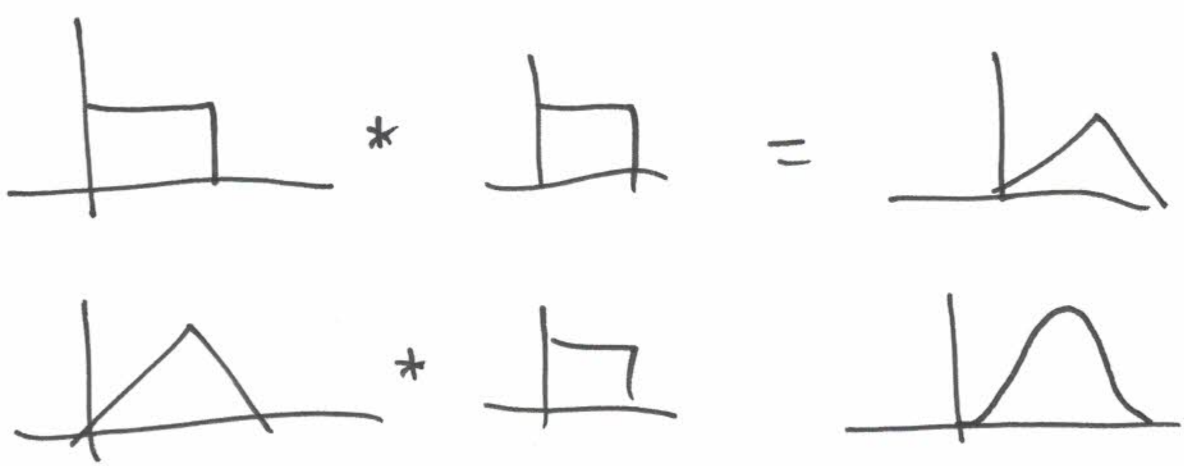
CENTRAL LIMIT THEOREM

LAST TIME:

MEAN, VARIANCE OF A SUM, AVERAGE OF I.I.D RANDOM VARIABLES.

ALSO THE PDF OF A SUM $(f_x * f_x * f_x * \dots)$



THE CENTRAL LIMIT THEOREM (CLT) TALKS ABOUT WHAT HAPPENS TO THE PDF / CDF OF THE SAMPLE MEAN AS n GETS LARGE.

SURPRISING RESULT: NO MATTER THE DISTRIBUTION OF X, IN THE LIMIT, THE DISTRIBUTION BECOMES GAUSSIAN!

LAST TIME:

X_i IID

MEAN $\mu < \infty$

VARIANCE $\sigma^2 < \infty$

$$S_n = X_1 + X_2 + \dots + X_n$$

$$M_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

$$E(S_n) = n\mu \qquad E(M_n) = \mu$$

$$Var(S_n) = n\sigma^2 \qquad Var(M_n) = \frac{1}{n}\sigma^2$$

CONSIDER A RELATED R.V.

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

$$E(Z_n) = 0$$

$$Var(Z_n) = \frac{1}{\sigma^2 n} \cdot Var(\overset{n\sigma^2}{S_n})$$

$$= 1$$

NO MATTER WHAT n IS,

Z_n IS ZERO MEAN

UNIT VARIANCE.

THE CENTRAL LIMIT THEOREM SAYS:

CDF OF Z_n.

$$\lim_{n \to \infty} P(Z_n \le u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-x^2/2}\, dx$$

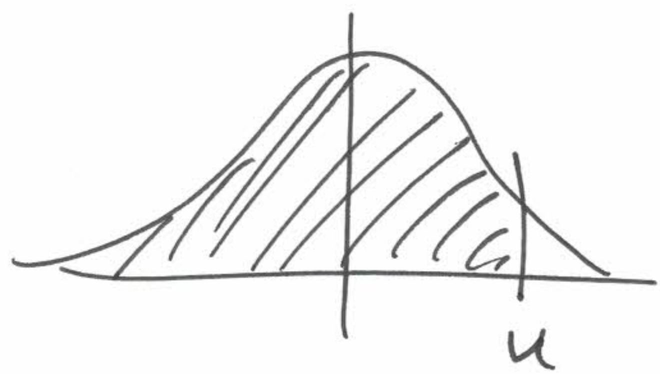
THIS IS THE CDF OF A GAUSSIAN WITH MEAN 0, VARIANCE 1.

$$f_{\text{GAUSSIAN}}(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

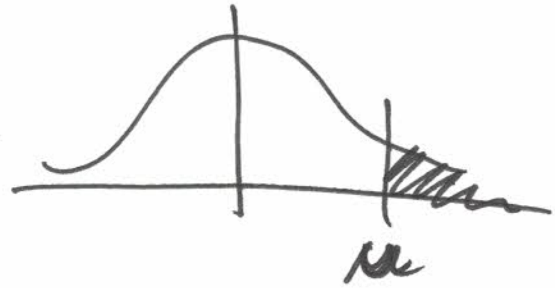
$$\lim_{n\to\infty} F_{Z_n}(u) = \Phi(u)$$

CDF OF Z_n

CDF OF A "STANDARD" GAUSSIAN WITH $\mu = 0$, $\sigma = 1$.

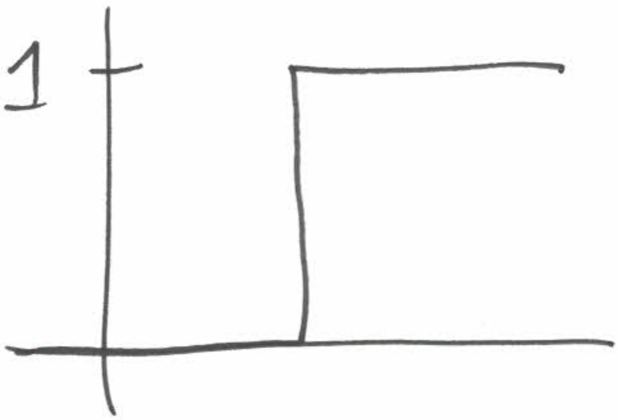


$$= 1 - Q(u)$$

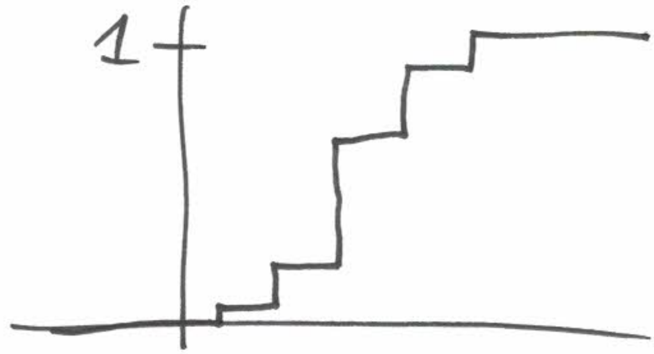
$1 -$ 

NO MATTER THE CDF OF

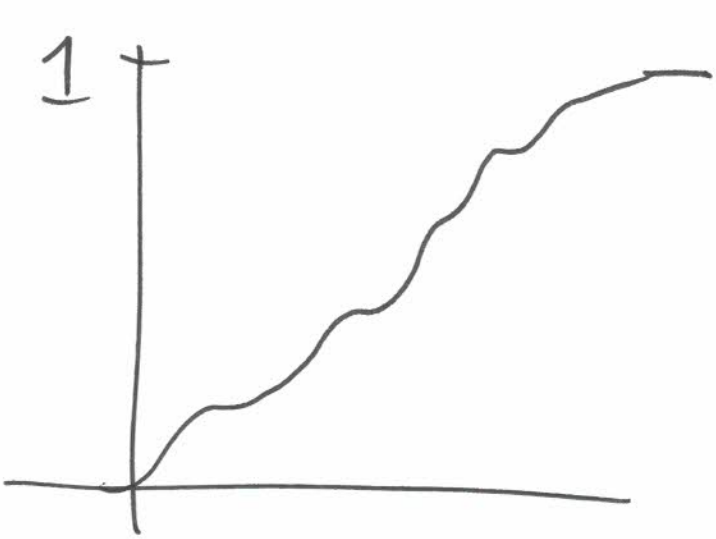
X_i!



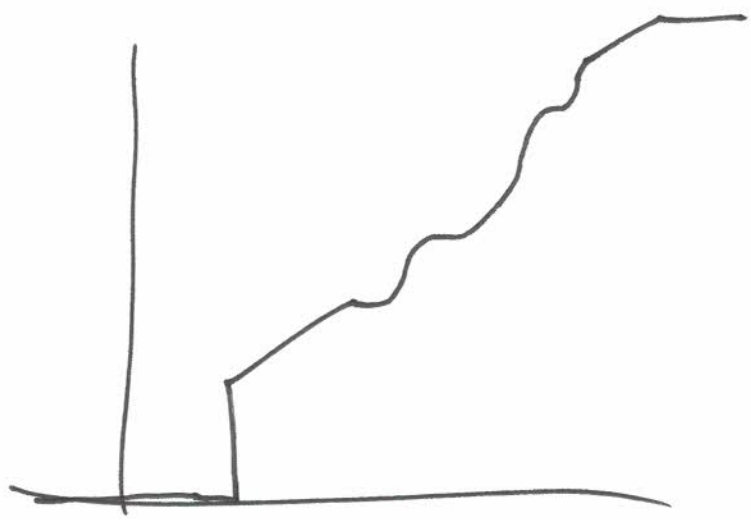
BERNOULLI



BINOMIAL

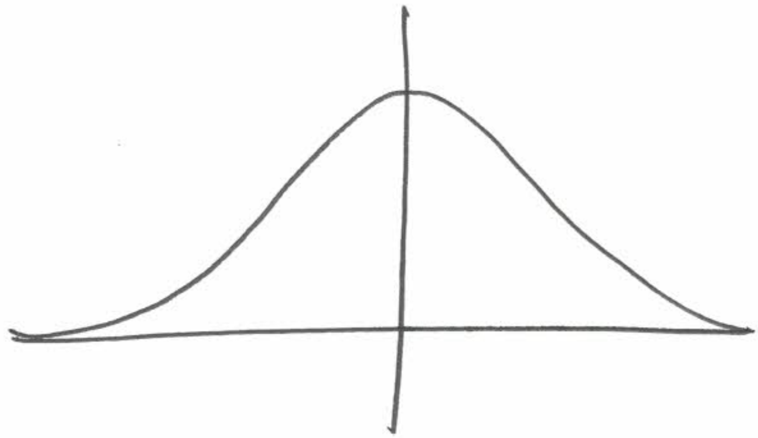


WEIRD



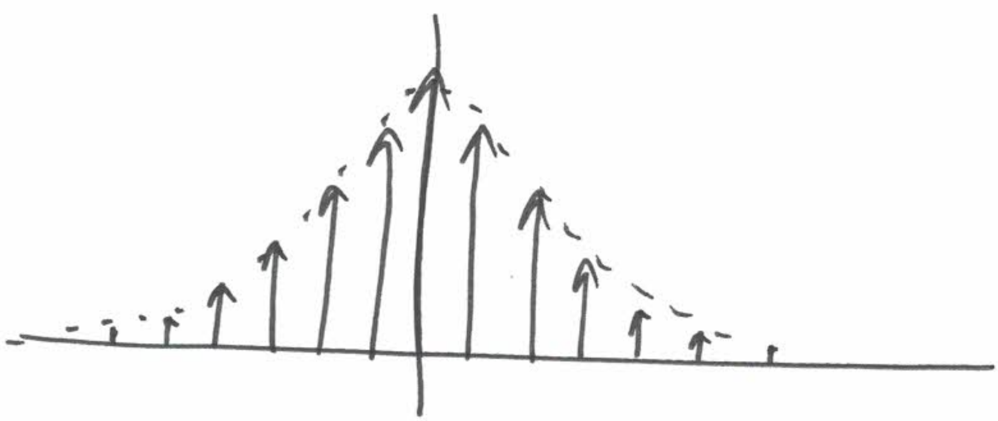
PSYCHO

FOR A CONTINUOUS PDF,



PDF OF Z_n

FOR A DISCRETE/MIXED PMF



IMPLICATION: WHEN n IS BIG, WE CAN USE THE CLT TO GET GOOD APPROXIMATIONS OF PROBABILITIES USING Q TABLE.

SPECIFICALLY:

CDF OF GAUSSIAN

$$P(Z_n \le u) = \Phi(u)$$
$$= 1 - Q(u)$$

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le u\right) = \Phi(u)$$

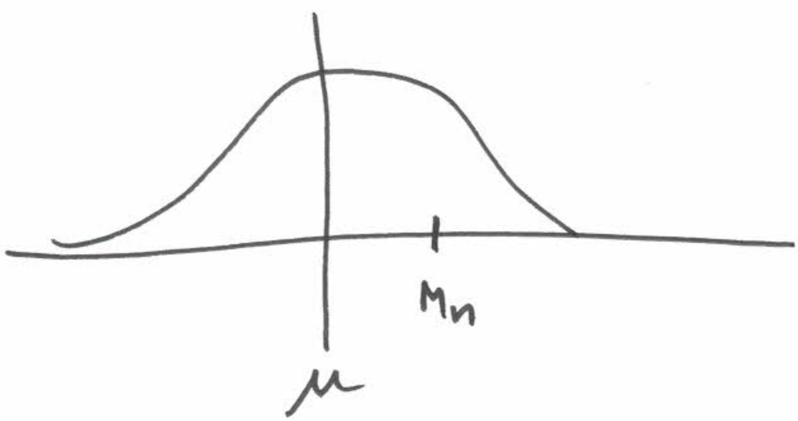
$$= P\left(S_n - n\mu \le u\sigma\sqrt{n}\right)$$

$$= P\left(M_n - \mu \le \frac{u\sigma}{\sqrt{n}}\right)$$

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right) = \phi(z)$$

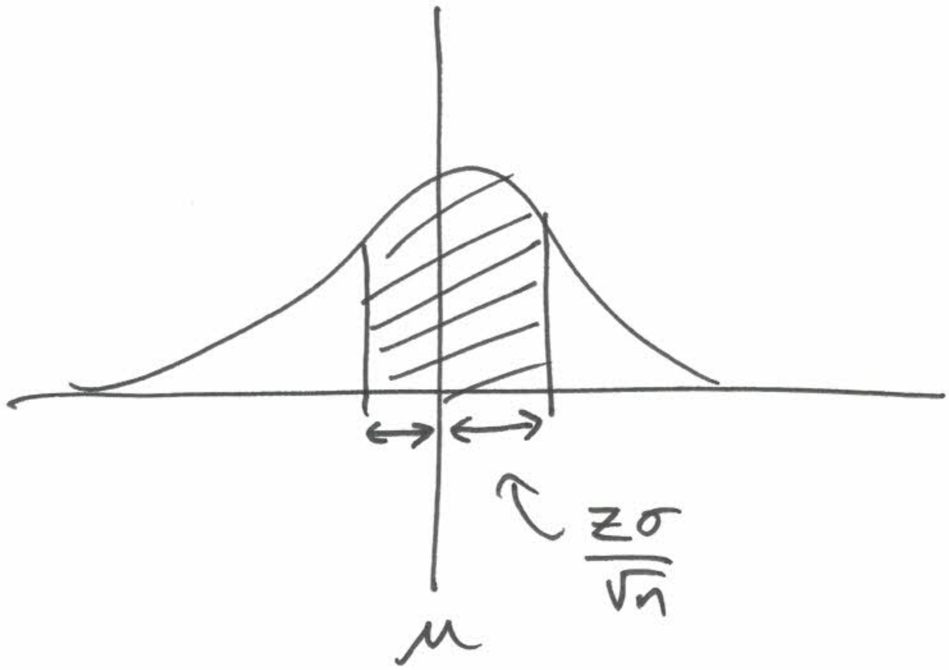
$$P\left(S_n - n\mu \le z\sigma\sqrt{n}\right) = \phi(z)$$

$$P\left(M_n - \mu \le z\sigma/\sqrt{n}\right) = \phi(z)$$



$$P\left(|M_n - \mu| \le \frac{z\sigma}{\sqrt{n}}\right)$$

$$\approx 1 - 2Q(z)$$


$\frac{z\sigma}{\sqrt{n}}$
μ

EX $X_1, X_2, \ldots X_n$

IID BERNOULLI VARIABLES

$$p = \frac{1}{2}.$$

$$\mu = \frac{1}{2}, \quad \sigma = \frac{1}{2}$$

$$\sigma^2 = \frac{1}{4}.$$

How CLOSE IS M_n TO $\frac{1}{2}$?

SAY $n = 10{,}000$

$z = 1$

$$\frac{z\sigma}{\sqrt{n}} = \frac{1 \cdot \frac{1}{2}}{100}$$

$$P\left(\left|M_{10000} - \frac{1}{2}\right| \leq \underset{}{\overset{0.005}{\overset{\shortparallel}{\frac{1}{200}}}}\right) =$$

$$1 - 2Q(\underset{\underset{z}{\uparrow}}{1}) = 1 - 2(0.159)$$

$$= 0.682$$

FOR WHAT VALUE OF n

IS

$$P\left(\left|M_n - \frac{1}{2}\right| \le \overset{0.005}{\overset{\shortparallel}{\frac{1}{200}}}\right) \ge 0.9$$

$$1 - 2Q(z) = 0.9$$

$$Q(z) = 0.05$$

$$z \approx 1.65$$

$$\frac{1}{200} = \frac{z\sigma}{\sqrt{n}} = \frac{(1.65)\left(\frac{1}{2}\right)}{\sqrt{n}}$$

$$\sqrt{n} = 200\,(1.65)\left(\tfrac{1}{2}\right) \simeq 165$$

$$n = (165)^2 = 27225.$$

A DIFFERENT SPIN:

SAY WE ~~WANT~~ HAVE 10000 SAMPLES AND WE WANT TO KNOW THE <u>CONFIDENCE INTERVAL</u> OF M_n'S APPROXIMATION TO μ.

FOR WHAT INTERVAL CAN WE BE 90% CONFIDENT?

$$P\left(\left|M_{10000} - \frac{1}{2}\right| \le ??\right) \approx 0.9$$

$\uparrow$

$z = 1.65$

$$?? = \frac{z\sigma}{\sqrt{n}} = \frac{(1.65)(\frac{1}{2})}{\sqrt{10000}}$$

$$= 0.00825$$

$$P\left(\left|M_n - \frac{1}{2}\right| \le 0.00825\right) \approx 0.9$$

$$P\left(M_n \in \left[\frac{1}{2} - 0.00825, \frac{1}{2} + 0.00825\right]\right) \simeq 0.9.$$

WHEN n IS LARGE, CLT WILL GIVE MUCH BETTER APPROXIMATIONS THAN MARKOV, CHEBYSHEV, ETC.

EX TOSS A FAIR DIE 20 TIMES. ADD UP DOTS. WHAT IS P(SUM OF DOTS IS IN [60, 80])?

FOR X, UNIFORM PMF ON $[1, 2, \ldots, 6]$

$$\mu = \frac{1+6}{2} = 3.5$$

$$\sigma^2 = \frac{35}{12} = 2.92$$

$$\sigma = 1.71.$$

CHEBYSHEV / WEAK LAW OF LARGE NUMBERS:

$$P\left(\left|S_{20} - 70\right| \le 10\right)$$

$$= P\left(\left|M_{20} - 3.5\right| \le \frac{10}{20} = \frac{1}{2}\right)$$

$$\overset{\text{WLLN}}{\geq} 1 - \frac{\sigma^2 = 2.92}{20 \cdot \left(\frac{1}{2}\right)^2} = \frac{5}{12}$$

$$= \boxed{0.42}$$

ON THE OTHER HAND,

THE CLT WOULD

TELL US.

$$P\left(\left|M_{20}-3.5\right| \le \tfrac{1}{2}\right) \approx 1-2Q(z)$$

$$\tfrac{1}{2} = \frac{z\sigma}{\sqrt{n}}$$

$$\frac{z\sigma}{\sqrt{n}} = \frac{1}{2}$$

$$\frac{z\sigma}{\sqrt{n}} = \frac{z(1.71)}{\sqrt{20}} = \frac{1}{2} \qquad z = 1.31$$

$$1-2Q(1.31) = 1-2(0.095)$$

$$= \boxed{0.81}$$

IN REAL WORLD ESTIMATION PROBLEMS, WE OFTEN WANT TO ESTIMATE AN UNKNOWN PARAMETER BY AVERAGING A BUNCH OF SAMPLES.

X_i IID

$$M_n = \overline{X_n} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

IN THE LIMIT

$$\overline{X_n} \to \mu = E(X_i)$$

LET $1-\alpha$ BE SOME HIGH PROBABILITY (e.g. 0.99)

α IS A SMALL NUMBER.

FIND ℓ AND u S.T.

$$P(\ell(x) \le \mu \le u(x)) = 1-\alpha$$

$$P(\mu \in [\ell(x), u(x)]) = 1-\alpha$$

THAT IS, WE "KNOW" THAT μ IS IN A DETERMINED RANGE WITH HIGH PROBABILITY.

CONF. LEVEL

$\boxed{(1-\alpha)\% \text{ CONFIDENCE INTERVAL.}}$

SUPPOSE $\{X_j\}$ ARE IID

UNKNOWN μ

~~UN~~ KNOWN σ

$$1 - 2Q(z) =$$

$$P\left(-z \le \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \le z\right)$$

$$= P\left(\frac{-z\sigma}{\sqrt{n}} \le \overline{X}_n - \mu \le \frac{z\sigma}{\sqrt{n}}\right)$$

$$= P\left(\overline{X}_n - \frac{z\sigma}{\sqrt{n}} \le \mu \le \overline{X}_n + \frac{z\sigma}{\sqrt{n}}\right)$$

THIS INTERVAL CONTAINS μ

WITH PROBABILITY $1 - 2Q(z)$

<u>Ex</u> X_j HAVE UNKNOWN MEAN AND VARIANCE 1.

$\mu = ?$ $\sigma = 1$.

MEASURE X 100 TIMES.

FIND OUT THAT $\overline{X}_{100} = 5.25$.

FIND THE 95% CONFIDENCE INTERVAL ON μ.

$n = 100$

$\sigma = 1$

$\alpha = 0.05 = 1 - 0.95$

Z s.t $Q(z) = \frac{0.05}{2} = 0.025$

$z_{0.025} = 1.96$ FROM TABLE.

SO IF WE SET α

AND COMPUTE $z_{\alpha/2}$ SO THAT

$$2Q(z) = \alpha \qquad Q(z) = \frac{\alpha}{2}$$

$$1 - \alpha =$$

$$P\left(\mu \in X_n \pm z_{\alpha/2}\,\sigma/\sqrt{n}\right)$$

THE 95% CONFIDENCE INTERVAL FOR THE MEAN IS THUS:

$$5.25 \pm 1.96 \cdot 1/\sqrt{100}$$

$$5.25 \pm 0.196$$

$$= [5.05, 5.45]$$

IF WE DON'T KNOW σ

AND n IS LARGE,

WE CAN STILL ESTIMATE

CONFIDENCE INTERVALS ON μ

USING THE

STUDENT'S t DISTRIBUTION.

OR IF WE KNOW μ

AND WANT TO FIND CONFIDENCE

INTERVALS ON σ^2, WE

CAN USE LOOKUP TABLES

BASED ON THE χ^2 (CHI-SQUARED)

DISTRIBUTION.

z	Q(z)	z	Q(z)
0.0	5.000e-01	3.0	1.350e-03
0.1	4.602e-01	3.1	9.677e-04
0.2	4.207e-01	3.2	6.872e-04
0.3	3.821e-01	3.3	4.835e-04
0.4	3.446e-01	3.4	3.370e-04
0.5	3.085e-01	3.5	2.327e-04
0.6	2.743e-01	3.6	1.591e-04
0.7	2.420e-01	3.7	1.078e-04
0.8	2.119e-01	3.8	7.237e-05
0.9	1.841e-01	3.9	4.812e-05
1.0	1.587e-01	4.0	3.17e-05
1.1	1.357e-01	4.5	3.40e-06
1.2	1.151e-01	5.0	2.87e-07
1.3	9.680e-02	5.5	1.90e-08
1.4	8.076e-02	6.0	9.87e-10
1.5	6.681e-02	6.5	4.02e-11
1.6	5.480e-02	7.0	1.28e-12
1.7	4.457e-02	7.5	3.19e-14
1.8	3.593e-02	8.0	6.22e-16
1.9	2.872e-02	8.5	9.48e-18
2.0	2.275e-02	9.0	1.13e-19
2.1	1.786e-02	9.5	1.05e-21
2.2	1.390e-02	10.0	7.62e-24
2.3	1.072e-02		
2.4	8.198e-03		
2.5	6.210e-03		
2.6	4.661e-03		
2.7	3.467e-03		
2.8	2.555e-03		
2.9	1.866e-03		