

STAT 157 HW1 Solutions

http://www.stat.ucla.edu/~dinov/courses_students.dir/10/Spring/Stats157.dir/

Problem 1.

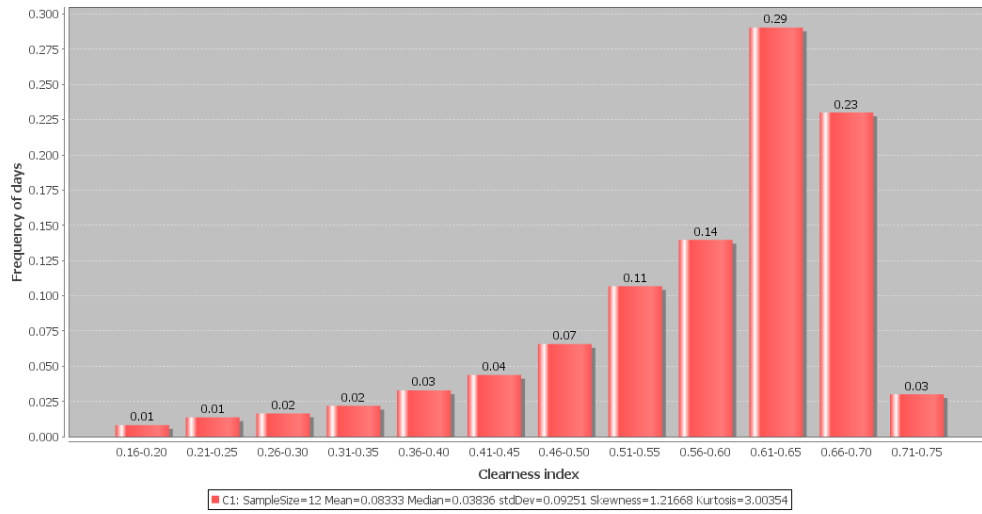
1.a: (6 points) Determine the Relative Frequency and the Cumulative Relative Frequency (fill in the two last columns of the table)

Clearness Index	Number of Days	Relative Freq	Cumulative Rel Freq	Model Probabilities
0.16-0.20	3	0.008219178	0.008219178	0
0.21-0.25	5	0.01369863	0.021917808	0
0.26-0.30	6	0.016438356	0.038356164	.0005
0.31-0.35	8	0.021917808	0.060273973	.0022
0.36-0.40	12	0.032876712	0.093150685	.0076
0.41-0.45	16	0.043835616	0.136986301	.0233
0.46-0.50	24	0.065753425	0.202739726	.0605
0.51-0.55	39	0.106849315	0.309589041	.1332
0.56-0.60	51	0.139726027	0.449315068	.2367
0.61-0.65	106	0.290410959	0.739726027	.2811
0.66-0.70	84	0.230136986	0.969863014	.2000
0.71-0.75	11	0.030136986	1	.0540

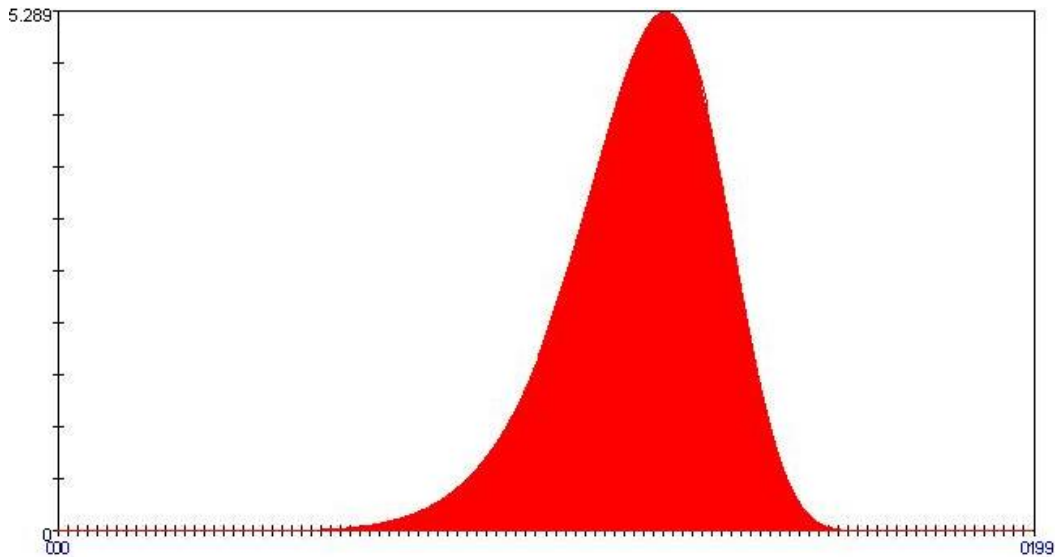
1.b: (6 points) Sketch the Relative Frequency histogram and comment on it

The relative frequency histogram is normalized such that the height of each column is equal to the probability that the clearness index falls within the bin of that column. The distribution is left-skewed, and unimodal.

Baghdad Sky Clarity



1.c: (6 points) Visually choose a model distribution for these data using SOCR Distributions, compute and enter the corresponding model probabilities for each range in the last column.



I visually chose the Weibull distribution with Shape=9 Scale =.63. These parameters matched the data reasonably close, so I did not rescale it. If you used a different distribution and a linear transformation to convert the x-axis units, that's perfectly acceptable. The probabilities for each interval are recorded in the above table using the distribution tool.

1.d: (7 points) Cloudy days are those with the clearness index < 0.35. What proportion of the days were cloudy? How different are the data and model probabilities?

Summing relative frequency entries between .16 and .35 = .0603. This is equivalent to the cumulative relative frequency you already calculated for the 0.31-0.35 range. Doing the same for my model probabilities gives .005, and the difference is .0553.

Clear days are those for which the clearness index is at least 0.66. What proportion of the days were clear?

Summing relative frequency entries between .66 and .75 = .2602. This is equivalent to one minus the cumulative relative frequency you already calculated for the 0.61-0.65 range. Doing the same for my model probabilities gives .2174, and the difference is .0428.

Problem 2.

2.a: (10 points) The operators want to know what escape time corresponds with a 1% chance of being exceeded.

First, obtain a line plot of the raw
Escape times

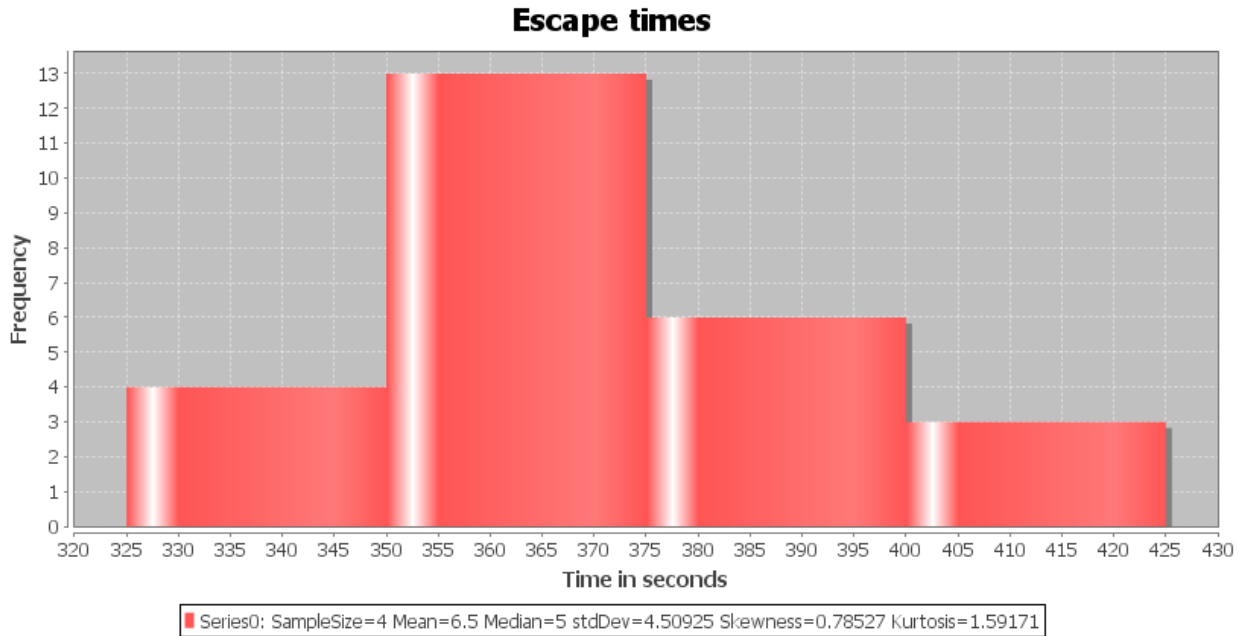


data.

This is not terribly informative, but we can experiment with making changes in the data to see how the mean, median, and standard deviation change. Doing the calculation for mean and standard deviation manually, we get the same result as in the chart.

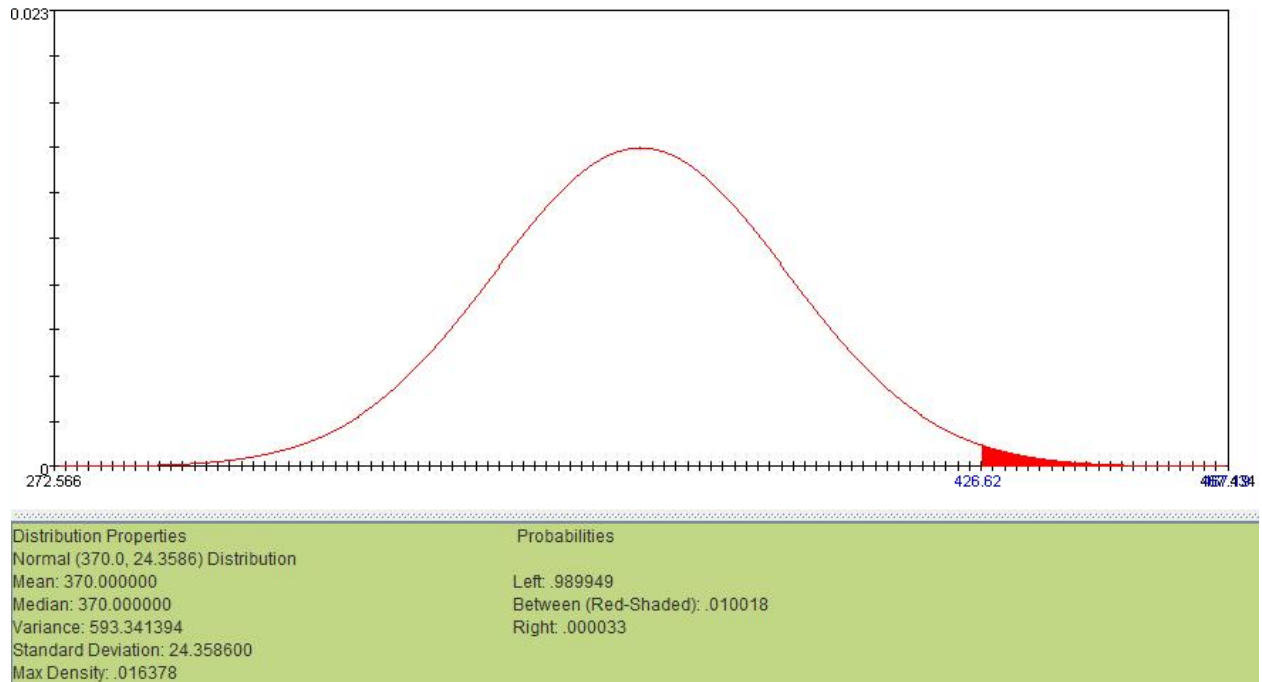
$$\mu = \frac{1}{26} \sum_{i=1}^{26} x_i = \frac{9638}{26} = 370.69231$$
$$\sigma = \sqrt{\frac{1}{26-1} \sum_{i=1}^{26} (x_i - \mu)^2} = \sqrt{\frac{14833.53846}{25}} = 24.3586$$

Next, summarize the data by plotting a histogram of the escape times to see the shape of the distribution. I used a bin size of 25.



Notice the mean, median, and standard deviation are different in this chart – this is because this chart is calculating these values for the frequency and not the data axis. Make sure you know what your results mean when you read anything from a computer program!

We don't really know what distribution this data is coming from, it looks roughly symmetric and somewhat bell shaped so a good initial choice is the normal distribution. As a bonus, we already calculated the parameters (mean and standard deviation). Plug those values into the distribution tool, and get the following chart:



The time corresponding to a 1% chance of being exceeded is equivalent to finding the threshold for which the area under the curve to the right (greater than the threshold) is 1%. This can easily be found by using the charting tool and adjusting the interval until the area is .01. Our x-axis is in units of seconds, and therefore our threshold is also in units of seconds. The value should be roughly 426.6 seconds using this distribution.

The question is a little bit ambiguous in the term “exceeded”. If interpreted as the magnitude of the escape time, then use the above answer. If interpreted as performance, then we obviously want short escape times. In this case, find the threshold for the leftmost 1% instead, which is roughly 313.37 seconds.

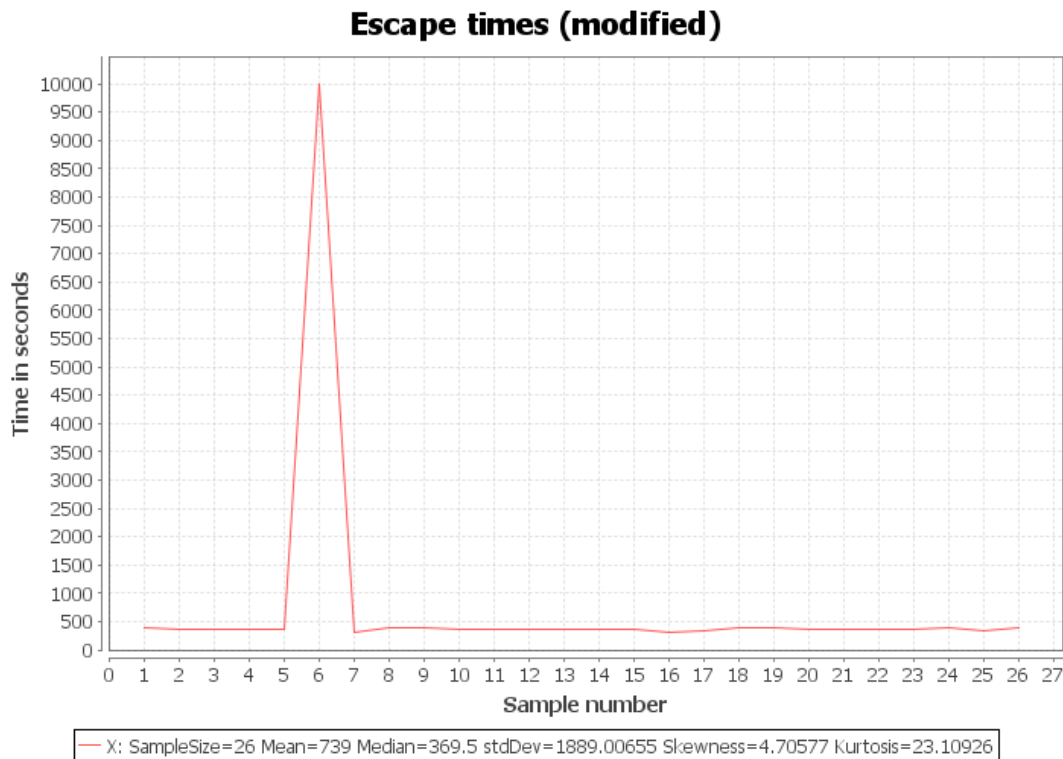
2.b: (5 points) How different are the sample mean and median?

Mean: 370.69231 Median: 369.5 Difference: 1.19231

The similarity between the mean and median further suggests that this distribution is symmetric.

2.c: (10 points) By how much should the largest time be increased so that the sample median is half the sample mean?

This is easily computed numerically as $n \cdot (2 \cdot \text{median} - \text{mean}) = 9576$. The largest escape time is 424, setting it to $424 + 9576 = 10000$ will make the mean 739. This can also be computed interactively using the line chart tool.



Problem 3.

3.a: (15 points) What is the five-number summary for this data?

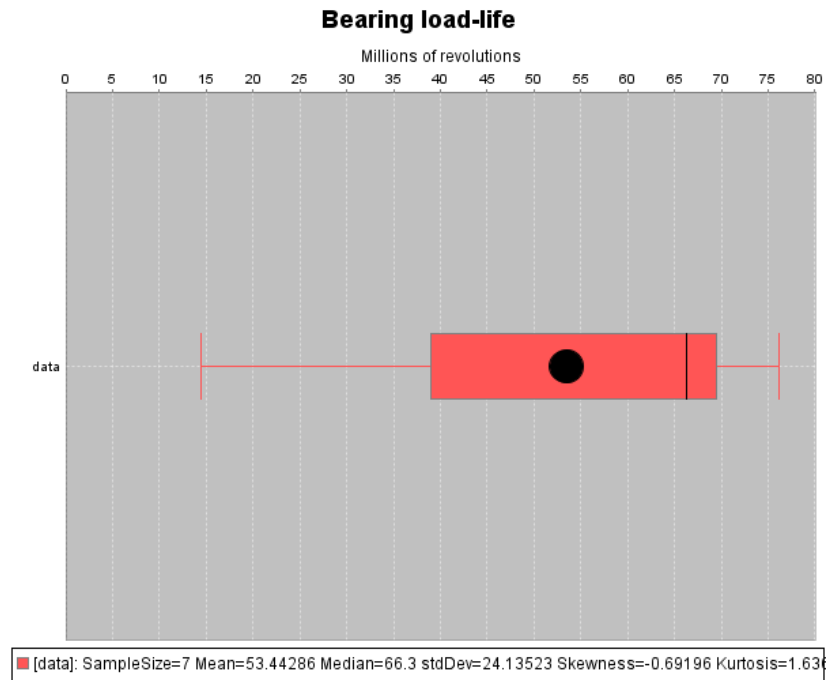
Minimum: 14.5

Lower quartile: $(25.6+52.4)/2 = 39$

Median: 66.3

Upper quartile: $(69.3+69.8)/2 = 69.55$

Maximum: 76.2



The box plot concisely illustrates the five-number summary. The whiskers at the ends indicate the minimum and maximum value, the center black line is the median, and the sides of the box are the lower and upper quartiles respectively. The black dot is the mean.

3.b: (10 points) Calculate the following sample measures of spread: variance, standard deviation and the mean-absolute-deviation.

Mean:

$$\mu = \frac{1}{7} \sum_{i=1}^7 x_i = \frac{374.1}{7} = 53.44286$$

Variance and standard deviation:

$$\sigma^2 = \frac{1}{7-1} \sum_{i=1}^7 (x_i - \mu)^2 = \frac{3495.057}{6} = 582.5095 \quad \sigma = \sqrt{582.5095} = 24.13523$$

Median: 66.3

Mean absolute deviation from median:

$$\frac{1}{6} \cdot |14.5 - 66.3| + |25.6 - 66.3| + |52.4 - 66.3| + |69.3 - 66.3| + |69.8 - 66.3| + |76.2 - 66.3| = 20.4666$$

Mean absolute deviation from mean:

$$\frac{1}{6} \cdot |14.5 - 53.44286| + |25.6 - 53.44286| + |52.4 - 53.44286| + |66.3 - 53.44286| + |69.3 - 53.44286| + |69.8 - 53.44286| = 20.4666$$

Problem 4.

Trial	Outcome
1	35
2	22
3	19
4	1
5	15
6	4
7	2
8	11
9	4
10	5

4.a (6 points) Should all 38 possible outcomes occur the same number of times? Why?

Having only run 10 trials, it is not even possible for each of the 38 outcomes to occur the same number of times. Even if we ran exactly 38 trials, we would not expect to get one outcome for each position on the wheel because the process is random.

4.b: (6 points) Does it appear as if some outcomes are just too frequent and some are too rare?

Yes, with so few trials we would expect the distribution of outcomes (imagine a histogram) and the expected probability distribution (which is uniform) to be pretty different. We might even get lucky/unlucky and get the same position twice or more. Similarly, the remaining positions account for 0% of the total outcomes. We would not expect this to become a trend if we perform more trials.

4.c: (6 points) How large is the difference between the even and odd outcomes in your 10 experiments? Is this expected to vary for different students?

Even: 4

Odd: 6

While now it is actually possible to get the same number of even and odd outcomes in our 10 trials, we would still expect the difference in the counts to vary between repeated experiments. Because each student runs the same number of trials, however, we would expect this difference to be subject to the same variance.

4.d: (7 points) Would the answers to the above questions change if we did 1,000 experiments, instead of just 10?

The total counts will continue to vary as we perform more trials. The relative proportions for each position, however, will converge toward the true probability.