Lecture 7

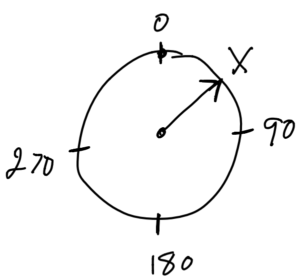
Topics:

- Continuous random variables
- Probability density functions (PDFs)
- Normal random variables
- Percentiles & z-scores

Up to now we have been studying R.V.s whose possible values are finite or discrete.

Now we study continuous R.V.s whose possible values lie within some interval of the real line.

Example: Random spinner



X can be any number in the interval $[0, 360)$

The problem is that PMFs don't work for continuous R.V.s

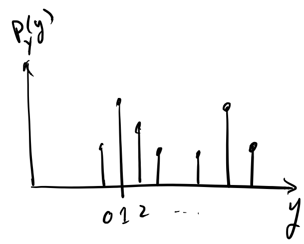
(The problem is that PMF sum to ∞ for continuous R.V.s)

To characterize such a R.V. X (i.e. find a blueprint) (which has infinitly precise possible values) we need something called a probability density function $P_X(x)$.

this is also a function like a PMF

Differences btwn PMFs & PDFs

Discrete Y

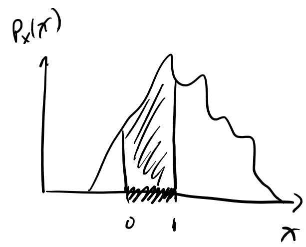


$$P(0 \le Y \le 1) = P_Y(0) + P_Y(1)$$

$$E(Y) = \sum_y y\, P_Y(y)$$

$$E(f(Y)) = \sum_y f(y)\, P_Y(y)$$

Continuous X



$$P(0 \le X \le 1) = \int_0^1 P_X(x)\,dx$$

$$E(X) = \int_{-\infty}^{\infty} x\, P_X(x)\,dx$$

$$E(f(X)) = \int_{-\infty}^{\infty} f(x)\, P_X(x)\,dx$$

Similarities btwn PMFs & PDFs

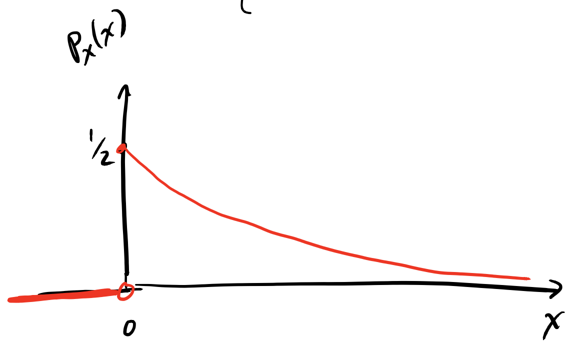
The following formulas still hold for both continuous & discrete R.V.s

- $var(X) = E(X^2) - (E(X))^2$
- $cov(X,Y) = E(XY) - E(X)E(Y)$
- $E(aX+bY) \overset{MF1}{=} aE(X) + bE(Y)$
- $var(aX+bY) \overset{MF2}{=} a^2 var(X) + b^2 var(Y) + 2ab\, cov(X,Y)$.

Example

Let X be a continuous R.V. with PDF:

$$P(x) = \begin{cases} \frac{1}{2} e^{-x/2} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$



Find $P(1 \leq X \leq 2)$, $E(X)$

Note: $P_X(x) \geq 0$

Note: $$\int_{-\infty}^{\infty} P_X(x)\,dx = \underbrace{\int_{-\infty}^{0} P_X(x)\,dx}_{=0} + \int_0^{\infty} \underbrace{P_X(x)}_{\frac{1}{2}e^{-x/2}}\,dx$$

$$= 0 + \int_0^{\infty} \frac{1}{2} e^{-x/2}\,dx$$

since $\frac{d}{dx} e^{-x/2} = (-\frac{1}{2}) e^{-x/2}$

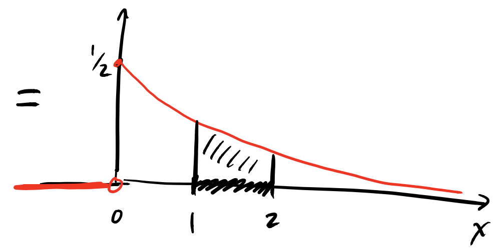
$$= \left[-e^{-x/2}\right]_0^{\infty}$$

$$= 0 - (-e^{-0/2}) = 1$$

all PDF must satisfy

$P_X(x) \geq 0$ & $\int_{-\infty}^{\infty} P_X(x)\,dx = 1$

$Pr(1 \leq X \leq 2) =$



$$= \int_1^2 \frac{1}{2} e^{-x/2}\,dx$$

$$= \left[-e^{-x/2}\right]_1^2$$

$$= -e^{-2/2} - (-e^{-1/2})$$

$$= -\frac{1}{e} + \frac{1}{\sqrt{e}}$$

$$= 0.2386$$

$$E(X) = \int_{-\infty}^{\infty} x P_X(x)\,dx$$

$$= \int_0^{\infty} x \underbrace{\frac{1}{2} e^{-x/2}\,dx}$$

$\frac{d}{dx}(-e^{-x/2}) = \frac{1}{2} e^{-x/2}$

$$= \int_0^{\infty} x\, \overbrace{d(-e^{-x/2})}$$

$$= x(-e^{-x/2})\Big|_0^{\infty} - \int_0^{\infty} (-e^{-x/2})\,dx$$

$$= 0 - 0 + \int_0^{\infty} e^{-x/2}\,dx$$

$$= \left[-2e^{-x/2}\right]_0^{\infty}$$

$$= 0 - (-2e^0)$$

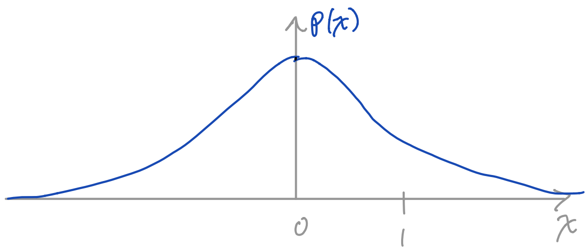
$$= 2$$

Normal Random variables

In some sense, the main reason we need to introduce continuous r.v.s is so that we can talk about Normal r.v.s. these are a special type of idealised continuous r.v. are is extremely important.

<u>example</u>: Let X be a random variable with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



This random variable satisfies

$$E(X) = 0$$
$$\mathrm{var}(X) = 1$$

and is called:

The standard Normal or the standard Gaussian or sometimes written $X \sim N(0,1)$

Now if $X \sim N(0,1)$ suppose we want to compute

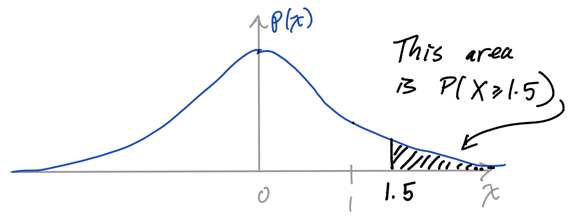
$$P(X \geq 1.5)$$

By the definition of PDF we can get this number by integrating the PDF:

$$P(X \geq 1.5) = \int_{1.5}^{\infty} p(x)\,dx$$

$$= \int_{1.5}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx$$

$=$



The problem is that $e^{-x^2/2}$ is basically impossible to integrate by hand.

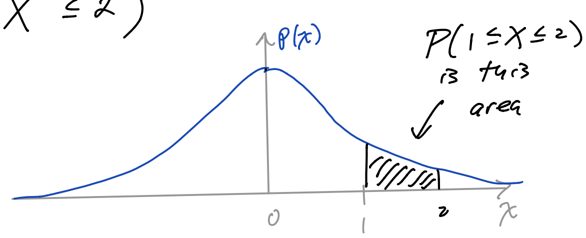
So we need to use computers or tables.

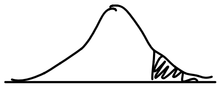
$$P(X \geq 1.5) = 0.0668$$

However most of the time you need to manipulate areas before entering values into a computer. For example,

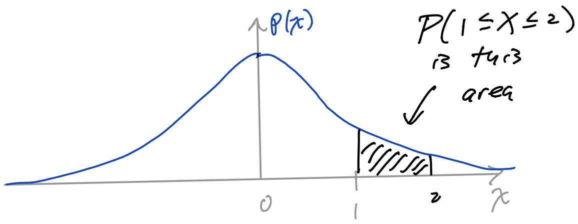
$$P(1 \leq X \leq 2)$$

$=$

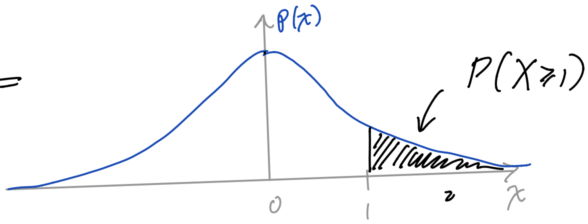


However my calculator only works for areas to the right: 

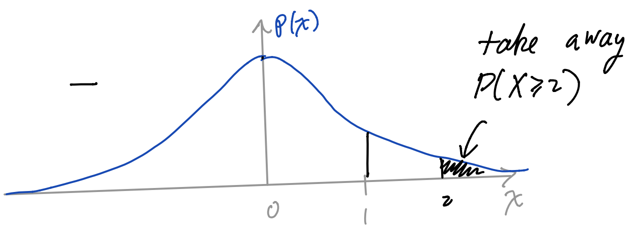
Here is how you can manipulate things to get something that works for the computer:



$=$



$-$



$= P(X \geq 1) - P(X \geq 2)$

$= 0.1587 - 0.0228$

$= \underbrace{0.1359}_{\text{this is } P(1 \le X \le 2).}$

In Fact the standard Gaussian is just one random variable in a whole class of Gaussian r.v.s:

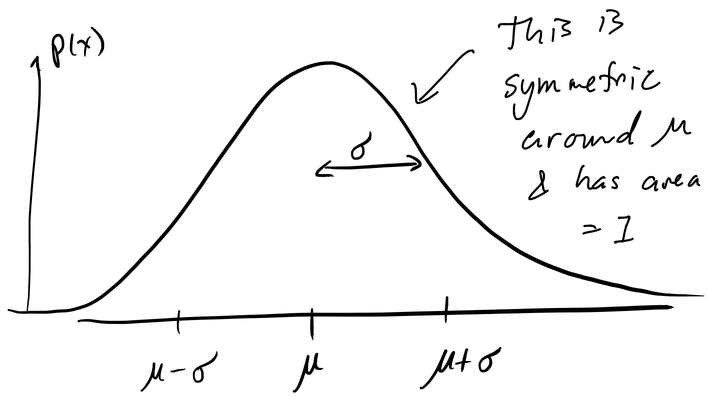
Definition: For any number μ and positive number σ^2 the three statements:

1) $X \sim N(\mu, \sigma^2)$ and

2) X is Gaussian with $E(X) = \mu$, $\mathrm{var}(X) = \sigma^2$

3) X is Normal with $E(X) = \mu$, $\mathrm{var}(X) = \sigma^2$

all mean the same thing: that X is a continuous r.v. with the following PDF:



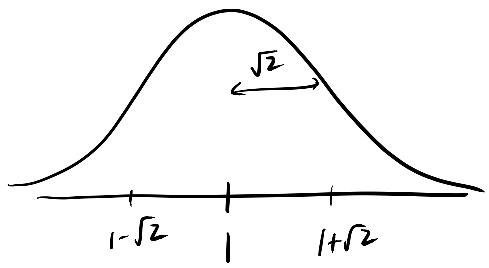
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

<u>Note:</u> If $X \sim N(\mu, \sigma^2)$ then the two numbers μ & σ^2 are called the <u>parameters of X</u>.

Example: If $X \sim N(1,2)$ then

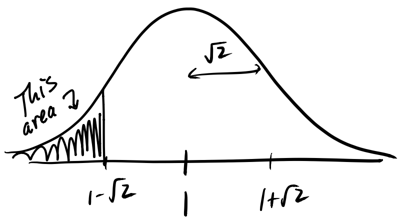
$E(X)=1$ & $var(X)=2$ and

X has PDF



So that

$P(X \leq 1-\sqrt{2})$ =



Fact 1: Any linear combination of two Gaussians is also Gaussian and the parameters can be found by Master Formulas 1 & 2.

Example:

If $X \sim N(7.2, 2.1)$ & $Y \sim N(-1.2, 10.5)$ and $cov(X,Y)=0$ then

$2X+Y \sim N(\mu, \sigma^2)$ where

$$\mu = E(2X+Y)$$
$$\overset{MF1}{=} 2E(X) + E(Y)$$
$$= 2\cdot(7.2) + (-1.2)$$
$$= 13.2$$

$$\sigma^2 = var(2X+Y)$$
$$\overset{MF2}{=} 4\,var(X) + var(Y)$$
$$= 4(2.1) + 10.5$$
$$= 18.9$$

$$\therefore \; 2X+Y \sim N(13.2, 18.9)$$

Fact 2: If $X \sim N(\mu, \sigma^2)$ then

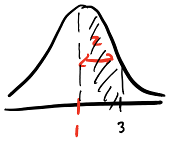
$$Z = \frac{X - E(X)}{sd(X)} = \frac{X-\mu}{\sigma} \sim N(0,1)$$

i.e. the z-score of any Gaussian r.v. is a standard Gaussian r.v.

This gives a way to compute probabilities for $X \sim N(\mu, \sigma^2)$ when X is not standard Gaussian and if your computer only works for standard Gaussian r.v.s

Example:

$X \sim N(1,4)$.

Find $P(1 \le X \le 3) =$ 

If your computer only works with $N(0,1)$ we need to convert to $N(0,1)$.

By Fact 2:

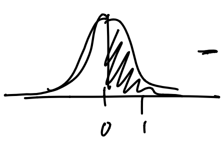
$$\frac{X-1}{2} = Z \sim N(0,1).$$

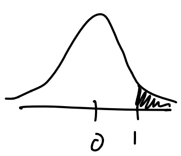
Therefore

$$P(\underbrace{1 \le X \le 3}_{\text{convert this event in terms of } Z}) = P\left(\frac{1-1}{2} \le \frac{X-1}{2} \le \frac{3-1}{2}\right)$$

$$= P(\underbrace{0 \le Z}_{\text{where } Z \sim N(0,1)} \le 1)$$

= area 

= 

− 

By symmetry $= \frac{1}{2} - P(Z \ge 1)$

$= \frac{1}{2} - 0.1587$

$= 0.3413$

Example

Suppose $Y \sim N(-5, 7)$

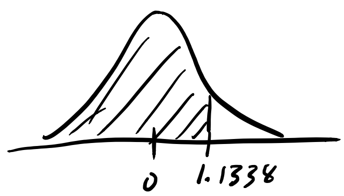
Find $P(Y < -2)$

Since $\frac{Y-(-5)}{\sqrt{7}} = Z \sim N(0,1)$ we have

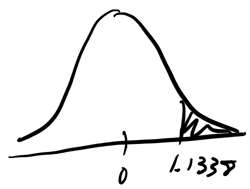
$$P(Y < -2) = P\left(\frac{Y-(-5)}{\sqrt{7}} < \frac{-2-(-5)}{\sqrt{7}}\right)$$

$$= P\left(Z < \frac{3}{\sqrt{7}}\right)$$

$$= P(Z < 1.1338)$$

= 

= 

− 

$= 1 - P(Z \ge 1.1338)$

$\approx 1 - 0.1292$

$= 0.8708.$

Percentiles vrs z-scores

Quick quiz:

Suppose $X_1, X_2, \dots, X_{100}$ are independent & identically distributed r.v.s with $E(X_1) = 0$ and $sd(X_1) = 1$.

would you be surprised if

$$\bar{X} = \frac{X_1 + \cdots + X_{100}}{100} = -1 \;?$$

Answer: You would be very surprised since $\bar{X} = -1$ corresponds to a z-score of

$$Z = \frac{\bar{X} - E(\bar{X})}{sd(\bar{X})} = \frac{(-1) - E(\bar{X})}{sd(\bar{X})}$$

where

$$E(\bar{X}) \overset{MFI}{=} \frac{1}{100}\big(\underbrace{E(X_1)}_{=0} + \cdots + \underbrace{E(X_{100})}_{=0}\big)$$

$$= 0$$

and

$$var(\bar{X}) \overset{MFI}{=} \frac{1}{(100)^2}\big(\underbrace{var(X_1)}_{=1} + \cdots + \underbrace{var(X_{100})}_{=1}\big)$$

$$= \frac{1}{100}$$

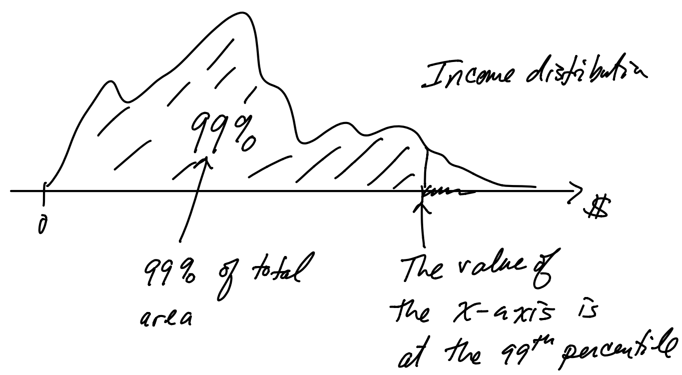
$\therefore$ The z-score for $\bar{X} = -1$ is

$$Z = \frac{(-1) - 0}{\sqrt{\frac{1}{100}}} = -10 \leftarrow \text{extremely rare.}$$

Most of the population likes to express rarity using Percentiles:

e.g.) "Her income is in the top 1%"

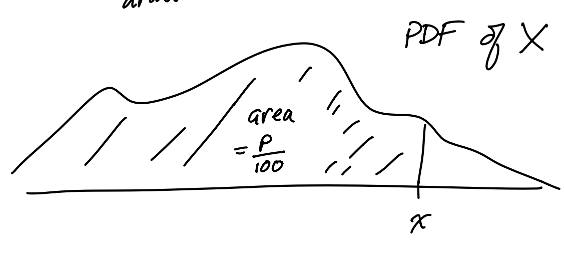
i.e. "The percentile of her income is greater than 99%"



If you observe a r.v. X to be x then we say x is at the p^{th} percentile if

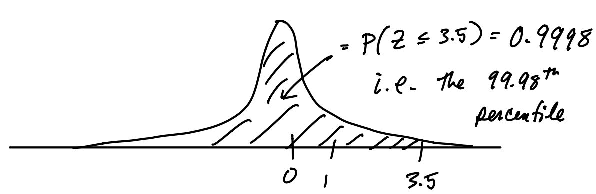
$$P(X \le x) = \frac{p}{100}$$

↑ a new draw from X



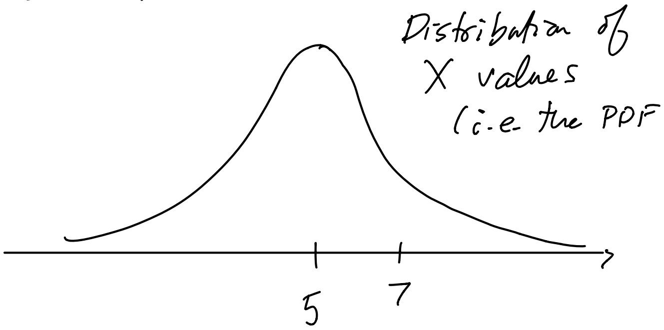
Here is how rare a z-score of 3.5 is when $X \sim N(\mu, \sigma^2)$.

z-scores for X are distributed like $N(0,1)$

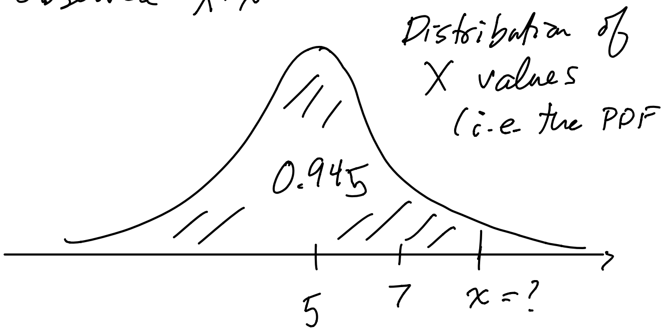


when $X \sim N(\mu, \sigma^2)$ there is a 1-to-1 correspondence btwn Z-scores and percentiles (using the fact that Z-scores are $N(0,1)$)

Example: Suppose a Gaussian $X \sim N(5,4)$ is observed to be at the 94.5^{th} percentile. What was the Z-score for that observation and what was the actual value x of the observed X?

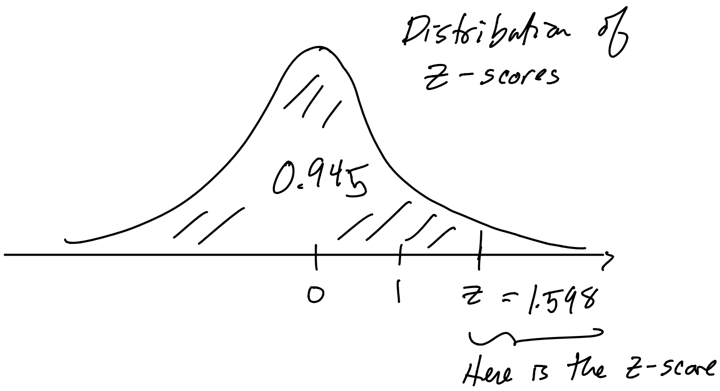


observed $X=x$



i.e. $P(X \leq x) = 0.945$

$\therefore$ the z-score for $X=x$ satisfies



So for $X \sim N(5,4)$

94.5^{th} percentile $\longleftrightarrow$ Z-score 1.598

To get the actual value just solve

$$1.598 = Z = \frac{x - E(X)}{sd(X)} = \frac{x-5}{2}$$

$$\therefore \quad x = 2(1.598) + 5$$
$$= 8.196$$

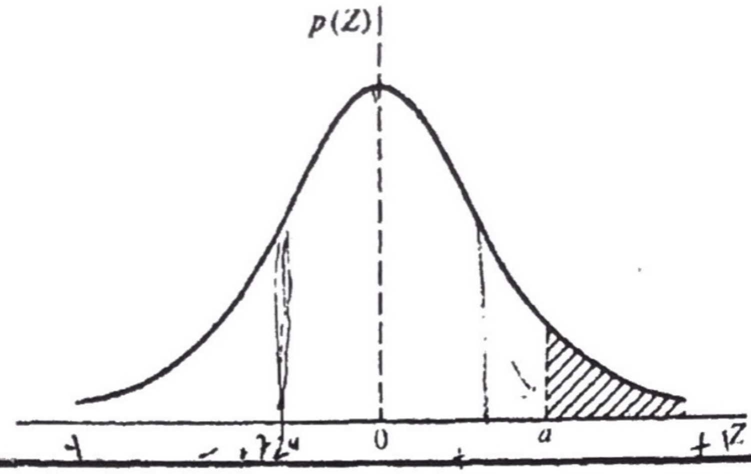
$\therefore$ for $X \sim N(5,4)$

94.5^{th} percentile $\longleftrightarrow$ Z-score 1.598

$\longleftrightarrow X = 8.196$

Table II PROBABILITIES FOR THE STANDARD NORMAL DISTRIBUTION

($\Pr(Z \geq a)$ when $Z \sim N(0, 1)$)



a	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	$\boxed{.0125}$	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

3.50	.0002326
4.00	.0000317
4.50	.0000034
5.00	.000000287