[1]. Вероятностные характеристики времени жизни: функция выживания, кривая смертей, интенсивность смертности, числовые характеристики.

Функция выживания $s(t) = P(T \geq t)$ – вероятность того, что человек проживет не меньше t лет.

Характеристики:

- $s(0) = 1$
- $s(\infty) = 0$
- Убывает
- Непрерывна

На практике часто работают не с s(t), а с $l_x = E[L(x)] = l_0 \cdot s(x)$, где l_0- размер выборки, $L(x)$ – число живых представителей группы в x лет

Кривая смертей – или график плотности продолжительности жизни $f(x) = -s'(x)$. Также рассматривают график $l_0 \cdot f(x)$. Функция выживания может быть восстановлена по плотности как $s(x) = \int_x^\infty f(u)du$

Интенсивность смертности – рассчитывается по формуле $\mu_x = \frac{f(x)}{s(x)} = \frac{d_x}{l_x}$, где d_x- число умерших в x лет, l_x – число доживших до x лет.

Числовые характеристики:

- Интенсивность
- Среднее время жизни = $\int_0^\infty x \cdot f(x)dx = \int_0^\infty s(x)dx$
- Дисперсия времени жизни = $\int_0^\infty x^2 \cdot f(x)dx = 2 \cdot \int_0^\infty x \cdot s(x)dx$
- Медиана – s(x) = 0.5
- Мода – максимум плотности

[3]. Аналитические законы смертности: модели Муавра и Вейбулла.

В модели *Муавра* время жизни распределено равномерно, т.е $f(x) = \frac{1}{w}$, где w – предельный возраст жизни. Тогда F(x) = $\frac{x}{w}$, s(x) = 1 - $\frac{x}{w}$ $\mu_x = \frac{1}{w-x}$. Эта модель плоха тем, что не соответствует реальным наблюдениям.

Модель *Вейбулла* учитывает влияние возраста на смертности и определяется как:

s(x) = $e^{\frac{-k \cdot x^{n+1}}{n+1}}$

f(x) = $k \cdot x^n \cdot$ s(x)

$\mu_x = k \cdot x^n$

$k > 0$

$n > 0$

Параметр k отвечает за форму кривой. Если k < 0, то кривая вогнута, а значит смертность уменьшается со временем. Если k = 1, то интенсивность смертности не меняется со временем. Если k > 1, то интенсивность смертности растет со временем.

Параметр n отвечает за скорость роста интенсивности смертности со временем.

[5]. Два основных способа построения вероятностных моделей. Пояснить на любом примере.

Первый способ основан на логическом анализе изучаемого явления. Объясним на примере вывода распределения времени жизни без учёта «прожитых лет» (далее копия 4 вопроса).

T – время жизни

A_T – человек проживет T лет

$G(T) = P(A_T)$ – человек живет в течение T лет

$G(T)$ строго убывает, непрерывна и строго монотонна

|----------|----------|

T S

$A_{T+S} = A_T \cap A_S$ (начало промежутков разное)

$P(A_{T+S}) = P(A_T \cap A_S) = P(A_T) * P(A_S|A_T)$

Если не учитываем эффект усталости, $P(A_S|A_T) = P(A_S)$

$P(A_T \cap A_S) = P(A_T) * P(A_S)$

$G(T+S) = G(T) * G(S), \quad \forall\, T, S$

$G(2T) = G(T)^2$

$G(nT) = G(T)^n$

Пусть $nt=u$, тогда $t=u/n$ и $G(u) = G^n(\frac{u}{n})$ или $G^{1/n}(u) = G(\frac{u}{n})$ для ∀u

Пусть $u=mt$, тогда $G(\frac{m}{n}t) = G^{\frac{m}{n}}(t)$, поэтому $G(rt) = G^r(t)$ выполняется для $\forall\ r \in Q$ (рациональных чисел)

$G(xt) = G(t)^x$

$t = 1 \ \rightarrow \ G(x) = G(1)^x$

$G(1) = a \rightarrow G(x) = a^x = \exp(\ln(a) * x), \quad ln(a) < 0 \ \rightarrow ln(a) = -\lambda, \lambda > 0$

$G(x) = \begin{cases} \exp(-\lambda x), & x \geq 0 \\ 0, & x < 0 \end{cases}$ – функция выживания

$S(x) = \begin{cases} \exp\left(-\frac{kx^{n+1}}{n+1}\right), & x > 0 \\ 0, & x \leq 0 \end{cases}$ – кривая смертей

$$f(x) = \begin{cases} kx^n exp\left(-\dfrac{kx^{n+1}}{n+1}\right) \\ 0 \end{cases}$$

$\mu_x = kx^n$ – интенсивность смерти

[6]. Тест суммы рангов Вилкоксона, распределение, пример применения.

Тест суммы рангов Вилкоксона – статистический тест, используемый для проверки гипотезы о равенстве распределений двух независимых случайных величин.

Пример: проводится тестирование нового лекарства. Требуется определить, имеет ли это лекарство какое-либо действие. Из N пациентов случайным образом выбираются $n \leq N$ пациентов, которым будут давать новое лекарство. Остальным $m = N - n$ пациентам даётся плацебо. В качестве случайной величины выбирается какой-либо количественный показатель "здоровья" пациентов, например, температура тела.

Алгоритм:

- Определяем нулевую гипотезу: эффекта от лекарства нет, выборки однородны. Альтернативная гипотеза – эффект есть. Более формально: для показателей здоровья $x_i, i \in \overline{1, N}$ и действий лекарства и плацебо $F_x(x), F_y(x)$, $H_0: F_x(x) = F_y(x)\ \forall x \in \overline{1, N}$
- Из N пациентов случайным образом выбираем n, которым даётся новое лекарство.
- По прошествии выбранного создателями лекарства времени производим измерение показателей здоровья пациентов x_i.
- Отсортируем x_i в порядке возрастания (считаем, что все x_i разные; если это не так, можно, например, искусственно задать порядок для одинаковых x_i). Таким образом получаем $x_{(1)} < x_{(2)} < \cdots < x_{(N)}$
- Каждому $x_{(i)}$ поставим в соответствие ранг X_i, равный порядковому номеру x_i в отсортированном массиве. Пациенту с самым низким показателем здоровья ставим ранг 1, пациенту с самым высоким – ранг N.
- Пусть $S_1, \dots, S_n$ – ранги пациентов, которым давали лекарство, $R_1, \dots, R_m$ – ранги тех, кому лекарство не давали. Тогда вероятность того, что $S_1, \dots, S_n$ примут соответственно произвольные значения $X_1, \dots, X_n$ при условии истинности нулевой гипотезы (лекарство не действует) $P_{H_0}(S_1 = X_1, \dots, S_n = X_n) = \frac{1}{C_N^n}$, где C_N^n – число способов выбрать n пациентов из N.
- Зададим статистику $w_S = \sum_{i=1}^{n} S_i$
- Тогда решающее правило будет иметь вид: $\delta = 1$, если $w_S \geq C$, и $\delta = 0$ иначе, где C – некоторое пороговое значение, выбираемое из правила:$P_{H_0}(w_S \geq C) \leq \alpha$ (α выбираем исходя из потребностей эксперимента). $\delta = 0$ – нулевая гипотеза принимается, $\delta = 1$ – гипотеза не принимается.

Этот тест часто выбирается потому, что:

- Соответствует "консервативной" точке зрения – не принимаем нулевую гипотезу только если сильно уверены в её ошибочности.
- Если нулевая гипотеза верна – мы знаем распределение.

[4]. Вывод распределения времени жизни без учета «прожитых лет».

T – время жизни

A_T – человек проживет T лет

$G(T) = P(A_T)$ – человек живет в течение T лет

$G(T)$ строго убывает, непрерывна и строго монотонна

|----------|----------|

T S

$A_{T+S} = A_T \cap A_S$ (начало промежутков разное)

$P(A_{T+S}) = P(A_T \cap A_S) = P(A_T) * P(A_S|A_T)$

Если не учитываем эффект усталости, $P(A_S|A_T) = P(A_S)$

$P(A_T \cap A_S) = P(A_T) * P(A_S)$

$G(T+S) = G(T) * G(S), \quad \forall\, T, S$

$G(2T) = G(T)^2$

$G(nT) = G(T)^n$

Пусть *nt=u,* тогда *t=u/n* и $G(u) = G^n(\frac{u}{n})$ или $G^{1/n}(u) = G(\frac{u}{n})$ для ∀u

Пусть *u=mt*, тогда $G(\frac{m}{n}t) = G^{\frac{m}{n}}(t)$, поэтому $G(rt) = G^r(t)$ выполняется для $\forall\ r \in Q$ (рациональных чисел)

$G(xt) = G(t)^x$

$t = 1 \ \rightarrow\ G(x) = G(1)^x$

$G(1) = a \rightarrow G(x) = a^x = \exp(\ln(a) * x), \quad ln(a) < 0 \ \rightarrow ln(a) = -\lambda, \lambda > 0$

$G(x) = \begin{cases} \exp(-\lambda x), & x \geq 0 \\ 0, & x < 0 \end{cases}$ – функция выживания

$S(x) = \begin{cases} \exp\left(-\frac{kx^{n+1}}{n+1}\right), & x > 0 \\ 0, & x \leq 0 \end{cases}$ – кривая смертей

$$f(x) = \begin{cases} kx^n exp\left(-\dfrac{kx^{n+1}}{n+1}\right) \\ 0 \end{cases}$$

$\mu_x = kx^n$ – интенсивность смерти

[2]. Вероятностные характеристики остаточного времени жизни: функция выживания, кривая смертей, интенсивность смертности.

Остаточное время жизни – сколько человек еще проживет лет, учитывая, что он уже дожил до x лет. То есть $T_x = T - x$

Функция выживания – вероятность того, что человек проживет еще не меньше t лет.

$$s_x(t) = P(T_x \geq t) = 1 - F_x(t) = \frac{s(x+t)}{s(x)}$$

$$F_x(t) = P(T_x < t) = P(T - x < t \mid T > x) = \frac{P(x < T < x+t)}{P(T > x)} = \frac{F(x+t) - F(x)}{1 - F(x)}$$

$$= \frac{s(x) - s(x+t)}{s(x)} = \frac{l_x - l_{x+t}}{l_x}$$

Кривая смертей:

$$f_x(t) = \frac{f(x+t)}{s(x)}$$

Интенсивность смертности – рассчитывается по формуле $\mu_x(t) = \frac{f_x(t)}{s_x(t)} = \frac{\frac{f(x+t)}{s(x)}}{\frac{s(x+t)}{s(x)}} = \frac{f(x+t)}{s(x+t)} =$

$\mu_{x+t} = \frac{d_{x+t}}{l_{x+t}}$

[7]. Статистика Манна-Уитни. Математическое ожидание, дисперсия и распределение при гипотезе однородности.

Имеем:

$Y_1, \dots, Y_n -$ характеристики объектов, которых обработали новым способом

$X_1, \dots, X_m -$ характеристики объектов, которых не обрабатывали

Статистика Манна-Уитни:

$$W_{XY} = \sum_{i=1}^{n} S_i - \frac{n(n+1)}{2}$$

$$W_{YX} = \sum_{j=1}^{m} R_i - \frac{m(m+1)}{2}$$

где S_i и R_i – ранги для объектов, которые подверглись и не подверглись обработке, соответственно.

В отличие от простой суммы рангов, такая статистика имеет несколько преимуществ:

1. Распределение статистики (при гипотезе однородности) при $n = k_1, m = k_2$ и $n = k_2, m = k_1$ будет совпадать:

Просто сумма рангов (при гипотезе однородности) имеет следующее распределение:

$$P_H(W_S = w) = \frac{\#(w; n, m)}{C_N^n}$$

где $\#(w; n, m)$ – количество различных комбинаций n рангов, образующих сумму w.

Получается, что, если поменять группы местами, то нужно заново находить распределение.

В случае статистики Манна-Уитни при любых $k_1 \le k_2$ распределение статистики будет одинаковым, если $n = k_1, m = k_2$ и в обратном случае. Поэтому можно составить 1 таблицу распределения статистики для пары m и n.

Пример:

Сумма рангов ($n = 2, m = 3$)	3	4	5	6	7	8	9
Сумма рангов ($n = 3, m = 2$)	6	7	8	9	10	11	12
Статистика Манна-Уитни	0	1	2	3	4	5	6
Количество комбинаций	1	1	2	2	2	1	1

2. Существует альтернативный способ подсчета:

$S_1 -$ перед объектом $(S_1 - 1)$ необработанных объекта

$S_2 -$ перед объектом $(S_2 - 2)$ необработанных объекта

...

$$W_{XY} = \sum_{i=1}^{n} S_i - \frac{n(n+1)}{2}$$

Значит можно считать статистику сравнивая объекты между собой:

Второй способ основан на подборе кривой плотности, более менее удачно описывающей реальные данные (можем брать любую с интегралом плотности = 1). Например распределение Вейбулла при а ≠1 может использоваться как математическая модель распределения времени жизни с учётом прожитых лет. Функция распределения и плотность распределения Вейбулла:

$$F(x) = \begin{cases} 1 - \exp\left(-\frac{x^a}{\sigma}\right), & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$f(x) = \begin{cases} \mathrm{a}\frac{x^{a-1}}{\sigma} \exp\left(-\frac{x^a}{\sigma}\right), & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$W_{XY} = \sum_{i=1}^{n}\sum_{j=1}^{m} I(Y_i > X_j)$$

где I – индикатор.

Распределение статистики W_{XY} при гипотезе однородности:

$$P_H(W_{XY} = w) = \frac{\#(w; n, m)}{C_N^n}$$

где $\#(w; n, m)$ – число различных комбинаций n чисел из отрезка $[1 - \frac{n(n+1)}{2}, n - \frac{n(n+1)}{2}]$, образующих сумму w.

При n и m больших 10 распределение стримится к нормальному (по Предельной Центральной Теореме).

Вероятность индикатора:

- $Y_i X_j$ из равновероятностного распределения на отрезке $[1, N]$
- Всего комбинаций: $n(n-1)$
- Комбинаций, в которых $Y_i > X_j$: $\sum_{i=1}^{n}(i-1) = \frac{(n-1)n}{2}$

$$P_H\big(Y_i > X_j\big) = \frac{1}{2}$$

Математическое ожидание W_{XY} (при гипотезе однородности):

$$E_H(W_{XY}) = \sum_{i=1}^{n}\sum_{j=1}^{m} E\left(I\big(Y_i > X_j\big)\right) = \frac{mn}{2}$$

Дисперсия индикатора:

$$D_H\left(I\big(Y_i > X_j\big)\right) = E_H\big(I^2(Y_i > X_j)\big) - \left[E_H\big(I(Y_i > X_j)\big)\right]^2 \ = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

Дисперсия W_{XY} (при гипотезе однородности). Для подсчета нужно рассмотреть разные ситуации:

$$D_H(W_{XY}) = E_H\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{m} I\big(Y_i > X_j\big)I(Y_k > X_l)\right) - \left[E_H\left(I\big(Y_i > X_j\big)\right)\right]^2 = \cdots$$
$$= \frac{1}{12}mn(N+1)$$

[9]. Знаково-ранговый тест Вилкоксона, распределение, пример применения.

Знаково-ранговый тест Вилкоксона – это тест для набора парных наблюдений, который учитывает знак и значение разности между характеристиками объектов в паре.

Алгоритм подсчета:

1. Для каждой пары наблюдений находим разницу между характеристиками d_i
2. Ранжируем модули разниц $|d_i|$ в порядке возрастания
3. Каждому рангу r_i приписываем знак (+ или -) соответствующей разницы
4. Считаем статистику V_S: сумму положительных рангов

Тест:

Имеем пары наблюдений $(X_1, Y_1), \dots, (X_N, Y_N)$

H: обработка не имеет влияния, распределения X и Y совпадают

K: обработка имеет влияние, после нее характеристики больше

$$\varphi(x) = \begin{cases} 1, & V_s \geq c \\ 0, & \text{иначе} \end{cases}$$

$$P_H(V_s \geq c) \leq \alpha$$

При гипотезе однородности знак каждого ранга равен + или – равновероятно (с вероятностью $\frac{1}{2}$) и все N знаков независимы. Тогда всего имеется 2^N возможных комбинаций знаков.

Вероятность конкретной комбинации при гипотезе однородности:

$$P_H(N_+ = n, S_1 = s_1, \dots, S_n = s_n) = \frac{1}{2^N}$$

где N_+ - количество положительных рангов.

Так как все комбинации равновероятны, то распределение статистики V_s при гипотезе однородности:

$$P_H(V_s = v) = \frac{\#(v; N)}{2^N}$$

где $\#(v, N)$ – количество комбинаций, сумма рангов которых равна v.

Аналогично можно ввести статистику V_r: сумма рангов со знаком – (сумма самих рангов, т.е. натуральных чисел), т.к. сумма этих статистик:

$$V_s + V_r = \frac{N(N+1)}{2}$$

Решающее правило для такой статистики:

$$\varphi(x) = \begin{cases} 1, & V_r \leq c \\ 0, & \text{иначе} \end{cases}$$

$$P_H(V_r \leq c) \leq \alpha$$

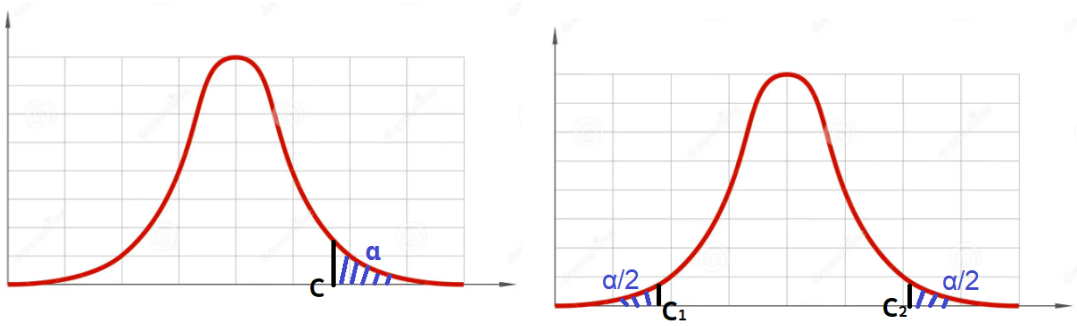
[10]. Какие выводы считаются статистически значимыми при проверке гипотез. Какие выводы можно сделать об альтернативе при не отклонении проверяемой гипотезы? Пояснить на любом примере.

Не нашел толковой инфы в записях. Материалов по этому вопросу в документе нет. Так что мог написать не то что нужно.

Статистически значимым является только отклонение гипотезы: если гипотеза не отвергается, то и близкие ей гипотезы также не отвергаются.

Не отклонении основной гипотезы означает непопадание статистики в критическую область, которая задается альтернативной гипотезой.

Критической областью называется область значений статистики критерия, при которых отвергается нулевая гипотеза. А критические значения — это границы критической области.



Критическая область бывает 3 видов: правосторонняя (слева), двусторонняя (справа) и левосторонняя. Вид задается альтернативной гипотезой.

Правосторонний случай: пусть H_0: эффект от нового лекарства отсутствует, H_1 – новое лекарство лучше. Функция выбора гипотезы $\varphi(x) = \begin{cases} 1, T(x) > C \\ 0, T(x) \leq C \end{cases}$, где T – статистика, C – критическое значение, находится из $P_H(T(x) > C) = \alpha, \alpha -$ уровень значимости (0.05)

Двусторонний случай: пусть H_0: эффект от нового лекарства отсутствует, H_1 – эффект от нового лекарства есть. Функция выбора гипотезы $\varphi(x) = \begin{cases} 1, \ T(x) < c_1 \text{ или } T(x) > c_2 \\ 0, \ c_1 < T(x) < c_2 \end{cases}$, где T – статистика, C_1 и C_2 – критические значения, находится из $P_H(T(x) < c_1 \text{ или } T(x) > c_2) = \alpha, \alpha -$ уровень значимости (0.05).

Двусторонняя альтернатива отвергается реже, так что получаем меньше значимых результатов.

[12]. Понятие условной независимости. Свойства и связь с маргинальной независимостью.

Частное (маргинальное) распределение:

$F_{X1,..Xn}(x_1,..x_n) = P(X_1 < x_1, X_2 < +\infty, \dots, X_n < +\infty)$

$f_{X1}(x_1) = \iint_{-\infty}^{+\infty} f_{X1,..Xn}(x_1, ..x_n)dx_2 \dots dx_n$

– чтобы получить плотность для 1 с.в., имея совместную плотность нескольких с.в., нужно интегрировать совместную плотность по всем переменным, кроме искомой (X_1).

Маргинальная независимость:

X, Y, Z – случайные величины

$f_{XY}(x, y) = f_X(x) * f_Y(y)$

$X_1, .., X_n$

$Y_1, .., Y_n$

$f_{X1,..Xn,\ Y1,..Yn} = f_{X1,..Xn} * f_{Y1,..Yn}$, если вектора независимые, т.е. $\forall$ случайная величина из 1 вектора независима со всеми случайными величинами из 2 вектора и наоборот.

Условное распределение:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Условная независимость:

$P(A \cap B|C) = P(A|C) * P(B|C)$

$f_{X,Y|Z}(x, y|z) = f_{X|Z}(x|z) * f_{Y|Z}(y|z)$

Из условной независимости не следует маргинальная независимость и наоборот!

Свойства:

1. Симметрия: $X \perp\!\!\!\perp Y \Rightarrow Y \perp\!\!\!\perp X$

2. Декомпозиция: $X \perp\!\!\!\perp (Y, Z) \Rightarrow X \perp\!\!\!\perp Y,\ X \perp\!\!\!\perp Z$ (док-во: третья переменная исчезает в процессе интегрирования)

3. Слабое объединение: $X \perp\!\!\!\perp (Y, Z) \Rightarrow X \perp\!\!\!\perp Y|Z, X \perp\!\!\!\perp Z|Y$

4. Условная независимость матожидания: $X \perp\!\!\!\perp Y|Z \Leftrightarrow E[X|Y, Z], E[X|Z]$

[13]. Графические модели. Примеры.

Графическая модель в статистике - это визуальная диаграмма, на которой наблюдаемые переменные идентифицируются точками (вершинами или узлами), соединенными ребрами, и соответствующим семейством распределений вероятностей, удовлетворяющих некоторым независимостям, заданным графом. Ребра могут быть не ориентированными

Данные статистики эквивалентны. На практике бывает легче считать какую-то статистику, т.к. она состоит из меньшего числа слагаемых.

Пример использования:

Для проверки нового удобрения 3 поля с клубникой поделили пополам и случайно распределили удобрение (новое и старое) для каждой половины. Взвесив урожай с каждой половины поля, с помощью знаково-рангового теста Вилкоксона можно понять имеет ли новое удобрение лучший эффект.

[8]. Тест суммы рангов Вилкоксона. Равные и неравные наблюдения.

При использовании теста суммы рангов Вилкоксона необходимо различать различные ситуации:

1. Неравные наблюдения

Все наблюдения различны, поэтому вероятность появления конкретных рангов у объектов, которые подверглись обработке, при гипотезе однородности:

$$P_H(S_1 = s_1, \dots, S_n = s_n) = \frac{1}{C_N^n}$$

В данном случае можно использовать заранее построенные таблицы распределения статистики.

2. Равные наблюдения

При ситуации, когда характеристики объектов могут повторяться есть 2 метода использования теста суммы рангов Вилкоксона:

- Искусственный порядок

Искусственно определяем порядок и используем метод с неравными наблюдениями.

- Мидранги

Для объектов с равными характеристиками поставим мидранг, равный среднему рангов, которые относятся к данным объектам.

В данном случае комбинации рангов перестают быть равновероятными, поэтому для каждого случая придется считать распределение.

Пример использования мидрангов:

$$n = m = 2$$

Наблюдения: $a = \{1.3, 1.7, 1.7, 2.5\}$

Ранги: $r = \{1, 2.5, 2.5, 4\}$

Вероятности комбинаций рангов объектов, которые подверглись обработке, при гипотезе однородности:

$$P_H(S_1^* = 1, S_2^* = 2.5) = \frac{2}{6}$$

$$P_H(S_1^* = 1, S_2^* = 4) = \frac{1}{6}$$

$$P_H(S_1^* = 2.5, S_2^* = 2.5) = \frac{1}{6}$$

$$P_H(S_1^* = 2.5, S_2^* = 4) = \frac{2}{6}$$

Получившееся распределение статистики:

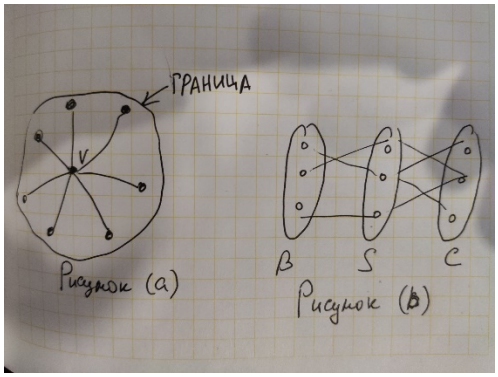
или ориентированными. Неориентированные ребра связаны с симметричной зависимостью и независимостью, в то время как ориентированные ребра могут отражать возможное направление действия или последовательность во времени. Эта независимость может исходить из априорного знания предмета или могут быть получены на основе тех или иных данных. Преимущества графического отображения включают простоту понимания, особенно сложных паттернов, простоту получения экспертного заключения и простоту сравнения вероятностей.

Ребра в таких моделях представляют наличие условной связи (зависимости) между переменными. Отсутствие же ребра показывает условную независимость 2 переменных.

Пусть $G \equiv (V, E)$ – граф, где V – множество вершин (наблюдаемых переменных, принадлежащих некому распределению), а E – ребра (наличие связи между переменными).

Марковские свойства:

- Распределение вероятностей на графе является попарно марковским относительно G, если для каждой пары вершин (u, v), которые не являются смежными, X_uи X_v условно независимы при всех других переменных в графе: $X_u \perp\!\!\!\perp X_v | X_{V\setminus(u,v)}$, где X_V-множестве переменных, соответствующих вершинам графа, X_v и X_u – переменные, соответствующие вершинам u и v.
- Распределение на графе является локально марковским, если для каждой вершины v переменная X_v не зависит от переменных, не входящих в $cl(v)$, обусловленных границей v: $X_u \perp\!\!\!\perp X_{V\setminus cl(v)} | X_{bd(v)}$, где $bd(A)$ - граница множества A, состоящая из вершин, не входящих в A, которые смежны с A; $cl(A) = A \cup bd(A)$. (Другими словами – v условно независимо со всеми вершинами вне ее границы/окрестности). Рисунок (a).
- Распределение на графе является глобально марковским, если для каждого кортежа непересекающихся множеств B, C и S, таких, что S разделяет B и C, векторные переменные X_B и X_C независимы от X_S. (Другими словами – если ∀ путь из B в C проходит через S, то B условно независимо от C при условии S). Рисунок (b)



Теорема Клифорда: $f(x_1,\ x_2, \dots, x_n) = \frac{1}{c}\prod_{cl_i \epsilon G}\ h_{cl_i}(X_{i1}, X_{i2})$, где $cl_i -$ клика i. Справедливо тогда и только тогда, когда есть условная независимость (когда выполняется хотя бы одно из марковских свойств (приведены выше)).

Примеры:

$$X \perp\!\!\!\perp Y|Z \ <=>\ f_{XY|Z}(x, y|z) = f_{X|Z}(x|z) * f_{Y|Z}(y|z)$$

[11]. Моделирование случайных величин.

Основным приемом моделирования случайных величин является метод обратных функций, который заключается в том, что на основании условия существования обратной функции и приведённой ниже теоремы, мы получаем конкретные формулы моделирования случайно величины $\mathscr{E}$.

Теорема. Случайная величина $\mathscr{E} \sim F(\gamma)$ имеет функцию распределения F(x), где ɣ - БСВ (базовая случай.

Алгоритм построения формулы моделирования случайной величины x на основе теоремы выглядит следующим образом:

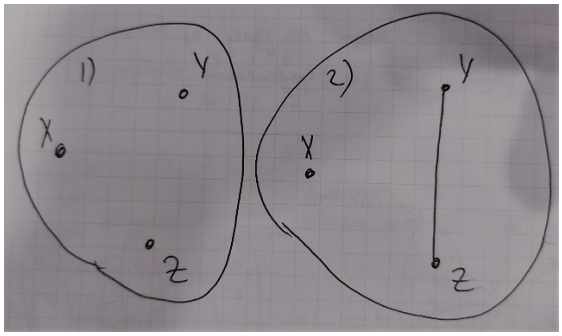
1. Генерируем ɣ - БСВ.
2. Составляем уравнение $F(\mathscr{E}) = \gamma$.
3. Находим $\mathscr{E} = F^{-1}(\gamma)$.

Подробно пример построения формулы моделирования случайной величины $\mathscr{E}$ рассмотрим на примере экспоненциального распределения. Случайная величина имеет экспоненциальное распределение с параметром $\lambda>0$, если её плотность распределения имеет вид $p(x)= \lambda e^{-\lambda x}$, $x>0$.

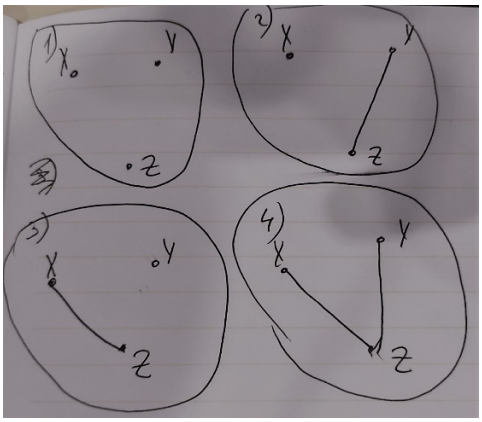
1. Генерируем ɣ - БСВ: $F(\mathscr{E}) = \gamma$, $F(\mathscr{E}) = \int_0^{\mathcal{E}} p(x)dx$
2. Составляем уравнение: $\int_0^{\mathcal{E}} \lambda e^{-\lambda x} dx = \gamma,\ -e^{-\lambda x}\,|_0^{\mathcal{E}} = \gamma$
3. Решаем уравнение относительно $\mathscr{E}$: $-e^{-\lambda\mathcal{E}} + 1 = \gamma,\ -\lambda\mathcal{E} = \ln(1-\gamma),\ \mathcal{E} = -\frac{1}{\lambda}\ln(1-\gamma)$

Аналогично можно получить формулы и для других распределений:

- Равномерное: $\mathcal{E} = a + (b-a)\gamma$



$X \perp\!\!\!\perp Y \mid Z, X \perp\!\!\!\perp Z \mid Y \iff X \perp\!\!\!\perp (Y, Z)$



[15]. Коэффициент корреляции Пирсона как показатель степени линейности корреляционной связи. Некоррелированность и независимость. Пример наличия функциональной связи и некоррелированности. Преимущества и недостатки коэффициента корреляции Пирсона.

Коэффициентом корреляции Пирсона между случайными величинами X, Y называется

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

Свойство: $|\rho(X,Y)| \leq 1$

Доказательство. Рассмотрим случайные величины $\mathrm{X}' = \frac{\mathrm{X}-\mathrm{E}(\mathrm{X})}{\sqrt{\mathrm{D}(\mathrm{X})}}$, $\mathrm{Y}' = \frac{\mathrm{Y}-\mathrm{E}(\mathrm{Y})}{\sqrt{\mathrm{D}(\mathrm{Y})}}$. Тогда $\mathrm{E}(\mathrm{X}') = \mathrm{E}(\mathrm{Y}') = 0, \mathrm{D}(\mathrm{X}') = \mathrm{D}\left(\frac{\mathrm{X}-\mathrm{E}(\mathrm{X})}{\sqrt{\mathrm{D}(\mathrm{X})}}\right) = \frac{1}{\mathrm{D}(\mathrm{X})}\mathrm{D}(\mathrm{x}) = 1$. Аналогично $\mathrm{D}(\mathrm{Y}') = 1$.

$$\mathrm{cov}(\mathrm{X}',\mathrm{Y}') = (\text{по свойству}) = \frac{1}{\sqrt{\mathrm{D}(\mathrm{X})}}\frac{1}{\sqrt{\mathrm{D}(\mathrm{Y})}}\mathrm{cov}(\mathrm{X},\mathrm{Y}) = \rho(\mathrm{X},\mathrm{Y})$$

По свойствам дисперсии $\mathrm{D}(\mathrm{X}+\mathrm{Y}) = \mathrm{D}(\mathrm{X}) + \mathrm{D}(\mathrm{Y}) + 2\mathrm{cov}(\mathrm{X},\mathrm{Y})$ и $\mathrm{D}(\mathrm{X}-\mathrm{Y}) = \mathrm{D}(\mathrm{X}) + \mathrm{D}(\mathrm{Y}) - 2\mathrm{cov}(\mathrm{X},\mathrm{Y})$:

$$\mathrm{D}(\mathrm{X}'+\mathrm{Y}') = 2\big(1+\rho(\mathrm{X},\mathrm{Y})\big), \mathrm{D}(\mathrm{X}'-\mathrm{Y}') = 2\big(1-\rho(\mathrm{X},\mathrm{Y})\big)$$

Так как для любых X, Y $\mathrm{D}(\mathrm{X}+\mathrm{Y}) \geq 0, \mathrm{D}(\mathrm{X}-\mathrm{Y}) \geq 0$, то $|\rho(\mathrm{X},\mathrm{Y})| \leq 1$

Свойство: Если коэффициент корреляции $|\rho(X,Y)| = 1$, то $P(Y = aX + b) = 1$, т.е. с вероятностью 1 между случайными величинами X, Y имеется линейная связь.

Доказательство:

Пусть $\rho(X,Y) = 1$. Тогда $D\left(\frac{X}{\sqrt{D(X)}} - \frac{Y}{\sqrt{D(Y)}}\right) = 0$. Следовательно по свойству дисперсии $P\left(\frac{X}{\sqrt{D(X)}} - \frac{Y}{\sqrt{D(Y)}} = c\right) = 1$ или $P\left(Y = \frac{\sqrt{D(Y)}}{D(X)}X - c\sqrt{D(Y)}\right) = 1$.

Пусть $\rho(X,Y) = -1$, Тогда $D\left(\frac{X}{\sqrt{D(X)}} + \frac{Y}{\sqrt{D(Y)}}\right) = 0$. Следовательно по свойству $P\left(\frac{X}{\sqrt{D(X)}} + \frac{Y}{\sqrt{D(Y)}} = c\right) = 1$ или $P\left(Y = -\frac{\sqrt{D(Y)}}{D(X)}X + c\sqrt{D(Y)}\right) = 1$.

Свойство: Если случайные величины X, Y линейно связаны, т.е. $Y = aX + b$, то $|\rho(X,Y)| = 1$.

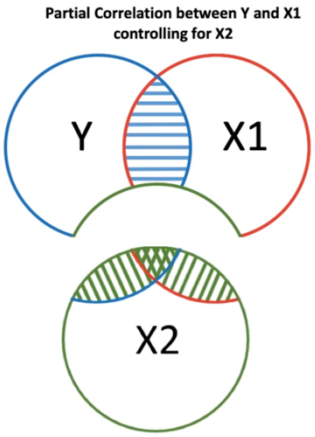
Доказательство. Пусть $Y = aX + b$. Тогда, $cov(X, aX+b) = a * cov(X,X)$, $D(Y) = a^2 D(X)$.

Поэтому $\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{D(X)D(Y)}} = \frac{aD(X)}{|a|D(X)} = \frac{a}{|a|} = \pm 1$

Свойство: Если случайные величины X, Y независимы, то $\rho(X,Y) = 0$.

Доказательство. Пусть X, Y независимы, тогда $cov(X,Y) = 0 \Rightarrow \rho(X,Y) = \frac{cov(X,Y)}{\sqrt{D(X)D(Y)}} = 0$.

[16]. Частный коэффициент корреляции и коэффициент корреляции в условном распределении. Специфика нормального распределения.



- смысл в том, что мы хотим посчитать корреляцию 2-х переменных без учета влияния остальных переменных на них, поэтому нам нужно вычесть линейную комбинацию остальных переменных, чтобы исключить их влияние.

$X_1, X_2, \dots X_N$

$$X_1{}^* = X_1 - \sum_{i=3}^{N} \alpha_i x_i$$

$$X_2{}^* = X_2 - \sum_{i=3}^{N} \beta_i x_i$$

α_i и β_i находим методом наименьших квадратов.

Частный коэффициент корреляции – это $\rho_{X_1{}^*, X_2{}^*}$

Коэффициент корреляции в условном распределении:

$$f_{X1,X2|X3,\dots XN} = \frac{f_{X1,\dots XN}}{f_{X3},..f_{Xn}}$$

Эти 2 коэффициента совпадают в нормальном законе:

$$\rho^{XY} = \frac{\rho_{XY} - \rho_{XZ} * \rho_{YZ}}{\sqrt{(1-\rho^2{}_{XZ})(1-\rho^2{}_{YZ})}}$$

[18]. Специфика сетевых моделей и сетевых структур в сетях случайных величин с эллиптическим распределением.

Рассмотрим различные сети со случайными переменными (X, γ), связанных с различными распределениями случайного вектора X и различными мерами подобия γ. Случайный вектор X принадлежит к классу эллиптических распределений, если его функция плотности имеет вид:

$f(x; \mu, \Lambda) = |\Lambda|^{-1/2} g\{(x - \mu)' \Lambda^{-1}(x - \mu)\}$

где $\Lambda = (\lambda_{i,j})$ i, j = 1, 2,…,N – положительно определённая симметричная матрица, а g(x) ≥ 0 и,

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(y'y) dy_1 dy_2 \dots dy_N = 1$$

Этот класс включает в частности многомерное гауссово распределение и многомерное распределение Стьюдента с ν степенями свободы.

Класс эллиптических распределений является естественным обобщением класса гауссовых распределений. Многие свойства гауссовых распределений имеют аналоги для эллиптических распределений, но этот класс намного больше, в частности он включает распределения с тяжёлыми хвостами. Известно, что если E(X) существует, то E(X) = μ. Одним из важных свойств эллиптического распределеия X является связь между ковариационной матрицей вектора X и матрицей Λ, а именно, если ковариационная матрица существует, то

$\sigma_{i,j} = Cov(X_i, X_j) = C \cdot \lambda_{i,j}$

где

$C = \frac{2\pi^{\frac{1}{2}N}}{\Gamma(\frac{1}{2}N)} \int_0^{+\infty} r^{N+1} g(r^2) dr$, где Γ – гамма функция

В частности для гауссова распределения $Cov(X_i, X_j) = \lambda_{i,j}$.

Для многомерного распределения Стьюдента с ν степенями свободы (ν > 2) имеем $\sigma_{i,j} = \nu / (\nu - 1) \lambda_{i,j}$.

Важное свойство эллиптических распределений - коэффициент корреляции не зависит от функции g.

$X = U_1 + U_2 + \dots + U_m + V_1 + \dots + V_n$

$Y = U_1 + U_2 + \dots + U_m + W_1 + \dots + W_n$

$$\rho = \frac{Cov(X, Y)}{\sqrt{D(X)\, D(Y)}} = \frac{m}{m + n}$$

Значение коэффициента корреляции – это пропорция общих факторов, если оно равно 0.5, значит случайные величины имеют половину общих факторов.

Наоборот не верно, т.е. их условия $\rho(X,Y)=0$ не следует, что случайные величины X, Y независимы.

Пример: $X \sim N(0,1), Y=X^3. \rho(X,Y) \neq 1$, хотя Y функционально зависит от X.

Случайные величины называются некоррелированными, если $\rho(X,Y)=0$.

Недостатки коэффициента корреляции Пирсона:

- Плохо интерпретируется
- (?) Неустойчив к выбросам
- Показывает только силу линейной связи
- Устойчив только в нормальных распределениях

[14]. Сетевые модели, сетевые структуры, сеть случайных величин. Меры связи случайных величин. Примеры.

Сетевая модель – совокупность (V, E, Y), где V −вершины, E − ребра, Y – веса или степень связи между вершинами.

Некоторые сетевые структуры:

- Клика – подмножество вершин графа, любые две из которых соединены ребром (полный подграф).
- Пороговый граф – граф, в множестве ребер которого есть только ребра с весами $y_{i,j} > y_0$, где y_0 − порог.
- MST (maximum spanning tree) – остовное дерево максимального веса в (планарном) графе. Т.е такое основное дерево, что сумма его весов максимальна.

  Остовное дерево это подграф исходного графа с тем же числом вершин. Остовное дерево получается из исходного графа удалением максимального числа рёбер, входящих в циклы, но без нарушения связности графа. Остовное дерево включает в себя все *n* вершин исходного графа и содержит *n-1* ребро.

Сеть случайных величин - граф, в котором вершины - случайные вектора, а веса ребер – корреляции между вершинами.

Меры связи:

- Коэффициент корреляции Пирсона $\rho = \frac{cov(X,Y)}{\delta_X * \delta_Y} = \frac{E[(X-E[X])*(Y-E[Y])]}{\sqrt{D(X)D(Y)}}$
- Коэффициент корреляции Кенделла $1 - \frac{2Q}{\frac{1}{2}n(n-1)}$

Коэффициент корреляции Спирмена $1 - \frac{6S(d^2)}{n^3-n}$

[19]. Коэффициенты ранговых корреляций Кендалла и Спирмена. Примеры.

Коэффициенты ранговой корреляции Кендалла и Спирмена являются статистической мерой силы зависимости признаков, представленных в порядковой (ранговой) шкале. Другими словами, коэффициенты измеряют тесноту ранговой корреляции.

Для обоих коэффициентов характерно:

- являются непараметрическими (не требуются предположения о распределении).
- значения коэффициентов инвариантны относительно изменений шкалы измерения (так как важны только ранги).
- значение 1 – полная положительная корреляция, -1 = полная отрицательная корреляция, 0 – ничего не говорит о независимости.

Пусть заданы ранги набор наблюдений 2 признаков одни и тех же n объектов.

ученики	A	B	C	D	E	F	G	H	I	J
музыка (X)	7	4	3	10	6	2	9	8	1	5
Матеша (Y)	5	7	3	10	1	9	6	2	8	4

Кендалл [τ]: будем говорить, что пары (x_i, x_j)и (y_i, y_j) согласованы, если одновременно $x_i < x_j$ и $y_i < y_j$ или $x_i > x_j$ и $y_i > y_j$. Иначе, если знак отношения рангов не совпадает – несогласованная пара. Например, AB несогласованная пара, AC AD - да.

$P -$ количество согласованных пар, $Q -$ несогласованных (инверсий). Тогда $\tau = \frac{P-Q}{\frac{1}{2}n(n-1)}$, где знаменатель – количество всевозможных пар, $P - Q$ также называют S – суммой приписанных значений.

Зная, что $P + Q = \frac{1}{2}n(n-1)$, получаем $\tau = \frac{P-Q}{\frac{1}{2}n(n-1)} = 1 - \frac{2Q}{\frac{1}{2}n(n-1)} = \frac{2P}{\frac{1}{2}n(n-1)} - 1$.

Посчитаем коэффициент Кендалла для примера. Для всех 45 пар нужно посчитать, согласованы они или нет, сложить эти количества и получить P и Q. Сделаем это более простым методом, для этого упорядочим первый признак по возрастанию:

ученики	I	F	C	B	J	E	A	H	G	D
музыка (X)	1	2	3	4	5	6	7	8	9	10
матеша (Y)	8	9	3	7	4	1	5	2	6	10

Все пары в 1 ряде входят со знаком $<$. Значит в P войдут те пары, у которых во 2 ряде также знак $<$. Для ученика I есть 2 такие пары (F и D), для F – 1 (D), для C – 5 (B, J, A, G, D). Таким образом P = 21, следовательно $\tau = \frac{2*21}{45} - 1 = -0.07$. Тау также называют коэффициентом неупорядоченности.

Спирмен [$\boldsymbol{\rho}$]: посчитаем квадраты разности рангов, обозначим сумму этих квадратов за $S(d^2)$. Тогда $\boldsymbol{\rho = 1 - \frac{6S(d^2)}{n^3-n}}$

ученики	A	B	C	D	E	F	G	H	I	J
музыка (X)	7	4	3	10	6	2	9	8	1	5

[17]. Вероятностные меры связи. Меры Блюмквиста-Краскала, Фехнера, Кендалла, Спирмена.

Пусть мы имеем 2 случайных вектора: X и Y. Коэффициент корреляции Пирсона определяется как:

$$\gamma^P(X,Y) = \frac{cov(X,Y)}{\sqrt{cov(X,X)}\sqrt{cov(Y,Y)}}$$

Общепринятая интерпретация корреляции Пирсона – мера линейной зависимости между X и Y. С другой стороны, есть большое семейство мер связи между X и Y, основанных на вероятностях. По определению:

$$P[(X > x_0 \text{ и } Y > y_0) \text{ или } (X < x_0 \text{ и } Y < y_0)] = P[(X - x_0)(Y - y_0) > 0]$$

Где x_0, y_0 – некоторые действительные числа. Различный выбор x_0 и y_0 ведёт к разным мерам похожести. Знаковая мера связи определяется при $x_0 = E(X), y_0 = E(Y)$.

$$\gamma^{Sg}(X,Y) = P[\big(X - E(X)\big)\big(Y - E(Y)\big) > 0]$$

Знаковая мера связи похожа на коэффициент корреляции Фехнера:

$$\gamma^{Fh}(X,Y) = 2\gamma^{Sg}(X,Y) - 1$$

Знаковые меры похожести показывают отклонение X и Y от их ожидаемых значений. Если $x_0 = median(X)$, а $y_0 = median(Y)$, то коэффициент корреляции Краскала записывается6 как:

$$\gamma^{Kr}(X,Y) = 2P\big[\big(X - median(X)\big)\big(Y - median(Y)\big) > 0\big] - 1$$

Можно заметить, что $\gamma^{Kr}(X,Y)$ остаётся неизменным при монотонных функциональных трансформациях аргументов (если X заменить на $f(X)$, а Y – на $g(Y)$, где f, g – монотонные, строго возрастающие (убывающие) функции, то $\gamma^{Kr}(X,Y)$ не изменится).

Чтобы избежать произвольности выбора x_0, y_0, предлагается рассмотреть разность двух независимых случайных векторов: $(X_1, Y_1), (X_2, Y_2)$, полученных из одного распределения (X, Y). Тогда мы получаем корреляцию, похожую на τ-Кендалла:

$$\gamma^{Kd}(X,Y) = 2P[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1$$

Есть очевидная связь между $\gamma^{Kd}(X,Y)$, $\gamma^{Kr}(X,Y)$ и $\gamma^{Fh}(X,Y)$:

$$\gamma^{Kd}(X,Y) = \gamma^{Kr}(X_1 - X_2, Y_1 - Y_2) = \gamma^{Fh}(X_1 - X_2, Y_1 - Y_2)$$

Если рассмотреть 3 независимых случайных вектора $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3)$, полученных из одного распределения (X, Y), то можно определить меру корреляции Спирмена:

$$\gamma^{Sp}(X,Y) = 6P[(X_1 - X_2)(Y_1 - Y_3) > 0] - 3$$

Классический коэффициент корреляции Спирмена может быть определён как несмещённая и состоятельная оценка $\gamma^{Sp}(X,Y)$. Обе вышеперечисленные меры связности также остаются неизменными после монотонных преобразований аргументов. Стоит ещё раз отметить, что $\gamma^{Kd}(X,Y)$ и $\gamma^{Sp}(X,Y)$ отличаются от своих традиционных определений. Данные размышления приведены для того, чтобы показать связь между этими мерами.

матеша (Y)	5	7	3	10	1	9	6	2	8	4
разности d	2	-3	0	0	5	-7	3	6	-7	1
разности d^2	4	9	0	0	25	49	9	36	49	1

Для нашего примера получаем $\boldsymbol{\rho = 1 - \frac{6*182}{990} = -0.103}$.

Можно доказать что нижний крайний случай спирмана = -1, но не думаю что это важно и нужно

Коэффициенты принимают очень похожие значения, можно сравнить их с разными шкалами измерения температуры. Одно из преимуществ Кенделла над Спирменом – пусть первый признак – временная шкала добавления данных, второй признак какой-нибудь численный. При добавлении новых измерений Спирмана приходится полностью пересчитывать разности, а для Кендалла только посчитать новый вклад в S.

$S(d^2) = \sum_{i=1}^{n} (p_i - q_i)^2 = 2\sum p_i^2 - 2\sum p_i q_i$

$\sum p_i^2 = \frac{1}{6}n(n+1)(2n+1)$, тогда

$$\sum (p_j - p_i)(q_j - q_i) = \frac{1}{6}n^2(n^2-1) - nS(d^2)$$

Знаменатель:

$$\sum (p_j - p_i)^2 = 2n\sum p_i^2 - 2\sum p_i p_j = 2n\sum p_i^2 - 2(\sum p_i)^2$$

$$= \frac{1}{6}n^2(n^2-1)$$

$\Gamma = \frac{\frac{1}{6}n^2(n^2-1) - nS(d^2)}{\frac{1}{6}n^2(n^2-1)} = 1 - \frac{6S(d^2)}{n^3-n} = S$

Коэффициент парной корреляции:

$$\{a_{ij} = x_j - x_i \; b_{ij} = y_j - y_i$$

$$\frac{1}{2}\sum \sum (x_j - x_i)(y_j - y_i) = n(\sum x_i y_i - \sum \sum x_i y_j) = cov(x,y)\, n$$

$$\frac{1}{2}\sum \sum (x_j - x_i)^2 = n(\sum {x_i}^2 - (\sum x_i)^2 var(x,y)\, n$$

$\Gamma = \frac{cov(x,y)}{\sqrt{var(x)\ var(y)}}$

Билет 22. Тест Кендала на проверку гипотезы зависимости, распределение, особенность с дисперсией статистики теста

Пусть имеется совокупность из N элементов и имеется 2 признака A и B, отранжируем признак A

A 1, 2 … N p_i – ранг i-го элемента

B $p_1, p_2 \dots p_n$

Пусть для двух выборок имеем $\tau_1 = 0{,}6$ и $\tau_2 = 0{,}8$

τ_2 – более существенно?

Выберем n наблюдений из N, тогда можем определить

t – выборочный коэффициент τ

Всего существует $C_N^n = \frac{N!}{n!(N-n)!}$ подвыборок

Распределение t стремится к нормальному при увеличении n, но t не очень близко к 1. Среднее распределение t=τ

Еще стандартное отклонение зависит не только от τ.

Пример:

A: 1 2 3 4 5 6 7 8 9 Всего $C_9^3 = 84$ варианта выборок.

B: 5 2 3 1 6 7 8 9 4 P – сумма положительных рангов.

P	кол-во
0	2
1	15
2	34
3	33
сум	84

$$\bar{P} = \sum p_i x_i = 0 \cdot \frac{2}{84} + 1 \cdot \frac{15}{84} + 2 \cdot \frac{34}{84} + 3 \cdot \frac{33}{84} = \frac{13}{6}$$

E(t) = $\frac{2p}{C_n^2} - 1 = 0.44$P+Q = C_n^2

τ = 2P - 1 = $\frac{52}{36} - 1 = 0.44$ = E(t) = $\frac{2Ep}{C_n^2}$ - 1

τ - для генеральной совокупности

t – оценка

[23]. Специфика рандомизационных и популяционных моделей. Пояснить на примерах тестов Вилкоксона и Кендалла.

Рандомизационная модель основывается на случайном распределении объектов между группами, случайном распределении объектов между группами и случайном назначении им значений. Например, в тесте Вилкоксона рандомизационная модель используется для проверки различий между двумя выборками. Значения в каждой выборке перетасовываются случайным образом, а затем рассчитывается статистика, которая показывает вероятность получения таких или более экстремальных результатов, если гипотеза о равенстве средних значений в выборках верна.

Популяционная модель основывается на предположении о распределении генеральной совокупности и использовании статистических методов для выводов о ней. Например, в тесте Кендалла популяционная модель используется для проверки связи между двумя переменными. Рассчитывается коэффициент корреляции Кендалла, который показывает степень связи между переменными на основе предположения о распределении генеральной совокупности.

Таким образом, рандомизационная модель используется для проверки гипотез о различиях между выборками или группами, а популяционная модель - для проверки связей между переменными на основе предположения о распределении генеральной совокупности.

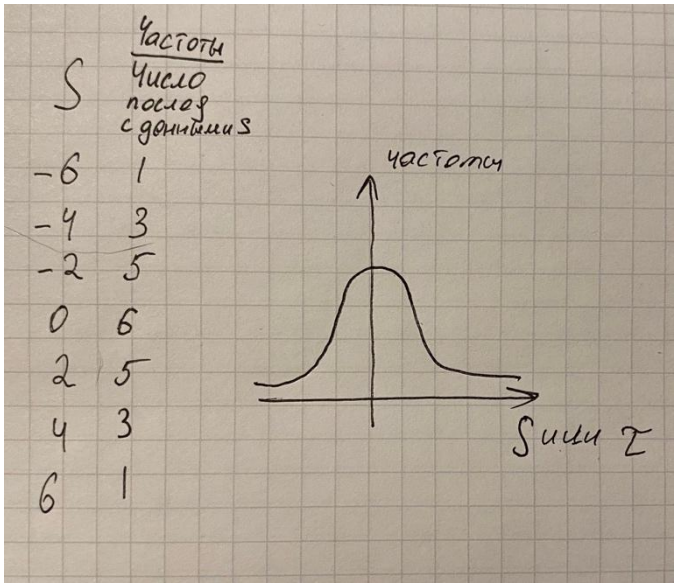
Билет 21. Тест Кендалла. Проверки гипотезы независимости, распределение, примеры

Предполагаем, что выборки A и B независимы, тогда как распределен коэффициент корреляции Кендалла?

Мы отранжируем 1 выборку, тогда у второй выборки перестановки равновероятны и равны $\frac{1}{n!}$.

У каждой возможной последовательности есть τ

Пример: n = 4, тогда есть n! = 24 перестановки, тогда частоты симметричны относительно центра S = 0. С увеличением n распределение близится к нормальному. В точках $\pm\frac{1}{2}n(n-1)$частоты равны 1.



Проверка значимости τ эквивалентна проверке S

Если между A и B нет взаимосвязи и значение τ соответствует наблюдаемым значениям S в хвостах распределения $+= \frac{1}{2}n(n-1)$ (то есть маловероятное событие), то гипотезу о независимости A и B отвергаем.

Пример: для n=10 имеем τ = 0.11, этому значению ставится $s = \pm 5$ и больше, что составляет 0.728 – долю от всех перестановок, это значительная величина (S не в хвостах), то есть гипотезу не отвергаем (о независимости A и B). А в случае τ = 0.56 этому соответствует S=25 -> 0.028 для незначительная -> гипотеза о независимости отвергается.

S = P-Q P—сумма положительный рангов, Q – сумма отрицательных рангов

Билет 20. Обобщенный коэффициент корреляции и его частные случаи

Пусть имеется n объектов с двумя свойствами x, y

x: $x_1 \dots x_n$

y: $y_1 \dots y_n$

Для каждой пары i, j имеем оценку a_{ij}, b_{ij}

a_{ij}=$-a_{ji}$, b_{ij}=$-b_{ji}$, при этом a_{ij}=0 если i=j

Коэффициент корреляции: $\Gamma = \frac{\sum_{ij=1}^{n} a_{ij}\, b_{ij}}{\sqrt{\sum_{ij=1}^{n} a_{ij}^2 \sum_{ij=1}^{n} b_{ij}^2}}$

Частные случаи:

τ – Кендала

Пусть p_i – ранг i-го элемента

Тогда $\{a_{ij} = +1;\ p_i < p_j\ a_{ij} = -1;\ p_i > p_j \quad \{b_{ij} = +1;\ p_i < p_j\ b_{ij} = -1;\ p_i > p_j$

Тогда:

$\Gamma = \tau = \frac{2S}{n(n-1)} = 1 - \frac{2L}{\frac{1}{2}n(n-1)}$, где $L = \frac{1}{2}\left[\frac{1}{2}n(n-1) - S\right]$

Коэффициент Спирмена

$a_{ij} = p_j - p_i$, p_i – ранг *i*-го значения по признаку *x*

$b_{ij} = q_j - q_i$, q_i – ранг *i*-го значения по признаку *y*

$p_i = \underline{1n}$, $q_j = \underline{1n}$, тогда

$$\sum_{ij=1}^{n} (p_j - p_i)^2 = \sum_{ij=1}^{n} (q_j - q_i)^2$$

$\Gamma = \frac{\sum (p_j - p_i)(q_j - q_i)}{\sum (p_j - p_i)^2}$

Числитель:

$$\sum \left(p_j - p_i\right)(q_j - q_i)$$

$$= \sum_{i}^{n} \sum_{j}^{n} p_i q_i + \sum_{i}^{n} \sum_{j}^{n} p_j\, q_j - \sum_{i}^{n} \sum_{j}^{n} (p_i q_j) + \left(p_j q_i\right)$$

$$= 2n \sum_{i=1}^{n} p_i q_i - 2 \sum_{i=1}^{n} p_i \sum_{j=1}^{n} q_j = 2n \sum_{i=1}^{n} p_i q_j - \frac{1}{2} n^2 (n+1)^2$$

$\sum (p_j - p_i)(q_j - q_i) = 2n \sum p_i^2 - \frac{1}{2} n^2 (n+1)^2 + nS(d^2)$, где

[24]. Коэффициент конкордации, распределение, пример.

Коэффициент конкордации Кендалла – некоторое число в промежутке $[0; 1]$, характеризующее степень согласованности мнений экспертов / корреляции нескольких выборок. Обычно используется для измерения статистической связи между несколькими выборками. В отличие от корреляции Пирсона не требует предположения о нормальности выборок и позволяет одновременно сравнивать любое их количество.

Пример: пусть имеется n объектов и k экспертов. Каждый эксперт выставляет оценки каждому объекту (различные целые числа от 1 до n). Требуется выяснить, насколько согласны между собой эксперты.

Вычисление:

- Пусть заданы $k \geq 2$ выборок $x_1 = (x_1^1, x_2^1, \dots, x_n^1), \dots, x_k = (x_1^k, x_2^k, \dots, x_n^k)$.
- Пусть также r_{ij} – ранг i-го объекта в j-й выборке.
- Ранговый коэффициент конкордации Кендалла W определяется по формуле:

$$W = \frac{12}{k^2(n^3 - n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{k} r_{ij} - \frac{k(n+1)}{2} \right)^2$$

Коэффициент конкордации равен 1 при максимальной согласованности (когда $\forall i, j, k\ r_{ij} = r_{ik}$) и равен 0 при максимальной несогласованности. Коэффициент не принимает отрицательных значений, поскольку для множества выборок не определена противоположность согласованности – упорядочения могут полностью совпадать, но не могут “полностью не совпадать”.

Если мы поменяем последовательность, но она будет с таким же τ, то распределение P поменяется -> будет разная дисперсия, то есть нужна последовательность рангов из генеральной совокупности, но для любой исходной совокупности есть ограничение на $var\ t\ \leq \frac{2}{n}\ (1-\tau^2)$. Но все это работает если выборка случайна.