Módulo 5

ESTUDOS DE COORTES

5.1 Estudos de coortes

Há vários tipos de estudos usados em epidemiologia para avaliar a associação entre factores de risco e doenças. Os principais são os estudos transversais, os de caso-controlo, os de coortes e os estudos interventivos. Em módulos anteriores vimos exemplos de estudos em que o aparecimento de casos de doença ocorre antes do estudo se iniciar: os estudos transversais e os estudos caso-controlo. Neste módulo, apresentam-se os **estudos de coortes**, nos quais os acontecimentos ainda não ocorreram antes do estudo se iniciar. Por vezes são também designados por estudos prospectivos, uma designação que evitarei por razões que se explanam mais adiante.

De um modo geral, uma coorte é um conjunto de indivíduos definidos segundo um critério qualquer, acompanhados ao longo de um período de tempo. Durante este período, que tanto pode ser toda a sua vida como apenas um intervalo de tempo pré-definido, é medida a incidência de casos de doença entre os indivíduos da coorte. Num estudo típico, são seleccionados dois grupos de pessoas que, à partida, não têm doença. Um grupo está exposto a um factor que se pensa poder estar associado com uma doença (os expostos), o outro grupo não está exposto. Ambos são seguidos ao longo do tempo e, no fim, a incidência da doença[1] é comparada entre os dois grupos (Fig. 5.1).

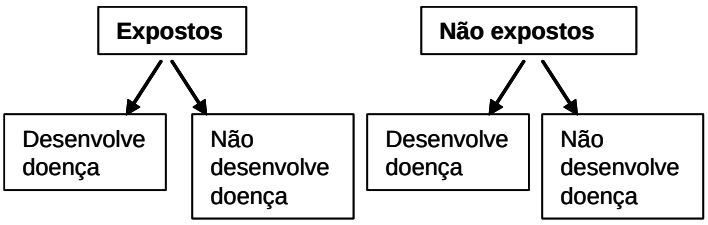


Figura 5.1. Conceptualização de um estudo de coortes. Dois grupos são formados à partida, segundo o critério de estarem ou não expostos a um factor considerado de risco. Mais tarde a incidência da doença é avaliada em cada um dos grupos.

[1] A "doença", nos estudos de risco, deve ser entendida como uma metáfora para qualquer acontecimento que se pretende estudar e que não tem de ser necessariamente ficar doente. Pode ser, por exemplo, um

Se existir uma associação entre exposição e doença, espera-se que a proporção de doentes dentro do grupo exposto seja maior que a proporção de doentes dentro do grupo não exposto.

Os estudos de coortes são por vezes ditos **longitudinais**. Os indivíduos são acompanhados ao longo do tempo à medida que envelhecem e a informação obtida reporta-se ao *intervalo* de tempo durante o qual se registam acontecimentos de "doença". Um estudo longitudinal bem planeado é o que de melhor se pode fazer experimentalmente em epidemiologia. Adopta-se o princípio de que a exposição ao factor de risco deve *anteceder* o desenvolvimento de doença para que não haja dúvidas de que a doença é, de alguma forma, consequência da exposição. Uma vez que num estudo de coortes, à partida, ninguém está doente, existe uma boa aproximação a este princípio. Além disso, como a coorte é acompanhada ao longo do tempo, pode-se estimar a taxa de incidência, um conceito que se apresenta adiante.

Os estudos de coortes distinguem-se dos de caso-controlo, porque em coortes os dois grupos de pessoas a comparar são formados com base no critério de estar ou não estar exposto ao factor de risco, antes mesmo de qualquer pessoa estar doente. Nos estudos caso-controlo, o processo é exactamente o contrário – o critério de formação dos grupos é estar ou não estar doente, antes de se saber se foram expostos ao factor de risco. Vejamos o seguinte exemplo.

Exemplo 5.1 (de Selwyn *et al* 1989)

Num estudo em que se pretende investigar se existe associação entre o HIV e o desenvolvimento de tuberculose (TB), tomou-se amostras de sangue de 513 toxicodependentes intravenosos, as quais foram testadas para a presença de anticorpos contra o HIV; 215 eram HIV-positivos e 298 eram HIV-negativos. Estes dois grupos (ou coortes) foram acompanhados para ver se desenvolviam sintomas de TB activa durante um período médio de 2 anos. Os resultados do estudo foram os seguintes:

	Desenvolveu TB ?		
Estado inicial	sim	não	total
HIV +	8	207	215
HIV -	0	298	298

Neste exemplo, os toxicodependentes seronegativos (HIV–) são usados como grupo para comparação com o grupo que tem o factor de risco (HIV+). São, respectivamente, os não expostos e os expostos. Os dois grupos iniciais são formados por indivíduos que não possuem a doença a estudar. Para isso, devem ser excluídos à partida todos os indivíduos com sintomas de tuberculose antes do estudo ter início. Os dois grupos são então seguidos

tratamento fazer efeito, um indivíduo mudar de um estado fisiológico para outro, morrer etc.

tendo-se a certeza que o factor de risco antecede o aparecimento de doença, o que é fundamental se se pretende estabelecer uma relação de causalidade. Durante o seguimento, regista-se a incidência da doença em ambos os grupos.

O estudo descrito neste exemplo é uma experiência "natural", uma vez que o investigador tira partido de quem está e quem não está exposto. Numa experiência totalmente controlada, seria o próprio investigador a decidir quem deve ser e quem não deve ser exposto. Para uma experiência de coortes totalmente controlada, é reservada a designação de *ensaio clínico*, assunto que será tratado em outro módulo. Pode-se argumentar que um estudo 'natural' requer maior criatividade que uma experiência totalmente controlada, uma vez que o investigador tem de se aperceber da possibilidade de poder capitalizar uma situação de exposição natural já existente. Nos estudos de coortes, a exposição das pessoas ao factor de risco pode ser conhecida de imediato. Por exemplo, pode-se saber desde o primeiro momento quem trabalha num local considerado de risco. Mas pode também haver necessidade de investigar primeiro quem está e quem não está exposto, como foi o caso do Exemplo 5.1. Em geral, os estudos de coortes são mais fáceis de conduzir se o tempo decorrido entre a exposição e o desenvolvimento da doença for curto. Um exemplo seria a investigação da associação entre a infecção com o vírus da rubéola durante a gravidez e o desenvolvimento de malformações congénitas no feto. Um exemplo contrário são as doenças de longa latência, como o HIV ou a tuberculose, as quais obrigam a estudos de coortes muito longos.

Variantes ao plano tradicional

Há algumas variantes ao plano básico dos estudos de coortes, usadas, em geral, por apresentarem vantagens financeiras ou de poupança de tempo. Por exemplo, nada impede que as mesmas pessoas que se subdividem em dois grupos para comparação (expostos e não expostos a um factor de risco) se subdividam de forma diferente quando se pretende considera outro factor de risco em simultâneo, definindo-se diferentes grupos de pessoas para cada factor.

Num segundo exemplo, em vez do grupo de controlo (os não expostos) pode-se usar um grupo externo, de carácter muito geral. Um grupo externo muito usado é a própria população onde se insere o grupo exposto ao factor de risco. Por exemplo, o grupo exposto podem ser os indivíduos de uma profissão e o grupo de controlo pode ser toda a população. A vantagem desta variante é que basta monitorizar o grupo exposto, mantendo-se a possibilidade de comparação, mas neste caso com a população geral. O principal problema

desta abordagem é que as estatísticas oficiais da população, em geral, não têm informação tão detalhada como aquela que é medida no grupo exposto ao risco, nomeadamente sobre possíveis variáveis de confundimento.

Por vezes é possível reconstruir as características relevantes de grupos de indivíduos existentes no passado, nomeadamente se estiveram ou não expostos ao factor de risco, e depois reconstruir o seu percurso até ao presente. Esta abordagem tem a vantagem de não obrigar a esperar muitos anos para completar o estudo. Estes estudos designam-se por **retrospectivos** e são também estudos de coortes, uma vez que acompanham grupos de indivíduos ao longo da sua vida, embora o façam "desde lá de trás", razão pela qual prefiro não usar o termo 'prospectivo' de forma generalizada a todos os estudos de coortes. Mais apropriado é considerar que os estudos de coortes se dividem em prospectivos, retrospectivos e mistos, uma vez que é possível combinar um estudo retrospectivo com um prospectivo. Para a abordagem retrospectiva funcionar, é indispensável que existam dados completos. Por exemplo, para estudar a relação entre comportamento sexual e infecção por HIV, não é aceitável limitarmo-nos a pedir a uma amostra de pessoas que recordem o seu comportamento sexual há 10 anos, e depois comparar a prevalência do HIV entre os diferentes tipos de comportamento. Para além das deficiências de memória que as pessoas têm, pode acontecer que grande parte dos que tinham um comportamento particular já tenha morrido.

5.2 Análise de uma coorte fixa

Se o estudo se iniciar com todos os indivíduos ao mesmo tempo e forem seguidos durante o mesmo intervalo de tempo, está-se numa situação dita de **coorte fixa** e a análise pode ser feita como descrito no módulo sobre risco – uma análise de risco tradicional. Situações de **coorte variável**, nas quais um grupo de indivíduos muda ao longo do estudo, fazendo com que os tempos de seguimento não sejam iguais para todos os indivíduos, devem ser analisadas utilizando taxas de incidência em vez de risco, tal como descrevo adiante. Mas vejamos para já o caso da coorte fixa retomando o Exemplo 5.1.

Exemplo 5.1 (continuação)
Retome-se a tabela de dados, agora com todos os totais marginais,

	Desenvolveu TB ?		
Estado inicial	sim	não	total
HIV +	8	207	215
HIV -	0	298	298
	8	505	513

O risco de desenvolver TB entre os seropositivos foi 8/215 = 0.04 e entre seronegativos foi 0/298 = 0. Neste exemplo particular, o risco relativo (0.04/0) e o *odds ratio* não podem ser calculados, dado que implicam uma divisão por zero.

Neste exemplo é apropriado calcular risco e RR como habitualmente, pois as coortes de indivíduos são acompanhadas ao longo de um intervalo de tempo relativamente curto, durante o qual ocorrem os episódios de doença e *todos os indivíduos chegam ao fim desse intervalo*. O risco é a proporção de indivíduos que adoeceram na respectiva categoria de exposição (HIV^+, HIV^-). Neste exemplo, o RR só não foi calculado por haver uma divisão por zero.

Se o tempo de acompanhamento das coortes fosse suficientemente longo para se fazerem sentir riscos em competição, por exemplo o risco de morte por causas que nada têm a ver com a exposição ao factor sob estudo, teria havido desistências de alguns indivíduos e o tamanho da amostra de indivíduos diminuiria ao longo do estudo. Nesta situação seria necessário calcular taxas de incidência em vez de risco.

5.3 Riscos em competição, pessoas-tempo e taxa de incidência

Riscos em competição

Suponhamos que estamos interessados em estudar se a vacina contra o HPV (human papiloma virus) é eficaz a conferir protecção contra o cancro do colo do útero[2]. Uma vez que os casos de cancro são raros, planeamos acompanhar uma grande amostra de N=10000 mulheres que à partida são seronegativas para o HPV, administrando a vacina contra o HPV a 5000 mulheres (as não-expostas) e administrando uma outra vacina contra as restantes 5000 (as expostas, ou grupo de controlo). Pretendemos acompanhar estas 10000 mulheres ao longo de 5 anos, a fim de verificar se o risco de lesões cancerígenas surgidas nas vacinadas para HPV é inferior ao risco nas vacinadas para outra infecção.

Será exequível recrutar todas as mulheres para este estudo ao mesmo tempo ? será

[2] O HPV pode ser contraído por contacto cutâneo (em geral relações sexuais). Na maioria dos casos origina verrugas ou alterações citológicas benignas, mas alguns tipos do vírus causam lesões neoplásicas nas

que todas serão seguidas durante o mesmo tempo ? é muito provável que a resposta a estas perguntas seja negativa. Provavelmente as 10000 mulheres vão entrar e sair do estudo em alturas diferentes e, consequentemente, serão acompanhadas durante intervalos de tempo diferentes. Poucas serão seguidas durante 5 anos. O que fazer com uma mulher que permaneça sem doença mas morra por outra razão um ano após ter entrado no estudo? Se a contarmos entre as não-doentes estamos a contribuir para subestimar o risco de doença, pois não temos forma de saber se ela viria a adoecer nos 4 anos seguintes. Se não contarmos com ela, estamos a ignorar o facto de ela ter-se mantido sem doença durante o ano em que foi seguida.

A saída de pessoas do estudo por razões alheias à doença que estamos a estudar, é consequência daquilo que por vezes é designado por *riscos em competição* (Rothman 2002). Trata-se dos riscos de saída do estudo, por razões que nada têm a ver com o fenómeno em estudo – por exemplo, a morte do indivíduo, a sua perca de contacto devido a mudança de morada, a aquisição de um estado de saúde que impede a ocorrência da doença em estudo, etc.. Num curto período de tempo, os riscos em competição têm pouca influência e são em geral negligenciados. Porém, quando o período de seguimento das coortes é longo, as consequências dos riscos em competição aumentam e não podem ser ignorados.

Nos estudos de coortes, a medição do risco de doença por meio da proporção de incidência, tal como é usada em estudos transversais, confronta-se com uma dificuldade séria. Se o período de tempo considerado fôr suficientemente longo, e/ou se houver muitas pessoas envolvidas, não é possível medir risco como fizemos até aqui, pela simples razão de que há pessoas a entrar e sair do estudo em alturas diferentes, por motivos que nada têm a ver com a doença em estudo. Umas razões prendem-se com os diferentes *timings* de recrutamento de pessoas ao estudo, outras prendem-se com os riscos em competição.

Pessoas-tempo e taxa de incidência

Para lidar com o problema colocado por riscos em competição, os epidemiologistas recorrem ao conceito de taxa de incidência, também conhecida por taxa pessoas-tempo. Na **taxa de incidência** (TI) divide-se a incidência cumulativa (X) por um denominador que mede tempo (Z). Z é o tempo total, somado indivíduo a indivíduo, durante o qual os indivíduos acompanhados estão em risco de contrair a doença.

mulheres, as quais podem progredir para cancro do colo do útero.

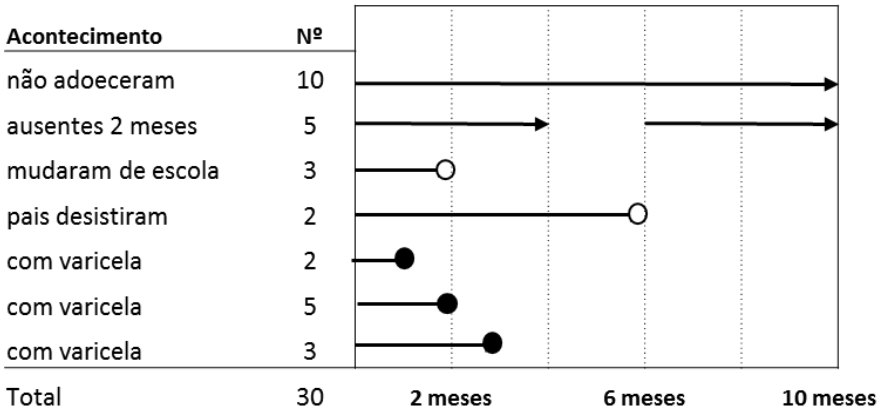
$$Taxa\ de\ incidência = \frac{Incidência}{Soma\ dos\ tempos\ de\ exposição\ à\ doença} = \frac{X}{Z} \qquad [5.1]$$

Suponhamos, por exemplo, que pretendemos conhecer o risco de contrair varicela numa escola, acompanhando ao longo do ano lectivo um grupo de 30 crianças que nunca antes teve varicela. A varicela é uma doença causada pelo vírus herpes zoster, para o qual não é administrada vacinação, razão pela qual a maioria das crianças contrai a doença antes de completar 10 anos de idade. Suponhamos que a maioria das crianças que integram o nosso estudo permanecem todo o ano na escola, mas algumas mudam para outra escola, outras ausentam-se temporariamente devido a outras doenças, e outras desistem do estudo porque os pais não têm mais paciência para responder às nossas perguntas. Suponhamos que no fim do ano lectivo somámos 10 casos de varicela, os quais serão colocados no numerador da taxa de incidência. Mas que número deve ser colocado no denominador?

As crianças que desistiram do estudo por mudar de escola ou porque os pais perderam a paciência não contrairam varicela enquanto foram acompanhadas. Se não as colocarmos no denominador, iremos sobreestimar a taxa de incidência. Mas não podem ser colocadas no denominador em pé de igualdade com crianças que permaneceram no estudo durante todo o ano. E como ter em atenção as crianças que se ausentaram temporariamente devido a outras doenças ?

Para lidar com estes problemas, os epidemiologistas recorrem ao conceito de **pessoas-tempo**. Cada indivíduo é multiplicado pelo tempo que permaneceu no estudo *sem contrair* a doença em estudo e estes produtos são adicionados para obter a quantidade total de pessoas-tempo acompanhadas. Este total será colocado no denominador. Completemos o nosso exemplo, agora com os cálculos.

Exemplo 5.2: O estudo iniciou-se com 30 crianças, das quais 10 foram acompanhadas durante 10 meses e nunca adoeceram, 5 estiveram ausentes 2 meses devido a outras doenças que não a varicela, 3 foram acompanhadas só durante 2 meses e abandonaram o estudo porque mudaram de escola, os pais de 2 crianças também abandonaram a autorização para o estudo ao fim de 6 meses e 10 crianças contrairam varicela após 1 mês (2 crianças), 2 meses (5 crianças) e 3 meses (3 crianças). A figura ilustra o estudo:



As linhas horizontais representam o percurso temporal (em meses) das crianças. As setas indicam interrupção do percurso por decisão do investigador (censura), sem que tenha havido varicela ou abandono. Quando as linhas terminam em bolas a cheio, o percurso parou devido a ocorrência de varicela. As bolas brancas indicam abandonos do estudo sem se ter tido varicela. O total de pessoas-tempo é calculado multiplicando os indivíduos pelo seu tempo de permanência no estudo e adicionando tudo, como a tabela indica,

Nº crianças		Meses		Pessoas-tempo
10	x	10	=	100
5	x	8	=	40
3	x	2	=	6
2	x	6	=	12
2	x	1	=	2
5	x	2	=	10
3	x	3	=	9
		Total	=	179

Conclui-se que houve um total de 179 pessoas-tempo ou, neste caso, crianças-mês. Dividindo a incidência de varicela pelas pessoas-tempo, obtem-se a taxa de incidência de varicela neste estudo: 0,0559 (=10/179) por crianças-mês.

_____

Formalizemos um pouco melhor o que fizemos. Para calcular a **taxa de incidência** divide-se a incidência absoluta pelo somatório dos tempos passados por cada indivíduo no estado de "poder contrair a doença". Os epidemiologistas em geral substituem esta última expressão por "estar exposto à doença":

$$Taxa\ de\ incidência = \frac{Incidência}{Soma\ dos\ tempos\ de\ exposição\ à\ doença} = \frac{X}{Z} \quad [5.1]$$

Onde o numerador e o denominador se referem ao mesmo intervalo de tempo.

A contagem do tempo de exposição à doença é feita indivíduo a indivíduo e tem sempre em consideração as características biológicas da doença. No caso da varicela, quando uma criança contrai a doença a contagem do seu tempo de exposição pára e, mesmo depois de regressar à escola já curada, não conta mais para o estudo. Ter varicela confere imunidade contra a reinfecção e, por isso, não se pode considerar que essa criança continua exposta à doença. Contudo, se se tratasse de medir a taxa de incidência de uma doença recorrente, como o ataque de asma, uma reacção alérgica, ou o herpes labial, o mesmo indivíduo poderia continuar a ser acompanhado depois de se considerar estar recuperado do episódio de doença anterior. Em alternativa, o investigador pode calcular uma taxa de incidência para o 1º episódio de doença, no qual o numerador só conta os primeiros casos, e pode calcular outra taxa de incidência para o 2º episódio, onde o numerador só conta segundos episódios e o denominador fica limitado ao acompanhamento dos que já tiveram a doença uma vez.

5.4 Interpretação da taxa de incidência

Interpretação e unidades

A taxa de incidência é expressa em termos de número de casos por pessoas-tempo. Pode também ser interpretada como número de indivíduos (que adoeceu), por indivíduo, por unidade de tempo. A taxa é portanto uma quantidade *per capita* e por unidade de tempo. Como o número de indivíduos não tem unidades, as unidades da taxa são "por unidade de tempo". No exemplo 5.2, obteve-se 0,0559 casos por crianças-mês. Isto é o mesmo que dizer 0,0559 casos por criança por mês. Como as crianças não têm unidades, podemos simplesmente escrever 0,0559 $mês^{-1}$, as unidades são "por mês".

De um modo geral, uma taxa de incidência tem unidades "por unidade de tempo", ou simplesmente $tempo^{-1}$. Esta unidade pode parecer um pouco abstracta, mas é característica habitual das taxas. Voltaremos a este assunto mais adiante, para discutir o significado do inverso da taxa de incidência.

Por 1000 ou por 10^5

Concluimos que, por mês, contraem varicela 0,0559 crianças por cada criança. Isto soará porventura estranho, porque estamos a dizer que uma fracção de criança será convertida em doente no decorrer de 1 mês. Talvez nos sintamos mais confortáveis se multiplicarmos a taxa de incidência por 1000. Poderemos então dizer que, por cada 1000 crianças, aproximadamente 56 contraem varicela em 1 mês. Frequentemente as taxas de incidência são apresentadas por 1000 ou 100 mil pessoas, para que o seu significado seja mais intuitivo.

O denominador e as estatísticas oficiais

A epidemiologia investiga os níveis de risco de doença a que diferentes grupos de indivíduos estão sujeitos. As medidas de risco baseiam-se sempre num quociente entre o número de ocorrências (o numerador) e o número de indivíduos expostos num determinado intervalo de tempo (o denominador). Contudo, há várias escolhas possíveis para o denominador e estas podem originar resultados muito divergentes. O denominador deve ter em atenção o tempo total de acompanhamento dos indivíduos e as alterações na constituição do grupo de indivíduos expostos ao longo do tempo.

É comum encontrar nas estatísticas de saúde publicadas por organismos estatais, taxas de incidência anuais expressas na forma de, por exemplo, 50 casos por 100 mil. Obtiveram-se dividindo o total de casos ocorridos durante o ano pelo número de pessoas que se estimou existirem a meio do ano (ou no inicio do ano) e multiplicando o resultado por 100 mil. As estatísticas denominam-nas de “incidência anual”, uma terminologia que suscita dúvidas. Poderão considerar-se uma aproximação à taxa de incidência, se se assumir que o denominador representa a observação de 100 mil pessoas-ano. Seria neste exemplo uma taxa de 50 por 100 mil pessoas-ano, ou 0,0005 por pessoa-ano. Contudo, o tamanho da população em denominador certamente não se manteve constante durante o ano e em geral não houve contabilização do tempo de exposição de cada indivíduo, pelo que a aproximação é de validade discutível.

A taxa de incidência não é um risco

Ao contrário do risco, a taxa de incidência não pode ser interpretado como uma probabilidade (Tabela 1). A taxa de incidência não está superiormente limitada por 1 e, teoricamente, pode ter um valor infinitamente grande. Pode parecer estranho que uma medida de ocorrência de doença possa exceder 1, como é possível que mais de 100% da

população seja afectada? A resposta é que a taxa de incidência *não é uma proporção*, logo esta pergunta está mal formulada. Como o denominador é medido em unidades de tempo, podemos conceber um denominador cada vez mais pequeno, manipulando estas unidades, e fazendo com que a taxa seja progressivamente maior. No exemplo 5.2, se usarmos como unidade de tempo o ano (12 meses) em vez do mês, o denominador passa a ser 14,92 pessoas-tempo (=179/12) e a taxa de incidência de varicela passa a ser igual a 0,67 (=10/14,92) crianças-ano. O valor numérico da taxa depende portanto da unidade de tempo que o investigador decide utilizar. Se se escolher usar uma unidade de tempo suficientemente grande (ano, década, século...), a taxa pode ultrapassar 1.

	Proporção de Incidência ou Risco	**Taxa de Incidência**
Expressão	Incidência cumulativa / Núm indivíduos expostos	Incidência cumulativa/ Tempo total de exposição
Mínimo	0	0
Máximo	1	infinito
Unidades	sem unidades	$tempo^{-1}$
Interpretação	Probabilidade	Inverso do tempo de espera

Tabela 1. Comparação entre o risco e a taxa de incidência

Carácter instantâneo da taxa de incidência

É frequente efectuar uma analogia entre a taxa de incidência e a velocidade. A taxa de incidência, tal como a velocidade, mede algo que se passa instantâneamente e, tal como a velocidade, pode ser aproximada por uma média calculada de forma não instantânea. Imaginemos um automóvel na estrada. Em qualquer instante, o automóvel tem uma certa velocidade e o velocímetro do carro dá uma medida contínua desta velocidade instantânea, embora a expresse para um período de tempo não-instantâneo, em geral em Kilómetros por hora (Km/h). O velocímetro divide continuamente a distância percorrida pelo carro num intervalo de tempo muito pequeno por esse mesmo intervalo, e depois converte o resultado em Km/h.

Embora a unidade de tempo do denominador seja 1h, não é necessário esperar 1h

para saber a velocidade a que o carro vai e o resultado *numérico* seria diferente se se usasse outra unidade de tempo. Por exemplo, é o mesmo dizer que o carro se desloca a 100 Km/h ou a 1,667 km/minuto (1,667=100/60min). A taxa de incidência também mede a incidência instantânea a que os casos de doença estão a ocorrer num grupo de pessoas, embora na prática seja calculada usando uma unidade de tempo grande (mês, ano, etc.). Para medir a taxa é necessário acompanhar pessoas em intervalos não-instantâneos de tempo mas, tal como a velocidade do carro, devemos pensar na taxa como algo que se aplica continuamente em cada instante de tempo, é uma taxa instantânea.

O valor numérico da taxa de incidência, só por si, não pode ser interpretado. É indispensavel saber qual a unidade de tempo usada para obter esse valor. Muitas vezes, o investigador escolhe a unidade de forma a forçar que o valor numérico da taxa seja maior que 1. Por exemplo, se a taxa for 0,004 casos por pessoas-dia, pode ser multiplicada por 1000 para ser apresentada como 4 casos por 1000 pessoas-dia ou pode ser apresentada em termos de 1,46 casos por pessoas-ano (1,46=0,004 x 365dias). Esta unidade pode ser usada, quer as observações tenham sido recolhidas durante 1 dia, 1 semana, ou 1 ano, da mesma forma que podemos medir a velocidade de um carro em termos de Km/hora, mesmo que o velocímetro esteja a efectuar a medição durante breves segundos.

O inverso da TI - tempo médio de espera

Enquanto o risco é facilmente entendido, desde que o intervalo de tempo a que se refere esteja claramente definido, a taxa de incidência apresenta portanto mais dificuldades. Há, contudo, mais uma forma de a interpretar. Como a unidade da TI é “por tempo”, caso invertamos a TI, obtemos uma quantidade cuja unidade é “tempo”. Qual o significado desta quantidade de tempo ? Se a taxa de incidência permanecer aproximadamente constante durante algum tempo, o inverso da taxa de incidência é o tempo médio que leva até que o acontecimento ocorra pela primeira vez. Este tempo é o chamado “tempo médio de espera” pela primeira ocorrência de doença.

Tome-se o exemplo dado acima em que TI=3,57 casos de doença por pessoas-ano. O seu inverso é 1/3,57 = 0,28 anos. Este valor deve ser interpretado como o tempo que, em média, se tem de esperar até aparecer o primeiro caso de doença. Num segundo exemplo, considere-se a taxa de mortalidade. Por exemplo, 11 mortes por 1000 pessoas-anos ou 11/1000 ano^{-1}. Se isto é a taxa de mortalidade de uma população, o seu inverso é o tempo que, em média, um recém-nascido espera até à morte, habitualmente conhecido por

esperança média de vida à nascença ou longevidade média dos indivíduos. Calculando o recíproco, obtem-se 90,9 anos. Isto, repito, assumindo que a taxa de mortalidade se mantem constante ao longo dos anos (o que é pouco provavel que aconteça ao longo de uma escala temporal tão longa).

Acontecimentos repetidos

Em geral a TI contabiliza no numerador apenas a primeira ocorrência de doença em cada indivíduo. Para muitas doenças, como as que originam imunidade, as doenças crónicas e a propria morte, cada indivíduo só pode ter a doença uma vez. Para doenças que se repetem, como a renite alérgica, o herpes labial, ou uma infecção com macroparasitas, podemos medir apenas a primeira ocorrência, ou então a primeira ocorrência após um período pré-definido livre de doença. Se o investigador pretende contar todos os casos de doença, incluindo repetições na mesma pessoa, em geral existem razões biomédicas suficientes para distinguir o 1º episódio de doença do 2º episódio, etc.. Para o 1º episódio de doença, o denominador da TI contabilizará todos os indíviduos expostos que ainda não tiveram a doença; para a TI do 2º episódio de doença, o denominador fica limitado aos que já tiveram a doença uma vez, etc. Por outro lado, estes últimos deixam de contribuir com tempo para o denominador da TI do 1º episódio de doença.

5.5 Relação entre risco e taxa de incidência

Como vimos, a taxa de incidência mede a 'velocidade' com que a doença incide sobre um conjunto de indivíduos. É intuitivo que o risco de um indivíduo contrair a doença num intervalo de tempo deve ser tanto maior quanto maior a taxa que actua ao longo desse intervalo e quanto mais longo o próprio intervalo. Se se designar por Δt a duração do intervalo de tempo e por λ o valor da taxa de incidência durante Δt, então o risco, r, de contrair a doença, é dado por:

$$r = 1 - e^{-\lambda \Delta t} \qquad [5.2]$$

Onde *e* é a base dos logaritmos neperianos (*e=2,71828).* A equação [5.2] pressupõe que a taxa λ permanece constante ao longo do intervalo Δt. Se tal não acontecer, o valor numérico de λ a utilizar em [5.2] deve ser a média da taxa que actua durante Δt. A equação mostra que quanto mais elevada fôr a taxa ou o tempo durante o qual ela actua, maior deve ser o

risco, embora o valor de r nunca possa ser maior que 1, o que está de acordo com o facto de r ser uma probabilidade. Se, pelo contrário, a taxa for nula ($\lambda\Delta$t =0) então a equação mostra que os risco de doença é nulo, r=0. A dedução da equação [5.2] está na caixa [Teoria 2].

Exemplo 5.3

Suponhamos que numa população se regista uma taxa de incidência de cancro do pulmão de 8 casos por 10 mil pessoas-ano. A expressão [5.2] indica que o risco médio de contracção deste cancro em 1 ano deveria ser aproximadamente igual a

$$r = 1 - e^{-8/10000 \times 1} = 0{,}0008$$

O risco pode também ser calculado para seis meses. O expoente altera-se e o risco é recalculado:

$$r = 1 - e^{-8/10000 \times 0.5} = 0{,}0004$$

o risco diminuiu quando o tempo de exposição diminuiu, como seria de esperar.

Exemplo 5.4

Suponhamos que, em determinada população, a partir dos 50 anos de idade os indivíduos se caracterizam por uma taxa de mortalidade de 11 mortes por 1000 pessoas-ano. Assumindo que esta taxa se mantém constante a partir dos 50 anos, qual o risco de morte até aos 70 anos ? O intervalo de tempo em causa é de 20 anos (=70-50) e de acordo com [5.2] o risco de morte é:

$$r = 1 - e^{-11/1000 \times 20} = 0{,}198$$

O resultado indica que por cada 1000 pessoas vivas com 50 anos de idade, 198 morrerão durante os 20 anos seguintes.

Exemplo 5.5

Num grupo de 100 pessoas, registou-se um caso de gripe em uma semana. Admitindo que esta taxa de incidência se mantém constante, quantos casos de gripe se esperam em 1 milhão de indivíduos por dia ? Expressemos a taxa obervada em termos de incidência por pessoas-dia:

1 caso /(100 pessoas x 7 dias) = 1 caso/700 pessoas-dia = 0,0014 casos por pessoas-dia

Se se esperam 0,0014 casos por pessoa por dia, então em 1 milhão de pessoas em 1 dia, esperam-se,

1000000 pessoas-dia x 0,0014 dia-1 = 1429 casos

Exemplo 5.6 (adaptado de Rothman 2002)

Há um velho quebra-cabeças que pergunta: “Se uma galinha e meia põe 1 ovo e meio em 1 dia e meio, quantos ovos põe uma galinha em 1 dia ?”

Resposta: 2/3 de ovo. Porquê ?

A forma como a galinha põe ovos pode ser representada em termos de taxa de “incidência dos ovos”, usando como unidade “galinhas-dia”. Usando os dados fornecidos, a incidência é de 1,5 ovos e o denominador é 2,25 galinhas-dia (=1,5 galinhas x 1,5 dias). Logo, a taxa é 1,5/2,25 = 0,667 ovos por galinha por dia, ou seja, 2/3 de ovo.

Os cálculos nestes exemplos assumem que λ, a taxa de incidência, permanece constante ao longo do intervalo de tempo considerado. O que fazer se λ mudar ao longo do tempo, como provavelmente acontece no mundo real ? Pode-se ainda calcular risco por meio de [5.2], mas é aconselhável fazê-lo para intervalos de tempo pequenos, dentro dos quais λ possa ser considerada aproximadamente constante. Teoricamente, quanto mais pequenos os intervalos melhor, mas também não podem ser tão pequenos que não haja observações de casos de doença em número suficiente para calcular λ dentro de cada intervalo. Existe toda uma panóplia de técnicas para lidar com este problema e voltaremos ao assunto a propósito dos estudos longitudinais em epidemiologia.

TEORIA 2

O risco de contrair doença resulta da acção contínua de uma taxa de incidência (λ) que, ao longo do tempo, incide sobre um grupo de indivíduos susceptíveis de desenvolver a doença. À medida que o tempo passa, o risco aumenta, tendendo para o limite em que r=1. A representação matemática deste fenómeno deve ter em consideração as suas características contínuas no tempo. Designemos por S_t o número de indivíduos susceptíveis à doença e pensemos na forma como este número varia à medida que o tempo passa. Quantos indivíduos ainda susceptíveis teremos ao fim de algum tempo ? Matematicamente, a variação de S_t à medida que o tempo passa, é a derivada de S_t em ordem ao tempo, representada por:

$$\frac{dS_t}{dt}$$

Se não adicionarmos novos susceptíveis, e se S_t fôr um número grande, a variação de S_t ao longo do tempo só pode ser negativa, uma vez que S_t vai diminuindo devido ao progressivo adoecimento dos indivíduos. Esta variação deve ser proporcional ao próprio S_t, por outras palavras, quanto mais indivíduos susceptíveis há mais devem adoecer. Representando por λ o coeficiente de proporcionalidade entre a variação e S_t, obtemos uma equação muito simples para representar a variação do número de susceptíveis,

$$\frac{dS_t}{dt} = -\lambda S_t \qquad \text{[T2-1]}$$

A equação [T2-1] é uma equação diferencial ordinária, uma vez que tem uma variável dependente (o número de susceptíveis, S), uma única variável independente (o tempo, t) e a derivada da primeira em ordem à segunda. O lado esquerdo da equação representa a variação instantânea dos susceptíveis à medida que o tempo passa. Se isolarmos λ na equação, poderemos compreender melhor as suas unidades,

$$\lambda = -\frac{1}{S_t}\frac{dS_t}{dt}$$

As unidades de λ devem ser iguais às do termo à direita do sinal igual, ou seja, número de susceptíveis por susceptível por unidade de tempo. Note-se que são as mesmas unidades já atrás definidas para a taxa instantânea de incidência – número de casos por pessoas-tempo – onde os casos não são mais do que os susceptíveis que adoecem e se tornam casos. Na verdade, [T2-1] é uma forma de definir em termos teóricos a taxa instantânea de incidência a qual, anteriormente, aprendemos a calcular em termos empíricos.

A equação [T2-1] não permite responder directamente à seguinte pergunta – se no início do intervalo de tempo (t, t+Δt) houver S_t indivíduos susceptíveis à doença, quantos susceptíveis ainda teremos ao fim de Δt tempo ? contudo, a solução matemática da equação diferencial conduz-nos à resposta. A solução é:

$$S_{t+\Delta t} = S_t e^{-\int_t^{t+\Delta t} \lambda_u \, du}$$

O número de susceptíveis no fim do intervalo, $S_{t+\Delta t}$, depende do número que havia no início e da acção acumulada da taxa λ ao longo do intervalo de tempo, representada pelo integral em expoente. Se assumirmos que λ se mantem constante durante o intervalo, o integral pode ser calculado e a solução simplifica-se:

$$S_{t+\Delta t} = S_t e^{-\lambda \Delta t} \quad \text{ou seja} \quad \frac{S_{t+\Delta t}}{S_t} = e^{-\lambda \Delta t}$$

O lado esquerdo desta equação é a proporção de susceptíveis que *não* adoeceram, relativamente ao total inicial de susceptíveis – é a probabilidade de não adoecer durante o intervalo (t, t+Δt). A probabilidade de adoecer, ou risco, deve ser então,

$$1 - \frac{S_{t+\Delta t}}{S_t} = 1 - e^{-\lambda \Delta t}$$

O lado esquerdo é o risco de adoecer, que temos representado por r,

$$r = 1 - e^{-\lambda \Delta t}$$

obtendo-se assim a equação [5.2].

O inverso da taxa de incidência é o "tempo médio de espera pela doença"

A taxa de incidência tem unidades "por unidade de tempo", porque envolve divisão directa por uma quantidade de tempo. O inverso da taxa deve então ter unidades de tempo. Por exemplo, suponhamos que a taxa de incidência do sarampo num país é 0,14 casos por pessoas-ano. As unidades desta taxa são "por ano" porque casos e pessoas não têm unidades. Quando calculamos o seu inverso obtemos 7,1 anos (=1/0,14). Qual o significado destes 7,1 anos ? Qual o significado do período de tempo obtido por inversão de uma taxa ?

O inverso da taxa de incidência é o tempo que passa, em média, até que o acontecimento ocorra pela primeira vez. É o chamado "tempo médio de espera" pela primeira ocorrência de doença. Por exemplo, se a taxa de incidência do sarampo é 0,14 ano^{-1}, então o tempo que decorre, em média, entre o nascimento duma criança e a sua infecção pelo sarampo é de 7,1 anos. Dito de outra forma, a idade média com que as crianças contraem sarampo é 7,1 anos. A caixa Teoria 3 explica porquê.

Considere-se um segundo exemplo em que o problema é colocado de forma inversa. Se, em média, o tempo decorrido entre ser-se diagnosticado com imunosupressão pelo HIV e passar-se ao estado de SIDA fôr 24 meses, qual o valor da taxa de incidência de SIDA entre os imunosuprimidos com HIV ? A taxa deve ser igual a 0,0417 $mês^{-1}$ (=1/24meses).

Estes cálculos pressupõem que a taxa de incidência se mantém aproximadamente constante durante todo o período de tempo em causa. Nos nossos exemplos, desde o nascimento até à infecção com sarampo e desde a imunosupressão com HIV até à passagem a SIDA.

Há outra aplicação desta interpretação. O inverso da taxa de mortalidade é a longevidade média, ou seja, o tempo que em média decorre entre o nascimento e a morte. A aplicação directa aos humanos, contudo, é bastante grosseira quando aplicada a períodos de tempo tão longos, porque a taxa de mortalidade varia bastante ao longo da vida. Por exemplo, a longevidade média em Portugal ronda os 79 anos de idade. Se fizermos os seu inverso obtemos 0,01266 ano^{-1}, a taxa de mortalidade média. Contudo, nos países industrializados, a taxa de mortalidade é muito baixa até aos 65 anos e depois cresce exponencialmente, fugindo muito ao pressuposto de se manter constante, pelo que 0,01266 não é uma estimativa fiável.

TEORIA 3

Considere-se um grupo de indivíduos que estão num estado de saúde pré-definido e que, à medida que o tempo passa, mudam desse estado de saúde para outro diferente. Quanto tempo, em

média, permanecem os indivíduos no primeiro estado ? Em epidemiologia é muitas vezes necessário calcular o tempo médio de estadia num estado de saúde. Dois exemplos ilustram o problema. Considere-se, por exemplo, o conjunto de indivíduos que em 2014 contrairam tuberculose. A partir do fim de 2014, à medida que o tempo passa, o número destes indivíduos só pode diminuir e há várias razões para isso. Uns indivíduos recuperam com o tratamento e deixam de estar doentes, outros morrem por razões relacionadas com a doença, outros morrem devido a causas independentes da doença (acidentes, etc.). A questão que se procura responder é – qual foi o tempo que, em média, um indivíduo pertencente à categoria "adoeceu com tuberculose em 2014" permaneceu doente ?

Se o número de indivíduos, N, que forma qualquer dos grupos acima exemplificados, fôr suficientemente grande, pode-se pressupôr que a sua diminuição ao longo do tempo decorre de forma aproximadamente contínua. Pressupondo também que o número de indivíduos que sai do grupo no instante de tempo t é directamente proporcional ao número de indivíduos que lá estava no referido instante, isto é N_t, então o mesmo raciocínio que conduziu à equação [T2-1] permite exprimir a diminuição do número de indivíduos no grupo:

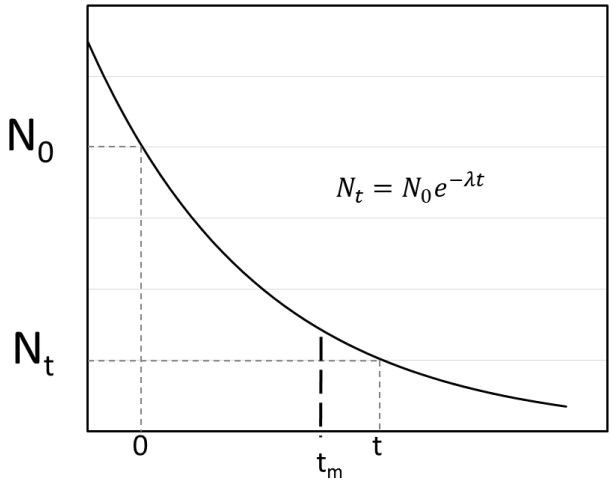
$$\frac{dN_t}{dt} = -\lambda N_t \qquad \text{[T3-1]}$$

O lado esquerdo de [T3-1] é a variação instantânea de N_t, representado pela derivada de N_t em ordem ao tempo. Do lado direito, λ é a taxa instantânea de saída do estado em que os indivíduos foram inicialmente pré-definidos; λ é uma taxa com características idênticas à taxa de incidência, mas pode representar morte, recuperação por tratamento, infecção por microorganismo, ou o que for apropriado. Pode representar o efeito de uma causa ou o efeito combinado de várias causas. Neste último caso seria uma soma: $\lambda = (\lambda_1 + \lambda_2 + \lambda_3 + \ldots)$, onde λ_i representaria o efeito da i-ésima causa de saída do estado pré-definido para os indivíduos. O sinal “menos” do lado direito de [T3-1] assegura que a variação de N_t é negativa, ou seja, ao fim de um intervalo de tempo (t, t+Δt), há menos indivíduos no grupo do que havia no inicio: $(N_{t+\Delta t}-N_t)<0$.

Suponhamos que no instante inicial t=0, há N_0 indivíduos no grupo. A solução da equação diferencial [T3-1] permite calcular o número de indivíduos que ainda permanecem no grupo em qualquer instante t, simbolicamente N_t :

$$N_t = N_0 e^{-\lambda t} \qquad \text{[T3-2]}$$

A equação [T3-2] pressupõe que a taxa λ se manteve constante ao longo de todo o intervalo (0, t) e mostra que, sob este pressuposto, o decréscimo do número de indivíduos se dá de acordo com uma lei exponencial negativa. Ou seja, se uma taxa instantânea de saída de estado actua de forma constante sobre um conjunto de indivíduos, o seu número decresce de forma exponencial, como a figura ilustra.



Em média, os indivíduos estão um certo tempo dentro do grupo, simbólicamente t_m. Este t_m não é igual a metade do período de tempo durante o qual o grupo existe, porque há muito mais indivíduos no inicio que no fim e a diminuição é exponencial. Tem de ser calculado como uma média ponderada. Cada período de tempo decorrido desde t=0 deve ser ponderado (multiplicado) pelo número de indivíduos que existe ao fim desse período: tN_t . O somatório destes produtos é então dividido pelo total de indivíduos que estiveram presentes no grupo ao longo de todo o tempo. Uma vez que o tempo é uma variável contínua, matemáticamente a média ponderada é calculada da seguinte forma:

$$t_m = \frac{\int_0^\infty t N_t \, dt}{\int_0^\infty N_t \, dt} = \frac{\int_0^\infty t \, N_0 \, e^{-\lambda t} \, dt}{\int_0^\infty N_0 \, e^{-\lambda t} \, dt}$$

Os integrais são superiormente indefinidos, pois assume-se tratar-se de um decréscimo exponencial de um grupo infinitamente grande, até não haver mais nenhum indivíduo no grupo. Primitivando entre 0 e $+\infty$ e após um pouco de manipulação algébrica, obtem-se:

$$t_m = \frac{1}{\lambda}$$

Conclui-se portanto que o tempo médio de permanência dentro do grupo é dado pelo inverso da taxa instantânea total com que os indivíduos saem do grupo.

Estes cálculos assumem que a taxa de incidência usada na equação [5.2] permanece constante ao longo dos intervalos de tempo considerados. O que fazer se a TI mudar ao longo do tempo, como provavelmente acontece no mundo real ? Nesse caso, pode-se ainda calcular risco, mas é aconselhável fazê-lo para intervalos de tempo pequenos, dentro dos quais a TI possa ser considerada aproximadamente constante. Teoricamente, quanto mais

pequenos os intervalos melhor, mas eles também não podem ser tão pequenos que não haja observações de casos de doença (ou morte) em número suficiente para calcular a TI dentro de cada intervalo. Existe toda uma panóplia de técnicas para lidar com este problema numa área da estatística denominada *Análise de Sobrevivência* (Survival Analysis). Mais abaixo retorno a este assunto.

5.6 Relação entre o risco relativo e as taxas de incidência

Definimos anteriormente o risco relativo (RR) como o quociente entre o risco de doença entre expostos e não-expostos. Como calcular o RR a partir das taxas de incidência de expostos e não expostos? será o RR igual ao quociente entre as taxas de incidência ?

Vimos já que o risco e a taxa de incidência têm unidades diferentes e só são comparáveis através de uma relação não-linear, a equação [5.2]. Designemos o risco nos expostos por r_1 e nos não expostos por r_2. Designemos também as respectivas taxas de incidência por TI_1 e TI_2. De acordo com a equação [5.2], o RR pode ser expresso em termos das taxas de incidência por,

$$RR = \frac{r_1}{r_2} = \frac{1 - e^{TI_1 t_1}}{1 - e^{TI_2 t_2}} \qquad [5.3]$$

onde t_1 e t_2 são, respectivamente, os tempos de exposição de expostos e não-expostos.

A equação [5.3] pode ser simplificada, caso os expoentes (TI_i t_i) sejam quantidades pequenas. É fácil verificar que para um valor de (TI t) pequeno, verifica-se a seguinte igualdade aproximada:

$$1 - e^{-TI\ t} \approx TI\ t$$

Os estudantes mais cépticos podem confirmar experimentando com a máquina de calcular, ou observando a Tabela 2.

TI t	*1-e* $^{TI\ t}$
0.01	0.01
0.05	0.05
0.10	0.10
0.15	0.14
0.20	0.18

Tabela 2. Comparação entre o valor de (TI t) e o resultado da expressão (1-$e^{-TI\ t}$) para valores baixos

de (TI t).

Ou seja, se o produto da taxa instantânea pelo tempo decorrido fôr pequeno (aproximadamente <0.15), a equação [5.3] simplifica-se:

$$RR = \frac{r_1}{r_2} \approx \frac{TI_1 t_1}{TI_2 t_2}$$

Se o tempo usado para medir a incidência cumulativa em expostos e não expostos fôr o mesmo, $t_1=t_2$, obtém-se

$$RR = \frac{r_1}{r_2} \approx \frac{TI_1}{TI_2} \qquad [5.4]$$

Esta relação mostra que o RR é igual ao quociente entre as taxas de incidência *quando o tempo de exposição é pequeno e/ou as taxas de incidência são pequenas*. Para períodos de tempo suficientemente prolongados, os riscos podem tornar-se tão elevados que o quociente entre riscos começa a divergir do quociente entre taxas de incidência. Ao contrário dos riscos, que estão superiormente limitados pelo valor 1, as taxas não têm limite superior (Tabela 1). Quando as taxas de incidência são elevadas e/ou o tempo de exposição é elevado, os riscos em competição fazem sentir o seu efeito e a equação [5.4] torna-se um mau substituto da equação [5.3].

Frequentemente, o termo risco relativo (RR) é usado na literatura epidemiológica para designar o quociente entre taxas de incidência, sem aviso prévio por parte dos autores. Isto pode originar confusão, pelo que é desejável que seja sempre explicado o que está no numerador e denominador do RR.

5.7 Intervalos de confiança para a taxa de incidência e o RR com pessoa-tempo

Intervalo de confiança para a TI com amostra grande

A taxa de incidência dada por [5.1] tem implicacções estatísticas diferentes do conceito de risco, por causa da natureza do denominador Z. Não se está em presença de uma proporção, mas sim de uma verdadeira taxa, porque o denominador tem unidades de "tempo", e a taxa não está limitada superiormente pelo valor 1. É necessário invocar um modelo estatístico que permita conceptualizar a TI adequadamente.

Se os casos de doença ocorrerem aleatória e independentemente ao longo do

tempo, é razoável assumir que estas ocorrências são bem descritas por uma distribuição de Poisson. Contudo, se durante o intervalo pessoa-tempo o número acumulado de casos for elevado (X > 20), a distribuição Normal constitui uma razoável aproximação à distribuição de Poisson, tendo a vantagem de permitir construir intervalos de confiança com muito maior facilidade.

Uma vez que média e variância são idênticas na distribuição de Poisson, um intervalo de confiança para [5.1] baseado na distribuição Normal, assume que X tem distribuição Normal com média μ e variância μ, i.e. $X \sim N(\mu, \mu)$, sendo μ estimada pelo próprio X. Uma vez que Z é uma constante, então a taxa de incidência, TI= X/Z, também tem distribuição aproximadamente Normal, com média μ/Z e variância igual a μ/Z^2, respectivamente estimadas por X/Z e X/Z^2. O erro padrão do risco é então $(X/Z^2)^{1/2}$ e os limites, inferior e superior, do intervalo de 95% para a taxa de incidência são simplesmente:

$$\left(\frac{X}{Z}\right) \pm 1.96\sqrt{\frac{X}{Z^2}} \qquad [5.5]$$

Exemplo

O registo de um hospital oncológico que serve uma grande cidade, permitiu estimar a ocorrência de 8 casos de cancro do colo do útero por 85000 mulheres-ano. A TI é, portanto, TI = 8/85000, o que equivale a 9,4 casos por 100000 mulher-ano. Um intervalo de confiança a 95% para esta taxa é obtido por:

$$\left(\frac{8}{85000}\right) \pm 1.96\sqrt{\frac{8}{85000^2}}$$

Ou seja, um intervalo de [2,46, 13,54] por 85000 mulher-ano, ou ainda, [2,89, 15,93] por 100000 mulher-ano.

Intervalo de confiança para a TI com amostra pequena

Há doenças para as quais é muito difícil observar um número de casos grande em estudos de coortes, mesmo quando se acompanham muitas pessoas durante muito tempo. É, por exemplo, o caso dos estudos de certos tipos de cancro. Se o número de casos for pequeno (X< 20), a aproximação dada pela equação [5.5] não funciona bem. O ideal seria obter limites de confiança de Poisson exactos, mas isso requer a utilização de técnicas iterativas que saem fora do âmbito deste curso (ver, por exemplo, Ahlbon 1993). Existe, contudo, uma expressão aproximativa que é quase tão boa como os métodos exactos mas é

muito mais fácil de utilizar (Rothman 2002). Conhecida por intervalos de Byar, para um intervalo de confiança de 95%, escreve-se assim,

$$\frac{(X+0.5)\left(1-\frac{1}{9(X+0.5)}\pm\frac{1.96}{3}\sqrt{\frac{1}{X+0.5}}\right)^3}{Z} \qquad [5.6]$$

Onde, como habitualmente, Z é a quantidade pessoa-tempo durante a qual foram observados X casos. O valor 1,96, no numerador a seguir ao sinal +/-, corresponde a um intervalo de 95%, e deve ser ajustado caso se pretenda um intervalo com outro nível de confiança (por exemplo, para intervalos de 90% e 99% seria, respectivamente, 1,645 e 2,58. Vejamos como se aplicaria a fórmula de Byar no Exempo da secção anterior, no qual X=8, situação em que a aproximação pela Normal não é ideal.

Exemplo (Contin.)

Registaram-se 8 casos de cancro por 85000 mulheres-ano, o que corresponde a 9,4 casos por 100000 mulher-ano. Pela fórmula de Byar, um intervalo de confiança a 95% para a taxa de incidência é obtido por:

$$\frac{(8+0.5)\left(1-\frac{1}{9(8+0.5)}\pm\frac{1.96}{3}\sqrt{\frac{1}{8+0.5}}\right)^3}{85000}$$

Originando um intervalo de [3,77, 15,10] por 85000 mulher-ano, ou ainda, [4,44, 17,76] por 100000 mulher-ano. Note-se que os intervalos não são simétricos em torno da TI, pois prolongam-se mais para a direita da estimativa pontual da taxa.

Intervalo de confiança para o RR calculado a partir de duas taxas de incidência

Suponhamos agora que se pretende comparar duas taxas, fazendo o seu quociente para estimar o risco relativo como na equação [5.4]. Um grupo 1 de indivíduos está exposto a um factor de risco e um grupo 2 não está exposto. No grupo 1 há X_1 casos de doença contabilizados em Z_1 pessoa-tempo, no grupo 2 há X_2 casos de doença contados em Z_2 pessoa-tempo. O risco relativo é estimado por,

$$RR=\frac{r_1}{r_2}\approx\frac{X_1/Z_1}{X_2/Z_2} \qquad [5.4]$$

Este quociente não pode ser inferior a zero e não tem limite superior. A sua distribuição na amostragem é assimétrica e é pouco dada a aproximações pela distribuição Normal. Um intervalo de confiança será portanto calculado trabalhando com o logaritmo do RR, uma vez que a distribuição do Ln(RR) é melhor aproximada pela Normal. Para calcular os limites de confiança, é necessário ter um estimador do erro padrão do Ln(RR) na amostragem. Assumindo que os casos de doença no contínuum pessoa-tempo é bem descrito por um processo de Poisson, é possível demonstrar (Ahlbon 1993, sec 6.1.2) que a variância do Ln(RR) é aproximativamente dada por:

$$\mathrm{var}\left(Ln\hat{RR}\right) \approx \frac{1}{X_1} + \frac{1}{X_2} \qquad [5.7]$$

Os limites de confiança aproximativos para o Ln(RR) são então obtidos com o erro-padrão,

$$Ln\hat{RR} \pm 1.96\sqrt{\frac{1}{X_1} + \frac{1}{X_2}} \qquad [5.8]$$

Os limites para o RR, obtêm-se deslogaritmizando estes últimos,

$$RR_{inf,sup} = e^{LnRR \pm 1.96\sqrt{1/X_1 + 1/X_2}} \qquad [5.9]$$

Vejamos um exemplo de aplicação.

Exemplo

8 pessoas seropositivas para HBs Ag, o antigénio de superfície da hepatite B, foram seguidas ao longo do tempo, registando-se se adoeciam com cirrose. Os tempos de entrada e saída das pessoas no estudo, bem como o respectivo tempo de seguimento, foram registados como se indica na seguinte tabela, verificando-se que houve seguimento de um total de 2126 pessoa-mês:

Indivíduo	Seguimento Início	Fim	Tempo de seguimento	Doença ?
1	05-Out-56	01-Dez-93	446	sim
2	10-Out-69	31-Dez-97	339	sim
3	10-Jun-79	31-Dez-97	223	não
4	30-Ago-84	28-Set-94	121	não
5	08-Mai-62	08-Jul-91	350	sim
6	01-Nov-66	10-Mai-79	150	sim
7	21-Mar-54	30-Jun-91	447	não
8	08-Jun-61	29-Jul-65	50	sim
		total:	2126	sim= 5 não=3

Houve 3 pessoas que deixaram de ser seguidas (por morte, abandono do estudo, outra razão) sem terem tido cirrose e 5 que tiveram cirrose. Logo, $X_1/Z_1 = 5/2126 = 0.00235$ por pessoa-mês, ou ainda, 2.35 por mil pessoa-mês. Caso se pretendesse ter a taxa numa base diária, bastaria fazer 0.00235/30 = 0.000078 por pessoa-dia.

Um outro grupo de 14 indivíduos HBs Ag-negativos, foi também seguido durante um total de 4725 pessoa-mês, tendo 3 adoecido e 11 não adoecido, a sua taxa de incidência é $X_2/Z_2 = 3/4725 = 0.00063$ por pessoa-mês. O risco relativo é estimado por RR=0.00235/0.00063 = 3.7. Os intervalos de confiança para os riscos (expostos, não-expostos, e total) são obtidos por aplicação de [5.5], originando os seguintes resultados,

	Cirrose	não cirrose	total	risco	L inf	L sup
HBsAg +	5		2126	0.00235	0.00029	0.00441
HBs Ag -	3		4725	0.00063	-0.00008	0.00135
	8		6851	0.00117	0.00036	0.00198

A variância do logaritmo do RR é (1/5+1/3=0.5333) e a sua raiz quadrada é o erro padrão, 0.730. A aplicação de [5.8] e [5.9] conduz aos limites inferior (LI) e superior (LS) de, respectivamente, o Ln(RR) e o próprio RR. Estes últimos são [0.89, 15.5]. A tabela seguinte resume os cálculos,

RR	Ln RR	RaizQ($1/X_1 + 1/X_2$)	LI Ln(RR)	LS (RR)	LI RR	LS RR
3.70	1.309	0.730	-0.122	2.741	0.88521	15.500

A tabela indica que o risco dos seropositivos desenvolverem cirrose é 3.7 vezes superior ao dos seronegativos. Note-se, contudo, que o IC a 95% para o RR, além de ser muito largo, inclui o valor RR=1, sugerindo que não se pode rejeitar a hipótese de não-associação apenas com estes dados. Isto é consequência de o número de indivíduos seguidos (nas células da tabela de contingência) ser bastante pequeno. Se, por exemplo, todos os números nas células interiores da tabela de contingência duplicarem, o RR continua a ser 3.7, mas os limites do IC passam para [1.346, 10.192], o qual é muito mais estreito e já não inclui o valor 1 (experimentar fazer !).

5.8 Análise de uma coorte variável: dados censurados e análise de sobrevivência

Num estudo de coorte é em geral possível que, a meio do seguimento, haja indivíduos que abandonem o estudo antes de ficarem doentes – esses indivíduos dizem-se "**abandonos**". O termo aplica-se apenas a indivíduos que *ainda poderiam* adoecer. Os abandonos causam problemas semelhantes aos diferentes tempos de seguimento. Suponhamos que se pretende efectuar análise de risco tradicional e que estes abandonos acontecem por razões não relacionadas com o factor de risco ou com a doença em estudo mas sim, por exemplo, porque os indivíduos deixaram de ser contactáveis, embora estejam vivos. Há duas formas de lidar com o problema numa análise de risco clássica (i.e. assumindo coorte fixa). Uma é ignorar os abandonos, a outra é contabilizá-los como casos negativos (não doença). No primeiro caso, está-se a ignorar a informação de que eles sobreviveram sem adoecer antes de abandonar e, dessa forma, está-se a sobrestimar o risco. No segundo caso, ignora-se o facto de que eles poderiam ainda vir a ter doença, caso não tivessem abandonado e, dessa forma, está-se a subestimar o risco. Nenhuma das soluções é satisfatória e só são aceitáveis se o número de abandonos for muito pequeno relativamente ao número inicial de indivíduos.

Num estudo de coortes pode também haver indivíduos que ainda não experimentaram a doença e também não abandonaram o estudo, contudo, parte do seu período de estadia na coorte não é utilizado pelo estudo. Estes indivíduos dizem-se "**censurados**"[3]. Por exemplo, num estudo de 20 anos de seguimento de mineiros que se decide terminar hoje, os mineiros que começaram a trabalhar há menos de 20 anos, não abandonaram o estudo e estão de boa saúde, são censurados. O termo aplica-se apenas a indivíduos que *ainda poderiam* adoecer.

A maneira ideal de lidar com abandonos e censuras, não é, evidentemente, ignorá-los. Existe uma vasta área da estatística conhecida sob a designação de **análise de sobrevivência***,* no âmbito da qual existem vários métodos ideais para lidar com coortes variáveis (e.g. Klein and Moeschberger 1997, Woodward 2004, Cap 5). Os métodos da análise de sobrevivência têm outra vantagem interessante. Permitem comparar as probabilidades de um acontecimento ocorrer em qualquer altura, ao longo do período de seguimento dos indivíduos, por oposição ao que se fez, por exemplo, no Exemplo 5.1, em que os indivíduos foram comparados quanto ao desenvolvimento de tuberculose apenas ao fim de 2 anos, no fim do estudo.

Em análise de sobrevivência, o tempo máximo de seguimento dos indivíduos é

[3] O termo "censurados", é muitas vezes usado num sentido mais lato, incluindo nele os abandonos.

subdividido em intervalos mais pequenos, dentro dos quais é calculada a probabilidade de adoecer. Há em geral duas formas de efectuar a subdivisão:

a) O tempo total é subdividido em intervalos arbitrários, em geral de igual duração (por exemplo, meses ou anos), dentro dos quais pode haver vários casos de doença. Se não existirem censurados (abandonos ou censurados propriamente ditos), a metodologia usada costuma ser apelidada de **análise da Life Table da coorte**, por analogia com procedimentos análogos feitos pelos demógrafos. Esta análise *não* distingue os censurados dos casos de doença dentro de cada intervalo. Se houver um número significativo de censurados, estes têm de ser tidos em consideração nos cálculos e o método utilizado é vulgarmente designado por **método actuarial**.

b) O tempo total é subdividido de acordo com os próprios acontecimentos de doença. Cada caso de doença define, simultaneamente, o fim de um intervalo de tempo e o início do intervalo seguinte. As técnicas usadas neste caso costumam ser apelidadas pelo termo geral de **métodos Kaplan-Meier**. Estes métodos são apropriados para lidar com dados censurados. Uma vez que dentro de cada intervalo só há 1 caso de doença, todos os outros indivíduos não sobreviventes são abandonos ou censurados e podem ser tidos em consideração na análise. Nos métodos Kaplan-Meier os indivíduos começam por ser agrupados de acordo com o tempo de seguimento que estiveram no estudo, independentemente dos instantes em que entraram e saíram do estudo. Por exemplo, um indivíduo que começou a ser seguido no início do estudo e foi seguido 2 semanas até ter tido doença ou até ter abandonado, é reunido com um indivíduo que entrou no fim do estudo, foi seguido apenas 2 semanas e foi censurado porque o estudo terminou. A estimação da probabilidade de adoecer em duas semanas, tem em atenção a contribuição destes dois indivíduos, embora eles possam não ter estado em simultâneo no estudo.

Globalmente, estes métodos são demasiado variados e uma discussão muito abrangente dos mesmos está fora do âmbito deste curso[4]. Uma descrição abrangente dos estudos de coortes, no âmbito da epidemiologia de doenças não-transmissíveis, pode ser encontrada em Breslow and Day (1987). Os métodos para lidar com diferentes tempos de seguimento e abandonos cabem no âmbito da análise de sobrevivência (e.g. Klein and Moeschberger 1997, Woodward 2004).

Seguidamente, resumem-se as vantagens e inconvenientes dos estudos de coortes.

[4] Os alunos que tenham de aplicar estas técnicas podem, contudo, encontrar no site da disciplina um texto que descreve o essencial das técnicas de Life Table da coorte, método actuarial e método Kaplan Meier. Ver Tema 5 em **http://webpages.fc.ul.pt/~mcgomes/aulas/ddi/TEMAS/index_temas.html**

5.9 Vantagens e desvantagens dos estudos de coortes

Pode-se já fazer um balanço das principais vantagens e desvantagens dos estudos de coortes no contexto das doenças transmissíveis, por comparação, nomeadamente, com os estudos transversais e com os estudos de caso-controlo.

Vantagens

1. Os estudos de coortes têm em atenção a sequência de acontecimentos, o que é fundamental para estabelecer relações de causalidade: primeiro existe a exposição ao risco e depois surge a doença. À partida, todos os indivíduos estão livres de doença, mas uma parte deles está ou estará sujeita ao factor de risco. Isto contrasta com os estudos em que uma parte dos indivíduos, à partida, já está doente, investigando-se depois se foram expostos ao risco (estudos transversais e caso-controlo).
2. É possível estudar várias doenças simultaneamente. Basta para isso que se registe a incidência dessas doenças durante o seguimento dos indivíduos.

Desvantagens

1. Os estudos de coortes são em geral longos e caros. Requerem o seguimento de muitos indivíduos ao longo, em geral, de vários anos. Este inconveniente é particularmente sério em doenças de longa latência, como a tuberculose, a zona pelo vírus varicela-zoster, o HIV/SIDA etc., ao ponto de ser impraticável para estas doenças, pelo menos com uma só equipe de investigadores.
2. Não são apropriados para doenças raras. Estas doenças requerem o seguimento de um número muito grande de indivíduos à partida, dezenas ou centenas de milhares, o que em geral é impraticável.
3. Pode haver mudanças de comportamento dos indivíduos incluídos no estudo durante o seguimento. Estas mudanças podem ser devidas ao facto de os indivíduos saberem estar a ser seguidos ou por razões independentes do estudo. Exemplos são as mudanças de dieta, de hábitos de higiene, de práticas sexuais, de hábitos de risco como o álcool ou o tabaco, etc.

Literatura Citada

Ahlbon, A, 1993. *Biostatistics for Epidemiologists*. Lewis Publications

Breslow, NE, NE Day. 1987. *Statistical Methods in Cancer Research. Vol II – Thje Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

Cameron DW, JN Simonsen, JD Lourdes et al. 1989. Female-to-male transmission of human immunodeficiency virus type I: risk factors for seroconversion in men. *Lancet* **2**:403-7

Clayton, D, and M Hills. 1993. *Statistical Models in Epidemiology*. Oxford Univ Press, Oxford.

Giesecke, J. 2002 (2nd ed). *Modern Infectious Disease Epidemiology*. Arnold, London.

Klein, JP and ML Moeschberger 1997. *Survival Analysis. Techniques for Censored and Truncated Data*. Springer, NY.

Rothman, KJ. 2002. *Epidemiology: An Introduction*. Oxford Univ Press

Selwyn PA, D Hartel, VA Lewis *et al.* 1989. Prospective study of tuberculosis among intravenous drug users with human immunodeficiency virus infection. *N Engl J Med* **320**:545-50.

Woodward, M. 2004, 2nd Ed. *Epidemiology, Study Design and Data Analysis*. Chapman & Hall