Accelerating Impact: Immersive Summer Bootcamp in Implementation Science and Biostatistics

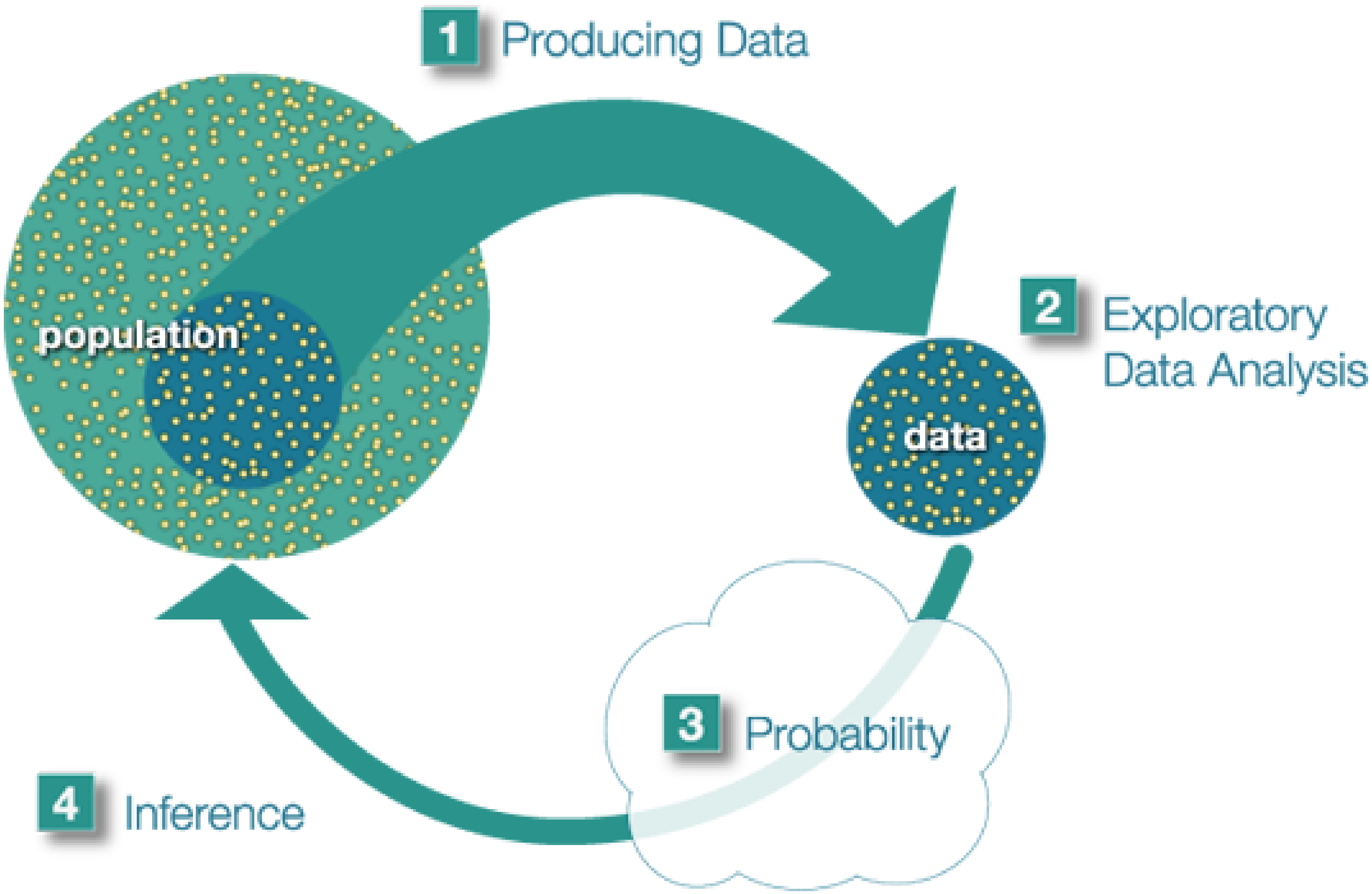
Georgian Implementation Science Fogarty Training (GIFT) Program

Ilia State University & Yale University

Yale Center for Analytical Sciences

Probability

The Big Picture



http://bolt.mph.ufl.edu/6050-6052/preliminaries/the-big-picture/

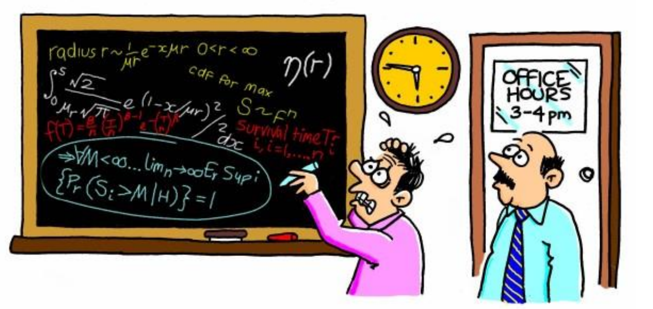
Inferential Statistics

- The process of drawing conclusions about a population based on a sample from that population
  - **Population** – entire group of subjects to which we'd like our conclusions inferred
  - **Sample** – a representative subset of the population

What is a p-value?

Where Statisticians go to get their P-values.

Probability



He watched patiently as his student battled to try and calculate *"a snowball's chance in hell"*.

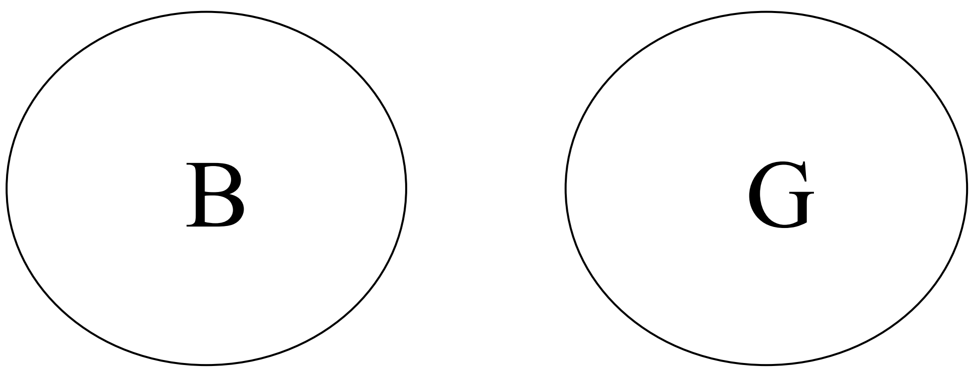
- The proportion of times a given outcome will occur if we repeated an experiment or observation a large number of times
  - Example: We can estimate the probability of a baby being a boy by observing what proportion of a sample of babies are boys
    - 100 babies, 48 boys
      - P(B) = 48/100 = 0.48
  - http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html
  - By definition, probabilities are b/w 0 and 1

Rules of Probability

- Addition Rule – for a given event, for any 2 or more outcomes that may happen the probability of either occurring is the sum of the individual probabilities
  - $P(B \text{ or } G) = P(B) + P(G)$
  - The sum of the probabilities for all possible outcomes must equal 1.

Mutually Exclusive

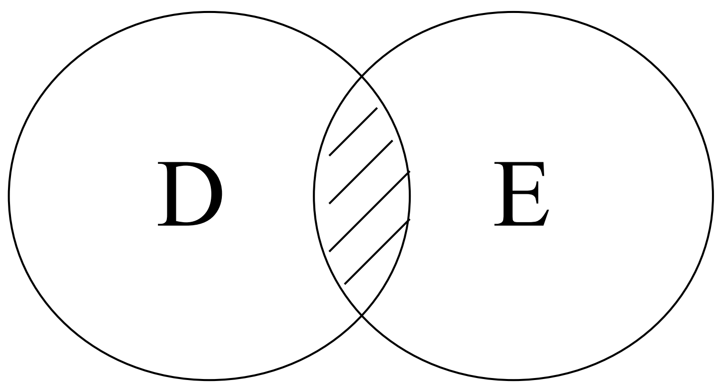
- The addition rule is valid only when possible outcomes are mutually exclusive – that is, the occurrence of one event precludes the occurrence of another



- P(B and G) = 0

Non Mutually Exclusive

- When the occurrence of one event does not preclude the occurrence of another



P(D and E) $\neq$ 0

- Modified Addition Rule
  - P(D or E) = P(D) + P(E) – P(D and E)

Rules of Probability

- Subtraction Rule – the probability of an event occurring is 1 minus the probability of it not occurring.
    - $P(B) = 1 - P(\text{not } B)$

Independence

- The outcome of one event has no effect on the outcome of other events
  - Example: Flip a coin twice, what you get on the 1^{st} flip doesn't influence what you get on the 2^{nd} flip

Rules of Probability

- Multiplication Rule – the probability of 2 or more independent events occurring is the product of their individual probabilities

- Example
  - Event 1: Disease – P(D) = 0.10
  - Event 2: Gender – P(M) = 0.50
  - Event 3: P(M and D) = P(M) x P(D)

Joint Prob Marginal Probs

Non-Independence

- The outcome of one event depends on the outcome of another event
  - Example: The probability of breast cancer depends on gender.
- Modified multiplication rule
  - $P(D \text{ and } E) = P(D|E) \times P(E) = P(E|D) \times P(D)$

Conditional Prob

	Disease	No Disease	
Exposure	5	20	25
No Exposure	3	72	75
	8	92	100

$$P(D \mid E) = \frac{5}{25} = 0.20 \qquad P(D \mid E) \times P(E) = 0.20 \times 0.25 = 0.05$$

$$P(E) = \frac{25}{100} = 0.25$$

	Disease	No Disease	
Test Positive	5	20	25
Test Negative	3	72	75
	8	92	100

$$P(T+ \mid D) = \frac{5}{8} = 0.625$$ Sensitivity

$$P(D) = \frac{8}{100} = 0.08$$ Prevalence

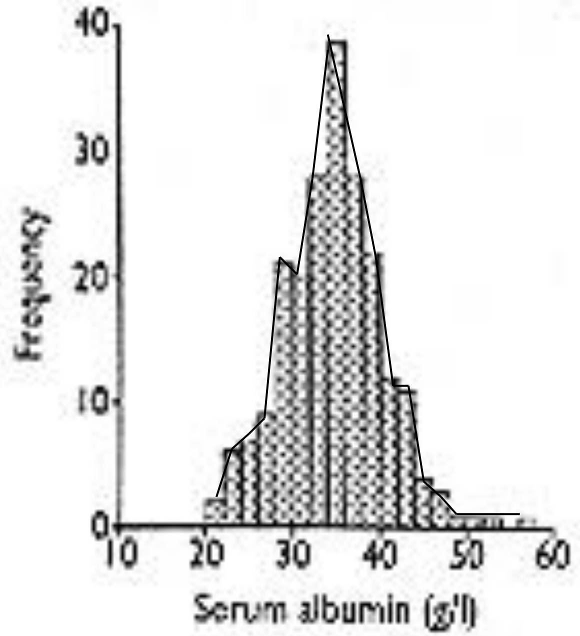
$$P(T- \mid D-) = \frac{72}{92} = 0.78$$ Specificity

Probability Distributions

- Empirical – based on actual observed data
- Theoretical – specified by a mathematical function
  - Can be used to calculate the theoretical probability of observing different values
  - Parametric statistics

Normal Distribution

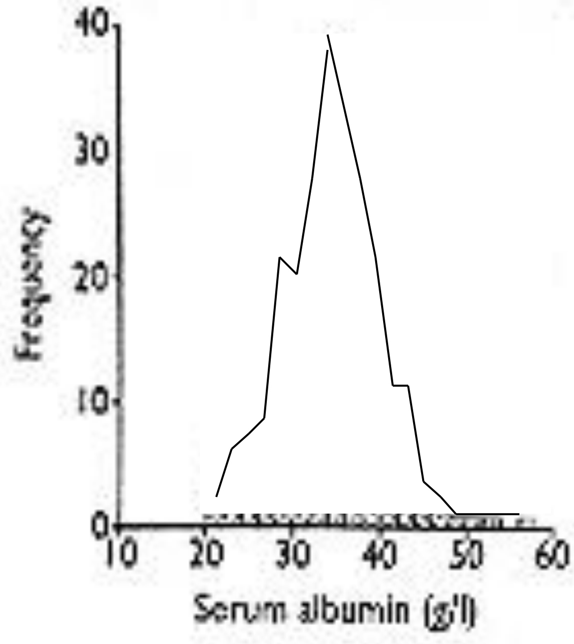
FIG 1 (left)--Serum albumin values in 248 adults FIG 2 (right)--Normal distribution with the same mean and standard deviation as the serum albumin values

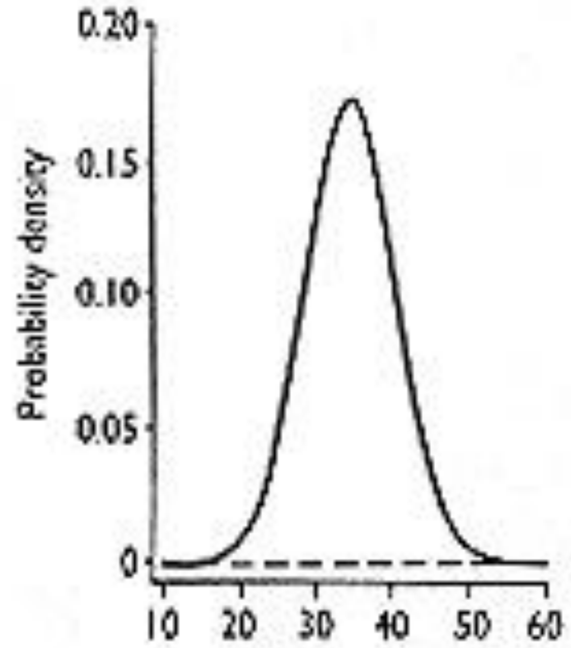


Altman, D. G et al. BMJ 1995;310:298

Normal Distribution

FIG 1 (left)--Serum albumin values in 248 adults FIG 2 (right)--Normal distribution with the same mean and standard deviation as the serum albumin values





Altman, D. G et al. BMJ 1995;310:298

Normal Distribution

- Continuous distribution
- Smooth, bell-shaped and symmetric about the mean
- Area under the curve = 1
- Mathematical Function

mean

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\times\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Standard deviation

Standard Normal Distribution

- Mean=0, SD=1
- http://davidmlane.com/hyperstat/z_table.html

- To use either the website need to calculate ‘z’

$$z = \frac{x - \mu}{\sigma}$$

Standard Normal Distribution Example

- Assume fasting plasma glucose is normally distributed with a mean of 92 and a SD of 9 mg/dl.
  - Find the probability that a patient will have a FPG of 110 or greater.

$$z = \frac{110 - 92}{9} = 2.0$$

$$P(> z) = 0.023$$

Guidelines for Normal Distribution

- Mean ± 1 SD –contains 66.7% of AUC
- Mean ± 2 SD – contains 95% of AUC
- Mean ± 3 SD – contains 99.7% of AUC

Binomial Distribution

- Discrete distribution
- For an event that has two outcomes
- Describes the number of successes (X) observed in n independent trials, each with the same probability of occurrence
  - Example: Flip a coin 4 times what's the probability of observing 1H, 2H, 3H or 4H

Binomial Distribution

$$P(x) = \frac{n!}{x!\,(n-x)!}\,\pi^{x}\,(1-\pi)^{n-x}$$

$$\uparrow$$

# of combinations

$$n! = n \times (n-1) \times (n-2) \times \ldots 1$$

Binomial Distribution Blood Type Example

- P(B) = 0.08 ; P(O or A or AB) = 0.92

- We observe 2 unrelated people.
  - What's the probability that both will be type B?

$$0.08 \text{ x } 0.08 = 0.0064$$

  - What's the probability that neither will be type B?

$$0.92 \text{ x } 0.92 = 0.8464$$

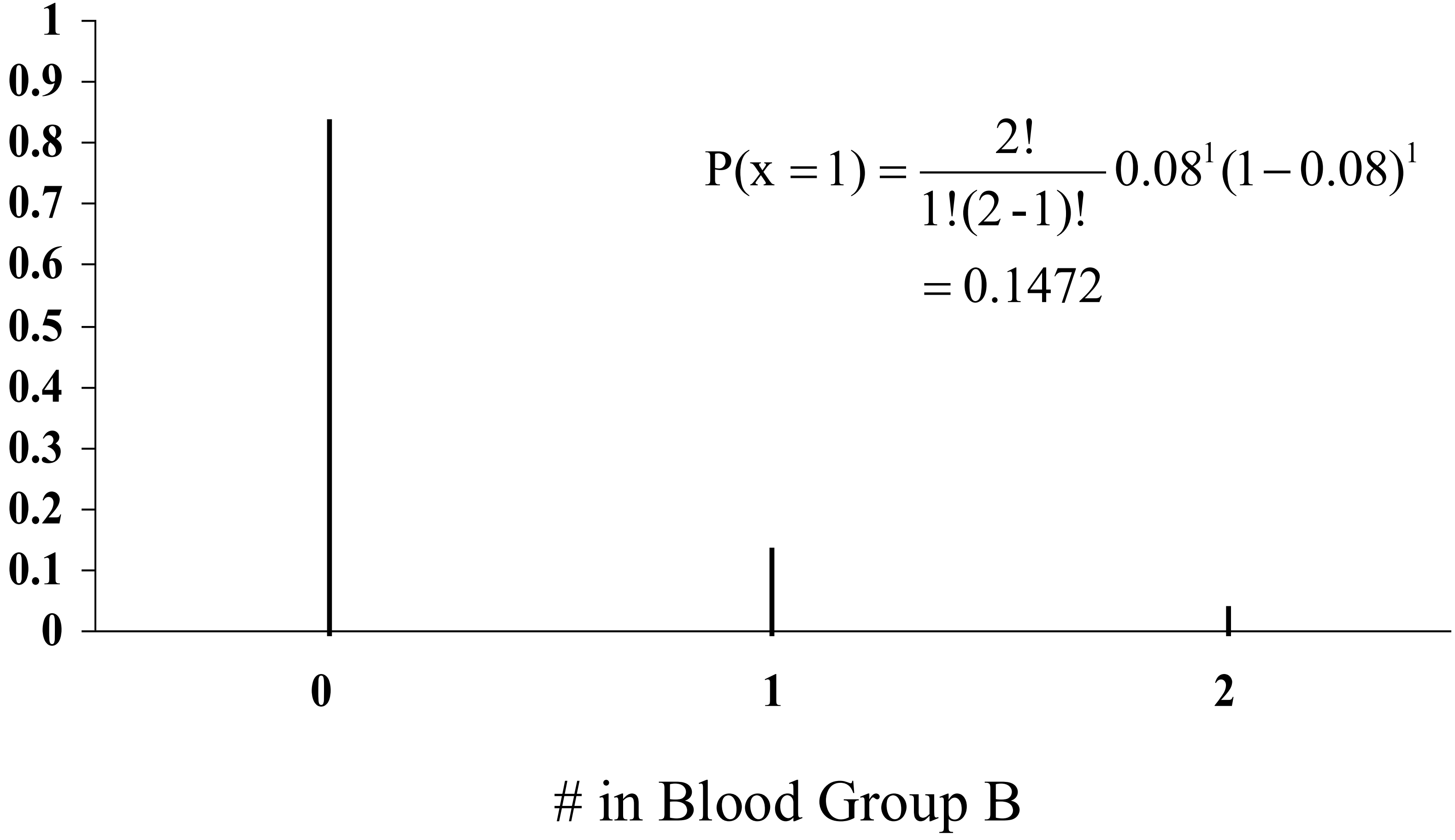
  - What's the probability that one of the two will be type B?

$$2 \text{ x } 0.08 \text{ x } 0.92 = 0.1472$$

Binomial Distribution Blood Type Example

Sample		# in BG B	Prob
B	B	2	0.08 x 0.08 = 0.0064
Not B	B	1	0.92 x 0.08 = 0.0736
B	Not B	1	0.08 x 0.92 = 0.0736
Not B	Not B	0	0.92 x 0.92 = 0.8464

Binomial Distribution Blood Type Example



$$P(x=1) = \frac{2!}{1!(2-1)!} 0.08^1 (1-0.08)^1$$

$$= 0.1472$$

Binomial Distribution

- When n becomes large, the binomial distribution looks more like the normal distribution

- Mean $n\pi$

- Standard Deviation $\sqrt{n\pi(1-\pi)}$

Poisson Distribution

- Discrete
- Gives the probability an outcome occurs a certain number of times when the number of trials is large and the probability of occurrence is low (i.e. rare events)
  - Example – Daily number of new cases of breast cancer reported to a registry or the number of abnormal cells in a fixed area of a slide from a series of liver biopsies

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

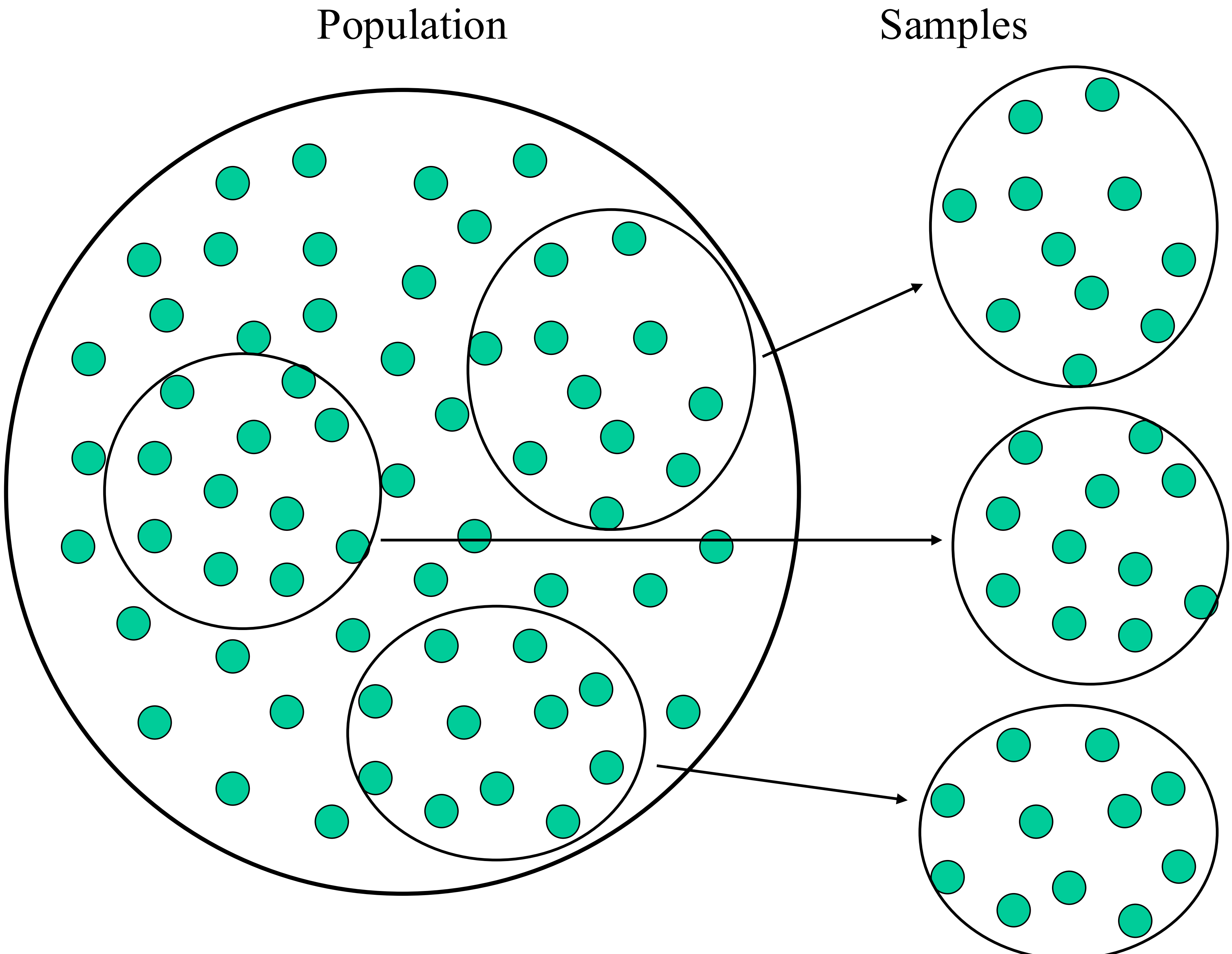
Poisson Distribution Example

- If the mean number of breast cancer cases reported daily to a tumor registry are 2.2, what is the probability of observing 7 on a given day?

$$P(7) = \frac{2.2^7 e^{-2.2}}{7!}$$
$$= 0.0055$$

Sampling

- Why sample?
  - Quicker
  - Cheaper
  - More accurate – can use better methods of measurement
  - Impossible to get whole population
- In practice we only take one sample (i.e. conduct the experiment once). Consequently, much effort should be taken to assure that this sample is representative of the population.
- We use statistics to tell us how likely our population would have produced our sample given certain assumptions.


Population
Samples

Sampling Distributions

- http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Central Limit Theorem

- Central Limit Theorem

As the sample size increases the sampling distribution of the mean becomes closer and closer to a normal distribution regardless of the underlying population distribution

If the population distribution is normal, the sampling distribution of the mean will be normal regardless of the sample size

Central Limit Theorem

- Mean of means = population mean μ

- Standard deviation of the sampling distribution of the mean

$$\frac{\sigma}{\sqrt{n}}$$

Standard Error of the Mean

(SEM)

SD vs SEM

- SD – describes variation of individual observations within a sample or population
- SEM – describes the variation of means from sample to sample