BIOF2014: Statistical Modelling in Bioinformatics

David J. H. Shih

Intended learning outcomes

- Describe the motivations for establishing and applying statistical models

How is statistical modeling useful?

- We want to model the real world
- Many quantities of interest have uncertainty
- Deterministic models can lead to unrealistic predictions
- Stochastic models account for randomness in real-world phenomenons

A world with absolute certainty: Newtonian physics

All variables are measurable to arbitrary precision.

Newton's laws of motion

$$F = 0 \;\Rightarrow\; \frac{dv}{dt} = 0$$

$$\frac{dv}{dt} m = F$$

$$F_{1\to 2} = F_{2\to 1}$$

A world with uncertainty: quantum mechanics

At the atomic scale, some quantum variables cannot be measured precisely at the same time.

Heisenberg's Uncertainty Principle

For uncertainty in momentum Δp and uncertainty in position Δx,

$$\Delta p \, \Delta x \geq \frac{h}{4\pi},$$

where h is Planck's constant.

A simple deterministic model

Let $x \in (0, 1)$ be a quantity of interest (e.g. position of a particle).

Define a function f as

$$f(x) \triangleq rx(1 - x) - x$$

We model the dynamics of x under continuous time $t \geq 0$ by

$$\frac{dx}{dt} = f(x).$$

We simplify the model to discrete time $t \in \{0, 1, 2, ...\}$ by

$$\frac{\Delta x}{\Delta t} = f(x).$$

We can simplify further by setting $\Delta t = 1$, giving

$$x_{t+1} = g(x_t) \triangleq x_t + f(x_t).$$

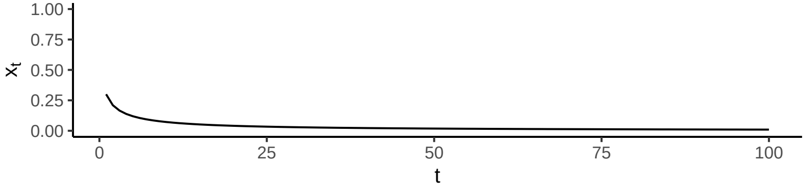
Also,

$$x_\infty \triangleq \lim_{t \to \infty} x_t$$

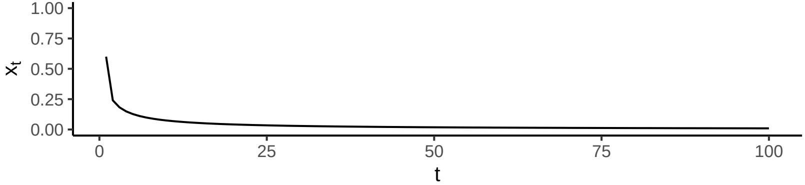
Simulation under deterministic model

```r
library(ggplot2)

# x \in (0, 1)
logistic_map <- function(x, r) {
  r * x * (1 - x)
}

trace_plot <- function(g, x0, T=100) {
  xs <- numeric(T);
  xs[1] <- x0;
  for (i in 2:T) {
    xs[i] <- g(xs[i-1]);
  }
  qplot(1:T, xs, geom="line") + theme_classic() +
    xlab("t") + ylab(expression(x[t])) + ylim(0, 1)
}
```

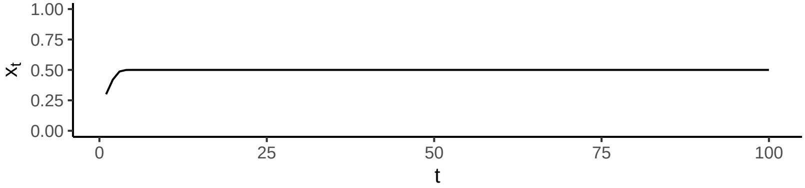
```r
g <- function(x) logistic_map(x, r=1.0);
trace_plot(g, x0=0.3)
```



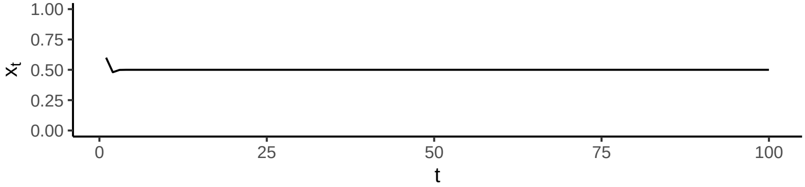

```r
trace_plot(g, x0=0.6)
```



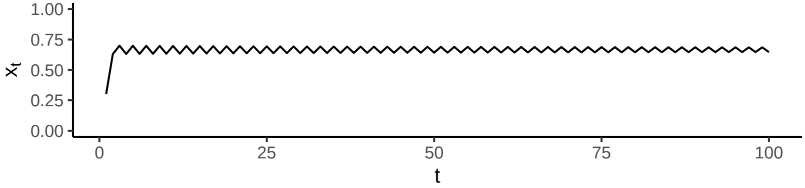
```r
g <- function(x) logistic_map(x, r=2.0);
trace_plot(g, x0=0.3)
```



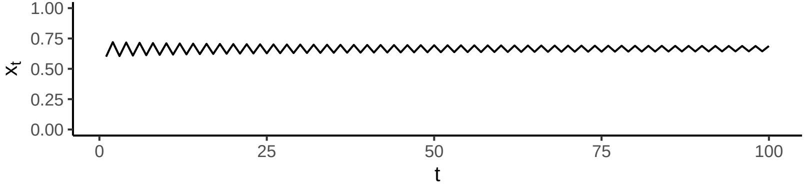

```r
trace_plot(g, x0=0.6)
```



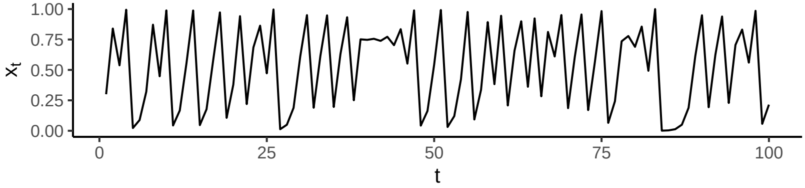
```r
g <- function(x) logistic_map(x, r=3.0);
trace_plot(g, x0=0.3)
```



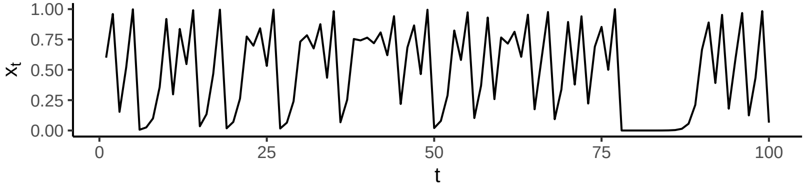

```r
trace_plot(g, x0=0.6)
```



```r
g <- function(x) logistic_map(x, r=4.0);
trace_plot(g, x0=0.3)
```



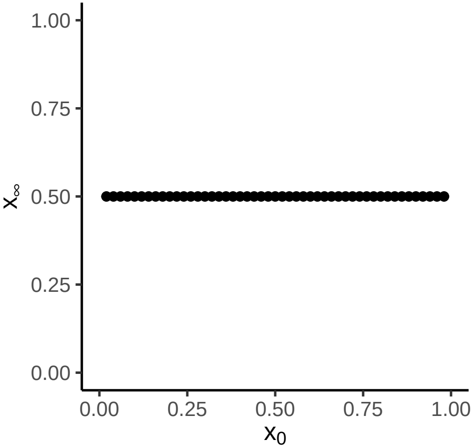

```r
trace_plot(g, x0=0.6)
```



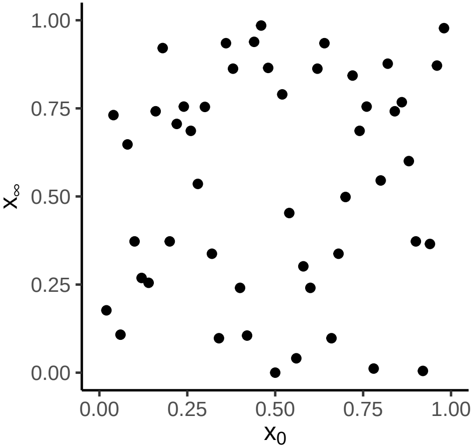
```r
iterate <- function(g, x0, T=100) {
  x <- x0;
  for (i in 1:T) {
    x <- g(x);
  }
  x
}

simulate <- function(g, x0) {
  data.frame(
    x.0 = x0,
    x.inf = vapply(x0,
      function(x0) iterate(g, x0), 0
    )
  )
}
```

```r
predict_plot <- function(g, x0) {
  d <- simulate(g, x0);
  qplot(d$x.0, d$x.inf, geom="point") +
    theme_classic() +
    xlab(expression(x[0])) +
    ylab(expression(x[infinity])) +
    xlim(0, 1) + ylim(0, 1) + coord_fixed()
}

step <- 0.02;
x0 <- seq(0 + step, 1 - step, by=step);
```

```r
g <- function(x) logistic_map(x, r=2.0);
predict_plot(g, x0)
```



```r
g <- function(x) logistic_map(x, r=4.0);
predict_plot(g, x0)
```



Chaos

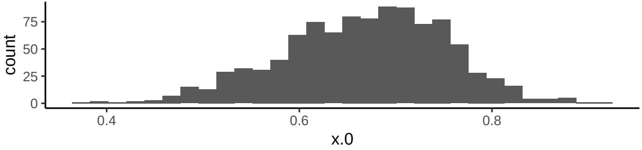
Depending on setting of the parameters, dynamical systems can converge or exhibit

- periodic behaviour
- *chaotic* behaviour

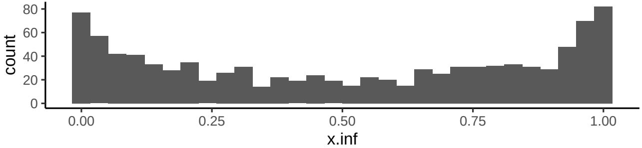
In the field of differential equations, periodic and chaotic systems are *pathological*.

A simple stochastic model

```
set.seed(1337); N <- 1000;
g <- function(x) logistic_map(x, r=4.0);
x.0 <- rbeta(N, 20, 10); x.inf <- iterate(g, x.0);
qplot(x.0, ylab="count") + theme_classic()
```




```
qplot(x.inf, ylab="count") + theme_classic()
```



Statistics

Study of uncertainty

- establish statistical models
- estimate parameters of the models
- quantify uncertainty in parameter estimates
- predict new quantities with uncertainty

Equations

Deterministic

$$y = mx + b,$$

where x and y are variables; m and b are constants.

Probabilistic

$$p\left(Y = y \mid \mu, \tau\right) = (2\pi)^{-\frac{1}{2}} \tau^{\frac{1}{2}} \, \exp\left(-\frac{1}{2}\tau(y-\mu)^2\right)$$

where Y is a *random* variable and y its *realized value*; μ and τ are parameters.

Atomic models

Bohr model

Electrons travel in different *circular* orbitals around a nucleus.

Quantum mechanical model

Electron are distributed in different *cloud* orbitals around a nucleus.

Applications of statistical models

- Demography
- Gambling
- Quantum mechanics
- Biology
- Epidemiology
- Medicine
- Psychology
- Genetics
- Machine learning
- Data science
- Bioinformatics
- ...

Analogy

Newton physics: equations of particles
Quantum mechanics: equations of wavefunctions

Algebra: equations of variables
Statistics: equations of probability distributions

Summary

Albert Einstein: “God does not play dice with the universe.”

Niels Bohr: “Einstein, stop telling God what to do.”

Intended learning outcomes

- Describe the motivations for establishing and applying statistical models