Biology 644

Old Title: Bioinformatics for Molecular Biologists

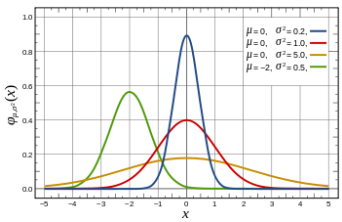
Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Outline

- Binomial Distribution
- Poisson Distribution
- Normal Distribution
- R Lab
  - Problem Set 1
  - Chapter 3
  - Chapter 3 Supplemental
  - Problem Set 2

Probability Distributions

- We use probability distributions that correctly model real world phenomena over and over again



Random Variable

- The mathematical rule (or function) that assigns a given numerical value to each possible outcome of an experiment in the sample space of interest
- Discrete random variables
  - Bernoulli Random Variables
  - Poisson Random Variables
- Continuous random variables
  - Normal Random Variables

Bernoulli Random Variables

- A trial is repeated n times
  - A coin is tossed 5 times
  - A die is rolled 25 times
  - 50 patients are screened

- Assumes p remains constant from trial to trial and that the trials are statistically independent of each other
- The Binomial Distribution models the probability of obtaining x successes in n trials for all values of x, when the probability of a success is p.

Binomial Distribution

- Example
  - What is the probability of obtaining 2 heads from a coin that was tossed 5 times?
  - $P(HHTTT) = (1/2)^5 = 1/32$

  - But there are more possibilities:

HHTTT	HTHTT	HTTHT	HTTTH
	THHTT	THTHT	THTTH
		TTHHT	TTHTH
			TTTHH

  - $P(2\ heads) = 10 \times 1/32 = 10/32$

Binomial Distribution

- If trials result in a series of success and failures:

  FFSFFFFSFSFSSFFFFFSF…

  Then the probability of x successes in that order is

  $$P(x) = (1-p)\cdot(1-p)\cdot p\cdot(1-p)\cdot \ldots = p^x\cdot(1-p)^{n-x}$$

- However, if order is not important then

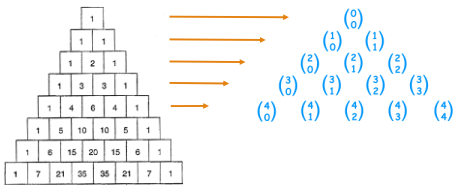
$$P(x) = \left(\frac{n!}{x!\,(n-x)!}\right)\cdot p^x\cdot(1-p)^{n-x}$$

where $\frac{n!}{x!\,(n-x)!}$ is the number of ways to obtain x successes in n trials, and $x! = x\cdot(x-1)\cdot(x-2)\cdot\ldots\cdot 2\cdot 1$

Binomial Coefficients and Pascal's Triangle

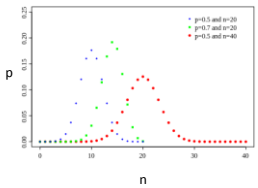
Tosses	Possible Results (Grouped)	Pascal's Triangle
1	H T	1, 1
2	HH HT TH TT	1, 2, 1
3	HHH HHT, HTH, THH HTT, THT, TTH TTT	1, 3, 3, 1
4	HHHH HHHT, HHTH, HTHH, THHH HHTT, HTHT, HTTH, THHT, THTH, TTHH HTTT, THTT, TTHT, TTTH TTTT	1, 4, 6, 4, 1
⋮	⋮	⋮

Binomial Coefficients and Pascal's Triangle

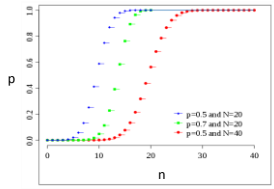


Binomial Distribution

Probability mass function



Cumulative distribution function



The Poisson Distribution

- When there are a large number of trials, but a small probability of success, binomial calculations become impractical and the Poisson approximation can be used
  - Example: Number of deaths from horse kicks in the Prussian Army over a certain time interval
- Uses the mean number of successes from n trials: $\lambda = np$ over a relatively long time frame
  - Example: 64 deaths in 20 years from thousands of soldiers

The Poisson Distribution

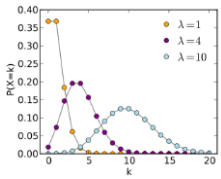
- If we substitute λ/n for p, and let n tend to infinity, the binomial distribution becomes the Poisson distribution:

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

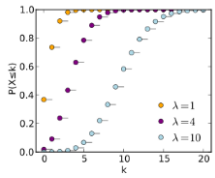
- Poisson distribution is applied where random events in space or time are expected to occur
- Deviation from Poisson distribution may indicate some degree of non-randomness in the events under study
- Investigation of cause may be of interest

The Poisson Distribution

Probability mass function



Cumulative distribution function



Normal Distribution

- Discovered by de Moivre in 1733 as an approximation to the binomial distribution when the number of trails n is large
- Derived by Gauss in 1809
- Importance lies in the Central Limit Theorem, which states that the sum of a large number of independent random variables (binomial, Poisson, etc.) will approximate a normal distribution
- Example: Human height is determined by a large number of factors, both genetic and environmental, which are additive in their effects. Therefore, it follows a normal distribution.

Normal Distribution

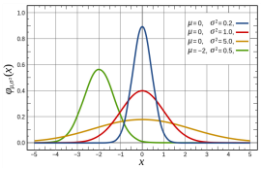
- A continuous random variable is said to be normally distributed with mean μ and variance σ^2 if its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

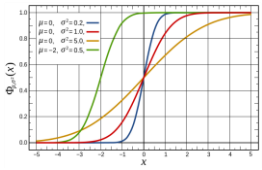
- f(x) is not the same as P(x)
- P(x) would be 0 for any specific x because the normal distribution is continuous
- $P(x_1 < x < x_2) = \int_{x_1}^{x_2} f(x)\, dx$

Normal Distribution

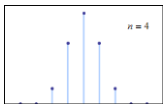
Probability density function



Cumulative distribution function

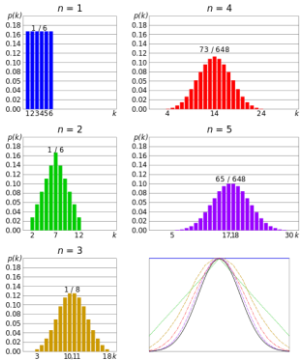


Normal Distribution



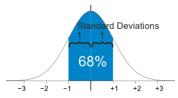
As the number of discrete events increases, the function begins to resemble a normal distribution

Comparison of probability density functions, $p(k)$ for the sum of n fair 6-sided dice to show their convergence to a normal distribution with increasing n

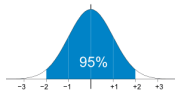


Standard Deviations

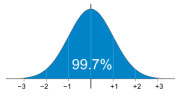
68% of values are within **1 standard deviation** of the mean



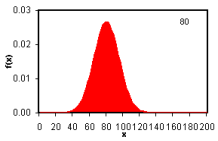
95% of values are within **2 standard deviations** of the mean



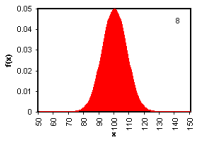
99.7% of values are within **3 standard deviations** of the mean



Normal Distribution

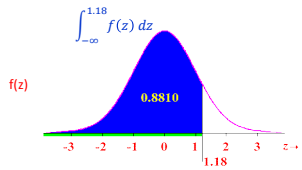


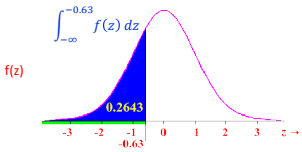
Mean changes



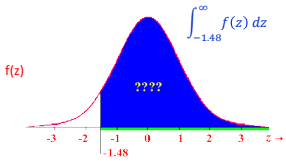
Variance changes

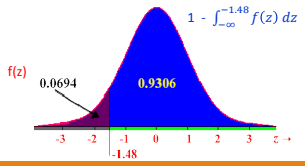
Normal Distribution





Normal Distribution



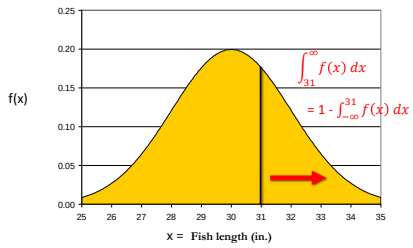


Normal Distribution

- A sample of cod in Boston Harbor suggests that the mean length of these fish is $\mu = 30$ in. and $\sigma^2 = 4$ in.
- Assume that the length of cod is a normal random variable
- If we catch one of these fish in Boston Harbor what is the probability that…
  - it will be at least 31 in. long?
  - it will be no more than 32 in. long?
  - its length will be between 26 and 29 inches?

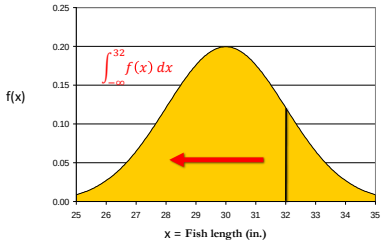
Normal Distribution

What is the probability that it will be at least 31 in. long?



Normal Distribution

…that it will be no more than 32 in. long?

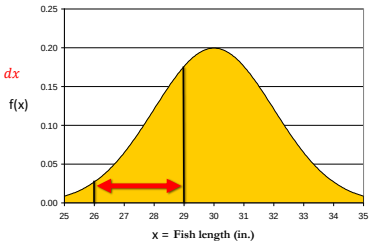


Normal Distribution

...that its length will be between 26 and 29 inches?

$$\int_{26}^{29} f(x)\,dx$$

$$= \int_{-\infty}^{29} f(x)\,dx - \int_{-\infty}^{26} f(x)\,dx$$



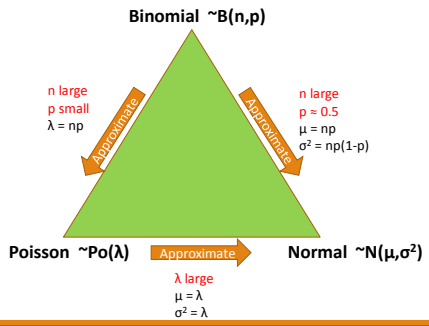
Useful Properties of the Normal Distribution

- 2 normal distributions X and Y can be added
- $E(X+Y) = E(X)+E(Y)$ and $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y)$
- A normal distribution can be transformed with shift and change of scale operations
- Let $X \sim N(\mu,\sigma)$ and let $Y = aX+b$ where a and b area constants
- Change of scale is the operation of multiplying X by a constant "a" whereby one unit of X becomes "a" units of Y.
- A shift is the operation of adding a constant "b" to X whereby we simply move our random variable X "b" units along the x-axis.
- If X is a normal random variable, then the new random variable Y created by these operations on X is also a random normal variable

For $X \sim N(\mu,\sigma)$ and $Y = aX+b$

- $E(Y) = a\mu+b$
- $\sigma^2(Y) = a^2\,\sigma^2$
- A special case of a change of scale and shift operation in which $a = 1/\sigma$ and $b = -1(\mu/\sigma)$
  - $Y = (1/\sigma)X - \mu/\sigma = (X-\mu)/\sigma$
  - $E(Y) = 0$ and $\sigma^2(Y) = 1$
  - Y is now a Standard Normally Distributed random variable!

The Central Limit Theorem

- Standardizing any random variable that itself is a sum or average of a set of independent random variables results in a new random variable that is nearly the same as a standard normal one.
- The only caveats are that the sample size must be large enough and that the observations themselves must be independent and all drawn from a distribution with common expectation and variance.

Approximating