



Review of Biostatistics

Montreal Spring School of Population
Genomics and
Genetic Epidemiology

Nathalie Malo, Ph.D.

May 30th, 2011



Learning Objectives

1. Sampling and exploratory data analysis
2. Concepts of probability
 - Rules of probability
 - Union and intersection
 - Conditional probability and Independence
 - Bayes' Theorem
3. Common probability distributions
 - Binomial, Poisson, Geometric, Exponential, Normal
4. Concept of likelihood
5. Statistical inference
 - Estimation (maximum likelihood)
 - Hypothesis testing
 - Modelisation



Biostatistics

- **Biostatistics** – Application of statistical reasoning and methods to the solution of biological, medical, and public health problems
- *Biostatistical tools* are involved throughout a study in:
 1. Formulating ideas, questions, and hypotheses in quantitative terms
 2. Study design
 3. Data collection and monitoring
 4. Data analysis
 5. Interpretation of results

Epidemiologists use biostatistical tools...
Genetic epidemiologists use them in the search for genes influencing human diseases

1. Sampling and exploratory data analysis





Sampling

- A study population is a collection of entities under study
- A sample is a part or subset of the population



- Random sample from the whole province
- Random sample from one region
- Sample of cases (from a hospital) with age-matched controls
- Random sample of families
- Sample of families ascertained through a proband
- ...

- Parameters are numbers describing a population
- Statistics are numbers describing a sample



Sampling

- Sample size ' n ' is the number of elements in the sample.
- Bias in sampling occurs when the sample was selected in a manner that guaranteed that it would not be representative of the population.
- Principles of Experimental Design:
 - Control
 - Randomization
 - Replication



What is statistics?

- Descriptive statistics (or exploratory data analyses) include the collection, presentation, and description of sample data.
- Inferential statistics refer to the technique of interpreting the values resulting from the descriptive techniques and making decisions and drawing conclusions about the population.



Exploratory data analysis

Methods for organizing data:

- Displaying data:
 - Tables
 - Histograms or bar charts
 - Scatter plots, Boxplots, etc.
- Grouping data:
 - Frequency distributions
- Summarizing data:
 - Measures of central tendency (mean, median, mode)
 - Measures of dispersion (variance, range)



Leprosy data

From Alexandre Alcaïs presentation

'QTassoc.ped' file

Columns:

1. Family ID
2. Individual ID
3. Father ID
4. Mother ID
5. Gender (1=Male, 2=Female)
6. Affection Status (1=unaffected, 2=affected)
7. Simulated Quantitative Trait
8. Genotypes...



Introduction to R

`read.table()`

`dim()`

`length()`

`sum()`

etc.



Exploratory data analysis in R

`table()`

`mean()`

`var()`

`summary()`

`hist()`

`boxplot()`

etc.

2. Concepts of Probability





Probability

- Probability asks you about the chance that something specific will happen when you know the possibilities (that is, you know the population).
 - *Example:* How likely it is that a head will result when a coin is tossed?
- Statistics, in contrast, asks you to draw a sample, describe the sample (*exploratory data analysis*), and then make inferences about the population based on the information found in the sample (*inferential statistics*).
 - *Example:* determining if a new drug shortens the recovery time from a certain illness.



Probability

- One of the main objectives of a statistician is to draw conclusions about a population of objects by conducting an experiment. The first step is to identify the *listing of all possible outcomes*, i.e. the sample space 'S'.

- *Example: Rolling a die*

$$S = \{ 1, 2, 3, 4, 5, 6 \}$$

- The concept of probability provides a measure of the uncertainty associated with an event and is the key to most statistical methods. Probability of some event E is usually expressed as $P(E)$.

- *Example: $P(\text{number 6 on top}) = 1/6$*



Probability

- The complement of an event, E^C , is the event that E does not occur, so

$$P(E) = 1 - P(\text{not } E) \text{ or}$$

$$P(E) + P(E^C) = 1$$

- *Example:* Coin toss

$$P(\text{head}) = 0.5$$

$$P(\text{tail}) = P(\text{not head}) = 1 - 0.5 = 0.5$$

$$P(\text{head}) + P(\text{tail}) = 0.5 + 0.5 = 1$$





Probability

- Classical interpretation of probability:
 - If there are N equally likely possible outcomes, of which m have the characteristic E then

$$P(E) = \frac{m}{N}$$

- *Example:* There are 50 black, 45 white, and 5 gold marbles in a bag, what is the probability that you will draw a gold one?

$$P(E) = \text{picking a gold marble} = \frac{5}{(50 + 45 + 5)} = 0.05$$



Probability

- Frequentist interpretation of probability:
 - If an experiment is repeated N times under identical situations, and event E occurs m times out of those n trials, then

$$P(E) \approx \frac{m}{n}$$

- *Example:* 10 coin tosses of a fair coin

$n = 10$, $m = 4$, and estimate of $P(\text{head})$ is 0.4

- But with 1000 coin tosses:

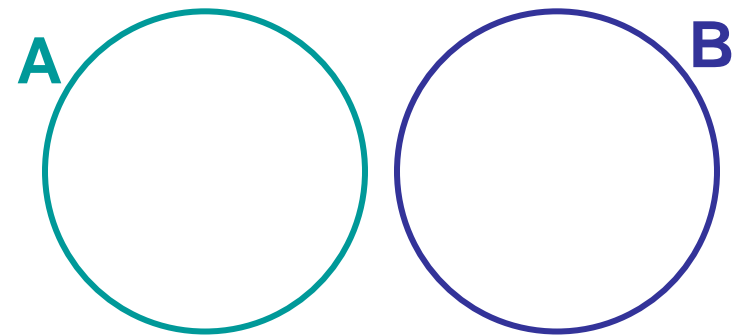
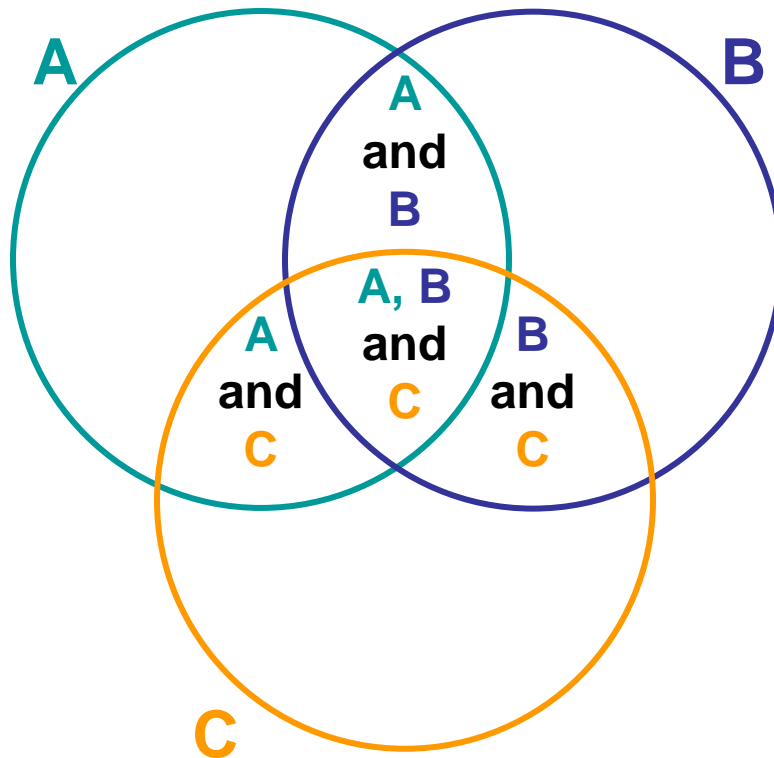
$n = 1000$, $m = 501$, and estimate is 0.501

- In the limit, as n approaches infinity, $P(\text{head}) = 0.5$



Union and Intersection

- Union symbol is \cup and means “OR” or either
- Intersection symbol is \cap and means “AND” or both



A and B are mutually exclusive: $P(A \cap B) = 0$



Rules of Probability

1. $P(E) \geq 0$, for any event E in the sample space S
2. $P(S) = 1$
3. For mutually exclusive events E_i and E_j (i.e., events i and j cannot occur at the same time):

$$P(E_i \cup E_j) = P(E_i) + P(E_j)$$

where \cup represents the union of events i and j

More generally, for any set of n pairwise mutually exclusive events:

$$P(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n P(E_i)$$

- Examples:

- $P(\text{head OR tail}) = P(\text{head}) + P(\text{tail}) = 0.5 + 0.5 = 1$
- $P(1 \text{ or } 2 \text{ or } 3 \text{ for a single roll of a die}) =$
 $P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$



Rules of Probability

4. Additive rule:

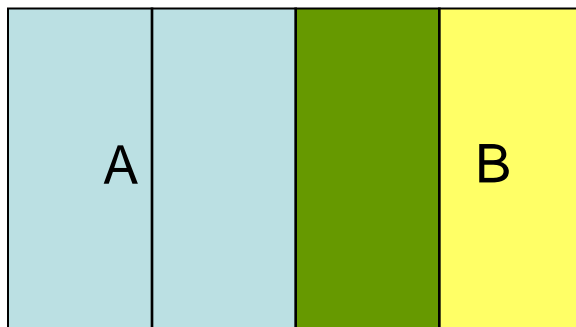
$$P(E_i \cup E_j) = P(E_i) + P(E_j) - P(E_i \cap E_j)$$

0 for mutually exclusive events

5. Conditional probability of an event E_i given an event E_j :

$$P(E_i | E_j) = \frac{P(E_i \cap E_j)}{P(E_j)}, \text{ where } P(E_j) \neq 0$$

Example:



$$P(A) = 3/4$$

$$P(B) = 1/2$$

$$P(A | B) = 1/2$$

$$P(B | A) = 1/3$$

$$P(A \cap B) = P(A | B) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$$P(A \cap B) = P(B | A) \cdot P(A) = \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{4}$$



Rules of Probability

6. Multiplicative rule:

$$\begin{aligned}P(E_i \cap E_j) &= P(E_i | E_j) \cdot P(E_j) \\ &= P(E_j | E_i) \cdot P(E_i)\end{aligned}$$

Dividing by $P(E_j)$ or $P(E_i)$ on each side we get back the conditional probabilities of E_i given E_j and E_j given E_i :

$$P(E_i | E_j) = \frac{P(E_i \cap E_j)}{P(E_j)}, \text{ and } P(E_j | E_i) = \frac{P(E_i \cap E_j)}{P(E_i)}$$

The conditional probability is the probability of both events, divided by the probability of the conditioned event.



Rules of Probability

6. (continued)

When events E_i and E_j are independent:

$$P(E_i \cap E_j) = P(E_i) \cdot P(E_j)$$

Two events are independent if the occurrence of one has no impact on the occurrence of the other so that

$$P(E_i | E_j) = P(E_i) \quad \text{and} \quad P(E_j | E_i) = P(E_j)$$

and

$$P(E_i \cap E_j) = P(\cancel{E_i} | E_j) \cdot P(E_j) = P(E_i) \cdot P(E_j)$$

Note: A red arrow points from the red $P(E_j)$ above to the $P(E_j)$ in the equation, and a red line is drawn through the E_i in the conditional probability term.



Joint, marginal, conditional

Example: If you flip a coin twice, what is the probability of 2 heads? Assuming it's a fair coin:

possibilities

H_1H_2	Marginal probabilities: $P(H_1) = P(T_1) = \frac{1}{2}$ $P(H_2) = P(T_2) = \frac{1}{2}$
H_1T_2	
T_1H_2	
T_1T_2	

Joint probabilities:

$$P(H_1 \cap H_2) = P(H_1) \cdot P(H_2) = \frac{1}{4}$$

$$P(H_1 \cap T_2) = P(H_1) \cdot P(T_2) = \frac{1}{4}$$

$$P(T_1 \cap H_2) = P(T_1) \cdot P(H_2) = \frac{1}{4}$$

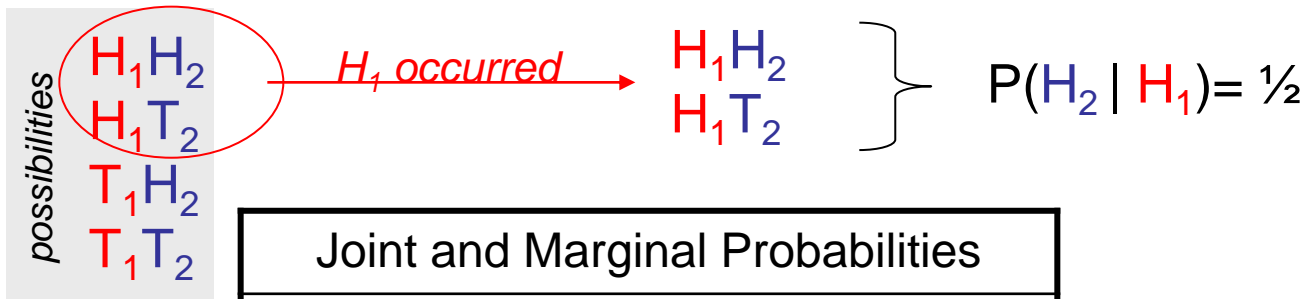
$$P(T_1 \cap T_2) = P(T_1) \cdot P(T_2) = \frac{1}{4}$$

		Toss 1		
		H_1	T_1	
Toss 2	H_2	HH $\frac{1}{4}$	TH $\frac{1}{4}$	$\frac{1}{2}$
	T_2	HT $\frac{1}{4}$	TT $\frac{1}{4}$	$\frac{1}{2}$
		$\frac{1}{2}$	$\frac{1}{2}$	1



Joint, marginal, conditional

Example (continued): what is the probability of a second head turning up, given the first flip is a head?



		Toss 1		
		H_1	T_1	
Toss 2	H_2	HH $\frac{1}{4}$	TH $\frac{1}{4}$	$\frac{1}{2}$
	T_2	HT $\frac{1}{4}$	TT $\frac{1}{4}$	$\frac{1}{2}$
		$\frac{1}{2}$	$\frac{1}{2}$	1

$$P(H_2 | H_1) = P(H_2 \cap H_1) / P(H_1)$$

$$= (\frac{1}{4}) / (\frac{1}{2}) = \frac{1}{2}$$

The two events are independent



Bayes' Theorem

- Suppose there is some *prior* information regarding the occurrence of event A that can be quantified by $P(A)$
- How is the information about A changed by observing event B? That is:

What is the posterior probability of A given B, $P(A|B)$?

- If A and B are *independent* the information is not changed at all since $P(A|B) = P(A)$
- When there is dependence, we can incorporate the information that B occurred by using:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$



Bayes' Theorem

- What is the posterior probability of A given B?

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B | A) \cdot P(A)}{P(B)}$$

Prior probability of A before observing B

- General form of Bayes' Theorem:

If events A_1, A_2, \dots, A_n form a partition of S and B is an event with $P(B) > 0$ then for each $i = 1, 2, \dots, n$

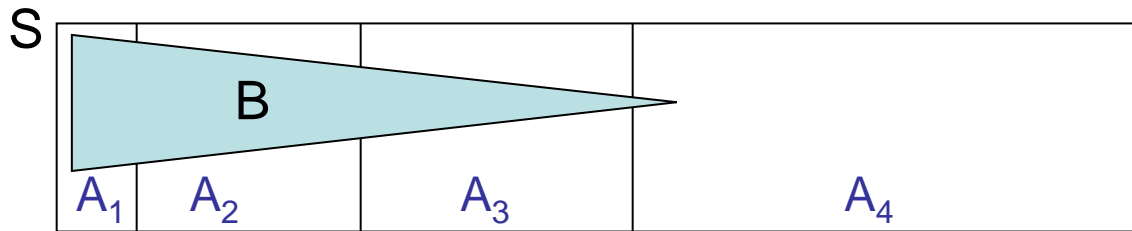
$$P(A_i | B) = \frac{P(B | A_i) \cdot P(A_i)}{\sum_{j=1}^n P(B | A_j) P(A_j)}$$

Events A_1, A_2, \dots, A_n form a partition of S if they are pairwise mutually exclusive and $\bigcup_{i=1}^n A_i = S$



Bayes' Theorem

Example: Suppose that a disease state that can be classified into 4 severity levels: A_1 = severe, A_2 = moderate, A_3 = mild, A_4 = healthy (no disease), and B represents a hospitalization event

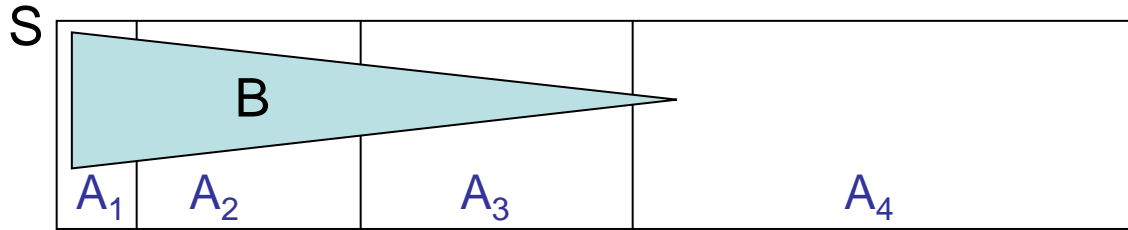


A_1 , A_2 , A_3 , and A_4 form a partition of S so that

$$P\left(\bigcup_{i=1}^n A_i\right) = P(A_1) + P(A_2) + P(A_3) + P(A_4) = P(S) = 1$$



Bayes' Theorem



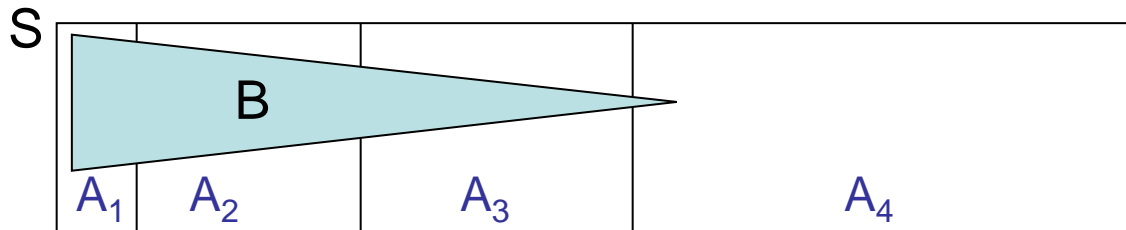
Event B (hospitalization) necessarily happens under one of the disease conditions (one of the A_i events):

$$\begin{aligned} P(B) &= P(B \cap S) \\ &= P(B \cap (\bigcup_{i=1}^n A_i)) \\ &= P(\bigcup_{i=1}^n (B \cap A_i)) \\ &= \sum_{i=1}^n P(B \cap A_i) \\ &= \sum_{i=1}^n P(B | A_i)P(A_i) \end{aligned}$$

Sum of joint probabilities or
Weighted sum of conditional probability



Bayes' Theorem



Suppose that we know the prevalence of the different disease state:

$$P(A_1)=0.05, P(A_2)=0.10, P(A_3)=0.20, P(A_4)=0.65$$

and that the conditional probabilities of a hospitalization given a disease state are:

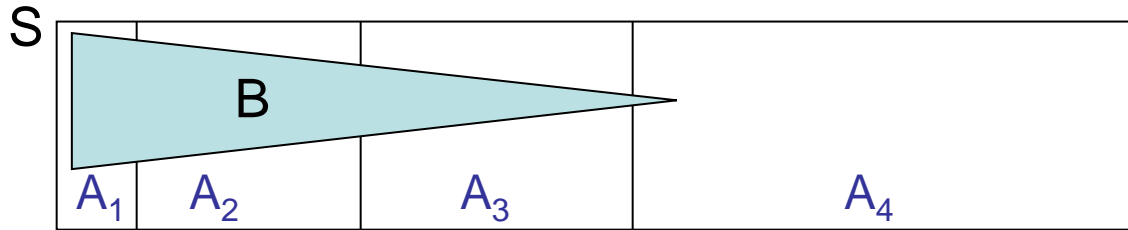
$$P(B | A_1)=0.8, P(B | A_2)=0.6, P(B | A_3)=0.25, P(B | A_4)=0.05$$

What is the probability that an individual has the severe state of the disease once we know that he (she) has been hospitalized?

We can answer that question using Bayes' theorem



Bayes' Theorem



$$\begin{aligned} P(A_1 | B) &= \frac{P(B | A_1) \cdot P(A_1)}{\sum_{j=1}^n P(B | A_j) P(A_j)} \\ &= \frac{(0.80) \cdot (0.05)}{(0.80) \cdot (0.05) + (0.60) \cdot (0.10) + (0.25) \cdot (0.20) + (0.05) \cdot (0.65)} \\ &= \frac{0.04}{0.18} = \frac{P(B \cap A_1)}{P(B)} \\ &= 0.22 \end{aligned}$$



Bayes' Theorem

A - Disease State					
B	Severe	Moderate	Mild	No disease	Marginal
Hospital stay	$P(B A_1)=0.80$	$P(B A_2)=0.60$	$P(B A_3)=0.25$	$P(B A_4)=0.05$	0.18
	$P(B \cap A_1)=0.04$	$P(B \cap A_2)=0.06$	$P(B \cap A_3)=0.05$	$P(B \cap A_4)=0.03$	
	$P(A_1 B)=0.22$	$P(A_2 B)=0.33$	$P(A_3 B)=0.27$	$P(A_4 B)=0.18$	
No Hospital stay	$P(B^c A_1)=0.20$	$P(B^c A_2)=0.40$	$P(B^c A_3)=0.75$	$P(B^c A_4)=0.95$	0.82
	$P(B^c \cap A_1)=0.01$	$P(B^c \cap A_2)=0.04$	$P(B^c \cap A_3)=0.15$	$P(B^c \cap A_4)=0.62$	
	$P(A_1 B^c)=0.01$	$P(A_2 B^c)=0.05$	$P(A_3 B^c)=0.18$	$P(A_4 B^c)=0.76$	
Marginal	0.05	0.10	0.20	0.65	1

- Marginal probability = Weighted sum of conditional probabilities
- Joint probability = Product of conditional by the weight for that stratum
- Conditional probability can be calculated based on the joint or conditional probabilities using Bayes' theorem



Bayes' Theorem

A - Disease State					
SNP Allele	Severe	Moderate	Mild	No disease	Marginal
C	$P(C A_1)=0.80$	$P(C A_2)=0.60$	$P(C A_3)=0.25$	$P(C A_4)=0.05$	0.18
	$P(B \cap A_1)=0.04$	$P(B \cap A_2)=0.06$	$P(B \cap A_3)=0.05$	$P(B \cap A_4)=0.03$	
	$P(A_1 C)=0.22$	$P(A_2 C)=0.33$	$P(A_3 C)=0.27$	$P(A_4 C)=0.18$	
T	$P(T A_1)=0.20$	$P(T A_2)=0.40$	$P(T A_3)=0.75$	$P(T A_4)=0.95$	0.82
	$P(B^c \cap A_1)=0.01$	$P(B^c \cap A_2)=0.04$	$P(B^c \cap A_3)=0.15$	$P(B^c \cap A_4)=0.62$	
	$P(A_1 T)=0.01$	$P(A_2 T)=0.05$	$P(A_3 T)=0.18$	$P(A_4 T)=0.76$	
Marginal	0.05	0.10	0.20	0.65	1

What is the probability that someone with a T allele does not have the disease?

$$P(A_4 | T) = \frac{P(T \cap A_4)}{P(T)} = \frac{P(T | A_4) \cdot P(A_4)}{\sum_{j=1}^n P(T \cap A_j)} = \frac{P(T | A_4) \cdot P(A_4)}{\sum_{j=1}^n P(T | A_j) P(A_j)} = 0.76$$

3. Common Probability Distributions





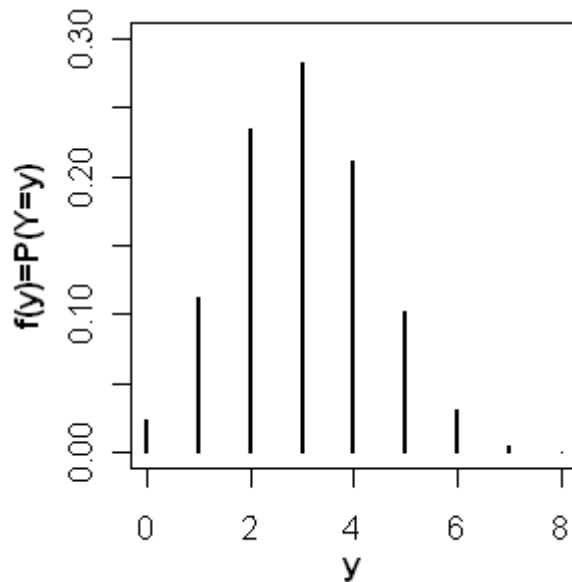
Probability Distributions

- A random variable is a function from S to the real numbers.
 - *Discrete* random variable: takes a countable (finite or infinite) number of values
 - *Continuous* random variable: takes a continuum of values (interval or real numbers)
- A probability distribution (or probability density function) is a function describing the relative frequency distribution of a random variable



Probability Distributions

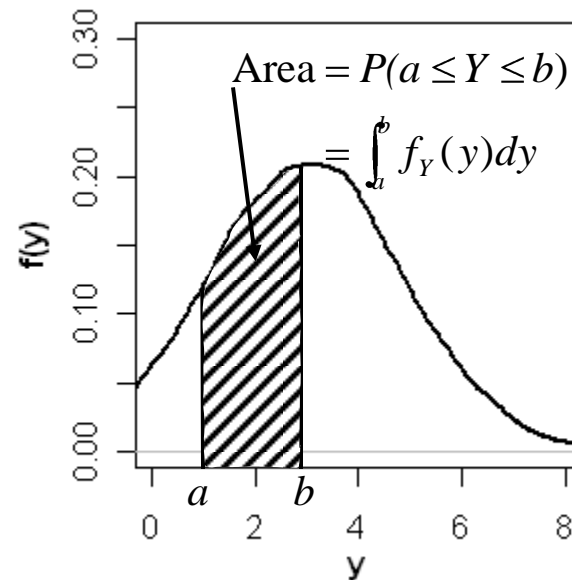
Discrete : $f_Y(y) = P(Y = y)$



$$\sum_{\text{all } y} P(Y = y) = 1$$

Continuous : for all a and b

$$P(a \leq Y \leq b) = \int_a^b f_Y(y) dy$$



$$\int_{-\infty}^{\infty} f_Y(y) dy = 1$$



Probability Distributions

Features of a probability distribution:

1. Expected value ($E[Y] = \mu$) measures the central tendency or location of the distribution

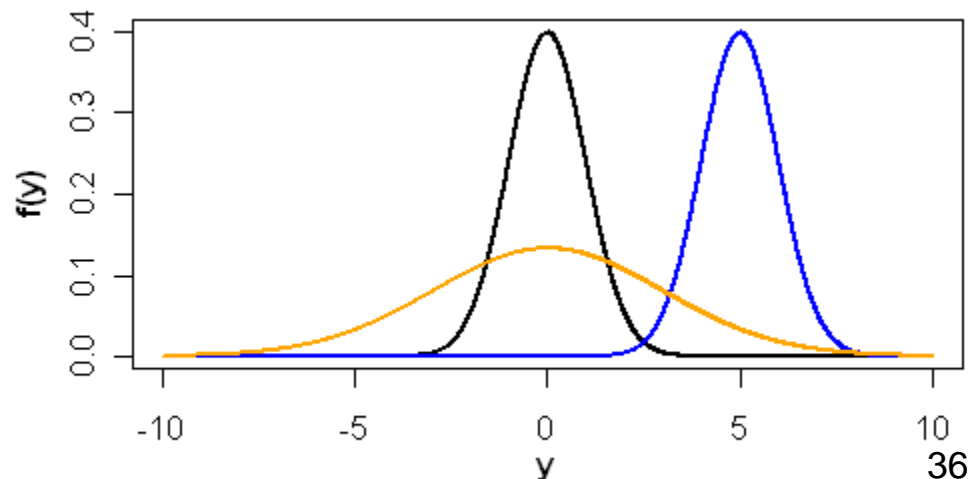
$$\text{Discrete : } E[Y] = \sum_{\text{all } y} yf_Y(y), \text{ Continuous : } E[Y] = \int_{-\infty}^{\infty} yf_Y(y)dy$$

2. Variance ($Var[Y] = \sigma^2$) measures the dispersion or how spread out the distribution is

$$Var[Y] = E[(Y - E[Y])^2]$$

Standard deviation :

$$\sigma = \sqrt{Var[Y]}$$





Probability distributions in R

d = density function $P(Y = y)$

p = cumulative distribution function $P(Y \leq y)$

q = quantile

r = random generation

punif, pbinom, ppois, pgeom, pexp, pnorm, pchisq, pt, pf



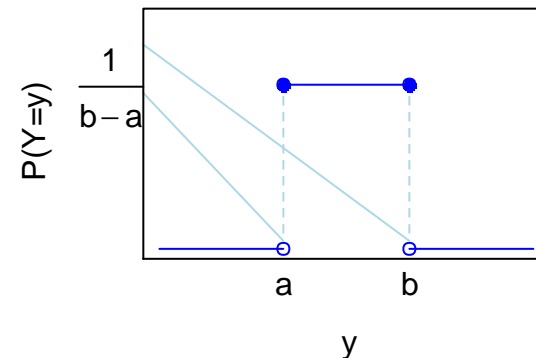
Uniform Distribution

- A random variable Y follows a discrete uniform distribution if each of the finite set of values that Y can take is equally probable
 - If Y has n different values then for all of these n values

$$P(Y = y) = 1/n$$

- Continuous uniform distribution, parameters $a, b \in (-\infty, \infty)$:

$$P(Y = y) = \frac{1}{b-a} \text{ for } a \leq y \leq b, 0 \text{ otherwise}$$





Binomial Distribution

- Assume n independent trials, each with 2 possible, mutually exclusive outcomes: *yes (success)* or *no (failure)*
- For each trial: $P(\text{yes}) = p$ and $P(\text{no}) = 1 - p = q$
- $Y =$ number of successes
 Y is a binomial random variable with parameters n, p

$$Y \sim \text{binomial}(n, p)$$

$$P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y} \quad \text{for } y = 0, 1, \dots, n$$

$$\binom{n}{y} = \frac{n!}{y!(n-y)!} \quad \begin{array}{l} \# \text{ possible ways of} \\ \text{obtaining } y \text{ successes} \\ \text{and } n-y \text{ failures} \end{array}$$

Probability associated with each way

n factorial ($n!$) = number of possible arrangements of n objects
 $= n(n-1)(n-2)(n-3)\cdots(1)$



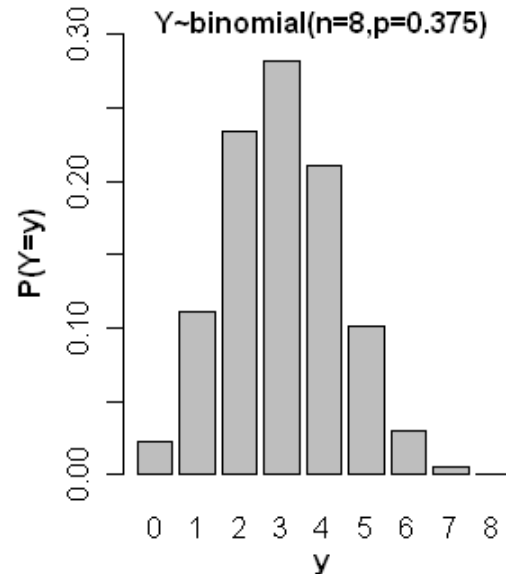
Binomial Distribution

$$Y \sim \text{binomial}(n, p)$$

$$E[Y] = np$$

$$\text{Var}[Y] = np(1 - p)$$

Each trial is called a
Bernoulli trial



$$E[Y] = 3$$

$$\text{Var}[Y] = 1.875$$



Computer Laboratory

Example: Wright-Fisher finite population model

- $N=4$ diploid individuals (parent generation) produce an infinite pool of gametes from which $2N=8$ gametes are randomly sampled to give the daughter generation
- If there are 3 copies of allele A in the parent generation, what is the probability that there will also be 3 copies in the daughter generation?

$Y = \#$ copies of A in the daughter generation = $\#$ copies randomly sampled

$n = \#$ trials = total number of allele sampled = 8

$p =$ probability of sampling an A allele = $3/8$

$Y \sim \text{binomial}(8, 3/8)$

$$P(Y = 3) = \binom{8}{3} 0.375^3 (1 - 0.375)^{8-3} = 0.28$$



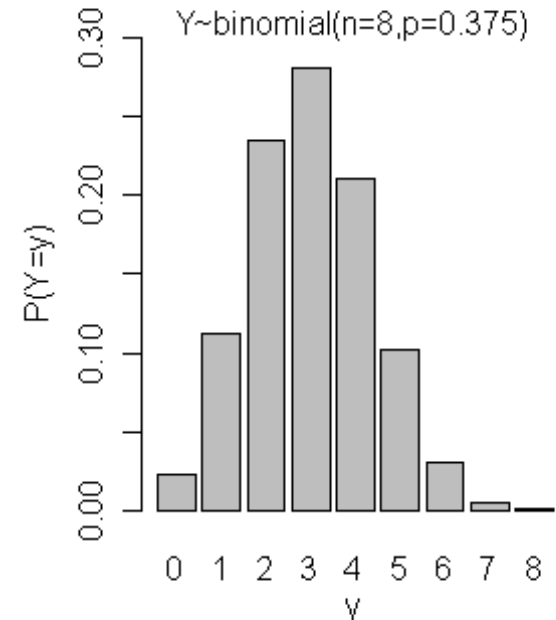
Computer Laboratory

Example (continued):

- What is the probability that the A allele will be lost in the daughter generation? What is the probability that the A allele will be fixed in the daughter generation?

$$P(Y = 0) = \binom{8}{0} 0.375^0 (1 - 0.375)^{8-0} = 0.02$$

$$P(Y = 8) = \binom{8}{8} 0.375^8 (1 - 0.375)^{8-8} = 0.0004$$





Poisson Distribution

- Same assumptions as binomial but for rare events and with an infinite number of trials
- Describes occurrences of events which are distributed randomly in space or time
- Y = number of events (successes)

$$Y \sim \text{Poisson}(\lambda)$$

$$E[Y] = \lambda$$

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!} \text{ for } y = 0, 1, 2, \dots$$

$$\text{Var}[Y] = \lambda$$

- When the binomial parameter n is large and p is small, the Poisson distribution with parameter $\lambda = np$ approximates the binomial



Computer Laboratory

Example: Crossover process

- What is the probability of observing 0 crossover in a single meiosis over a $1/3$ Morgan interval assuming no crossover interference?

$Y = \# \text{ crossovers}$

$\lambda = \text{expected \# crossovers} = 1/3$

(1 Morgan = genetic distance over which the number of expected crossover per meiosis is 1)

$Y \sim \text{Poisson}(1/3)$

$$P(Y = 0) = \frac{(1/3)^0 e^{-1/3}}{0!} = 0.72$$



Geometric Distribution

- Assume a series of Bernoulli trials, how many trials will be needed to achieve the 1st success?
- Y = number of trials until 1st success

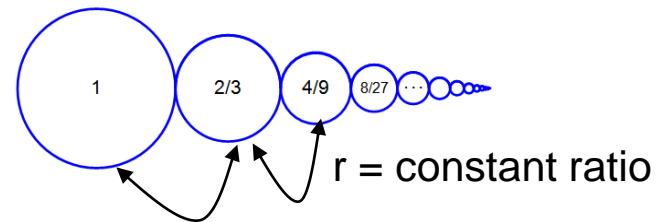
$$Y \sim \text{geometric}(p)$$

$$E[Y] = \frac{1}{p}$$

$$P(Y = y) = p(1 - p)^{y-1} \quad \text{for } y = 1, 2, \dots$$

$$\text{Var}[Y] = \frac{1 - p}{p^2}$$

- Geometric series: $\sum_{n=0}^{\infty} ar^n = \frac{a}{1 - r}$



- Memoryless (Markov) property: $P(Y = m + k \mid Y > m) = P(Y = k)$



Computer Laboratory

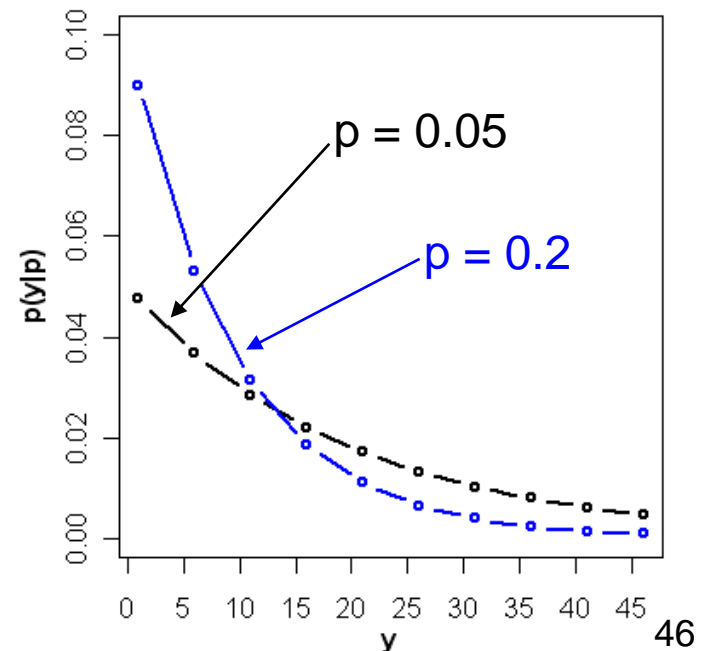
Example: There are 50 black, 45 white, and 5 gold marbles in a bag, what is the average number of draws needed until you pick a gold one?

$Y = \#$ draws before a gold marble is drawn
 $p =$ probability of drawing a gold marble $= 0.05$

$Y \sim \text{geometric}(0.05)$

$P(Y = y) = 0.05(0.95)^{y-1}$ for $y = 1, 2, \dots$

Expected value $= E[Y] = \frac{1}{0.05} = 20$





Exponential Distribution

- Models time or distance between independent events
- Continuous analog of the geometric distribution
- $1/\lambda$ is the rate of events

$$Y \sim \text{Exponential}(\lambda)$$

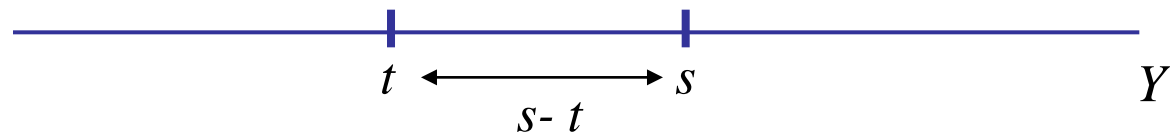
$$E[Y] = 1/\lambda$$

$$f_Y(y) = (1/\lambda)e^{-y/\lambda} \quad \text{for } y \geq 0$$

$$\text{Var}[Y] = 1/\lambda^2$$

- Memoryless (Markov) property:

$$P(Y > s \mid Y > t) = P(Y > s - t) \quad \text{for } s > t \geq 0$$





Computer Laboratory

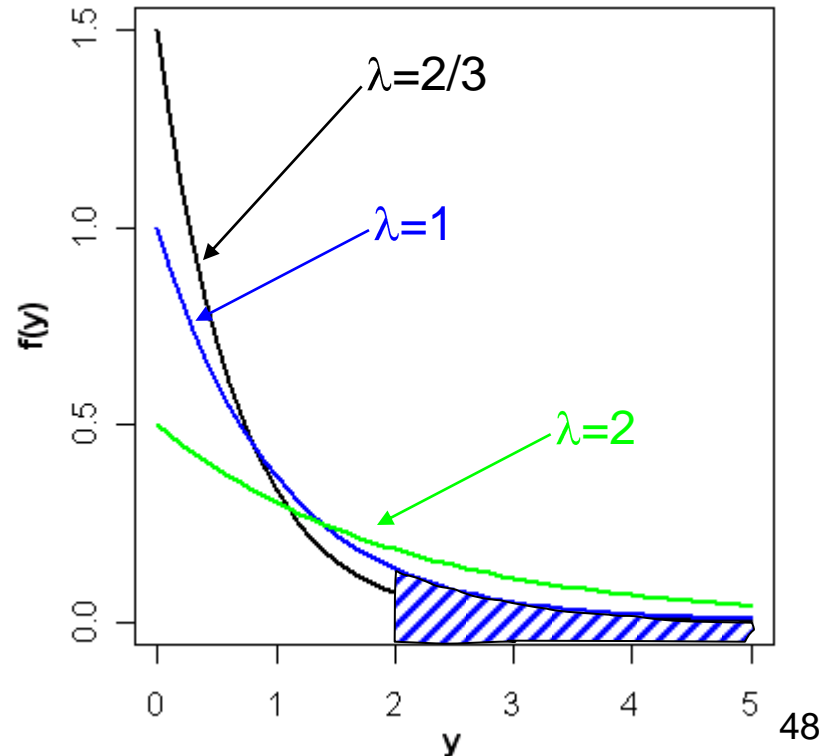
Example: Suppose that you usually get phone calls every hour on average. What is the probability that you can go shopping for 2 hours without getting a phone call?

$$\lambda = 1$$

$$f(y) = e^{-y}$$

$$P(Y > 2) = \int_2^{\infty} f_Y(y) dy$$

$$= \int_2^{\infty} e^{-y} dy = 0.135$$





Normal Distribution

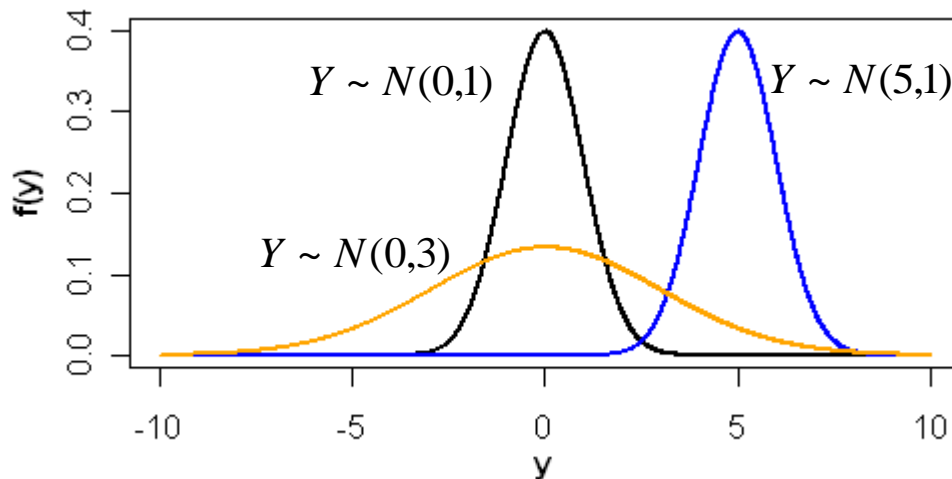
$$Y \sim N(\mu, \sigma)$$

$$E[Y] = \mu$$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad \text{for } -\infty < y < \infty$$

$$\text{Var}[Y] = \sigma^2$$

($\pi = 3.14159 = \text{a constant}$)



Characteristics:

- Infinite tails
- Symmetrical about its mean
- $P(\mu - 2\sigma < Y < \mu + 2\sigma) \cong 0.95$

Standard normal distribution $\longrightarrow Z = \frac{Y - \mu}{\sigma} = N(0,1)$



Normal Distribution

- Key role because of the Central Limit Theorem:
Let Y_1, \dots, Y_n follow the same non-normal distribution with mean μ and variance σ^2 , then for n large

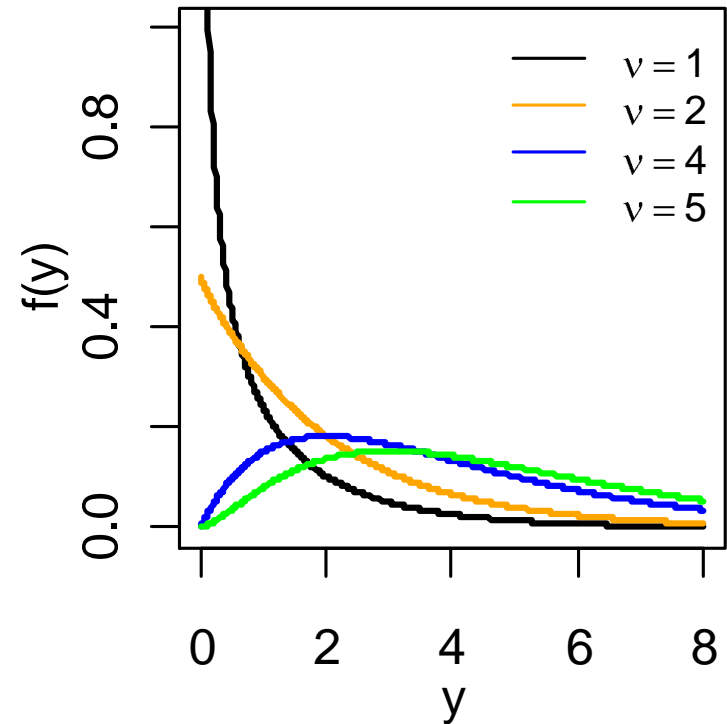
$$\bar{Y} \approx N(\mu, \sigma^2/n) \quad \text{where } \bar{Y} = \sum_{i=1}^n Y_i / n$$

- Approximates the distribution of any sample mean for large samples
- Represents the sum of a large number of small (finite variance), independent contributions
- Approximates the binomial and other distributions
- Describes many characteristics measured in a population



Chi-square distribution

- parameter
 $k > 0$ degrees of freedom (df)

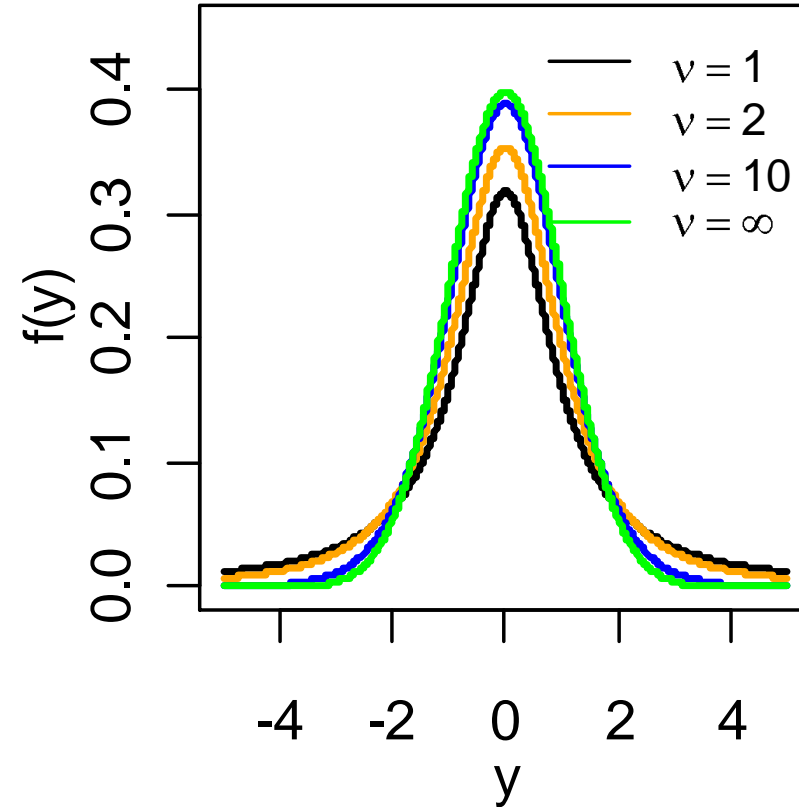


$$Y = \sum_{i=1}^k Z_i^2 \sim \chi_k^2, Z_i \text{'s are independent } N(0,1)$$



Student's t distribution

- parameter $\nu > 0$ df



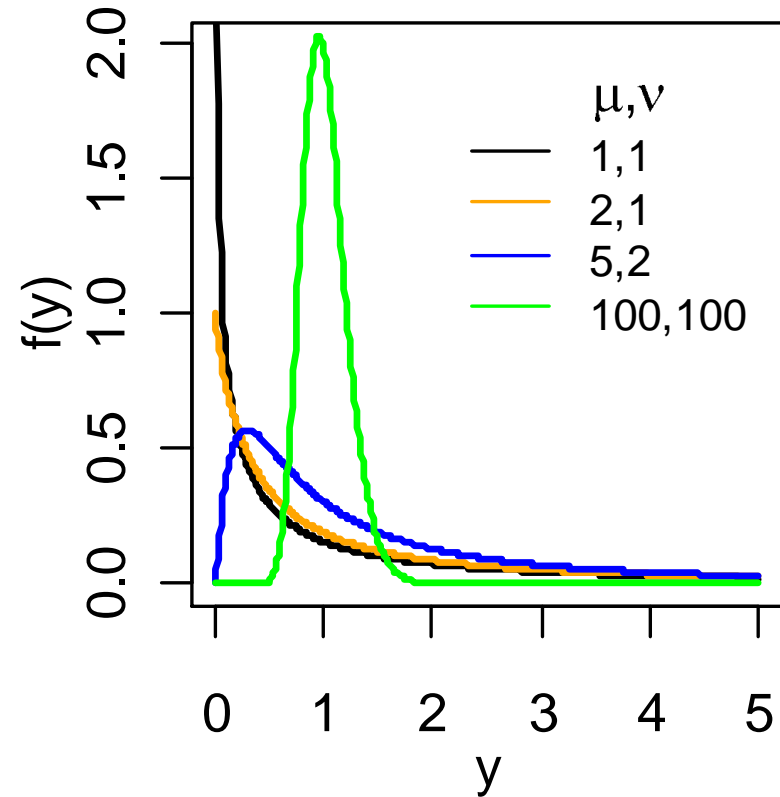
$$Y = \frac{Z}{\sqrt{V/\nu}} \sim t_\nu, Z \sim N(0,1) \text{ and } V \sim \chi_\nu^2 \text{ are independent } t$$



F distribution

- parameters

$\mu > 0$ and $\nu > 0$ dfs



$$Y = \frac{U/\mu}{V/\nu} \sim F_{\mu,\nu}, U \sim \chi_{\mu}^2 \text{ and } V \sim \chi_{\nu}^2 \text{ are independent}$$



Computer laboratory

For each distribution:

1. Simulate a vector variable sampled from the specific distribution
2. Calculate the mean and variance of the sampled variable
3. Compare to the expected mean and variance from the probability distribution
4. Histogram of the sampled variable
5. Compare to the probability distribution
6. Calculate the examples on the previous slides using R

4. Concept of Likelihood



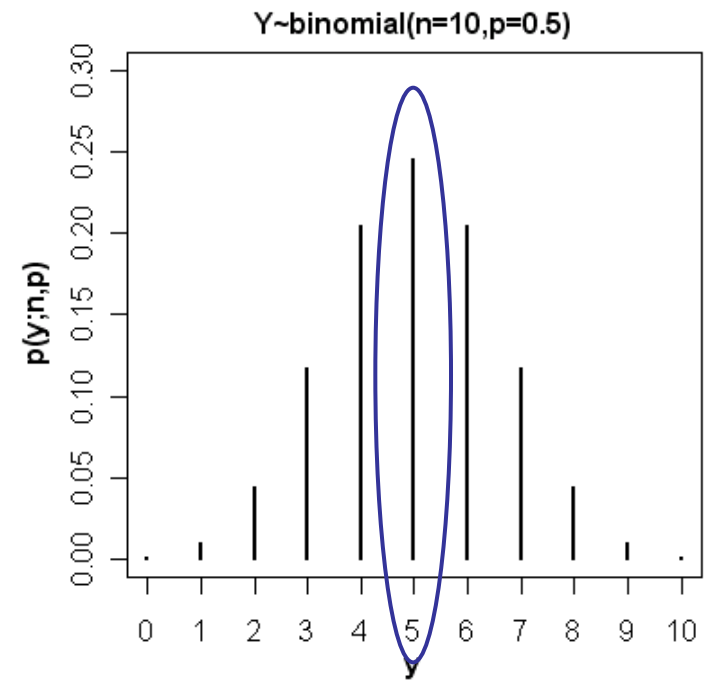


Likelihood

- Reconsider the binomial distribution:

$$\begin{aligned}f_Y(y) &= P(Y = y) \\ &= P(y; n, p) \\ &= P(y | n, p) \\ &= \binom{n}{y} p^y (1-p)^{n-y}\end{aligned}$$

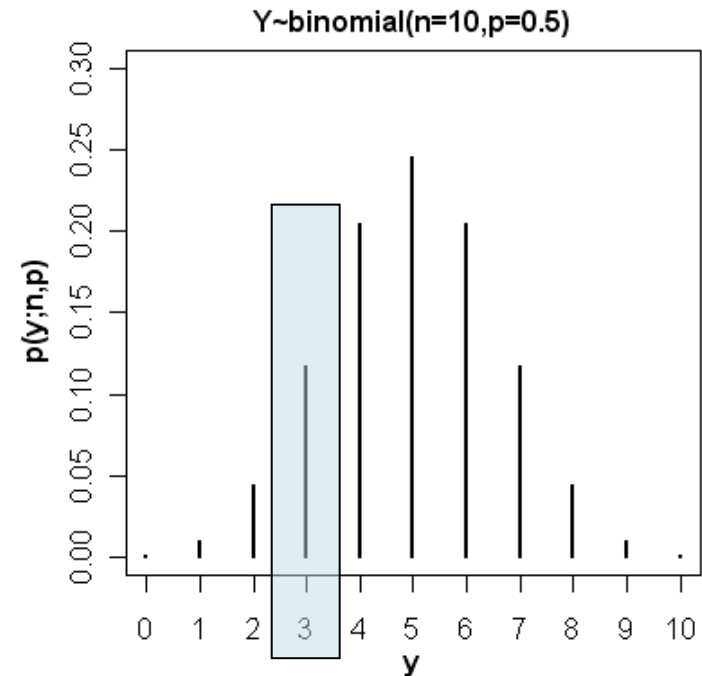
- Probability distribution given $n = 10$ and $p = 0.5$ are known
- Most likely value of y observed is 5 successes out of the 10 trials





Likelihood

- So far we've calculated the probability that the random variable Y takes on the value y
- Imagine the reverse...
 - p is not known
 - You perform 10 trials
 - 3 of those are successes



- What is the most likely value of p given we observed 3 successes?

For a fixed (known) observation of $Y = y$, treat p as a *random variable* and choose the value of p that would make the observation of $y = 3$ most likely



Likelihood

- *Probability*: Y is a random variable, with the probability of a particular realization of y calculated from the probability distribution, assuming p is known
- *Likelihood*: y is observed (known, fixed), p is variable

$$L(p | y, n) = \binom{n}{y} p^y (1-p)^{n-y}$$

The likelihood of the parameter being equal to p when the data y is observed



Likelihood

- In general, for a random sample Y_1, \dots, Y_n from the distribution $f_Y(y; \theta)$ and if y_1, \dots, y_n are the observed (fixed) values of the Y_i 's, the likelihood function is

$$\begin{aligned} L(\theta) &= L(\theta \mid y_1, \dots, y_n) \\ &= f_Y(y_1; \theta) f_Y(y_2; \theta) \dots f_Y(y_n; \theta) \\ &= \prod_{i=1}^n f_Y(y_i; \theta) \end{aligned}$$

The likelihood is the joint probability distribution of the data viewed as a function of the parameters

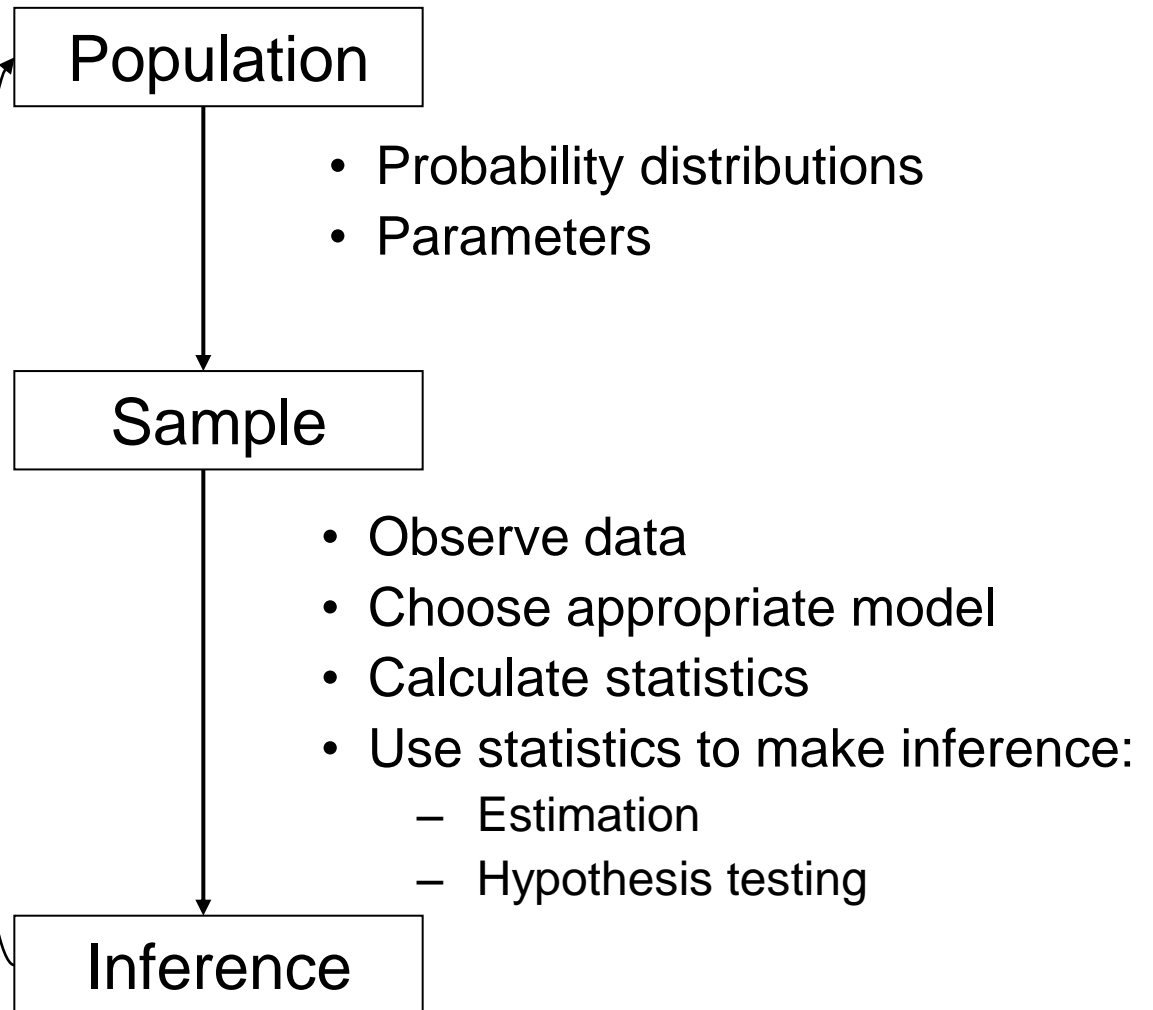
5. Statistical Inference





Statistical inference

Statistical inference:
Making a statement about a population based on data from a sample and describing the uncertainty associated with the statement





Estimation

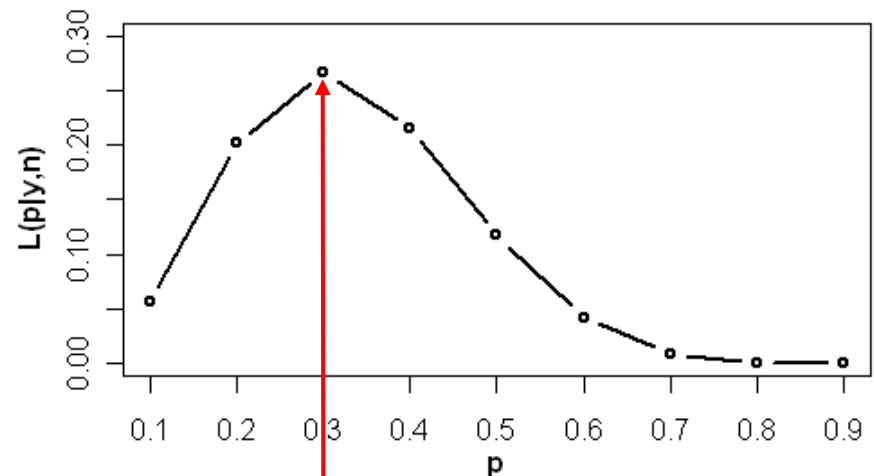
Maximum likelihood

- Reconsider the 10 Bernoulli trials with 3 successes observed:

Choose the value of p that would make the observation “3 successes” most likely by trying many different values of p

Find the value of p that provides the highest (maximum) likelihood for the observed data

$$L(p | 3,10) = \binom{10}{3} p^3 (1-p)^{10-3}$$



Maximum Likelihood Estimate
(MLE) = 0.3



Estimation

- How do we find the value of the parameter that maximizes the likelihood?
 - For simple problems, set the 1st derivative of the likelihood, with respect to the variable p , equal to 0 and solve for p (also check that the 2nd derivative is negative)
 - Value of p which maximizes $L(p)$ also maximizes the natural log of L and it is easier to work with the log-likelihood

$$\ln L(p | y, n) = \ln \binom{n}{y} + y \ln(p) + (n - y) \ln(1 - p)$$

- 1st derivative: $\frac{d[\ln L(p | y, n)]}{dp} = \frac{y}{p} + \frac{(n - y)}{(1 - p)} = 0$

- MLE: $\hat{p} = \frac{y}{n}$

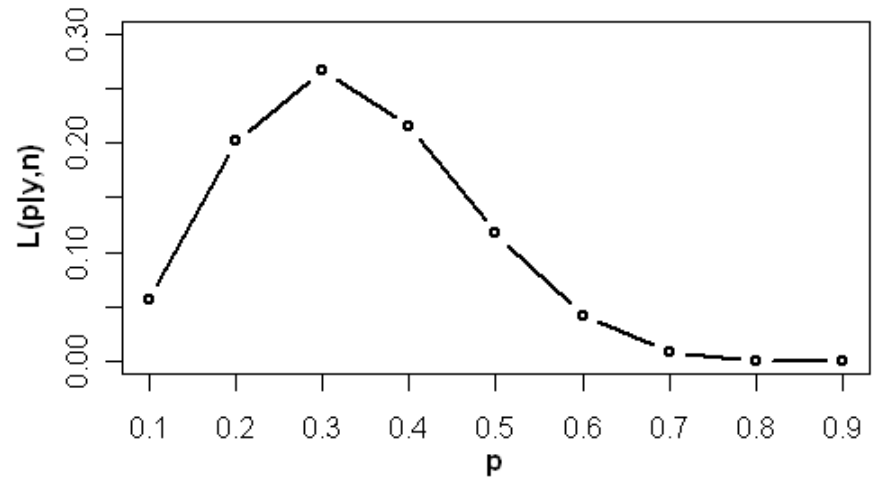


Estimation

- MLE: $\hat{p} = \frac{y}{n}$
- Previous example:
 $n = 10, y = 3$

$$\hat{p} = \frac{y}{n} = \frac{3}{10} = 0.3$$

$$L(p | 3, 10) = \binom{10}{3} p^3 (1-p)^{10-3}$$





Estimation

- How do we find the value of the parameter that maximizes the likelihood?
 - Calculating the derivative is difficult/impossible for more complex likelihood function
 - Iterative methods are needed to try many different values of p
 - Try all possible values of p
 - Not feasible
 - Could try increments, like values at every .05 interval
 - Not very efficient
 - Maximization algorithms
 - Newton-Raphson
 - Expectation-Maximization (EM)



Estimation

In large samples (i.e., asymptotically), MLEs have the following desirable properties:

1. **Sufficiency** – The estimator contains all available information about the parameter of interest
2. **Consistency** – The estimator converges on the parameter as the number of observations increases (related to unbiasedness)
3. **Efficiency** – Minimum variance among a class of estimators
4. **Asymptotic normality** – Under certain regularity conditions



Estimation

- Assuming a probability model for the data is required with the likelihood methods
 - The example given assumed each trial were Bernoulli so that Y followed a binomial distribution
- If this model is incorrect, the likelihood estimators are not meaningful
- When the probability model is correct (or close), likelihood methods are very useful to make inference



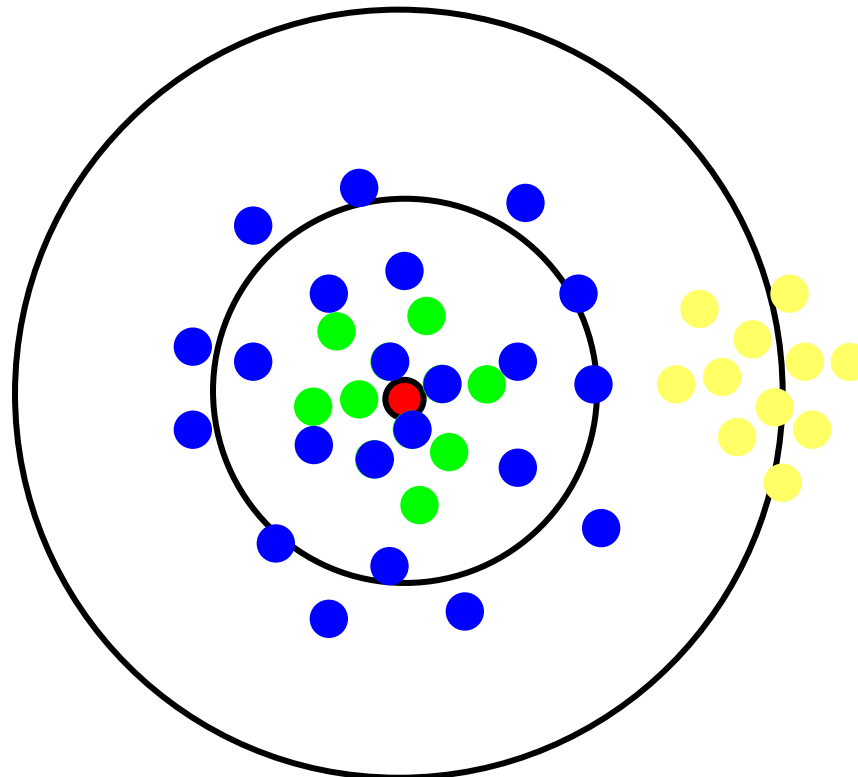
Estimation

- Other estimation procedures exist
- For example:
 - Least squares
 - Methods of moments:
Estimators are obtained by equating sample moments (such as mean or variance) with unobservable population moments and solving the resulting equations for the quantities to be estimated



Interval Estimation

- Point estimates provide no indication of their precision





Interval Estimation

- The standard deviation provides information about the variability of the estimate
- Confidence intervals around the point estimate provide a measure of how confident we are that the true parameter value lies within the interval
- A 95% confidence interval for a parameter θ is such that we have 95% chance of covering θ
 - If we draw a large number of samples and compute the confidence interval in each of them, the interval will include θ in 95% or more of the samples



Computer Laboratory

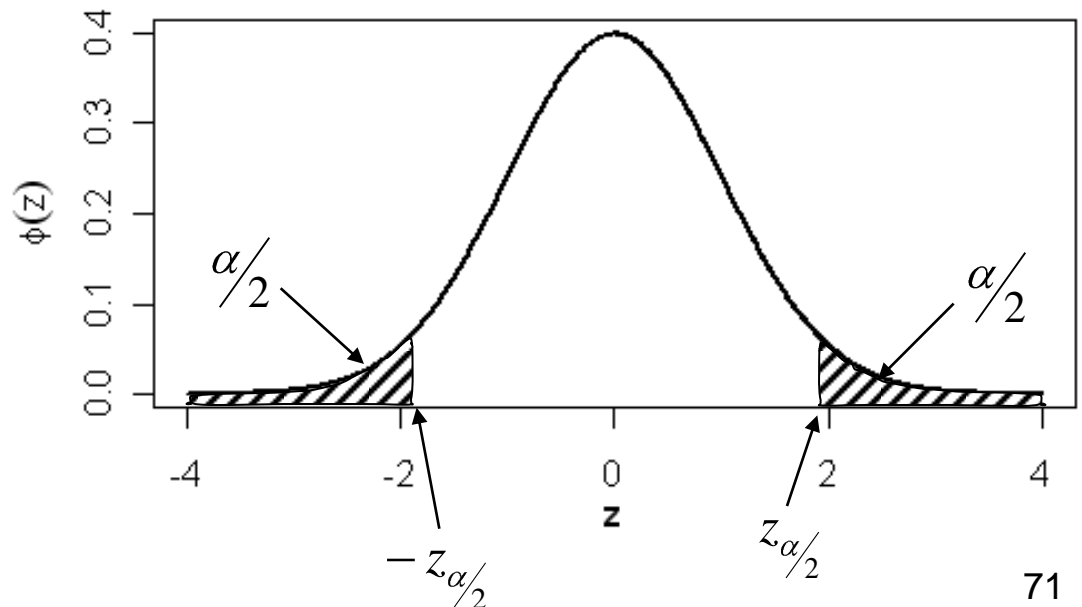
Example: Confidence interval for the sample mean

$$\bar{Y} = \sum_{i=1}^n Y_i / n \sim N(\mu, \sigma^2/n)$$

$$P\left(\bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

95% confidence
interval:

$$\left[\bar{Y} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{Y} + 1.96 \frac{\sigma}{\sqrt{n}} \right]$$





Hypothesis Testing

- Purpose: reaching a decision concerning a population by examining a sample of that population
- Null hypothesis (H_0): a statement about the hypothesized value of population parameters or about the relationship between parameters
- Alternative hypothesis (H_1)
- Test statistic calculated in the sample
- How likely is the value of the observed test statistic if H_0 is true?
- P-value: probability that the test statistic is as extreme or more as that observed in the sample when H_0 is true



Hypothesis Testing

- Steps to perform a hypothesis test:
 1. Select the probability model for the observed data
 2. Determine H_0 based on the problem being investigated (the opposite of what you would like to prove)
 3. Determine H_1
 4. Select a test statistic
 5. Select a decision rule and rejection (critical) region
 6. Determine if the test statistic falls into the critical region and make a statistical decision (inference)



Hypothesis Testing

Examples:

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu \neq \mu_0$$

(two-sided)

$$H_0: \mu \leq \mu_0 \quad \text{vs.} \quad H_1: \mu > \mu_0$$

(one-sided)

$$H_0: \mu \geq \mu_0 \quad \text{vs.} \quad H_1: \mu < \mu_0$$

(one-sided)



Hypothesis Testing

- Possible results of a hypothesis testing:

	Truth	
Conclusion based on data (sample)	H_0 true	H_0 false
Reject H_0	Type I error	Correct conclusion
Fail to reject H_0	Correct conclusion	Type II error



Hypothesis Testing

Conclusion based on data (sample)	Truth	
	H_0 true	H_0 false
Reject H_0	Type I error	Correct
Fail to reject H_0	Correct	Type II error

- $\alpha = P(\text{type I error}) = P(\text{rejecting } H_0 \mid H_0 \text{ is true})$
= level of significance
Aim: to keep Type I error small (small rejection region)
- $\beta = P(\text{type II error}) = P(\text{failing to reject } H_0 \mid H_0 \text{ is false})$
Power = $1 - \beta = P(\text{rejecting } H_0 \mid H_0 \text{ is true})$
Aim: to keep power high (type II error small)



Hypothesis Testing

p -value

- The p -value for a hypothesis test is the probability of obtaining by chance alone, when H_0 is true, a value of the test statistic as extreme or more extreme than the one computed from the sample
- The rejection region is determined by α , the desired level of significance (probability of committing a Type I error)
- The p -value gives an indication of how common or rare the computed value of the test statistic is if H_0 is true



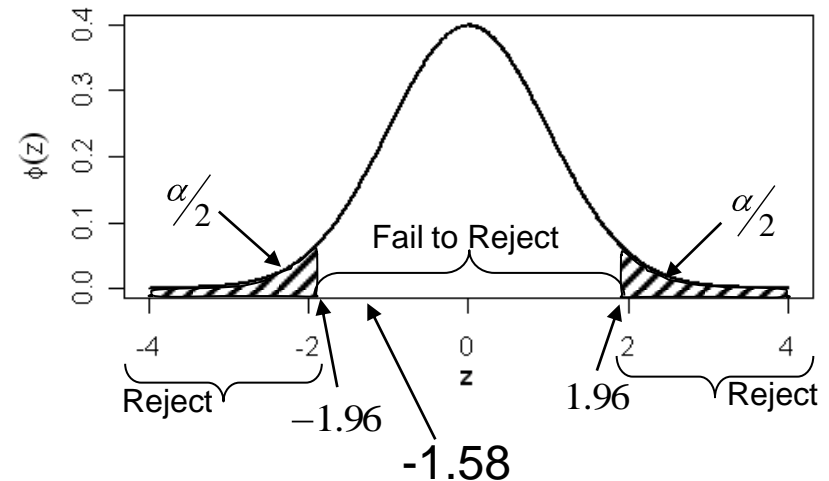
Computer Laboratory

Example: A sample of 100 infants from a study population gives an average birthweight of 2500g (s = sample standard deviation = 1000). We would like to know if the mean birthweight in this population is different from 3000g.

1. Two-sided test of
 $H_0: \mu = 3000$ vs. $H_1: \mu \neq 3000$
2. Test statistic

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \quad t \rightarrow z \text{ for large } n$$

$$t_{obs} = \frac{2500 - 3000}{1000/\sqrt{100}} = -1.58$$



$$p\text{-value} = P(z \leq -1.58) + P(z \geq 1.58) = 0.11$$



Hypothesis Testing

- Likelihood ratio tests:
 - Related to maximum likelihood estimators
 - Compare the likelihood under H_0 , i.e., the maximum likelihood with restriction $\theta = \theta_0$, to the unrestricted maximum likelihood (at the MLE for θ)

$$\lambda(y) = \frac{\sup_{\theta_0} L(\theta | y)}{\sup_{\theta} L(\theta | y)}$$

LRT statistic
Reject H_0 when small



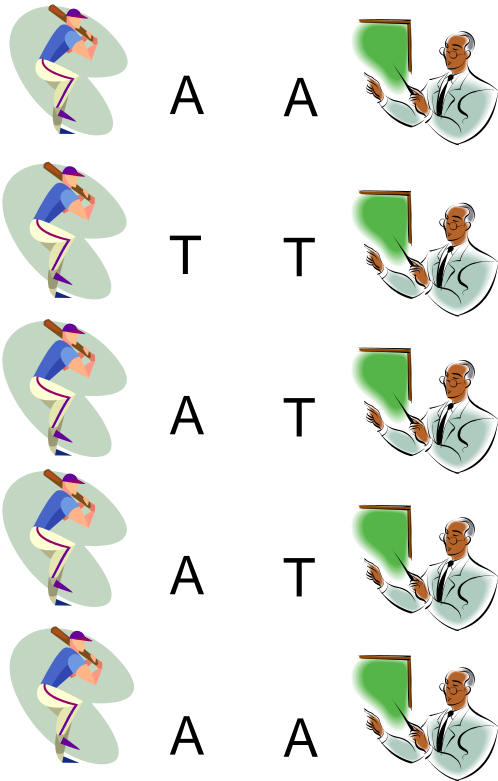
Hypothesis Testing

- Asymptotic tests and confidence intervals assume a distribution of the test statistic for large sample sizes
- What if we have a small sample size?
What if we are unsure of the probability model assumed?
 - Exact tests (e.g. Fisher's exact test)
 - Resampling techniques: permutation tests, bootstrap
(Permutation tests are exact tests if all possibilities are resampled)
 - Monte Carlo simulations
- Multiple testing → Inflation of the type I error
 - Corrections (e.g. Bonferroni), permutation tests, simulations

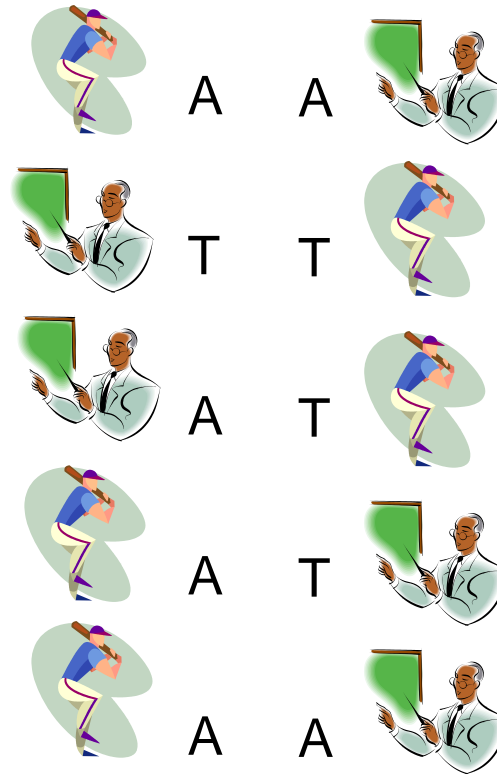


Hypothesis Testing

Example of permutation test



Calculate estimates and test statistic



Shuffle the labels and re-calculate
Repeat many times to obtain the null distribution, i.e., labels do not matter

p -value is the proportion of permutations where the original statistic is more extreme than the permuted one



Measures of Association

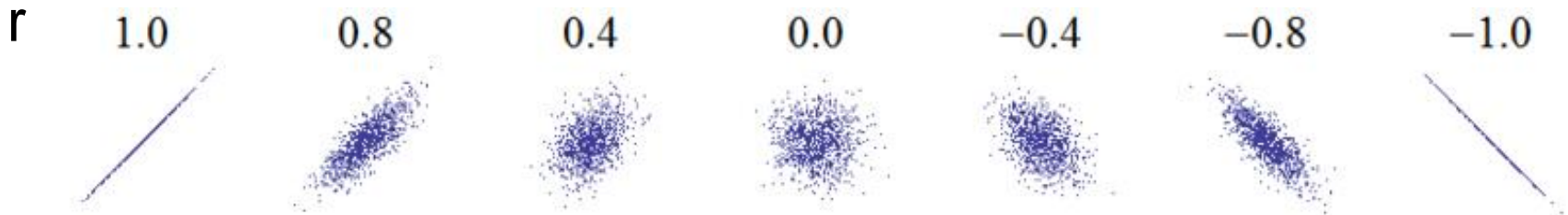
- Two continuous variables
 - Pearson's correlation coefficient
 - Simple linear regression
- Two categorical variables – Frequency tables
 - Relative risk
 - Odds ratio
 - Chi-square
 - Fisher's exact test
- A continuous and a categorical variable
 - ANOVA



Pearson's correlation coefficient

- Measures the strength of the linear relationship between 2 continuous variables
- $-1 \leq r \leq 1$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y}$$





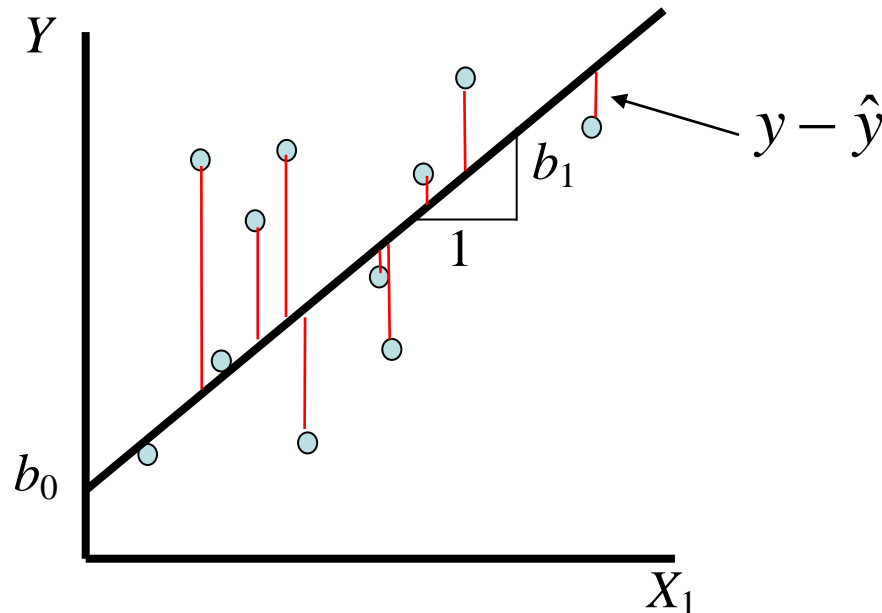
Simple Linear Regression

- Goal: Describe data or predict response
- Find the best fitting line to the data by minimizing the sum of squares of the residuals $\sum (y_i - \hat{y}_i)^2$ (least squares)

$$Y = b_0 + b_1 X_1$$

Hypothesis test
of $H_0: b_1 = 0$
using a t-test

$$r = \frac{s_x}{s_y} b_1$$





Simple Linear Regression

- Assumptions (LINE)
 - Linearity
 - Independence
 - Normality
 - Errors with homogenous variance
- Interpretation of b_1
 - Average change in Y for a 1-unit change in X
- Interpretation of b_0
 - Average Y when $X=0$
 - Can center X to make more interpretable



Odds Ratio

- Odds = $P/1-P$
- Odds Ratio (OR) = 1 means no association
- Example: AMD (age-related macular degeneration)
 - OR = odds of AMD_{C allele} / odds of AMD_{T allele}
 - OR = $(121/107)/(65/181) = 3.1$

	AMD	Control	Total
C allele	121	107	228
T allele	65	181	246
Total	186	288	474



Pearson's chi-square test

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

O_i = observed frequency

E_i = expected frequency

n = number of possible outcomes of each event

Use χ^2 to calculate p-value by comparing to chi-square distribution with $df = n - p$, where p is 1 + the number of parameters estimated to get the expected frequencies



Pearson's chi-square test

AMD example:

	AMD	Control	Total
C allele	121(89)	107(139)	228
T allele	65 (97)	181(149)	246
Total	186	288	474

$$E = (228/474)288 = 139$$

$$\chi^2 = (121-89)^2/89 + (107-139)^2/139 + (65-97)^2/97 + (181-149)^2/149 = 11.5+7.4+10.6+6.9 = 36.4$$

$$df = 1$$

$$p\text{-value} = 1.6E-09$$



Fisher's Exact Test

- An alternative to chi-square test when large sample approximation is not appropriate (if cell has expected frequency less than 10)
- Based on exact probabilities
 - Given the observed marginals, calculate the probability of all possible table configurations and determine how the extreme is the observed configuration
- Computed by most software packages

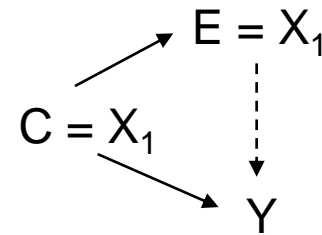


Multiple linear regression

- Multiple independent variables

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

- Estimates obtained using least squares
- Control for confounding
 - A confounder introduces a spurious association or distorts the association
- Interpretation of b_1
 - Average change in Y for a 1-unit change in X_1 controlling for X_2, \dots, X_n





Multiple logistic regression

- Dichotomous outcome (example: mortality)
- Model:

$$\log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \log(\text{odds}(Y = 1)) = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

- Why use odds rather than probability
 - Probability constrained between 0 and 1
 - Log odds goes from $-\infty$ to $+\infty$
- Estimate coefficients obtained using maximum likelihood
- Test using a Wald test, likelihood ratio test, or score test



Multiple logistic regression

- Interpretation of b_1
 - Change in log odds for a 1-unit change in x_1 controlling for the other x 's
- Interpretation of e^{b_1}
 - Change in odds for a 1-unit change in x_1 controlling for the other x 's (adjusted odds ratio)

- Example:

$$\text{Log odds(Death)} = b_0 + b_1(\text{age}) + b_2(\text{gender})$$

e^{b_1} = odds ratio of death for those who are 1 year older compared to 1 year younger controlling for gender



General linear models

- Family of regression models comprising linear and logistic regression models among others:

<u>Outcome</u>	<u>Model</u>
Continuous	Linear regression
Counts	Poisson regression
Survival	Cox model
Binomial	Logistic regression

- These models are used to:
 - Estimate the strength of and test for association between outcome and covariates while controlling for confounding
 - Model building
 - Risk prediction



Computer Laboratory

`cov()`

`cor()`

`chisq.test()`

`fisher.test()`

`lm()`

`glm()`

`anova()`

etc.

GeneticsBase Package:

- Allele frequency description
- Hardy-Weinberg Equilibrium (HWE) test



Learning Objectives

1. Sampling and exploratory data analysis
2. Concepts of probability
 - Rules of probability
 - Union and intersection
 - Conditional probability and Independence
 - Bayes' Theorem
3. Common probability distributions
 - Binomial, Poisson, Geometric, Exponential, Normal
4. Concept of likelihood
5. Statistical inference
 - Estimation (maximum likelihood)
 - Hypothesis testing
 - Modelisation

Questions?

nathalie.malo@gmail.com