Tratamento Estatístico de Dados em Física Experimental

quarta-feira, 2 de março de 2022 10:58

Fichamento do livro Tratamento Estatístico de Dados em Física Experimental de Otaviano A. M. Helene e Vito R. Vanin.

Capítulo I - Introdução

quarta-feira, 2 de março de 2022 11:13

A) **Origem e tipos de erros**

Erros sistemáticos: Está relacionado com falhas nos equipamentos e método de extração de dados. Eles podem ser eliminados ou reduzidos.
Erros estatísticos: Causados por variações incontroláveis e aleatórias dos instrumentos de medida, condições externas e etc. Ao eliminarmos os erros sistemáticos de uma medida nos sobram os erros estatísticos.

D) **Independência dos Dados**

Para garantirmos que existe uma relação de independência entre os dados obtidos, devemos nos certificar que durante o processo de extração de dados não interfira na extração dos próximos dados, com isso conseguimos obter uma ideia clara de erro experimental dos nossos dados.
Dado: resultado de uma observação da grandeza.
Medida: um conjunto de dados obtidos.
Valor Verdadeiro: valor que representa melhor o resultado de uma medida.

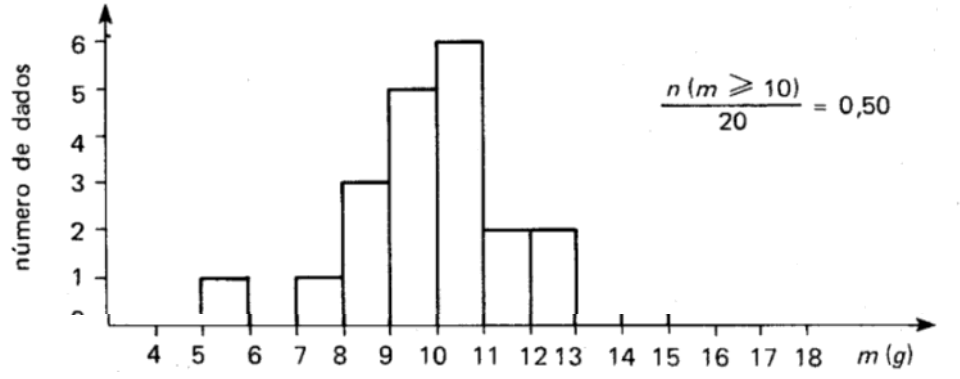
Capítulo II - Função Densidade de Probabilidade

quarta-feira, 2 de março de 2022 11:13

A) **Histogramas**

Histograma: Forma comum de representar uma colação de dados obtidos em uma medida. Normalmente, os dados de um histograma estão distribuídos em torno de uma valor central, que pode ser o nosso valor verdadeiro da medida. Os histogramas tentam se aproximar de um formato de "montanha" e conseguem isso quanto maior for o número de dados usados. O que governa um histograma é a probabilidade de uma dado estar em um determinado intervalo. Assim, há uma certa probabilidade de uma dado obtido estar em um determinado intervalo do histograma.

- *Flutuação estatística*: A variação no número de dados em um intervalo do histograma. Quando falamos que existe uma certa probabilidade de obter um dado num intervalo do histograma estamos esperando essa flutuação estatística - algo que é probabilístico, como o conjunto de observações de uma grandes e que não pode dar origem a histogramas exatamente predeterminados.



Probabilidade: A probabilidade de uma grandeza é igual à relação entre o número de vezes que obtivemos esse resultado dividido pelo número total de dados, quando este é suficientemente grande (tende a infinito). Em linguagem matemática,

$$\left(\begin{array}{c}\text{probabilidade de obtermos}\\ \text{determinado resultado}\end{array}\right) = \begin{array}{c}\text{limite quando o número}\\ \text{de dados tende a infinito}\end{array}\left(\frac{\begin{array}{c}\text{número de vezes que}\\ \text{que obtivemos}\\ \text{determinado resultado}\end{array}}{\text{número de dados}}\right).$$

Há determinadas propriedades da distribuição dos dados obtidos em uma medida que são características da própria medida (ou seja, do objeto da medida e do arranjo experimental usado), onde essas propriedades podem ser representadas por uma *função*.

B) **Função Densidade de Probabilidade**

Função densidade de probabilidade: Função que governa a distribuição de dados em uma experiência, uma forma matemática de representar uma função contínua que governa um conjunto de medidas é dada por: $P(x_1, x_2) = \int_{x_1}^{x_2} F(x)\,\mathrm{d}x$, onde F(X) é a função densidade de probabilidade e $P(x_1, x_2)$ é a probabilidade que um dado qualquer obtido durante a medida pertença ao intervalo $[x_1, x_2]$.
A forma da função densidade de probabilidade depende da grandeza medida e das condições experimentais. Evidentemente a forma do histograma depende não apenas da distribuição que governa a medida, mas também da forma que agrupamos os dados para construir o histograma.

C) **Propriedades da Função Densidade de Probabilidade**

A expressão $P(x_1, x_2) = \int_{x_1}^{x_2} F(x)\,\mathrm{d}x$ nos dá a probabilidade de obtermos um valor dentro de um certo intervalo. Ao somarmos essa probabilidade sobre todos os intervalos possíveis, teremos a probabilidade de obter qualquer valor, portanto, 1. Assim temos: $\sum P(x_1, x_2) = 1$. Como a função é contínua, podemos escrever esta última expressão em termos da função densidade de probabilidade F(X), $\int F(x).\,\mathrm{d}x = 1$. A função densidade de probabilidade, por ter o significado de densidade, não pode ser negativa e, portanto, $F(x) \geq 0$ para qualquer valor de x.

D) **Média e Desvio Padrão**

Média: Valor usado pra representar um conjunto de medições de uma mesma medida: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, onde x_i são os váriips dados obtidos, N é a quantidade desses dados e $\bar{x}$ o valor médio.
Desvio padrão: O valor verdadeiro pode ser associado a média do conjunto de dados por meio de um desvio padrão que nos mostra de maneira quantitativa o quanto estamos errando em uma medida, mais precisamente mostrando o grau de dispersão do conjunto de dados. Assim, temos: $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$.

Capítulo III - Distribuição Binomial e Distribuição de Poisson

quarta-feira, 2 de março de 2022 11:13

A) **Distribuição Binomial**

Antes de entrarmos no tópicos das distribuições, é importante remos claro alguns pontos sobre a função densidade de probabilidade. Com isso, sendo a probabilidade de se obter um dado no intervalo $[x_a, x_b]$ se relaciona com a função densidade de probabilidade, $f(x)$, por: $P(x \in [x_a, x_b]) = \int_{x_a}^{x_b} f(x)\,\mathrm{d}x$. Para uma função ser considerada uma densidade de probabilidade só é necessário que: $f(x) \geq 0$ e $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$. Além disso, os principais parâmetros de uma função densidade de probabilidade são:
Para variáveis continuas, temos:

- Valor médio (verdadeiro), x_0: $x_0 = \langle x \rangle = \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x$
- Variância de x, σ^2: $\sigma^a = \left\langle \left(x - x_{\dot{\Theta}}\right)^2 \right\rangle = \int_{-\infty}^{+\infty} \left(x - x_0\right)^2 f(x)\,\mathrm{d}x$

Para variáveis discretas, temos:

- Valor médio (verdadeiro), x_0: $x_0 = \langle x \rangle = \sum_{i=1}^{N_x} x_i F(x_i)$
- Variância de x, σ^2: $\sigma^2 = \left\langle \left(x - x_0\right)^2 \right\rangle = \sum_{i=1}^{N_x} \left(x_i - x_0\right)^2 F(x_i)$

Distribuição binomial: Distribuição do número de vezes, **n**, que se observam resultados de um tipo escolhido em **N** tentativas independentes, sendo **p** a probabilidade de se obter, em cada tentativa, um resultado do tipo escolhido: $F_{N,p}(n) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$.

B) **Média e Desvio padrão para a distribuição binomial**

- A binomial é normalizada: $\sum_{n=0}^{N} F_{N,p}(n) = \sum_{n=0}^{N} \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} = \left[p + (1-p)\right]^N = 1$
- O valor médio (verdadeiro) do número de ocorrências, n_0: $n_0 = \langle n \rangle = \sum_{n=}^{N} n \frac{N!}{n!(N-n)!} p^n \cdot (1-p)^{N-n} = Np$
- O desvio-padrão (verdadeiro) do número de ocorrências, σ_n: $\sigma_n = \sqrt{\langle n^2 \rangle - n_0^2} = \sqrt{Np(1-p)}$

D) **Distribuição de Poisson**

Distribuição do número de ocorrências de eventos, **n**, em uma situação em que a probabilidade de sucesso em cada tentativa, **p**, é muito baixa e o número de tantativas, **N,** é muito grande, mas o produto **a = Np** é um número finito: $P_a(n) = \frac{a^n}{n!} \mathrm{e}^{-a}$, que corresponde ao caso limite da Binomial com $N \to \infty$, mas com o produto **Np = a** mantido constante.

- A Poisson é normalizada: $\sum_{n=0}^{\infty} P_a(n) = \sum_{n=0}^{\infty} \frac{a^n}{n!} \mathrm{e}^{-a} \left(\sum_{n=0}^{\infty} \frac{a^n}{n!} \right) \mathrm{e}^{-a} = 1$
- O valor médio (verdadeiro) do número de ocorrências, n_0: $n_0 = \sum_{n=0}^{\infty} n \frac{a^n}{n!} \mathrm{e}^{-a} = \sum_{n=1}^{\infty} \frac{a^n}{(n)-1!} \mathrm{e}^{-a} = \sum_{m=0}^{\infty} \frac{a^{m+1}}{m!} \mathrm{e}^{-a} = a$
- O desvio-padrão (verdadeiro) do número de ocorrências, σ_n: $\sigma_n = \sqrt{\langle n^2 \rangle - n_0^2} = \sqrt{a}$

Fichamento do livro Tratamento Estatístico de Dados em Física Experimental de Otaviano A. M. Helene e Vito R. Vanin.

Capítulo IV - Distribuição Normal

quarta-feira, 2 de março de 2022 11:13

A) **Função densidade de probabilidade normal**

Função densidade de probabilidade normal: A curva contínua que representa a função densidade de probabilidade normal é dada pela expressão: $\mathrm{F}(x) = \frac{1}{\sqrt{2\pi}\,\sigma_0}\exp\left(-\frac{(x-x_0)^2}{2\,\sigma_0^2}\right)$. No entanto, existe uma outra forma reduzida de apresentar a mesma função densidade de probabilidade como função do erro, onde definimos o erro de uma observação de uma grandeza que tem um valor verdadeiro como a diferença entre o valor observado e o verdadeiro, $y = x - x_0$. Assim, a função densidade de probabilidade normal dos erros é $G(y) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp(-y^2/2\,\sigma_0^2)$.
Essas funções também são conhecidas como funções densidade de probabilidade normais ou como funções de Gauss ou Gaussianas.

C) **Um significado para o desvio padrão**

A função densidade de probabilidade permite-nos calcular a probabilidade de obter um dado num certo intervalo, ou seja, calcular: $P(x_a \le x \le x_b) = \int_{x_a}^{x_b} F(x)\,\mathrm{d}x$.

D) **Desvio padrão da média**

Desvio padrão da média: Sendo N o número de dados do conjunto e σ_0 o desvio padrão, temos que: $\sigma_m^2 = \frac{1}{N}\sigma_0^2$.[

F) **Propagação de erros**

Lei geral de propagação de incertezas: A incerteza de uma grandeza w, calculada a partir de grandezas x, y com valores verdadeiros x_0, y_0 e incertezas σ_x, σ_y pode ser determinada a partir da definição de σ_w, usando: $\sigma_w^2 = \langle \varepsilon_w^\alpha \rangle = \left\langle (w - w_0)^2 \right\rangle$, onde $w = w(x,y)$ e $w_0 = w(x_0, y_0)$.
A lei geral de propagação de incertezas é obtida pela expansão, em série de Taylor até primeira ordem, de w na vizinhança de (x_0, y_0): $w(x,y) \cong w(x_0,y_0) + \frac{\partial w}{\partial x}(x - x_0) + \frac{\partial w}{\partial y}(y - y_0)$, onde $\varepsilon_x = x - x_0$ e $\varepsilon_y = y - y_0$ $\quad\rightarrow\quad$ $w - w_0 \cong \frac{\partial w}{\partial y}\varepsilon_y + \frac{\partial w}{\partial y}\varepsilon_y$
Com isso, temos que para $w = w(x,y)$, obtêm-se: $\sigma_n^2 = \left\langle (w - w_0)^2 \right\rangle \cong \left\langle \left(\frac{\partial w}{\partial y}\varepsilon_y + \frac{\partial w}{\partial y}\varepsilon_y\right)^2 \right\rangle$, que resulta em:

$\sigma_w^2 \cong \left(\frac{\partial w}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial w}{\partial y}\right)^2 \sigma_y^2 + 2\left(\frac{\partial w}{\partial x}\right)\left(\frac{\partial w}{\partial y}\right) cov(x,y)$, onde $\sigma_x^2 = \langle \varepsilon_x^2 \rangle = \left\langle (x - x_0)^2 \right\rangle$ é a variância de x (e o mesmo para y) e $\mathrm{cov}(x,y) = \langle \varepsilon_x \varepsilon_y \rangle = \langle (x - x_0)(x - y_0) \rangle$ é a covariância entre x e y.

Covariâncias: As covariÂncias podem aumentar ou diminuir a incerteza de w. Com o efeito das covariâncias depende do sinal das derivadas parciais e da própria covariância. Em ajustes e em resultados de medições de muitas grandezas é útil fornecer a matriz de covariâncias, $V_{i,j} = \langle \varepsilon_i, \varepsilon_j \rangle$: $\mathrm{V} = \begin{bmatrix} \sigma_x^2 & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & \sigma_y^2 \end{bmatrix}$. As covariâncias tem dimensão igual ao produto das dimensões das grandezas envolvidas.

Correlações: Correlações são covariâncias normalizadas pelo produto dos desvios-padrões correspondentes, $\rho_{i,j} = \frac{\mathrm{cov(i,}j)}{\sigma_i \sigma_j}$, onde as correlações são adimentsionais e limitadas ao intervalo entre -1 e +1. Para dados independentes temos uma correlação zero e quanto mais próximo de 1 for o módulo da correlação, mais importante é a correlação entre as grandezas. Assim, é usual fornecer também a matriz de correlações, mais importante é a correlação entre as grandezas. Além disso, é usual fornecer também a matriz de correlações, $C_{i,j} = \rho_{i,j}$: $\mathrm{C} = \begin{bmatrix} 1 & p(x,y) \\ p(x,y) & 1 \end{bmatrix}$

Apêndice

Função de distribuição acumulada: A função de distribuição acumulada, $g(x)$, corresponde à integral da função densidade de

probabilidade, $f(x)$, desde $-\infty$ até x: $g(x) = \int_{-\infty}^{x} f(x')\,\mathrm{dx}'$. As funções de distribuição acumulada são usadas em testes estatísticos para avaliar se o valor obtido para x no experimento, x_{Exp}, não nem é muito pequeno ($g\left(x_{Exp}\right) \sim 0$) nem muito grande ($g\left(x_{Exp}\right) \sim 1$).

Teorema Central do Limite: A função densidade de probabilidade da soma, S, de variáveis aleatórias independentes x_i, que sigam qualquer função densidade de probabilidade, $f_i(x)$, com valores médios verdadeiros μ_i e variâncias σ_i^2 finitos, tende a uma gaussiana quando o número de variáveis somadas, N, tende ao infinito. Além disso, S terá valor verdadeiro igual à soma das variâncias:

$S = \sum_{i=1}^{N} x_i$, $\mu_0 = \langle s \rangle = \sum_{i=1}^{N} \mu_i$, $\sigma_0^2 = \left\langle (S - \langle s \rangle)^2 \right\rangle = \sum_{i=1}^{N} \sigma_i^2$, $s^* = \frac{s - \mu_0}{\sigma_0}$ (variável soma normalizada).

Parâmetros de localização: Outros parâmetros de localização de funções densidade de probabilidade.

- **Moda,** x_{m_P}: valor de x em que $f(x)$ é máximo.
- **Mediana,** x_M: valor de x tal que a probabilidade de se obter um dado com $x \leq x_M$ é igual ao de se obter $x \geq x_M$**.** Ou seja:

$$\int_{-\infty}^{x_m} f(x)\,dx = \int_{x_m}^{+\infty} f(x)\,dx = 0.5$$

Momentos de funções densidade de probabilidade: Os momentos de uma variável aleatória podem ser obtidos pela derivação da função geradora.

- **Momento de ordem n,** μ_n: $\mu_n = \langle x^n \rangle = \int_{-\infty}^{+\infty} x^n f(x)\,dx$**.**
- **Momento central de ordem n,** $\mu_n^{(0)}$: $\mu_n^{(0)} = \left\langle (x - x_0)^n \right\rangle = \int_{-\infty}^{+\infty} (x - x_0)^n f(x)\,dx$

Parâmetros de caracterização: Outros parâmetros para caracterizar funções densidade de probabilidade, considerando momentos centrais normalizados.

- **Obliquidade ou assimetria ("skewness"), S**: é o grau de assimetria observada em uma distribuição de probabilidade que se desvia da distribuição normal simétrica (sino curva) em um determinado conjunto de dados: $S = \frac{\mu_3^0}{\sigma_0^3} = \frac{\left\langle (x - x_0)^3 \right\rangle}{\sigma_0^3}$
- **Curtose ("kurtosis"), K**: grau de presença de outliers na distribuição: $K = \frac{\mu_4^0}{\sigma_0^4} = \frac{\left\langle (x - x_0)^4 \right\rangle}{\sigma_0^4}$**.**

Capítulo V - Princípio da máxima probabilidade

quarta-feira, 2 de março de 2022 11:13

A) **Princípio da máxima probabilidade**

Tendo em mente que quanto maior a quantidade de dados de uma medida, melhor é a estimativa da função probabilidade que governa a medida. No limite quando o número de dados tende a infinito podemos conhecer exatamente a distribuição.

Método da máxima Verossimilhança: Para estimar valores dos parâmetros de funções (densidade) de probabilidade que regem o experimento. Os valores dos parâmetros são estimados como sendo os que maximizam a função (densidade) de probabilidade de todo o conjunto de dados obtidos no experimento (que recebe o nome de função verossimilhança, L). O M.M.V. pode ser usado com qualquer função densidade de probabilidade ou função de probabilidade e pode ser usado para estimar quaisquer parâmetros dessas funções.
A função verossimilhança, $\mathcal{L}(\{x_i\}|\vec{a})$ de que o conjunto de dados independentes $\{x_i\} = \{x_1, x_2, \dots, x_N\}$ seja obtido em experimentos regidos por funções (densidade) de probabilidade com parâmetros $\vec{a}$ (a serem estimados) é:
$\mathcal{L}(\{x_i\}|\vec{a}) = \prod_{i=1}^{N} f(x_i|\vec{a}) = f(x_1|\vec{a}) * \cdots * f(x_n|\vec{a})$, onde $f(x_i|\vec{a})$ é a função (densidade) de probabilidade de se obter o valor , sendo que o vetor $\vec{a}$ representa os parâmetros desconhecidos dessa função. O uso do Método da Máxima Verossimilhança consiste em três etapas:

1. Escrever a função verossimilhança $\mathcal{L}(\{x_i\}|\vec{a})$
2. Determinar os parâmetros $\tilde{\vec{a}}$ que maximizam $\mathcal{L}$ (na prática, se maximiza $l = \ln(\mathcal{L})$, por simplificar os cálculos).
3. Determinar as incertezas dos parâmetros estimados $\tilde{\vec{a}}$ por propagação de incertezas.

Após isso, é conveniente avaliar se as expressões obtidas para os estimadores não são tendenciosas, especialmente nos casos e m que o N é pequeno. Para N → ∞ o M.M.V. nunca é tendencioso.

B) **Método do mínimo quadrado**

Método dos mínimos quadrados: O método dos mínimos quadrados consiste em determinar os parâmetros $\vec{a}$ que, para o caso de dados independentes, minimizam a seguinte somatória: $Q(\vec{a}) = \sum_{i=1}^{N} \left(\frac{y_i - F(i,\vec{a})}{\sigma_i^-}\right)^2$, onde a função modelo, $F(i,\vec{a})$, descreve a relação entre o i-ésimo dado e os parâmetros $\vec{a}$ a serem estimados. No caso do ajuste de uma reta, $\vec{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$, e $F(\mathrm{i},\vec{a}) = a_1 + a_2 x_i$.

C) **Ajuste de polinômios**

No caso de funções lineares nos parâmetros, $\frac{\partial}{\partial a_j}\left(\frac{\partial F}{\partial a_k}\right) = 0$, a função modelo pode ser escrita como: $F(\mathrm{i},\vec{a}) = a_1 g_1(\mathrm{i}) + \cdots + a_P g_p(\mathrm{i})$, e os parâmetros que minimizam a variável $Q(\vec{a})$ correspondem às soluções do seguinte sistema linear:

$$\begin{aligned}
\sum_{i=1}^{N} \frac{y_i g_1(i)}{\sigma_i^2} &= \tilde{a}_1 \sum_{i=1}^{N} \frac{g_1(i) g_1(i)}{\sigma_i^2} + \tilde{a}_2 \sum_{i=1}^{N} \frac{g_1(i) g_2(i)}{\sigma_i^2} + \cdots \\
\sum_{i=1}^{N} \frac{y_i g_2(i)}{\sigma_i^2} &= \tilde{a}_1 \sum_{i=1}^{N} \frac{g_2(i) g_1(i)}{\sigma_i^2} + \tilde{a}_2 \sum_{i=1}^{N} \frac{g_2(i) g_2(i)}{\sigma_i^2} + \cdots \\
&\vdots
\end{aligned}$$

O sistema de equações do MMQ pode ser escrito de forma matricial como: $\vec{D} = M\vec{A}$, onde $A_l = a_l$, e: $D_l = \sum_{i=1}^{N} \frac{y_i g_l(i)}{\sigma_i^2}$ e $M_{l,c} = \sum_{i=1} \frac{g_l(i) g_c(i)}{\sigma_i^2}$.

- A solução é dada por: $\vec{A} = \left(M^{-1}\right)\vec{D}$
- E a matriz de covariância de A é dada por: $V_A = \left(M^{-1}\right)$

- O χ^2 do ajuste é calculado por: $\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - F\left(i, \vec{A}\right)}{\sigma_i} \right)^2$

Capítulo VI - Testes de Significância

quarta-feira, 2 de março de 2022 11:13

A) **Distribuição de qui-quadrado (χ^2)**

Análise de resíduos: Os resíduos do ajuste são definidos como: $R_i = y_i - F(x_i, \vec{a})$. Se o modelo for adequado, os resíduos não devem apresentar estrutura clara (pois os dados devem se distribuir aleatoriamente ao redor da função ajustada). No caso de ajustes com muitos dados, erros gaussianos, incertezas corretas e modelo adequado, resíduos de módulo maior que as incertezas correspondentes, $|R_i| > \sigma_i$, devem ocorrer em cerca de 1/3 dos pontos:

- $|R_i| > 2\sigma_i$ deve ocorrer em ~1/20 dos pontos;
- $|R_i| > 3\sigma_i$ é improvável, mas possível (~1/300 dos pontos).

B) **Teste de χ^2**

O teste de χ^2 avalia se a dispersão dos pontos ao redor da função ajustada é consistente com a incerteza dos dados. No caso, de dados estatisticamente independentes, o χ^2 é calculado por: $\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - F(i,\vec{A})}{\sigma_i} \right)^2$. O valor esperado (o valor médio verdadeiro) do χ^2 é $\langle \chi^2 \rangle = L$, onde L = N - P é o número de graus de liberdade do ajuste (N é o número de dados e P o número de parâmetros ajustados). O χ^2 reduzido é definido como $\chi^2_{\mathrm{Re}\,d} = \frac{\chi^2}{L}$, tem valor esperado 1.