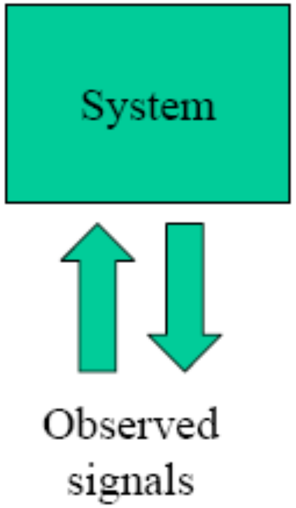
ENGINEERING STATISTICS

Lale Alatan, Elif Vural

(Based on previous talk by Sencer Koç)

Middle East Technical University
Department of Electrical and Electronics Engineering
Spring 2017-2018

Engineering System Analysis

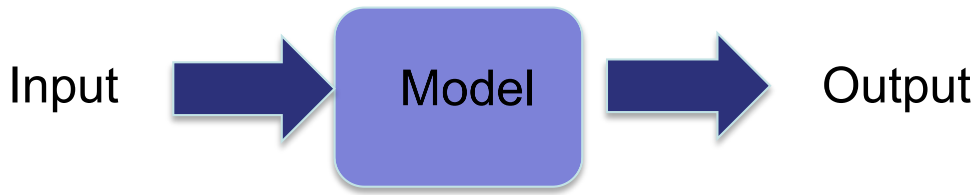


- Use observations to qualitatively and quantitatively **understand** a system.
- Use mathematics to determine how a set of interconnected components behave in response to a given input

Questions

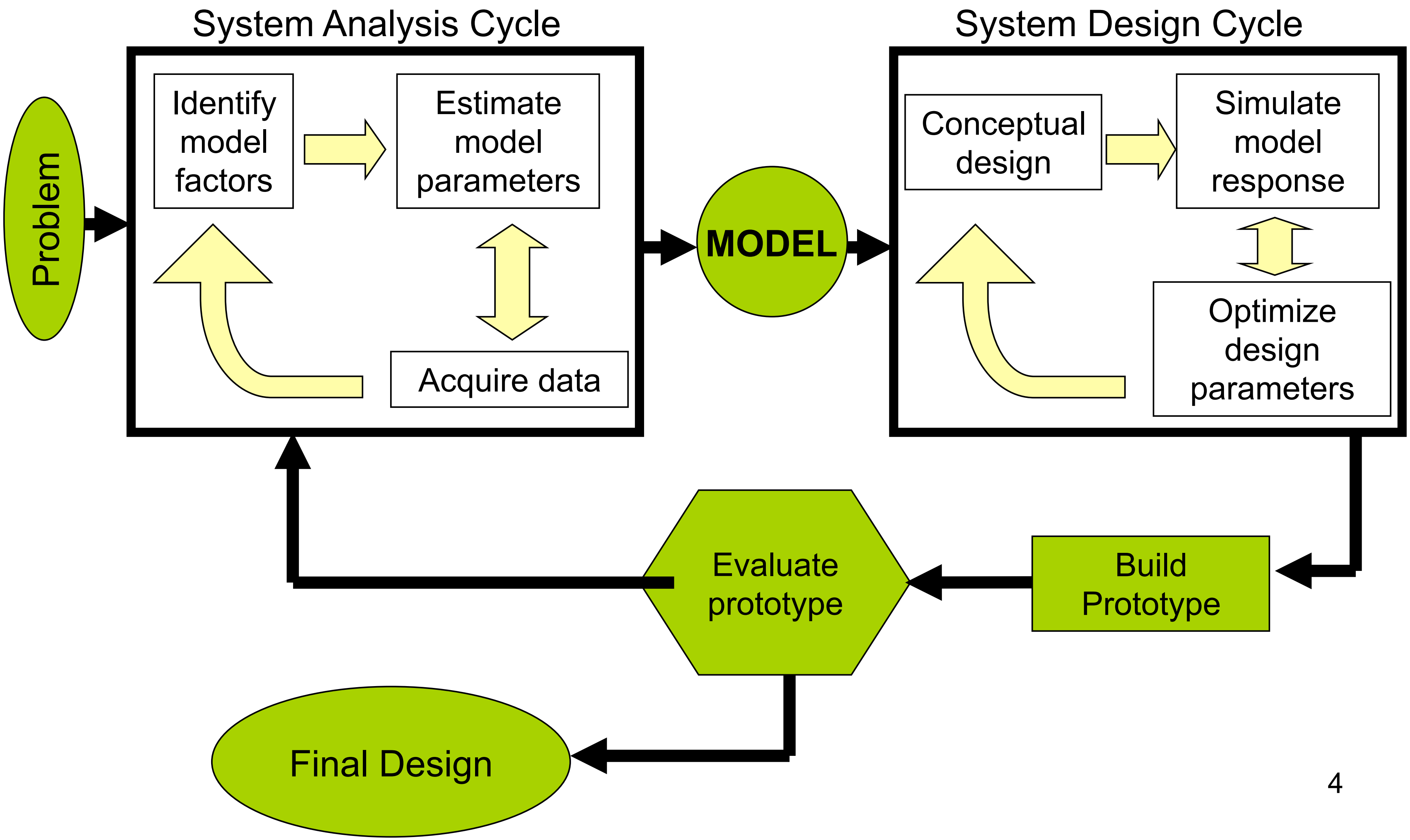
1. What is meant by “understanding a system”?
   - Predict **future outcomes** from the system based on **hypothetical inputs.**

2. How to formalize this?
   - By a **model** that maps **input signals** to **output signals**.

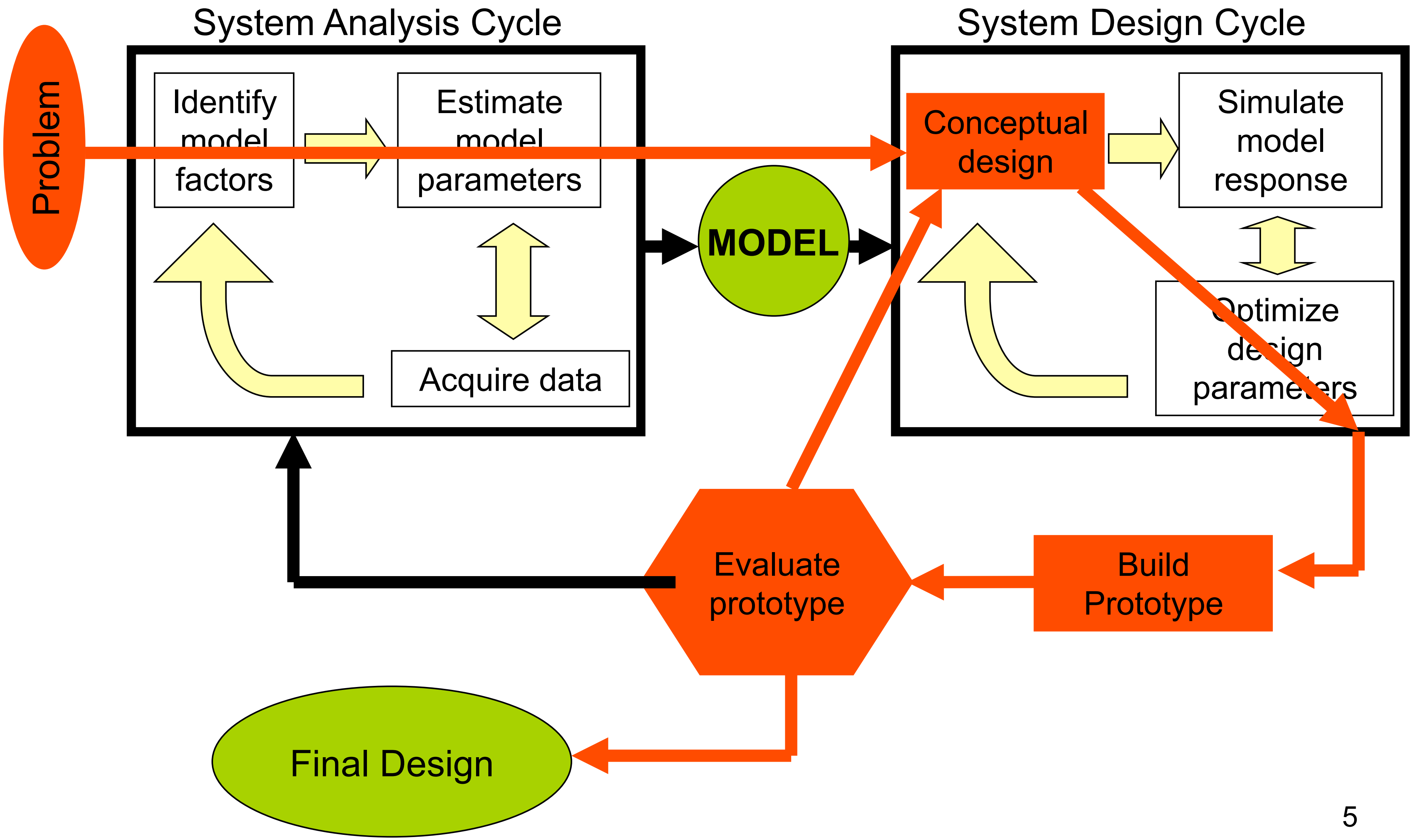


3. Why is this important?
   - A system model is a key component in the systems engineering design cycle.

Systems Engineering Cycle



How **not** to solve a design problem

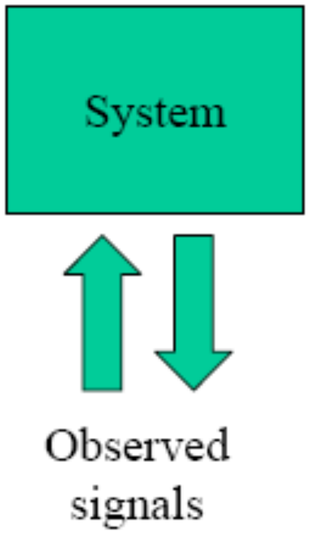


Models

- **Model:** A comprehensible simplified description of a real world system

- Engineering systems analysis:
  – Process of using observations to identify a model of a system

- Modeling a system:
  – Find correlations or patterns in the observed signals.



Statistical framework

- Measuring real signals is a statistical process:
  - Observed signals will be noisy
  - Noise must be included in the modeling process.

  ➡ All modeling is inherently a statistical process.

  - Models of systems are uncertain approximations of the real world.
  - The modeling error itself is interpreted as a statistical process.

A systems engineer should have a good understanding of **statistical** modeling and **statistical** decision methodology

Measurement issues

- **Validity:** Faithfully representing the aspect of interest; i.e.: usefully or appropriately represents the feature of an object or system

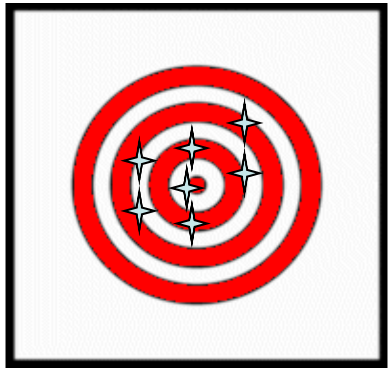
- **Precision:** Small variation in repeated measurements

- **Accuracy (unbiasedness):** Producing the “true value” “on average”

Precision and accuracy

Not Accurate
Not Precise

Accurate, Not
Precise



Precise, Not
Accurate

Accurate and
Precise

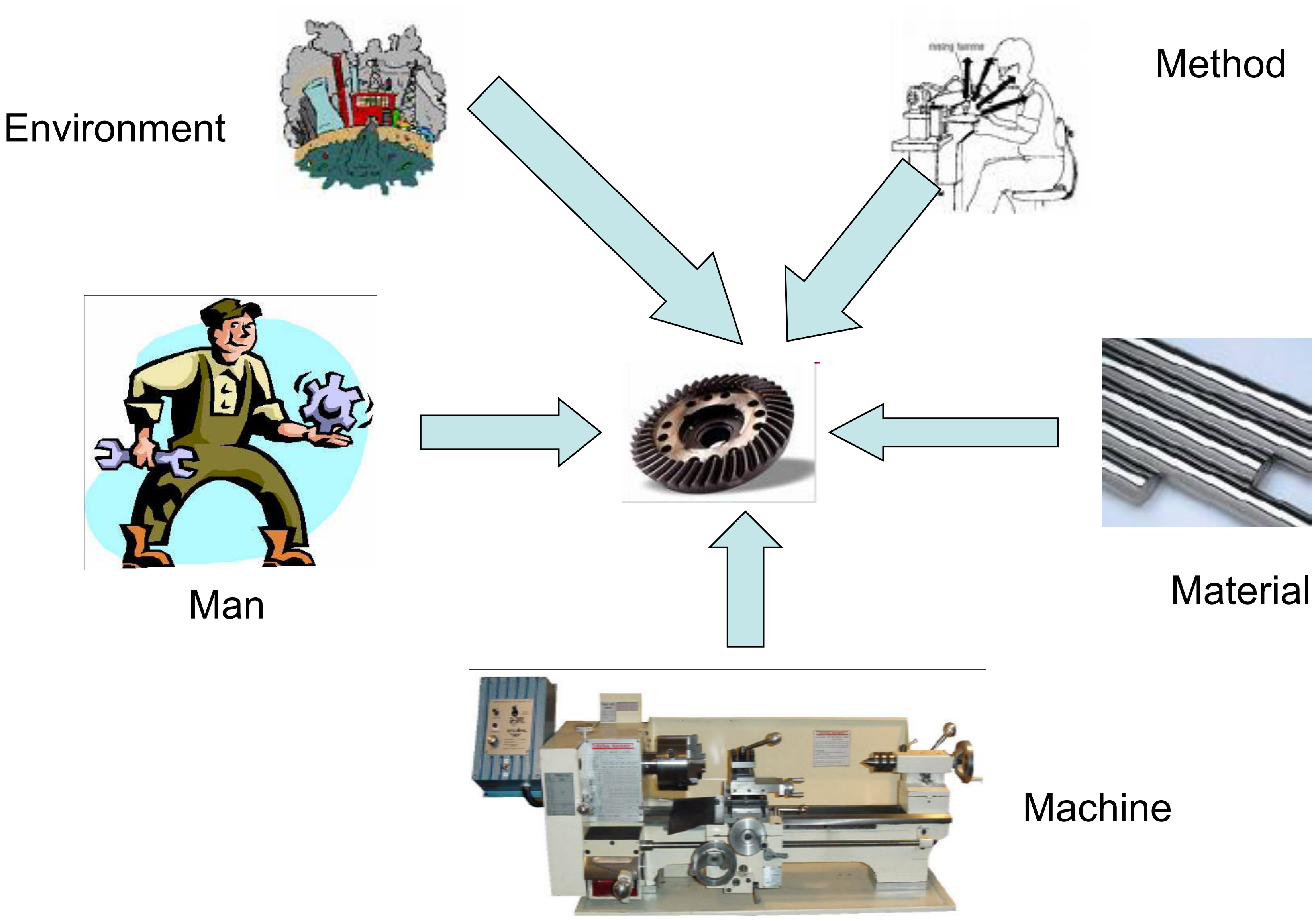
Statistical thinking

- Statistical methods are used to help us describe and understand variability.
- By variability, we mean that successive observations of a system or phenomenon do not produce exactly the same result.

Are these gears produced exactly the same size?

NO!

Sources of Variability



Random variables

- We often model a measurable quantity X as a random variable.
- The probability density function is assumed to be known.
  - A common choice is the Gaussian distribution (Central Limit Theorem)

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

Estimation of model parameters

- Our purpose is to estimate certain parameters of $f(x)$, (mean, variance) from observation of the samples.
- Observe samples from the distribution:

$$R = \{X_1, X_2, X_3, \ldots, X_i, \ldots, X_N\}$$

Sample mean: $M = \frac{1}{N}\sum_{i=1}^{N} X_i$ ➡ Point estimate of μ

Sample variance: $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - M\right)^2$ ➡ Point estimate of σ^2

Examples

Sample (N = 10)	M	S
{55,41,50,44,55,56,48,29,51,66}	49.5	10.01
{60,34,49,43,40,38,53,46,51,46}	46	7.69
{45,54,57,71,36,40,60,46,36,53}	49.8	11.29
{66,57,70,55,69,47,39,48,62,39}	55.2	11.64
{56,44,56,39,51,30,45,55,47,62}	48.5	9.49
{44,27,38,61,49,54,59,29,44,43}	44.8	11.47

Point estimates as random variables

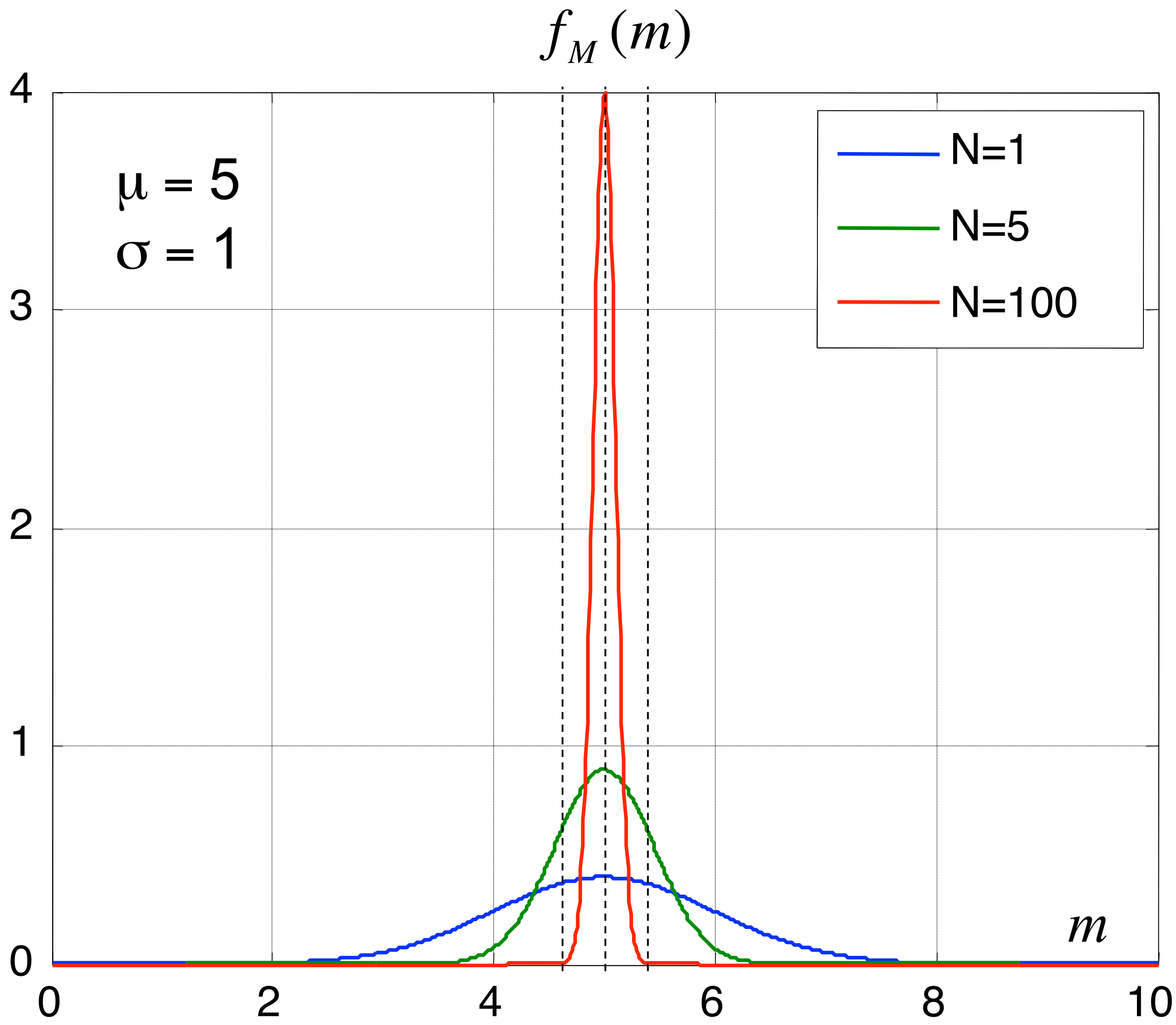
- Sample mean and standard deviation depend on the random samples chosen
  - M and S are random variables

Sample mean: $M = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad E\{M\} = \mu,\ \sigma_M^2 = \sigma^2 / N$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$f_M(m) = \frac{1}{\sqrt{2\pi}\sigma_M}\exp\left(-\frac{(m-\mu)^2}{2\sigma_M^2}\right)$$

Distribution of Sample Mean

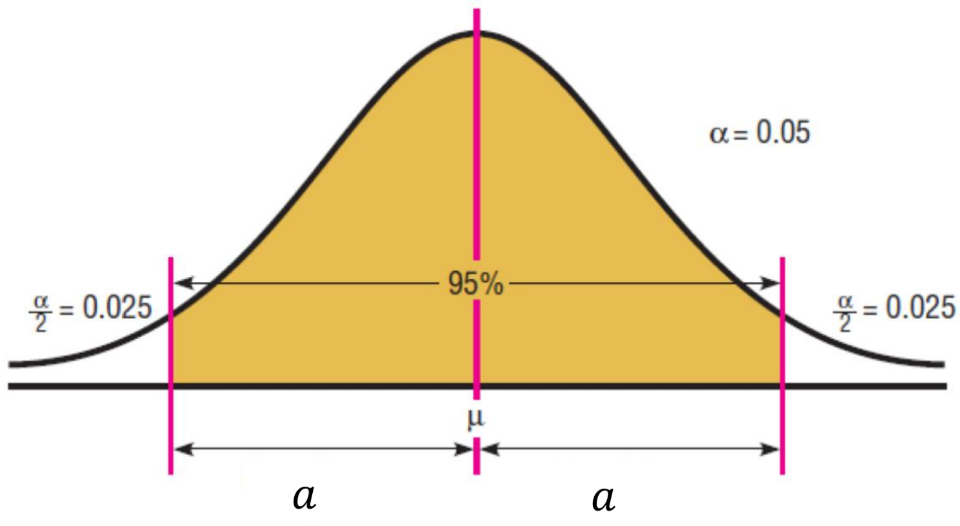


For larger sample sizes N the mean estimate is closer to the mean with higher probability.

$$\sigma_M^2 = \frac{\sigma^2}{N}$$

Confidence intervals

We want to determine an interval I for the actual mean μ so that $P(\mu \in I) = 1 - \alpha$



$$P(\mu - a \le M \le \mu + a) = \int_{\mu - a}^{\mu + a} f_M(m)\,dm$$

$$= P(M - a \le \mu \le M + a)$$

Confidence intervals

- Given that X is a Gaussian random variable with mean μ and variance σ^2:

$$R:\{X_1, X_2, X_3, \ldots, X_i, \ldots, X_N\}$$

$$M = \frac{1}{N}\sum_{i=1}^{N} X_i; \quad S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - M)^2$$

$$Z = \frac{(M-\mu)}{\sigma/\sqrt{N}}$$ has distribution N(0,1)

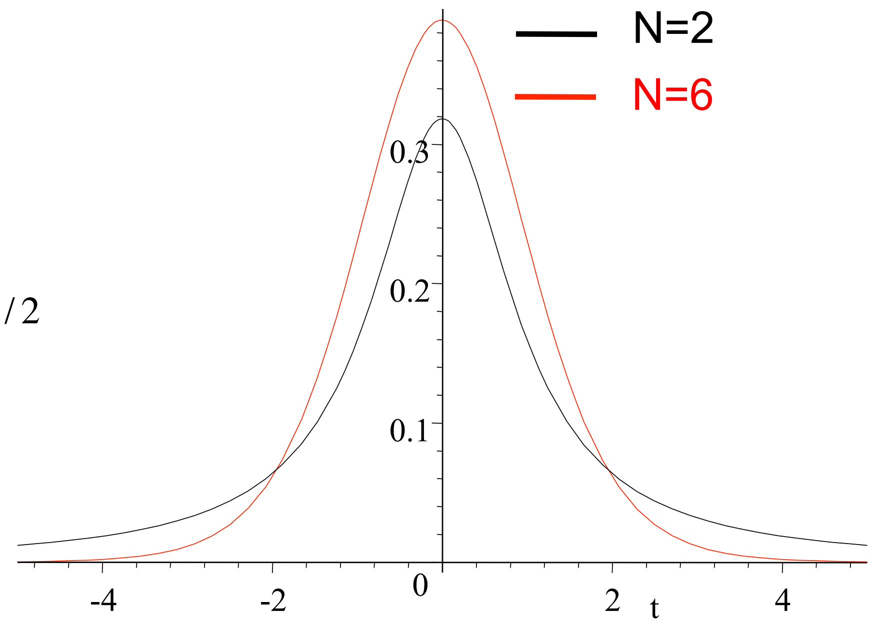
$$T = \frac{(M-\mu)}{S/\sqrt{N}}$$ has Student’s t-distribution

Student's t-distribution

$$T = \frac{(M - \mu)}{S/\sqrt{N}}$$ has pdf

$$f_T(t) = \frac{\Gamma[(k+1)/2]}{\Gamma(k/2)\sqrt{\pi k}}\left(1 + \frac{t^2}{k}\right)^{-(k+1)/2}$$

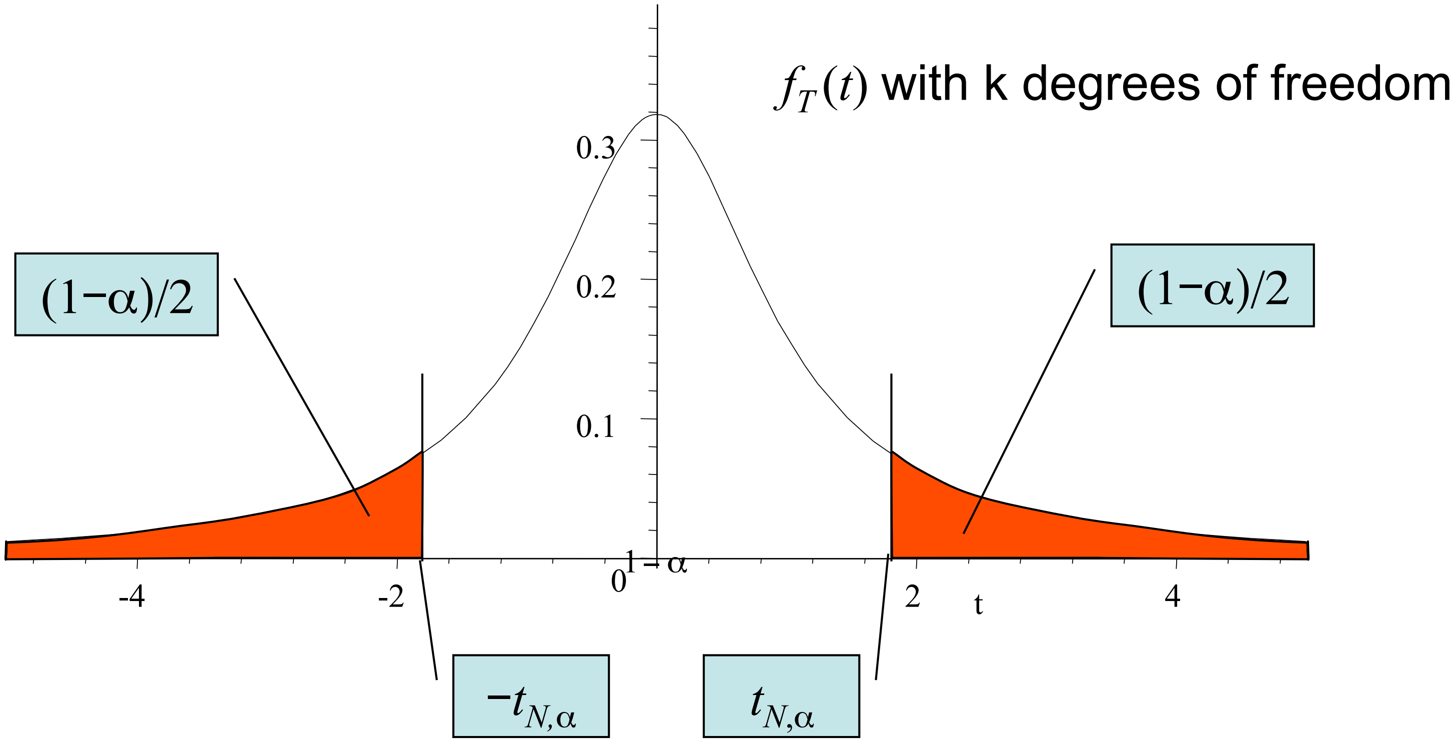
$$k = N - 1$$



This distribution is known as *Student's t-distribution* with k degrees of freedom.

The distribution is named after the English statistician W.S. Gosset, who published his research under the pseudonym "Student."

Confidence intervals



$$\alpha = P\left(-t_{N,\alpha} \leq T \leq t_{N,\alpha}\right) =$$

$$P\left(M - \frac{S}{\sqrt{N}} t_{N,\alpha} \leq \mu \leq M + \frac{S}{\sqrt{N}} t_{N,\alpha}\right)$$

Confidence intervals

- When we obtain the estimates M and S from the sample set, the actual mean μ will lie in the interval

$$\left[M - \frac{S}{\sqrt{N}} t_{N,\alpha}, \quad M + \frac{S}{\sqrt{N}} t_{N,\alpha} \right]$$

  with probability α. This is called a $\alpha \times 100$ percent confidence interval.

- The values for Student's t-distribution are tabulated.

Confidence coefficients of intervals

	Confidence coefficient α		
N	0.90	0.99	0.995
10	1.8331	3.2498	3.6897
50	1.6766	2.6800	2.9397
100	1.6604	2.6264	2.8713
500	1.6479	2.5857	2.8196

Example 1: Wire resistance

- Ten measurements were made on the resistance of a certain type of wire. Suppose that M=10.48 Ω and S=1.36 Ω. We want to obtain a confidence interval for μ with confidence coefficient 0.90. From the table

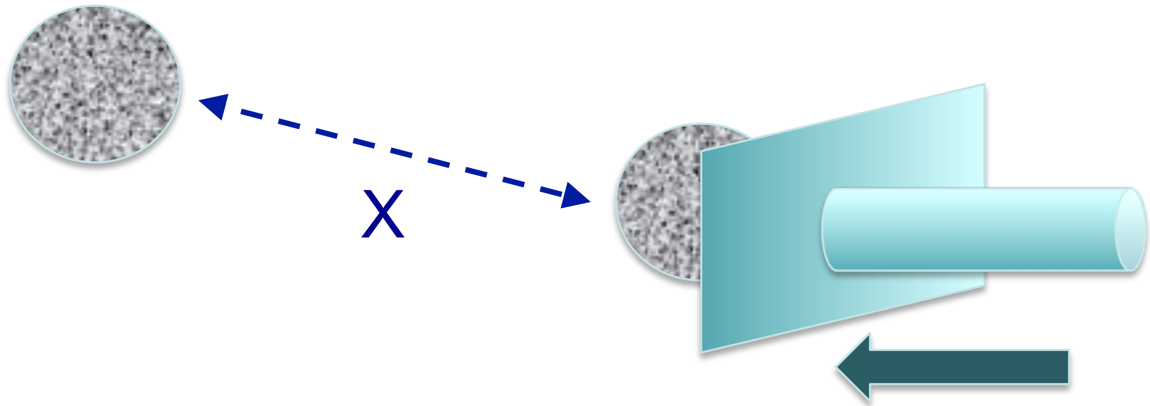
$$N = 10,\ \alpha = 0.9 \Rightarrow t_{10,0.9} = 1.83$$

$$\mu \in \left[10.48 - \frac{(1.36)}{\sqrt{10}}(1.83), 10.48 + \frac{(1.36)}{\sqrt{10}}(1.83)\right]$$

$= [9.69, 11.27]$ with probability 90%

Example 2: Robot rolling an object

- You design an actuator system whose purpose is to kick and roll an object. You are interested in estimating the distance the object travels before stopping.



Assume Gaussian distribution:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

1. Estimate the mean distance
2. How large can X be?

Example 2: Estimating average distance

Approach 1: Take actual measurements with a physically implemented system.

You take 10 measurements and find M=51.3 cm, S=6.4 cm

$$N = 10, \quad \alpha = 0.99 \Rightarrow t_{10,0.99} = 3.25$$

$\mu =$ Average distance the object travels

$$\mu \in \left[51.3 - \frac{6.4}{\sqrt{10}}(3.25),\ 51.3 + \frac{(6.4)}{\sqrt{10}}(3.25) \right]$$

$$= \left[44.72,\ 57.88 \right]$$ with probability 99%

Example 2: Estimating average distance

Approach 2: Form a system model and prepare a simulation setting

You simulate with 500 realizations and find M=54.2 cm, S=6.7 cm

$$N = 500, \quad \alpha = 0.99 \Rightarrow t_{500,0.99} = 2.58$$

$\mu =$ Average distance the object travels

$$\mu \in \left[54.2 - \frac{6.7}{\sqrt{500}}(2.58),\ 54.2 + \frac{(6.7)}{\sqrt{500}}(2.58) \right]$$

$$= \left[53.42,\ 54.97 \right]$$ with probability 99%

Worst-case analysis

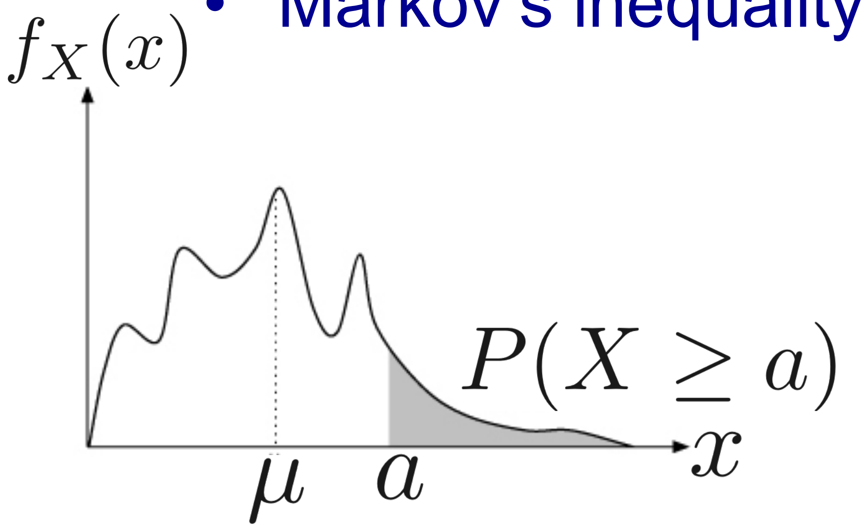
- Say we can estimate the mean and variance with high confidence:

$$M \approx \mu, \quad S \approx \sigma$$

- What about the maximum/minimum values the random variable can realistically take?
  - e.g. the “maximum distance” the ball is likely to travel

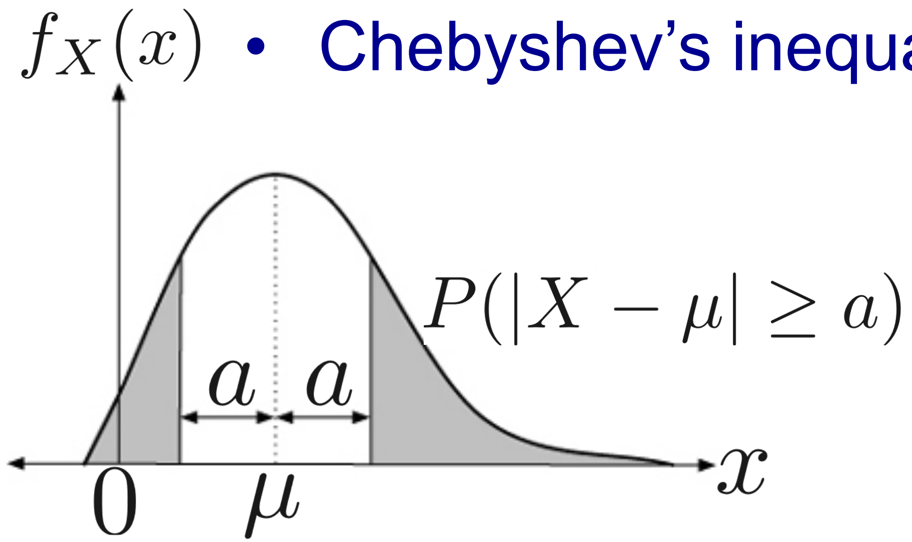
Probability inequalities

- Markov's inequality: For a random variable $X \geq 0$



$$P(X \geq a) \leq \frac{E[X]}{a}$$

- Chebyshev's inequality: For a random variable X



$$P(\,|X-\mu| \geq a) \leq \frac{\sigma_X^2}{a^2}$$

Example 2: “Maximum distance”

- Assume after many measurements you found:

$$\mu \approx 54.2, \quad \sigma \approx 6.7$$

- Apply Chebyshev’s inequality (no Gaussian assumption needed):

$$P(\,|X - 54.2| \geq a) \;\leq \frac{6.7^2}{a^2} = \overbrace{0.1}^{\text{For 90\% confidence}} \quad \Rightarrow \quad a = 21.2$$

  - So with probability at least 90%

$$|X - 54.2| < 21.2 \quad \Rightarrow \quad X < 75.4$$

Conclusion

- Problem ➡ Model Identification ➡ Design
- Prepare setups/simulations, take measurements, use tools from statistics to
  - Estimate important parameters: **Mean, variance**, …
  - Find confidence intervals
  - Take care of typical/worst cases, maximum/minimum parameter values in your design

QUIZ

List some of the important statistical parameters related to the distribution of a random variable.