جلسه‌ی بیست و چهارم آمار و احتمال مهندسی — دکتر سروش زارچی - دانشگاه شریف

مهم‌ترین جلسه‌ی آمار و احتمال مهندسی . مهم‌ترین تست‌های آماری را بیان می‌کنیم.

χ^2-Test

T_Test

تست کای-2 (χ^2 Test)

$X_1 \sim N(0,1) \Rightarrow X_1^2 \sim \chi^2(1)$ ← درجه‌ی آزادی (degree of freedom)

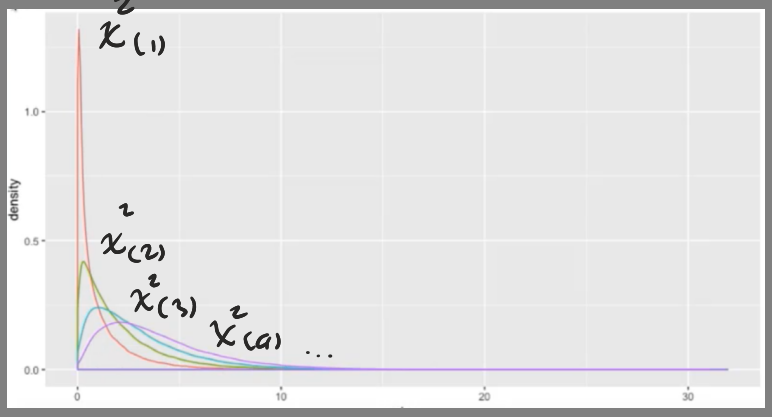
$X_2 \sim N(0,1) \Rightarrow X_1^2 + X_2^2 \sim \chi^2(2)$

$\vdots$

$X_n \sim N(0,1) \Rightarrow X_1^2 + X_2^2 + \dots + X_n^2 \sim \chi^2(n)$

همواره یک توزیع مثبت است.

منطقا چون χ^2 مربعی است هرگز صفر نخواهد بود.



آن یک سری نمونه داشته باشیم که در چندین دسته قرار بگیرند و هر نمونه در یک دسته قرار بگیرد.

و کل نمونه‌ها m تا باشد.

$\underbrace{\quad}_{\text{با احتمال } p_1 \text{ در این دسته}}$ $\underbrace{\quad}_{p_2}$ $\underbrace{\quad}_{p_3}$ $\dots$ $\underbrace{\quad}_{p_k}$

$mp_1 \quad mp_2 \quad mp_3 \quad \dots \quad mp_k \sim$ Expected

$X_1 \quad X_2 \quad X_3 \quad \dots \quad X_k \sim$ Observed.

مشاهده شده توسط نمونه‌گیری عینی (واقعی)

اگر H_0 برقرار باشد. یعنی داده‌ها به همان احتمالات داده شده در هر دسته قرار گرفته باشند.

$$\sum_{i=1}^{k} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i} \xrightarrow[m\to\infty]{} \chi^2(k-1)$$

برای $k=5$:

$$\sum_{i=1}^{5} \frac{(X_i - mp_i)^2}{mp_i} \sim \chi^2(4)$$

مثلاً اگر یک کالا توسط 3 کمپانی تولید میشود و شخصی اظهار کند که α_1 درصد از برند 1 و α_2 درصد از برند 2 و α_3 درصد از برند 3 خریده است آن گاه ما می‌رویم و از جامعه نمونه گیری می‌کنیم. فرض را بر H_0 می‌گذاریم و به محاسبه χ^2 می‌پردازیم.
بعد از محاسبه χ^2، مقدار P_Value را محاسبه می‌کنیم و P_Value صحت حرف شخص را مشخص می‌کند.

کاربرد اصلی χ^2

ما برای جداول وقوع 2×2 تست فیشر را داشتیم و برای جداول وقوع بزرگتر ما از تست کای-2 استفاده می‌کنیم. به نوعی کای-2 تعمیم یافته‌ی تست فیشر است.

	α_1	α_2	α_3
برند 1	0	0	0
برند 2	∩	∩	0
برند 3	0	∩	0

	بهبود ✓	بهبود ✗
دارو ✓	0	0
دارو ✗	0	0

تست فیشر ← ... تست کای-2 ←

نکته تست کای برای $n \to \infty$ و m مناسب است.

نکته برای m های کوچک بهتر است که از تست فیشر استفاده کنیم.

محاسبه‌ی درجه‌ی آزادی

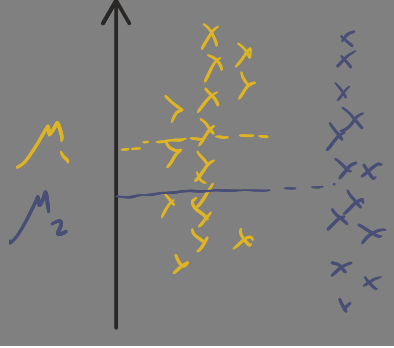
به جدول وقوع با اندازه‌ی $\alpha \times \beta$، درجه‌ی آزادی (k) از رابطه‌ی روبه رو به دست می‌آید.

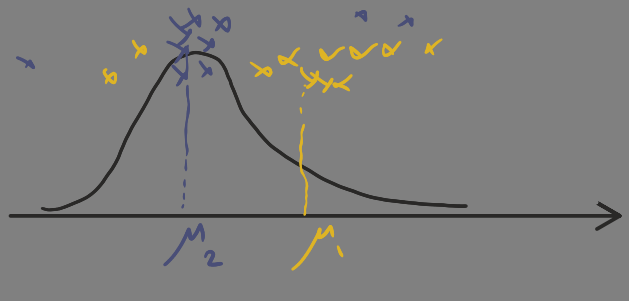
$$k = (\alpha - 1)(\beta - 1)$$

خیلی وقت‌ها پیش میاد که ما می‌خواهیم متوجه بشیم که آیا تفاوت دو معنادار هست یا نه. برای مثال یا تاثیر عمر کارمندهای دو شرکت را مقایسه کنیم.

تست-ی (t-test)

دو نمونه داریم (هر نمونه‌ای تواند یک حالت Sample باشد). آیا این دو نمونه از توزیع‌های با میانگین یکسان آمده‌اند یا خیر؟





در شرایطی که بالا به تصادیر دیده‌ایم. ما روی یک توزیع یکسان دو سمپل جدا گرفته‌ایم که میانگین‌های متفاوت دارند. دقیقاً همین جاست که اهمیت و کاربرد t-test برای ما مشخص می‌شود. با استفاده از تست-ی های توزیع دو تا سمپل رو به خوبی با هم مقایسه کنیم و به یکسان بودن توزیع اصلی هر دو سمپل پی ببریم.

هنر اصلی طراحی یک آماره در این است که تحت برقراری H_0 ما توزیع sample داشته باشیم.

در واقع آن‌ها μ_1 و μ_2 های نزدیک به هم داشته باشیم، خوبختانه σ_1^2 و σ_2^2 های کوچک داشته باشیم ریاضیاً این دو تا sample از یک توزیع اومدن چون پراکندگی داده‌های اونا هم کم هست.

پس فرض کنیم که H_0 برقراره $\Rightarrow \sigma^2 = \sigma_1^2 = \sigma_2^2$

بریم سراغ توزیع

توزیع t

$$t = \frac{Z}{\sqrt{\frac{V}{n}}}, \quad \begin{matrix} Z \sim N(0,1) \\ V \sim \chi^2(n) \end{matrix} \quad \Rightarrow t = \frac{Z}{S}$$

n: اندازه‌ی نمونه

t توزیع

تنها پارامتر توزیع t، درجه‌ی آزادی آن هست
$t(\text{degree of freedom})$

استنباط‌های واحدی در عمل؟

آماره‌ی t

تک نمونه

یک سری داده داریم. فرض ضمن داده‌ها از توزیعی با میانگین μ_0 آمده‌اند.

میانگین فرض ضمن — میانگین sample

$$t = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

تخمین واریانس از روی sample

درجه‌ی آزادی $= n-1$
توزیع t_{n-1}

خواص توزیع t خیلی شبیه به توزیع نرمال هست اما دانته امیریه ضمن میل مکینه وقتی مرکزی کوچکتری داره.
اما درجه‌ی که درجه‌ی آزادی نزدیک بشه به 30 و یا میل کنه به ∞ در اون حالت توزیع t میاد به توزیع نرمال نزدیک میشه.

دو نمونه

زمانی که می خواهیم دو نمونه داده را مقایسه کنیم. H_0 : این دو نمونه از توزیع هایی با میانگین یکسان آمده اند.

$$t = \frac{\overline{X_1} - \overline{X_2}}{S_p\sqrt{\frac{2}{n}}} \quad , \quad S_p = \frac{\sqrt{S_{X_1}^2 - S_{X_2}^2}}{2} \Rightarrow t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_{X_1}^2 + S_{X_2}^2}{n}}}$$

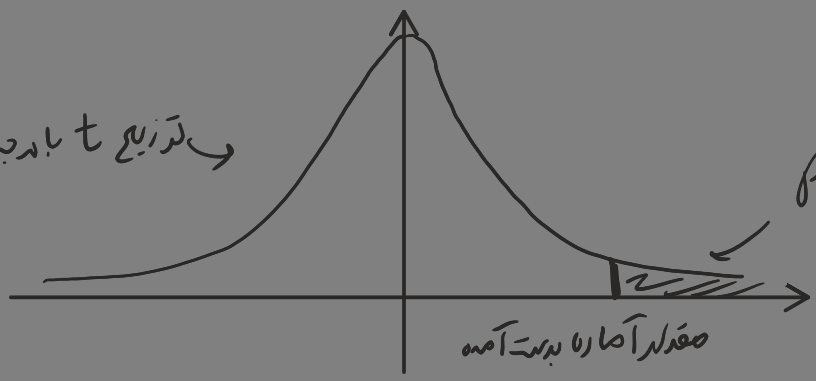
S^2 : واریانس sample های ما

درجه آزادی : $2n-2$

توزیع t دو

حالا خاصیه که باید توجه به آماره داری که داریم رو بدست بیاریم و ببینیم که Pvalue برای شباهت آماره‌ی این دو با توزیع t چقدره.

اگر که Pvalue برای شباهت صفر آماره توزیع t به اندازه‌ی کافی کم چیک باشه $\Leftarrow H_0$ واقعاً برقراره است.



توزیع t با درجه آزادی $2n-2$ تو نمونه

$n-1$ تو نمونه

اصلاً برای Pvalue ، α برابر 0.05 هست که در مجامع علمی عدد قابل قبولی هست.

در زمان استفاده از تست t نکات زیر باید رعایت بشوند :

برقرار بودن این نکات شرط استفاده از تست t هست.

1. میانگین های دو sample باید توزیع نرمال داشته باشند.
2. وقتی به هست 2 نمونه استفاده می‌کنیم مهم این هست که واریانس دو نمونه باهم برابرند. برای همین واریانس هم یکی از حل های هست.
3. توزیع t باید از هم مستقل باشند.