Probability Distributions
Part II

Session 3

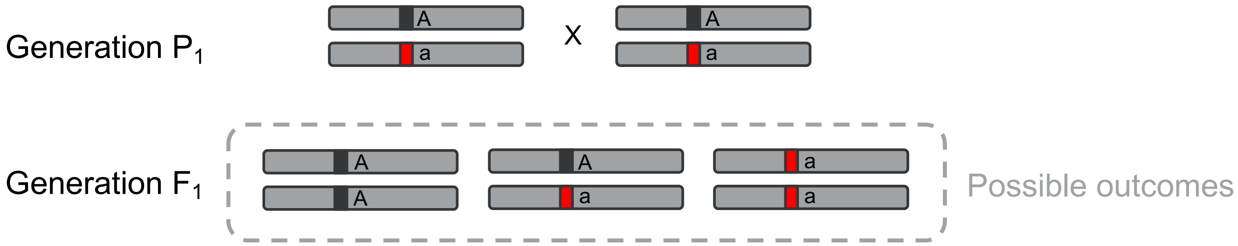
Module 1 Probability & Statistical Inference

The Summer Institutes
DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY of WASHINGTON

Multinomial Distribution

This distribution generalizes beyond 2 outcomes of the binomial distribution.

For example, let's consider a cross between two parents that are heterozygous carriers for a recessive trait:



We can then use the multinomial distribution to calculate the probability of observing various outcomes in the offspring generation.

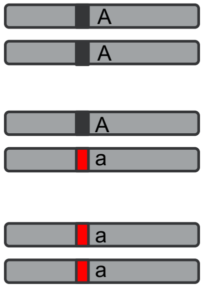
Multinomial Distribution

For any given offspring, the 3 possible outcomes can be represented by:

Y_{i1} = 1 if i^{th} offspring is **unaffected** (AA),
= 0 otherwise

Y_{i2} = 1 if i^{th} offspring is a **carrier** (Aa),
= 0 otherwise

Y_{i3} = 1 if i^{th} offspring is **affected** (aa),
= 0 otherwise



Only one of Y_{i1}, Y_{i2}, Y_{i3} can be equal to 1, so $Y_{i1} + Y_{i2} + Y_{i3} = 1$.

For the binomial distribution with 2 outcomes, there are 2^n unique outcomes in n trials. With n=3 offspring, there are $2^3 = 8$ unique outcomes.

For the multinomial distribution with 3 outcomes, the number of unique outcomes in n trials is 3^n. With n=3 offspring, there are 3^3=27 unique outcomes.

Multinomial distribution

To calculate probabilities of interest, we can use **combinations**. For the multinomial distribution, the combinations are calculated as:

$$C_k^n = \frac{n!}{k_1!k_2 \dots k_J!}$$

where k_j (j=1, 2, …, J) correspond to the totals for the different outcomes

Let's consider a scenario where:

n = 2 offspring
J = 3 possible outcomes (unaffected, carrier, affected)

The possible outcomes are:

Offspring 1	Offspring 2	Outcome
AA	AA	2 unaffected, 0 carrier, 0 affected
AA	Aa	1 unaffected, 1 carrier, 0 affected
Aa	AA	1 unaffected, 1 carrier, 0 affected
AA	aa	1 unaffected, 0 carrier, 1 affected
aa	AA	1 unaffected, 0 carrier, 1 affected
Aa	Aa	0 unaffected, 2 carrier, 0 affected
aa	Aa	0 unaffected, 1 carrier, 1 affected
Aa	aa	0 unaffected, 1 carrier, 1 affected
aa	aa	0 unaffected, 0 carrier, 2 affected

Number of unique outcomes
$3^n = 3^2 = 9$

Multinomial distribution

For a defined number of offspring, what is the probability of a specific outcome? E.g., for n=2, what is the probability of observing two unaffected individuals? Or two affected? Or…?

First, write the outcomes in terms of k's:

k_1=number of unaffected
k_2=number of carriers
k_3=number of affected

Offspring 1	Offspring 2	Outcome	# ways
p_1	p_1	$k_1=2, k_2=0, k_3=0$	1
p_1	p_2	$k_1=1, k_2=1, k_3=0$	2
p_2	p_1	$k_1=1, k_2=1, k_3=0$	
p_1	p_3	$k_1=1, k_2=0, k_3=1$	2
p_3	p_1	$k_1=1, k_2=0, k_3=1$	
p_2	p_2	$k_1=0, k_2=2, k_3=0$	1
p_3	p_2	$k_1=0, k_2=1, k_3=1$	2
p_2	p_3	$k_1=0, k_2=1, k_3=1$	
p_3	p_3	$k_1=0, k_2=0, k_3=2$	1

Formula for # ways

$$\frac{n!}{k_1!k_2!k_3!}$$

Probability for each possible outcome

$$\Pr[Y_1=k_1, Y_2=k_2, Y_3=k_3] = p_1^{k1}p_2^{k2}p_3^{k3}$$

The probability of a **specified outcome** is going to be:

[prob of the possible outcome] x [number of ways for that outcome]

Multinomial distribution

The probability that a multinomial random variable with n trials and success probabilities $p_1, p_2, \ldots, p_J$ will yield exactly $k_1, k_2, \ldots k_J$ successes is:

Probability of a specific scenario | # ways | Probability of the outcome

$$P(Y_1 = k_1, Y_2 = k_2, \ldots, Y_J = k_J) = \frac{n!}{k_1! k_2! \ldots k_J!} p_1^{k_1} p_2^{k_2} \cdots p_J^{k_J}$$

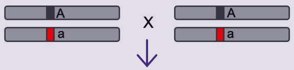
Defines what we are asking, e.g., for n=2, what is the probability of observing 1 unaffected (k_1=1), 1 carrier (k_2=1), and 0 affected (k_3=0)

Recall:
k_1=number of unaffected
k_2=number of carriers
k_3=number of affected

These again are the k values, which are given based on what we are asking

The p probabilities define the "baseline" probability of success for each of the J outcomes =

For heterozygous cross:



p_1 = P(AA) = 0.25
p_2 = P(Aa) = 0.5
p_3 = P(aa) = 0.25

Multinomial distribution

The probability that a multinomial random variable with n trials and success probabilities $p_1, p_2, \ldots, p_J$ will yield exactly $k_1, k_2, \ldots k_J$ successes is:

$$P(Y_1 = k_1, Y_2 = k_2, \ldots, Y_J = k_J) = \frac{n!}{k_1! k_2! \ldots k_J!} p_1^{k_1} p_2^{k_2} \cdots p_J^{k_J}$$

Assumptions:

- J possible outcomes; only one can be a success, 1, in a given trial.
- The probability of success for each possible outcome, p_j, is the same for each trial.
- The outcome of one trial has no influence on other trials (independent trials).
- Interest is in the (sum) total number of successes over all the trials.

Multinomial distribution

What is the probability that one of n=3 offspring will be unaffected (AA), one will be affected (aa) and one will be a carrier (Aa)?

Our k values for this scenario are:

k_1 = number of unaffected = 1
k_2 = number of carriers = 1
k_3 = number of affected = 1

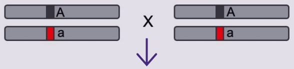
$$P(Y_1 = k_1, Y_2 = k_2, \ldots, Y_J = k_J) = \frac{n!}{k_1!k_2!\ldots k_J!} p_1^{k_1} p_2^{k_2} \cdots p_J^{k_J}$$

$$P(Y_1 = 1, Y_2 = 1, Y_3 = 1) = \frac{3!}{1!1!1!} p_1^1 p_2^1 p_3^1$$

$$= \frac{(3)(2)(1)}{(1)(1)(1)} \left(\frac{1}{4}\right)^1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{4}\right)^1$$

$$= \frac{3}{16} = 0.1875.$$

For heterozygous cross:



p_1 = P(AA) = 0.25
p_2 = P(Aa) = 0.5
p_3 = P(aa) = 0.25

Multinomial distribution

Calculating the mean and variance

The marginal outcomes of the multinomial distribution are binomial.

We can obtain the means for each outcome, e.g, $Y_j = k_j$, the j^{th} outcome, as follows:

Mean:

$$E[k_j] = E\left[\sum_{i=1}^{n} Y_{ij}\right] = \sum_{i=1}^{n} E[Y_{ij}]$$

$$= \sum_{i=1}^{n} p_j = np_j$$

Variance:

$$V[k_j] = V\left[\sum_{i=1}^{n} Y_{ij}\right] = \sum_{i=1}^{n} V[Y_{ij}]$$

$$= \sum_{i=1}^{n} p_j(1-p_j) = np_j(1-p_j)$$

Multinomial distribution

Multinomial distribution summary

- Multinomial random variables are discrete
- Parameters are $n, p_1, p_2, \ldots, p_J$
- Each outcome $Y_j = k_j$ is the sum of n independent Bernoulli outcomes
- Extends binomial distribution
- Seen in contingency tables, polytomous regression

Paws

Work through questions 1-2

Continuous distributions

Continuous distributions

For measurements like height or weight, it does not make sense to talk about the probability of any single value. Instead, we talk about the probability for an **interval**.

P[weight = 70.000kg] ≈ 0

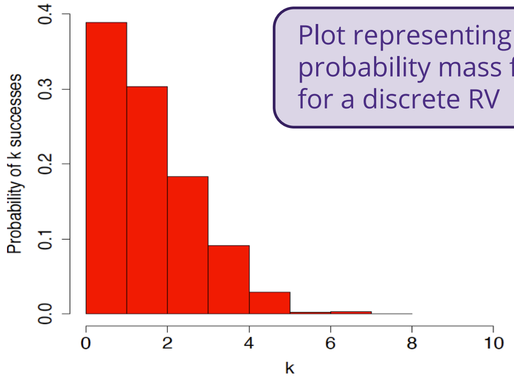
P[69.0kg ≤ weight ≤ 71.0kg] = 0.08

For discrete random variables, a **probability mass function** gives the probability of each possible value.

For continuous random variables, we require a **probability density function** to tell us about the probability of obtaining a value within an interval.

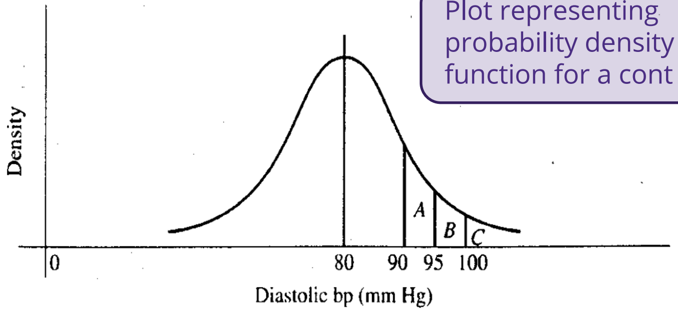
Continuous distributions

With discrete probability distributions, we can determine the probability of a single outcome:



Plot representing probability mass function for a discrete RV

With continuous probability distributions, we determine the probability across a range of outcomes:



Plot representing probability density function for a cont RV

For any interval, the **area under the curve** represents the probability of obtaining a value in that interval.

Continuous distributions

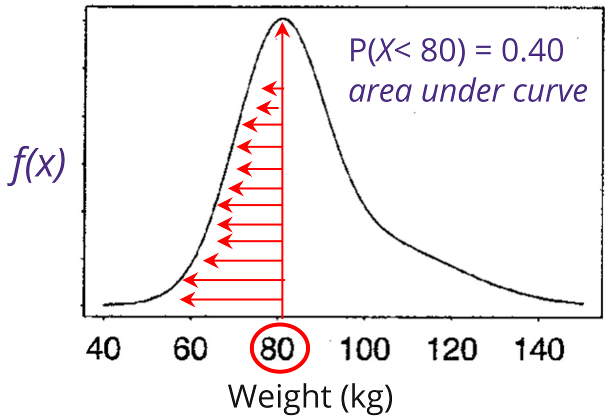
Probability density function

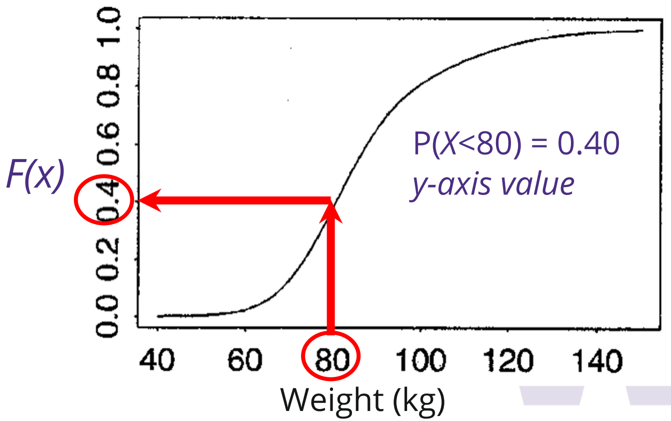
- Given by *f(x)*
- Gives probability that *X* falls within an interval: *f(x)* = P(value 1 < *X* < value 2)
- Probability represented by area under curve
- Total area under curve is 1: $\int f(x)dx = 1.0$

Cumulative distribution function

- Given by *F(x)*
- Gives cumulative probability that *X* is less than some value *x:*
- *F(x)* = P(*X* ≤ *x*)
- y-axis ceiling is 1

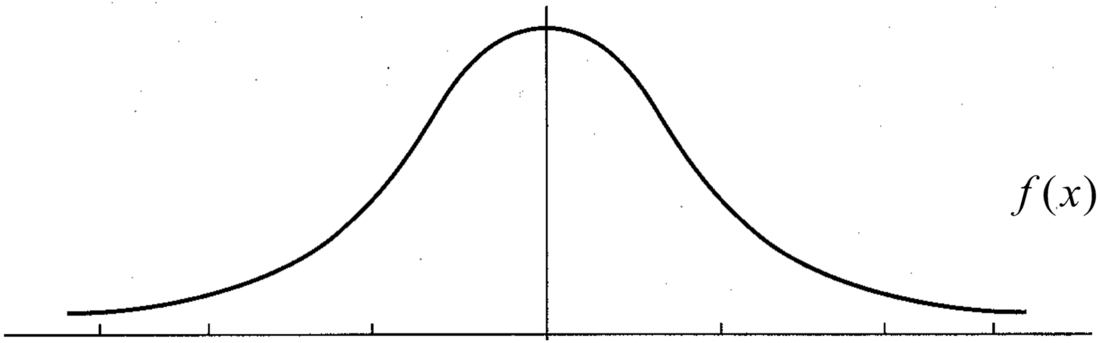
The **PDF** and **CDF** represent the **same information**.





Normal distribution

The normal distribution is a well-known probability model for continuous data. It is unimodal with a “bell-shaped curve”.



PDF of the Normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Random variable values range from -∞ to + ∞.

Symmetric about mean: **mean = median = mode**

Common examples include human height, birth weight, blood pressure.

Normal distribution

The NOR distribution is defined by its mean and variance. Note the appearance of μ and σ in the probability density function for the NOR distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right)$$

Thus, the normal distribution has two parameters:

μ = the mean of X

σ = the standard deviation of X

$X \sim N(\mu , \sigma^2)$

"X is normally distributed with mean μ and variance σ^2"

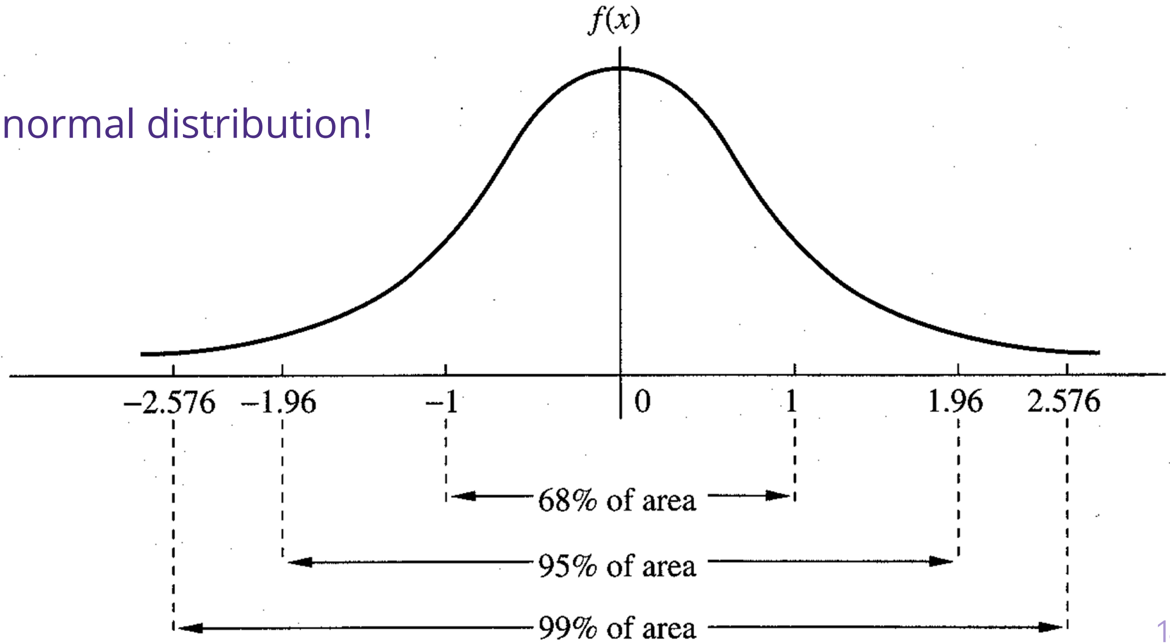
Normal distribution

The standard normal distribution

The standard normal is a special case of the NOR!

- AKA "z" distribution
- $Z \sim N(0, 1)$: $\mu = 0$ and $\sigma^2 = 1$
- There is only ONE standard normal distribution!

Under the STD NOR, 95% of the area lies between ~2 standard deviations of the mean. Useful!



Standard Normal distribution

Calculating probabilities for the standard normal distribution

Using ***z*** notation, the **probability density function** of *Z*, the random variable of the standard normal distribution is:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2 \right)$$

The **cumulative distribution function** of Z is:

$$P(Z \le x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2 \right) dz$$

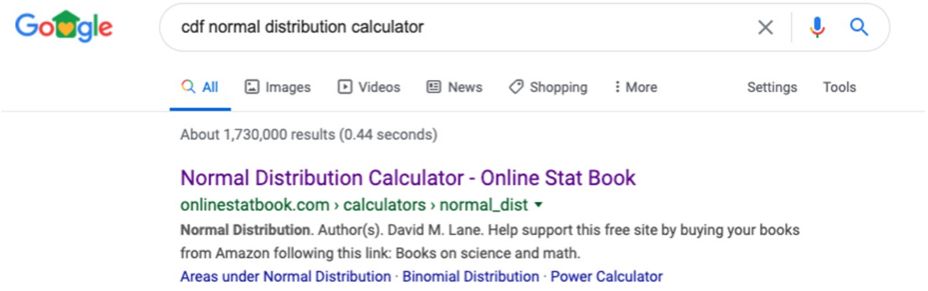
Any computing software will give the values of *f(z)* and *Φ(x).*

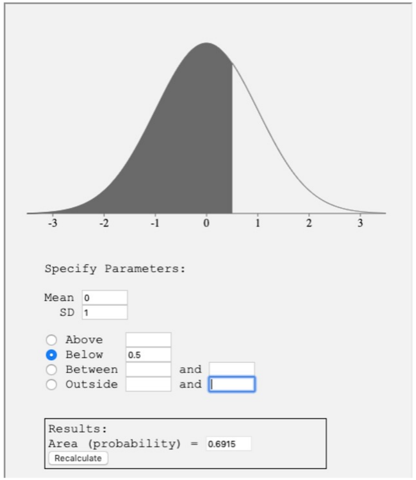
Standard Normal distribution

Online calculators for standard Normal distribution probabilities

Suppose we wish to solve a probability statement for the standard Normal variable. For example, what is the probability that Z takes a value less than 0.05?

$P(Z \leq 0.5) = ?$





$P(Z \leq 0.5) = 0.6915$

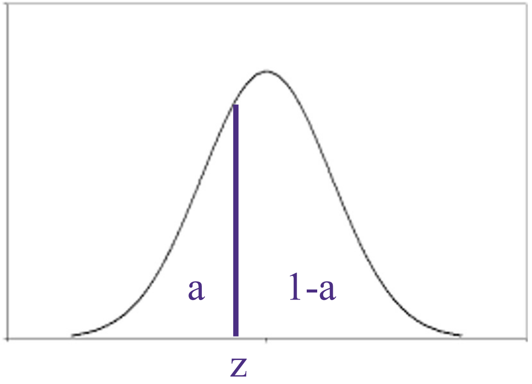
Standard Normal distribution

Properties

If the probability of ($Z <$ some value z) is equal to a, then the probability that ($Z > z$) is equal to 1 - a.

$$P(Z < z) = a$$

$$P(Z > z) = 1 - a$$



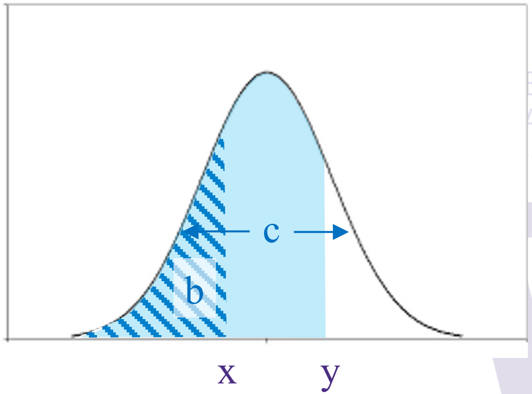
If the probability of ($Z <$ some value x) is equal to b, and the probability that ($Z > y$) is equal to c, then the probability that (Z lies between x and y) is equal to $c – b$.

$$P(Z < x) = b$$

$$P(Z < y) = c$$

$$P(x < Z < y) = c - b$$

Note that this is true of all distributions!



Standard Normal distribution

Properties

Because the standard Normal distribution is symmetrical around 0, the probability that ($Z < -y$) is equal to the probability that ($Z > y$), shown here as d.

$$P(Z < -y) = P(Z > y) = d$$

Note that this is NOT true of all distributions!