Программа экзамена

1. Критерии согласия (Пирсона и Колмогорова)

Критерий Пирсона (хи-квадрат) помогает оценить значимость различий между фактическим и теоретическим значением в каждой категории. Часто используется. Критерий для проверки гипотезы о принадлежности наблюдений выборки к некоторому распределению. По заданному уровню значимости и числу степеней свободы по таблице находится критическое значение, меньше которого должно быть наблюдаемое значение критерия.

Наблюдаемое и ожидаемое

- Разность наблюдаемого (observed, O) и ожидаемого (expected, E) имеет в предположении H_0 нормальное распределение со средним 0 и дисперсией E : $(O - E)/E^{1/2} \sim N(0; 1)$
- Разные O зависимы, но зависимость линейна: $\sum O = nE$
- Величина $\chi^2 = \sum \frac{(O-E)^2}{E}$ распределена как χ^2 с $n - 1$ степенью свободы

Критерий согласия хи-квадрат применим, только если ожидаемое значение в **каждой** ячейке больше 5, а общее количество объектов не менее 20.

Нередко бывает нужно установить отличие распределения наблюдений от некоторого заранее заданного распределения.

Например, H_0 может состоять в том, что распределение равномерное, а H_1 — что любое другое.

В случае дискретных распределений (и достаточно больших чисел для каждого исхода) следует использовать критерий хи-квадрат Пирсона.

Как быть в случае непрерывных распределений?

Самый простой выход: разбить область значений на несколько подобластей, для каждой рассчитать теоретическую вероятность попадания в неё, потом посчитать реальные числа попаданий и применить хи-квадрат.

Недостаток такого подхода: зависимость от разбиения.

Колмогорова:
Обозначим через Ф функцию распределения, отвечающую нулевой гипотезе (то есть $\Phi(x) = P(\xi<x)$ при $\xi \sim H_0$).
Обозначим через F эмпирическую функцию распределения нашей выборки:

$F(x) = \#(x_i < x)/n$

Статистика: $D = \max|\Phi(x) - F(x)|$

При нулевой гипотезе (независимо от исходного распределения!) и при больших n величина $n^{1/2}D$ распределена приблизительно по Колмогорову (см. https://ru.wikipedia.org/wiki/Распределение_Колмогорова), для малых n существуют таблицы
(напр., http://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/), но в наше время проще оценить p-значение вычислительным экспериментом.
Важное дополнение:

> Нельзя (по крайней мере без специальных корректировок) использовать критерий Колмогорова, если параметры того распределения, с которым вы сравниваете выборку, подбирались исходя из той же выборки

2. Таблицы сопряженности

Это способ представления совместного распределения двух переменных для исследования связи между ними. Строки - значения одной переменной, столбцы - другой.

Нулевая гипотеза — строки и столбцы независимы. Для проверки составляется «таблица ожидаемых значений».

Наблюдаемое

n_{1111}	n_{1212}
n_{2121}	n_{2222}

Ожидаемое

E_{1111}	E_{1212}
E_{2121}	E_{2222}

Ожидаемое значение на пересечении строки и столбца равно доле строки, умноженной на долю столбца и умноженной на общее количество примеров.

Статистика $\chi^2 = \Sigma_{ij} (n_{ijij} - E_{ij})^2/E_{ijij}$

Такая статистика **при нулевой гипотезе (= независимость строк от столбцов) распределена по закону «хи-квадрат с одной степенью свободы»**

3. Одновыборочный критерий Стьюдента. Условия его применимости. Парный критерий Стьюдента. Непараметрические парные критерии (знаков и Уилкоксона)

Критерий Стьюдента:

Применяется, если n невелико **и мы твёрдо уверены**, что разности $Y - X$ распределены нормально. Если такой уверенности нет, применять нельзя — используйте критерий Уилкоксона или знаков.

Статистика выглядит так же, как для Z-теста : $\boldsymbol{t = (\bar{X} - \bar{Y})/(s^2/n)^{1/2}}$ **.**

Здесь уже существенно при подсчёте s^2 делить на $(n - 1)$, а не на n.

Распределение этой статистики при малом n другое, это «распределение Стьюдента с $n - 1$ степенью свободы»

Чем больше n , тем сильнее распределение t похоже на распределение Z (то есть стандартное нормальное, N(0,1)).

Для малых n оно существенно отличается от нормального: чем больше t (и меньше n), тем сильнее вероятность получить такое или большее значение будет отличаться (в бóльшую сторону) от аналогичной вероятности для Z.

"Плотность t убывает медленнее, чем плотность Z"

"У распределения t тяжёлые хвосты"

Парный критерий знаков:

Есть **пары** наблюдений (X_i, Y_i). Считаем, что у нас есть генеральная совокупность пар чисел. Критерий знаков рассматривает не разницу значений в каждой паре, а знак этой разницы.

Нулевая гипотеза — медиана разностей $X_i - Y_i$ равна 0 (эквивалентная формулировка: **вероятность того, что $X > Y$, равна вероятности того, что $X < Y$**).

Альтернативная гипотеза: медиана $X_i - Y_i$ меньше нуля (односторонняя) или медиана не равна 0 (двусторонняя).

Статистика: число случаев, когда $X_i < Y_i$.

P-value может быть посчитано, исходя из бернуллиевского распределения. *Сводим анализ пары измерений к двум числам «успехов»*

Парный критерий Уилкоксона:

Пары наблюдений (X_i, Y_i). Для каждой пары вычисляется величина изменения признака (разница). Все разницы упорядочиваются (по модулю). Затем каждому рангу приписывают знак разницы и суммируют. Полученное значение - значения критерия Уилкоксона.

Нулевая гипотеза — разности $X_i — Y_i$ распределены симметрично относительно нуля. Альтернативная гипотеза содержательно состоит в том, что на Y действует некоторый фактор, приводящий к увеличению Y по сравнению с соответствующим X.

Точная формулировка H_1 : значения разностей $X_i — Y_i$ распределены симметрично относительно некоторого $M < 0$.

Двусторонний вариант H_1 : разности распределены симметрично относительно некоторого $M \neq 0$.

Упорядочим абсолютные значения разности $|X_i — Y_i|$:

$\{|X_1-Y_1|, \ldots, |X_l-Y_l|\} = \{Z_1, \ldots, Z_l\}$, причём $Z_1 < Z_2 < \ldots < Z_l$.

Статистика W равна сумме рангов тех пар, для которых $X < Y$.

При больших l и нулевой гипотезе величина W распределена нормально со средним $l(l+1)/4$ и дисперсией $l(l+1)(2l+1)/24$

Z-тест (только для большого числа наблюдений!):

Пары наблюдений (X_i, Y_i).

Нулевая гипотеза — **математическое ожидание** (= среднее по ген. совокупности) разностей $ER_i = EX_i - EY_i$ равно нулю, Альтернативная гипотеза: среднее меньше 0 (односторонняя) или среднее не равно 0 (двусторонняя).

Посчитаем среднее разностей по выборке: $\bar{R} = \bar{X} - \bar{Y} = \sum (X_i - Y_i)/n$

Если наблюдений много, то это среднее при нулевой гипотезе распределено нормально со средним 0

Оценим дисперсию разностей s^2 как среднее величин $(R_i - \bar{R})^2$

(лучше в качестве оценки брать не $\sum_i (R_i - \bar{R})^2/n$, а $\sum_i (R_i - \bar{R})^2/(n-1)$, хотя при больших n это не очень существенно)

Тогда дисперсия среднего при нулевой гипотезе равна s^2/n (почему?).Квадратный корень из s^2/n называется **стандартной ошибкой** SE. Статистика: $Z = \bar{R}/SE$, при нулевой гипотезе имеет распределение N(0,1)

Можно применять только при n > 100 !!!

4.Двухвыборочные критерии (Смирнова, Уилкоксона, Стьюдента, Фишера, Z-тест)

Уилкоксона:

Объединим и упорядочим выборки: $\{X_1, \ldots, X_k\} \cup \{Y_1, \ldots, Y_l\} = \{Z_1, \ldots, Z_{k+l}\}$, причём $Z_1 < Z_2 < \ldots < Z_{k+l}$.

Таким образом, номер i значения Z_i — это его ранг в объединенной выборке, а каждое Z_i равно либо какому-нибудь X_j, либо какому-нибудь Y_j.

Обозначим ранг каждого Z через $r(Z)$: $r(Z_i) = i$

Статистика Уилкоксона (Wilcoxon) R — это сумма рангов X в объединенной выборке:

$R = \Sigma_{Zi = Xj}\, i = \Sigma\, r(X_j)$

Нулевая гипотеза H_0 отклоняется, если $R < C$, где C — критическое значение (разумеется, есть и двусторонний вариант этого теста)

При больши́х k и l величина R при условии H_0 имеет нормальное распределение со средним $k(k+1)/2 + kl/2$ и дисперсией $kl(k + l + 1)/12$. При малых k или l нужно смотреть в специальные таблицы для распределения Уилкоксона (или провести вычислительный эксперимент). При совпадении части значений нужно усреднить ранги этих значений. При совпадении большого числа значений критерий неприменим.

Стьюдента (Можно применять только если данные распределены нормально!):

Оцениваем средние каждой выборки: $\bar{X} = (X_1 + \ldots + X_k)/k \quad \bar{Y} = (Y_1 + \ldots + Y_l)/l$

Оцениваем дисперсию: $s^2 = (\sum_i (X_i - \bar{X})^2 + \sum_j (Y_j - \bar{Y})^2) / (k + l - 2)$

Теперь величина $t = (\bar{X} - \bar{Y}) / (s^2/k + s^2/l)^{1/2}$ распределена по Стьюденту с $k + l - 2$ степенями свободы (*мы вынуждены предполагать равные дисперсии, иначе нельзя использовать распределение Стьюдента!*)

Z-test (**с осторожностью, т.к. вывод не точен - нужно еще поискать материал**) Вроде как совпадает с t-test.
Оцениваем средние каждой выборки: $\bar{X} = (X_1+\ldots+X_k)/k \quad \bar{Y} = (Y_1+\ldots+Y_l)/l$
Оцениваем дисперсию: $s^2 = (\sum_i (X_i - \bar{X})^2 + \sum_j (Y_j - \bar{Y})^2) / (k + l - 2)$
$Z = (\bar{X} - \bar{Y}) / (s^2/k + s^2/l)^{1/2}$

Смирнова:

Универсальный критерий для проверки **несовпадения** распределений, представленных двумя выборками размеров k и l (без предположения о характере различий: положительном сдвиге и т.п.)
Сравниваем две эмпирические функции распределения: $F_X(t) = \#(X < t)/k$ и $F_Y(t) = \#(Y < t)/l$
Статистика $D_{kl} = \max_t | F_X(t) - F_Y(t) |$

Величина $K = (kl/(k+l))^{1/2} D_{kl}$ при больших k и l (и справедливости H_0, то есть совпадении распределений), распределена по Колмогорову (критические значения: $K = 1{,}949$ для $\alpha = 0{,}001$; $K = 1{,}358$ для $\alpha = 0{,}05$).

Для малых k и l существуют таблицы, но проще оценить P вычислительным экспериментом.

Фишера:

О разности дисперсий. используется редко. Смотрим отношение двух дисперсий. Распределение Фишера исп в Анове.
Имеются две выборки: $X_1, \ldots, X_k$ и $Y_1, \ldots, Y_l$ (не обязательно одного размера)

H_0: они происходят из нормальных распределений с равными дисперсиями.
H_1: они происходят из нормальных распределений с разными дисперсиями

Статистика: $\kappa = s_X^2/s_Y^2$, где $s_X^2 = \sum_i (X_i - \bar{X})^2 / (k - 1)$, $s_Y^2 = \sum_j (Y_j - \bar{Y})^2 / (l - 1)$
— две выборочные дисперсии

Величина κ при нулевой гипотезе подчиняется так называемому F-распределению (распределению Фишера) с параметрами $k - 1$ и $l - 1$.

Поскольку $\sum_i (X_i - \bar{X})^2$ распределена по закону хи-квадрат, распределение Фишера определяется как распределение отношения двух независимых с.в., распределённых по хи-квадрат, умноженного на $(l-1)/(k-1)$

5.Дисперсионный анализ (ANOVA): постановка задачи, схема решения, условия применимости

Поиск зависимостей в экспериментальных данных за счет исследования значимости различий в средних значениях. В отличие от t-критерия можно рассматривать более 2-х групп. Разработан Фишером. Многофакторный анализ. Если есть несколько выборок разного размера (не нужно подравнивать размеры выборок). Это обобщение двустороннего t-теста, нужно нормальное распределение. Сводится к двухвыборочному критерию Стьюдента.

H0: все выборки из одного нормального распределения, H1: одно нормальное распределение с одинаковыми дисперсиями, но имеют разные средние (не знаем, кто именно отличается)

При параметрическом анализе:

1) находим k выборочных средних и общее среднее
2) оцениваем дисперсию выборок (по предположению она одинакова)
3) находим межгрупповую дисперсию
4) отношение межгрупповой дисперсии к дисперсии выборок при H0 будет иметь распределение Фишера с k-1

Анова для 2х выборок эквивалентна двустороннему тесту Стьюдента.

Непараметрический вариант (без требования нормальности)(обобщение Манна-Уитни на случай нескольких выборок) :

H0: все выборки из одного распределения, H1: из разных

Критерий Краскелла-Коллиса:

1) Считаем ранги всех наблюдений в объединенной выборке
2) Рассчитывается средний ранг по каждой выборке
3) В качестве статистики берем среднее квадратичное отклонение этих средних рангов от ожидаемого среднего = (N+1)/2

6. Проблема множественного тестирования при проверке статистических гипотез

- Тестируется новое лекарство. *Группой лечения* будем называть группу больных, которым выдают новое лекарство, а *группой контроля* — группу больных, которым его не выдают. Будем считать, что эффективность лекарства заключается в ослаблении симптомов заболевания (понижении температуры, нормализации давления, ослабления боли и т.д.). *Чем больше симптомов рассматривается, тем более вероятно, что найдётся хотя бы один симптом, который в силу случайных причин окажется достоверно слабее у «группы лечения».*
- Рассмотрим аналогичную ситуацию, но теперь будем считать лекарство безвредным, если оно не вызывает побочных эффектов. *Чем больше возможных побочных эффектов рассматривается, тем более вероятно, что найдётся хотя бы один, который будет больше проявляться у «группы лечения» в конкретном исследовании .*
- Опять аналогичная ситуация, но пусть теперь лекарство должно ослаблять симптом (например, понижать давление). Лекарство тестируется на пациентах разного возраста и пола, в разных географических зонах — пациенты разбиты на категории. *Чем больше категорий, тем более вероятно, что в одной из категорий ситуация будет выглядеть так, как если бы лекарство помогало (даже если оно реально не действует).*

показатели, вычисляемые при множественном тестировании:

- FWER — family-wise error rate, вероятность ошибки первого рода хотя бы в одном варианте
- FDR — false discovery rate, средняя доля ложных отклонений нулевой гипотезы (среди всех отклонений)
- FCR — false coverage rate, средняя доля ложных покрытий, то есть не покрытие верных параметров в пределах выбранных интервалов.

FWER:

Это фактически P-value, но вычисляемое с учётом множественного тестирования.На практике вычисляются P-value для каждого варианта (показателя, категории) отдельно, а затем к полученным числам применяется **поправка на множественное тестирование**

Варианты поправки:

- Поправка Бонферрони
- Поправка Шидака
- Поправка Холма – Бонферрони

FDR:

FDR — это средняя доля ложных отклонений нулевой гипотезы.

Метод Бенджамини — Хохберга (Benjamini–Hochberg procedure)

Пусть мы готовы «сделать ложное открытие» в 5% вариантов. Положим $\alpha = 0{,}05$.

Опять упорядочим P_i по возрастанию: $P_1 < P_2 < P_3 < \ldots < P_N$.

Найдём наибольшее k такое, что $P_k < k\alpha / N$

После этого считаем, что в вариантах 1, …, k эффект есть (имея в виду, что среди них можно ожидать 5% вариантов без эффекта)

7. Теория точечного оценивания. Состоятельность, несмещённость и эффективность оценок. Принцип максимального правдоподобия при получении оценок параметров.

Предполагается, что выборка происходит из распределения, зависящего от параметра θ (неизвестный параметр, но фиксированый).
Например:
– из экспоненциального распределения с неизвестным средним θ;
– из нормального распределения с неизвестным средним θ;
– из нормального распределения с известным средним (например, 0), но неизвестной дисперсией θ;
– …

Результатом является **оценка θ’ (случайная величина)(^/*)** — функция от наблюдений, значение которой принимается за предполагаемое значение параметра θ. В мат. статистике оценка рассматривается как функция от случайных величин, поэтому сама является случайной величиной с распределением, зависящим от θ.

- Оценка θ’ называется **состоятельной**, если θ’ сходится к θ с ростом числа наблюдений. Т.е. оценка должна сходиться к реальному числу. Важно еще и то, насколько быстро оценка сходится.
- Оценка θ’ называется **несмещенной**, если математическое ожидание оценки θ’ равно θ.
- **Эффективностью** оценки θ’ (неформально) называется то, **насколько хорошо** θ’ приближает θ. Мерой эффективности обычно служит среднее квадратичное отклонение θ’ от θ: $E(\theta' - \theta)^2$. Для несмещенной оценки эта мера равна дисперсии θ’.

*В мат. статистике оценка называется **эффективной**, если она несмещенная и ее дисперсия **при любом** θ наименьшая среди всех оценок.*

Эффективной (в любом смысле) оценкой **среднего значения генеральной совокупности** почти всегда служит **среднее по выборке**. В математической статистике среднему значению генеральной совокупности соответствует матожидание случайной величины.

Эффективной оценкой **математического ожидания** большинства распределений является **среднее по выборке.** Для выборки из нормального распределения среднее по выборке тоже распределено нормально с тем же матожиданием и с дисперсией

$D(\theta') = D/n$, где D — дисперсия исходного распределения,
n — число наблюдений.

Для выборки из другого распределения (с конечной дисперсией) среднее по выборке для больших n распределено почти нормально (распределение среднего по выборке стремится к нормальному с ростом n)

Квадратный корень из дисперсии среднего по выборке принято называть **стандартной ошибкой.** Насколько мы ошибаемся, когда используем n?

Стандартное отклонение = корень квадратный из дисперсии исходного распределения (то есть среднее квадратичное отклонение от среднего значения по генеральной совокупности). Ст. отклонение не зависит от выборки, это свойство генеральной совокупности.

Стандартная ошибка = корень квадратный из дисперсии среднего по выборке. Ст. ошибка равна ст. отклонению, деленному на квадратный корень из размера выборки

Несмещенная оценка дисперсии:

При **известном** матожидании μ несмещенной (и эффективной) оценкой дисперсии является $D' = \sum(x_i - \mu)^2/n$;

При **неизвестном** матожидании сначала оцениваем его средним по выборке: $\bar{x} = \sum x_i / n$, после этого оцениваем дисперсию как $s^2 = \sum(x_i - \bar{x})^2/(n-1)$. Если вместо $(n-1)$ поставить n, получится смещенная оценка: ее матожидание будет несколько меньше реального значения дисперсии σ^2. Это происходит потому, что x_i всегда в среднем ближе к $\bar{x}$, чем к μ. Стандартное отклонение и стандартная ошибка оцениваются из этих оценок дисперсии как квадратный корень из s^2 и квадратный корень из s^2/n, соответственно. *Эти оценки — смещенные, но достаточно эффективные, ими все пользуются.* ***Обычно именно эти оценки называются стандартным отклонением и стандартной ошибкой выборки.***

Оценка частоты:

При бернул. испытаниях тета = число успехов к числу испытаний.

Принцип максимального правдоподобия: из конкурирующих гипотез выбираем ту, что дала бы максимальную вероятность того, что мы наблюдаем. Правдоподобность: вероятность наблюдаемого при какой-либо гипотезе.

Оценка по максимальному правдоподобию:

Дискретный случай:

Пусть сначала наши наблюдения $X_1, X_2, \ldots, X_n$ происходят из дискретного распределения. Это значит, что каждое X может принимать лишь значения из какого-то дискретного множества, и каждое такое значение имеет ненулевую вероятность $P_\theta(X)$. Вероятность каждого значения зависит от θ, которого мы не знаем и хотим оценить.

Посчитаем (зависящую от θ) вероятность наших наблюдений: $P(X_1, X_2, \ldots, X_n \mid \theta) = \prod_i P_\theta(X_i)$ (правдоподобие)

Получилась функция от θ. Если мы можем найти такое $\theta = \theta'$, при котором данная вероятность достигает максимума, то это значение и будет максимально правдоподобной оценкой θ. Правдоподобием (likelihood) называется вероятность сделанных наблюдений.

8. Байесовы оценки: по максимуму апостериорной вероятности, по среднему и медиане апостериорного распределения. Связь с псевдоотсчётами.

Апостериорная вероятность: $P(\theta \mid X_1, X_2, \ldots, X_n) = P(X_1, X_2, \ldots, X_n \mid \theta) \cdot P(\theta) / P(X_1, X_2, \ldots, X_n)$

В дискретном случае единственной байесовской оценкой является MAP (maximum of a posterior probability) — в качестве оценки берётся значение θ, максимизирующее апостериорную вероятность

Если все априорные вероятности $P(\theta)$ равны между собой, то MAP-оценка превращается в ML-оценку (ML = maximum likelihood = максимальное правдоподобие)

Пример: определить, какой кубик ты взял по частоте выпадений.

9. Доверительные интервалы и их интерпретация. Получение симметричного и одностороннего доверительного интервала из точечной оценки.

по выборке $x_1, \ldots, x_n$ найти **два** числа θ' и θ'' (границы доверительного интервала) такие, что при любом значении параметра θ вероятность того, что θ попадёт внутрь интервала (θ', θ'') больше, чем $1 - \alpha$ (где α — заранее заданная маленькая вероятность, например $\alpha = 0{,}05$).

Вероятность нужно понимать правильно: в рамках стандартной мат. статистики θ — константа (то есть **неслучайное** число), а вот θ' и θ'' — случайные величины.
Стандартный способ построения доверительного интервал — так называемый **симметричный** доверительный интервал.

Начинаем с точечной оценки $\theta^* = \theta^*(x_1, \ldots, x_n)$. Это случайная величина, чьё распределение зависит от θ. Обозначим её функцию распределения через F_θ .

Теперь обозначим:
$a(\theta) = F^{-1}_\theta(\alpha/2)$
$b(\theta) = F^{-1}_\theta(1 - \alpha/2)$
Таким образом, с вероятностью $1 - \alpha$ имеем $a(\theta) < \theta^* < b(\theta)$.

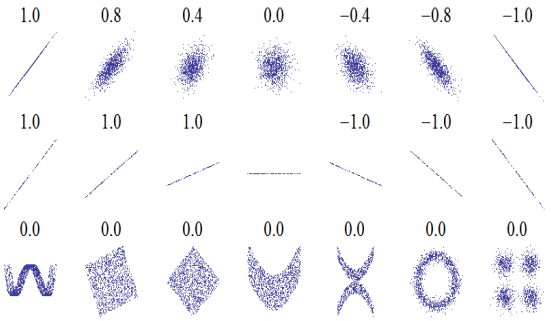
Теперь θ' и θ'' определяются из уравнений $\theta^* = b(\theta')$ и $\theta^* = a(\theta'')$.

10. Коэффициент корреляции Пирсона, его интерпретация. Связь скореллированности и независимости. Проверка гипотезы о нескореллированности. Корреляция Спирмена.

Коэф кор Пирсона = ковариация (мат. ожидание от произведения разниц значения и среднего для всех значений), разделенная на среднее геометрическое их дисперсий.
Спирмена = ранговая корреляция,т.к. вместо значений мы берем их ранги (не зависит от распределения).

Коэффициентом корреляции двух случайных величин ξ и η называется число:

$$R = \mathrm{E}((\xi - \mathrm{E}\xi)(\eta - \mathrm{E}\eta))/(\mathrm{D}\xi\mathrm{D}\eta)^{1/2}$$



R принимает значения от -1 до 1 (включительно). Если величины ξ и η независимы, то их коэффициент корреляции равен 0. Обратное, вообще говоря, неверно.

Пусть $X_1, \ldots, X_n$; $Y_1, \ldots, Y_n$ — две выборки чисел одинаковой длины. **Выборочной корреляцией (также корреляцией Пирсона, Pearson correlation)** этих выборок называется число:

$r = \sum(X_i - \bar{X})(Y_i - \bar{Y})/(\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2)^{1/2}$ (являющееся оценкой коэффициента корреляции соответствующих случайных величин).

Корреляция Спирмена (*Spearman*) — это корреляция между **рангами** измерений
(ранг измерения – это число измерений в выборке, не превосходящих по величине данное измерение).
Если среди измерений есть одинаковые, то их ранги надо усреднить.
Например, для выборки (3; 7; 1; 3; 9) надо взять в качестве рангов числа (2,5; 4; 1; 2,5; 5)

Если все измерения различные, то корреляция Спирмена легко вычисляется по формуле: $\rho = 1 - 6\sum_i(\text{rank}_X X_i - \text{rank}_Y Y_i)^2/n(n^2-1)$ (но при наличии усреднённых рангов она не годится).Важное преимущество ранговой корреляции: независимость критерия независимости от исходных распределений.

Не уверены в нормальности распределений — используйте ранги!

Существует ещё ранговая корреляция Кендалла (Kendall), но никаких преимуществ по сравнению с корреляцией Спирмена она не имеет, используют её редко.

11. Регрессия. Выбросы и влиятельные значения. Коэффициент детерминации. Разложение суммы квадратов. Статистические задачи, связанные с линейной регрессией.

Регрессия

$x_1, \ldots, x_n$ — условия
$y_1, \ldots, y_n$ — наблюдения
$y = F(x) + \varepsilon$ — **модель**
здесь $F(x)$ — некоторая функция, а ε — «ошибка измерения» (случайная величина)
(например, если $F(x) = bx + a$, то это «линейная регрессия»)
$\hat{y}_i = F(x_i)$ — предсказанные значения.
$r_i = \hat{y}_i - y_i$ — невязки.

Функция F может быть подобрана из разных соображений. Чаще всего стараются минимизировать сумму квадратов невязок.
Пусть $x_1, \ldots, x_n$; $y_1, \ldots, y_n$ — две выборки чисел одинаковой длины.

Гипотеза состоит в том, что значения y зависят от значений x линейно с точностью до ошибки измерения, которая (ошибка) нормально распределена со средним 0 :
$y = bx + a + \varepsilon$
$\varepsilon \sim N(0; \sigma^2)$
(σ неизвестно, но предполагается постоянным)
Наша задача — оценить a и b.

Для этого минимизируем сумму квадратов невязок:
$Q(a, b) = \sum(bx_i + a - y_i)^2$

решение:
$\hat{b} = \sum (x_i - \overline{x})(y_i - \overline{y}) / \sum (x_i - \overline{x})^2$

$\hat{a} = \overline{y} - \hat{b}\,\overline{x}$
При линейной регрессии $\hat{b} = r\, s_y/s_x$
где r — коэффициент корреляции между $y_1, \ldots, y_n$ и $x_1, \ldots, x_n$,
а s_y и s_x — стандартные отклонения выборок $y_1, \ldots, y_n$ и $x_1, \ldots, x_n$

Стандартная ошибка для b :
$$SE_b = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{(n-2)\sum_i (x_i - \bar{x})^2}}$$

Пусть мы определили параметры линейной регрессии по наблюдениям
$x_1, \ldots, x_n$; $y_1, \ldots, y_n$ — и пусть нам хочется предсказать значение y для какого-либо нового x.
На самом деле тут **две разные** задачи:
(1) Оценить **среднее** значение y при таком x

(2) Оценить границы доверительного интервала значений *y*

Интуитивно понятно, что точечная оценка для среднего значения *y* — это $\hat{y} = \hat{a} + \hat{b}\,x$

Но вот границы доверительного интервала для **самого** *y* и для его среднего разные. Для среднего это

$$\hat{y} \pm T^{-1}_{n-2}(1-\alpha/2)\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

$$\hat{y} \pm T^{-1}_{n-2}(1-\alpha/2)\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}+1\right)}$$

А для самого *y* :

Коэффициент детерминации

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \qquad \text{total sum of squares}$$

$$SS_{res} = \sum_i (y_i - \hat{y}_i)^2 \qquad \text{residual sum of squares}$$

$$SS_{ex} = \sum_i (\hat{y}_i - \bar{y})^2 \qquad \text{explained sum of squares}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

десь $\hat{y}_i = F(x_i)$ — предсказанные моделью значения, $\bar{y}$ — среднее значение y_i

Величина R^2 называется «коэффициент детерминации»

$R^2 = 1$ означает идеальное соответствие модели наблюдениям (все $\hat{y}_i = y_i$)

Близкое к 0 или тем более отрицательное значение R^2 означает, что модель ничего не объяснила (остаточная сумма квадратов практически равна полной сумме квадратов)

Разложение суммы квадратов

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

Если модель линейная: $\hat{y}_i = bx_i + a$

и **оба коэффициента** получены методом наименьших квадратов (!) ,

то полная сумма квадратов равна сумме объяснённой суммы квадратов и остаточной суммы квадратов.

Поэтому R^2 в этом случае равен доле объяснённой суммы квадратов в полной сумме квадратов (или, что то же самое, доле объяснённой дисперсии в полной дисперсии — дисперсия получается из суммы квадратов делением на *n*):

$R^2 = SS_{ex}/SS_{tot}$

Кроме того, R^2 в этом случае равен квадрату коэффициента корреляции между $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$.

Как следствие, в этом случае $0 \le R^2 \le 1$

Выбросы и влиятельные значения

В регрессии выбросом (outlier) называется пара x, y с большой (по модулю) невязкой.

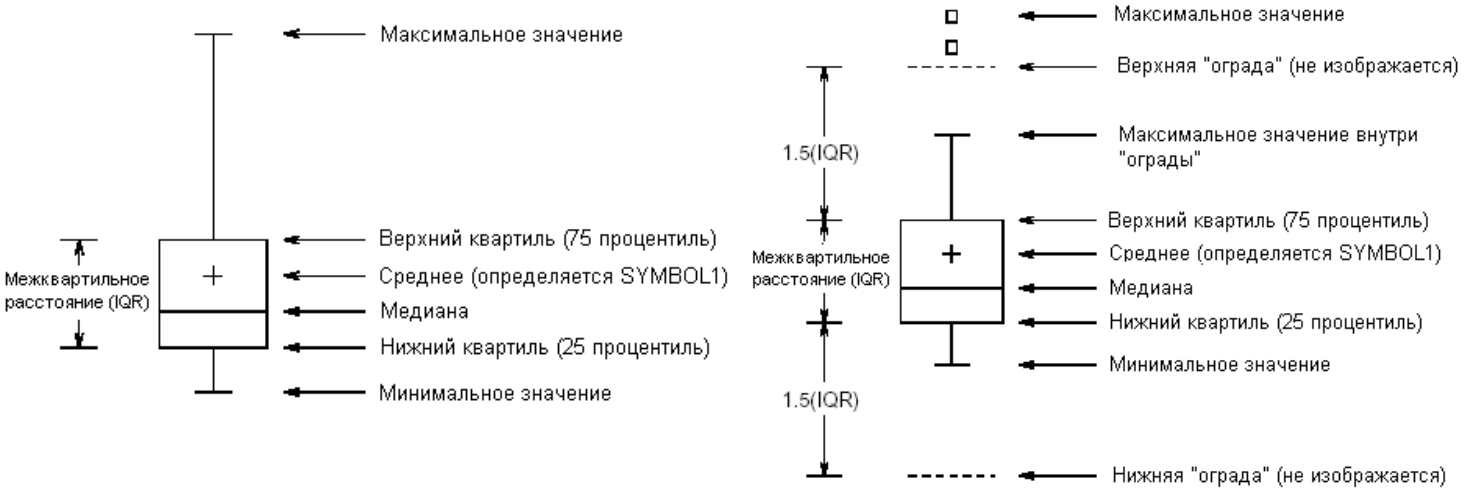
Точнее: обычно выбросами называются те пары, для которых невязки выбиваются из

нормального распределения, в то время как для большинства пар они хорошо соответствуют гипотезе нормальности.

Влиятельным (influential) значением называется пара x, y, сильно влияющая на параметры регрессии, то есть такая, после выкидывания которой параметры регрессии существенно меняются.

Дополнительные вопросы

1. Ящики с усами (box plots), какие элементы что могут означать



2. Что такое p-значение?
   “Если бы то, что мы предполагаем в нулевой гипотезе, было верно, то какова была бы вероятность видеть то, что мы видим в выборке (это, или еще «хуже»)?”
   - Чтобы определить P-value, нужна только статистика (без порога).
   - Вместо порога на статистику часто задаётся порог на P-value
   - Малые P-value показывают, что вы видите что-то очень необычное с точки зрения H_0

3. Как распределено p-значение при нулевой гипотезе?
   на отрезке 0, 1, если гипотеза принята верно, оно будет распределено нормально
4. Формула Байеса
   Если нам известна вероятность B при условии A, мы можем посчитать A при условии B.
   Формула следует из формулы условной вероятности. Нам известны обстоятельства B и мы хотим посчитать вероятность A с учетом этих обстоятельств. Например, свойства теста на наркотики.

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

- P(A) – априорная (a priori) вероятность
- P(A|B) – апостериорная (a posteriori) вероятность

5. Отношение правдоподобия
   это **отношение** вероятности получить положительный результат для положительного исхода к вероятности получить положительный результат для отрицательного исхода.
6. Стандартное отклонение и стандартная ошибка

Стандартное отклонение — это мера количества вариаций или дисперсии набора значений. Низкое стандартное отклонение указывает на то, что значения имеют тенденцию быть близкими к среднему значению

набора, в то время как высокое стандартное отклонение указывает на то, что значения разбросаны по более широкому диапазону.

Пусть $X_1, X_2, \ldots, X_n$ - выборка.

Стандартное отклонение (или среднеквадратичное отклонение выборки) оценивается по формуле:

$$SD = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}$$

где $\overline{X}$ - среднее значение выборки.

Синоним: sample standard deviation – выборочное среднеквадратичное отклонение.

7. Выбор между односторонней и двусторонней альтернативами
   Предположение, принимаемое в случае отклонения нулевой гипотезы (H_0). Как правило, альтернативная гипотеза (H_1) — это единственное утверждение, являющееся логическим отрицанием нулевой гипотезы. Часто альтернативная гипотеза означает наличие связи между изучаемыми переменными.

   Пример:

   Нулевая гипотеза (H_0): связи между признаками нет, $H_0: \ \overline{X_1} = \overline{X_2}$

   Альтернатива (H_1): связь между признаками есть.

   двусторонняя альтернатива: $H_1: \ \overline{X_1} \neq \overline{X_2}$

   односторонняя альтернатива: $\overline{X_1} < \overline{X_2}$ или $\overline{X_1} > \overline{X_2}$.

   Стандартная ошибка среднего (SEM) - теоретическое стандартное отклонение всех средних выборки размера n, извлекаемое из совокупности.

   Стандартная ошибка среднего подсчитывается следующим образом:

$$SEM = \frac{s}{\sqrt{n}},$$

   где s - стандартное отклонение, подсчитанное по выборке,

   n – число наблюдений в выборке.

8. Распределение Пуассона (формула и где возникает)
   Дискретное распределение. Самое частое. Предельный случай биномиального распределения.
   Случайная величина, распределенная по Пуассону = число достаточно редких событий за достаточно долгий промежуток времени или в достаточно большой области пространства.
   Имеет один параметр - лямбда - среднее число событий. Важно, чтобы события были независимыми. Мутации, распределение лейкоцитов в крови, полет ласточек - распределены по Пуассону.
   Дисперсия равна среднему! поэтому стандартное отклонение можно оценить как корень из среднего. Удобно оценить разброс.
   f(k,λ) = λ^k e^-λ / k! - вер-ть наблюдать k событий

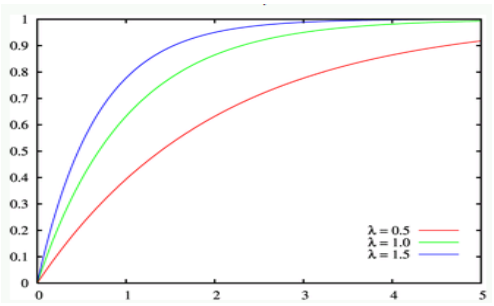
9. Экспоненциальное распределение (формула для функции распределения, формула для плотности распределения, где возникает)

Встречается в природе. Удобная формула и для плотности, и для функции.

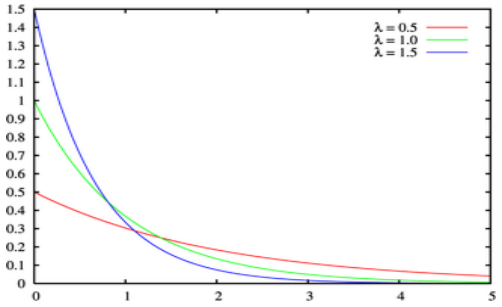
Непрерывное распределение. Каждое значение принимается с вероятностью 0 (как и у всех непрерывных). Вероятность попасть в небольшой интервал пропорциональна функции. Функция распределения - вероятность оказаться меньше этого значения. Плотность - вероятность попасть в такой интервал, деленная на длину этого интервала. Среднее равно 1/лямбда.

Кол-во независимых событий за промежуток времени распределено по Пуассону (это признак независимых событий), а интервалы событий распределены экспоненциально (тоже признак независимости событий).

Функция распределения



Плотность распределения

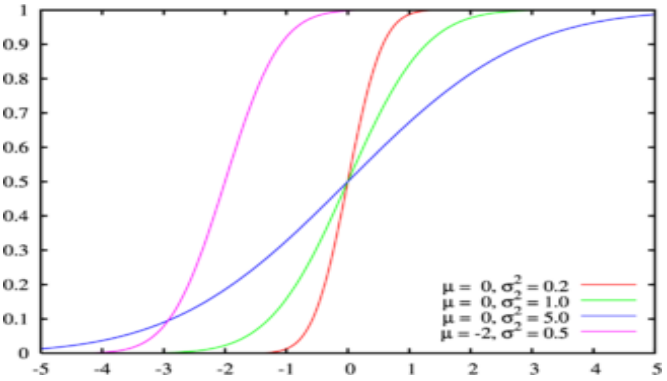


$$1 - e^{-\lambda x}$$

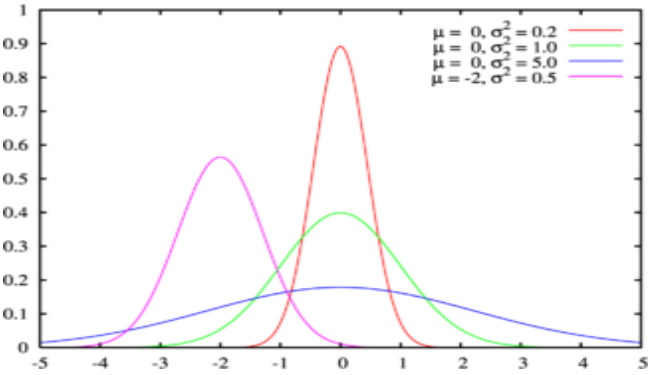
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

10. Нормальное распределение (формула для плотности распределения, в каких ситуациях возникает)

Функция распределения



Плотность распределения



Формулы нет :(

Единственно возможная формула — интеграл от плотности.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

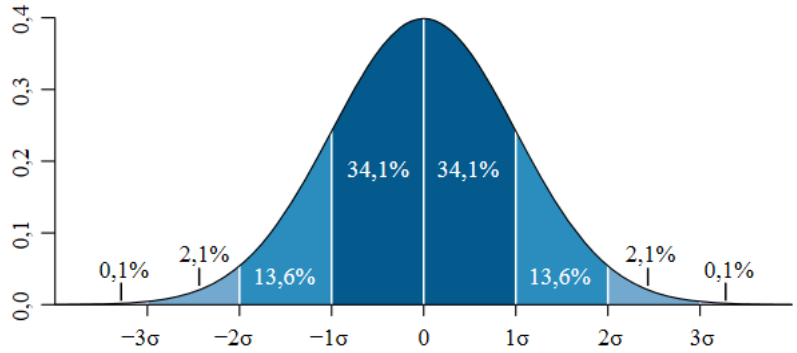
Смысл параметров: μ — мат. ожидание, σ^2 — дисперсия. σ часто называют стандартным отклонением

Нормальное распределение возникает везде, где величина представляет собой сумму большого количества элементов, вносящих приблизительно одинаковый вклад (без сильного доминирования небольшого числа из них)

Например:

- длина тела животных одной популяции, как правило, распределена нормально
- ошибки измерений в большинстве экспериментов распределены нормально

- количество крупинок в 1 кг сахарного песка распределено нормально
- число выпадений «орла» при бросании монеты 1 млн. раз распределено нормально
- и т. д.



«Правило трёх сигм»: вероятность удалиться от среднего (в заранее заданную сторону)
более чем на три стандартных отклонения – около одной тысячной

в каких ситуациях возникает, при каких значениях *x* функция распределения *F*(*x*) принимает значения 0,05 и 0,025
$F(-1{,}96) \approx 0{,}025$
$F(-1{,}65) \approx 0{,}05$

11. Перестановочные тесты

Идея: достаточно придумать хорошую статистику, а её распределение нам знать не надо!

Чтобы узнать p-value, достаточно придумать, как перемешать данные, чтобы при нулевой гипотезе ничего не изменилось, а при альтернативной — изменилось.

Тогда p-value (или как минимум его надёжная верхняя оценка) получается вычислительным экспериментом с многократным перемешиванием.

Вместо перемешивания можно моделировать выборку, исходя из нулевой гипотезы.

Перестановочные тесты – позволяют сравнивать несколько выборок между собой при помощи любых критериев.

12. Простая альтернатива и оптимальный критерий Неймана – Пирсона

Предположим, что нам известно распределение вероятностей **и для H_0, и для H_1**.

Для дискретных величин:

Пусть сначала наши наблюдения $X_1, X_2, \ldots, X_n$ происходят из дискретного распределения.

Это значит, что каждое наблюдение может принимать лишь значения из некоторого дискретного множества, и каждое такое значение X при H_0 имеет ненулевую вероятность $P_0(X)$, а при H_1 — ненулевую вероятность $P_1(X)$.

Тогда по **теореме Неймана – Пирсона** оптимальной статистикой (т.е., статистикой, порождающей самый мощный критерий при любом заданном уровне значимости) является **отношение правдоподобия** (likelihood ratio):

$L = P_1(X_1) \cdot P_1(X_2) \cdot \ldots \cdot P_1(X_n)/P_0(X_1) \cdot P_0(X_2) \cdot \ldots \cdot P_0(X_n)$

Часто, чтобы иметь дело с суммами вместо произведений, используют log *L* вместо самого *L*

Для непрерывных

Пусть теперь наблюдения $X_1, X_2, \ldots, X_n$ происходят из непрерывного распределения.

Это значит, что каждое X может принимать любое действительное значение (или любое значение из какого-то интервала: $(0,\infty)$ или (a, b)), вероятность каждого конкретного значения равно нулю

и для любого числа определена плотность вероятности в окрестности этого числа:

$p(x) = \lim_{\varepsilon \to 0} P(x - \varepsilon < X < x + \varepsilon)/2\varepsilon$

Тогда в качестве статистики используют отношение совместных плотностей вероятности при двух гипотезах:

$L = p_1(X_1) \cdot p_1(X_2) \cdot \ \dots \ \cdot p_1(X_n)/p_0(X_1) \cdot p_0(X_2) \cdot \ \dots \ \cdot p_0(X_n)$ или логарифм этой величины.

13. Генерация псевдослучайных чисел по заданной функции распределения

14. Гистограмма, график оценки плотности распределения, скрипичная диаграмма
    Скрипичная диаграмма (violin plot) — как бы гибрид ящика с усами и графика оценки плотности

15. Сведение сравнения выборок к таблице сопряженности

Сведение к таблице сопряжённости

Возьмём какой-то порог S и получим таблицу сопряжённости:

Число $X < S$	Число $X \geq S$
Число $Y < S$	Число $Y \geq S$

При нулевой гипотезе зависимости между строками нет, так что если мы её обнаружим (например, с помощью точного критерия Фишера), можно будет отклонить H_0.

Недостаток: произвол в выборе порога

(*у рецензента может возникнуть подозрение, что вы его подбирали, стараясь минимизировать P-value*)

Можно взять в качестве S, например, медиану объединённой выборки.