University of Washington

STATISTICS

Applied Statistics and Experimental Design

Fritz Scholz

Fall Quarter 2006

Introduction: Census and Samples $\longrightarrow$ Induction

Statistics originally served to describe matters of the state (status of state) by capturing numerically various aspects of full populations.

Today this is called a census, from Latin censere (to count or estimate).
Recall the historical census of Emperor Augustus.

Nowadays the field of statistics focusses mainly on samples, i.e., part of the total.

The goal is to make statements or conclusions about the whole population.

This process is referred to as induction, from Latin inducere (to lead to).

Its validity depends crucially on the process of sampling.

Sample $\leftarrow$ example from Latin exemplum.

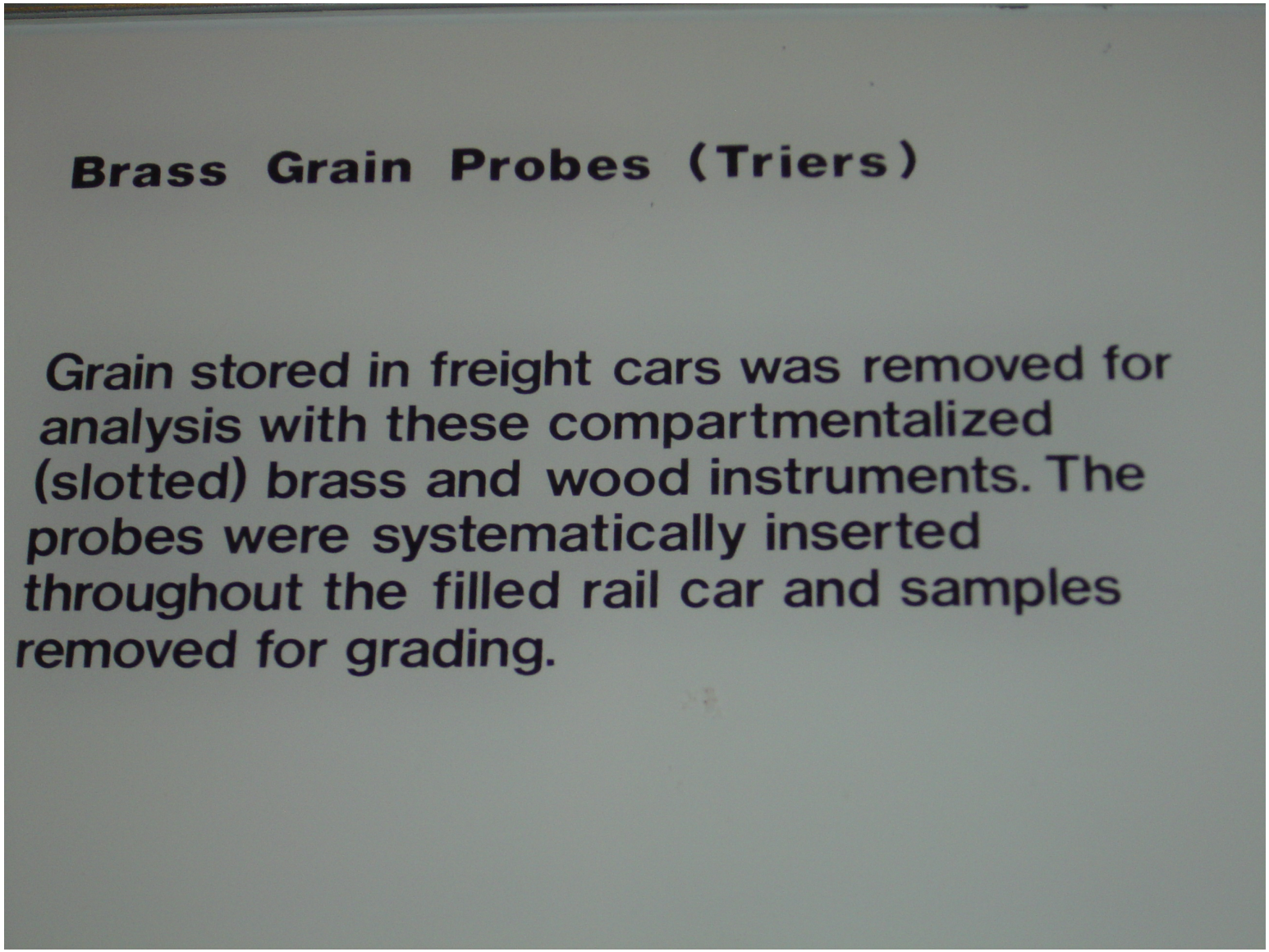
Brass Grain Probes–Stick Probe

Stichprobe

German

for Sample

Chicago Board of Trade (CBOT)



Experiments and Observational Studies

In observational studies we obtain measurements on several variables.

Sampling could be random or not.

It is not clear which variables have an effect on which other variables
if we observe any correlations.

There may be unmeasured factors that affect seemingly correlated variables.

In a "controlled" experiment we control certain input variables and
determine their effect on response variables.

We have to guard against subconscious effects when "controlling" inputs.

Effect of Estrogen Treatment on Post-Menopausal Women?

Population: Healthy post-menopausal women in the U.S.
Women's Health Initiative (WHI)

Input variables:
estrogen treatment (yes/no)
demographic variables (age, race, diet, family history, ...)
unmeasured variables

Output variables (responses):
coronary heart disease (CHD)
invasive breast cancer
others

Question: How does estrogen treatment affect health outcomes?

Results of Observational Study

In the observational study the variables of interest were measured for each subject in the given available sample.

The subjects possibly were a random sample.

Findings: good health and low rates of CHD are more prevalent in the estrogen portion of the sample.

Is this a valid conclusion?

What could be wrong?

Experimental Study (WHI Randomized Controlled Trial)

373,092 women were determined to be eligible, 18,845 consented to take part (not knowing whether treatment would be estrogen or placebo)

16,608 were included in the experiment.

These women were divided into different blocks
K clinics by 3 age groups 50-59, 60-69, 70-79.

Within each block half the women (randomly chosen within each block) were assigned to the estrogen treatment, the other half was given a placebo control.

This is a randomized block design.

Why blocking and randomization? Why not randomly split the 16,608 subjects?

Results (JAMA, July 17, 2002)

Compared with the control group, women on the estrogen treatment had higher rates of

CHD

breast cancer

stroke

pulmonary embolism

and lower rates of

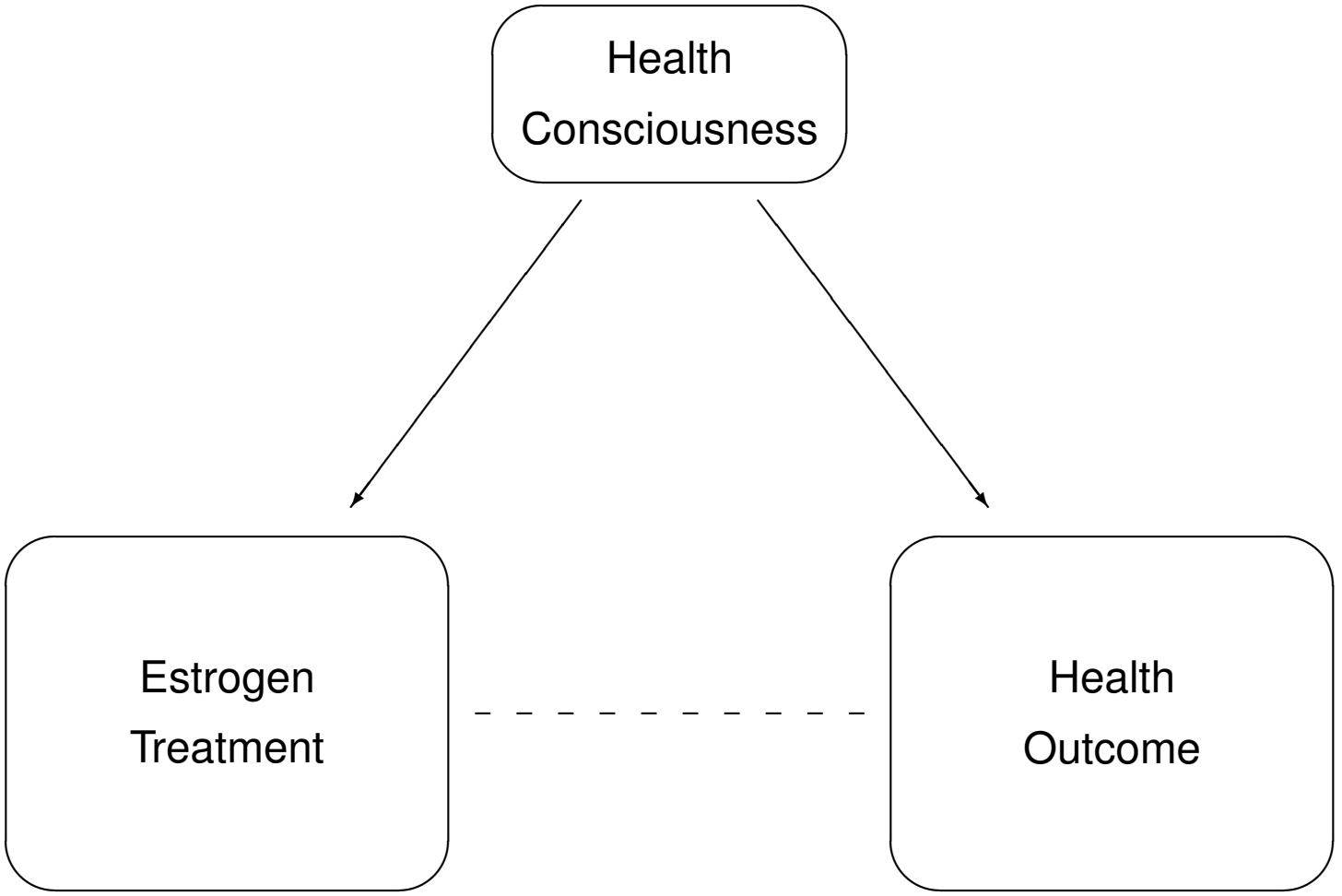
colorectal cancer

hip fracture

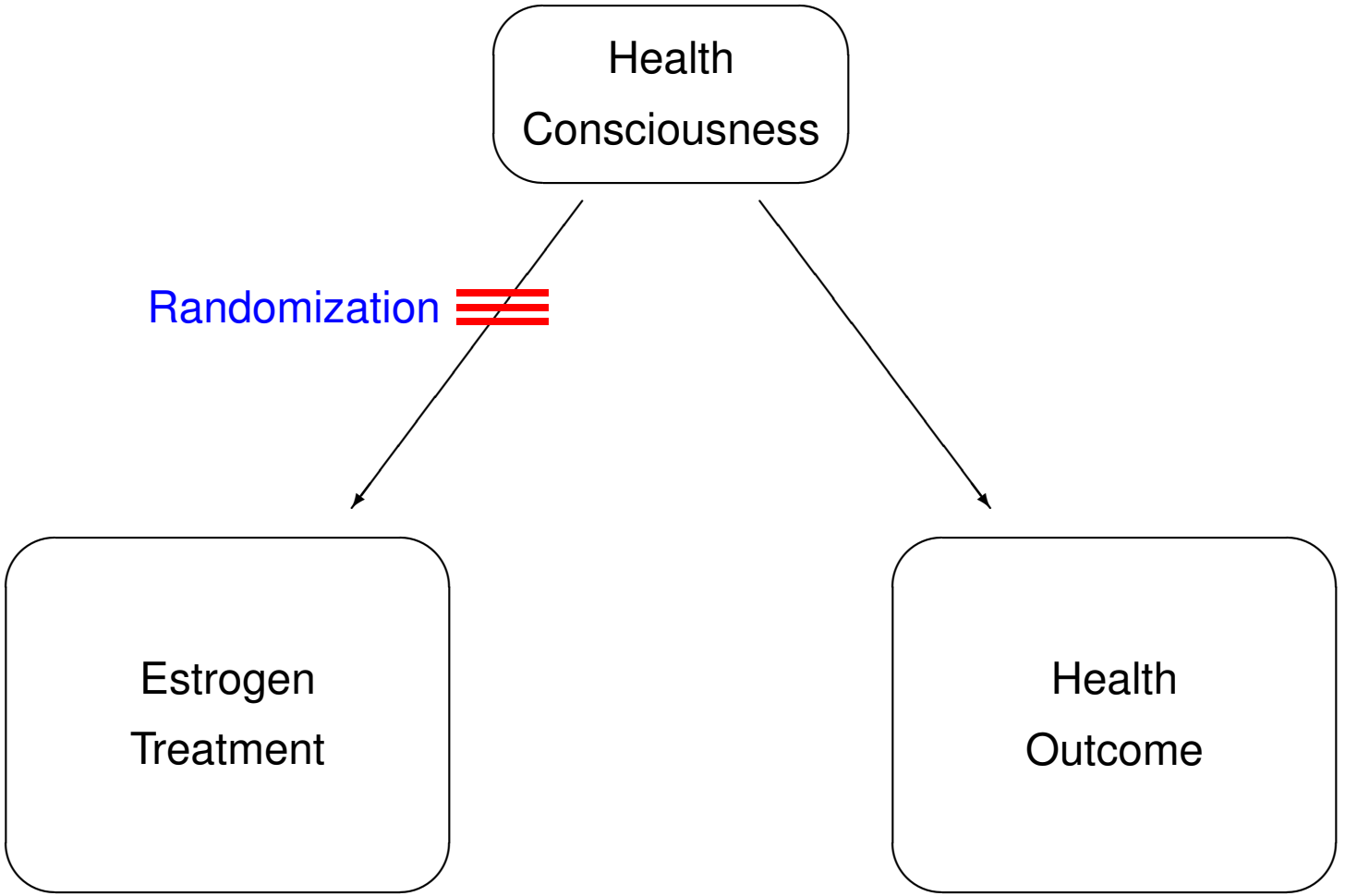
Conclusion: Estrogen cannot be viewed as a viable preventative measure for CHD in the general population.

Why the different conclusion?

Possible Explanation



Break Link by Randomization



Ethical Randomization Issues

Smoking and lung cancer.

Both could be linked to stress, an unmeasured variable.

Assigning smoking randomly is not a viable approach.

One could attempt to measure stress on some scale.
However, this again runs into the hidden factor problem,
health conscious people dealing with stress differently.

Smoking, Personality and Stress:
Psychosocial Factors in Prevention of Cancer & Heart Disease
By Hans J Eysenck

Steps in Designing of Experiments (DOE)

1. Be clear on the goal of the experiment. Which questions to address?
   Set up hypotheses about treatment/factor effects, a priori.
   Don't go fishing afterwards! It can only point to future experiments.
2. Understand the experimental units over which treatments will be randomized.
   Where do they come from? How do they vary? Are they well defined?
3. Define the appropriate response variable to be measured.
4. Define potential sources of response variation
   a) factors of interest
   b) nuisance factors
5. Decide on treatment and blocking variables.
6. Define clearly the experimental process and what is randomized.

Three Basic Principles in Experimental Design

Replication:

repeat experimental runs under same values for control variables.

$\Rightarrow$ understanding inherent variability

$\Rightarrow$ better response estimate via averaging.

Repeat all variation aspects of an experimental run.

Randomization:

Confounding between treatment and other factors (hidden or not) unlikely.

Removes sources of bias arising from factor/unit interaction.

Provides logical/probability basis for inference about treatment effects.

Blocking:

Randomized treatment assignment within blocks

Separates variation between blocks from treatment effect (variation within blocks)

Most effective when blocks are homogeneous.

Makes treatment effect more clearly visible, i.e., increases test power.

Circuit Board Solder Flux Experiment (Courtesy Denis Janky)

Flux is material used to facilitate soldering.

Moisture condensing on the boards can interact with soldering contaminants (usually residual solder flux) and result in thin filaments (dendrites) on the surface.

The dendrites can carry current and disrupt the circuits or short the board.

Measure Surface Insulation Resistance (SIR) between two electrically isolated sites.

Electrical problems on aircraft could occur during flight, possibly intermittent.

Troubled circuit boards would be yanked, sometimes without findings.

Cleaning soldering contaminants is a leading contaminator of ozone layer.

Some fluxes are easier to clean than others.

The Experimental Process

18 boards are available for the experiment,
not necessarily a random sample from all boards.

Test flux brands X and Y: randomly assign 9 boards each to X & Y (FLUX)

The boards are soldered and cleaned. Order randomized. (SC.ORDER)

Then the boards are coated and cured to avoid handling contamination.
Order randomized. (CT.ORDER)

Then the boards are placed in a humidity chamber and measured for SIR.
Position in chamber randomized. (SLOT)

The randomization at the various process steps avoids unknown biases.
When in doubt, randomize! (Denis Janky)

Randomization of flux assignment gives us a mathematical basis
for judging flux differences with respect to the response SIR.

DOE Steps Recapitulated

1. Goal of the experiment. Answer question: Is Flux X different from Flux Y?
   If not we can use them interchangeably. One may be cheaper than the other.
   Test null hypothesis H_0: No difference in fluxes.
2. Understand the experimental units:
   Boards with all processing steps up to measuring response.
3. Define the appropriate response variable to be measured. SIR
4. Define potential sources of response variation
   a) factors of interest: flux type
   b) nuisance factors: boards, processing steps, testing.
5. Decide on treatment and blocking variables.
   Treatment = flux type, no blocking.
   With 2 humidity chambers we might have wanted to block on those.
6. Define clearly the experimental process and what is randomized.
   Treatments and all nuisance factors are randomized.

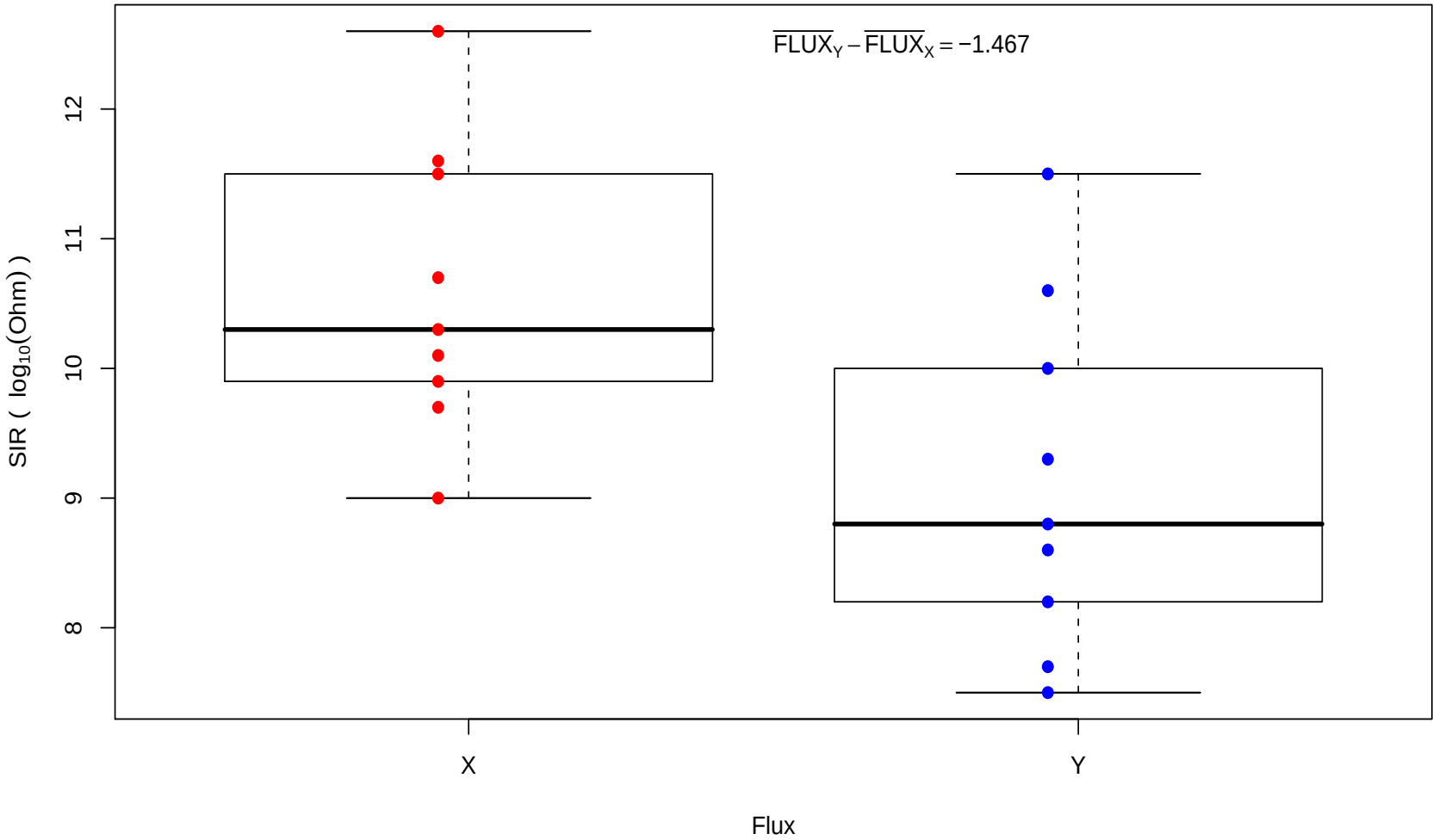
Flux Data

BOARD	FLUX	SC.ORDER	CT.ORDER	SLOT	SIR
1	Y	13	14	5	8.6
2	Y	16	8	6	7.5
3	X	18	9	15	11.5
4	Y	11	11	11	10.6
5	X	15	18	9	11.6
6	X	9	15	18	10.3
7	X	6	1	16	10.1
8	Y	17	12	17	8.2
9	Y	5	10	13	10.0
10	Y	10	13	14	9.3
11	Y	14	5	10	11.5
12	X	12	17	12	9.0
13	X	4	7	3	10.7
14	X	8	6	1	9.9
15	Y	3	2	4	7.7
16	X	7	3	2	9.7
17	Y	1	16	8	8.8
18	X	2	4	7	12.6

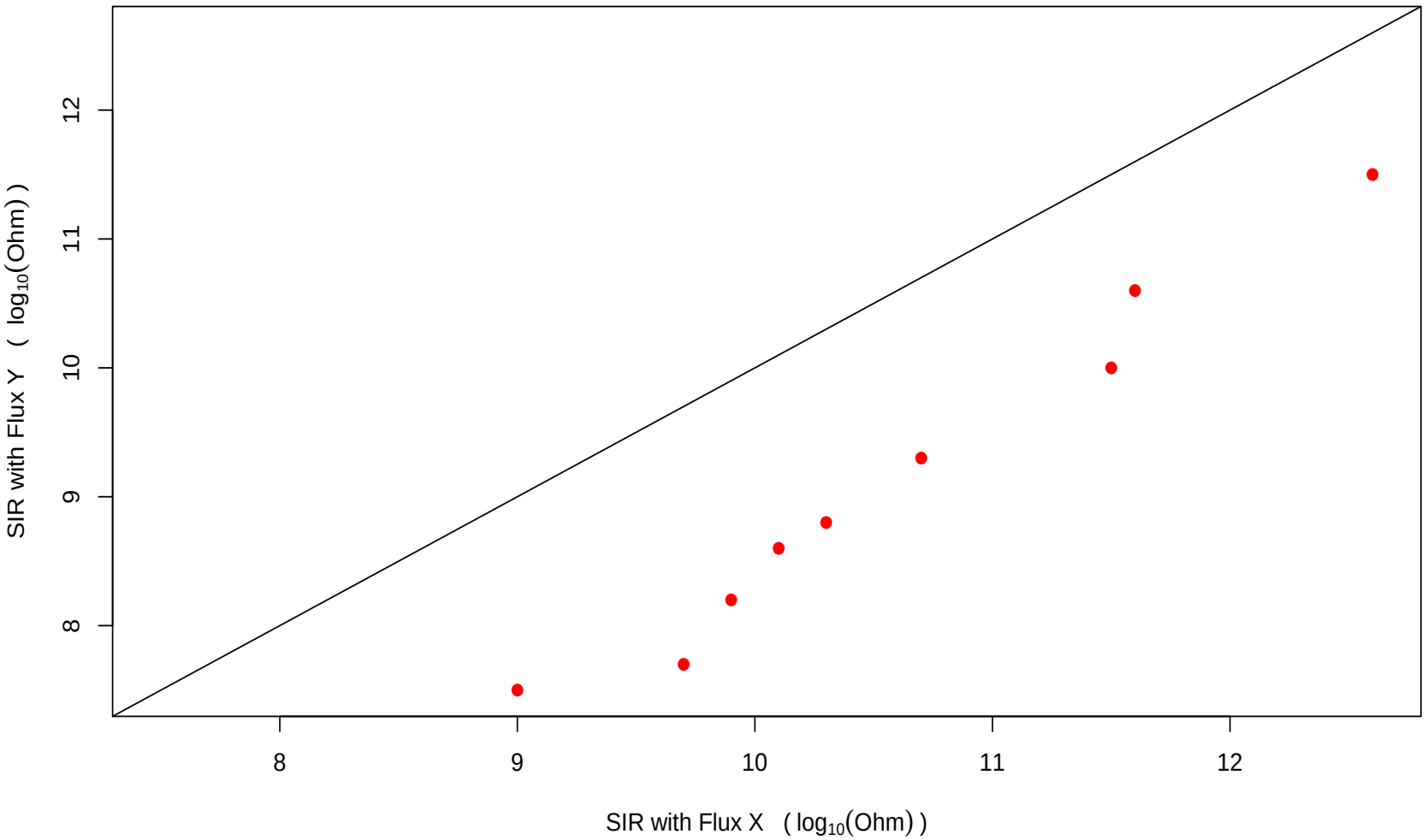
see `Flux.csv`

or `flux`

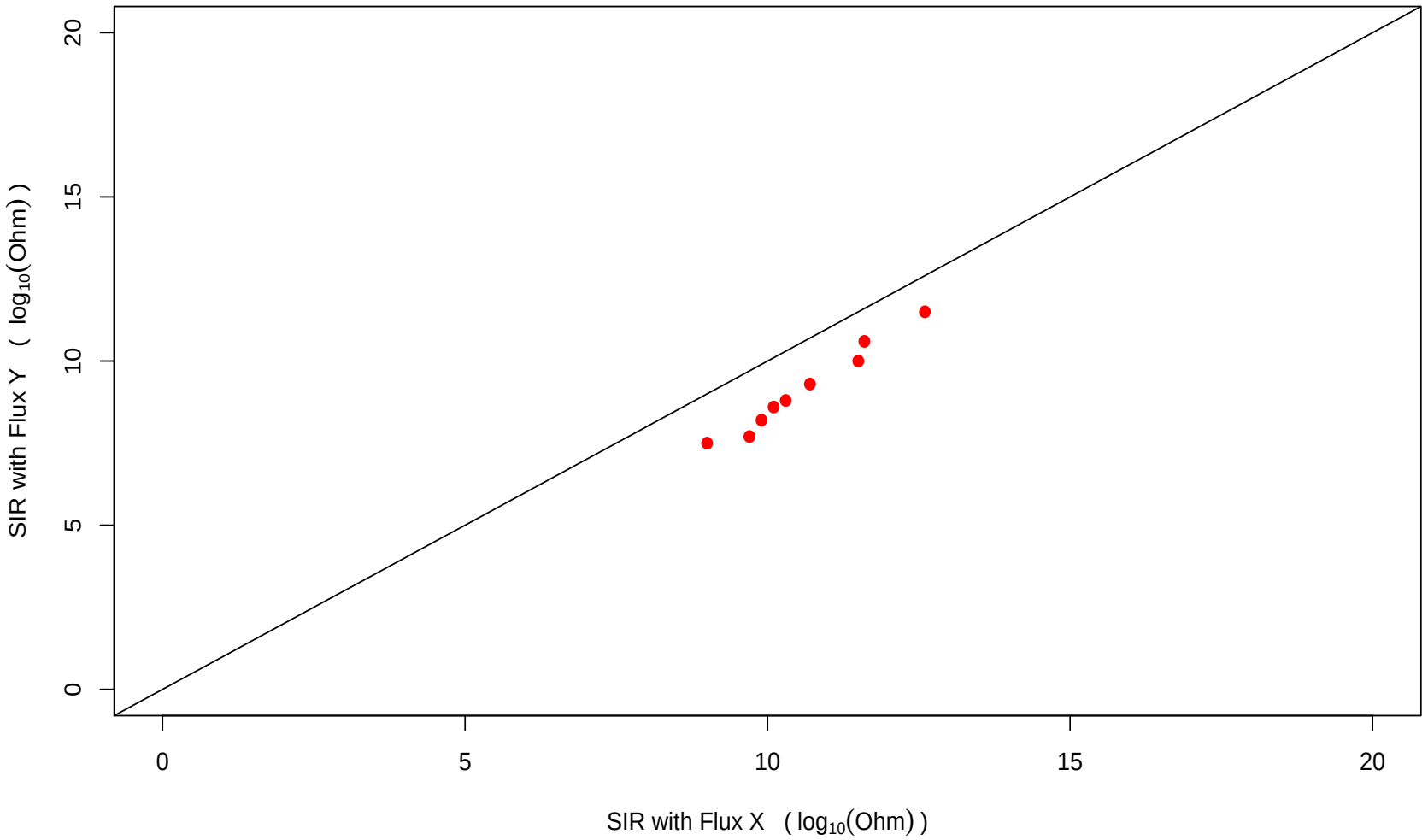
Flux Experiment: First Boxplot Look at SIR Data



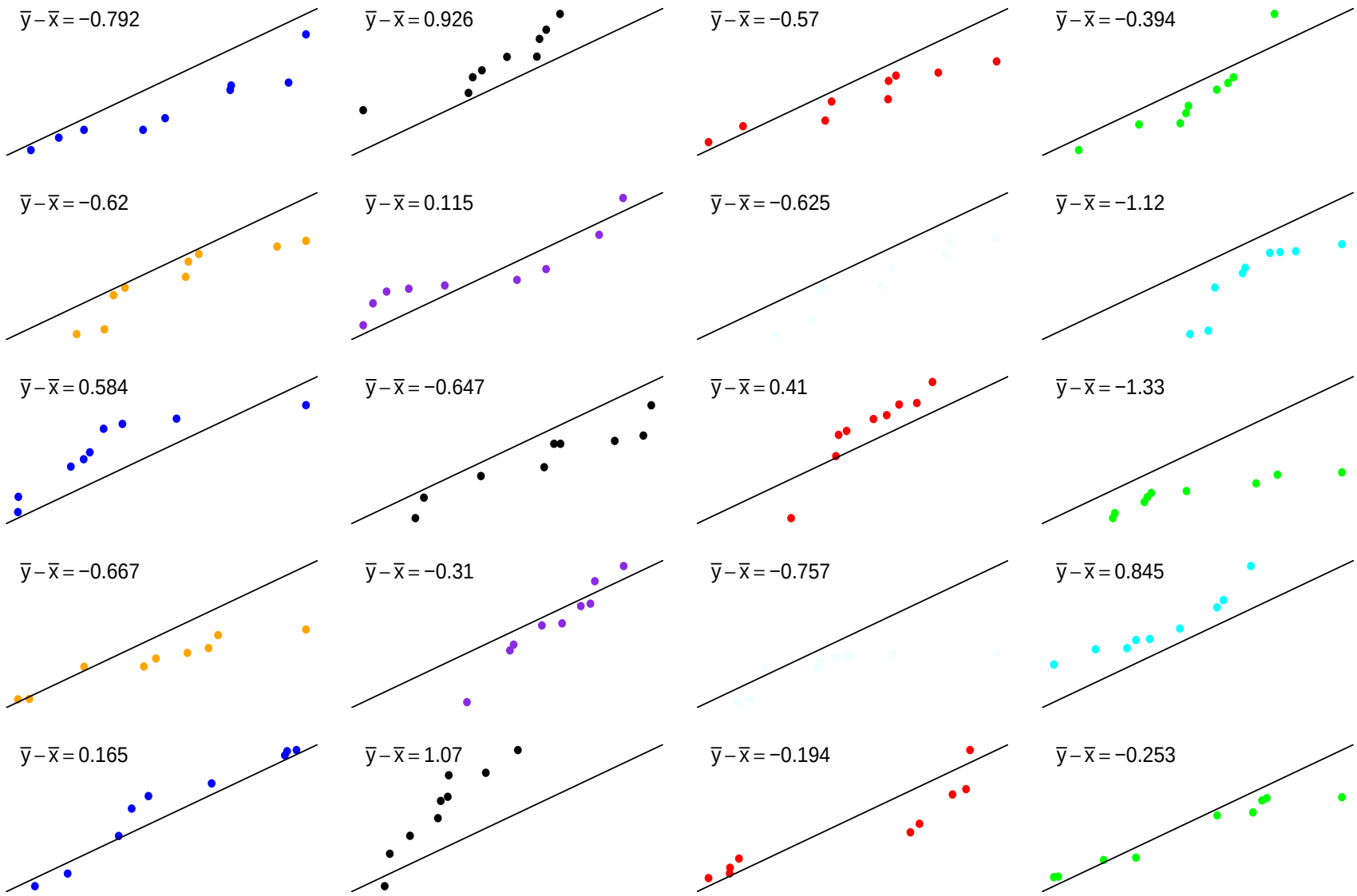
Flux Experiment: QQ-Plot of SIR Data



QQ-Plot of SIR Data (Higher Perspective?)



Some QQ-Plots from N(0,1) Samples (m=9, n=9)



Is the Difference $\bar{Y} - \bar{X} = -1.467$ Significant?

In comparing SIR for the two fluxes let us focus on the difference of means $\overline{\text{FLUX}}_Y - \overline{\text{FLUX}}_X = \bar{Y} - \bar{X}$.

If the use of flux X or flux Y made no difference then we should have seen the same results for these 18 boards, no matter which got flux X or Y.
X or Y is just an artificial "distinguishing" label with no consequence.

For other random assignments of fluxes, or random splittings of 18 boards into two groups of 9 & 9, we would have seen other differences of means.

There are $\binom{18}{9} = 48620$ such possible splits. For each split we could obtain $\bar{Y} - \bar{X}$.

Was our observed difference of -1.467 from an unusual random split?

Need the reference distribution of $\bar{Y} - \bar{X}$ for all 48620 splits.

Some Randomization Examples of $\bar{Y} - \bar{X}$

8.6	8.6	8.6	8.6	8.6	8.6	8.6
7.5	7.5	7.5	7.5	7.5	7.5	7.5
11.5	11.5	11.5	11.5	11.5	11.5	11.5
10.6	10.6	10.6	10.6	10.6	10.6	10.6
11.6	11.6	11.6	11.6	11.6	11.6	11.6
10.3	10.3	10.3	10.3	10.3	10.3	10.3
10.1	10.1	10.1	10.1	10.1	10.1	10.1
8.2	8.2	8.2	8.2	8.2	8.2	8.2
10	10	10	10	10	10	10
9.3	9.3	9.3	9.3	9.3	9.3	9.3
11.5	11.5	11.5	11.5	11.5	11.5	11.5
9	9	9	9	9	9	9
10.7	10.7	10.7	10.7	10.7	10.7	10.7
9.9	9.9	9.9	9.9	9.9	9.9	9.9
7.7	7.7	7.7	7.7	7.7	7.7	7.7
9.7	9.7	9.7	9.7	9.7	9.7	9.7
8.8	8.8	8.8	8.8	8.8	8.8	8.8
12.6	12.6	12.6	12.6	12.6	12.6	12.6
$\bar{Y} - \bar{X}$	$\bar{Y} - \bar{X}$	$\bar{Y} - \bar{X}$	$\bar{Y} - \bar{X}$	$\bar{Y} - \bar{X}$	$\bar{Y} - \bar{X}$	$\bar{Y} - \bar{X}$
1.1778	0.4222	-0.0889	-0.4000	0.5778	0.7778	0.2000

Randomization Distribution of $\bar{Y} - \bar{X}$

The randomization or reference distribution $\mathcal{D}(\bar{Y} - \bar{X})$ of $\bar{Y} - \bar{X}$ is in 1-1 correspondence with that of $\bar{Y}$ since

$$m(\bar{Y} - \bar{X}) = m\bar{Y} - \sum X_j = m\bar{Y} + \sum Y_i - \sum X_j - \sum Y_i = (m+n)\bar{Y} - \sum Z_k = N(\bar{Y} - \bar{Z})$$

with $N = m + n$, $\sum Z_k = \sum Y_i + \sum X_j$ and $\bar{Z} = \dfrac{\sum Y_i + \sum X_j}{n+m} = \dfrac{\sum Z_k}{N}$

Note that $\bar{Z}$ does not change for all $\binom{N}{m}$ combinations of treatment assignments over the $N = m + n$ positions.

$$\Longrightarrow \quad \mathcal{D}(\bar{Y} - \bar{X}) = \frac{N}{m}(\mathcal{D}(\bar{Y}) - \bar{Z})$$

Reference Distribution of $\bar{Y} - \bar{X}$

Compute $\bar{Y} - \bar{X}$ for each of the 48620 possible splits and determine how unusual the observed difference of -1.467 is.

This seems like a lot of computing work but it takes just a few seconds in R using the function `combn` of the package `combinat`.

Download and install that package first from the contributed packages in `CRAN` or from R packages under STAT 421 site and invoke `library(combinat)` prior to using `combn`.

```
randomization.ref.dist=combn(1:18,9,fun=mean.fun,y=SIR)
```

gives the vector of all 48620 such averages, where

```
mean.fun <- function(ind,y) { mean(y[ind]) }
```

Here `SIR` is the vector of all 18 SIR values,
passed as secondary argument to `mean.fun`.

Reference Distribution of $\bar{Y} - \bar{X}$ (continued)

The function `combn` goes through all combinations (referred to as `ind` in `mean.fun`) of 9 values taken from `1:18` and evaluates the mean of those SIR values.

This vector of averages only gives the reference distribution of $\bar{Y}$.

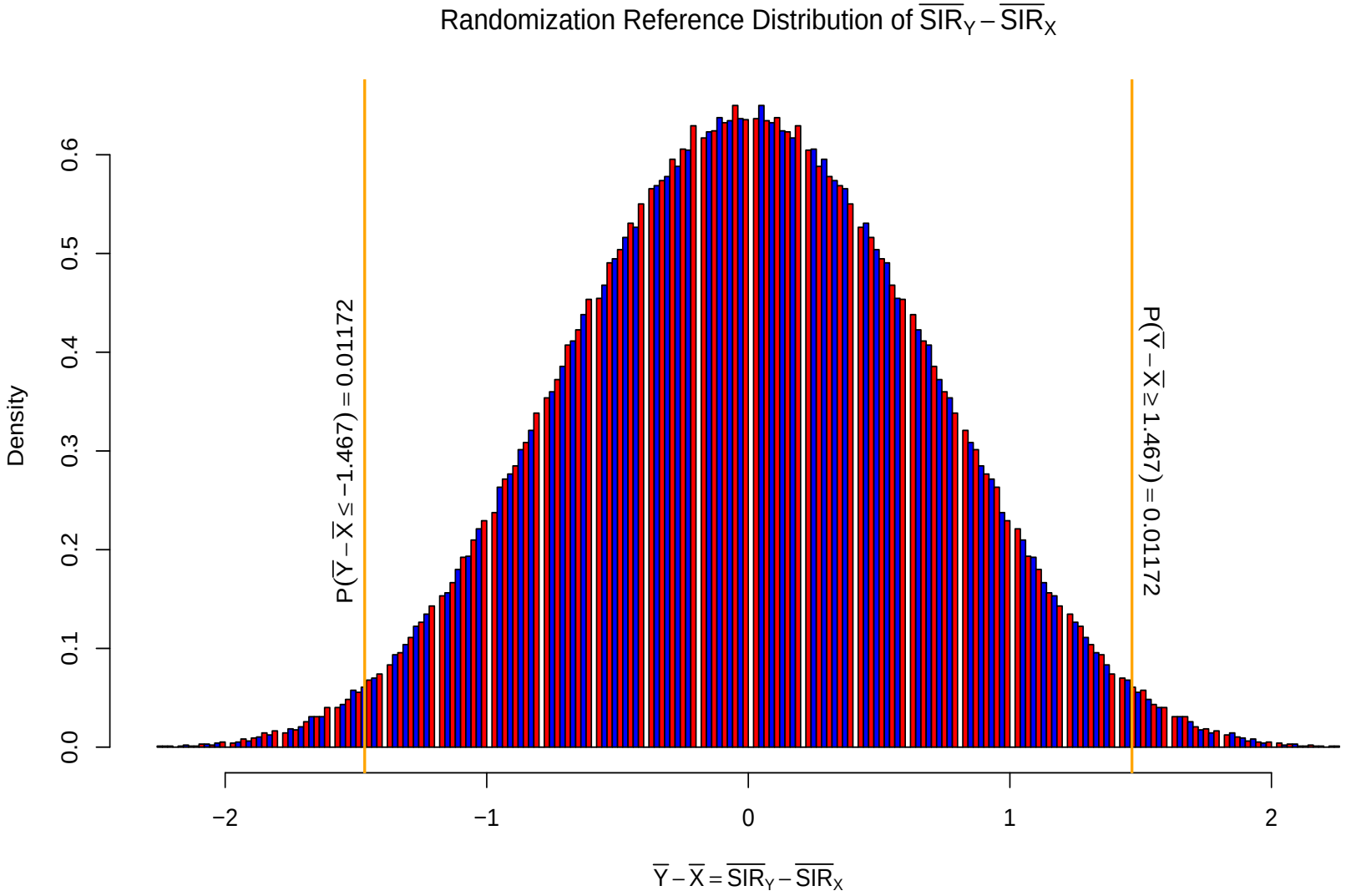
However, `(randomization.ref.dist-mean(SIR))*18/9`
gives the reference distribution of $\bar{Y} - \bar{X}$.

We find a (two-sided) p-value of .02344 for our observed $\bar{Y} - \bar{X} = -1.467$.
That is the probability of seeing an $|\bar{Y} - \bar{X}|$ value as or more extreme than $|\bar{y} - \bar{x}| = 1.467$ when in fact the hypothesis H_0 holds true,
i.e., under the randomization reference distribution.

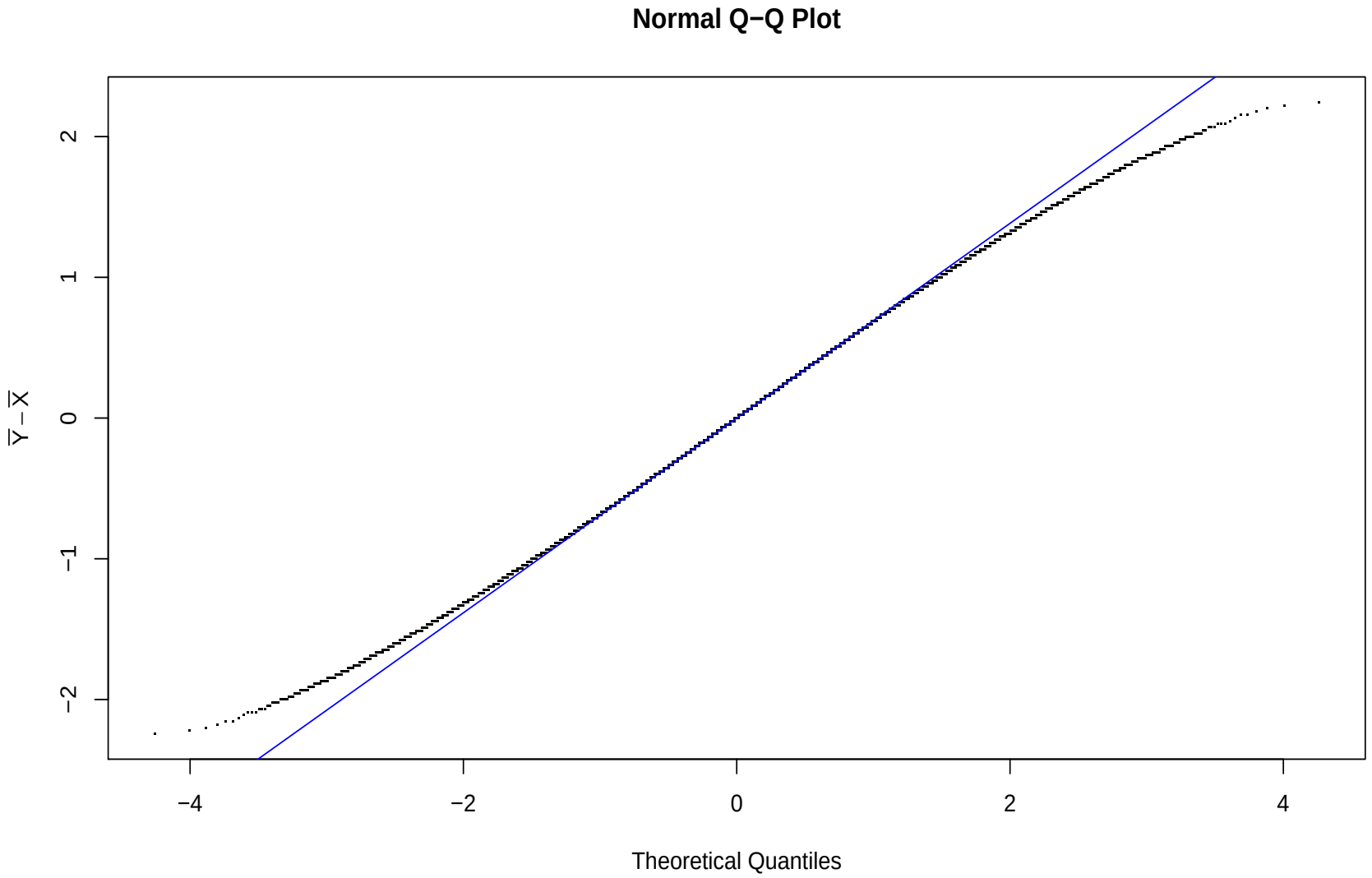
Randomization of fluxes enables probability statements!

Why is this reference distribution symmetric around zero?
Would this still hold if $m \neq n$ with $m + n = 18$? (Look at $m = 1$!).

Randomization Reference Distribution of $\bar{Y} - \bar{X}$



Normal QQ-Plot of $\bar{Y} - \bar{X}$ Randomization Reference Distribution



Approximation to Randomization Reference Distribution

For moderate to large m and n the number of combinations $\binom{m+n}{m}$ becomes so large that it taxes the computing power or storage capacity of the average computer.

I have not tested the limits of `combn`. You may want to try that.

A simple way out is to generate a sufficiently large sample, say $M = 10,000$, of combinations from this set of all $\binom{m+n}{m}$ combinations.

Compute the statistic of interest, $s(\underline{X}_i, \underline{Y}_i) = \bar{Y}_i - \bar{X}_i,\ i = 1, \ldots, M$ for each sampled combination and approximate the randomization reference distribution

$$F(z) = P(s(\underline{X}, \underline{Y}) \le z) \qquad \text{by} \qquad \widehat{F}_M(z) = \frac{1}{M} \sum_{i=1}^{M} B_i(z)$$

with $B_i(z) = 1$ when $s(\underline{X}_i, \underline{Y}_i) \le z$ and $B_i(z) = 0$ else.
By the law of large numbers (LLN) we have for any z

$$\widehat{F}_M(z) \longrightarrow F(z) \quad \text{as} \quad M \to \infty$$

Sample Simulation Program

This can be done in a loop using the `sample` function in R.

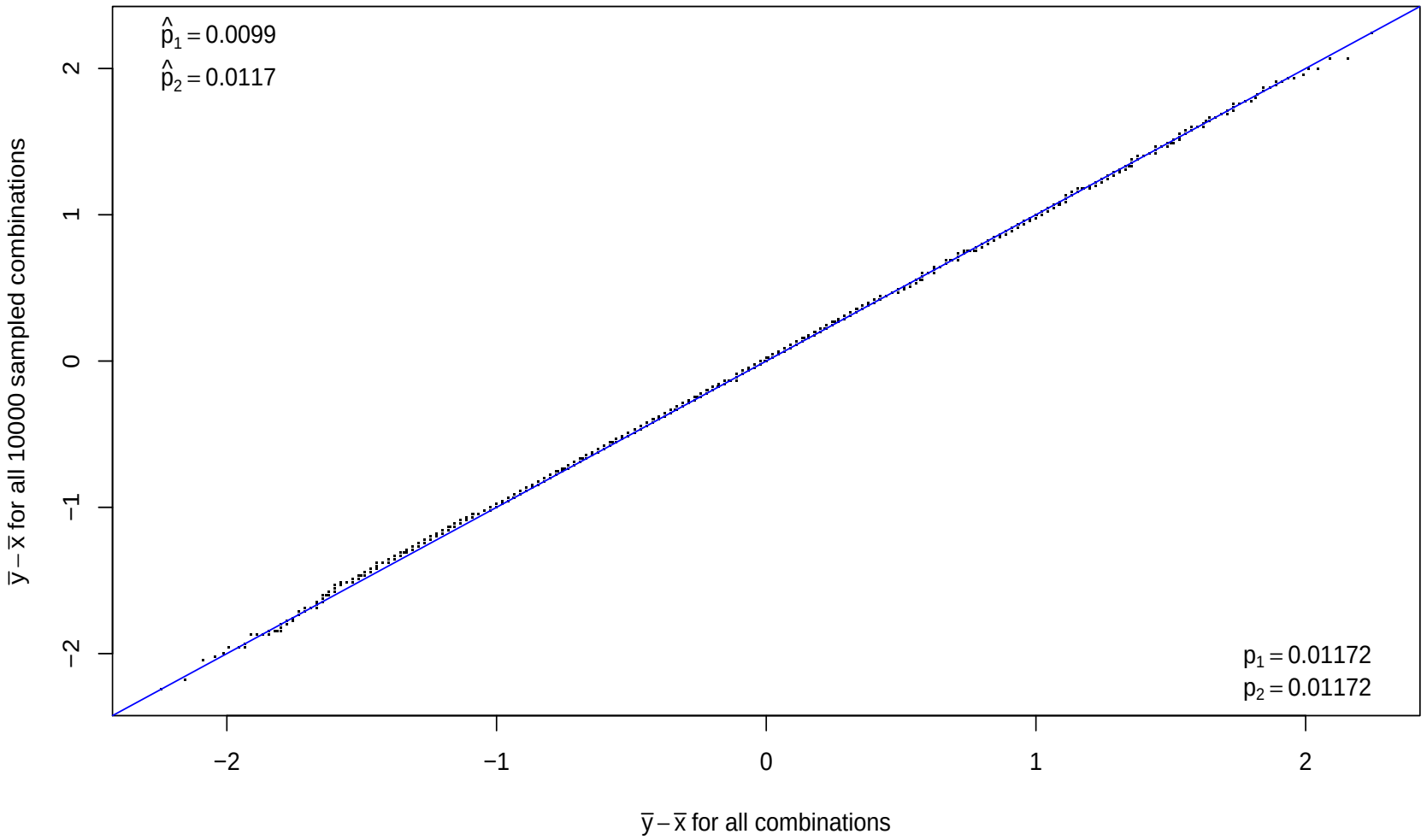
```
simulated.reference.distribution=function(M=10000){
  sim.ref.dist=NULL
  for(i in 1:M){
    SIR.star=sample(SIR)
    sim.ref.dist=c(sim.ref.dist,mean(SIR.star[1:9]))
  }
D.star=2*(sim.ref.dist-mean(SIR))
D.star}
```

Here we simulated the reference distribution of $\bar{Y}$ and then turned it into the simulated reference distribution sample of $\bar{Y} - \bar{X}$ via `D.star`.

The following slide shows the QQ-plot comparison with the full randomization reference distribution, together with the respective p-values.

This approach should suffice for practical purposes.

QQ-Plot of $\bar{Y} - \bar{X}$ for Simulated & Full Randomization Reference Distribution



Randomization Distribution of 2-Sample t-Test

$$t(\underline{X},\underline{Y}) = \frac{(\bar{Y}-\bar{X})/\sqrt{1/n+1/m}}{\sqrt{[\sum_{i=1}^{n}(Y_i-\bar{Y})^2+\sum_{j=1}^{m}(X_j-\bar{X})^2]/(m+n-2)}}$$

$\mathcal{D}(t(X,Y))$ is in 1-1 correspondence with that of $\bar{Y}-\bar{X}$ and thus with that of $\bar{Y}$.

$$\sum(Z_k-\bar{Z})^2-\frac{mn}{N}(\bar{Y}-\bar{X})^2=\sum Y_i^2+\sum X_j^2-\frac{1}{N}(m\bar{X}+n\bar{Y})^2-\frac{mn}{N}(\bar{Y}-\bar{X})^2$$

$$=\sum Y_i^2+\sum X_j^2-n\bar{Y}^2-m\bar{X}^2=\sum(Y_i-\bar{Y})^2+\sum(X_i-\bar{X})^2$$

Note that $\dfrac{t(\underline{X},\underline{Y})\sqrt{1/n+1/m}}{\sqrt{m+n-2}}=\dfrac{\bar{Y}-\bar{X}}{\sqrt{\sum(Y_i-\bar{Y})^2+\sum(X_i-\bar{X})^2}}=\dfrac{W}{\sqrt{1-\frac{mn}{N}W^2}}$ $\nearrow$ in W

with $W=(\bar{Y}-\bar{X})/\sqrt{\sum(Z_k-\bar{Z})^2}$

$$\Longrightarrow \mathcal{D}(t(\underline{X},\underline{Y}))=\frac{\sqrt{m+n-2}}{\sqrt{1/n+1/m}}\,\mathcal{D}\left(W\Big/\sqrt{1-\frac{mn}{N}W^2}\right)$$

t-Approximation for $t(\underline{X}, \underline{Y})$ Randomization Reference Distribution

The randomization reference distribution of $t(\underline{X}, \underline{Y})$ is very well approximated by a t-distribution with $16 = 18 - 1 - 1$ degrees of freedom.

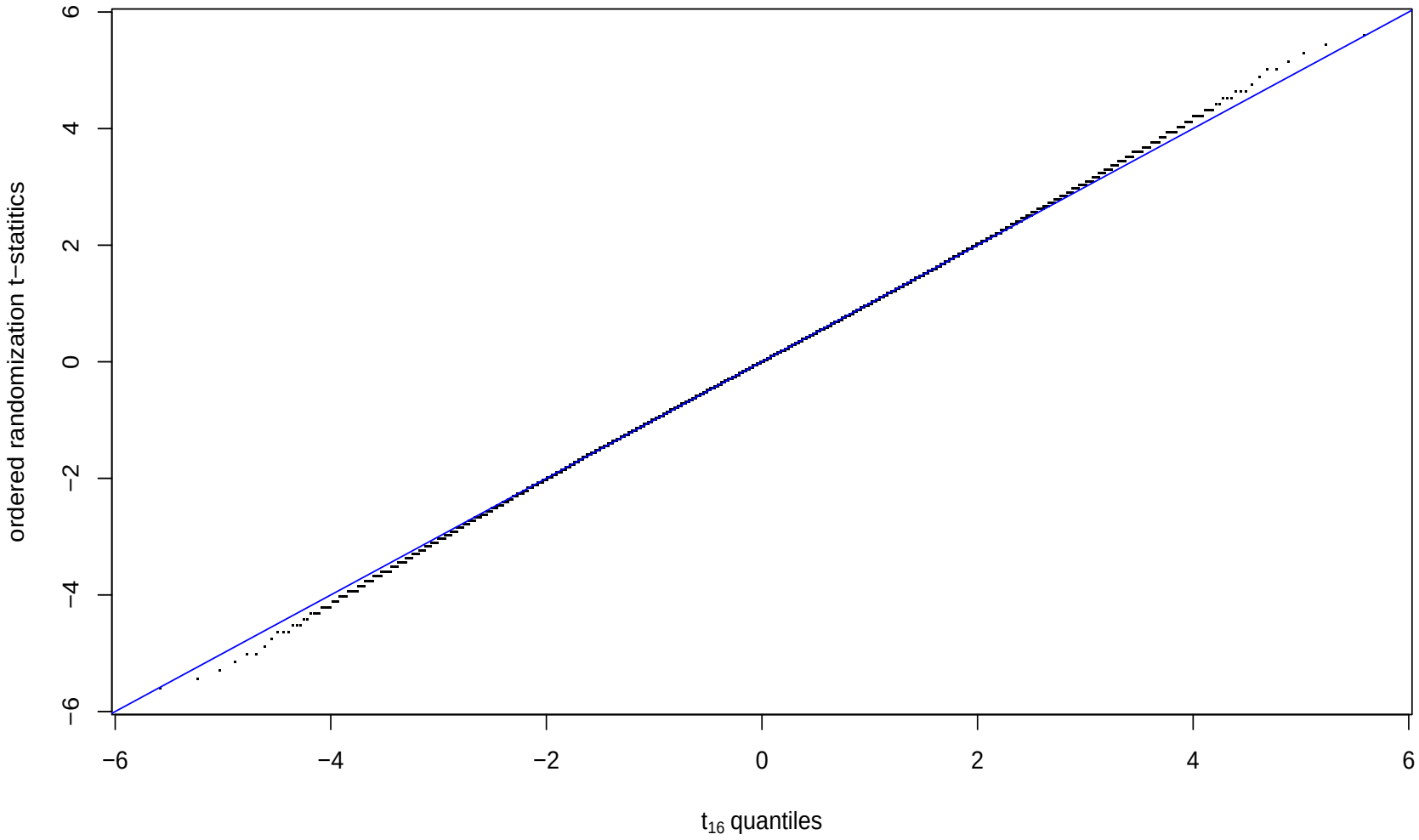
This was very useful before computing and simulation were readily available.
All one needs to do is compute $t(\underline{X}, \underline{Y})$ for the observed data and calculate its p-value from the t_{16} distribution. Here $t(\underline{x}, \underline{y}) = -2.513$.

Reference: G.E.P Box and S.L. Anderson, "Permutation theory in the derivation of robust criteria and the study of departures from assumptions,"
J. Roy. Stat. Soc., Ser. B, Vol 17 (1955), pp. 1-34.

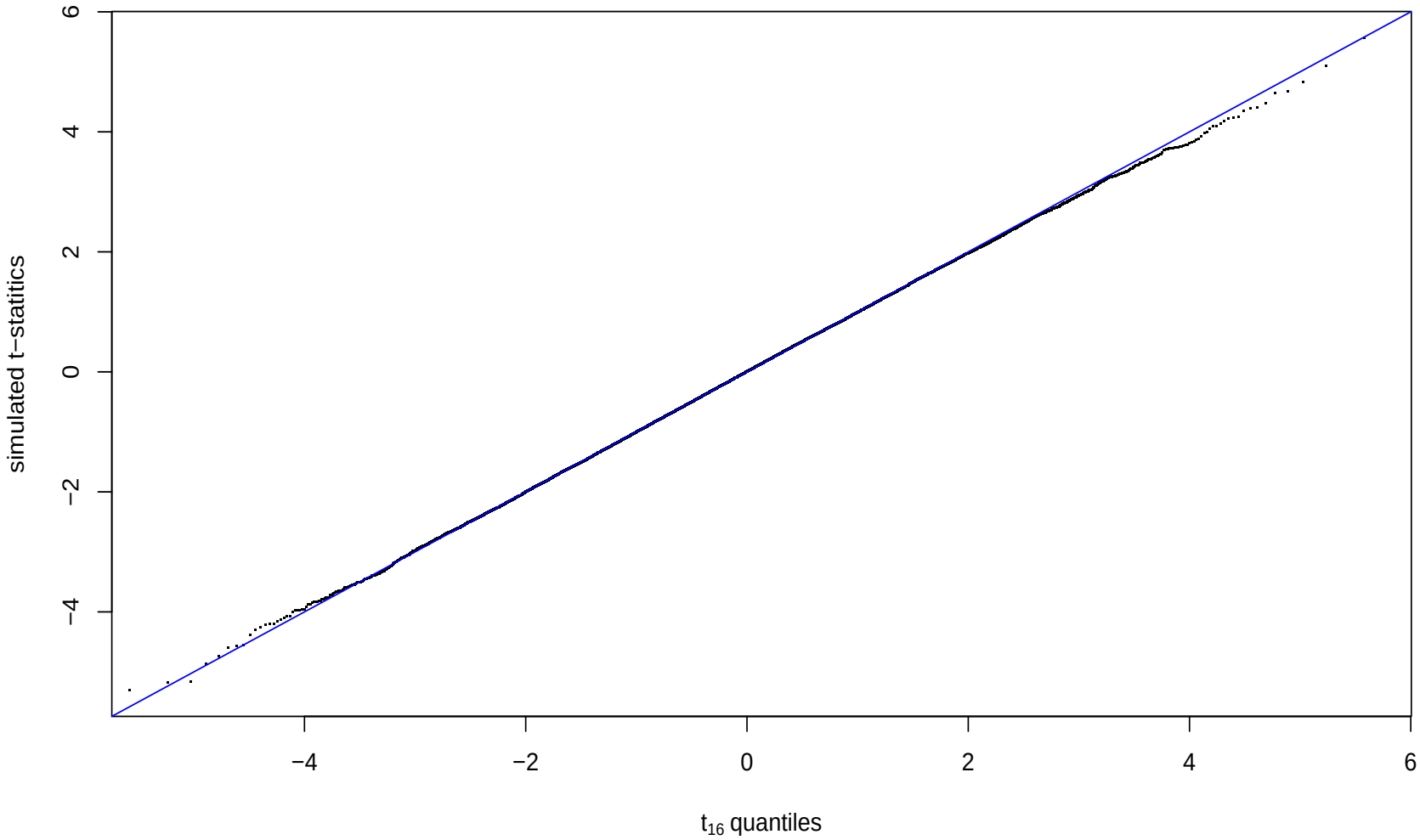
The test based on $t(\underline{X}, \underline{Y})$ and its t-distribution under H_0 also shows up in the normal-model-based approach to this problem.

More discussion of this later.

t-QQ-Plot of $t(X, Y)$ Randomization Reference Distribution



Comparison for t_{16}-Sample of Same Size



The Randomization Test

We have obtained our full or simulated randomization reference distribution.

Thus any extreme value of $|\bar{Y} - \bar{X}|$ could either come about due to a rare chance event or due to H_0 actually being wrong.

We have to make a decision: Reject H_0 or not?

We may decide to reject H_0 when $|\bar{Y} - \bar{X}| \geq C$, where C is some critical value.

To determine C one usually sets a significance level α which limits the probability of rejecting H_0 when in fact H_0 is true (Type I error). The requirement

$$\alpha = P(\text{reject } H_0 \mid H_0) = P(|\bar{Y} - \bar{X}| \geq C \mid H_0) \qquad \text{then determines } C = C_\alpha \,.$$

Significance Levels and P-Values

When we reject H_0 we would say that the results were significant at the (previously chosen) level α.

Commonly used values of α are $\alpha = .05$ or $\alpha = .01$.

Rejecting at smaller α than these would be even stronger evidence against H_0.

For how small an α would we still have rejected?

This leads us to the observed significance level or p-value of the test for the given data, i.e., for the observed discrepancy value $|\bar{y} - \bar{x}|$

$$\text{p-value} = P(|\bar{Y} - \bar{X}| \geq |\bar{y} - \bar{x}| \mid H_0)$$

How to Determine the P-Value

We have stated p-values obtained from the full and the simulated (M=10000) reference distributions. How are they obtained?

Note the following:

```
> x=1:10
> x>3
 [1] FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> sum(x>3)
[1] 7
> mean(x>3)
[1] 0.7
```

Note that `x>3` produced a logic vector with same length as x.

The logic values `FALSE` and `TRUE` are also interpreted numerically as `0` and `1`, respectively, in arithmetic expressions.

How to Determine the P-Value (continued)

We view the reference distribution as a vector x of numbers for all the differences of means, $\bar{Y} - \bar{X}$, obtained either for all 48620 possible splits or for the $M = 10000$ simulated splits.

`mean(x<=-1.467)` and `mean(x>=1.467)` would give us the respective p-values $p_1 = .01172$ and $p_2 = .01172$ for the full reference distribution,
and $\hat{p}_1 = .0099$ and $\hat{p}_2 = .0117$ for the simulated reference distribution.

The simulated distribution is obviously not quite symmetric.

Rather than adding these 2 p-values to get a 2-sided p-value we can also do this directly via `mean(abs(x)>= 1.467)=.02344`.
Here `abs(x)` gives the vector of absolute values of all components in `x`.

How to Determine the Critical Value C.crit for the Level α Test

For $\alpha = .05$ we want to find `C.crit` such that `mean(abs(x)>=C.crit)=.05`.

Equivalently, find the .95-quantile of `abs(x)` via `C.crit=quantile(abs(x),.95)`.

From the full reference distribution we get $\texttt{C.crit}(\alpha = .05) = 1.288889$
and $\texttt{C.crit}(\alpha = .01) = 1.644444$.

From the simulated reference distribution we get $\texttt{C.crit}(\alpha = .05) = 1.311111$
and $\texttt{C.crit}(\alpha = .01) = 1.666667$.

Note that we avoided the use of `C` in place of `C.crit` because in R the letter `C` has a predetermined meaning.

Letters to avoid as object names are: `c, q, t, C, D, F, I, T`.
Try them out with `?` in front to see what they stand for, e.g., `?T`.

What Does the t-Distribution Give Us?

What does the observed t-statistic $t(\underline{x},\underline{y}) = -2.513$ give as 2-sided p-value?

We find $P(|t(\underline{X},\underline{Y})| \geq 2.513) = 2 * (1 - \texttt{pt}(2.513, 16)) = .02306$, pretty close to the $.02344$ from the full randomization reference distribution.

What are the critical values $t_{\text{crit}}(\alpha)$ for $|t(\underline{X},\underline{Y})|$ for level $\alpha = .05, .01$ tests?

We find $t_{\text{crit}}(\alpha = .05) = \texttt{qt}(.975, 16) = 2.1199$ and
$t_{\text{crit}}(\alpha = .01) = \texttt{qt}(.995, 16) = 2.9208$, respectively.

With $|t(\underline{x},\underline{y})| = 2.513$ we would reject H_0 at $\alpha = .05$ since $|t(\underline{x},\underline{y})| \geq 2.1199$ but not at $\alpha = .01$ since $|t(\underline{x},\underline{y})| < 2.9208$.