**java关于float,double浮点数的处理**[**http://www.docin.com/p-877573409.html**](http://www.docin.com/p-877573409.html)

float：首先将整数部分转化为2进制，再加上转化成2进制的小数部分，最后采用科学计数法表示（例如：20.5，其整数部分为10100，小数部分为0.1，所以其二进制的科学计数法为：1.01001\*2^4，即1.01001E100）后进行存储，float存储包含4个字节，32位，其中有1位是符号位(sign:0->正；1->负)，还有8位指数位(取值0~255)，偏移位为127（2^7-1）， 故其真正的取值范围为-127~128（另外全0和全1做特殊处理，故真正的取值范围为-126~127），最后就是23位尾数位，一般情况下默认前导为1（1.f），故上述20.5的存储结构（存储时指数会加上偏移位127进行存储）为：

0-10000 011-01001 00000 00000 00000 000（前导默认为1，尾数只存储了小数部分）

double：和float 一样也需要先转化为二进制的科学计数法，然后进行存储，double存储包含8个字节，64位，其中有1位是符号位（sign:0->正；1->负），还有11位指数位（取值0~2047），偏移位为1023（2^10-1），故其真正的取值范围为-1023~1024（全0和全1做特殊处理，所以真正的取值范围为-1022~1023），最后就是52位尾数位，一般情况下默认前导是1(1.d)，故上述20.5的存储结构（存储时指数会加上偏移位1023进行存储）为：

0-10000 000011-01001 00000 00000 00000 00000 00000 00000 00000 00000 00000 00（前导默认为1，尾数只存储了小数部分）