آنتروپی درواقع معیاری برای تفاوت ها و شباهت ها بوده و نشان دهنده مقدار بی نظمی و آشفتگی است.

اگر به عنوان مثال ما یک ایونت روزمره را درنظر گرفته که هرروز اتفاق می افتد اگر این اتفاق روزمره روزی اتفاق نیفتد یا حالت دیگری از آن اتفاق بیفتد باعث سورپرایز شدن یا تعجب میشود و احتمال رخدادی که کم است درحین وقوع باعث سورپرایز بیشتری میشود و این معیار سورپرایز را برای نمایش ریاضی با لگاریتم معکوس احتمال رخداد کار میکنیم. (معکوس احتمال مطلوب نیست چون یک را یک نشان میدهد)

به طور واضح مفهوم سورپرایز شدن با آشفتگی رابطه دارد و انگیزه استفاده لگاریتم در آن هم مشابه و مرتبط با مفهوم surprise است. به طور دقیق تر آنتروپی میانگین سورپرایز ما یا امید ریاضی سورپرایز های ما میباشد.

اگر مفهوم بالا را به صورت فرمول نشان دهیم با تغییر و ساده سازی فرمول داخل جزوه نیز بدست می آید.

یک مثال خوب شهودی درنظر گرفتن زندگی روز به روز یک فرد منظم و یک فرد هیجانی است که زندگیش هرروز متفاوت است. اگر هرروز را با رنگ نشان دهیم، تفاوت رنگ ناگهانی در زندگی فرد منظم بیشتر باعث میشود ما بیشتر سورپرایز شویم ولی میانگین آنتروپی فرد هیجانی همچنان بیشتر از فرد منظم است.

شاخص جینی برای محاسبه و مقایسه diversity و یا همان تنوع دیتاست ها و مجموعه ها بکار میرود بطوریکه هرچه دیتاست از اعضای متنوع تری تشکیل شده باشد به 1 نزدیک تر بوده و یک مجموعه با اعضای کاملا یکسان شاخص صفر را دارا خواهد بود.

این شاخص براساس انتخاب جفت های مرتب از مجموعه و نسبت جفت های یکسان و غیریکسان بدست می آید.

برای این کار میتوان از جدول استفاده کرد که هر سطر آن نشان دهنده یک عضو مجموعه و همچنین هر ستون آن نیز نشان دهنده یک عضو مجموعه باشد. (جدولی n*n)

حال میتوان خانه های جدول را معادل انتخاب شدن جفت مرتب عضو نظیر سطر و عضو نظیر ستون درنظر گرفت این گونه برای اندازه گیری احتمال یکسان نبودن اعضای جفت مرتب منتخب کافی است مساحت خانه های جفت مرتب یکسان را از مساحت کل کم کرده و جواب را تقسیم بر مساحت کل کنیم یا به طور مشابه احتمال انتخاب شدن جفت های یکسان را از 1 کم کنیم.

- انحراف معیار محاسبه شده با مقسوم علیه n-1 یک انحراف استاندارد است که از نمونه به عنوان تخمینی از انحراف استاندارد جامعه ای که نمونه از آن گرفته شده است محاسبه می شود. از آنجایی که مقادیر مشاهده شده به طور متوسط به میانگین نمونه نزدیک تر از میانگین جامعه می شوند، انحراف استاندارد که با استفاده از انحراف از میانگین نمونه محاسبه می شود، انحراف استاندارد مورد نظر جامعه را دست کم می گیرد. استفاده از n-1 به جای n به عنوان مقسوم علیه آن را با کمی بزرگتر کردن نتیجه تصحیح می کند.

توجه داشته باشید که تصحیح زمانی که n کوچک است نسبت به زمانی که بزرگ است اثر متناسب بیشتری دارد، این همان چیزی است که ما می خواهیم زیرا وقتی n بزرگتر باشد میانگین نمونه احتمالاً برآورد خوبی برای میانگین جامعه خواهد بود. وقتی نمونه کل جامعه باشد، از انحراف معیار با n به عنوان مقسوم‌کننده استفاده می‌کنیم، زیرا میانگین نمونه، میانگین جامعه است.

- Population(جمعیت) کل گروهی است که می خواهید درباره آن نتیجه گیری کنید. Sample(نمونه) گروه خاصی است که از آن داده ها را جمع آوری خواهید کرد. حجم نمونه همیشه کمتر از حجم کل جامعه است.
- نماد های هر خصوصیت برای سمپل و جمعیت به شرح زیر است:

Mean of general population μ		2.5 mmol/l
Mean of sample $\bar{x}$		3.2 mmol/l
Standard deviation of sample , SD		1.1 mmol/l
Standard error of sample mean,	$SD/\sqrt{n} = 1.1/\sqrt{18}$	0.26 mmo1/l
Difference between means $\mu - \bar{x}$ = 2.5 - 3.2		-0.7 mmol/l

- همچنین در لینک 2 نمودار هایی برای تفاوت استفاده از n,n-1,n-2 نشان داده شد که در نمودارهای سمت چپ همگرایی، کمتر بودن یا بیشتر شدن

از واریانس حقیقی را براساس تعداد نمونه مشاهده میکنیم. برای تقسیم n مشاهده میشود که باوجود اندازه مناسب نمونه هنوز هم واریانس به وضوح کمتر از واریانس حقیقی است. برای n-1 میبینیم که با نمونه مناسب به واریانس حقیقی بسیار نزدیک میشویم. ولی با n-2 واریانس محاسبه شده از واریانس حقیقی بیشتر میشود (در نمونه هایی که سایز کوچکی ندارند)

- در سمت راست نیز نمودارها بیانگر تفاضل واریانس محاسبه شده و واریانس حقیقی براساس تفاضل نمونه و میانگین جمعیت میباشد. لازم به ذکر است که در اکثر مواقع ما اطلاعات جمعیت را نداریم و این نمودارها فقط برای نشان دادن تفاوت تغییر در فرمول ها است. در اینجا هم مشاهده میشود که با n ما همواره داریم مقدار را کم محاسبه نموده و به اصطلاح underestimate میکنیم ولی برای n-2 مقدار را در برخی جاها بیشتر و برخی جاها کمتر محاسبه کرده و در مجموع نتیجه نزدیک به واریانس حقیقی در خواهد آمد. N-2 نیز بدلیل overestimate کردن زیاد نتیجه مطلوبی به ما نمیدهد.