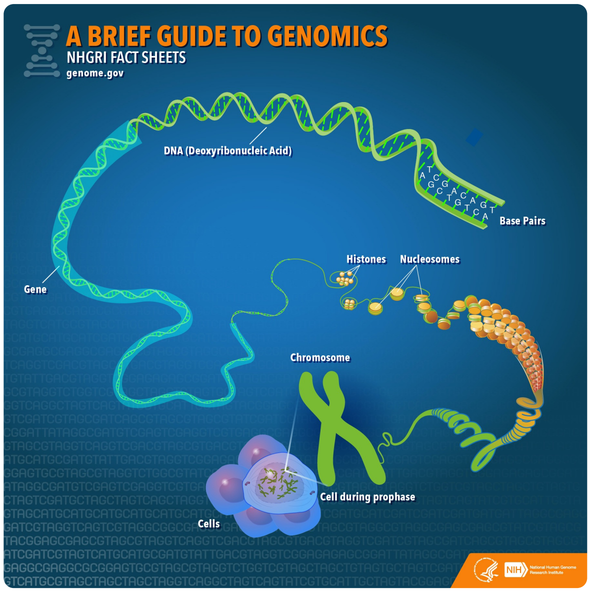
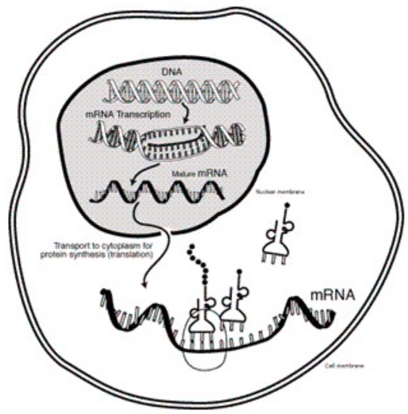
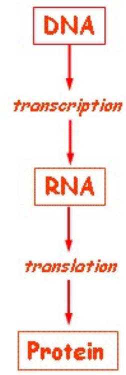
Week 1 QCB 408 / 508 Spring 2020



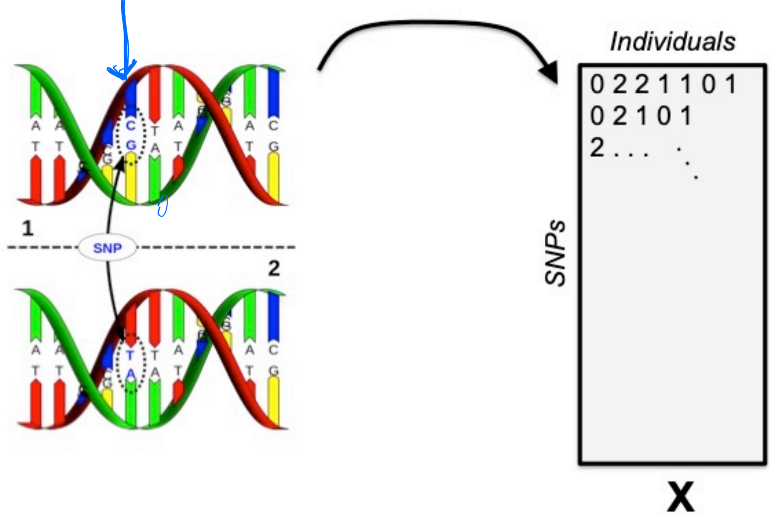




www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics
www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/central_dogma.html

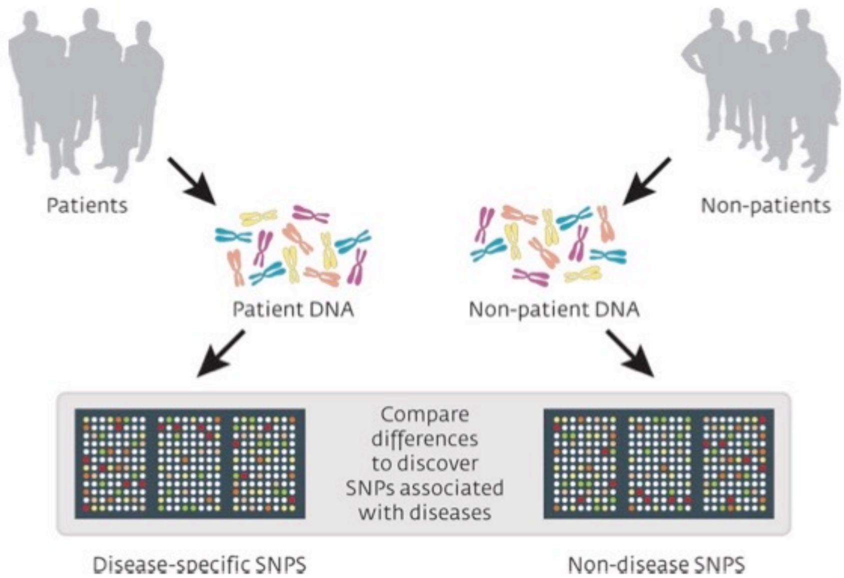
## SNP data

SNP

| Example: | Genotype | CC | CT | TT |
|---|---|---|---|---|
| | $x_{ij}$ | 0 | 1 | 2 |



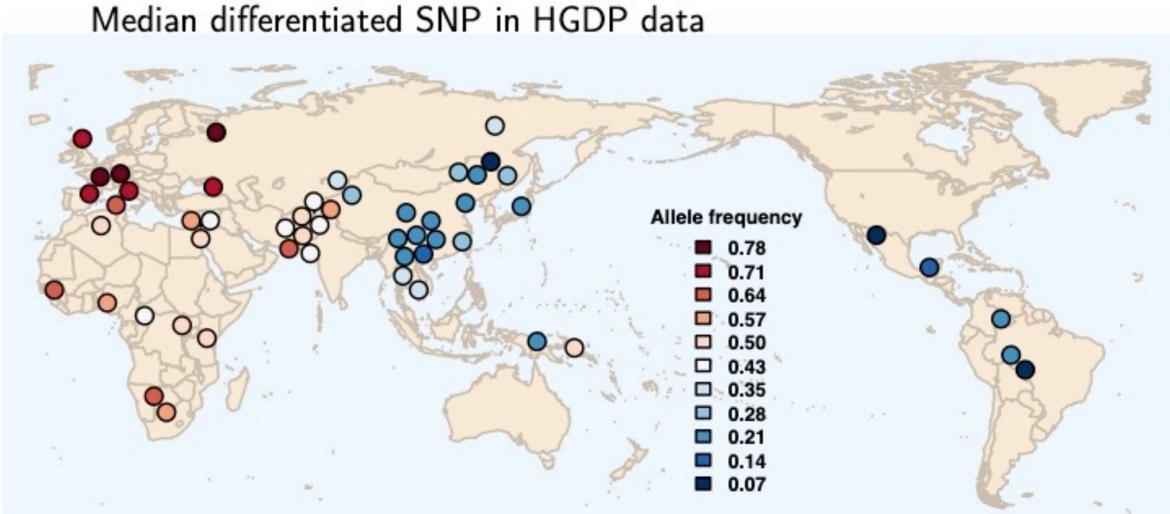
## Genome-wide association studies

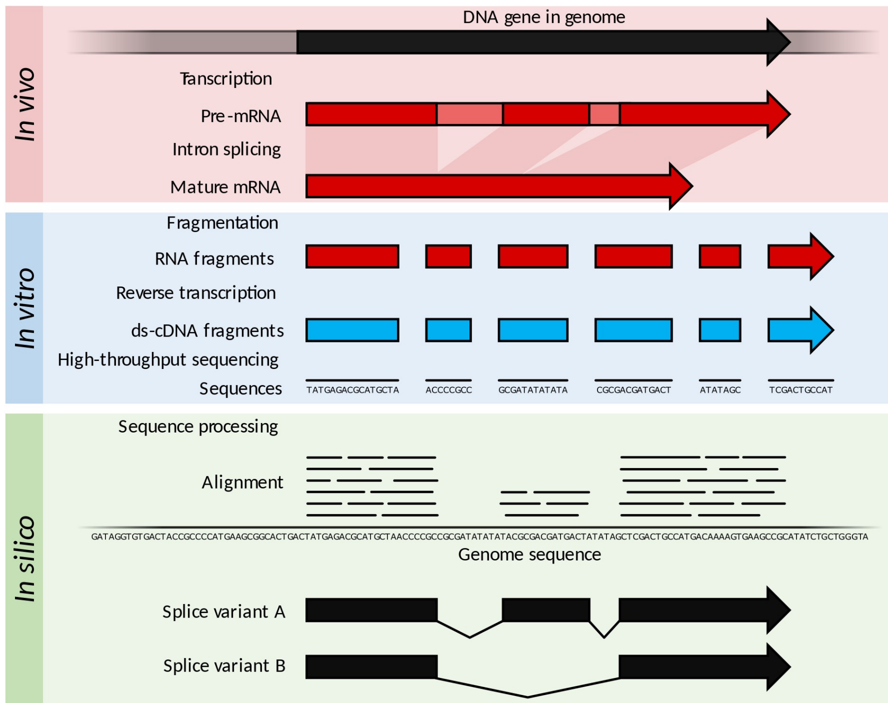


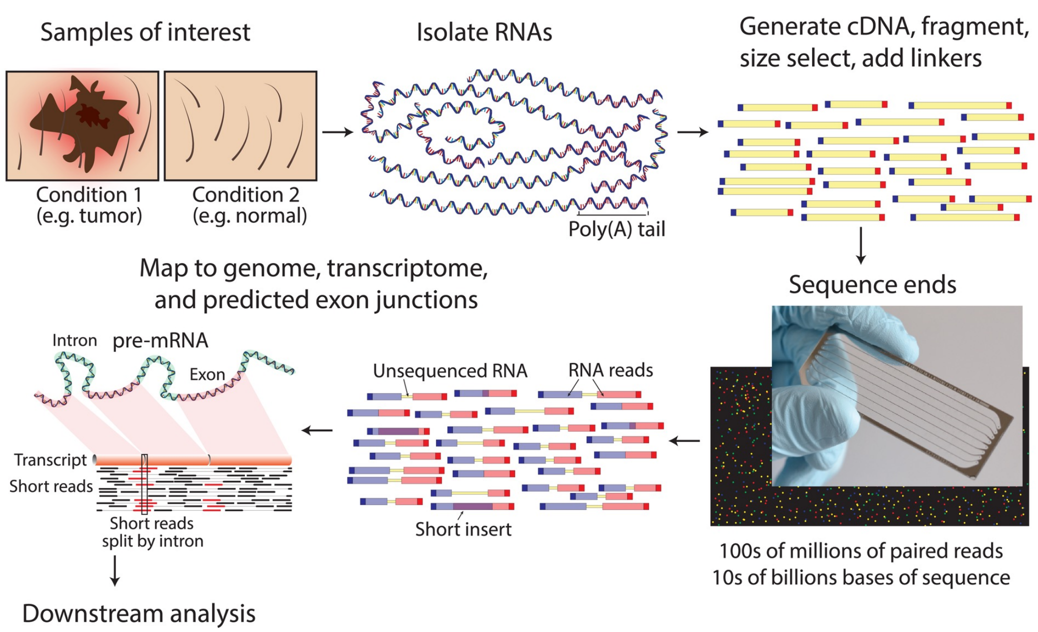



## Allele frequencies in human populations

Median differentiated SNP in HGDP data

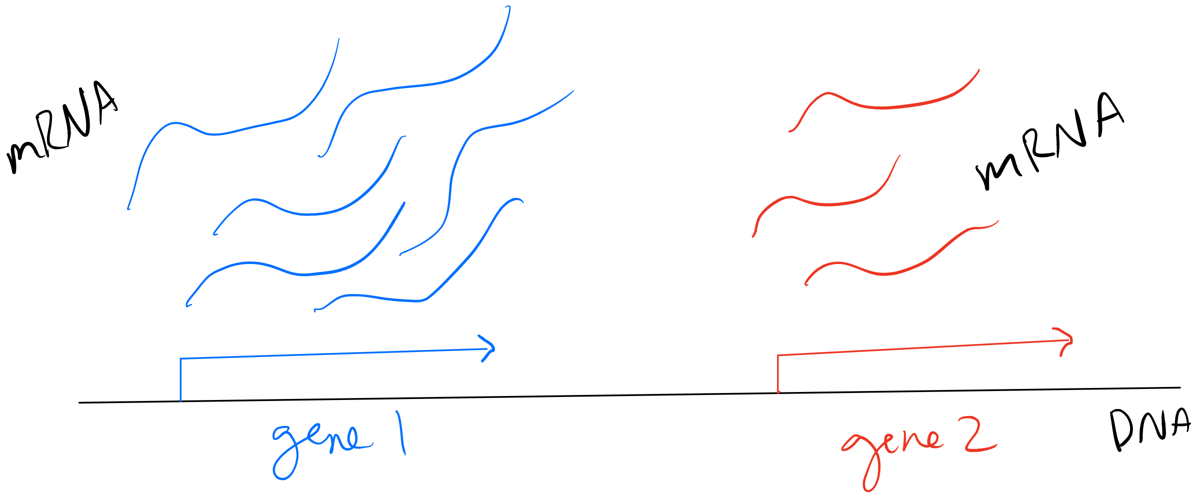




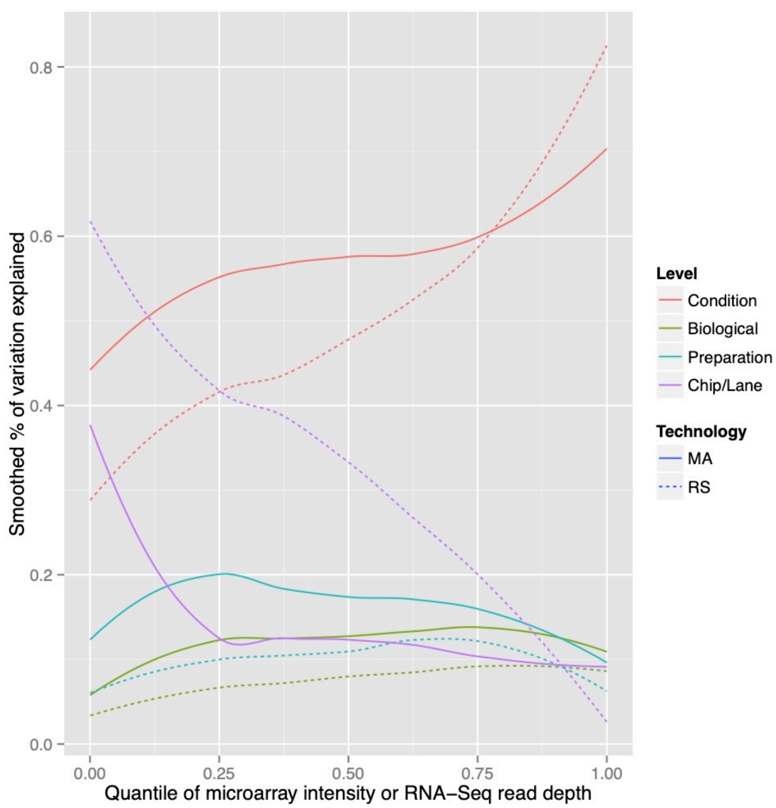



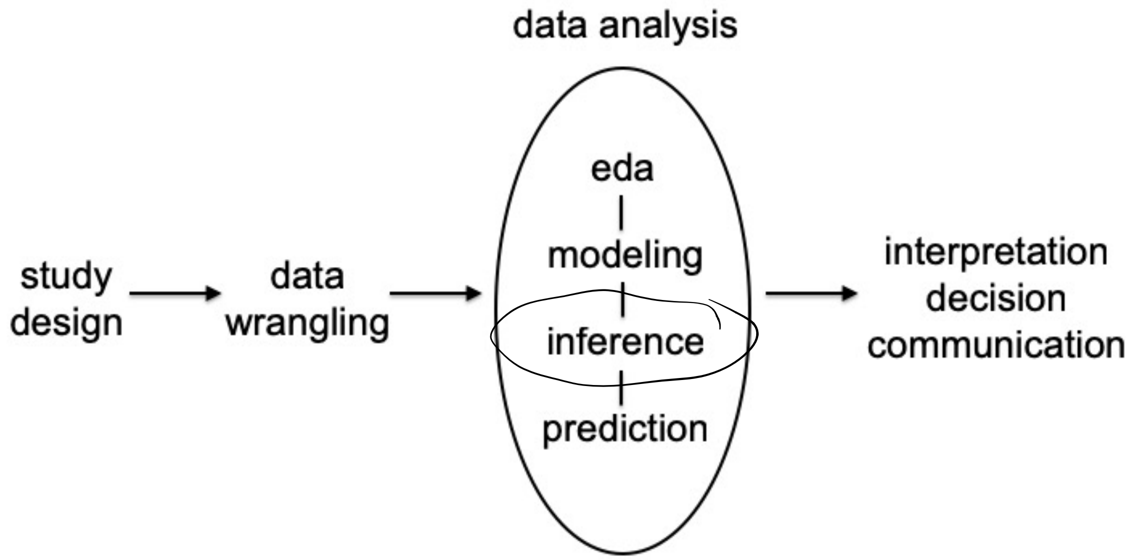

en.wikipedia.org/wiki/RNA-Seq

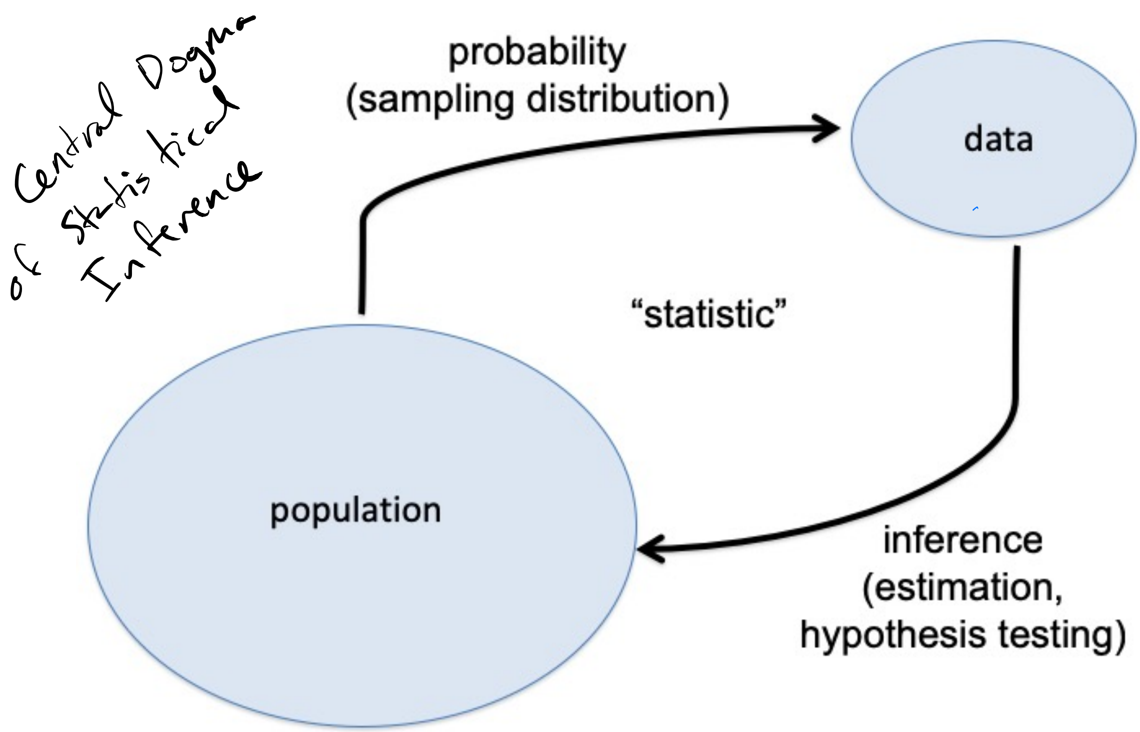
# RNA-Seq — Gene Expression Quantification




Robinson et al.
(2015) NAR






*Foundations of Applied Statistics* -- Storey -- jdstorey.org/fas

## Probability

Probability space $(\Omega, \mathcal{F}, Pr)$

$\Omega$ = set of outcomes, sample space

$Pr$ = probability measure

Events $A \subseteq \Omega$, calculate $Pr(A)$

$\mathcal{F}$ = $\sigma$-algebra, all events A where $Pr(A)$ is meaningful

## Examples of $\Omega$

$\Omega = \{TT, HT, TH, HH\}$ coin flips

$\Omega = \{CC, CT, TT\}$ diploid genotypes

$\Omega = \{C, T\}$ haploid genotypes

$\Omega = \mathbb{R}$ stock returns

$\Omega = [0, \infty)$ height

## Mathematical Probability

1. The probability of any event A is such that $0 \leq Pr(A) \leq 1$.

2. $\Pr(\Omega) = 1$

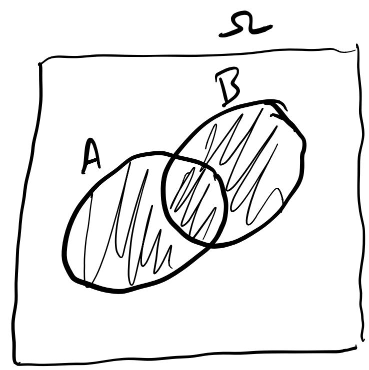
3. Let $A^c$ be the complement of $A$, then $\Pr(A^c) + \Pr(A) = 1$.

4. For any $n$ events such that $A_1, A_2, \ldots, A_n$

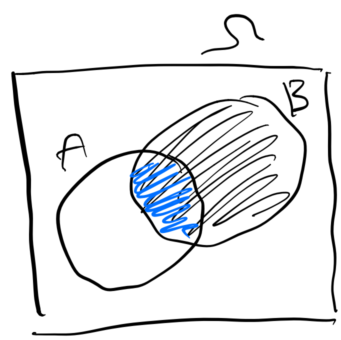
$A_i \cap A_j = \phi \quad \forall i \neq j$, then

$$\Pr\left(\bigcup_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} \Pr(A_j)$$

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$




$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$




Independence

Events $A$ and $B$ are independent if (all equivalent):

- $Pr(A|B) = Pr(A)$
- $Pr(B|A) = Pr(B)$
- $Pr(A \cap B) = Pr(A) Pr(B)$

## Bayes Theorem

$$Pr(B|A) = \frac{Pr(A|B)\, Pr(B)}{Pr(A)}$$

$$Pr(A \cap B) = Pr(B|A)\, Pr(A) = Pr(A|B)\, Pr(B)$$

## Law of Total Probability

Events $A_1, A_2, \dots, A_n$ such that $A_i \cap A_j = \emptyset$ $\forall\, i \neq j$ and $\cup A_i = \Omega$, then for any event $B$,

$$Pr(B) = \sum_{i=1}^{n} Pr(B|A_i)\, Pr(A_i)$$

$A_i \cap B \quad i = 1, \dots, n$

$\bigcup_{i=1}^{n} \{A_i \cap B\} = B$ and disjoint

$$Pr(B) = \sum_{i=1}^{n} \underline{Pr(B \cap A_i)}$$

$$\hookrightarrow = Pr(B \mid A_i) Pr(A_i)$$

## Random Variable

A random variable (rv) $X$ is a function:

$$X : \Omega \longrightarrow \mathbb{R}$$

Take any outcome $\omega \in \Omega$, the $X(\omega)$ produces a real value.

The "range" of $X$ is:

$$\mathcal{R} = \{ X(\omega) : \omega \in \Omega \}$$

where $\underline{\mathcal{R}} \subseteq \underline{\mathbb{R}}$.

## Example

$\Omega = \{CC, CT, TT\}$ SNP genotypes

$X(CC) = 0$ $\qquad$ $Pr(X=0) = Pr(\{CC\})$

$X(CT) = 1$

$X(TT) = 2$

## Distribution of rv's

cumulative distribution function (cdf):

$$F(y) = Pr(X \leq y)$$

Example: $F(1) = Pr(X \leq 1) = Pr(\{CC, CT\})$

$F(1.1) = F(1)$

Discrete rv's have a discrete $\mathcal{X}$

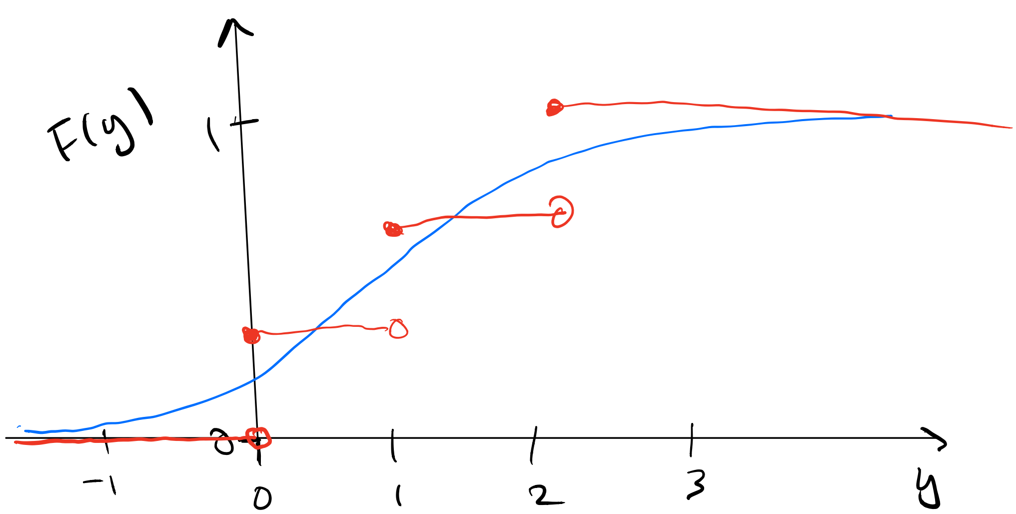
E.g. $\mathcal{X} = \{0, 1, 2, \ldots, 10\}$

$\mathcal{X} = \{0, 1, 3, 4, \ldots\}$

Continuous rv's have a continuous $\mathcal{X}$

E.g. $\mathcal{X} = [0, 1]$

$\mathcal{X} = \mathbb{R}$



# Probability mass or density functions

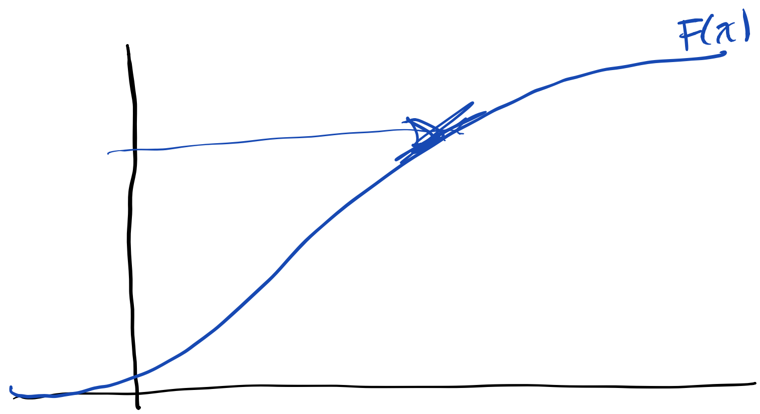
Discrete → Probability mass function (pmf) is

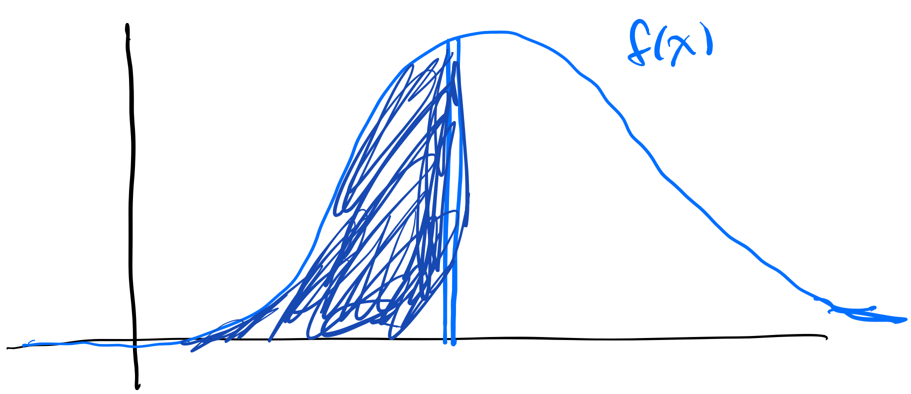
$$f(x) = \Pr(X = x) \quad \text{for all } x \in \mathbb{R}$$

$$f(x) = F(x) - F(b) \quad \text{as } b \uparrow x$$

Continuous → Probability density function (pdf)

$$f(x) = \frac{d}{dx} F(x)$$



Discrete

$$F(y) = \sum_{\substack{x \leq y \\ x \in \mathbb{R}}} f(x) = \Pr(X \leq y)$$

Continuous

$$F(y) = \int_{-\infty}^{y} f(x)\, dx = \Pr(X \leq y)$$

Note that $\Pr(X = x) = 0$

Median of a distribution (aka rv):

A value $y$ s.t. $F(y) = 0.5$

Expected value or "population mean"

$$E(X) = \sum_{x \in \mathbb{R}} x\, f(x) \quad \text{discrete}$$

$$= \int x\, f(x)\, dx \quad \text{for continuous}$$

$$= \int x\, dF(x) dx \quad \text{for measure theory}$$

## Population Variance

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

$$= \sum (x - \underset{\uparrow \text{number}}{E[X]})^2 f(x) \qquad \text{discrete}$$

$$= \int (x - E[X])^2 f(x)\,dx \qquad \text{continuous}$$

$$SD(X) = \sqrt{Var(X)}$$

## Covariance

rv's X and Y

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

$$Var(X) = Cov(X,X)$$

$$Cor(X,Y) = \frac{Cov(X,Y)}{SD(X)\,SD(Y)}$$

$$-1 \le Cor(X,Y) \le 1$$

## Discrete rv's

### Uniform

$X \sim \text{Uniform}(\{1, 2, \dots, n\})$

$\mathcal{X} = \{1, 2, \dots, n\}$

$f(x; n) = \frac{1}{n}$ for $x \in \mathcal{X}$

Calculate $E[X]$, $Var(X)$

"sample" in R

### Bernoulli

$X \sim \text{Bernoulli}(p)$

$\mathcal{X} = \{0, 1\}$

$f(x; p) = (1-p)^{1-x} p^x$

$f(0) = 1-p$, $f(1) = p$

$E[X] = p = 0 \cdot f(0) + 1 \cdot f(1)$

$Var(X) = p(1-p)$

### Binomial

$X \sim \text{Binomial}(n, p)$

Sum of $n$ independent Bernoulli($p$)

$R = \{0, 1, \ldots, n\}$

$$f(x; p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$ number ways to choose $x$ from $n$ w/out order

$$E[X] = np, \quad Var(X) = np(1-p)$$

Example: Under Hardy-Weinberg equilibrium, $X$ = # of T alleles,

$X \sim$ Binomial$(2, p)$

where $p$ is the allele frequency of T.

CC: $Pr(X=0) = (1-p)^2$

$Pr(X=1) = 2p(1-p)$ ←

$Pr(X=2) = p^2$

Poisson

$X \sim$ Poisson$(\lambda)$

$\mathcal{R} = \{0, 1, 2, \dots\}$

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

$E[X] = \lambda$ , $Var(X) = \lambda$

In R

dpois $\rightarrow$ pmf
ppois $\rightarrow$ cdf
qpois $\rightarrow$ quantile
rpois $\rightarrow$ random draws

?Distributions

Continuous

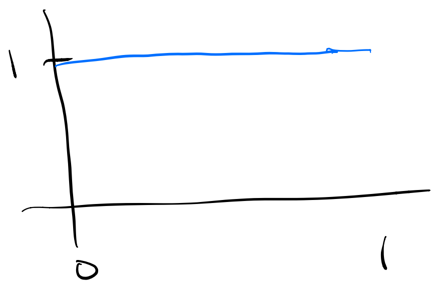
Uniform(0,1) , Normal$(\mu, \sigma^2)$ , Beta$(\alpha, \beta)$

$X \sim$ Uniform(0,1)

$\mathcal{R} = [0,1]$

$f(x) = 1 \qquad x \in [0,1]$

$F(y) = y \qquad y \in [0,1]$




$E[X] = 1/2 \quad Var(X) = \frac{1}{12}$

Uniform $(0, \theta)$

$f(x;\theta) = \frac{1}{\theta} \quad F(y;\theta) = \frac{y}{\theta}$

$\mathcal{X} = [0, \theta]$

<u>Beta</u>

$X \sim Beta(\alpha, \beta) \quad \alpha, \beta > 0$

$\mathcal{X} = (0,1)$

$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$

$x \in (0,1)$

$\int_0^1 f(x;\alpha,\beta)\,dx = 1$

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \, dx$$

$$E[X] = \frac{\alpha}{\alpha+\beta} \qquad Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Beta pdf's