Homework 2: Tools of the Trade - Statistics

No Group - David Guzmán

November, 9th 2016

Contents

1 Computing variables and checking assumptions

1.1 What is the ECDF F (x) of the data set?

Data set: $[-10.1, -1.2, -9.5, -1, -1, -1, 0.1, 5, 7, 7, 7, 7, 2, 2, 2]$.

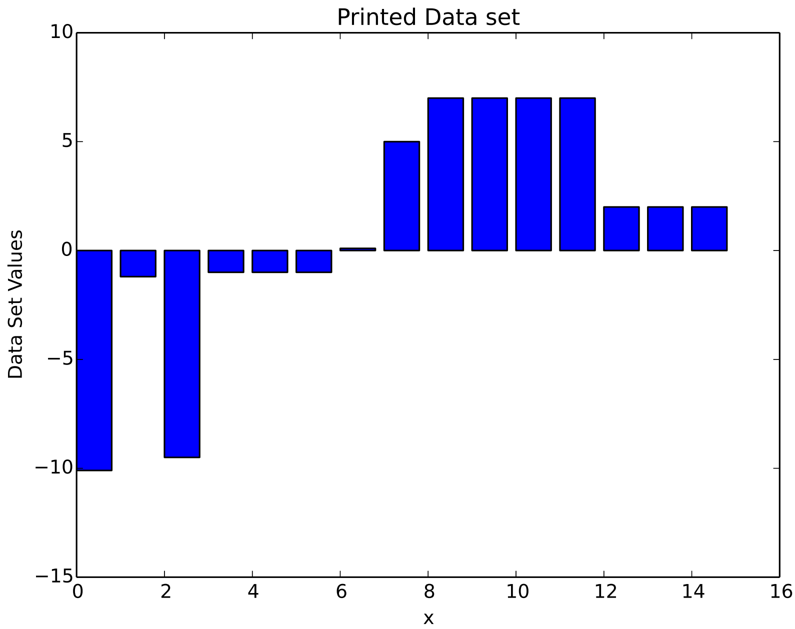

Figure 1: Dataset

The ECDF defined as $F(x) = \frac{1}{n}\sum_{n=1}^{n} 1_{\{x_i \leq x\}}$.

For the case: $x = -100$, we get $F(-100) = \frac{1}{n}\sum_{n=1}^{n} 1_{\{x_i \leq -100\}} = 0$

For the case: $x = -10$, we get $F(-10) = \frac{1}{n}\sum_{n=1}^{n} 1_{\{x_i \leq -10\}} = 1/15$

For the case: $x = 0$, we get $F(0) = \frac{1}{n}\sum_{n=1}^{n} 1_{\{x_i \leq 0\}} = 6/15$

For the case: $x = 7$, we get $F(7) = \frac{1}{n}\sum_{n=1}^{n} 1_{\{x_i \leq 7\}} = 1$

For the case: $x = 10$, we get $F(10) = \frac{1}{n}\sum_{n=1}^{n} 1_{\{x_i \leq 10\}} = 1$

1.2 What is the median?

The median defined as $x_{(\frac{n}{2})}$ iff n is odd and $\frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$ for the rest [1].
Let $xi < xj$ if $i < j$
Median of: x_{-2}, x_1, x_0 is equal to: x_0
Median of: $x_{-2}, x_1, x_0, x_2, x_4, x_{-3}$ is equal to: $\frac{x_0+x_1}{2}$
Median of: $[-10.1, -9.5, -1.2, -1, -1, -1, 0.1, 2, 2, 2, 5, 7, 7, 7, 7]$ is equal to: 2

1.3 What is the main necessary assumption for the calculation of a confidence interval?

We assume that the data is coming from an Independent Identically Distributed stochastic model (random variables) [1].

1.4 Compute the confidence interval for the median

Dataset: $[-10.1, -9.5, -1.2, -1, -1, -1, 0.1, 2, 2, 2, 5, 7, 7, 7, 7]$

Using the tables [1] to determine confidence intervals for quantiles (including the median), according to Theorem Confidence Interval for the Median and other Quantiles. For a sample of n *iid* data points $x_1, ..., x_n$, the tables give a confidence interval at the confidence level $\lambda = 0.95$ or $\lambda = 0.99$ for the q-quantile with $q = 0.5$ (median) [1].

For the sample Dataset: $n = 10$ and median $m = x_{(8)} = 2$, and using apendix A from [1] we get for $n = 15$, $j = 4$, $k = 12$ and $p = 0.965$, the confidence interval given by the Figure 3 is $[x_{(4)} = -1, x_{(12)} = 7]$, as we can observe in Figure 2.

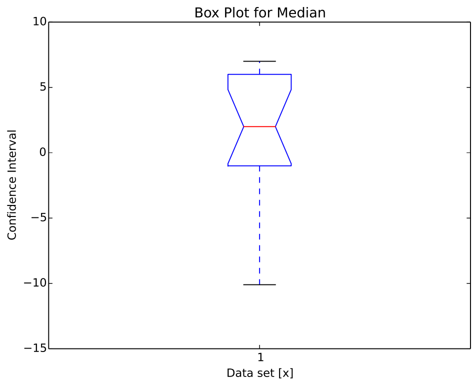

Figure 2: Box Plot Confidence Intervals

Table A.1 Quantile $q = 50\%$, Confidence Levels $\gamma = 95\%$ (left) and 0.99% (right).

n	j	k	p
$n \leq 5$: no confidence interval possible			
6	1	6	0.969
7	1	7	0.984
8	1	7	0.961
9	2	8	0.961
10	2	9	0.979
11	2	10	0.988
12	3	10	0.961
13	3	11	0.978
14	3	11	0.965
15	4	12	0.965
16	4	12	0.951
17	5	13	0.951

n	j	k	p
$n \leq 7$: no confidence interval possible			
8	1	8	0.992
9	1	9	0.996
10	1	10	0.998
11	1	11	0.999
12	2	11	0.994
13	2	12	0.997
14	2	12	0.993
15	3	13	0.993
16	3	14	0.996
17	3	15	0.998
18	4	15	0.992
19	4	16	0.996

Figure 3: Confidence Intervals [1]

1.5 Does it make sense to compute a confidence interval for the mean on the data set in (a)? Why (not)?

The mean m of a data set $x_1, ..., x_n$ is $m = \frac{1}{n}\sum_{i=1}^{n} x_i$, and gives information about the average [1], for this case $m = 1.02$. Even though, the assumptions, data source is coming from iid, large number of samples, and it is a common distribution with a finite variance, and for $n < 30$ data must come from *iid* and a *normal distribution* are strictly neccesary, *the computation of a confidence interval for the mean does make sense, for it characterizes both the variability of the data and the accuracy of the measured average* [1]. Although, we can inspect in Figure 4 , which shows the quantiles of the measurements against the corresponding quantiles of the normal distribution, data is not coming from a normal distribution. As a result, we should re-scale through for example Box-Cox transformation to compute a better approach to the CI for mean[1].

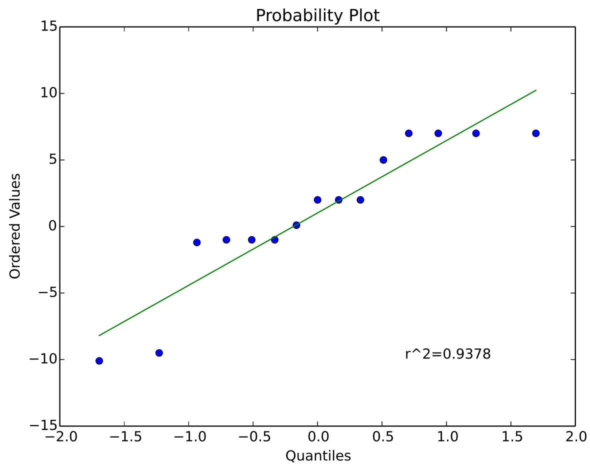

Figure 4: QQPlot Normal Distribution for Data Set [x]

1.6 If the confidence interval of the difference between the means of two data sets includes 0, what does it imply?

When we claim that an interval I is a confidence interval at the level 0.95 for a certain parameter θ, we mean the following. Repeating the experiment many times would lead to, in about 95% of the cases, the interval I indeed containing the true value θ [1]. Hence, if a confident interval for a mean difference *includes* 0, *the data set is consistent with another data set (population) mean difference of* 0 [2].

1.7 What is the simple moving average of the data set in (a) with n = 3?

For SMA defined as: $SMA = \frac{1}{n}\sum_{i=0}^{n-1} p_{(M-i)}$

for $n = 3$: $SMA = \frac{1}{3}\sum_{i=0}^{3-1} p_{(M-i)} = 1.02$

2 Reading a plot

The following plot shows the result of a performance test over two different wireless links:

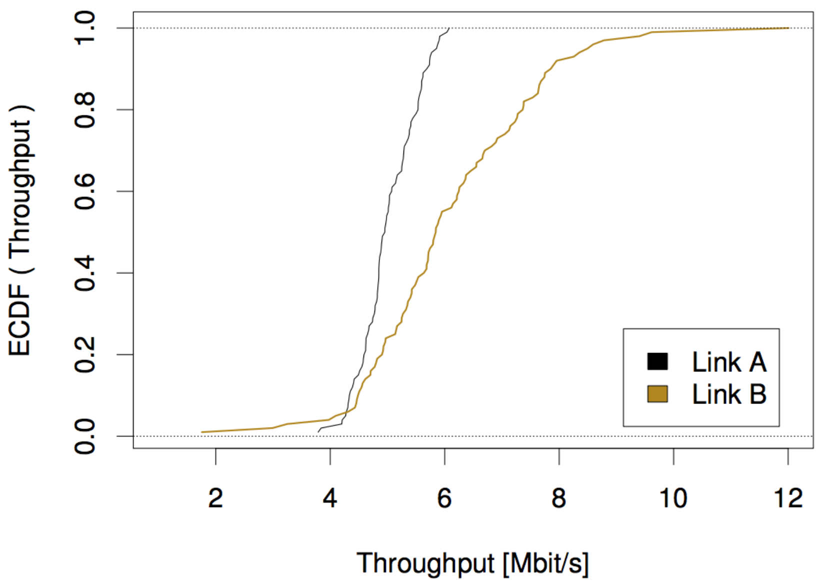

Figure 5: Throughput

2.1 What is the metric here?

Average Throughput

2.2 What are the median, the quartiles and the 95% quantiles for both links?

According to Figure 5:

Median Link A = $5Mbps$, Median Link B = $6Mbps$
25% Quartil Link A = $4.8Mbps$, 25% Quartil Link B = $5.2Mbps$
75% Quartil Link A = $5.3Mbps$, 75% Quartil Link B = $7Mbps$
95% Quantil Link A = $5.8Mbps$, 95% Quantil Link B = $8Mbps$
5% Quantil Link A = $4.3Mbps$, 5% Quantil Link B = $4.3Mbps$

2.3 Which link do you think has the higher mean? Which one has the higher standard deviation? Why?

Link B has the higher mean because the measurements in the different quartiles and quantiles are always higher, and due to the fact that the balance point for the area under the curve for Link B is higher (between 7 and 9) than the balance point for Link A (bet ween 4 and 6).

2.4 Based on this data, which link do you think had the better performance? Why?

Link B, measurements of throughput are higher than Link A.

2.5 What possible factors could have influenced the performance?

Range between the two points establishing the communication measured (Proximity to AP), we could reference the Figure 6, to have an idea of the rate variation according to the distance for the IEEE 802.11 standard.

IEEE 802.11: data rate

	Outdoor Range (m)	Indoor Range (m)
1 Mbps DSSS	550	50
2 Mbps DSSS	388	40
5.5 Mbps CCK	235	30
11 Mbps CCK	166	24
5.5 Mbps PBCC	351	38
11 Mbps PBCC	248	31
6 Mbps OFDM	300	35
12 Mbps OFDM	211	28

Figure 6: 802.11 Data Rate [3]

Receiving Sensivity in the client card that could affect the EIRP(Effective Isotropic Radiated Power [4]) in the total link budget, as we could infer from Figure 8.

- Typical values are -85 dBm for maximum data rate in 802.11b
- *Example: Orinoco cards PCMCIA Silver/Gold*
  11Mbps => -82 dBm ; 5.5Mbps => -87 dBm;
  2Mbps=> -91 dBm; 1Mbps=> -94 dBm.
- *Example: Senao 802.11b card*
  11 Mbps => -89dBm; 5.5 Mbps =>-91dBm
  2 Mbps => -93dBm; 1 Mbps => -95dBm

Figure 7: Sensitivity in Cards [4]

Noise and weather affecting free space propagation.

Change in antennas gain configuration.

Different Coding Schemes according to the version of the protocol, for example 802.11b systems offer 11, 5.5, 2 ,or 1 Mbps, with maximun user data rate approx 6Mbps. The lower data rates use DBPSK or DQPSK [4].

Data rate [Mbit/s]	Modulation	Coding rate	Coded bits per subcarrier	Coded bits per OFDM symbol	Data bits per OFDM symbol
6	BPSK	1/2	1	48	24
9	BPSK	3/4	1	48	36
12	QPSK	1/2	2	96	48
18	QPSK	3/4	2	96	72
24	16-QAM	1/2	4	192	96
36	16-QAM	3/4	4	192	144
48	64-QAM	2/3	6	288	192
54	64-QAM	3/4	6	288	216

Figure 8: Rate Dependent Parameters for IEEE 802.11a [4]

References

[1] Jean-Yves Le Boudec. *Performance Evaluation of Computer and Communication Systems*. EPFL Press, 2010. ISBN: 978-2-940222-40-7.

[2] Jerry Dallal. *Confidence Intervals*. URL: `http://www.jerrydallal.com/lhsp/ci.htm`.

[3] Rami Langar. *Telecommunication Systems*. 2015.

[4] Jochen Schiller. *Mobile Communications*. Addison-Wesley, 2003. ISBN: 0 321 12381 6.