CH4rLIE
Meeting

SOME COMMENTS ABOUT THE DATA

17/07/25 P.Vitulo

File → Analisi CH4 Donalisio def.xlsx

7417×7 table

	Time	Point	CH4	TemperaturaMediaC
First date	05-Sep-2023 10:42:55	P1	2.3611	25.185
	05-Sep-2023 10:44:17	P1	2.4192	25.156
	05-Sep-2023 10:45:38	P1	3.8303	25.099
	05-Sep-2023 10:47:01	P1	4.8497	25.07
	05-Sep-2023 10:48:24	P1	5.359	25.078
	:	:	:	:
Last date	17-Apr-2024 00:52:14	P5	6.8939	NaN
	17-Apr-2024 00:53:35	P5	6.518	NaN
	17-Apr-2024 00:54:58	P5	7.0015	NaN
	17-Apr-2024 00:56:20	P5	5.7064	NaN
	17-Apr-2024 00:57:42	P5	7.9563	NaN

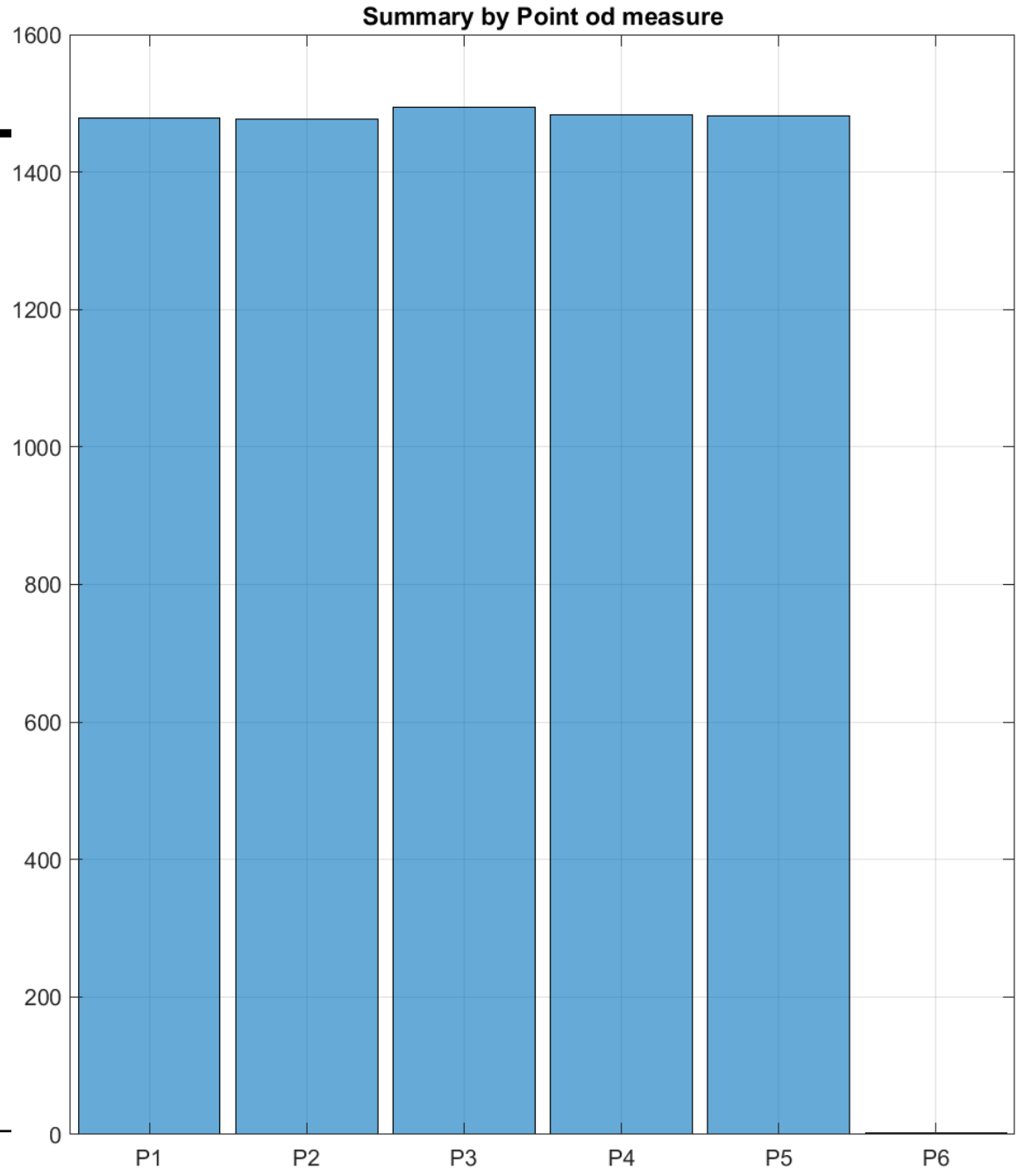
Display all 7417 rows.

- Taken over 24 hours
- Data rate: about 1 ½ minute (except when switching from P_i to P_{i+1} → about 3 min)
- Sequential P1 P2 P3 P4 P5
- 9 measurements/Point (repeated over the day) :

File → Analisi CH4 Donalisio def.xlsx

HOW THE DATA LOOK LIKE: SUMMARY BY POINTS

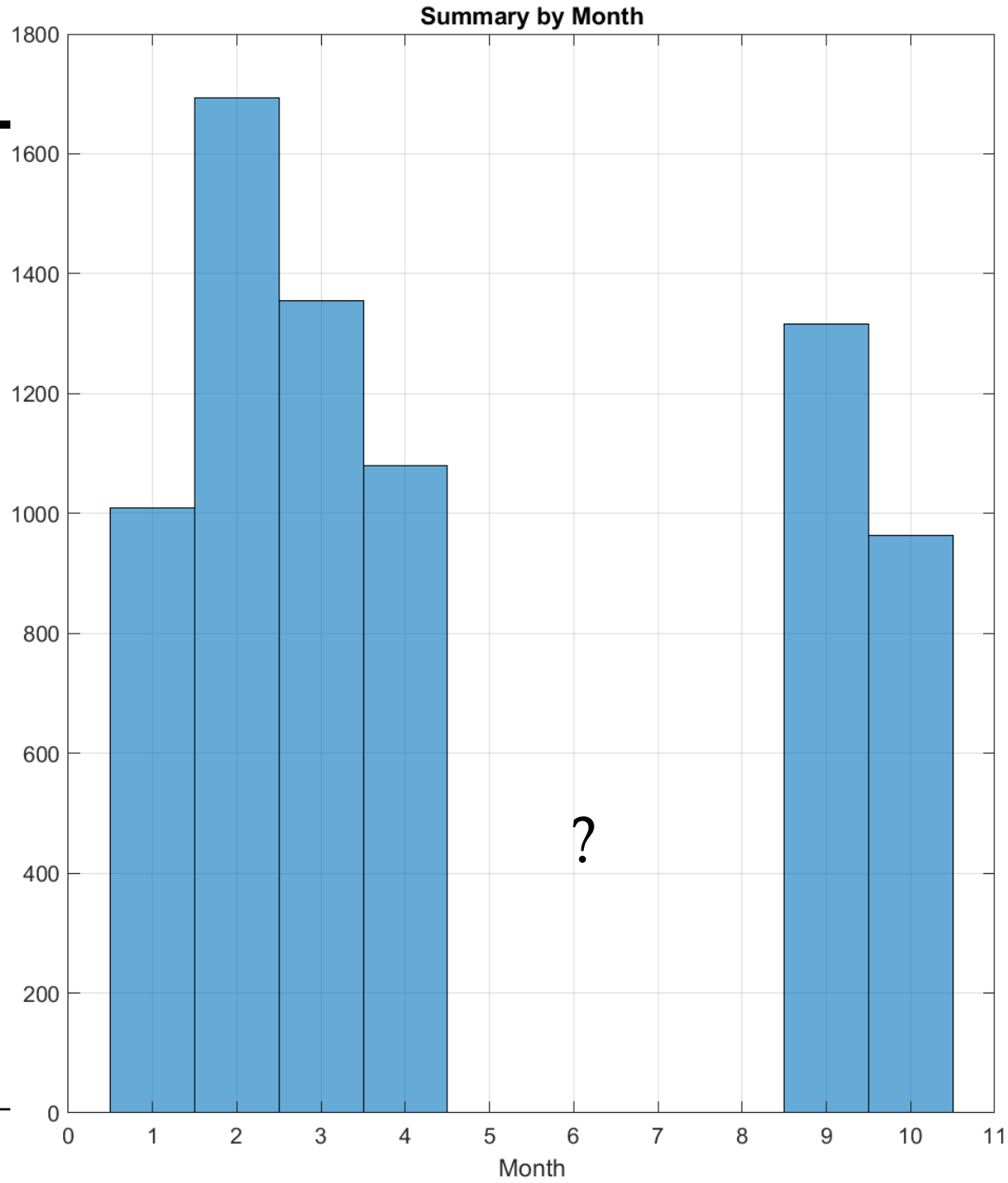
Point of measure	Total counts
P1	1478
P2	1477
P3	1494
P4	1484
P5	1482
P6	2
Total	7417



P6 REFERENCE (OUTSIDE ?)

HOW THE DATA LOOK LIKE: SUMMARY BY MONTH

Month	Total counts
Jan	1010
Feb	1693
Mar	1355
Apr	1080
Sep	1316
Oct	963
Total	**7417**

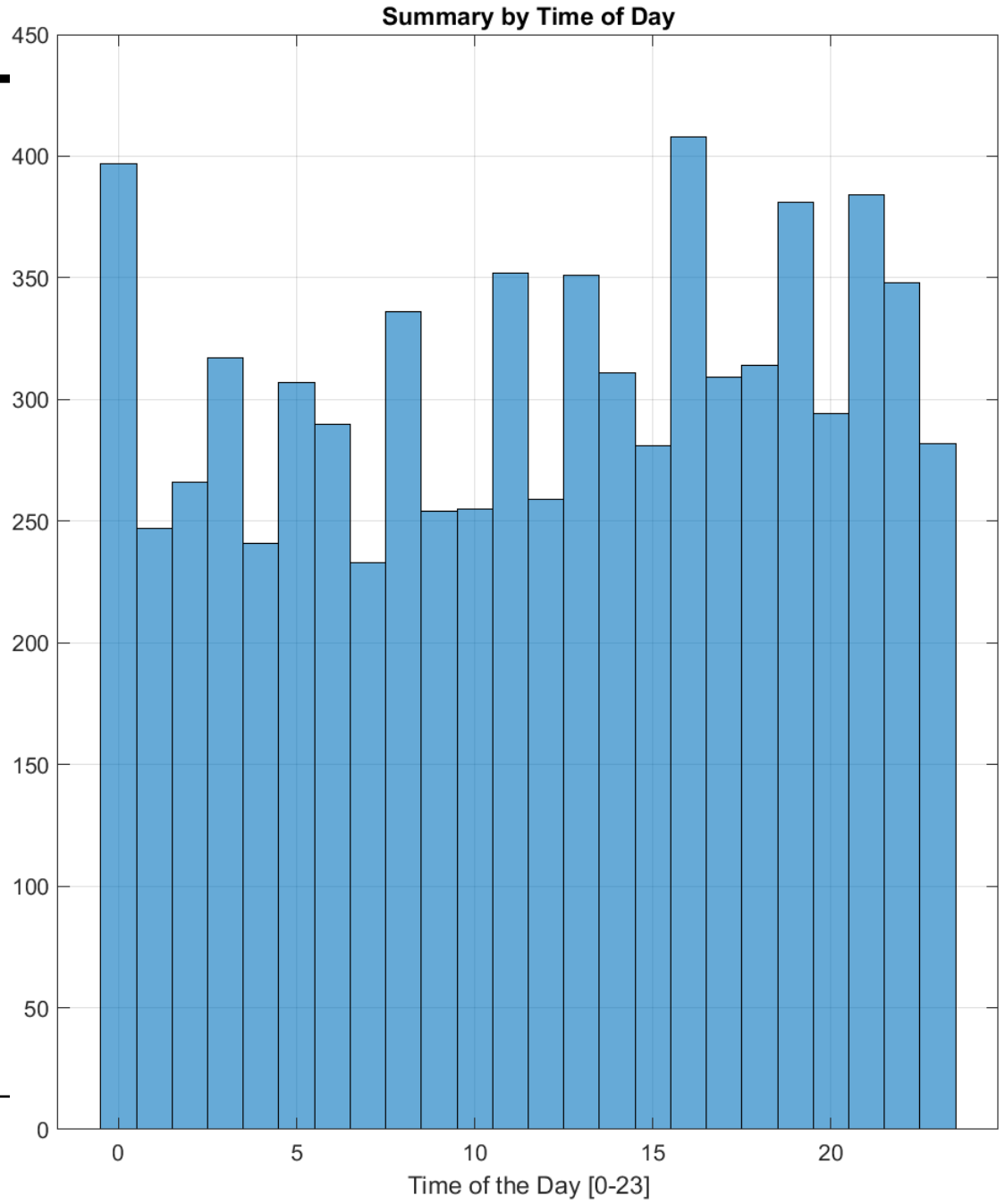


SINCE WE HAVE MEASUREMENTS EVERY MINUTE AND HALF, WE NEED TO DECIDE HOW MUCH TO AGGREGATE TO EVALUEATE SOME VARIABLES AND DEDUCE SOME INFORMATION. WE CHOSE TO AGGREGATE DATA HOURLY, BUT THIS CHOICE IS ONLY DICTATED BY THE NEED TO HAVE REASONABLE PLOTS AND STATISTICS.

0	397
1	247
2	266
3	317
4	241
5	307
6	290
7	233
8	336
9	254
10	255
11	352
12	259
13	351
14	311
15	281
16	408
17	309
18	314
19	381
20	294
21	384
22	348
23	282

HOW THE DATA LOOK LIKE: SUMMARY BY TIME OF DAY

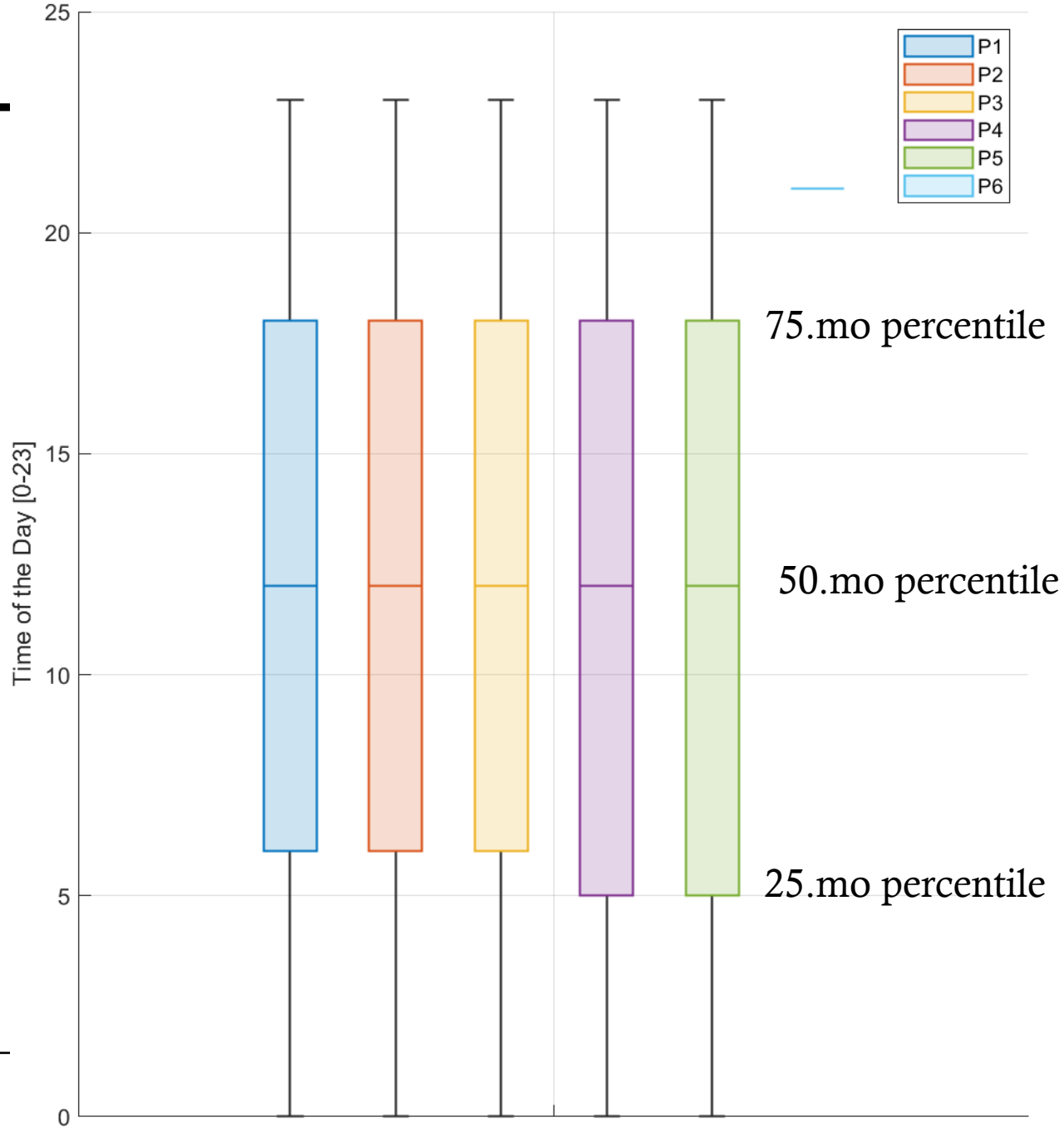
> 240 measures/ Hour of the Day



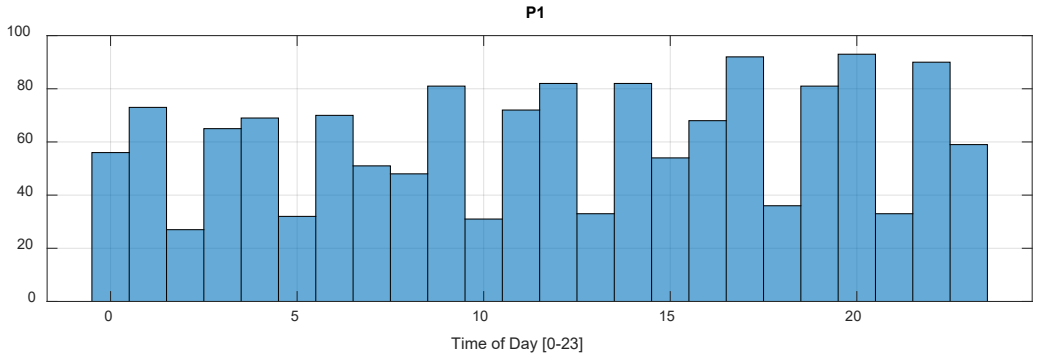
HOW THE DATA LOOK LIKE:
NO BIAS IN THE HOURLY MEASUREMENTS AMONG POINTS OF MEASURE

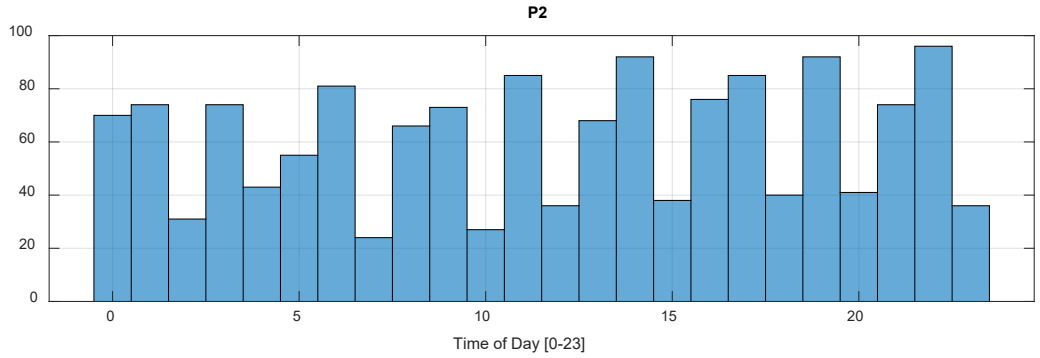
i.e. ALL THE HOURLY MEASUREMENTS HAVE ABOUT THE SAME DISTRIBUTION IN ALL THE POINTS.

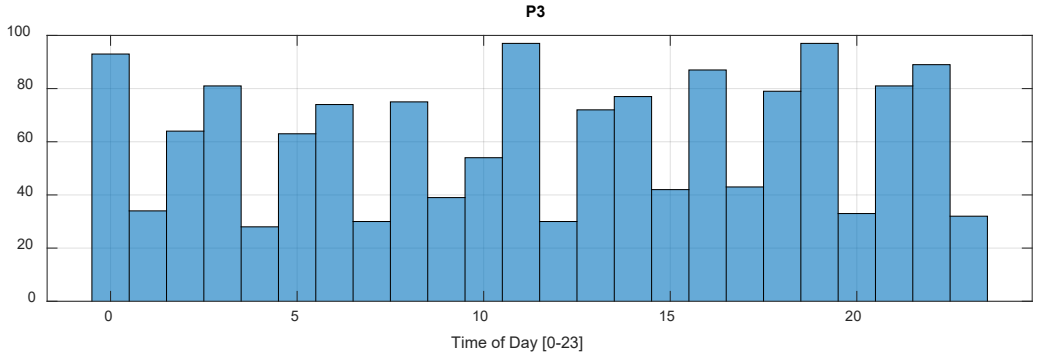
- MEDIAN OF DIST. → 12:00
- 50% OF THE DATA ARE WITHIN 5:00 AND 18:00

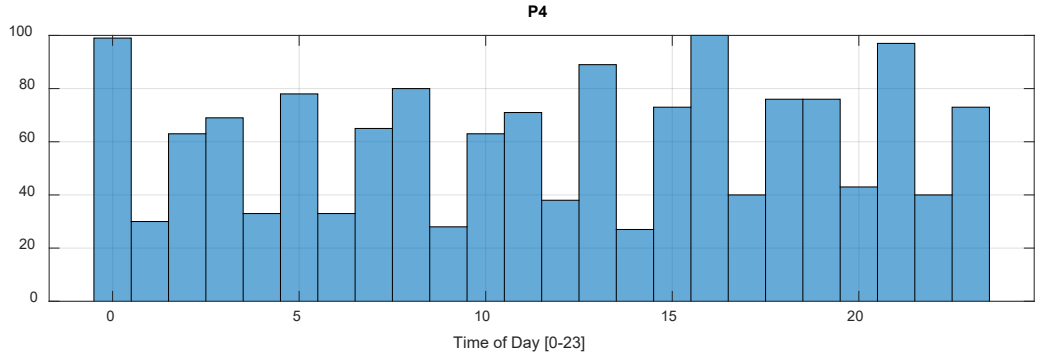


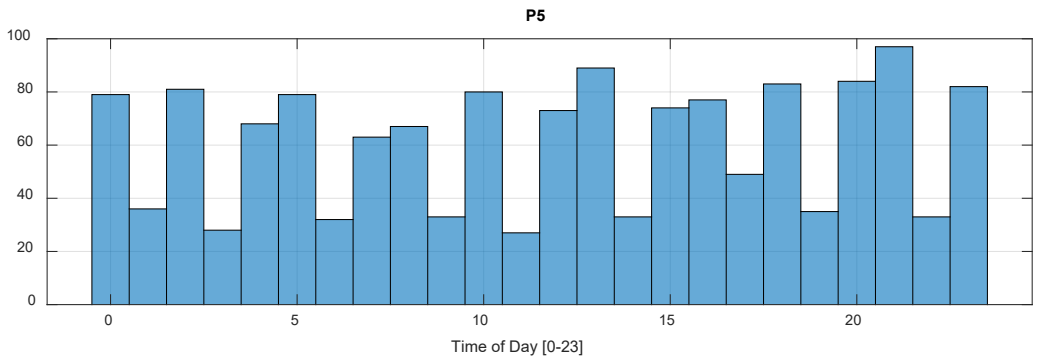
DISTRIBUTION OF THE HOURLY MEASUREMENTS AT THE POINTS OF MEASURE





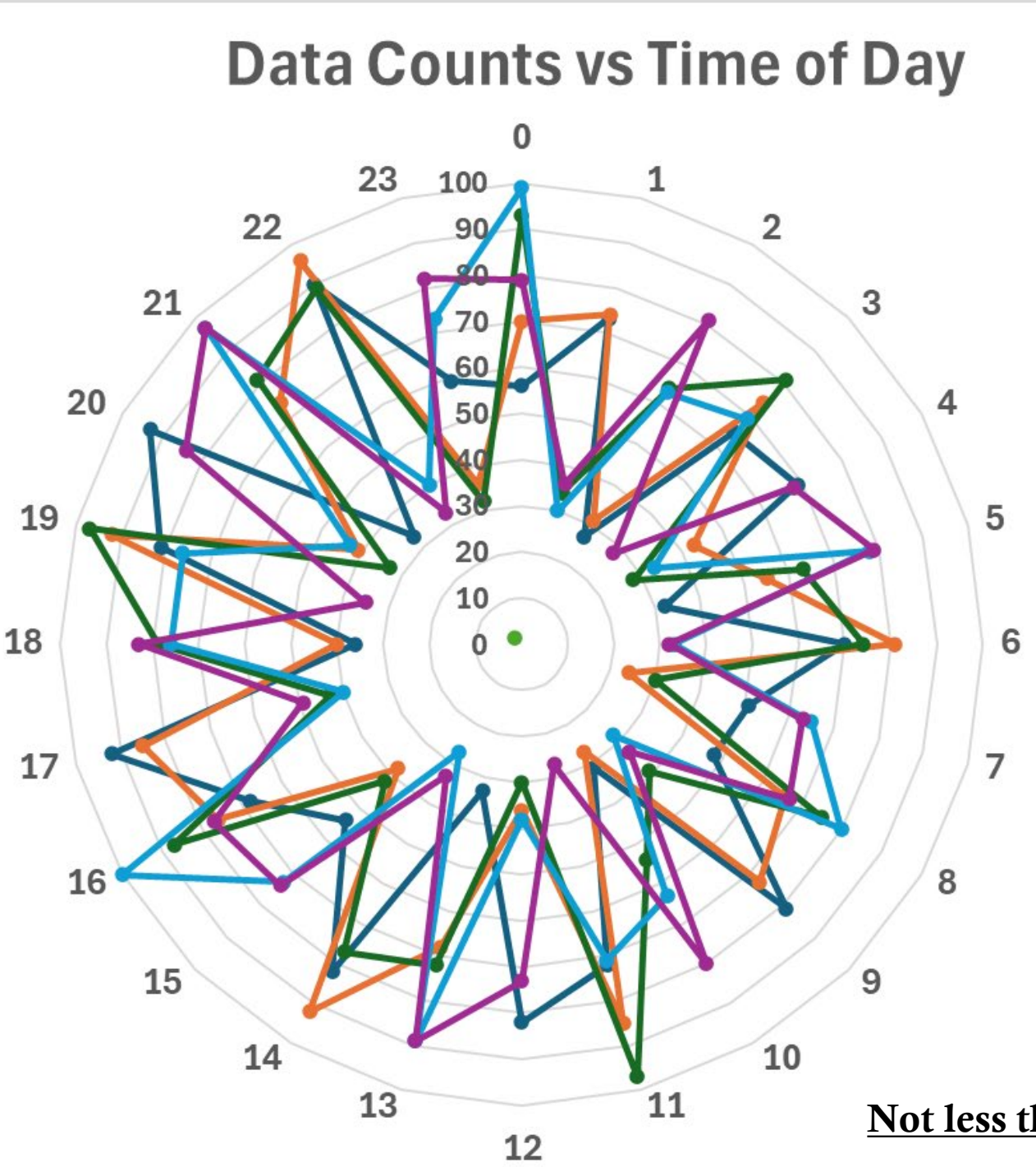




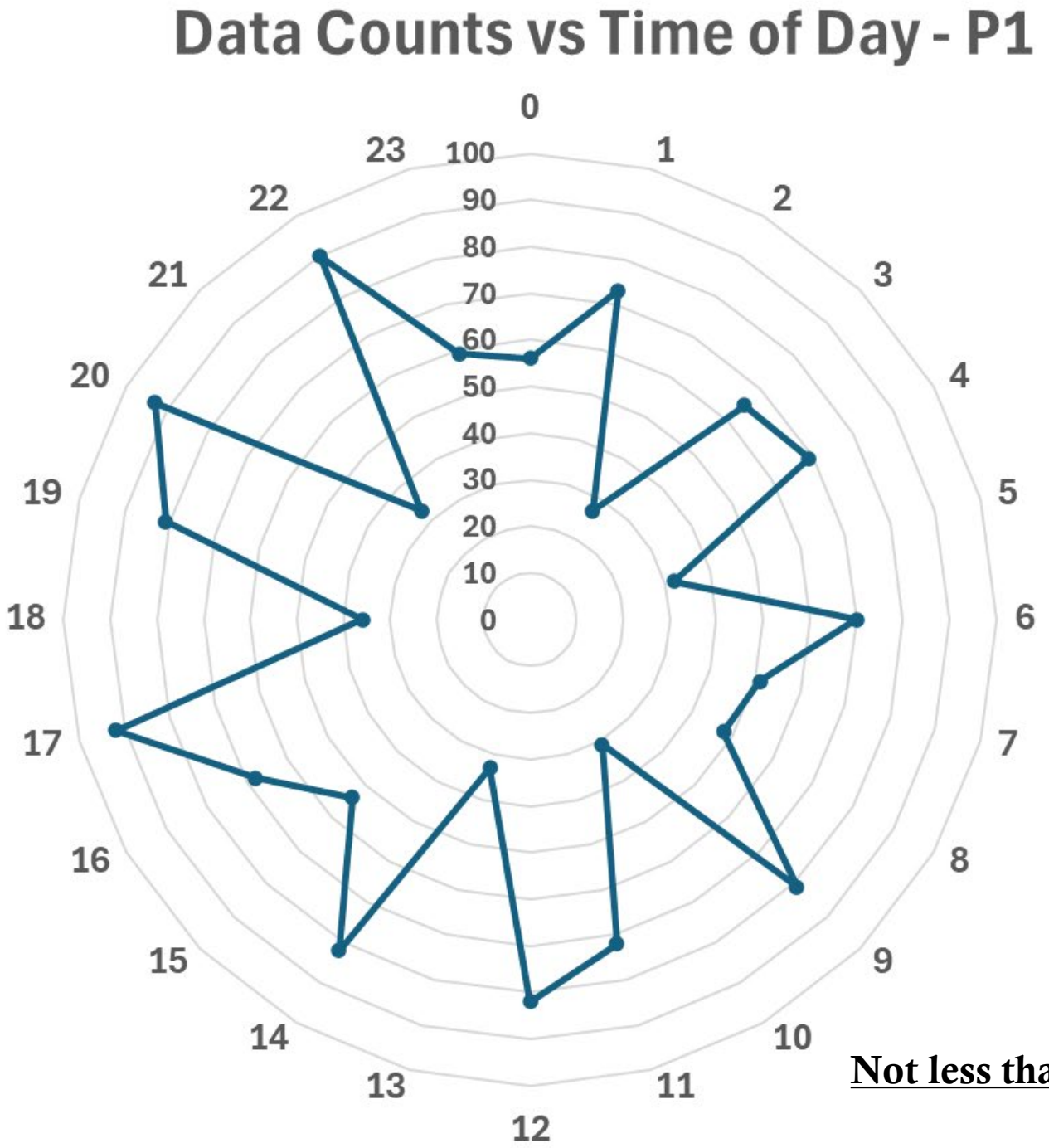


…however the statistics of the counts in different hours is not uniform (but overall not less than 20 measurements in each bin). Obviously this conclusion depends on the preceding aggregation choice…

HOW THE DATA LOOK LIKE: SUMMARY BY TIME OF DAY VS POINTS

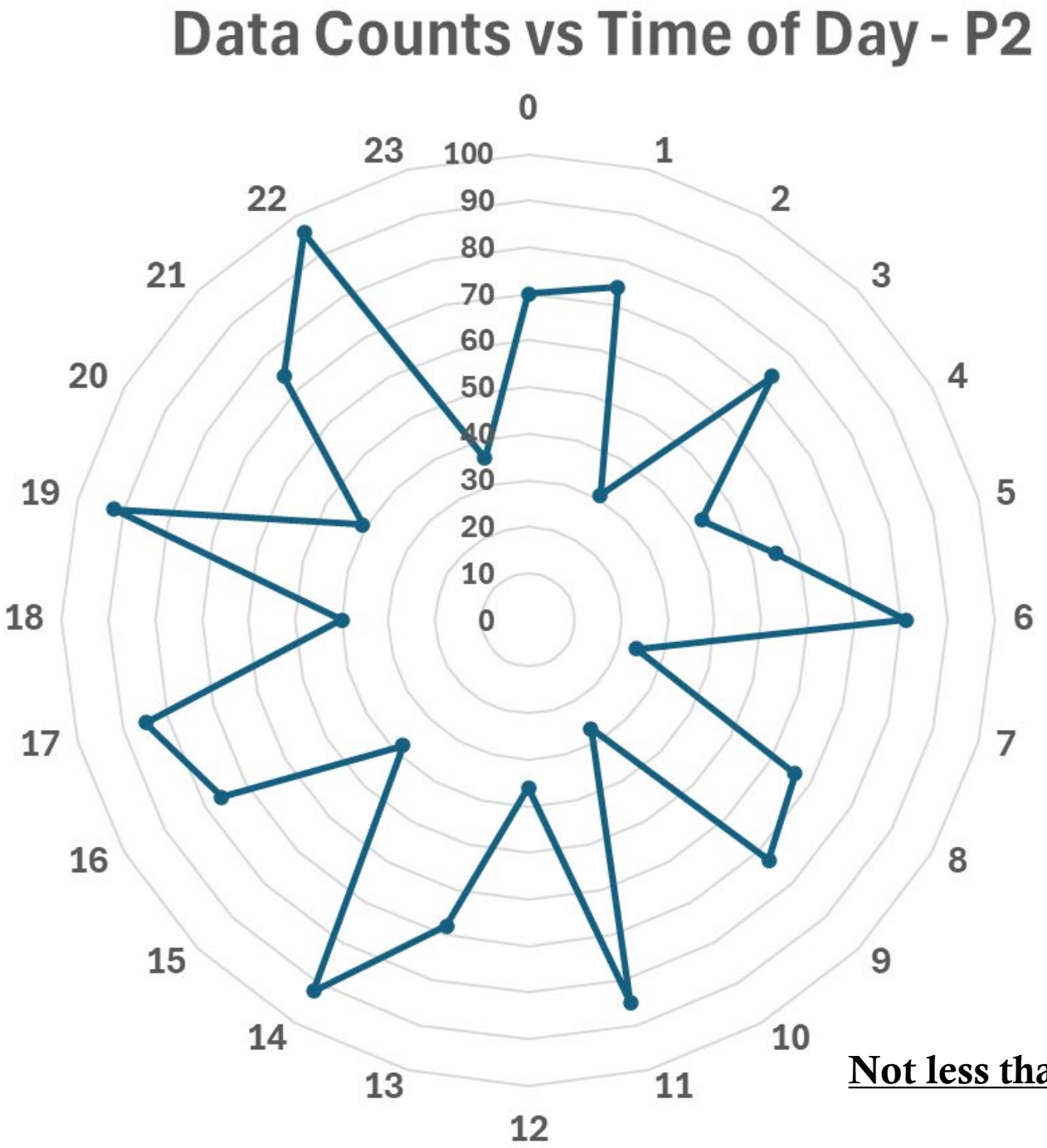


Not less than 20÷ 30 hourly measurements/point



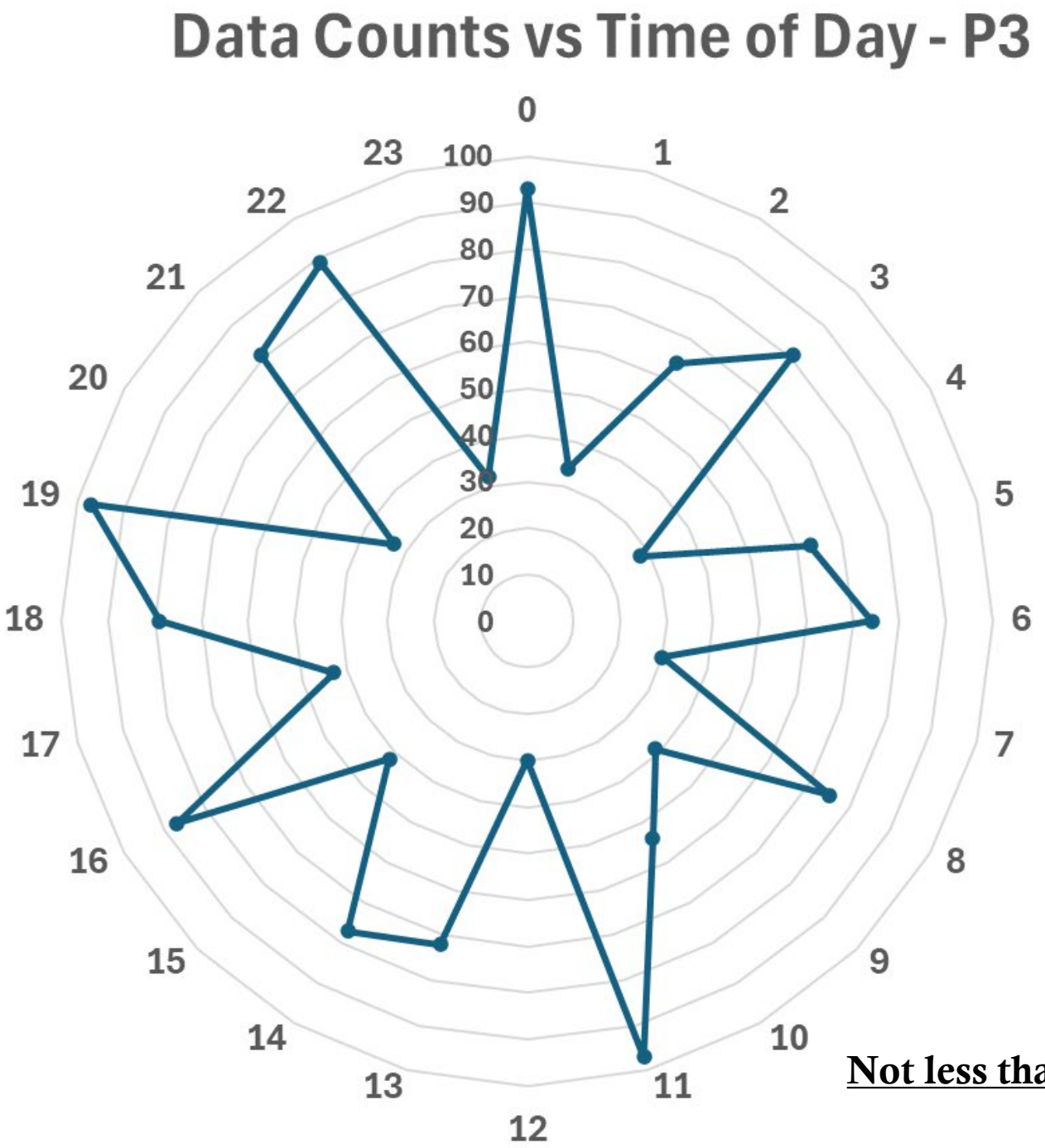
HOW THE DATA LOOK LIKE: SUMMARY BY TIME OF DAY VS POINTS

Not less than about 30 hourly measurements/point



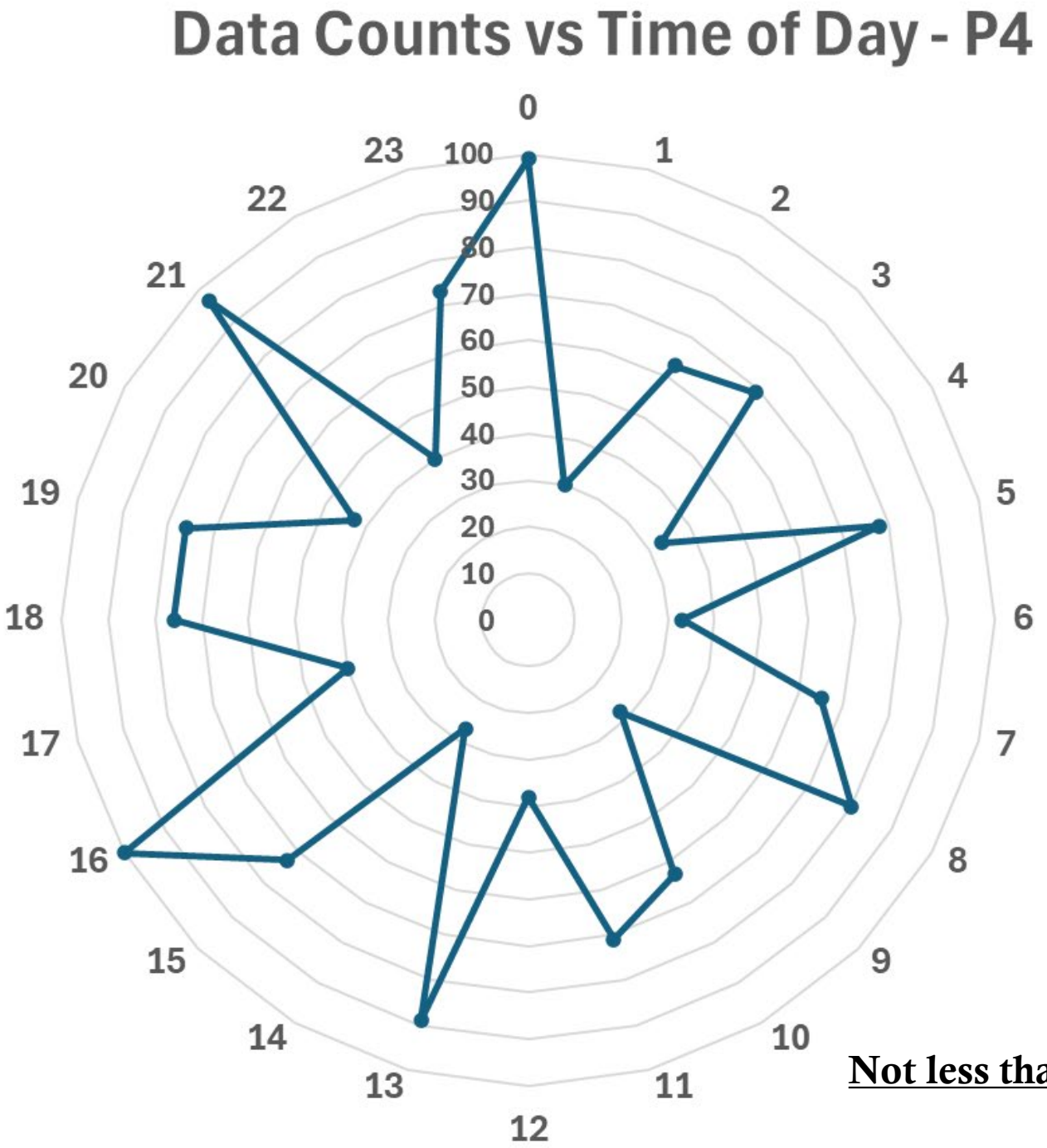
HOW THE DATA LOOK LIKE: SUMMARY BY TIME OF DAY VS POINTS

Not less than about 30 hourly measurements/point



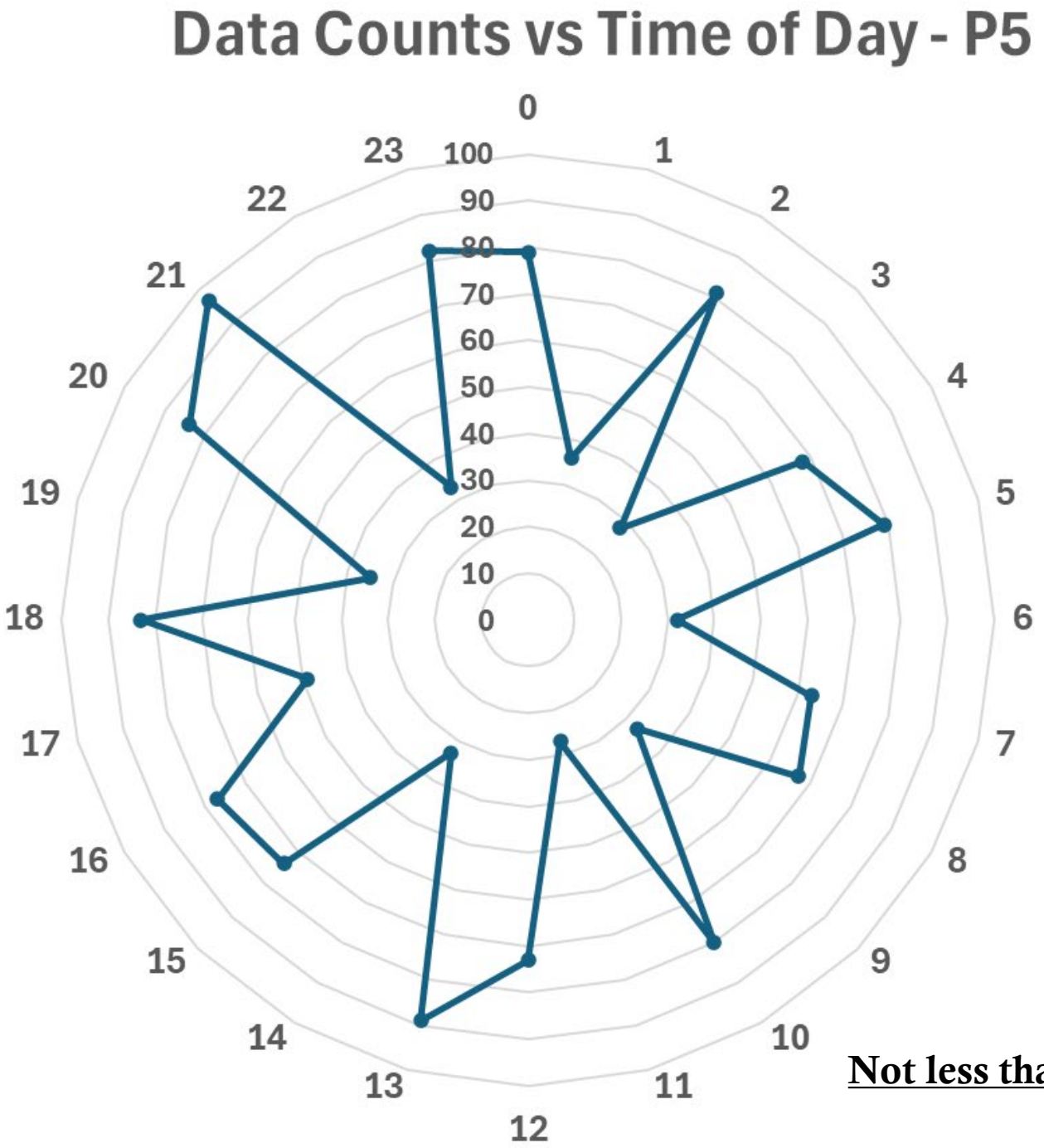
HOW THE DATA LOOK LIKE: SUMMARY BY TIME OF DAY VS POINTS

Not less than about 30 hourly measurements/point



HOW THE DATA LOOK LIKE: SUMMARY BY TIME OF DAY VS POINTS

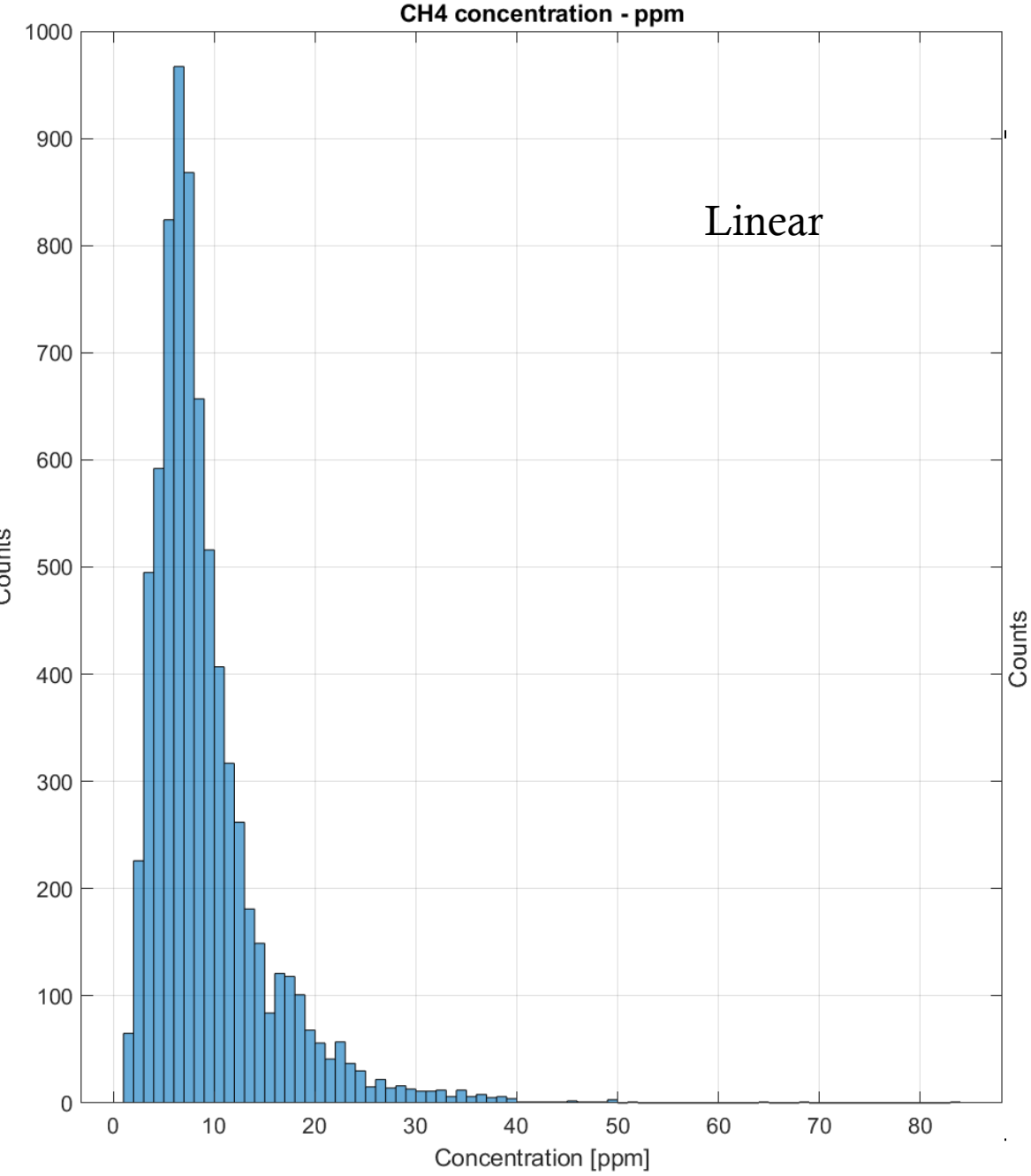
Not less than about 30 hourly measurements/point

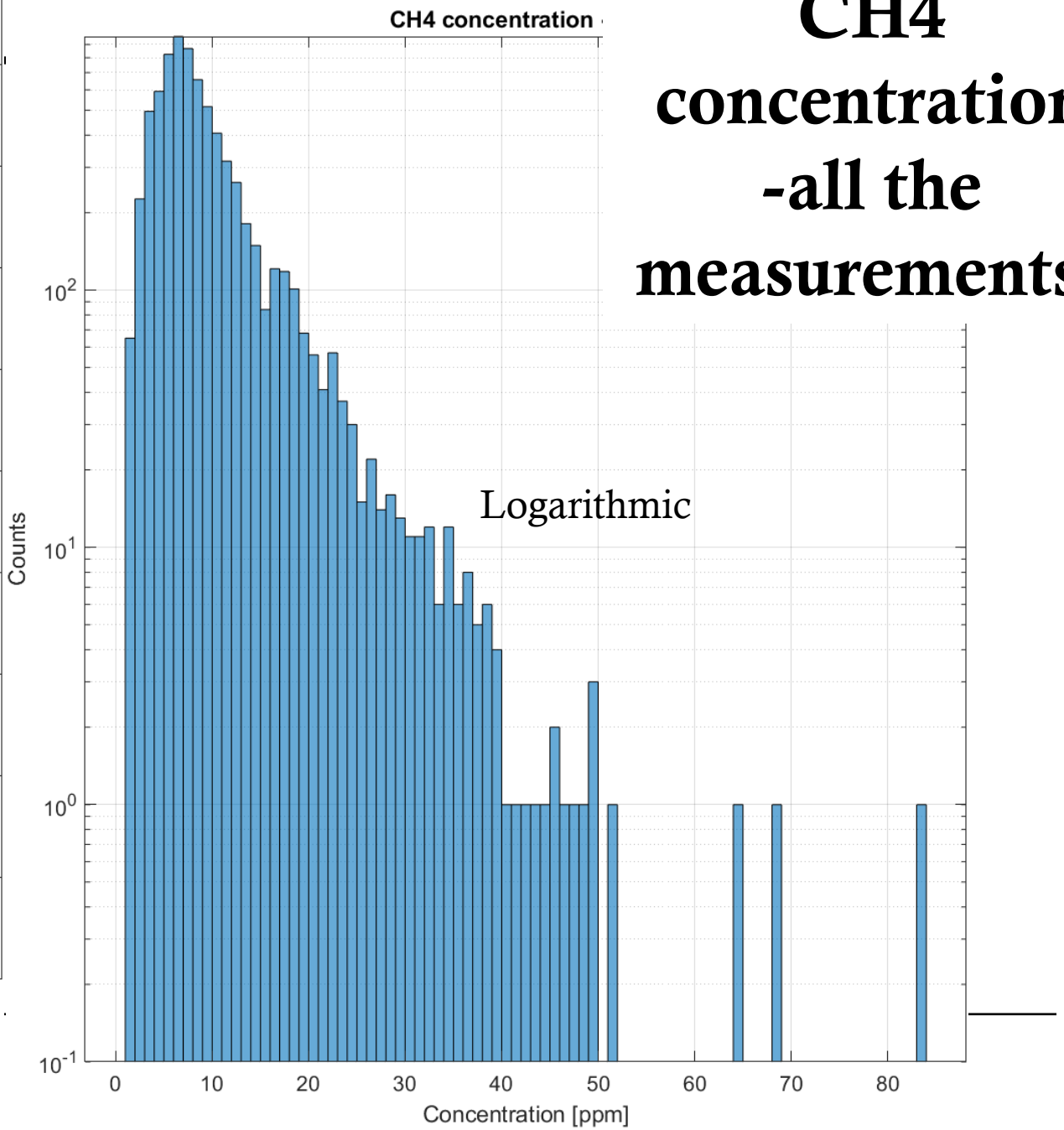


HOW THE DATA LOOK LIKE: SUMMARY BY TIME OF DAY VS POINTS

Not less than about 30 hourly measurements/point

CH4 concentration -all the measurements-



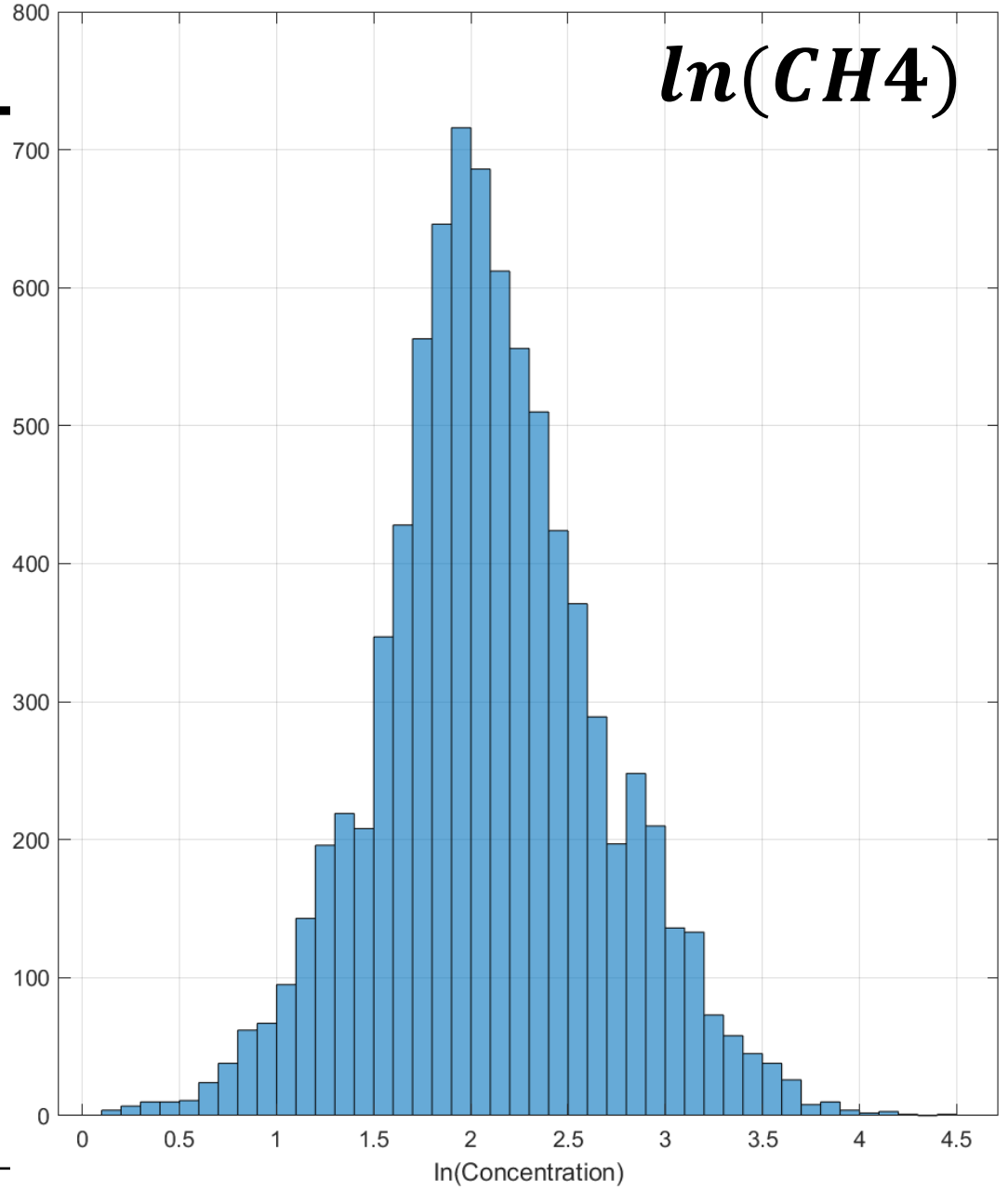


Raw Data concentration is not normal distributed

But…

The logarithm of the concentration is distributed as (about) a gaussian.
This means that the raw concentration is distributed as (about) a log-Normal …

A concentration log-normal of methane indicates that the underline physics variables are the result of multiplicative effects : emission x diffusion x …
And not the results of a summation effect



To compare different measurements we need to standardize/normalize the values then aggregate data.

Take the logarithm of concentration X and standardize it:

$$Z = (\ln(X) - \mu\ )/\sigma$$

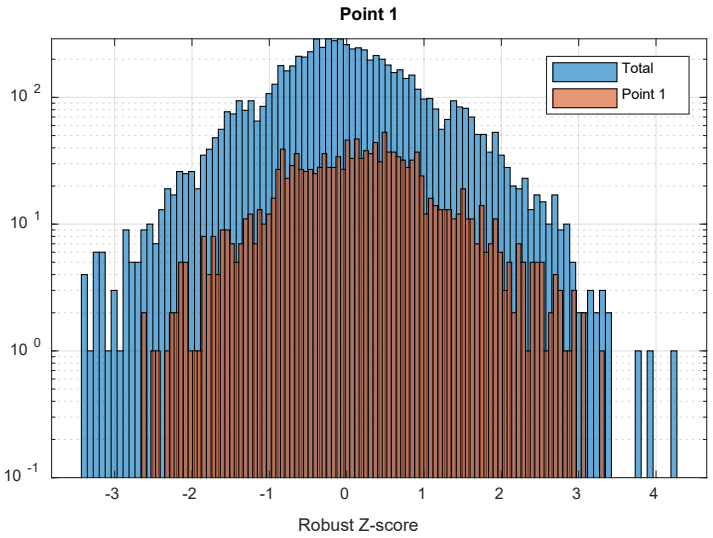
μ and σ are average and standard deviation of the distribution of $\ln(X)$

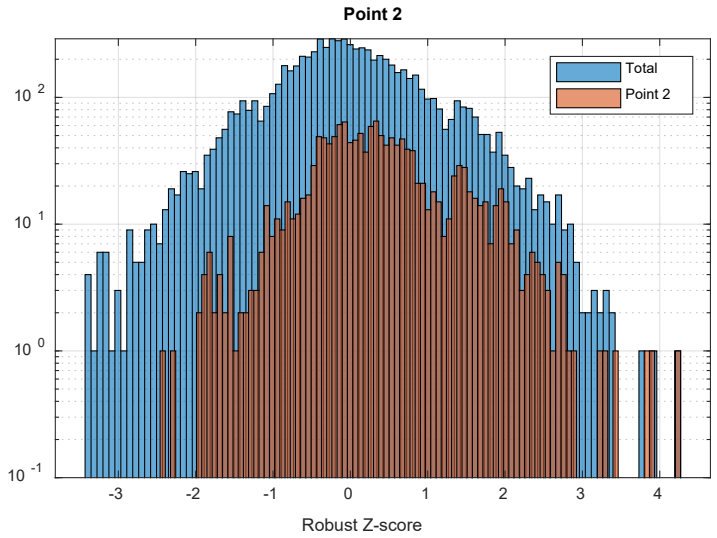
From now on (unless otherwise specified) Z-score (or Robust Z-score) refers to that formula

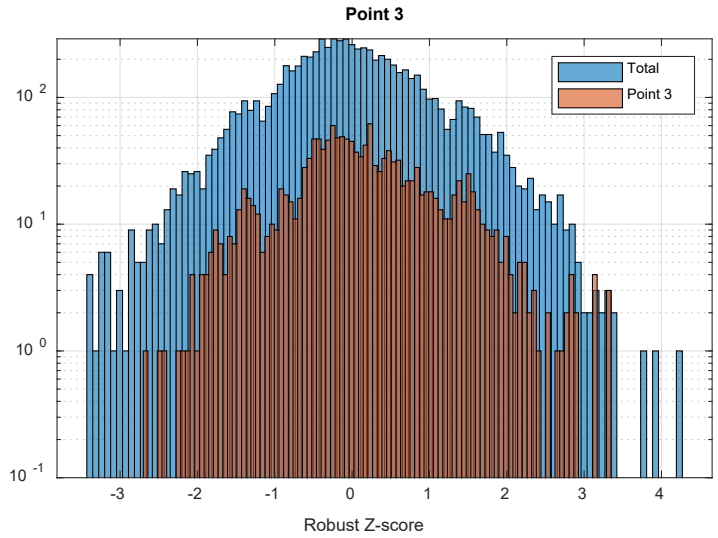
From Z-score to ppm: $ppm = e^{\mu + Z\sigma}$ ($\mu = 2.1\ \sigma = 0.6$)

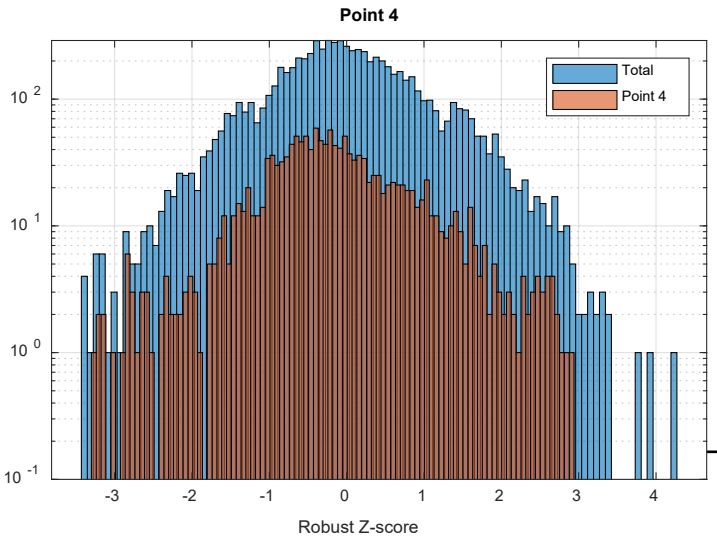
(i.e. $Z = -1 \rightarrow ppm = e^{\mu + (-1)\sigma} = 4.5\ ppm$)

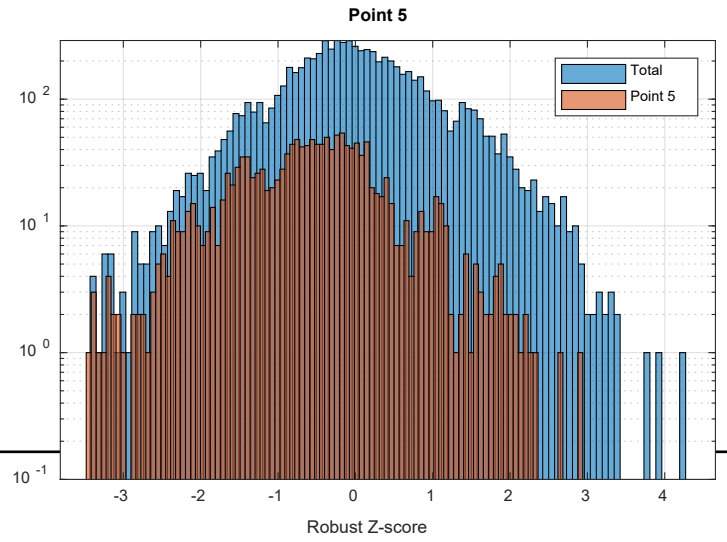
Z-SCORE spectra in different Points (compared to total)

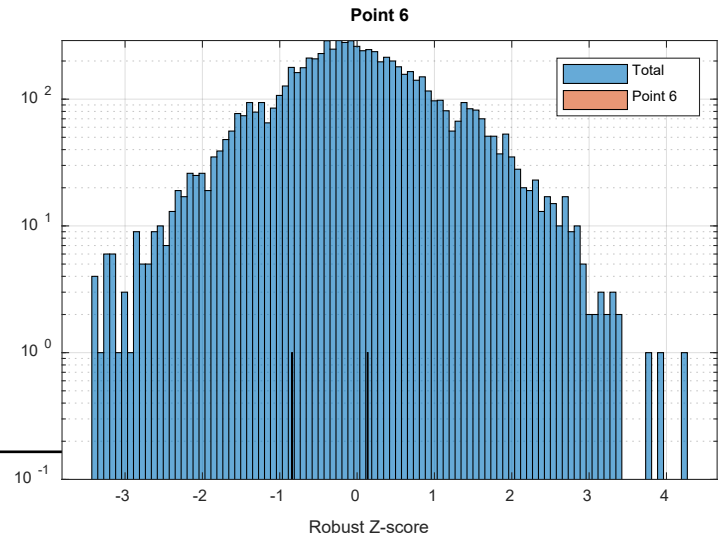


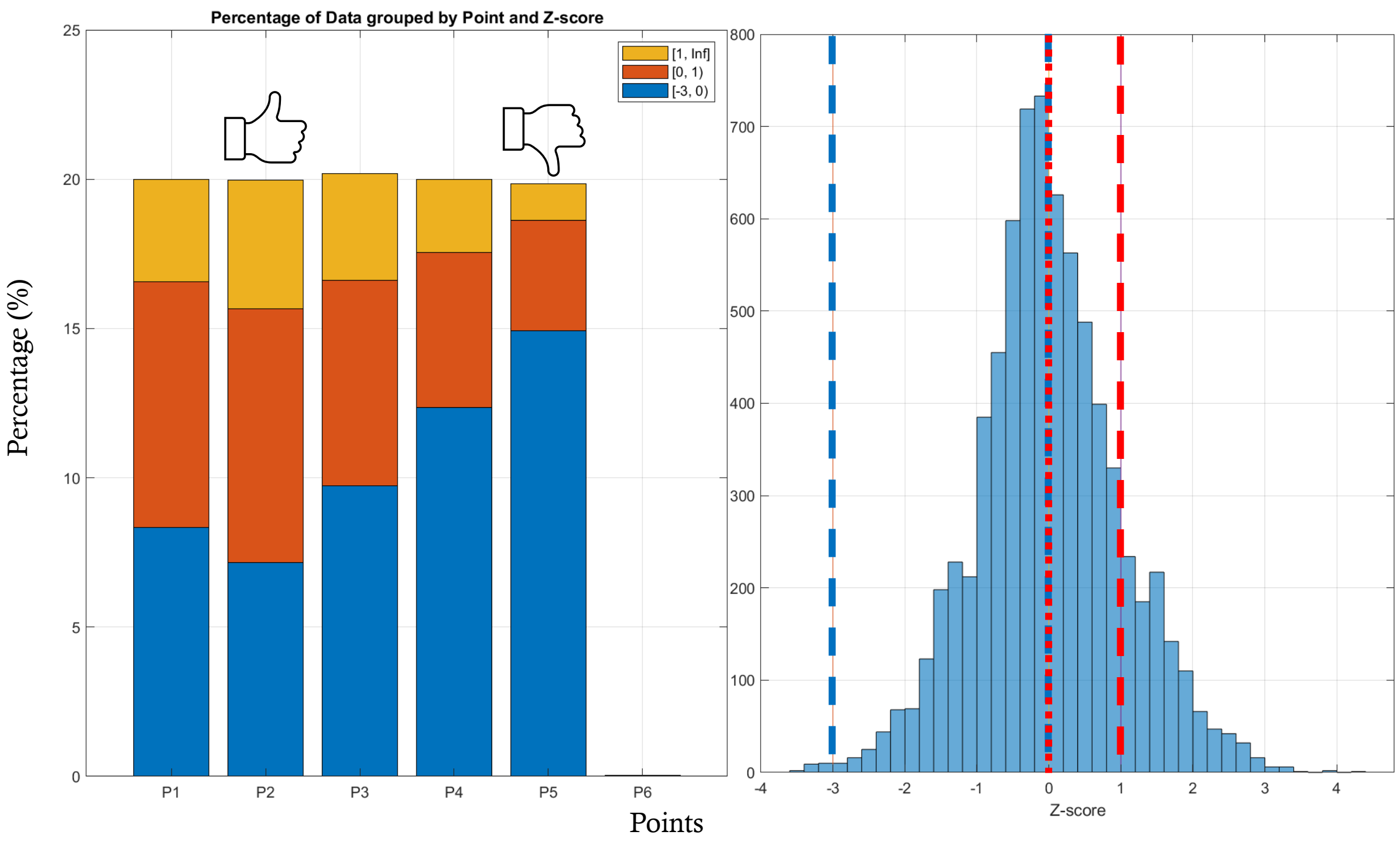


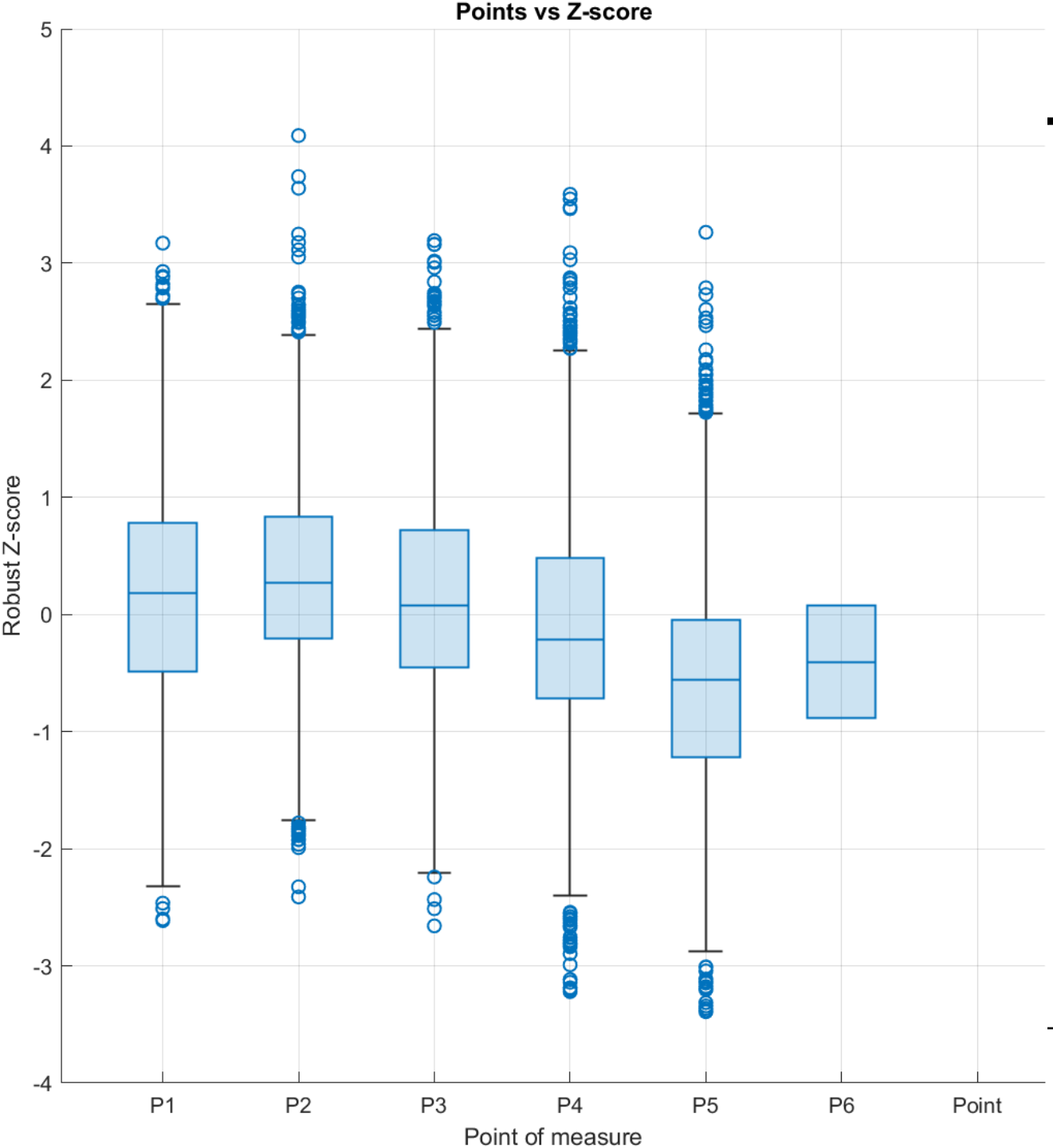









Percentage of Data grouped by Point and Z-score
[1, Inf]
[0, 1)
[-3, 0)
Percentage (%)
Points
P1
P2
P3
P4
P5
P6
Z-score



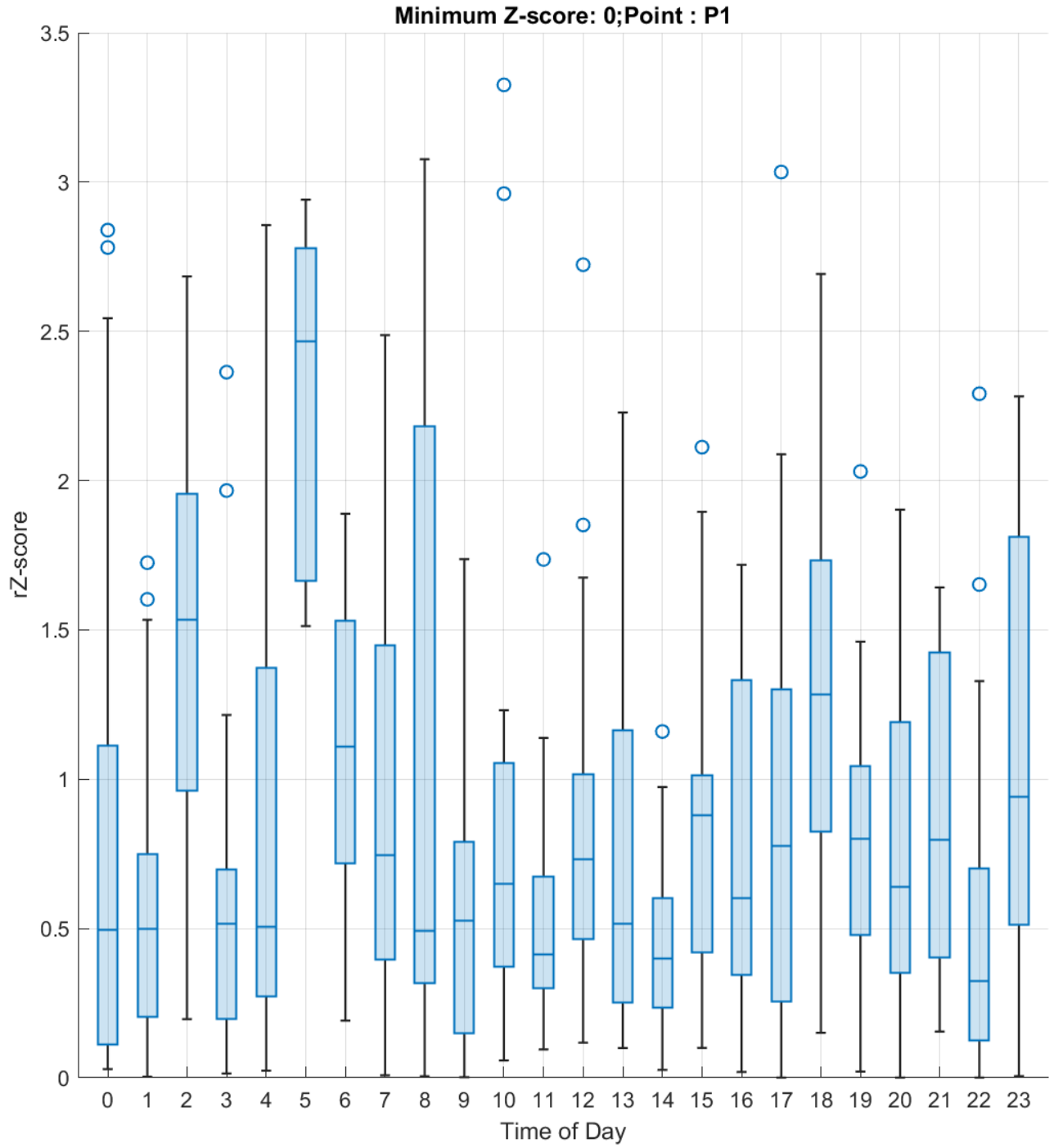
POINT 'S CONTRIBUTION TO Z-SCORE

The higher the score , the higher the concentration

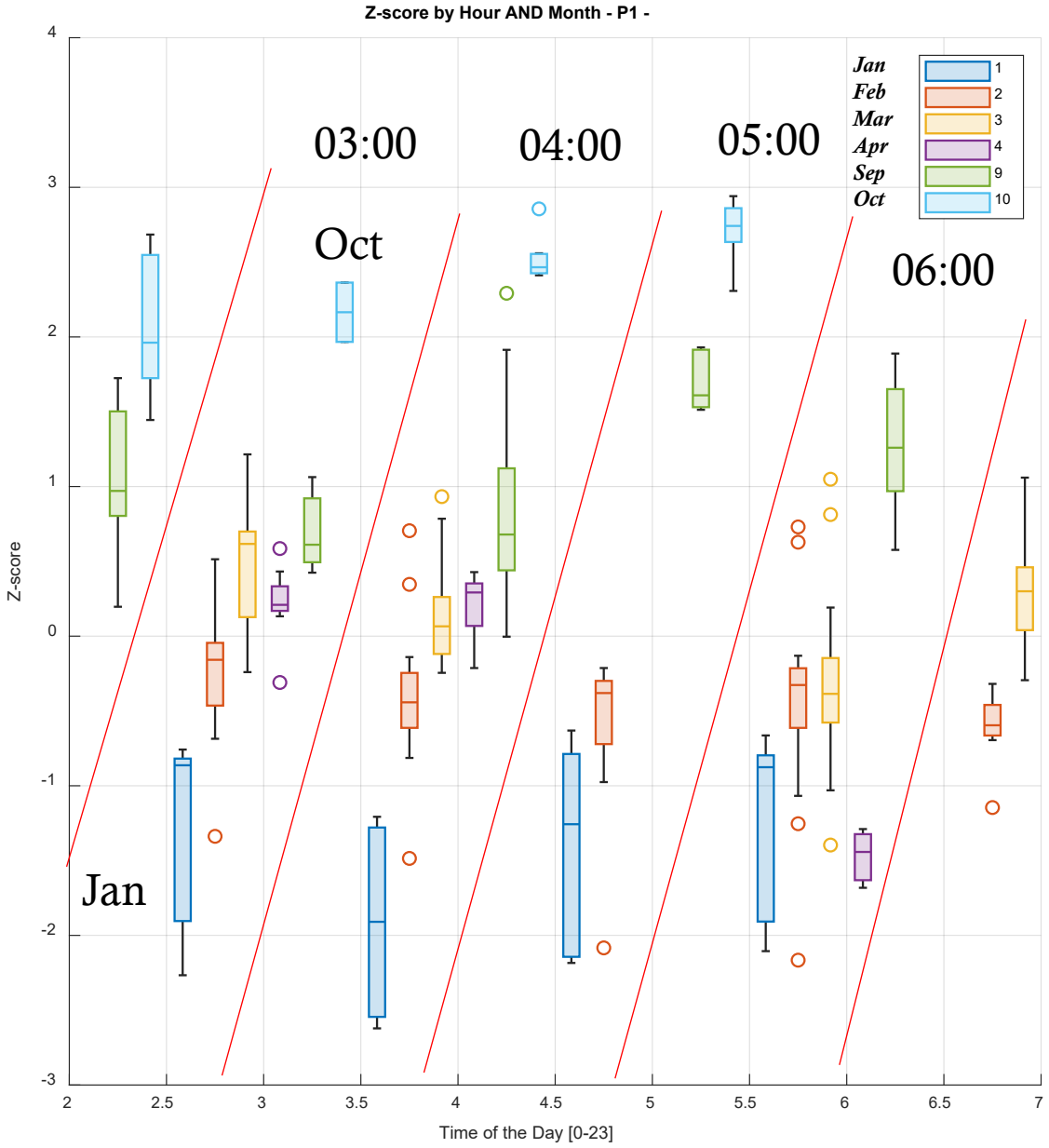
HOW THE HOURLY CONCENTRATION VARIES DURING THE DAY IN THE DIFFERENT POINTS ?

TAKE THE Z-SCORE DISTRIBUTION AT EVERY POINT , MAKE A CUT AND SEE THE FINAL HOURLY DISTRIBUTION. i.e. take a minimum value of Z-score=0 (CH4 concentration about 7.8 ppm) and see each hour how its distribution varies…

Z-SCORE VS TIME OF DAY AT POINT 1

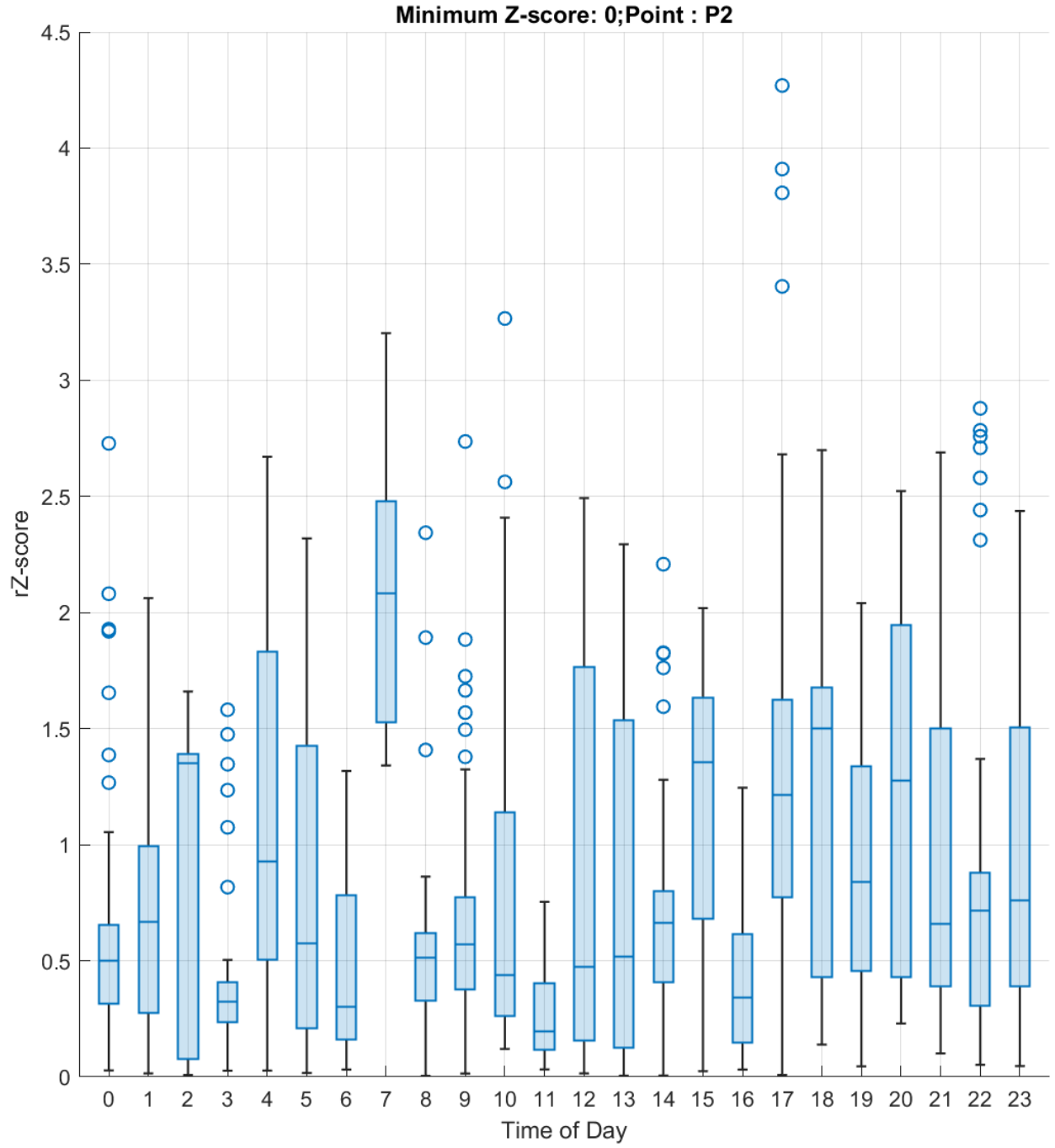


Z-SCORE VS TIME OF DAY (FROM 3:00 TO 6:00) AND MONTH AT POINT 1

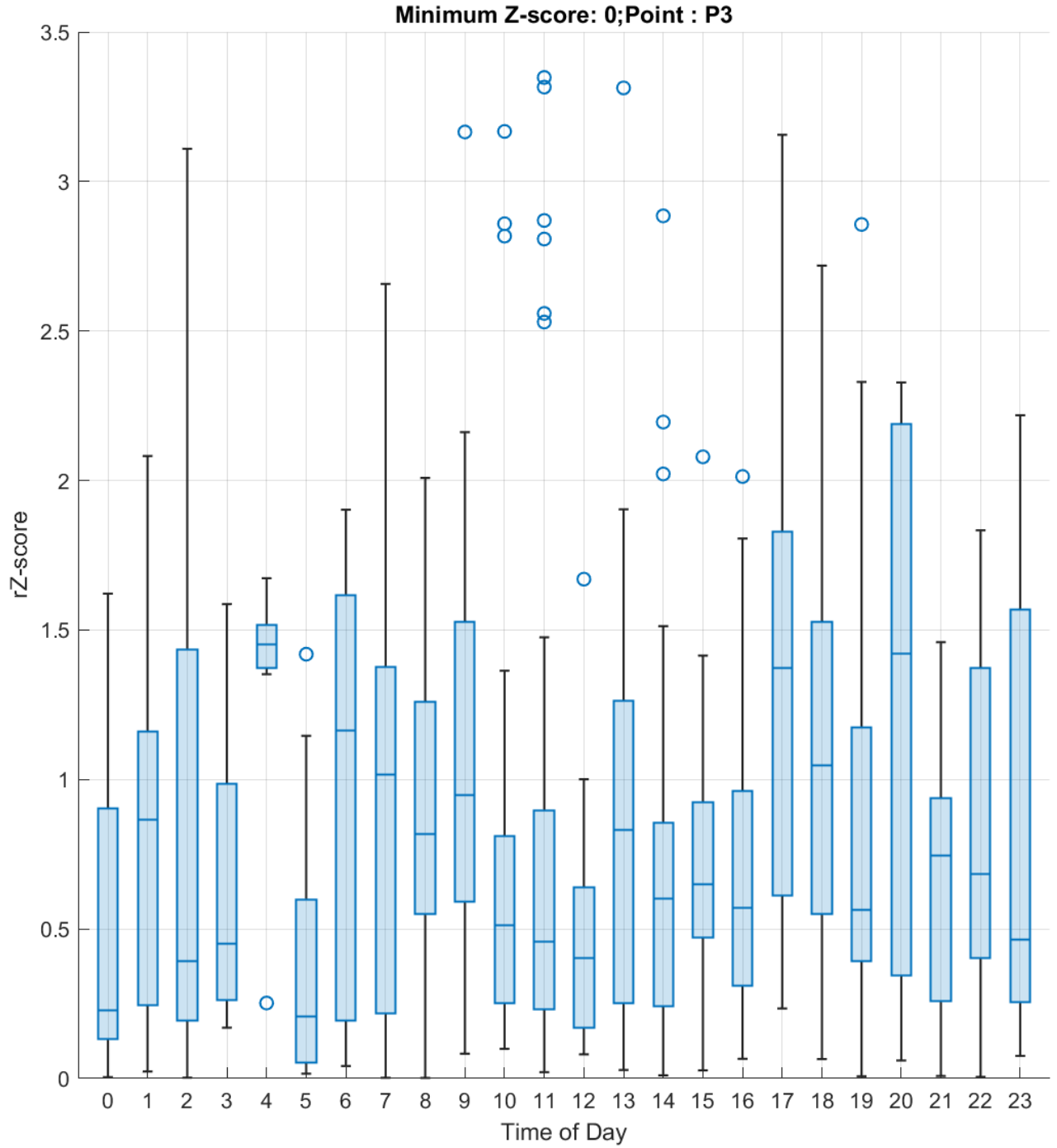


Only as an example

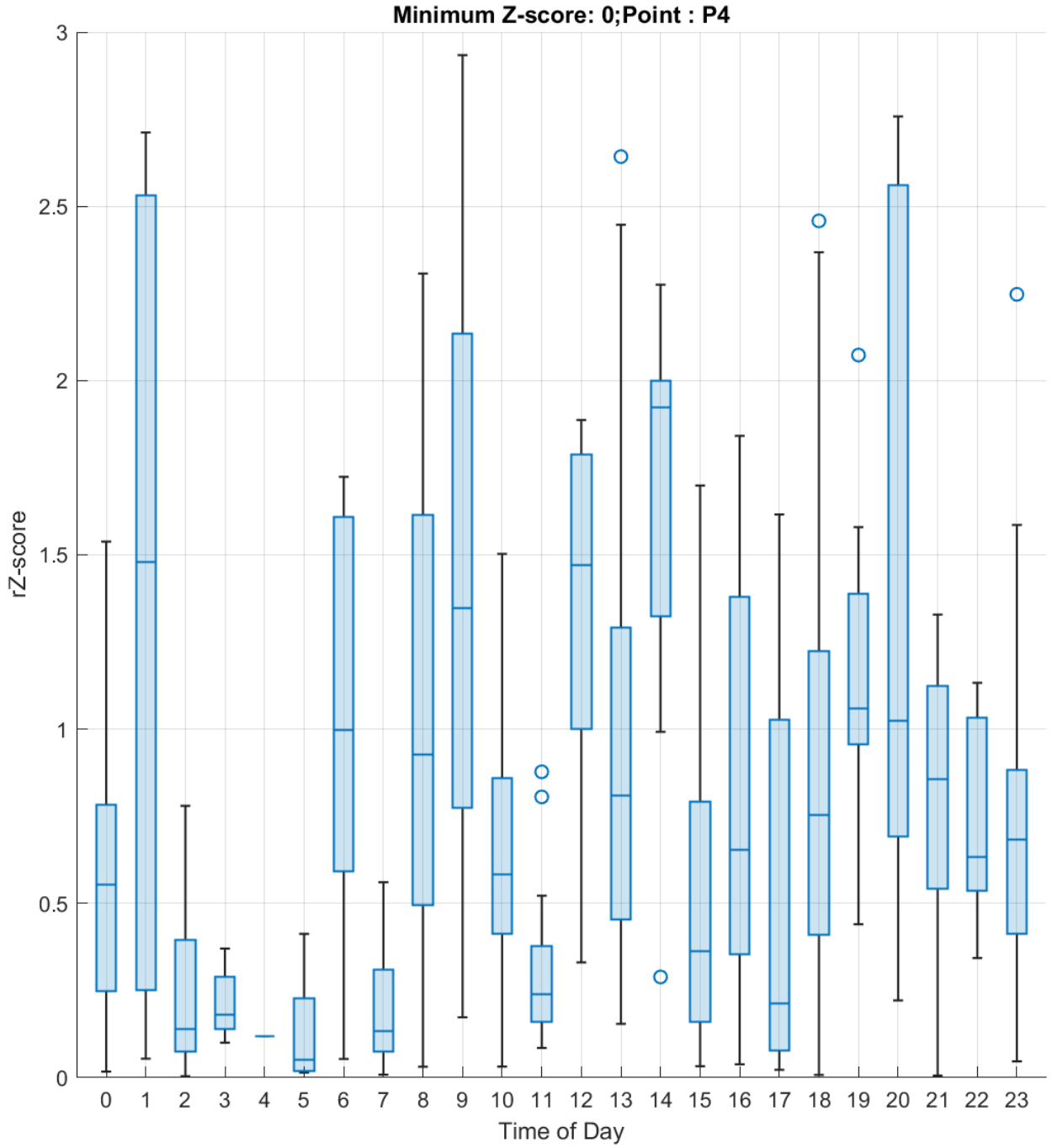
Z-SCORE VS TIME OF DAY AT POINT 2



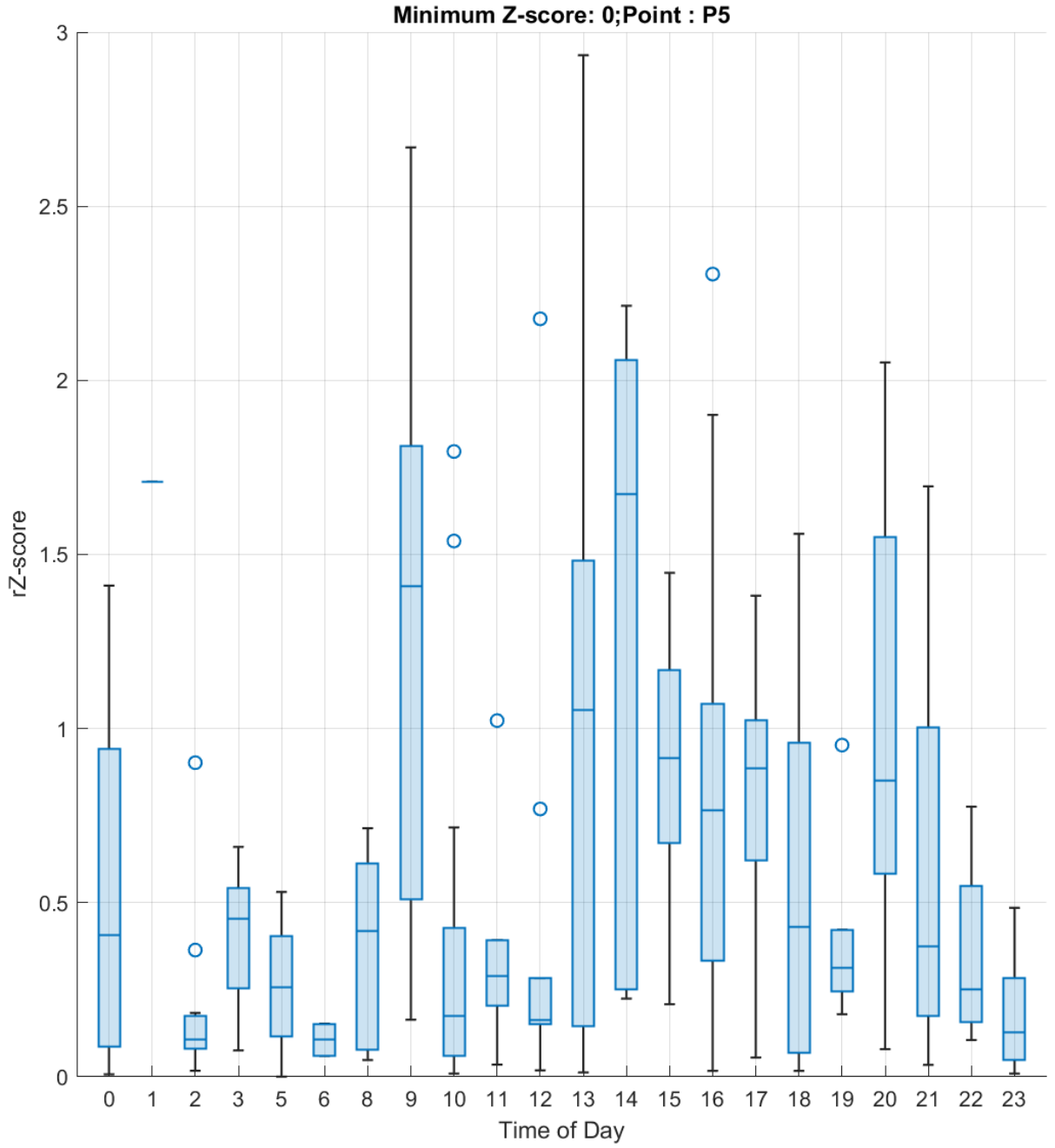
Z-SCORE VS TIME OF DAY AT POINT 3



Z-SCORE VS TIME OF DAY AT POINT 4



Z-SCORE VS TIME OF DAY AT POINT 5



UP TO HERE WE WERE INTERESTED INTO THE RELATIONS AMONG THE POSITION MEASURE WITHIN THE BARN.

NOW LET’S COMPARE DIFFERENT MONTHS

Z-SCORE (NO CUTs)
VS
MONTH

Feb has a many outliers
Half of the Jan data are below the median
Almost half of Sep and Oct data are above the median