Estimation

Presented By:
Dr. Vinay Kulkarni

Estimation related Topics

- Sampling
- Point Estimation
- Sample Mean and Sample Variance
- Interval Estimation
- Confidence interval: one parameter
- Sample Size

Samples, Parameters & Statistics

- Sampling
  - Allows us to make inferences about a population based on a sample of that population
- Parameters
  - Numerical characteristics about the population that are of interest
- Statistics
  - Parameters cannot be exactly determined
  - They can only be estimated from samples
  - These estimates or summaries, based on the sample, are known as **Statistics**
- Major aspects of samples and statistics:
  - How accurate are the estimators (statistics)?
  - Is the sample truly representative of the population?

Statistics as Estimates for Parameters

- We use statistics to estimate parameters
  - Proportions
  - Arithmetic averages
  - Ranges
  - Quartiles
  - Deciles
  - Percentiles
  - Variances
  - Standard deviations

Sampling Methods

- Non-probability sampling
  - Convenience sampling
    - Randomly pick-up the easily accessible apples
  - Judgment or subjective sampling
  - Volunteer sampling
    - Especially used in clinical trials and research
- Probability sampling methods
  - This involves the planned use of **chance**
  - There is no selection bias
- Assignment:
  - **Review of sampling techniques and when they should be used**

Point Estimation

- Estimation
  - First step of Inferential Statistics
    - (Second step is Hypothesis Testing)
  - Two types:
    - Point estimation
    - Interval estimation
- Point Estimate
  - Value of a statistic
  - Calculated from a sample
  - Estimates the parameter of the population

Discussion

- Different possible samples can be drawn from the same population
- Each of those samples can yield a different value of a statistic
  - Example: Mean and Variance
- It becomes important to investigate the sampling distributions of estimators
- **Sampling Distribution** for a given sample size, n:
  - Collection of all the estimators of that parameter
  - Of all possible samples of size ‘n’ from the population

Central Limit Theorem

- Statement
  - The **central limit theorem** states that given a distribution with a mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean (μ) and a variance σ^2/N as N, the sample size, increases

Sampling Distribution: Mean

- The sampling distribution of the Mean
  - Becomes approximately normal as the size of the sample 'n' increases
  - **Regardless of the shape of the population**
- Standard deviation of the sampling distribution of the mean: Also known as the **standard error**
- Where σ is the standard deviation of the population

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- http://onlinestatbook.com/

Exercise

- Savings account in a bank are normally distributed with mean Rs 2000 and standard deviation Rs 600
- The bank conducts a study by selecting 100 random accounts
- Find the probability that the mean of the sample will lie between Rs 1900 and Rs 2050

- Answer: 0.7492

Exercise

- Distribution of income of a certain category of bank employees has a mean of Rs 1,50,000 and a standard deviation of Rs 20000
- If a random sample of 30 is selected, what is the probability that the mean salary of the sample will exceed Rs 1,57,500?

Interval Estimation

- Definition:
  - It is a range of values related to a parameter
  - Calculated based on the sample
  - Such that
    - The parameter will be within that range
    - With some degree of confidence
- Use
  - A statistic, such as the mean, can be presented
    - As a Point Estimate, X
    - As an interval, X $\pm E$ where ***E* is the margin of error**

Point v/s Interval Estimates

- Point estimate is often insufficient
  - It is either right or wrong!
  - It also does not indicate the confidence level in that estimate

- Interval estimate
  - Better option : report an interval estimate
  - It provides the range, as well as the degree of confidence

Confidence Interval : Mean

- Margin of error for the Mean
  - Where the population SD is known

$$\bar{x} - Z_{\alpha/2}\sigma / \sqrt{n} < \mu < \bar{x} + Z_{\alpha/2}\sigma / \sqrt{n}$$

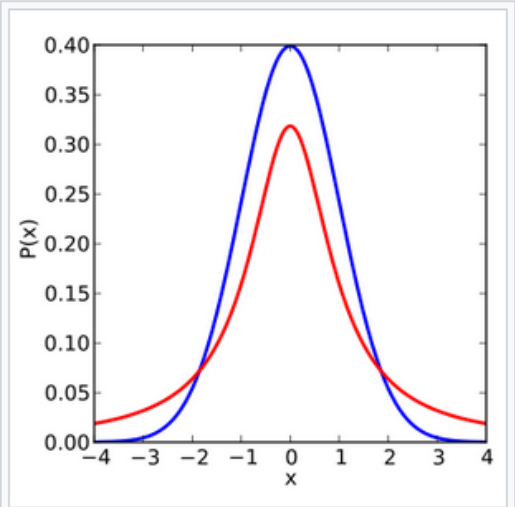
  - Where sample size is large, and population SD is not known

$$\bar{x} - Z_{\alpha/2}S / \sqrt{n} < \mu < \bar{x} + Z_{\alpha/2}S / \sqrt{n}$$

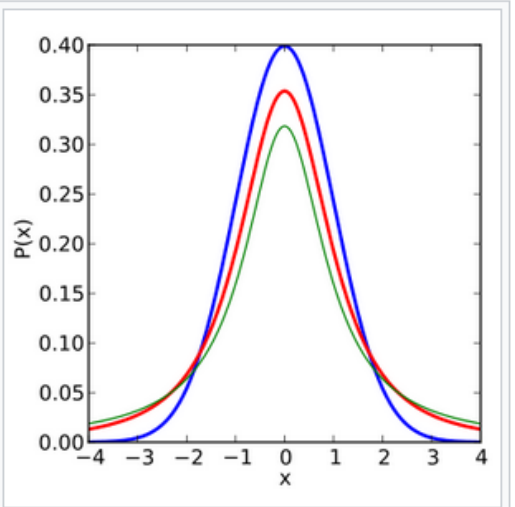
  - Where sample size is < 30, student t-distribution is used (**degrees of freedom = n-1**)

$$\bar{x} - t_{\alpha/2}S / \sqrt{n} < \mu < \bar{x} + t_{\alpha/2}S / \sqrt{n}$$

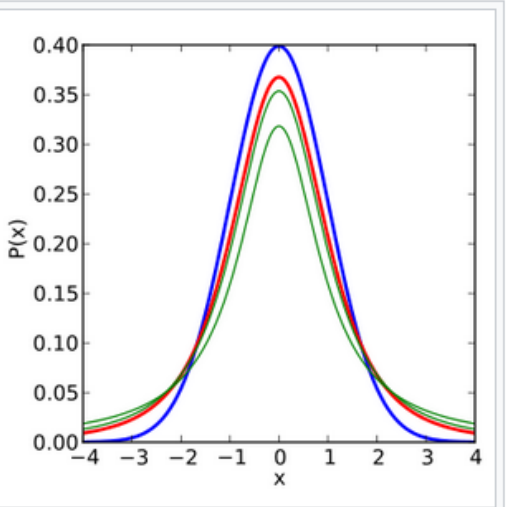
The t-distribution



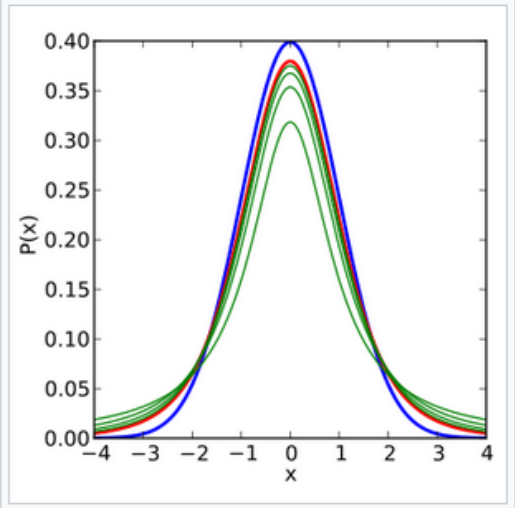
1 degree of freedom



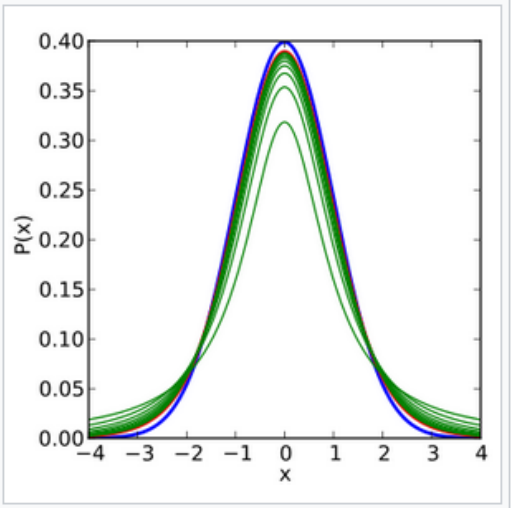
2 degrees of freedom



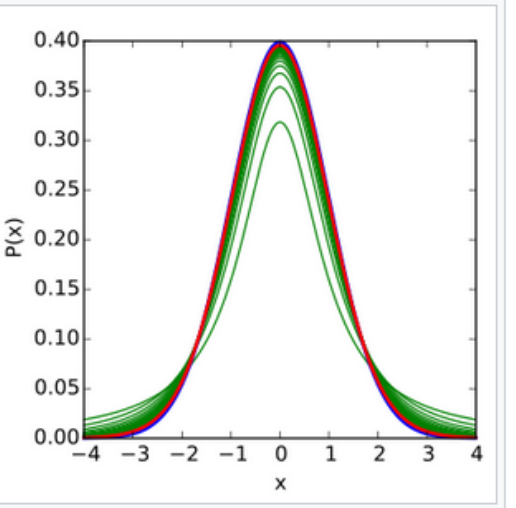
3 degrees of freedom



5 degrees of freedom



10 degrees of freedom



30 degrees of freedom

Exercise

- Mercury needs to be estimated in the water of a certain area. 16 samples are collected and their ppm values of mercury are as given below

409, 400, 406, 399, 402, 406, 401, 403, 401, 403, 398, 403, 407, 402, 410, and 399

- Manually calculate the mean, variance and standard deviation
- Estimate the ppm levels of mercury as an interval for 95% confidence level

Answer

- Sample size small
  - n <= 30
  - Hence students t-distribution will be used
- n = 16
- Sample mean = 403.063
- Variance = 12.996
- Standard deviation = 3.605
- Confidence level 95%, DOF = 16-1 = 15
- From t charts $t_{.025}$ = 2.131
- **Margin of error = 2.131 * 3.605 / sqrt(4) = 1.92**
- Based on this the interval (401.143, 404.983)

Problem

A random sample of 100 families from a large city is chosen to estimate the current average annual demand for milk in that city. The mean family demand from the sample is 150 gallons with a standard deviation of 40 gallons.

a) Construct a 95% C.I. for the mean annual demand of milk by all families in the city.

Exercise

- In a factory that used coal as fuel, the consumption was observed for 10 consecutive weeks. It was found that an average value of 11400 tons of coal was consumed per day with a standard deviation of 700 tons.
- From this data, the plan manager wants to estimate an interval for the mean consumption such that he can be 95% confident about the coal requirement. Can you help him out?

Sample Size Determination

- Why determine sample size?
  - To ensure that the error in estimating a population parameter is less than a desired threshold
- When sample is too small:
  - Required precision is not achieved
- When sample size is too large
  - Wastage of resources required to estimate the parameter

Sample Size Determination

- In the case of sampling distribution of the mean, the margin of error is:

$$E = Z_{\alpha/2}.\sigma / \sqrt{n}$$

- Therefore

$$n = (Z_{\alpha/2}.\sigma / E)^2$$

Exercise

- In a normal distribution with mean 375 and SD 48, what should be the size of the sample to ensure that the mean will be between 370 and 380 with a probability of 0.95

- Answer: More than 355

Exercise

- Goal: To estimate the average weight of watermelons
  - With 95% certainty that the error is 0.1Kg
- Problem
  - How many watermelons should be included in the sample?

Solution

- E = 0.1Kg
- $(1-\alpha) = 0.95$; $\alpha = 0.05$;
- $\alpha/2 = 0.025$; $1-\alpha/2 = 0.975$
- $Z_{\alpha/2} = 1.96$

- How to estimate σ ?
- What will ‘n’ turn out to be?

Confidence Interval: Variance and SD

- The Variance is a “sum of squares of items”

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N}$$

- The Chi-squared distribution : best represents the probability distribution of such “sum of squared items”
- Therefore Chi-squared distribution is used to derive the confidence interval of sampling distribution of variance (and, hence, the standard deviation)

Confidence Interval: Variance and SD

- Given $S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- Chi-squared CDF is given by:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

- If (1-α) is the desired confidence level
- From the CDF Table, this will be the region between (α/2) and (1- α/2)

$$P\left(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}\right) = 1 - \alpha$$

- Hence: $\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$

Confidence Interval: Variance and SD

- Confidence interval: Standard Deviation

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right)^{1/2} < \sigma < \left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right)^{1/2}$$

Exercise

- 100 healthy adults were subject to driving hazards
- Their response times were measured, and the variance calculated was
  - 0.0196 seconds squared
- For 95% confidence level, find the interval within which this value will lie

Solution

- n = 100; n-1 = 99
- S^2 = 0.0196; SD = 0.14
- Confidence level = 95% = 0.95
- (1-α) = 0.95; α = 0.05;
- $\alpha/2$ = 0.025; 1-$\alpha/2$ = 0.975
- $\chi_{0.025}$ = 73.36; $\chi_{0.975}$ = 128.422

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$$

- (100-1) * 0.0196 / (128.422) = 0.0151
- (100-1) * 0.0196 / (73.36) = 0.02645
- 0.015 < σ^2 < 0.0264
- 0.123 < σ < 0.1625

Confidence Interval: Proportions

- Another important “statistic” is the “proportion”
- We are often interested in:
  - Proportion of the population satisfying a certain criteria
    - Proportion of population above / below poverty line
    - Proportion of travellers reporting sick on arrival
    - Proportion of population using public transport
    - Proportion of kids dropping out of school by the age of 15

Confidence Interval: Proportions

- Let ‘P’ be the TRUE value of the proportion in the population
- Let ‘n’ be the size of the sample drawn from the population
- Let ‘X’ be the number of elements in the sample that exhibit the attribute under study
- The ‘estimated value’ of the TRUE proportion of the population is given by : $\hat{p} = X/n$

Confidence Interval: Proportions

- It can be shown that Z (it is normally distributed)

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

- and

$$\hat{p} - Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

- Thus, if number of elements ‘n’ in the sample and the proportion p^ in the sample are known, one can estimate the interval of the TRUE proportion of the population at a given confidence level

Example

- In a sample of 500 employees, 160 preferred taking training classes in the morning. What would be the 95% confidence interval for the TRUE proportion of employees preferring morning classes?

Solution

- x = 160
- n = 500
- p^ = 160 / 500 = 0.32
- alpha (for 95%) = 0.05
- alpha/2 = 0.025
- Substituting in

$$\hat{p} - Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

- $Z_{0.025}$ = 1.96
- 1.96 * sqrt(0.32 * (1-0.32)/500) = 0.041
- 0.32 – 0.041 < p < 0.32 + 0.041
- 0.279 < p < 0.361

Estimation: Additional Exercises

3.1 The percentage of copper in a certain chemical element is measured 6 times. The standard deviation of repeated measurements in such an analysis is known to be 2.5%. The sample mean is 14.1%. Construct a 95% C.I. for the true percentage of copper, assuming that the observations are approximately normally distributed.

3.2 25 measurements are made on the speed of light. Those averaged to 300007 with an SD of 10, the units being in Kilometers per second. Report your estimate of the speed of light as a 95% C.I. (1 Km = (5/8) mile).

3.3 A laboratory has a method for measuring lengths, using modern laser technology. The operator's job is to calibrate a yardstick. Measurements were taken 25 times, resulting in an average of 0.910835 meters, with a standard deviation of 45 microns (a micron is one millionth of a meter). Find an approximate 95% C.I. for the exact length of this stick. (Using this modern laser technology, the length can be measured to within one wave length of visible light which is about half a micron.)

Estimation: Additional Exercises

3.4 An investigator made 10 measurements of a metric standard and obtained an average of 1.0002 meters, with a standard deviation of 0.0001 meters. Construct a 90% C.I. for the exact length.

3.5 The weight of v7 similar containers of sulfuric acid is: 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 ounces. Find an 85% C.I. for the mean of all such containers assuming an approximate normal distribution.

3.6 An efficiency expert wishes to determine the average time it takes to drill three holes in a certain clamp. How large a sample will he need to be 95% confident that his sample mean will be within 15 seconds of the true mean? Assume that it is known from previous studies that sigma is 40 seconds.

3.7 A random sample of 8 cigarettes of a certain brand has an average nicotine content of 1806 milligram and a standard deviation of 2.4 milligram. Construct a 99% C.I. for the true average of nicotine content of this particular brand of cigarettes.

Estimation: Additional Exercises

3.8 A random sample of 100 families from a large city is chosen to estimate the current average annual demand for milk in that city. The mean family demand from the sample is 150 gallons with a standard deviation of 40 gallons.

a) Construct a 95% C.I. for the mean annual demand of milk by all families in the city.
b) If the rage you obtained in a) is larger than you are willing to accept, in what way can you narrow it?

3.9 In a part of a large city in which houses were rented, an economist wishes to estimate the average monthly rent correct to within US$50, a part from a 1-in-20 chance. If he guesses from past experience that sigma is about US$40, how many houses must he include in his sample?

3.10 The yield of alfalfa fro 9 plots were 0.8, 1.3, 1.5, 1.7, 1.7, 1.8, 2.0, 2.0, and 2.2 tons per acre. Set a 95% C.I. for the true average yield.

3.11 A manufacturer of batteries guarantees them to last for a specified period of time and wants to know how much variability there is in the lifetime of the batteries. A sample of 20 batteries was tested for longevity and S2 was found to be 53 hours. Suppose that the lifetimes are normally distributed, estimate the true variability in the life time as a 99% C.I.

Estimation: Additional Exercises

3.14 An electrical engineer wishes to estimate the variation in the amount of heat generated by a certain type of electronic component in order to design an appropriate heat dissipater for it. He took a sample of 16 components and observed the following units of heat generated: Find a 95% confidence interval for the true variation in heat generation.

4.260, 3.882, 4.741, 3.897, 4.925, 4.021, 4.822, 4.113, 4.628, 4.013, 4.728, 4.224, 4.171, 4.585, 4.509, 4.419

3.15 The following are the weights, below, in ounces, of 10 packages of grass seeds distributed by a certain company: Find a 95% confidence interval for the variation in all such packages distributed by this company.

16.9, 15.2, 16.0, 16.4, 16.1, 15.8, 17.0, 16.1, 15.9, 15.8

3.16 The vitamin C concentration (in mg per 100 gm) in a sample of size 17 of canned orange juice is: 16, 22, 21, 20, 23, 21, 19, 15, 13, 23, 17, 20, 29, 18, 22, 16, and 25. Find a 90% C.I. for the true variation in Vitamin C concentrations.

Estimation: Additional Exercises

3.17 In a sample of 31 patients, the amount of an anesthetic required to produce anesthesia suitable for surgery was found to have a standard deviation (from patient to patient) of 10.2 mg. Compute a 98% confidence interval on the population standard deviation.

3.18 Five out of 50 randomly selected time sharing terminals give incorrect character response. A firm has 800 of these terminals.

a) Estimate the proportion of terminals that give incorrect response.
b) eport your estimate as a 95% confidence interval on the true population proportion.

3.19 A manufacturer of flashcubes wants to estimate the probability that a flashcube will work. Since, destructive testing is involved; he wants to keep the sample size as small as possible. Find the number of observations that must be taken to estimate the probability within 0.04 and with 95% confidence of that if

a) He has no idea of the percent defective.
b) He believes that the percent defective is no more than 6%.

3.20 A public Library wants to estimate the percentage of books in its collection that have publication dates of 1970 or earlier. How large a random sample must be taken to be 90% sure of coming within 5% of the actual proportion?

ADDITIONAL SLIDES

Random Variables

- **Random experiment**
  - Process of measurement or observation in which the outcome **cannot** be completely determined in advance
- **Sample space**
  - All possible outcomes of a random experiment
- **Random Variable**
  - A real-valued quantity, or numerical measure, whose value depends on the outcomes of a random experiment
  - Can be **Discrete** or **Continuous**

Random Variable

- The probability that a Random Variable may assume a particular value is governed by a Probability Function:
  - For Discrete variables
    - **Probability Mass Function (PMF)**
  - For Continuous variables
    - **Probability Density Function (PDF)**
  - For Discrete / Continuous variables
    - **Cumulative Distribution Function**

Random Variable and Probabilities

- Let X be the Random Variable
- Let x be one of its possible values
- Let P(x) be the probability that X=x
- Then

$$0 \leq \mathrm{P}(\mathrm{x}) \leq 1$$

$$\sum_{\text{all values of } x} P(x) = 1$$

Random Variable: Expected Value

- Expected value : E(X)
  - Weighted average, of all possible values, considering their probabilities

- For Discrete random variable

$$\mathrm{E(X)} = \mu_X = \sum_{all\ values\ of\ x} [x.p(x)]$$

- For Continuous random variable

$$\mathrm{E(X)} = \mu_X = \int_{-\infty}^{\infty} x.f(x)dx$$

Random Variable: Variance and Standard Dev

- Variance of a Random Variable

$$\sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$$

- Standard Deviation of a Random Variable

$$\sigma_x = +\sqrt{\sigma_X^2}$$

- Where

$$E(X^2) = \sum_{\text{all values of } x} [x^2 . p(x)], \text{ in the discrete case, and}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 . f(x) dx, \text{ in the continuous case}$$

- Variance and Standard Deviation reflects the extent to which the Random variable is close to its mean

Sample – to – Population

- Goal of all these definitions:
  - Given the nature of the phenomena
    - Probability distributions
  - And a sample with certain deductions
    - Measured observations
  - Predict additional properties and confidence levels
- We therefore need to
  - Understand generic phenomena
  - And the probabilistic nature of their events

Probability Distributions

- Binomial Distribution
- Poisson Distribution
- Students T-Distribution
- Chi-Square Distribution
- F-Distribution
- Normal Distribution
- Log normal Distribution
- Other Distributions
  - Bernoulli
  - Geometric
  - Hypergeometric
  - Multinomial
  - Exponential
  - Beta
  - Gamma

Use of Probability Functions of “R”

- dxxxx
  - Probability density function
  - Given a ‘number’ (quantile) this function will tell you the probability of getting that number (quantile)
- pxxxx
  - Cumulative probability distribution function
  - Given a ‘number’ (quantile) this function will tell you the cumulative probability of getting all values up to that number
- qxxxx
  - Given a cumulative probability, this function returns the ‘number’ (quantile) associated with that probability
- rxxxx
  - This function generates the required number of data points conforming to the desired probability distribution (as per the specified parameters)

Normal Distribution

- The Cumulative Distribution Function for Normal Distribution is available
  - In the form of a Standard Normal Distribution Table
  - Based on $Z = (x-\mu)/\sigma$
- The Standard Normal Distribution table is used in solving problems

Some properties of probability distributions

- Central Moments
  - the expected value of a specified integer power of the deviation of the random variable from the mean

$$\mu_n = \mathrm{E}\left[(X - \mathrm{E}[X])^n\right] = \int_{-\infty}^{+\infty} (x-\mu)^n f(x)\,\mathrm{d}x$$

- The moments and their interpretations
  - Zeroth central moment (= 1)
  - First central moment (the mean)
  - Second central moment (measure of variance)
  - Third central moment (measure of skewness)
  - Fourth central moment (measure of kurtosis)