INTRODUCTORY APPLIED MACHINE LEARNING

Yan-Fu Kuo
Dept. of Biomechatronics Engineering
National Taiwan University

Today:

- Review of probability
- Review of inferential statistics

Outline

- Goal of the lecture
- Univariate statistics
- Distributions and significance
- Multivariate statistics

Goals

- After this, you should be able to:
  - Be familiar with probability terminologies
  - Understand basic random variable operation
  - Conduct univariate and multivariate statistical analysis
  - Perform hypothesis test

Univariate Statistics

- Univariate means a single variable
- For example, the height, weight, and test score, of a population

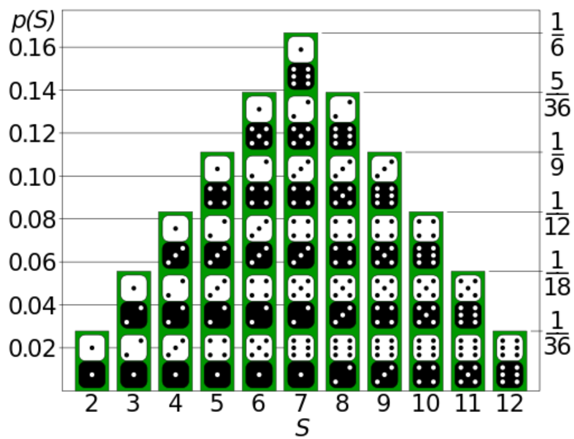
Random Variable

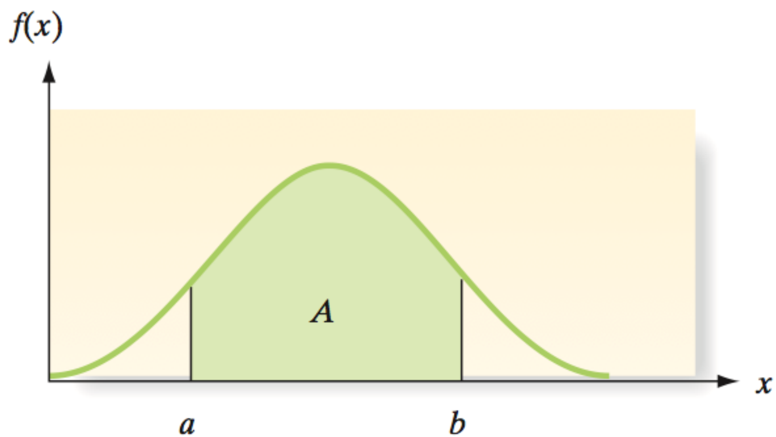
- A variable whose value is subject to variations due to chance (i.e., randomness, in a mathematical sense)
- Also called stochastic variable
- Two types of random variable: discrete and continuous
- Example: die roll

https://en.wikipedia.org/wiki/Random_variable

Probability Density Function (PDF)

- The area under a probability distribution
- The graphical form:
  - Histogram for a discrete random variable
  - Smooth curve for continuous random variable





https://en.wikipedia.org/wiki/Random_variable

Measures of Random Variables

- Expected value μ (mean of probability distribution)

$$\mu = E[x] = \int x\, p(x)$$

- Variance σ^2
  - A measure of how far a set of data is spread out
  - Defined as the expected value of $(x-\mu)^2$, i.e.,

$$\sigma^2 = \mathrm{Var}(x) = E[(x-\mu)^2] = \int (x-\mu)^2\, p(x)$$

- Standard deviation σ
  - Square root of the variance

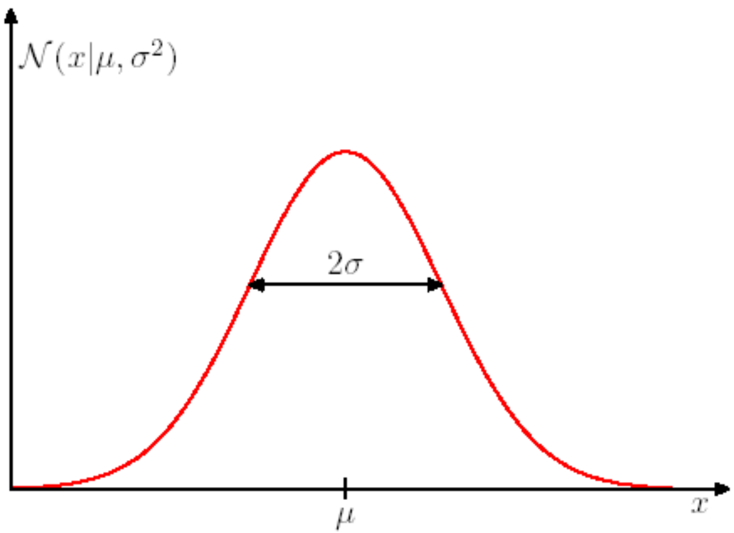
Normal Distribution

- Also called Gaussian distribution

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

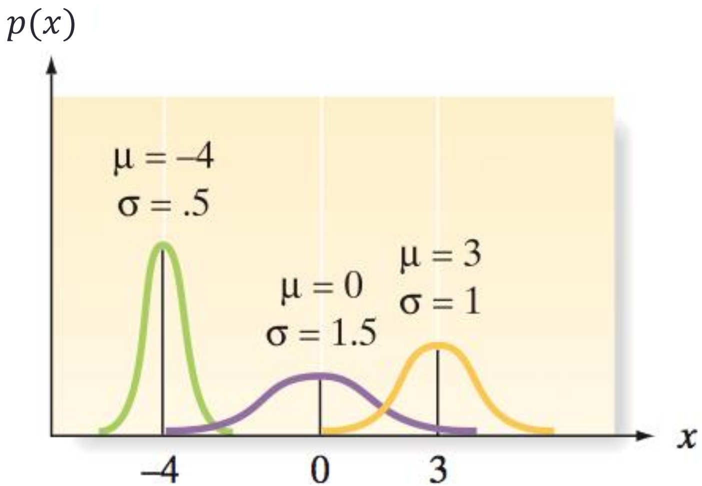
where μ is the mean and σ^2 is the variance

- Probability density function (PDF) of normal distribution



Effect of Varying Parameters μ & σ

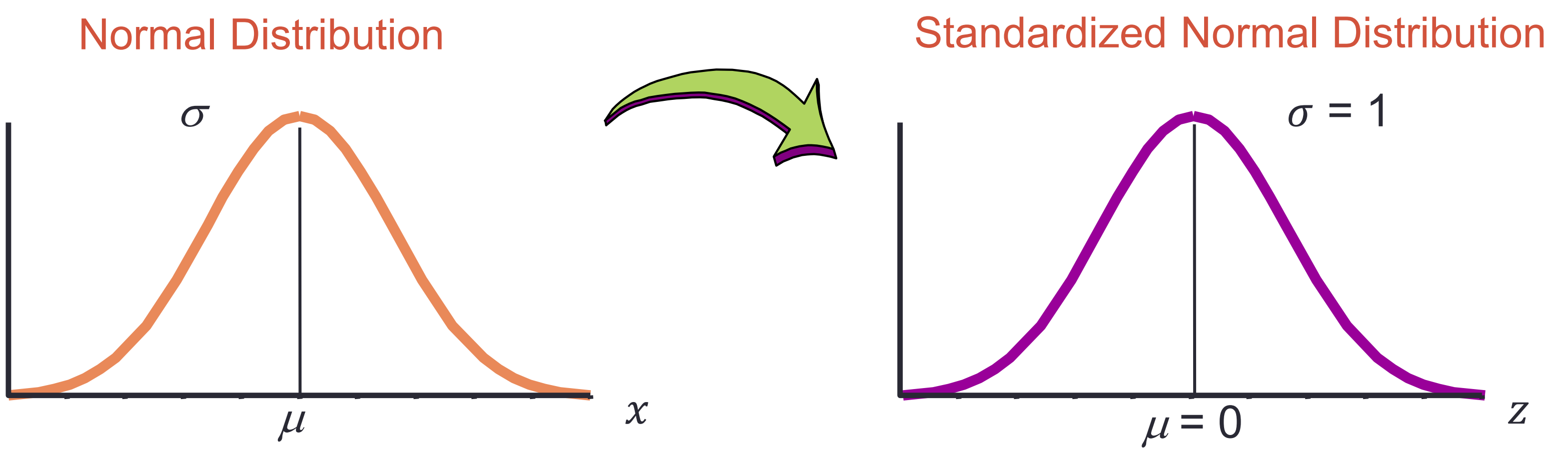
- Normal distributions differ by mean and standard deviation



- This was a problem back to the time when there was no computer – properties of normal distribution was pre-calculated and printed on table

Standard Normal Distribution

- A normal distribution with μ = 0 and σ = 1
- A random variable with a standard normal distribution is usually denoted by the symbol z
- Standardization can be performed by the formula $z = \frac{x-\mu}{\sigma}$



Expected Value of New Random Variable

- Suppose there are two random variables w and x, and their relationship is $w = ax + b$
- Knowing that the expected value of x is μ_x, what is the expected value of w?

$$\mu_w = E[w] = E[ax+b] = \int (ax+b)p(x)$$

$$= \int axp(x) + \int bp(x) = a\int xp(x) + b\int p(x)$$

$$= aE[x] + b = a\mu_x + b$$

Variance of New Random Variable

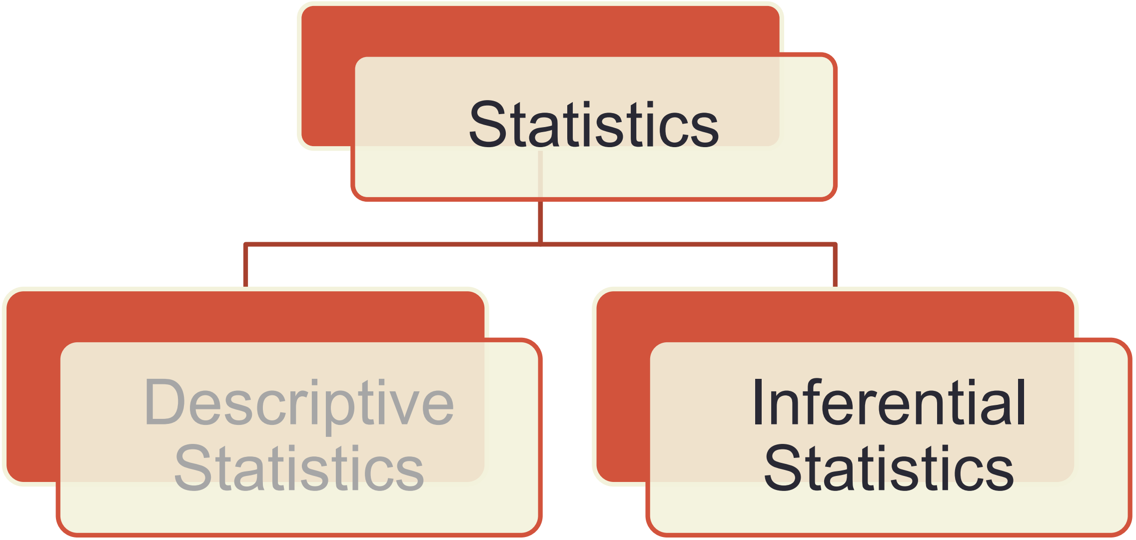
- Suppose the relationship between two random variables, w and x, is $w = ax + b$
- Knowing that the expected value of x is μ_x, what is the expected value of the variance of w?

$$\sigma_w^2 = \int (ax + b - (a\mu_x + b))^2 p(x)$$

$$= \int (a(x - \mu_x))^2 p(x) = \int a^2 (x - \mu_x)^2 p(x)$$

$$= a^2 \int (x - \mu_x)^2 p(x) = a^2 \sigma_x^2$$

Statistical Methodologies



Methods to make estimates, decisions, and predictions using sample data

Statistical Hypothesis Testing

- A method of making decisions using data
- Example: The mean of a population is equal to θ_0?
- Typical hypothesis:
  - $H_0\colon \theta = \theta_0 \quad v.s. \quad H_1: \theta \neq \theta_0$
  - $H_0\colon \theta \geq \theta_0 \quad v.s. \quad H_1: \theta < \theta_0$
  - $H_0\colon \theta \leq \theta_0 \quad v.s. \quad H_1: \theta > \theta_0$

  where H_0 is null hypothesis, and H_1 is alternative hypothesis

Decision Rules and Terminology

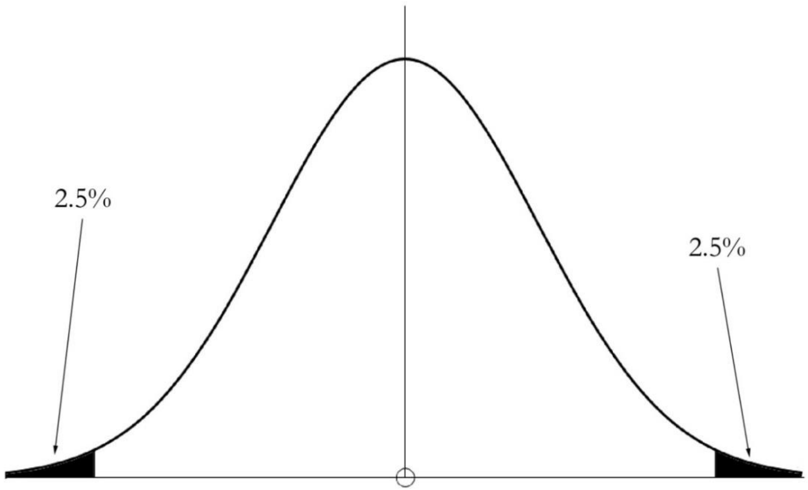
- The hypothesis testing checks if samples randomly from the population are consistent with the statistics or not
- Base upon the sample statistic, one can
  1. Either reject null hypothesis H_0 and conclude that alternative hypothesis is substantiated
  2. Or retain null hypothesis H_0 and conclude that alternative hypothesis fails to be substantiated

Hypothesis Testing Procedure

1. Determine a probability, say 0.95, for the hypothesis test
2. Find the 95% “confidence Interval” of the H_0
3. Check if your score falls into the interval

Terminology in Hypothesis Testing

- Determine a probability, say 0.95, for the hypothesis test
- Find the 95% "confidence Interval" of the H_0
- Check if your score falls into the interval
- Terminology:
  - Confidence interval
  - Confidence level $(1-\alpha)$
  - Significance level α
  - p-value



Distinguishing 2 Populations

Normals

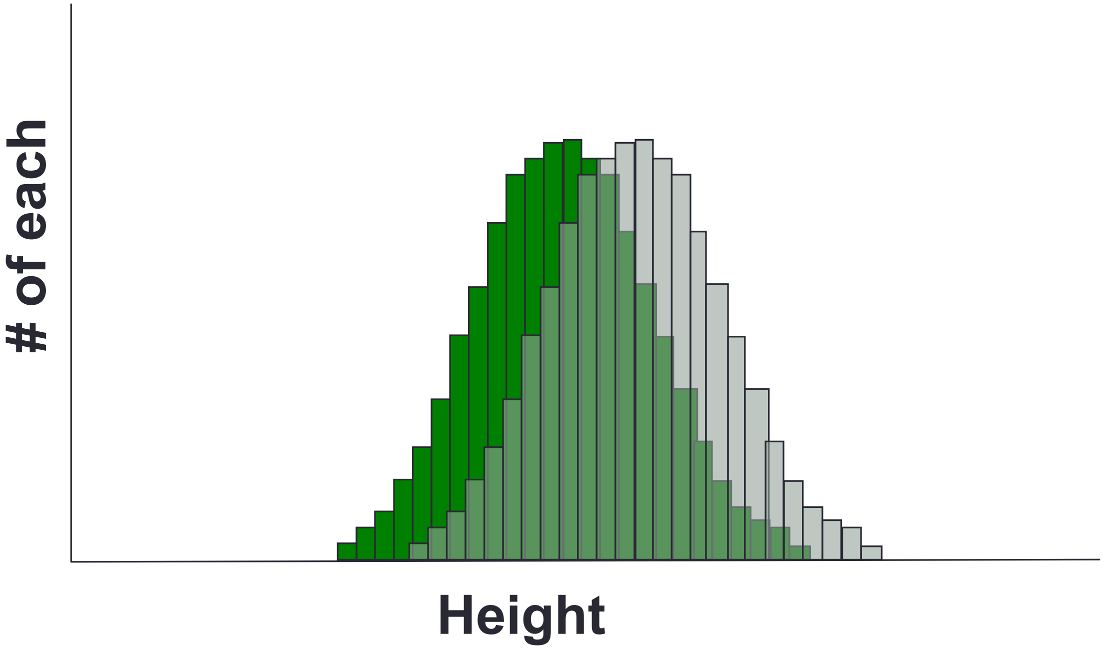
Dwarfs

The Result

Are they different?

What about these 2 Populations?

The Result



Are they different?

Student's t-test

- Formally allows you to calculate the probability that 2 sample means are the same
- If the t-Test statistic gives you a

  p = 0.4, and the α = 0.05, the mean of the 2 populations are <u>the same</u>
- If the t-Test statistic gives you a

  p = 0.04, and the α = 0.05, the mean of the 2 populations are <u>different</u>
- Paired and unpaired t-Tests are available

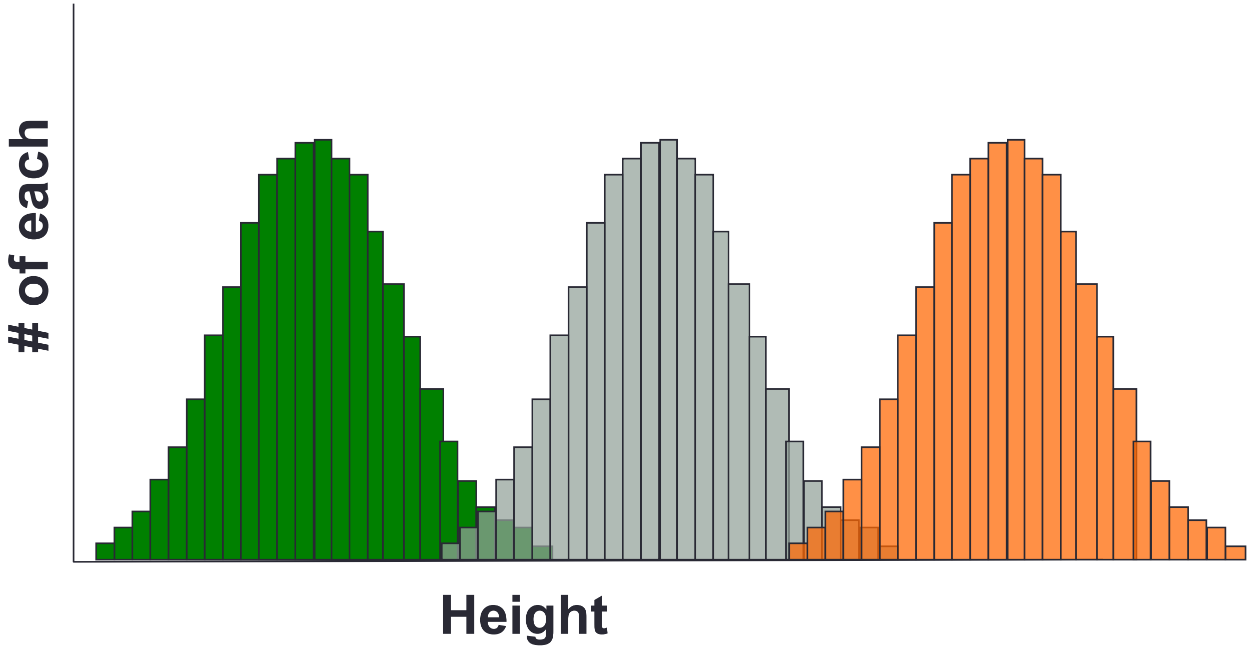
Distinguishing 3+ Populations

Normals

Dwarfs

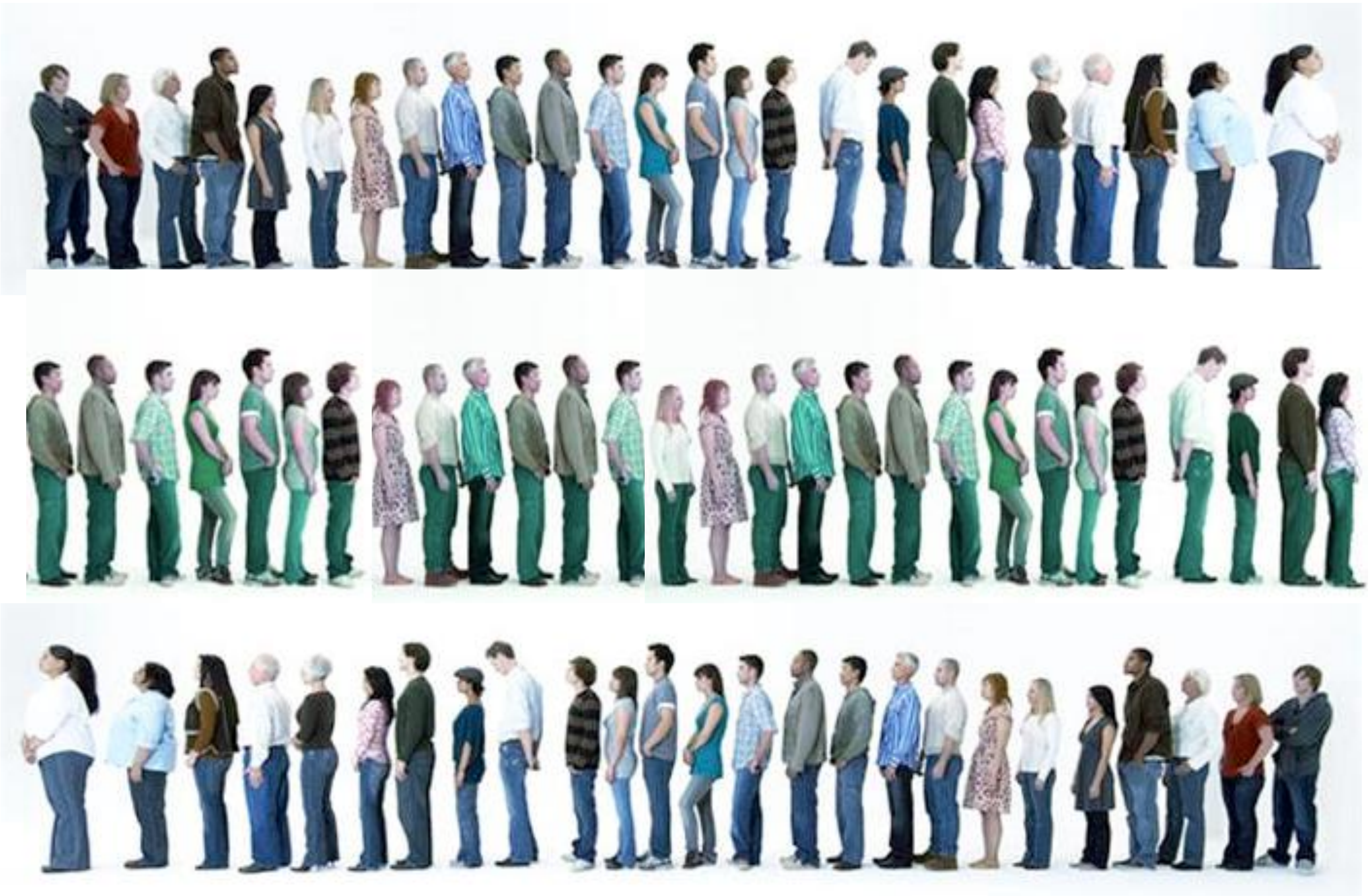
Elves

The Result

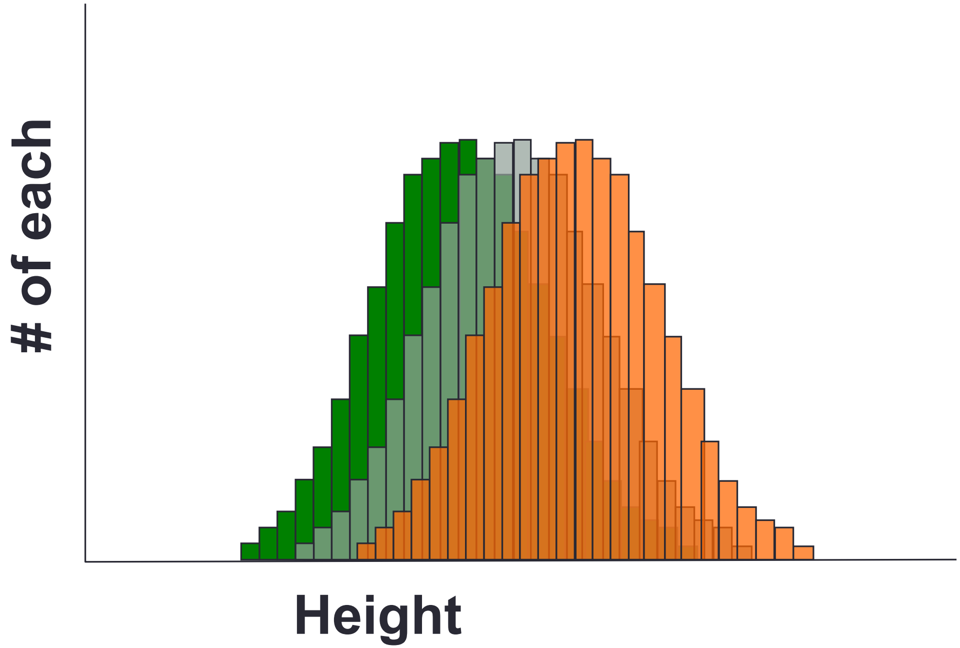


Are they different?

Distinguishing 3+ Populations



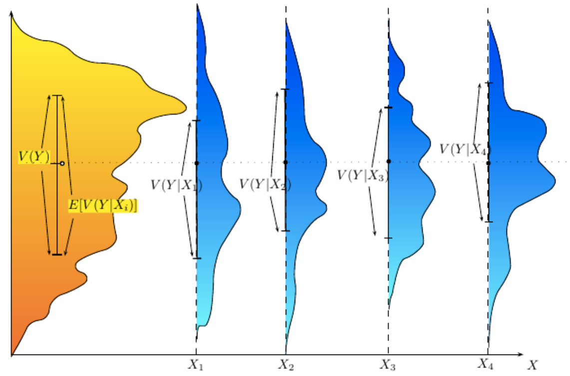
The Result

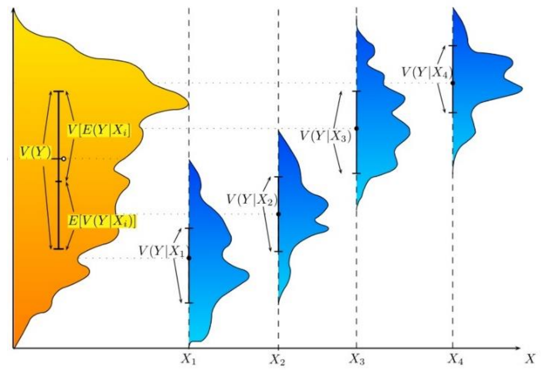


Are they different?

ANOVA

- Analysis of variance $F = \dfrac{Variance\ between\ groups}{Variance\ within\ groups}$
- Used to determine if the means of 3 or more populations are different





Multivariate Statistics

- Multivariate means multiple variables
- If you measure a population using multiple measures at the same time such as height, weight, hair color, etc., you are performing multivariate statistics
- Multivariate statistics requires more complex, multidimensional analyses or dimensional reduction methods

Bivariate Gaussian

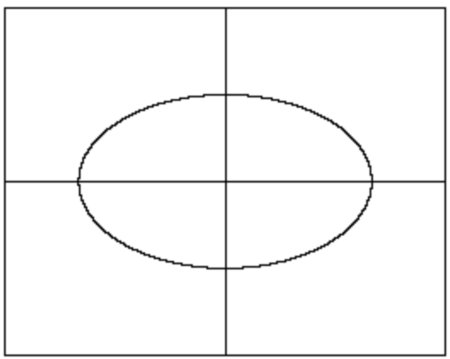
- Let $x_1 \sim N(\mu_1, {\sigma_1}^2)$ and $x_2 \sim N(\mu_2, {\sigma_2}^2)$
- Suppose x_1 and x_2 are independent

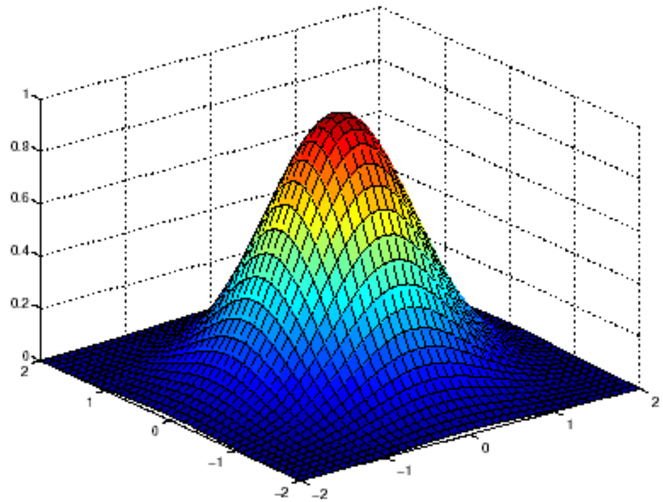
$$p(x_1, x_2) = \frac{1}{2\pi({\sigma_1}^2{\sigma_2}^2)^{1/2}} \exp\left(-\frac{1}{2}\{\frac{(x_1-\mu_1)^2}{{\sigma_1}^2} + \frac{(x_2-\mu_2)^2}{{\sigma_2}^2}\}\right)$$

Let $\boldsymbol{x} = \binom{x_1}{x_2}$, $\boldsymbol{\mu} = \binom{\mu_1}{\mu_2}$, $\boldsymbol{\Sigma} = \begin{bmatrix} {\sigma_1}^2 & 0 \\ 0 & {\sigma_2}^2 \end{bmatrix}$

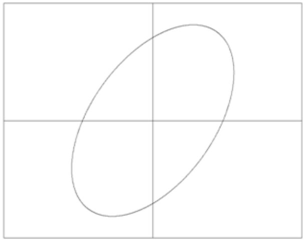
$$p(\boldsymbol{x}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp -\frac{1}{2}\{(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

Bivariate Gaussian





What If There Is Correlation?



- Two random variables might not be independent
- Example: plot of weight vs. height for a population
- Let ρ be the correlation between x_1 and x_2
- Covariance between two random variables:

$$\boldsymbol{\Sigma} = E[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T] = \begin{bmatrix} {\sigma_1}^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & {\sigma_2}^2 \end{bmatrix}$$

- Covariance between two random variables:

$$p(\boldsymbol{x}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}\right)$$

Summary

- In stochastic models, variable states are not described by unique values, but rather by probability distributions
- T-tests and ANOVA are parametric statistical techniques that are widely used to compare group means
- ANOVA is used to test differences in means between more than three groups
- In multivariate statistical analysis, there can exist interaction between variables

References

- D. Montgomery and G. Runger, *Applied Statistics and Probability for Engineers*