Probability Distributions

II

Multinomial Distribution - Motivation

Suppose we modified assumption (1) of the binomial distribution to allow for more than two outcomes.

For example, we might be interested in calculating the following probabilities for the offspring of parents that are heterozygote carriers of a recessive trait.

Q_1: One of their n=3 offspring will be unaffected (AA), 1 will be affected (aa) and one will be a carrier (Aa),

Q_2: All of their offspring will be carriers,

Q_3: Exactly two of their offspring will be affected (aa) and one will be a carrier.

Multinomial Distribution - Motivation

For each child, we can represent each of these possible outcomes with three indicator variables for the i^{th} child as:

Y_{i1} = 1 if i^{th} child is unaffected (AA) & =0 otherwise
Y_{i2} = 1 if i^{th} child is a carrier (Aa) & =0 otherwise
Y_{i3} = 1 if i^{th} child is affected (aa) & =0 otherwise

Notice only one of the three Y_{i1}, Y_{i2}, Y_{i3} can be equal to 1 ($Y_{i1} + Y_{i2} + Y_{i3} = 1$).

For the binomial distribution with 2 outcomes (e.g., unaffected vs. carrier/affected), there are 2^n unique outcomes in n trials. In the family with n=3 children, 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. For the family of 3, that's 3^3=27 unique outcomes.

Possible Outcomes

Combinations: As with the binomial distribution, when order doesn't matter, the total number of possible outcomes, can be straightforwardly calculated. For the multinomial distribution, the combinations are calculated as

$$\frac{n!}{k_1!k_2!\ldots k_J!}$$

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

e.g. n=2 children
J=3 possible outcomes (unaffected/carrier/affected)

Child number

1	2	Outcomes
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

For n=2 children, what are the probabilities of various outcomes?

E.g. (n=2, k_1=number of unaffected, k_2=number of carrier, k_3=number of affected)

Child number 1	2	Outcomes	# 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

For each possible outcome, the probability $\Pr[Y_1=k_1, Y_2=k_2, Y_3=k_3]$ is $p_1^{k_1} p_2^{k_2} p_3^{k_3}$

There are $\frac{n!}{k_1!k_2!k_3!}$ sequences for each probability, so in general…

Multinomial Probabilities

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:

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

k_1	k_2	k_3	k_4	$\cdots$	k_{J-1}	k_J

$n = \Sigma_j k_j$ is the total number of trials.

Multinomial Probabilities - Examples

Returning to the original questions:

Q$_1$: One of n=3 offspring will be unaffected (AA), one will be affected (aa) and one will be a carrier (Aa) (recessive trait, carrier parents)?

Solution: For a given child, the probabilities of the three outcomes are:

p_1 = Pr[AA] = 1/4
p_2 = Pr[Aa] = 1/2
p_3 = Pr[aa] = 1/4

We have
$$P(Y_1 = 1, Y_2 = 1, Y_3 = 1) = \frac{3!}{1!1!1!} p_1^1 p_2^1 p_3^1$$
$$= \frac{(3)(2)(1)}{(1)(1)(1)} \left(\frac{1}{4}\right)^1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{4}\right)^1$$
$$= \frac{3}{16} = 0.1875.$$

Questions 1, 2

Recall, carrier=Aa with Pr(Aa) = ½

unaffected=AA with Pr(AA) = ¼

affected=aa with Pr(aa) = ¼

1. What is the probability that all three offspring will be carriers?

2. What is the probability that exactly two offspring will be affected and one a carrier?

Example - Mean and Variance

The (marginal) outcomes of the multinomial distribution are binomial. We can immediately obtain the means for each outcome, e.g, $Y_j = k_j$, the j^{th} outcome:

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 Summary

1. Multinomial RVs are discrete
2. Parameters - $n, p_1, p_2, \ldots, p_J$
3. Each outcome $Y_j = k_j$ is the sum of n independent Bernoulli outcomes
4. Extends binomial distribution
5. Contingency tables, polytomous regression

Continuous Distributions

Continuous Distributions

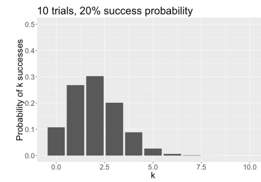
For measurements like height and weight which can be measured with arbitrary precision, it does not make sense to talk about the probability of any single value. Instead we talk about the probability for an **interval**.

P[weight = 70.000kg] $\approx$ 0

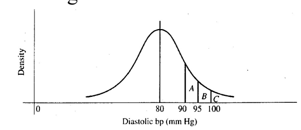
P[69.0kg $\leq$ weight $\leq$ 71.0kg] = 0.08

For discrete random variables, we had a probability mass function to give us the probability of each possible value. For continuous random variables we use a **probability density function** to tell us about the probability of obtaining a value within some interval.

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



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



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

Probability density function

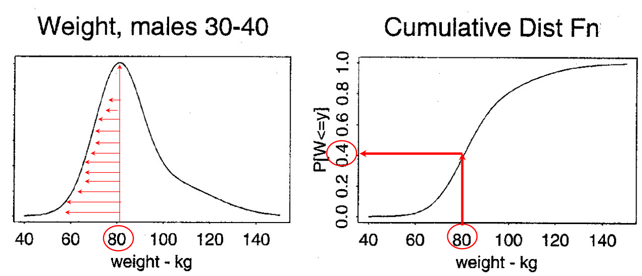
1. A function, typically denoted f(x), that gives probabilities based on the **area** under the curve.
2. f(x) $\geq$ 0
3. Total area under the function f(x) is 1. $\int f(x)dx = 1.0$

Cumulative distribution function

The **cumulative distribution function**, F(t), tells us the total probability that X is less than some value t.

F(t) = P(X $\leq$ t)

f(t) F(t)



Prob[weight < 80] = 0.40
Area under the curve

Normal Distribution

- A well-known probability model for continuous data
- Bell-shaped curve

⇒ takes values between -∞ and + ∞

⇒ symmetric about mean

⇒ mean = median = mode

- Common examples (that don't always hold):
  - birthweights
  - blood pressure
  - CD4 T cell counts (transformed)

The normal distribution is more useful as a derived distribution, as we will see when we talk about the central limit theorem.

Normal Distribution

Specifying the mean and variance of a normal distribution completely determines the probability distribution function and, therefore, all probabilities.

The **normal probability density function** is:

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

where

$\pi \approx 3.14$ (a constant)

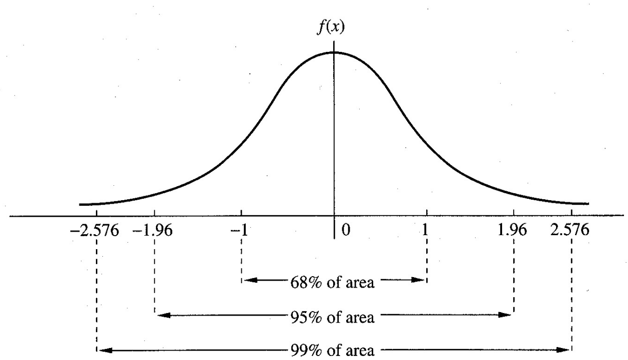
Notice that the normal distribution has two *parameters*:

μ = the mean of X

σ = the standard deviation of X

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

The **standard normal** distribution, N(0, 1), is a special case where $\mu = 0$ and $\sigma^2 = 1$.



Normal Distribution - Calculating Probabilities

Example 5.20: Rosner, Fundamentals of Biostatistics

Serum cholesterol is approximately normally distributed with mean 219 mg/mL and standard deviation 50 mg/mL. If the clinically desirable range is < 200 mg/mL, then what proportion of the population falls in this range?

X = serum cholesterol in an individual.

μ = 219 mg/mL

σ = 50mg/mL

$$P[x < 200] = \int_{-\infty}^{200} \frac{1}{50\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-219)^2}{50^2} \right) dx$$

negative values for cholesterol ??

Standard Normal Distribution - Calculating Probabilities

First, let's consider the **standard normal** - N(0,1). We will usually use Z to denote a random variable with a standard normal distribution. The probability density function of Z is:

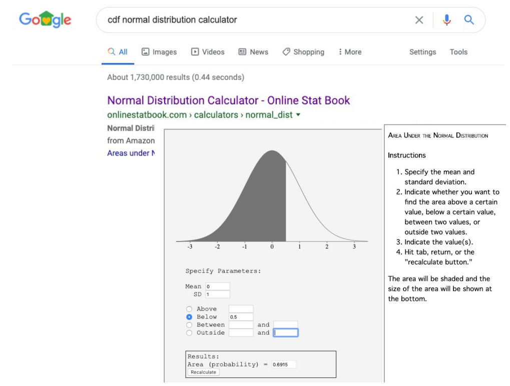
$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

and the **cumulative distribution** 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 can give you the values of f(z) and Φ(z)

Standard Normal Distribution - Calculating Probabilities

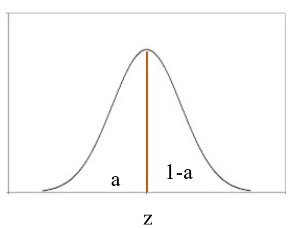


Pr(Z ≤ 0.5) = 0.6915

Facts about probability distributions

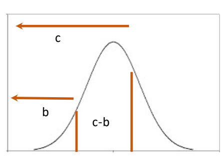
P(Z ≤ z) = a

P(Z > z) = 1- a



P(Z ≤ x) = b, P(Z ≤ y) = c

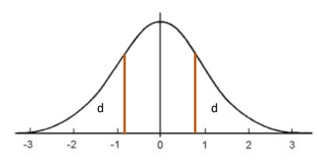
Pr(x < Z ≤ y) = c - b



Facts about the standard normal distribution

Because the N(0,1) distribution is symmetric around 0,

Pr(Z ≤ -y) = Pr(Z ≥ y) = d



Pause- break time then work on exercises

Questions 3, 4, 5, 6

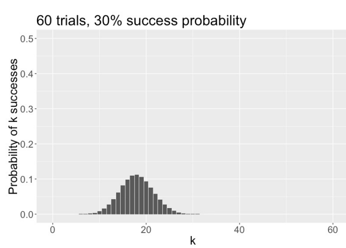
Google "cdf normal distribution calculator" and find the following if $Z \sim N(0,1)$

3. $P(Z \leq 1.65) =$
4. $P(Z \geq 0.5) =$
5. $P(-1.96 \leq Z \leq 1.96) =$
6. $P(-0.5 \leq Z \leq 2.0) =$

Normal Approximation to Binomial Distribution

Example:

Suppose the prevalence of HPV in women 18-22 years old is 0.30. What is the probability that in a sample of 60 women from this population that 9 or fewer would be infected?



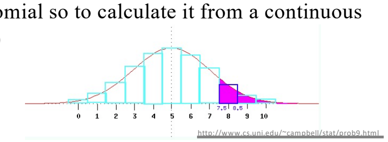
Normal Approximation to Binomial Distribution

Binomial

- When **np(1-p)** is "large" (e.g. ≥ 3), the normal distribution may be used to approximate the binomial distribution.
- $X \sim \text{bin}(n,p)$

  $E(X) = np$

  $V(X) = np(1-p)$
- X is approximately $N(np, np(1-p))$
- Apply continuity correction for discreteness:
  - $P(X \leq x)$ is a discrete binomial so to calculate it from a continuous normal, use $P(X \leq x + 0.5)$



Application of Normal Approximation to Binomial Distribution

Example:

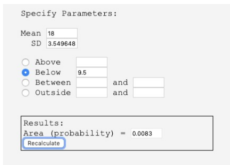
Suppose the prevalence of HPV in women 18 -22 years old is 30%. What is the probability that in a sample of 60 women from this population that 9 or less would be infected?

Solution

X = number infected out of 60

X ~ Binomial(n=60, p =0.3)

X close to Normal distribution with mean 60*0.3=18 and variance 60*0.3*(1-0.3)=12.6



Therefore, $P(X \leq 9.5) = 0.0083$