# Statistics 

IND4080, Research Methods
Elisabet Syverud
Sept 2019

## Overview

BLOCK 1 DATA, POPULATION AND SAMPLES

- Mean, mode, median
- Spread, range, standard deviation, variance
- Exercise

BLOCK 2 PROBABILITY DISTRIBUTION AND SAMPLING ERRORS

- Normal distribution
- Sampling errors and estimating mean
- Non-sampling errors
- Exercise


## Why do we measure data?

- Empirical based research rely on test data.
- Measurements makes it possible to test the theory.
- Statistics requires data as values.


## Continuous and Discrete Data

## Quantitative, numerical information based on measurements and surveys



## Continuous data

Values in to the data set can take any value (within the measurement range)


## Discrete data

Discrete data has only a certain number of data and can be counted

## Data Organization Guidelines

- Screen your data before you start your analysis and ensure commonsense consistency
- continuous variables - histograms, scatterplots, boxplots
- discrete variables - tabulate, bar charts
Garbage in = Garbage out



## Population and Samples

- Population = all possible observations of concern (finite or infinite)
- Sample = subset of population

We usually work with samples, not the whole population.

- To avoid bias we need to choose a random sample



## Statistics

Descriptive statistics allows us to summarize information about a sample.
$\rightarrow$ It is the simplest way to understand your data sample.

You have a sample of 14 readings of the same system.

What is the general tendency of the data?
Mean, median, mode, max, mean

What is the variation in the data set?
Range, quartile, standard deviation, variance

| Reading | Value |
| :---: | :---: |
| 1 | 6.4 |
| 2 | 7.3 |
| 3 | 4.1 |
| 4 | 5.1 |
| 5 | 5.3 |
| 6 | 4.7 |
| 7 | 6.2 |
| 8 | 3.9 |
| 9 | 4.8 |
| 10 | 5.6 |
| 11 | 7.5 |
| 12 | 4.7 |
| 13 | 4.2 |
| 14 | 7.1 |

## Descriptive statistics

## Step 1: Arrange and plot

Arrange the data in ascending order helps to see what is there.
Plotting data helps in your intuitive understanding of the typical data value and range.

| Reading | Value |
| :---: | :---: |
| 8 | 3.9 |
| 3 | 4.1 |
| 13 | 4.2 |
| 6 | 4.7 |
| 12 | 4.7 |
| 9 | 4.8 |
| 4 | 4.9 |
| 5 | 5.2 |
| 10 | 5.6 |
| 7 | 6.2 |
| 1 | 6.4 |
| 14 | 7.1 |
| 2 | 7.3 |
| 11 | 7.5 |



## Descriptive statistics

## Step 2a: Mean

The sample mean is the arithmetic average of the data

$$
\bar{x}=\frac{\sum x_{i}}{n}=\frac{x_{1}+x_{2}+\cdots+x_{n}}{n}
$$

| Reading | Value |
| :---: | :---: |
| 1 | 6.4 |
| 2 | 7.3 |
| 3 | 4.1 |
| 4 | 5.1 |
| 5 | 5.3 |
| 6 | 4.7 |
| 7 | 6.2 |
| 8 | 3.9 |
| 9 | 4.8 |
| 10 | 5.6 |
| 11 | 7.5 |
| 12 | 4.7 |
| 13 | 4.2 |
| 14 | 7.1 |



## Descriptive statistics

## Step 2a: Median

Median is the middle data value in a sorted list of data.

For an even set of numbers in a data set, the median is the arithmetic average of the two middle values.

| Reading | Value |
| :---: | :---: |
| 1 | 6.4 |
| 2 | 7.3 |
| 3 | 4.1 |
| 4 | 5.1 |
| 5 | 5.3 |
| 6 | 4.7 |
| 7 | 6.2 |
| 8 | 3.9 |
| 9 | 4.8 |
| 10 | 5.6 |
| 11 | 7.5 |
| 12 | 4.7 |
| 13 | 4.2 |
| 14 | 7.1 |



## Descriptive statistics Step 2c: Mode

Mode is the most frequently occurring data value.



## Descriptive statistics

## Step 3a: Range

Range is the range from maximum to minimum in the data set.

Range $=$ Max value $-\min$ value


## Descriptive statistics

## Step 3d: Standard Deviation

Standard deviation expresses how much the members of the group deviates from the sample mean. It is a measure of the data scatter:

- Low standard deviation: data points are close to the mean, low scatter.
- High standard deviation: data points are further away from mean, high data scatter.

For estimating the standard deviation based on a sample, we use n-1 degrees of freedom (independent pieces of information):

| Reading | Value |
| :---: | :---: |
| 1 | 6.4 |
| 2 | 7.3 |
| 3 | 4.1 |
| 4 | 5.1 |
| 5 | 5.3 |
| 6 | 4.7 |
| 7 | 6.2 |
| 8 | 3.9 |
| 9 | 4.8 |
| 10 | 5.6 |
| 11 | 7.5 |
| 12 | 4.7 |
| 13 | 4.2 |
| 14 | 7.1 |

$$
s=\sqrt{\frac{\sum(x-\bar{x})^{2}}{n-1}}
$$



## Descriptive statistics

## Step 3e: Sample Variance

Variance tells us how far the data set is spread out relative to the mean, and relative to each other.

$$
s^{2}=\frac{\sum(x-\bar{x})^{2}}{n-1}
$$

A high variance tells us that the data points are further spread out from one another.

You cannot plot variance on the same line as the data - the unit is squared.

| Reading | Value |
| :---: | :---: |
| 1 | 6.4 |
| 2 | 7.3 |
| 3 | 4.1 |
| 4 | 5.1 |
| 5 | 5.3 |
| 6 | 4.7 |
| 7 | 6.2 |
| 8 | 3.9 |
| 9 | 4.8 |
| 10 | 5.6 |
| 11 | 7.5 |
| 12 | 4.7 |
| 13 | 4.2 |
| 14 | 7.1 |

Standard deviation is 1.23.
Variance is 1.50 .

## Exercise 1: Statisctical Process Control



- You need to know how long time it takes to manufacture a part made in an established production line.
- The production process has seven steps.
- Determine the central tendencies of the data registered for 10 production serial numbers.


## Exercise 1: Descriptive Statistics <br> Registered hours per production step for 10 equal parts

Plot the data

Calculate:

- Mean
- Mediane
- Mode
- Range
- Standard deviation
- Variance

|  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Estimated | 0.80 | 2.00 | 3.25 | 4.90 | 1.20 | 1.30 | 1.30 |
| S/N 1 | 0.6 | 2.48 | 2.72 | 4.31 | 1.22 | 1.5 | 1.3 |
| S/N 2 | 0.6 | 1.75 | 6.36 | 4.85 | 1.32 | 1.5 | 1.3 |
| S/N 3 | 0.6 | 2.12 | 2.55 | 0.85 | 1.7 | 1.5 | 1.4 |
| S/N 4 | 0.6 | 3.22 | 4.87 | 8.2 | 1.18 | 2.5 | 1.4 |
| S/N 5 | 0 | 3.45 | 2.55 | 3.9 | 1.67 | 1.5 | 1.2 |
| S/N 6 | 0.6 | 1.63 | 3.37 | 3.09 | 1.47 | 1.5 | 1.3 |
| S/N 7 | 0.6 | 2.74 | 3.01 | 7.65 | 0.27 | 1.5 | 1.3 |
| S/N 8 | 0.6 | 1.95 | 2.4 | 4.1 | 1.94 | 1.5 | 1.3 |
| S/N 9 | 0.6 | 2.38 | 2.76 | 5.58 | 1.6 | 1.5 | 1.2 |
| S/N 10 | 0.6 | 2.42 | 2.13 | 4.74 | 1.2 | 1.5 | 1.3 |

What do you see in the data?

## Exercise 1:

## Registered hours per production step for 10 equal parts

|  |  |  |  |  |  | $\begin{aligned} & \text { N } \\ & \stackrel{0}{0} \\ & \text { O } \end{aligned}$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Estimated | 0.80 | 2.00 | 3.25 | 4.90 | 1.20 | 1.30 | 1.30 |
| S/N 1 | 0.6 | 2.48 | 2.72 | 4.31 | 1.22 | 1.5 | 1.3 |
| S/N 2 | 0.6 | 1.75 | 6.36 | 4.85 | 1.32 | 1.5 | 1.3 |
| S/N 3 | 0.6 | 2.12 | 2.55 | 0.85 | 1.7 | 1.5 | 1.4 |
| S/N 4 | 0.6 | 3.22 | 4.87 | 8.2 | 1.18 | 2.5 | 1.4 |
| S/N 5 | 0 | 3.45 | 2.55 | 3.9 | 1.67 | 1.5 | 1.2 |
| S/N 6 | 0.6 | 1.63 | 3.37 | 3.09 | 1.47 | 1.5 | 1.3 |
| S/N 7 | 0.6 | 2.74 | 3.01 | 7.65 | 0.27 | 1.5 | 1.3 |
| S/N 8 | 0.6 | 1.95 | 2.4 | 4.1 | 1.94 | 1.5 | 1.3 |
| S/N 9 | 0.6 | 2.38 | 2.76 | 5.58 | 1.6 | 1.5 | 1.2 |
| S/N 10 | 0.6 | 2.42 | 2.13 | 4.74 | 1.2 | 1.5 | 1.3 |

## Probability distributions

## Normal Distribution

$\qquad$

The normal distribution is completely determined by the mean $(\mu)$ and the standard deviation $(\sigma)$.

For populations characterized by the normal distribution, a random sample size of 30 is generally a good approximation (Central Limit Theorem).


## Probability distributions

## T-Distribution and F-Distribution





## Data Sampling Errors

## Sampling errors

- The sample does not represent the population

Non-sampling errors

- Systematic or random errors in the sample

Bias is when the information is skewed due to preconception

Sampling Errors

## Estimating the mean



Sampling Errors

## Estimating the mean




## Sampling Errors

## Estimating the mean




## Non-Sampling Errors in Measurements

$\qquad$
ASME PTC 19.1-1998
TEST UNCERTAINTY

- Measurement Uncertainties tells us how accurate a measurement is.
- It is the researcher's best estimate of how far an experimental quantity might be from the true value.
- Measurement Uncertainty depends on the measurement method.
- Systematic error (bias, fixed error)
- Random (precision error)


Measured Values

## Example:

## Measurement Errors

- Temperature sensor
- Sampling frequency 200ms
- Stable conditions



## Exercise 1: Descriptive Statistics <br> Registered hours per production step for 10 equal parts

1. Any strange data?
2. Any potentially polluted data series?
3. Can you still use the data, and why?

|  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Estimated | 0.80 | 2.00 | 3.25 | 4.90 | 1.20 | 1.30 | 1.30 |
| S/N 1 | 0.6 | 2.48 | 2.72 | 4.31 | 1.22 | 1.5 | 1.3 |
| S/N 2 | 0.6 | 1.75 | 6.36 | 4.85 | 1.32 | 1.5 | 1.3 |
| S/N 3 | 0.6 | 2.12 | 2.55 | 0.85 | 1.7 | 1.5 | 1.4 |
| S/N 4 | 0.6 | 3.22 | 4.87 | 8.2 | 1.18 | 2.5 | 1.4 |
| S/N 5 | 0 | 3.45 | 2.55 | 3.9 | 1.67 | 1.5 | 1.2 |
| S/N 6 | 0.6 | 1.63 | 3.37 | 3.09 | 1.47 | 1.5 | 1.3 |
| S/N 7 | 0.6 | 2.74 | 3.01 | 7.65 | 0.27 | 1.5 | 1.3 |
| S/N 8 | 0.6 | 1.95 | 2.4 | 4.1 | 1.94 | 1.5 | 1.3 |
| S/N 9 | 0.6 | 2.38 | 2.76 | 5.58 | 1.6 | 1.5 | 1.2 |
| S/N 10 | 0.6 | 2.42 | 2.13 | 4.74 | 1.2 | 1.5 | 1.3 |

