Z1/
Epsylon maszynowe:
fl(1.0 = eps) = 1.0+ eps
eps = 2^{1-t}

Biasy
Float16 : 15
Float32 : 127
Float64 : 1023

FloatMax = $+m_{max}*2^{cmax}$

$m_{max}=m_{max}*2^0=2^1$-eps(Float)
2^{cmax} – wyznaczamy iteracyjnie
eps – wyznaczony wcześniej
FloatMax = $m_{max}*2^{cmax}=(2^1-eps)*2^{cmax}=(2^1-2^{1-t})*2^{cmax}$

EPS for Float64 = 2.220446049250313e-16 ~
EPS for Float32 = 1.1920929e-7 ~
EPS for Float16 = 0.000977 ~
F64:Float64
Macheps64: 2.220446049250313e-16
1.0000000000000002
F32:Float32
Macheps32: 1.1920929e-7
1.0000001
F16:Float16
Macheps16: 0.000977
1.001
NextFloat64: 5.0e-324 ~
NextFloat32: 1.0e-45 ~
NextFloat16: 6.0e-8 ~
E64:Float64
eta: 5.0e-324
E32:Float32
eta: 1.0e-45
E16:Float16
eta: 6.0e-8
MaxFloat64: 1.7976931348623157e308 ~
MaxFloat32: 3.4028235e38 ~
MaxFloat16: 6.55e4 ~
Max 64:Float64
Max: 1.7976931348623157e308
Max 32:Float32
Max: 3.4028235e38
Max 16:Float16
Max: 6.55e4

Float.h:

=== FLOAT ===
FLT_MAX   = 3.4028234664e+38
FLT_MIN   = 1.1754943508e-38
FLT_EPSILON = 1.1920928955e-07
=== DOUBLE ===
DBL_MAX   = 1.7976931348623157081 5e+308
DBL_MIN   = 2.2250738585072013830 9e-308
DBL_EPSILON = 2.2204460492503130808 5e-16
=== LONG DOUBLE ===
LDBL_MAX   = 1.1897314953572317650212638530 31e+4932
LDBL_MIN   = 3.3621031431120935062626778173 22e-4932
LDBL_EPSILON = 1.0842021724855044340074528008 70e-19

Jaki związek ma liczba macheps z precyzją arytmetyki?
Precyzja arytmetyki $\mathcal{E} = 0.5\beta^{1-t}$. Wynika to z tego, że błąd musi spełniać założenia $|\delta| \le \mathcal{E}$ i $fl(1.0 + \delta) = 1.0$. Oznacza to, że nie może zostać zaokrąglona w górę, ponieważ wtedy będzie już następną liczbą.

Jaki związek ma liczba eta z liczbą MINsub (zob. wykład lub raport [1])?
$Min_{sub}=m_{min}*\beta^{cmin}=\beta^{1-t}*\beta^{cmin}=\beta^{cmin+(1-t)}$
$MIN_{nor}=1*\beta^{cmin}$
Liczba eta = Min_{sub}

Dlaczego $Min_{sub} \neq$ Macheps
$\mathcal{E}$ oznacza precyzję arytmetyki, więc jest najmniejszą liczbą w zakresie cyfr mantysy t. (Tłumaczenie dla mnie żebym zrozumiał)

Co zwracają funkcje floatmin(Float32) i floatmin(Float64) i jaki jest związek zwracanych wartości z liczbą MIN_{nor}
Floatmin jest MIN_{nor} dla danych typów

biasy- https://en.wikipedia.org/wiki/Exponent_bias

Z2/
eps64: 2.220446049250313e-16 ~
-2.220446049250313e-16
eps32: 1.1920929e-7 ~
1.1920929e-7
eps16: 0.000977 ~
-0.000977

Czasami minus wynika z zaokrągleń

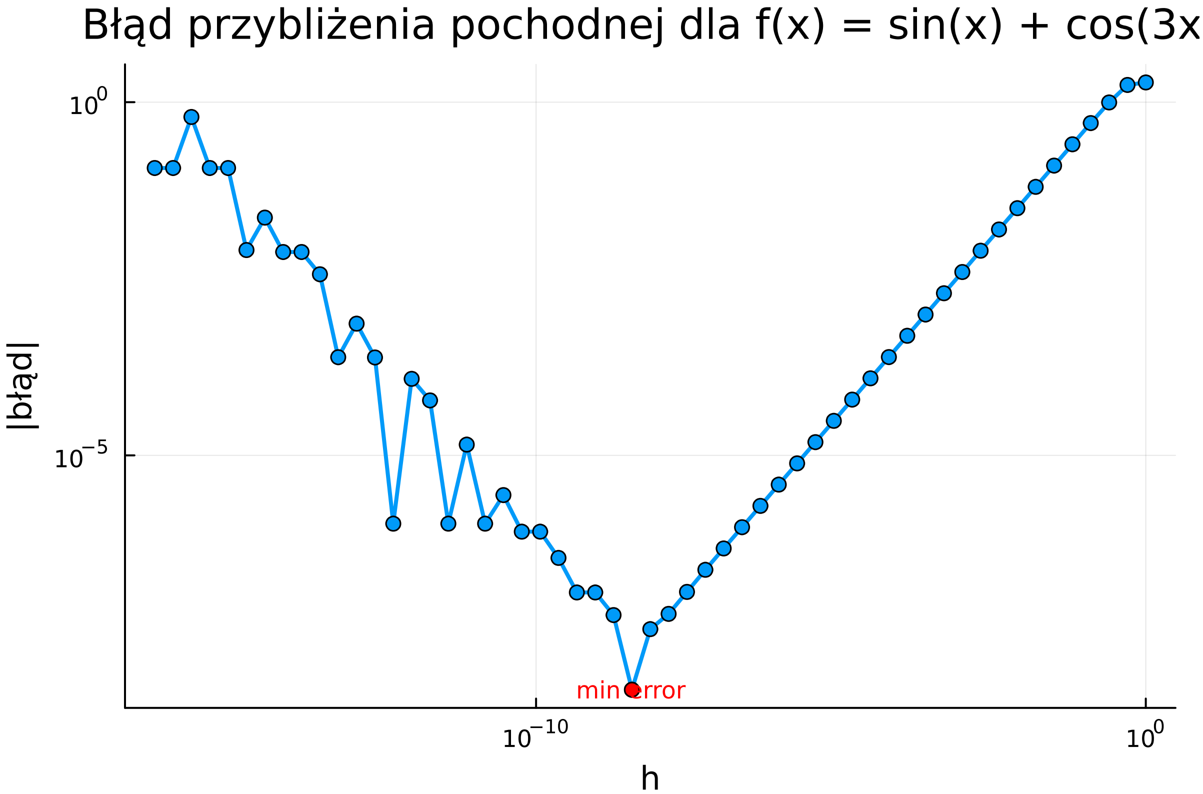
Z3/
$\delta = 2^{-52}$
dla [0.5,1] δ jest dwukrotnie za duże
dla [1,2] δ jest idealne
dla [2,4] δ jest dwukrotnie za małe

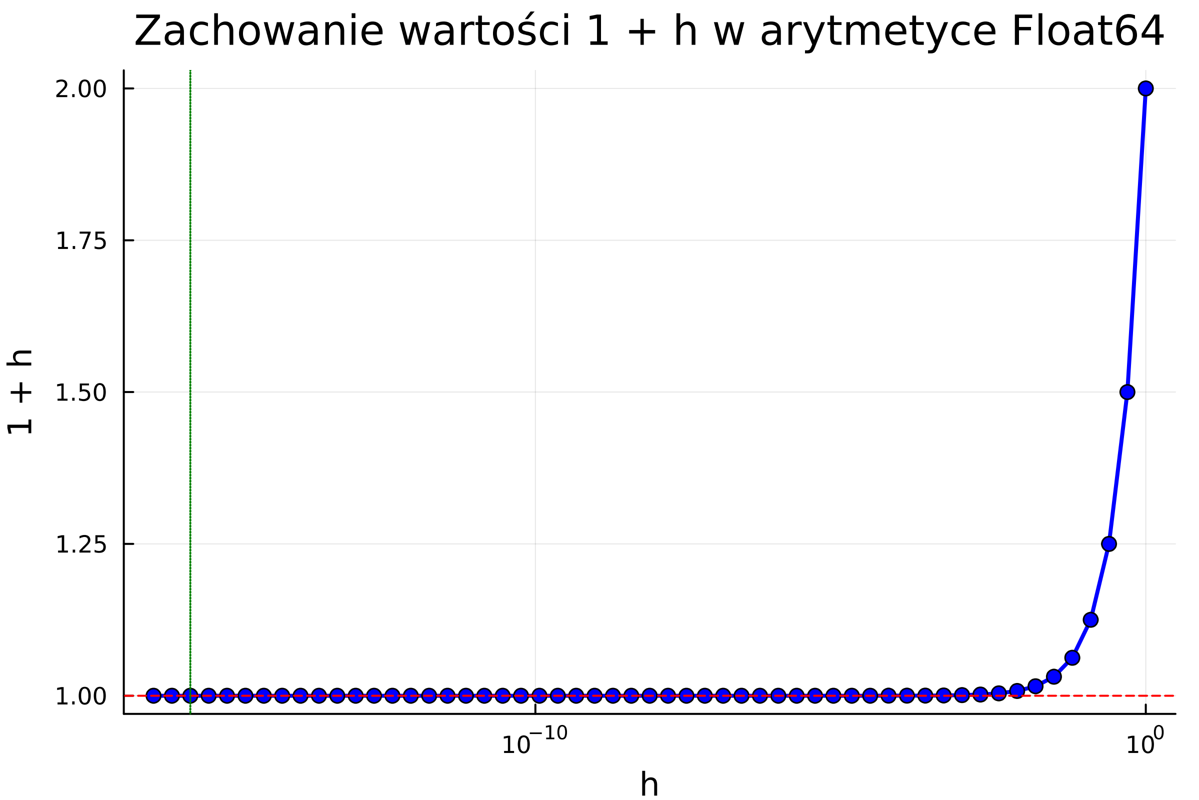
Z4/

Z5/
Wyniki:
pa: 1.0251881368296672e-10
pb: -1.5643308870494366e-10
pc: 5.653547379013446e6
pd: 5.653547379013446e6
pa32: -0.4999443
pb32: -0.4543457
pc32: 5.653547e6
pd32: 5.6535475e6

Z6/
g jest bardziej wiarygodne pewnie przez odejmowanie przy którym musi następować redukcja cyfr znaczących

Z7/





Kod do wykresów wygenerował ChatGPT