University of Washington

STATISTICS

Applied Statistics and Experimental Design

Two Sample Experiments: Treatment vs Control

Fritz Scholz

Fall Quarter 2008

Circuit Board Solder Flux Experiment (Courtesy Denis Janky)

Flux is material used to facilitate soldering on circuit boards.
`http://en.wikipedia.org/wiki/Flux_(metallurgy)`

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 circuit 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 the ozone layer.

Some fluxes are easier to clean than others. We have 2 flux candidates.

Treatments and Controls

Here the experimental situation was laid out in terms of two different and competing flux treatments.

It could be that one of the treatments is what has been used in the past.

It is the familiar treatment and often it would then be called the control.

The other treatment would be the newer treatment. Why should we change?

Some possible reasons:

1. the promise of better results

2. equal results with a cheaper or cleaner process

3. treatment flexibility, spreading the risk of treatment availability

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 on the effect of any process step, randomize!

The randomization of flux assignments gives us a mathematical basis for
judging flux differences with respect to the response SIR (to become clear later).

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.

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, various 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,
   since variation from chamber to chamber may be substantial.

6. Define clearly the experimental process and what is randomized.

   Treatments and all nuisance factors are randomized.

Why Are Treatments and Nuisance Factors Randomized?

The treatments are randomized among boards to avoid confounding and to provide a mathematical basis for inference concerning treatment effects.

We don't want any board peculiarities all ganging up on one treatment.

This would make it difficult to distinguish treatment from board effect.

The nuisance factors all come into play after the flux treatments are applied.
If not, such nuisance factor effects could be lumped with the board peculiarities.

Any effect that they have could align with the treatments, i.e., the 9 highest nuisance factor effects could be aligned with flux X. Randomization makes that unlikely.

Randomizing over all nuisance factors separately and independently makes any combined confounding effects extremely unlikely. It reduces variability since $+$ and $-$ effects cancel out to a large extent.

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

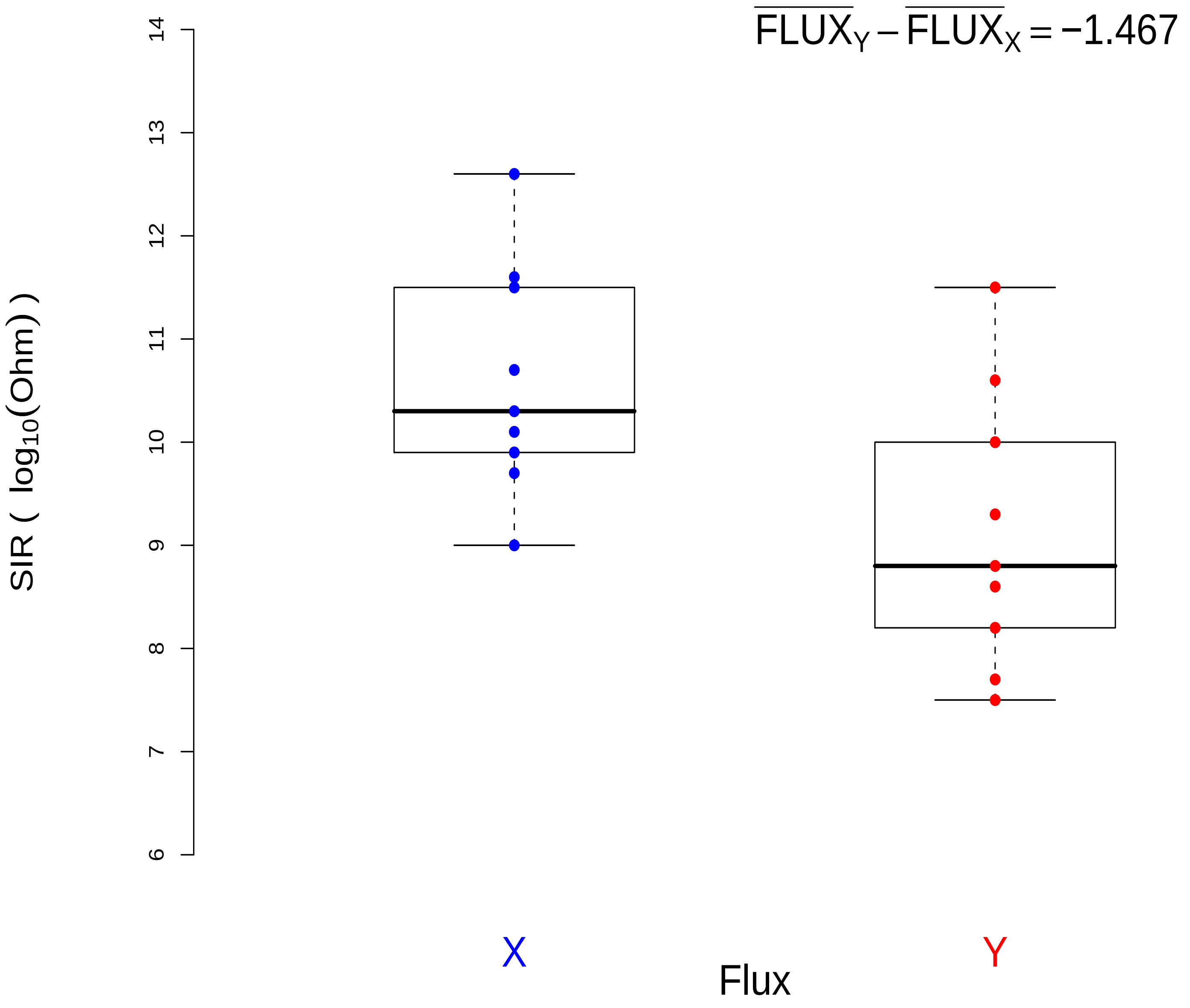
see `Flux.csv`

or `flux`

Note the 4 levels

of randomization

Flux Experiment: First Box Plot Look at SIR Data



What Does a Box Plot Show?

In the box plot (Tukey 1977) the upper and lower quartiles, $Q(.75)$ and $Q(.25)$, of the sample are portrayed by the top and bottom edges of a rectangle, and the median, $Q(.5)$, is portrayed by the horizontal line segment within the rectangle.

Dashed lines extend from the ends of the box to the adjacent values, defined below.

Compute the interquartile range $IQR = Q(.75) - Q(.25)$.

The upper adjacent value is the largest observation $\leq Q(.75) + 1.5 \times IQR$.

The lower adjacent value is the smallest observation $\geq Q(.25) - 1.5 \times IQR$.

Any observations beyond the adjacent values are shown individually.

For a normal population $\approx .35\%$ of the values fall beyond each adjacent value.

The dots shown superimposed on our SIR box plots are just the added data values.

Sample quantiles, $Q(p)$, determined by one of many possible schemes, see `?quantile` in R.

The Utility of Box Plots

Succinct description of data sets, showing median, upper and lower quartiles.

This avoids looking at too many data points that may be hard to distinguish.

Median $\approx$ center of the box $\Longrightarrow$ the middle 50% of the data appear symmetric.

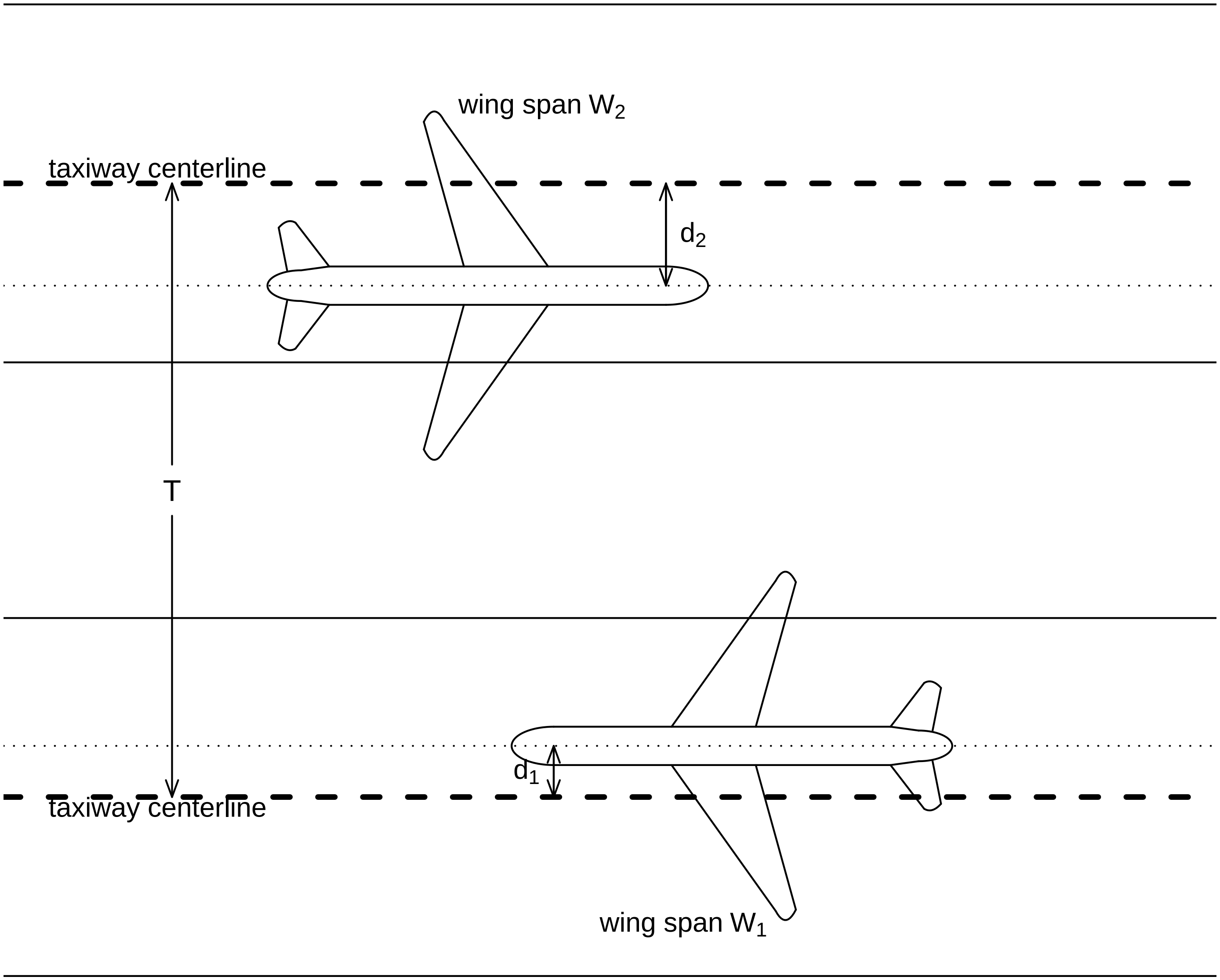
If in addition adjacent values $\approx$ symmetric relative to the box

$\Longrightarrow$ this symmetry appears to extend to the tails, otherwise data are skewed.

Any values beyond the adjacent values are singled out as potential outliers.

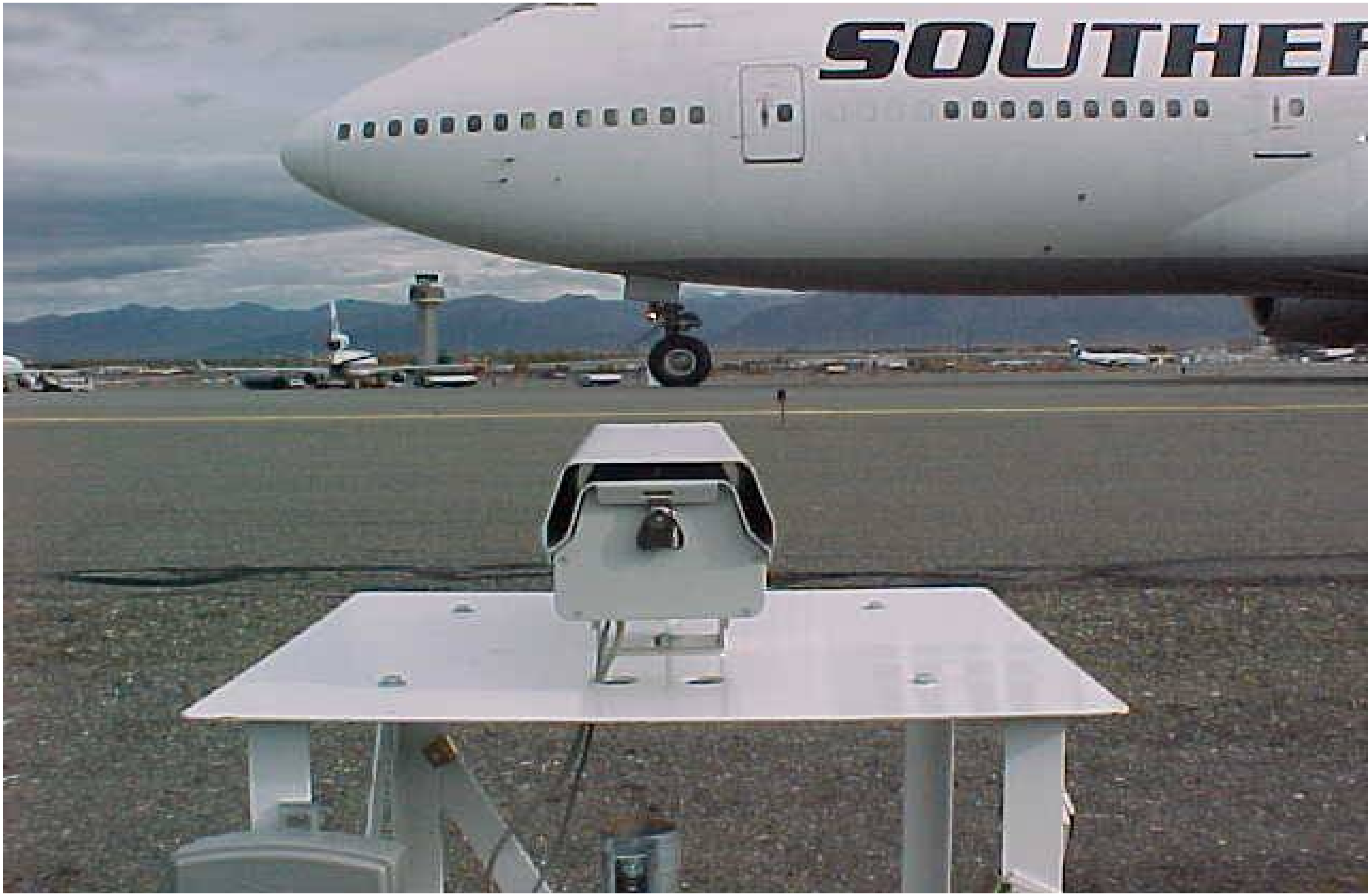
Box plots are especially effective for comparing (large) data sets.

Detour on Taxiway Deviations

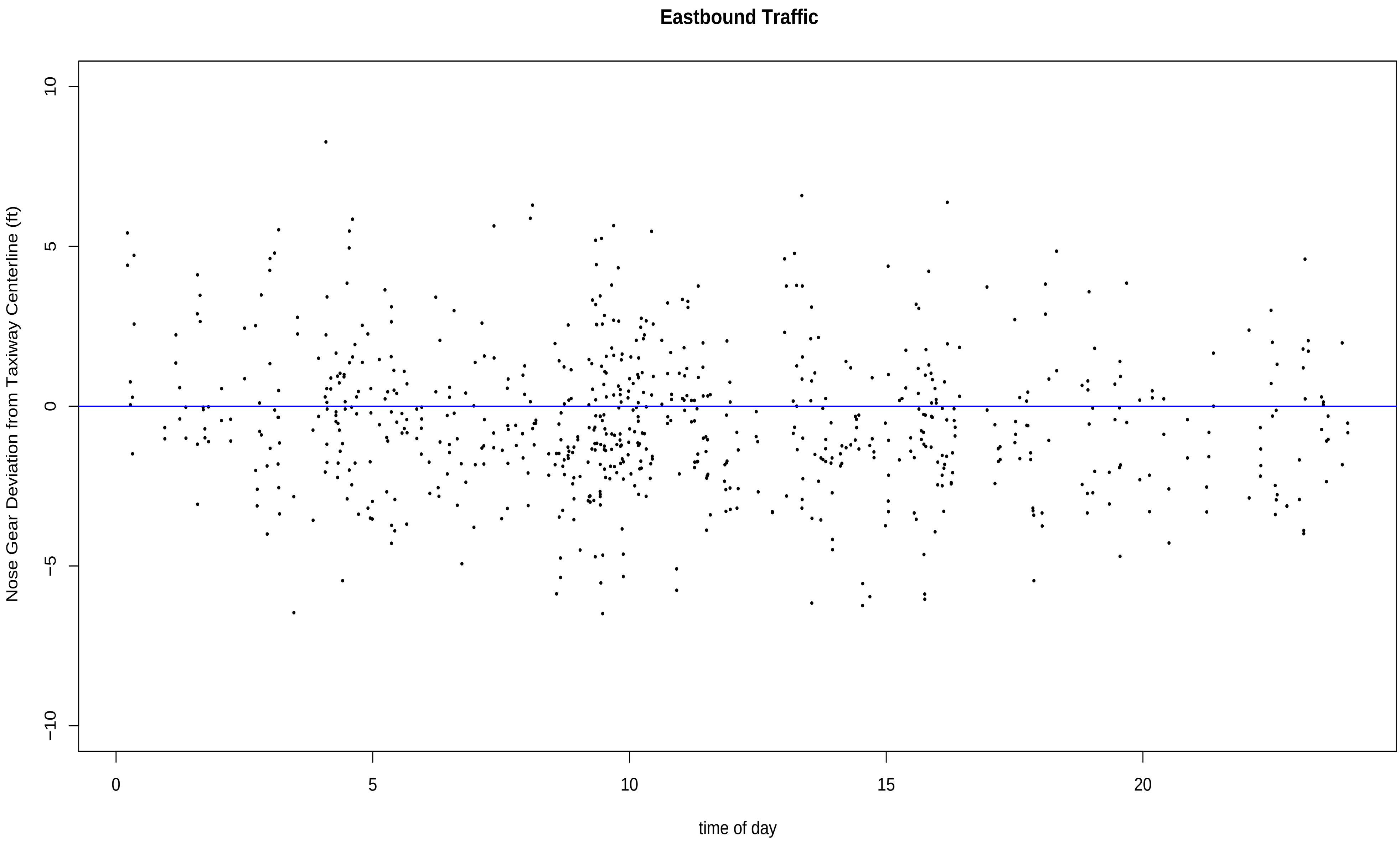


Risk of collision and running off taxiway, getting stuck in the mud.

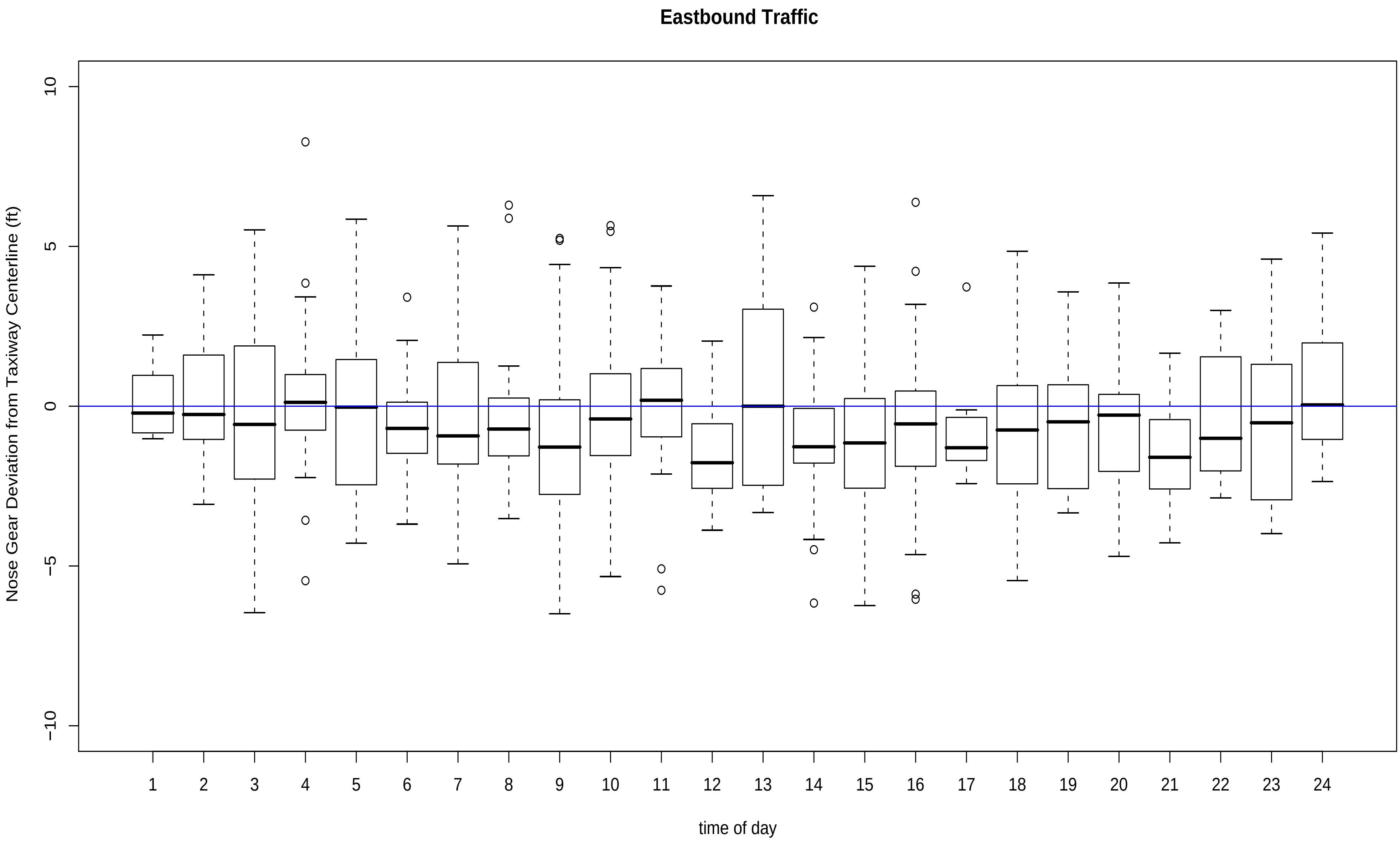
Two Lasers Measure Distances to Nose & Main Gears



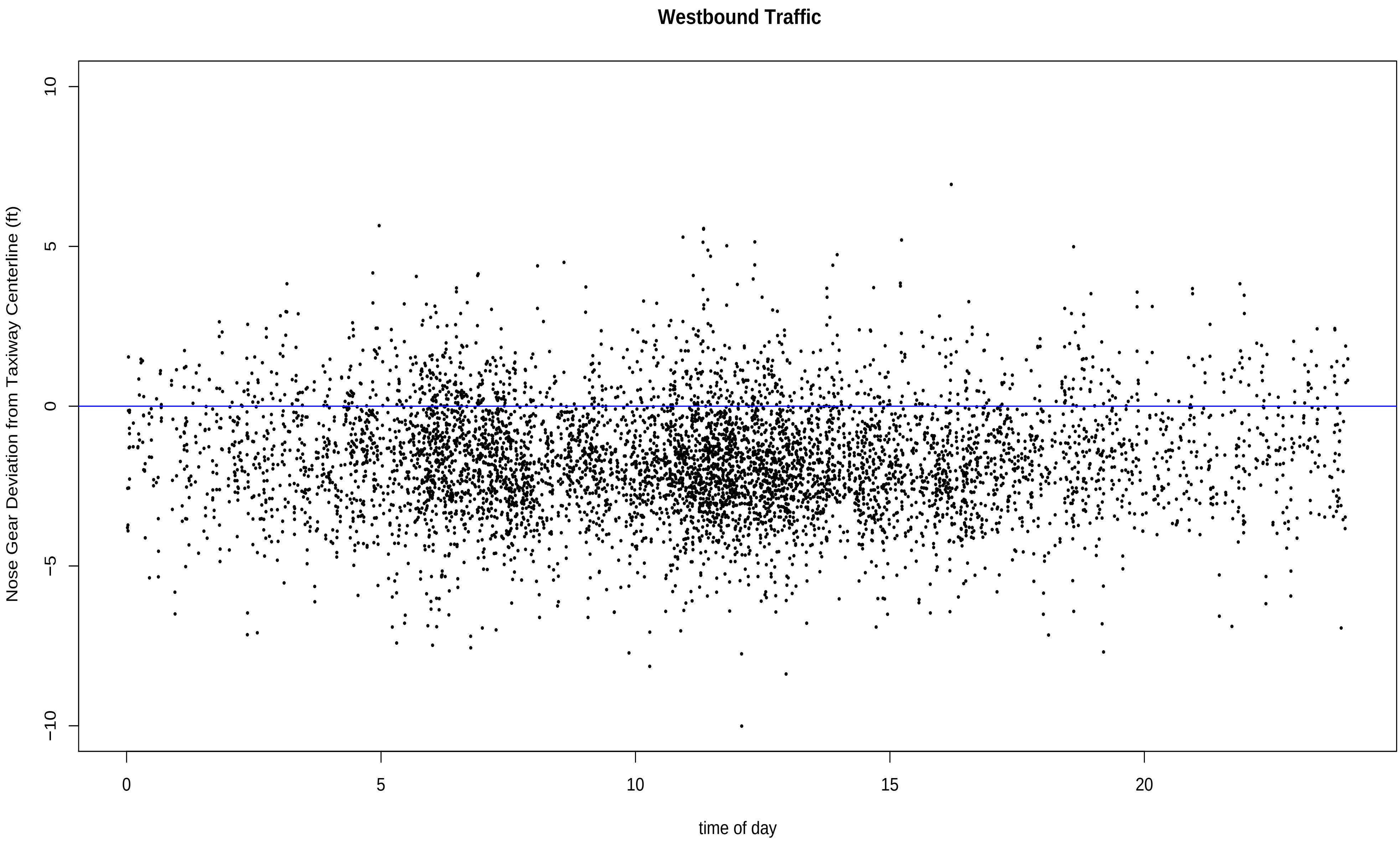
747 Taxiway Deviations



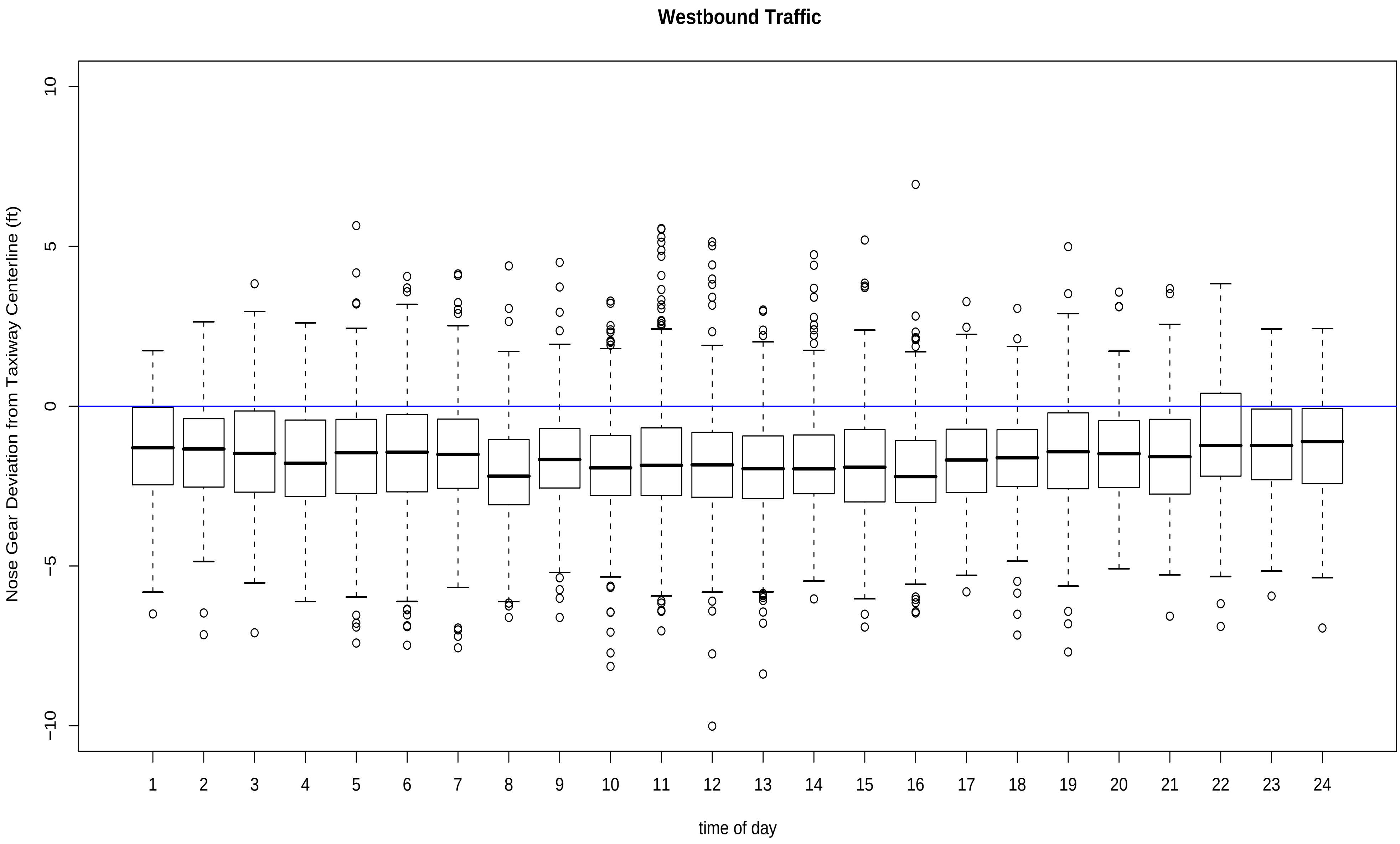
Box Plots for 747 Taxiway Deviations



747 Taxiway Deviations



Box Plots for 747 Taxiway Deviations



Some Comments

Notice that the first set of box plots (Eastbound Traffic) hinted at a preponderance of medians on one side of the taxiway centerline.

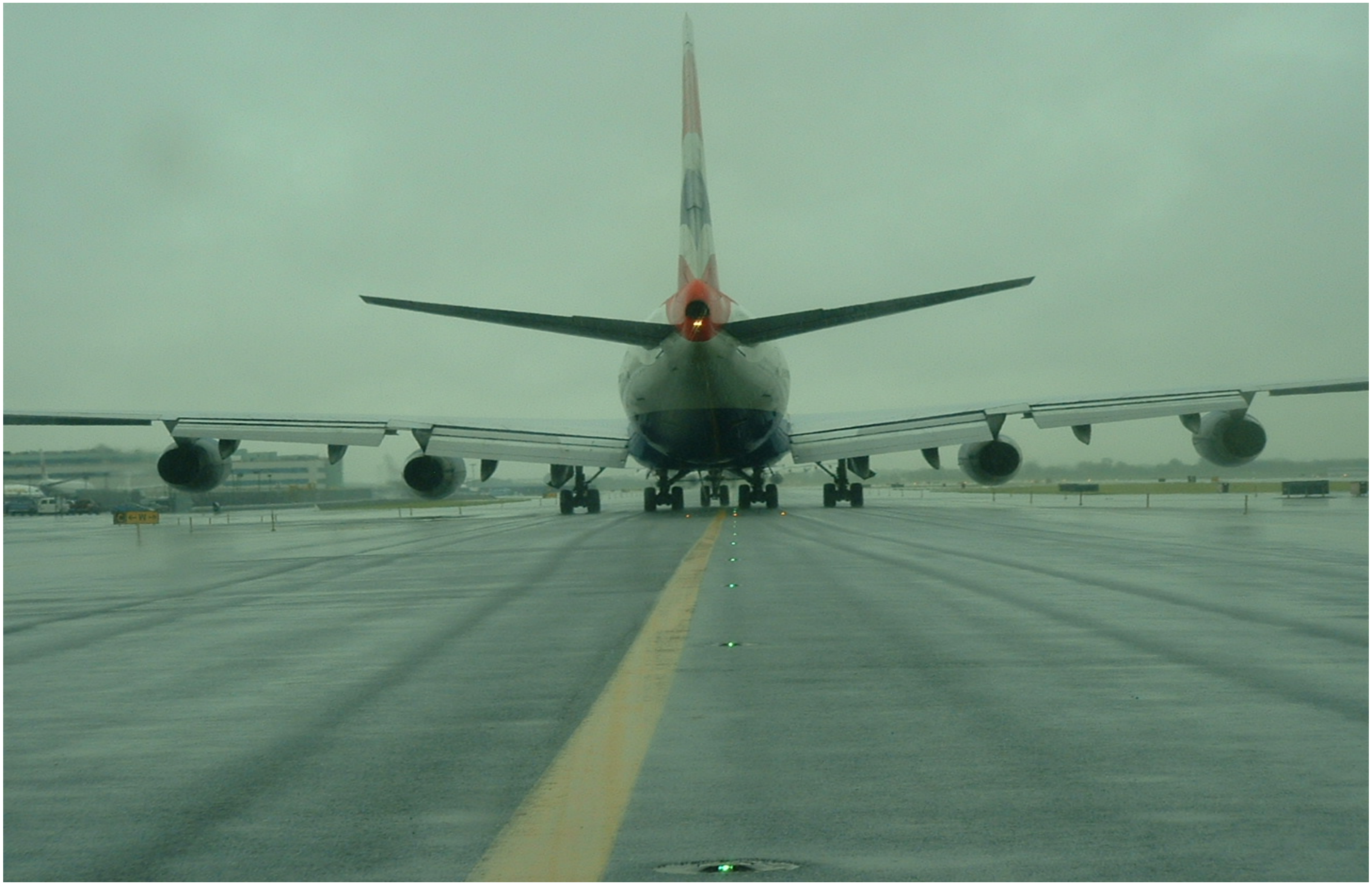
The second set (Westbound Traffic) made the same point very clearly because of the much larger traffic volume in that direction.

The consistent direction of the bias suggests that it is not due to pilot position relative to aircraft centerline (parallax effect).

Bias size in the Eastbound Traffic seems less than that in the Westbound Traffic.

This suggests two bias components, one switching direction with traffic direction (parallax effect), the other does not.

Taxiway Centerline and Centerlights



Pilots avoid hitting the bumps ⟹ consistent bias (in same direction)

in addition to pilot parallax bias which changes with direction of travel.

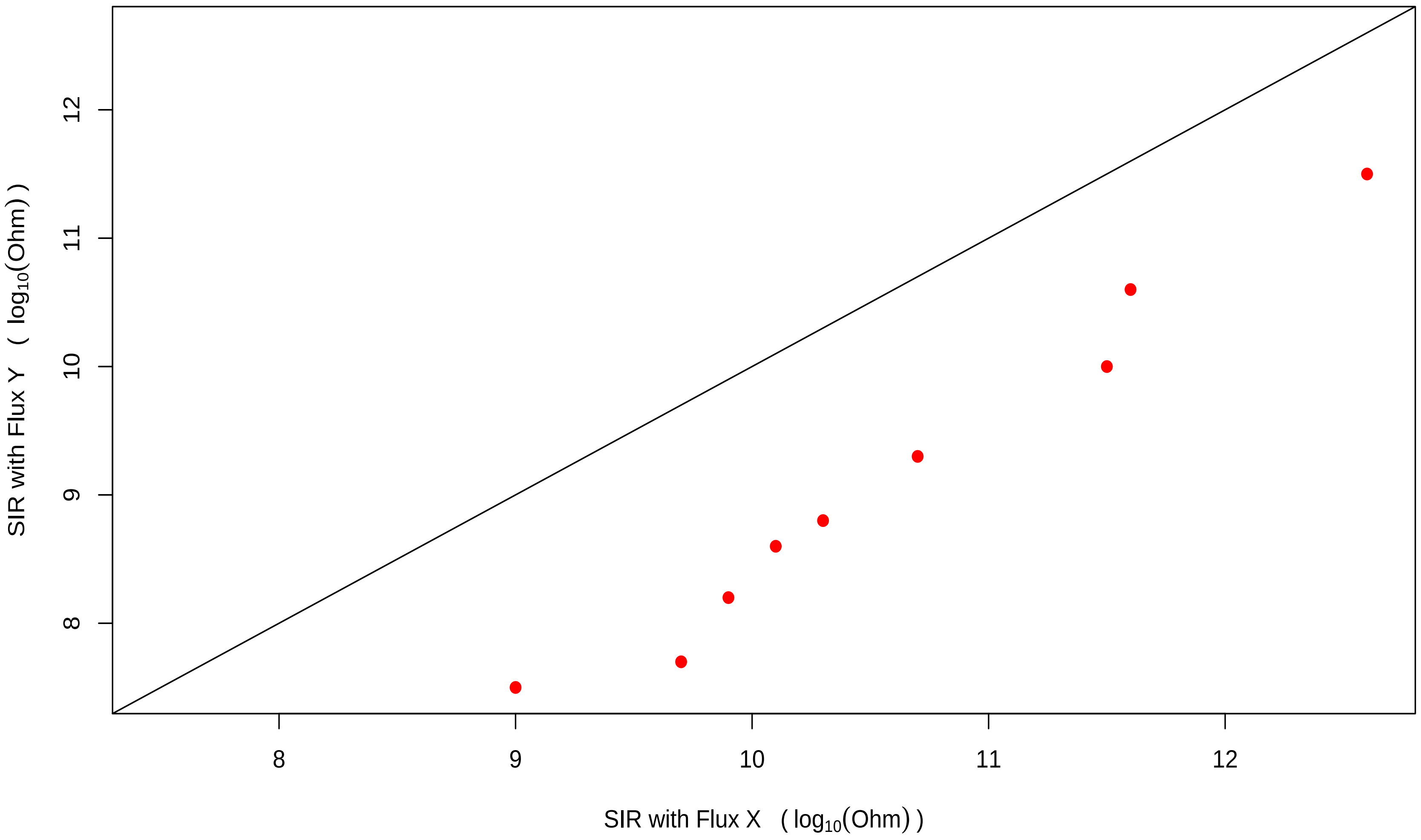
Three Studies on Taxiway Deviations

http://www.airporttech.tc.faa.gov/Design/Downloads/anc.report.091003.pdf

http://www.airporttech.tc.faa.gov/Design/Downloads/jfk.report.pdf

http://www.airtech.tc.faa.gov/Design/Downloads/separation%20new.pdf

Returning to the Flux Experiment: QQ-Plot of SIR Data



What is a QQ-Plot?

A QQ-plot is a another convenient way of comparing two data sets visually.

It is easiest to construct when both samples have the same size $m = n$, in which case you plot the ordered values of one sample against the corresponding ordered values of the other sample.

For $m > n$ you linearly interpolate n values from the sorted larger sample and use those values to plot against the sorted smaller sample.

`out=qqplot(x,y)` will do this for two sample vectors `x` and `y`.

The `out=` is optional. `out` is a list with plotting positions `out$x` and `out$y`.

The command `abline(lsfit(out$x,out$y))` will fit a line to the QQ-plot.

See the function body of `qqplot` and `?approx` for the interpolation scheme.

What Can a QQ-Plot Tell You?

If the points scatter around the main diagonal

$\Longrightarrow$ samples appear to come from the same distribution.

An $\approx$ linear point pattern $\Longrightarrow$ the two samples differ at most in location and scale.

An $\approx$ linear point pattern parallel to the main diagonal

$\Longrightarrow$ the two samples differ mainly in location (not scale).

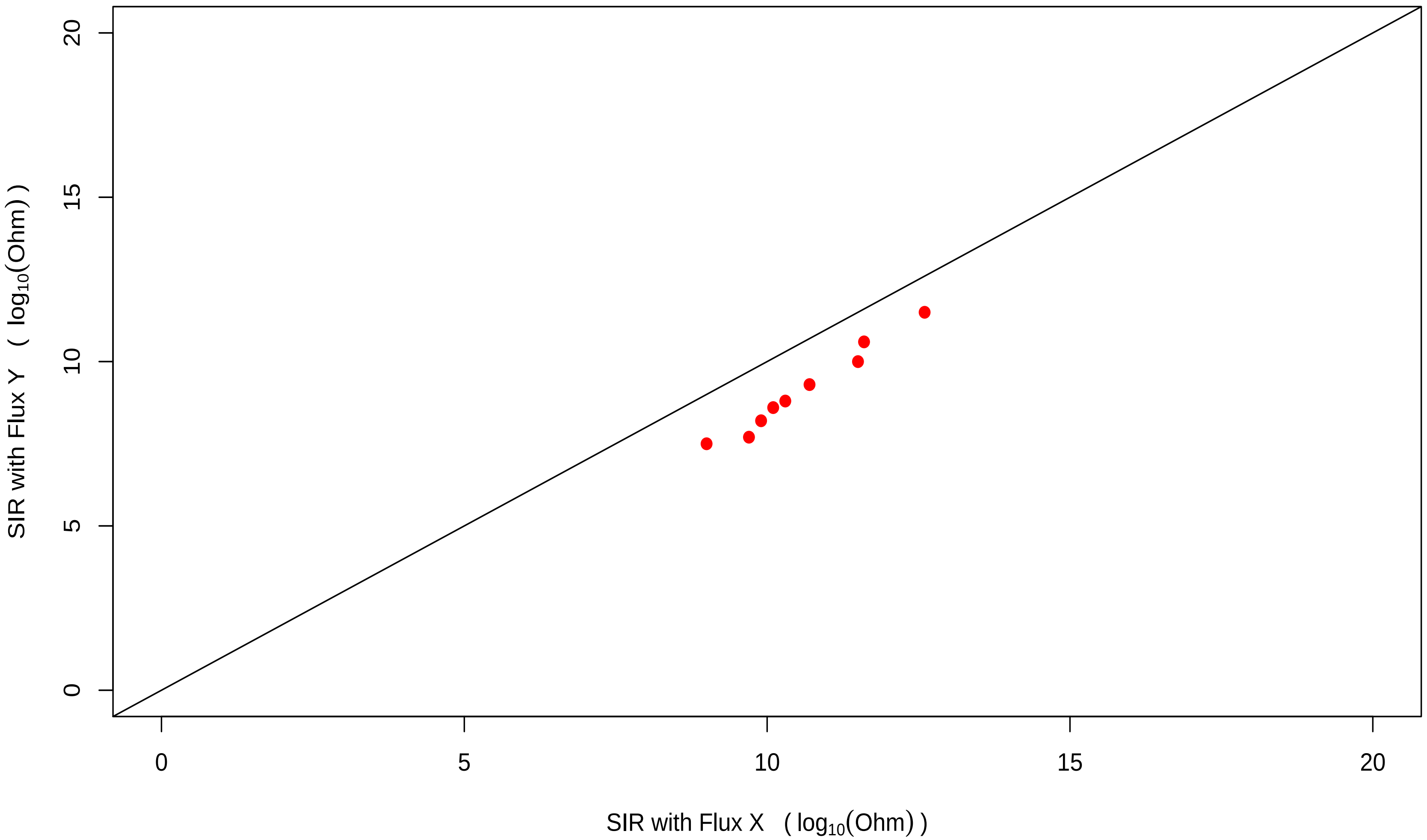
If the point pattern is $\approx$ linear and crosses the diagonal near the middle of the point pattern, then the two samples differ mainly in scale (not location).

If the point pattern does not look $\approx$ linear, then the samples differ in aspects other than location and scale. For this there are many possibilities.

Location is captured by mean or median, scale by standard deviation or IQR.

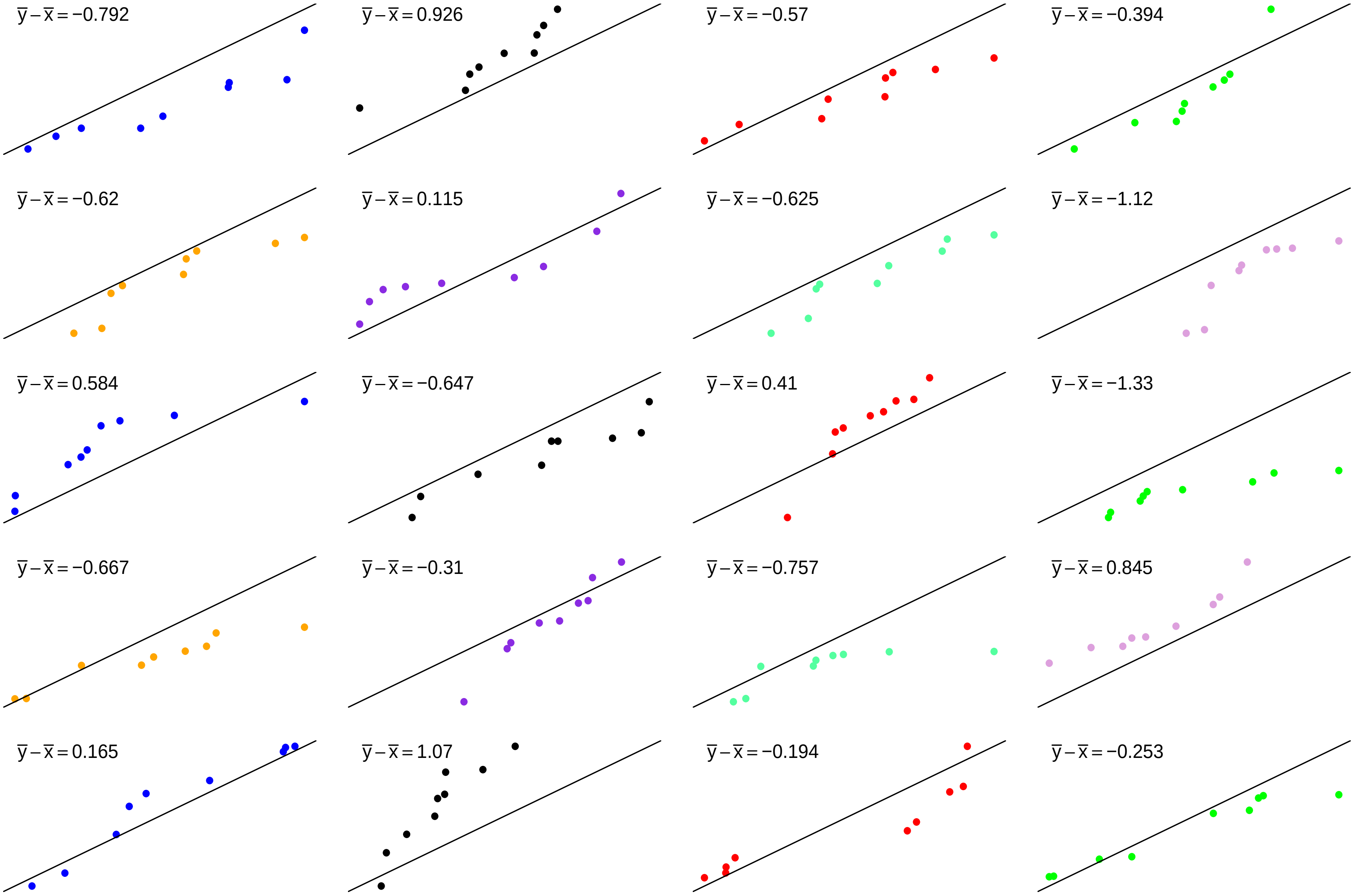
Judgment is subjective and takes experience. Simulation is one way to get it.

QQ-Plot of SIR Data (Higher Perspective?)



Here the same point pattern looks more linear and closer to the main diagonal.

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



What Do the 20 Simulated QQ-Plots Tell Us?

Very few patterns seem to vary along the main diagonal, the “expected” situation.

We say “expected” because both samples come from the same population/distribution.

Clear linear patterns are also rare. This is due to the small sample size.

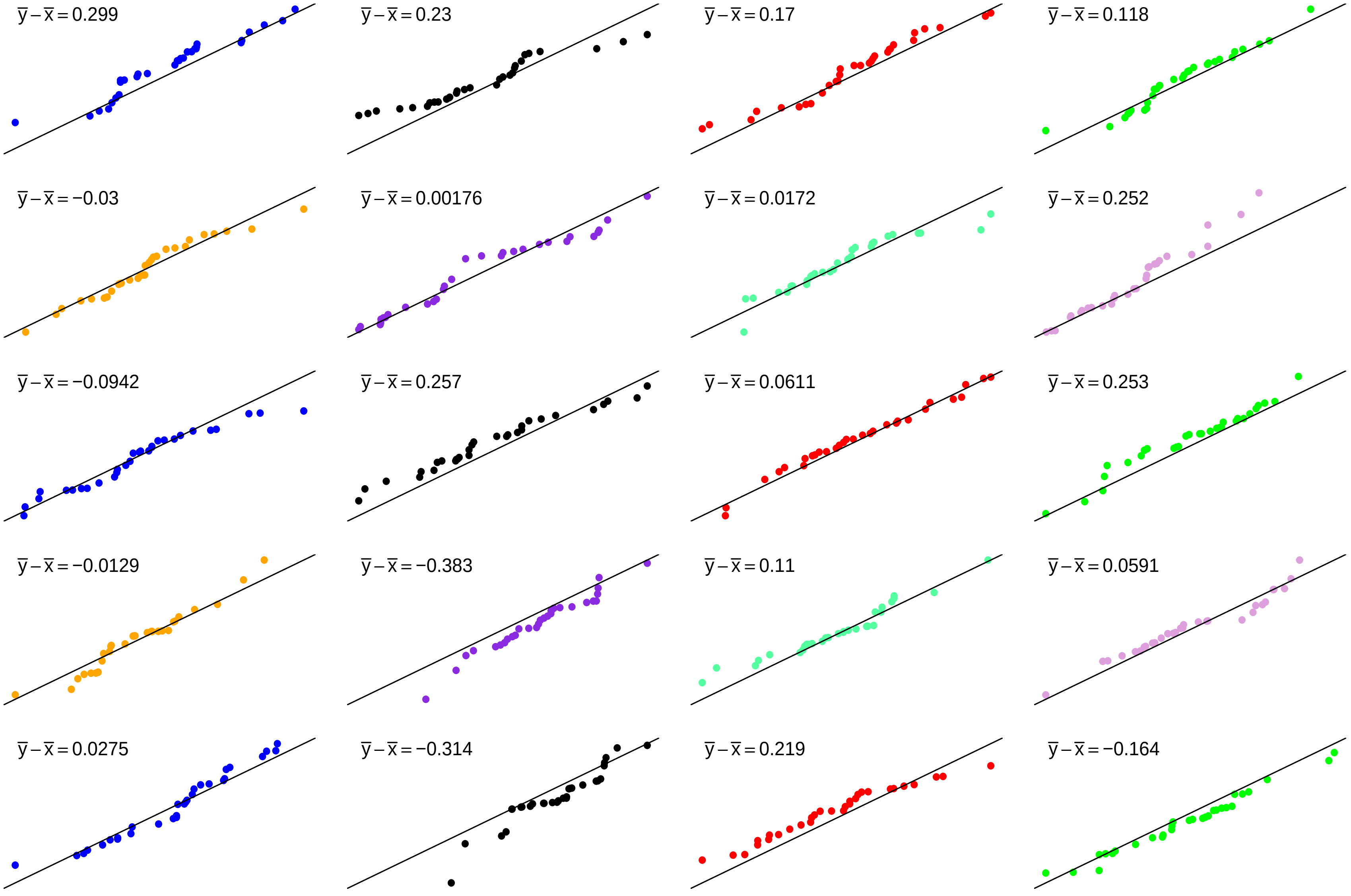
12 patterns out of 20 have points on both sides of the main diagonal.

Some patterns are clearly offset from the main diagonal.

Thus one should be rather forgiving when judging with small sample sizes.

The next slide illustrates how matters improve when both samples are of size 30.

Some QQ-Plots from N(0,1) Samples (m=30, n=30)



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 (H_0), 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 splitting of 18 boards, the 18 values would remain unchanged, but due to the different splitting into two groups of 9 & 9, we would see other differences of means.

There are $\binom{18}{9} = 48620 = \texttt{choose}(18,9)$ 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 randomization or reference distribution of $\bar{Y} - \bar{X}$ for all possible 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
10.0	10.0	10.0	10.0	10.0	10.0	10.0
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
9.0	9.0	9.0	9.0	9.0	9.0	9.0
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

The Range of $\bar{Y} - \bar{X}$

What is the most extreme difference $\bar{Y} - \bar{X}$ possible?

We get it by sorting all 18 SIR values and taking the highest 9 as one sample and the lowest 9 as the other.

7.5	7.7	8.2	8.6	8.8	9.0	9.3	9.7	9.9
10.0	10.1	10.3	10.6	10.7	11.5	11.5	11.6	12.6

Then $\bar{Y} = 8.744444$ and $\bar{X} = 10.98889$. Thus $\bar{Y} - \bar{X} = -2.244446$ would be the lowest possible value for $\bar{Y} - \bar{X}$ among all splits. Chance $1/48620 = 2.1 \cdot 10^{-5}$

Correspondingly, $\bar{Y} - \bar{X} = 2.244446$ would be the highest possible value.
Just interchange the X and Y choices, $Y' = X$ and $X' = Y$, and we get
$\bar{Y}' - \bar{X}' = \bar{X} - \bar{Y} = -(\bar{Y} - \bar{X}) = -(-2.244446) = 2.244446.$

This interchange works because $m = n = 9$. For $m \neq n$ it would not work.

Where does the observed -1.467 fall relative to these extremes ± 2.244446 and the vast majority of splits for $\bar{Y} - \bar{X}$? $\Longrightarrow$ reference distribution of $\bar{Y} - \bar{X}$.

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. Here it takes just a few seconds in R using the function `combn`, but the computing work can grow rapidly, since $\binom{m+n}{m}$ gets large in a hurry. For example $\binom{26}{13} = 10400600$ (10 million$+$).

Let `SIR` be the vector of all 18 SIR values in your work space (in any order).

In your work space define the following function of the 2 arguments `ind` and `y`

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

Then the command

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

gives the vector of all 48620 such differences of averages.

What does `combn` Do?

The function `combn` goes through all combinations of 9 indices taken from 1:18.

Each such combination is fed as the value for the first argument `ind` to the function `mean.diff.fun`. Note that `ind` is not explicitly named in `combn(...)`.

Instead of `ind` any other unused variable name could be used, e.g., `IndianaJones`.

The second argument `y` to the function `mean.diff.fun` is explicitly named in the assignment `y=SIR` in `combn(1:18,9,FUN=mean.diff.fun,y=SIR)`.

For each combination `ind` and assigned vector `SIR` $= (X_1,\ldots,X_9,Y_1,\ldots,Y_9) = (Z_1,\ldots,Z_{18})$ the function `mean.diff.fun(ind,y)` is evaluated.

For example,

`ind=c(3,5,7,9,10,12,14,17,18)` $\longrightarrow$ `mean(y[ind]) - mean(y[-ind])`

$= (Z_3+Z_5+Z_7+\ldots+Z_{17}+Z_{18})/9-(Z_1+Z_2+Z_4+\ldots+Z_{15}+Z_{16})/9.$

Using `IndianaJones`

Using

```
mean.diff.fun = function(IndianaJones,y){mean(y[IndianaJones])-
                                           mean(y[-IndianaJones])}
```

and

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

would have produced the same output vector `mean.diff.fun`.

If `mean.diff.fun = function(z,y1,y2)` we would need to explicitly indicate the sources for the two variables `y1` and `y2` in `combn(...,y1=...,y2=...)`.

combn

The default function call (with no FUN specification)

`combn(x, m, FUN = NULL, simplify = TRUE, ...)`

or just `combn(x, m)` results in a matrix with columns representing all possible combinations of m elements taken from the vector x.

Don't use `x=` in `...` since `x` is already implied as first argument.

Here is an example with `x=1:5=c(1,2,3,4,5)` and `m=3`

```
> combn(1:5,3)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    1    1    1    1    1    2    2    2     3
[2,]    2    2    2    3    3    4    3    3    4     4
[3,]    3    4    5    4    5    5    4    5    5     5
```

Note $\binom{5}{3} = \binom{5}{2} = 5!/(3!2!) = 5 \cdot 4/(1 \cdot 2) = 10.$

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

This vector, `randomization.ref.dist`, gives the reference distribution of $\bar{Y} - \bar{X}$.

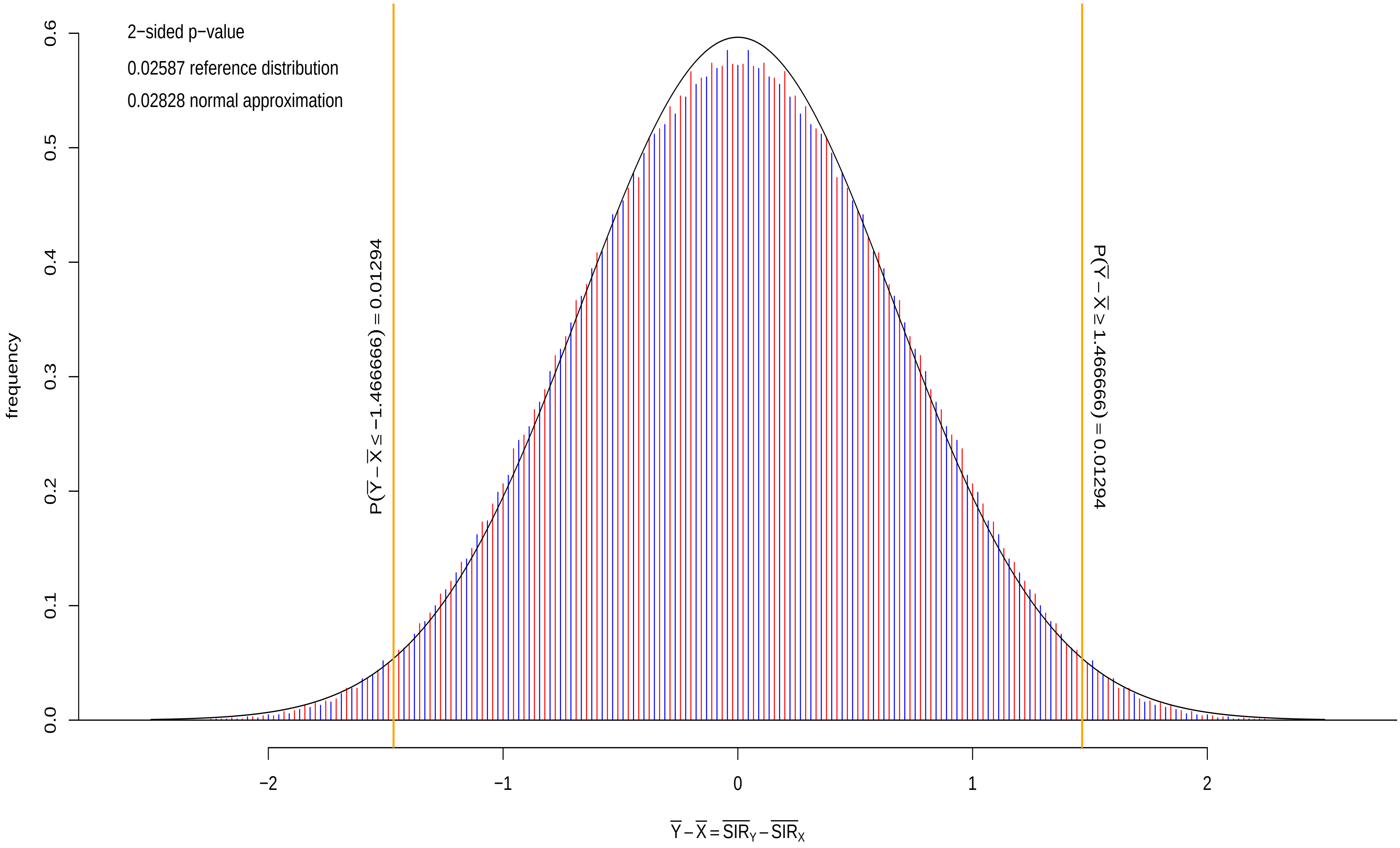
We find a (two-sided) p-value of `p.val=.02587` for our observed $\bar{Y} - \bar{X} = -1.467$ via `p.val = mean(abs(randomization.ref.dist)>= 1.46666)` $=$ proportion of `T` or `1` in the logic vector `abs(randomization.ref.dist)>= 1.46666`.

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

The upfront randomization of fluxes and the assumption of H_0 make this a valid probability statement! Without that randomization we cannot speak of chance.

With upfront randomization and assuming H_0 to be true, we can view the observed difference as having arisen as a result of one of 48620 equally likely splits.

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



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

Why is this reference distribution symmetric around zero? ($\Longrightarrow$ slide 28)

Would this still hold if $m \neq n$ with $m + n = 18$? Look at $m = 1$ and $n = 17$!

$$\bar{Y} - \bar{X} = \frac{1}{17} \sum_{i=1}^{17} Y_i - X_1 = \frac{1}{17} \left(\sum_{i=1}^{18} Z_i - X_1 \right) - X_1 = \frac{18}{17} \bar{Z} - \frac{18}{17} X_1 = \frac{18}{17} (\bar{Z} - X_1)$$

$\bar{Z}$ will stay fixed under all splits of $Z_1, \ldots, Z_{18}$ into groups of 17 and 1,

while X_1 will take on any value Z_i with equal chance $1/18$ (why $1/18$?).

The distribution of these 18 differences $\bar{Y} - \bar{X}$ (under all splits) will look symmetric

only if the distribution of all 18 Z_i values is symmetric.

How to Determine the P-Value

How to obtain the p-value from the reference distributions?

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) # the proportion of x values > 3
[1] 0.7
```

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

In arithmetic expressions or arithmetic functions the logic values `FALSE` and `TRUE` are also interpreted numerically as `0` and `1`, respectively, .

How to Determine the P-Value (continued)

`x = randomization.ref.dist` is the vector of all the differences of means, $\bar{Y} - \bar{X}$, obtained for all 48620 possible splits.

`mean(x<=-1.46666)` and `mean(x>=1.46666)` give us the respective one-sided p-values $p_1 = .01294$ and $p_2 = .01294$ for the randomization reference distribution.

Rather than adding these 2 p-values to get a two-sided p-value we can also do this directly via `mean(abs(x)>= 1.46666)=.02587`.

Here `abs(x)` is the vector of absolute values of all components in `x`.

Some Computational Issues

Note that in the p-value calculation we switched from 1.467 to 1.46666 in

`p.val = mean(abs(randomization.ref.dist)>= 1.46666)` Why?

The number -1.467 is obtained by rounding the true $\bar{y}-\bar{x} = -1.46666.....$

Thus the logic vector `abs(randomization.ref.dist)>= 1.467`

would give an `F` to all those splits with $|\bar{Y} - \bar{X}| = 1.46666....$

`p.val = mean(abs(randomization.ref.dist)>= 1.467)=.02345`

would compute the chance of seeing a result for $|\bar{Y} - \bar{X}|$ that is more extreme than the observed $|\bar{y}-\bar{x}| = 1.46666.....$

However, the p-value is the chance of seeing a $|\bar{Y} - \bar{X}|$ that is as extreme as or more extreme than the observed $|\bar{y}-\bar{x}| = 1.46666....$, i.e., `p.val = .02587`.

Interpreting the P-Value

The calculation of the p-value is based on two premises

1. The fluxes were assigned randomly to the 18 units upfront. TRUE!!

2. The hypothesis H_0 of no flux effect. TRUE or FALSE.

If the p-value is very small, then we can either believe that we saw a rather rare result for our test statistic $\bar{Y} - \bar{X}$ while H_0 is true, or we should be induced to disbelieve H_0 and reject it. When is the p-value sufficiently small?

When p-value $\leq .05$ the result is said to be significant at level $.05$, and we should reject H_0, if $.05$ is our prearranged cut-off or significance level.

The actual p-value is more informative than stating a significant result at .05 or .01.

One-Sided or Two-Sided P-Value?

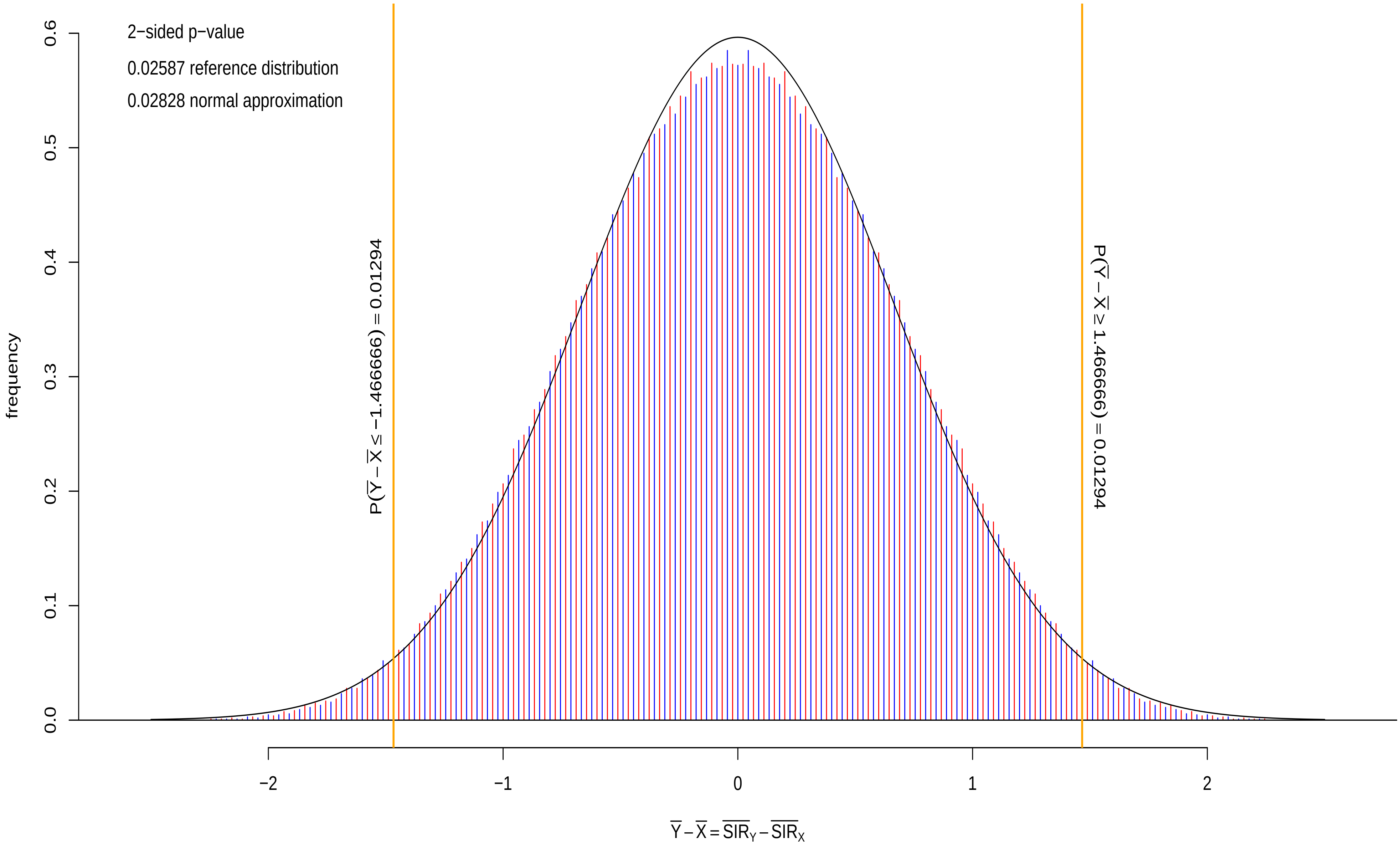
A two-sided p-value is appropriate if we test the hypothesis of no treatment difference against the alternative that there is a difference, without specifying in which direction.

Here we may hope for an insignificant result so that we may be able to use both treatments interchangeably.

A one-sided p-value would be appropriate if we test the hypothesis of no treatment difference against an alternative that specifies the direction in which the means may differ, say the newer treatment, flux X, yields higher SIR values than flux Y, the previous treatment.

However, the test result should not influence the choice of the targeted alternative.

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



How Many Distinct Values in the Reference Distribution?

How many distinct values of $\bar{Y}-\bar{X}$ can we have?

We saw previously (from the split below) that the max/min values are ± 2.244446.

7.5	7.7	8.2	8.6	8.8	9.0	9.3	9.7	9.9
10.0	10.1	10.3	10.6	10.7	11.5	11.5	11.6	12.6

The smallest non-zero change in $\bar{Y}-\bar{X}$ is attained when exchanging a Y_j with an X_i that differ by .1, in which case the change in $\bar{Y}-\bar{X}$ is $\pm .1 \cdot 2/9 = \pm .02222222$.

For example, we could exchange colors among 9.9 and 10.0.

At most $202 = (2.244446 - (-2.244446))/(.02222222)$ such increments fit between the two extremes ± 2.244446.

This means that there are at most 203 distinct values of $\bar{Y}-\bar{X}$ (the correct count), but with widely varying frequencies as shown on the previous slide.

A Matter of Numerical Accuracy

```
> length(unique(randomization.ref.dist))
[1] 261
> length(unique(round(randomization.ref.dist,6)))
[1] 203
```

Numerical inaccuracies can arise from limits on internal representation of numbers in binary form.

For example,

$$\frac{7.7+10.1+10.0}{3}-\frac{7.5+10.3+8.2}{3}=\frac{17.8+10.0}{3}-\frac{17.8+8.2}{3}=\frac{1.8}{3}=.6$$

yet

```
> mean(c(7.7,10.1,10))-mean(c(7.5,10.3,8.2))-.6
[1] -3.330669e-16
```

Which Normal Distribution Do We Use as Approximation?

When taking n values $Y_1,\ldots,Y_n$ randomly and without replacement from $Z_1,\ldots,Z_N$, then sampling theory gives the mean and variance of $\bar{Y}=\sum_{i=1}^n Y_i/n$ as

$$E(\bar{Y})=\bar{Z} \quad \text{and} \quad \mathrm{var}(\bar{Y})=\frac{S^2}{n}\left(1-\frac{n}{N}\right) \quad \text{with} \quad S^2=\frac{1}{N-1}\sum_{i=1}^{N}(Z_i-\bar{Z})^2 \,.$$

$(1-n/N)$ is called the finite population correction factor. See Appendix A.

Further theory, a version of the Central Limit Theorem (CLT), suggests that $\bar{Y}$ has an approximate normal distribution with this mean and variance.

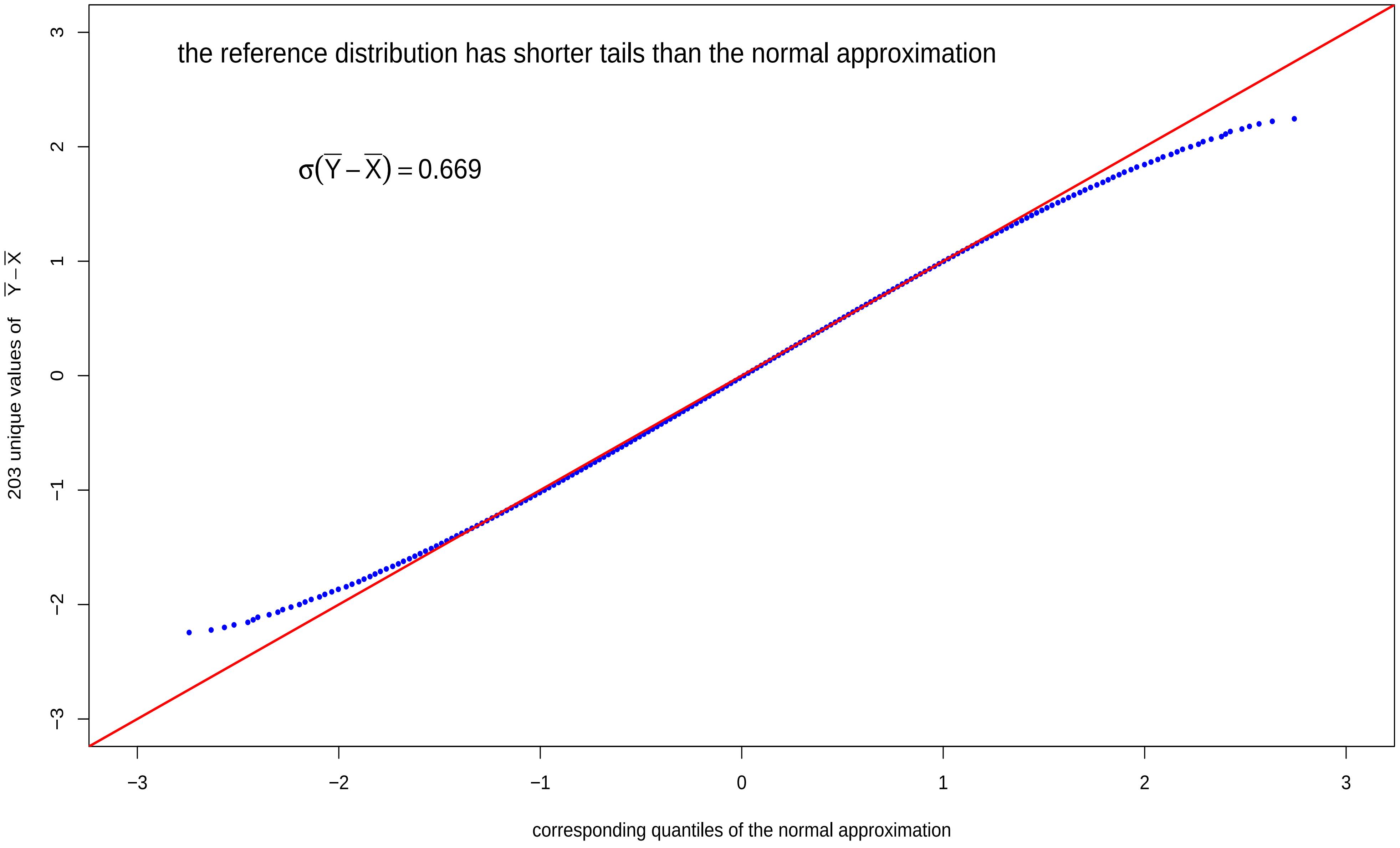
If $\bar{X}$ is the average of the $N-n=m$ Z-values not contained in the Y-sample, then

$$\bar{Z}=\frac{n\bar{Y}+m\bar{X}}{N} \quad \Longrightarrow \quad \bar{Y}-\bar{Z}=\bar{Y}-\frac{n\bar{Y}+m\bar{X}}{N}=\frac{m}{N}(\bar{Y}-\bar{X}) \quad \Longrightarrow \quad \bar{Y}-\bar{X}=\frac{N}{m}(\bar{Y}-\bar{Z})\,.$$

Since $\bar{Z}$ remains constant during any of this random sampling, it follows that $\bar{Y}-\bar{X}$ is approximately normally distributed with mean and variance given by

$$E(\bar{Y}-\bar{X})=0 \qquad \text{and} \qquad \mathrm{var}(\bar{Y}-\bar{X})=\frac{N^2}{m^2}\frac{S^2}{n}\left(1-\frac{n}{N}\right)=S^2\frac{N}{mn}=S^2\left(\frac{1}{m}+\frac{1}{n}\right) \,.$$

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



Construction of the Previous Normal QQ-Plot

Let $D_1 < D_2 < \ldots < D_{203}$ denote the distinct possible values of $\bar{Y} - \bar{X}$ and let n_i be the frequency of $\bar{Y} - \bar{X} = D_i$ among the $M = 48620$ possible splits.

The proportion of $\bar{Y} - \bar{X} \leq D_i$ is $p_i = (n_1 + \ldots + n_i)/M$ and one could regard D_i as the p_i-quantile of the reference distribution and plot it against the p_i-quantile of the approximating normal distribution with mean zero and standard deviation $\sigma = .669$.

Unfortunately $p_{203} = 1$ and the 1-quantile of the normal distribution $= \infty$.
To remedy this and also to treat p_1 and p_{203} in a more symmetric fashion we modify p_i to $\tilde{p}_i = p_i \cdot M/(M+1)$ and plot D_i against the $\tilde{p}_i$-quantile of the above normal distribution. This was done in the previous plot.

Another option, with roughly the same effect, is to use $\tilde{p}_i = (n_1 + \ldots + n_i - .5)/M$.

We note that the QQ-plot shows the discrepancy in the normal approximation much more clearly (in the tails) than the density plot superimposed on the histogram.

Some Experimenting with Randomization Tests

In the following we will take our SIR data and subtract $\bar{Y}-\bar{X}$ from all the Y values, i.e., obtain $Y_i' = Y_i - (\bar{Y}-\bar{X}), \;\; i = 1,2,\ldots,9$.

This altered Y' sample will have as average $\bar{Y}' = \bar{Y} - (\bar{Y}-\bar{X}) = \bar{X}$, i.e., both samples are aligned on their common average.

Then we take Y_1' and add $9\times(\bar{Y}-\bar{X})$, i.e., obtain $Y_1'' = Y_1' + 9\times(\bar{Y}-\bar{X})$, and leave the other $Y_i'' = Y_i'$, $i = 2,\ldots,9$ unchanged.

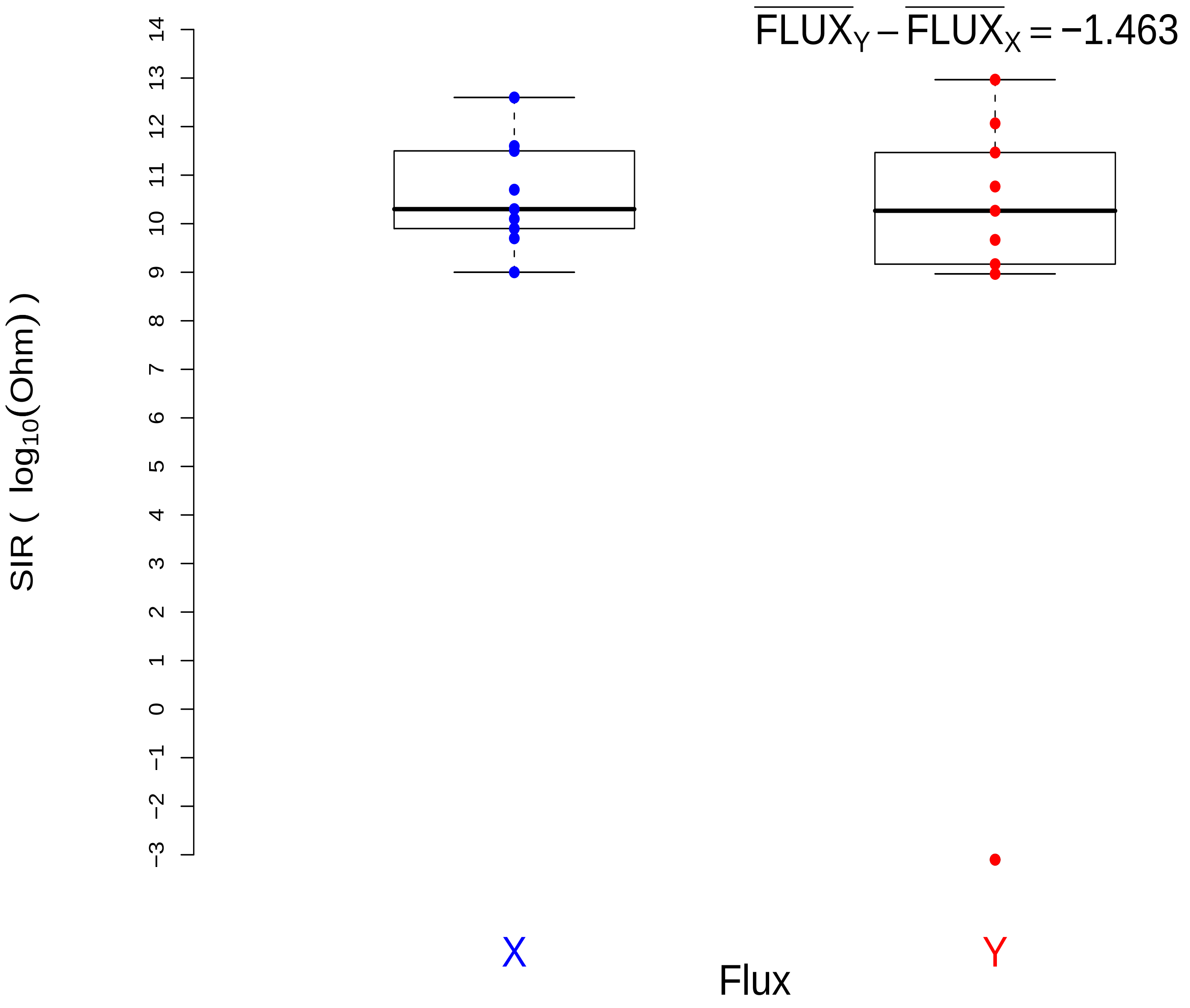
$$\Longrightarrow \quad \frac{1}{9}\left(Y_1'' + \ldots + Y_9''\right) = \frac{1}{9}\left(Y_1' + \ldots + Y_9'\right) + (\bar{Y}-\bar{X}) = \bar{X} + (\bar{Y}-\bar{X}) = \bar{Y}$$

$$\Longrightarrow \quad \bar{Y}'' - \bar{X} = \bar{Y} - \bar{X} \quad \text{same difference as before.}$$

However, we round $Y_1'',\ldots,Y_9''$ to nearest decimal. This changes $\bar{Y}''-\bar{X}$ slightly.

What is the effect of this data change on the randomization reference distribution?

Box Plot View of Altered SIR Data



Comments and Questions

The two samples seem well aligned when comparing medians.

The Y'' sample has an outlier, which does not affect the median.

$\bar{Y}'' - \bar{X} = -1.463$, which is ≈ -1.467, our previous value (rounding).

It seems that there is no difference in Flux, if we disregard the outlier.

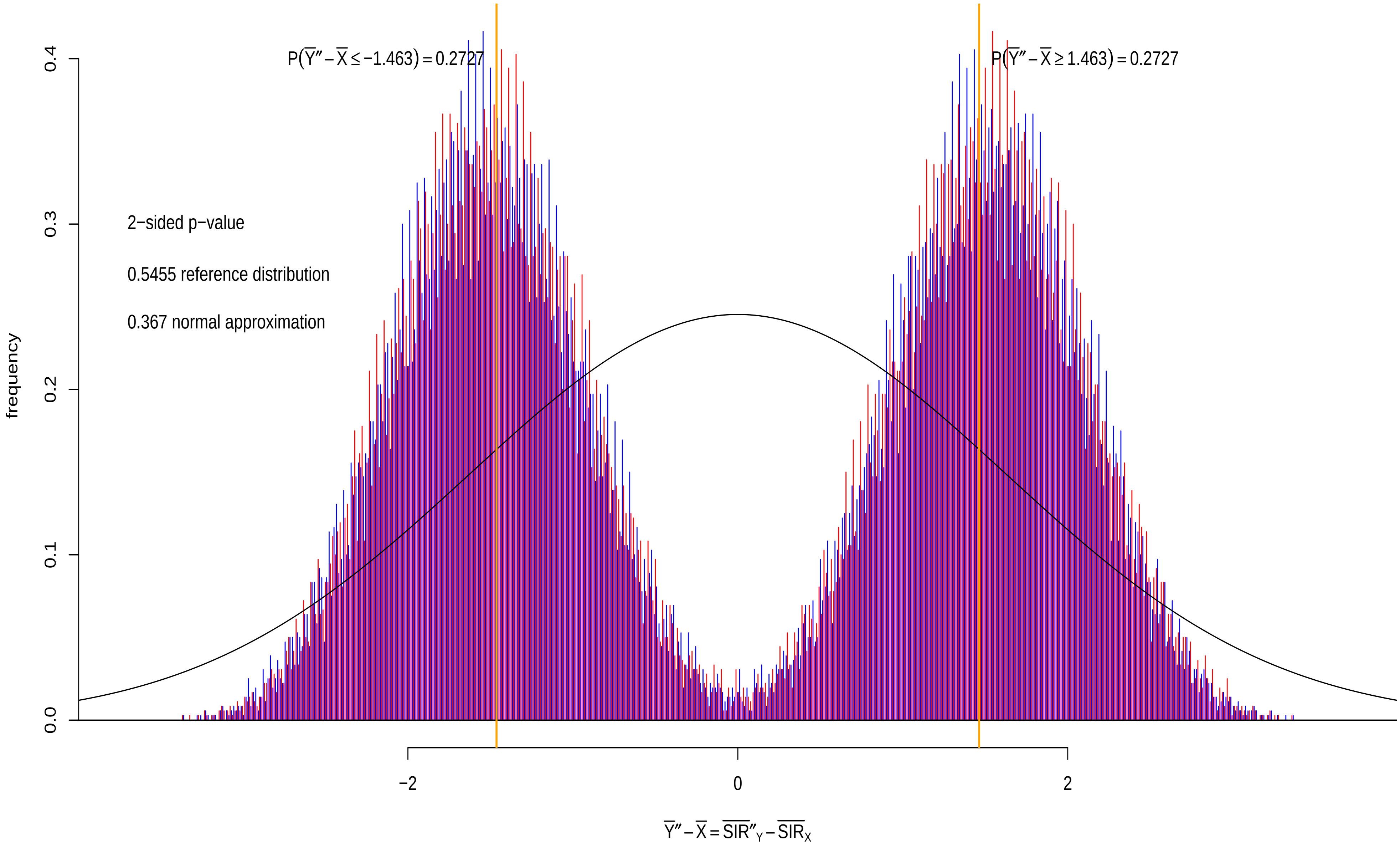
The outlier is presumably a defective circuit board or something gone wrong with this board and is not a treatment effect (otherwise should see more such cases).

What will happen to $\bar{Y}'' - \bar{X}$ as we compute it for all 48620 splits?

Where will -1.463 be positioned relative to the reference distribution?

What is its two-sided p-value?

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



Comments

The randomization reference distribution does not judge this outlier sample an unusual split. The randomization test is not fooled by the outlier.

The two-sided p-value of .55 shows no significance and basically reflects the fact that the outlier has equal chance of being in either the X-sample or the Y-sample under random splits.

The normal approximation is lousy and should not be used here.

In fact, the randomization reference distribution looks like a $(.5, .5)$ mixture of two approximately normal distributions, centered at ± 1.463, respectively.

The CLT effect still shows in these two constituent distributions, generated as randomization distributions from samples of 8 and 9 or 9 and 8, respectively.

Some More Experimenting with Randomization Tests

In the following we will take our SIR data and subtract $\bar{Y} - \bar{X}$ from all the Y values, i.e., obtain $Y_i' = Y_i - (\bar{Y} - \bar{X}), \;\; i = 1, 2, \ldots, 9$.

This altered Y' sample will have average $\bar{Y}' = \bar{Y} - (\bar{Y} - \bar{X}) = \bar{X}$,
i.e., both samples are aligned on their common average.

Then we take $Y_1', \ldots, Y_4'$ and add $(9/4) \times (\bar{Y} - \bar{X})$ to each, i.e., obtain
$Y_i'' = Y_i' + (9/4) \times (\bar{Y} - \bar{X})$ for $i = 1, 2, 3, 4$, and leave the other
$Y_i'' = Y_i'$, $i = 5, \ldots, 9$ unchanged.

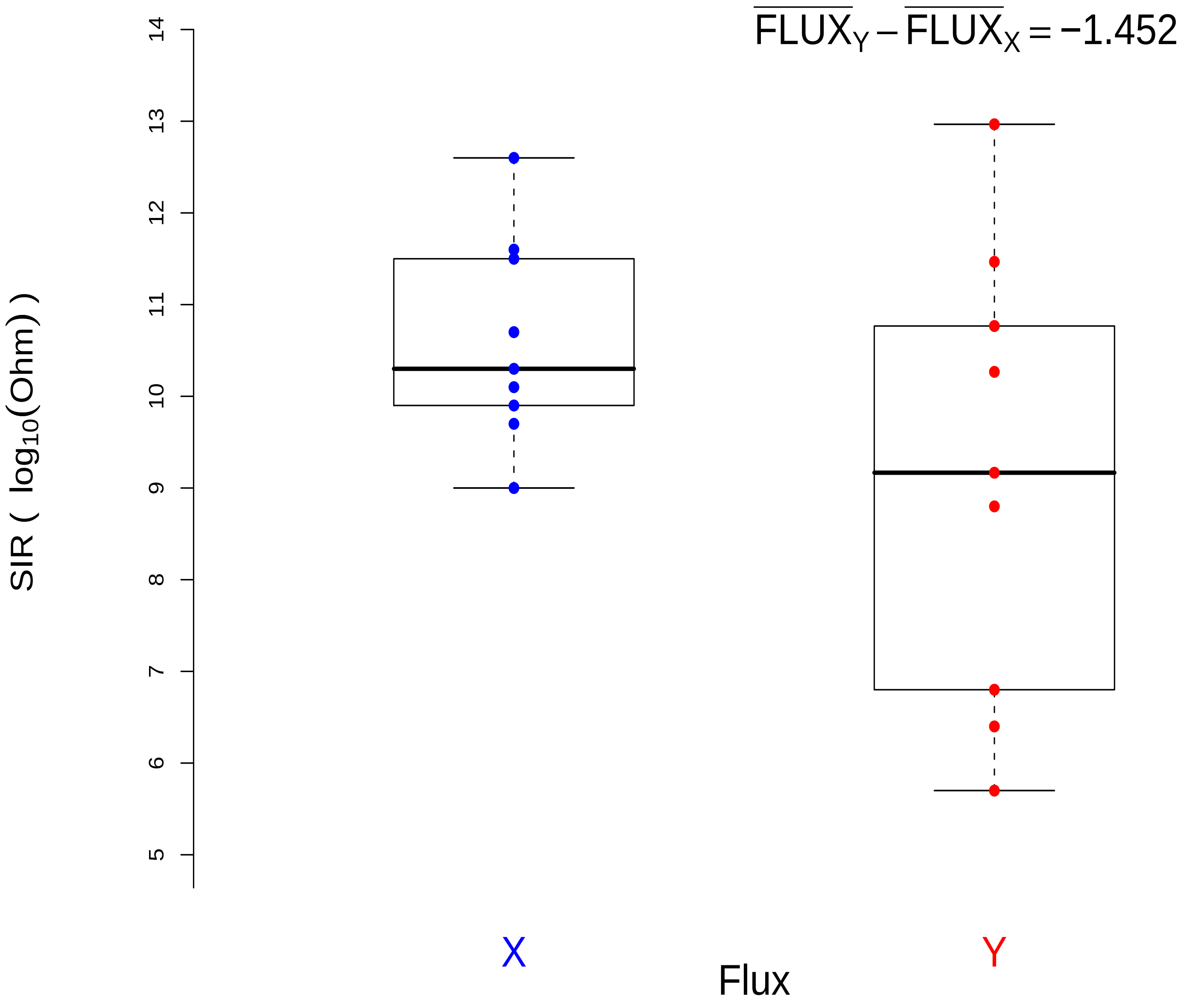
$$\Longrightarrow \quad \frac{1}{9}\left(Y_1'' + \ldots + Y_9''\right) = \frac{1}{9}\left(Y_1' + \ldots + Y_9'\right) + (\bar{Y} - \bar{X}) = \bar{X} + (\bar{Y} - \bar{X}) = \bar{Y}$$

$$\Longrightarrow \quad \bar{Y}'' - \bar{X} = \bar{Y} - \bar{X} \quad \text{same difference as before.}$$

Again we will round $Y_1'', \ldots, Y_9''$ to nearest decimal.

How significant will this be in the randomization reference distribution?

Box Plot View of Altered SIR Data



Comments and Questions

The two samples seem not so well aligned when comparing medians.

Three Y values seem to fall apart from the other 6, which results in
$\bar{Y}'' - \bar{X} = -1.452$ which is ≈ -1.467, our previous value.

It seems that there is a difference in Flux, but it shows in a difference
in mean and scale (center and spread).

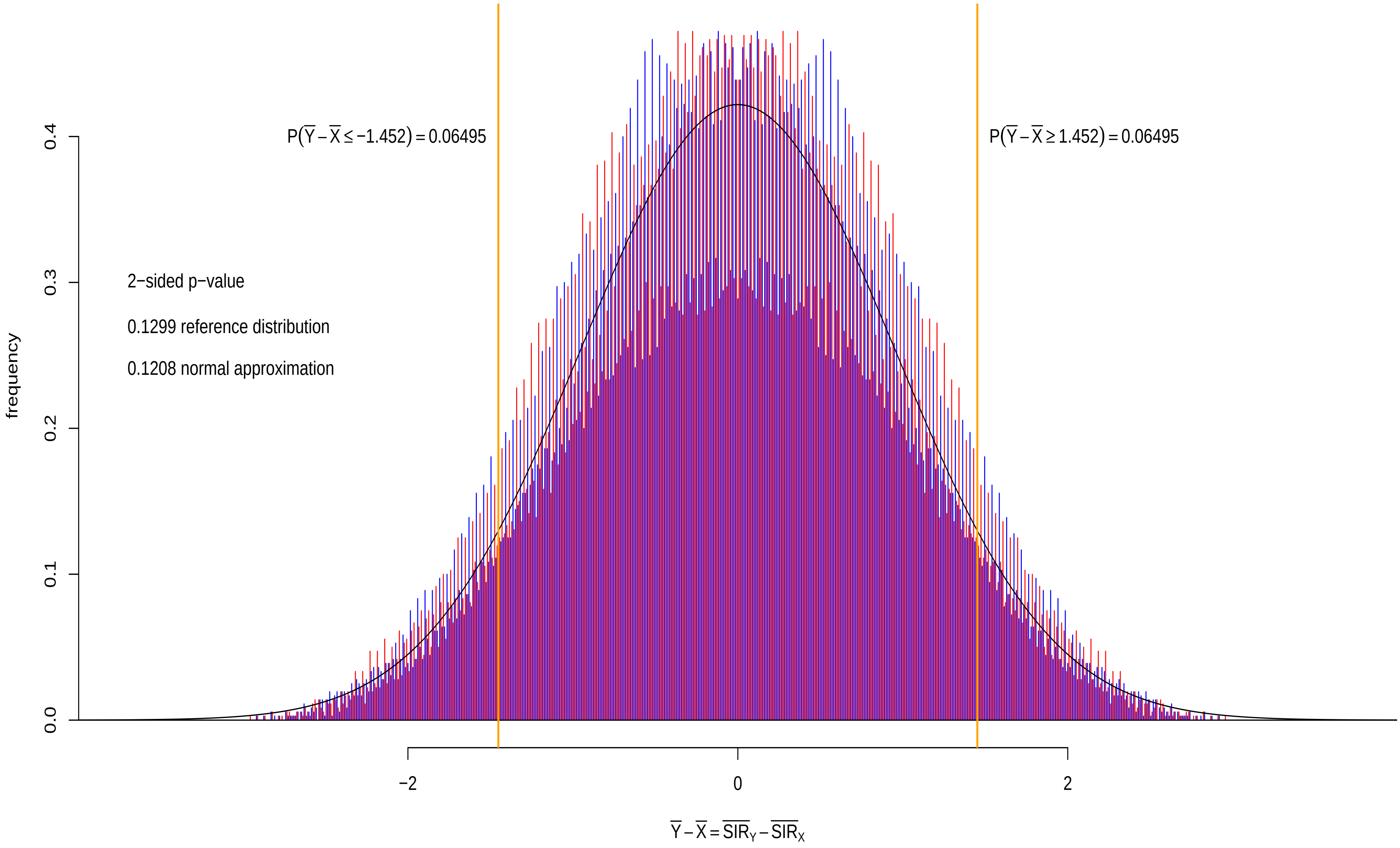
The three "outliers" could presumably be defective circuit boards
or represent more strongly varying treatment effects (treatment/unit interactions).

What will happen to $\bar{Y}'' - \bar{X}$ as we compute it for all 48620 splits?

Where will -1.452 be positioned relative to the reference distribution?

What is its two-sided p-value?

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



Comments

The randomization reference distribution does not judge this more dispersed

and shifted Y-sample a very unusual split.

The randomization test looks only at the difference in averages and not at scale changes, thus loses some efficacy.

The two-sided p-value of .13 shows no significance.

The normal approximation is reasonable when compared to the previous case,

especially when smoothing over the very local jaggedness,

i.e., using a wider histogram bin width.

A difference in sample dispersion has a negative impact when comparing for means.

The Effect of Spread

In the following we will take our SIR data and transform them as follows

$$X_i' = \bar{X} + .5 \times (X_i - \bar{X}) \qquad \text{and} \qquad Y_i' = \bar{Y} + .5 \times (Y_i - \bar{Y})$$

Note that $\bar{X}' = \bar{X}$ and $\bar{Y}' = \bar{Y}$, and thus $\bar{Y}' - \bar{X}' = \bar{Y} - \bar{X}$ is unchanged.

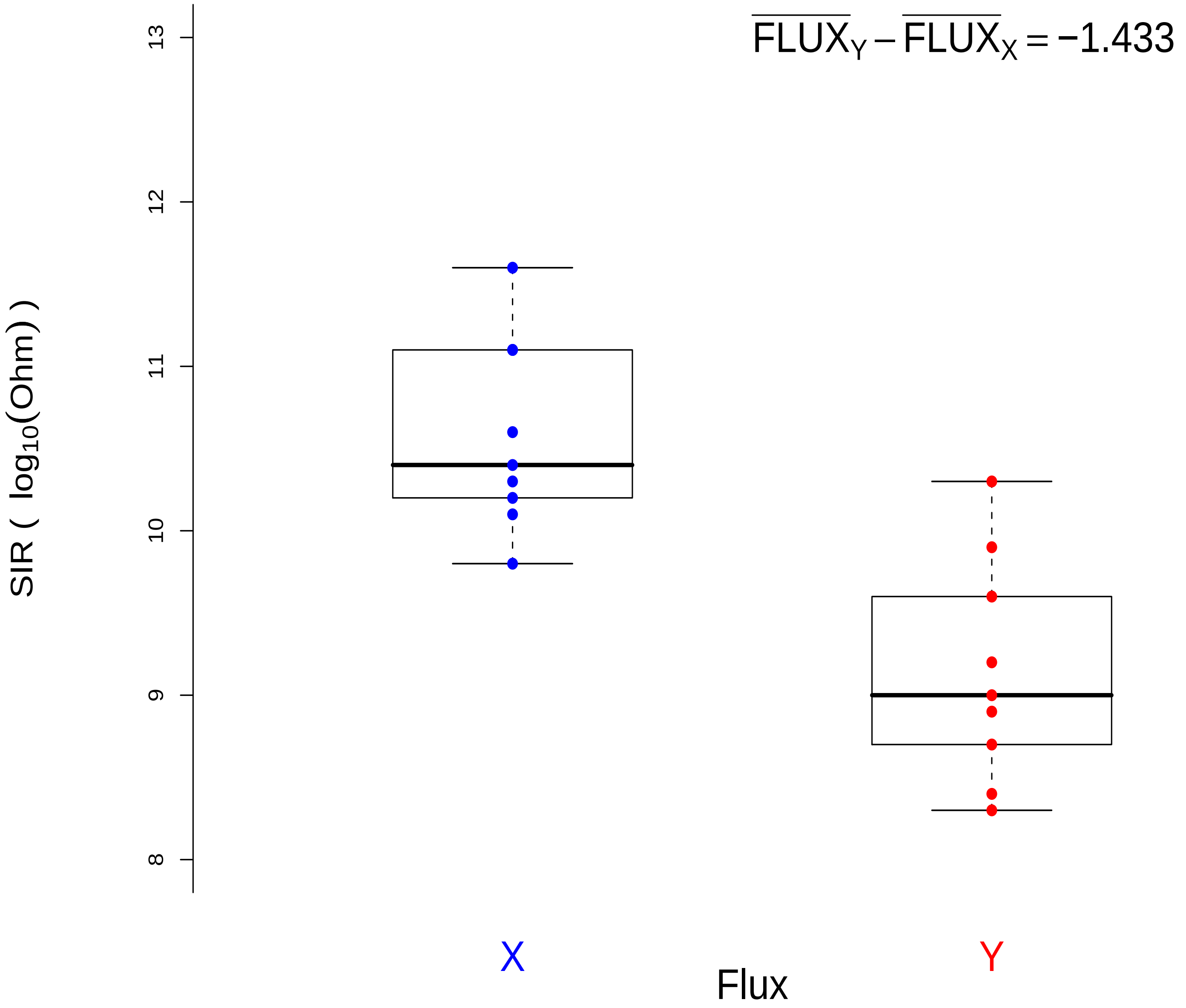
However, we have reduced the spread in both samples by a factor $.5$

$$\sqrt{\frac{\sum_i (X_i' - \bar{X}')^2}{m-1}} = .5 \times \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{m-1}} \qquad \text{and similarly for the } Y_i'.$$

Again we will round $X_1', \ldots, X_9'$ and $Y_1', \ldots, Y_9'$ to nearest decimal.

What is the effect of this change on the randomization reference distribution?

Box Plot View of Altered SIR Data



Comments and Questions

The two samples seem to have roughly the same difference in sample means

$\bar{Y}' - \bar{X}' = -1.433$ as opposed to the $\bar{Y} - \bar{X} = -1.467$ in the original data.

The difference comes from the rounding.

It seems that here the difference in flux is more clearly defined,

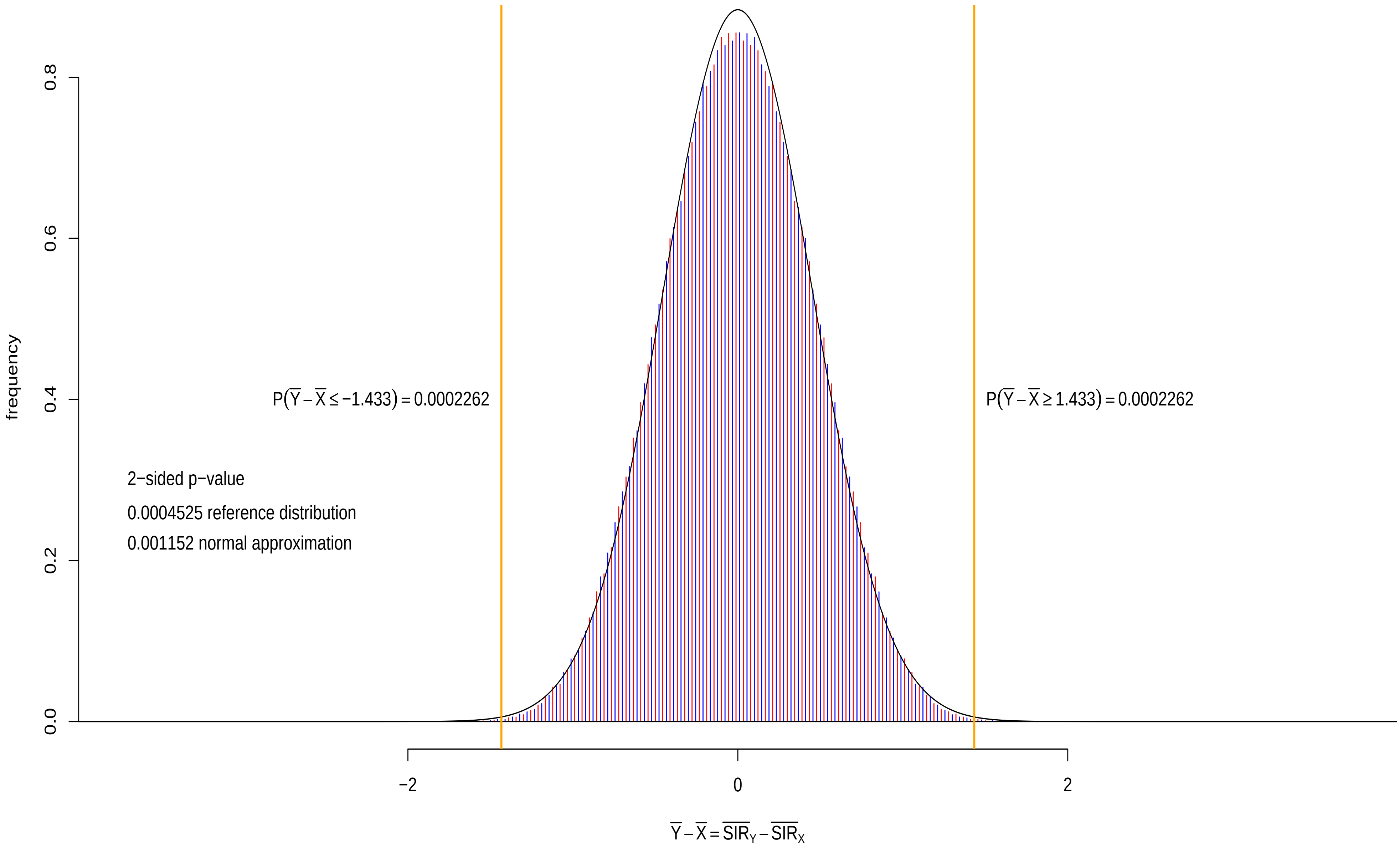
because each sample is more tightly scattered around its center.

Clearer separation of signal from noise!

Where will -1.433 be positioned relative to the reference distribution?

What is its two-sided p-value?

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



Comments and Questions

The two-sided p-value `.00045` is much smaller, as one would expect, because the evidence for a shift seems stronger against the reduced background noise.

The normal approximation seems somewhat reasonable but the resulting two-sided p-value `.0012` is more than twice as large as the p-value `.00045` based on the randomization reference distribution. (Tail discrepancy! Slide 45)

If instead of tightening the spread of each sample by a factor .5 we enlarge the spread by a factor of 2 what would be the result?

If instead of tightening the spread of each sample by a factor .5 we tighten the spread by a factor of .5 in only one of the samples, what would be the result?

Computation Time and Storage Issues

```
 ref.dist2 = function ()
{
SIR=flux$SIR
randomization.ref.dist=combn(1:18,9,FUN=mean.diff.fun,y=SIR)
randomization.ref.dist
}
> system.time(ref.dist2())
   user  system elapsed
   4.23    0.00    5.49
```

5.49 seconds to compute the $\binom{18}{9} = 48620$ values of the reference distribution.

30 experimental units split into 15 and 15 $\Longrightarrow \binom{30}{15} = 155,117,520$ splits

and it would take at least $5.49 \cdot 155117520/48620 = 17515.33$ seconds

or 4.9 hours.

Computation times and storage requirements get out of hand in a hurry.

We need an alternative computational approach.

Approximation to Randomization Reference Distribution

A simple way out of this dilemma of exploding computation times and storage requirements is to generate a sufficiently large random sample (with replacement), say $K = 10,000$, of combinations from this set of all $\binom{m+n}{m}$ combinations.

Compute the statistic of interest, $S_i = S(\underline{X}_i, \underline{Y}_i) = \bar{Y}_i - \bar{X}_i,\ i = 1, \ldots, K$ for each sampled combination and approximate the randomization reference distribution by the distribution of the $S_1, \ldots, S_K$. In particular, approximate the probability

$$F(z) = P(S(\underline{X}, \underline{Y}) \le z) = P(\bar{Y} - \bar{X} \le z)$$

by the proportion of simulated sample statistics $S_1, \ldots, S_K$ that are $\le z$, i.e.,

$$F(z) = P(S(\underline{X}, \underline{Y}) \le z) \approx \widehat{F}_K(z) = \frac{1}{K} \sum_{i=1}^{K} B_i(z)$$

with $\quad B_i(z) = 1 \quad$ when $\quad S_i = S(\underline{X}_i, \underline{Y}_i) \le z \quad$ and $\quad B_i(z) = 0 \quad$ else.

The Law of Large Numbers (LLN)

For any z we have $E(B_i(z)) = F(z)$. Since the $B_i(z)$ are independent and identically distributed (iid), the Law of Large Numbers (LLN)

$$\Longrightarrow \qquad \widehat{F}_K(z) = \frac{1}{K}\sum_{i=1}^{K} B_i(z) \longrightarrow F(z) \quad \text{as} \quad K \rightarrow \infty\,.$$

Similarly, we can approximate $\tilde{p}(z) = P(|S(\underline{X},\underline{Y})| \geq |z|) = P(|\bar{Y} - \bar{X}| \geq |z|)$ by

$$\widehat{p}_K(z) = \frac{1}{K}\sum_{i=1}^{K} \tilde{B}_i(z)$$

with $\tilde{B}_i(z) = 1$ when $|S_i| = |S(\underline{X}_{\,i}, \underline{Y}_{\,i})| \geq |z|$ and $\tilde{B}_i(z) = 0$ else.

For $z = \bar{y} - \bar{x}$, the observed sample mean difference, $\widehat{p}_K(z)$ gives us a good approximation for the true two-sided p-value of z.

Sample Simulation Program

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

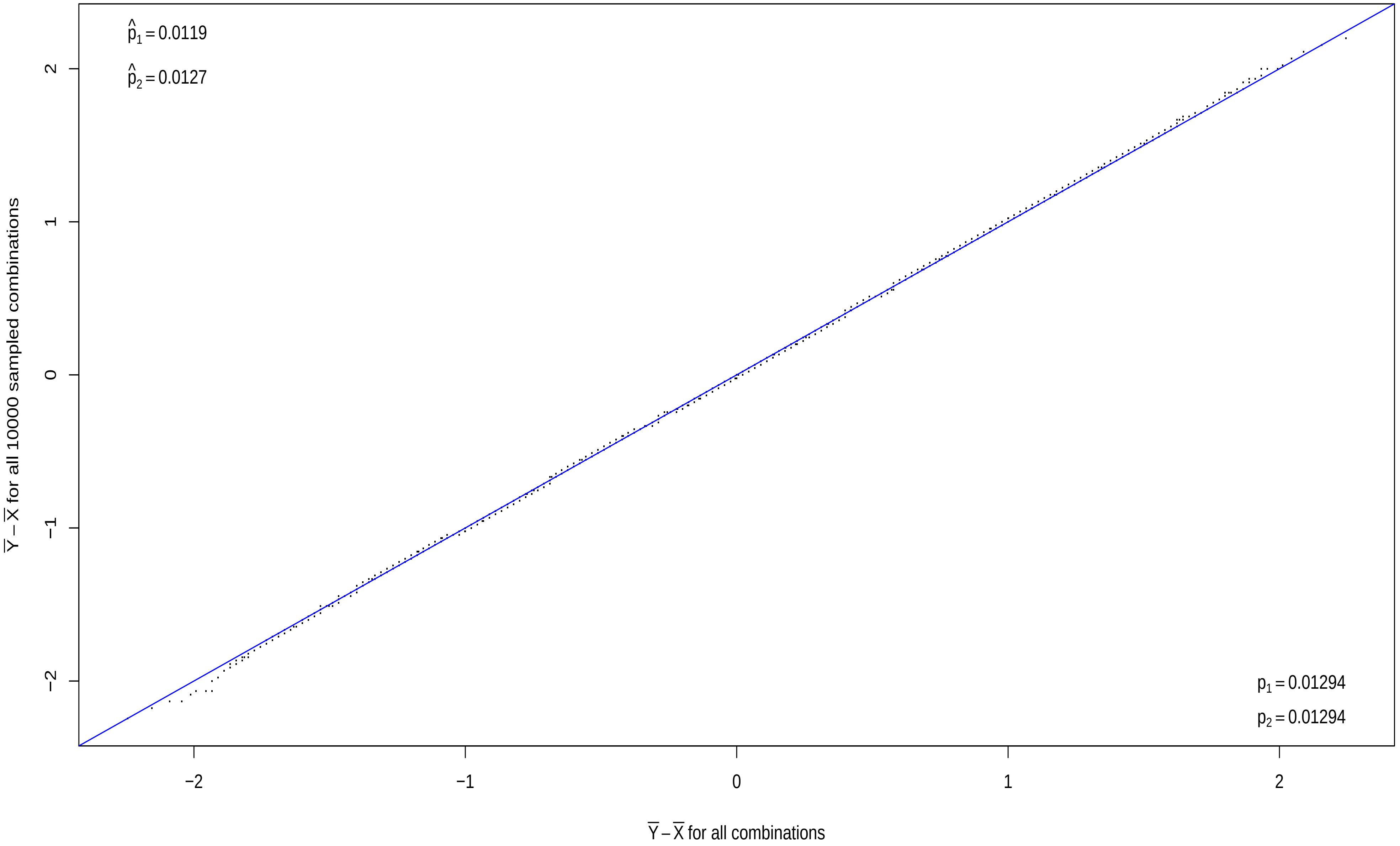
```
simulated.reference.distribution=function(K=10000){
  D.star=NULL
  for(i in 1:K){
    SIR.s=sample(SIR)
    D.star[i]=mean(SIR.s[1:9])-mean(SIR.s[10:18])
  }
D.star}
```

The output vector `D.star` holds the sampled reference distribution of $\bar{Y}-\bar{X}$ from which estimated p-values can be computed for any observed $\bar{y}-\bar{x}$.

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



The 2-Sample Student t-Test Statistic

When testing the equality of means of two normal populations it is customary and optimal to compare the respective sample means relative to a measure of the dispersion in the two samples $\Longrightarrow$ Student's 2-sample t-statistic

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

In the presence of substantial inherent variation "small" differences in means could easily be due to this variation and thus will not appear significant.

Thus one should judge mean differences relative to a measure of this variation.

So far we have not assumed that we sample normal populations.
In fact, we only guaranteed that we sample from a finite population $Z_1,\ldots,Z_N$.

Randomization Distribution of the 2-Sample t-Statistic

The values of $t(\underline{X},\underline{Y})$ are in 1-1 monotone increasing correspondence with those of $\bar{Y}-\bar{X}$ under all splits. See Appendix B.

In fact,

$$\Longrightarrow \quad t(\underline{X},\underline{Y}) = \frac{\sqrt{m+n-2}}{\sqrt{1/n+1/m}} \frac{W}{\sqrt{1-\frac{mn}{N}W^2}} \quad \nearrow \text{ in } W$$

Here

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

with $\sum(Z_k-\bar{Z})^2$ staying constant over all splits.

Thus p-values for $t(\underline{x},\underline{y})$ and $\bar{y}-\bar{x}$ will be the same under each respective randomization reference distribution.

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

In regular* situations the randomization reference distribution of $t(\underline{X},\underline{Y})$ is very well approximated by a Student 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 Student t_{16}-distribution. Here $t(\underline{x},\underline{y}) = -2.512236$.

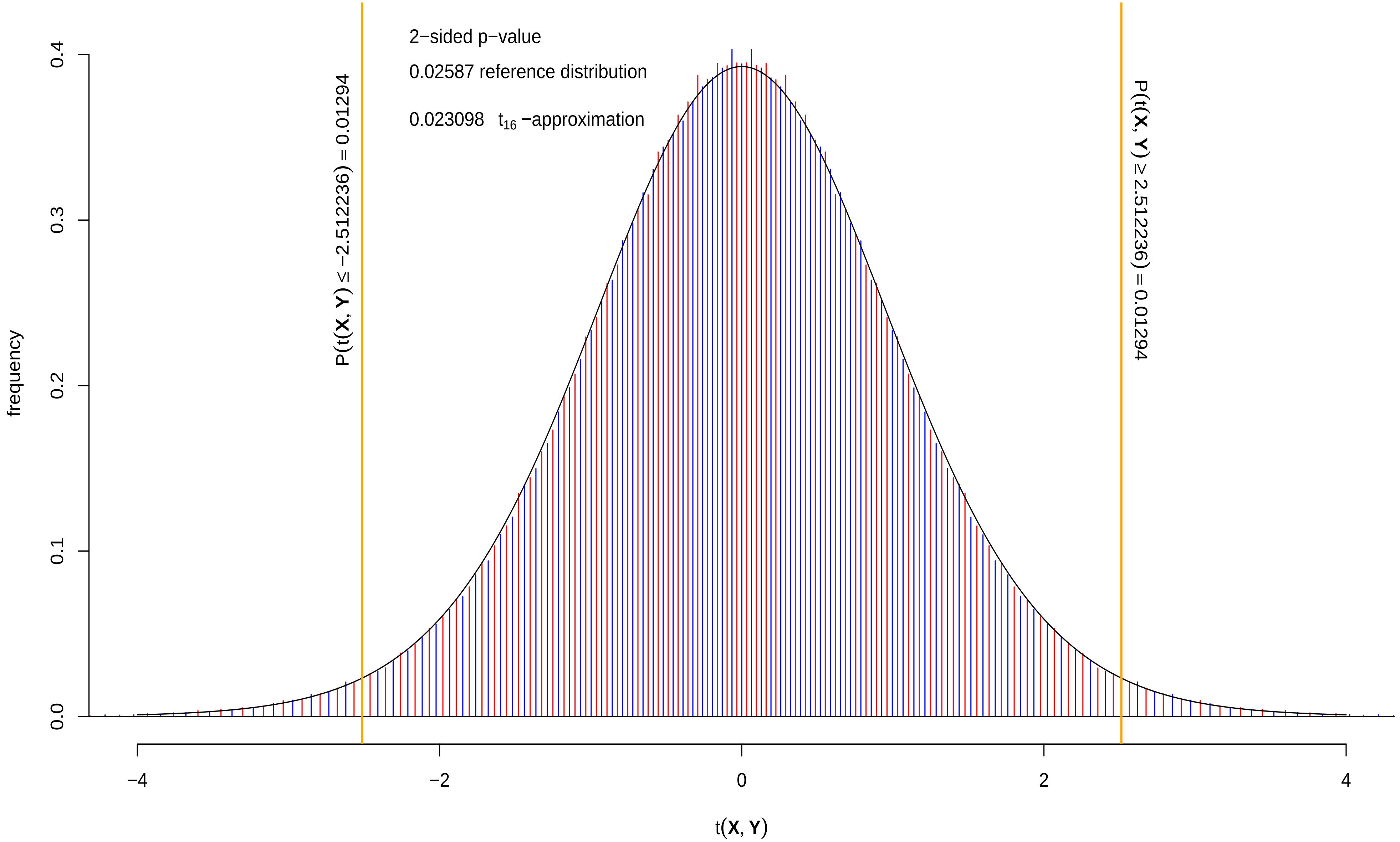
Reference: G.E.P Box and S.L. Anderson, "Permutation theory in the derivation of robust criteria and the study of departures from assumptions,"
J. Roy. Stat. Soc., Ser. B, Vol 17 (1955), pp. 1-34.

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 on this later.

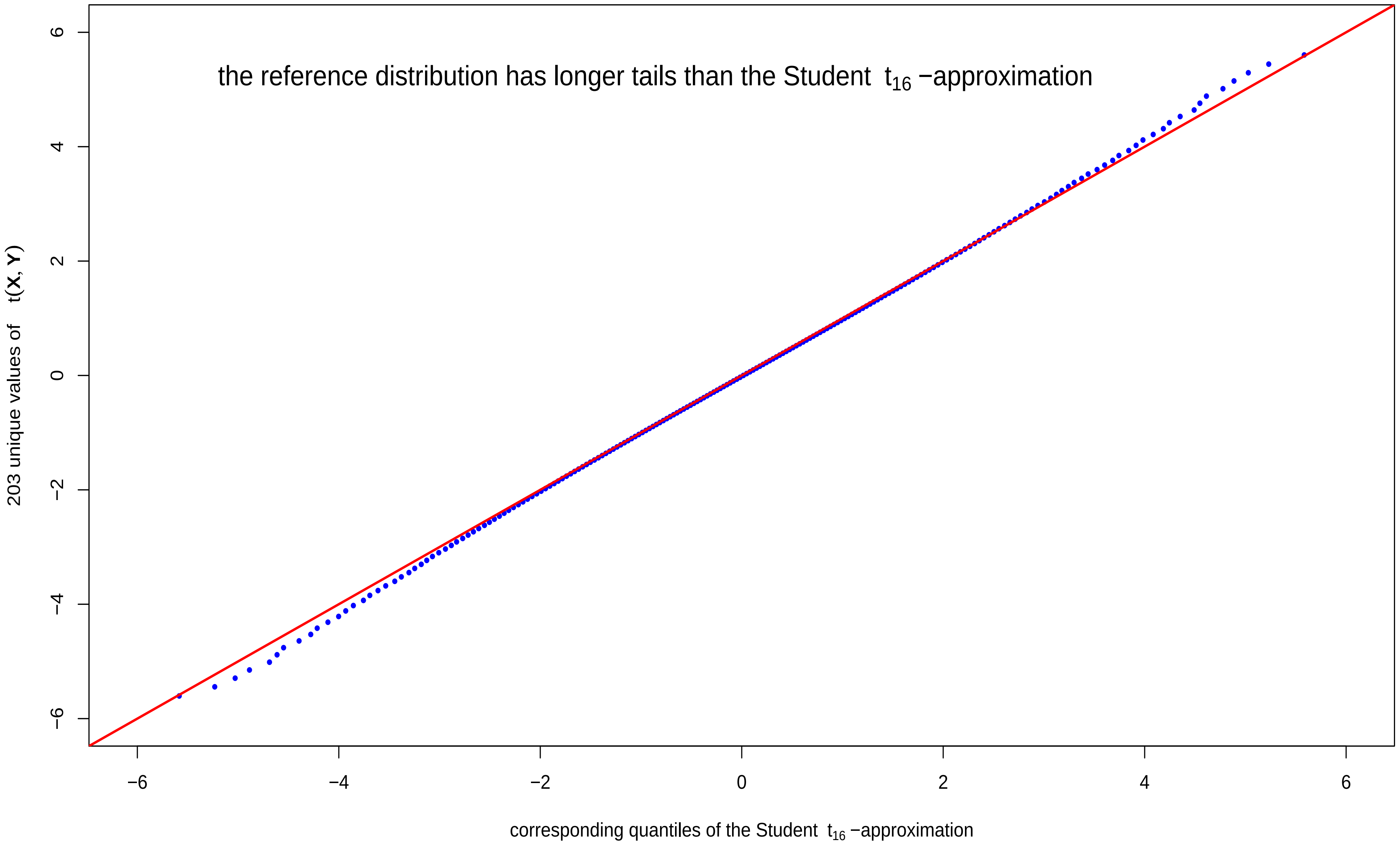
It is the concept of the $t(\underline{X},\underline{Y})$ statistic (not $\bar{Y}-\bar{X}$) that generalizes to more complex experimental situations.

*Our outlier experiment shows what can happen

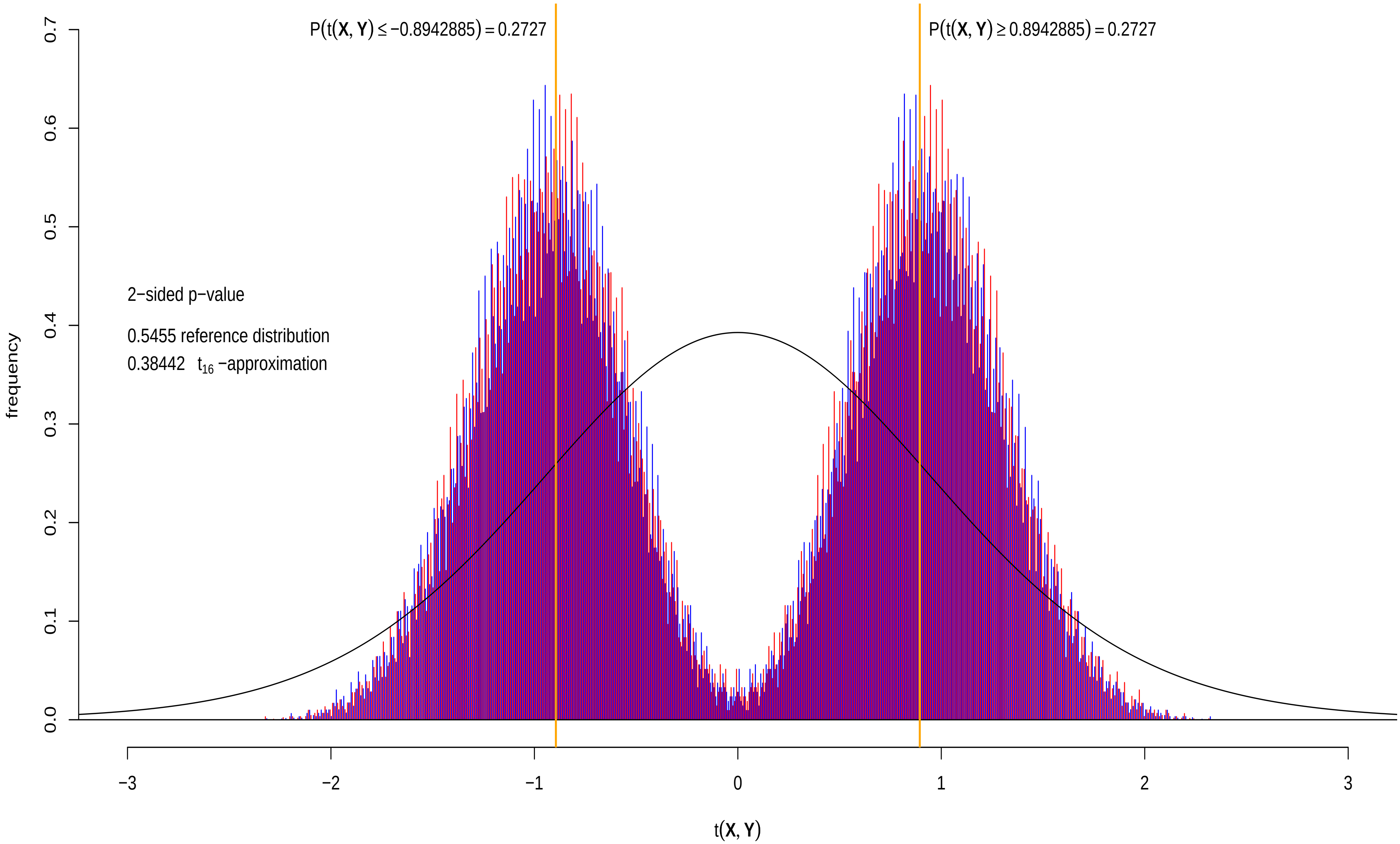
$t(X,Y)$ Randomization Reference Distribution



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



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



The outlier example shows that the t-distribution is not always a good fit.

Critical Value, Significance Level & Type I Error

Any extreme value of $|\bar{Y} - \bar{X}|$ or $|t(\underline{X}, \underline{Y})|$ could either come about due to a rare chance event (via the randomization) 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 $|\bar{Y} - \bar{X}| \geq C$ or $|t(\underline{X}, \underline{Y})| \geq \tilde{C}$, where C or $\tilde{C}$ are respective critical values.

To determine C or $\tilde{C}$ 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

$$\alpha = P(\text{reject } H_0 \mid H_0) = P(|\bar{Y} - \bar{X}| \geq C \mid H_0) = P(|t(\underline{X}, \underline{Y})| \geq \tilde{C} | \, H_0)$$

then determines $C = C_\alpha$ or $\tilde{C} = \tilde{C}_\alpha$. As $\alpha \searrow$ we have $C_\alpha \nearrow$ or $\tilde{C}_\alpha \nearrow$.

Here $P(\ldots | H_0)$ refers to the randomization reference distribution.

Significance Levels and P-Values

When we reject H_0 we would say that the results were significant at the level α, agreed upon prior to the experiment.

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 $\hat{\alpha}$ or p-value of the test for the given data, i.e., for the observed discrepancy value $|\bar{y} - \bar{x}|$

$$\text{p-value} = P(|\bar{Y} - \bar{X}| \geq |\bar{y} - \bar{x}| \mid H_0) = \hat{\alpha} \qquad \text{i.e.,} \quad C_{\hat{\alpha}} = |\bar{y} - \bar{x}|$$

$\hat{\alpha}$ is the smallest α at which we would have rejected H_0 with the observed $|\bar{y} - \bar{x}|$.

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`.

Here `x = randomization.ref.dist` is the reference distribution vector.

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

From the full reference distribution we get $\texttt{C.crit}(\alpha = .05) = 1.28889$ and $\texttt{C.crit}(\alpha = .01) = 1.64444$.

From the simulated reference distribution we get $\texttt{C.crit}(\alpha = .05) = 1.31111$ and $\texttt{C.crit}(\alpha = .01) = 1.66667$.

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

R used to warn you when your chosen object name masks a system object name. This warning no longer seems to happen.

What Does the t-Distribution Give Us?

What does the observed t-statistic $t(\underline{x},\underline{y}) = -2.5122$ give as 2-sided p-value?

$P(|t(\underline{X},\underline{Y})| \geq 2.5122) = 2 * (1 - \texttt{pt}(2.5122, 16)) = 2 * \texttt{pt}(-2.5122, 16) = .02310,$

as compared to the `.02587` 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 $t_{\text{crit}}(\alpha = .05) = \tilde{C}_{.05} = \texttt{qt}(.975, 16) = 2.1199$ and

$t_{\text{crit}}(\alpha = .01) = \tilde{C}_{.01} = \texttt{qt}(.995, 16) = 2.9208$, respectively.

With $|t(\underline{x},\underline{y})| = 2.5122$ we would reject H_0 at $\alpha = .05$ since $|t(\underline{x},\underline{y})| \geq 2.1199$

but not at $\alpha = .01$ since $|t(\underline{x},\underline{y})| < 2.9208$.

The same message derives from the p-value: `.01 < .02310 < .05`.

Comparing P-Values

The randomization reference distribution gave us a p-value of: `.02587`

The normal approximation for the $\bar{Y} - \bar{X}$ reference distribution gave us: `.02828`

The t_{16} approximation for $t(\underline{X}, \underline{Y})$ gave us: `.02310`.

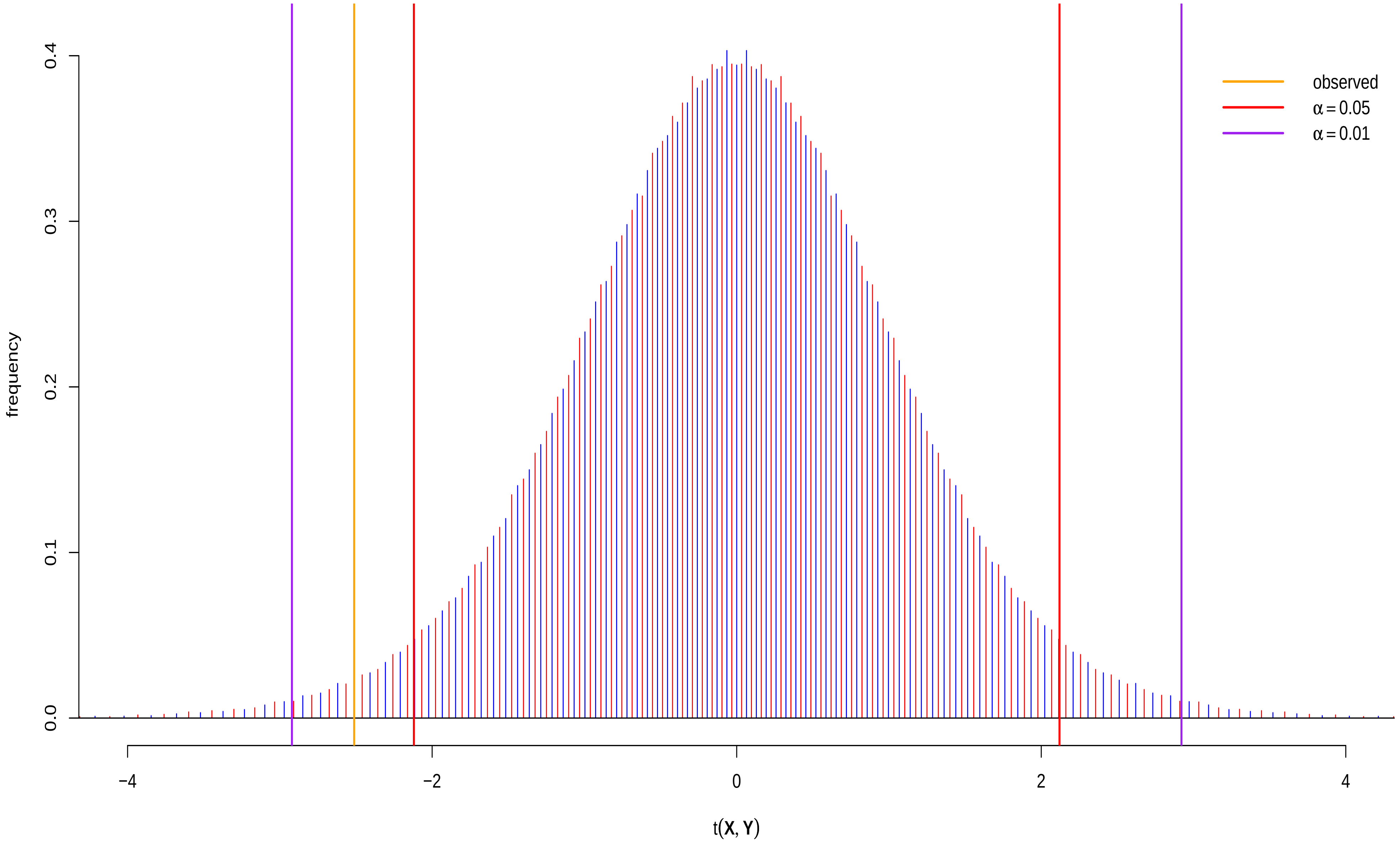
The approximations deviate in opposite directions with roughly the same magnitude `.02828-.02587 = .00241` and `.02310-.02587 = -.00277`.

It is not clear whether the t approximation should be preferred based on approximation quality.

The t approximation involves a single distribution
while the normal approximation still requires $\sigma(\bar{Y} - \bar{X})$.

$t(\underline{X}, \underline{Y})$ will fit better into the larger picture of what lies ahead.

Critical Values of the Randomization Reference Distribution



Some Concluding Comments

Critical values and fixed significance levels were motivated by

1. theoretical considerations that allowed the comparison of different level α tests and to find the best one. Neyman-Pearson optimality theory.

2. the planning of sample sizes to achieve sufficient power to detect treatment effects of specified size.

3. easy tabulation of critical values for a limited set of α values.

4. the fact that computing was not what it is today.

   Statistical tables are following the path of logarithm tables.

Calculator Simulations in 1980

Back in 1980 I asked colleagues Piet Groeneboom and Ron Pyke to look at the large sample distribution of a particular statistic S_n (a very difficult problem).

They succeeded in showing that the asymptotic distribution of S_n should be normal

$$S_n \approx \mathcal{N}\left(n\log(n), 3n^2\log(n)\right) \qquad \text{for large } n.$$

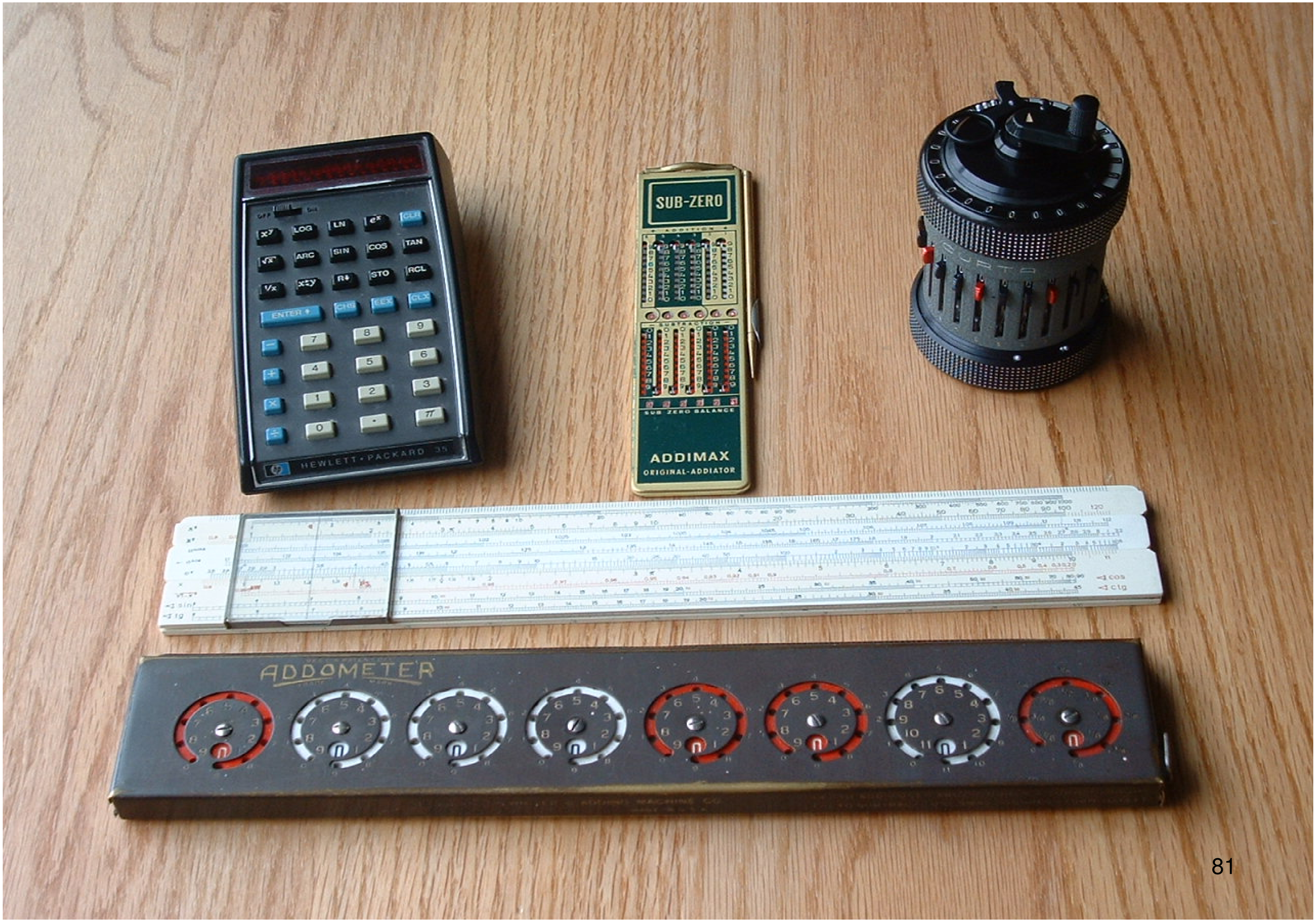
I simulated the distribution for S_n on my programmable calculator HP-41 CV.

It ran for days, weeks, and the intermediate results did not look normal at all.

Since it was known that $S_n \geq 0$, a natural heuristic for decent normality would be

$$0 \leq n\log(n) - 3 \times \sqrt{3n^2\log(n)} \quad \Longrightarrow \quad \log(n) \geq 27 \text{ or } n \geq 5.3 \times 10^{11} \, .$$

Calculators in My Time



Appendix A: Sampling Theory Proofs

The following two slides establish the mean and variance of $\bar{Y}$ when $Y_1,\ldots,Y_n$ are randomly drawn without replacement from $Z_1,\ldots,Z_N$ with $N = m + n$.

$$E(\bar{Y}) = \bar{Z}$$

and

$$\mathrm{var}(\bar{Y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) \quad \text{with} \quad S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Z_i - \bar{Z})^2$$

Note the evident correctness when $n = N$ $(m = 0)$.

Sampling Theory

Write $\bar{Y} = \sum_{i=1}^{N} I_i Z_i / n$, where the random variable $I_i = 1$ whenever the Y-sample contains Z_i and $I_i = 0$ otherwise. Then

$$E(I_i) = P(I_i = 1) = \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{(N-1)!n!(N-n)!}{N!(n-1)!(N-n)!} = \frac{n}{N} \,.$$

and for $i \neq j$

$$E(I_i I_j) = P(I_i = 1, I_j = 1) = \frac{\binom{2}{2}\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{(N-2)!n!(N-n)!}{N!(n-2)!(N-n)!} = \frac{n(n-1)}{N(N-1)} \,.$$

The $Z_1, \ldots, Z_N$ are the finite population from which we sample.

$$\Longrightarrow \quad E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{N} Z_i \, E(I_i) = \frac{1}{n}\sum_{i=1}^{N} Z_i \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} Z_i = \bar{Z} \,.$$

Further note

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{N} I_i Z_i = \frac{1}{n}\sum_{i=1}^{N} I_i (Z_i - \bar{Z}) + \bar{Z} \quad \text{since} \quad \sum_{i=1}^{N} I_i = n \,.$$

$\text{var}(\bar{Y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$

Writing $z_i = Z_i - \bar{Z}$ with $\sum_{i=1}^N z_i = 0$ and dropping the constant $\bar{Z}$ in the variance

$$\Longrightarrow \ \text{var}(\bar{Y}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^N I_i z_i\right) = \text{cov}\left(\frac{1}{n}\sum_{i=1}^N I_i z_i, \frac{1}{n}\sum_{j=1}^N I_j z_j\right) = \frac{1}{n^2}\sum_{i=1}^N \sum_{j=1}^N z_i z_j \text{cov}(I_i, I_j)$$

For $i = j$ we have $\text{cov}(I_i, I_j) = \text{var}(I_i) = (n/N)(1 - n/N) = nm/N^2$

and for $i \neq j$ we have

$$\text{cov}(I_i, I_j) = E(I_i I_j) - E(I_i)E(I_j) = P(I_i = 1, I_j = 1) - \frac{n}{N}\frac{n}{N} = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2}$$

$$= \frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right) = \frac{n}{N}\frac{N(n-1) - (N-1)n}{N(N-1)} = -\frac{n}{N}\frac{N-n}{N(N-1)} = -\frac{nm}{N^2(N-1)}$$

$$\text{var}(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^N z_i^2 \frac{nm}{N^2} - \frac{1}{n^2}\sum_{i \neq j} z_i z_j \frac{nm}{N^2(N-1)} = S^2\left(\frac{nm(N-1)}{n^2 N^2} + \frac{nm}{n^2 N^2}\right) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$$

since $\quad 0 = \sum_i z_i \sum_j z_j = \sum_i \sum_j z_i z_j = \sum_i z_i^2 + \sum_{i \neq j} z_i z_j, \quad$ i.e., $\quad \sum_{i \neq j} z_i z_j = -\sum_i z_i^2.$

Appendix B: Monotone Equivalence of $\bar{Y}-\bar{X}$ and $t(\underline{X},\underline{Y})$

The values of $t(\underline{X},\underline{Y})$ are in 1-1 monotone increasing correspondence with those of $\bar{Y}-\bar{X}$ under all splits.

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

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

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

with $W=(\bar{Y}-\bar{X})/\sqrt{\sum(Z_k-\bar{Z})^2}$, where $\sum(Z_k-\bar{Z})^2$ is constant over all splits.

$$\Longrightarrow\quad t(\underline{X},\underline{Y})=\frac{\sqrt{m+n-2}}{\sqrt{1/n+1/m}}\frac{W}{\sqrt{1-\frac{mn}{N}W^2}}\qquad\text{or}\qquad W=\frac{t(\underline{X},\underline{Y})}{\sqrt{\frac{m+n-2}{1/n+1/m}+\frac{mn}{N}t(\underline{X},\underline{Y})^2}}$$