

## IEEE Arithmetik – Standard 754 (1985)

Datentyp	$p$	$t$	$e_{\min}$	$e_{\max}$	$u$	$x_{\min}$	$x_{\max}$
single	2	$23 + 1$	-125	128	$\approx 5.96 \cdot 10^{-8}$	$\approx 10^{-38}$	$\approx 10^{38}$
double	2	$52 + 1$	-1021	1024	$\approx 1.11 \cdot 10^{-16}$	$\approx 10^{-308}$	$\approx 10^{308}$

### Exception Handling

Flag	Beispiel	Ergebnis
<i>invalid</i>	$0/0, 0 \cdot \infty, \sqrt{-1},$ $\infty/\infty, +\infty + (-\infty)$	NaN (“not a number”)
<i>overflow</i>	$x_{\max} * x_{\max}$	$\pm\infty$ in MATLAB: Inf
<i>division by zero</i>	$x/0$ für $x \neq 0$	$\pm\infty$
<i>underflow</i>	$x_{\min}/p^s, 1 < s < t$	subnormale Zahlen
<i>inexact</i>	$\text{rd}(x \circ y) \neq x \circ y$	korrekt gerundetes Resultat

## Speicherung von single-bzw. double-Variablen:

single: (32 bit)

```
V EEEEEEEE MMMMMMMMMMMMMMMMMMMMMMMMMMMM
0 1      8 9                      31
```

### Beispiele:

0 11111111 000000000000000000000000 =  $+\infty$

1 11111111 000000000000000000000000 =  $-\infty$

0 11111111 000010000000000000000000 = NaN

1 11111111 001000100010010101010101 = NaN

0 10000000 000000000000000000000000 =  $+1.0 * 2^{128-127} = 2$

0 10000001 101000000000000000000000 =  $+1.101 * 2^{129-127} = 6.5$

1 10000001 101000000000000000000000 =  $-1.101 * 2^{129-127} = -6.5$

0 00000001 000000000000000000000000 =  $+1.0 * 2^{1-127} = 2^{-126} = x_{\min}$

0 00000000 100000000000000000000000 =  $+0.1 * 2^{-126} = 2^{-127}$

0 00000000 000000000000000000000001 =  $+0.0\dots01 * 2^{-126} = 2^{-149}$   
 = kleinste darstellbare Zahl

double: (64 bit)

```
V EEEEEEEEEEE MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM
0 1          11 12                                          63
```