Чтобы $u(r)$ была распределена по нормальному закону (так удобно считать квантили), с $r$ производится преобразование Фишера:

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r}$$

$$M(Z) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}\ ; \quad D(Z) = \frac{1}{n-3}.$$

Тогда

$$u = \sqrt{n-3}\left(Z - \frac{1}{2}\ln\frac{1+\rho}{1-\rho}\right) \sim N(0;1)$$

распределена по стандартному нормальному закону.

Она с вероятностью $1-\varepsilon$ будет заключена в пределах $\pm u_{1-\varepsilon/2}$. Если решить полученное неравенство

$$P\left(-u_{1-\varepsilon/2} < \sqrt{n-3}\left(Z - \frac{1}{2}\ln\frac{1+\rho}{1-\rho}\right) < u_{1-\varepsilon/2}\right) = 1-\varepsilon,$$

то получатся страшные и огромные формулы для вычисления $\rho_{min}$ и $\rho_{верх}$.

Этот доверительный интервал чувствителен к отклонениям от нормальности и годится только для нормально распределенных двумерных случайных величин.

**Билет № 15.** Проверка статистических гипотез. Вероятность ошибки 1-го и 2-го рода. Уровень значимости критерия и мощность критерия.

Пусть есть некоторая случайная выборка, распределенная по неизвестному закону:

$x_1; x_2; \ldots; x_n = \xi; \quad \xi \sim F(x)$.

И пусть есть некая характеристика этого закона $\theta$. (мат. ожид., дисперсия или что-нибудь еще).

Мы предполагаем, что она равна некоторому числу $\theta_0$: вводим нулевую гипотезу:

$$H_0: \theta = \theta_0.$$

И альтернативную ей гипотезу $H_1$:

$H_1: \theta = \theta_1$ простая альтернатива

$H_1: \theta \neq \theta_0$ двусторонняя альтернатива

$H_1: \; \theta \neq \theta_0$
$\theta > \theta_0$
$\theta < \theta_0$ } сложные альтернативы

$H_1: \; \theta > \theta_0$
$\theta < \theta_0$ } односторонние альтернативы

Выбрав нулевую гипотезу, мы берем нашу выборку и вводим для нее некоторую функцию – статистику критерия $g = g(x_1; x_2; \ldots; x_n)$.

Если $H_0$ верна, то эта функция будет иметь одно распределение, если не верна – то другое. В области значений этой ф-ии выбирается некоторая область $W$ (критическая область), такая, что вероятность попадания значения ф-ии $g$ в эту область не превосходит заданного малого значения $\alpha$ – уровня значимости критерия. Т.е. если верна $H_0$, то мы не будем выходить за критическую область, а уж если мы все время попадаем за нее, значит, $H_0$ не верна, а верна $H_1$.

Ошибки.

- Ошибка 1-го рода: верна $H_0$, но мы ее отвергаем (наша гипотеза верна, но мы говорим, что не верна).
- Ошибка 2-го рода: верна $H_1$, но мы ее отвергаем (наша гипотеза не верна, но мы говорим, что она верна).

Вероятность ошибки 1-го рода $\alpha$ и есть уровень значимости критерия (вероятность попасть в критическую область при верной $H_0$).

Вероятность ошибки 2-го рода $\beta$.

$1-\beta$ – вероятность правильно принять $H_1$, когда верна $H_1$. $1-\beta$ – мощность критерия.

Чем меньше вероятность ошибки 2 рода, тем больше $1-\beta$, тем более мощный критерий.

Пример: проверка гипотез о математических ожиданиях.

Пусть $\xi \sim N(\mu; \sigma^2)$. $H_0: \mu = \mu_0;$ $H_1: \mu = \mu_1;$ и есть выборка $x_1; x_2; \ldots; x_n$ значений случайной величины $\xi$.

Но если $\xi \sim N(\mu; \sigma^2)$, то $\frac{\xi - \mu}{\sigma} \sim N(0; 1)$ – стандартное распределение.

А поскольку выборочное среднее имеет математическое ожидание исходной случайной величины, и дисперсию, равную дисперсии исходной случайной величины, поделении на $\sqrt{n}$, то мы заменяем $\xi$ на $\bar{x}$, а $\sigma$ на $\sigma/\sqrt{n}$ и получаем стандартизованное выборочное среднее:

$$u = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \qquad D(u)=1 \quad M(u)=0$$

А т.к. выборочное среднее - оценка для математического ожидания, то мы его выбираем в качестве статистики критерия.

$H_0: \mu=\mu_0$; распределение $U$ стандартно.

$H_1: \mu=\mu_1$; распределение отличается от стандартного.

В качестве статистики критерия выбирается величина, характеризующая степень отклонения от нулевой гипотезы.

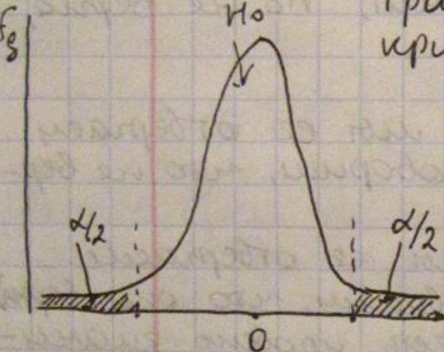
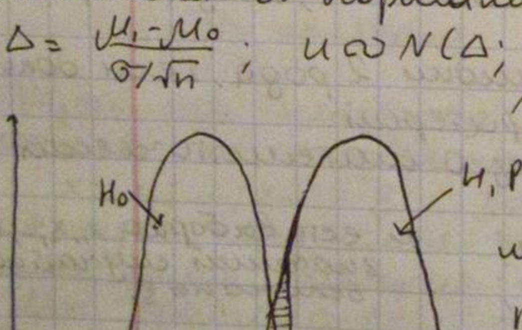



График функции плотности статистики критерия: при верной $H_0$ $\mu=\mu_0$

Если мы получаем значения выборочного среднего и такие, что они попадают в критическую область, то есть повод задуматься, что $H_0$ не верна, а верна $H_1$. Но в то же время $H_0$ может быть и верна с вероятностью $\alpha$, а мы ее отвергли, потому что и попало в критическую область. Поэтому $\alpha$ - вероятность ошибки 1-го рода.

Если верна $H_1$, то распределение статистики $U$ отличается от нормального на некоторую величину $\Delta$:

$$\Delta = \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}; \quad u \sim N(\Delta; 1)$$




Чем больше $\Delta$, тем меньше вероятность ошибки 2 рода $\beta$.

Чем меньше дисперсия, тем меньше вероятность $\beta$.

Чем больше $\alpha$, тем меньше $\beta$.

Поскольку ошибки 1-го рода опаснее, то мы в первую очередь следим за $\alpha$ и стараемся его уменьшить. А $\beta$ - как получится.

Мощность критерия - способность обнаружить имеющееся отклонение от нулевой гипотезы.

Ошибка 1 рода - ошибка ложного обнаружения несуществующего эффекта
Ошибка 2 рода - ошибка ложного необнаружения существующего эффекта.

## Билет № 16. Одновыборочный t-критерий.

Проверка гипотезы о равенстве заданному числу математического ожидания нормально распределенной случайной величины с неизвестной дисперсией.

В предыдущем случае нам была известна дисперсия, и мы брали в качестве статистики критерия стандартизованное выборочное среднее $u = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.

Оно у нас было распределено по стандартному закону. Если дисперсия неизвестна, то вместо $\sigma$ придется брать ее оценку $S$, но $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$, поэтому $u = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ будет иметь t-распределение с $(n-1)$ степенями свободы:

$$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$ (если верна $H_0$).

Тогда критическая область для проверки $H_0: \mu = \mu_0$ при альтернативе $H_1: \mu \neq \mu_0$ будет состоять из двух бесконечных полуинтервалов

$$(-\infty;\, t_{n-1;\frac{\alpha}{2}}] \cup [t_{n-1;\,1-\frac{\alpha}{2}};\, \infty),$$

где $t_{n-1;\frac{\alpha}{2}}$ и $t_{m;\,1-\alpha/2}$ - квантили t-распределения с $(n-1)$ степенью свободы порядков $\alpha/2$ и $1-\alpha/2$ (а в силу симметричности t-распределения $t_{\alpha/2} = -t_{1-\alpha/2}$).

Тогда если $t$ попадет в интервал $[t_{n-1;\alpha/2};\, t_{n-1;1-\alpha/2}]$, то $H_0$ согласуется с экспериментальными данными.

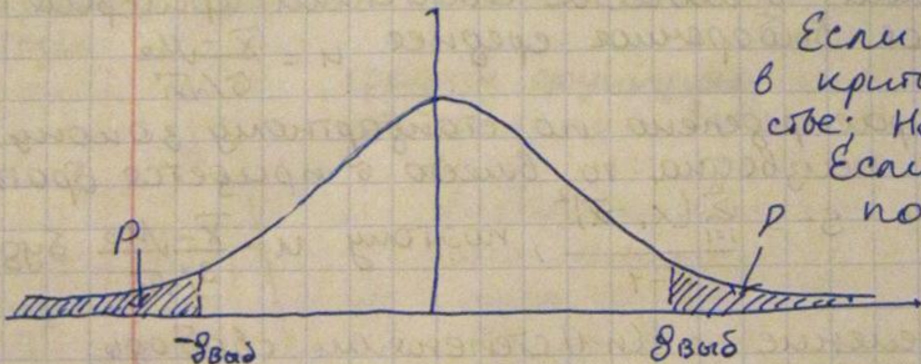
Для случаев, когда статистика выборки попала близко к границе и очень легко может через нее перейти, если поварьировать данными, а значит, резко возрастает вероятность ошибки 2 рода, существует P-значение.

P-значение - вероятность того, что статистика критерия $g$ по модулю превзойдет $g$, вычисленное по выборке:

$$P = P\{|g| > g_{выб.}\}.$$

Т.е. мы считаем $g$ для некоторой выборки, и вероятность того, что $g$ для любой другой выборки будет больше, чем для этой, будет P-значением.

Чем меньше P, тем дальше "заехало" $g$ для нашей выборки, тем меньше шансов его превзойти и тем больше шансов у нашей выборки вообще выпасть за границу в критическое множество:




Если $P < \alpha$, то мы в критическом множестве; $H_0$ не верна

Если $p > \alpha$, то мы не попали в критическое множество. $H_0$ верна.

Т.е. P-значение показывает, как близко мы от границы.

Применение t-критерия требует нормальности исходной случайной величины, но его можно использовать и при умеренных отклонениях от нормальности.

**Билет № 17.** Двухвыборочный t-критерий для независимых и связанных выборок.

Проверка гипотезы о равенстве математических ожиданий двух нормально распределенных случайных величин (двухвыборочный t-критерий).

1) Независимые выборки

Пусть есть 2 случайные выборки $x_1; x_2; \dots; x_n$ и $y_1; y_2; \dots; y_n$ значений двух независимых нормально распределенных случайных величин $\xi \sim N(M(\xi); D(\xi))$ и $\eta \sim N(M(\eta); D(\eta))$.

Гипотеза $H_0: M(\xi) = M(\eta)$.

$H_1: M(\xi) \neq M(\eta)$.

1) Известно, что $D(\xi) = D(\eta) = \sigma^2$, значение $\sigma^2$ неизвестно. Тогда для $\sigma^2$ можно получить объединенную несмещенную оценку:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{j=1}^{m}(y_j - \bar{y})^2}{n+m-2}.$$

Тогда $s^2/n$ и $s^2/m$ будут несмещенными оценками для дисперсии выборочных средних $\bar{x}$ и $\bar{y}$, а сумма $\frac{s^2}{n} + \frac{s^2}{m}$ будет несмещенной оценкой для дисперсии разности средних $\bar{x} - \bar{y}$. Тогда статистика

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s^2}{n} + \frac{s^2}{m}}}$$

имеет $t$-распределение с $n+m-2$ степенями свободы.

Критическая область будет состоять из двух бесконечных полуинтервалов

$(-\infty; t_{n+m-2;\frac{\alpha}{2}}]$ и $[t_{n+m-2;1-\alpha/2}; \infty)$, где

$t_{n+m-2;\alpha/2}$ и $t_{n+m-2;1-\alpha/2}$ — квантили $t$-распределения порядков $\alpha/2$ и $1-\alpha/2$ со степенями свободы $n+m-2$.

2) Если $D(\xi) \neq D(\eta)$, то для каждой из дисперсий вычисляется своя оценка:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \text{и} \quad s_y^2 = \frac{\sum_{j=1}^{m}(y_j - \bar{y})^2}{m-1}$$

Тогда статистика критерия

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}};$$

имеет $t$-распределение с числом степеней свободы, равным целой части от $1/k$, где

$$k = \frac{\left(\frac{s_x^2/n}{s_x^2/n + s_y^2/n}\right)^2}{n-1} + \frac{\left(\frac{s_y^2/m}{s_x^2/n + s_y^2/m}\right)^2}{m-1}.$$

2). Связанные выборки.

Пусть $x_1; x_2; \dots; x_n$ и $y_1; y_2; \dots; y_n$ — связанные случайные выборки из нормальных распределений $\xi \sim N(M(\xi); D(\xi))$ и $\eta \sim N(M(\eta); D(\eta))$.

$H_0: M(\xi) = M(y)$;
$H_1: M(\xi) \neq M(y)$.

Статистика критерия: $t = \dfrac{(\bar{x}-\bar{y})\sqrt{n}}{\sqrt{S_x^2+S_y^2-2S_{xy}}}$, где

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}).$$

$t$ имеет $t$-распределение с $(n-1)$ степенями свободы; критическая область состоит из двух интервалов $(-\infty; -t_{n-1;1-\alpha/2})$ и $(t_{n-1;1-\alpha/2};\infty)$, где

$t_{n-1;1-\alpha/2}$ - квантиль $t$-распределения порядка $1-\alpha/2$ с числом степеней свободы $n-1$.

Билет № 18. Двухвыборочный F-критерий.

Проверка гипотезы о равенстве дисперсий двух независимых нормально распределенных случайных величин.

Пусть есть 2 случайные выборки $x_1; x_2; \dots; x_n$ и $y_1; y_2; \dots; y_m$ значений двух независимых нормально распределенных случайных величин

$\xi \sim N(M(\xi); D(\xi))$; $y \sim N(M(y); D(y))$;
$H_0 = D(\xi) = D(y)$
$H_1 = D(\xi) \neq D(y)$.

Статистика критерия - отношение несмещенных оценок дисперсий этих случайных величин:

$F = \dfrac{S_x^2}{S_y^2}$. имеет $F$-распределение с $(n-1)$ и $(m-1)$ степенями свободы

($n$ и $m$ - объемы двух выборок).

Критическая область состоит из двух интервалов:
$(0; F_{n-1;m-1;\alpha/2}]$ и $[F_{n-1;m-1;1-\alpha/2};\infty)$, где

$F_{n-1;m-1;\alpha/2}$ и $F_{n-1;m-1;1-\alpha/2}$ - квантили порядка $\alpha/2$ и $1-\alpha/2$ $F$-распределения с $n-1$ и $m-1$ степенями свободы.

В отличие от $t$-критерия, $F$ критерий чувствителен к отклонениям исходных случайных величин от нормальности.

# Билет №19. Критерий согласия $\chi^2$ и Колмогорова-Смирнова.

Критерий согласия — критерий для проверки согласия между распределением выборочных значений и заданным теоретическим распределением.

Пусть есть выборка $x_1; x_2; \dots; x_n$ значений случайной величины $\xi$ с неизвестной ф-ей распределения $F(x)$

$H_0: F(x) = F_0(x)$ $\quad$ $F_0(x)$ — некоторое заданное распределение.

$H_1: F(x) \neq F_0(x)$

Если $F_0(x)$ задано полностью, то мы проверяем простую нулевую гипотезу, если о $F_0(x)$ нам известно только то, что она принадлежит к некоторому виду ($F_0(x)$ - нормальное распределение, или $F$-распределение, или $t$-распределение), но больше мы о $F_0(x)$ ничего не знаем (ни ее дисперсии, ни мат. ожидания, ничего) - то это сложная нулевая гипотеза.

1) Критерий согласия $\chi^2$.

$F(x)$ может быть и непрерывным, и дискретным.

1) Простая нулевая гипотеза

Область изменения значения выборки разбивается на $k$ интервалов так, чтобы число наблюдений $n_i$, попавших в $i$-ый интервал, было не менее 10.

$n_i$ - эмпирическое число наблюдений в $i$-ом интервале.

$np_i$ - теоретическое число наблюдений в $i$-ом интервале, т.е. каким оно должно было быть, если бы $H_0$ была верна, и выборка принадлежала к $F_0(x)$.

Здесь $n$ - объем выборки, $p_i$ - вероятность попадания в $i$-ый интервал, рассчитанная исходя из известного распределения $F_0(x)$.

Тогда статистика критерия

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}.$$

Если $H_0$ верна, то распределение этой статистики приближается $\chi^2$-распределением с $k-1$ степенями свободы. Критическое множество состоит из одного полуинтервала $[\chi^2_{k-1;1-\alpha}; +\infty)$, где $\chi^2_{k-1;1-\alpha}$ - квантиль $\chi^2$-распределения с числом степеней свободы $k-1$ порядка $1-\alpha$.

2) Сложная нулевая гипотеза.

Мы разбиваем выборку на $k$ интервалов, оцениваем выборочные характеристики и из них вычисляем оценки вероятностей попадания в тот или иной интервал.

$n_i$ - эмпирическое число наблюдений в $i$-м интервале

$n\hat{p}_i$ - теоретическое число наблюдений в $i$-м интервале, где $n$-объем выборки, $\hat{p}_i$ - оценка вероятности попадания в $i$-ой интервал

Тогда
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i},$$
и если $H_0$ верна, то $\chi^2$ при $n \to \infty$ распределена асимптотически (т.е. стремится к распределению как $\chi^2$ с числом степеней свободы $k-r-1$, где $r$ - число параметров $F_0(x)$.

$W_{крит}$: $[\chi^2_{k-r-1;1-\alpha}; +\infty)$, где

$\chi^2_{k-r-1;1-\alpha}$ - квантиль $\chi^2$-распределения с числом степеней свободы $k-r-1$ порядка $1-\alpha$.

## Критерии Колмогорова и Смирнова.

Применяются для проверки соответствия непрерывного распределения $F(x)$ заданному $F_0(x)$.

$H_0: F(x) = F_0(x)$

$H_1: F(x) \neq F_0(x)$

Статистика Колмогорова - мера близости эмпирической ф-ии распределения $\hat{F}(x)$ и теоретической $F_0(x)$:
$$D_n = \sup_x |\hat{F}(x) - F_0(x)|.$$

Чем ближе друг к другу $\hat{F}(x)$ и $F_0(x)$, тем меньше верхняя грань их разницы.

Статистика Смирнова:

$H_0: F(x) = F_0(x)$

$H_1^+: F(x) > F_0(x)$
$$D_n^+ = \sup_x |\hat{F}(x) - F_0(x)|.$$

Критическое множество для проверки $H_0$ против $H_1$:
$$[D_{n;1-\alpha}; +\infty)$$

Для проверки $H_0$ против $H_1^+$: $[D^+_{n;1-\alpha}; +\infty)$,

где $D_{n,1-\alpha}$; $D^+_{n,1-\alpha}$ – критические значения статистик $D_n$ и $D_n^+$.
В случае сложной нулевой гипотезы

$$\hat{D}_n = \sup_x \left| \hat{F}(x) - F_0(x; \hat{\theta}_1; \dots; \hat{\theta}_r) \right|,$$

где $\hat{\theta}_1; \dots; \hat{\theta}_r$ – оценки неизвестных параметров. Если для простой нулевой гипотезы распределение статистик $D_n$ и $D_n^+$ при справедливости $H_0$ не зависит от типа $F_0(x)$, то в случае сложной нулевой гипотезы при верной $H_0$ распределение $D_n$ и $D_n^+$ уже зависит от конкретного вида распределения $F_0(x)$.

## Билет №20. Критерий знаков и ранговых знаков.

Непараметрические критерии – не требуют знания вида исходного распределения $F(x)$ за исключением предположения о непрерывности $F(x)$.
Используются, если выборка не принадлежит к нормальному распределению.
Ранг – порядковый номер наблюдения при их упорядочении по возрастанию.
Одновыборочные непараметрические критерии – для проверки гипотезы о равенстве медианы заданному значению.
Пусть есть выборка $x_1; x_2; \dots; x_n$ значений случайной величины $\xi$ с неизвестной непрерывной ф-ей распределения $F(x; Me)$, где $Me$ – неизвестная медиана.

1) Критерий знаков

$H_0: Me = Me_0$ Статистика критерия:

$H_1: Me \neq Me_0$ $n^+$ – число положительных разностей $x_i - Me_0$; $i = 1; \dots; n$.

При верной $H_0$ $P(x_i > Me_0) = P(x_i < Me_0) = 1/2$.
Статистика $n^+$ – дискретная случайная величина, распределенная по биномиальному закону,
$p = 1/2$; ... выборки

**Билет № 19.** Проверка гипотезы о равенстве заданному числу коэффициента корреляции

Пусть $(x_1; y_1); (x_2; y_2); \dots; (x_n; y_n)$ – случайная выборка пар значений двумерной случайной величины $(\xi; \eta)$, имеющей двумерное нормальное распределение.

$H_0: \rho = \rho_0 \qquad H_1: \rho \neq \rho_0$.

Коэффициент корреляции $\rho = \rho_0$ – заданное число.

Для этого можно использовать статистику:

$$u = \frac{\sqrt{n-3}}{2}\left(\ln\frac{1+r}{1-r} - \ln\frac{1+\rho_0}{1-\rho_0}\right).$$

При верной $H_0$ и при большом $n$ распределение $u$ приближается к стандартному нормальному.

$W_{крит}: (-\infty; u_{\alpha/2}] \cup [u_{1-\alpha/2}; \infty)$.

Если проверяется $H_0: \rho = 0$, то это эквивалентно проверке гипотезы о независимости $\xi$ и $\eta$.

Тогда

$$u = \frac{\sqrt{n-3}}{2}\left(\ln\frac{1+r}{1-r}\right).$$

**Билет № 23.** Классификация методов многомерного статистического анализа.

Данные, описываемые любым числом переменных – многомерные данные.

Представляются в виде матрицы, строки – наблюдения, столбцы – переменные.

Независимые переменные – факторы.

Анализ данных

↙ ↘

| Анализ зависимостей | анализ факторов |
|---|---|
| анализ взаимосвязей между переменными | анализ структуры многомерных данных |

Переменные

↙ ↘

| количественные | качественные |
|---|---|

$y$ – зависимые переменные, $x$ – независимые

Анализ факторов

количественные переменные — Факторный анализ (н-р, метод главных компонент)

качественные переменные — Кластерный анализ (агломеративно-иерархический метод — разбиение и систематизация всей совокупности наблюдений).

Анализ зависимостей

Y-количественные

- X-колич. Регрессионный анализ — поиск функциональной зависимости Y от X.
- X-качеств. Дисперсионный анализ Установление связи между X и Y.

Y-качественные

- X-колич. Дискриминантный анализ — получение правила, позволяющего на основе X предсказывать Y.
- X-качеств. Сегментационный анализ Последоват. разбиение совокупности наблюдений

**Билет №24.** Регрессионный анализ.

Пусть есть матрица наблюдений и переменных, и есть переменные $Y_1; X_1; X_2; \dots; X_m$.

$$\begin{pmatrix} y_1 & x_{11} & x_{12} & \dots & x_{1m} \\ y_2 & x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ y_n & x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}$$ — матрица экспериментальных данных

В регрессионном анализе рассматривается связь между переменной $Y$ (зависимой) и переменными $X_1; X_2; \dots; X_n$ — независимыми, эта связь описывается моделью:

$$Y = F(X_1; X_2; \dots; X_m; \beta_0; \beta_1; \dots; \beta_k) + \varepsilon,$$

где $\beta_0; \beta_1; \dots; \beta_k$ — неизвестные параметры — коэффициенты регрессии;

$\varepsilon$ — случайное отклонение $Y$ от $F$ (ошибка измерения). У каждого результата измерения $X_i$ есть своя ошибка измерения, но ошибки $X_i$ несоизмеримы с ошибками $Y$, поэтому ошибками $X_i$ мы пренебрегаем. Ортогональная регрессия — ошибки $X$ и $Y$ соизмеримы.

Виды регрессии.

1 независимая переменная – простая линейная регрессия
$$y_i = \beta_0 + \beta_i x_{i1} + \varepsilon_i.$$

Много независимых переменных – множественная линейная регрессия:
$$y_i = \beta_0 + \beta_i x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im} + \varepsilon_i.$$

$\varepsilon$ – независимые случайные величины, распределенные по нормальному закону:
$$\varepsilon_i \backsim N(0; \sigma^2).$$
Если модель по независимым переменным нелинейна, а по параметрам линейна, то она все равно остается линейной. Н-р, полиномиальная:
$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{i1}^k + \varepsilon_i.$$

Она сводится к линейной путем введения новых независимых переменных.

Экспоненциальная модель вида
$$y = a\varepsilon e^{bx}$$

нелинейна по параметрам, но может быть приведена к линейной логарифмированием.

Но
$$y = a_1 \varepsilon_1 e^{b_1 x} + a_2 \varepsilon_2 e^{b_2 x}$$

к линейной уже свести нельзя.

$y_1 = \beta_0 (1 - e^{-\beta_1 x_{i1}})^{\beta_2}$ – S-образная кривая.

Нахождение параметров регрессии.

1. Оценивание параметров
2. Нахождение доверительных интервалов для параметров
3. Проверка гипотез на эти параметры.

Т.е. мы ищем не ф-ию от параметров, а подбираем какую-либо модель и проверяем ее применимость.

Простой линейный регрессионный анализ

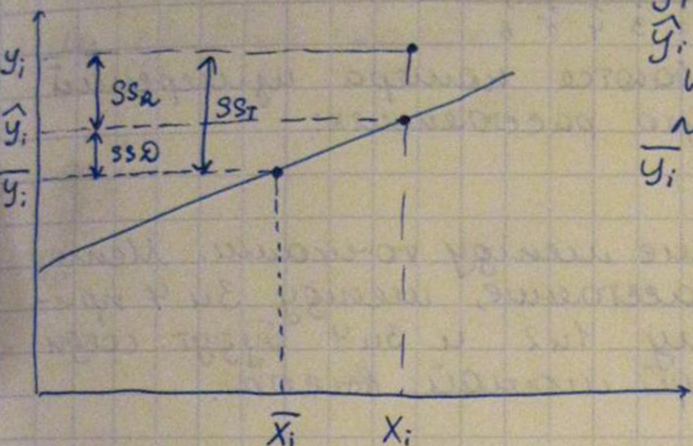
Строим двумерную диаграмму рассеяния и прямую, такую, чтобы расстояния от точек, соответствующих наблюдениям, до прямой, были наименьшими.

Получаем ур-е прямой:
$$y_i = \beta_0 + \beta_i x_i + \varepsilon_i.$$

$\beta_0$ - точка пересечения линии с $Oy$
$\beta_1$ - tg угла наклона.

Нахождение расстояний между прямой и точками (метод наименьших квадратов)

Пусть у нас не ортогональная регрессия, где расстояния меряются по перпендикуляру.




$y_i$ - наблюдение
$\hat{y}_i$ - соответствующее ему значение на прямой регрессионного анализа, зависит от параметра $\beta$
$\bar{y}_i$ - среднее наблюдение

Тогда

$$SS_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ должно быть минимальным.

Другие параметры:

$$SS_D = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

$$SS_T = SS_D + SS_R.$$

Если $SS_D = 0$, то прямая идет по линии $y = \bar{y}$. т.е. $y$ не зависит от $x$. Поэтому это не очень хорошая ситуация.

$R^2 = \frac{SS_D}{SS_T}$ - коэффициент детерминации (определяет качество предсказания).

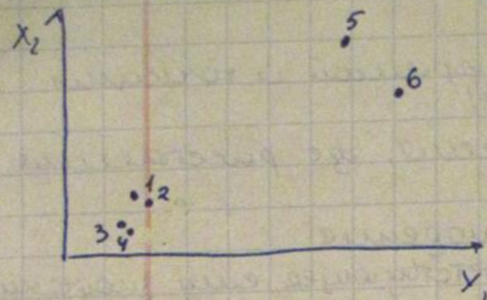
$R^2 = 0$ плохое предсказание $SS_D = 0$ $R^2 = 1$ $SS_R = 0$ хорошее предсказание.

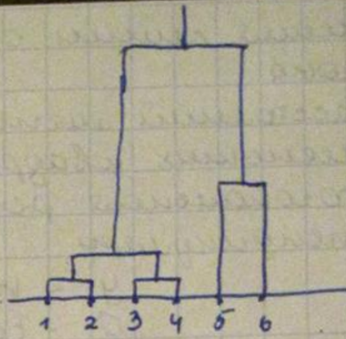
Смысл $R^2$ - какой процент общего разброса данных объяснен с помощью линейной регрессии.

## Билет №25. Кластерный анализ.

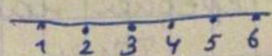
Пусть есть n переменных и некоторое число наблюдений. Кластерный анализ оценивает, какие из этих наблюдений близки друг к другу, а какие - далеки, находит взаимосвязи и классифицирует данные.

Если 2 переменных, то можно построить двумерную диаграмму рассеяния, а в многомерном случае строится дендрограмма

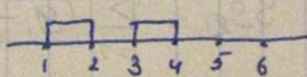


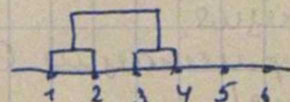

На дендрограмме указываются номера измерений на одинаковых друг от друга расстояниях:




Потом измеряется расстояние между точками. Между 1 и 2 самое маленькое расстояние, между 3 и 4 примерно такое же. Поэтому 1 и 2 и 3 и 4 будут соединяться столбиками самой малой высоты.




Затем измеряется расстояние между группами 1-2 и 3-4. Оно уже побольше, и столбик между ними будет повыше.




5 и 6 вообще далеко друг от друга - между ними высокий столбик. Ну а самый высокий будет между 1234 и 56.

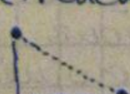
Оценка расстояния между точками:

· Евклидово расстояние

$$\rho_{1,2} = \sqrt{(x_{11}-x_{12})^2+(x_{21}-x_{22})^2}$$

Но у разных переменных разные единицы измерения, и складывать различия в возрасте и в цвете глаз как-то нелогично.

· Манхэттенское расстояние



сумма катетов прямоугольного треугольника, гипотенуза которого соединяет точки

Оценка расстояния от точки до совокупности точек

- Метод средней связи - расстояние до точек 1 и 2 считается как среднее между расстоянием до точки 1 и точки 2.
- Метод ближнего соседа - расстояние до группы точек считается как расстояние до ближайшей из этих точек

Для разделения на группы достаточным считается тройной или двойной скачок.