Глоссарий статистических терминов для химфака мгу. Версия 0.5.

Первое слово в глоссарии – если вдруг вы решили начать читать этот глоссарий с самого начала, то возможно вам нужно сначала посмотреть на слово Выборка, либо изучить игру в классики, которая находится в самом низу глоссария.

Асимметрия (коэффициент асимметрии) – Теоретический показатель. Отношение модуля третьего центрального момента к дисперсии в степени 3/2 ($(DX_1)^{3/2} = \sigma^3$). Формула $\gamma_1 = \frac{E(X_1 - EX_1)^3}{(DX_1)^{3/2}}$. Асимметрия показывает насколько случайная величина отличается от нормальной. Для нормальной случайной величины $\gamma_1 = 0$. Используя оценку асимметрии можно прикинуть является ли данная выборка нормальной или нет.

Вариационный ряд – последовательность значений заданной выборки, расположенных в порядке неубывания: Так, для выборки 1,5,2,7,5,6,9,3,1 вариационным рядом является 1, 1, 2, 3, 5, 5, 6, 7, 9.

Выборка – Выборка это набор наблюдений $X_1, \ldots, X_n$. На выборку можно смотреть с двух точек зрения теоретической: в этом случае $X_1, \ldots, X_n$ являются независимыми одинаково распределенными случайными величинами (пример: X_i – н.о.р. монетки с вероятностью выпадения единицы p, и нуля 1-p), и практической: в этом случаем мы рассматриваем $X_1, \ldots, X_n$ как набор чисел, который уже получен в результате исследований (пример: мы воздействовали на 10 мышей воздействием A, записали в табличку напротив каждой из мышей 1, если она погибла, или 0 если выжила). Замечание: элемент выборки может быть многомерное, пусть например X_1 двумерный вектор, состоящий из компонент: мышь сдохла/выжила, температура на момент наблюдения. На выборке построены понятия статистики и оценки.

Выборочная дисперсия – Оценка для теоретической дисперсии случайной величины из выборки (DX_i). Обозначается s^2. Формула может отличаться нормирующим коэффициентом в зависимости от условий на выборку. Стандартный вид $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Форма оценки, дающей ей свойство несмещенности $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Выборочная дисперсия также может быть расчитана при известном среднем $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - MX_1)^2$.

Выборочная функция распределения – это приближение теоретической функции распределения, построенное с помощью выборки из него. В то время как функция распределения F(x) есть вероятность, что значение случайной величины не превзойдет x, выборочная функция распределения $\hat{F(x)}$ есть доля (от 0 до 1) элементов выборки не превосходящих x.

Выбросы – элементы выборки, резко выделяющиеся из нее. Выброс может являться следствием ошибки измерения, неверной записи наблюдения, неоднородности исследуемых данных. Например, если наугад измерять температуру предметов в комнате, получим цифры от 18 до 22 °C, но радиатор отопления будет иметь температуру в 70°.

Выборочное среднее – Оценка для теоретического математического ожидания случайной величины из выборки (MX_i). Обозначается $\overline{X}$. Формула $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Выборочное среднее, рассмотренное как случайная величина имеет среднеквадратическое отклонение, эту величину (точнее её оценку) модуль Анализ Данных->Описательная статистика называет стандартная ошибка.

Выборочное среднеквадратическое отклонение (СКО) – Выборочный параметр, равный корню из выборочной дисперсии. Оценивает теоретическое среднеквадратическое отклонение одной случайной величины из выборки.

Доверительный интервал – Интервал значений, построенный по наблюдениям с неизвестным параметром, накрывающий этот неизвестный параметр с заданной вероятностью (так называемой доверительной вероятностью, обычно выбирающейся как 90%, 95% или 99%). Доверительный интервал с одной и той же доверительной вероятностью обычно сужается при увеличении количества наблюдений, т.е. возможно тем точнее оценить неизвестный параметр, чем большее количество наблюдений мы имеем.

Достаточная численность – оценка количества наблюдений (элементов выборки), которые требуется произвести для получения заданной точности.

Дисперсионный анализ – статистический метод, направленный на поиск различий в средних значениях различных выборок. В случае однофакторного дисперсионного анализа, мы изучаем несколько выборок полученных из одной общей совокупности изменением какого-либо фактора, и пытаемся определить имеет ли фактор влияние на среднее значение в этих выборках, при этом гипотеза H_0: средние во всех выборках равны, против альтернативы: в какой-то паре выборок средние различны. Важно отметить, что однофакторный дисперсионный анализ работает только в предположении равенства дисперсий в выборках. Либо такое предположение должно быть сделано исходя из постановки эксперимента, либо соответствующая гипотеза может быть проверена с помощью критериев Бартлетта и Кокрена.

Значимость коэффициента линейно регрессии – Оценка коэффициента регрессии может быть близкой к нулю, но ненулевой даже в том случае когда истиный коэффициент равен нулю. Для оценки значимости коэффициента следует использовать его p-значение, расчитываемое модулем анализ данных. Это p-значение соответствует проверке гипотезы о равенстве нулю коэффициента регрессии против альтернативы о неравенстве его нулю.

Изотерма Ленгмюра – зависимость вида $\frac{x}{ax+b}$. Один из видов нелинейной зависимости, применяемый как пример нелинейной регрессии. Вывод о наличии такого вида нелинейности можно сделать исходя из физического смысла. Изотерма Лэнгмюрена соответствующая процессу, при котором происходит только адсорбция на поверхности и отсутствует капиллярная конденсация. Эту особенность обнаруживают молекулярные сита всех типов, что проявляется в высокой адсорбционной емкости в области низких давлений и достижении соответствующей насыщению емкости при низких парциальных давлениях адсорбата.

Изотерма Фрейндлиха – зависимость вида ax^b (степенная зависимость). Еще один пример нелинейной зависимости. Изотерма фрейндлиха соответсвует процессу, когда присутствуют и адсорбция на поверхности и капилярная фильтрация в равновесии.

Интервальное оценивание – один из видов статистического оценивания, предполагающий построение интервала, в котором с некоторой вероятностью находится истинное значение оцениваемого параметра. Суть интервального оценивания в нахождении таких чтобы по выборке, содержащей информацию об интересующем нас параметре θ, построить доверительный интервал, содержащий θ с заданной вероятностью.

Корреляция теоретическая – величина $Corr(X,Y) = \frac{E(X-EX)(Y-EY)}{\sqrt{DX}\sqrt{DY}}$, характеризует зависимость случайных величин X и Y. Идея ковариации очень проста: если рассмотреть величину $(X-EX)(Y-EY)$, очевидно что она больше нуля в тех случаях когда X и Y одновременно большие (больше своих средних) и когда они одновременно маленькие (меньше своих средних), в остальных случаях величина $(X-EX)(Y-EY)$ меньше нуля.

Коэффициент детерминации (R^2) – коэффициент оценивающий качество аппроксимации данных регрессионной зависимостью. Формула $R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(\overline{y} - y_i)}$, где $\hat{y}_i = \hat{f}(x_i)$ – это оценка значения функции в точке x_i (в случае линеной регрессии это $\hat{a}x_i + \hat{b}$). Коэффициент детерминации показывает насколько разброс (строго говоря дисперсия) данных увеличился по сравнению с примитивным предсказанием, что Y равна константе (оценивается средним значением) + шум.

Критерий Бартлетта -– статистический критерий, позволяющий проверять равенство дисперсий нескольких (двух и более) выборок. Нулевая гипотеза предполагает, что рассматриваемые выборки получены из генеральных совокупностей, обладающих одинаковыми дисперсиями. Критерий Бартлетта является параметрическим и основан на дополнительном предположении о нормальности выборок данных. Критерий Бартлетта очень чувствителен к нарушению данного предположения.

Критерий Кокрена (G-Кокрена) – статистический критерий, позволяющий проверять равенство дисперсий нескольких (двух и более) выборок. Нулевая гипотеза предполагает, что рассматриваемые выборки получены из генеральных совокупностей, обладающих одинаковыми дисперсиями. Критерий Кокрена является параметрическим и основан на дополнительном предположении о нормальности выборок данных. Критерий Кокрена очень чувствителен к нарушению данного предположения.

Критерий Колмогорова – критерий согласия Колмогорова (также известный, как критерий согласия Колмогорова — Смирнова) используется для того, чтобы определить, подчиняются ли два эмпирических распределения одному закону, либо определить, подчиняется ли полученное распределение предполагаемой модели Критерий Колмогорова достаточно чувствителен к различиям в исследуемых выборках. Например, использования критерия возможно для проверки нормальности, если предварительно оценить среднее и дисперсию (в этом случае необходимо использовать модифицированные значения статистики).

Критерий Стьюдента (t-тест) – критерий для проверки равенства средних значений в двух выборках. Критерий Стьюдента предполагает нормальность выборок. Существуют версии критерия как предполагающие равенство дисперсий в выборках, так и не использующие такое предположение. Для применения наиболее эффективного в данной ситуации критерия стьюдента предварительно проводят тест равенства дисперсий.

Критерий Фишера (F-Критерий) – Критерий сравнения дисперсий двух выборок, проверяет гипотезу о равенстве дисперсий против альтернативы о различии дисперсий. Статистика – отношение выборочной дисперсии первой выборки и второй выборки. Среднее значение статистики при выполненинии гипотезы о равенстве дисперсий равно 1. Отклонения от 1 говорят о различии дисперсий.

Критерий χ^2 (Хи-квадрат пирсона) – критерий хи-квадрат (также известный, как критерий согласия Пирсона) используется для того, чтобы определить, подчиняется ли эмпирическое распределение выборки предполагаемой модели. В частности, возможно использование критерия для проверки гипотезы нормальности.

Мода – значение во множестве наблюдений, которое встречается наиболее часто. Так, если в классе на контрольной учениками получены оценки 5, 5, 5, 4, 4, 3, 3 то модой является оценка 5. Возможна ситуация, когда в выборке несколько мод (так, если из предыдущего примера исключить пятерки, то в оставшейся выборке как 4 так и 3 будет модой).

Медиана – значение во множестве наблюдений, которое делит упорядоченную выборку на две равные части: половина данных будут иметь значение не больше, чем медиана, а другая половина — значения не меньше, чем медиана.Так, если в классе на контрольной учениками получены оценки 5, 5, 5, 4, 4, 3, 3 то медианой является оценка 4. Возможна ситуация, когда в выборке много медиан (так, если из предыдущего примера исключить одну тройку, то в оставшейся выборке любое число между 4 и 5 будет медианой).

Непараметрическое оценивание – метод статистического оценивания, когда мы не предполагаем ничего про исследуемое распределение, таким образом всю информацию о распределении мы получаем из наблюдений. Данный метод обладает большей гибкостью, чем параметрическое оценивание, но требует бо'льших объемов выборки для получения результатов аналогичной точности (при тех же самых объемах выборки метод менее точен).

Нормальная вероятностная бумага(номральный вероятностны масштаб) – выборка нарисованная как набор точек $(X_i, \Phi^{-1}(F_n(X_i))$, где Φ – функция распределения станартного нормального распределения ($\mathcal{N}(0,1)$), F_n – выборочная функция распределения. Представленная в таком виде выборка обладает ценным свойством – в случае если выборка обладает нормальным распределением ($X_i \sim \mathcal{N}(a, \sigma^2)$), график должен быть близок к прямой. Это связано с тем фактом, что набор точек $(X_i, \Phi^{-1}(F(X_i))$ (для истиной функции распределения вместо выборочной) представляет из себя прямую. Числовым аналогом проверки нормальности при помощи нормальной вероятностной бумаги является критерий Колмогорова-Смирнова.

Однофакторный дисперсионный анализ – статистический метод, направленный на поиск различий в средних значениях нескольких выборок. В случае однофакторного дисперсионного анализа, мы изучаем несколько выборок полученных из одной общей совокупности изменением какого-либо фактора, и пытаемся определить имеет ли фактор влияние на среднее значение в этих выборках, при этом гипотеза H_0: средние во всех выборках равны, против альтернативы: в какой-то паре выборок средние различны. Важно отметить, что однофакторный дисперсионный анализ работает только в предположении равенства дисперсий в выборках. Либо такое предположение должно быть сделано исходя из постановки эксперимента, либо соответствующая гипотеза может быть проверена с помощью критериев Бартлетта и Кокрена.

Оценка (точечная) – число, вычисляемое на основе наблюдений, предположительно близкое к оцениваемому параметру.

Ошибка первого рода – ошибка в проверке гипотезы, при которой отвергается верная гипотеза. Вероятность ошибки первого рода вычисляется по формуле $P(f(X_1, \ldots, X_n) > T|H_0) = \alpha$, где $f(X_1, \ldots, X_n) > T$ – критерий отброса гипотезы H_0. В примере с 10 лабораторными мышами ошибка первого рода, это вероятность того, что при безопасном воздействии A (вероятность смерти одной мыши 0.1), совершенно случайно погибнет больше 5 мышей ($P(\sum_{i=1}^{10} X_i > 5|H_0)$). Ошибку первого рода часто называют ложной тревогой, ложным срабатыванием или ложноположительным срабатыванием — например, анализ крови показал наличие заболевания, хотя на самом деле человек здоров, или металлодетектор выдал сигнал тревоги, сработав на металлическую пряжку ремня.

Параметрическое оценивание – метод статистического оценивания, когда мы предполагаем, что наблюдаем выборку из распределения, зависящую от параметра (возможно многомерного) $theta$. Пример: мы облучаем мышей, далее смотрим погибла мышь или нет, мы считаем, что мыши гибнут независимо с неизвестной вероятностью погибнуть $\theta \in (0, 1)$. Наше параметрическое распределение – несимметриченая монетка (Бернуллиевская случайная величина), нам остается только определить вероятность гибели мыши. Другой пример: мы мерим не гибель мыши, а время её жизни до гибели. Мы предполагаем, что время смерти мыши у всех мышей имеет определенный пик, поэтому принимаем модель $X \sim \mathcal{N}(\theta, 1)$ нормального распределения для времени смерти мыши. Наша задача вновь оценить параметр θ. Замечание 1: Параметрическое оценивание это очень мощный метод, так как мы можем использовать многомерные параметры, а значит "вшивать"в модель много неизвестной информации, которую мы хотим определить. Замечание 2: кроме параметрического оценивания бывает еще непараметрическое – когда мы не предполагаем ничего про модель, параметрическое оценивание более эффективно, чем непараметрическое в случае если правильно выбрана параметрическая модель.

Парный критерий Стьюдента – критерий применяемый к парным выборкам. Пример: выборка 1 представляет из себя время, которое требуется студентам химфака чтобы применить критерий колмогорова-смирнова до занятия статпрактикума, выборка 2 представляет из себя время, которое требуется тем же самым студентам химфака, чтобы применить критерий колмогорова-смирнова после занятия. Парный

критерий стьюдента в таком случае покажет, имеются ли статистически значимые изменения (матожидание случайных величин $Z_i = X_i - Y_i$ отличается от нуля).

Проверка гипотез – Проверка статистической гипотезы производится следующим образом: есть гипотеза H_0, которая проверятся на предмет отбросить её и принять альтернативу H_1, для принятия решения об отбросе гипотезы и принятии альтернативы существует статистический критерий вида $f(X_1, \ldots, X_n) > T$ (где $f(X_1, \ldots, X_n)$ – некоторая статистика. Пример: имеется выборка из 10 мышей, на каждую мышь производится воздействие А. H_0 – воздействие безопасно (р смерти мыши равно 0.01), альтернатива H_1 – $p > 0.01$, статистика критерия $f(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$, критерий проверки гипотезы $\sum_{i=1}^{n} X_i > 6$ (если сдохло более 6 мышей отбрасываем гипотезу о безопасности воздействия). Замечание: важнейшим свойством критерия является его ошибка первого рода, на практике величину Т в критерии часто выбирают таким образом, чтобы ошибка первого рода не превосходила заданное значение α, в этом случае вместо Т пишут T_α.

Ранг – номер элемента выборки по порядку, или другими словами – номер элемента в вариационном ряду. Пример: рангами для выборки (5, 6, 8, 7) являются числа (1, 2, 4, 3)

Ранг и персентиль (Excel) – Модуль, позволяющий построить вариационный ряд и расставить ранги для элементов выборки. Имеет ряд особенностей: рассчитывает ранги в обратном порядке (самый большой элемент имеет ранг 1), кроме того считает процентное отношение номера элемента к общему количеству элементов, эти данные удобно использовать для построения выборочной функции распределения и нормальной вероятностной бумаги.

Регрессия – зависимость вида $Y = f(X) + \varepsilon$, где ε – случайная ошибка. По сути регрессия это случайная функция. В большинстве случаев, а также в случае малого количества наблюдений рассматривается линейная регрессия $Y = aX + b + \varepsilon$ как самый простой из возможных вид зависимости. Имеет место также многомерная регрессия (аналог функции множества переменных), самая простая – многомерная линейная регрессия $Y = a_1X_1 + \cdots + a_kX_k + b$. Для многомерной регрессии важна значимость коэффициентов $a_1, \cdot, a_N$. Замечание: в Excel оценка свободного члена регрессии называется Y-пересечение.

Среднее отклонение (Excel, Анализ данных) – выборочный параметр, оценивающий среднее отклонение случайной величины ($M|\xi - M\xi|$). Важно не путать этот термин со среднеквадратическим отклонением.

Среднеквадратическое отклонение (СКО, стандартное отклонение) – Теоретический параметр, равный корню из дисперсии. Называется так из-за своей формулы: $\sigma = (M(\xi - M\xi)^2)^{1/2}$. Наряду с ним можно, например, расмотреть среднекубическое отклонение $(M|\xi - M\xi|^3)^{1/3}$, или среднее отклонение $M|\xi - M\xi|$. Так как наиболее часто используемым и удобным является среднеквадратичное отклонение, оно также называется стандартным отклонением.

Статистика – Функция от выборки $f(X_1, \ldots, X_n)$, пример: $f_1(X_1, \ldots, X_n) = \sum X_i$, $f_2(X_1, \ldots, X_n) = max(x_i)$. Важнейшим свойством статистики является то, что она

зависит только от выборки, таким образом когда у вас уже есть результаты наблюдений (то есть колонка цифр) вы всегда можете вычислить статистику, для этого не требуется знать параметры модели или свойства распределений.

Эксцесс (коэффициент эксцесса) – Теоретический показатель. Отношение модуля четвертого центрального момента к квадрату дисперсии. Формула $\gamma_2 = \frac{E(X_1 - EX_1)^4}{(DX_1)^2}$. Применяемая в экселе формула немного отличается от этой, а именно $\gamma_2' = \frac{E(X_1 - EX_1)^4}{(DX_1)^2} - 3$. Эксцесс показывает насколько случайная величина отличается от нормальной. Для нормальной случайной величины $\gamma_2 = 3$ ($gamma_2' = 0$). Используя оценку эксцесса можно прикинуть является ли данная выборка нормальной или нет.

Эксцесс (Excel, Анализ данных) – выборочная характеристика, оценивающая коэффициент эксцесса. Особенностью вывода Анализа Данных в Excel является то, что выдаваемое значение эксцесса меньше настоящего значения эксцесса на 3. Это сделано для удобства сравнения выборки с нормальным распределением – истиное значение экцесса для нормальной выборки равно 3, таким образом чем ближе значение выдаваемого Excel к нулю, тем ближе выборка к нормальному распределению.

p-значение – Вероятность ошибиться, отбрасывая гипотезу. Более подробно, p-значение это величина ошибки первого рода, которая получается если в критерии отброса гипотезы выбрать T_α равной текущему значению статистики $f(x_1, \ldots, x_n)$, где $x_1, \ldots, x_n$ это набор получившихся в результате эксперимента наблюдений. Таким образом p-значение, это минимально возможная вероятность ошибки первого рода, для которой срабатывает критерий отброса гипотезы.

Игра в классики

Начало

Выборка Статистика

Параметрическое оценивание; оценка.

Среднее, Среднеквадратическое отклонения, Мода, Медиана, Асимметрия, Эксцесс.

Доверительный интервал.

Проверка гипотез, критерий.

Проверка нормальности распределения. Критерий колмогрова; Критерий Хи-Квадрат.

Нормальная вероятностная бумага.

Корреляция. Диаграмма рассеивания

Корреляционная матрица.

Значимость коэффициента корреляции.

Критерий стьюдента (t-test).

Криетрий фишера.

Однофакторный дисперсионный анализ. Критерии Бартлетта и Кокрена

Регрессия. Линейная регрессия.

Коэффициент детерминации.

Многомерная линейная регрессия; значимость коэффициентов регрессии.

Нелинейная регрессия, изотерма Фрейндлиха, изотрема Ленгмюрена.

Конец