Appendix C:

The Standard Normal Distribution

Forsberg, Ole J. (September 27, 2015). "The Standard Normal Distribution." In *R for Starters*. Version 0.57721. Retrieved from http://rfs.kvasaheim.com/.

One of the most important theorems in statistics is the Central Limit Theorem. At its core, it states that sums (and averages) of independent and identically distributed random variables will converge in distribution to a Normal distribution. How well the Normal distribution approximates the true distribution depends on the sample size and the underlying distribution. As n increases, the approximation becomes better. This result holds for *all* distributions with finite means and variances.

This implies that if all we care about is the population mean then we only need to study the Normal distribution. This is very powerful. It is also very abused.

For the approximation brought about by the Central Limit Theorem to be close, one needs the original distribution to be close to Normal *or* one needs n to be large. If the original distribution is unimodal and symmetric, the n can be on the order of 30 for the approximation to be "close enough." If the original distribution is highly skewed, then n needs to be on the order of 500.

❧❧❧

To estimate the proportion of people who use a smart phone, a researcher decides to call people until she comes across the 50th person with a smart phone. After performing this experiment, she found that she needed to call 127 people. What is the estimated proportion of people with a smart phone?

C.1: The Standard Normal Distribution

Given that your variable of interest is Normally distributed with $\mu = 0$ and $\sigma^2 = 1$, the probability that a random value is less than a specified value, z_0, is equal to the area under the standard Normal pdf curve, to the left of the value.

standard Normal

$$\phi(z_0) := \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z_0^2}{2}\right] \tag{C.1}$$

Figure C.1 is a plot of the standard Normal pdf, $\phi(z)$, with the shaded area representing the probability that Z is less than (or equal to) z_0: $\mathbb{P}\left[Z \leq z_0\right]$. Note that the specified value is z_0, not Z. This is because Z is a random variable, whereas z_0 is a specified number.

There is an alternate way of representing the area. If we integrate the area between $z = -\infty$ and z_0, we have the cumulative distribution function (CDF) of the standard Normal distribution. While the pdf of the standard Normal distribution *can* be written explicitly (see Eqn C.1), the Normal CDF cannot. As such, we represent it by $\Phi(z)$.

cumulative distribution function

$$\Phi(z_0) := \int_{\infty}^{z_0} \phi(t)\, dt \tag{C.2}$$

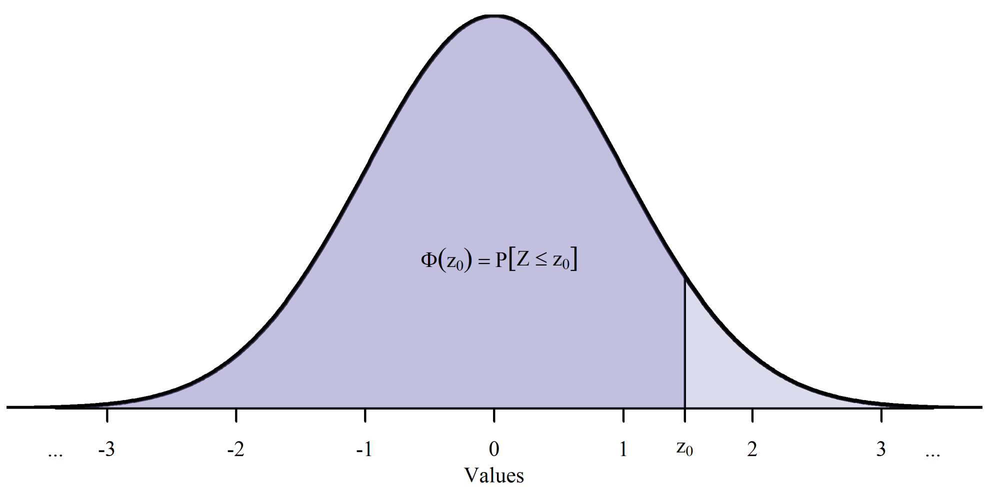


Figure C.1: *A plot of the Standard Normal probability density function, with the cumulative probability shaded and designated by the* $\Phi(z)$ *function.*

approximate

One cannot calculate the values of $\Phi(z_0)$ exactly; they must be approximated using advanced methods. However, because of the great utility of the standard Normal distribution, its values were estimated long ago.

error function

Historical note: While working with astronomical data, Laplace (1783) discovered that the integral $\int_0^x e^{-t^2}\,dt$ appeared and reappeared in his work. This function later became known as the error function (erf), and it is essentially equivalent to the CDF for the standard Normal distribution.

Unfortunately, while he was an excellent mathematician, the two approximations Laplace used to construct his table were poor for values of x between $\frac{1}{4}$ and 4. A little over a decade later, in the newly-minted French Republic, Kramp (1799) used Euler's formula to create a table of values from $x = 0$ to 3 up to eight decimal places. Values beyond this interval suffered from Laplace's approximation until Burgess (1898) improved upon them, producing estimates to 15 decimal places beyond $x = 1$.

Biometrika tables

Today's Z-Tables are based on the work of the Australian statistician William Fleetwood Sheppard (1903), which provided estimates from $x = 0$ to 6. These tables became known as the Biometrika tables, named after the journal in which he published them.

false precision

> ***Note***: Realize that in most applied work, one does not need so many decimal places. This is especially true when your underlying distribution is not Normal and you are using a Normal approximation. In such (common) cases, 15 decimal places constitutes false precision. As with most things in life, when making claims, your choice is **to have humility or to suffer humiliation**.

C.2: The Z-Transform

standardize

Before we explore the utility of the Central Limit Theorem, let us examine the Normal distribution in greater detail than we did in Appendix B.3. The importance of this section rests on the conversion of any Normally-distributed random variable into a *standard* Normally-distributed random variable. This is important to do if you rely on statistical tables as they are only for the standard Normal distribution. If you are using technology to determine the cumulative probabilities, then this section is of historical interest—if that.

The standard Normal tables only list cumulative probabilities for the *standard* Normal distribution—i.e., only for the distribution $Z \sim \mathcal{N}(0,1)$. If your variable has a different distribution, such as $X \sim \mathcal{N}(\mu, \sigma^2)$, then you will first have to standardize X using the z-transformation (Eqn C.3). **CDF** **z-transform**

$$Z := \frac{X - \mu}{\sigma} \tag{C.3}$$

For example, if $X \sim \mathcal{N}(\mu = 3, \sigma^2 = 16)$, then

$$\frac{X-3}{4} = Z \sim \mathcal{N}(0,1)$$

As such, if we want to calculate the probability that X is less than 2, we would do the following:

$$\begin{aligned}\mathbb{P}[X < 2] &= \mathbb{P}[X \leq 2] \\ &= \mathbb{P}\left[Z \leq \frac{2-3}{4}\right] \\ &= \mathbb{P}[Z \leq -0.25]\end{aligned}$$

Table C.1 gives the probability 0.4031, thus

$$\mathbb{P}[Z \leq -0.25] = 0.4013$$

and, *finally*:

$$\mathbb{P}[X < 2] = 0.4013$$

Note: The dancing we had to do to find the appropriate probability was simply due to us using a table of values instead of a computer. **table**

Were we to use a computer in lieu of the table, as statisticians usually do, we would simply run either of the following `R` commands

```
pnorm(-0.25)
```

or

```
pnorm(2, m=3, s=4)
```

In general, if we need to calculate $\mathbb{P}[X \le x]$, we use the command

```
pnorm(x, m=mu, s=sd)
```

where *mu* is the expected value of x and *sd* is its standard deviation.
If we need to calculate $\mathbb{P}[X > x]$, we use the command

```
1-pnorm(x, m=mu, s=sd)
```

If we need to calculate $\mathbb{P}[| X | < x]$, we use the command

```
2*pnorm(x, m=mu, s=sd)-1
```

To see why, note the following

$$\begin{aligned}\mathbb{P}[| X | < x] &= \mathbb{P}[-x < X < x] \\ &= \mathbb{P}[X < x] - \mathbb{P}[X < -x] \\ &= F(x) - F(-x) \\ &= F(x) - \big(1 - F(x)\big) \\ &= 2F(x) - 1\end{aligned}$$

For a graphic, refer to Figure C.1: the graph of $\Phi(z)$.

Finally, if we need to calculate $\mathbb{P}[a < X < b]$, we use the line

```
pnorm(b, m=mu, s=sd) - pnorm(a, m=mu, s=sd)
```

Example C.1: The scores on an Intelligence Quotient test are distributed approximately Normal with a mean of 100 and a standard deviation of 15. You scored 145 on the test. What proportion of the population has a higher IQ score than you?

Solution: If we let Q represent one's score on the IQ test, then we are given $Q \sim \mathcal{N}(\mu = 100, \sigma = 15)$. Your score is $q_0 = 145$, and we are asked to find $\mathbb{P}[Q > 145]$.

We first need to transform your score into an equivalent score on a standard Normal distribution: $z_0 = \frac{145-100}{15} = 3$. Now, we just have to find $\mathbb{P}[Z > 3]$. Looking in the standard Normal table (Table C.1), we see $\mathbb{P}[Z \le 3] = 0.9987$. Thus, we have $\mathbb{P}[Z > 3] = 1 - 0.9987 = 0.0013$. Thus, approximately 0.13% of the population has an IQ score as high or higher than you.

Alternatively, we could have just used the computer to get

```
1-pnorm(145,m=100,s=15)
```

And reached the same conclusion. ◇

Example C.2: American male heights are distributed approximately Normal with a mean of 70 inches and a standard deviation 3 inches. The author is 74 inches tall. What proportion of American males are shorter than the author?

Solution: Let H represent the height of an American male. We are given that $H \sim \mathcal{N}(\mu = 70, \sigma^2 = 9)$. The author's height is $h_0 = 74$ inches. We are to calculate $\mathbb{P}[H \leq 74]$. To wit:

$$\begin{aligned}\mathbb{P}[H \leq 74] &= \mathbb{P}\left[Z \leq \frac{74-70}{3}\right] \\ &= \mathbb{P}[Z \leq 1.333] \\ &= 0.9082\end{aligned}$$

Thus, 90.82% of American males are shorter than the author. In `R`, this is just

```
pnorm(74, m=70,s=3)
```

◇

C.3: The Central Limit Theorem

There is a finding in the field of Statistics that provides an approximate distribution of averages (and sums) of random variables, *no matter their underlying distribution* (with two minor requirements).

CLT for Means

This result is called the Central Limit Theorem: If random samples of size n are taken from a distribution with finite variance $\sigma^2 < \infty$, then

$$\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n) \tag{C.4}$$

The Central Limit Theorem is one of the most useful (and surprising) findings in statistics. No matter how the individual values of X are distributed, their average is *always* approximately distributed Normally.[1]

Example C.3: Let us suppose that the wait-time in minutes for a bus is exponentially distributed with rate parameter $\lambda = 0.05$. If 50 buses are in circulation at any one time, what is probability that the average wait-time is greater than 25 minutes?

Solution: We are given that $W \sim Exp(\lambda = 0.05)$. Section B.5 tells us we have $\mu = 20$ and $\sigma^2 = 400$. Thus:

$$\begin{aligned}
\mathbb{P}\left[\overline{W} > 25\right] &\approx \mathbb{P}\left[Z > \frac{25-20}{\sqrt{400/50}}\right] \\
&= 1 - \mathbb{P}\left[Z \leq 1.7678\right] \\
&= 1 - 0.9616
\end{aligned}$$

Thus, there is approximately a 3.84% chance that the average wait-time is more than 25 minutes.

Alternatively, the computer gives

```
1-pnorm(25, m=20, s=20/sqrt(50))
```

Which has a bit better accuracy and precision. ◇

[1] This approximation gets better as the sample size increases. Even with the most highly skewed distributions, if $n > 500$, the approximation is usually quite good for the usual social science work.

There is an equivalent way of writing the Central Limit Theorem. Instead of focusing on the average value, we can focus on the total value. Under the assumptions of the Central Limit Theorem, if $T := \sum X_i$ is the sum of the sample elements, then

CLT for Sums

$$T \sim \mathcal{N}(n\mu, n\sigma^2) \tag{C.5}$$

The use of this formula is the same as for the usual formulation of the Central Limit Theorem (Eqn C.4) except for the meanings of the variables. In fact, you have the skills to move between the two formulations.

Note: The proof of the Central Limit Theorem is beyond the scope of this book. While one proof is elementary, it requires knowledge of either characteristic functions or moment generating functions.

C.4: End of Chapter Materials

C.4.1 Exercises and Extensions

1. Noting that Z has a standard Normal distribution, find the following probabilities:

   a) $\mathbb{P}[Z \leq 1.53]$

   b) $\mathbb{P}[Z \leq -1.53]$

   c) $\mathbb{P}[Z \leq -2.11]$

   d) $\mathbb{P}[Z > 2.11]$

   e) $\mathbb{P}[Z \leq 1.53] + \mathbb{P}[Z \leq -1.53]$

   f) $\mathbb{P}[Z \leq 2.11] + \mathbb{P}[Z \leq -2.11]$

2. Given that $X \sim \mathcal{N}(\mu = 1, \sigma^2 = 2)$, find the following probabilities:

   a) $\mathbb{P}[X > 1.96]$

   b) $\mathbb{P}[X > -1.96]$

   c) $\mathbb{P}[|X| > 1.96]$

   d) $\mathbb{P}[-3 \leq X \leq -2]$

3. Given that $X \sim \mathcal{N}(\mu = 1, \sigma^2 = 2)$, find the following probabilities:

   a) $\mathbb{P}\left[X^2 \leq 1\right]$

   b) $\mathbb{P}\left[X^2 \leq 0\right]$

   c) $\mathbb{P}\left[X^2 \geq 0\right]$

   d) $\mathbb{P}\left[X^2 \leq 2\right]$

4. Given that $X \sim \mathcal{N}(\mu = 3, \sigma^2 = 4)$, find $\mathbb{P}[X \leq 2]$.

5. Given that a certain item has a lifetime $T \sim Exp(\lambda = 2)$, what is the approximate probability that a sample of 100 items will have an average lifetime of more than 0.55?

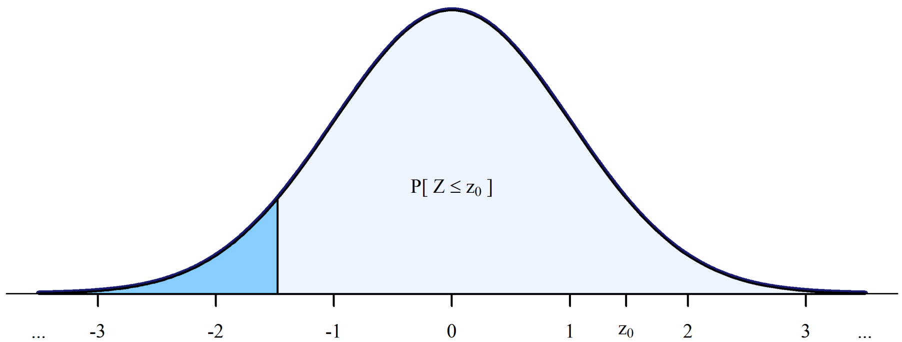
6. You select 100 Uniform random numbers between 0 and 1. What is the approximate probability that they add up to more than 75? What is the probability that their average is more than 0.05 from their expected value?

7. Intelligence Quotient tests are approximately distributed Normal with mean 100 and standard deviation 15. What proportion of the population has an IQ between 85 and 115? What is the probability that the average IQ of a class of 30 is below 90? What is the probability that the average IQ of a class of 5 is above 125?

8. Recall the *Quercus fusiformis* example from Appendix B (page 563). I randomly select 50 leaves. What is the probability that the average leaf thickness is greater than 5.25? What is the probability that the average leaf thickness is within 0.5 microns of the expected value? What is the probability that the total thickness exceeds 275 microns?

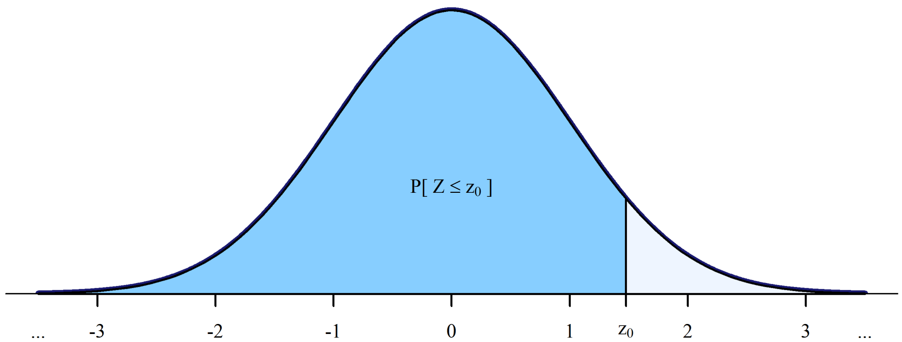
9. Dry erase markers have a lifetime that is approximately Gamma with expected lifetime 95 minutes and standard deviation 30 minutes. Exactly, what is the probability that this dry marker has a lifetime less than 90 minutes? Using the Normal approximation, what is the probability that this dry marker has a lifetime less than 90 minutes?

10. The time between when the journal editor receives your publication and when the editor informs you of the decision has an average of four months and a standard deviation of a half month. You send in 25 articles. What is the probability that the average decision time for these articles is greater than 4.5 months? What is the probability that the average decision time for these articles is less than two months?



$\mathbb{P}[Z \le z_0]$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.90	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
-3.80	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.70	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.60	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.50	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.40	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.30	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.20	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.10	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.40	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.30	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.20	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.10	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.00	0.0227	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.90	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.80	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.70	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.60	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.50	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.40	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.30	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.20	0.1151	0.1131	0.1112	0.1094	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.10	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.00	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.80	0.2119	0.2090	0.2061	0.2033	0.2004	0.1977	0.1949	0.1921	0.1894	0.1867
-0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2207	0.2177	0.2148
-0.60	0.2742	0.2709	0.2676	0.2643	0.2611	0.2579	0.2546	0.2514	0.2482	0.2451
-0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.20	0.4207	0.4168	0.4129	0.4091	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.00	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table C.1: *Table of values for the standard Normal distribution.*



$\mathbb{P}[Z \le z_0]$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5754
0.20	0.5793	0.5832	0.5871	0.5909	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7258	0.7291	0.7324	0.7357	0.7389	0.7421	0.7454	0.7486	0.7518	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7793	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8943	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9485	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9648	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9865	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9930	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.30	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	0.9998
3.50	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.60	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.70	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.80	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	1.0000	1.0000	1.0000
3.90	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000