فهرست مطالب

فهرست تصاویر

فصل ۱

مفاهیم اولیه و کلیات

۱-۱ مقدمه

مفصل یک مدل احتمالی است که یک توزیع یکنواخت چند متغیره را نشان می دهد که ارتباط یا وابستگی بین بسیاری از متغیرها را بررسی می کند. به عبارت دیگر، یک مفصل به جداسازی احتمالات مشترک (حاشیه ای) یا ترکیب احتمالات مشترک متغیرها که در یک سیستم چند متغیره پیچیده تر کمک می کند.

سپس مفصل شاخص منحصر به فرد یا مجموعه ای از دستورالعمل ها برای توصیف نحوه قرار گرفت جفت ها در سیستم پیچیده تر است. این روش می تواند به شناسایی همبستگی های جعلی (ساختگی) مشاهده شده در داده ها کمک کند در نتیجه این روش مفید است. همچنین در مدل‌های قیمت‌گذاری مشتقات تنظیم دقیق که در آن قیمت اوراق بهادار به قیمت برخی از اوراق بهادار بستگی دارد، مفید است.

اگرچه محاسبه آماری مفصل در سال ۱۹۵۹ توسط سکلار ارایه و توسعه یافت، اما تا اواخر دهه ۱۹۹۰ در بازارهای مالی و امور مالی اعمال نشد. مفصل یک روش آماری برای درک احتمالات مشترک یک توزیع چند متغیره است. توضیح توزیع توام و توزیع های حاشیه ای ... کلمه copula از لاتین به معنای "پیوند" یا "مفصل یا ارتباط" به هم می آید، جایی که این اصطلاح در زبان شناسی برای توصیف چنین کلمات یا عبارات پیوند دهنده استفاده می شود. امروزه، مفصل ها در تجزیه و تحلیل مالی پیشرفته برای درک بهتر نتایجی که شامل دم ضخیم (پهن) و چولگی است، استفاده می شود. در تئوری گزینه‌ها، مفصل ها می‌توانند به قیمت‌گذاری بهتر سبدی از گزینه‌ها کمک کنند، زیرا قیمت یک گزینه به قیمت یک دارایی اساسی بستگی دارد. آبه اسکالار در سال ۱۹۵۹ [1] تجزیه و تحلیل احتمال مفصل را توسعه داد اما تا مدت ها بعد از آن در امور مالی استفاده نشد.

یک سوال عالی!

در آمار، کوپول یک تابع ریاضی است که رابطه بین دو یا چند متغیر تصادفی را توصیف می کند. این یک مفهوم اساسی در نظریه احتمال است و برای مدل سازی وابستگی بین متغیرها استفاده می شود.

- C یک تایع مفصل دو بعدی است:

$$C : [۰۱] \times [۰۱] \to [۰, ۱]$$

ویژگی های تابع مفصل:

- شرایط مرزی

$$C(۰, u) = ۰$$

$$C(۱, u) = u; \forall u \in [۰, ۱]$$

[1] Abe Skalar

- تداعی

$$C(C(u,v),w) = C(u,C(v,w)); \forall u,v \in [۰,۱]$$

- تقارن

$$C(u,v) = C(v,u); \forall u,v \in [۰,۱]$$

ویژگی اول تضمین می کند که copula بین ۰ و ۱ محدود شده است. خاصیت دوم نشان می دهد که copula تداعی کننده است، به این معنی که ترتیب ترکیب متغیرها مهم نیست. ویژگی سوم متقارن بودن مفصل را تضمین می کند، به این معنی که رابطه بین دو متغیر بدون توجه به ترتیب آنها یکسان است.

از مفصل‌ها برای مدلسازی انواع مختلف وابستگی بین متغیرهای تصادفی استفاده می شود، مانند:

۱. وابستگی مثبت: دو متغیر به طور مثبت وابسته هستند اگر مقادیر آنها با هم افزایش یا کاهش یابد.
۲. وابستگی منفی: دو متغیر به صورت منفی وابسته هستند اگر مقادیر آنها در جهت مخالف حرکت کنند.
۳. استقلال: دو متغیر مستقل هستند اگر مقادیر آنها نامرتبط باشد.

مفصل کاربردهای متعددی در آمار و امور مالی دارد، از جمله:

۱. مدیریت ریسک: ز مفصل‌ها برای مدلسازی وابستگی ریسک در امور مالی و بیمه استفاده می شود.
۲. مدلسازی چند متغیره: از کوپولها برای مدلسازی روابط بین متغیرهای تصادفی چندگانه استفاده می شود.
۳. تخمین وابستگی: از مفصل ها برای تخمین قدرت و نوع وابستگی بین متغیرها استفاده می شود.

به طور خلاصه، یک مفصل آماری یک تابع ریاضی است که رابطه بین دو یا چند متغیر تصادفی را توصیف می کند و امکان مدل سازی انواع مختلف وابستگی و استقلال بین آنها را فراهم می کند.

یکی از شاخه‌های مهم و کابردی در علم آمار تحلیل بقاء است. برای این منظور در بخش دوم از این فصل به تشریح مقدماتی از این شاخه می‌پردازیم و حالت‌های مختلف آن را توضیح می‌دهیم. علاوه بر این برخی از تعاریف و مفاهیم اولیه که درک آن‌ها برای فهم کامل موضوع ضروری است را به طور کامل و تحت عنوان مثال و قضایا شرح می‌دهیم. همچنین با توجه به اهمیت و کابرد داده‌های شکست در مطالعات تحلیل بقاء پس از اشاره به این داده‌ها، دو نمونه از مدل‌های پارامتری این قبیل داده‌ها مانند مدل وایبل، را به تفصیل بیان کرده بحث را خاتمه می‌دهیم.

۱-۲ مقدمات آنالیز بقاء

یکی از انواع مسائلی که مورد علاقه‌ی شدید محققین است، اهمیت فاصله زمانی تا وقوع بعضی حوادث مانند مرگ و میر و ... می‌باشد؛ یعنی پرداختن و توجه نمودن به گروهی از افراد به‌طوری که پس از مدتی

برای هر کدام از آن‌ها یک نقطه‌ی زمانی به نام شکست یا وقوع حادثه تعریف می‌گردد، که شکست یا حادثه مورد نظر حداکثر یکبار برای هر فرد می‌تواند اتفاق بیفتد. از جمله مواردی که می‌توان مصداق شکست یا حادثه‌ی مورد نظر باشد،

- طول عمر یک ماشین صنعتی،
- زمان بقاء یک بیمار پس از درمان یا انجام یک عمل جراحی،
- زمان طلاق یک زوج پس از ازدواج،
- روی آوردن دوباره یک معتاد پس از ترک اعتیاد

و مثال‌هایی از این قبیل هستند.
از آنجایی که این روش‌ها در ابتدا غالبا برای مطالعات مرگ و میر به کارده می‌شدند و برای این منظور طراحی شده‌بودند، عنوان " تجزیه و تحلیل بقاء" بر آن نهاده شده است.

تحلیل بقاء:

از نظر علم آمار، مجموعه‌ای از روش‌های مختلف آماری در تحلیل متغیرهای تصادفی نامنفی است که مقدار آن می‌تواند زمان شکست یک مولفه فیزیکی یا زمان مرگ یک واحد زنده باشد.
جالب است بدانید، آنالیز بقاء اولین بار توسط ستاره‌شناس معروف " ادموند هالی" صورت گرفت؛ به این ترتیب که، او در دوره‌ای از زندگی‌اش به ثبت سن مرگ هم‌شهری‌های خود پرداخت و جدولی از اطلاعات را فراهم نمود که نشان می‌دهد چند درصد از این آدم‌ها در هر دوره‌ی سنی می‌میرند. امروزه به چنین جدولی **جدول طول عمر** [۲] گفته می‌شود. از آن‌جایی که در اولین تلاشهای این چنینی، پیامدی که بررسی می‌شد مردن یا زنده ماندن افراد بود، به روش تحلیلی این داده‌ها **آنالیز بقاء** [۳] نام دادند. برای بقاء افراد مورد مطالعه از نمودار پلکانی استفاده می‌شود که در این نمودار محورهای افقی و عمودی به‌ترتیب نشاندهنده‌ی زمان ورود به مطالعه و بقاء است.

---
[۲] Life table
[۳] Survival analysis

زمان بقاء:

فاصله زمانی ورود به مطالعه تا وقوع حادثه، یکی از متغیرهای مهم در تجزیه و تحلیل داده‌های بقاء است. زمان بقاء فرد iام مورد مطالعه یک متغیر تصادفی نامنفی است که با T_i نشان داده می‌شود که فاصله زمانی نقطه ورود فرد i ام نمونه به مطالعه تا زمان وقوع حادثه مورد نظر مانند مرگ و ... را شامل می‌شود؛ واضح است که این متغیر برای هر فرد متفاوت است. برای مشخص کردن زمان بقاء بایستی سه مولفه زیر به‌طور دقیق مشخص شوند و هیچگونه ابهامی برای آن‌ها وجود نداشته باشد:

- *مبدأ زمان:* مبدا زمانی در زمان مطالعه بایستی برای هر فرد مشخص شود. البته یکسان بودن تقویم مبدا زمان برای هر فرد ضروری نیست و بسیاری از اوقات امکان‌پذیر نیست چرا که در بسیاری از مطالعات زمان ورود شخص به مطالعه را به عنوان نقطه ورود در نظر می‌گیرند. به عنوان مثال در یک مطالعه بالینی، طبیعی است که زمان انتساب بیماران به درمان‌های مختلف را به عنوان مبدأ زمانی در نظر می‌گیرند یا در حوادثی مانند طلاق چون زوجین پیوسته در معرض چنین واقعه‌ای هستند، زمان ازدواج به عنوان مبدأ در نظر گرفته می‌شود.
  یکی از حالتها این است که همه آزمودنی ها را به سیستم وارد کنیم و منتظر رخداد مورد نظر باشیم. و پایان مطالعه را از قبل مشخص کنیم یا اینکه بگوییم مطالعه را زمانی پایان می دهیم که تعداد مورد رخداد به عدد خاصی برسد. در این حالت اولی را سانسور از راست نوع ۱ و حالت دوم را سانسور از راست نوع ۲ می نامند. تابع درستنمایی را برای آنها می توان نوشت.

- *مقیاس و واحد اندازه‌گیری گذشت زمان:* اندازه‌گیری زمان در حقیقت بوسیله وقت ساعتی، که همان وقت حقیقی است یا بر حسب میزان استفاده از یک سیستم مانند تعداد رادیوگرافی‌های انجام شده توسط یک دستگاه، تعداد کپی، تعداد کیلومتر طی شده ... محاسبه می‌گردد.

- *مفهوم شکست یا وقوع حادثه:* وقوع حادثه یا شکست نیز باید مانند سایر موارد دقیقا مشخص شود چراکه این زمان وقوع حادثه است که به عنوان پیامد مورد نظر بایستی تجزیه و تحلیل روی آن انجام شود. این حوادث ممکن است گاه ناگوار باشند مانند مرگ به دلیلی خاص مانند بروز یک بیماری جدید، یا عود بیماری و بازگشت به شرایط اولیه بیماری و گاه نیز خوشایند باشند مانند مرخص شدن یک بیمار از بیمارستان، ازدواج، یافتن شغل، فارغ التحصیلی ...

به عنوان مثال در مطالعات بالینی زمان بقاء، زمان مرگ به پیشرفت یک نشانه خاص یا عود بیماری پس از بهبود آن اشاره دارد. در اینگونه مطالعات نقطه پایان به طور دقیق مشخص و تعریف شده‌است اما نقطه شروع به وضوح مشخص نیست و بهترین حالت آن زمانی است که بیماری تشخیص داده شده‌است.

زمان شروع مطالعه و ورود بیمار به سیستم:

در یک مطالعه بقاء همه افرادی که مورد مطالعه قرار می‌گیرند، همزمان با هم وارد مطالعه نمی‌شوند و در نتیجه همزمان با هم درمان نشده و تحت مراقبت قرار نمی‌گیرند، بلکه ممکن است مطالعات ماه‌ها و حتی سال‌ها به طول انجامد. اما بعد از درمان، آن‌ها را برای مدت مشخصی تحت پیگیری و مراقبت قرار می‌دهند تا اینکه زمان تقویمی تعریف شده برای مطالعه سپری گشته و به پایان برسد یا او را از دست داده و یا با وجود زنده بودنش از دسترس مطالعه خارج شود. این دوره زمانی تقویمی مشخص که افراد نمونه و بیماران بایستی طی آن زمان مورد پیگیری و مراقبت قرار بگیرند، **زمان مطالعه** نامیده می‌شود و زمانی را که یک بیمار در مطالعه می‌گذراند (یعنی از زمان ورودش به مطالعه تا زمانی که مورد پیگیری و مراقبت قرار می‌گیرد) **زمان بیمار** می‌نامند که حداکثر مقدار این زمان، زمان مطالعه است.

مفهوم سانسور:

ویژگی مهمی که همواره تجزیه و تحلیل داده‌های بقاء را با مشکل مواجه می‌سازد، سانسور است. سانسور به مواردی اطلاق می‌شود که آزمودنی در طول مدت پیگیری از دست رفته و گم شده‌اند و قادر به ثبت وضعیت نهایی آن‌ها بر حسب وقوع حادثه نیستیم. بعضی اوقات قبل از اتمام زمان مطالعه، حادثه مورد نظر برای برخی از افراد نمونه اتفاق نمی‌افتد؛ بدین معنی که قبل از اینکه حادثه موردنظر برای همه افراد مورد مطالعه اتفاق بیفتد، دوره مشاهده سپری شده و به پایان می‌رسد. لذا مشخص نیست که پیشامد مورد نظر برای این دسته از افراد نمونه چه زمانی اتفاق می‌افتد یا اینکه اصلا اتفاق خواهد افتاد یا خیر و فقط فقط می دانیم که حادثه مورد نظر تا پایان مطالعه اتفاق نیفتاده است. از طرف دیگر ممکن است برخی از افراد مورد مطالعه قبل از سپری شدن دوره مشاهده حاضر به همکاری بیشتر نباشند و یا اینکه به هر دلیل دیگری مانند مهاجرت، مرگ و ... از مطالعه خارج شوند و دیگر در دسترس نباشند. به همه این موارد سانسور اطلاق می‌گردد.
همانطور که گفته شد، سانسور در داده‌های بقاء رایج بوده و روش‌های آماری متداول نمی‌تواند پاسخگوی آنالیز این اطلاعات گم شده یا به اصطلاح سانسور شده، باشند. از آنجایی که این اطلاعات گم شده مهم و ارزشمند بوده و حاوی این نکته است که تا زمان گم شدن (سانسور شدن) فرد نمونه، حادثه مورد نظر برایش اتفاق نیفتاده است، بنابراین نه می‌توان از آن صرف‌نظر کرد و نه می‌توان زمان گمشدگی را با زمان وقوع حوادث و زمان بقاء یکسان پنداشت؛ لذا تجزیه و تحلیل آن‌ها روش‌های خاصی را می‌طلبد.

انواع سانسور

- **سانسور از راست نوع I:** در این نوع سانسور، ابتدا طول دوره آزمایش را تعیین می شود و بعد از این زمان هر تعداد از آزمودنی ها که اتفاقی برای آنها نیافتاده است به‌عوان سانسور از راست محسوب می شوند.

  ۱) تعداد آزمودنی ها n و تعداد شکستها r است: $x_1, x_2, ..., x_n$.

  ۲) C زمان پایان دوره از قبل تعیین می شود.

  ۳) r تعداد افرادی که شکست می خورند یک متغیر تصادفی است.

- **سانسور از راست نوع II:** در این نوع سانسور، ابتدا تعداد افرادی را باید رخداد موردنظر برای آنها اتفاق بیافتد تعیین می شود و بعد از این هر تعداد از آزمودنی ها که اتفاقی برای آنها نیافتاده است به‌عوان سانسور از راست محسوب می شوند.

  ۱) تعداد آزمودنی ها n و تعداد شکستها r است: $x_1, x_2, ..., x_n$.

  ۲) r تعداد افراد شکست دوره ثابت و از قبل تعیین می شود.

  ۳) C_T پایان دوره از قبل معلوم نیست و یک متغیر تصادفی است.

- **سانسور از راست تصادفی:** اگر از دست دادن فرد نمونه پس از ورود در مطالعه اتفاق بیفتد آن مشاهده را سانسور از راست می‌گویند. دلیل این نام‌گذاری این است که فرد تحت مراقبت روی محور در طرف راست نقطه‌ی ورود به مطالعه می‌باشد؛ یعنی $T_i \geq C_i$ است. (t_0= زمان ورود فرد به مطالعه)

- **سانسور از چپ تصادفی:** اگر زمان واقعی بقاء فرد نمونه قبل شروع زمان مطالعه باشد یعنی حادثه مورد نظر قبل از ورود فرد به مطالعه رخ داده شده باشد، آن مشاهده را سانسور از چپ می‌نامند و این نام بدان جهت است که زمان از دست دادن فرد تحت مراقبت و پیگیری، در طرف چپ نقطه ورود به مطالعه یعنی $T_i < C_i$ است.

- **سانسور فاصله‌ای:** در حقیقت ترکیبی از سانسور راست و چپ با هم است و زمانی رخ می‌دهد که حادثه مورد نظر بین زمان L و R اتفاق افتاده باشد. بنابراین اگر متغیر T مشاهده‌ای باشد که در فاصله زمانی L و R به وقوع بپیوندد، یعنی برای فرد ام i $L_i < T_i < R_i$ باشد، آن را سانسور فاصله‌ای می‌نامند. بدیهی است اگر $L_i = 0$ باشد سانسور از چپ و اگر $R_i = \infty$ باشد سانسور از راست اتفاق افتاده‌است.

- **سانسور تصادفی:** به مواردی اطلاق می‌شود که توقف مشاهدات تحت کنترل و اختیار محقق نباشد. دلایل متعددی برای این کار وجود دارد؛ از جمله، افراد مورد بررسی به مکان دیگری مهاجرت کنند و

برقراری تماس با آن‌ها امکان پذیر نباشد یا از ادامه همکاری و شرکت در مطالعه منصرف شده‌باشند و یا حتی به دلایل دیگری از جمله تصادف و ...فوت کرده باشند. حالت دیگر سانسور تصادفی هنگامی است که زمان توقف مطالعه برای همه افراد نمونه یکسان و از قبل مشخص شده باشد ولی زمان ورود آن‌ها به مطالعه متفاوت و تصادفی باشد.
در این نوع بررسی‌ها به جای تعیین یک دوره زمانی مشخص، تاریخ اتمام بررسی را از قبل تعیین می‌کنند. بنابراین تمامی افرادی که در آن تاریخ زنده هستند، سانسور تصادفی می‌باشند، چراکه زمان ورودشان به مطالعه تحت اختیار و کنترل محقق بوده است.

۱-۳ تعاریف:

تعریف ۱-۱. تابع توزیع: تابع توزیع F، یک تابع با دامنه $\overline{R}$ است به قسمی که:

۱ F نانزولی است

۲ $F(-\infty) = ۰$ و $F(\infty) = ۱$.

تعریف ۱-۲. یک متغیر تصادفی پیوسته نامیده می‌شود اگر تابع توزیع تجمعی آن پیوسته باشد.

تعریف ۱-۳. تابع بقاء تابع بقاء عبارت است از احتمال اینکه فرد با مؤلفه‌ی بیشتر از t واحد زمانی زنده بماند. این تابع را با نماد $S(t)$ نمایش می‌دهیم و بصورت زیر محاسبه می‌کنیم:

$$S(t) = P(T > t) = ۱ - F(t)$$

این تابع یک تابع غیرصعودی و پیوسته از راست با ویژگی‌های زیر می‌باشد:

- $S(۰) = ۱ - F(۰) = ۱$
- $lim_{t\to\infty} S(t) = lim_{t\to\infty} ۱ - F(t) = ۰$

این رابطه نشان می‌دهند که در لحظه‌ی اول، تمامی واحدها زنده هستند و بالاخره به پایان می‌رسند؛ پس زمان شکست بی‌نهایت نداریم.

- $S(t)$ نرخ تابع بقاء را نشان می‌دهد؛ به‌طور مثال اگر زمان بر حسب سال باشد، $S(۲)$ نرخ بقاء دو ساله را نشان می‌دهد.

- در عمل وقتی از داده‌های واقعی استفاده می‌کنیم، تابع بقاء پله‌ای است. علاوه بر این چون طول دوره مطالعه هیچ‌گاه بی‌نهایت نیست و ممکن است مؤلفه‌هایی در معرض شکست باشند، این احتمال وجود دارد که هیچ فردی پیشامد مورد نظر را تجربه نکند.

تعریف ۱-۴. تابع بقاء تجربی: تابع بقاء n متغیر تصادفی مستقل و هم توزیع $x_1, \cdots, x_n$ را با $S_n(x)$ نشان می‌دهیم که بصورت زیر تعریف می‌شود:

$$S_n(x) = \frac{\#\{x_i > x\}}{n} = \frac{1}{n}\sum I_{(x,\infty)}(x_i)$$

تعریف ۱-۵. دو مشاهده (x_1, y_1) و (x_2, y_2) هم‌شیب هستند اگر $x_2 > x_1$ و $y_2 > y_1$ و یا اینکه $x_2 < x_1$ و $y_2 < y_1$ باشد؛ یا به‌عبارت دیگر اگر $(x_2 - x_1)(y_2 - y_1) > 0$ باشد. و بالعکس دو مشاهده ناهم شیب نامیده می‌شوند اگر $(x_2 - x_1)(y_2 - y_1) < 0$.

تعریف ۱-۶. $\delta_{X,Y}$ اندازه‌ای از وابستگی نامیده می‌شود اگر:

۱ $0 \leq \delta_{X,Y} \leq 1$

۲ $\delta_{X,Y} = \delta_{Y,X}$

۳ $\delta_{X,Y} = 0$ اگر X و Y مستقل از هم باشند و بالعکس، $\delta_{X,Y} = 1$ است اگر X و Y توابعی یکنوا از یکدیگر باشند.

۴ اگر α و β توابعی اکیدا یکنوا روی برد X و برد Y باشند، آنگاه $\delta_{XY} = \delta_{\alpha(X)\beta(Y)}$.

تعریف ۱-۷. فرض کنید $\phi \neq S_1, S_2 \subset \overline{\mathbf{R}}$ و فرض کنید H یک تابع $S_1 \times S_2 \longrightarrow \mathbf{R}$ باشد، Hـحجم [۴] از $B = [x_1, x_2] \times [y_1, y_2]$ به‌صورت زیر تعریف می‌شود:

$$V_H(B) = H(x_2, y_2) - H(x_2, y_1) - H(x_1, y_2) + H(x_1, y_1)$$

تعریف ۱-۸. H یک تابع ۲ ـصعودی [۵] است اگر و تنها اگر برای همه $B \subset S_1 \times S_2$، $V_H(B) \geq 0$ توجه شود که $\overline{\mathbf{R}} = \mathbf{R} \cup \{-\infty, +\infty\}$ می‌باشد.

مثال ۱-۹. فرض کنید H یک تابع، تعریف شده روی I^2 با ضابطه $H(x, y) = (2x - 1)(2y - 1)$ باشد. H ۲ـصعودی است با این وجود تابعی نزولی از x برای هر $y \in (0, \frac{1}{2})$ و یک تابع نزولی از y برای هر $x \in (0, \frac{1}{2})$ می‌باشد.

---

[۴] $H - Volume$

[۵] ۲-increasing

مثال ۱-۱۰. فرض کنید H یک تابع، تعریف شده روی $I^۲$ با ضابطه $H(x,y) = max(x,y)$ باشد. آنگاه تابع H یک تابع نانزولی از x و y می‌باشد. با این وجود $V_H(I^۲) = -۱$ و این بدان معنی است که H یک تابع ۲ـصعودی نیست.

از این دو مثال نتیجه می‌شود که اگر تابع H یک تابع ۲ـصعودی باشد نمی‌توان این برداشت را کرد که در هر مقداری نانزولی است.

تعریف ۱-۱۱. فرض کنید$maxS_۱ = b_۱$ و $maxS_۲ = b_۲$ وجود داشته باشند، آنگاه حاشیه‌ای‌های Fو G از H بصورت زیر داده می‌شوند:

$$F : S_۱ \longrightarrow \mathbf{R}, \quad F(x) = H(x, b_۲)$$
$$G : S_۲ \longrightarrow \mathbf{R}, \quad G(y) = H(b_۱, y)$$

توجه شود که $b_۱$ و $b_۲$ هر دو می‌توانند $+\infty$ باشند.

تعریف ۱-۱۲. فرض کنید $minS_۱ = a_۱$ و $minS_۲ = a_۲$ وجود داشته باشند. H، grounded نامیده می‌شود اگر برای هر $(x,y) \in S_۱ \times S_۲$

$$H(a_۱, y) = H(x, a_۲) = ۰$$

($a_۱$و $a_۲$ می‌توانند $-\infty$ باشند.)

مثال ۱-۱۳. فرض کنید H یک تابع با دامنه $[-۱,۱] \times [۰,\infty]$ بصورت زیر داده شده باشد:

$$H(x,y) = \frac{(x+۱)(e^y - ۱)}{x + ۲e^y - ۱}$$

H، grounded است چون، $H(x,۰) = ۰$ و $H(-۱,y) = ۰$. همچنین توابع حاشیه‌ای $F(x)$ و $G(y)$ بصورت زیر داد می‌شوند:

$$F(x) = H(x,\infty) = \frac{(x+۱)}{۲}$$
$$G(y) = H(۱,y) = ۱ - e^{-y}$$

اگر H، یک تابع ۲ـصعودی باشد آنگاه بنا به تعریف (۸-۱) داریم:

$$H(x_۲,y_۲) - H(x_۱,y_۲) \geq H(x_۲,y_۱) - H(x_۱,y_۱) \qquad (۱-۱)$$

و

$$H(x_۲, y_۲) - H(x_۲, y_۱) \geq H(x_۱, y_۲) - H(x_۱, y_۱) \qquad (۲-۱)$$

برای هر $[x_۱, x_۲] \times [y_۱, y_۲] \subset S_۱ \times S_۲$ با قرار دادن $x_۱ = a_۱$ و $y_۲ = a_۲$ در رابطه (۲) داریم:

لم ۱-۱۴. *هر تابع grounded و ۲ـ صعودی، $H : S_۱ \times S_۲ \longrightarrow \boldsymbol{R}$، در هر مقداری نانزولی است. برای همه $x_۱ \leq x_۲$ در $S_۱$ و $y_۱ \leq y_۲$ در $S_۲$*

$$H(., y_۲) \geq H(., y_۱), \quad H(x_۲, .) \geq H(x_۱, .)$$

**از لم (۱-۱۴) نتیجه می‌شود که رابطه (۱-۲) و (۱-۳) تحت قدر مطلق مقادیر نیز برقرار است.*

لم ۱-۱۵. *فرض کنید $S_۱$ و $S_۲$ زیرمجموعه‌هایی ناتهی از $\overline{\boldsymbol{R}}$ و H یک تابع ۲ـ صعودی با دامنه $S_۱ \times S_۲$ باشند. همچنین فرض کنید $x_۱$ و $x_۲$ دو نقطه از $S_۱$ باشند به قسمی که $x_۱ < x_۲$ و نیز $y_۱$ و $y_۲$ دو نقطه از $S_۲$ باشند که $y_۱ < y_۲$ باشد. آنگاه تابع $t \to H(t, y_۲) - H(t, y_۱)$ نانزولی روی $S_۱$ و تابع $t \to H(x_۲, t) - H(x_۱, t)$ نیز نانزولی روی $S_۲$ می‌باشد.*

لم ۱-۱۶. *فرض کنید، $H : S_۱ \times S_۲ \longrightarrow \boldsymbol{R}$، یک تابع ۲ـ صعودی و grounded باشد. آنگاه برای هر $[x_۱, x_۲] \times [y_۱, y_۲] \subset S_۱ \times S_۲$ داریم:*

$$|H(x_۲, y_۲) - H(x_۱, y_۱)| \leq |F(x_۲) - F(x_۱)| + |G(y_۲) - G(y_۱)|$$

اثبات. بنا به نامساوی مثلثی داریم:

$$H(x_۲, y_۲) - H(x_۱, y_۱) \leq | H(x_۲, y_۲) - H(x_۱, y_۲) | + | H(x_۱, y_۲) - H(x_۱, y_۱) |$$

حال فرض کنید که $x_۱ \leq x_۲$ باشد. چون H، ۲ـ صعودی و grounded دارای حاشیه‌ای است و بنا به لم‌های (۱-۱۴) و (۱-۱۵)داریم:

$$۰ \leq H(x_۲, y_۲) - H(x_۱, y_۲) \leq F(x_۲) - F(x_۱)$$

نابرابری مشابه نیز وقتی برقرار خواهد بود که $x_۲ \leq x_۱$ بنابراین نتیجه می‌شود، برای هر $x_۱$ و $x_۲$ در $S_۱$، $| H(x_۲, y_۲) - H(x_۱, y_۲) | \leq | F(x_۲) - F(x_۱) |$ و به طور مشابه برای هر $y_۱$ و $y_۲$ در $S_۲$ نیز، $| H(x_۱, y_۲) - H(x_۱, y_۱) | \leq | G(y_۲) - G(y_۱) |$. □

تعریف ۱-۱۷. تابع دو قسمتی: تابع دو قسمتی H تابعی است که دامنه آن زیرمجموعه‌ای از $\overline{R}^{۲}$ و برد آن نیز زیرمجموعه‌ای از R باشد.

تعریف ۱-۱۸. فضای متری: یک فضای متری شامل یک مجموعه S و یک متر d است که فاصله بین نقاط p و q را اندازه می‌گیرد.

داده‌های زمان شکست

منظور از داده‌های زمان شکست [۶] متغیرهای تصادفی مثبتی است که زمان‌های مربوط به وقوع یک رخداد را شرح می‌دهند. مثال‌هایی از این قبیل مربوط به رخداد شکست یا بقاء می‌باشد که می‌تواند زمان خراب یک قطعه مکانیکی از یک دستگاه یا زمان بهبودی یک بیمار باشد. این قبیل داده‌ها در مطالعات جمعیت شناسی، جامعه شناسی، بیولوژیکی، اپیدمیولوژیک و... به وفور دیده می‌شوند.
مدل‌های پارامتری زمان شکست شامل مدل نمایی، مدل وایبل، مدل لوگ نرمال و مدل لوگ لجستیک می‌باشد که در این بخش دو مدل نمایی و وایبل را تعریف می‌کنیم. توزیع های لگ نرمال، گاما و ...

مدل نمایی

این مدل ساده‌ترین مدل زمان شکست است که در آن فرض می‌شود، نرخ شکست مستقل از زمان t است. تحت این مدل تابع بقاء و تابع توزیع به‌ترتیب به صورت زیر تعریف می‌شوند:

$$S(t) = e^{-\lambda t} \qquad ; f(t) = \lambda e^{-\lambda t}$$

مدل وایبل

توزیع وایبل [۷] یکی از مهمترین توزیع‌های طول عمر در مطالعات قابلیت اعتماد و تحلیل بقا است. متغیر تصادفی T دارای توزیع وایبل با پارامترهای (λ, β) گوییم هرگاه تابع توزیع آن به صورت زیر باشد:

$$F(t) = ۱ - e^{(\lambda t)^{\beta}}, t > ۰, \lambda > ۰, \beta > ۰$$

[۶] failur time
[۷] Weibull Distribution

در این‌جا، λ پارامتر مقیاس و β پارامتر شکل در توزیع است. تابع بقاء و تابع چگالی توزیع وایبل به ترتیب به فرم زیر می‌باشند:

$$S(t) = e^{-(\lambda t)^{\beta}}$$

$$f(t) = \lambda^{\beta}\beta t^{\beta-1}e^{-(\lambda t)^{\beta}}, t > 0, \lambda > 0, \beta > 0$$

واضح است درحالت، $\beta = 1$ توزیع وایبل به توزیع نمایی تبدیل می‌شود. از جمله دلایل اهمیت توزیع وایبل در تحلیل بقا به موارد زیر می‌توان اشاره کرد:
الف) نرخ خطر توزیع وایبل دارای فرم بسته است. ($h(t) = \lambda^{\beta}\beta t^{\beta-1}$) و همین امر باعث می‌شود تحلیل در مورد داده‌ها با این توزیع راحتتر صورت گیرد.
ب) توزیع وایبل مدلی است که با آن می‌توان انواع داده‌ها را تفسیر کرد. داده‌های با نرخ خطر صعودی ($\beta > 1$)، نرخ خطر نزولی ($\beta < 1$) و نرخ خطر ثابت ($\beta = 1$). ج) اگر T دارای توزیع وایبل باشد آنگاه لگاریتم آن دارای توزیعی است که متعلق به خانواده مقیاس‌ـ‌مکانی است که در استنباط آماری یک ویژگی مهم محسوب می‌شود. د) به طور کلی اگر $T_1, \ldots, T_n$ متغیرهای تصادفی پیوسته باشند که تابع توزیع آن‌ها در نزدیکی صفر به فرم زیر باشد:

$$Z_n = a_n min(T_1, \cdots, T_n)$$

که در آن $a_n = n^{\frac{1}{d}}$ می‌باشد و برای nهای بزرگ ($n \to \infty$)در توزیع به سمت توزیع وایبل با پارامترهای ($c^{\frac{1}{d}}, d$)) میل می‌کند.

فصل ۲

تابع مفصل

۲-۱ مقدمه

اکثر مدل‌بندی‌های آماری بر اساس استقلال بین داده‌ها انجام می‌پذیرد. با این وجود در بسیاری از موارد متغیرهای مورد مطالعه به نوعی وابسته به یکدیگرند. ساختارهای متنوعی برای توصیف و تحلیل چنین وابستگی‌هایی معرفی شده‌اند که مدل‌های بر مبنای مفصل از مهمترین و پرکاربردترین این مدل‌ها می‌باشند.

واژه مفصل یا Copula یک اسم لاتین به معنای یک رابطه یا پیوند است (Cassell's Latin Dictionary).

واژه مفصل نخستین بار در قضیه‌ی مشهور اسکلار (Sklar,1959) نام برده شد. این قضیه بیان می‌کند که هر تابع توزیع توأمی در قالب تابع مفصل مربوطه‌اش بیان می‌شود و بالعکس. بنا به این قضیه، یک مفصل تابعی است که حاشیه‌ای‌های یک متغیره را به توابع توزیع چندمتغیره متصل می‌کند به قسمی که ساختار وابستگی در توزیع چندمتغیره حفظ شود. همچنین این تابع در کار افراد بسیاری چون، Frechet, Dall' Aglio, Feron و بسیاری نویسندگان دیگر که توابع توزیع چندمتغیره با توابع حاشیه‌ای یک متغیره ثابت، اساس مطالعه‌ی آنهاست نیز به چشم می‌خورد. علاوه بر این نتایج اساسی بسیاری راجع به مفصل را می‌توان در کاری از WAssily Hoeffding جستجو کرد. بنا به مقاله‌ی Hoeffding. 1940, 1941 ، فرد می‌تواند به توابع استاندارد دومتغیره‌ای که تکیه‌گاه‌شان در واحد مربع $[-\frac{1}{2},\frac{1}{2}]^2$ و توابع توزیع حاشیه‌ای‌هایشان یکنواخت روی فاصله $[-\frac{1}{2},\frac{1}{2}]^2$ می‌باشند، دست پیدا کند. در سال ۱۹۹۱، Hoeffding به همراه Schwizer واحد مربع $[0,1]^2$ را به جای $[-\frac{1}{2},\frac{1}{2}]^2$ برای نرمال شدنش انتخاب کردند و از این‌جا بود که مفصل کشف شد. Hoeffding همچنین ممکن‌ترین حدود نابرابری برای چنین توابعی را به تصویر کشید، توابعی پیوسته را مطابق با حدودشان توصیف کرد و اندازه‌هایی از وابستگی را بررسی کرد که مقیاس‌ـ پایا [1] ، به عنوان مثال پایا تحت تبدیلات صعودی، بودند.

جدای از کار Hoeffding ، Frechet در سال ۱۹۵۱ مستقلا همان نتایج را نشان داد که با نام حدود Frechet و یا به عبارت دیگر، کلاس Frechet عنوان شد. برای به رسمیت شناختن این ایده قابل اهمیت، ما به حدود Frechet-Hoeffding یا کلاس Frechet-Hoeffding اشاره خواهیم کرد.

بعد ازFrechet ، Hoeffding و Sklar توابعی تحت عنوان مفصل توسط چندین نویسنده کشف شد که Sampson, Kimeldoref.1975 b از آنها با عنوان نمایش‌های یکنواخت و نیز Deheuvels, 1978 آنها را توابع وابستگی نامیدند.

در آن هنگام که اسکلار مقاله‌اش را با عنوان" **مفصل** "نوشته بود، در حال مساعدت با Berthold

[1] invariant-Scale

Schwezer جهت توسعه نظریه فضای متری احتمال یا فضای متری PM بود؛ این امر منجر به رسیدن به بخش اعظمی از نتایج پیرامون مفصل‌ها در مسیر مطالعه فضاهای متری PM در سال‌های ۱۹۵۸ تا ۱۹۷۵ شد.

در فضای متری احتمال، ما فاصله $d(p,q)$ را با یک تابع توزیع F_{pq} عوض می‌کنیم، که مقدار $F_{pq}(x)$ برای هر عدد حقیقی احتمالی این است که فاصله بین p و q کمتر از x است. اولین مشکل در ساخت فضاهای متری وقتی نمایان می‌شود که فرد علاقه‌مند است تا یک قیاس احتمالی از نابرابری مثلثی $d(p,r) \leq d(p,q) + d(q,r)$ داشته باشد.حال این سوال مطرح می‌شود که رابطه‌ی بین F_{pq}، F_{qr} و F_{pr} برای همه p، q و r چه خواهد بود؟

Karl Menger در سال ۱۹۴۲ $F_{pr}(x+y) \geq T(F_{pq}(x), F_{qr}(y))$ را پیشنهاد کرد؛ در حالیکه T، یک اندازه مثلثی یا t مقدار است. همانند یک مفصل، یک tمقدار، نگاشت $[۰,۱]^۲$ به $[۰,۱]$ است و توابع توزیع را به هم متصل می‌کند. برخی از t مقدارها مفصل هستند و بالعکس. بنابراین در مفهوم، بدیهی است که مفصل‌ها خواهند توانست در مطالعه فضاهای متری رخ دهند. برای مطالعه کامل رفتار نظریه فضاهای متری PM و تاریخچه آن Schweizer and Sklar.1983 , Schweizer, 1991 را ببینید.

در زمره‌ی نتایج بسیار مهم در فضاهای متری، برای آماردان‌ها کلاسی از t مقدارها که آنان t مقدارهایی هستند که برای همه uها در (۱،۰)، در $T(u,u) < u$ صدق می‌کنند. این t مقدارها، مفصل‌هایی هستند که مفصل ارشمیدسی یا Archimedean نامیده می‌شوند. این مفصل‌ها به دلیل فرم ساده‌شان، سهولت در ساخت و بسیاری ویژگی‌های زیبای دیگر به وفور در بحث‌های توزیع‌های چندمتغیره دیده می‌شوند. به عنوان مثال Genest and Mackay, 1986 a,b، Marshall and Olkin,1988 و Joe , 1993,1997 را ببینید.

حال به ارتباط بین مفصل و اندازه‌های وابستگی می‌پردازیم؛ مفصل‌ها در سال ۱۹۵۹ در مفهومی از فضاهای متری احتمال و بعدها به عنوان ابزاری برای فهم رابطه بین برآمدهای چندمتغیره گسترش یافتند.

اولین مقاله که صریحا مفصل‌ها را به مطالعه وابستگی بین متغیرها متغیرهای تصادفی ارتباط می‌دهد در کاری از Schweizer and Wolff درسال ۱۹۸۱ دیده شده است. در این مقاله،Schwezer و Wolff به بحث راجع به تحلیل مقدار بحرانی Renyi (1959)، برای اندازه‌گیری وابستگی بین جفت متغیرهای تصادفی می‌پردازند. آنها نیز ویژگی‌های اساسی پایایی مفصل‌ها را تحت تبدیلات اکیدا یکنوا از متغیرهای تصادفی و همچنین اندازه‌ای از وابستگی را که با عنوان سیگما Schweizer and Wolff معروف می‌باشد، بیان می‌کنند.

به منظور دسترسی به تئوری مفصل‌ها و کاربردهای آن می‌توان علاقه مندان را به Nelsen,2006 ارجاع دهیم. در بخش 2-2 به تعریف مفصل و برخی ویژگی‌های اساسی آن می‌پردازیم. بخش 2-3 نیز به قضیه اسکلار و نقش آن اختصاص داده شده است که به ما این اجازه را می‌دهد تا مفصل را به عنوان یک تابع وابستگی در نظر بگیریم. در بخش 2-4 به بیان ویژگی‌های مفصل، مفاهیم وابستگی و دو اندازه از وابستگی که مشهور به ضرایب همبستگی رتبه‌ای τ کندال و ρ اسپیرمن می‌باشند بر حسب تابع مفصل می‌پرازیم. همچنین مثال‌هایی از چند خانواده پرکاربرد از خانواده مفصل‌ها را نیز در بخش 2-5 به تصویر می‌کشیم.

۲-۲ تعاریف و ویژگی‌های اساسی

تعریف ۲-۱. هر تابع ۲ـصعودی و grounded، $C' : S_{۱۲} \longrightarrow \mathbf{R}$، برای همه $(u,v) \in S_۱ \times S_۲$ یک زیر مفصل (۲ بعدی) می‌باشد اگر

$$C'(u, ۱) = u, \quad C'(۱, v) = v$$

(در حالیکه $S_۱$ و $S_۲$ زیرمجموعه‌هایی از $[۰, ۱]$ می‌باشند که شامل ۰ و ۱ می‌باشند.)

تعریف ۲-۲. یک مفصل (۲ بعدی)، زیرمفصلی است که دامنه آن $[۰, ۱]^۲$ می‌باشد.

تعریف تئوری تابع مفصل:

برای تعریف تئوری از تابع مفصل ابتدا یک جفت از متغیرهای تصادفی x و y با توابع توزیع به ترتیب$F(x) = P[X \leq x]$ و $G(x) = P[Y \leq y]$ و تابع توزیع توام

$$H(x, y) = P[X \leq x, Y \leq y]$$

در نظر بگیرید. برای هر جفت از مقدارهای حقیقی (x, y) می‌توانیم سه مقدار $F(x)$، $G(y)$ و $H(x, y)$ را نسبت دهیم(توجه کنید که هر یک از این سه مقدار بین $[۰, ۱]$ قرار می‌گیرند)؛ به عبارت دیگر هر جفت (x, y) از مقادیر حقیقی به یک نقطه $(F(x), G(y))$، در واحد مربع $[۰, ۱] \times [۰, ۱]$ میل می‌کند و این جفت مرتب شده، به ترتیب متعلق به یک مقدار از $H(x, y)$ در $[۰, ۱]$ می‌باشند. ما نشان خواهیم داد این مطابقت که مقدار تابع توزیع توام را به هر جفت مرتب شده از مقادیر توابع توزیع انفرادی نسبت می‌دهد، یک تابع است که چنین توابعی مفصل نامیده می‌شوند

تعریف ۲-۳. تابع چگالی مربوط به مفصل C بصورت زیر محاسبه می‌شود:

$$c(u, v) = \frac{\partial^۲ C(u, v)}{\partial u \partial v}$$

۲-۳ قضیه اسکلار

قضیه‌ای که در این بخش مورد بررسی قرار گرفته است اولین بار توسط اسکلار در سال ۱۹۵۹ بیان شده‌است و به این سوال پاسخ می‌دهد که چرا مفصل‌ها در مدل‌سازی‌ها مشهور می‌باشند. این قضیه بیان می‌کند که هر تابع توزیع توأم با حاشیه‌ای‌های پیوسته می‌تواند به طور یکتا توسط یک مفصل از توابع حاشیه‌ای نوشته شود. در این قسمت ابتدا به یادآوری تعاریفی اولیه که برای درک این قضیه ضروری است می‌پردازیم و در نهایت این قضیه را بر حسب زیرمفصل و سپس بر اساس مفصل بیان می‌کنیم.

تعریف ۲-۴. یک فضای احتمال (Ω, F, P) را در نظر بگیرید که در آن Ω فضای نمونه، $\mathbf{P}$ یک اندازه است به قسمی که $\mathbf{P}(\Omega) = ۱$ ، F یک سیگما ـ جبر است؛ یک متغیر تصادفی X بصورت زیر قابل تعریف است

$$X : \Omega \longrightarrow \mathbf{R}$$

به قسمی که X، F ـ اندازه‌پذیر است.

تعریف ۲-۵. فرض کنید X یک متغیر تصادفی باشد. تابع توزیع X (CDF) بصورت زیر تعریف می‌شود:

$$F : \mathbf{R} \longrightarrow [۰, ۱], F(x) = \mathrm{P}[X \leq x]$$

تعریف ۲-۶. اگر مشتق تابع توزیع تجمعی X وجود داشته باشد، آن تابع چگالی احتمال X (pdf) نامیده می‌شود.

تعریف ۲-۷. فرض کنید X و Y متغیرهای تصادفی باشند. تابع توزیع توأم این دو متغیر بصورت

$$H(x, y) = \mathrm{P}[X \leq x, Y \leq y]$$

می‌باشد. همچنین توابع حاشیه‌ای H، $F(x) = lim_{y\to\infty}H(x, y)$ و $G(y) = lim_{x\to\infty}H(x, y)$ می‌باشند.

لم ۲-۸. *فرض کنید H یک تابع توزیع توأم با حاشیه‌ای‌های F و G باشد، آنگاه وجود دارد یک زیرمفصل یکتا C' به قسمی که*

$$DomC' = RanF \times RanG, \qquad (۱-۲)$$

$$H(x, y) = C'\big(F(x), G(y)\big) \qquad \forall (x, y) \in \overline{R} \qquad (۲-۲)$$

اثبات. اگر C' یکتا باشد، بایستی هر $(u,v) \in Ranf \times RanG$ تنها یک تصویر ممکن از $C'(u,v)$ داشته باشد و بنابراین رابطه(۲-۲) برقرار است؛ اما بالعکس اگر فرض کنید که $C'_1(u,v) \neq C'_2(u,v)$ هر دو در رابطه (۲-۲) صدق کنند، به عنوان مثال وجود دارند $(x_1,y_1),(x_2,y_2) \in \overline{R}^2$ به قسمی که

$$C'_1(u,v) = C'_1\big(F(x_1),G(y_1)\big) = H(x_1,y_1)$$
$$C'_2(u,v) = C'_2\big(F(x_2),G(y_2)\big) = H(x_2,y_2)$$

بنابراین بایستی $u = F(x_1) = F(x_2)$ و $v = G(y_1) = G(y_2)$ برقرار باشد.
بنا به لم (۲-۱۶) یک تابع توزیع توأم H در شرط زیر صدق می‌کند:

$$|H(x_2,y_2) - H(x_1,y_1)| \leq |F(x_2) - F(x_1)| + |G(y_2) - G(y_1)| = 0$$

بنابراین C'_1 و C'_2 روی (u,v) یکسان می‌باشند.
حال فرض کنید C' تابعی یکتا و برای همه $(x,y) \in \overline{R}^2$نگاشتی از جفت‌های $\big(F(x),G(y)\big)$ روی $H(x,y)$ باشد؛ کافی است نشان دهیم C'، یک زیرمفصل دو بعدی است.

•

$$C'\big(0,G(y)\big) = C'\big(F(-\infty),G(y)\big) = H(-\infty,y) = 0$$
$$C'\big(F(x),0\big) = C'\big(F(x),G(-\infty)\big) = H(x,-\infty) = 0$$

- فرض کنید $u_1 \leq u_2$ در برد F و $v_1 \leq v_2$ در برد G باشند. چون توابع توزیع تجمعی نانزولی می‌باشند، وجود دارد $x_1 \leq x_2$ و $y_1 \leq y_2$ یکتا با $F(x_1) = u_1$، $F(x_2) = u_2$و $G(y_1) = v_1$، $G(y_2) = v_2$.

$$\text{C'}(u_2,v_2) - C'(u_1,v_2) - C'(u_2,v_1) + C'(u_1,v_1) =$$
$$C'\big(F(x_2),G(y_2)\big) - C'\big(F(x_1),G(y_2)\big) - C'\big(F(x_2),G(y_1)\big) + C'\big(F(x_1),G(y_1)\big) =$$
$$H(u_2,v_2) - H(u_1,v_2) - H(u_2,v_1) + H(u_1,v_1) \geq 0$$

نابرابری اخیر از خاصیت sigma-aditivity، P نتیجه می‌شود.

- همچنین توابع حاشیه‌ای نیز نگاشتی یکسان دارند

$$C'\big(1,G(y)\big) = C'\big(F(\infty),G(y)\big) = H(\infty,y) = G(y)$$
$$C'\big(F(x),1\big) = C'\big(1,G(\infty)\big) = H(x,\infty) = F(x)$$

□

⋆ *عکس لم نیز برقرار است؛ یعنی هر تابع تعریف شده H بصورت رابطه (۲-۲) یک تابع توزیع توأم است و این امر با استفاده از ویژگی‌های زیرمفصل نتیجه می‌شود.*

قضیه ۲-۹. (قضیه اسکلار) *فرض کنید H یک تابع توزیع توأم با حاشیه‌ای‌های F و G باشد؛ آنگاه وجود دارد یک مفصل دو بعدی یکتا C به قسمی که برای همه $(x,y) \in \overline{R}^2$*

$$H(x,y) = C\big(F(x), G(y)\big) \qquad (۳-۲)$$

اگر F و G پیوسته باشند آنگاه C یکتا خواهد بود؛ در غیراینصورت C بر روی $RanF \times RanG$ به طور یکتا تعیین می‌شود. عکس این قضیه نیز برقرار است؛ یعنی اگر F و G توابع توزیع و C یک مفصل باشد آنگاه H تعریف شده در رابطه (۳-۲) یک تابع توزیع با حاشیه‌ای‌های F و G می‌باشد.

اثبات. بنا به لم (۲-۸) زیر مفصل یکتا C' در رابطه (۳-۲) صدق می‌کند. اگر F و G پیوسته باشند، آنگاه $RanF \times RanG = I^2$ است، بنابراین $C = C'$ یک مفصل می‌باشد. در غیر اینصورت ثابت شده‌است که C' می‌تواند به C بسط داده شود.(Nelsen,1999) □

مثال ۲-۱۰. فرض کنید متغیرهای تصادفی X و Y با تکیه‌گاه I^2 دارای تابع توزیع توأم $H(x,y) = \frac{1}{2}(x+y)$، $(x,y) \in I^2$ می‌باشند. توزیع‌های حاشیه‌ای عبارتند از $F_X(x) = \frac{1}{2}(x+1)$ و $G_Y(y) = \frac{1}{2}(y+1)$. متغیرهای X و Y پیوسته نیستند، با این حساب بنا به قضیه اسکلار می‌توان بر روی ناحیه$[\frac{1}{2}, 1] \times [\frac{1}{2}, 1]$ مفصل منحصر به فرد C را بصورت زیر تعیین کرد:

$$H(x,y) = max\big(\frac{1}{2}(x+1) + \frac{1}{2}(y+1) - 1, 0\big) = C\big(F(x), G(y)\big).$$

(خارج از این فاصله، C یکتا نخواهد بود)

حال با استفاده از قضیه اسکلار ارتباط بین متغیرهای تصادفی و مفصل‌ها را بیان می‌کنیم.

قضیه ۲-۱۱. *پایایی مفصل‌ها تحت تبدیلات صعودی از متغیرهای تصادفی*

فرض کنید $X \sim F$ و $Y \sim G$ متغیرهای تصادفی با مفصل C باشند. اگر α و β به‌ترتیب توابعی صعودی روی برد X و برد Y باشند آنگاه $\alpha(X) \sim F_\alpha$ و $\beta(Y) \sim G_\beta$، مفصل $C_{\alpha\beta} = C$ را دارا می‌باشند.

اثبات. $C_{\alpha\beta}\big(F_\alpha(x), G_\beta(y)\big) =$

$P[\alpha(X) \le x, \beta(Y) \le y] = \mathrm{P}[X \le \alpha^{-1}(x), Y \le \beta^{-1}(y)] = C(F\big(\alpha^{-1}(x)\big), G\big(\beta^{-1}(y)\big)) =$

$$C(\mathrm{P}[X \le \alpha^{-1}(x)], \mathrm{P}[Y \le \beta^{-1}(y)]) = C(\mathrm{P}[\alpha(X) \le x], \mathrm{P}[\beta(Y) \le y]) =$$
$$C\big(F_\alpha(x), G_\beta(Y) \le y]\big)$$

□

فرض کنید $X \sim F$ و $Y \sim G$ متغیرهای تصادفی پیوسته با تابع توزیع توأم H باشند. X و Y مستقل از هم هستند اگر $H(x,y) = F(x).G(y)$ باشد. بر حسب تابع مفصل استقلال دو متغیر تصادفی بصورت زیر تعریف می‌شود:

قضیه ۲-۱۲. *متغیرهای تصادفی X و Y مستقل از هم هستند اگر مفصل آنها بصورت $C^{\perp}(u,v) = u.v$ باشد.* (*$C^{\perp}$ مفصل حاصلضرب نامیده می‌شود*)

۲-۴ خانواده‌های مفصل ارشمیدسی

یک کلاس مهم از مفصل‌ها، مشهور به مفصل ارشمیدسی [۲] است و نسبت به سایر مفصل‌ها پرکاربردتر هستند چرا که به راحتی ساخته می‌شوند، خانواده عظیمی از مفصل‌ها به این کلاس تعلق دارند و ساختار وابستگی‌شان تنها توسط یک تابع حقیقی مقدار یک متغیره مشخص می‌شود. این کلاس از مفصل‌ها نه تنها در آمار بلکه در فضاهای متری احتمال نیز ظاهر می‌شوند (Schweizer,1991).

تعریف ۲-۱۳. فرض کنید ϕ یک تابع پیوسته و اکیدا نزولی از I به $[0,\infty]$ باشد به قسمی که $\phi(0) = 1$. شبه ـ معکوس ϕ تابع $\phi^{[-1]}$ است با دامنه $[0,\infty]$ و برد I که بصورت زیر داده می‌شود:

$$\phi^{[-1]}(t) = \begin{cases} \phi^{-1}(t) & 0 \le t \le \phi(0), \\ 0 & \phi(0) \le t \le \infty. \end{cases}$$

تعریف ۲-۱۴. (مفصل ارشمیدسی) فرض کنید ϕ تابعی پیوسته، اکیدا نزولی از $I \to [0,\infty]$ باشد به طوری که $\phi(1) = 0$. مفصلی به فرم

$$C(u,v) = \phi^{[-1]}\big(\phi(u), \phi(v)\big) \qquad (۲-۴)$$

را مفصل ارشمیدسی و ϕ را تابع مولد مفصل ارشمیدسی گویند. Genest, Mackay(1986)

[۲] Archimedean Copula

سه زیر کلاس مهم از مفاصل ارشمیدسی عبارتند از:

- **مفصل کلیتون**: این مفصل توسط Clayton, (1978) معرفی و بعدها توسط Cook, Johnson(1981) و Oaks(1982) مورد مطالعه و بررسی قرار گرفت و در این خانواده $\alpha \in [-۱, \infty) \setminus \{۰\}$ و $\phi(t) = \frac{t^{-\alpha}-۱}{\alpha}$ و لذا

$$C_\alpha(u,v) = \left(u^{-\alpha} + v^{-\alpha} - ۱\right)^{-\frac{۱}{\alpha}}$$

- **مفصل فرانک:** این مفصل توسط Frank (1979) مطرح و ویژگی‌های آن توسط Nelson (1986) و Genest(1987) مورد ارزیابی قرار گرفته است. در این خانواده، $\alpha \in (-\infty, \infty) \setminus \{۰\}$ و $\phi(t) = -ln\frac{e^{-\alpha t}-۱}{e^{-\alpha}-۱}$ و لذا

$$C_\alpha(u,v) = \frac{-۱}{\alpha} ln\left(۱ + \frac{(e^{-\alpha u}-۱)(e^{-\alpha v}-۱)}{(e^{-\alpha}-۱)}\right)$$

- **مفصل گمبل:** این مفصل توسط Hougaard (1986) مورد برسی قرا گرفته است. در این خانواده $\varphi(t) = \big(-ln(t)\big)^{\alpha+۱}$ که در آن $\alpha \geq ۰$ می‌باشد. بنابراین

$$C_\alpha(u,v) = exp\left([(-lnu)^\alpha + (-lnv)^\alpha]^{\frac{۱}{\alpha}}\right)$$

۲-۴-۱ ویژگی‌های مفصل ارشمیدسی

توابع مفصل ارشمیدسی، توابعی به‌طور کامل یکنوا می‌باشند؛ بدین معنی که،

- این توابع مقیاس پایا می‌باشند؛ یعنی برای هر $c > ۰$، ϕ و $c\phi$ مفصل مشابه تولید می‌کنند.

$$\phi^{(-۱)} = \big(\phi(u) + \phi(v)\big) = (c\phi)^{(-۱)} = \big(c\phi(u) + c\phi(v)\big)$$

- تابع مولد ϕ همواره پیوسته و مشتق‌پذیر بوده و $۱ - \phi^{-۱}(u)$ تابع توزیعی تک‌مدی روی $[۰, \infty)$ می‌باشد.
- مفصل C ارشمیدسی است اگر و تنها اگر تابع f از $[۰, ۱]$ به $[۰, ۱]$ موجود باشد به قسمی که:

$$\frac{\partial C(u,v)/\partial u}{\partial C(u,v)/\partial v} = \frac{f(u)}{f(v)}$$

در این‌صورت تابع ϕ با ضابطه $\phi(t) = \int_t^۱ f(u)du$ محاسبه می‌شود.

توابع بقاء و توابع توزیع مفصل‌های متفاوتی را تولید می‌کنند؛ به جز خانواده مفصل فرانک که آن هم بدلیل شرط تقارن بخصوصی است که توسط Genest(1978) به ثبت رسیده است.

۲-۵ مفصل و تابع وابستگی

مفصل‌ها در آمار به دو دلیل عمده مورد توجه می‌باشند؛ اول به این دلیل که یک روش آزاد ـ مقیاس برای مطالعه وابستگی می‌باشند و دوم اینکه نقطه شروعی برای ساخت خانواده های دو یا چند متغیره از توزیع‌ها هستند. بنا به نظریه Schweizer:Wolff (۱۹۸۱) یک اندازه مناسب از وابستگی $\kappa(X,Y)$ بایستی تابعی از مفصل باشد؛ چنین اندازه‌هایی همواره وجود دارند، چرا که درون یک مجموعه‌ی کراندار قرار میگیرند که توسط فرشه ـ هافدینگ [۳] معرفی شده است و معروف به نابرابری F-H است که قابل تعریف بصورت زیر می باشد:

۲-۵-۱ حدود فرشه ـ هافدینگ

برای هر مفصل C

$$W(u,v) = max(۰, u+v-۱) \leq C(u,v) \leq min(u,v) = M(u,v) \qquad (۵-۲)$$

(این رابطه برای همه $(u,v)\epsilon[0,1]$ به طور نقطه‌وار برقرار است)
همچنین بنا به نتیجه قضیه اسکلار Sklar:1959، اگر X و Y متغیرهای تصادفی با با تابع توزیع توام H و توابع توزیع حاشیه‌ای F و G باشند، آنگاه برای همه X و Y ها،

$$max(۰, F(x)+G(y)-۱) \leq H(x,y) \leq min(F(x), G(y))$$

حدود W و M در رابطه (۵-۲)، حدود Frechet-Hoeffding برای تابع توزیع توام H با توابع حاشیه‌ای F و G می‌باشد.

۲-۵-۲ اندازه‌هایی از وابستگی بر اساس مفصل

بنا به اندازه همبستگی پیرسون، واضح است که هر اندازه‌ای از همبستگی به صورت $\kappa(X,Y) = \kappa(C)$ $\kappa(H)$ می‌باشد که،

- در حالت استقلال، $\kappa(\sqcap) = ۰$

[۳] Hoeffding and Frechet

- در مواردی که همبستگی اساسی و یکنوا برقرار است، مقادیر ± ۱ را به خود اختصاص می‌دهد.

بنابراین همبستگی را بصورت زیر تعریف می‌کنیم:

$$\rho(X,Y) = Corr(U,V) \qquad (۶-۲)$$

که در آن، $V = G(Y)$ و $U = F(X)$ می‌باشند. به عبارت دیگر، ρ برابر است با ضریب همبستگی پیرسون بین $G(Y)$و $F(X)$که؛ البته این ایده اولین بار توسط پیرسون در سال ۱۹۰۴ مطرح شده است.

بنا به تحقیقات Trivedi and Zimmer(2005) ضرایب همبستگی خطی نمی‌توانند میزان همبستگی توابع غیر خطی از متغیرهای تصافی مانند داده‌های بقاء را اندازه بگیرند و این جا است که نقش همچنین دو اندازه پایا از وابستگی به نام τ کندال و ρ اسپیرمن، که در قالب مفهوم هم‌شیبی تعریف می‌شوند پررنگ‌تر می‌گردد.

در این قسمت نشان خواهیم داد که می‌توان این ضرایب همبستگی رتبه‌ای، را بر اساس مفصل نوشت.

۲-۵-۲-۱ ρ اسپیرمن

چون $V = G(Y)$و$U = F(X)$ متغیرهای تصادفی یکنواخت می‌باشند و $E(U) = E(V) = \frac{۱}{۲}$ و $Var(U) = Var(V) = \frac{۱}{۱۲}$ بنابراین:

$$\rho(X,Y) = Corr(U,V) = \frac{E(UV) - \frac{۱}{۲} \times \frac{۱}{۲}}{\sqrt{\frac{۱}{۱۲} \times \frac{۱}{۱۲}}}$$

چون تابع توزیع توام (U,V) برابر است با $Pr(U,V) = C(u,v)$نتیجه می‌شود که:

$$\rho(X,Y) = ۱۲E(UV) - \frac{۱۲}{۴} = -۳ + ۱۲ \int\int uvdC(u,v)$$

بنابراین:

$$\rho_C = -۳ + ۱۲E_C(UV) \qquad (۷-۲)$$

⋆وقتی C=M، حد بالایی F-H باشد، آنگاه $U = V$ و داریم: $E_C(UV) = E(U^۲) = \frac{۱}{۳}$ ونتیجه می‌شود که $\rho_M = -۳ + ۴ = ۱$.

⋆ اما اگر $C = W$، حد پایینی F-H باشد، آنگاه $E_C(UV) = E(U - U^۲) = \frac{۱}{۲} - \frac{۱}{۳} = \frac{۱}{۶}$ و $V = ۱ - U$ در نتیجه $\rho_W = ۲ - ۳ = -۱$.

۲-۵-۲-۲ کندال τ

توءکندال اولین بار توسط Kendall:1938 معرفی شده است که بصورت زیر تعریف می‌شود:

$$\tau(X,Y) = Pr\{(X_۱ - X_۲)(Y_۱ - Y_۲) > ۰\} - Pr\{(X_۱ - X_۲)(Y_۱ - Y_۲) < ۰\} \quad (۸-۲)$$

در حالیکه $(X_۱, Y_۱) \sim C(F,G)$ و $(X_۲, Y_۲) \sim C(F,G)$ دو مشاهده مستقل از $H = C(F,G)$ می‌باشند.

با توجه به اینکه، $Pr(concordance) + Pr(discordance) = ۱$ داریم:

$$\Pr(\text{concordance}) = \Pr(X_۱ - X_۲)(Y_۱ - Y_۲) > ۰ = ۲Pr(X_۱ < X_۲, Y_۱ < Y_۲)$$

بنابراین $\tau(X,Y)$ را بصورت زیر می‌توان محاسبه کرد،

$$\tau(X,Y) = -۱ + ۴Pr(X_۱ < X_۲, Y_۱ < Y_۲) = -۱ + ۴\int\int H(X,Y)dH(X,Y)$$

و بر اساس $H(X,Y) = C\{F(X), G(Y)\}$ می‌توان نوشت:

$$\tau(X,Y) = -۱+۴\int\int C(F(X),G(Y))dH\{F(X),G(Y)\} = -۱+۴\int\int C(u,v)dC(u,v)$$

و در نهایت می‌توان آن را بصورت زیر خلاصه کرد:

$$\tau_C = -۱ + ۴E_C\{C(U,V)\} \quad (۹-۲)$$

⋆وقتی C=M، حد بالایی F-H باشد، آنگاه $V = U$ و داریم: $E_C\{C(U,V)\} = E\{min(U,V)\} = \frac{۱}{۲}$ ونتیجه می‌شود که $\tau_M = \frac{۴}{۲} - ۱ = ۱$.

اما اگر $C = W$، حد پایینی F-H باشد، آنگاه $V = ۱ - U$ و $E_C\{C(U,V)\} = E\{max(۰, U+V-۱)\}$ ، در نتیجه $\tau_W = ۰ - ۱ = -۱$.

$\rho \in [-۱, ۱]$ و τ می‌باشد، در حالیکه ۱ نشاندهنده‌ی هم‌شیبی کامل، -۱ نشاندهنده‌ی ناهمشیبی کامل و ۰ هم‌شیبی صفر را بیان می‌کند. بنابراین با توجه به رابطه اساسی ρ و τ با تابع مفصل، به راحتی می‌توان با داشتن مقادیر ضریب همبستگی اسپیرمن و توءکندال ، پارامتر مفصل را برآورد کرد و بالعکس.

جدول زیر نشاندهنده مقادیر τ کندال، بر اساس پارامتر توابع مفصل ارشمیدسی مذکور می‌باشد:

Copula	Clayton	Frank	Gumbel
τ	$\frac{\alpha}{(\alpha+۲)}$	$۱+\frac{۴}{\alpha}\{D_۱(\alpha)-۱\}$	$\frac{\alpha}{(\alpha+۱)}$

(که در آن ، $D_۱(\alpha)=\int_۰^\alpha \{t/\alpha(e^t-۱)\}dt$ می‌باشد.)