Задачи дисперсионного анализа (проверка однородности в основном) чаще всего встречаются в производстве, когда нужно сравнить выработанные изделия, к примеру, с контрольным образцом. Т.е. есть 2 выборки: одна - из рабочей совокуп., другая - контрольная. Ну и хотим, например, понять, есть ли эффект обработки, т.е. отличается ли рабочая выборка от контрольной по распределению.

Итак, пусть есть $X=(X_1 \dots X_m)$ - н.о.р. сл.в. с ф.р. $F(x)$; $Y=(Y_1 \dots Y_n)$ - н.о.р. сл.в. с ф.р. $G(x)$. Практически везде будем считать, что $F(x)$ и $G(x)$ - непрерывны.

Важным для практики случаем является случай нормально распред выборок, т.е. $X_i \sim N(\mu_1, \sigma_1^2)$, $Y_j \sim N(\mu_2, \sigma_2^2)$. Проверим эти выборки на однородность при условии, что эти выборки независимы.

Помним, что $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ несмещённо оцен. σ_1^2 и ~~$\frac{S_1^2}{\sigma_1^2}$~~ $\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n-1}$, а $ES_2^2=\sigma_2^2$ и $\frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2_{m-1}$.

Напоминание. Сл.в. ξ имеет распр. Фишера с k_1 и k_2 степ. свободы, если $\xi \overset{d}{=} \frac{\frac{1}{k_1}\cdot\zeta}{\frac{1}{k_2}\eta}$, где $\zeta \sim \chi^2_{k_1}$; $\eta \sim \chi^2_{k_2}$ и $\zeta \perp\!\!\!\perp \eta$.

Обозн.: $\xi \sim F_{k_1,k_2}$.

Критерий Фишера Если верна H': $\sigma_1=\sigma_2$, то $\frac{S_1^2}{S_2^2} \sim F_{n-1,m-1}$, тогда критерий: $\left\{\frac{S_1^2}{S_2^2} > f_{1-\alpha}\right\}$, $f_{1-\alpha}$ - квантиль ур. $1-\alpha$ ч $F_{n-1,m-1}$

здесь большая делится на меньшую.

Если крит. Фишера не отвергает H', то проверяем H'': $\mu_1=\mu_2$

Поскольку $S_1^2 \perp\!\!\!\perp S_2^2$, то $\frac{1}{\sigma^2}[(n-1)S_1^2+(m-1)S_2^2] \sim \chi^2_{n+m-2}$, кроме того, $E\xi=k$, если $\xi \sim \chi^2_k$. Тогда $S^2_{tot} = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$ несмещённо оценивает σ^2. Далее, $\bar{X}-\bar{Y} \sim N\left(0, \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right)$ и не зависит от S^2_{tot} (т. об ортогональном разложении гаусс вектора)

Напоминание Пусть $\xi_0 \dots \xi_n$ - нез. $N(0,1)$, тогда $\xi = \frac{\xi_0}{\sqrt{\xi_1^2+\dots\xi_n^2}}$ имеет распред Стьюдента с n степ. свободы. (обозн. $\xi \sim t_n$)

Имеем $T = \frac{\bar{X}-\bar{Y}}{\sqrt{S^2_{tot}\left(\frac{1}{n}+\frac{1}{m}\right)}} = \sqrt{\frac{nm}{n+m}}\,\frac{\bar{X}-\bar{Y}}{S_{tot}} \sim t_{n+m-2}$

Это и есть критерий Стьюдента проверки однород нормальных выборок.

Замечание 1) Оптимален для норм. выборок, но его мощность сильно снижается при отклон. от строгой нормальности

2) Лучше брать ≈ равные выборки.

Рассмотрим ещё несколько методов проверки однородности независимых выборок, если уже они не являются нормальными.

Критерий Смирнова

Определим $D_{n,m} = \sup_x |F_n^*(x) - G_m^*(x)|$, где $F_n^*(x)$ и $G_m^*(x)$ – эмпирические ф.р. выборок $\{X$ и Y соотв.

Теорема (Смирнов) Если верна $H_0: F = G$, то $\sqrt{\frac{nm}{n+m}} D_{n,m} \xrightarrow{d} K$, где K – сл.в., имеющая ф.р. Колмогорова $F_K(z) = \sum_{i \in \mathbb{Z}} (-1)^i e^{-2i^2 z^2}$.

Приближение хорошо работает при $n, m \geq 20$, при $n, m < 20$ искать ~~таблицы~~ критич. знач. в книгах или с пом. моделирования.

Омега-квадрат критерий Розенблатта

Статистика критерия $\omega^2_{n,m} = \int_{-\infty}^{+\infty} [F_n^*(x) - G_m^*(x)]^2 \, d H^*_{n+m}(x)$, где $H^*_{n+m}(x) = \frac{n}{n+m} F_n^*(x) + \frac{m}{m+n} G_m^*(x)$ – эмпир. ф.р. выборки $(X_1 \dots X_n, Y_1 \dots Y_m)$

(есть явный вид $\omega^2_{n,m} = \frac{1}{nm}\left[\frac{1}{6} + \frac{1}{m}\sum_{i=1}^{n}(R_i - i)^2 + \frac{1}{n}\sum_{j=1}^{m}(S_j - j)^2\right] - \frac{2}{3}$, где R_i – ранг $X_{(i)}$, S_j – ранг $Y_{(j)}$ в объед. вариационном ряду)

Теорема (Розенблатт) Если $H_0: F = G$ верна, то при $\frac{n}{n+m} \to \lambda \in (0,1)$ $\frac{nm}{n+m} \omega^2_{n,m} \xrightarrow{d} \zeta$, где ζ имеет табличное распред.

Замечание При малых n, m статистику $Z = \frac{nm}{n+m} \omega^2_{n,m}$ следует заменить на $Z^* = \frac{Z - EZ}{\sqrt{45 DZ}} + \frac{1}{6}$, где $EZ = \frac{1}{6}\left(1 + \frac{1}{n+m}\right)$, $DZ = \frac{1}{45}\left(1 + \frac{1}{n+m}\right)\left(1 + \frac{1}{n+m} - \frac{3}{4}\left(\frac{1}{n} + \frac{1}{m}\right)\right)$, это обеспеч. хорошую точность уже при $n, m \geq 7$.

Критерий ранговых сумм Уилкоксона (или крит. Манна–Уитни–Уилкоксона, англ. обозн. MWW)

Проверяем тут гипотезу о сдвиге: предполагаем, что $G(y) = F(y - \theta)$ и $H_0: \theta = 0$ (альтернативы бывают $H_+: \theta > 0$, $H_-: \theta < 0$, $H_1: \theta \neq 0$, для них всех критерий явл. состоятельным)

Рассм. объединённую совокупность $(X_1 \dots X_n, Y_1 \dots Y_m)$, перейдем к рангам сл.в. в этой совокупности: ~~$R(X_i)$~~ $R_i := R(X_i)$ и $S_j := R(Y_j)$. $N = n+m$. Определим $W_{m,n} = \sum_{j=1}^{m} S_j$ – статистика ранг. суммы Уилкоксона.

Теорема При верной H_0 $\frac{W_{m,n} - m \cdot \frac{n+m+1}{2}}{\sqrt{nm \frac{n+m+1}{12}}} \xrightarrow{d} N(0,1)$.

Замечание При $n, m \leq 50$ статистика $W^* = \frac{W_{m,n} - m \cdot \frac{n+m+1}{2}}{\sqrt{nm \frac{n+m+1}{12}}}$ плохо аппроксимируется норм. законом, в этом случае стоит применять $\widetilde{W} = \frac{1}{2} W^* \cdot \left(1 + \sqrt{\frac{N-2}{N-1-(W^*)^2}}\right)$, где в кач. критич. значения брать $z_{1-\alpha} = (x_{1-\alpha} + y_{1-\alpha})/2$, $x_{1-\alpha}$ и $y_{1-\alpha}$ – квантили ур. $1-\alpha$ ~~из~~ $N(0,1)$ и t_{N-2} соотв. Если $n, m \geq 10$, то применять табл. значения критич.

Представляет интерес также ситуация, когда выборки $(X_1 \dots X_n)$ и $(Y_1 \dots Y_n)$ зав. парными, т.е. $(Z_1 \dots Z_n)$, где $Z_i = (X_i, Y_i)^T$, зав. выборка из векторов, зависимость между X_i и Y_i неизвестна.

Например, такая задача возникает, когда мы делаем измерения на нек. приборах и делаем замеры дважды, пытаясь понять, изменилось ли что-нибудь.

Из-за неизвест. зависимости между X_i и Y_i приходится пользоваться ранговыми методами, причём будем проверять гипотезу об отсутствии сдвига, как и в крит. Уилкоксона.

Рассмотрим $V_i = Y_i - X_i = \theta + \varepsilon_i$, где θ – интерес. нас эффект воздействия. Будем считать, что $P(\varepsilon_i \le 0) = P(\varepsilon_i \ge 0) = \frac{1}{2}$ и распред. ε_i непрерывно. Кроме того, ε_i неодинак. по построению.

Критерий знаков

Рассм. $S = \sum_{i=1}^{n} I(V_i > 0)$. Тогда при верной $H_0: \theta = 0$ $\frac{S - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \xrightarrow{d} N(0,1)$.

При $n < 15$ нормал. аппроксимация работает плохо, то можно вычислить квантили из след.: $P_0(S \ge k) = 2^{-n} \sum_{i=k}^{n} C_n^i$.

Замечание Если всё же распред. ε не зав. непрерывным и среди значений V_i есть нулевые, то их надо отбросить и уменьшить n до числа ненул. эл-тов.

Критерий знаковых рангов Уилкоксона

Теперь вообще считаем, что $\varepsilon_1 \dots \varepsilon_n$ – н.о.р. с симметр. непрер. распред.

Статистика: $T = \sum_{i=1}^{n} R_i I(V_i > 0)$, где R_i – ранг $|V_i|$ в вариац. ряду выборки $(|V_1|, \dots |V_n|)$.

Теорема: при верной $H_0: \theta = 0$ $T^* = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{d} N(0,1)$

При $n \le 15$ следует пользоваться табл. крит. значениями, т.к. нормальное приближение работает плохо.

Замечание 1) Поправка: $\widetilde{T} = \frac{1}{2} T^* \left(1 + \sqrt{\frac{n-1}{n - (T^*)^2}}\right)$, уточняет аппроксимацию при малых n. Тогда крит. знач. $z_\alpha = \frac{x_{1-\alpha} + y_{1-\alpha}}{2}$, где $x_{1-\alpha}$ и $y_{1-\alpha}$ – $(1-\alpha)$-квантили ч/у $N(0,1)$ и t_{n-1} соотв.

2) Если всё-таки распред. ε_i не непрерывно, то если $V_i = 0$ для нек. i, то отбрас. их и уменьш. n.

А если среди $|V_i|$ есть равные, то используем средние ранги, а дисперсия меняется так: $\frac{1}{24}\left[n(n+1)(2n+1) - \frac{1}{2}\sum_{k=1}^{g} l_k(l_k^2 - 1)\right]$, где g – число групп совпадений, а l_k – кол-во эл-тов в k-той группе.

3) Чтобы применять критерий рановых знаков, нужно убедиться хотя бы в симметричности ε; делается это так: должны примерно совпадать $\zeta_i = -V_{(i)} + \hat{\mu}$ и $\eta_i = V_{(n+1-i)} - \hat{\mu}$, $i = 1 \dots [n/2]$, $\hat{\mu}$ – выборочн. медиана.

Мы вот для норм. распределений проверяли только гипотезу однород., а иногда хочется и гипотезу о равенстве средних проверить.

Итак, пусть $X_1 \dots X_n \sim N(\mu_1, \sigma_1^2)$; $Y_1 \dots Y_m \sim N(\mu_2, \sigma_2^2)$ и выборки неавис.

Тогда при верной H_0 $T(X,Y) = \dfrac{\bar X - \bar Y}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}} \overset{\text{примерно}}{\sim} t_k$, где

$$k = \frac{\left(\frac{s_1^2}{n} + \frac{s_2^2}{m}\right)^2}{\frac{s_1^4}{n^2(n-1)} + \frac{s_2^4}{m^2(m-1)}}.$$

Критерий Асина-Уэлча.

(С этой задачей связана известная проблема Беренса-Фишера о несуществовании наилучшего критерия в этой задаче)

Пусть теперь наблюдения парные (ещё говорят, что выборки связанные) и $n=m$.

При верной $H_0: \mu_1 = \mu_2$ $T(X,Y) = \dfrac{\bar X - \bar Y}{S/\sqrt{n}} \sim t_{n-1}$, где

$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar D)^2}$, $D_i = X_i - Y_i$.

(Это называется t-критерий Стьюдента для связанных выборок)

Кроме того, есть множество непараметрических критериев, которые проверяют частные случаи гипотезы однородности – например, гипотезу о равенстве средних (перестановочный критерий) или гипотезу о равенстве дисперсий (критерий Зигеля-Тьюки).

Приведём критерий Зигеля-Тьюки.

Пусть выборки X и Y независимы, проверяем гипотезу $H_0: DX_1 = DY_1$ vs. $H_1: DX_1 \neq DY_1$.

Пусть $(Z_{(1)} \dots Z_{(n+m)})$ – вариац. ряд объединённой выборки $(X_1 \dots X_n, Y_1 \dots Y_m)$.

Статистика критерия $R_1(X,Y) = \sum_{i=1}^{n} \widetilde{rank}(X_i)$, где $\widetilde{rank}$ присваивается так: (т.е. по Y_j ранги не суммируем.)

$Z_{(i)}$:	$Z_{(1)} \le$	$Z_{(2)} \le$	$Z_{(3)} \le$	…	$\le Z_{(n+m-2)} \le$	$Z_{(n+m-1)} \le$	$Z_{(n+m)}$
$\widetilde{rank}$	1	4	5	…	6	3	2

$R_1(X,Y)$ при верной H_0 имеет табличное распределение.