آمار

در یکی از آزمایش های مندل، گیاهان نخود فرنگی هتروزیگوت در F_1 که همه آنها بلند بودند درون آمیزی انجام دادند. در سنل بعد (F_2) او 787 زاده بلند و 277 زاده کوتاه بدست آورد به صورت نسبت 2/ 84:1 مندل این نسبت را به صورت 3:1 گرد کرد که از قانون او در توارث صفات حمایت می کند در واقع آیا می توان 787:277 را به صورت 3:1 گرد کرد؟ با توجه به بحث کوتاه در مورد احتمال ما مقداری انراف از نسبت دقیق 3:1 را انتظار داریم اما چه مقدار انحراف قابل قبول است؟ آیا 787:277 از قانون مندل حمایت می کند؟ 785:279 چه طور؟ آیا 709:355 یک نسبت 2:1 است یا 532:532 یک نسبت 1:1 ؟ در این جاست که نظم آمار به ما کمک می کند.

ما هیچ وقت نمی توانیم با اطمینان در مورد پدیده های اتفاقی و تصادفی صحبت کنیم برای مثال مورد مندل را در نظر بگیرید. اگر چه یک نسبت 3:1 بر اساس فرضیه ی او. انتظار می رفت شانس می توان نسبت 1:1 را در مورد داده های 532:532 بوجود آورد در حالی که مکانیزم همان مکانیزم پیشنهادی مندل باشد. ما می توانیم یک سکه ی قابل اعتماد را پرتاب کنیم و ده بار پشت سرهم رو بیاید بر عکس مندل می توانست دقیقاً نسبت 3:1 (798:266) را در نسل F_2 بدست آورد اما فرضیه ی او در مورد جدایی الل ها غلط باشد. نکته این جاست که هر وقت ما با پدیده های احتمالی سروکار داریم مقداری شانس وجود دارد که داده ها ما را به سمت حمایت از یک فرضیه ی بد یا رد یک فر ضیه ی خوب هدایت کنند. آمار این احتمالات را کمی می کند ما نمی توانیم با قطعیت بگوییم که نسبت 2/84:1 نشان دهنده ی 3:1 است اما می توانیم بگوییم که یک درجه اطمینان مشخص به این نسبت داریم این آمار است که به ما کمک می کند تا این حدهای اطمینان را تعیین کنیم.

آمار یک شاخه از تئوری احتمالات است که ژنتیک دانان را به سه صورت یاری می دهد. اول شاخه ای از آمار که طراحی آزمایش نام دارد. مقداری اندیشیدن قبل از اجرای یک آزمایش می تواند ما را در طراحی آزمایشی موثر تر یاری دهد اگر چه مندل آمار نمی دانست اما طراحی آزمایش او بسیار خوب بود. راه دومی که آمار می تواند کمک دهنده باشد در جمع بندی و خلاصه کردن داده هاست واژه های آشنایی مثل میانگین و انحراف معیار قسمت هایی از آمار توصیفی هستند که حجم زیادی از داده ها را به صورت یک یا دو مقدار معنی دار خلاصه می کنند ما از ایـن واژه ها و مفاهیم در فصل وراثت کمی استفاده بیشتری خواهیم کرد.

آزمون فرضیه :

راه سومی که آمار برای ژنتیک دانان مفید واقع می شود در آزمودن فرضیه است: مشخص شدن اینکـه آیـا یک فرضیه رد می شود یا حمایت می شود برا ساس میزان نزدیکی داده ها به پیش بینی های آن فرضیه است ایـن قسمت بیشتر به بحث جاری ما مربوط می شود به عنوان مثال آیا نسبت 787:277 واقعاً بیانگر نسبت 3:1 بـود؟ از آنجایی که اکنون می دانیم نمی توان یک جواب قطعی به این سوال داد چگونه می توانیم یک سطح از پـشتیبانی و حمایت را به جوابمان نسبت بدهیم؟

آمار ما را به صورتی که در اینجا می آید به پیش می برد. برای شروع ما نیازمند به فراوانی ای که با آن پدیده های مختلف ممکن می توانند در یک آزمایش خاص اتفاق بیفتد برای مثال اگر ما یک گیاه هترزویگـوت بلنـد را در شرایط درون آمیزی قرار دهیم در میان زاده های تولید شده انتظار نسبت 3:1 برای گیاهان بلند در مقابل کوتاه قد را داریم (نسبت 3:1 فرضیه ی ماست که بر اساس فرض کنترل ژنتیکی بلندی به وسیله ی یک *Locas* با دو آلـل

بدست آمده) اگر ما به چهار زاده ی اول نگاه کنیم احتمال اینکه 3 گیاه بلند و یک گیاه کوتاه داشته باشیم چقـدر است؟ جواب با استفاده از فرمول جمله های بسط دو جمله ای محاسبه می شود

$$p=\frac{4!}{3!1!}\left(\frac{3}{4}\right)^{3}\left(\frac{1}{4}\right)^{1}=\frac{108}{256}=0/42$$

به طریق مشابه می توانیم این احتمال ها را محاسبه کنیم : همگی بلند قد شوند $\left(\frac{81}{256}=0/32\right)$ دو تا بلند قدر و دو تا کوتاه قد شوند $\left(\frac{54}{256}=0/21\right)$ یکی بلند قد و سه تا کوتاه قد شود $\left(\frac{12}{256}=0/05\right)$ و همگی کوتاه قد شوند $\left(\frac{1}{256}=0/004\right)$

این پخش احتمال به همراه پخش های احتمال مربوط به نمونه هــای 40 زاده ای و 8 زاده ای در جــدول زیــر نشان داده شده است

$n = 4$		$n = 8$		$n = 40$	
No. Tall Plants	*Probability**	*No. Tall Plants*	*Probability**	*No. Tall Plants*	*Probability**
4	$\frac{81}{256}=0.32$	8	0.10	40	0.00001
3	$\frac{108}{256}=0.42$	7	0.27	39	0.0001
2	$\frac{54}{256}=0.21$	6	0.31	38	0.0009
1	$\frac{12}{256}=0.05$	5	0.21		. . .
0	$\frac{1}{256}=0.004$	4	0.09	30	0.14
		3	0.02		. . .
		2	0.004	2	0.59×10^{-20}
		1	0.0004	1	0.10×10^{-21}
		0	0.00002	0	0.83×10^{-24}

$sampling\ \ distribution$ برای اندازه های نمونه 4،8 و 40 برای نسبت 3:1 بلند قد به کوتاه قد داده شده در این آزمایش

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

* احتمالات تئوری دو جمله ای محاسبه شده اند

$n =$ تعداد کل زاده ها

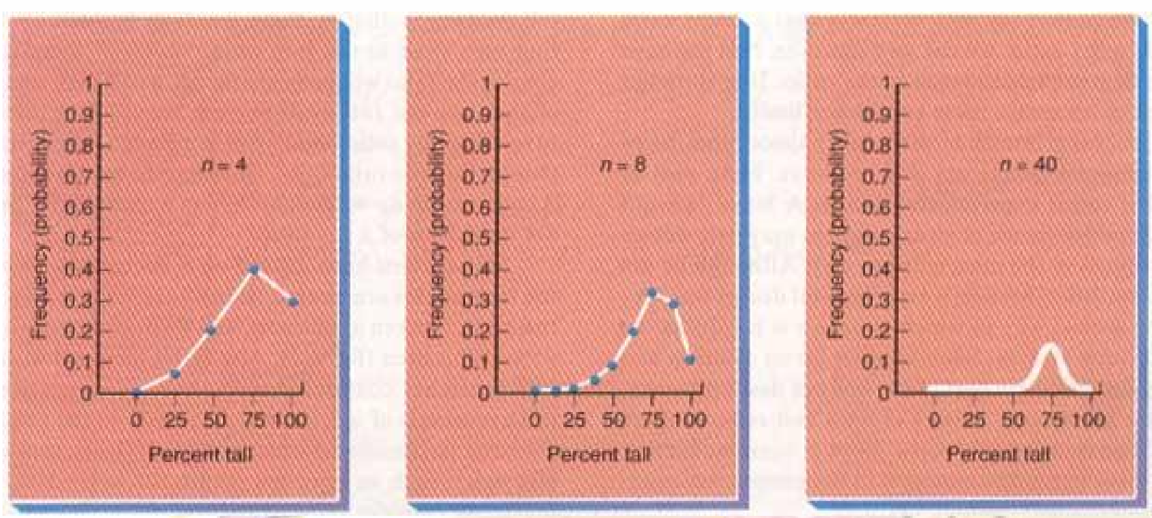
$s =$ تعداد زاده های بلند قد

$t =$ تعداد زاده های کوتاه قد

$p = \left(\frac{3}{4}\right)$ احتمال بلند قد بودن یک زاده

$q = \left(\frac{1}{4}\right)$ احتمال کوتاه قد بودن یک زاده

این پخش ها همچنین در نمودار شکل 1 نمایش داده شده اند.



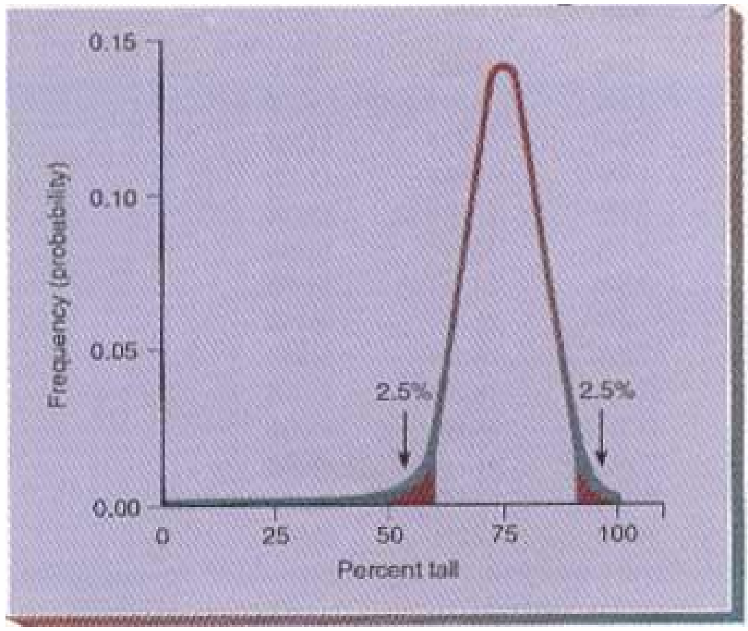
زیر نویس شکل 1: $sampling\ \ distribution$ از یک آزمایش نسبت مورد انتظار سه بلند قد در مقابل یک کوتاه قد هنگامی که

اندازه نمونه، 8، بزرگتر می شود پخش صاف تر می شود این پخش ها جمله های بسط دو جمله ای است توجه کنید اگر چه n بزرگتر می

شود پیک منحنی پایین تر می آید چون همان طور که نقاط بیشتری (نسبت های ممکن) در طول محور X فشرده می شوند احتمال هر کدام از آن ها کمتر می شود.

هنگامی که اندازه ی نمونه بزرگ تر می شود (از 4 بـه 8 و از 8 بـه 40 در شـکل 1) پخـش نمونـه گیـری ($sampling\ \ distribution$) شکل یک نمودار صاف با یک پیک در نسبت درست 3:1 (75٪ زاده ی بلند قد) را به خود می گیرد این بدان معناست که احتمال بالایی وجود دارد که به نسبتی خیلی نزدیـک بـه نـسبت درسـت 3:1 بدست آوریم اما احتمال کمی هم وجود دارد که به نسبت بسیار دوری برسیم پس در زمان های کمی ممکـن اسـت نسبت بدست آمده دور از نسبت 3:1 باشد مهم است، بدانیم. هر نسبتی ممکن است در یک آزمایش بدست آید اگر چه نسبت درست 3:1 است. پس ما خط را کجا می کشیم؟ در چه نسبتی تصمیم مـی گیـریم کـه یـک نتیجـه ی آزمایشگاهی را بیانگر نسبت 3:1 ندانیم؟

آمار دانان با یک قرار داد موافقند وقتی که مثل شکل 2 همه ی فراوانی ها رسم شد سطح زیـر نمـودار بـه عنوان یک واحد گرفته می شود



$sampling\ \ distribution$ مربوط به شکل 50 $(n = 40)$ 5٪ از سطح زیر منحنی علامت گذاری شده

و ما خط را به نحوی رسم می کنیم که 95٪ این سطح را در بر گیرد هر نسبتی که در این فاصله قرار گیـرد تایید کننده ی فرضیه ی نسبت 3:1 است هر نسبتی در5٪ سطح باقی مانده غیر قابل قبـول اسـت. (قراردادهـای دیگری هم وجود دارد که 1٪ یا 10٪ سطح باقی مانده را به عنوان نسبت های غیر قابل قبول معرفی می کند که مـا در پایان فصل به آن ها می پردازیم) بنابراین می توانیم بفهمیم که یک داده ی حاصل از آزمایش فرضـیه ی مـا را تایید می کند یا نه. به عبارتی اگر در 95٪ سطح بین دو خط قرار بگیرد تایید کننده ی فرضیه ی ما و اگر خـارج از آن باشد رد کننده ی فرضیه است. یک بار از بیست بار ما یک خطای نوع I خواهیم داشت یعنی با یـک فرضـیه ی درست را رد می کنیم (خطای نوع II شکست در رد کردن یک فرضیه ی غلط است.)

آیا محاسبه ی $sampling\ \ distribution$ هر بار که آزمایشی انجام می دهـیم لازم اسـت؟ بـرای مـشخص کردن اینکه آیا یک فرضیه رد یا قبول می شود یک نمودار پخش فراوانی برای هر نوع آزمایش باید رسم شود. مندل

توانسته بود از پخش نشان داده شده در شکل 1 برای پوشش دانه ها یا رنگ دانه ها استفاده کند تا زمانی که یک نسبت 3:1 را انتظار داشت و اندازه ی نمونه ی مشابهی داشت. اما در مورد یک دسته بندی مستقل چطور؟ جایی که انتظار نسبت 9:3:3:1 را داریم یک ژنتیک دان باید یک *sampling distribution* براساس نسبت 9:3:3:1 و یک اندازه ی نمونه ی خاص را محاسبه کند. آمار دانان روش های میانبری ابداع کرده اند که با استفاده از پخش های استاندارد شده احتمال ها را محاسبه می کنند تعداد زیادی از آنها کاربرد دارند مثل پخش t، پخش دو جمله ای و پخش مربع کای که هر کدام برای نوع خاصی از داده ها مفید است. ژنتیک دانان معمولاً پخش مربع کای را برای آزمایش فرضیه ها با توجه به داده های تولید مثل مورد استفاده قرار می دهند.