Estimação para um parâmetro μ que representa uma proporção p

O parâmetro p , $0 \leqslant p \leqslant 1$, representa a probabilidade de algum evento de interesse.

Relembrando resultados importantes

RESULTADO ***Valor esperado e variância da média amostral.***

Seja uma amostra aleatória de tamanho n de uma variável $\boldsymbol{X}$, com média μ e variância σ^2. Temos que:

$$E(\overline{X}) = \mu \qquad \text{e} \qquad Var(\overline{X}) = \frac{\sigma^2}{n}$$

RESULTADO ***Médias de Normais independentes tem distr. Normal.***

Se $\boldsymbol{X} \sim N(\mu,\sigma^2)$, para uma amostra aleatória de tamanho n de X, temos que

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Logo,

$$Z = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1).$$

RESULTADO *Distribuição t; para n grande: Normal Padrão.*

Se $X \sim N(\mu,\sigma^2)$, sendo σ^2 desconhecida, então, para uma amostra aleatória de tamanho n de X,

$$T = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1},$$

onde, $S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$ é a variância amostral e t_{n-1} representa a distribuicão *t de Student* com *n-1 graus de liberdade*.

Para n grande, temos: $T \sim N(0;1)$

Aproximadamente

Teorema Limite Central (*TLC*)

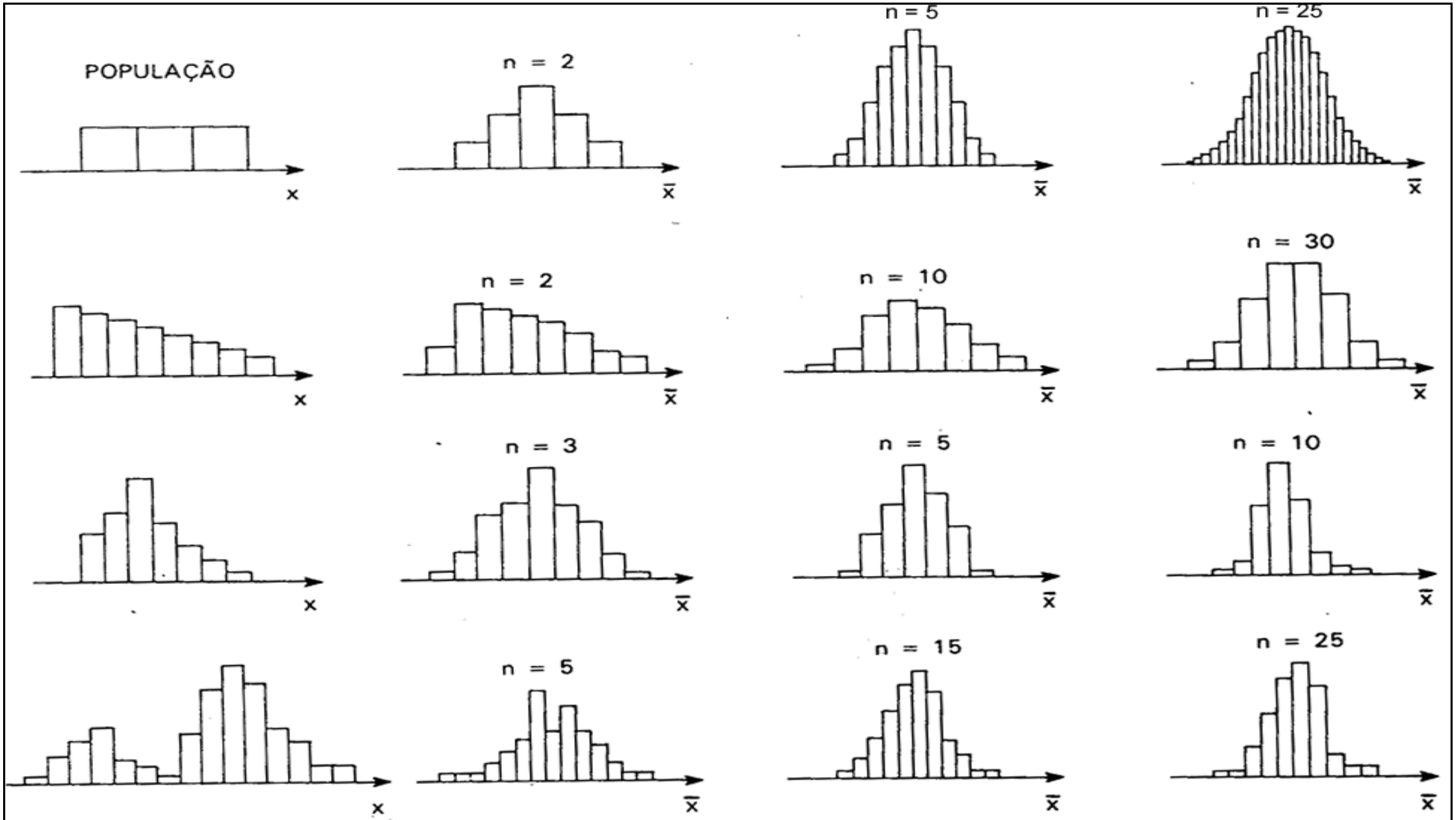
RESULTADO

Seja X uma variável com média μ e variância σ^2.

Para amostras X_1, X_2, ..., X_n , retiradas ao acaso e com reposição de X, a distribuição de probabilidade da média amostral $\overline{X}$ *aproxima-se*, *para n grande*, de uma distribuição normal, com média μ e variância σ^2/n , ou seja,

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \text{ para } n \text{ grande, aproximadamente.}$$

Figura 2: Histogramas correspondentes às *distribuições de* $\bar{X}$ para amostras de algumas populações.



Fonte: BUSSAB & MORETTIN. ***Estatística Básica***. São Paulo, Atual, 4ª edição, 2002, pp. 273.

Portanto, se a variável X na população não tem distribuição normal, e n é grande, usando o TLC o **intervalo de confiança aproximado para** μ, com nível de confiança γ, é

para σ **conhecido**: $$\left[\overline{X} - z\frac{\sigma}{\sqrt{n}}\ ;\ \overline{X} + z\frac{\sigma}{\sqrt{n}}\right],$$

Estimador de σ

para σ **desconhecido**: $$\left[\overline{X} - z\frac{S}{\sqrt{n}}\ ;\ \overline{X} + z\frac{S}{\sqrt{n}}\right]$$

sendo z tal que $\gamma = P(-z \leq Z \leq z)$, com $Z \sim N(0, 1)$, σ é o desvio padrão da população, e S é o desvio padrão amostral.

Estimação da proporção

Caso particular com $X \sim Bernoulli(p)$

$P(X=1) = p$ $\quad$ p = *probabilidade de "sucesso"*

$P(X=0) = 1-p$ $\quad$ $1-p$ = *probabilidade de "fracasso"*

$$0 \leqslant p \leqslant 1$$

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 um 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.

- Vamos observar ***n*** elementos, extraídos ao acaso da população, de forma independente;

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

Neste caso, temos uma amostra aleatória (*a.a.*) de tamanho n de X, sendo X uma *v.a.* com distribuição de *Bernoulli*, que representamos por

$$X_1, X_2, \ldots, X_n,$$

onde X_i vale "1", se ocorre sucesso, ou "0", se ocorre fracasso para o i-ésimo elemento da amostra.

Estimador pontual

O ***estimador pontual para*** $\boldsymbol{p}$, também denominado ***proporção amostral***, é definido como

$$\hat{p} = \frac{X_1 + \ldots + X_n}{n}.$$

Note que:

- $X_1 + \ldots + X_n$ é o número de elementos na amostra que apresentam a característica ***e tem distribuição Binomial(n,p).***
- $\hat{p} = \bar{X}$

Se observamos k elementos na amostra com a característica, obtemos $\hat{p}_{obs} = k / n$, que denominamos ***estimativa pontual para*** $\boldsymbol{p}$.

Exemplo

Seja p a proporção de alunos da USP que foram ao teatro pelo menos uma vez no último mês.

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

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

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

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.

$\widehat{p}$ *é uma* ***variável aleatória.***

Intervalo de confiança para p

Vimos que, para qualquer variável aleatória X, quando n é grande, usando o Resultado 4 (TLC), um ***intervalo de confiança*** para μ tem a forma

$$\left[\overline{X} - \varepsilon \,;\, \overline{X} + \varepsilon\right]$$

onde $\varepsilon = z\dfrac{\sigma}{\sqrt{n}}$, sendo σ^2 a variância de X.

Neste caso, como $X \sim Bernoulli(p)$, com $\sigma^2 = Var(X)=p(1-p)$
o estimador intervalar para p é dado por $E(X)=p$

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

Normal Padrão

com $\varepsilon = z\sqrt{\dfrac{p(1-p)}{n}}$ e z tal que $\gamma = P(-z \le Z \le z)$

Intervalo de confiança para p

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

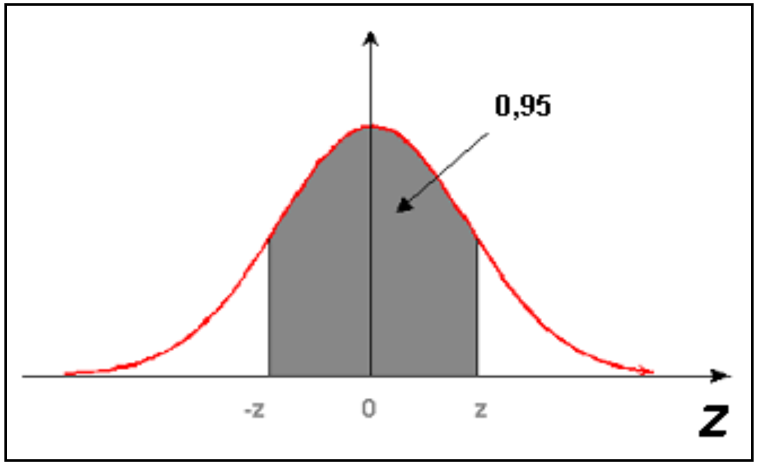
$$IC(p\,;\gamma) = \left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\;;\;\; \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Exemplo (continuação):

proporção amostral observada

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

Construir um intervalo de confiança para p com coeficiente de confiança γ = 0,95.



Como γ = 0,95 fornece z = 1,96, o intervalo é dado por:

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

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

$$= \left[0{,}20 - 0{,}035\ ;\ 0{,}20 + 0{,}035\right] = \left[0{,}165\ ;\ 0{,}235\right].$$

Interpretação do IC com γ = 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, o erro diminui com o aumento de n.
- para n fixado, o erro aumenta com o aumento de γ.

Dimensionamento da amostra

"Qual é o n?"

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

segue que o **tamanho amostral** $\boldsymbol{n}$, dados γ e a margem de erro ε, tem a

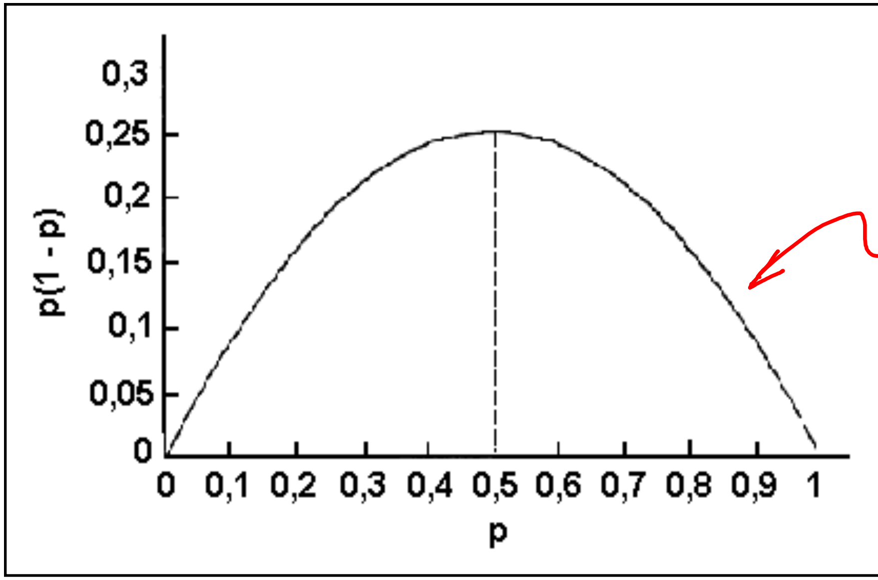
$$\boxed{n = \left(\frac{z}{\varepsilon}\right)^2 p(1-p),}$$

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 calcular o valor de n?

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



Parte da parábola

$y = x - x^2$

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, na prática, substituímos $p(1-p)$ por seu valor máximo, obtendo

$$n = \left(\frac{z}{\varepsilon}\right)^2 0{,}25\,, \qquad Var(X) = p(1-p) \leqslant \frac{1}{4}$$

que pode fornecer um valor de n maior "do que o necessário"

Exemplo (continuação):

No exemplo da *USP* 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 probabilidade de 0,95?

Dados do problema:

ε = 0,02 (erro da estimativa);

γ = 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

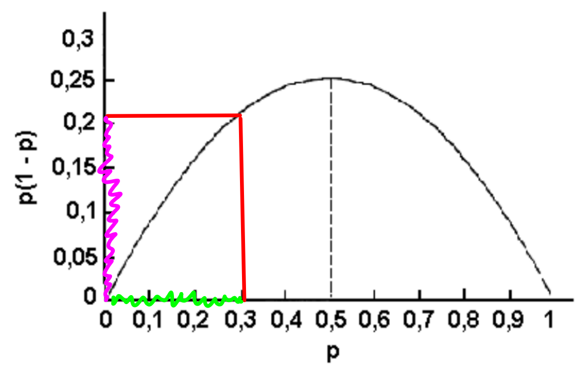
Vimos que, se nada sabemos sobre o valor de p, no cálculo de n, substituímos $p(1\text{-}p)$ por seu valor máximo, e calculamos

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

Se temos a informação de que ***p* é no máximo 0,30** ($p \leq 0{,}30$) então o valor máximo de $p(1\text{-}p)$ será dado por 0,3x0,7 = 0,21.

Logo, reduzimos o valor de n para

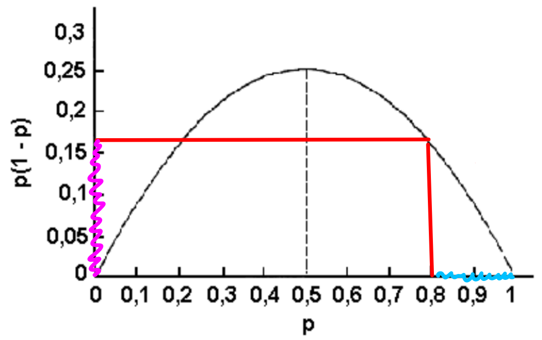
$$n = \left(\frac{z}{\varepsilon}\right)^2 \times 0{,}21.$$



$p \leqslant 0{,}3 \implies p(1\text{-}p) \leqslant 0{,}21$

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

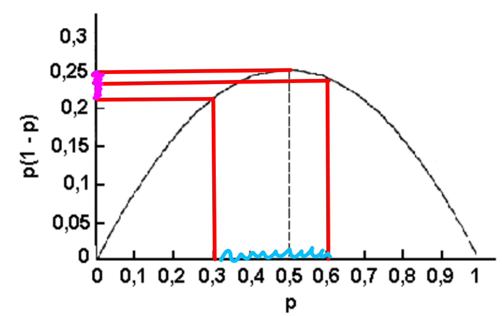
$$n = \left(\frac{z}{\varepsilon}\right)^2 \times 0{,}16\,.$$



$p \leqslant 0{,}8 \Longrightarrow p(1-p) \leqslant 0{,}16$

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

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



$0{,}3 \leqslant p \leqslant 0{,}6 \Longrightarrow 0{,}21 \leqslant p(1-p) \leqslant 0{,}25$

Exemplo (continuação):

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

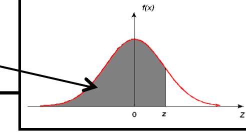
Portanto, temos que $p \leq 0{,}30$ e, como vimos, o máximo valor de $p(1-p)$ neste caso é 0,21.

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.

Distribuição Normal : Valores de P(Z ≤ 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

Volta