Basic probability and statistics for geneticists

Timothy Mak

Postdoctoral Fellow

Centre for Genomic Sciences

Aims

By the end of the lecture, you should know about the following:

- What is a probability
- What is a statistic
- What is a random variable (r.v.)
- probability density/mass function (pdf/pmf)
- Expectation and variance
- Mixture distributions
- Common notations for probability and statistics
- Principles of statistical testing
- Principles of estimation
- Principles of statistical modeling

Goals

- The main goal is to help you read papers with statistical content.
- A better understanding of statistics and probabilities should also help us avoid misunderstanding of tests and results from statistical analyses.

Some notations

x, y, z: variables – what you encounter in high school, e.g. $x + 2 = 5 \Rightarrow x = 3$

X, Y, Z: capital letters are usually used to denote random variables. I'll introduce these later

f(x, y, z), g(x), …: f() and g() are commonly used to denote functions.

p(x), Pr(x), f(x): These are often used to denote probabilities and densities.

PROBABILITY

What is a probability?

- Probability that a dice shows up 6 on a throw
- Probability that the statue of liberty is over 70m tall
- Probability that a mutation in X is causal of disease D.
- Difference between a probability and a proportion?
- Is probability subjective?

Frequentist vs. Bayesian view of probability

Frequentists: Probability is NOT subjective

Bayesians: Probability IS subjective

(although it doesn't mean that Bayesian analyses always involve subjective probabilities)

Variables vs. Random variables (R.V.)

Variables:

Let x denote the genotype of Peter: x = AA| x = AT| x = TT

- Fixed value, although possibly unknown.

Random variables:

Let X denote the genotype of a randomly selected person

X = AA with probability (w.p.) 0.01

X = AT w.p. 0.18

X = TT w.p. 0.81

- Can take more than 1 different values at the same time.
- Associated with a probability

Continuous vs. discrete R.V.

Discrete (random) variables:

- Diseased/non-disease
- Genotype (A/C, G/T)
- Sex
- Scores on a Likert-type scale: 1, 2, 3, 4, 5
- Number of siblings

Continuous (random) variables:

- height
- time taken to perform a task

Discrete variables – some notations

Let X be a genotype for a randomly selected person for the SNP rs123456.

Probabilities associated with the genotypes are:

- Pr(X=AA) = 0.04
- Pr(X=AG) = 0.32
- Pr(X=GG) = 0.64

Often shortened to P(AG), P(AA), etc., or:

$$p(x) = P(X = x) = \begin{cases} 0.04 \text{ if } x = AA \\ 0.32 \text{ if } x = AG \\ 0.64 \text{ if } x = GG \end{cases}$$

p(x) is then called a probability mass function or pmf.

Continuous variables

Pr(Y = any number) = 0

e.g.

Pr(Height = 120) = 0

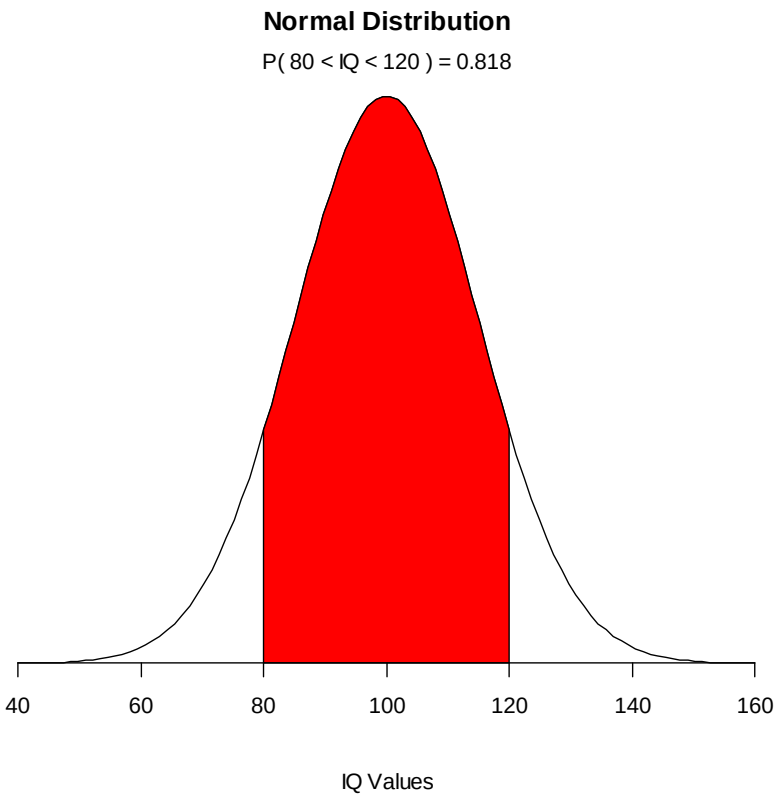
However,

Pr(120 < Height < 120.0001) = 0.0001

i.e. a range must be given, however small it is

Continuous variables

- We represent a continuous distribution by a probability density function (pdf).
- Probability = area under the pdf function
- A density can go beyond 1, therefore it's not a probability!
- But we still use the same notation for them as pmfs, e.g. f(x), p(x), ...



Some common distributions

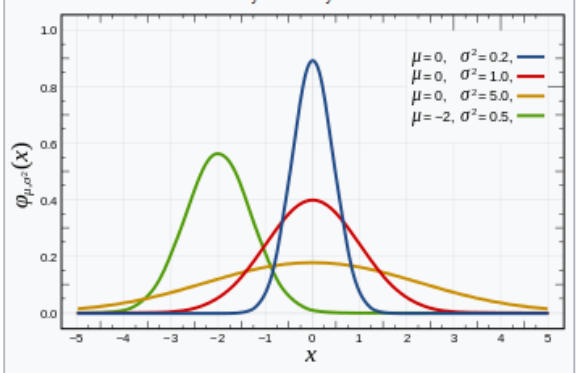
Normal (Gaussian) distribution:

Examples:

- Height
- IQ

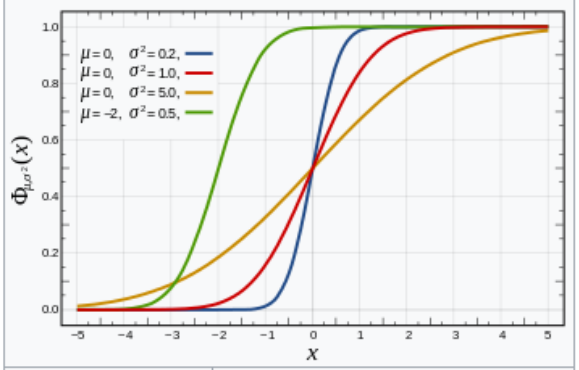
Normal distribution

Probability density function



The red curve is the *standard normal distribution*

Cumulative distribution function



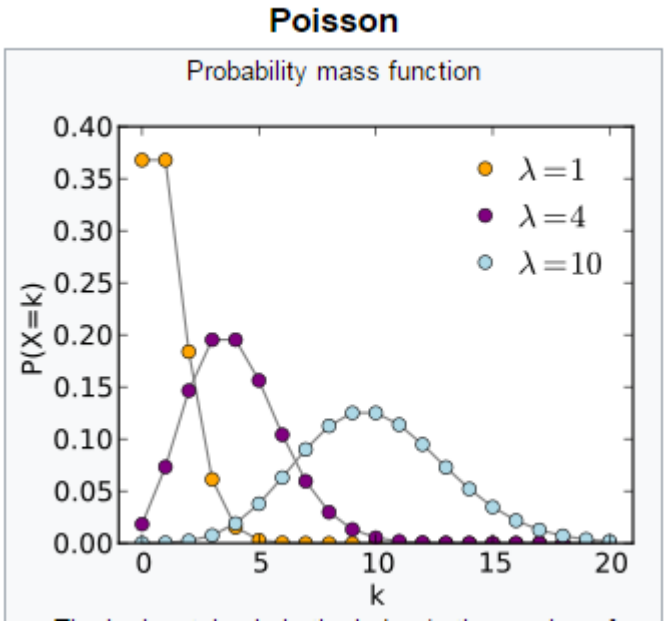
Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbf{R}$
PDF	$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Some common distributions

Poisson distribution

Examples:

- Number. of times something "random" happens in a time interval
- However, often used for any kind of count data, e.g. number of disease episodes, number of siblings... (more on this later)



of k and flat everywhere else because a variable that is Poisson distributed takes on only integer values.

Parameters	$\lambda > 0$ (real)
Support	$k \in \mathbb{N} \cup 0;$
pmf	$\frac{\lambda^k e^{-\lambda}}{k!}$
CDF	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$, or

Some common distributions

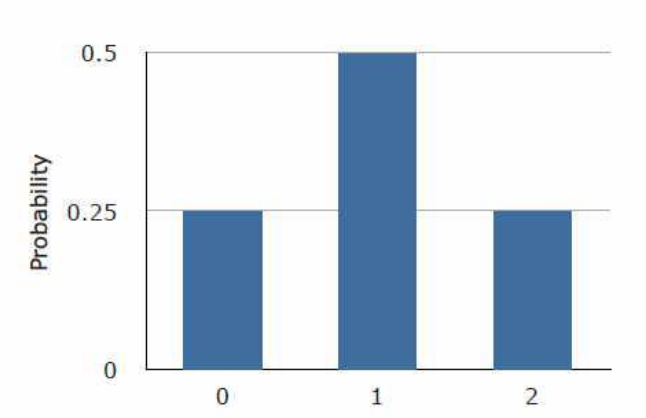
Bernoulli distribution

Examples:

- Disease/non-disease status
- Yes/no answers
- haplotype in a bi-allelic locus

Some common distributions

Binomial distribution



http://onlinestatbook.com/2/probability/graphics/binomial1.jpg

Examples:

- Genotypes in a bi-allelic locus, under Hardy-Weinberg Equilibrium

Notation	$B(n, p)$
Parameters	$n \in \mathbb{N}_0$ — number of trials $p \in [0,1]$ — success probability in each trial
Support	$k \in \{0, \ldots, n\}$ — number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$

Quiz

Let's say the SNP rs123456 has two alleles (A or T).

Frequency (proportion) of A in the population is 0.1.

Let X be a randomly sampled allele from the population

- What distribution does X follow?
- Pr(X = A) = ?

Let's say SNP rs123456 is in Hardy Weinberg Equilibrium.

Let G denote the genotype of a randomly sampled person.

- What distribution does G follow?
- Pr(G = AA) = ?
- Pr(G = AT) = ?

Other distributions

- Uniform (p-values under the null)
- Gamma
- Chi-squared
- t (or Student's t)
- F
- Negative binomial
- logNormal
- and many more…

Use of distributions in tests

- Often a particular test/procedure assumes your data follows a particular distribution
- But it’s often the case your data doesn’t follow any

What to do?

What to do?

- Know which variable the assumption applies to
- Sometimes it helps to transform the data, e.g. by taking logs.
- Sometimes we choose a non-parametric test instead. (Mann-Whitney U test)
- Sometimes we make it a permutation test, or apply bootstrapping
- Sometimes it doesn’t matter a lot!

Mixture distributions

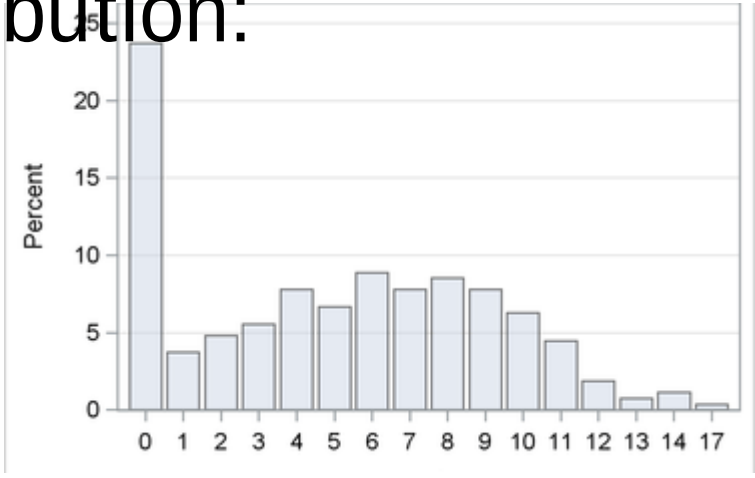
Many more distributions can be formed by "mixing" common distributions.

e.g.

Number of RNA transcripts for a gene: Many zeros.

A zero-inflated Poisson distribution:

$$X \sim \begin{cases} 0 & \text{w.p. } p \\ \text{Poisson}(\mu) & \text{w.p. } 1 - p \end{cases}$$



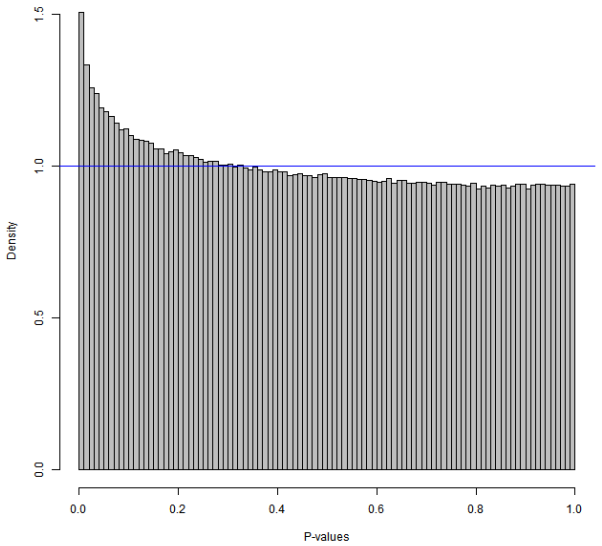
https://support.sas.com/rnd/app/stat/examples/GENMODZIP/images/roots.png

Mixture distributions

Another example:

Distribution of p-values in GWAS

$$p \sim \begin{cases} \text{Uniform}(0,1) & \text{w.p. } p & \text{(Null distribution)} \\ f() & \text{w.p. } 1 - p & \text{(Non-null distribution)} \end{cases}$$



http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/histogram.png

Mixture distributions

Another example:

The negative-Binomial distribution

$X \sim \text{Poisson}(\mu)$

$\mu \sim \text{Gamma}(\alpha, \beta)$

Expectation and variance

Expectation = Mean:

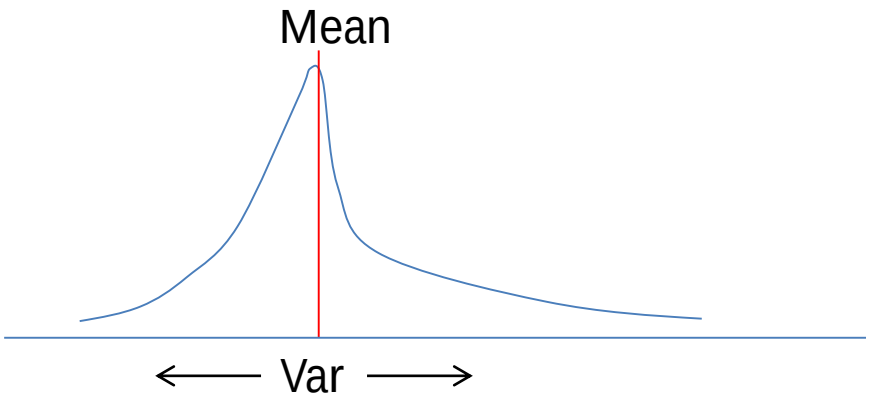
often denoted μ

$$E(X) = \int x\, f(x)\, dx$$

Variance:

often denoted σ^2

$$Var(X) = E(X - E(X))^2 = \int (x - E(X))^2\, f(x)\, dx$$



These are NOT random

Other distribution-related properties

- Median
- Mode
- Skewness
- Kurtosis
- Correlation (for 2 variables)
- Covariance (for 2 variables)

STATISTICS

What is a “statistic”

statistic = f(data)

E.g. our data = {0, 1.2, 2.5, 2.8, 5}
Examples of statistics:

- The sum = 11.5
- The mean = sum/5 = 2.3
- The standard deviation (SD) = 1.876
- The standard error (SE) = SD/√5 = 0.84
- t = mean/SE = 2.74
- The z-score = mean / SD = 1.22

What's the connection between "statistics" and "probability"?

In their simplest forms, statistics are simply deterministic functions of data and not probabilistic.

However, if we regard the data as "random", then because

$$\text{statistic} = f(\text{data})$$

statistics become random.

How is it that data are “random”?

Several paradigms:

1. Data is a random sample from a population
2. Data are results from an experiment influenced by other random, uncontrollable factors.
3. Some data are arbitrarily considered random because we want to use a “model” to represent the reality.

The distinction between 1 and 3, and 2 and 3 is not always clear.

An example

Relationship between gene X and disease D:

Paradigm 1: Random sampling

Among 10 people with genotype AB, 3 were observed with the disease. Proportion = 0.3.

Among 10 people with genotype AA, 2 were observed with the disease. Proportion = 0.2.

Within this sample, X is clearly associated with D.

However, we're not content with inference within the sample. We want to know whether there is an association in the bigger population also.

So we do some statistical calculations, and end up with a chi-squared statistic of $X^2(1)=0.27$, p-value = 0.61, and conclude that gene X is not associated with disease D in the bigger population.

An example

Relationship between gene X and disease D:

Paradigm 2: Experiment

Among 10 mice with gene X knocked out, 3 were observed with the disease. Proportion = 0.3.

Among 10 mice with gene X normal, 2 were observed with the disease. Proportion = 0.2.

The difference between this example and the last is that before, both X and D are random. They're both drawn randomly from the underlying popluation.

Here, we fixed X beforehand, and consider D to be a random result.

An example

Relationship between gene X and disease D:

Paradigm 3: Modeling

L = (Number of alleles in gene X) * β + E

D = 1 if L > 1.2. D = 0 if L < 1.2

Our goal is to estimate the parameter β, although we may also test whether β is significantly different from 0.

Statistical testing

- Deriving useful statistical tests is a complicated and difficult business. Not going to go into details here.
- However, I want to cover some principles of statistical testing.

Statistical testing

- The goal of test is to try and gain some understanding in the problem at hand. e.g. Does a mutation in gene X cause the observed phenotype?
- Ideally, we want to be able to say

  Pr(It does) = x.
- However, frequentists would not allow us to answer this question, because whether or not gene X cause the phenotype is not a random thing. (Only things like random sampling are truly “random”.)
- Bayesians would be able to answer it. However, it has its difficulties.

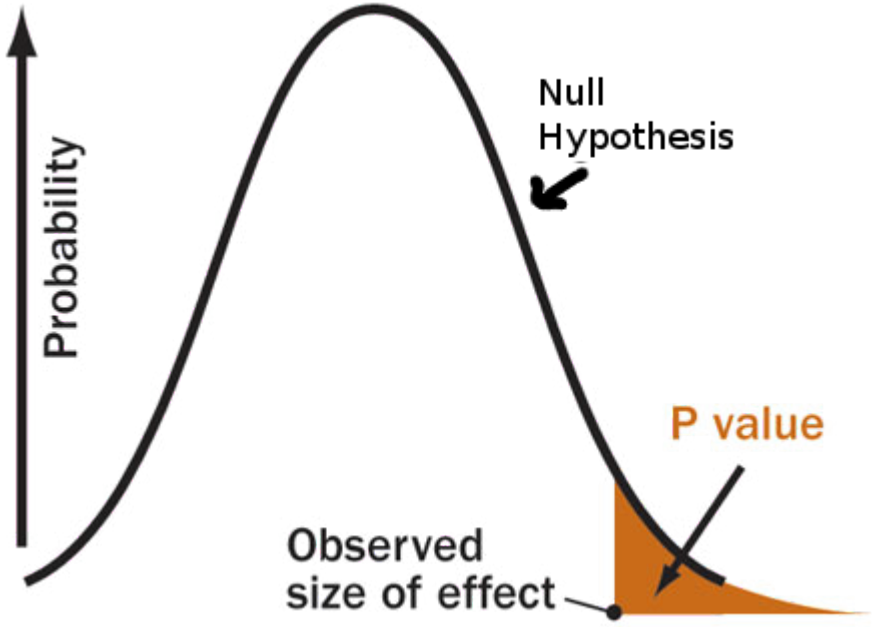
Statistical testing

- Statistical tests are largely developed by frequentists.
- Instead of answering Pr(X causes Y), we ask:

  If X does not cause Y, what is the probability of us observing data that is as extreme as what we've observed? (the p-value)
- How do we define "extreme"?

Statistical testing

- The idea is to try and find a statistic that will tend to have very extreme values if X does cause Y, and values that follow a known distribution if X does not cause Y.
- Looking up the statistic according to the known distribution gives us a p-value.



https://bayesianbiologist.files.wordpress.com/2011/08/p_value.png

Modeling

- Basically a set of equations linking together various variables, e.g.

$$Y \sim f(a,b)$$

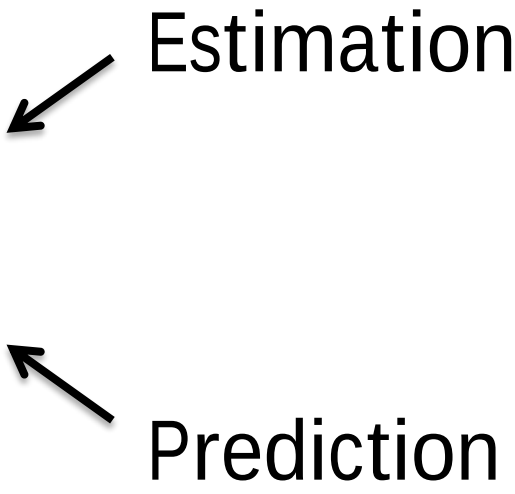
$$a = g(c,d,e)$$

$$c \sim h(X)$$

- The key is to know
  - which variables are random/fixed
  - which variables are known/unknown

4 types of variables

	Known	Unknown
Fixed	4	8
Random	4	8



Common notations

	Known	Unknown
Fixed	X, x	α,β, … (Parameters)
Random	Y, y	α,β, … Y, y …

Estimation

Prediction

Other clues:

- X ~ Distribution(…)
- Y follows a distribution of …

Bayesian thinking

	Known	Unknown
Fixed	X, x	
Random	Y, y	α,β, ... Y, y ...

Prediction

Estimation

2 scenarios

1. A sample from a population

   We know the sample mean, variance, Odds Ratios,

   But we want to estimate the population mean, variance, Odds ratio, etc.

2. Assume a model with random and fixed components, e.g.:

$$Y_i \sim \text{Binomial}(1, p_i)$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i$$

   Assume some of the random elements (Y_i) and/or fixed elements (x_i) are observed.

   Make “best guesses” for the unknown fixed components (β_0, β_1, p_i).

Estimation

Difference between:

- sample mean and population/true mean

$$\bar{x} = \sum x_i/n \text{ vs. } E(X)$$

- sample MAF and population MAF
- sample variance and population/true variance

$$\hat{\sigma}^2 = \sum (x_i - \bar{x})^2/n \text{ vs. } Var(X)$$

- Estimate vs. Parameter, e.g. $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$

$$\hat{\beta}_1 = f(x, y) \text{ vs. } \beta_1$$

Estimation

An estimate is random.

A population parameter is fixed.

An estimate has a distribution, a pdf, a mean, and variance.

However, these are often unknown, although we can estimate them (either through theory or bootstrapping).

The standard deviation (√ Variance) of an estimate is called the standard error.

Quiz: Is the standard error random or fixed?

Confidence intervals

- 95% confidence intervals often constructed as:

$$\hat{\beta} \pm 1.96 * \text{s.e.}$$

  but not always.
- In Normal distribution with mean 0 and variance 1, [-1.96, 1.96] covers 95% of the distribution.
- Note that the frequentist interpretation of CIs is quite awkward.

Caveats

- Interpretation correct only subject to assumptions of models being met.
- Some common pitfalls:

1. Sample not random (applies to paradigm 1)
2. Relevant (confounding) variables not taken into account of (applies to paradigm 2 and 3)
3. Cherry-picking of most significant results.

Conclusions

Statistical methods rarely give us definitive answers on the probability of events.

Consider all tests and results as clues to help us interpret data.