Characterizing Measured Data

Overview

- **Basics** of data set characterization
  - Indices of central tendency
  - Indices of dispersion
- Analyzing the data distribution
- **Confidence intervals**
  - Zero mean
  - Comparing different alternatives
- Determining the sample size

Characterizing a data set

- Summarizing measured data is a common problem;
- Experimental campaigns usually consist of *hundred* or *million* observations of a given variable.

- For example, let us consider a **system A**, whose response time is measured over 5 days

	system A (*ms*)
1	10
2	9
3	11
4	10
5	10

Sample Mean

- **Sample Mean** is obtained by taking the sum of all observations and dividing this sum by the number of observations in the sample.

	System A (*ms*)
1	10
2	9
3	11
4	10
5	10
sum	*50*
mean	**10**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Is the mean always good?

- Mean might not be the best option in all the cases
  - e.g., **skewness of the distribution**.
- E.g., 10*ms* represents none of the responses of System B!

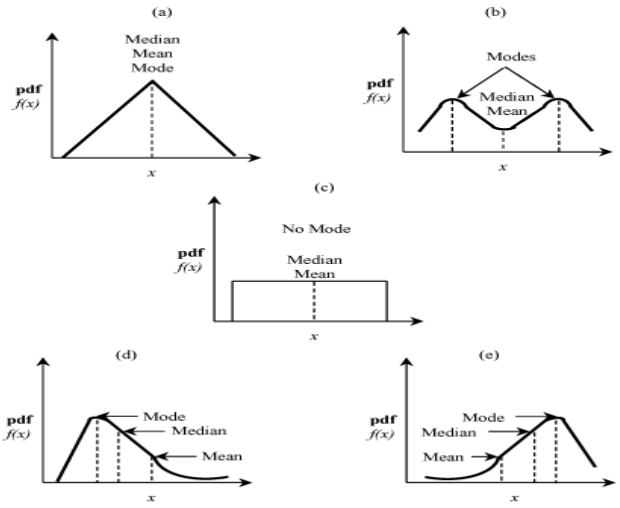
	System A (*ms*)	System B (*ms*)
1	10	5
2	9	5
3	11	5
4	10	4
5	10	31
sum	*50*	*50*
mean	**10**	**10**

Other indices of central tendency

- **Sample Median** is obtained by sorting the observations in an increasing order and taking the observation that is in the **middle** of the series. If the number of observations is even, the mean of the middle two values is used as a median.
- **Sample Mode** is obtained by plotting a histogram and specifying the midpoint of the bucket where the histogram peaks. For categorical variables, mode is given by the category that occurs most frequently.

- Some observations:
  - **mean** is affected more by outliers than the median or mode;
  - **mean** and **median** always exist and are unique;
  - **mode**, on the other hand, may not exist.

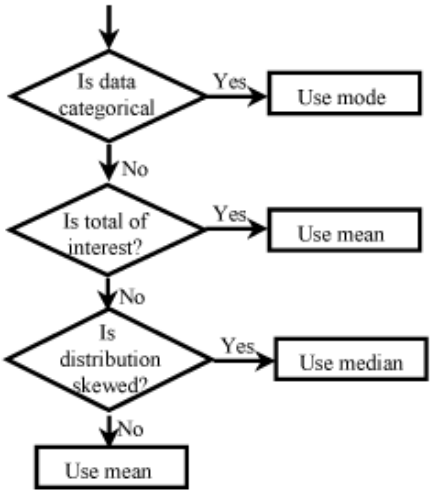
Mean, Median, and Mode



Examples

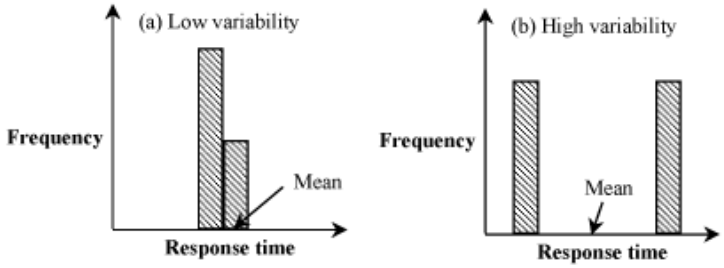
- *"Most used resource in a system"* **mode**
- *"Inter-arrival time"* total time is of interest and so **mean** is the proper choice
- "*Load on a computer*" **median** is preferable if the distribution is skewed
  - a simple way to determine **skewness** for small samples is examining the ratio y_{MAX} / y_{min}. If the ratio is **large** the data is skewed.

Selecting Mean, Median, and Mode



Summarizing variability

- Using a **single number** is usually not enough to characterize the whole data set;
- It is important to provide information about the **variability** of the measured data data.



Indices of Dispersion

- Range: *minimum* and *maximum* of the values observed
- Variance or standard deviation
- 10- and 90- percentiles
- Semi inter-quartile range
- Mean absolute deviation

Range

- Range = $y_{MAX} - y_{min}$
- Larger range => higher variability
- The minimum often comes out to be zero and the maximum comes out to be an outlier far from typical values. Unless the data is bounded:
  - y_{MAX} goes on *increasing* with the number of observations
  - y_{min} goes on *decreasing* with the number of observations
- Range is **useful** if, and only if, there is a reason to believe that the variable is bounded

Variance

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

- the divisor for s^2 is *n-1* and not *n*
- only *n-1* of the *n* differences $(x_i - \bar{x})$ are independent
- the sum of all *n* differences must be **zero**

- the unit of the variance is the square of the unit of the observations: **standard deviation** is preferable

Quantile and Percentile

- The x-value at which the CDF of the observations takes a value α is called α–quantile or 100*α percentile
- α-quantiles can be estimated by sorting the n observations and taking the [(n-1)*α + 1]*th* element ([.] denotes the nearest integer)
- Fractile=quantile. Percentiles at multiples of 10% are called deciles. (e.g., the first decile is 10-percentile)

{1.9, 2.7, 2.8, 2.8, 2.8, 2.9, 3.0, 3.1, 3.1, 3.2, 3.2} n=11

- 0.1-quantile [10*0.1 + 1]=2 → 2.7 (10-percentile)
- 0.5-quantile [10*0.5 + 1]=6 → 2.9 (50-percenitle, or median)

Semi Inter-Quartile Range

- **Quartiles** divide the data into four parts at 25%, 50%, and 75%
  - 25% of the observations are less than or equal to the first quartile Q_1
  - 50% of the observations are less than or equal to the second quartile Q_2 (again, Q_2 is also the median)
  - 75% are less than the third quartile Q_3

$$\text{SIQR} = \frac{Q_3 - Q_1}{2} = \frac{x_{0.75} - x_{0.25}}{2}$$

Mean Absolute Deviation

$$\text{Mean absolute deviation} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

- No multiplication or square root is required

Example

- In an experiment, which was repeated **32 times**, the measured CPU time was found to be
  {1.9, 2.7, 2.8, 2.8, 2.8, 2.9, 3.1, 3.1, 3.2, 3.2, 3.3, 3.4, 3.6, 3.7, 3.8, 3.9, 3.9, 3.9, 4.1, 4.1, 4.2, 4.2, 4.4, 4.5, 4.5, 4.8, 4.9, 5.1, 5.1, 5.3, 5.6, 5.9}.

- 10-percentile = [1+(31)(0.10)] = 4.1 → 4th element = 2.8
- 90-percentile = [1+(31)(0.90)] = 28.9 →29th element = 5.1
- 1st quartile Q_1= [1+(31)(0.25)] = 8.75 → 9th element = 3.2
- 2nd quaritle Q_2 =[1+(31)(0.50)] = 16.5 → 16th element = 3.9
- 3rd quartile Q_3=[1+(31)(0.75)] = 24.25 → 24th element = 4.5
- SIQR = (4.5 – 3.2) / 2 = 0.65

Some remarks

- Range is affected considerably by **outliers**
- Sample variance is also affected by outliers but the affect is less
- Semi inter-quantile range is very resistant to outliers
- If the distribution is highly skewed, outliers are highly likely and SIQR is preferred over standard deviation
- In general, SIQR is used as an index of dispersion whenever median is used as an index of central tendency

Data Distribution

- Histogram of the observations
  - MAX / min of the observed values; range is divided in a number of **cells**;
  - the count of each cell is divided by the total number of observations.
- The key problem is determining the **cell size**
  - *small* cells =>large variation in of observations per cell
  - *large* cells => details of the distribution are completely lost
- One guideline: if any cell has less than five observations, the cell size should be **increased**.

Quantile-Quantile plots

- $y_{(i)}$ is the *observed* q_ith quantile
  x_i is the *theoretical* q_ith quantile
- $(x_i, y_{(i)})$ plot should be a straight line

- To determine the q_i^{th} quantile x_i, need to **invert** the cumulative distribution function

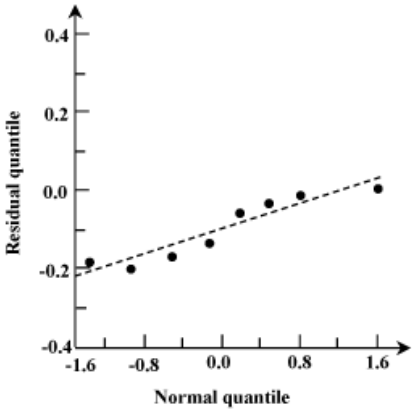
$$q_i = F(x_i) \qquad x_i = F^{-1}(q_i)$$

Example: normal distribution

- We want to test if a sample y comes from a **normal** distribution
  - $q_i = (i-0.5) / n$ — n = number of observations
  - $x_i = 4.91 [q_i^{0.14} - (1-q_i)^{0.14}]$ — approximation for N(0,1)

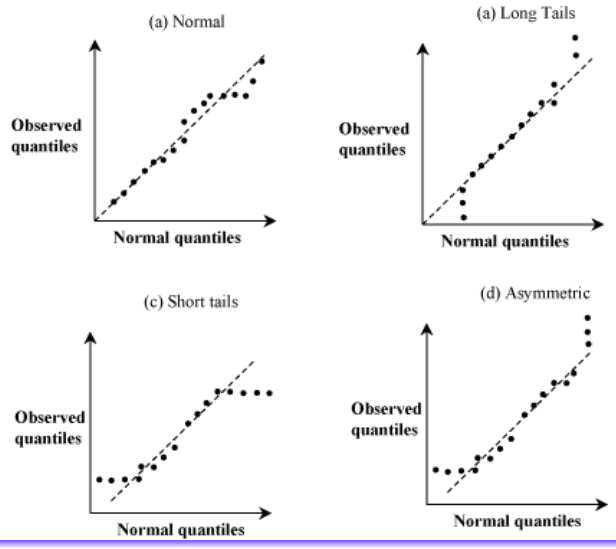
	y_i	q_i	x_i
1	-0.19	0.0625	-1.535
2	-0.14	0.1875	-0.885
3	-0.09	0.3125	-0.487
4	-0.04	0.4375	-0.157
5	0.04	0.5625	0.157
6	0.04	0.6875	0.487
7	0.14	0.8125	0.885
8	0.19	0.9375	1.535

Example: normal distribution



- Normal quantiles have been obtained using the **unit normal N(0,1)** distribution
- **Intercept** and **slope** of the resulting line give the values of the shape parameters of the distribution
- Knowing the **name** of the distribution is enough in a quantile-quantile plot

Interpretation of Quantile-Quantile Plots



Confidence Intervals and Sample Size

Sample and Population

- Let us consider several million random numbers with a given **mean** μ and **standard deviation** σ in an urn
- A sample is a set of *n* observations (*n* drawings from the urn, i.e., the population)

  sample: $\{x_1, x_2, \ldots, x_n\}$
- A fact: the sample mean $\bar{x}$ is very likely to be different from μ
  - (μ, σ), Greek letters – **parameters** (fixed, they are the population characteristics);
  - $(\bar{x}, s)$, English letters – **statistics** (depend on the samples)

Confidence Interval

- If we draw *k* samples, we obtain *k* sample means $\bar{x}$
- Determine 2 bounds, i.e., c_1 and c_2, such that there is a high probability that the mean is in the interval $[c_1, c_2]$

$$\textbf{Probability}\{c_1 \le \mu \le c_2\} = 1 - \alpha$$

- Confidence interval: $[c_1, c_2]$
- Significance level: α is usually small, e.g., 0.1, 0.05.
- Confidence level: $\mathbf{100(1-\alpha)}$ is usually high, e.g., 90%, 95%.

Writing a Confidence Interval

- First option: 5-percentile and 95-percentile of the sample means to get 90% confidence interval
  - many samples are required (**n > 30**)
- However, just one sample is enough because the mean is approximately **normally** distributed (central limit theorem)

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$$

where μ = population mean, σ = population standard deviation

- a 100(1-α)% confidence interval for μ is:

$$(\bar{x} - z_{1-\alpha/2}s/\sqrt{n}, \bar{x} + z_{1-\alpha/2}s/\sqrt{n})$$

$z_{1-\alpha/2}$ = (1-α/2)-quantile of N(0,1)

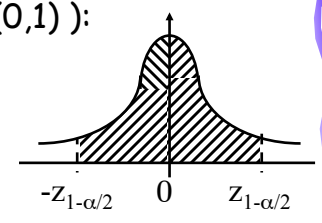
Writing a Confidence Interval

- If a random variable x has a N(μ, σ), then (x-μ)/σ has a N(0,1) distribution

$$x \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \quad \Rightarrow \quad \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

- For each probability values 1-α, it is possible to write (z_α is the α–quantile of N(0,1)):

$$P(-z_{1-\frac{\alpha}{2}} \le \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}} \le +z_{1-\frac{\alpha}{2}}) = 1-\alpha$$



Example

- $\bar{x}$ = 3.90, s = 0.95 and n = 32
- A 90% confidence interval for the mean

$$3.90 \mp (1.645)(0.95)/\sqrt{32} = (3.62,\ 4.17)$$

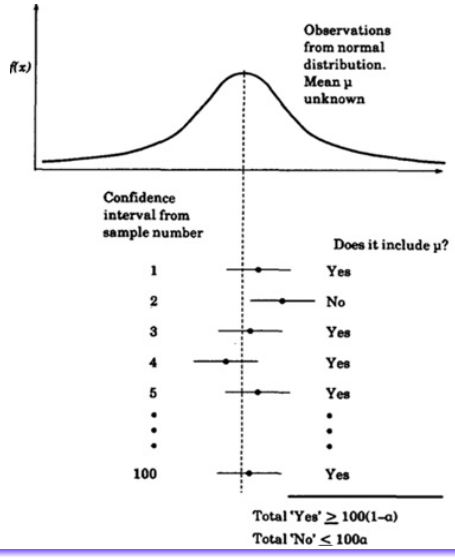
The population mean is between 3.62 and 4.17 with 90% confidence.

Commonly used normal quantiles

Confidence Level (%)	α	α/2	$z_{1-\alpha/2}$
20	0.8	0.4	0.253
40	0.6	0.3	0.524
60	0.4	0.2	0.842
68.26	0.3174	0.1587	1.000
80	0.2	0.1	1.282
90	0.1	0.05	1.645
95	0.05	0.025	1.960
95.46	0.0454	0.0228	2.000
98	0.02	0.01	2.326
99	0.01	0.005	2.576
99.74	0.0026	0.0013	3.000
99.8	0.002	0.001	3.090
99.9	0.001	0.0005	3.29
99.98	0.0002	0.0001	3.72

What does it mean?

- If we take 100 samples and construct confidence interval for each sample, the interval would include the population mean in 100(1-α) cases.



CI for small samples

- If the sample is **small** (i.e., n is less than 30), the 100(1-α)% confidence interval is

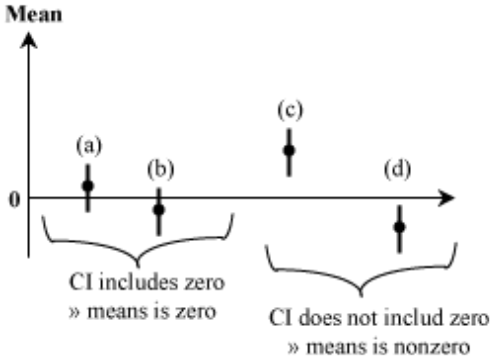
$$(\bar{x} - t_{[1-\alpha/2;n-1]}s/\sqrt{n}, \bar{x} + t_{[1-\alpha/2;n-1]}s/\sqrt{n})$$

  – only if the observations come from a normally distributed population

- $t_{[1-\alpha/2;\ n-1]}$ = (1-α/2)-quantile of a t-variate with n-1 degrees of freedom (tabled values)

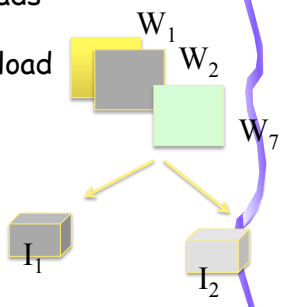
Zero mean

- Confidence intervals are used to test whether a value is **significantly** different from zero.
- The test consists of checking if the interval includes zero



Example

- Performance of 2 implementations of the same algorithm is measured with 7 similar workloads
- Difference in processor times for each workload
  {1.5, 2.6, -1.8, 1.3, -0.5, 1.7, 2.4}
- Is one implementation better than the other with 99% confidence?



$\bar{x}$ = 1.03 s = 1.60 n = 7 α = 0.01

$t_{[1- 0.01/2;\ 6]} = t_{[0.995;\ 6]} = 3.707$

CI = 1.03 ± 3.707 * (1.60 / √7) = 1.03 ± 2.24 = (-1.21, 3.27)

Comparing two alternatives

- **Paired observations**: *n* experiments on each system such that there is a one-to-one correspondence between the i^{th} test on system *A* and the i^{th} test on system B
  - **confidence interval** of the difference
- **Unpaired observations**: No correspondence. Two sample of size n_A and n_B are available from system *A* and B respectively
  - **t-test** procedure
  - approximate visual test

Example: paired observations

- 6 similar workloads are used on 2 systems. Is one system better than the other with **90% confidence**?

(5.4, 19.1), (16.6, 3.5), (0.6, 3.4), (1.4, 2.5), (0.6, 3.6), (7.3, 1.7)

differences {-13.7, 13.1, -2.8, -1.1, -3.0, 5.6}

$\bar{x} = -0.32 \quad s = 9.03 \quad n = 6 \quad \alpha = 0.1$

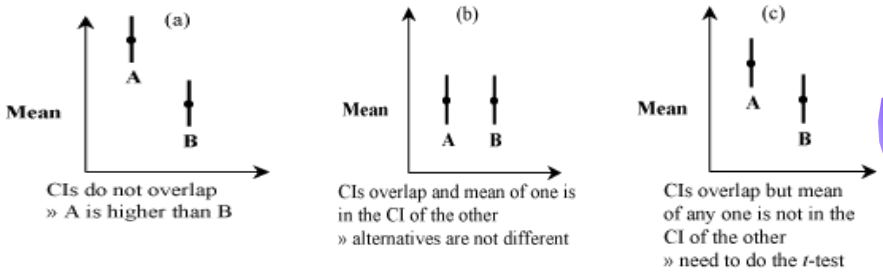
$t_{[1-\ 0.1/2;\ 5]} = t_{[0.95;\ 5]} = 2.015$

$CI = -0.32 \pm 2.015 * (9.03 / \sqrt{6}) = 0.32 \pm 7.43 = (-7.75, 7.11)$

***CI includes 0**: the systems are not significantly different*

Unpaired observations

- A simple test consists of computing each confidence interval separately and comparing them with a visual approach



Example: unpaired observations

- System A: {5.36, 16.57, 0.62, 1.41, 0.64, 7.26}

  $\bar{x}$ = 5.31 s = 6.16 n = 6
- System B: {19.12, 3.52, 3.38, 2.50, 3.60, 1.74}

  $\bar{x}$ = 5.64 s = 6.64 n = 6
- CI with 90% confidence interval ($t_{[1- 0.1/2;\ 5]} = t_{[0.95;\ 5]} = 2.015$)

  CI_A = **(0.24, 10.38)** CI_B = **(0.18, 11.10)**
- Answer: the **systems are not different** at this level of confidence

One Sided Confidence Intervals

- **Two** sided intervals: if the confidence level is 100(1-α)% there is a 100α/2% that the difference is more (less) than the upper (lower) confidence limit.
- **One** sided interval
  - e.g., is the mean greater than a certain value?
- One sided **lower** and **upper** confidence interval for μ:

$$(\bar{x} - t_{[1-\alpha;n-1]}\frac{s}{\sqrt{n}}, \bar{x}) \qquad \left(\bar{x}, \bar{x} + t_{[1-\alpha;n-1]}\frac{s}{\sqrt{n}}\right)$$

t is taken at 1-α (not 1-α/2)
(for large samples: z values are used instead of t)

Sample Size

- Problem: How many observations **n** we need to get an **accuracy** of r % with a **confidence level** of 100(1-α)%?
- r% accuracy implies that confidence interval should be

$$\text{CI} = (\bar{x}(1 - r/100), \bar{x}(1 + r/100))$$

$$\bar{x} \mp z\frac{s}{\sqrt{n}} = \bar{x}\left(1 \mp \frac{r}{100}\right)$$

$$z\frac{s}{\sqrt{n}} = \bar{x}\frac{r}{100}$$

$$n = \left(\frac{100zs}{r\bar{x}}\right)^2$$

Example

- Based on a preliminary sample, the mean of a response time is 20*s* and the standard deviation 5*s*
- Determine the number of repetitions that are needed to get the response time **accurate within 1s at 95% confidence**

- accuracy = 1s => r= 100% / 20 = 5% z=1.960

$$n = \left(\frac{(100)(1.960)(5)}{(5)(20)}\right)^2 = (9.8)^2 = 96.04$$

- accuracy = 0.5s => r=2.5% n= 384.1

Sample size to compare alternatives

- Algorithm A loses 0.5% of packets and algorithm B loses 0.6%
- Question: How many packets do we need to observe to state with 95% confidence that algorithm A is better than the algorithm B?
- Answer:

$$\text{CI for algorithm A} = 0.005 \mp 1.960\left(\frac{0.005(1-0.005)}{n}\right)^{1/2}$$

$$\text{CI for algorithm B} = 0.006 \mp 1.960\left(\frac{0.006(1-0.006)}{n}\right)^{1/2}$$

Example

- Non-overlapping interval constraint

$$0.005 \mp 1.960 \left(\frac{0.005(1-0.005)}{n}\right)^{1/2}$$
$$\leq 0.006 \mp 1.960 \left(\frac{0.006(1-0.006)}{n}\right)^{1/2}$$

n = 84,340

we need to observe around 85,000 packets

Remarks

- What confidence level to use?
  - based on the **loss** that you would sustain if the parameter is outside the range and the gain you would have if the parameter is inside the range
  - low loss ⇒ Low confidence level is fine
- Confidence interval provides possible **range**
  - narrow confidence interval ⇒ **high** degree of precision
  - wide confidence interval ⇒ **low** precision
- Confidence intervals indicate not only **what** to say but also **how loudly** to say it
- CI is easy to explain to decision makers