

Chapter 3: Statistical Data Analysis

Part I

Topics:

Review of Descriptive Statistics

From Histograms to Probability Distributions

Uniform and Normal Distributions

Reference: Holman, CH 3.

Cleveland State University

Mechanical Engineering

Hanz Richter, PhD

MCE380 – p.1/11

Random Variables, Populations and Samples

When a quantity X is subject to uncertainty, its values are not fixed, but *randomly distributed*. We obtain information about X through descriptions of the most probable values and their spread (for example by looking at the mean and the standard deviation). These variables are known as *random variables*.

The weight of any paper clip from the same box can be thought of as a random variable. The values that we obtain as we extract the clips from the box and weigh them are *samplings* of the random variable. The set of all weights is the *population*, while the ones we obtain with the scale are the *sample*. Cost and other factors prohibit sampling of the full population.

MCE380 – p.2/11

Min, Max, Mode, Median

The minimum, maximum, mode and median are commonly-used statistics. The median is the value that divides the sample in half in terms of the numbers of data points at both sides of it. It may differ from the mean (example: 10, 12, 13, 14, 15; median is 13, while mean is 12.8). Upper 25%, 5%, etc. are generalizations of the median for fractions other than 1/2.

The median is useful when outliers (absurd data points) may cause the mean to be shifted.

The mode is the data item that repeats more often. This can be useful when dealing with non-numeric data. In Matlab, given a data vector d , the min, max, mode and median are calculated as $\min(d)$, $\max(d)$, $\text{mode}(d)$ and $\text{median}(d)$ (a little hard to remember).

MCE380 – p.3/11

Mean and Variance for Population and Samples

The mean (average) of a whole population of size N , when known, is calculated as

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

Similarly, the population variance is calculated as

$$\sigma_N^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

Often, we don't sample the whole population, but just n points. In this case, we must estimate the mean and the variance of the population using a statistical estimator. We want the estimator to be *unbiased*, which simply means that our mean and variance estimates will converge to the population mean and variance as the sample size n grows. The mean and variance estimators are

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{and} \quad \sigma_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

MCE380 – p.4/11

Exercise

Find out how your calculator finds all of the above statistics. Have the information ready for the exam. For homeworks and lab reports, Matlab is strongly suggested.

The sample and population means are calculated in Matlab by using `mean(d)`

The population variance (normalized by N) is calculated with `var(d)`

The sample variance (normalized by $n - 1$) is calculated with `var(d, 1)`

The standard deviation is the square root of the variance. It can be calculated for populations and samples with `std`, following a similar syntax.

From Histograms to Probability Distributions

Suppose we construct a histogram for a given data set. We divide the data in ranges (bins) and count how many data points fall in each bin. The histogram is a plot of the bin counts.

In Matlab, if the data vector is d we create a histogram using the syntax `hist(d, M)`, where M indicates the number of bins (of equal size).

Suppose a die is thrown sixty times, and we get 10 occurrences of each number. The histogram would be flat. If we get 10 occurrences of each except for 1 and 6, which occur 8 and 12 times, respectively, the histogram would show a dip and a hump indicating that the die is loaded.

As the size of the bins is decreased, the bin counts tend to a function called the *probability distribution*.

Simulated Data Example

The following code will be run to illustrate how histograms converge to a distribution as the number of bins increases:

```
d=1+2.*randn(1000); %create a vector of 1000 random points
for i=10:10:100, %start with 10 bins and add 10 until we have 10
hist(d,i); %display the histogram
pause(1); %wait one second
end
```

MCE380 – p.7/11

Using Histograms to Characterize Measurement Uncertain

If we can afford to take a large number of data points, we can use the histogram to make certain statements regarding future measurements made under the same conditions.

1. The area under the histogram, from the lower limit up to point x , divided by the total area gives us the probability (relative frequency) that the measurement will be less than or equal to x .
2. The area under the histogram, from point x to the upper limit, divided by the total area gives us the probability (relative frequency) that the measurement will be greater than or equal to x .
3. The area under the histogram between points x_1 and x_2 divided by the total area gives us the probability (relative frequency) that the measurement will fall between these two values.

MCE380 – p.8/11

Questions

If we have a histogram (or distribution), the point which divides it into equal areas matches the

1. Mode
2. Mean
3. Median

and the point at which the histogram has a peak would be....?

Example with Simulated Data

Compare the value of the normalized area under the histogram with the actual relative frequency of the data range $[-2 \ 4]$.

Probability Distributions

Probability distributions are often assumed for a population, so that we don't have to sample it all and construct the histogram. Instead, a distribution is given by a *probability density function* or pdf, which is continuous (an infinite number of zero-width bins).

PDF's are often normalized to have a total area of one, so that the integral between two points gives us the relative frequency (probability) of the random variable falling in that interval.

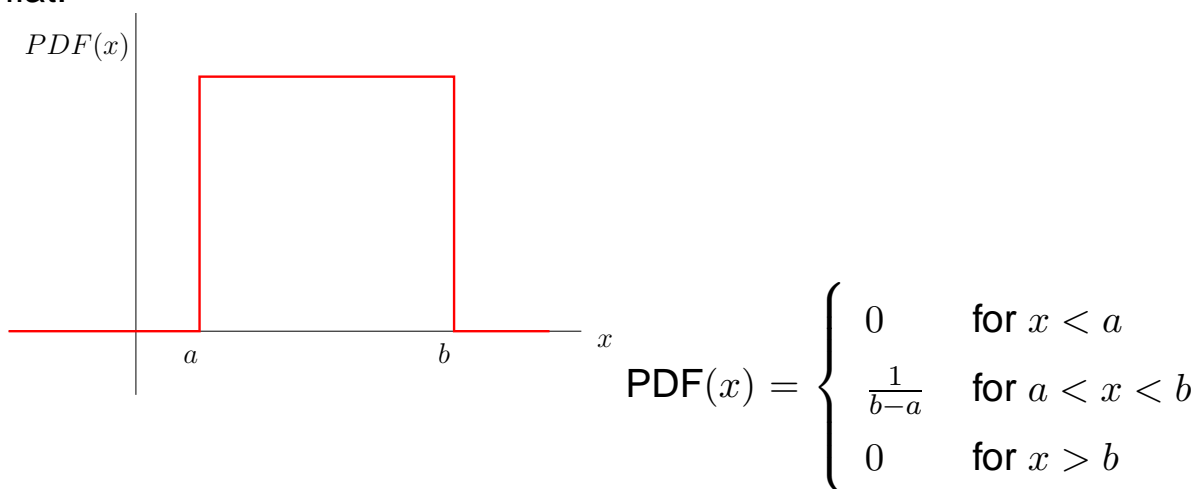
Two distributions will be introduced at this time: the uniform distribution and the normal (or Gaussian) distribution.

MCE380 – p.11/11

The Uniform Distribution

The uniform distribution corresponds to a variable whose values can occur with equal probability within a range. The throws of a perfect die are uniformly distributed between 1 and 6 with probability of 1/6.

The histogram and PDF corresponding to uniformly-distributed data are flat.



MCE380 – p.12/11

The Normal Distribution

When the overall error is the sum of many individual ones, the distribution of the sum approximately follows a normal, or Gaussian, distribution.

This fact was initially discovered by Gauss in the early 1800. He wrote down the formula for the normal distribution and verified that astronomical measurement errors were normally distributed.

Later, statisticians proved that in fact, the sum of many independent random variables is likely to follow a normal distribution, regardless of their individual distributions (this is called the Central Limit Theorem).

The normal distribution has PDF

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-x_m)^2}{2\sigma^2}}$$

Here x_m is the true population mean and σ is the true standard deviation.

When we assume that measurements are normally distributed and calculate \bar{x} and σ_n^2 , we are estimating the true values belonging to the underlying distribution.

MCE380 – p.13/14

Evaluating range probabilities

The formula for the Gaussian PDF is normalized so that the total area under the PDF is 1:

$$\int_{-\infty}^{\infty} P(x) dx = 1$$

This means the integral between two points directly gives us the chances that the variable will fall in that range. In the past, tables had to be used to evaluate the PDF. Your book still has a table.

In Matlab, assuming the Statistics Toolbox is not available, we need to use the *error function* `erf`:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

To use `erf` to evaluate $P = \int_a^b P(x) dx$, we use the change of variables $\eta = \frac{x-x_m}{\sqrt{2}\sigma}$ so that

$$P = \frac{1}{\sqrt{\pi}} \int_{\frac{a-x_m}{\sigma}}^{\frac{b-x_m}{\sigma}} e^{-\eta^2} d\eta$$

The code is `(erf((b-xm)/sigm/sqrt(2))-erf((a-xm)/sigm/sqrt(2)))`

MCE380 – p.14/14

Examples

Use Matlab's `erf` to calculate the probability that the simulated data presented earlier falls in the $[-2 \ 4]$ range. Compare with the histogram method with 100 bins.

Calculate which percentage of the data falls within 3 standard deviations of the mean.

