CONTEXT
образовательно-исследовательский центр

Дистанционно-очный учебный курс
Математические основы проектирования и анализа результатов эмпирических социально-экономических исследований

Тема 4. Проверка статистических гипотез

И.Н. Дубина, доцент кафедры информационных систем в экономике АлтГУ
656049 г.Барнаул, пр. Социалистический, 68
Тел.: (385 2) 246558
din *at* econ.asu.ru cic@econ.asu.ru

Содержание

- Базовые термины и идеи
- Статистическая значимость и обоснованность
- Статистические гипотезы
- Уровни статистической значимости
- Логика и процедура проверки гипотез
- Статистические методы проверки гипотез: виды и границы применения
- Примеры использования статистических методов проверки гипотез
- Литература и ресурсы

«Статистика — это прежде всего способ мышления, и для ее применения нужно лишь иметь немного здравого смысла и знать основы математики»

Мак-Коннелл

Базовые термины и идеи

- Генеральная совокупность (population) (иногда используется калька с англоязычного термина – «популяция) – все множество объектов, в отношении которых формулируется исследовательская гипотеза
- Выборка (sample) – ограниченная по численности группа объектов (респондентов), отбираемая из генеральной совокупности для изучения ее свойств
- Сплошное и выборочное исследование
- Репрезентативность выборки (representativeness of sample) – способность выборки представлять изучаемые явления достаточно полно с точки зрения их изменчивости в генеральной совокупности
- Любое исследование направлено на определение некоторой характеристики или выявление связи между признаками
- Связь может характеризоваться не только величиной (степенью связи) и направлением, но также и надежностью или статистической достоверности (statistical confidence)
- Эта характеристика связи показывает, можно ли распространить результаты, полученные на данной *выборке,* на всю генеральную совокупность*,* из которой взята эта выборка

Базовые термины и идеи

- Статистическая гипотеза – утверждение относительно неизвестного параметра генеральной совокупности на основе выборочного исследования
- Любое заключение, полученное из статистического наблюдения / исследования / анализа, – индуктивно и строится на конечном числе наблюдений, поэтому оно не полно и может быть не достоверно
- Необходимо обоснование заключения, т.е. тестирование результатов, на которых строится гипотеза, на статистическую достоверность
- Надежность (достоверность) непосредственно связана с репрезентативностью выборки, т.е. с тем, насколько уверенно данные, полученные по выборке, позволяют судить о соответствующих параметрах генеральной совокупности
- Надежность определяется тем, насколько вероятно, что обнаруженная в выборке связь подтвердится (будет вновь обнаружена) на другой выборке той же генеральной совокупности
- Какова вероятность случайного получения результата, подтверждающего наличие связи, которой нет в генеральной совокупности?

Подход к проверке статистической гипотезы, основанный лишь на «здравом смысле»

- Формулировка гипотезы
- Измеряемые значения по выборке всегда отличаются от измеряемых значений по генеральной совокупности, поэтому нужно определить, насколько велико (значимо) это различие
- Разница в измерениях может считаться значимой, если есть основания считать, что она не представляет случайную флуктуацию
- Отклонение или констатация невозможности отклонения сформулированной гипотезы на основе имеющихся данных

- Пример 1 (как проверить монетку по правилу «орел-решка»: фальшивая или настоящая?)
- Пример 2 (производительность труда в разных возрастных группах)

Статистическая значимость и обоснованность

Пример: Проверяется гипотеза о том, что женщины тратят больше времени на разговоры по телефону, чем мужчины. Предположим, что в исследовании принимали участие 52 мужчины и 43 женщины. Среднее время разговора составило 37 мин. в день у мужчин и 41 мин. в день у женщин. На первый взгляд, различия обнаружены, и эти результаты подтверждают гипотезу.

Однако такой результат может быть получен случайно, даже если в генеральной совокупности различий нет, как и наоборот, когда различия на самом деле существуют.

Поэтому закономерен вопрос: достаточно ли полученного различия в средних значениях для того, чтобы утверждать, что вообще *все* женщины в среднем говорят по телефону дольше, чем *все* мужчины? Какова вероятность, что это не так? Является ли это различие *статистически значимым*?

Статистическая значимость и обоснованность

- Точный ответ о различиях или связях в отношении генеральной совокупности по результатам выборочного исследования получить невозможно
- Необходимо определить, достаточно ли велика разность между средними двух распределений для того, чтобы можно было объяснить ее действием независимой переменной, а не случайностью, связанной с малым объемом выборки
- Многократное проведение исследования на разных выборках трудоемко, иногда не возможно и не может обеспечить точного ответа, пока не проведено сплошное исследование

- Методы статистики позволяют оценить вероятность *случайного* получения такого различия при условии, что на самом деле различий в генеральной совокупности нет

Статистические гипотезы

- Нулевая гипотеза (null hypothesis) – гипотеза об отсутствии различий (утверждение об отсутствии различий в значениях или об отсутствии связи в генеральной совокупности)
- Согласно нулевой гипотезе (H_0), различие между значениями недостаточно значительно, а независимая переменная не оказывает никакого влияния
  Альтернативная гипотеза (alternative hypothesis) – гипотеза о значимости различий (утверждает наличие различий или существование связи)
- Альтернативная гипотеза (H_A) является «рабочей» гипотезой исследования. В соответствии с этой гипотезой, различия достаточно значимы и обусловлены влиянием независимой переменной
- Ненаправленная и направленная альтернативы

$H_0: \mu=50$

$H_A: \mu\neq50$

$H_A: \mu>50$

$H_A: \mu<50$

- Нулевая и альтернативная гипотезы представляют полную группу несовместных событий: отклонение одной влечет принятие другой
- Основной принцип метода проверки гипотез состоит в том, что выдвигается нулевая гипотеза H_0, с тем чтобы попытаться опровергнуть ее и тем самым подтвердить альтернативную гипотезу H_A. Если результаты статистического теста, используемого для анализа разницы между средними, окажутся таковы, что позволят отклонить H_0, это будет означать, что верна H_A, т.е. выдвинутая рабочая гипотеза подтверждается
- Не можем отклонить нулевую гипотезу - не значит «принять» альтернативную (нулевая гипотеза никогда не может быть абсолютно подтверждена!)

Статистические ошибки при принятии решений Ошибки первого и второго рода

- Статистическая ошибка первого рода (Type I Error) – ошибка обнаружить различия или связи, которые на самом деле не существуют

  «Истинная нулевая гипотеза отклоняется»

- Статистическая ошибка второго рода (Type II Error) - не обнаружить различия или связи, которые на самом деле существуют

  «Ложная нулевая гипотеза не может быть отклонена»

- Более «критичной» ошибкой считается статистическая ошибка первого рода

- «Судебная» аналогия: Вердикт «Не виновен» или «Виновен»

  Ошибка первого рода - невинный обвинен

  Ошибка второго рода - виновный освобожден

Уровни статистической значимости

- Тот или иной вывод с некоторой вероятностью может оказаться ошибочным, и обычно вероятность ошибки тем меньше, чем больше выборка. Таким образом, чем больше получено результатов, тем в большей степени по различиям между двумя выборками можно судить о том, что действительно имеет место в той генеральной совокупности, из которой взяты эти выборки
- Однако обычно используемые выборки относительно невелики, и в этих случаях вероятность ошибки может быть значительной
- *Уровень значимости* (level of significance) (уровень достоверности, уровень надежности, доверительный уровень, вероятностный порог) - это пороговая (критическая) вероятность ошибки, заключающейся в отклонении (не принятии) нулевой гипотезы, когда она верна. Другими словами, это допустимая (с точки зрения исследователя) вероятность совершения статистической ошибки первого рода – ошибки того, что различия сочтены существенными, а они на самом деле случайны
- Обычно используют уровни значимости (обозначаемые *α*), равные 0,05, 0,01 и 0,001
- Например, уровень значимости, равный 0,05, означает, что допускается не более чем 5%-ая вероятность ошибки. Т.е. нулевую гипотезу можно отвергнуть в пользу альтернативной гипотезы, если по результатам статистического теста вероятность ошибки, т.е. вероятность случайного возникновения обнаруженного различия (*p-уровень*) не превышает 5 из 100, т.е. имеется лишь 5 шансов из 100 ошибиться. Если же этот уровень значимости не достигается (вероятность ошибки выше 5%), считают, что разница вполне может быть случайной и поэтому нельзя отклонить нулевую гипотезу
- Таким образом, *p*-уровень значимости *(p-value)* соответствует риску совершения ошибки первого рода (отклонения истинной нулевой гипотезы). Если $p < \alpha$, нулевая гипотеза отклоняется

Уровни статистической значимости: содержательная интерпретация

- Вопрос о приемлемом значении *α*, т.е. вопрос о том, при каком уровне можно отклонить Но, не имеет однозначного ответа
- Для установленного значения *α* вероятность ошибки второго рода уменьшается с ростом объема выборки
- При увеличении значения *α* (например, с 0,01 до 0,05) вероятность ошибки второго рода уменьшается
- Значение *α* устанавливается исходя из «научных конвенций» - соглашений, принятых в научном сообществе на основе практического опыта в различных областях исследования. Традиционная интерпретация различных уровней значимости исходит из *α* = 0,05 и приведена в табл. Такое значение *α* рекомендовано для небольших выборок (когда высока вероятность ошибки второго рода). Если объемы выборок n≥100, то порог отклонения Но целесообразно снизить до *α*=0,01 и принимать решение о наличии связи (различий) при $p \leq 0{,}01$(Наследов, 2004)

Уровень значимости	Решение	Возможный статистический вывод
p>0,1	H_o не может быть отклонена	«Статистически достоверные различия не обнаружены»
p <= 0,1	сомнения в истинности H_o, неопределенность	«Различия обнаружены на уровне статистической тенденции»
p<=0,05	значимость, отклонение H_o	«Обнаружены статистически достоверные (значимые) различия»
p<=0,01	высокая значимость, отклонение H_o	«Различия обнаружены на высоком уровне статистической значимости»

Логика проверки гипотез

- Для принятия решений о том, какую из гипотез (нулевую или альтернативную) следует принять, используют *статистические критерии*, которые включают в себя методы расчета определенного показателя, на основании которого принимается решение об отклонении или принятии гипотезы, а также правила (условия) принятия решения
- Этот показатель называется *эмпирическим значением критерия*
- Это число сравнивается с известным (например, заданным таблично) эталонным числом, называемым *критическим значением* критерия.
- Критические значения приводятся, как правило, для нескольких уровней значимости: 5% (0,05), 1% (0,01) или еще более высоких
- Если полученное исследователем эмпирическое значение критерия оказывается меньше или равно критическому, то нулевая гипотеза не может быть отклонена – считается, что на заданном уровне значимости (то есть при том значении *a*, для которого рассчитано критическое значение критерия) характеристики распределений совпадают
- Если эмпирическое значение критерия оказывается строго больше критического, то нулевая гипотеза отвергается и принимается альтернативная гипотеза – характеристики распределений считаются различными с достоверностью различий $1 - \alpha$.
- Например, если $\alpha = 0{,}05$ и принята альтернативная гипотеза, то достоверность различий равна 0,95 или 95%

Логика проверки гипотез

- Если эмпирическое значение критерия для данного *числа степеней свободы (df=n-1)* оказывается ниже критического уровня, соответствующего выбранному значению *α* (порогу вероятности), то нулевая гипотеза не может считаться опровергнутой, и это означает, что выявленная разница (или связь) недостоверна
- Чем эмпирическое значение меньше критического значения критерия, тем больше степень совпадения характеристик сравниваемых объектов
- Чем эмпирическое значение критерия больше критического значения, тем сильнее различаются характеристики сравниваемых объектов
- Если эмпирическое значение критерия оказывается меньше или равно критическому, то можно сделать вывод, что характеристики экспериментальной и контрольной групп совпадают на уровне значимости *α*
- Если эмпирическое значение критерия оказывается строго больше критического, то можно сделать вывод, что достоверность различий характеристик экспериментальной и контрольной групп равна *α*

Процедура проверки статистической гипотезы

- Сформулировать нулевую и альтернативной гипотезы
- Выбрать соответствующий статистический тест
- Выбрать требуемый уровень значимости (α=0.05, 0.01, 0.025, …)
- Вычислить эмпирическое значение критерия по тесту
- Сравнить с критическим значением критерия по тесту
- Принять решение (для большинства тестов приемлемо правило: если вычисленное значение больше, чем критическое, нулевая гипотеза отклоняется)

Статистические тесты

- Для того чтобы судить о том, какова вероятность ошибиться, принимая или отвергая нулевую гипотезу, применяют статистические методы, соответствующие особенностям выборки
- Для данных, полученных в метрических шкалах (интервальных или относительных) при распределениях, близких к нормальным, используют *параметрические* методы, основанные на таких показателях, как среднее и стандартное отклонение
- В частности, для определения достоверности разницы средних для двух выборок применяют метод Стьюдента, а для того чтобы судить о различиях между тремя или большим числом выборок, — F-тест или дисперсионный анализ (ANOVA)
- Если исследователь имеет дело с данными, полученными в неметрических (номинативных или порядковых) шкалах или выборки слишком малы для уверенности в том, что ген. совокупности, из которых они взяты, подчиняются нормальному распределению, используют *непараметрические* методы — критерий *χ2* (хи-квадрат), Манна-Уитни, Уилкоксона и др. Эти методы очень просты с точки зрения как расчетов, так и применения
- Выбор статистического метода также зависит от того, являются ли выборки, средние которых сравниваются, *независимыми* (т. е., например, взятыми из двух разных групп испытуемых) или *зависимыми* (т. е. отражающими результаты одной и той же группы испытуемых до и после воздействия или после двух различных воздействий)
- В зависимости от тестируемой выборки, возможно использование свыше 100 возможных вариантов тестирования

Статистические тесты: Приложения

- Сравнение отказов компьютеров после 20-часового тестирования
- Уровень доходов разных групп населения
- Предпочтения товаров в разных демографических группах
- Сравнение числа подписчиков на журналы
- Психологические характеристики (тревожность, IQ, коммуникативность, агрессивность, …) в разных группах
- Сравнение производительности труда разных групп работников предприятия
- Общественное мнение (выборы и др.)

Пример параметрических тестов: Z-тест

Согласно одной из основных теорем статистики — *центральной предельной теореме*, распределение средних значений выборок, извлекаемых из одной и той же совокупности при достаточно большом *n* соответствует нормальному распределению. Среднее значение всех выборочных средних будет равно среднему значению совокупности (μ), а ст. отклонение выборочных средних составит величину

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

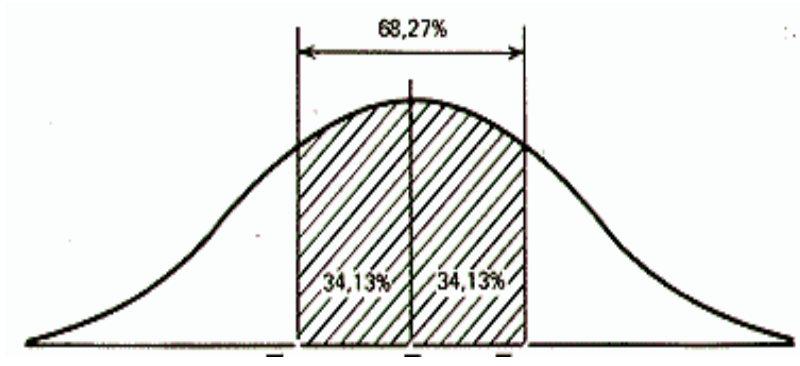
где s – стандартное отклонение выборочной совокупности

Эмпирическое значение z-критерия показывает, насколько выборочное среднее отличается от среднего ген. совокупности в единицах стандартного отклонения и определяется по формуле

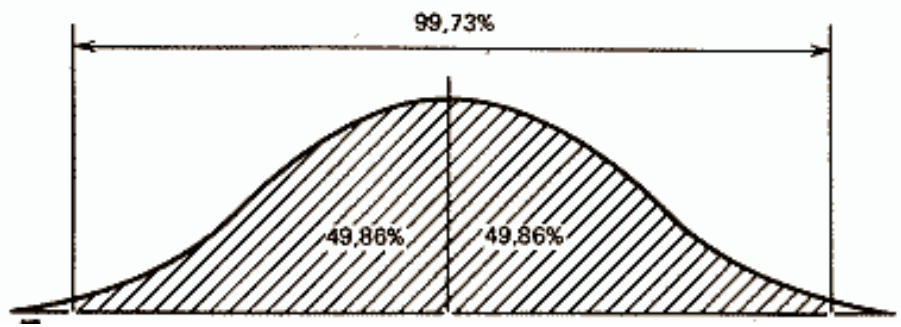
$$z = \frac{\overline{X} - m}{s_{\overline{X}}}$$

Z-тест

При нормальном распределении 68.27% результатов, располагается в пределах одного стандартного отклонения по обе стороны от среднего значения, вне зависимости от величины стандартного отклонения.



В пределах трех стандартных отклонений умещается почти вся генеральная совокупность — 99,73%.



Z-тест

- Вычислив критическое значение z, по таблице параметров нормального распределения можно определить площадь под кривой (%), которая соответствует вероятности того, что случайное отклонение разности средних выборочной и генеральной совокупности от 0 будет меньше z.
- Пусть, например, по данным выборки получено значение z=2. Тогда вероятность того, что данная выборка принадлежит генеральной совокупности со средним μ (то есть, что верна Но), составляет p=1 —0,954 = 0,046.
- Это значение соответствует *p*-уровню значимости, т.е. вероятности того, что данный выборочный результат мог быть получен случайно, когда на самом деле в генеральной совокупности верна Но
- Таким образом, при α=0.05 нулевая гипотеза отклоняется, поскольку p<α

Z	Площадь покрытия, %	Вероятность попадания в интервал, %
1.00	68.27	68.27
1.65	90.10	90.10
1.96	95.00	95.00
3.00	99.73	>99

Пример параметрических тестов: t-тест

- Для выборок меньшего объема (n<100) распределение средних соответствует другому теоретическому распределению – t-распределению Стьюдента, но общая логика проверки и формула расчета эмпирического критерия остаются теми же, что и при использовании Z-теста
- Для больших выборок Z и t-тесты обеспечивают почти идентичные результаты

H_0: μ=50 – нулевая гипотеза
H_A: μ>50 – альтернативная гипотеза
X=52.5 – среднее по выборке
s=14 – стандартное отклонение по выборке
n=100 – объем выборки
t=1.786 – эмпирическое значение критерия
α=0.05 – уровень значимости
df=99 – число степеней свободы
t_{cr}=1.66 – критическое значение критерия

$$t = \frac{\overline{X} - m}{s / \sqrt{n}}$$

Вывод: 1.786>1.66 **è** H_0 отклоняется

Пример непараметрических тестов: критерий c^2 Пирсона

- Примеры использования
  - Кто чаще обращается в службу занкомств: мужчины или женщины?
  - Кто чаще совершает аварии: мужчины или женщины?
  - Зависит ли количество аварий от дня недели?
  - Повлияла ли рекламная компания на выбор одного из двух товаров?
- Используется для номинативных шкал, но может использоваться и для шкал более высокого уровня
- Тестируется значимость различия между наблюдаемыми данными и ожидаемыми данными (основанными на H_0)
- Сравнивается наблюдаемое (эмпирическое) распределение частот (O) и ожидаемое (теоретическое) распределение (E)

$$c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

df=(k-1)(l-1) – число степеней свободы (необходимо для определения критического значения критерия при заданном α)

k – число градаций

l – количество сопоставляемых распределений

Пример использования критерия c^2

- Опрос 200 студентов о желании посещать «Ланч-клуб»
- Анализ результатов в зависимости от условий проживания
- Ho: Желание посещать «Ланч-клуб» не зависит от условий проживания, т.е наблюдаемое распределение (observed) частот соответствует ожидаемому (expected) распределению (O=E)

	Намерены посещать (O)	Опрошено всего	Ожидаемое (E)
Общежитие	16	90	27
Квартира в центре	13	40	12
Квартира на окраине	16	40	12
За городом	15	30	9
	60	200	60

χ^2=9.89 – эмпирическое значение критерия

df= (4-1)(2-1)=3 α=0.05 χ^2_{cr}=7.82 – критическое значение критерия

Вывод: H_o отклоняется

Литература и ресурсы

- Cooper, D.R., Shindler, P.S. (1995) *Business Research Methods*. Irwin/McGraw-Hill.
- Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. Учебное пособие. СПб.: Речь, 2004.
- Статистика и обработка данных в психологии. Электронный ресурс http://psyfactor.org/lib/
- В.А. Дюк. Конструирование психодиагностических тестов: традиционные математические модели и алгоритмы. Электронный ресурс http://psyfactor.org/lib/
- Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. М.: ИНФРА-М, 1998.

- StatSoft Russia. www.statsoft.ru