Probability Review

Manpreet Katari

Foundational and Applied Statistics for Biologists using R - Aho

Set Theory

A set is a collection of objects, often defined in { }

A = { H,T }

To show that an element is part of a set you use $\in$ or $\notin$

$H \in A$ or $J \notin A$

If A is a subset of B or A is not a subset of C

$A \subset B$ or $A \not\subset C$

S is often the sample space (all possible values)

An empty set $B = \varnothing$

Probability

Frequentist interpretation -

P(A) is the number of times A was observed in all trials.

P(H) = 0.5 - when you flip the coin, you get H 50% of the time.

Frequentist assume that there is a true value, but there may be none.

There is no prior knowledge used in the calculation of P(A)

Probability

Degrees of belief (bayesian)

P(A) the confidence that A is true.

P(A) does not have to be fixed and it is not being determined, it's just the confidence that A is true.

Prior information is expected to help augment the data.

Classical Probability

N	all possible outcomes
$N(A)$	outcomes resulting in A
$P(A) = N(A)/N$	probability of A is outcomes resulting in A divided by all outcomes
$0 \leq P(A) \leq 1$	probability of A is between 0 and 1
$P(A') = 1-P(A)$	probability of of not A is 1 - probability of A
$P(S) = 1$	probability of getting all possible outcomes is 1
$P(\varnothing) = P(S') = 0$	probability of of getting nothing something not in set is 0

Disjoint (mutually exclusive)

Two events cannot occur simultaneously they are **mutually exclusive or disjointed**.

If we want to know the probability of either outcome when the events are disjointed, we need to determine the **union**

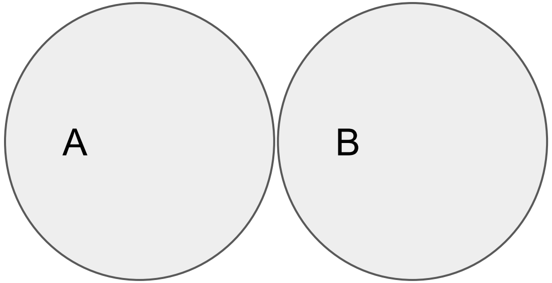
$P(A \cup B) = P(A) + P(B)$

If the events are not mutually exclusive and we want to when they are both true we are looking for **intersect**

$P(A \cap B) = P(A) \times P(B)$

If the event are **not mutually exclusive** and we want to calculate the union it is

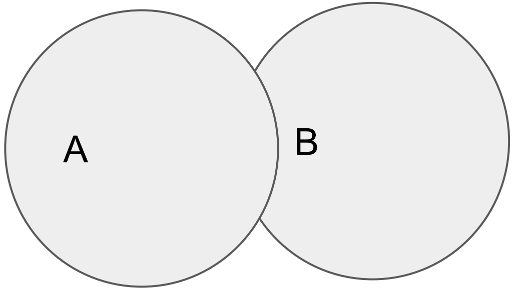
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ # *inclusion-exclusion principle*



Disjoint

$P(A \cup B) = P(A) + P(B)$

$P(A \cap B) = 0$



NOT Disjoint

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

$P(A \cap B) = P(A) \times P(B)$

Conditional Probability

If A and B are not independent, we can use conditional probability to calculate P(A ∩ B). P(A|B) is read probability of A given B.

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

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

Bayes Rule

The Bayes rule allows us to look at the alternative conditional probability.

Posterior distribution

Data distribution

Prior distribution

$$P(\theta \mid \text{data}) = \frac{P(\text{data} \mid \theta) P(\theta)}{P(\text{data})}.$$

Total probability

$$P(\text{data}) = \sum_{k=1}^{c} P\left(\text{data} \mid \theta_k\right) P(\theta_k),$$

Bayes Example

Cystic fibrosis, a deadly lung disorder, is one of the most common autosomal recessive diseases. This is true in part because there are more than 1500 mutational forms of the cystic fibrosis transmembrane regulatory (CFTR) gene (Ratbi et al. 2007). With current diagnostic methods, the true-positive rate (i.e., sensitivity) for detecting CFTR mutations is approximately 0.99; thus the false-negative rate is 0.01. The true-negative rate (i.e., specificity) of this test is generally assumed to be 100%, but it ignores human error. As a result, we will let this probability be 0.998, and let the false-negative probability be 0.002.

True-positive rate = $P(POS|CF+) = 0.99$ False-positive rate = $P(POS|CF-) = 0.002$
True-negative rate = $P(NEG|CF-) = 0.998$ False-negative rate = $P(NEG|CF+) = 0.01$

A question of great interest is “what is the probability that a person is a carrier of some mutational variety of CFTR but will not know it because their test was negative.” “what is $P(CF+|NEG)$?”

$P(CF+) = 0.04$

Combinatorial Analysis

When an event include multiple selection from S, pay attention to:

1) Will the element be replaced to be selected.
2) Does the order of selection matter.

Multiplication principle (with replacement)

- If there are $\mathbf{n_1}$ ways to organize first operation and $\mathbf{n_2}$ ways to organize the second than there are $\mathbf{n_1 * n_2}$ ways of representing all possibilities
- If there are **n** possibilities performed **r** times, than there are $\mathbf{n^r}$ possible outcomes

Permutations (without replacement - order matters)

The number of permutations of any S of size n is n!

n! = n * (n-1) * (n-2) * … 1 # 0! = 1

If the selection is a subset of S

nPerm = n! / (n-r)! # where r is the number of S taken at a time.

How many different ways can you draw cards from a deck? (order matters)

How many ways to draw 5 cards from a deck ? (order matters)

Combinations - without replacement & order doesn’t matter

Binomial coefficient

(n choose r)

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

How many ways to select 2 out of the 5 items

(without replacement & order doesn’t matter)?

Probability Density

What is Density?

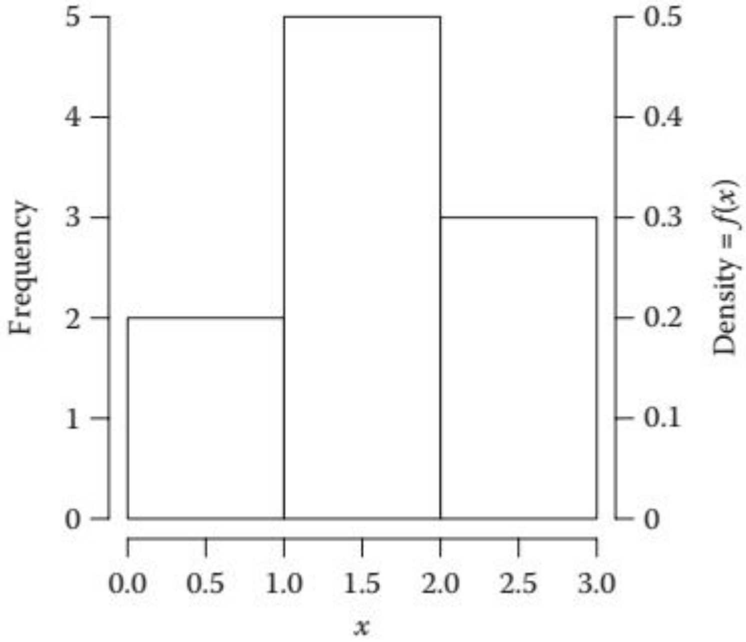
Density is the output generated by the pdf function

A pdf function is the result of the function when x varies.

A pdf is best illustrated using a histogram.

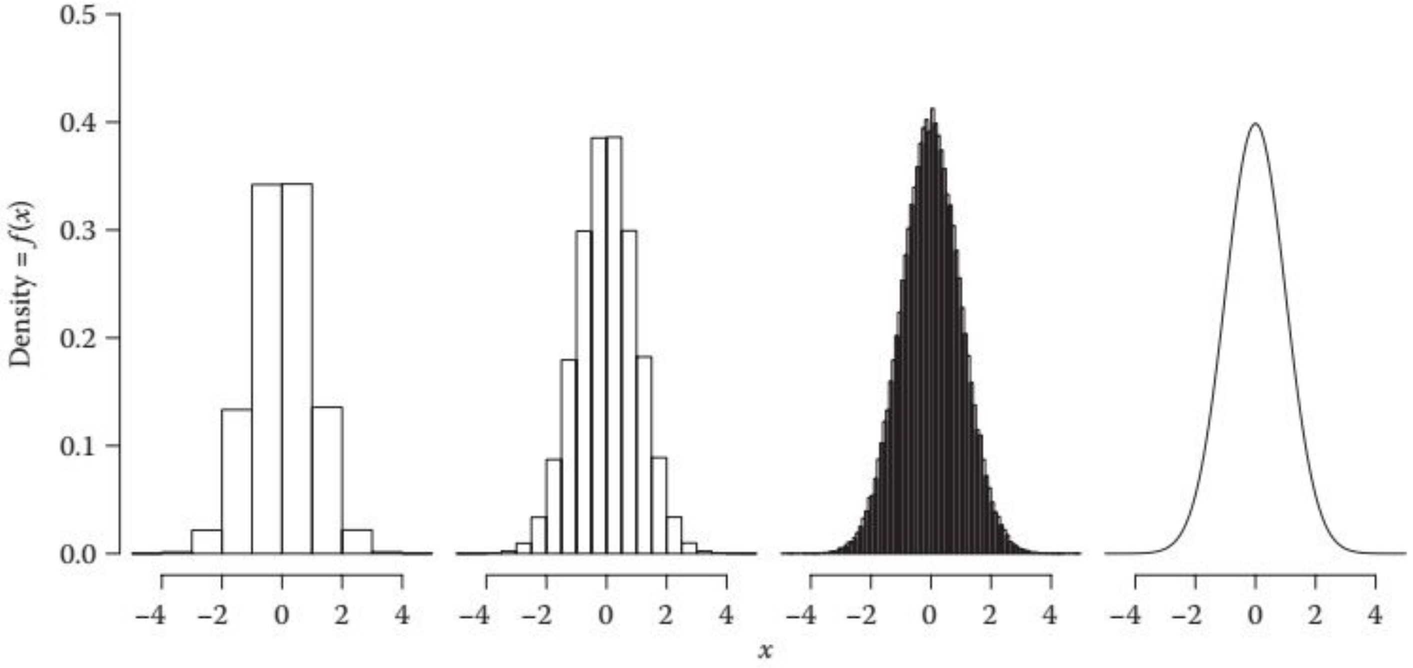
If the numerical value is discrete, the height of the bar at a given values of x, is the probability of x.

If the numerical value is continuous, this not possible.



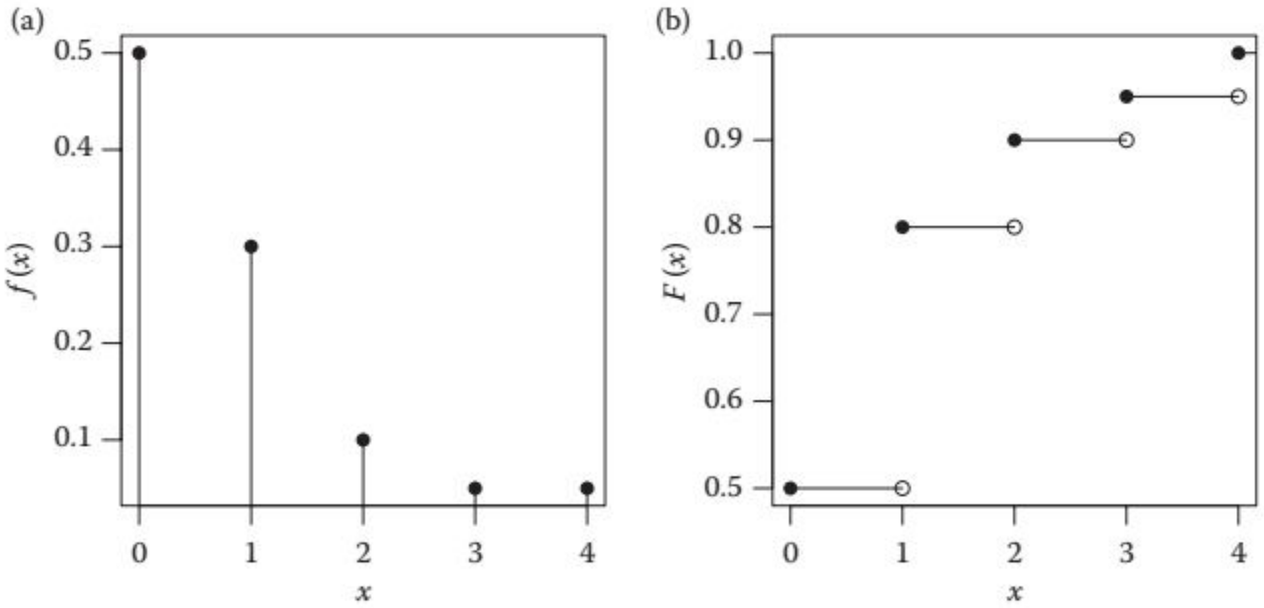
Density of Continuous values

As we shrink the size of the bins, the discrete values will be more continuous.



Cumulative Density Function (CDF)

The cdf is simply the sum of the density upto x.



Bernoulli Distribution

- Probability of success for a single binary trial (the output is binary)
- π = probability of success, x = number of times you want x to occur, in this case 1 or 0.

$$f(x) = \pi^x (1-\pi)^{1-x}$$

For a coin toss π is 0.5, for a dice roll, π is 0.167

Binomial Distribution

- Is a combination of Bernoulli trials that are independent and identically distributed

$$f(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x},$$

Binomial Distribution Example

Native cutthroat trout (Salmo clarki) in Yellowstone Lake (Yellowstone National Park, Wyoming, USA) are currently being threatened by nonnative lake trout (Salvelinus namaycush). Juvenile cutthroats are highly vulnerable to the larger piscivorous lake trout. In fact, one mature lake trout can eat 41 juvenile cutthroats per year (Ruzycki et al. 2003). In 1994, the year that the first confirmed lake trout was caught in Yellowstone Lake, a gill net survey found that approximately 68% of fish in the lake were cutthroats (Kaeding et al. 1995). Suppose that recent catch and release data from anglers showed that of 50 fish caught in the lake only 25 were cutthroats. * What is the probability of this outcome given 1994 fish proportions?

Binomial Distribution Example continued

We might also want to know $P(X \leq 25)$, the probability of catching 25 or fewer cutthroats, that is, 0 fish, or 1 fish, up to and including 25 fish.

R commands for statistical distributions

rbinom - generate random numbers from a binomial distribution

dbinom - density (height) of distribution at x

pbinom - cumulative density of distribution at x

qbinom - quantile of x in the defined distribution