

Module Overview

Refresher:

Genetic Data and Probability

Summer Institute in Statistical Genetics 2014

Module 10

Topic 1

Topics

- 1 Introductory Material: genetic terminology, probability, Mendelian genetics (Kerr)
- 2 Allele frequencies and Hardy-Weinberg Equilibrium (Kerr)
- 3 Introduction to Linkage and Linkage disequilibrium (Kerr)
- 4 Case-Control Association (Thornton)
- 5 Quantitative traits and heritability (Thornton)
- 6 Identity by Descent, Kinship Coef, Coef of Fraternity (Kerr)
- 7 Testing with Related Samples (Thornton)
- 8 Estimating Relatedness (Thornton)
- 9 Population structure (Thornton)
- 10 Methods for Cryptic structure (Thornton)

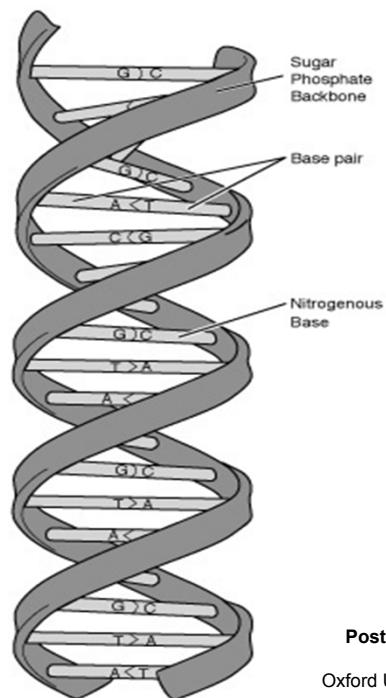
Software for Genetic Studies: R, Haploview, Locuszoom, PLINK

Instructors

Kathleen Kerr
Tim Thornton

katiek@u.washington.edu
tathornt@u.washington.edu

3

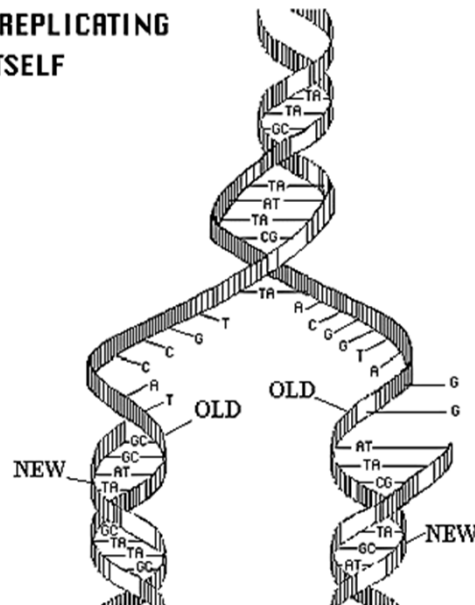


...ATGATGATCCAGC...
...TACTACTAGGTCG...

Post-genome Informatics
Minoru Kanehisa,
Oxford University Press, 2000

4

DNA REPLICATING ITSELF

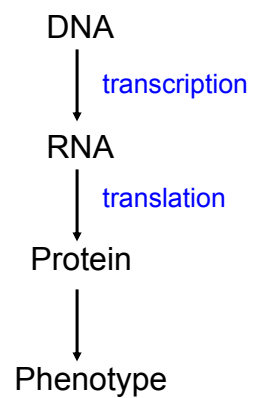


5

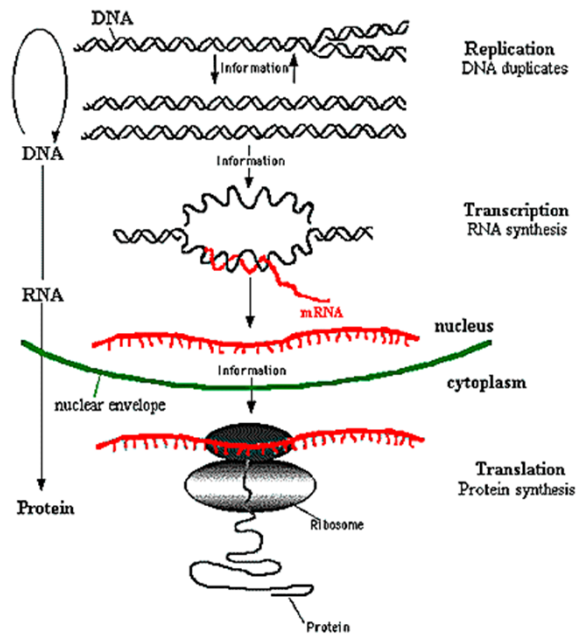
Central Dogma

“Genes are perpetuated as sequences of nucleic acid, but function by being expressed in the form of proteins.”

(Lewin, GENES VII, page 31)



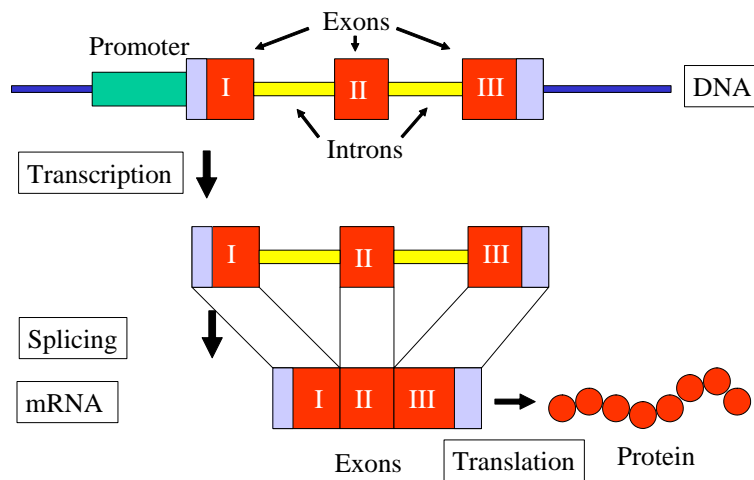
6



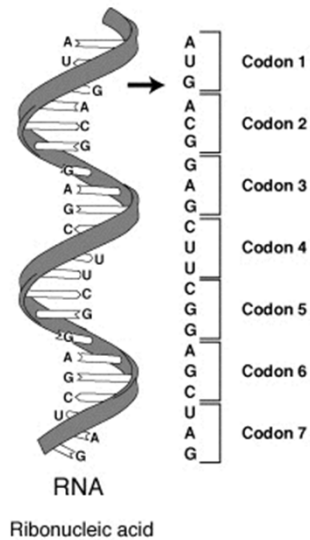
The Central Dogma of Molecular Biology

Post-genome Informatics
 Minoru Kanehisa,
 Oxford University Press, 2000

7



8



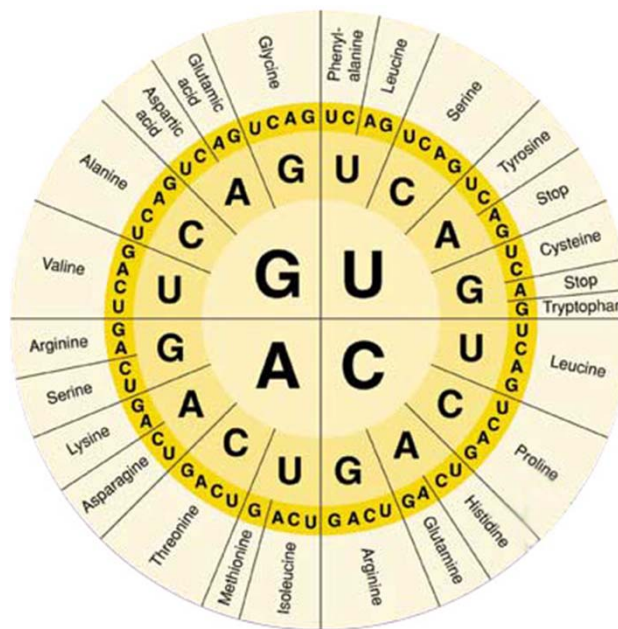
The genetic code defines a mapping between 3-tuples of nucleotides (called **codons**) and the 20 amino acids that comprise proteins.

There are also “start” and “stop” codons.

The genetic code has redundancy but no ambiguity. For example, although codons GAA and GAG both specify glutamic acid (redundancy), neither of them specifies any other amino acid (no ambiguity).

9

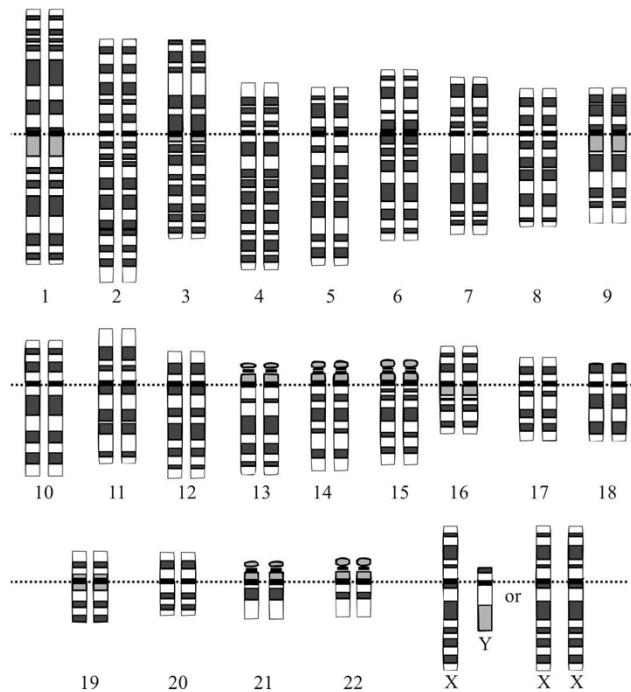
The Genetic Code



Some Terminology

- DNA is organized into **chromosomes**. Humans have 22 pairs of **autosomes** plus the sex chromosomes. Mammalian cells are **diploid**, meaning that chromosomes come in pairs.
- Chromosomes have two “arms” connected by the **centromere**. The shorter of the two arms extending from the centromere is called **p**. The the longer is called **q**.

11



12

More Terminology

- A specific location on a chromosome, for instance the location of a gene, a **SNP** (single-nucleotide polymorphism), or another genetic marker, is a **locus** (plural: **loci**). There can be more than one form of a locus. These forms are called **alleles**. When there is more than one allele at a locus, the locus is said to be **polymorphic**.

13

Still More Terminology

- **Mitosis** is cell division that yields two identical **diploid** cells, which have two of each chromosome. **Meiosis** is a special type of cell division that happens in reproductive tissue yielding **haploid** cells, which have one of each chromosome, called **gametes**. In females, the gametes are the egg cells and in males the gametes are the sperm cells.

14

and Some More Terminology

- When two haploid gametes unite, the complete diploid number of chromosomes is reinstated. In sexual reproduction, each individual has one chromosome of **maternal** origin and one chromosome of **paternal** origin. Thus at any locus an individual has one allele of maternal origin and one allele of paternal origin. These define the individual's **genotype** at that locus. If an individual has two copies of the same allele, then that individual is **homozygous** at that locus. If an individual has two different alleles at a locus, then s/he is **heterozygous**.

15

Genetics Data Explosion

- Sequence data
 - Human Genome Project
 - 1000s genomes project
 - Mouse Genome Project
 - etc. etc.
- Genotype data
 - HapMap Project
 - SNP chips
 - Copy Number variation

16

Data Explosion

- Gene Expression data
 - Microarray gene expression data
 - RNA-seq data
- Proteomics data
- Metabalomics data

17

The Foundation of Statistics:
Probability

Probability

Probability provides the language of data analysis. We should distinguish between a precise definition of probability, e.g.

- **Equiprobable outcomes definition**
 - Probability of E is number of outcomes in E divided by the total number of equally likely outcomes, e.g. probability of a head = $1/2$
- **Long-run frequency definition**
 - If E occurs n times in N identical experiments, the probability of E is the limit of n/N as N goes to infinity

...and a subjective probability (“I’m 90% sure I turned off the lights”)

19

Conditional Probabilities

All probabilities are conditional – probabilities depend on the context.

If E is an event or proposition of interest, and I is information or data, then $P(E|I)$ is the “probability of E conditional on I” or “probability of E given I.”

20

Conditional Probability Example

Roll a fair die.

E =get a 4

I =get an even number

$P(E \mid \text{roll a fair die}) = 1/6$

$P(E \mid \text{roll a fair die and } I) = 1/3$

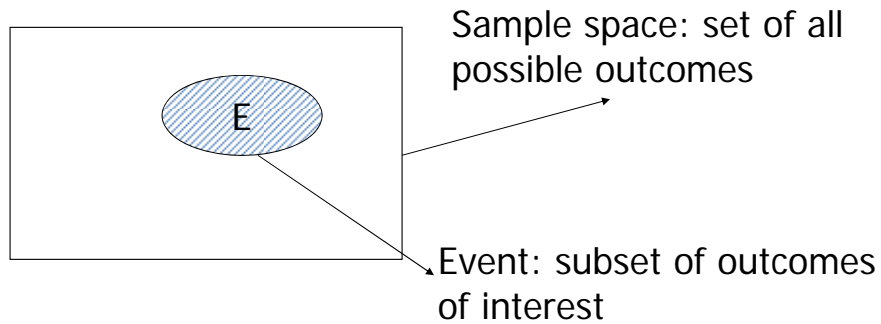
21

Comment: Conditional Probabilities and Hypothesis Testing

- A p-value results from statistical hypothesis testing
- Loosely, $p\text{-value} \approx P(\text{data} \mid \text{null hypothesis})$
- Incorrectly, a p-value is sometimes interpreted as $P(\text{null hypothesis} \mid \text{data})$

22

Rules of Probability



Probability of event E:

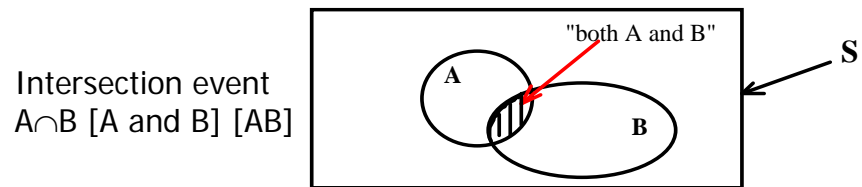
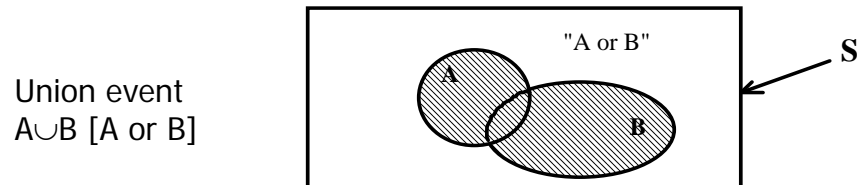
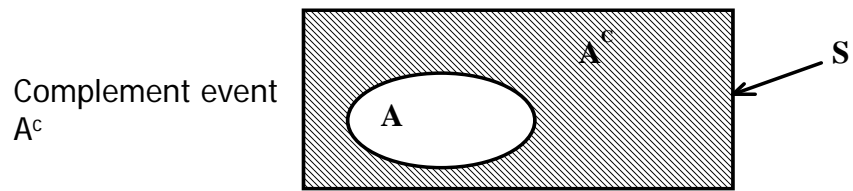
$$0 \leq P(E) \leq 1$$

23

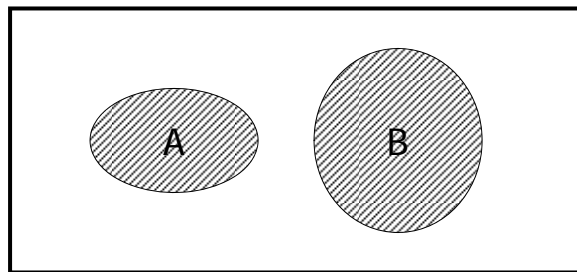
Example: Events

- A father's blood type is A and a mother's blood type is B.
- Event: "child has blood type A"
- Event: "child has an 'O' allele"

24



25



Disjoint or Mutually Exclusive events
[$A \cap B = \emptyset$ ("empty set")]

Example: A father's blood type is A and a mother's blood type is B.
A="first child has blood type A"
B="first child has blood type B"

Example: A and A^c are always mutually exclusive events

26

Rules of Probability

- Let A and B denote events and S denote the sample space

- Rule 1: $0 \leq P(A) \leq 1$
- Rule 2: $P(S) = 1$
- Rule 3: Addition rule for disjoint events

$$P(A \text{ or } B) = P(A) + P(B)$$

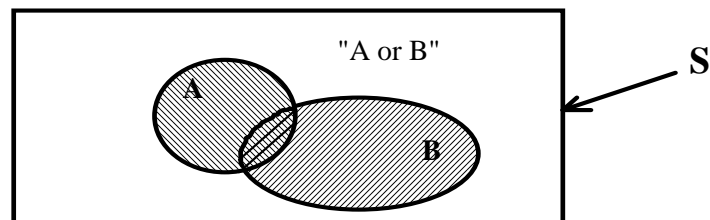
- Rule 4: Complement rule

$$P(A^c) = 1 - P(A)$$

27

Rules of Probability

- General addition rule:** What is $P(A \text{ or } B)$?



$$P(A \text{ or } B) = \frac{\#(A \text{ or } B)}{\#S} = \frac{(\#A) + (\#B) - (\#AB)}{\#S}$$

$$= P(A) + P(B) - P(AB)$$

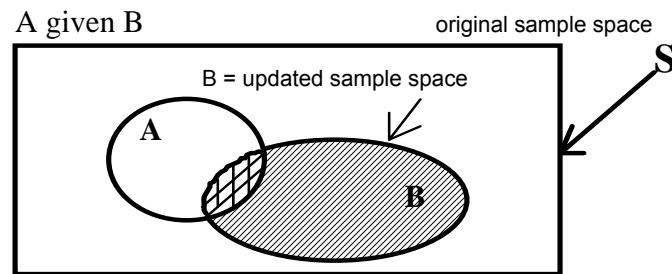
If events are disjoint, then $P(AB) = 0$ and

$$P(A \text{ or } B) = P(A) + P(B)$$

28

Rules of Probability

- **Conditional probability:** What is $P(A|B)$?
“the probability of A given B”



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

29

Rules of Probability

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

Equivalently: $P(A \cap B) = P(A|B) P(B)$

30

Rules of Probability

Definition: Two events A and B are independent if

$$P(A|B)=P(A) \text{ or, equivalently, if } P(B|A)=P(B) .$$

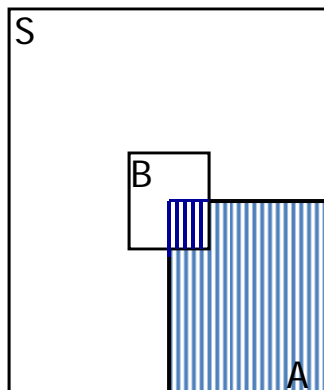
Interpretation:

If two events do not influence each other (that is, if knowing one has occurred does not change the probability of the other occurring), the events are independent.

Using the multiplicative rule, if two events A and B are independent, then

$$P(AB) = P(A) P(B).$$

31



Independence is harder to visualize

Independence means that $P(A|B)=P(A)$

[area of A in relation to S is the same as the area of AB in relation to B]

32

Caution

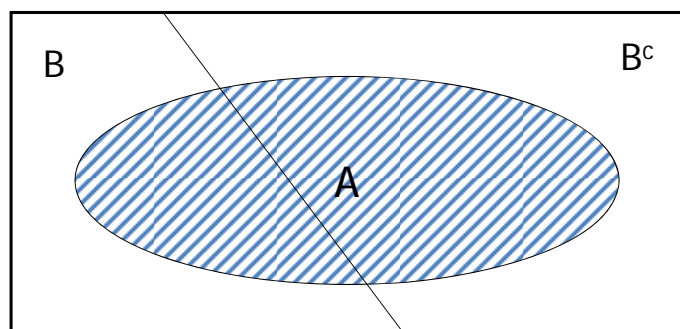


- “Independent” and “mutually exclusive” sound similar in everyday language. But in probability they mean VERY different things.

33

Rules of Probability

- Partitioning



$$A = (A \text{ and } B) \text{ or } (A \text{ and } B^c)$$

→ Disjoint!

$$P(A) = P(A | B)P(B) + P(A | B^c)P(B^c)$$

34

Rules of Probability

- Bayes' theorem/Bayes' rule

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A | B)P(B)}{P(A | B)P(B) + P(A | \text{not } B)P(\text{not } B)}$$

Definition of conditional probability

Partitioning

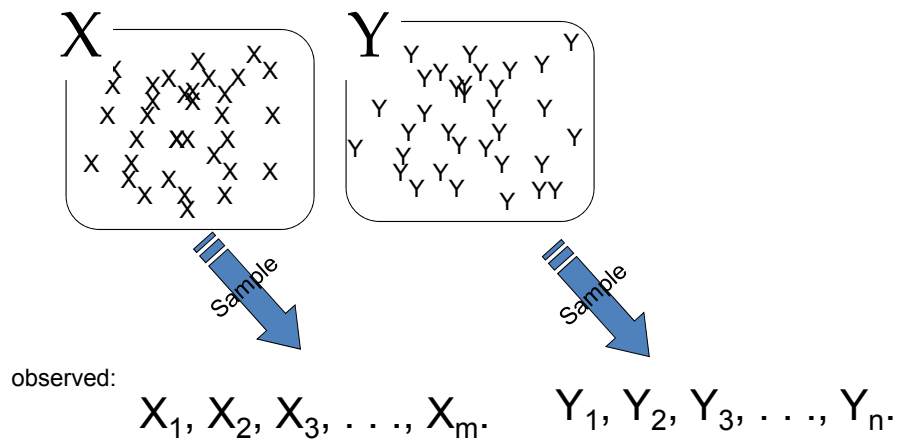
35

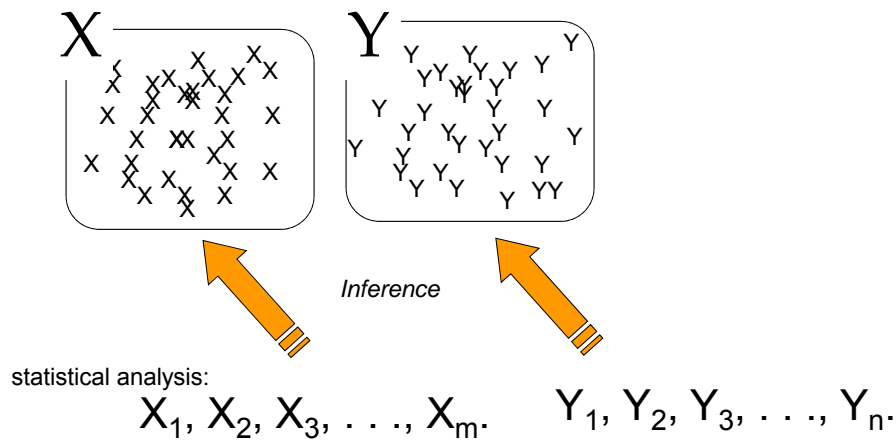
- Bayes' rule... not just for Bayesians!
- “The term ‘controversial theorem’ sounds like an oxymoron, but Bayes' theorem has played this part for two-and-a-half centuries. Twice it has soared to scientific celebrity, twice it has crashed, and it is currently enjoying another boom.” -Bradley Efron, *Science* 2013

36

The Paradigm of Statistics: Statistical Inference

Basic Statistics: Two Populations to Compare (unobserved)
For example, we may wish to compare smokers to non-smokers; or wild-type individuals with mutant genotypes





39

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

— Sir Ronald A. Fisher
 1890-1962

40

Elements of Statistical Inference

- Estimator (point estimates)
- Confidence Intervals
- Hypothesis Testing

41

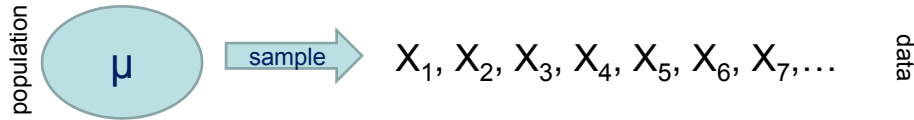
Estimator

- In statistics, we use data on a sample to estimate population parameters. For example, for a specific genetic variant, a population parameter is the frequency of that variant in the population.
- We like estimators that are:
 - unbiased: expected value is the parameter
 - consistent: increasing accuracy as sample size increases
 - efficient: small variance

42

Statistical Estimates

- Point estimates: We like point estimates that have low bias, are consistent, make efficient use of the data



Estimate of μ , the population mean	Unbiased?	Consistent?	Efficient?
median	Biased*	No*	NA
X_1	unbiased	No	No
X_{odd}	unbiased	Yes	No
\bar{X}	unbiased	Yes	Yes

43

Components of Inference

- Confidence intervals: The set of parameter values for which the observed data are “typical” / the set of parameter values for which the observed data are “consistent”
- Example: I test a drug on 100 mice with a particular disease. 79% of the mice are cured. I estimate that the drug has a cure rate of 0.79 with 95% confidence interval 0.70 to 0.86.

44

Components of Inference

- p-values: What is the probability of observing data like our sample when there is nothing going on?

- “Nothing going on” = null hypothesis
- $p\text{-value} \approx P(\text{data} \mid \text{null hypothesis})$

45

Hypothesis Tests and P-values

- Null hypothesis H_0 and Alternative hypothesis H_A .
 - E.g., H_0 : SNP is not associated with height.
 - H_A : SNP is associated with height.
 - (Hypotheses are statements about nature or about the population, not about the sample or about data)
- Test statistic T
- Statistical theory tells us the sampling distribution of T if the null hypothesis is true
- Sometimes, we do not know the distribution of T under H_0 and try to approximate it using the data (permutations, etc)
- The result of a Hypothesis Test is “reject H_0 ” or “fail to reject H_0 ”

46

Example, continued: an existing treatment for a disease has a well-established cure rate of 0.75. Do I have compelling evidence that my treatment is better?

H_0 : cure rate=0.75

H_A : cure rate \neq 0.75

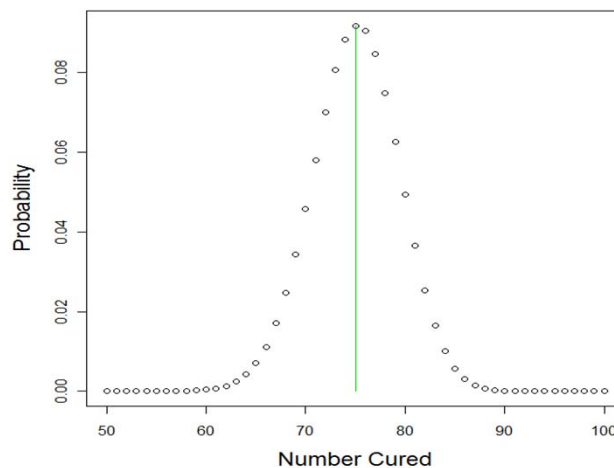
If the true cure rate is 0.75 and I observe 100 treated individuals, then the number cured has a Binomial distribution 100 “trials” and probability 0.75.

p-value=P(observe 79 or more extreme | H_0)

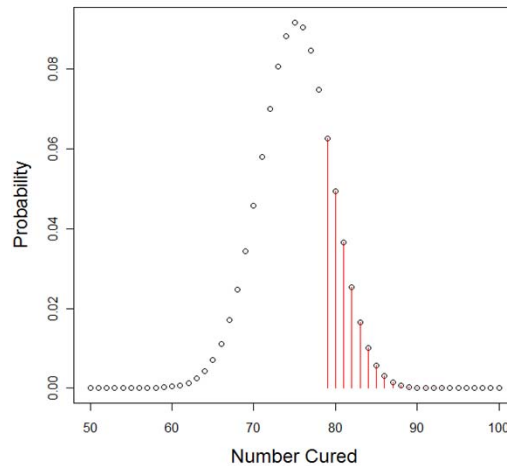
47

Binomial distribution for n=100 “trials” with probability 0.75 of “success”

For example, the probability of 75 successes is about 0.092.



The one-sided p-value is the probability of getting a result as extreme or more extreme than the observed data, 79. In this case the p-value ≈ 0.21 .



49

Hypothesis Testing: errors

- When we perform a hypothesis test, there are two ways we can make a mistake.

- Type I error:** Reject the null hypothesis when it is true. We usually conduct a test in a way that ensures that this type of error is not too large. A popular threshold is 5%. In this case the “**size**” or “ **α -level**” of a hypothesis test is 0.05 and we reject the null hypothesis if the p-value < 0.05 .

- Type II error:** Fail to reject the null hypothesis when it is false.

50

- Next, we “warm up” by applying the basic rules of probability to some early experiments in genetics.

51

Module Warm-Up: Mendelian Genetics

Mendel's Laws

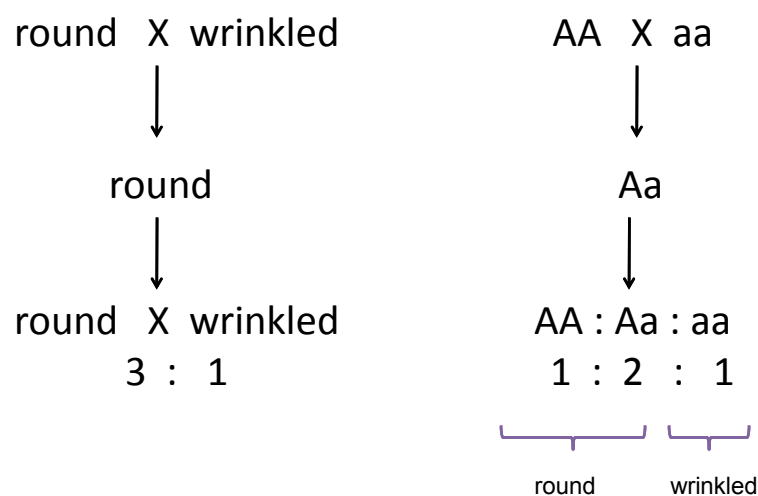
Mendel's first law: each character is controlled by a pair of genes. Two genes segregate into different gametes during meiosis.

Mendel's second law: When two or more pairs of genes segregate, they do so independently.

Classical Mendelian experiments use inbred strains of animals or self-fertilized plants so that individuals in the starting generation are homozygous at every locus and genetically identical.

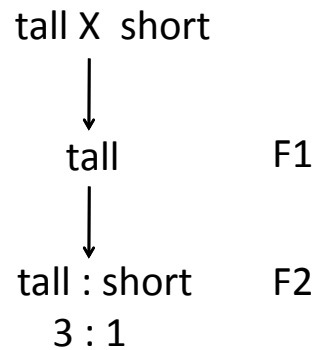
53

Example 1: Peas experiment



54

Example 2 (Fisher): Plant stem length



F2 generation: AA : Aa : aa

55

Example 2

Question: How to distinguish the AA and Aa genotypes?

Fisher's answer: grow 10 offspring from each tall F2 plant. If all offspring are tall, classify as AA; otherwise classify as Aa.

$$P(\text{an offspring is tall} \mid Aa) = \frac{3}{4}$$

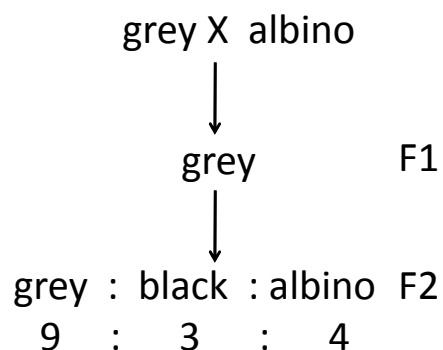
$$P(\text{all 10 offspring are tall} \mid Aa) = \left(\frac{3}{4}\right)^{10} \approx 0.056$$

Therefore, about 5% of Aa plants in F2 will be misclassified as AA.

Among tall plants in the F2 generation, the expected ratio of AA:Aa is 1:2. However, among those labeled "AA" and "Aa," the expected ratio is about 37:63

56

Example 3: Rabbits



F2 generation: AA : Aa : aa – single gene model with two alleles
cannot explain these data

57

Example 3: Rabbits

Color is controlled by two genes
1. Presence of color C, no color c
2. Grey G or black b

CC or Cc → GG or Gb grey rabbit
bb black rabbit

cc → albino rabbit regardless of other locus

58

Example 3: Rabbits

CCGG X ccbb



CcGb



F ₂	CG	Cb	cG	cb
CG	grey	grey	grey	grey
Cb	grey	black	grey	black
cG	grey	grey	albino	albino
cb	grey	black	albino	albino

59

Example 4: Mice

grey X chocolate



grey

F₁



grey : black : chocolate F₂
12 : 3 : 1

60

Example 4: Mice

Genetic Model

1. Grey allele G dominates allele g
2. Black allele B dominates chocolate allele b

GG or Gg	→	grey mouse
gg and (BB or Bb)	→	black mouse
gg and bb	→	chocolate mouse

61

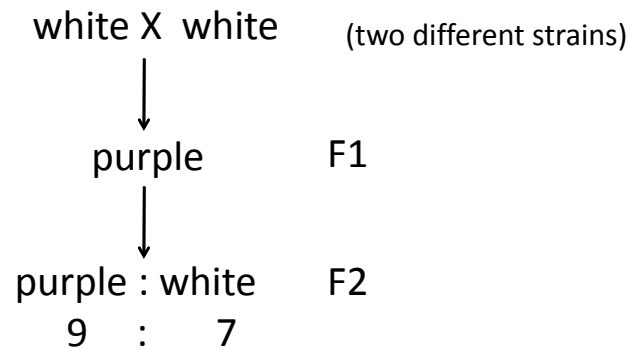
Example 4: Exercise

Show how this genetic model explains the observed phenotype relative frequencies.

F ₂	GB	Gb	gB	gb
GB	grey	grey	grey	grey
Gb	grey	grey	grey	grey
gB	grey	grey	black	black
gb	grey	grey	black	chocolate

62

Example 5: Sweet Pea Flower Color



Genetic Model: A dominates a
B dominates b
Need at least one A and one B to have purple; otherwise, white

63

Example 5: Sweet Pea Flower Color

AAbb X aaBB

↓

AaBb

↓

F ₂	AB	Ab	aB	ab
AB	purple	purple	purple	purple
Ab	purple		purple	
aB	purple	purple		
ab	purple			

64

Example 6: Bean Flower Color

Flowers come in shades from white to purple.
Quantify color: white (0) to purple (10)

10 X 0
purple X white

↓

5

↓

color	10	9	8	7	6	5	4	3	2	1	0
relative count	1	0	2	2	1	4	1	2	2	0	1

Genetic Model: Two genes with additive effects:

Gene 1: A=3, a=0

Gene 2: B=2, b=0

65

Example 6: Bean Flower Color

AABB X aabb

↓

AaBb

↓

F ₂	AB	Ab	aB	ab
AB	10	8	7	5
Ab	8	6	5	3
aB	7	5	4	2
ab	5	3	2	0

66

Example 7: Human Blood Groups

Four phenotypes in the ABO blood group: A, B, AB and O

Two loci theory (Von Dungen & Hirszfeld 1910, Ottenberg 1923).

Locus 1 has A dominant to a

Locus 2 has B dominant to b

phenotype	genotype
O	aabb
A	AAbb or Aabb
B	aaBB or aaBb
AB	AABB or AaBB or AABb or AaBb

67

Example 7: Human Blood Groups

Four phenotypes in the ABO blood group: A, B, AB and O

One locus theory

One locus with three alleles A, B, and O

phenotype	genotype
O	OO
A	AA or AO
B	BB or BO
AB	AB

68

Example 7: Exercise

What are the possible blood types among offspring from each type of mating?

Parental Blood Types	Two loci model	Single locus model
O x O	O	O
O x A	O, A	O, A
O x B	O, B	O, B
A x A	O, A	O, A
A x B	O, A, B, AB	O, A, B, AB
B x B	O, B	O, B
O x AB	(O), A, B, (AB)	A, B
A x AB	(O), A, B, AB	A, B, AB
B x AB	(O), A, B, AB	A, B, AB
AB x AB	(O), A, B, AB	A, B, AB

Parentheses indicate predictions that do not appear in actual data.