Podstawy probabilistyki i statystyki w kojarzeniu ryb akwariowych

Piotr Łapa[1]

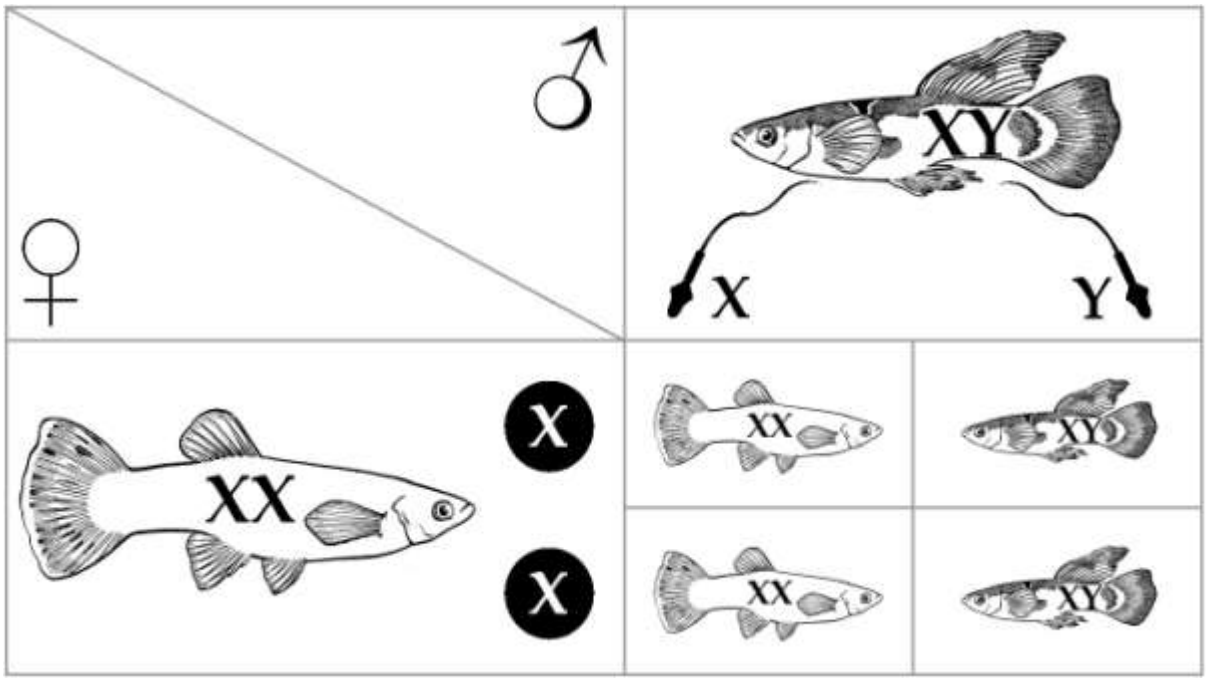
[1]Towarzystwo Naukowe Branży Zoologicznej „Animalian”

W niniejszym opracowaniu przedstawione zostaną elementarne podstawy rachunku prawdopodobieństwa i testowania hipotez. Wykład nie będzie jednak opierał się o wzory matematyczne i będą one przedstawiane sporadyczne, lecz opisywane zjawiska prawdopodobieństwa, ilustrowane będą rzeczywistymi przykładami mechanizmów, które są dla większości osób oczywiste, choć nie zawsze uświadomione.

1. Rozkłady genotypów i fenotypów w grupach potomstwa

Najprostsza definicja prawdopodobieństwa mówi, że jest to stosunek liczby przypadków sprzyjających określonemu zjawisku, do liczby wszystkich możliwych przypadków. Ale co to właściwie znaczy? Prawdopodobieństwem nazywamy spodziewaną częstość zajścia konkretnego zdarzenia. Rozważmy to na przykładzie rozkładu płci u gupików (*Poecilia reticulata* Peters, 1859) u których to dziedziczenie płci odbywa się za pośrednictwem pary homologicznych heterochromosomów, oznaczanych jako X i Y (podobnie jak u ssaków). W kariotypie samic występuje układ homozygotyczny XX, natomiast u samców układ chromosomów jest heterozygotyczny XY. Dziedziczenie chromosomów płci odbywa się zgodnie z prawami Mendla i dla ilustracji tego procesu można posłużyć się również tzw. szachownicą Punneta (rycina 1). W procesie mejozy, przebiegającym podczas podziałów komórkowych prowadzących do powstania komórek rozrodczych (gametogeneza), chromosomy homologiczne przekazywane są do różnych gamet w taki sposób, że jeden chromosom z każdej pary przechodzi do jednej komórki potomnej, a drugi chromosom do drugiej gamety. Identycznie odbywa się w przypadku chromosomów płci, z tym, że w przypadku samic posiadających dwa chromosomy X, wszystkie komórki jajowe posiadają jeden z możliwych chromosomów, tj. właśnie X, a u samców posiadających układ XY do połowy spośród produkowanych plemników, przekazywany jest chromosom X, a do drugiej połowy alternatywny wariant czyli Y. Właśnie istnienie dwóch równych części

plemników, wyposażonych w chromosom X lub Y jest powodem tego, że wśród potomstwa znajdzie się 50% samic i 50% samców. Z szachownicy Punneta wynika, że prawdopodobieństwo wystąpienia płci żeńskiej wynosi dwa przypadki z czterech, czyli 2/4=0,5 (50%). Podobnie występują tam dwa samce na cztery możliwości tj. 2/4=0,5 (50%). Jest to praktyczny przykład definicji prawdopodobieństwa. Znając rozkład płci u potomstwa wynoszący 1:1 wiemy, że prawdopodobieństwo wystąpienia każdej z dwóch alternatywnych płci wynosi po 0,5 (50%).



Rycina 1. Schemat dziedziczenia chromosomów płciowych u gupików w systemie zwanym *Drosophila*

W hodowli ryb akwariowych spotykamy się najczęściej z licznym potomstwem, czyli od jednej pary tarlaków otrzymujemy jednorazowo dziesiątki czy setki sztuk narybku. Zgodnie z powyższym wywodem w grupie 20 nowonarodzonych gupików spodziewamy się, że 10 spośród nich to samice, a pozostałe 10 to samce. Nie można jednak wykluczyć sytuacji, że wśród narybku będzie 11 samic i 9 samców, co potwierdzają obserwacje praktyczne hodowców, w których to najczęściej obserwowane proporcje płci są bardzo bliskie teoretycznemu stosunkowi, a rzadziej stwierdzana jest sytuacja modelowa. Dzieje się tak dlatego, iż proces zapłodnienia jest zjawiskiem losowym, a zapłodnienie jednej komórki jajowej w danej chwili konkretnym plemnikiem, w żaden sposób nie ma wpływu na to, jaki inny plemnik zapłodni inną z komórek jajowych. Są to więc zdarzenia niezależne, gdyż wystąpienie jednego zdarzenia, np. epizodu zapłodnienia nie ma żadnego wpływu na

występowanie drugiego zapłodnienia. W procesie rozrodu jednej pary ryb, najczęściej zapłodnień jest wiele, czyli mamy wiele zdarzeń niezależnych, czyli za każdym razem prawdopodobieństwo wystąpienia jednego z alternatywnych wariantów płci, wynosi 0,5 (50%).

Jakie jest więc prawdopodobieństwo, że wybrana losowo para ryb ma wśród dwojga potomków samych synów? Często spotkać się można z błędnym rozumowaniem wychodzącym z założenia, że skoro istnieją trzy rodzaje możliwych kombinacji: dwa samce, dwie samice oraz rodzeństwo różnopłciowe, to prawdopodobieństwo każdej z tych kombinacji wynosi 1/3 (≈33%). Choć może się to w pierwszym momencie wydawać racjonalne, to jednak nie jest zgodne z rzeczywistością. Natomiast uwzględniając kolejność narodzin, są nie trzy a cztery kombinacje rodzeństwa czyli: dwa samce, dwie samice, samiec i samica oraz samica i samiec. Czyli tym razem prawdopodobieństwo każdej z tych kombinacji to 1/4 (25%). Zauważyć należy, że rodzeństwo różnopłciowe to dwie z możliwych kombinacji, czyli napotkanie takiej sytuacji to 2/4 (50%).

Jakie jest więc prawdopodobieństwo, że wybrana losowo para ryb, ma wśród dwojga potomków samych synów jeśli wiadomo, że jeden z nich na pewno jest samcem? Spotykane dość często błędne rozumowanie sugeruje , że skoro drugi z potomków może być samcem lub samicą, to możliwości są dwie, czyli prawdopodobieństwo to wynosi 1/2 (50%) . Powyższe wnioskowanie wydaje się być poprawne, ale tak nie jest. Skoro wśród rodzeństwa jest już jeden samiec to z możliwych kombinacji trzeba wykluczyć rodzeństwo składające się dwóch samic. Skoro tak, to możliwe kombinacje to albo dwóch synów albo para różnopłciowa (samiec i samica lub samica i samiec). Oznacza to, że prawdopodobieństwo jednopłciowej pary rodzeństwa składającej się z dwóch samców w tym przypadku wynosi 1/3 ((≈33%)

Rozkład prawdopodobieństwa występowania płci w grupie potomstwa jest przykładem rozkładu dwumianowego, którego nazwa jest pochodną faktu, że istnieją dwa alternatywne warianty zdarzenia czyli w tym przypadku płeć męska lub żeńska. Jest to doskonały przykład ilustrujący proste zasady rachunku prawdopodobieństwa. Wszelkie obliczenia związane z tym rozkładem. przeprowadzić można jedynie operując pojedynczymi prawdopodobieństwami zdarzeń, bez korzystania ze wzorów.

Dla ułatwienia wywodu i zrozumienia omawianego zagadnienia rozpatrywany „miot" nowonarodzonego potomstwa pary gupików będzie liczył 6 sztuk. W takiej sytuacji można sobie wyobrazić, iż istnieje 7 możliwych rozkładów płci. Zgodnie z tym co już napisano powyżej, najbardziej prawdopodobna jest sytuacja, gdy samców i samic jest tyle samo czyli po 3 sztuki. Jednak nie można całkowicie wykluczyć żadnej z pozostałych możliwości czyli przewagi liczebnej którejś z płci, nawet takiej, że cały „miot" będzie jednopłciowy, choć jest to sytuacja, którą intuicyjnie określamy jako najmniej prawdopodobną. Można jednak w swych rozważaniach być bardziej precyzyjnym, czyli zamiast operować ogólnikowymi stwierdzeniami „mniej prawdopodobne" lub „bardziej prawdopodobne", określić dokładnie prawdopodobieństwo każdej z możliwych sytuacji. Kluczem do rozwiązania jest tu wspomniany już rozkład dwumianowy a konkretnie algebraiczne rozwinięcie dwumianu, znane najczęściej pod postacią $(a+b)^n$, czyli przykładowo $(a+b)^2=a^2+2ab+b^2$. W rozważaniach genetycznych rozkład dwumianowy najczęściej zapisywany jest w postaci $(p+q)^n$ gdzie literą „p" oznaczono prawdopodobieństwo wystąpienia jednego zdarzenia (zapłodnienia komórki jajowej przez plemnik z chromosomem X czyli powstania samicy) a literą „q" oznaczono prawdopodobieństwo wystąpienia zdarzenia przeciwnego (zapłodnienia komórki jajowej przez plemnik z chromosomem Y czyli powstania samca). Symbole „p" i „q" nazywane są parametrami rozkładu dwumianowego. Po dodaniu do siebie obu tych prawdopodobieństw dostajemy wynik 1 (100%) gdyż suma wszystkich zdarzeń czyli 50% samic i 50% samców to suma wszystkich osobników.

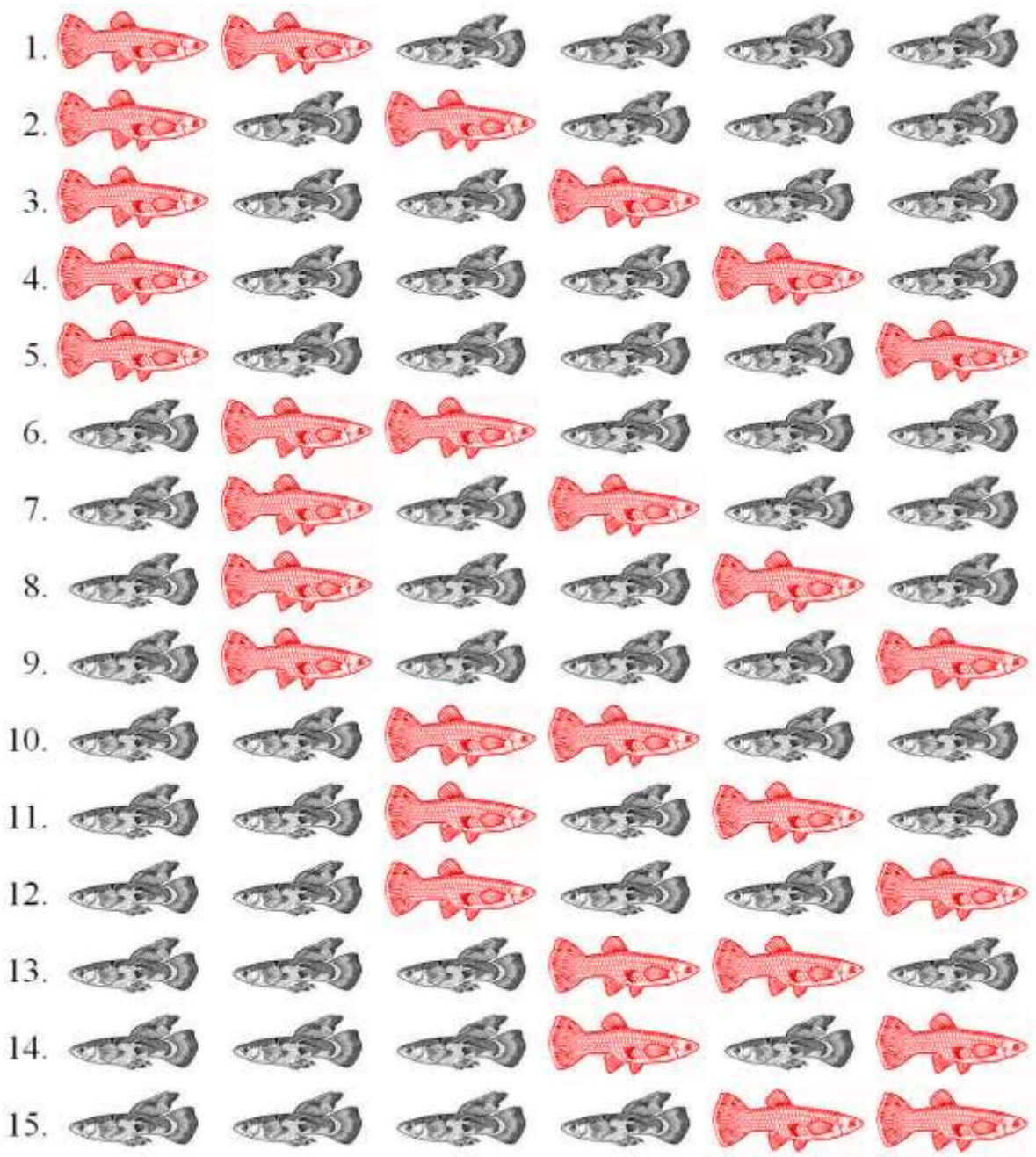
Prawdopodobieństwo pojawienia się wszystkich możliwych frekwencji płci w grupie 6 osobników odpowiada rozwinięciu $(p+q)^n$ a konkretnie $(p+q)^6$. Jak go wyliczyć? Najprostszym sposobem nie wymagającym korzystania ze wzorów matematycznych jest przeprowadzenia całej operacji w formie tabelarycznej (tabela 1).

Tabela 1. Tabelaryczne obliczanie prawdopodobieństw. Prawdopodobieństwo każdej z płci to 50% a liczebność grupy potomstwa to 6 osobników (p=0,5, q=0,5, n=6)

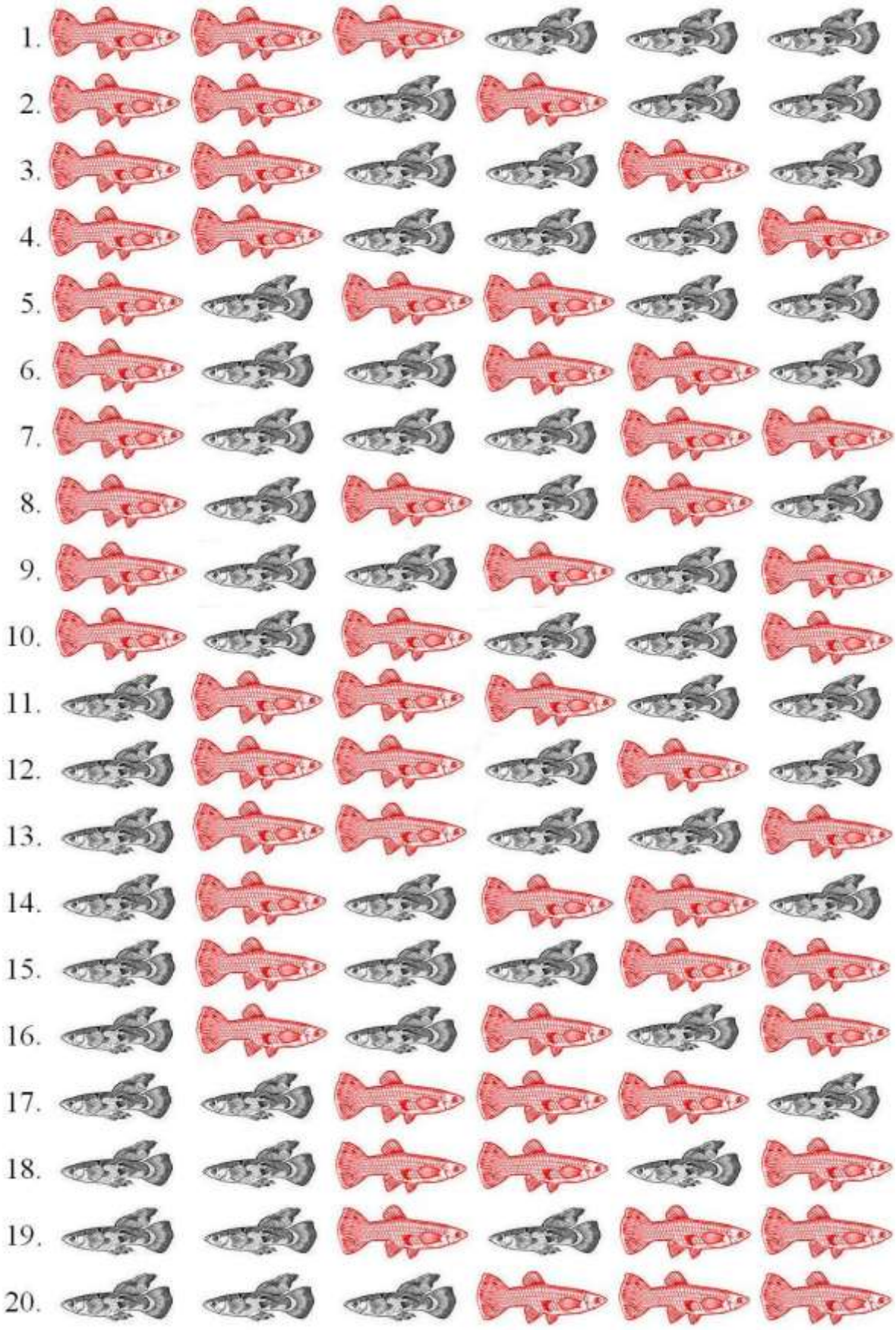
L.p.	Liczebność		Współczynniki dwumianu	Wyrażenia dwumian	Prawdopodobieństwo	
	Samce (s)	Samice (t)				% (≈)
1	6	0	1	$p\times p\times p\times p\times p\times p = p^6$	0,015625	2%
2	5	1	6	$p\times p\times p\times p\times p\times q = p^5q^1$	0,09375	9%
3	4	2	15	$p\times p\times p\times p\times q\times q = p^4q^2$	0,234375	23%
4	3	3	20	$p\times p\times p\times q\times q\times q = p^3q^3$	0,3125	31%
5	2	4	15	$p\times p\times q\times q\times q\times q = p^2p^4$	0,234375	23%
6	1	5	6	$p\times q\times q\times q\times q\times q = p^1q^5$	0,09375	9%
7	0	6	1	$p\times p\times p\times p\times p\times p = q^6$	0,015625	2%

Najpierw w tabeli rozpisuje się w osobnych kolumnach możliwe liczebności samic i samców. Pierwsza sytuacja to urodzenie 6 samców, czyli 6 kolejno rodzących się rybek to samce, a prawdopodobieństwo takiej sytuacji to iloczyn sześciu prawdopodobieństw narodzin samca czyli $p\times p\times p\times p\times p\times p=p^6$. Kolejna możliwość to narodzenie 5 samców i 1 samiczki ($p\times p\times p\times p\times p\times q$) przy czym samiczka może urodzić się jako pierwsza (a potem kolejno po sobie 5 samców) albo jako druga (czyli najpierw samiec, potem samiczka a następnie kolejno 4 pozostałe samce), albo trzecia, albo czwarta, albo piąta i wreszcie może też urodzić się jako ostatnia 6 rybka (po 5 braciach). Stwierdzono więc, że sytuacja w której grupa potomstwa to 5 samców i 1 samiczka może powstać na 6 różnych równie możliwych sposobów co trzeba uwzględnić w obliczeniach jako „współczynnik dwumianu" ($6\times p\times p\times p\times p\times p\times q=6p^5q^1$). Jeśli w „miocie" są 2 samiczki to mogły się one urodzić jako: pierwsze dwie, albo druga i trzecia itd. aż do piątej i szóstej w kolejności narodzin. Może się jednak zdarzyć, że samiczka urodzi się jako pierwsza i piata albo trzecia i szósta, a wszystkich możliwości jest 15 (rycina 2). W przypadku 3 samiczek w grupie 6 nowonarodzonych ryb, kolejność narodzin samic i samców może przyjąć 20 różnych sekwencji (rycina 3). Gdy samiczek jest 4 to ilość możliwych kombinacji wynosi ponownie 15, a gdy samiczek jest 5 to możliwe jest 6 różnych kolejności narodzin. Natomiast tylko jeden wariant to wszystkie 6 osobników płci żeńskiej.

W ten sposób wydedukowaliśmy jakie są współczynniki rozwinięcia dwumianu Newtona i można je wpisać do tabeli w kolejnej kolumnie.

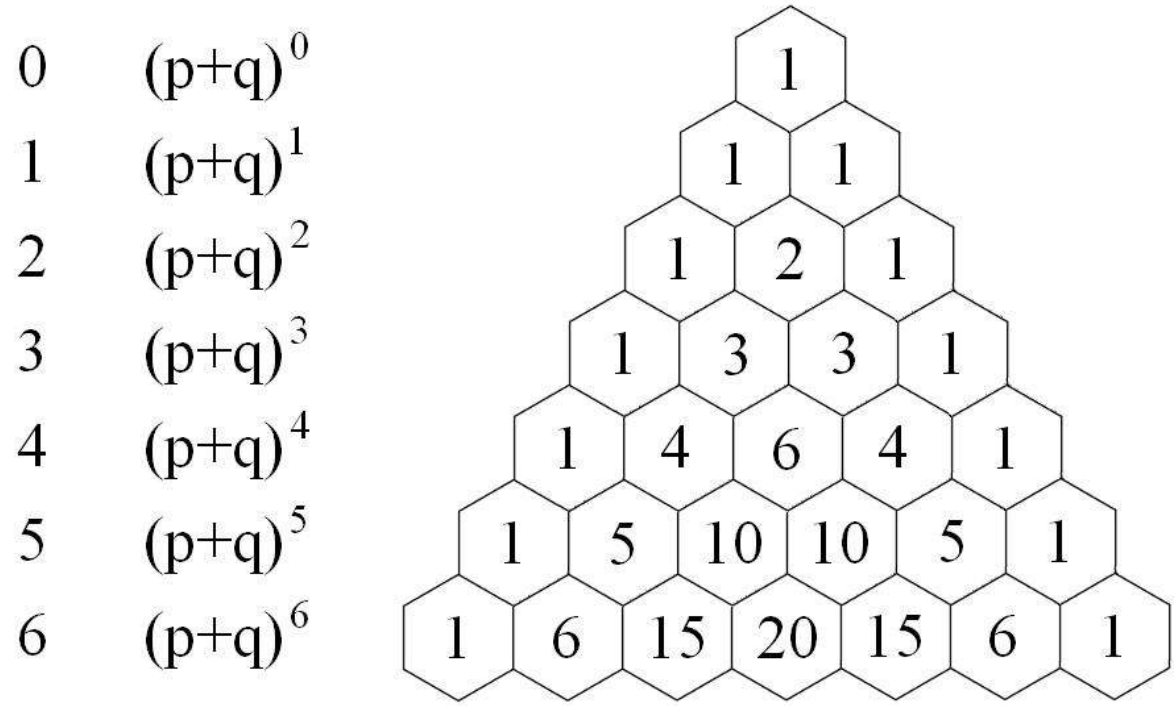


Rycina 2. Możliwa kolejność narodzin 2 samic w miocie liczącym 6 sztuk ryb. Jest 15 różnych sposobów kolejności narodzin takiego rozkładu płci. (czerwone zwrócone pyskiem w prawo osobniki to samice a czarne zwrócone pyskiem w lewo to samce, kolejne ponumerowane wiersze to kolejne potencjalne grupy rodzeństwa)



Rycina 3. Możliwa kolejność narodzin 3 samic w miocie liczącym 6 sztuk ryb. Jest 20 różnych sposobów kolejności narodzin takiego rozkładu płci.

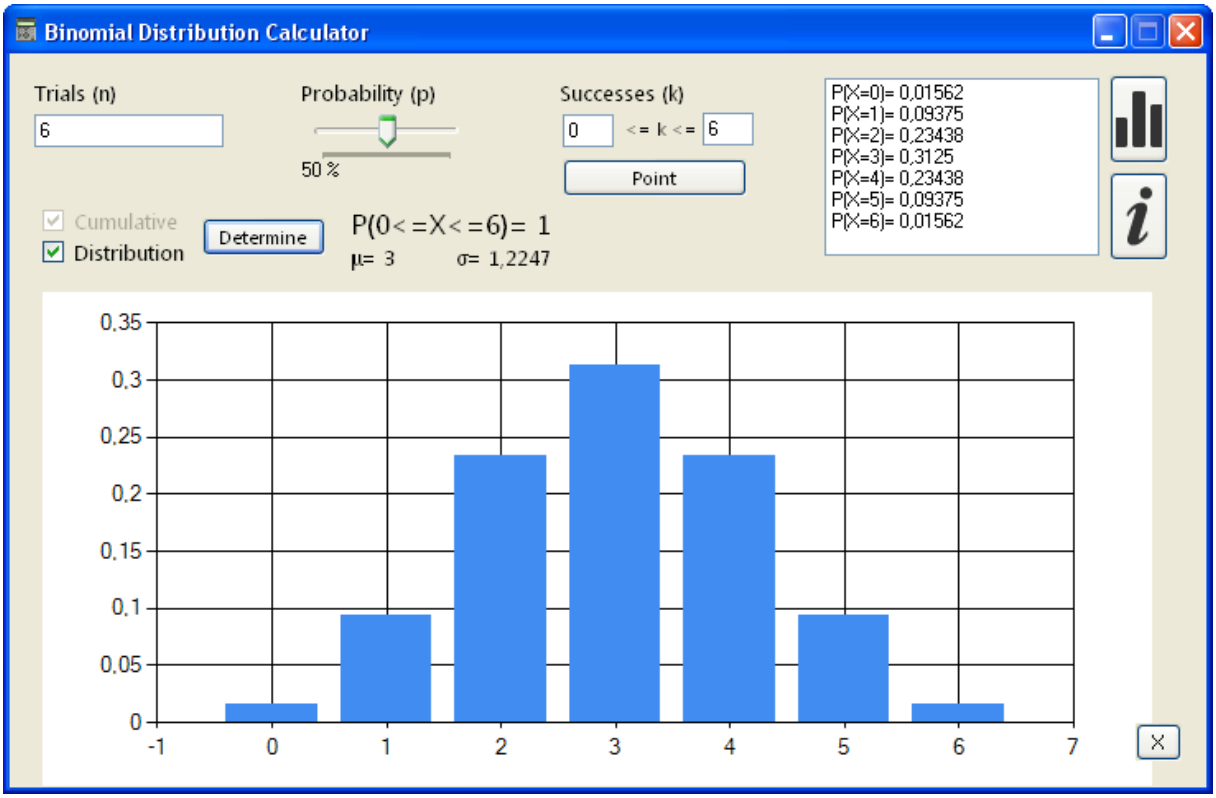
W praktyce najłatwiej współczynniki rozwinięcia dwumianu Newtona dla dowolnego „n” określa się na podstawie trójkąta Pascala. W trójkącie Pascala, każda liczba oprócz jedynek na początku i na końcu każdego wiersza, jest równa sumie dwóch liczb nad nią. Każdy wiersz trójkąta Pascala to nic innego jak współczynniki rozwinięcia odpowiedniej potęgi dwumianu (rycina 4).



Rycina 4. Trójkąt Pascala

W kolejnej kolumnie tabeli umieszczamy wyrażenia dwumianu czyli dla wariantu narodzin 6 samców jest to prawdopodobieństwo narodzin samca „p” podniesione do potęgi 6 (prawdopodobieństwa samic nie uwzględniamy, bo w tym wariancie nie występują, a jeśli chcielibyśmy je uwzględnić to byłoby to q do potęgi 0 czyli wynik zawsze wynosi 1). Dla sytuacji gdy w potomstwie jest 5 samców i 1 samica prawdopodobieństwo narodzin samca podnosi się do potęgi 5, a prawdopodobieństwo narodzin samicy do potęgi 1 (p^5q^1). Gdy samców będzie 4 a samic 2, prawdopodobieństwo narodzin samca podnosi się do potęgi 4, a prawdopodobieństwo narodzin samicy do potęgi 2 (p^4q^2). Natomiast gdy samców i samic będzie po 3, prawdopodobieństwo narodzin samca podnosi do potęgi 3 podobnie jak prawdopodobieństwo narodzin samicy (p^3q^3). Analogicznie postępuje się w dalszych etapach, aż dochodzimy do sytuacji gdy cała grupa to samice, więc tylko prawdopodobieństwo narodzin samicy podnosi się do potęgi 6 (q^6). Podstawiając wartość 0,5 (50%) pod „p” i q”

można wyliczyć już cząstkowe wyniki. Jako, że prawdopodobieństwo narodzin samca i samicy jest identyczne i wynosi 0,5 wynikiem wszystkich tych działań jest liczba 0,015625 (jest tak dlatego, że prawdopodobieństwo obu zdarzeń jest równe sobie, ale tak nie musi być, o czym mowa będzie w dalszej części). Trzeba teraz jedynie pomnożyć ten wynik przez współczynnik dwumianu dla każdej z rozpatrywanych sytuacji, a otrzymana wartość to prawdopodobieństwa dla poszczególnych rozkładów płci. Jest to np. 20×0,015625=0,3125 czyli około 31% szans, że obserwowany rozkład będzie 3 samce i 3 samice i jest to najwyższy z otrzymanych wyników a 1×0,015625=0,015625 czyli około 2% szans na to, że cały „miot" będzie jednopłciowy męski i tyle samo, że będzie jednopłciowy żeński. Otrzymano więc konkretne odpowiedzi na pytanie, jakiego rozkładu płci u potomstwa można się spodziewać wśród 6 osobników. Tak jak przewidywano z najwyższym prawdopodobieństwem będzie to rozkład 3 samce i 3 samice ale teraz wiadomo dokładnie jakie jest to prawdopodobieństwo. Wiadomo też konkretnie jakie są prawdopodobieństwa pozostałych hipotetycznych sytuacji włącznie z jednopłciowością całego potomstwa. Rozkład obliczonych prawdopodobieństw z tabeli 1 w formie graficznej prezentowany jest na rycinie 5. Jest to rozkład symetryczny względem najczęstszego wyniku tj. proporcji 1 : 1.



Rycina 5. Rozkład prawdopodobieństw z tabeli 1 obliczony za pomocą programu Binomial Distribution Calculator.

Współczynniki dwumianów to nic innego jak kombinacja bez powtórzeń, czyli można je wyliczyć za pomocą symbolu Newtona. Przy pomocy tabelarycznych obliczeń krok po kroku odtworzono mimo chodem wzór na obliczanie tego prawdopodobieństwa, czyli:

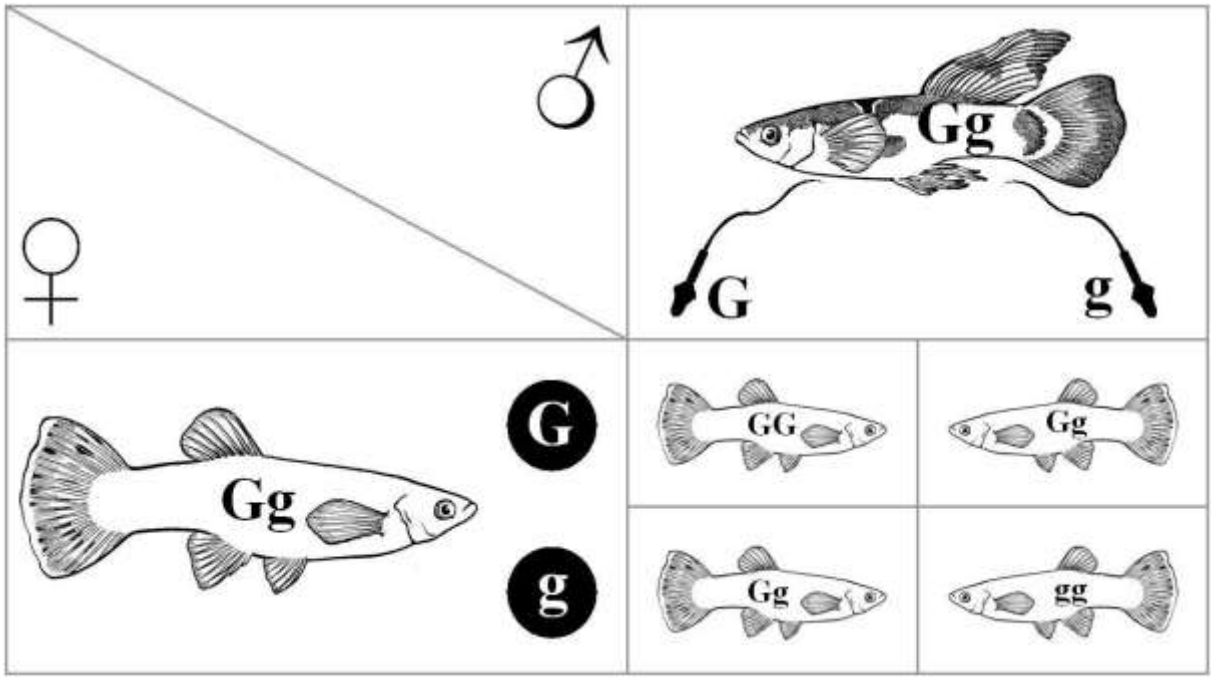
$$\Pr. = \frac{n!}{s! \times t!} \times p^s \times q^t$$

gdzie:
n – liczebność próby np. liczebność miotu
s – liczba osobników jednej z grup np. samców
t – liczba osobników grupy alternatywnej np. samic
p – częstość (prawdopodobieństwo) występowania osobników jednej grupy np. samców
q – częstość (prawdopodobieństwo) występowania osobników grupy alternatywnej np. samic

We wzorze tym pojawia się znak silni czyli „!" i oblicza się ją w ten sposób, że mnoży się przez siebie wszystkie liczby naturalne od 1 do tej konkretnej liczby np. dla 3! będzie to 1×2×3 = 6. Wzór powyższy jest niczym innym jak uogólnieniem obliczeń wykonywanych tabelarycznie ale nie ma konieczności korzystaniaz niego, a tym bardziej zapamiętywania lecz wystarczy postępować krok po kroku zgodnie z metodą prezentowaną w tabeli 1.

Prawdopodobieństwo w rozkładzie dwumianowym można więc obliczyć korzystając z prostych reguł rachunku prawdopodobieństwa czyli stosując mnożenie prawdopodobieństwa zdarzeń niezależnych (metoda w tabeli), albo korzystać z rozwinięcia dwumianu albo stosując powyższy wzór. Wynik uzyskany każdą z tych metod będzie identyczny.

Jak już wspomniano prawdopodobieństwa obu alternatywnych zdarzeń nie muszą być sobie równe czyli wynosić po 50% jak to ma miejsca w przypadku płci. Za przykład posłuży model dziedziczenia złotego ubarwienia u gupików warunkowany recesywnym allelem ***g*** podczas gdy dominujący wariant ***G*** odpowiada za dzikie umaszczenie (szare). Nie można odróżnić homozygot dominujących ***GG*** od heterozygot ***Gg*** po fenotypie, gdyż wszystkie one są szare. Natomiast homozygoty ***gg*** są wyróżnialne na podstawie fenotypu czyli złotej barwy. Kojarząc dwa heterozygotyczne osobniki otrzymuje się w potomstwie wszystkie trzy możliwe genotypy oraz dwa fenotypy tj. osobniki szare i złote. Stosunek fenotypów w potomstwie to 3 osobniki szare przypadające na jednego osobnika złotego (rycina 6). Prawdopodobieństwo narodzin ryby szarej to 3/4=0,75 (75%) a prawdopodobieństwo narodzin ryby złotej to 1/4=0,25 (25%). Więc tym razem nie są to prawdopodobieństwa sobie równe, ale ich suma nadal wynosi 1 (100%).

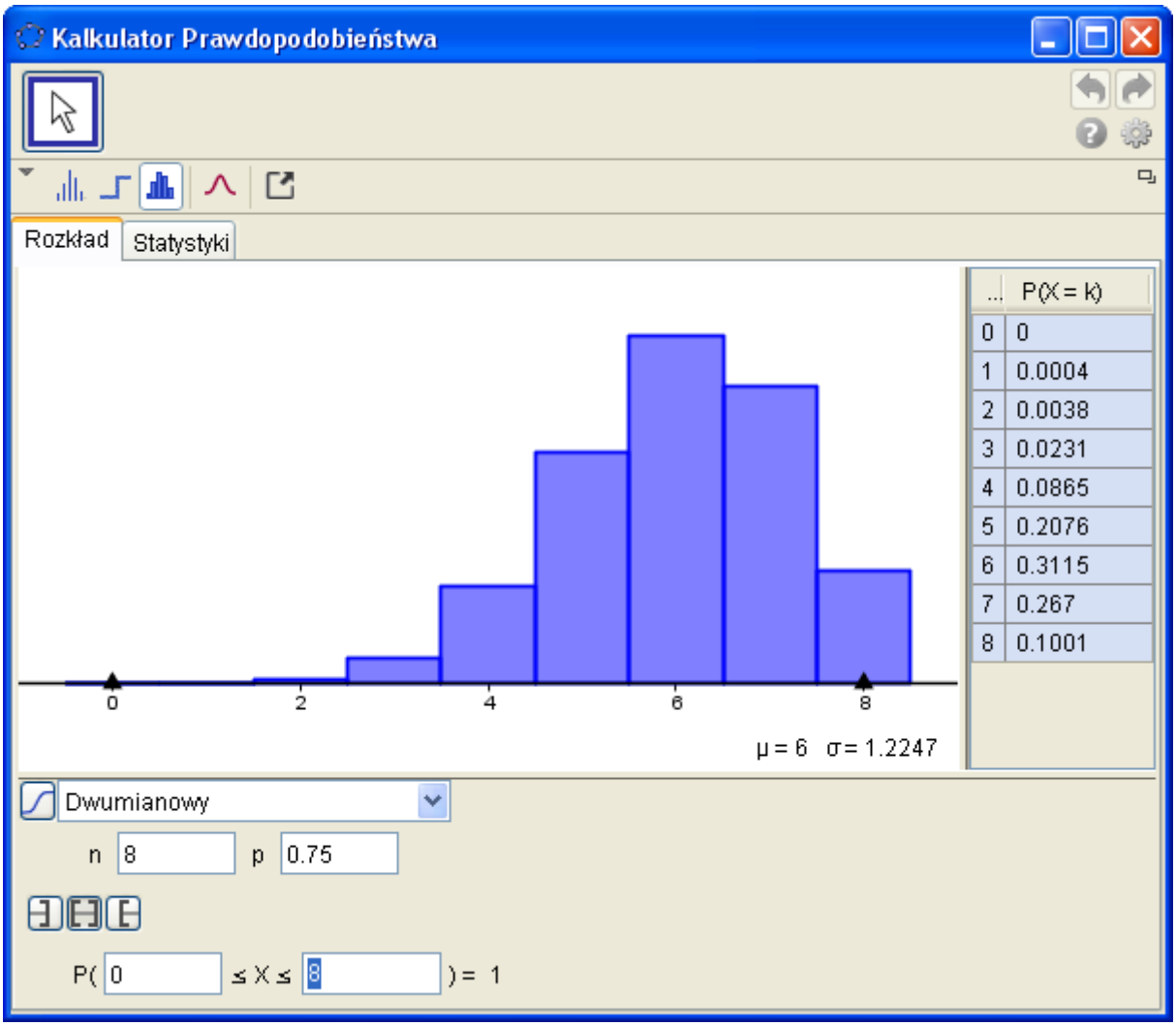


Rycina 6. Schemat dziedziczenia złotego ubarwienia u gupików w systemie zwanym *pisum* (pełna dominacja)

W grupie rodzeństwa liczącej 8 ryb spodziewać się można z największym prawdopodobieństwem 6 ryb szarych i 2 ryb złotych. Nikogo jednak raczej nie zdziwi, gdy będzie to 7 szarych ryb i 1 złota. Nadal możliwe jest, że wszystkie 8 będzie szarych lub, że wszystkie 8 będzie złotych. Teraz jednak intuicja podpowiada, że bardziej prawdopodobna jest przewaga szarych nad złotymi, a mniej prawdopodobna jest sytuacja odwrotna czyli przewaga złotych nad szarymi. Uzasadnieniem tego jest fakt, że prawdopodobieństwo narodzin szarej ryby jest wyższe niż osobnika złotego. Obliczeń ponownie można dokonać metodą tabelaryczną (tabela 2). Tym razem prawdopodobieństwo „p" wynosi 0,75 a prawdopodobieństwo „q" to 0,25. Liczba osobników to 8 sztuk, więc w trójkącie Pascala współczynniki znajdują się dwa wiersze niżej, niż w poprzednim przykładzie dotyczącym płci. Otrzymane wyniki pozwalają stwierdzić że najbardziej prawdopodobna (31%) jest sytuacja, w której 6 ryb jest szarych, a 2 są złote. Prawdopodobieństwo, iż wszystkie 8 rybek będzie szarych wynosi 10% czyli co dziesiąty „miot" będzie reprezentował taki rozkład fenotypów. Natomiast rozkład fenotypów 1 do 1 czyli 4 osobniki szare i 4 złote, jest już mniej prawdopodobny (9%) niż obecność tylko szarych ryb (10%). Sytuacja, w której wszystkie 8 ryb będzie miało złoty fenotyp, występuje 1 raz na 50 tysięcy takich kojarzeń czyli prawdopodobieństwo takiego rozkładu płci w potomstwie wynosi 0,002%. Rozkład prawdopodobieństw z tabeli 2 w formie graficznej prezentowany jest na rycinie 7.

Tabela 2. Tabelaryczne obliczanie prawdopodobieństw. Prawdopodobieństwo narodzin osobnika szarego to 75% a złotego to 25% a liczebność grupy potomstwa to 8 osobników (p=0,75, q=0,25, n=8)

L.p.	Liczebność		Współczynniki dwumiany	Wyrażenia dwumian	Prawdopodobieństwo	
	Samce (s)	Samice (t)				% (≈)
1	8	0	1	$p \times p \times p \times p \times p \times p \times p \times p = p^8$	0,100113	10%
2	7	1	8	$p \times p \times p \times p \times p \times p \times p \times q = p^7 q^1$	0,266968	27%
3	6	2	28	$p \times p \times p \times p \times p \times p \times q \times q = p^6 q^2$	0,311462	31%
4	5	3	56	$p \times p \times p \times p \times p \times q \times q \times q = p^5 q^3$	0,207642	21%
5	4	4	70	$p \times p \times p \times p \times q \times q \times q \times q = p^4 q^4$	0,086517	9%
6	3	5	56	$p \times p \times p \times q \times q \times q \times q \times q = p^3 q^5$	0,023071	2%
7	2	6	28	$p \times p \times q \times q \times q \times q \times q \times q = p^2 q^6$	0,003845	0,4%
8	1	7	8	$p \times q \times q \times q \times q \times q \times q \times q = p^1 q^7$	0,000366	0,04%
9	0	8	1	$q \times q \times q \times q \times q \times q \times q \times q = q^8$	0,000015	0,002%

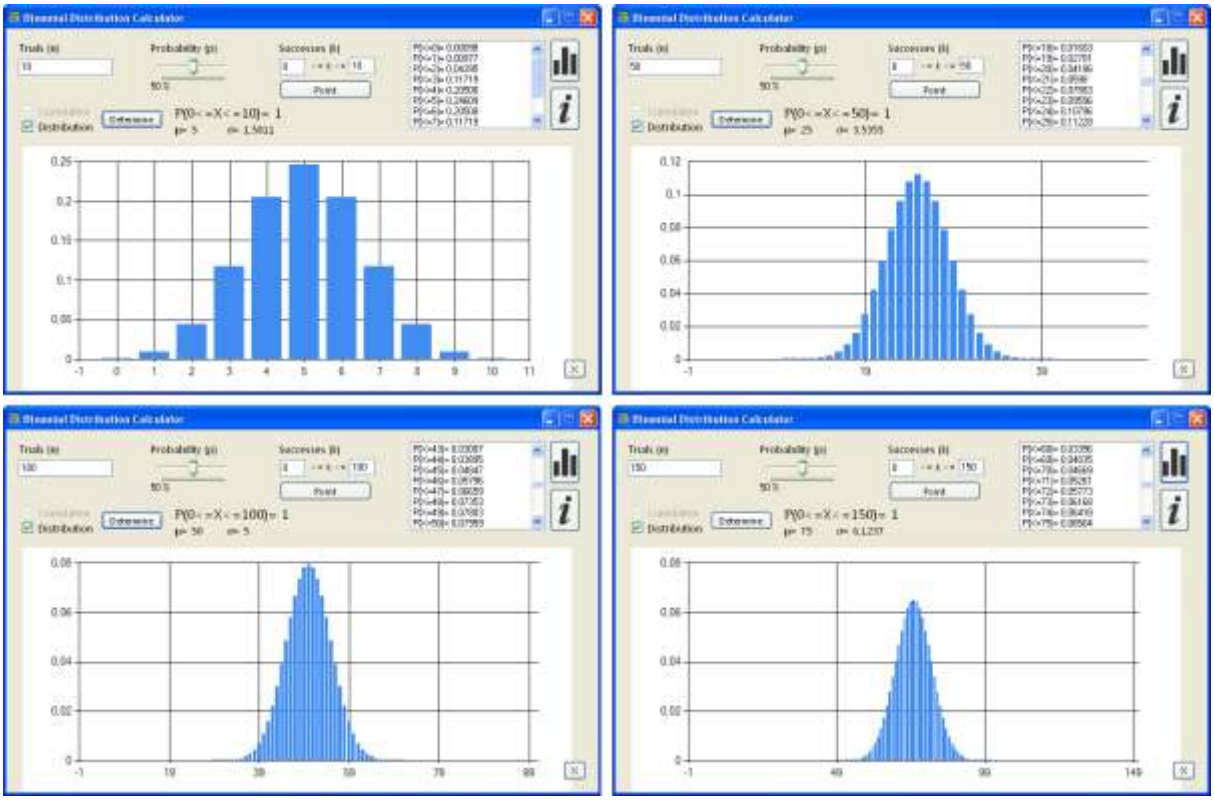


Rycina 7. Rozkład prawdopodobieństw z tabeli 2 obliczony za pomocą programu GeoGebra

Wyniki z tabeli 2 przedstawione na rycinie 7 uwidaczniają asymetryczny rozkład prawdopodobieństwa względem najczęstszej spodziewanej proporcji tj. 3 : 1. Jest to asymetria lewostronna, czyli rozkład nazywamy lewoskośnym lub ujemnie skośnym. Lewostronna asymetria dotyczy sytuacji gdy $p>0,5$ a jeśli $p<0,5$ to asymetria będzie prawostronna.

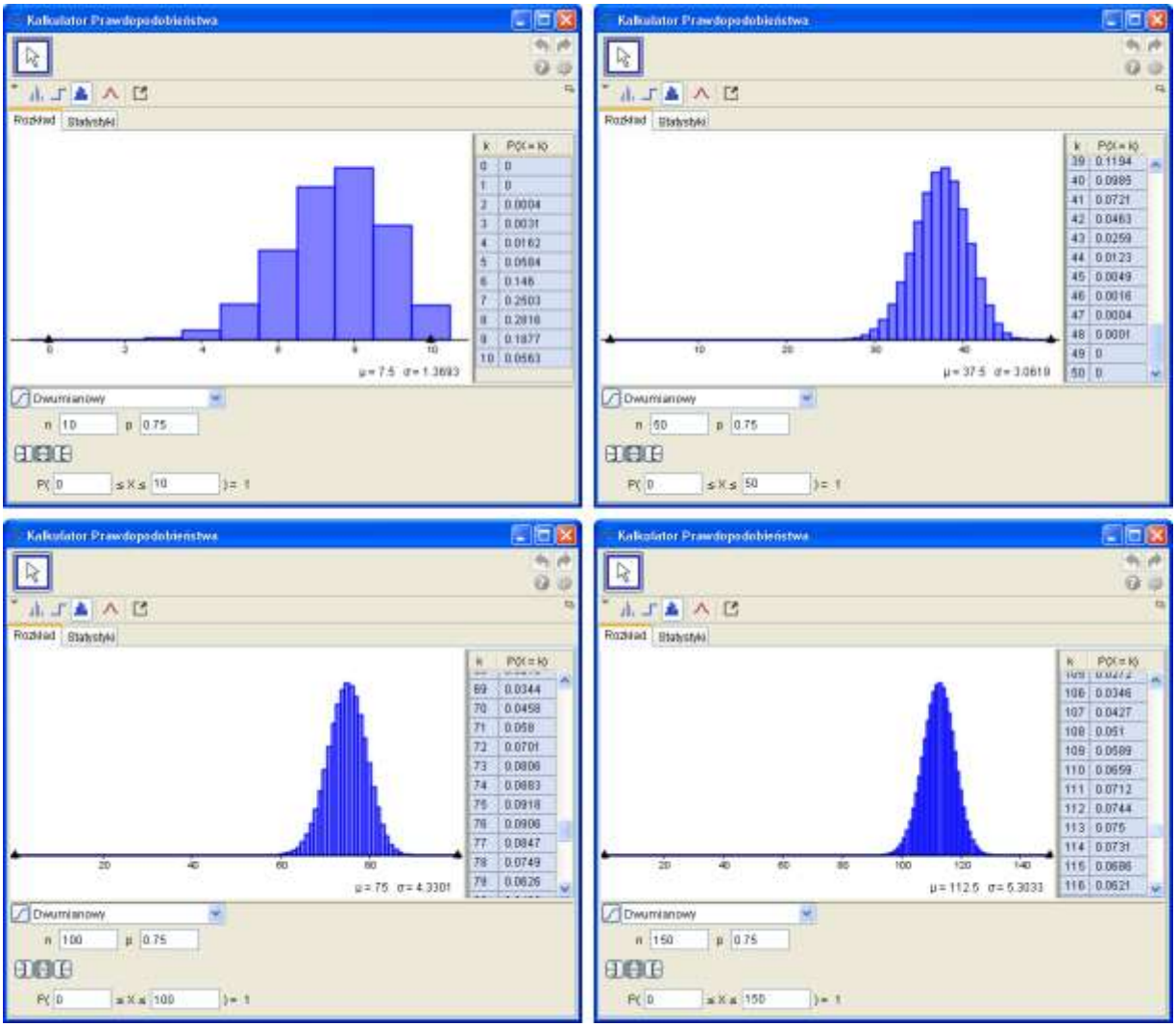
Pojawić się może pytanie jak zmieniały się będą rozkłady prawdopodobieństwa wraz ze wzrostem liczby potomstwa. Intuicja podpowiada, że im liczniejsza jest grupa, tym bardziej prawdopodobne jest otrzymanie wyniku „idealnego" z modelem, a coraz mniej stają się prawdopodobne rozkłady skrajnie odbiegające od teoretycznego. By to sprawdzić nie trzeba wykonywać żmudnych obliczeń w tabeli lub za pomocą wzoru, lecz skorzystać można z odpowiedniego oprogramowanie komputerowego np. „Binomial Distribution Calculator" lub „GeoGebra".

Wyniki prezentowane na rycinie 8 to rozkłady prawdopodobieństwa przy zwiększającej się liczbie potomstwa czyli, n=10, n=50, n=100 i n=150 oraz przy p=q=0,5 (czyli przykład z proporcją płci 1 : 1). Jak widać prawdopodobieństwo modelowego rozkładu identycznej liczby samic i samców, staje się coraz mniej prawdopodobne wraz ze wzrostem liczby potomstwa.



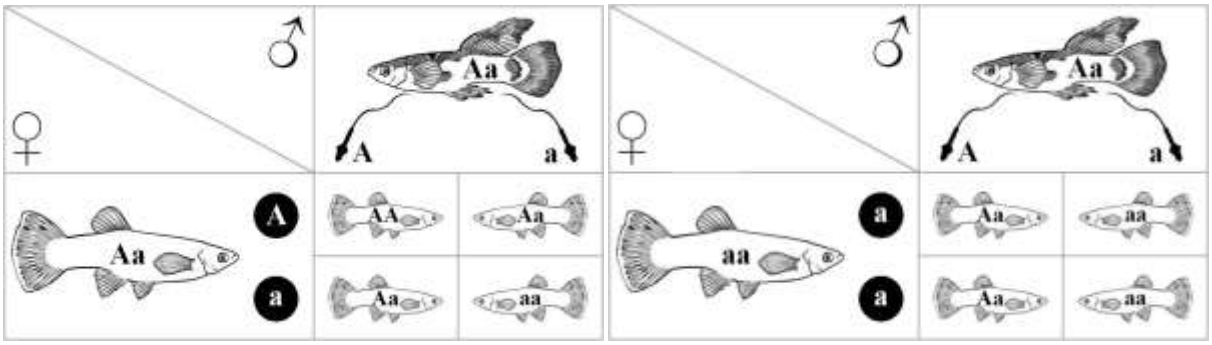
Rycina 8. Rozkład prawdopodobieństw z tabeli 1 przy zwiększającej się liczbie potomstwa n=10, n=50, n=100 i n=150 obliczone za pomocą programu Binomial Distribution Calculator

Zarówno dla rozkładu symetrycznego (rycina 8) jak i asymetrycznego (rycina 9) zauważalna jest ogólna prawidłowość, czyli wraz ze wzrostem liczby zdarzeń (liczby potomstwa) obserwowany rozkład coraz bardziej przypomina znany każdemu przyrodnikowi rozkład normalny, zwany krzywą Gaussa. Dzieje się tak dlatego, że rozkład normalny jest przybliżeniem wartości trójkąta Pascala dla bardzo dużej liczby prób, czyli rozkład normalny jest uogólnieniem rozkładu dwumianowego. Jest to przykład centralnego twierdzenia granicznego mówiącego, że przy liczbie powtórzeń dążącej do nieskończoności, rozkład sum wartości niezależnych doświadczeń losowych po standaryzacji jest zbieżny do standardowego rozkładu normalnego. Wraz ze wzrostem liczebności, rozkład prawdopodobieństwa jest coraz mniej dyskretny co znaczy „oddzielone od siebie" a coraz bardziej ciągły.



Rycina 9. Rozkład prawdopodobieństw z tabeli 2 przy zwiększającej się liczbie potomstwa n=10, n=50, n=100 i n=150 obliczony za pomocą programu GeoGebra

Jakie praktyczne znaczenie mogą mieć powyższe rozważania na temat prawdopodobieństwa? Jednym z zastosowań jest testowanie tarlaków pod kątem nosicielstwa recesywnych alleli. Przykładowo tym recesywnym genem może być allel warunkujący albinizm.



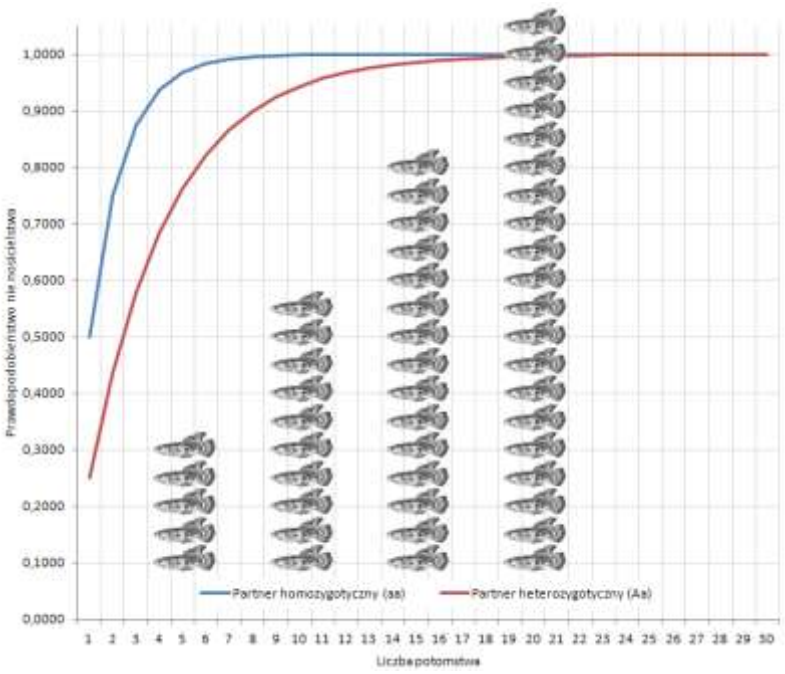
Rycina 10. Kojarzenie testowe tarlaków pod kątem nosicielstwa recesywnych alleli (kojarzenie z partnerem heterozygotycznym (A) oraz partnerem homozygotycznym (B))

Testowanie czy dziko ubarwiony samiec gupika jest nosicielem recesywnego genu albinizmu polega na skojarzeniu go z samicą o znanym genotypie. Może to być samica o której już wiadomo, że jest nosicielką recesywnego genu albinizmu (rycina 10A). Jeszcze lepiej kojarzyć tego samca z homozygotyczną albinoską (rycina 10B). W pierwszym przypadku, jeśli samiec jest heterozygotą, to wśród potomstwa spodziewać się można 25% albinosów. W drugim przypadku odsetek albinosów byłby wyższy i wynosił 50%. Znając prawdopodobieństwo występowania poszczególnych fenotypów wśród potomstwa oraz uwzględniając fakt, iż można obliczyć prawdopodobieństwo obserwowanych odstępstw od teoretycznej modelowej proporcji, można obliczyć na podstawie fenotypów potomstwa prawdopodobieństwo nosicielstwo genu recesywnego przez ich ojca.

Jeżeli skojarzono dziko ubawionego samca o nieznanym genotypie (***AA*** lub ***Aa***) z samicą albinotyczną (***aa***) to prawdopodobieństwo narodzin homozygotycznego albinosa (***aa***) wynosi 0,5 a co za tym idzie prawdopodobieństwo, iż nowo narodzony osobnik będzie dziko ubarwiony (***AA*** lub ***Aa***) wynosi również 0,5. Gdy para tarlaków ma tylko jednego potomka dziko ubarwionego to testowany ojciec jest nosicielem genu albinizmu z prawdopodobieństwem równym prawdopodobieństwu narodzin takiego osobnika czyli 0,5. Gdy w potomstwie uzyskano dwa dziko ubarwione osobniki to prawdopodobieństwo takiej sytuacji dla heterozygotycznego ojca wynosi $0,5\times0,5=0,5^2=0,25$ co oznacza, że jest on wolny od genu recesywnego z prawdopodobieństwem 1-0,25=0,75. Analogicznie gdy potomstwo stanowią 3 ryby dziko ubarwione, to sytuacja taka występuje z prawdopodobieństwem $0,5\times0,5\times0,5=0,5^3=0,125$ a to oznacza, że samiec nie jest nosicielem genu recesywnego

z prawdopodobieństwem 1-0,125=0,875. Im więcej będzie ryb dziko ubarwionych w potomstwie tym mniej prawdopodobne jest, iż jest to wynik losowego zapłodnienia a bardziej staje się prawdopodobne, że ojciec nie przekazuje genu recesywnego, gdyż go nie posiada (rycina 1). Jeżeli skojarzono dziko ubawionego samca o nieznanym genotypie (***AA*** lub ***Aa***) z samicą heterozygotyczną (***Aa***) to prawdopodobieństwo narodzin homozygotycznego albinosa (***aa***) wynosi 0,25 a co za tym idzie prawdopodobieństwo, iż nowo narodzony osobnik będzie dziko ubarwiony (***AA*** lub ***Aa***) wynosi 0,75. Gdy para tarlaków ma tylko jednego potomka i jest od dziko ubarwiony to testowany ojciec jest nosicielem genu albinizmu z prawdopodobieństwem równym prawdopodobieństwu narodzin osobnika dziko ubarwionego czyli 0,75 czyli nie jest nosicielem tego genu z prawdopodobieństwem 1-0,75=0,25. Gdy potomków jest dwóch i obie ryby są dziko ubarwione to sytuacja taka występuje z prawdopodobieństwem $0,75\times0,75=0,75^2=0,5625$ czyli samiec nie jest nosicielem recesywnego genu z prawdopodobieństwem 1-0,5625=0,4375. Analogicznie gdy potomstwo stanowią 3 ryby dziko ubarwione to sytuacja taka występuje z prawdopodobieństwem $0,75\times0,75\times0,75=0,75^3=0,4219$ a to oznacza, że samiec ten jest wolny od recesywnego genu z prawdopodobieństwem 1-0,4219=0,5781. Z każdym kolejnym dziko ubarwionym potomkiem mniej prawdopodobne jest, że to wynik losowego zapłodnienia, a bardziej staje się prawdopodobne, że ojciec nie przekazuje genu recesywnego, gdyż go nie posiada (rycina 11).

Liczba potomstwa	Partner homozygotyczny	Partner heterozygotyczny
1	0,5000	0,2500
2	0,7500	0,4375
3	0,8750	0,5781
4	0,9375	0,6836
5	0,9688	0,7627
6	0,9844	0,8220
7	0,9922	0,8665
8	0,9961	0,8999
9	0,9980	0,9249
10	0,9990	0,9437
11	0,9995	0,9578
12	0,9998	0,9683
13	0,9999	0,9762
14	0,9999	0,9822
15	1,0000	0,9866
16	1,0000	0,9900
17	1,0000	0,9925
18	1,0000	0,9944
19	1,0000	0,9958
20	1,0000	0,9968
21	1,0000	0,9976
22	1,0000	0,9982
23	1,0000	0,9987
24	1,0000	0,9990
25	1,0000	0,9992
26	1,0000	0,9994
27	1,0000	0,9996
28	1,0000	0,9997
29	1,0000	0,9998
30	1,0000	0,9998



Rycina 11. Testowanie tarlaków pod kątem nosicielstwa recesywnych alleli przy kojarzeniu ich partnerem homozygotycznym oraz partnerem heterozygotycznym

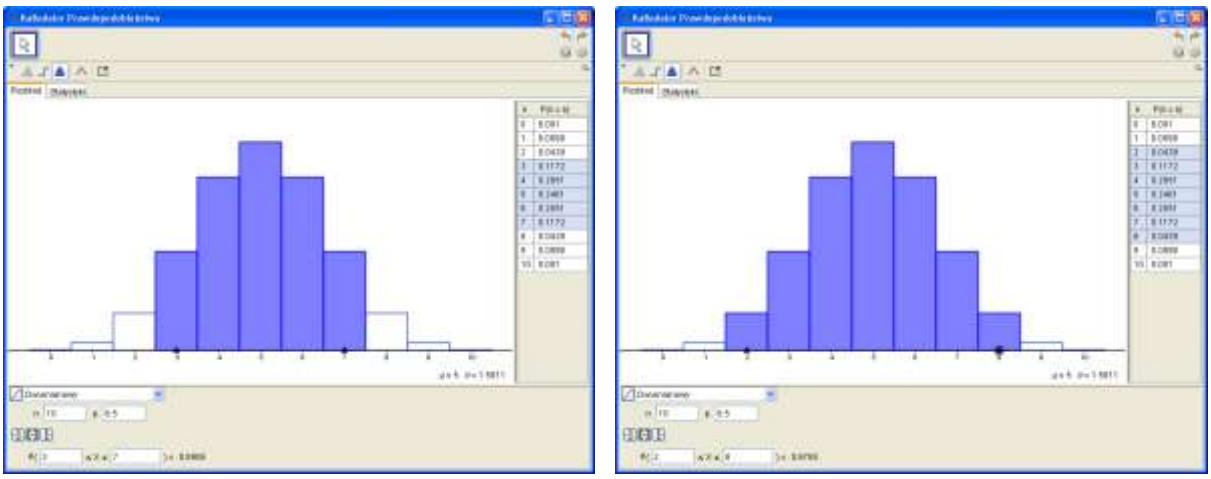
Kojarzenia testowe mające na celu określić genotyp interesującego nas osobnika, są doskonałym przykładem testowania hipotez statystycznych. Hipotezy można weryfikować przez ich potwierdzenie (konfirmacje) czyli szukając dowodu, że jest prawdziwa lub przez jej zaprzeczenie (falsyfikacje) czyli wykazując, że hipoteza jest fałszywa. Wbrew pozorom różnica między obydwoma sposobami jest zasadnicza i determinuje samą celować weryfikacji. Stawiamy więc hipotezę „Wszystkie gupiki są barwne (nie ma białych gupików)" czyli antytezą jest stwierdzenie „Nie wszystkie gupiki są barwne (istnieją białe gupiki)". Weryfikując tę hipotezę poprzez konfirmacje, szuka się dowodów potwierdzających jej prawdziwość. Gdy w grupie kilkunastu gupików wszystkie są barwne nie można stwierdzić jednoznacznie, że hipoteza jest prawdziwa, gdyż nie ma żadnej pewności, że gdzieś nie żyje biały gupik. Zwiększamy obserwowaną grupą do kilkuset osobników i mimo, że nadal wszystkie będą barwne, to jednak nie możemy z całym przekonaniem powiedzieć, że nie istnieją białe gupiki. Można zwiększać obserwowaną grupę do tysięcy i mimo, że nie znajdzie się w niej biały gupik to jednak nie wykażemy ostatecznie że nie ma żadnych białych gupików. Oznacza to, że konfirmacja nigdy nie jest konkluzywna ponieważ nigdy ostatecznie nie można wykazać prawdziwości postawionej hipotezy. Dlaczego? Dlatego, że zawsze może zdarzyć się jeden jedyny fakt czyli jeden biały gupik który to zakwestionuje prawdziwość tej hipotezy. Szukanie tego choćby jedynego białego gupika czyli faktu przeczącego hipotezie to falsyfikacja. Czyli jeśli znajdziemy białego gupika to z całą pewnością możemy stwierdzić, że hipoteza powyższa jest fałszywa. Jeśli takiego gupika nie znajdziemy to nie możemy hipotezy powyższej potwierdzić, a jedynie stwierdzamy, że jest ona mniej lub bardziej prawdopodobna. Dlatego hipotezy zawsze poddaje się falsyfikacji. Jest to dokładnie sytuacja, z jaką mamy do czynienia podczas testowania na nosicielstwo genu recesywnego. Stawiamy hipotezę „Osobnik ten nie jest nosicielem genu recesywnego" a potem szukamy dowodu na jej nieprawdziwość, czyli szukamy choćby jednego albinosa wśród potomstwa. Jeśli albinos taki zostanie odnaleziony zbędne są wszelkie obliczenia, gdyż jest to dowód, że testowany osobnik przekazuje recesywny allel. Jeśli natomiast wśród potomstwa brak jest albinosów to nie jest to dowodem że osobnik ten nie jest nosicielem genu recesywnego ale wraz ze wzrostem liczebności obserwowanego potomstwa staje się to coraz bardziej prawdopodobne. W którymś momencie należy podjąć decyzje czy zwiększać liczbę potomstwa czy uznać już, że osobnik ten jest wolny od recesywnego allelu. W naukach przyrodniczych najczęściej przy poziomie 95% pewności, uznajemy, że nie ma podstaw do odrzucenia hipotezy.

Rozpatrując uogólnioną prawidłowość rachunku prawdopodobieństwa, płynnie przechodzi się z probabilistyki do statystyki i jej najważniejszego zadania, jakim jest testowanie hipotez. Przykładem takiej hipotezy, może być pytanie czy rzeczywiście stosunek płci w populacji gupików wynosi 1:1 to znaczy czy proporcja płci w populacji wynosi p=0,5 (50%). Model dziedziczenia płci zakłada taki właśnie rozkład. W statystyce taką hipotezę nazywa się zerową, a jej zaprzeczenie czyli twierdzenie, że udział każdej z płci w populacji nie jest równy 50% to hipoteza alternatywna. Zadaniem testu statystycznego jest wybór jednej z nich na podstawie obiektywnej analizy. Analizy statystyczne przeprowadza się na podstawie obserwacji w próbie losowej, czyli nie odnotowuje się płci u wszystkich osobników populacji, lecz odławia się ich pewną część, zwaną próbą i w niej określa się ilość samic i samców. Przykładowo będzie to 10 osobników czyli liczba samic może wahać się w zakresie od 0 do 10, przy czym najbardziej prawdopodobna jest liczba 5 samic. Kluczową kwestią jest teraz określenie przy jakich liczbach samic w rozpatrywanej próbie, należy odrzucić hipotezę zerową, to znaczy przy jakich obserwowanych liczebnościach samic uznać można, że odstępstwo od zakładanego rozkładu jest na tyle duże, że modelowy rozkład nie występuje w populacji. Gdy liczba samic w próbie wynosi 5 to przyjmując prawdziwość hipotezy zerowej (p=0,5) unikamy pomyłki z prawdopodobieństwem 0,2461 (≈25%) bo jeśli proporcja samic w populacji to naprawdę 0,5 to z prawdopodobieństwem 25% odnajduje się ją w próbie. Odrzucając wszystkie inne proporcje płci w próbie przyjmujemy prawdopodobieństwo odrzucenia prawdziwej hipotezy zerowej na poziomie 0,7539 czyli na 75%. W statystyce błąd taki, czyli odrzucenie prawdziwej hipotezy zerowej nazywa się błędem pierwszego rodzaju. Odrzucenie prawdziwej hipotezy mówiącej o proporcji płci równej 1:1 z prawdopodobieństwem 75% to stanowczo za wiele. Zmniejsza się je akceptując w próbie również inną obserwowaną ilość samic np. o jedną więcej lub o jedną mniej, niż idealna ilość 5 czyli będą to wyniki 4 i 6. W takiej sytuacji unikamy pomyłki z prawdopodobieństwem 0,2461+0,2051+0,2051=0,6563 (≈66%) czyli bierzemy pod uwagę przypadki gdy samic jest dokładnie 5 oraz te gdy samic jest 4 i 6. Prawdopodobieństwo pomyłki polegającej na odrzuceniu prawdziwej hipotezy zerowej wynosi teraz 1-0,6563=0,3437 (34%). Jest to już znacznie lepszy wynik, choć nadal wysoki. Poszerzając zakres akceptowanych ilości samic w próbie o dodatkowo 3 i 7 unikamy pomyłki z prawdopodobieństwem 0,8906 (89%) czyli prawdopodobieństwo pomyłki spada do 0,1094 (11%) co jest już bardziej akceptowalne. Ogólnie przyjęło się w naukach biologicznych, że prawdopodobieństwo popełnienia błędu pierwszego rodzaju powinno być nie większe niż 5% tzn. dopuszczona jest 1 pomyłka na 20 razy, a wartość tą nazywa się poziomem istotności.

Na histogramie rozkładu prawdopodobieństwa poszukuje się takiego przedziału wyników których suma prawdopodobieństw wyniesie nie mniej niż 95%. W tym przypadku jest to zakres od 2 do 8 samic w próbie. Oznacza to, że jeśli w próbie znajduje się tylko 2 samice albo jest ich aż 8 to odrzucając hipotezę o proporcji płci 1:1 i przyjmując hipotezę o nierówności proporcji płci w populacji prawdopodobieństwo pomyłki wynosi 0,0215 (2%). Zakres tych wyników w statystyce nazywa się obszarem krytycznym. Jeśli w próbie znajduje się od 3 do 7 samic to nie ma podstaw by odrzucić hipotezę o równości frakcji płci w populacji. Jeśli w próbie znajduje się mniej niż 3 lub więcej niż 7 samic to są podstawy by odrzucić hipotezę o równości frakcji płci w populacji i przyjąć stwierdzenie, że w populacji stosunek płci odbiega od teoretycznego rozkładu czyli po 50% (tabela 3).

Tabela 3. Prawdopodobieństwa uniknięcia pomyłki oraz popełnienia błędu pierwszego rodzaju przy testowaniu hipotezy o proporcji płci równej 1:1 (n=10)

Zakres		prawdopodobieństwo uniknięcia pomyłki	prawdopodobieństwo błędu
od	do		
5	5	0,2461	0,7539
4	6	0,6562	0,3438
3	7	0,8906	0,1094
2	8	0,9785	**0,0215**
1	9	0,9980	**0,0020**
0	10	1,0000	**0,0000**



Rycina 12. Obszar krytyczny przy testowaniu hipotezy o proporcji płci równej 1:1 (n=10)

Dalsze rozszerzanie akceptowalnych wyników drastycznie zmniejsza prawdopodobieństwo błędu pierwszego rodzaju, ale jest bez sensu bo powoduje, że praktycznie każdy wynik będzie akceptowalny i wtedy wątpliwy jest cel całych analiz. Przyjmując hipotezę zerową czyli przyjmując prawdziwość twierdzenia, że proporcja samic w populacji to 50% można popełnić inny błąd zwany błędem drugiego rodzaju, polegający na tym, że przyjęto hipotezę zerową, która w rzeczywistości nie jest prawdziwa. Generalnie rozpatrując przykład z próbą składającą się z 10 osobników błąd taki jest stosunkowo wysoki i w celu jego minimalizacji należy zwiększać liczebność analizowanej próby. Jest to bardzo ważne by zapamiętać, że liczebność próby na poziomie 10 osobników jest mała i jeśli to możliwe, należy ją zwiększyć.

W praktyce do testowania hipotezy, iż stosunek płci w populacji gupików wynosi 1:1 na podstawie 10 elementowej próby wykorzystuje się test statystyczny zwany testem zgodności chi^2. Obliczenie testu zgodności to podstawienie liczb do prostego wzoru.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

gdzie:
Σ – suma
O – liczebność obserwowana
E – liczebność oczekiwana

Dobra wiadomość dla wszystkich, którzy unikają wszelkich nawet najprostszych wzorów matematycznych jest taka, że można dokonać stosownego obliczenia w tabeli (tabela 4). Można też odpowiednie obliczenia zautomatyzować przenosząc je do Excela albo użyć programu GeoGebra.

Tabela 4. Obserwowane i oczekiwane częstości genotypów oraz test zgodności chi^2

Oczekiwane (E)		Obserwowane (O)		Różnica O – E		Kwadrat różnicy $(O – E)^2$		$((O – E)^2)/E$		Suma (chi^2)	Tablicowe chi^2	Prawdopodobieństwo błędu jeśli odrzuca się H_0
Samice	Samce	Samice	Samce	Samice	Samce	Samice	Samce	Samice	Samce			
5	5	0	10	-5	5	25	25	5	5	10	3,84	**0,0016**
5	5	1	9	-4	4	16	16	3,2	3,2	6,4	3,84	**0,0114**
5	5	2	8	-3	3	9	9	1,8	1,8	3,6	3,84	0,0578
5	5	3	7	-2	2	4	4	0,8	0,8	1,6	3,84	0,2059
5	5	4	6	-1	1	1	1	0,2	0,2	0,4	3,84	0,5271
5	5	5	5	0	0	0	0	0	0	0	3,84	1,0000
5	5	6	4	1	-1	1	1	0,2	0,2	0,4	3,84	0,5271
5	5	7	3	2	-2	4	4	0,8	0,8	1,6	3,84	0,2059
5	5	8	2	3	-3	9	9	1,8	1,8	3,6	3,84	0,0578
5	5	9	1	4	-4	16	16	3,2	3,2	6,4	3,84	**0,0114**
5	5	10	0	5	-5	25	25	5	5	10	3,84	**0,0016**

Obliczenia te pozwalają analitycznie stwierdzić czy obserwowany rozkład odbiega od teoretycznych proporcji, czy też nie ma jeszcze podstaw do odrzucenia hipotezy modelu dziedziczenia płci w proporcji 1 do 1.

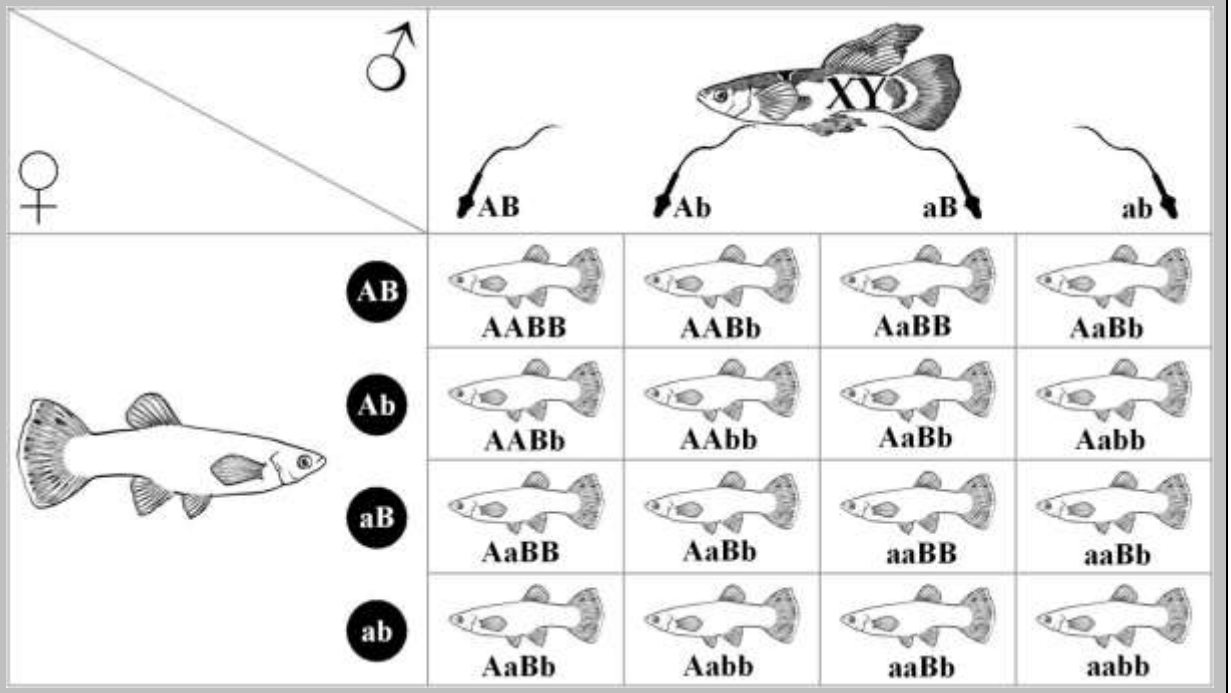
Od obserwowanej liczebności każdej z płci odejmujemy liczebność oczekiwaną. Uzyskane wyniki różnic są dodatnie lub ujemne, ale po podniesieniu do kwadratu każdy z nich jest już zawsze dodani. Dzieli się obliczone kwadraty różnicy przez liczebność oczekiwaną i uzyskuje cząstkowy wynik dla każdej z płci. Suma wyników dla obu płci jest statystyką testową testu chi^2. W ten sposób bez świadomego skorzystania ze wzoru można dokonać obliczeń statystycznych. Uzyskany wynik testu porównuje się teraz do granicznej wartości zawartej w tablicy rozkładu chi^2. W przypadku jednego stopnia swobody i poziomu istotności $\alpha=0,05$ graniczna wartość to 3,84 czyli każdy wynik wyższy od tej wartości

świadczy o tym, że obserwowany rozkład różni się od teoretycznego. Jeśli jednak obliczona statystyka testowa jest mniejsza od tablicowej to nie ma podstaw, by uznać obserwowany rozkład za odbiegający od rozkładu teoretycznego.

Dla obserwowanego rozkładu 3 samice i 7 samców obliczona statystyka testowa chi^2= 1,6 i jest mniejsza od granicznej wartości 3,84 czyli jeśli w wylosowanej próbie taki jest rozkład płci to nie ma jeszcze podstaw by twierdzić, ze w populacji rozkład płci nie odpowiada teoretycznemu 1:1. Gdybyśmy jednak mimo to odrzucili prawdziwość tej hipotezy to popełnilibyśmy błąd na wysokim poziomie 21% (0,2059). Gdy w próbie proporcje płci to 1 samica i 9 samców to statystyka testowa wynosi 6,4 czyli jest wyższa od tablicowej wartości 3,84 co oznacza, że można odrzucić twierdzenie o równej ilości samców i samic w populacji, a prawdopodobieństwo popełnienia w ten sposób błędu to tylko 1% (0,0114). Specyficzny jest tu przypadek proporcji 2 do 9 dla którego obliczony wynik testu to 3,6 czyli mniejszy od granicznego 3,84 co oznacza, że taka proporcja w próbie nie jest traktowana jako argument za odrzuceniem rozkładu teoretycznego. Gdyby jednak mimo to odrzucić teoretyczny rozkład płci 1:1 to prawdopodobieństwo popełnienia błędu wynosi jedynie niecałe 6% (0,0578) czyli jest bliskie granicznej wartości 5%.

Tym razem jako nie odbiegające od normy przyjęto również proporcje płci 2 do 8. Nie będziemy się w tym miejscu bardziej wgłębiać w zagadnienia testów statystycznych ale warto nadmienić, że przyjęta przez statystyków empiryczna reguła mówi, że jeśli wśród wartości obserwowanych są takie liczebności poniżej 10 to do analizy testem chi^2 stosuje się poprawkę na ciągłość. Dodatkowo gdy liczebności są bardzo małe jak w przypadku tabeli 4 to stosuje się analizy permutacyjne takie jak dokładny test Fishera. Czytelnik chcący poznać bardziej szczegółowo zagadnienia testowania hipotez statystycznych znajdzie te informacje w licznych podręcznikach statystyki.

Pierwszą rzeczą której uczy się na statystyce matematycznej jest rachunek prawdopodobieństwa oraz to, że jeśli dwa zdarzenia są niezależne (tzn. pojawienie się jednego z tych zdarzeń nie ma żadnego wpływu na pojawienie się drugiego z nich) to prawdopodobieństwo zajścia obu jednocześnie jest ilorazem prawdopodobieństwa każdego z nich. Przykładowo gdy mamy 2 niezależne pary genów ***Aa*** i ***Bb*** to każda para genów dziedziczy się zgodnie z pierwszym prawem Mendla. Zatem kojarząc ze sobą dwie podwójne heterozygoty ***AaBb*** uzyskuje się (1/4) potomstwa o genotypie ***AA***, (1/2) potomstwa z genami ***Aa*** oraz (1/4) osobników o genotypie ***aa***. Podobnie będzie wyglądała sytuacja z drugą parą genów czyli uzyskuje się (1/4) potomstwa o genotypie ***BB***, (1/2) potomstwa z genami ***Bb*** oraz (1/4) osobników o genotypie ***bb***. Dziedziczenie niezależnych par genów pozwala nam na obliczenie prawdopodobieństwa uzyskania poszczególnych genotypów dla dwóch par genów. Prawdopodobieństwo uzyskania genotypu ***AABB*** będzie równe iloczynowi prawdopodobieństwa uzyskania genotypu ***AA*** i genotypu ***BB*** tj. (1/4)×(1/4)=(1/16). W taki też sposób obliczyć można prawdopodobieństwo otrzymania innych genotypów:

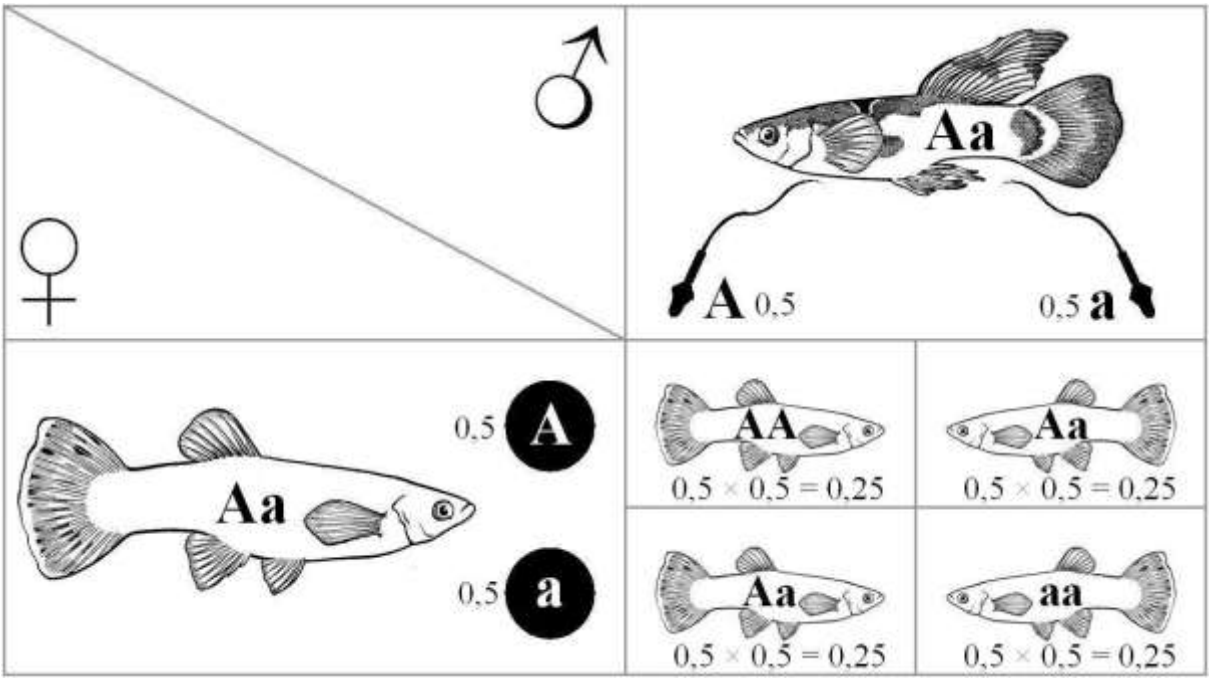


Rycina 13. Genotypy w potomstwie dihybrydów – 1:2:1:2:4:2:1:2:1

AABB – (1/4)×(1/4)=(1/16)=(1/16)
AABb – (1/4)×(1/2)=(1/8) =(2/16)
AAbb – (1/4)×(1/4)=(1/16)=(1/16)
AaBB – (1/2)×(1/4)=(1/8) =(2/16)
AaBb – (1/2)×(1/2)=(1/4) =(4/16)
Aabb – (1/2)×(1/4)=(1/8) =(2/16)
aaBB – (1/4)×(1/4)=(1/16)=(1/16)
aaBb – (1/4)×(1/2)=(1/8) =(2/16)
aabb – (1/4)×(1/4)=(1/16)=(1/16)

2. Rozkłady genotypów i fenotypów w populacji

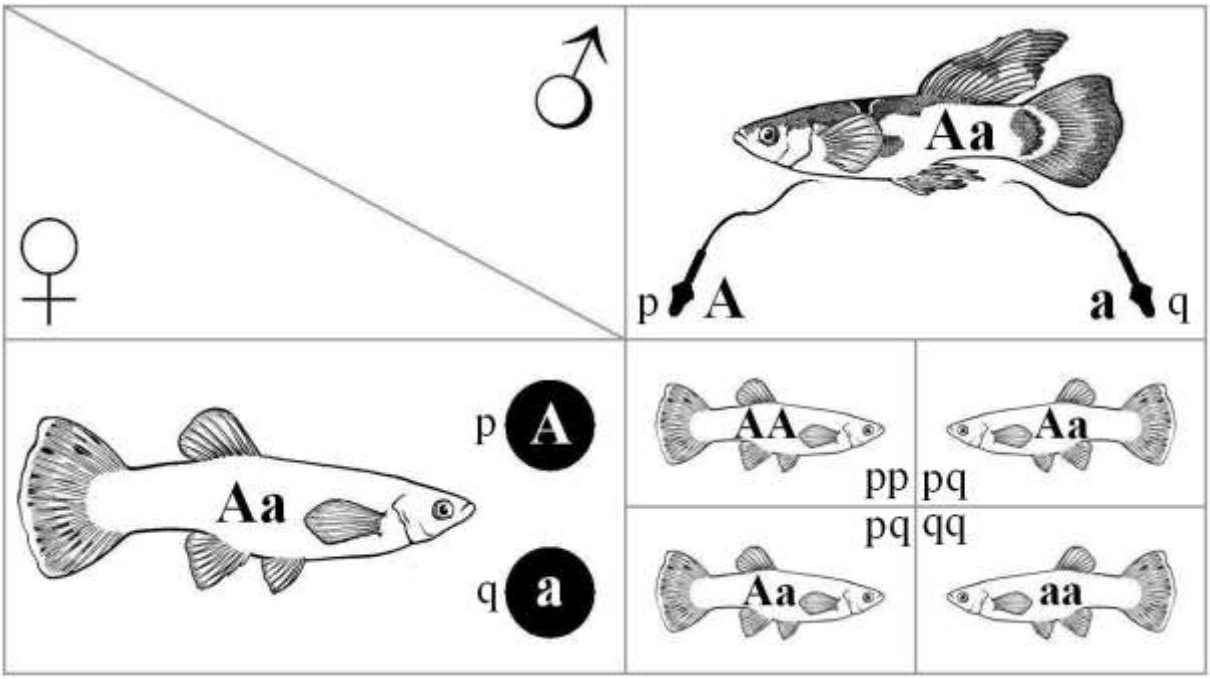
Powyższe rozważania na temat prawdopodobieństwa dotyczyły rozkładu fenotypów i genotypów w pojedynczych grupach rodzeństwa i opierały się o prawa dziedziczenia analizowane przez tzw. genetykę klasyczną zwana Mendlowską. Dalszym etapem rozwoju genetyki było przeniesienie praw dziedziczenia na grunt całej populacji. Dyscyplina, która zajmuje się matematyczną analizą i opisem praw dziedziczenia, nie dla pojedynczych rodzin lecz całych zbiorowości osobników jednego gatunku, nazywana jest genetyką populacji.



Rycina 14. Częstość występowania alternatywnych typów gamet (***A*** lub ***a***) oraz prawdopodobieństwo pojawienia się w potomstwie osobników o poszczególnych genotypach

Prawidłowości opisywane przez genetykę populacji w bezpośredni sposób wynikają z genetyki Mendlowskiej. Kojarząc dwa heterozygotyczne osobniki o genotypach ***Aa*** w pokoleniu potomnych genotypy maja rozkład 1 (***AA***) do 2 (***Aa***) do 1 (***aa***). Dokładnie taki sam rozkład genotypów będzie w potomstwie będącym produktem kojarzenia wielu heterozygotycznych par. W przypadku jednej pary heterozygot samiec produkuje połowę plemników z allelem ***A*** i drugą połowę plemników z allelem ***a*** czyli frekwencja allelu ***A*** wynosi 0,5 podobnie jak frekwencja allelu ***a.*** Frekwencje alleli najczęściej oznacza się literami „p" i „q". Czyli frekwencja allelu A to p=0,5 a frekwencja allelu ***a*** to również q=0,5. Identyczny rozkład gamet występuje u heterozygotycznej samicy. Można teraz obliczyć frekwencje genotypów wśród potomstwa. Potomek o genotypie ***AA*** powstaje po zapłodnieniu

komórki jajowej z allelem ***A*** przez plemnik niosący allel ***A***. Czyli prawdopodobieństwo spotkania się takiego plemnika i komórki jajowej to p×p=0,5×0,5=0,25. Jedna czwarta (25%) nowo powstałych zarodków będzie wiec o genotypie ***AA***. Osobnik o genotypie ***Aa*** czyli heterozygota powstaje w dwóch alternatywnych zdarzeniach. Pierwszy wariant to zapłodnienie komórki jajowej z allelem ***A*** przez plemnik będący nosicielem allelu ***a***. Drugi wariant to zapłodnienie komórki jajowej z allelem ***a*** przez plemnik będący nosicielem allelu ***A***. Oznacza to, że prawdopodobieństwo powstania zygoty heterozygotycznej to 2×p×q=2×0,5×0,5=0,5. Połowa (50%) nowo powstałych zarodków będzie wiec o genotypie ***Aa***. Potomek o genotypie ***aa*** powstaje po zapłodnieniu komórki jajowej z allelem ***a*** przez plemnik niosący allel ***a***. Czyli prawdopodobieństwo spotkania się takiego plemnika i komórki jajowej to q×q=0,5×0,5=0,25. Jedna czwarta (25%) nowo powstałych zarodków będzie o genotypie ***aa***. Rozpatrując frekwencje genotypów w szachownicy Punnetta można więc operować frekwencjami alleli.



Rycina 14. Prawdopodobieństwo pojawienia się w potomstwie osobników o poszczególnych genotypach - $pp=p^2$, $pq+pq=2pq$ i $qq=q^2$ to składniki rozwinięcia wyrażania $(p+q)^2$

Identyczne rozważania jak dla jednej pary ryb przeprowadzić można dla teoretycznej populacji składającej się z wielu osobników. Różnica polega na tym, że reguły mendlowskie dotyczą przypadku, gdy frekwencje alternatywnych alleli wynoszą dokładnie po 0,5 (rycina 14) a w modelach genetyki populacji frekwencje genów mogą przyjmować dowolne wartości (rycina 15). Przykładowo wyobraźmy sobie nowo stworzoną przez akwarystę grupę ryb

składającą się z 40 homozygot dominujących i 10 homozygot recesywnych. Wszystkich osobników jest 50 (co oznaczamy jako n) a wśród nich 40 sztuk to homozygoty dominujące (oznaczane jako d) a 10 sztuk to homozygoty recesywne (oznaczane jako r). W pokoleniu rodzicielskim brakuje heterozygot (oznaczanych jako h) ale przewidujemy, że pojawią sie one w pokoleniu potomnym przy losowym kojarzeniu pokolenia wyjściowego. Jaka jest frekwencja poszczególnych genotypów w pokoleniu rodzicielskim? Homozygoty dominujące to d/n czyli 40/50=0,8 (80%) i frekwencję tę oznaczamy jako P. Homozygoty recesywne to r/n czyli 10/50=0,2 (20%) i frekwencję tę oznaczamy jako Q. Suma częstości występowania homozygot recesywnych, heterozygot i homozygot dominujących to 1 (100%) czyli jest to suma wszystkich osobników tego pokolenia D+H+R=1.

Jaka jest frekwencja poszczególnych genów w pokoleniu rodzicielskim? Skoro homozygoty dominujące ***AA*** stanowią 0,8 populacji i nie ma żadnych heterozygot to wszystkie geny ***A*** są zawarte w genotypie homozygot czyli frekwencja genu ***A*** to p=P=0,8 (80%). Skoro homozygoty recesywne ***aa*** stanowią 0,2 populacji i nie ma żadnych heterozygot to wszystkie geny ***a*** są zawarte w genotypie homozygot, czyli frekwencja genu ***a*** to q=Q=0,2 (20%). W jakich frekwencjach wystąpią genotypy w kolejnym pokoleniu? Homozygoty dominujące to p×p=0,8×0,8=0,64 (64%), heterozygoty to 2×p×q=2×0,8×0,2=0,32 (32%), a homozygoty recesywne to q×q=0,2×0,2=0,04 (4%).

W powyższy przykładzie wyjściowe pokolenie składało się jedynie z homozygot obu wariantów a heterozygoty pojawiły się dopiero w kolejnym pokoleniu. Podobne rozważania można przeprowadzić dla populacji, która od początku składa się z homozygot i heterozygot. Przykładowo w populacji 50 osobników homozygoty dominujące ***AA*** to d=20, heterozygoty to h=25 a homozygot recesywnych jest r=5. Frekwencja osobników o dominującym genotypie w populacji to P=d/n=20/50=0,4 (40%). Frekwencja osobników heterozygotycznych to H=h/n=25/50=0,5 (50%). Frekwencja osobników homozygotycznych recesywnych to Q=r/n=5/50=0,1 (10%). Frekwencja genu dominującego ***A*** to frekwencja osobników dominujących AA tj.0,4 ale dodatkowo heterozygoty też są nosicielami genu dominującego i połowa ich gamet będzie ten gen przekazywała kolejnemu pokoleniu, czyli ich wkład we frekwencje genu dominującego to 0,5×H=0,5×0,5=0,25. Suma obu wyników daje nam ostateczną frekwencje genu dominującego czyli p=P+0,5×H=0,4+0,25=0,65 (65%).

Frekwencja genu recesywnego ***a*** to frekwencja osobników dominujących ***aa*** tj.0,1 ale dodatkowo heterozygoty też są nosicielami genu dominującego i połowa ich gamet będzie ten gen przekazywała kolejnemu pokoleniu czyli ich wkład we frekwencje genu dominującego to 0,5×H=0,5×0,5=0,25. Suma obu wyników daje nam ostateczną frekwencje genu

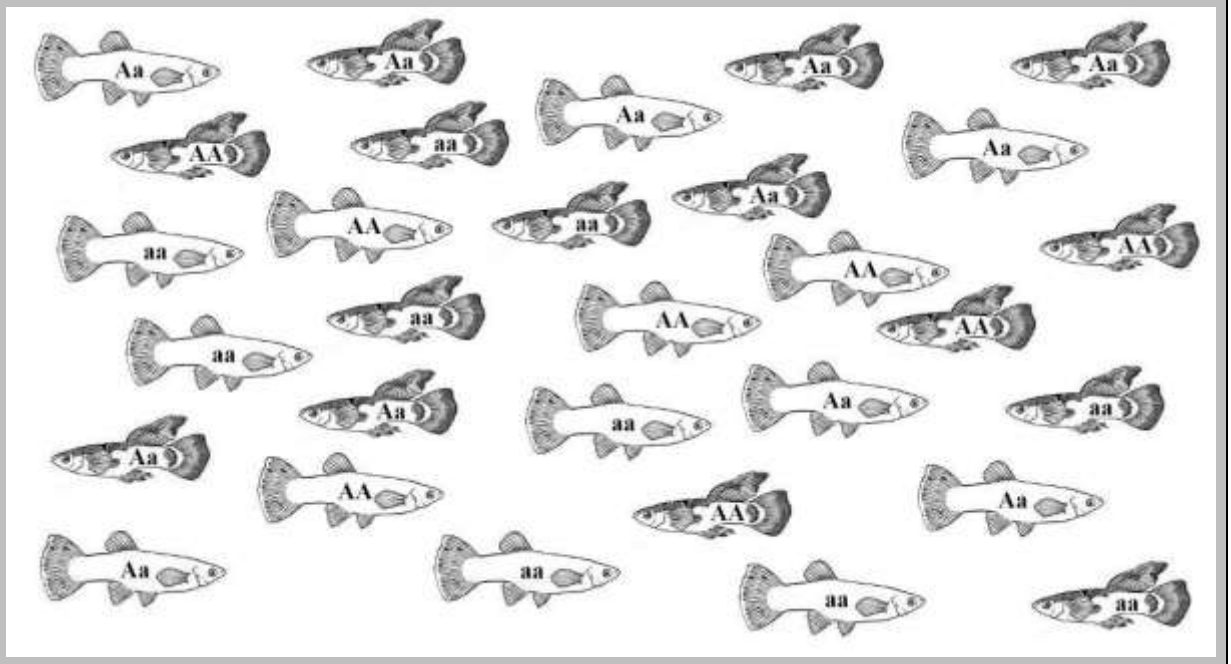
dominującego czyli p=Q+0,5×H=0,1+0,25=0,35 (35%). W pokoleniu potomnym przy losowym zapłodnieniu komórek jajowych przez plemniki osobniki homozygotyczne dominujące stanowić będą p×p=0,65×0,65=0,4225 (42%), heterozygotyczne to 2×p×q=2×0,65×0,35=0,455 (46%) a homozygoty recesywne to q×q=0,35×0,35= 0,1225 (12%).

Jeżeli w dużej losowo kojarzącej się populacji nie działają czynniki zmieniające frekwencje alleli oraz frekwencja alleli u obu płci jest jednakowa, to populacja taka pozostaje w stanie równowagi genetycznej. Jest to tzw. prawo Hardy'ego-Weinberga od nazwisk dwóch genetyków, którzy niezależnie od siebie opisali taką prawidłowość.

Frekwencja p i q to razem 1 (100%) czyli gdy nie możemy odróżnić homozygot dominujących od heterozygot, to jedynie frekwencja homozygot recesywnych może być policzona, a z niej uzyskać można frekwencje genu recesywnego q, a dzięki temu wiemy, że pozostałe to geny dominujące czyli 1-q=p a z tego możemy oszacować liczbę homozygot dominujących i heterozygot przy założeniu równowagi genetycznej.

W populacji będącej w równowadze genetycznej frekwencja homozygot recesywnych R jest równa iloczynowi frekwencji allelu recesywnego czyli $q \times q = q^2$ czyli $q^2 = Q$ a więc $q = \sqrt{Q}$. Jest to bardzo ważne, gdyż pozwala nam oszacować ilość homozygot dominujących i heterozygot w populacji nawet jeśli nie różnią się one fenotypowo. Niezbędne jest założenie że populacja ta znajduje się w równowadze genetycznej. Wtedy liczymy częstotliwość homozygot recesywnych, którą oznaczamy jako Q. Z frekwencji homozygot recesywnych wyliczamy frekwencje recesywnego allelu $q = \sqrt{Q}$. Znając frekwencje allelu recesywnego wyliczamy frekwencje allelu dominującego p=1-q. Dzięki temu możemy już policzyć frekwencje homozygot dominujących $P = p^2$ oraz heterozygot H=2pq.

Jak już wielokrotnie wspominano natura nie działa jednak pod ścisłe dyktando prawidłowości matematycznych, lecz zjawiska zapłodnienia są losowe, więc w rzeczywistości mogą nastąpić pewne odstępstwa od teoretycznych wyników. Ocena stopnia tej nieścisłości między wynikami teoretycznymi i rzeczywistymi, pozwala na określenie czy obserwowany rozkład fenotypów świadczy o równowadze genetycznej populacji czy też jakieś zjawisko zakłóciło tę równowagę. Głównym czynnikiem mogącym wpływać na równowagę genetyczną jest selekcja preferująca pewne genotypy, a jednocześnie eliminująca inne. Do sprawdzenia czy obserwowane i oczekiwane wyniki są zbliżone, służy wspominany już test zgodności chi^2.



Powyższa ilustracja prezentuje populacje 30 osobników (n=30). Homozygoty ***AA*** występują w ilości 8 (d=8) a homozygoty ***aa*** to 10 osobników (r=10) natomiast heterozygot ***Aa*** jest 12 (h=12). Bardzo łatwo można więc wyliczyć frekwencje poszczególnych genotypów dzieląc liczebności kolejnych genotypów przez ogólną liczbę osobników. Frekwencja występowania homozygot ***AA*** to P=d/n=8/30=0,27 a frekwencja homozygot ***aa*** to Q=r/n=10/30=0,33, natomiast frekwencja heterozygot ***Aa*** to H=h/n=12/30=0,40. Suma frekwencji wszystkich trzech genotypów to P+H+Q=1 gdyż stanowią one w sumie całość populacji.

W przypadku 30 diploidalnych osobników ilość genów to 60. Geny dominujące ***A*** występują u homozygot ***AA*** i heterozygot ***Aa*** czyli skoro jest 8 homozygot dominujących to ilość genów dominujących to 2×8=16 a dodatkowe 12 takich genów występuje u 12 heterozygot. Łącznie ilość genów dominujących to 16+12=28 a frekwencja to p=28/60=0,47. Geny recesywne ***a*** występują u homozygot ***aa*** i heterozygot ***Aa*** czyli skoro jest 10 homozygot recesywnych to ilość genów dominujących to 2×10=20 a dodatkowe 12 takich genów występuje u 12 heterozygot. Łącznie ilość genów dominujących to 20+12=32 a frekwencja to q=32/60=0,53. W ten sposób scharakteryzowaliśmy genetycznie obserwowaną populację. Przy losowym kojarzeniu takiej populacji zakładamy, że gamety niosące dwa alternatywne warianty tego genu, mają równe szanse na spotkanie się i stworzenie zygoty, a potem osobnika kolejnego pokolenia. Można więc obliczyć, jaka będzie struktura genetyczna kolejnego pokolenia ryb.

Homozygoty dominujące ***AA*** powstają po zapłodnieniu komórki jajowej z allelem ***A*** plemnikiem niosącym ten sam allel i prawdopodobieństwo takiego zdarzenia to iloraz prawdopodobieństwa, że każda z tych gamet jest wyposażona właśnie w ten allel (p) czyli prawdopodobieństwo powstania homozygoty dominującej to $p \times p = p^2 = 0{,}47 \times 0{,}47 = 0{,}22$. Heterozygoty ***Aa*** powstają po zapłodnieniu komórki jajowej przez plemnik niosącym allel przeciwny. Czyli komórkę jajową z ***A*** zapładnia plemnik z ***a*** lub komórkę jajowa z ***a*** zapładnia plemnik z ***A*** co oznacza, że istnieją dwie alternatywne drogi powstania heterozygot co trzeba uwzględnić podczas obliczeń. Prawdopodobieństwo takiego zdarzenia to podwojony iloraz prawdopodobieństwa, iż jedna z gamet jest wyposażona w allel dominujący (p) a druga w allel recesywny (q) czyli prawdopodobieństwo powstania homozygoty dominującej to $2 \times p \times q = 2 \times 0{,}47 \times 0{,}53 = 0{,}50$. Homozygoty recesywne ***aa*** powstają po zapłodnieniu komórki jajowej z allelem ***a*** plemnikiem niosącym ten sam allel i prawdopodobieństwo takiego zdarzenia, to iloraz prawdopodobieństwa, że każda z tych gamet jest wyposażona właśnie w ten allel (q) czyli prawdopodobieństwo powstania homozygoty recesywnej to $q \times q = q^2 = 0{,}53 \times 0{,}53 = 0{,}28$.

W pokoleniu potomnym frekwencja genotypów to P=0,22, H=0,50 i Q=0,28 i są to wyniki przy zachowaniu równowagi genetycznej a w pokoleniu rodzicielskim było to P=0,27, H=0,40 i Q=0,33. Czy różnica ta jest istotna? Można to sprawdzić testem zgodności chi^2. Obliczenia te prowadzi się nie na frekwencjach, lecz rzeczywistych liczebnościach. W pokoleniu rodzicielskim d=8, h=12 i r=10. Natomiast przy zachowaniu równowagi genetycznej byłoby to d=0,22×30=7, h=0,50×30=15, r=0,28×30=8.

	AA (d)	Aa (h)	aa (r)
O --- Obserwowane (pokolenie rodzicielskie)	8	12	10
E --- Oczekiwane (pokolenie potomne)	7	15	8
(O-E) --- różnica	1	-3	2
(O-E)^2 --- kwadrat różnicy	1	9	4
((O-E)^2)/E	0,1429	0,6000	0,5000
Suma (chi^2)	1,24		
df	2	1	
p	0,5372	0,2649	
Tablicowe chi^2	5,99	3,84	

Nie stwierdzono istotnej różnicy między obserwowanymi i oczekiwanymi liczebnościami co oznacza, iż można przyjąć, że populacja ta jest w stanie równowagi genetycznej. Pojawić się może wątpliwość czy stopnie swobody to 2 (czyli liczba grup odjąć 1) czy 1 (czyli liczba

grupa minus 1 minus 1) w tym drugim przypadku uwzględniamy również liczbę parametrów oszacowanych na podstawie danych czyli uwzględniamy fakt, że znając p nie trzeba już liczyć q tylko korzysta się z zależności q=1-p.

Statystyka zazwyczaj nie należy do ulubionych przedmiotów zarówno podczas edukacji jak i samokształcenia, a najczęstszą opinią jest twierdzenie, że statystyka to jeden z najtrudniejszych przedmiotów na studiach. W niniejszym tekście podjęto skromną próbę przedstawienia zagadnień statystycznych w formie przyjaznej czytelnikowi, zgodnie z zasadą że zagadnienia te należy przedstawiać na przykładach osobiście interesujących odbiorcę. W tym przypadku przykłady są ichtiologiczne, więc być może będą przyczynkiem do zainteresowania się statystyką przez jakiegoś akwarystę. Cytując Williama I.B. Beveridge „*Nie jest rzeczą konieczną, aby biolog był biegły w statystyce matematycznej, jeśli nie ma zamiłowania do tego przedmiotu, powinien jednak poznać go na tyle, aby mógł uniknąć czy to niedoceniania go, czy też przecenienia. Powinien wiedzieć, kiedy jest mu potrzebne pomoc statystyka.*”.

Literatura:

1. Hartl D.L., Clark A.G.: 2007. Podstawy genetyki populacyjnej. Wydawnictwo Uniwersytetu Warszawskiego, Warszawa,
2. Krzanowska H., Łomnicki A., Rafiński J.: 1982. Wprowadzenie do genetyki populacji. Państwowe Wydawnictwo Naukowe, Warszawa,
3. Łomnicki A.: 2010. Wprowadzenie do statystyki dla przyrodników. Wydawnictwo Naukowe PWN, Warszawa,
4. Parker R.E.: 1978. Wprowadzenie do statystyki dla biologów. Państwowe Wydawnictwo Naukowe, Warszawa,
5. Sváb J.: 1978. Genetyka populacji. Państwowe Wydawnictwo Rolnicze i Leśne, Warszawa,
6. Watała C.: 2002. Biostatystyka – wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych. α-media Press, Bielsko-Biała,
7. Wołek J.: 2006. Wprowadzenie do statystyki dla biologów. Wydawnictwo Naukowe Akademii Pedagogicznej, Kraków.