II2202: Quantitative tools with Excel and R

prof. Gerald Q. Maguire Jr.
School of Information and Communication Technology (ICT)
KTH Royal Institute of Technology
http://web.ict.kth.se/~maguire

Prof. em. Marilyn E. Noz, Ph. D.
School of Medicine
New York University

II2202 Fall 2015 2015.08.03

Some statistical concepts

Independent versus dependent variables

Independent variable – a variable that you can change

Dependent variable:

- A response or outcome
- This is what you will measure

Types of data

Nominal data	unordered groups, e.g., male/female, left-handed/right-handed, …
Ordinal data	rank ordered; the difference between item numbered n and n+i does *not* tell you anything other than that one is ranked ahead of the other, e.g. Top 500 Universities, top 10 protocols in bytes, …
Interval data	continuous ranges mapped to some scale, without a clear zero
Ratio data	like interval data, but with a clear absolute zero value

Metrics

Type of data	Example Metrics	Common statistics
Nominal data	Success/failure	Frequencies, Chi-square
Ordinal data	Ranking	Frequencies, Chi-square, Wilcoxian rank sum tests, Spearman rank correlation
Interval data	Likert scale, System Useability Scale,	All descriptive statistics (average, median, std. dev., …), Student's t-test, ANOVA, correlation, regression, …
Ratio data	Task completion time, packet inter-arrival time, …	All of the previous + geometric mean

Adapted from Table 2.3 on page 23 of [Tullis2008]

Measures of Central Tendency

Three most common measures are:

Mean	arithmetic average
Median	mid point of the distribution (half the values are larger and half are smaller)
Mode	most common value

Selecting participants

Random sampling

Systematic sampling – e.g. every 3rd person

Stratified sampling – based upon a representative subset

Samples of convenience

- Who can you get?
- Are they representative of the target population?

Sample size

- What is the goal?
- Is the difference expected to be large or small?
- What is an acceptable margin of error?

Within- vs. between-subjects

Within-subjects

- Also known as repeated-measures
- The same subject, but repeated measurements

Between-subjects

- Comparing results of subject_i with subject_k
- Avoids carry-over effects (where the subject learns from one trial and this causes a difference in subsequent trials)

Mixed design

Counterbalancing

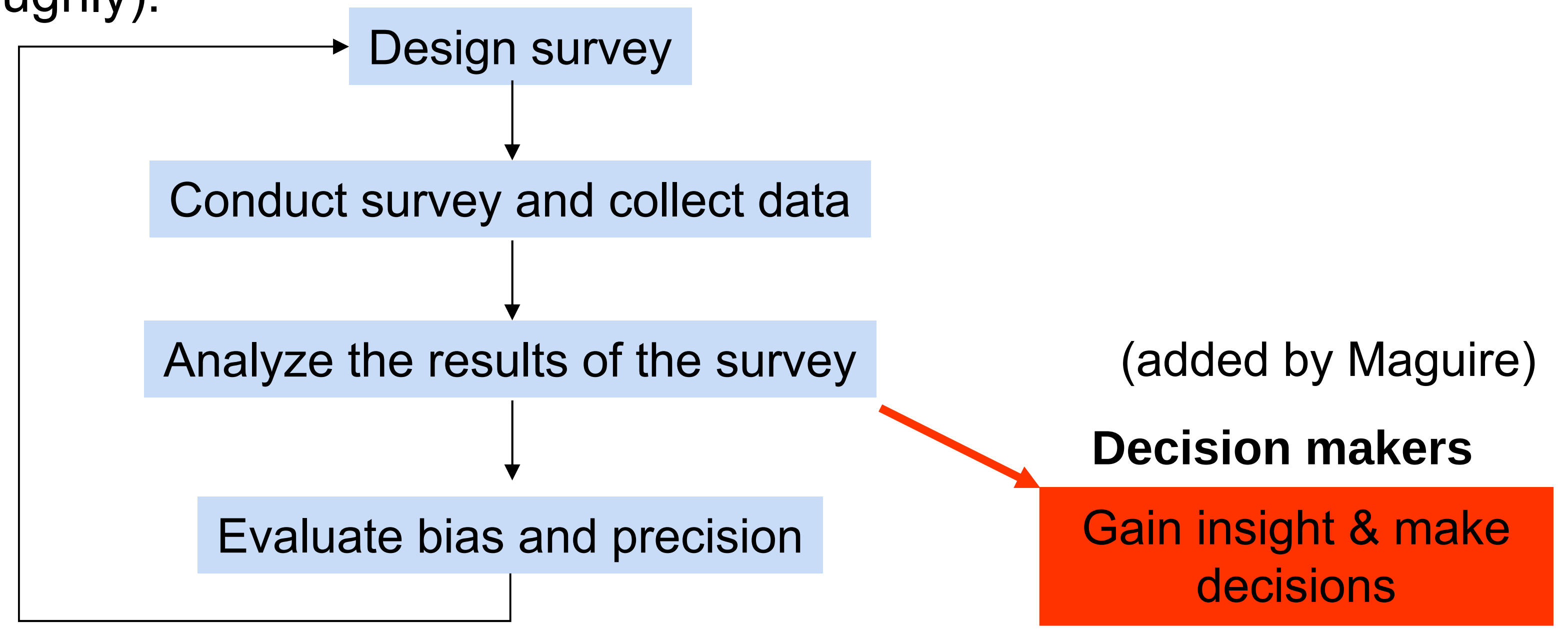
To avoid carryover effects *vary the order of the tasks*:

- Randomize order
- Sets of predefined orders – subject is randomly assigned to one of these sets

(Starting) Quantitative analysis of survey data

Overview

Gillian Raab, Professor of Applied Statistics at Napier University, shows the process of carrying out surveys as viewed by a statistician (roughly):



Adapted from the figure on his slide 7 in “Background to P|E|A|S project”, 9 September 2004, http://www2.napier.ac.uk/depts/fhls/peas/workshops/workshop1presentationGR.ppt

Objective

What is the object of the survey?

- Finding a **predictive** model
- Finding **hidden** relationships
- Segmenting a population into **strata**
- **Visualizing** responses
  (e.g., Distance from a park versus frequency of visits to this park)
- Making a **decision** (e.g., where to put a park)

What is (are) the research question(s)?

Considerations when designing studies

Ken Kelley and Scott E. Maxwell state:

"At a minimum, the following points must be considered when designing studies in the behavioral, educational, and social sciences:

(a) the question(s) of interest must be determined;

(b) the population of interest must be identified;

(c) a sampling scheme must be devised;

(d) selection of independent and dependent measures must occur;

(e) a decision regarding experimentation versus observation must be made;

(f) statistical methods must be chosen so that the question(s) of interest can be answered in an appropriate and optimal way;

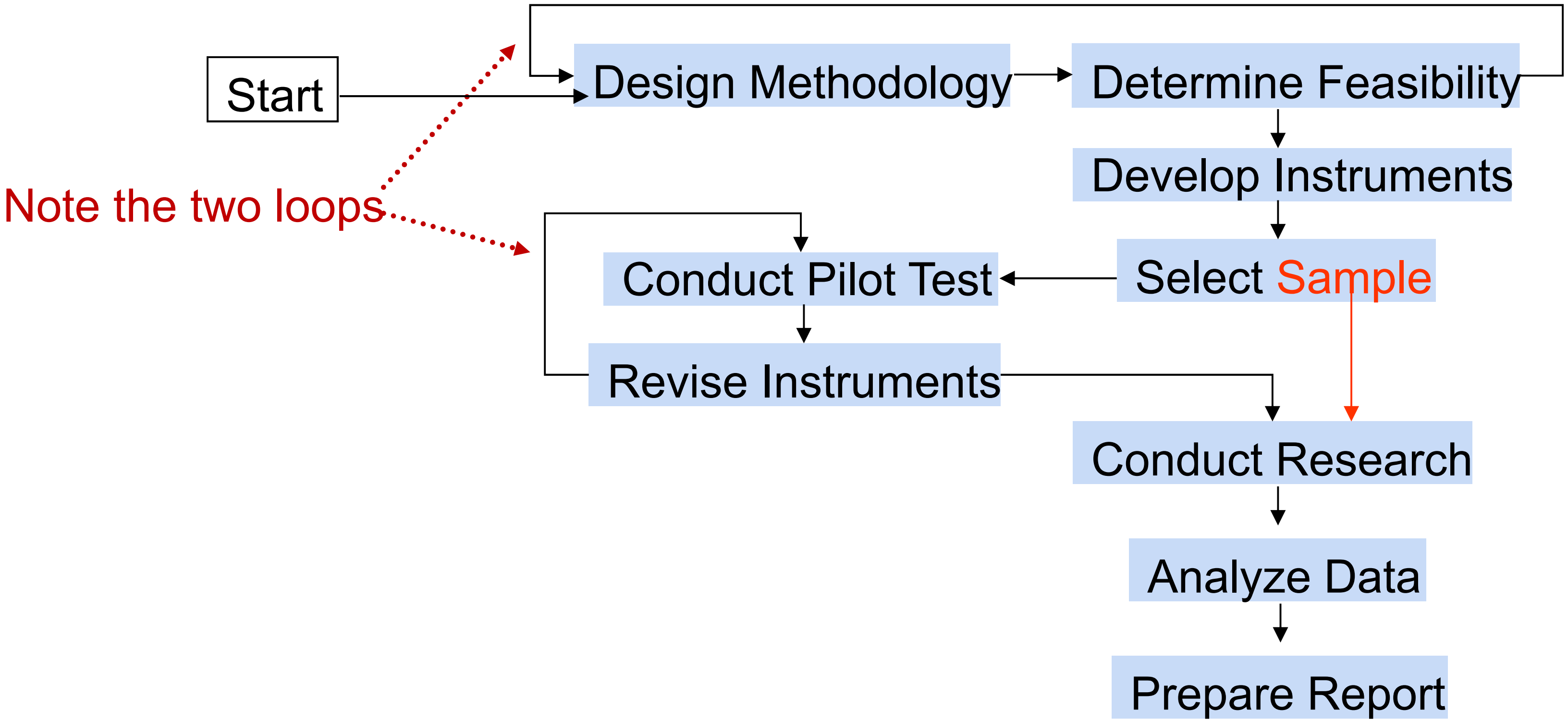
(g) sample size planning must occur so that an appropriate sample size given the particular scenario, as defined by points a through f, can be used;

(h) the duration of the study and number of measurement occasions need to be considered;

(i) the financial cost (and feasibility) of the proposed study calculated."

Ken Kelley and Scott E. Maxwell, Sample Size Planning with Applications to Multiple Regression: Power and Accuracy for Omnibus and Targeted Effects, [Kelley2008] http://nd.edu/~kkelley/publications/chapters/Kelley_Maxwell_Chapter_SSMR_2008.pdf

Questionnaire Research Flow Chart



Adapted from pg. 3 of David S. Walonick, *A Selection from Survival Statistics* [Walonick2010] , https://www.statpac.com/surveys/surveys.pdf

Sampling methods

Probability

- Random sampling & systematic sampling (every Nth person) ⇒ equal probability of selection
- Sampling proportional to size (PPS) – concentrates on the largest segments of the population
- Stratified sampling (members of each stratum (a sub-population) share some characteristic)
- Advantage: can calculate sampling error

Nonprobability

- Accidental, Haphazard, convenience sampling ⇒ these might not be representative of the target population
- Purposeful – sampling with a purpose in mind
  - Modal instance sampling –focused on 'typical' case
  - Expert sampling – choosing experts for your samples
  - Quota sampling - proportional vs. non-proportional
  - Heterogeneity sampling – to achieve diversity in samples
  - Snowball sampling – get recommendations of others to sample, from your samples

For further details of Nonprobability sampling see: [Trochim2006]
http://www.socialresearchmethods.net/kb/sampnon.php

Sample size

Choosing the size of your sample is related to your expected signal to noise ratio and your desired confidence.

Statisticians speak about **statistical power**, for details see [Trochim2010] http://www.socialresearchmethods.net/kb/power.php

See also: [Kelley2008] [Maxwell2008 [Kelley2003a] [Kelley2003b] [Kelley2008a]

Getting started with data analysis

Assuming that the survey has already be conducted and that the data has been entered into a computer system, what is the next step?

Preliminary analysis

- Descriptive statistics

Exploratory data analysis

- Plots (points, lines, scatterplots, …), histograms, …

Types of analysis

Design-based analysis

- In this approach randomness is **induced** by the random selection of sample or the assignment of samples to a subset
- Choice of a statistical model can be used for model-based inference

Model-based analysis

In this approach randomness is because of the **innate** randomness in the measurements (in the case of surveys – these are the responses)

Modeling techniques

Prediction, classification (using neural networks, Bayesian networks, trees, …), regression

Clustering, segmentation

Fitting to an *a priori* model

Factor analysis, principle components analysis

Weights

When we have samples, we need to make sure that these samples are representative of the total population – to do this we may need to establish weights

For details of weights see:

James R. Chromy and Savitri Abeyasekera, "Statistical analysis of survey data" [Chromy2005] http://www.cpc.unc.edu/projects/addhealth/data/guides/weight1.pdf

Significance

Significance is a statistical term indicating *your confidence* in your conclusion that a real difference exists or that a relationship actually exists, i.e., that the result is unlikely to be due simply to chance.

If your hypothesis states a direction of this difference – use a **One-Tailed** significance test, otherwise use a **Two-Tailed** significance test.

Note: Significant **does not** imply important, interesting, or meaningful!

Similarly not all observations that are not statistically significant are unimportant, uninteresting, …

Testing for significance

1. Decide on your significance level α
2. Calculate your statistical value p
3. If $p < \alpha$, then the result is significant, else it is not significant

An alternative view is:

confidence = (signal/noise) * √sample size

For details of the above equation see: David L. Sackett, Why randomized controlled trials fail but needn't:2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!). [Sackett2001] http://www.cmaj.ca/cgi/content/full/165/9/1226

See also: Understanding Hypothesis Testing: Example #1, Department of Statistics, West Virginia University, last modified 4 April 2000 http://www.stat.wvu.edu/SRS/Modules/HypTest/exam1.html

Next steps

1. Search the literature and read extensively

2. Consult a statistician to get help with your statistical analysis
   (In most cases this is going to cost you money, but can save you a lot of time and effort.)

3. Doing some statistical analysis yourself

Using Excel for statistics and plotting

Experiment 1

Captured packets using Wireshark during a long (2150.12 second) VoIP call

$\Rightarrow$ at least: 107,505 RTP packets in each direction

$\Rightarrow$ 429 RTCP packets in one direction

http://www.Wireshark.org

Load the data, then extract relevant RTP packets

Starting with a tab separated file of the form:

"No." "Time" "Source" "Destination“ "Protocol“ "RSSI" "Info“

"1443" "17685.760952" "90.226.255.70" "217.211.xx.xx" "RTP" "" "PT=ITU-T G.711 PCMA, SSRC=0x6E21893F, Seq=183, Time=46386 “

Extract the traffic going to me, i.e., with the 217.211.xx.xx destination and Protocol = RTP

⇒ Do this by sorting on column Protocol and Destination

⇒ Select the desired rows and move to new sheet

Note: Either preprocess the “Info” filed into separate columns (for PT, SSRC, Seq, and Time) or write an Excel function to do this for you

From network to local user agent

Raw output fom Microsoft Excel 2010 (Beta)

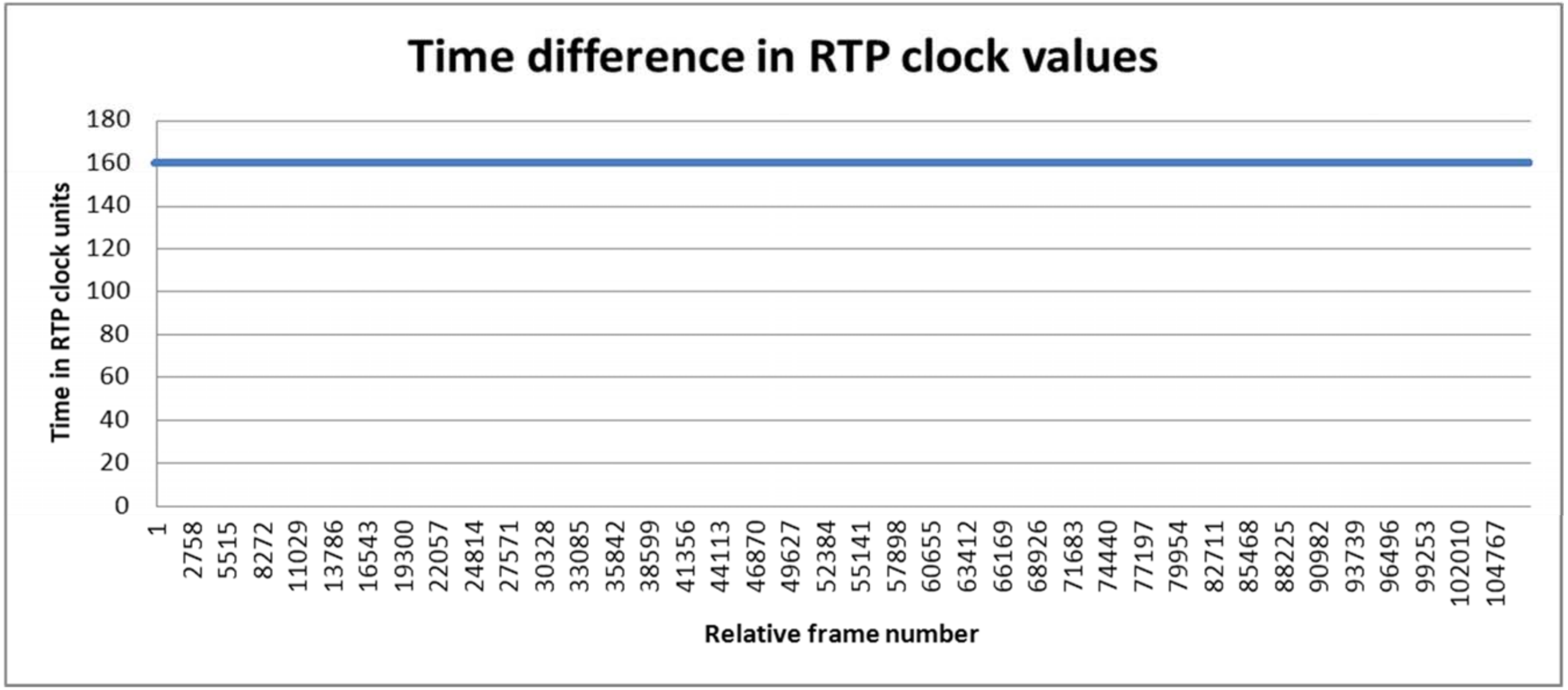
Difference in RTP clock from previous sample

Mean	160
Standard Error	0
Median	160
Mode	160
Standard Deviation	0
Sample Variance	0
Kurtosis	#DIV/0!
Skewness	#DIV/0!
Range	0
Minimum	160
Maximum	160
Sum	17200960
Count	107506
Confidence Level(95.0%)	0

Inter-arrival times (in seconds) of RTP packets

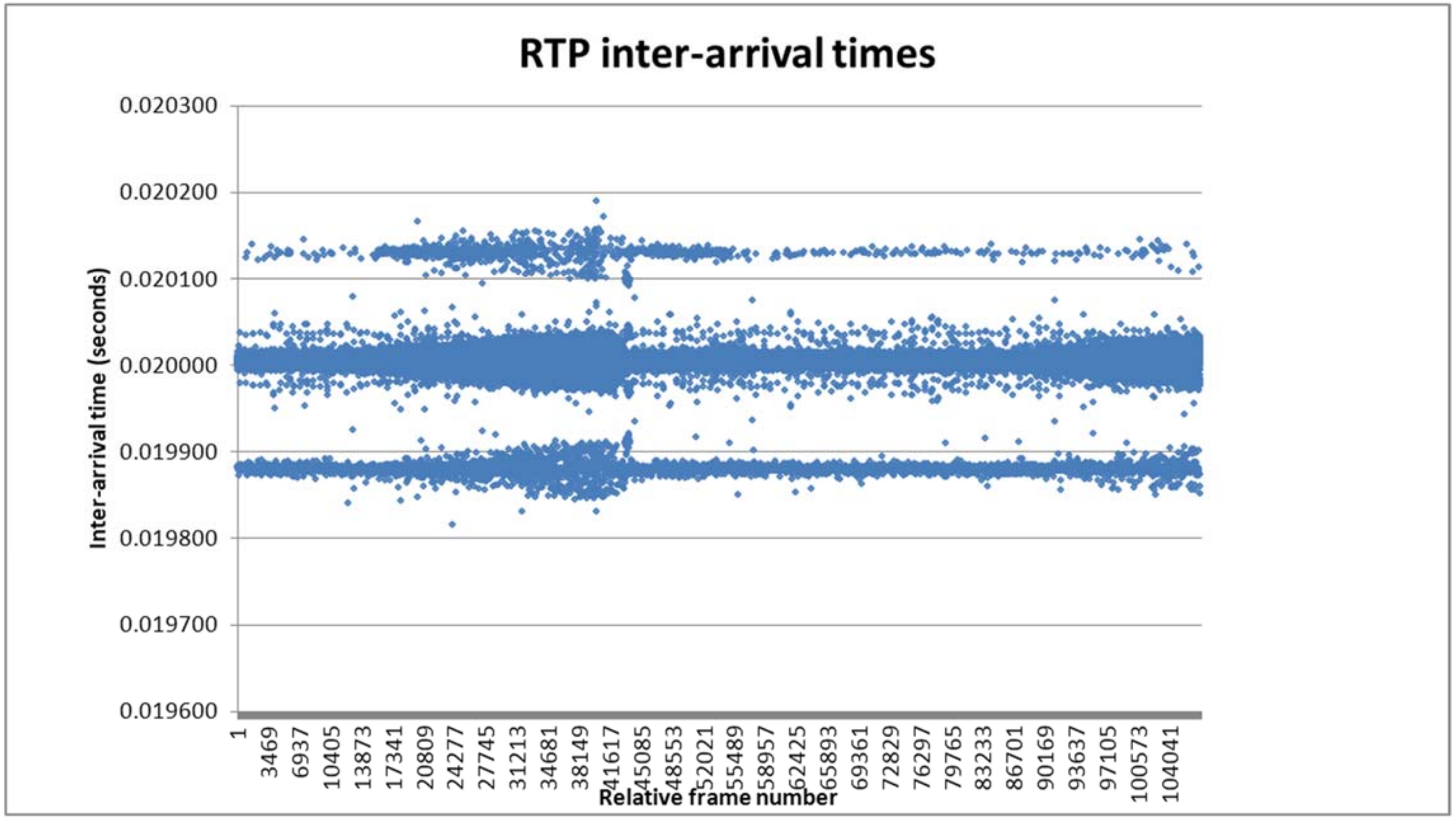
Mean	0.019999999
Standard Error	9.28526E-08
Median	0.020004
Mode	0.020005
Standard Deviation	3.04446E-05
Sample Variance	9.26874E-10
Kurtosis	12.36652501
Skewness	-2.054662184
Range	0.000374
Minimum	0.019815
Maximum	0.020189
Sum	2150.11991
Count	107506
Confidence Level(95.0%)	1.8199E-07

First look at the RTP clock (Time) differences

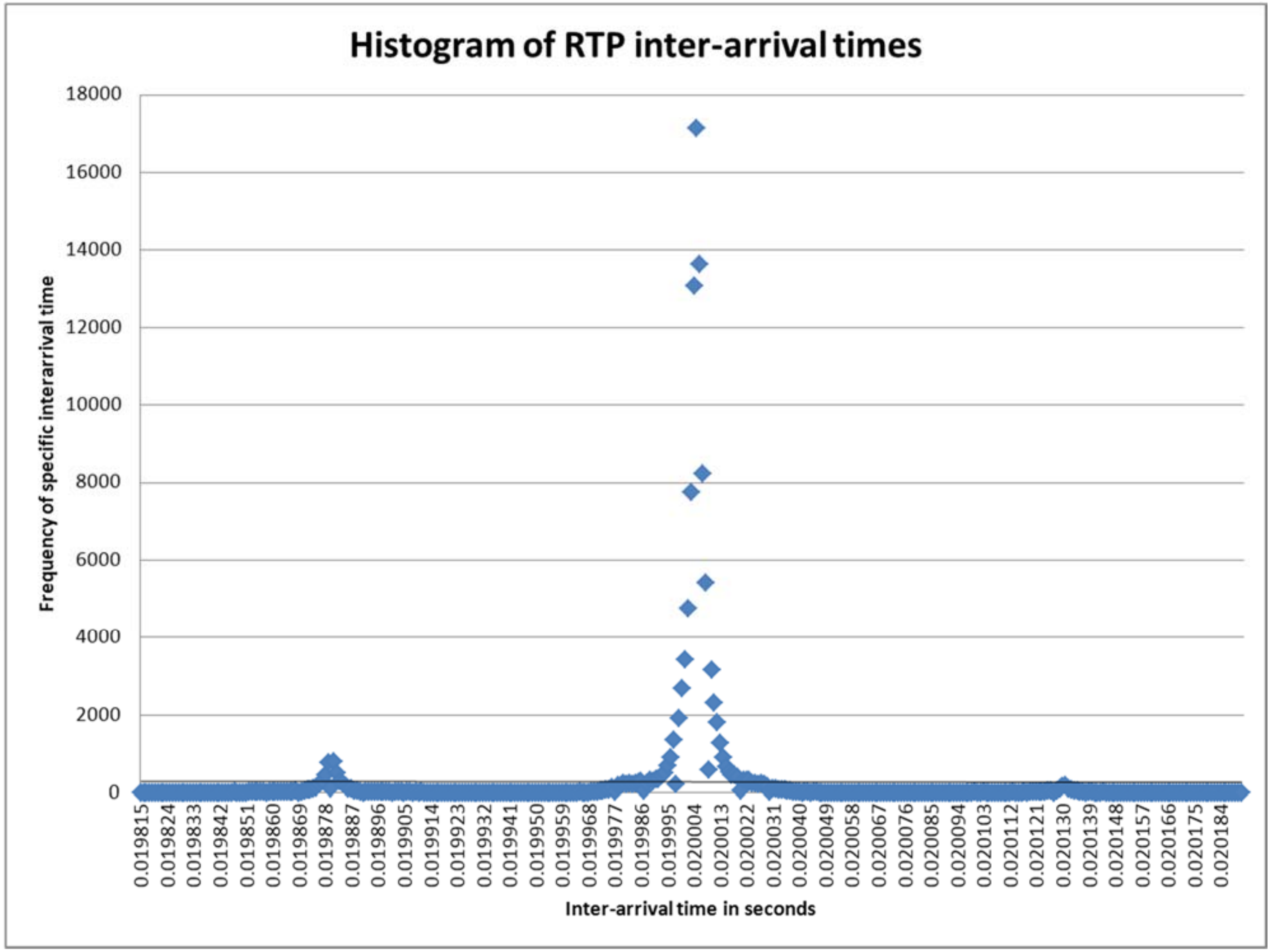


Conclusion: 160 audio samples per frame, with a frame time of 0.20 ms
⇒ 8 K sample/second sampling rate – consistent with ITU-T G.711 PCMA encoding

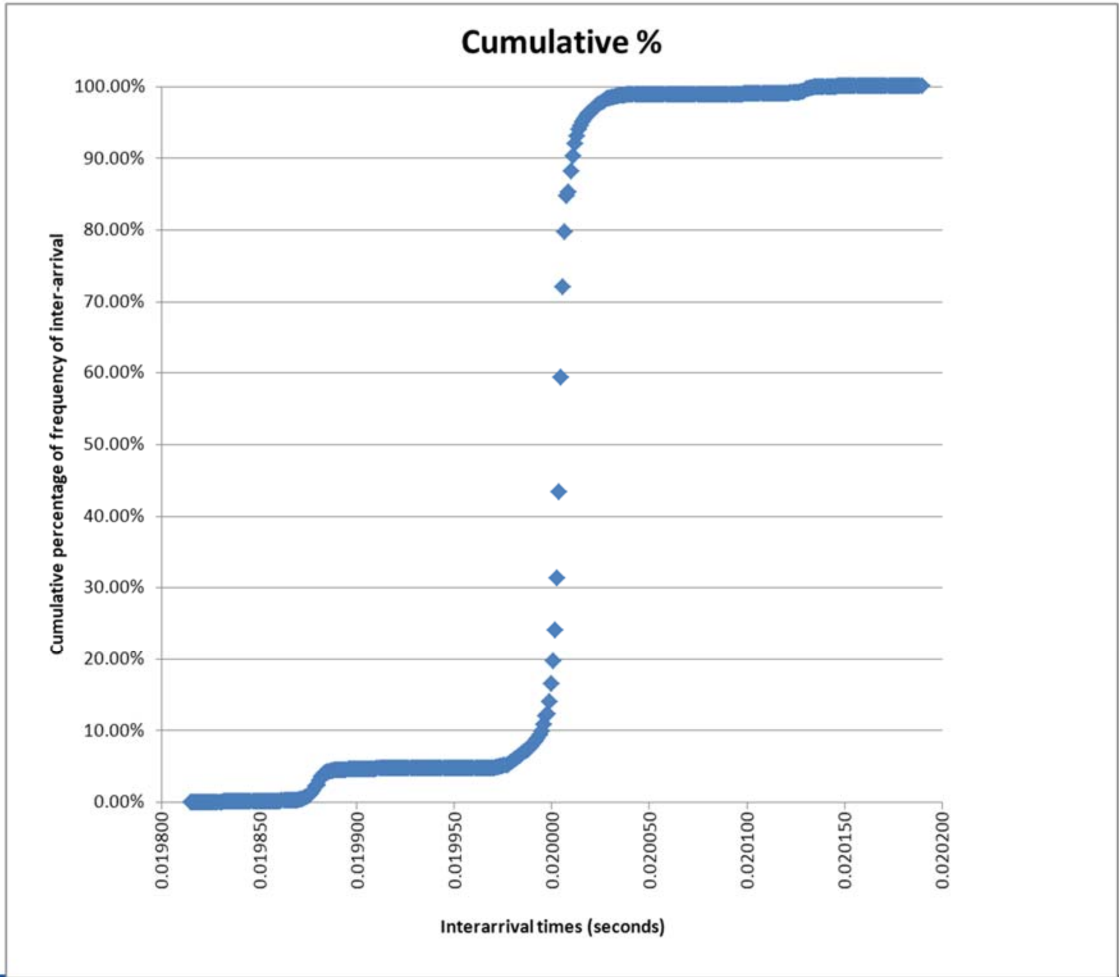
Plot RTP inter-arrival times as measured by Wireshark



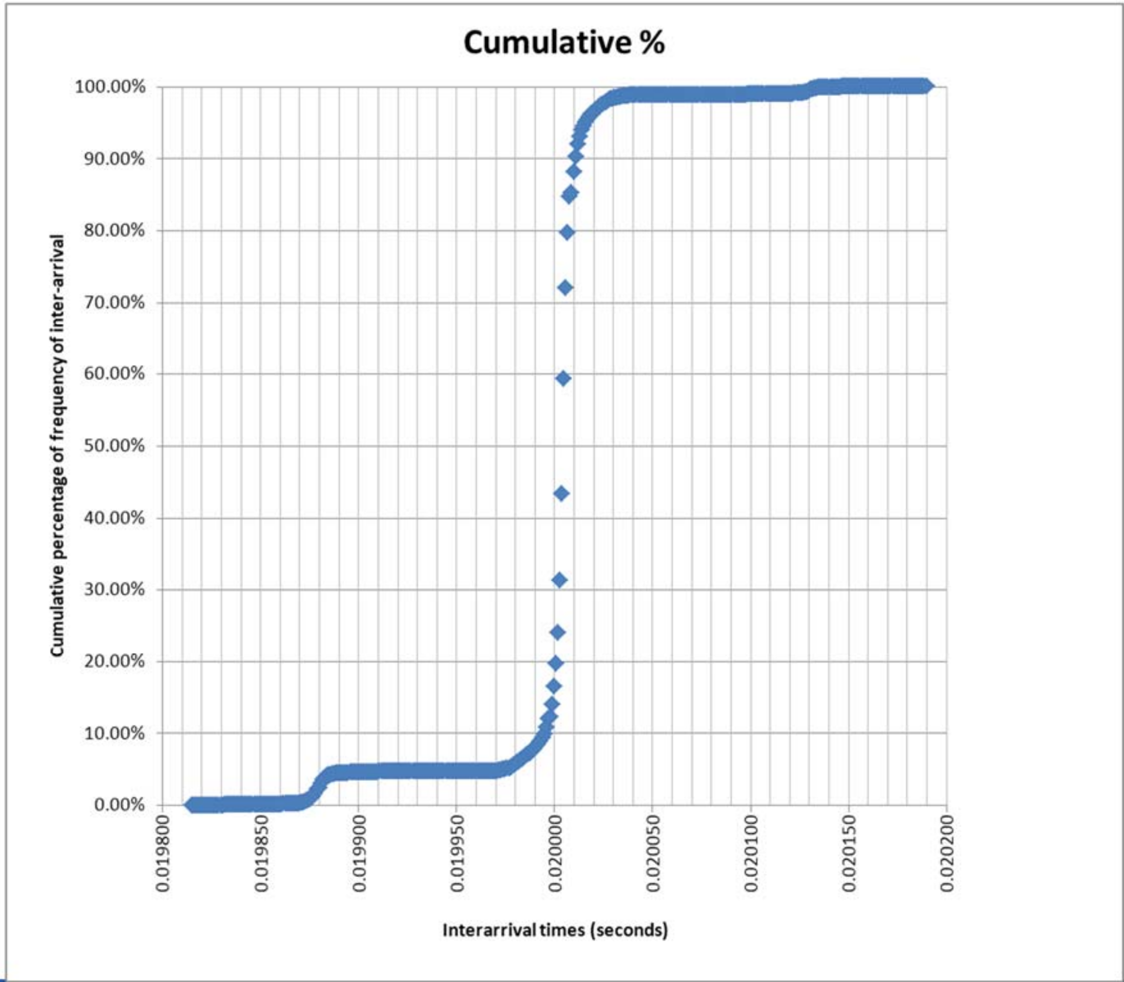
Compute histogram of inter-arrival times



Plot Cumulative Distribution of inter-arrival times



Add grid lines



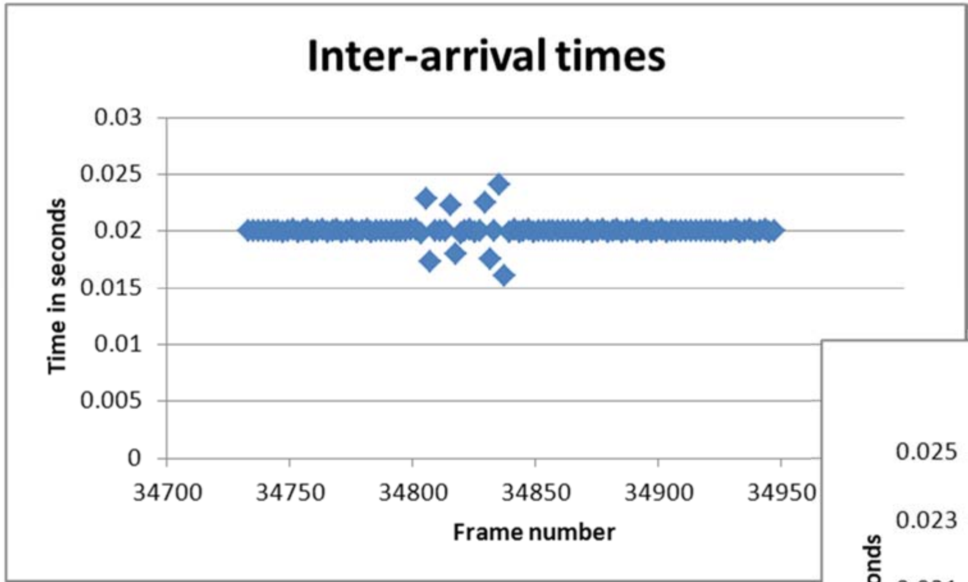
As numbers - near median

	seconds	frequency	Cumulative %
	0.019995	687	9.92%
	0.019996	895	10.75%
	0.019997	1334	11.99%
	0.019998	209	12.18%
Mean	0.019999	1898	13.95%
	0.020000	2671	16.44%
	0.020001	3403	19.60%
	0.020002	4747	24.02%
	0.020003	7742	31.22%
Median	0.020004	13059	43.37%
Mode	0.020005	17121	59.30%
	0.020006	13630	71.98%
	0.020007	8211	79.62%
	0.020008	5404	84.64%
	0.020009	570	85.18%
	0.020010	3158	88.11%
	0.020011	2305	90.26%
	0.020012	1787	91.92%
	0.020013	1262	93.09%
	0.020014	886	93.92%

With varying numbers of samples

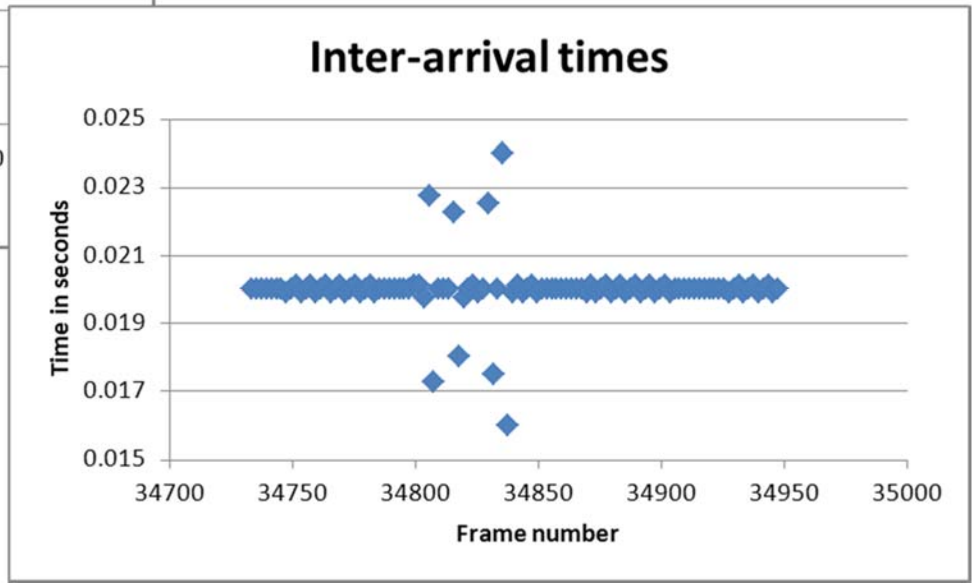
Descriptive Statistics	First 100	First 1K	First 10K	First 100K
Mean	0.02000071	0.020000066	0.020000004	0.02
Standard Error	2.12714E-06	7.53406E-07	2.51164E-07	9.69855E-08
Median	0.020005	0.020004	0.020004	0.020004
Mode	0.020005	0.020005	0.020005	0.020005
Standard Deviation	2.12714E-05	2.38248E-05	2.51164E-05	3.06695E-05
Sample Variance	4.52471E-10	5.67621E-10	6.30831E-10	9.40618E-10
Kurtosis	28.87137928	21.46428225	19.07376827	12.23083198
Skewness	-5.453831468	-4.509853108	-3.831289593	-2.003065575
Range	0.000135	0.000252	0.000277	0.000374
Minimum	0.01988	0.019872	0.019868	0.019815
Maximum	0.020015	0.020124	0.020145	0.020189
Sum	2.000071	20.000066	200.000044	1999.999951
Count	100	1000	10000	100000
Confidence Level(95.0%)	4.2207E-06	1.47844E-06	4.92331E-07	1.9009E-07

Zooming in on interesting behavior

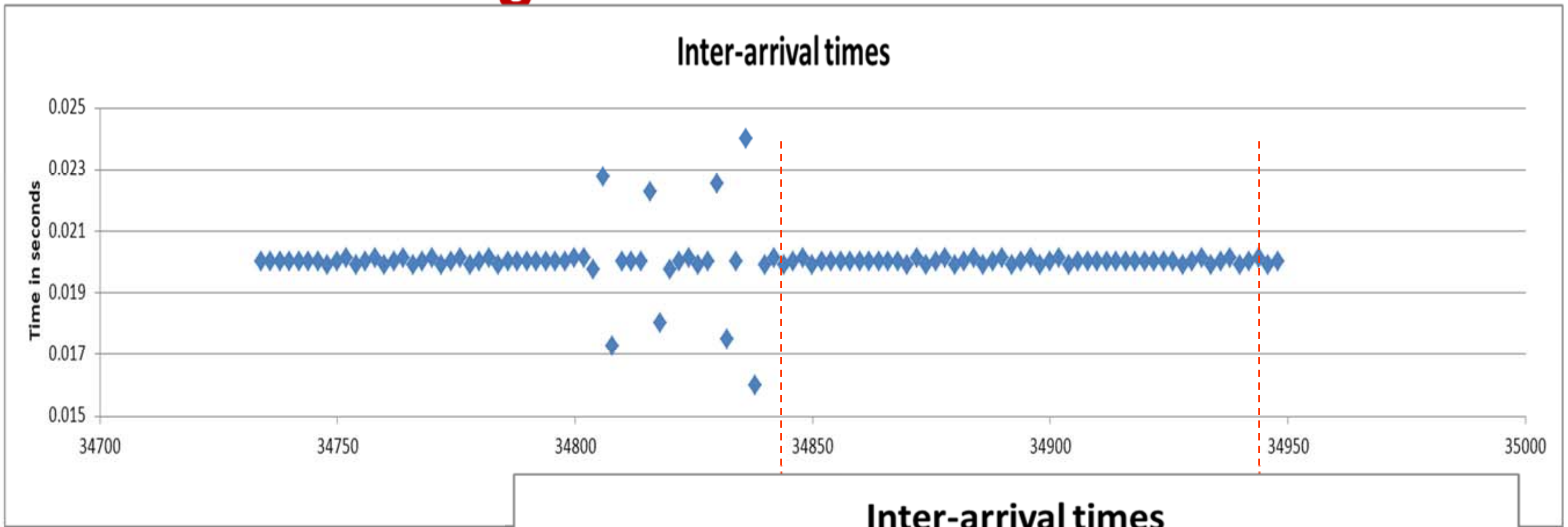


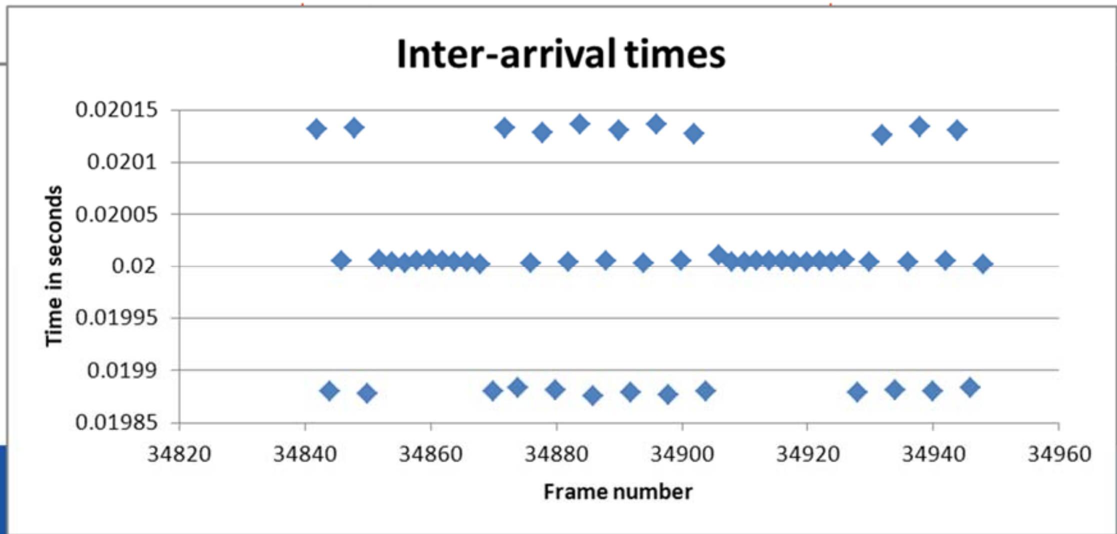
Note that the plot is now a **scatter plot**.

Re-scale

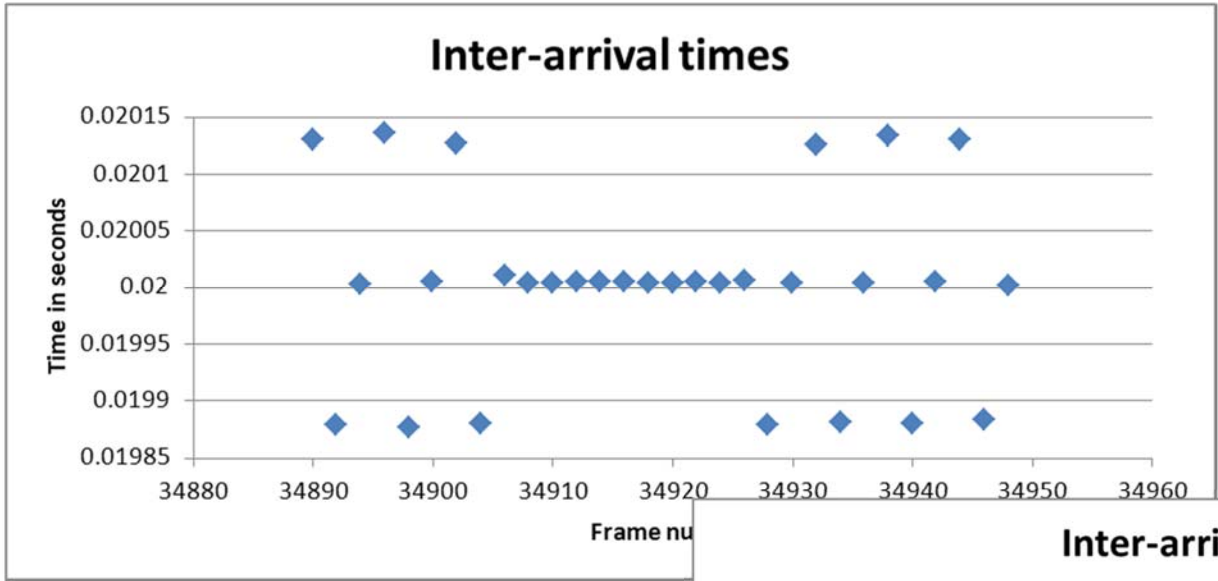


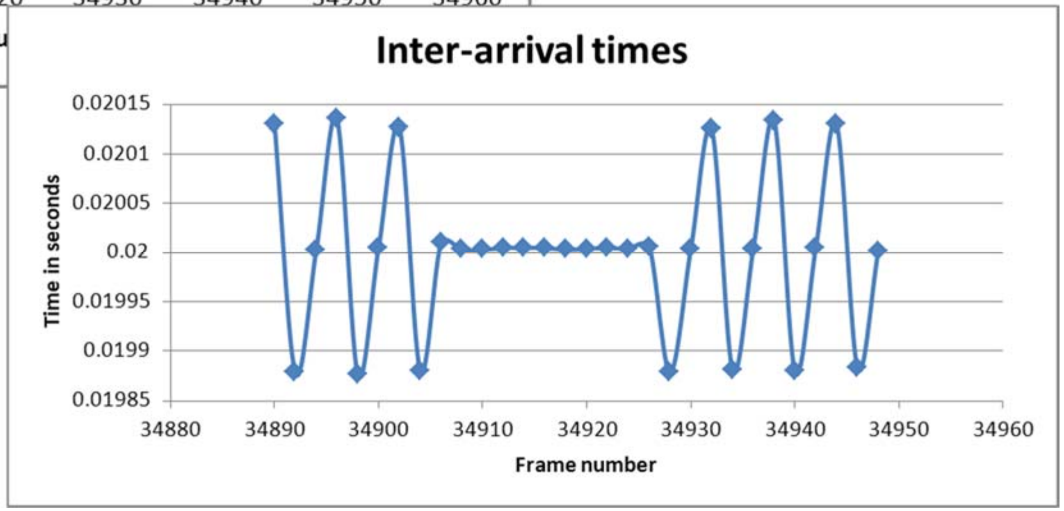
Looking in more detail at a relatively "flat" region





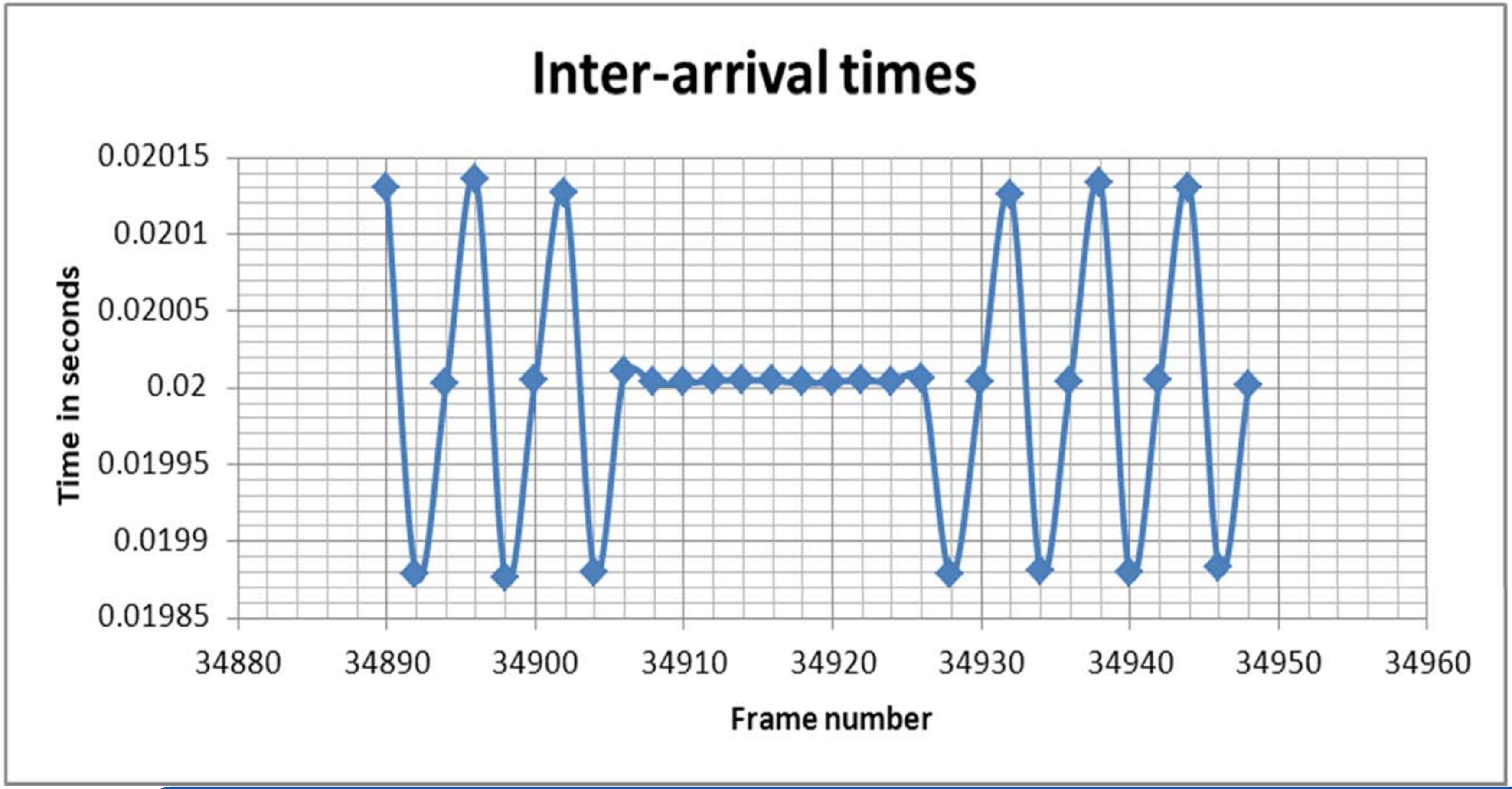
Is there some pattern?



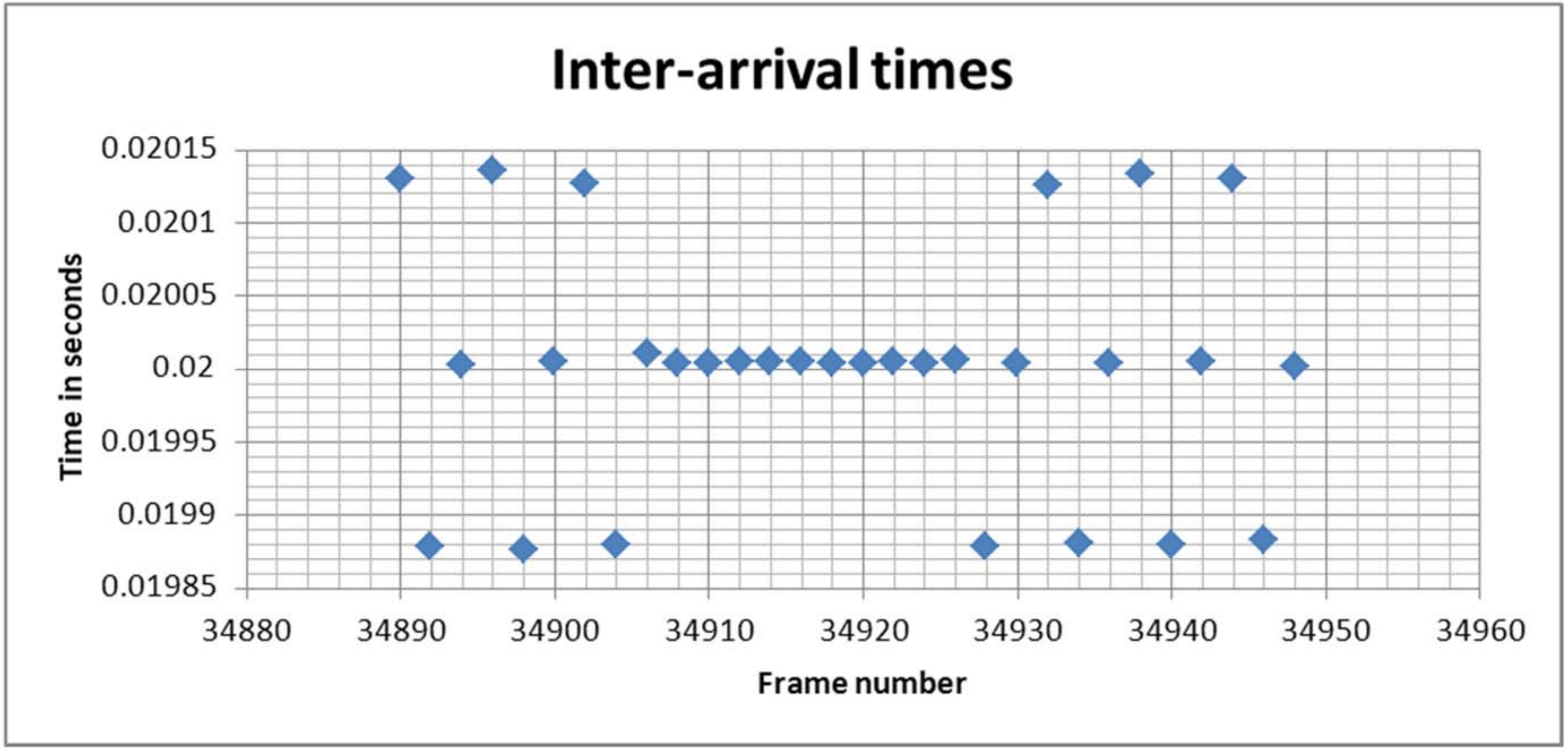


Although this is **not** continuous data, connecting with lines shows the values oscillate

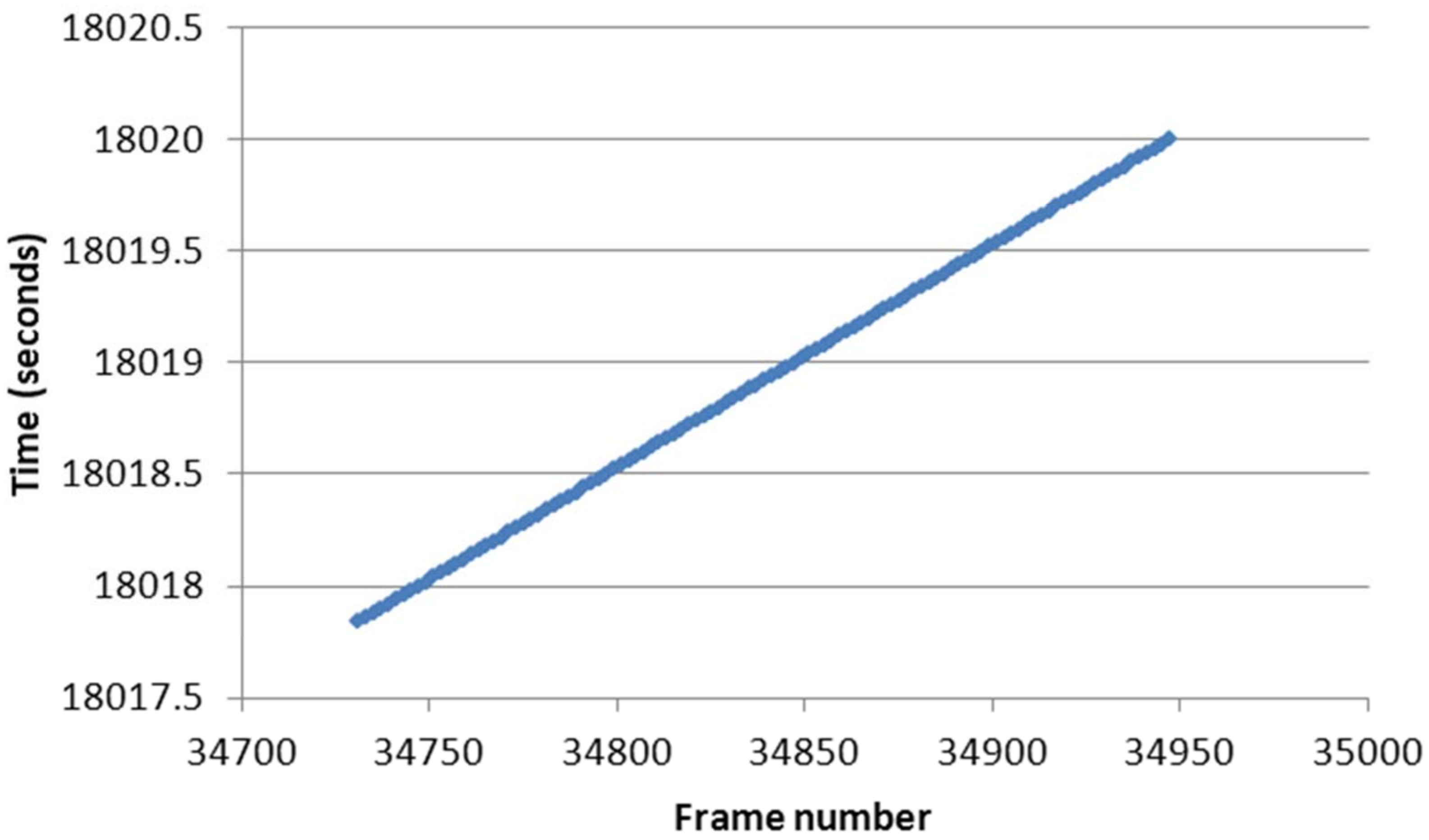
Adding grid lines



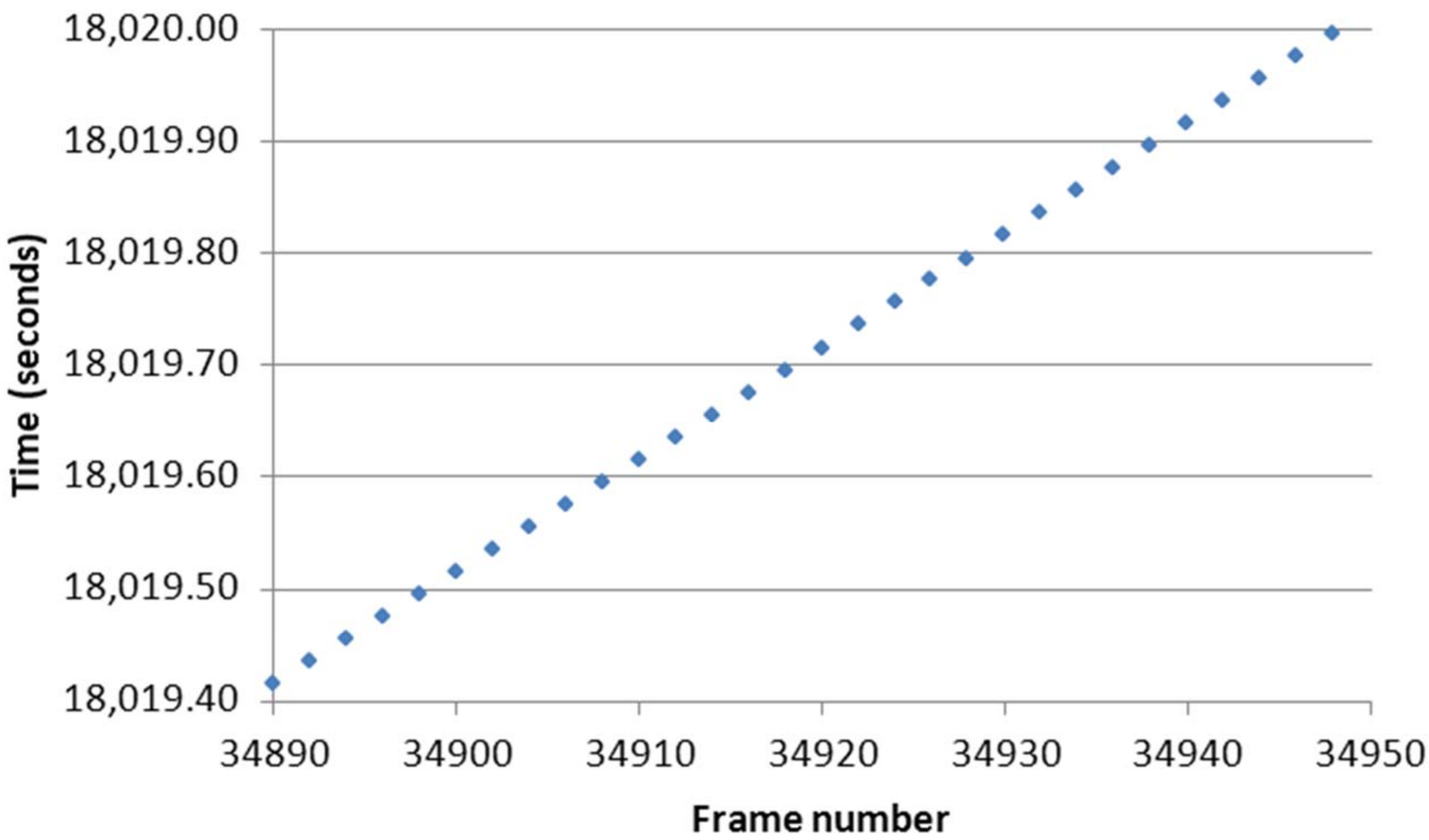
Are grid lines alone sufficient?



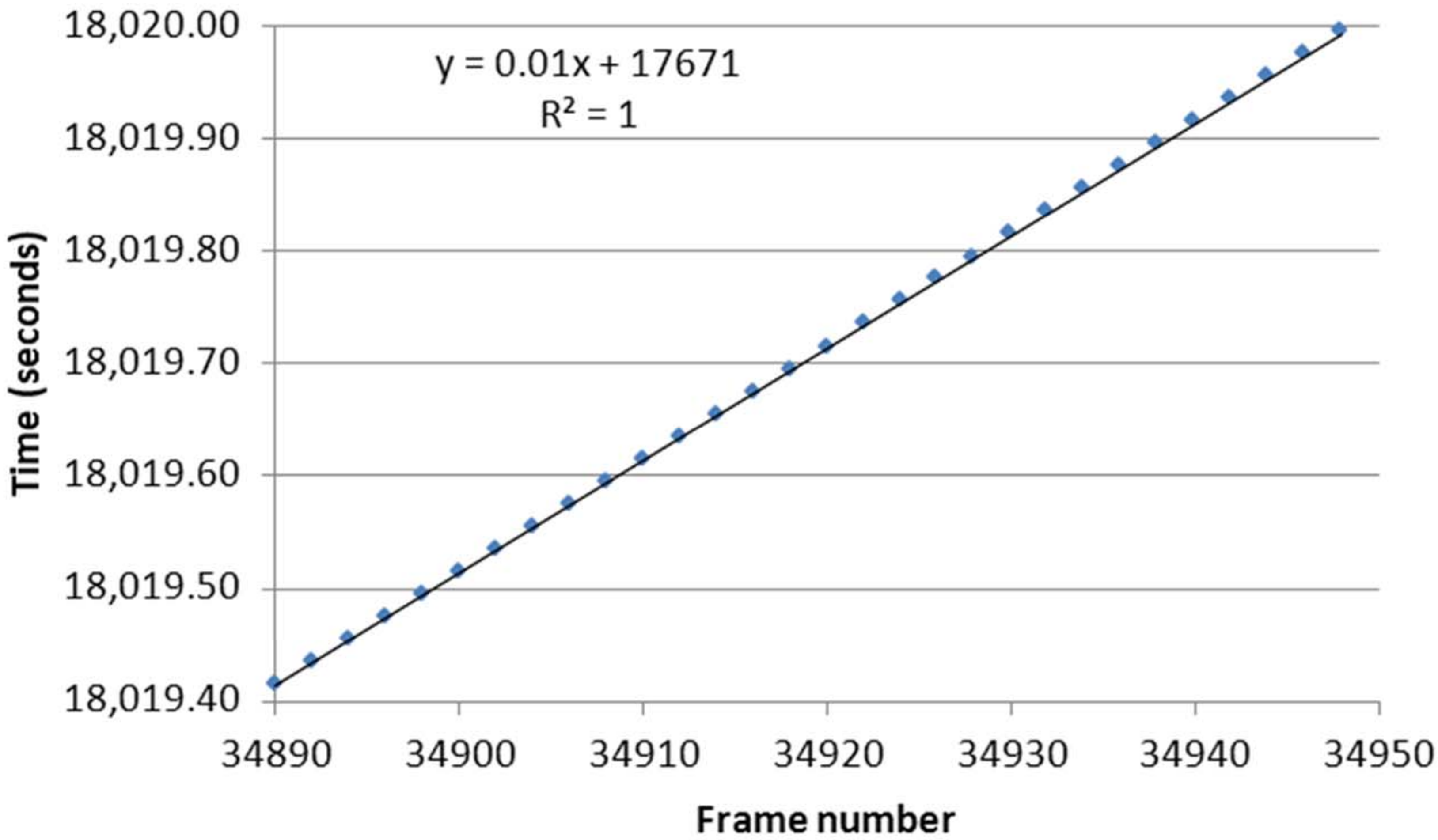
Scatter plots of frame # versus time



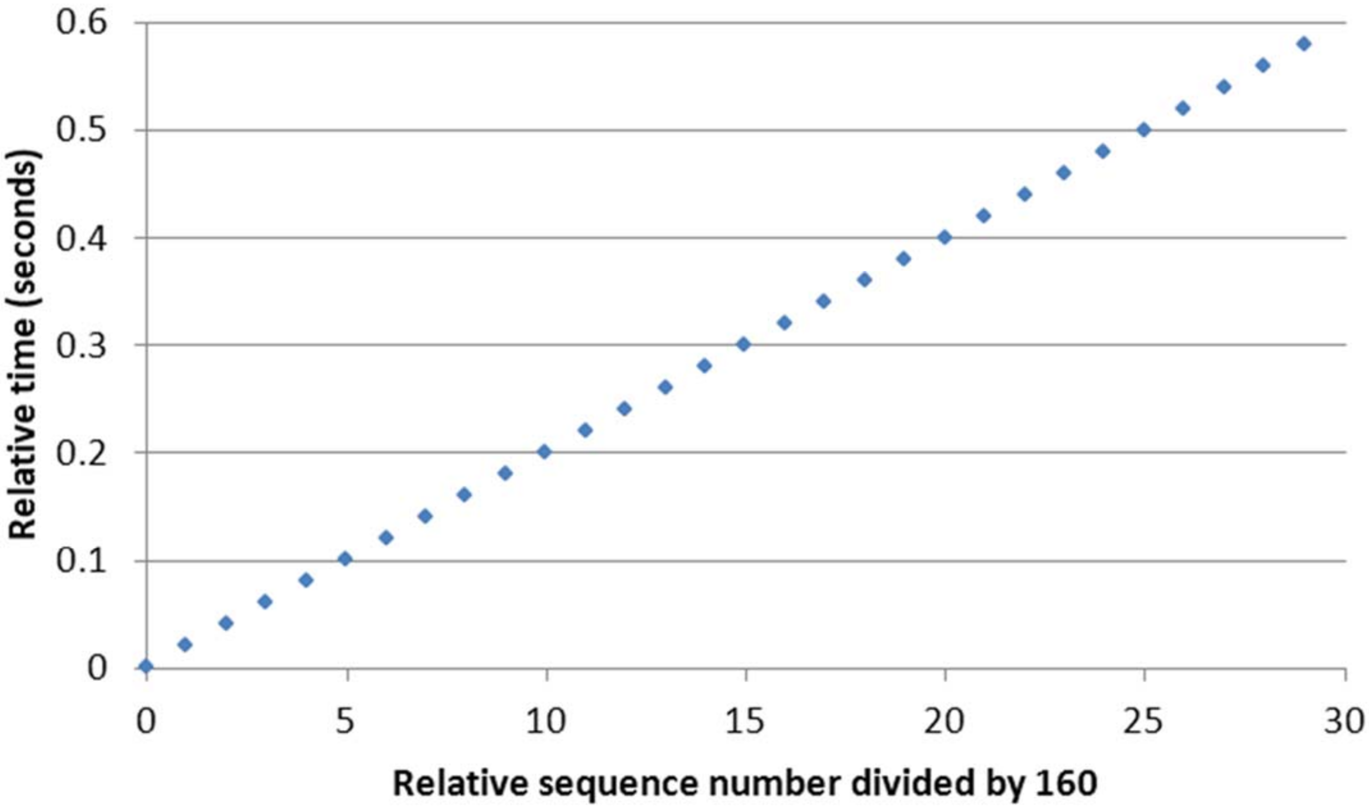
Zoom in on last few samples



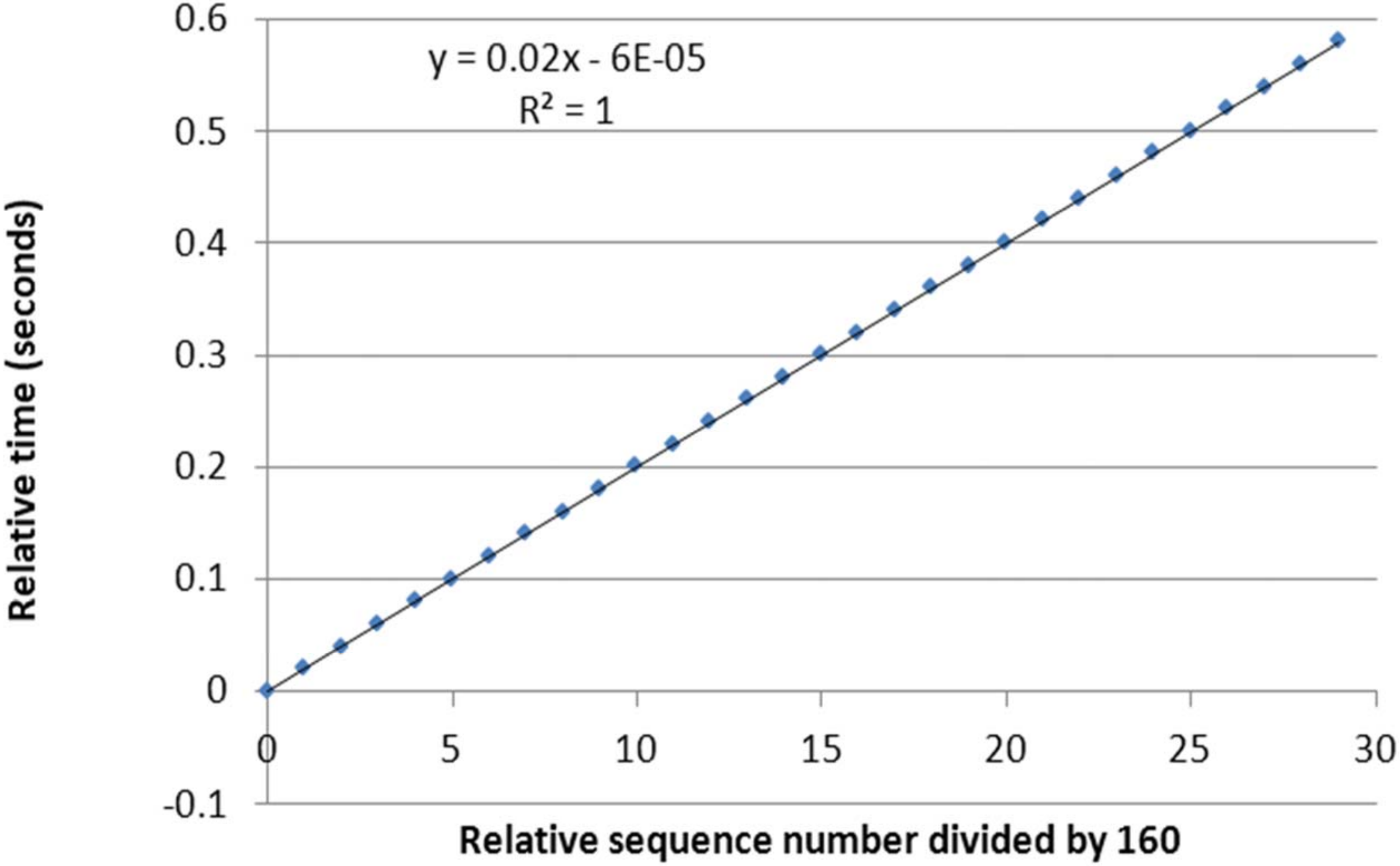
Add a trendline and show equation

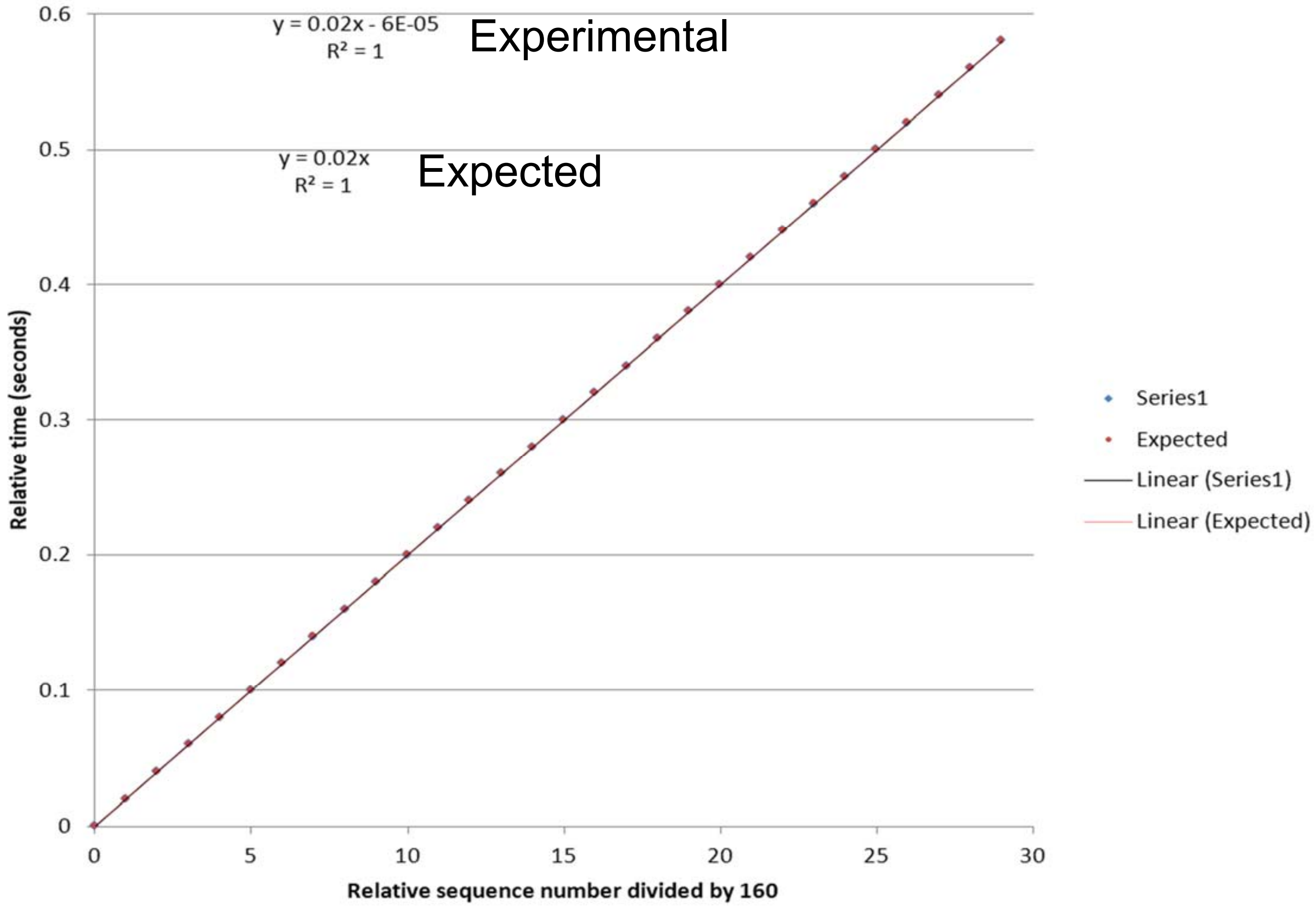


Computing new axis

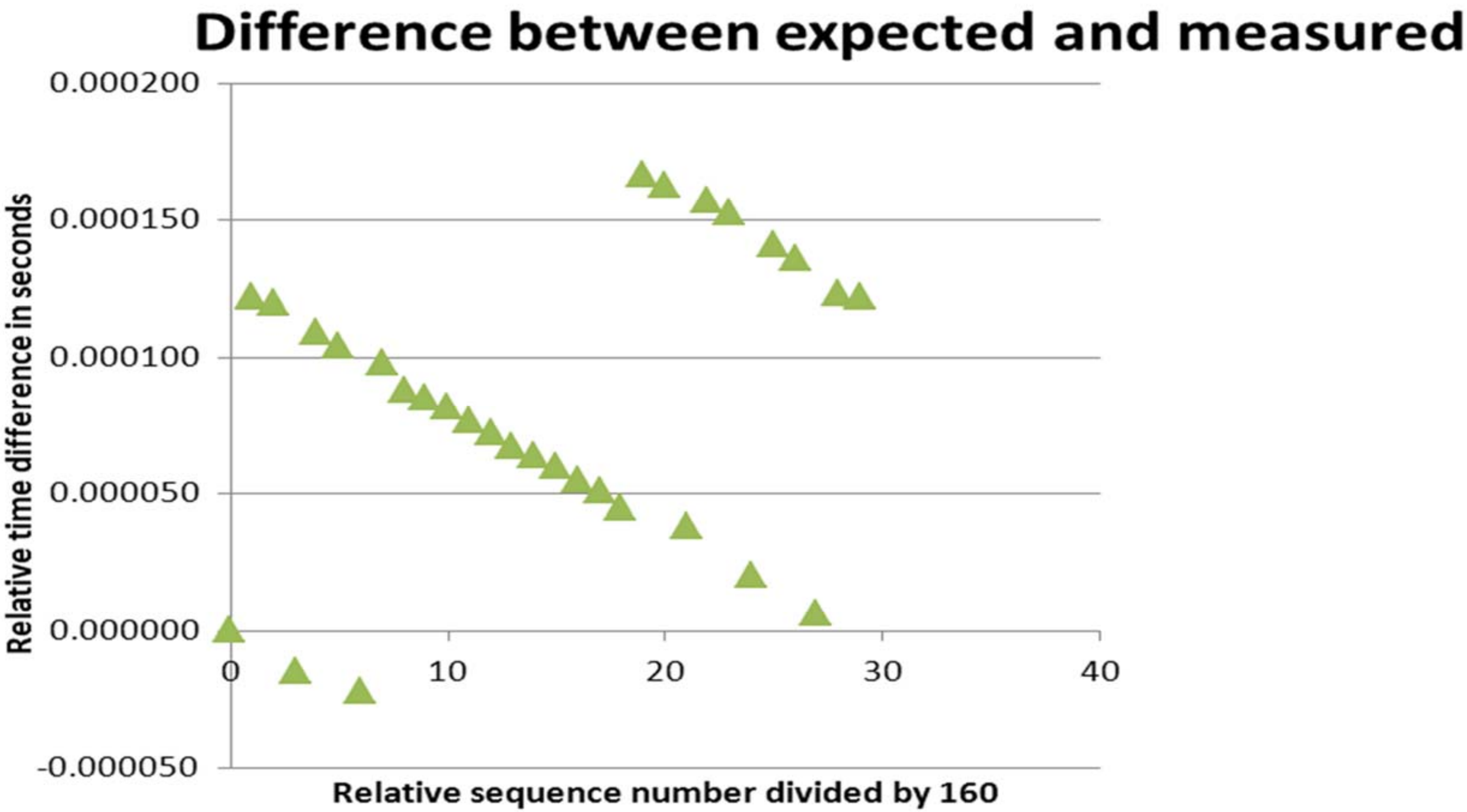


Now add the trendline

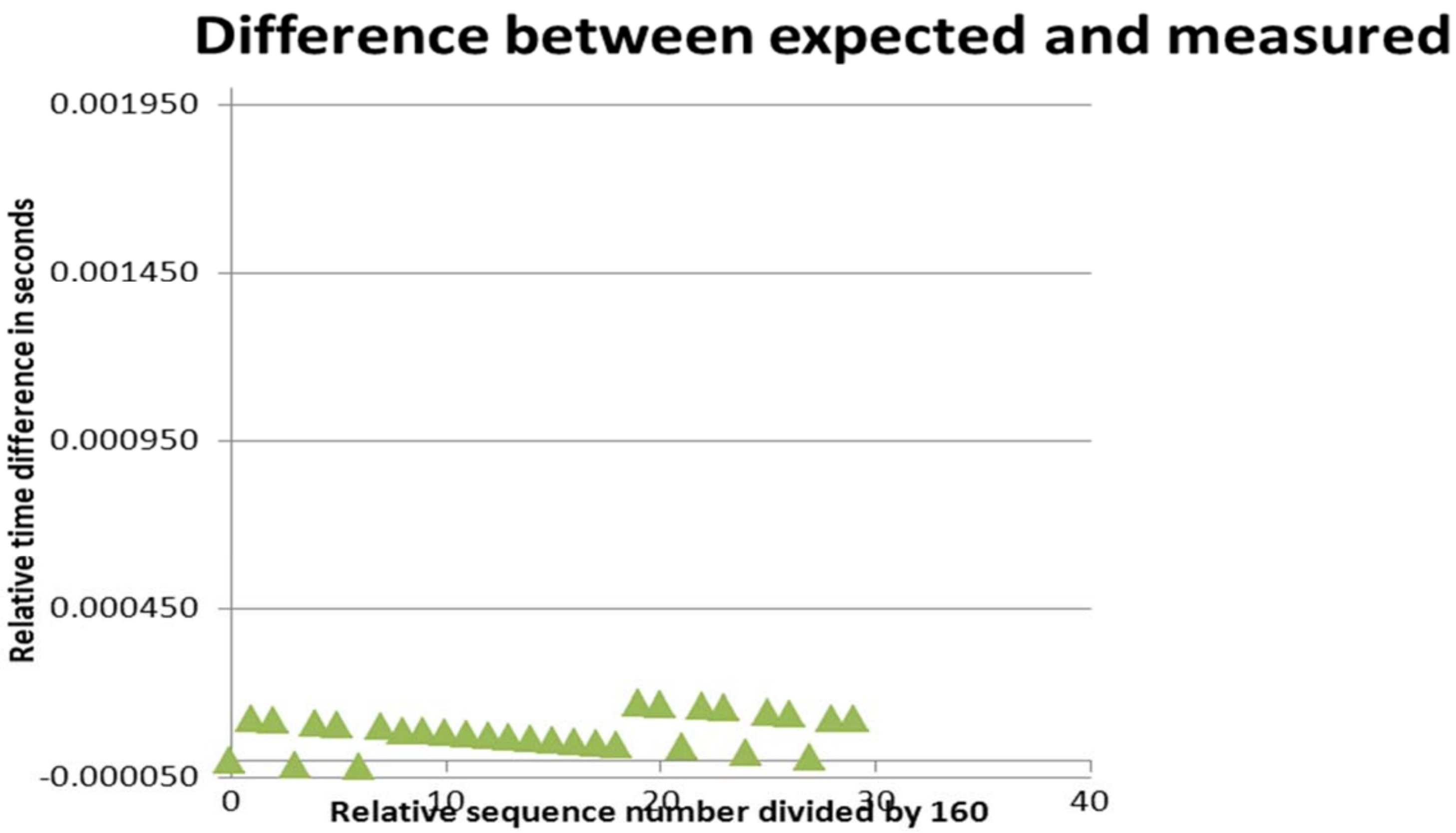



y = 0.02x - 6E-05
R² = 1
Experimental
y = 0.02x
R² = 1
Expected
0.6
0.5
0.4
0.3
0.2
0.1
0
Relative time (seconds)
0
5
10
15
20
25
30
Relative sequence number divided by 160
Series1
Expected
Linear (Series1)
Linear (Expected)

How does the measured data differ from the expected data?



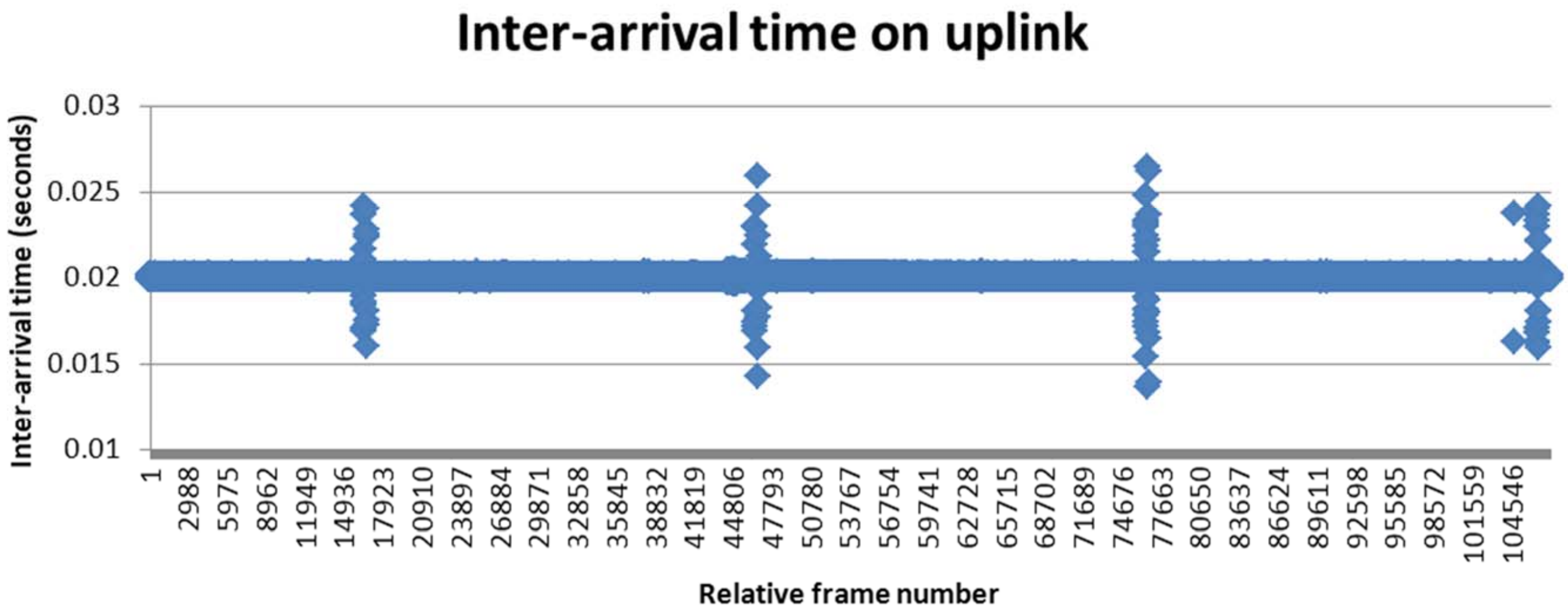
Does the difference matter? Plot scaled to 1/10 of the inter-arrival time period $\Rightarrow$ No



For traffic in the opposite direction

Mean	0.020000275
Standard Error	3.6743E-07
Median	0.020004
Mode	0.020005
Standard Deviation	0.000120472
Sample Variance	1.45135E-08
Kurtosis	670.0855429
Skewness	0.482218958
Range	0.012759
Minimum	0.013625
Maximum	0.026384
Sum	2150.109545
Count	107504
Confidence Level(95.0%)	7.20157E-07

Uplink inter-arrival times



What is going on?

Note the spikes near:		time in seconds	difference in time in seconds
16453		329.06	
46682		933.64	604.58
76657		1533.14	599.5
106512		2130.24	597.1

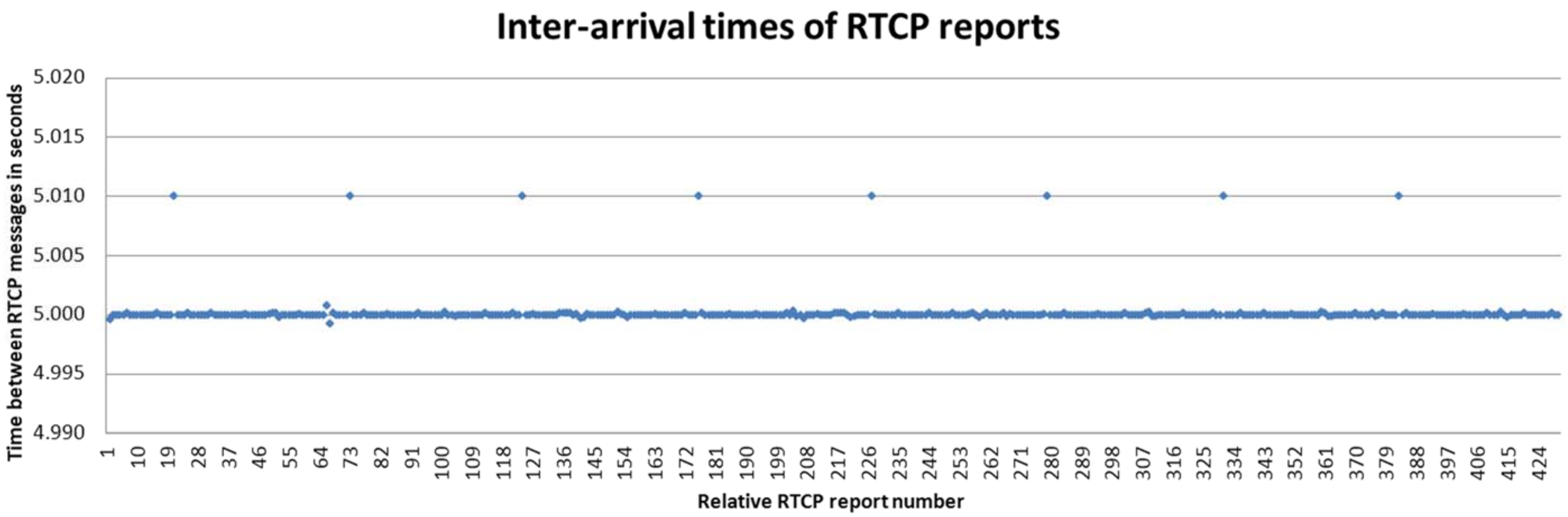
Q: What happens roughly every 600 seconds?

A: DHCP requests

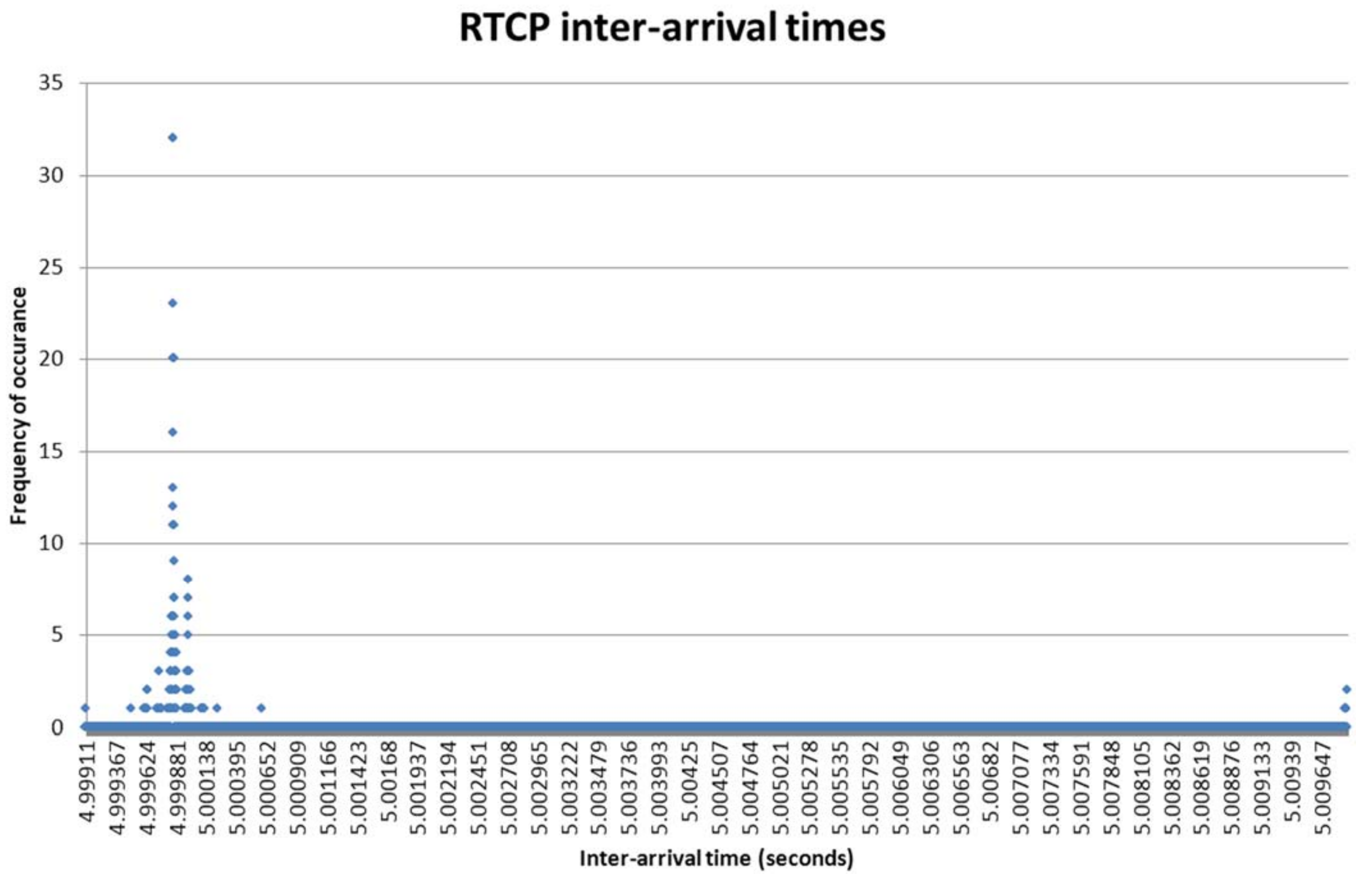
RTCP descriptive statistics

Mean	5.00006104
Standard Error	6.54393E-05
Median	4.999861
Mode	4.99986
Standard Deviation	0.001355399
Sample Variance	1.83711E-06
Kurtosis	48.80806181
Skewness	7.096344028
Range	0.010758
Minimum	4.99911
Maximum	5.009868
Sum	2145.026186
Count	429
Confidence Level(95.0%)	0.000128622

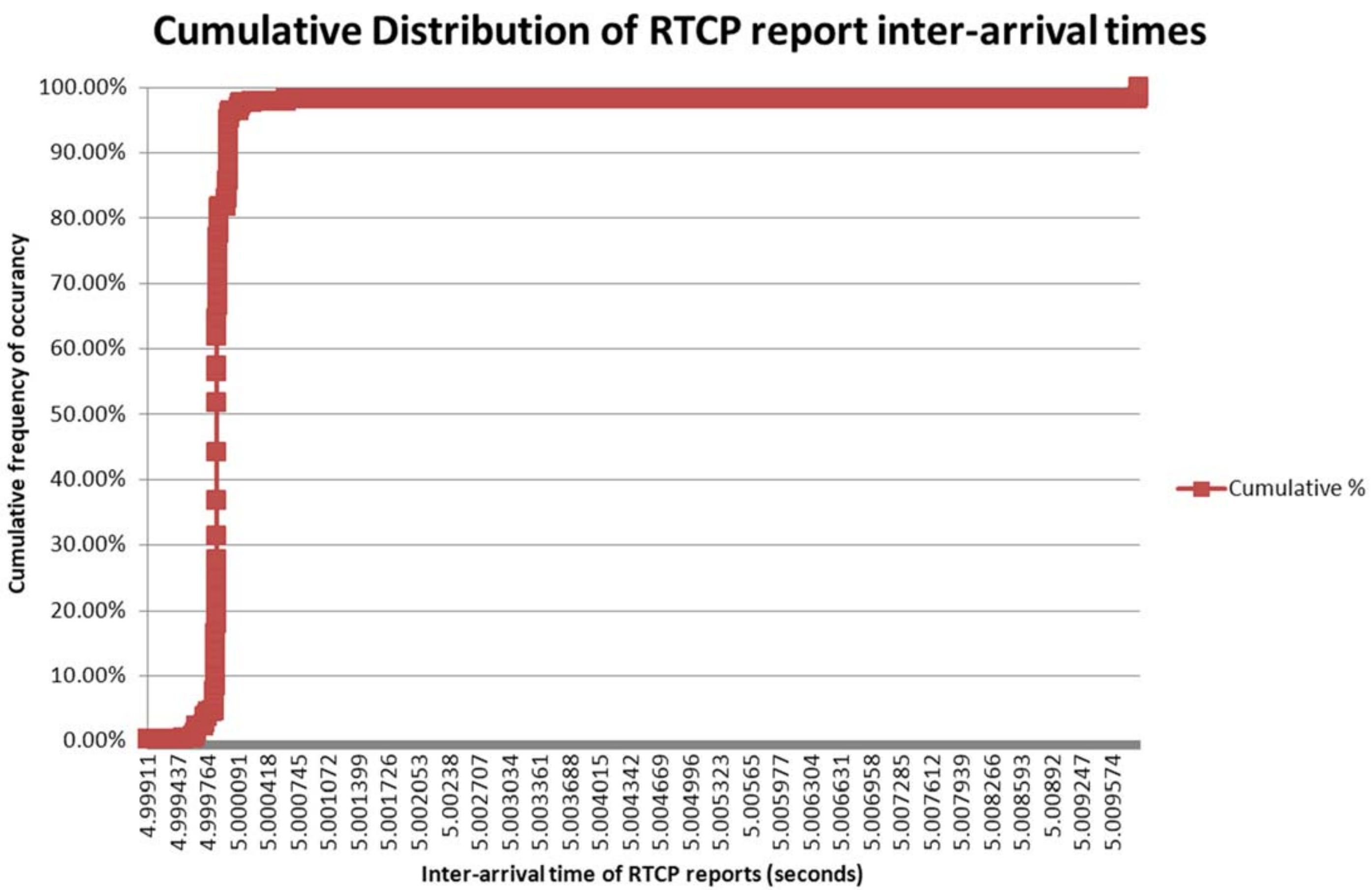
Plot of inter-arrival times of RTCP reports



Histogram of RTCP inter-arrivals



RTCP CDF



Cumulative Distribution of RTCP report inter-arrival times
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Cumulative frequency of occurancy
4.99911
4.999437
4.999764
5.000091
5.000418
5.000745
5.001072
5.001399
5.001726
5.002053
5.00238
5.002707
5.003034
5.003361
5.003688
5.004015
5.004342
5.004669
5.004996
5.005323
5.00565
5.005977
5.006304
5.006631
5.006958
5.007285
5.007612
5.007939
5.008266
5.008593
5.00892
5.009247
5.009574
Inter-arrival time of RTCP reports (seconds)
Cumulative %

Remarks

Some problems with Excel:

- It is not easy to change, add, or subtract data points without having to manual redo all the analysis
- It is not easy to write general functions in Excel and then use these over and over again as either more data comes in or the experiment is redesigned.

As an alternative, you might want to think about learning a different way of analyzing your data. While it might take some effort to learn this new method, it will stand you in good stead in your future work.

R

R is an open source successor to the statistics package S and Splus

- S was developed by the statisticians at Bell Labs to help them help others with their problems

Josef Freuwald (when a graduate student in Linguistics at the University of Pennsylvania, now Lecturer in Sociolinguistics in Linguistics and English at the University of Edinburgh) said:

> "Quite simply, R is the statistics software paradigm of our day. "

http://www.ling.upenn.edu/~joseff/rstudy/week1.html#why

And its free! Additionally, it supports Windows, Linux, and Mac OS

"As the Cantonese say, yauh peng, yauh leng, which means both inexpensive and beautiful."

– from Norman Matloff, The Art of R Programming: A Tour of Statistical Software Design [Matloff2011]

Commercial alternatives to R

Microsoft's Excel – we saw this earlier in the lecture

MathWorks' MATLAB – Statistics Toolbox™
http://www.mathworks.se/products/statistics/

Statistical Analysis with SAS/STAT® Software
http://www.sas.com/en_us/software/analytics/stat.html

IBM® SPSS® Advanced Statistics
http://www-03.ibm.com/software/products/en/spss-advanced-stats

Stata® http://stata.com/

TIBCO Spotfire S+® http://spotfire.tibco.com/

…

R Resources

Comprehensive R Archive Network (CRAN) http://cran.r-project.org/

Lots of tutorials:

- http://www.r-tutor.com/
- http://heather.cs.ucdavis.edu/~matloff/r.html
- …

R Packages

Lots of libraries called **packages**:

- Basic packages (included with the distribution): base, datasets, grDevices, graphics, grid, methods, splines, stats, stats4, tcltk, tools, utils

  http://cran.r-project.org/doc/FAQ/R-FAQ.html#Which-add_002don-packages-exist-for-R_003f

- Add-on packages from lots of others (including commercial packages such as https://plot.ly/)

…

Why use a programming language versus using a spreadsheet?

When you want to do something:

- **over and over again** and/or
- **systematically**

Experiment 1

Captured packets using Wireshark during a long (2150.12 second) VoIP call

⇒ at least: 107,505 RTP packets in each direction

⇒ 429 RTCP packets in one direction

http://www.Wireshark.org

Load the data, then extract relevant RTP packets

Starting with a tab separated file of the form:

"No."   "Time"   "Source"   "Destination“   "Protocol“   "RSSI"   "Info“

"1443"   "17685.760952"   "90.226.255.70"   "217.211.xx.xx"   "RTP"   ""   "PT=ITU-T G.711 PCMA, SSRC=0x6E21893F, Seq=183, Time=46386 “

```
data1<-read.table("one-call.tab", sep="\t", header=TRUE,
   stringsAsFactors = FALSE)
```

Extract the traffic going to me:

```
To_Chip<-subset(data1, Source == "90.226.255.70",
   drop=TRUE)
```

Extract only the RTP protocol packets:

```
To_Chip_RTP<-subset(To_Chip, Protocol == "RTP",
   drop=TRUE)
```

Summary

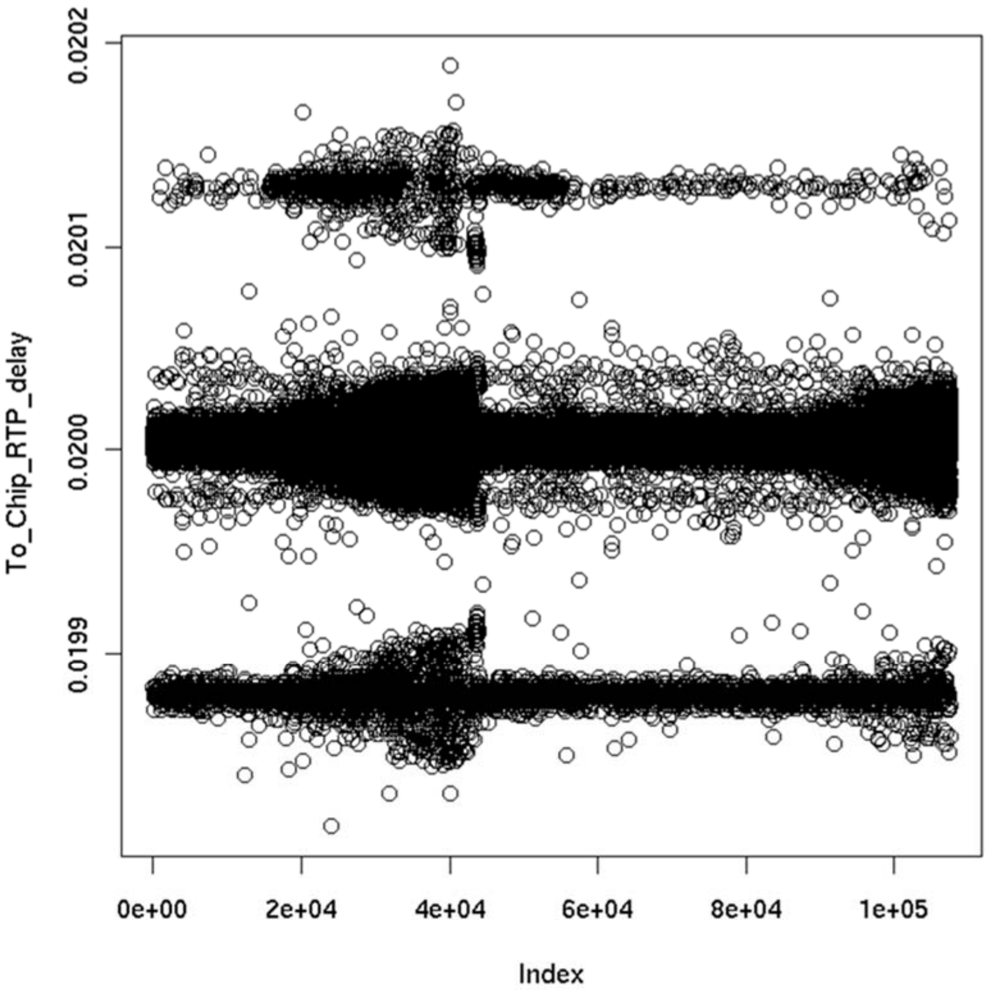
```
summary(To_Chip_RTP)
     No.                   Time                          Source
Min.   :  1443        Min.   :17686             90.226.255.70 :107515
1st Qu.: 55331        1st Qu.:18223             217.211.xx.xx :     0
Median :109224        Median :18761             41.209.78.223 :     0
Mean   :109223        Mean   :18761             62.20.251.42  :     0
3rd Qu.:163110        3rd Qu.:19298             81.228.11.66  :     0
Max.   :217022        Max.   :19836             90.226.251.20 :     0
                                                (Other)       :     0
Destination                Protocol         RSSI
217.211.47.125:107515      RTP    :107515   Mode:logical
41.209.78.223 :     0      ARP    :     0   NA's:107515
62.20.251.42  :     0      DHCP   :     0
81.228.11.66  :     0      ICMP   :     0
90.226.251.20 :     0      NTP    :     0
90.226.255.70 :     0      RTCP   :     0
(Other)       :     0      (Other):     0

Info
PT=ITU-T G.711 PCMA, SSRC=0x6E21893F, Seq=0, Time=10502866     :      1
PT=ITU-T G.711 PCMA, SSRC=0x6E21893F, Seq=10000, Time=12102866 :      1
PT=ITU-T G.711 PCMA, SSRC=0x6E21893F, Seq=10000, Time=1617106  :      1
PT=ITU-T G.711 PCMA, SSRC=0x6E21893F, Seq=10001, Time=12103026 :      1
PT=ITU-T G.711 PCMA, SSRC=0x6E21893F, Seq=10001, Time=1617266  :      1
PT=ITU-T G.711 PCMA, SSRC=0x6E21893F, Seq=10002, Time=12103186 :      1
(Other)                                                        :107509
```

Inter-arrival delays

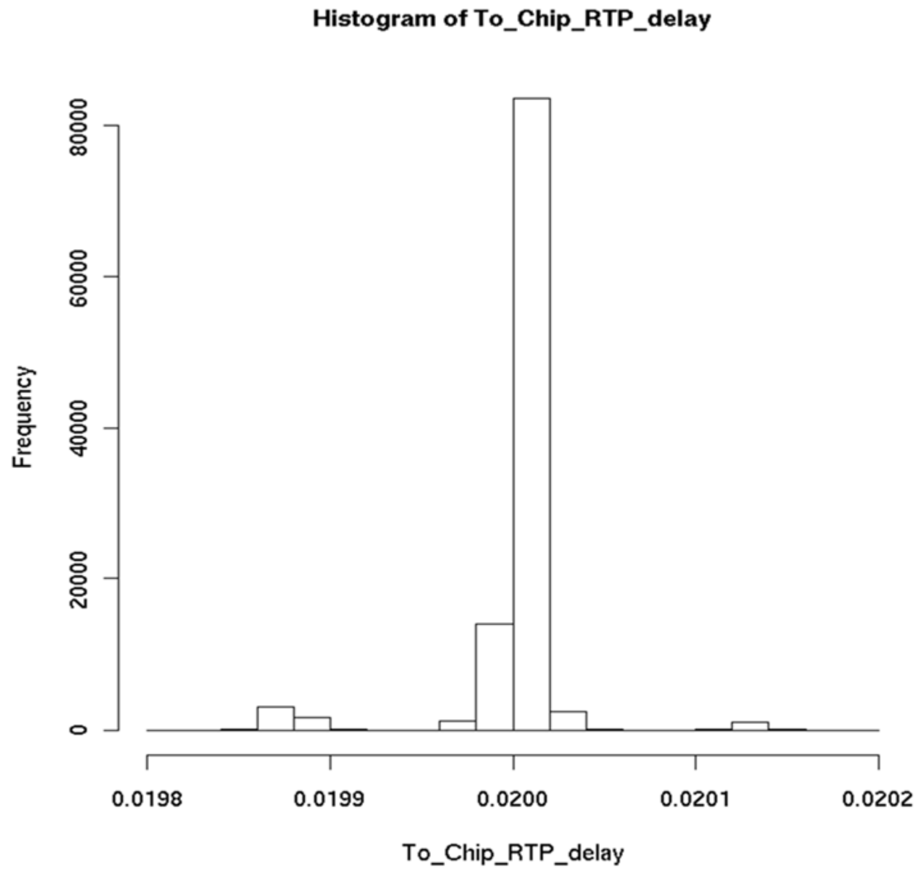
```
lvh<-nrow(To_Chip_RTP)
[1] 107515
lvh<-lvh-1> lvh
[1] 107514
To_Chip_RTP_delay=vector(length=(nrow(To_Chip_RTP)-1))
for (i in 1:lvh) {
To_Chip_RTP_delay[i]<-To_Chip_RTP$Time[i+1]-
  To_Chip_RTP$Time[i]
}

summary(To_Chip_RTP_delay)
   Min.    1st Qu.   Median     Mean    3rd Qu.    Max.
0.01981    0.02000   0.02000    0.02000  0.02001
  0.02019
```

(Note that these times are in seconds.)

plot(To_Chip_RTP_delay)



hist(To_Chip_RTP_delay)

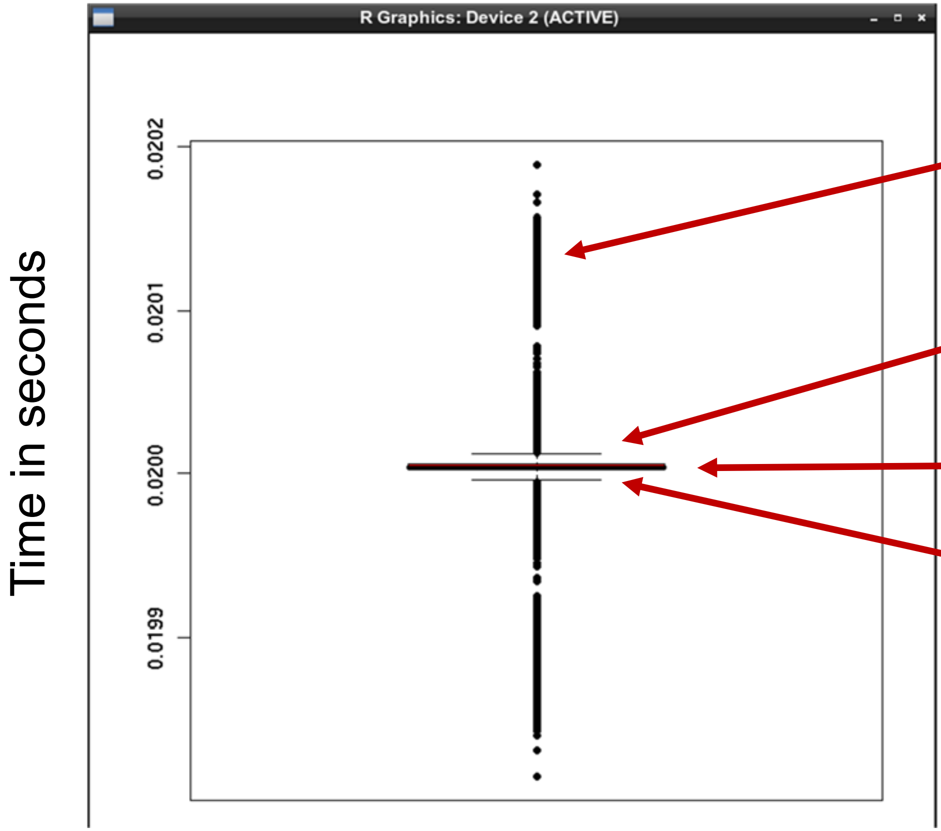


Time in Seconds

boxplot(To_Chip_RTP_delay, pch=20, col=3)

Symbol = bullet

Color 3 = Blue



Outliers

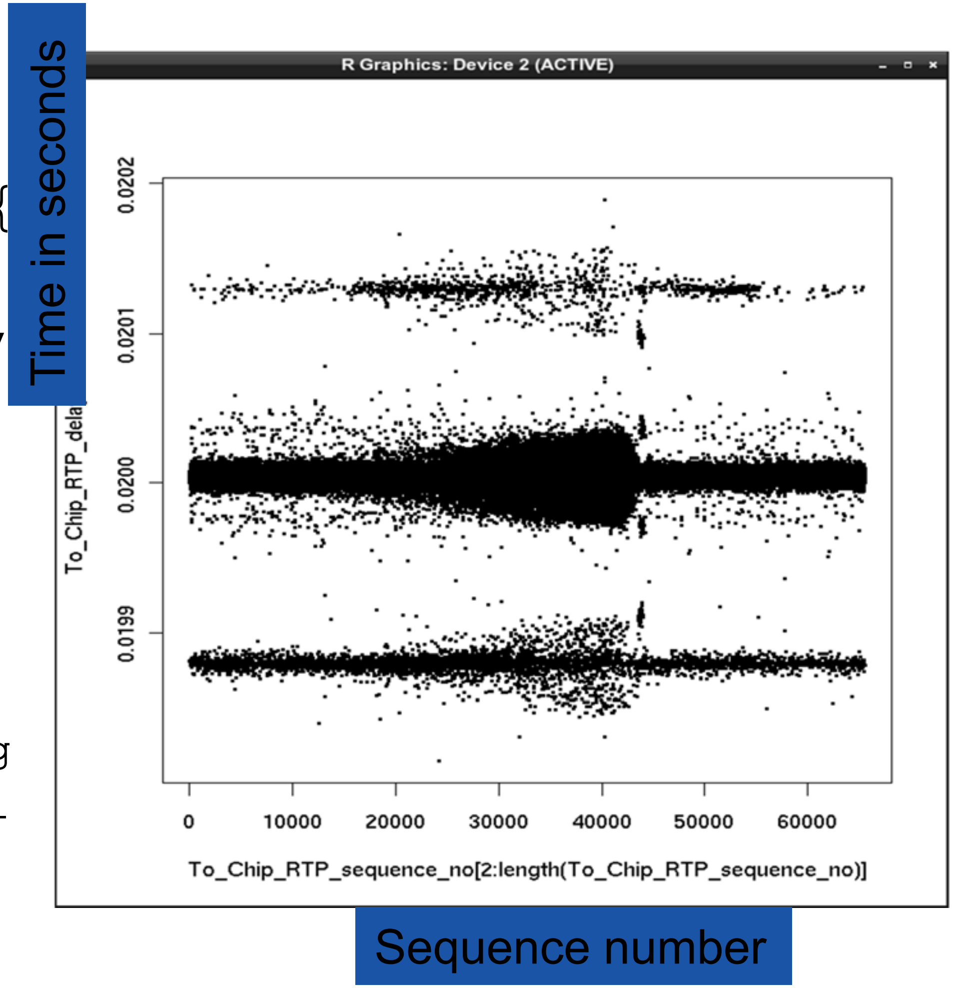
Upper quartile

Median

Lower quartile

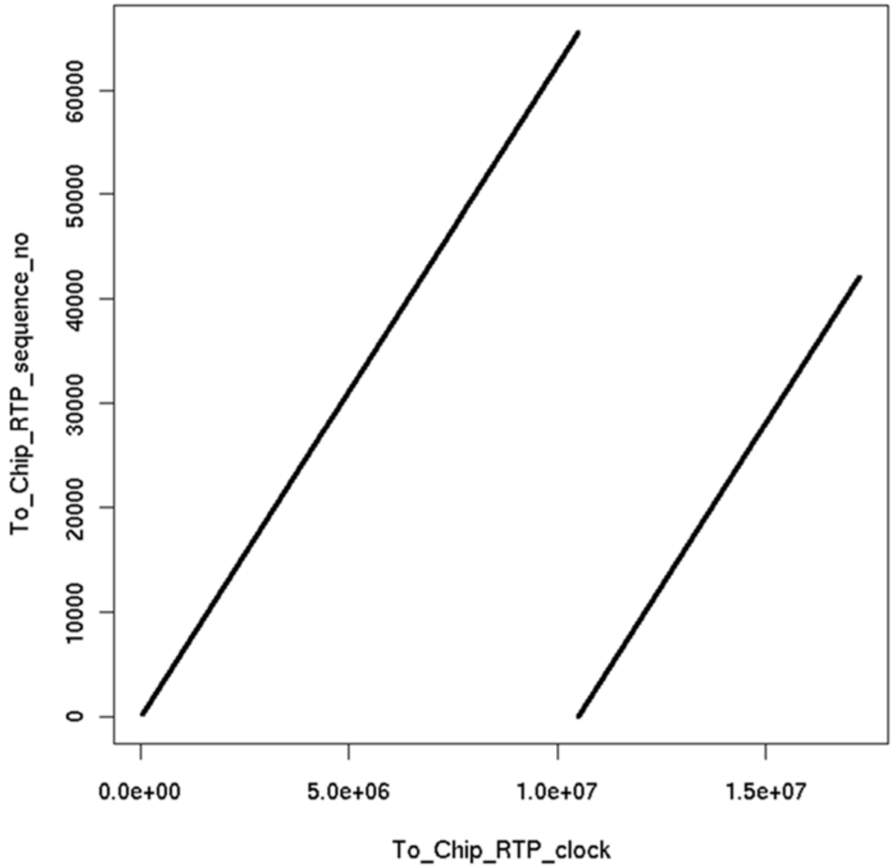
Interarrival delay vs. sequence

```
for (i in
   1:length(To_Chip_RTP$Info)) {
z1<-
   strsplit(To_Chip_RTP$Info[i],
   ",")
z2<-strsplit(z1[[1]][3], "=")
To_Chip_RTP_sequence_no[i]<-
   z2[[1]][2]
}

plot(To_Chip_RTP_sequence_no[2:leng
   th(To_Chip_RTP_sequence_no)],To_
   Chip_RTP_delay, pch=20,
   cex=0.25)
```



RTP Clock vs. sequence

```
To_Chip_RTP_clock<-1
for (i in
   1:length(To_Chip_RTP$Info)) {
z1<-
   strsplit(To_Chip_RTP$Info[i],
   ",")
z2<-strsplit(z1[[1]][4], "=")
To_Chip_RTP_clock[i] <-
   z2[[1]][2]
}
plot ( To_Chip_RTP_clock,
   To_Chip_RTP_sequence_no,
   pch=20, cex=0.25)
```



Inter-arrival times of RTP packets: From network to local user agent

Raw output fom Microsoft Excel 2010 (Beta)

Using Excel:

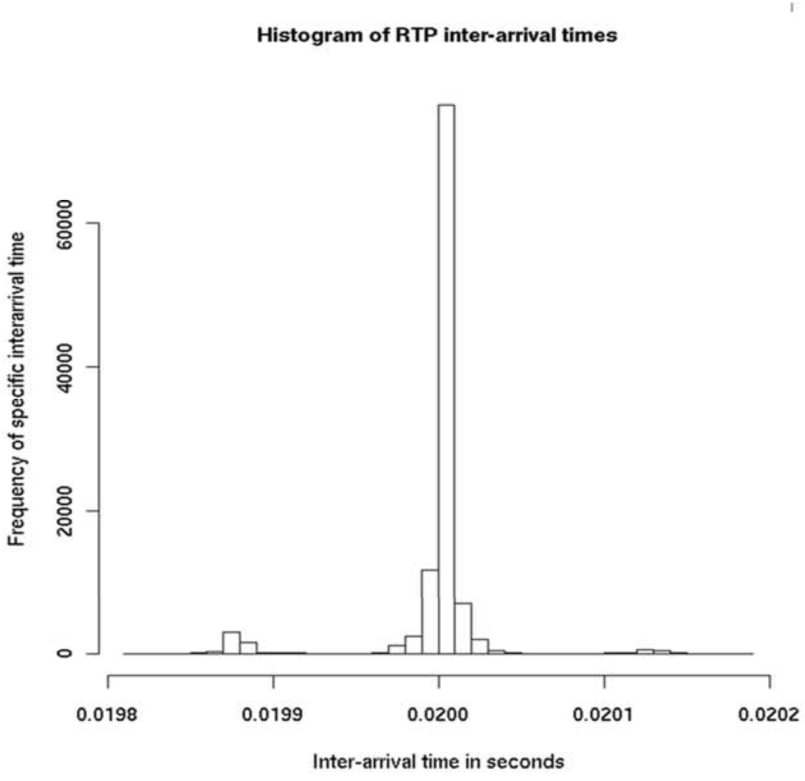
Mean	0.019999999
Standard Error	9.28526E-08
Median	0.020004
Mode	0.020005
Standard Deviation	3.04446E-05
Sample Variance	9.26874E-10
Kurtosis	12.36652501
Skewness	-2.054662184
Range	0.000374
Minimum	0.019815
Maximum	0.020189
Sum	2150.11991
Count	107506
Confidence Level(95.0%)	1.8199E-07

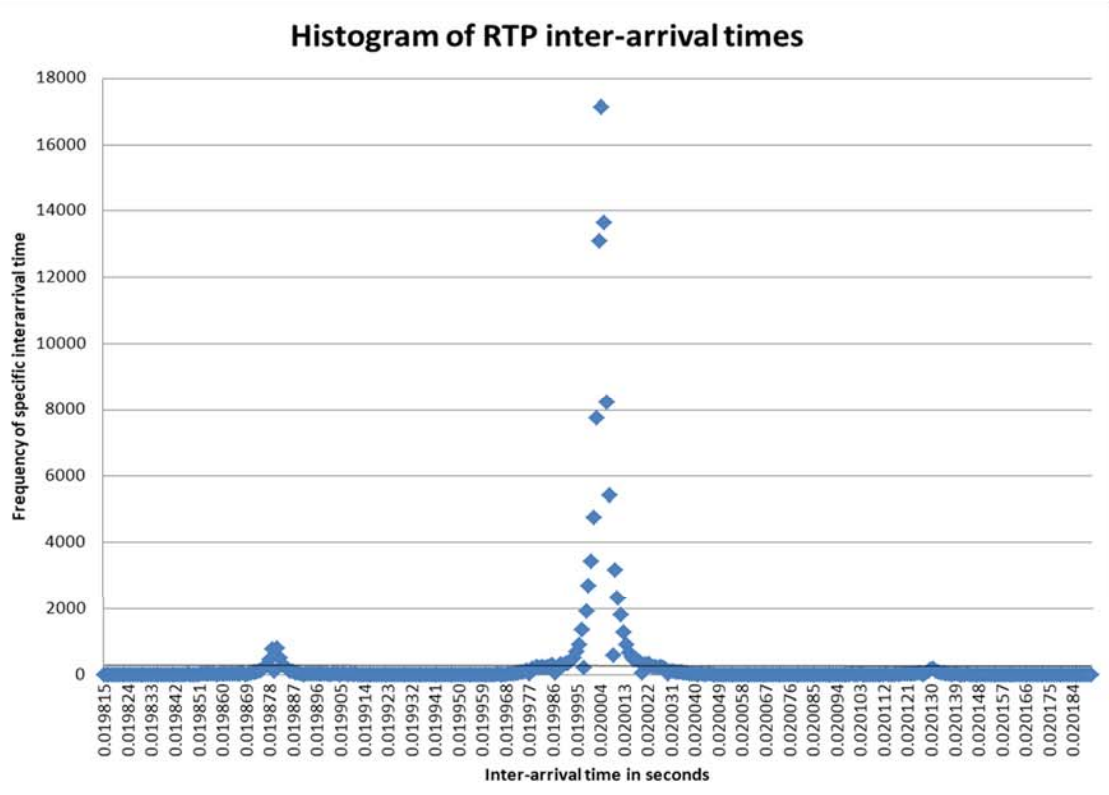
Note: count ≠ length and the two programs get a different value for kurtosis

Using R functions:

```
mean(To_Chip_RTP_delay): 0.02
library(plotrix);
std.error(To_Chip_RTP_delay): 9.284597e-08

The mode is the most frequently occurring
value (hence via
https://stat.ethz.ch/pipermail/r-help/1999-
December/005668.html):
names(sort(-table(To_Chip_RTP_delay)))[1]:
"0.0200049999984913"

sd(To_Chip_RTP_delay): 3.044357e-05
var(To_Chip_RTP_delay): 9.268109e-10
library(moments);
   kurtosis(To_Chip_RTP_delay): 15.36689
   skewness(To_Chip_RTP_delay): -2.054706
min(To_Chip_RTP_delay): 0.019815
max(To_Chip_RTP_delay): 0.020189
sum(To_Chip_RTP_delay): 2150.28
length(To_Chip_RTP_delay): 107514
qnorm(0.975)*std.error(To_Chip_RTP_delay):
1.819748e-07
```

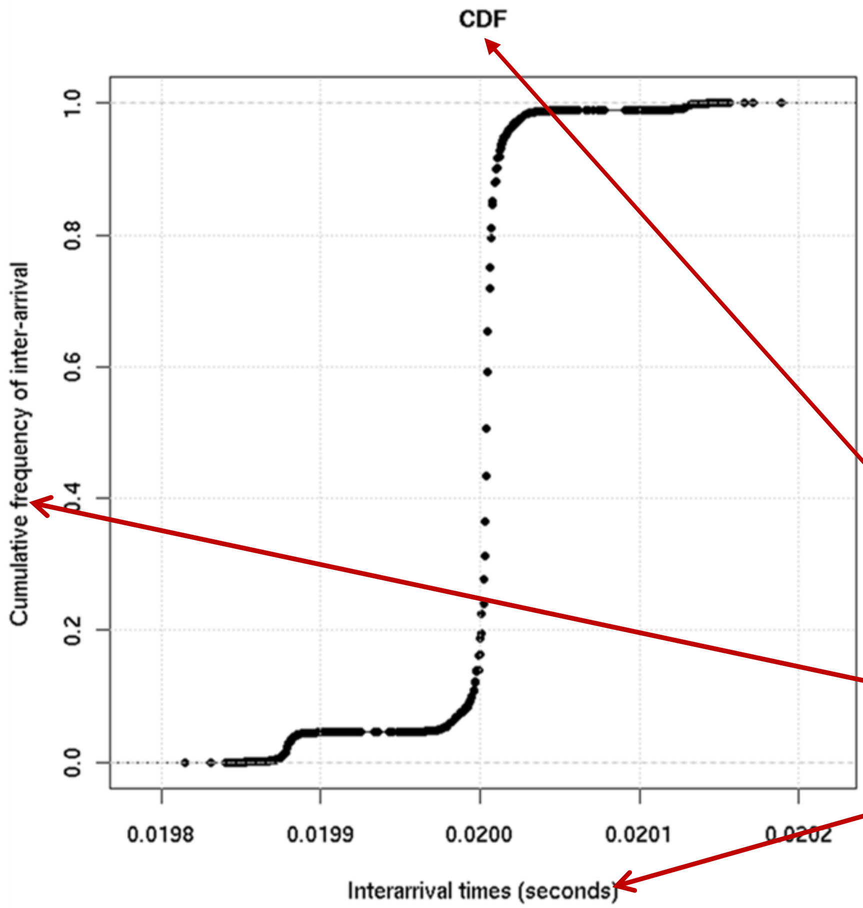
R vs. Excel histogram






```
hist(To_Chip_RTP_delay, ylab="Frequency of specific interarrival time",
xlab="Inter-arrival time in seconds", main="Histogram of RTP inter-arrival times", breaks=46)
```

Plot as a Cumulative Distribution (CDF)



plot(ecdf(To_Chip_RTP_delay), pch=20, cex=1, main="CDF", xlab="Interarrival times (seconds)", ylab="Cumulative frequency of inter-arrival"); grid()

cex = size of text or symbol for plot
1 = default
main = major label
ylab = y label
xlab = x label

grid() adds the grid in the background

With varying numbers of samples

Descriptive Statistics	First 100	First 1K	First 10K	First 100K
Mean	0.02000071	0.020000066	0.020000004	0.02
Standard Error	2.12714E-06	7.53406E-07	2.51164E-07	9.69855E-08
Median	0.020005	0.020004	0.020004	0.020004
Mode	0.020005	0.020005	0.020005	0.020005
Standard Deviation	2.12714E-05	2.38248E-05	2.51164E-05	3.06695E-05
Sample Variance	4.52471E-10	5.67621E-10	6.30831E-10	9.40618E-10
Kurtosis	28.87137928	21.46428225	19.07376827	12.23083198
Skewness	-5.453831468	-4.509853108	-3.831289593	-2.003065575
Range	0.000135	0.000252	0.000277	0.000374
Minimum	0.01988	0.019872	0.019868	0.019815
Maximum	0.020015	0.020124	0.020145	0.020189
Sum	2.000071	20.000066	200.000044	1999.999951
Count	100	1000	10000	100000
Confidence Level(95.0%)	4.2207E-06	1.47844E-06	4.92331E-07	1.9009E-07

With varying numbers of samples

Descriptive Statistics	First 100	First 1K	First 10K	First 100K
Mean				
Standard Error				
Median				
Mode				
Standard Deviation				
Sample Variance				
Kurtosis				
Skewness				
Range				
Minimum				
Maximum				
Sum				
Count				
Confidence Level(95.0%)				

```
foo<-function(n){
v <-1:12
v[1]=mean(To_Chip_RTP_delay[1:n])
v[2]=std.error(To_Chip_RTP_delay[1:n])
v[3]=names(sort(-table(To_Chip_RTP_delay[1:n])))[1]
v[4]=sd(To_Chip_RTP_delay[1:n])
v[5]=var(To_Chip_RTP_delay[1:n])
v[6]=kurtosis(To_Chip_RTP_delay[1:n])
v[7]=skewness(To_Chip_RTP_delay[1:n])
v[8]=min(To_Chip_RTP_delay[1:n])
v[9]=max(To_Chip_RTP_delay[1:n])
v[10]=sum(To_Chip_RTP_delay[1:n])
v[11]=length(To_Chip_RTP_delay[1:n])
v[12]=qnorm(0.965)*std.error(To_Chip_RTP_delay[1:n])
return(v)}
seq1<-c(foo(100),foo(1000),foo(10000),foo(100000))
mat1<-matrix(seq1,  ncol=4)
```

Applying a function to a list of arguments

Descriptive Statistics	First 100	First 1K	First 10K	First 100K
Mean	0.02000071	0.020000066	0.020000004	0.02
Standard Error				
Median				
Mode				
Standard Deviation				
Sample Variance				
Kurtosis				
Skewness				
Range				
Minimum				
Maximum				
Sum				
Count				
Confidence Level(95.0%)				

```
foo<-function(m,n){v <- 1:12
v[1]=mean(m[1:n])
v[2]=std.error(m[1:n])
v[3]=names(sort(-table(m[1:n])))[1]
v[4]=sd(m[1:n])
v[5]=var(m[1:n])
v[6]=kurtosis(m[1:n])
v[7]=skewness(m[1:n])
v[8]=min(m[1:n])
v[9]=max(m[1:n])
v[10]=sum(m[1:n])
v[11]=length(m[1:n])
v[12]=qnorm(0.965)*std.error(m[1:n])
return(v)}

fee<-function(n) {foo(To_Chip_RTP_delay, 10^n)}

lapply(c(2:5), fee)
[[1]] [1] "0.0200006800000119"   "2.12697347407497e-06" "0.0200049999984913“
      [4] "2.12697347407497e-05" "4.52401615941855e-10" "30.3672958382318“ …
```

Uplink inter-arrival times stats

```
library(plotrix);library(moments)foo
<-function(m,n){v <- 1:12
v[1]=mean(m[1:n])
v[2]=std.error(m[1:n])
v[3]=names(sort(-table(m[1:n])))[1]
v[4]=sd(m[1:n])
v[5]=var(m[1:n])
v[6]=kurtosis(m[1:n])
v[7]=skewness(m[1:n])
v[8]=min(m[1:n])
v[9]=max(m[1:n])
v[10]=sum(m[1:n])
v[11]=length(m[1:n])
v[12]=qnorm(0.965)*std.error(m[1:n])
return(v)}
```

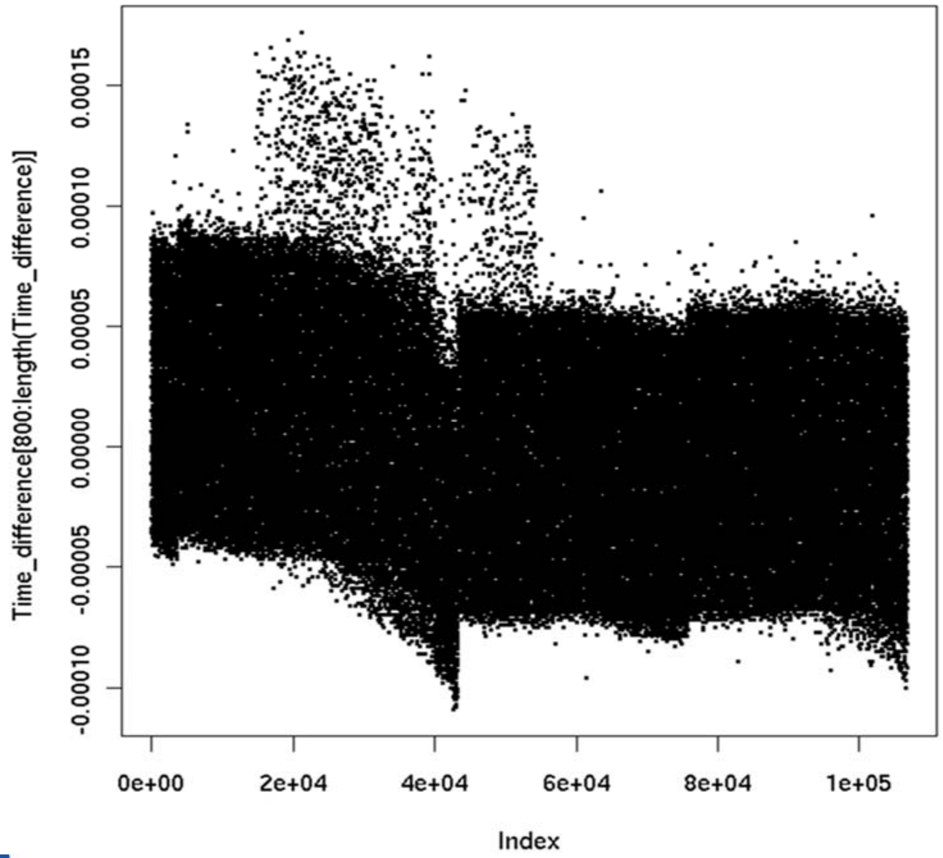
foo(From_Chip_RTP_delay, 10^5)	What to put into a report:
"0.02000027577"	0.020000 s
"3.63331229733734e-07"	3.63e-07 s
"0.0200049999984913"	0.020005 s
"0.000114895423102849"	0.000115 s
"1.32009582499827e-08"	1.32e-08 s
"742.58155664333"	742.58
"0.633658007213615"	0.634
"0.0136249999995925"	0.013625 s
"0.0263840000006894"	0.026384 s
"2000.027577"	2000.027577 s
"100000"	100000
"6.58323732971544e-07"	6.58e-07 s

Truncated to meaningful number of digits, added units, decimal align the numbers, set in fixed width font (Courier)

How does the measured data differ from the expected data?

```
for (i in
   1:length(To_Chip_RTP$Time)) {
Time_difference[i]=
(To_Chip_RTP$Time[i]-To_Chip_RTP$Time[1])-
   ((as.numeric(To_Chip_RTP_clock[i])-
   as.numeric(To_Chip_RTP_clock[1]))/8000)
}
plot( Time_difference[800:
        length(Time_difference)]
   , pch=20, cex=0.25)
```

Scale the bullet to ¼ size

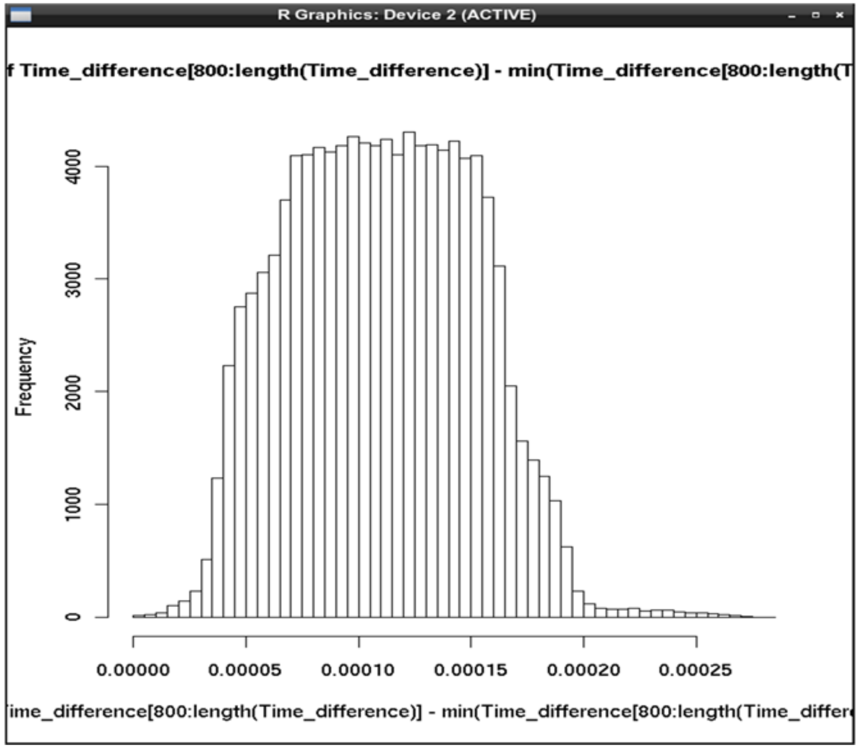


How does the measured data differ from the expected data?

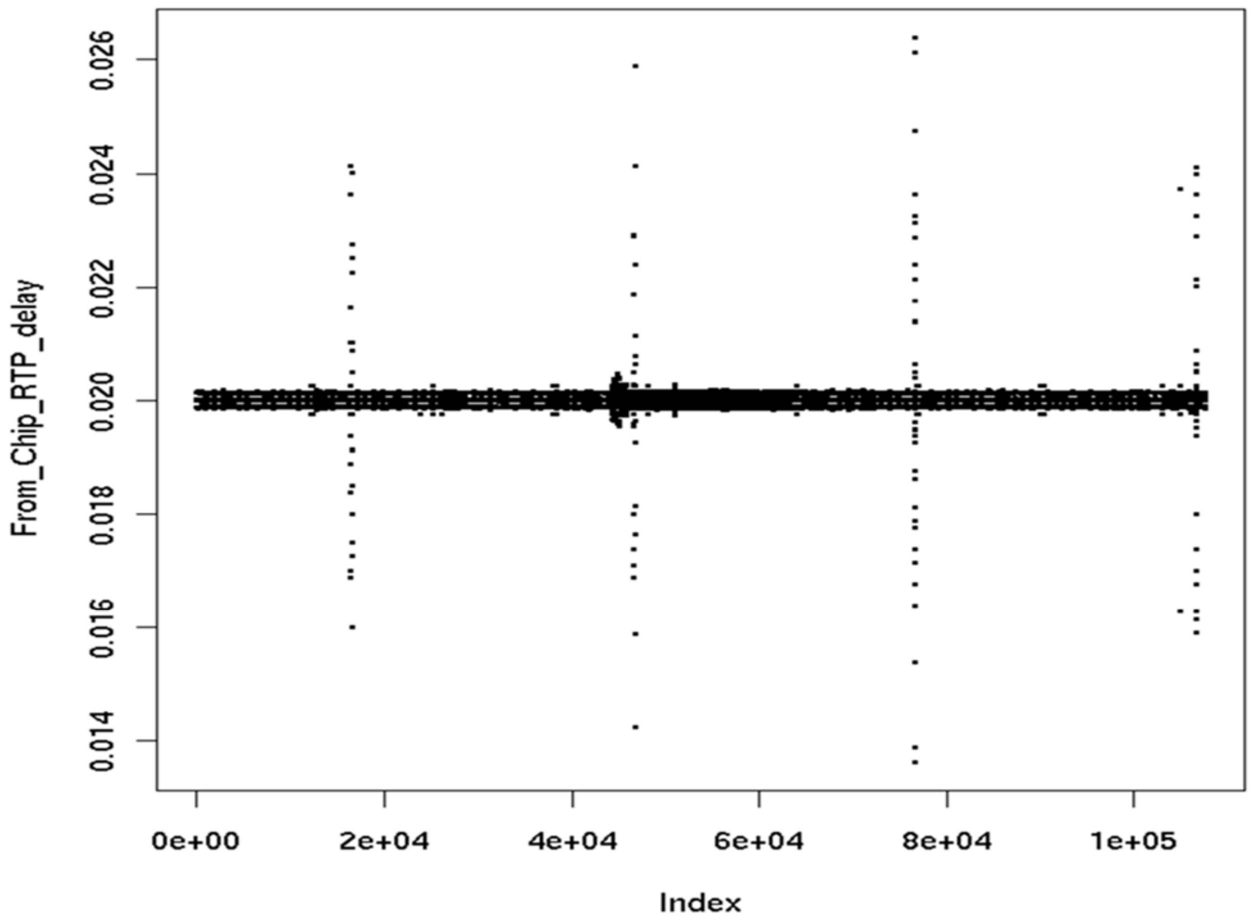
Since delay can not be negative, the real difference can be found by subtracting the min() ⇒

```
hist(
  Time_difference[800:length(Time_difference)]-
  min(Time_difference[800:
  length(Time_difference)]  ),
  breaks=100)
```

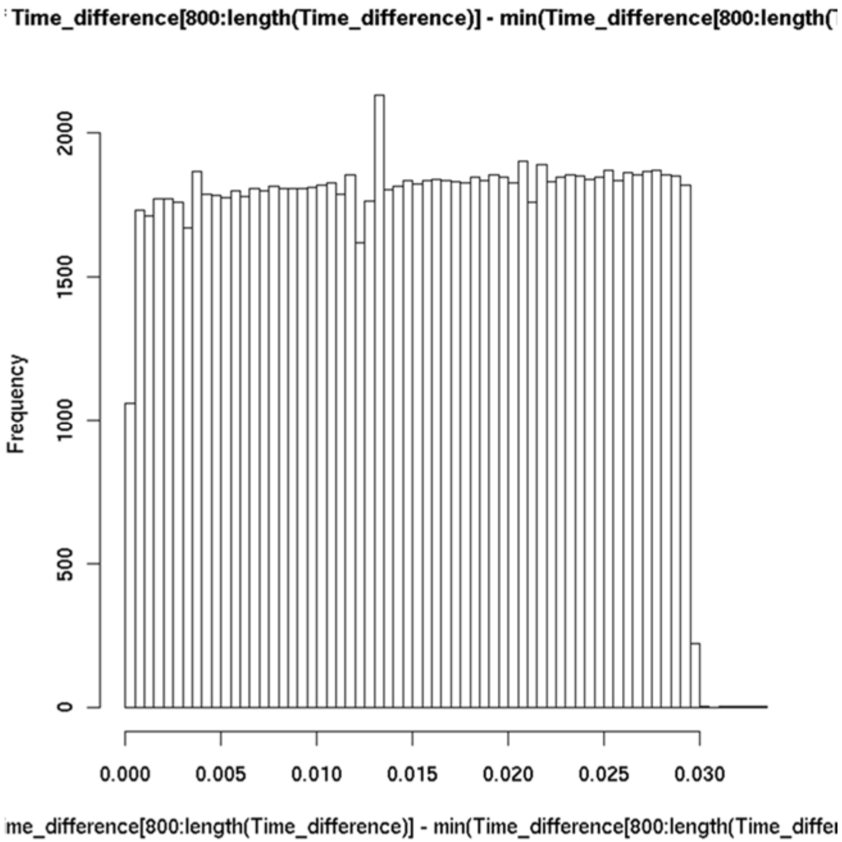
Number of bins to use

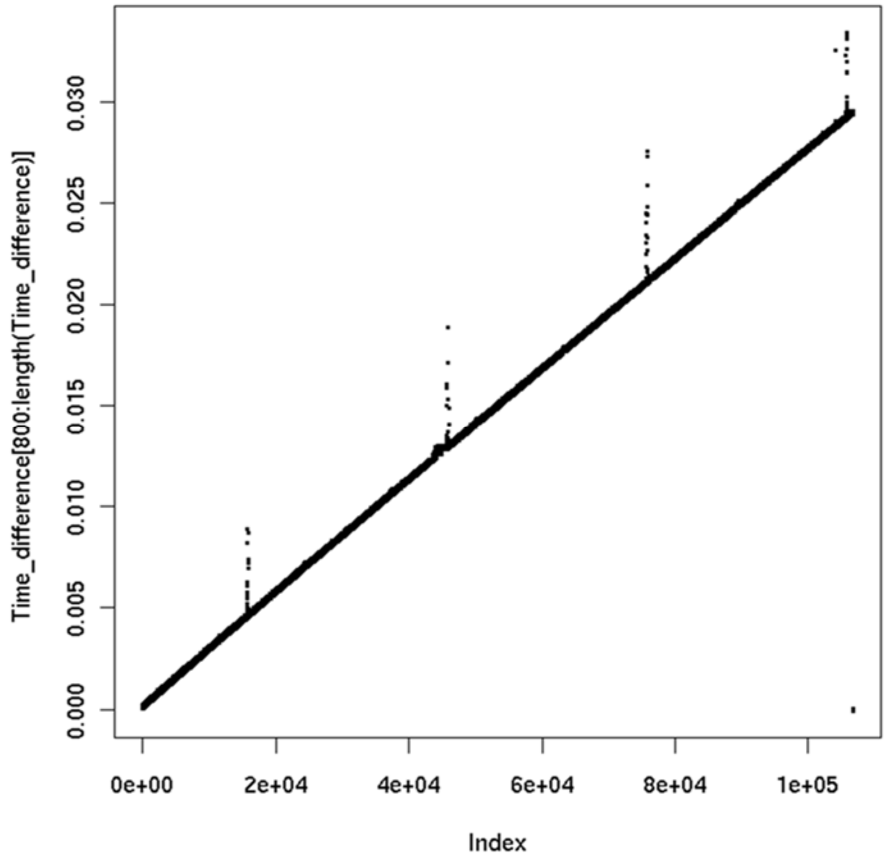


Uplink inter-arrival times



For traffic in the opposite direction





Difference histogram and difference plot ⇒ the clock is drifting wrt the Wireshark clock

References

[Chromy2005] James R. Chromy and Savitri Abeyasekera, 'Chapter XIX: Statistical analysis of survey data', in *Household Sample Surveys in Developing and Transition Countries*, New York, N.Y.: United Nations, 2005 [Online]. Available: http://www.cpc.unc.edu/projects/addhealth/data/guides/weight1.pdf

[Faraway2004] Julian J Faraway, *Linear Models with R.* London: Chapman & Hall/CRC, 2004, ISBN: 978-0-203-50727-8 [Online]. Available: http://www.myilibrary.com?id=23179. [Accessed: 03-Aug-2015]

[Goldvasser2012] Dov Goldvasser, Marilyn E. Noz, G.Q. Maguire, Henrik Olivecrona, Charles R. Bragdon, and Henrik Malchau, 'A New Technique for Measuring Wear in Total Hip Arthroplasty Using Computed Tomography', *The Journal of Arthroplasty*, vol. 27, no. 9, pp. 1636–1640.e1, Oct. 2012. DOI: 10.1016/j.arth.2012.03.053

[McCown2015] Frank McCown, 'Producing Simple Graphs with R', 30-Apr-2015. [Online]. Available: http://www.harding.edu/fmccown/r/ . [Accessed: 03-Aug-2015]

[Matloff2008] Norman Matloff, 'R for Programmers', 27-Nov-2008. [Online]. Available: http://heather.cs.ucdavis.edu/~matloff/R/RProg.html . [Accessed: 03-Aug-2015]

[Matloff2011] Norman Matloff, *The Art of R Programming: A Tour of Statistical Software Design*, 1st ed. San Francisco, CA, USA: No Starch Press, 2011, ISBN: 1-59327-384-3.

[Matloff2013] Norm Matloff, 'A Course in Probabilistic and Statistical Modeling in Computer Science', 23-Jun-2013. [Online]. Available: http://heather.cs.ucdavis.edu/~matloff/probstatbook.html . [Accessed: 03-Aug-2015]

[Matloff2015] Norm Matloff, 'From Algorithms to Z-Scores: Probabilistic and Statistical Modeling in Computer Science', 05-Jul-2015. [Online]. Available: http://heather.cs.ucdavis.edu/~matloff/132/PLN/ProbStatBook.pdf . [Accessed: 03-Aug-2015]

[Matloff2015a] Norm Matloff, 'Getting Started with the R Data Analysis Package', 02-Jul-2015. [Online]. Available: http://heather.cs.ucdavis.edu/~matloff/r.html . [Accessed: 03-Aug-2015]

References

[Raab2004] Gillian Raab, 'Background to P|E|A|S project', Napier University, 09-Sep-2004 [Online]. Available: http://www2.napier.ac.uk/depts/fhls/peas/workshops/workshop1presentationGR.ppt

[Sackett2001] David L. Sackett, 'Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!)', Canadian Medical Association Journal (*CMAJ)*, vol. 165, no. 9, pp. 1226–1237, Oct. 2001. PubMedID (PMID): 11706914

[Tullis 2008] Tom Tullis and Bill Albert, *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Amsterdam ; Boston: Elsevier/Morgan Kaufmann, 2008, ISBN: 978-0-12-373558-4.

[Walonick2010] David S. Walonick, 'A Selection from Survival Statistics', in *Survival statistics*, Bloomington, MN, USA: StatPac,Inc., 2010. ISBN 0-918733-11-1. https://www.statpac.com/surveys/surveys.pdf

[Walonick2003] David S Walonick, *Survival statistics*. Minneapolis, Minn.: StatPac, 2003, ISBN: 978-0-918733-11-5. http://www.statpac.com/surveys/

[Wexler2010] Michael Wexler, 'The Net Takeaway: R', Jun-2010. [Online]. Available: http://www.nettakeaway.com/tp/?s=R . [Accessed: 03-Aug-2015] (VP of Web Analytics at Barnes and Noble.com)

[Wickham2009] Hadley Wickham, *Ggplot2: elegant graphics for data analysis*. New York: Springer, 2009, ISBN: 978-0-387-98140-6. website for the book: http://had.co.nz/ggplot2/book/

[Wickham2015] Hadley Wickham, 'had.co.nz', 2015. [Online]. Available: http://had.co.nz/ . [Accessed: 03-Aug-2015]

[WVU2000] Department of Statistics, West Virginia University, 'Understanding Hypothesis Testing: Example #1', 04-Apr-2000. [Online]. Available: http://www.stat.wvu.edu/SRS/Modules/HypTest/exam1.html. [Accessed: 03-Aug-2015]