Simulation Wrap-up, Statistics

COS 323

Last time

- Time-driven, event-driven
- “Simulation” from differential equations
- Cellular automata, microsimulation, agent-based simulation
- Example applications: SIR disease model, population genetics

Simulation: Pros and Cons

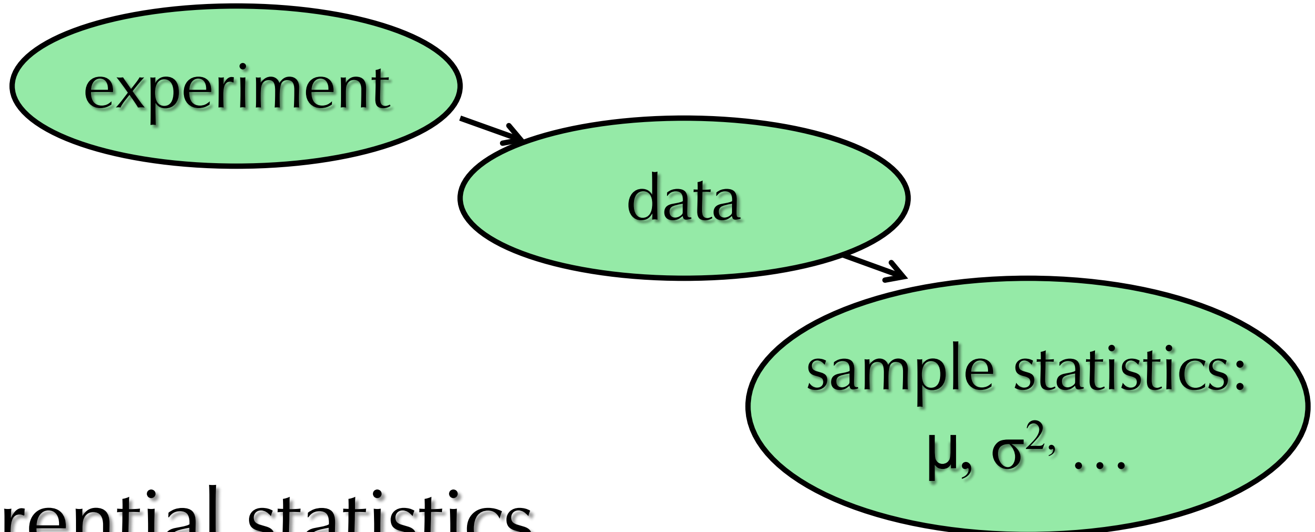
- Pros:
  - Building model can be easy (easier) than other approaches
  - Outcomes can be easy to understand
  - Cheap, safe
  - Good for comparisons
- Cons:
  - Hard to debug
  - No guarantee of optimality
  - Hard to establish validity
  - Can't produce absolute numbers

Simulation: Important Considerations

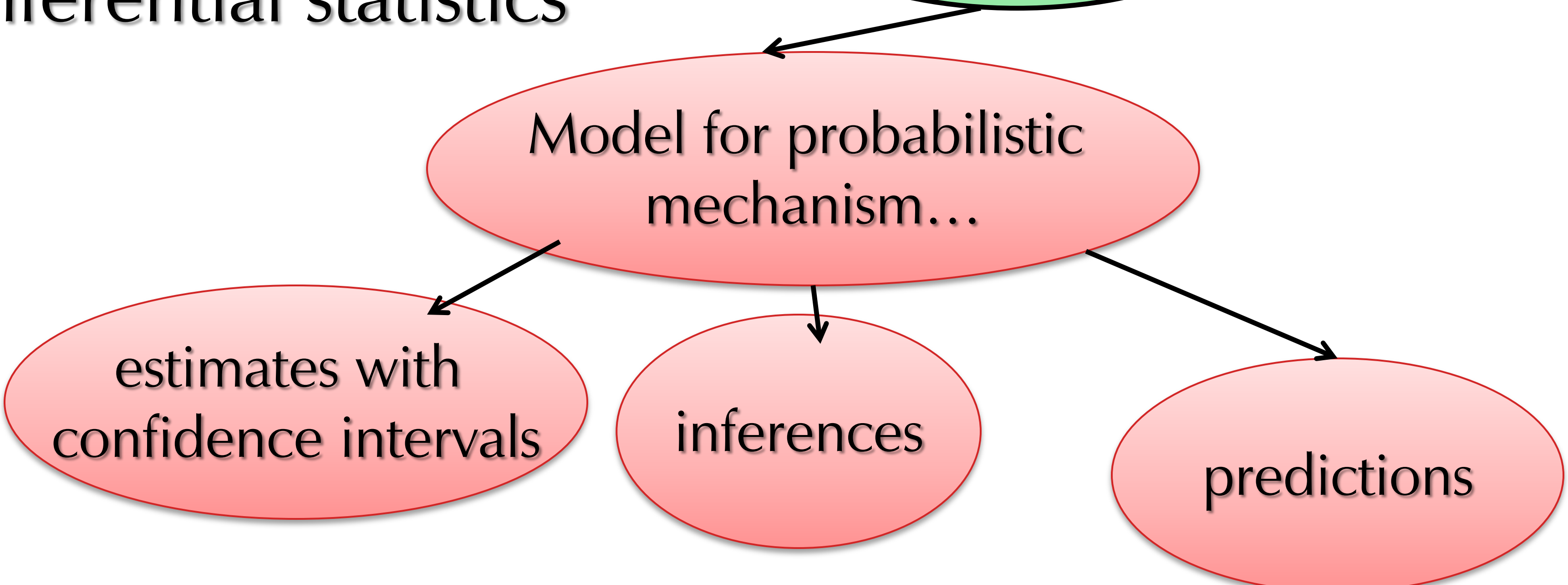
- Are outcomes statistically significant? (Need many simulation runs to assess this)
- What should initial state be?
- How long should the simulation run?
- Is the model realistic?
- How sensitive is the model to parameters, initial conditions?

Statistics Overview

Descriptive statistics



Inferential statistics



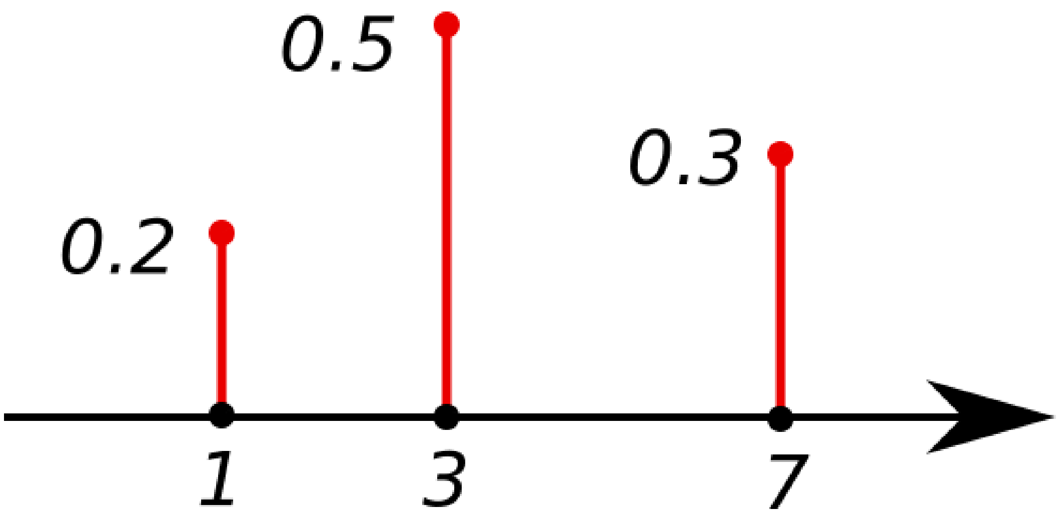
Random Variables

- A random variable is any "probabilistic outcome"
  - e.g., a coin flip, height of someone randomly chosen from a population
- A R.V. takes on a value in a *sample space*
  - space can be discrete, e.g., {H, T}
  - or continuous, e.g. height in (0, infinity)
- R.V. denoted with capital letter (X), a realization with lowercase letter (x)
  - e.g., X is a coin flip, *x* is the value (H or T) of that coin flip

Probability Mass Function

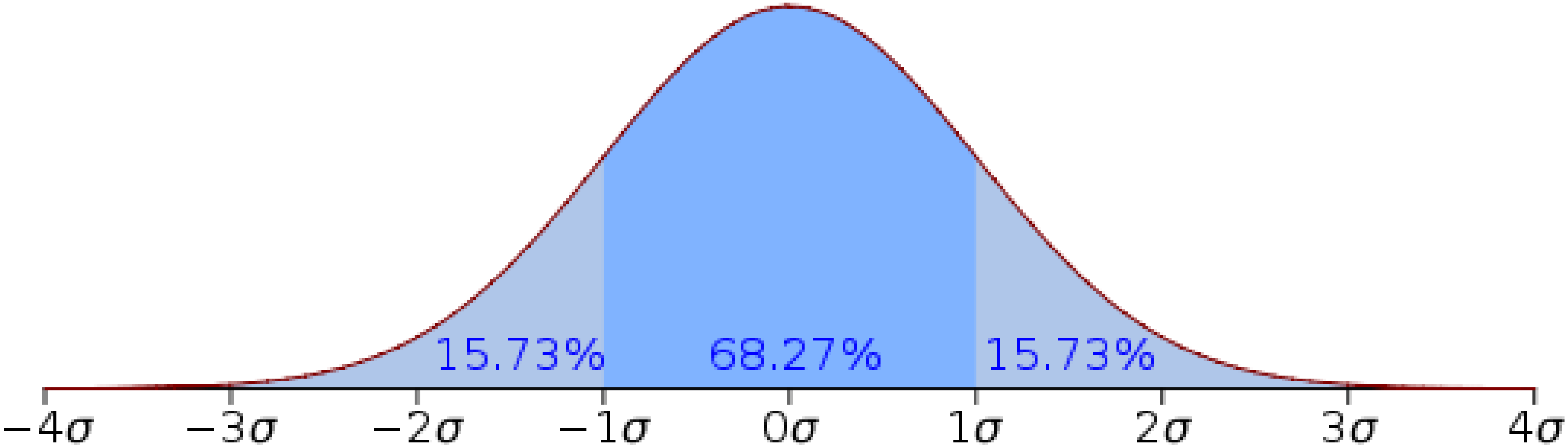
- Describes probability for a discrete R.V.
- e.g.,

$$f_X(x) = \begin{cases} \frac{1}{2}, & x \in \{0, 1\}, \\ 0, & x \notin \{0, 1\}. \end{cases}$$



Probability Density Function

- Describes probability for a continuous R.V.
- e.g.,



[Population] Mean of a Random Variable

- aka expected value, first moment
- for discrete RV: $\mathrm{E}[X] = \mu = \sum_i x_i p_i$

  $$\text{(requires that } \sum_i p_i = 1\text{)}$$

- for continuous RV: $\mathrm{E}[X] = \mu = \int_{-\infty}^{\infty} x\, p(x)\, dx$

  $$\text{(requires that } \int_{-\infty}^{\infty} p(x)\, dx = 1\text{)}$$

[Population] Variance

$$\sigma^2 = \mathrm{E}\left[(X-\mu)^2\right]$$
$$= \mathrm{E}\left[X^2 - 2X\mu + \mu^2\right]$$
$$= \mathrm{E}\left[X^2\right] - \mu^2$$
$$= \mathrm{E}\left[X^2\right] - \left(\mathrm{E}[X]\right)^2$$

- for discrete RV: $$\sigma^2 = \sum_{\mathrm{i}} p_i (x_i - \mu)^2$$
- for continuous RV: $$\sigma^2 = \int (x-\mu)^2 p(x)dx$$

Sample mean and sample variance

- Suppose we have N independent observations of X: $x_1, x_2, \ldots x_N$
- Sample mean:

Unbiased:

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

$$\mathrm{E}[\bar{x}] = \mu$$

- Sample variance:

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

$$\mathrm{E}[s^2] = \sigma^2$$

1/(N-1) and the sample variance

- The N differences $x_i - \bar{x}$ are not independent:

$$\sum (x_i - \bar{x}) = 0$$

- If you know N-1 of these values, you can deduce the last one
  - i.e., only N-1 degrees of freedom
- Could treat sample as population and compute population variance:

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

  - BUT this underestimates true population variance (especially bad if sample is small)

Computing Sample Variance

- Can compute as

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

- Prefer:

$$s^2 = \frac{\left(\sum_{i=1}^{N} x_i^{\ 2}\right) - N(\overline{x})^2}{N-1} = \frac{\left(\sum_{i=1}^{N} x_i^{\ 2}\right) - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2}{N-1}$$

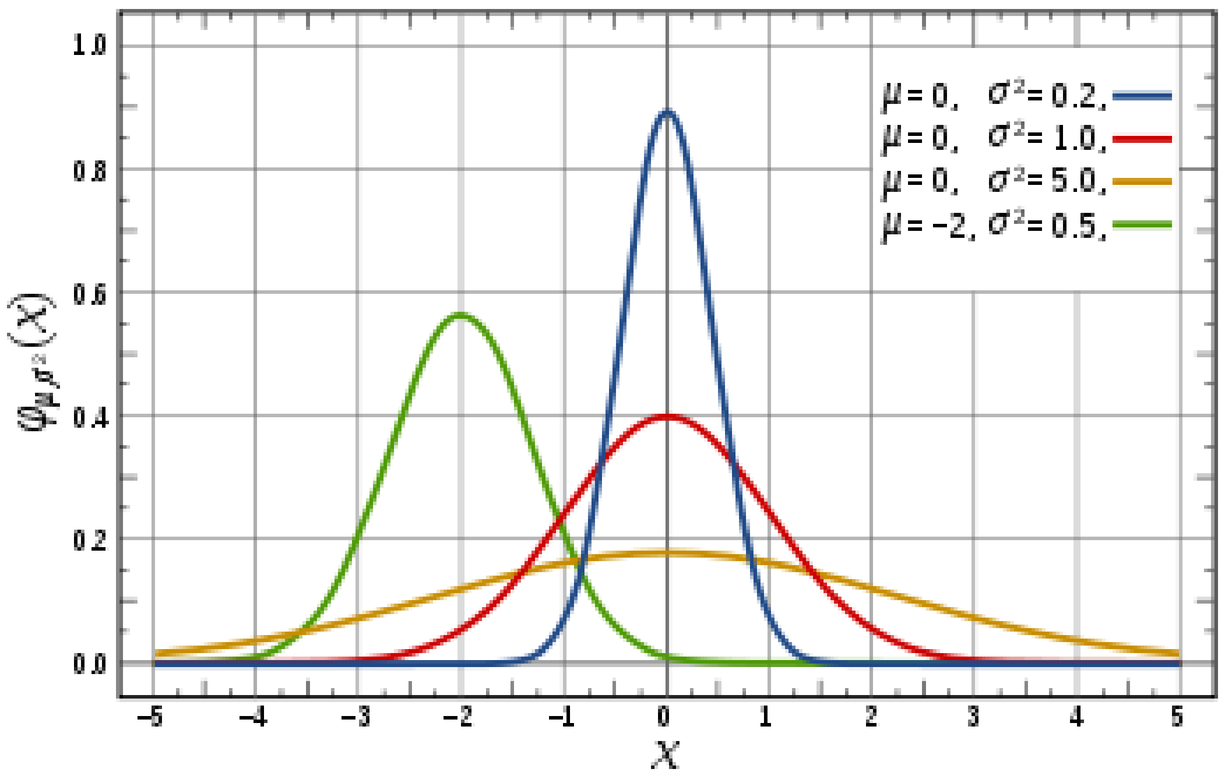
(one pass, fewer operations, more accurate)

The Gaussian Distribution

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

$$E[X] = \mu$$

$$Var[X] = \sigma^2$$



Why so important?

- **sum** of independent observations of random variables converges to Gaussian *(with some assumptions)
- **in nature**, events having variations resulting from many small, independent effects tend to have Gaussian distributions
  - demo: http://www.mongrav.org/math/falling-balls-probability.htm
  - e.g., measurement error
  - if effects are multiplicative, **logarithm** is often **normally** distributed

Central Limit Theorem

- Suppose we sample $x_1, x_2, \ldots x_N$ from a distribution with mean μ and variance σ^2
- Let

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

holds for *(almost) any parent distribution!

- then

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \rightarrow N(0,1)$$

- i.e., $\bar{x}$ distributed normally with mean μ, variance σ^2/N

Important Properties of Normal Distribution

1. Family of normal distributions closed under linear transformations:

   if $X \sim N(\mu, \sigma^2)$ then

   $(aX + b) \sim N(a\mu + b, a^2\sigma^2)$

2. Linear combination of normals is also normal:

   if $X_1 \sim N(\mu_1, \sigma_1^{\,2})$ and $X_2 \sim N(\mu_2, \sigma_2^{\,2})$ then

   $aX_1 + bX_2 \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^{\,2} + b^2\sigma_2^{\,2})$

Important Properties of Normal Distribution

3. Of all distributions with mean and variance, normal has **maximum entropy**

   Information theory: Entropy like "uninformativeness"

   Principle of maximum entropy: choose to represent the world with as uninformative a distribution as possible, subject to "testable information"

If we know x is in [a, b], then uniform distribution on [a, b] has least entropy

If we know distribution has mean μ, variance σ^2, normal distribution $N(\mu, \sigma^2)$ has least entropy

Important Properties of Normal Distribution

4. If errors are normally distributed, a least-squares fit yields the maximum likelihood estimator

Finding least-squares x st Ax ≈ b finds the value of x that maximizes the likelihood of data A under some model

$$P(data|model) \propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2}\left(\frac{y_i - y(x_i)}{\sigma_i} \right)^2 \right] \Delta y$$

$$P(model|data) \propto P(data|model) P(model)$$

Important Properties of Normal Distribution

5. Many derived random variables have analytically-known densities

   e.g., sample mean, sample variance

6. Sample mean and variance of n identical independent **samples** are independent; sample mean is a normally-distributed random variable

$$\overline{X}_n \sim N(\mu, \sigma^2 / n)$$

What if we don't know true variance?

- Sample mean is normally distributed R.V.

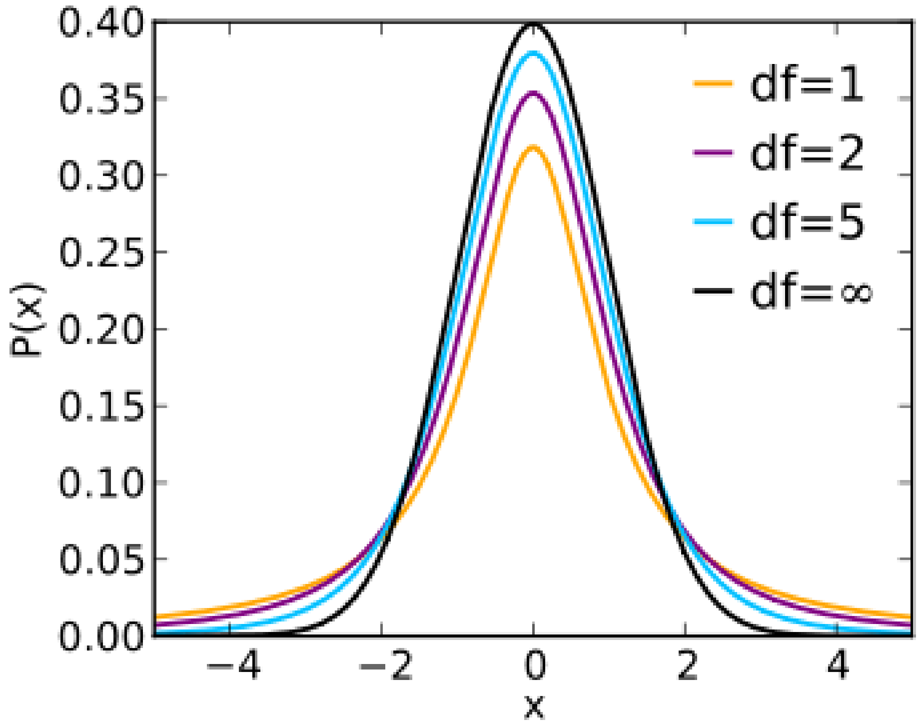
$$\overline{X}_n \sim N(\mu, \sigma^2 / n)$$

- Taking advantage of this presumes we know σ^2

- $\dfrac{\bar{x} - \mu}{s_n / \sqrt{n}}$ has a *t* distribution with (n-1) d.o.f.

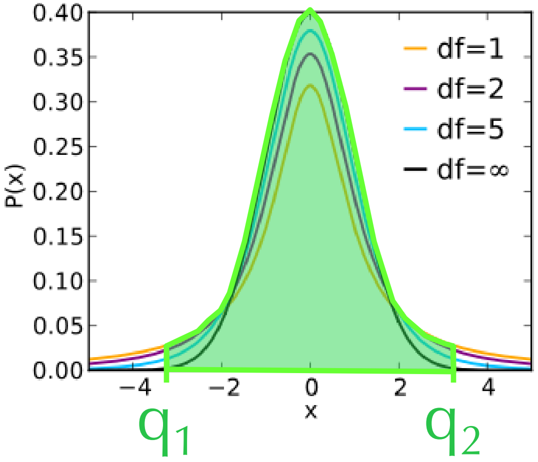
[Student's] t-distribution

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$



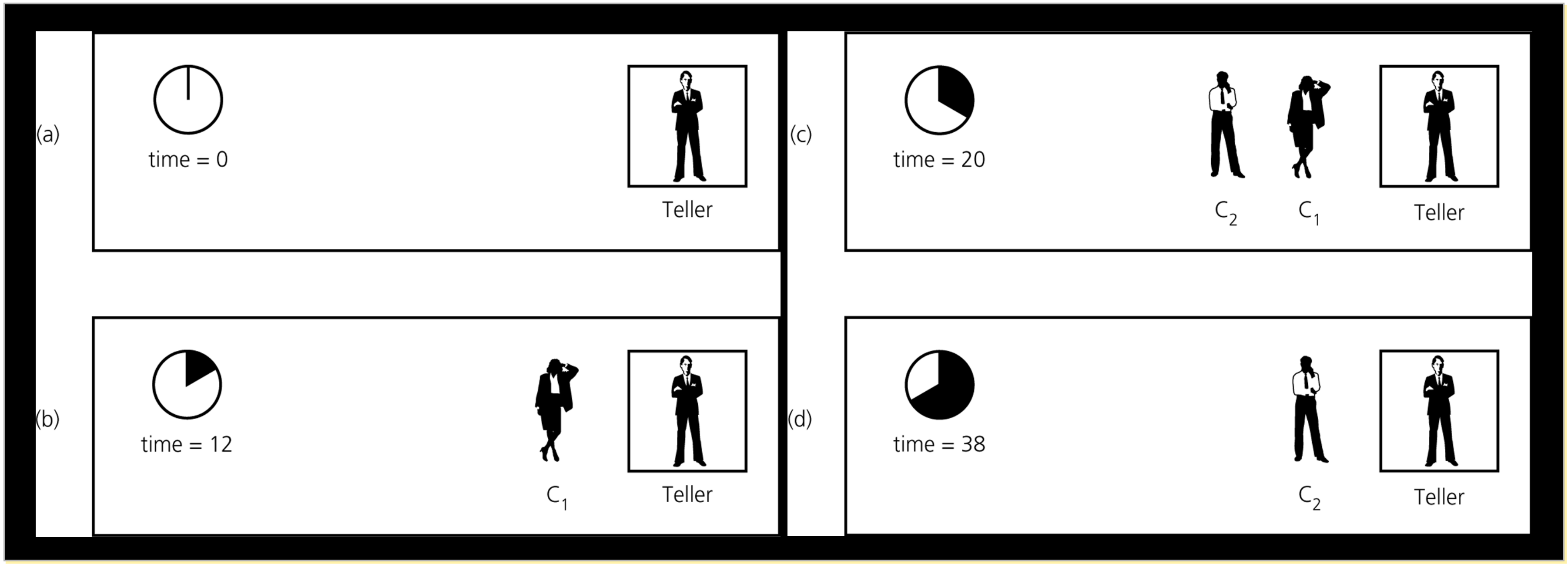
Forming a confidence interval

- e.g., given that I observed a sample mean of ____, I'm 99% confident that the true mean lies between ____ and ____.
- Know that $\dfrac{\bar{x}-\mu}{s_n/\sqrt{n}}$ has t distribution
- Choose q_1, q_2 such that student t with (n-1) dof has 99% probability of lying between q_1, q_2



Interpreting Simulation Outcomes

- How long will customers have to wait, on average?
  - e.g., for given # tellers, arrival rate, service time distribution, etc.



Interpreting Simulation Outcomes

- Simulate bank for N customers
- Let x_i be the wait time of customer i
- Is mean(x) a good estimate for μ?
- How to compute a 95% confidence interval for μ?
  - Problem: x_i are not independent!

Replications

- Run simulation to get M observations
- Repeat simulation N times (different random numbers each time)
- Treat the sample mean of different runs as approximately uncorrelated

$$\bar{X}_i = \frac{1}{m}\sum_{j=1}^{n} x_{ij} \qquad \bar{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} \bar{X}_i$$

$$s^2 = \frac{1}{n-1}\sum_i (\bar{X}_i - \bar{\bar{X}})^2$$