

ESTIMAÇÃO DA PROPORÇÃO POPULACIONAL p

Objetivo

Estimar a proporção p (desconhecida) de elementos em uma população apresentando certa característica de interesse, a partir da informação fornecida por uma amostra.

Exemplos:

p : proporção de alunos da *USP* que foram ao teatro pelo menos uma vez no último mês;

p : proporção de consumidores satisfeitos com os serviços prestados por uma empresa telefônica;

p : proporção de eleitores da cidade de São Paulo que votariam em determinado candidato, caso a eleição para presidente se realizasse hoje;

p : proporção de crianças de 2 a 6 anos, do estado de São Paulo, que não estão matriculadas em escola de educação infantil.

Dois aspectos de estimação:

- **Estimação pontual**
- **Estimação intervalar (c/ margem de erro)**

- Vamos começar por observar n elementos, extraídos ao acaso e com reposição da população;

- Para cada elemento selecionado, verificamos a presença (sucesso) ou não (fracasso) da característica de interesse.

Estimador pontual

Um **estimador pontual para p** , natural nesta situação, é a **proporção amostral**, definida naturalmente como

$$\hat{p} = \frac{X}{n},$$

sendo que,

X denota o número de elementos na amostra que apresentam a característica;

n denota o tamanho da amostra coletada.

Se observamos o valor k da v. a. X , obtemos $\hat{p} = k / n$ que denominamos **estimativa pontual para p** .

Exemplo 1: Sejam,

p : proporção de alunos da *USP* que foram ao teatro pelo menos uma vez no último mês, e

X : número de estudantes que respondem “sim” em uma pesquisa com n entrevistados.

Suponha que foram entrevistados $n = 500$ estudantes e que, desses, $k = 100$ afirmaram que foram ao teatro pelo menos uma vez no último mês.

A **estimativa pontual (proporção amostral) para p** é dada por:

$$\hat{p} = \frac{k}{n} = \frac{100}{500} = 0,20 ,$$

ou seja, 20% dos estudantes *entrevistados* afirmaram que foram ao teatro pelo menos uma vez no último mês.

→ Note que, outra amostra de mesmo tamanho pode levar a uma outra estimativa pontual para p .

Estimativa intervalar ou intervalo de confiança

- Para uma amostra observada, os estimadores pontuais fornecem como estimativa um único valor numérico para o parâmetro.
- Os estimadores pontuais são variáveis aleatórias e, portanto, possuem uma distribuição de probabilidade, em geral, denominada *distribuição amostral*.

Idéia: construir **intervalos de confiança**, que incorporem à estimativa pontual informações a respeito de sua variabilidade (erro amostral) através de margens de erro.

Intervalos de confiança são obtidos por meio da ***distribuição amostral do estimador pontual***.

A ***estimativa intervalar*** corresponde a um intervalo determinado da seguinte maneira:

$$[\hat{p} - \varepsilon; \hat{p} + \varepsilon],$$

sendo ε a **margem de erro**.

Pergunta: *Que valor de ε adotar ?*

Seja $P(\varepsilon)$ a probabilidade de a estimativa pontual estar a distância de, no máximo, ε da proporção verdadeira p , ou seja,

$$P(\varepsilon) = P(|\hat{p} - p| \leq \varepsilon).$$

A probabilidade $P(\varepsilon)$ é também denominada **coeficiente ou nível de confiança do intervalo**, que denotamos pela letra grega γ (gama).

Afirma-se ainda que a estimativa intervalar tem nível de confiança $\gamma = P(\varepsilon)$.

Queremos fixar um número γ (bastante próximo de 1) e adotar uma margem de erro ε que satisfaça a igualdade acima.

Vamos fazer algumas contas:

$$P(\varepsilon) = P(|\hat{p} - p| \leq \varepsilon) = P\left(\left|\frac{X}{n} - p\right| \leq \varepsilon\right)$$

$$= P(p - \varepsilon \leq \frac{X}{n} \leq p + \varepsilon)$$

$$= P(np - n\varepsilon \leq X \leq np + n\varepsilon)$$

$$= P\left(\frac{-n\varepsilon}{\sqrt{np(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{n\varepsilon}{\sqrt{np(1-p)}}\right)$$

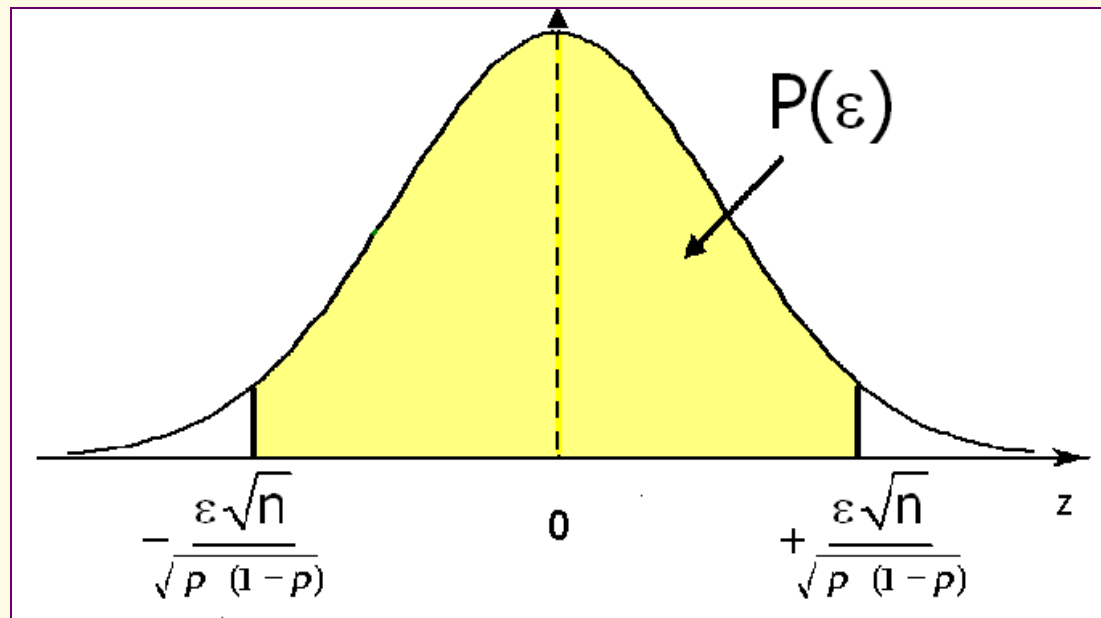
Como $X \sim b(n, p)$ temos que, para n **grande**,

a v. a. $Z = \frac{X - np}{\sqrt{np(1-p)}}$ tem distribuição aproximada $N(0,1)$.

Deste modo, para n grande,

$$P(\varepsilon) \cong P\left(\frac{-\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} \leq Z \leq \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right),$$

onde $Z \sim N(0,1)$.

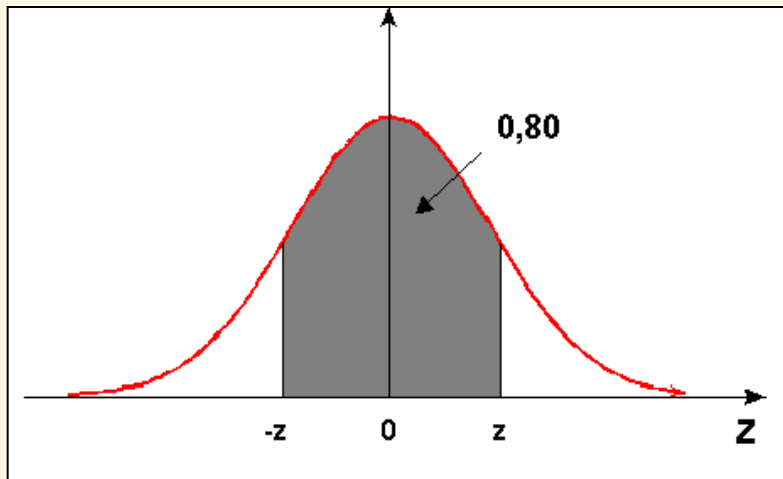


Denotando $\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} = z$, temos que

$$P(\varepsilon) = \gamma = P(-z \leq Z \leq z).$$

Assim, podemos obter z a partir de γ .

Por exemplo, vamos tomar $\gamma = 0,80$.



$\Rightarrow z$ é tal que $A(z) = 0,90$.

Pela tabela, temos $z = 1,28$.

Margem de erro

Da igualdade $z = \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}$,

se conclui que a **margem de erro** ε é dada por

$$\varepsilon = z\sqrt{\frac{p(1-p)}{n}},$$

onde z é tal que $\gamma = P(-z \leq Z \leq z)$, com $Z \sim N(0,1)$.

Dimensionamento da amostra

Da relação $\varepsilon = z \sqrt{\frac{p(1-p)}{n}}$

segue que o **tamanho amostral** n , fixados o nível de confiança γ e a margem de erro ε , deve satisfazer

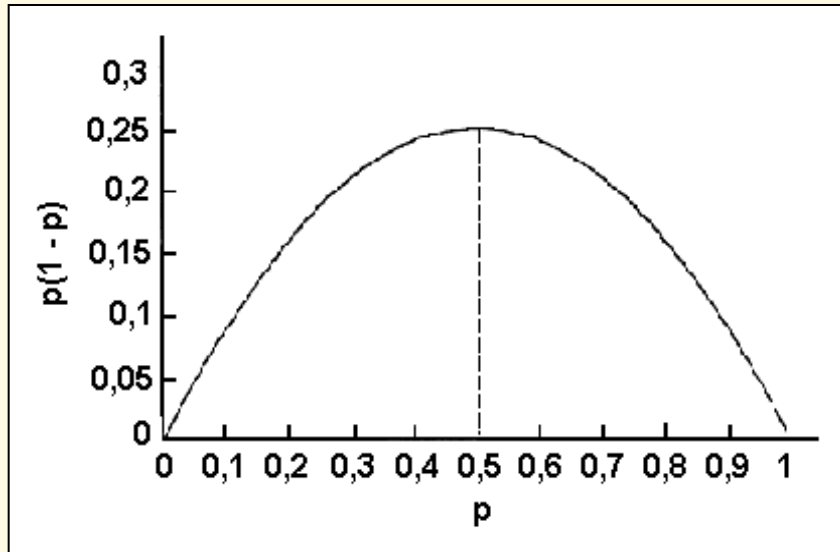
$$n = \left(\frac{z}{\varepsilon} \right)^2 p(1-p),$$

onde z é tal que $\gamma = P(-z \leq Z \leq z)$ e $Z \sim N(0,1)$.

Entretanto, nesta expressão, n depende de $p(1-p)$, que é desconhecido.

→ **Como determinar o valor de n ?**

Gráfico da função $p(1-p)$, $0 \leq p \leq 1$.



Pela figura

observamos que:

- a função $p(1-p)$ é uma parábola simétrica em torno de $p = 0,5$;
- o máximo de $p(1-p)$ é $0,25$, alcançado quando $p = 0,5$.

Assim, fazendo uma análise de pior caso, substituímos $p(1-p)$ por seu valor máximo, obtendo

$$n = \left(\frac{z}{\varepsilon}\right)^2 0.25,$$

o que pode fornecer um valor de n maior do que o necessário, mas em todo caso é suficiente para garantir o nível de confiança γ com a margem de erro ε .

Exemplo 2:

No exemplo da *USP* (Exemplo 1) suponha que nenhuma amostra foi coletada. Quantos estudantes precisamos consultar de modo que a estimativa pontual esteja, no máximo, a 0,02 da proporção verdadeira p , com uma probabilidade de 0,95?

Dados do problema:

$\varepsilon = 0,02$ (margem de erro);

$P(\varepsilon) = \gamma = 0,95 \Rightarrow z = 1,96.$

$$n = \left(\frac{1,96}{0,02} \right)^2 p(1-p) \leq \left(\frac{1,96}{0,02} \right)^2 0,25 = 2401 \text{ estudantes.}$$

Pergunta: *É possível reduzir o tamanho da amostra quando temos alguma informação a respeito de p ?*

Por exemplo, sabemos que:

- p não é superior a 0,30, ou
- p é pelo menos 0,80, ou
- p está entre 0,30 e 0,60.

Resposta: *Depende do tipo de informação sobre p .*

Em alguns casos, podemos substituir a informação $p(1-p)$, que aparece na expressão de n , por um valor menor que 0,25.

Redução do tamanho da amostra

Vimos que, se nada sabemos sobre o valor de p , no cálculo de n , substituímos $p(1-p)$ por seu valor máximo, e calculamos

$$n = \left(\frac{z}{\varepsilon} \right)^2 \times 0,25.$$

Se temos a informação de que p **é no máximo 0,30** ($p \leq 0,30$), então o valor máximo de $p(1-p)$ será dado por $0,3 \times 0,7 = 0,21$.

Logo, reduzimos o valor de n para

$$n = \left(\frac{z}{\varepsilon} \right)^2 \times 0,21.$$

Agora, se p é pelo menos **0,80** ($p \geq 0,80$), então o máximo valor de $p(1-p)$ é $0,8 \times 0,2 = 0,16$, e temos

$$n = \left(\frac{z}{\varepsilon} \right)^2 \times 0,16.$$

Mas, se **$0,30 \leq p \leq 0,60$** , o máximo valor de $p(1-p)$ é $0,5 \times 0,5 = 0,25$ e, neste caso, não há redução, ou seja,

$$n = \left(\frac{z}{\varepsilon} \right)^2 \times 0,25.$$

Exemplo 3:

No Exemplo 2, suponha que temos a informação de que no máximo 30% dos alunos da *USP* foram ao teatro no último mês.

Portanto, temos que $p \leq 0,30$ e, como vimos, o máximo de $p(1-p)$ neste caso é 0,21.

Assim, precisamos amostrar

$$n = \left(\frac{z}{\varepsilon} \right)^2 0,21 = \left(\frac{1,96}{0,02} \right)^2 0,21 = 2017 \text{ estudantes ,}$$

conseguindo uma redução de $2401 - 2017 = 384$ estudantes.

Intervalo de confiança para p

Vimos que a estimativa intervalar para p tem a forma:

$$\left[\hat{p} - \varepsilon ; \hat{p} + \varepsilon \right] ,$$

com $\varepsilon = z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ e z tal que $\gamma = P(-z \leq Z \leq z)$ na $N(0,1)$.

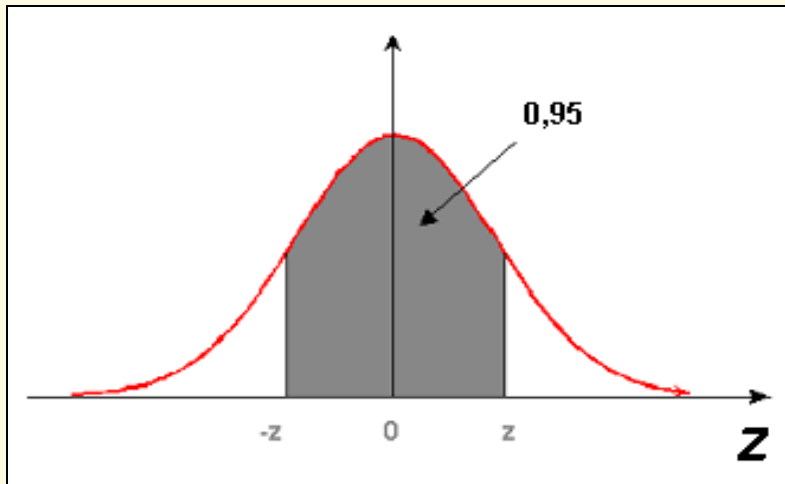
*Na prática, substituímos a proporção desconhecida p pela proporção amostral \hat{p} , obtendo o seguinte **intervalo de confiança com coeficiente de confiança γ** :*

$$IC(p; \gamma) = \left[\hat{p} - z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} ; \hat{p} + z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Exemplo 4:

No exemplo da *USP*, temos $n = 500$ e $\hat{p} = 0,20$.

Construir um intervalo de confiança para p com coeficiente de confiança $\gamma = 0,95$.



Como $\gamma = 0,95$ fornece $z = 1,96$, o intervalo é dado por:

$$\left[\hat{p} - z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} ; \hat{p} + z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right]$$

$$= \left[0,20 - 1,96 \sqrt{\frac{0,20 \times 0,80}{500}} ; 0,20 + 1,96 \sqrt{\frac{0,20 \times 0,80}{500}} \right]$$

$$= [0,20 - 0,035 ; 0,20 + 0,035] = [0,165 ; 0,235].$$

Neste intervalo, a estimativa pontual para p é 0,20, e a margem de erro ε é igual a 0,035.

Interpretação do IC com $\gamma = 95\%$:

Se sortearmos 100 amostras de tamanho $n = 500$ e construirmos os respectivos 100 intervalos de confiança, com coeficiente de confiança de 95%, esperamos que, aproximadamente, 95 destes intervalos contenham o verdadeiro valor de p .

Comentários:

Da expressão $\varepsilon = z \sqrt{\frac{p(1-p)}{n}}$, é possível concluir que:

- para γ fixado, a margem de erro diminui quando n aumenta;
- para n fixado, a margem de erro aumenta quando γ aumenta.

Exemplo 5:

Ainda no exemplo da USP, temos $k = 100$ e $n = 500$.

Qual a probabilidade de a estimativa pontual estar a uma distância de no máximo 0,03 da proporção verdadeira?

Dados do problema:

$$n = 500, \hat{p} = 0,20 \text{ e } \varepsilon = 0,03 \quad \Rightarrow \quad P(\varepsilon) = \gamma = ?$$

Como a proporção verdadeira p é desconhecida, utilizamos a estimativa pontual \hat{p} para calcular z e, assim, obter γ (ou $P(\varepsilon)$).

Cálculo de z :

$$z = \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} \cong \frac{0,03\sqrt{500}}{\sqrt{0,2 \times 0,8}} = 1,68 .$$

Logo, obtemos

$$\begin{aligned} P(\varepsilon) &\cong 2 \times A(z) - 1 \\ &= 2 \times A(1,68) - 1 \\ &= 2 \times 0,953 - 1 \\ &= 0,906 \quad (90,6\%). \end{aligned}$$

Exemplo 6: Suponha que estejamos interessados em estimar a proporção p de pacientes com menos de 40 anos, diagnosticados com câncer no pulmão, que sobrevivem pelo menos 5 anos.

Em uma amostra aleatoriamente selecionada de 52 pacientes, somente 6 sobreviveram mais de 5 anos.

- Estimativa por ponto para p : $\hat{p} = \frac{6}{52} = 0,115$ (proporção amostral)

- Intervalo de confiança aproximado de 95% para p :

$$\left(0,115 - 1,96 \sqrt{\frac{0,115(1 - 0,115)}{52}} ; 0,115 + 1,96 \sqrt{\frac{0,115(1 - 0,115)}{52}} \right)$$
$$= (0,028, 0,202)$$

Comentário final:

Embora este intervalo tenha sido construído usando a aproximação normal para a distribuição binomial, poderíamos ter gerado um intervalo de confiança *exato* para p usando a própria distribuição binomial.

Um intervalo exato é particularmente útil para pequenas amostras, em que o uso da aproximação normal não pode ser justificada.

Distribuição Normal : Valores de $P(Z \leq z) = A(z)$

Segunda decimal de z

Parte inteira e primeira decimal de z

	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000