

**Statistics Problems.**

Hypothesis Testing I:

1. The following are real data from an expression study: gene expression was measured in cells exposed to two conditions. For each condition mRNA levels were measured on 4 microarrays. The normalized data for the gene Metallothionein I is displayed below.

Condition/Gene	1	1	1	1	2	2	2	2
Metallothionein I	13.04	13.12	10.29	11.42	83.42	85.8	78.08	79.63

- a) What are the sample mean and standard deviation for Metallothionein expression under each condition?
- b) What are the Maximum Likelihood Estimates for the mean and standard deviation of Metallothionein under each condition (assume Gaussian distributions for conditions 1 and 2)?
- c) Do the data support a difference in expression level between condition 1 and 2? Assume that the expression levels are modeled as Gaussians with equal variances. Give a test statistic and a p-value for the null hypothesis that the gene has identical means under each condition.

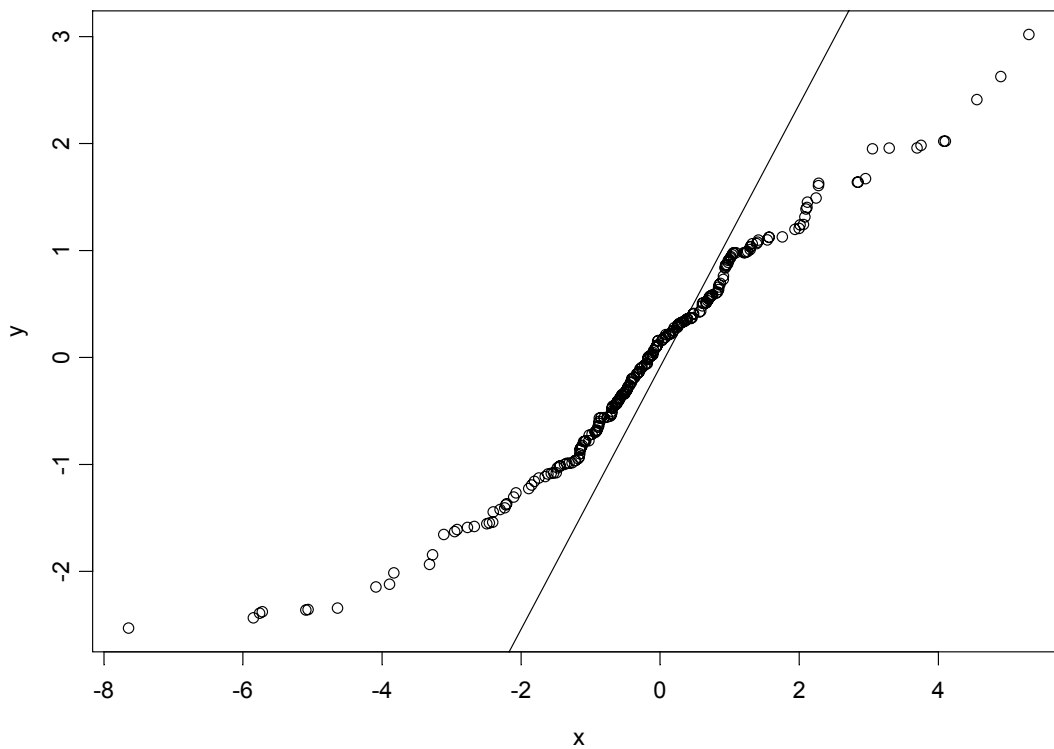
Hypothesis Testing II:

2. The result of hypothesis test is not black or white. We accept or reject hypotheses with a certain probability of error. Typically, if under the null hypothesis the probability of the test statistic is less than .05, we reject the null hypothesis. But this means that about 1 in 20 times we will be fooled. Consider the following example: you have two batches of cells: one batch is exposed to a drug, the other is exposed to “vehicle” (just the solvent that the drug is diluted in). You will use microarrays to measure gene expression in both batches and would like to evaluate which genes are induced by the drug. Suppose the arrays contain probes for 10000 mouse genes.
  - a) What is the danger of straightforward hypothesis testing?
  - b) What if by mistake we forgot to dissolve drug in to the tube labeled “drug”, and went ahead with the experiment – what would be the magnitude of the problem you identified in a)?
  - c) Can you suggest a simple one step correction to the hypothesis testing procedure so that the scenario in b) would be unlikely to present you with difficulty?
  - d) What is the disadvantage of permanently incorporating the correction from c) in to your studies of gene expression?

Percentile Plots for Comparing Distributions:

3.

- a) Give a rough sketch (the relative difference is what's important to capture) of the probability density functions corresponding to the two data sets,  $x$  and  $y$ , plotted against each other on the following percentile plot (commonly referred to as a qqplot for quantile-quantile plot)



- b) Draw a qqplot of  $N(1,1)$  vs.  $N(0,1)$  [ $N(a,b)$  refers to a normal distribution with mean  $a$  and stdev  $b$ ]. Label the axes.  
c) Draw a qqplot of  $N(1,1)$  vs.  $N(1,2)$ . Label the axes.

The Assumption of Gaussianity:

4. Using what you've learned, suggest a simple test statistic to evaluate whether a particular data set (let's say 100 measurements of gene expression for Metallothionein) has been drawn from a gaussian distribution.

