sistemáticos (viés) |

No entanto, a precisão e a validade estão muitas vezes lado a lado, e muitas das estratégias para aumentar a precisão aumentam também a validade.

É também importante verificar que um procedimento não preciso não pode ser válido; desta forma, não faz sentido fazer um teste de validade a um procedimento não reprodutível.

**Outros aspectos das medições:**

Todas as medições devem ser **sensíveis** (para conseguir detectar diferenças numa característica em estudo), **específicas** (representando apenas a característica de interesse), apropriadas (aos objectivos do estudo), **objectivas** (o que é conseguindo ao diminuir o envolvimento do observador e aumentando a utilização de instrumentos), e **devem providenciar uma distribuição adequada de respostas na população em estudo**.

**2)**

Os resultados obtidos serão diferentes porque os observadores poderão apertar mais ou menos a fita métrica, as medições poderão ser realizadas em zonas diferentes do braço do mesmo indivíduo, alguns observadores podem arredondar e outros não. Trata-se assim de **erros aleatórios**, sendo que os resultados não são precisos.

**3)**

**a)**

$$\textbf{Grau de concordância do diagnóstico} = \frac{\text{nº de vezes que concordaram com o diagnóstico}}{\text{nº máximo de vezes que podiam ter concordado}}$$

$$\textbf{Grau de concordância} = \frac{33+7+25}{90} = 72{,}2\%$$

Mas, por exemplo, eu não sabendo analisar mamografias e sabendo que a maior parte são normais, posso dizer que são todas normais e ter um grau de concordância de 90% com um

médico capaz de as analisar. Outro problema é o facto de ambos os observadores até terem o mesmo número de negativos e positivos, tendo assim um grande grau de concordância, mas os pacientes considerados positivos serem pessoas diferentes. Assim, o grau de concordância não é muito fiável. Quanto menor o número de variáveis, maior será o grau de concordância, porque maior será a probabilidade de acertarmos ao acaso. É então necessária outra medida (ver alínea b).

**b)**

Com o Kappa, nós pretendemos responder à seguinte questão:

"Qual a extensão em que os dois observadores concordaram sem ser devido ao acaso?" ou "Em que extensão é que a concordância dos dois observadores excede o nível de concordância que resulta apenas do acaso?"

$$\textbf{Kappa (K)} = \frac{\textbf{Cobs} - \textbf{Cesp}}{1 - \textbf{Cesp}}$$

$C_{Obs}$ – Concordância observada – *foi o que calculámos na alínea a).*

$C_{esp}$ – Concordância esperada *(o que nós esperávamos que eles concordassem, ao acaso)*

Assim, o numerador do kappa representa quão melhor é a concordância dos observadores do que seria esperada se eles só acertassem ao acaso e o denominador representa o número total de vezes que eles podiam acertar menos o número de vezes que podiam acertar ao acaso, ou seja, o número de vezes que poderiam acertar sem ser devido ao acaso.

$$\textbf{C}_{\textbf{esp}} = \frac{\text{nº de vezes que concordaram ao acaso}}{\text{nº de vezes que poderiam concordar ao acaso}}$$

$$\textbf{Nº de vezes que concordaram ao acaso} = \frac{\text{total da linha} \times \text{total da coluna}}{\text{total}}$$

Calcular apenas para as células em que eles concordam e não para todas (nesta caso para as células 33, 7 e 25)

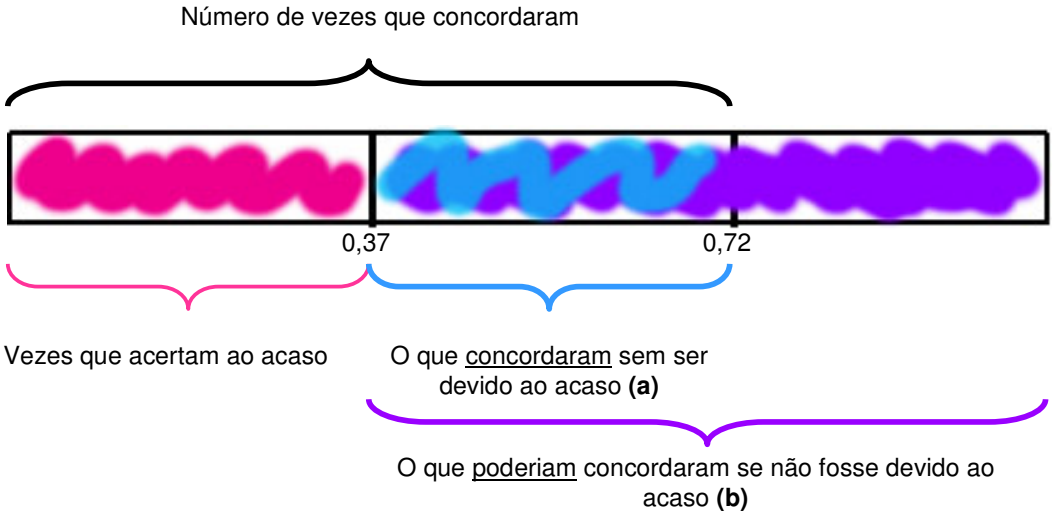
Então…

$$Cesp = \frac{\overset{\text{Não doloroso}}{\frac{41 \times 43}{90}} + \overset{\text{Duvidoso}}{\frac{21 \times 15}{90}} + \overset{\text{Doloroso}}{\frac{28 \times 32}{90}}}{90} = 0{,}37$$

Assim:

$$\textbf{K} = \frac{0{,}72 - 0{,}37}{1 - 0{,}37} = 0{,}56$$

**c)**




**Kappa** $= \frac{\mathbf{a}}{\mathbf{b}}$ → **Proporção das vezes que podiam ter concordado sem ser devido ao acaso**

**Se K = 1 →** concordam sempre

**Se K = 0 →** concordância entre eles é igual à concordância devido ao acaso

**Se K = negativo →** concordaram ainda menos vezes do que concordariam se fosse ao acaso (poderiam estar a usar critérios opostos)

**Quanto mais próximo de 1 →** Melhor **K →** Maior concordância
**Quanto mais próximo de 0 →** Pior **K →** Menor concordância

**Kappa –** é um método melhor que o grau de concordância para avaliar a concordância. É sempre menor que o grau de concordância porque lhe retiram sempre aquilo que corresponde ao que concordaram devido ao acaso. (*Na realidade o kappa também é algo afectado pela prevalência do sinal observado, o que é uma das razões pela qual os epidemiologistas clínicos estão ainda à procura de melhores maneiras de descrever a concordância)*

**4)**

**Concordância intra-observador –** a mesma pessoa a avaliar a mesma coisa duas vezes. É sempre a maior mas pode não ser sempre 100%. Os valores obtidos nas medições variam ao longo do tempo, sendo esta variabilidade considerável, mesmo durante um curto período de tempo. Para além disso, as próprias condições em que o teste é realizado (pós-prandial, pós-exercícicio, em casa ou no consultório) podem levar claramente a diferentes resultados no mesmo indivíduo. Assim, na avaliação de qualquer teste é importante considerar as condições nas quais o teste foi efectuado, incluindo a altura do dia.

**Concordância inter-observador –** duas pessoas a avaliar a mesma coisa. Dois observadores diferentes muitas vezes não obtêm o mesmo resultado. A extensão na qual os observadores concordam ou discordam é uma questão importante, já abordada na pergunta 3.

A variabilidade **intra-observador** é predominantemente aleatória, a variabilidade **inter-observador** pode ser aleatória ou sistemática.

## B. Provas de diagnóstico

O diagnóstico é um dos mais importantes actos em medicina. Fazer um **diagnóstico** é um processo probabilístico de decisão que visa classificar o doente dentro de uma determinada entidade nosológica a que corresponderá um determinado tratamento e um determinado prognóstico. Para levar a cabo um diagnóstico teremos então que utilizar métodos que permitam discriminar entre populações de doentes e de não doentes, sendo essa a definição de **teste diagnóstico**. O termo "testes diagnósticos" aplica-se geralmente aos exames complementares de diagnóstico; no entanto, ele deve ser entendido num sentido mais amplo, abrangendo não só os exames complementares de diagnóstico como também todos os dados provenientes da história clínica e exame físico.

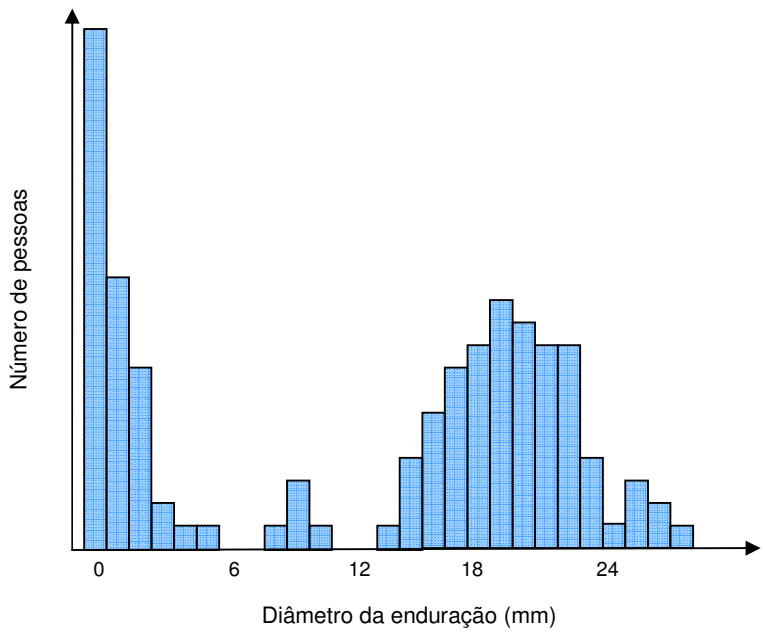
A avaliação da exactidão de um qualquer teste diagnóstico está dependente da comparação dos resultados a partir dele obtidos com o verdadeiro estado de cada indivíduo. Assim, para determinar o verdadeiro diagnóstico, é preciso que exista um teste, ou conjunto de testes, que dêem uma grande certeza sobre o mesmo. Ao teste que, em determinado momento, tem a maior exactidão na determinação de um diagnóstico dá-se o nome de **"gold standard"**. No entanto, o "gold standard" raramente tem uma exactidão de 100%, tornando-se, assim, difícil utilizá-lo como padrão para comparação com testes alternativos de que se desconhece a exactidão.

Os testes diagnósticos podem ser classificados em dois grandes grupos:

- Testes qualitativos: o resultado do teste, positivo ou negativo, é dado tendo em conta a presença ou ausência de uma determinada característica.
- Testes quantitativos: o resultado do teste é estabelecido numa escala contínua e é classificado como positivo ou negativo tendo em conta um determinado "cutoff point" arbitrariamente seleccionado.

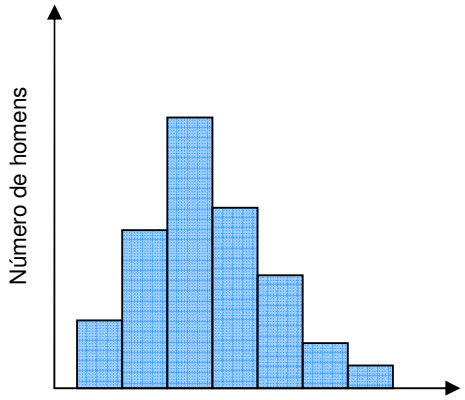
Ao usar testes diagnósticos é importante perceber como as características estão distribuídas nas populações humanas.

Por exemplo, o seguinte gráfico representa a distribuição dos resultados de testes de tuberculina numa população:



Um grande grupo de pessoas centra-se no valor dos 0 mm – sem enduração (é a área de dureza no local da injecção) – e outro grupo centra-se perto dos 20 mm de enduração. Este tipo de distribuição, na qual há dois picos, é chamada de curva bimodal. Esta distribuição bimodal permite a separação de indivíduos que não tiveram contacto anterior com tuberculose (sem enduração) daquelas que já tiveram contacto (enduração superior a 20 mm). Apesar de alguns indivíduos se situarem numa "zona cinzenta" no centro, podendo pertencer a qualquer uma das duas curvas mencionadas, a maioria da população pode ser facilmente distinguida usando as duas curvas. Assim, quando a característica tem uma distribuição bimodal é relativamente fácil separar a maioria da população em dois grupos (ex: doente/não doente).

No entanto, geralmente, a maioria das características humanas não são distribuídas de um modo bimodal. O seguinte gráfico mostra a distribuição de pressões arteriais sistólicas num grupo particular.




Neste gráfico não há uma curva bimodal; o que nós vemos é uma curva unimodal – um pico único. Assim, se quisermos separar deste grupo aqueles que são hipertensos dos que não são, deve-se estabelecer um nível de *cutt off* de pressão arterial, acima do qual as pessoas são hipertensas e abaixo do qual as pessoas são normotensas. Não há um nível óbvio de pressão arterial que distinga os hipertensos dos normotensos. Apesar de podermos escolher um *cutt off point* para a hipertensão baseado em considerações estatísticas, podemos idealmente escolher um cutoff com base em informações biológicas: isto é, nós queremos saber que um nível de pressão arterial acima do cutoff escolhido está associado com um risco aumentado de doença subsequente, tal como AVC, enfarte agudo do miocárdio ou mortalidade subsequente. Infelizmente, para muitas características humanas, não possuímos tal informação para servir como um guia para estabelecer este nível.

Em qualquer uma das distribuições – unimodal ou bimodal – é relativamente fácil distinguir os valores extremos do anormal e normal. No entanto, a incerteza permanece relativamente aos casos que se situam na zona cinzenta em qualquer tipo de curva.

## Teorema de Bayes

Thomas Bayes, um matemático inglês do século XVII legou-nos o seu teorema que estabeleceu que a probabilidade pós-teste de uma doença era função da sensibilidade e especificidade do exame e da prevalência da doença na população (a prevalência é a probabilidade pré-teste, é a probabilidade de ocorrência da doença).

O teorema de Bayes é usado na inferência estatística para actualizar estimativas da probabilidade de que diferentes hipóteses sejam verdadeiras, baseado nas observações e no conhecimento de como essas observações se relacionam com as hipóteses. Este teorema é uma das pedras angulares da estatística das probabilidades combinadas, e é largamente utilizada em áreas à primeira vista pouco relacionadas, como Medicina e Informática. Em Medicina, o paradigma baseado em evidências é todo construído com base no teorema de Bayes. Baseado na experiência acumulada de exames e testes para tentar diagnosticar uma

doença, o médico enquadra os seus pacientes e pode estimar qual a probabilidade de que uma dada doença se esteja a manifestar. Ou seja, dada uma probabilidade inicial (por exemplo, o paciente é fumador) e aplicado um exame em que, se sabe, há uma probabilidade de falsos-positivos e falso-negativos (por exemplo, uma biópsia de pulmão), o médico sabe qual a probabilidade resultante daquele paciente ter a doença (por exemplo, cancro de pulmão). Nós, médicos, ao formularmos as nossas hipóteses diagnósticas, ao interpretarmos os exames laboratoriais e ao prescrevermos um tratamento, intuitivamente, utilizamos o teorema de Bayes.

Explicando o teorema de Bayes:

## 1. Eventos

Independentes: Diz-se que dois eventos são independentes quando a ocorrência de um deles não influencia a probabilidade do outro ocorrer.

$$p(E_1 \ e \ E_2) = p(E_1) \times p(E_2)$$

sendo ***p*** ... probabilidade, e $E_1$ e $E_2$ dois eventos independentes, lê-se: a probabilidade de ocorrência conjunta dos dois eventos é o produto das probabilidades de ocorrência individuais.

Não Independentes:

$$p(E_1 \ e \ E_2) = p(E_2) \times p(E_1 | E_2)$$

sendo ***p*** ... probabilidade, e $E_1$ e $E_2$ dois eventos não independentes, lê-se: a probabilidade de ocorrência conjunta dos dois eventos é a probabilidade de ocorrência de um evento $E_2$ multiplicada pela probabilidade do outro evento $E_1$ dado que o $E_2$ ocorreu.

## 2. Probabilidade de Ocorrência

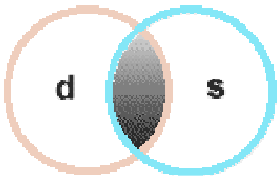
Se a prevalência de casos de Tuberculose em uma dada comunidade é dada por:

$$P(Tbc) = \frac{n^o \ de \ casos}{total \ da \ população}$$

então a probabilidade de ocorrência ***P*(*d*) = *P*(*Tbc*)**

## 3. Diagnóstico

O processo classificatório das doenças (***d***) relaciona-se com o conjunto de evidências (sinais, sintomas, exames auxiliares), denominado ***s***.




$$P(d \mid s) = \frac{|d \cap s|}{|s|}$$

Em outras palavras, para o diagnóstico precisamos conhecer a probabilidade condicional ***P*(*d*|*s*)** da doenca ***d*** para cada evidência ***s***, mas é preciso levar em conta também a probabilidade de ocorrência (prevalência) da doença na comunidade na qual estamos a actuar, então:

$$P(d \mid s) = \frac{P(d) \cdot P(s \mid d)}{P(s)}$$

Nota: na aula teórica foi utilizada a seguinte nomenclatura: **P(H|D)=P(D|H) x P(H) / P(D)**

**Aplicação do Teorema de Bayes**

| | | Doença | | |
|---|---|---|---|---|
| | | **Presente** | **Ausente** | |
| **Teste** | **Positivo** | a | b | $a+b$ |
| | **Negativo** | c | d | $c+d$ |
| | | $a+c$ | $b+d$ | |

$$Se = \frac{a}{a+c} \qquad Es = \frac{d}{b+d}$$

$$VPP = \frac{a}{a+b} \qquad VPN = \frac{d}{c+d}$$

Características dos testes diagnósticos: *Se - sensibilidade; Es - especificidade; VPP - valor preditivo positivo; VPN - valor preditivo negativo; P - prevalência; Ex - exactidão*

Como já referido anteriormente, para medições numa escala categórica, a comparação da validade da medição com o gold standart pode ser descrita em termos de sensibilidade e especificidade.

**Sensibilidade:** proporção de indivíduos doentes que têm um teste positivo ou a probabilidade de, estando doente, ter um teste positivo; dizemos que um teste é sensível quando tem a capacidade de detectar os doentes, isto porque o teste, geralmente, é positivo quando a doença está presente. Toda a gente que tiver o teste negativo é não doente. No entanto, coloca-se o problema dos falsos positivos, isto é, pessoas a quem o teste dá positivo mas que não estão doentes. Ou seja, se o teste der positivo, as pessoas podem ou não ter a doença. Conclui-se que um teste muito sensível é mais útil quando o resultado é negativo.
Este tipo de testes é útil nas seguintes situações:

- Quando existe uma penalização importante para a omissão do diagnóstico;
- Em programas de rastreio;
- No início da avaliação de um doente, quando estão a ser consideradas muitas possibilidades de diagnóstico, de modo a pôr de parte, com grande confiança, alguns diagnósticos, e assim, reduzir as possibilidades de diagnóstico.

**Especificidade –** proporção de indivíduos não doentes que têm um teste negativo ou a probabilidade de, não estando doente, ter um teste negativo; dizemos que um teste é específico quanto tem a capacidade de detectar com bastante certeza os não doentes, isto porque o teste, geralmente, é negativo quando a doença está ausente. Toda a gente que tiver o teste positivo tem a doença. No entanto coloca-se o problema dos falsos negativos, isto é, pessoas a quem o teste dá negativo mas que têm a doença. Ou seja, se o teste der negativo, as pessoas podem estar ou não doentes. Conclui-se que um teste muito específico é mais útil quando o resultado é positivo.
Este tipo de testes é útil nas seguintes situações:

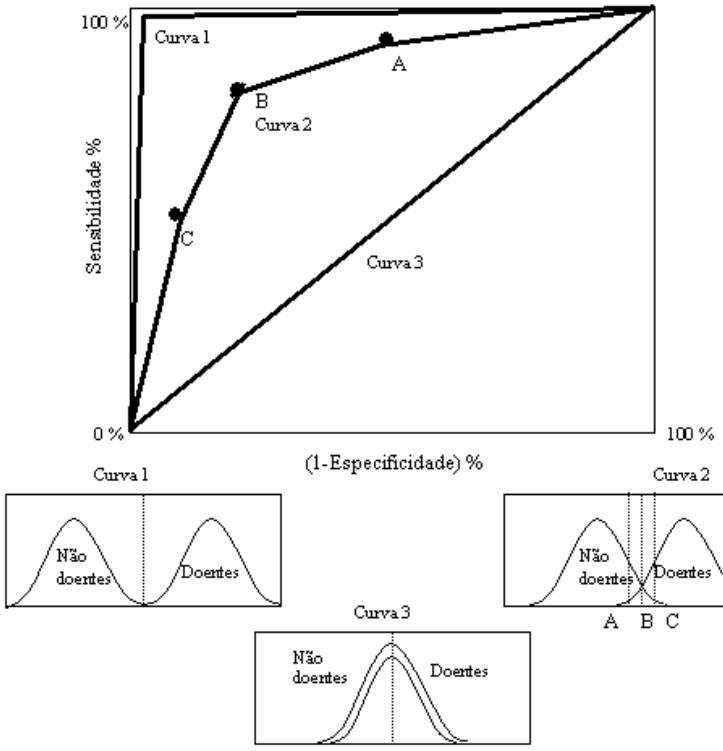
- Quando se pretende confirmar um diagnóstico que é sugerido por testes menos específicos;
- Quando a existência de um resultado falso positivo tem importantes implicações físicas, emocionais ou financeiras para o doente.

**Quando a sensibilidade aumenta, a especificidade <u>não</u> diminui obrigatoriamente (e vice-versa)!! Só podemos assumir que isto acontece (uma aumentar e a outra diminuir) quando as variáveis são <u>contínuas!!</u>**

**Curvas ROC**

Geralmente, a sensibilidade e a especificidade são características difíceis de conciliar, isto é, é complicado aumentar a sensibilidade e a especificidade de um teste ao mesmo tempo. As curvas ROC (receiver operator characteristic curve) são uma forma de representar a relação, normalmente antagónica, entre a sensibilidade e a especificidade de um teste diagnóstico quantitativo, ao longo de um contínuo de valores de "cutoff point".

Para construir uma curva ROC traça-se um diagrama que represente a sensibilidade em função da proporção de falsos positivos (1- Especificidade) para um conjunto de valores de "cutoff point".




Quando se tem uma variável contínua, resultado da aplicação de um teste diagnóstico quantitativo, e se pretende transforma-la numa variável dicotómica, do tipo doente / não doente, temos que utilizar um determinado valor na escala contínua que discrimine entre essas duas classes. A esse valor dá-se o nome de **"cutoff point"**.

O valor escolhido como "cutoff point" vai influenciar as características do teste, como exemplificado na figura (curva 2). No exemplo da figura, quanto maior é o "cutoff point" maior é a especificidade do teste mas menor é a sensibilidade (ponto C da curva 2); e quanto menor o "cutoff point" maior é a sensibilidade mas menor é a especificidade (ponto A da curva 2).

Assim, a intenção com que se utilizará o teste diagnóstico vai influenciar a escolha do "cutoff point", logo, das características do teste. No exemplo da curva 2, se pretendemos um teste muito sensível e menos específico, escolhe-se um "cutoff point" menor (ponto A), obtendo-se uma maior proporção de falsos positivos; se pretendemos um teste muito específico e menos sensível, escolhe-se um "cutoff point" maior (ponto C), obtendo-se uma menor proporção de falsos negativos.

As curvas ROC descrevem a capacidade discriminativa de um teste diagnóstico para um determinado número de valores "cutoff point". Isto permite pôr em evidência os valores para os quais existe maior optimização da sensibilidade em função da especificidade. O ponto, numa curva ROC, onde isto acontece é aquele que se encontra mais próximo do canto superior esquerdo do diagrama (ponto B da curva 2).

Por outro lado, as curvas ROC permitem quantificar a exactidão de um teste diagnóstico, já que, esta é proporcional à área sob a curva ROC, isto é, tanto maior quanto mais a curva se aproxima do canto superior esquerdo do diagrama. Sabendo isto, a curva será útil, também, na comparação de testes diagnósticos, tendo um teste uma exactidão tanto maior, quanto maior for a área sob a curva ROC.

Assim, resumindo, no caso das variáveis contínuas aumentamos a sensibilidade ao diminuir o cutoff point, diminuindo assim a especificidade; se aumentarmos a especificidade ao aumentar o nível de cutoff, diminuímos a sensibilidade. A escolha de um nível alto ou baixo de cutoff depende da importância que os falsos positivos e os falsos negativos tiverem na doença em questão.

A questão dos falsos positivos é importante porque todas as pessoas cujo resultado do teste for positivo têm de voltar a ser testadas com testes mais sofisticados e mais caros. Dos vários problemas que resultam, o primeiro é o encargo para o sistema de saúde. Outro é a ansiedade e preocupação induzidas nas pessoas a quem foi dito que o teste foi positivo, para além de nunca se livrarem do rótulo de o teste ter dado positivo, mesmo que subsequentemente os testes forem todos negativos.

A questão dos falsos negativos é importante porque se uma pessoa for erroneamente informada que o seu teste deu negativo, e se a doença for séria havendo uma intervenção eficaz disponível, o problema é de facto crítico, principalmente se a doença só for curável nos primeiros estádios. Assim, a importância dos falsos negativos depende da natureza e severidade da doença que está a ser rastreada, da eficácia das medidas de intervenção e de o facto de a intervenção só ser eficaz se administrada precocemente na história natural da doença.

Note-se que para calcular a sensibilidade e a especificidade de um teste, nós temos de saber quem realmente tem a doença e quem não tem, usando outra fonte para isso (o gold standart). No entanto, na vida real, quando usamos um teste para identificar doentes e não doentes numa população, nós claramente não sabemos quem tem a doença ou não (se isto já estivesse estabelecido fazer o teste não faria sentido nenhum).

Na clínica, e uma vez pedido um teste diagnóstico, a sensibilidade e a especificidade do teste deixam de ser importantes, passando a interessar só os **valores preditivos** do teste, isto é, a probabilidade de, perante um resultado positivo ou negativo, existir ou não doença.

**Valor preditivo positivo (VPP) –** proporção de indivíduos com o teste positivo que são doentes ou a probabilidade de, tendo um teste positivo, estar doente (probabilidade pós-teste). <u>É maior nos testes mais específicos</u>

**Valor preditivo negativo (VPN) –** proporção de indivíduos com o teste negativo que não são doentes ou a probabilidade de, tendo um teste negativo, não estar doente. <u>É maior nos testes mais sensíveis</u>

Na tabela 1 sumariza-se a relação entre o resultado de um teste diagnóstico e o verdadeiro diagnóstico:

| | | **Doença** | |
|---|---|---|---|
| | | **Presente** | **Ausente** |
| **Teste** | **Positivo** | Verdadeiro Positivo | Falso Positivo |
| | **Negativo** | Falso Negativo | Verdadeiro Negativo |

**<u>IMPORTANTE:</u>**

A **sensibilidade** e a **especificidade** <u>caracterizam um teste</u> (ou seja, são sempre iguais para um teste, independentemente da população a que são aplicados, <u>não dependem da prevalência da doença</u>). São elas que nos dão a **validade** de um teste (a capacidade de ele acertar).

O **VPP** e o **VPN** não são tão importantes para um epidemiologista, mas são muito importantes para um médico; no entanto, **não são** características dos testes. Para o mesmo teste, conforme a população a que aplicarmos o teste, o VPN e o VPP vão ser diferentes. Os valores preditivos de um teste diagnóstico dependem, essencialmente, de dois factores: especificidade do teste e a prevalência da doença. A sensibilidade e especificidade, pelo contrário, não dependem da prevalência da doença.

→ relação do valor preditivo com a prevalência (ver exemplo no exercício 5 e) )

Quanto maior a prevalência da doença na população testada, maior é o VPP e menor o VPN; quanto menor a prevalência da doença na população testada, menor é o VPP e maior o VPN.

Qual o interesse prático disto? Se quanto maior a prevalência, maior o VPP, um programa de rastreio é mais produtivo e eficiente se for direccionado para uma população de alto risco. Rastrear uma população para uma doença relativamente infrequente pode ser um grande desperdício de recursos. Para além disso, uma população de alto risco pode estar mais motivada para participar em tal rastreio e em tomar alguma acção se os seus resultados forem positivos.

A relação entre valor preditivo e prevalência da doença também demonstra que os resultados de qualquer teste devem ser interpretados no contexto da população da qual o indivíduo provém.

→ relação do valor preditivo com a especificidade

Se numa população com baixa incidência da doença, aumentarmos a especificidade, isto resulta num aumento muito maior no valor preditivo do que o mesmo aumento na sensibilidade. Porque é que isto acontece?


Como nós estamos a lidar com doenças pouco frequentes, a maioria da população encontra-se à esquerda da linha vertical. Consequentemente, qualquer mudança que ocorra à esquerda da linha vertical afecta um maior número de pessoas do que uma mudança comparável à direita. Assim, um aumento na especificidade tem um maior efeito no valor preditivo do que uma mudança na sensibilidade. (Se estivéssemos a lidar com uma doença muito prevalente a situação seria diferente.)

**As relações entre as variáveis mencionadas são:**

• Quanto maior a sensibilidade, maior será o valor preditivo negativo, isto é, maior será a probabilidade de, perante um resultado negativo, não haver doença.

• Quanto maior a especificidade, maior será o valor preditivo positivo, isto é, maior será a probabilidade de, perante um resultado positivo, haver doença.

• Quanto maior a prevalência da doença, maior será o valor preditivo positivo e menor será o valor preditivo negativo, isto é, quanto mais frequente é uma doença mais provável é encontrar verdadeiros positivos (aumentando o valor preditivo positivo), mas também é mais provável encontrar falsos negativos (diminuindo o valor preditivo negativo).

**Uso de testes múltiplos:**

Muitas vezes podemos usar vários testes, tanto sequencialmente como simultaneamente.

No caso dos testes sequenciais, podemos fazer primeiro um teste menos caro, menos invasivo e menos desconfortável, e chamar apenas para um segundo teste (mais caro, mais invasivo ou mais desconfortável, com maior sensibilidade ou especificidade) aqueles a quem o primeiro teste teve um resultado positivo. Espera-se que assim se reduza o problema dos falsos positivos.

No caso dos testes simultâneos, o indivíduo é submetido á uma bateria de testes sendo apenas considerado como "positivo" se tiver obtido um resultado positivo em um ou mais testes, e considerado como "negativo" se o resultado for negativo em todos os testes. O resultado desta técnica na sensibilidade e especificidade difere do resultado dos testes sequenciais. Nos testes sequenciais, quando apenas "re-testamos" aqueles que obtiveram resultado positivo no primeiro teste, há uma perda da sensibilidade geral e ganho de especificidade geral. Nos testes simultâneos, como o indivíduo que obtém um resultado positivo em *qualquer* um dos testes ou em múltiplos testes é considerado como "positivo" há um ganho de sensibilidade geral. No entanto, para ser considerado negativo, tem de obter resultados negativos em todos os testes efectuados, resultando numa perda de especificidade.

## Resolução dos exercícios – segunda parte:

**5)**

**a)**

**Sensibilidade =** 150 / **(**150 + 76**) = 66,4%**
**Especificidade =** 42 / **(**42 + 9**) = 82,3%**
**VP (+) =** 150 / **(**150 + 9**) = 94,3%**

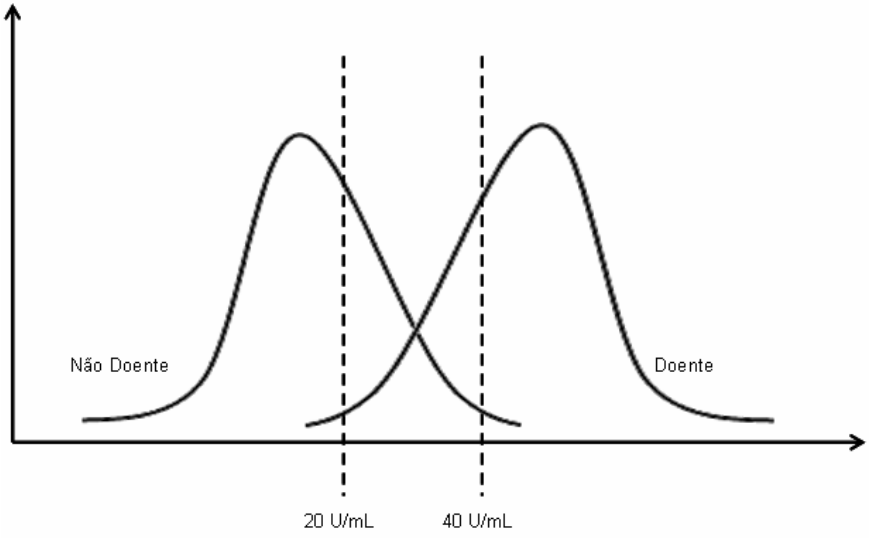
**b)**

↓ Falsos positivos
↑ Falsos negativos

} Teste mais específico e menos sensível
↑ VPP

Doentes com 35 U/mL, com doença, agora teriam teste negativo porque só a partir de 40 U/mL é que o teste daria positivo. → ↑ Falsos negativos

Contudo, pessoas com 25 U/mL, sem doença, agora já não teriam teste positivo (o que acontecia quando o ponto de corte era 20 U/mL) mas sim negativo. → ↓ Falsos positivos



Das pessoas com teste positivo, a percentagem de não doentes é maior para 20 U/mL *(parte da curva a pertencente a não doentes que está para a direita da linha tracejada correspondente a 20 U/mL)*.

**c)**

↑ **Sensibilidade:** usada nos rastreios (para excluir a doença) – *os que têm teste negativo são não doentes.*

↑ **Especificidade:** usada para confirmação de diagnóstico – *os que têm teste positivo são mesmo doentes.*

**d)**

**Sensibilidade =** 215 / **(**215 + 109**) = 0,67**

**Especificidade =** 280 / **(**55 + 280**) = 0,84**

**VPP =** 215 / **(**215 + 55**) = 0,80**

**e)**

A sensibilidade e a especificidade são sensivelmente iguais. Não esquecer que, como já foi referido, a sensibilidade e a especificidade são características do teste; sendo o teste o mesmo, é natural que se obtenham resultados semelhantes.

No entanto o VPP é menor para a alínea d). Isto porque o teste foi aplicado a indivíduos internados no hospital por qualquer motivo; a prevalência da doença (um dos factores que influencia os valores dos Valores preditivos) é menor nesta população do que na população com dor abdominal.

< Prevalência, < VPP.

## Outros exercícios:

**1)**

Pessoa no Inverno apanha chuva.

Teste usado: tosse sem expectoração (*os sintomas também são testes)*

A tosse será, neste contexto, um teste:

a) Sensível
b) Específico
c) com elevado VPP
d) com elevado VPN

**R:** relativamente à sensibilidade e à especificidade não podemos dizer nada. Mas tendo em conta o contexto (*Inverno e tendo apanhado chuva*), a probabilidade de uma pessoa estar constipada é grande → assim, o teste tem um elevado VPP, visto que a probabilidade de uma pessoa, tendo um teste positivo (*ter tosse),* estar doente (*estar constipada)* é elevada.

**2)**

Quem é doente tem teste positivo.

**R:** teste sensível.

**3)**

Pessoas com trombo embolismo pulmonar têm dendímeros positivos. Mas existem pessoas com dendímeros positivos sem doença.

**R:** Teste sensível *(detecta todos os doentes mas não consegue detectar todos os não doentes).*

**4)**

Biopsia da próstata:

- Se **Sim** → Tem adenocarcinoma
- Se **Não** → Pode ter ou não

**R:** Teste específico *(detecta os não doentes porque todos os não doentes têm teste negativo).*

**5)**

A mesma tosse de 1) mas no Verão e é fumador
Neste contexto como será o VPP da tosse?

**R:** Neste caso, a probabilidade de tendo tosse estar constipada é muito baixa. Assim, o VPP é baixo.

**6)**

Teste tuberculina + → doente ou não
Teste tuberculina - → não é doente

Como é a sensibilidade, especificidade, VPP e VPN?

**R**: Sensibilidade alta, especificidade baixa, VPP baixa, VPN negativo

**7)**

Ecografia:
Com derrame pleural: teste positivo
Sem derrame pleural: teste negativo

Como é a sensibilidade, especificidade, VPP e VPN?

R: É tudo alto. (aqui se vê o que foi referido no exercício 5 de que lá por a sensibilidade aumentar não significa que a especificidade diminua – isto excluindo os testes com variáveis contínuas em que se utilizam pontos de corte.)