

ВШБ Бизнес-информатика: ТВиМС 2025.
 Лист задач для самостоятельного решения #12.
 Проверка статистических гипотез.

Основные формулы

Распределения статистик

- Выборочное среднее, дисперсия известна:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- Выборочное среднее, дисперсия неизвестна:

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

- Выборочная доля:

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

- Разность долей:

$$D = (\hat{p}_1 - \hat{p}_2) \sim \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}\right)$$

- Разность средних, дисперсии известны:

$$D = (\bar{X} - \bar{Y}) \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$$

- Разность средних, дисперсии неизвестны, но предполагаются равными:

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

где $S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$ — объединённая выборочная дисперсия.

- Разность средних, дисперсии неравны (тест Уэлча):

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim t_k, \text{ где степень свободы равна } k \approx \frac{(V_X + V_Y)^2}{\frac{V_X^2}{n-1} + \frac{V_Y^2}{m-1}}, \quad V_X = \frac{S_X^2}{n}, \quad V_Y = \frac{S_Y^2}{m}$$

Формулы для score (статистик)

- Выборочное среднее, дисперсия известна:

$$z_{\text{score}} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

- Выборочное среднее, дисперсия неизвестна:

$$t_{\text{score}} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- Выборочная доля:

$$z_{\text{score}} = \frac{\tilde{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- Разность долей:

$$z_{\text{score}} = \frac{\tilde{p}_1 - \tilde{p}_2}{\sqrt{p_c(1-p_c) \left(\frac{1}{n} + \frac{1}{m}\right)}},$$

где $p_c = \frac{\tilde{p}_1 n + \tilde{p}_2 m}{n+m}$ — объединённая доля.

- Разность средних, дисперсии известны:

$$z_{\text{score}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

- Разность средних, дисперсии неизвестны, но предполагаются равными:

$$t_{\text{score}} = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

- Разность средних, дисперсии неизвестны:

$$t_{\text{score}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

- Пусть X — случайная величина, показывающая реальное количество кофе в банке "100 г кофе" $E[X] = \mu$. При выборке размера 16 вы хотите проверить нулевую гипотезу $H_0 : \mu = 100$ против альтернативы $H_1 : \mu > 100$ на уровне значимости 5%. Пусть $X \sim \mathcal{N}(\mu, \sigma^2)$ и дисперсия известна: $\sigma^2 = 1$. Найдите критическую (отклоняющую) область для этого теста в оригинальной шкале, т.е. при каком выборочном среднем содержания кофе мы начнем отклонять нулевую гипотезу. У вас нет фактических экспериментальных данных для выполнения теста, вам нужно только определить границу критической области.
- Исследователь проводил односторонний тест, но вместо использования (как должно было быть) верхней 5% критической точки стандартного нормального распределения, он использовал верхнюю 5% точку распределения Стьюдента с 6 степенями свободы. Каков истинный уровень значимости этого теста?
- Случайная выборка из десяти студентов показала следующие значения времени (в часах), потраченного на изучение в неделю перед финальными экзаменами:

28 57 42 35 61 39 55 46 49 38.

Предположим, что распределение генеральной совокупности нормальное.

- (a) Найдите выборочное среднее и выборочное стандартное отклонение.
- (b) Проверьте гипотезу о том, что среднее генеральной совокупности равно 40, против альтернативы, что оно больше.

Можно пользоваться следующим разбиением на подзадачи, чтобы лучше разобраться в теме:

- (a) Опишите распределение статистики \bar{X} в этой задаче. Известна ли вам дисперсия генеральной совокупности здесь?
 - (b) Сформулируйте нулевую и альтернативную гипотезы.
 - (c) Опишите все случайные величины, необходимые для процедуры проверки, их свойства и распределения при нулевой гипотезе, когда мы уверены, что H_0 полностью верна.
 - (d) Найдите границу критической области в оригинальной шкале и выполните проверку, используя $\alpha = 1\%, 5\%, 10\%$.
 - (e) Выполните проверку с помощью правильного *score*, т.е. путём преобразования \bar{x} в шкалу выбранного распределения.
 - (f) Покажите, что результаты проверки в двух подходах идентичны.
- Если вы живёте в Калифорнии, решение о покупке страховки от землетрясений является критически важным. Статья в научном журнале от июня 1992 года исследовала множество факторов, которые жители Калифорнии учитывают при покупке страховки от землетрясений. Опрос показал, что только 133 из 337 случайно выбранных домохозяйств в округе Лос-Анджелес были защищены страховкой от землетрясений.
 - (a) Каковы подходящие нулевая и альтернативная гипотезы для проверки утверждения, что менее 40% жителей округа Лос-Анджелес были защищены страховкой от землетрясений?
 - (b) Предоставляют ли данные достаточные доказательства в поддержку нулевой гипотезы? (Используйте $\alpha = 0.10$)
 - Американская ассоциация больниц сообщает в Hospital Statistics, что средняя стоимость для общих общественных больниц на одного пациента в день в больницах США составляла \$951 в 1998 году. В том же году случайная выборка из 30 дневных затрат в больницах Нью-Йорка дала среднее значение \$1185. Предполагая стандартное отклонение генеральной совокупности \$333 для больниц Нью-Йорка, предполагают ли данные достаточные доказательства для заключения, что в 1998 году средняя стоимость в больницах Нью-Йорка превышала национальное среднее \$951? Выполните требуемую проверку гипотез на уровне значимости 5%.

6. Во время ночной смены в пятницу из производственной линии случайным образом было отобрано $n = 28$ мятных конфет и взвешено. Они имели средний вес $\bar{x} = 21.45$ граммов. Известно, что стандартное отклонение веса конфеты равно $\sigma = 0.31$ грамма.

- (a) Проверьте нулевую гипотезу $\mu = 20$ против альтернативы $\mu > 20$ на уровне значимости 5%.
- (b) Какое в данном случае получилось P -значение?
- (c) Предположим, что вдруг *на самом деле* $\mu = 22$ (то есть верна альтернативная гипотеза). Также пусть K - граница критической области в оригинальной шкале из предыдущего пункта. Напишите, как будут выглядеть необходимые распределения статистик в таком случае (то есть распределения при $\mu = 22$). Найдите вероятность:

$$P_{H_1}(\bar{X} < K) = \beta$$

Это вероятность того, что мы не отклоним нулевую гипотезу, когда она на самом деле неверна. Это и есть вероятность ошибки II рода. Говоря более подробно, это вероятность такого случая, когда при верной альтернативной гипотезе мы волей случая получаем неубедительные данные, и, как следствие, не отклоняем нулевую гипотезу.

7. Были проведены два опроса в Москве и Твери. Из выборки 200 человек в Москве 125 были против курения в ресторанах. В Твери 52 из выборки 100 были против курения в ресторанах. Пусть p_1 и p_2 — доли генеральных совокупностей людей, которые против курения в Твери и Москве соответственно.

- (a) Постройте 95% доверительный интервал для разности долей $p_1 - p_2$.
- (b) На уровне значимости 5% проверьте нулевую гипотезу $H_0 : p_1 = p_2$ против $H_1 : p_2 > p_1$.
- (c) На уровне значимости 2.5% проверьте нулевую гипотезу $H_0 : p_2 = 0.55$ против $H_1 : p_2 > 0.55$.

8. Супермаркет провёл исследование, чтобы выяснить, одинаковы ли средние недельные продажи шоколадных батончиков при использовании обычного расположения на полке и при использовании витрины в конце прохода (дисплейная выкладка). Сводка данных представлена в таблице:

	Размер выборки	Выборочное среднее	Выборочная дисперсия
Обычное расположение на полке	11	5.3	2.4
Витрина в конце прохода	10	7.2	2.8

Предполагая, что недельные продажи распределены нормально, аналитический отдел хочет определить, действительно ли средние недельные продажи при использовании витрины в конце прохода выше, чем при обычном расположении товаров на полке.

Можно пользоваться следующим разбиением на подзадачи, чтобы лучше разобраться в теме:

- (a) Пусть первая выборка — это $\{X_1, \dots, X_{11}\}$, а вторая — $\{Y_1, \dots, Y_{10}\}$. Опишите распределения статистик \bar{X} , \bar{Y} и $D = \bar{X} - \bar{Y}$ в этой задаче.
- (b) Сформулируйте нулевую и альтернативную гипотезы.
- (c) Опишите все случайные величины, необходимые для процесса тестирования, их свойства и распределения при нулевой гипотезе, когда мы уверены, что H_0 полностью верна.
- (d) Выполните проверку с помощью правильного *score* теста, *t.e.* путём преобразования $\bar{x} - \bar{y}$ в шкалу выбранного распределения. Используйте уровни значимости $\alpha = 1\%, 5\%, 10\%$.
- (e) В задачах на разность параметров это может быть скорее в качестве доп. пункта для проверки себя. Быстрее решать конечно же с помощью *score*.

Найдите границу критической области в оригинальной шкале и выполните проверку, используя предыдущие уровни значимости. Покажите, что результаты проверки в двух подходах идентичны.

9. Данные в следующей таблице показывают количество ежедневных нарушений парковки в двух районах города. Идентификация дней неизвестна, и записи не обязательно были сделаны в одни и те же дни. Есть ли доказательства того, что районы имеют разные средние количества нарушений? Укажите необходимые предположения и выполните проверку гипотез на уровнях значимости $\alpha = 1\%, 5\%, 10\%$.

Район А	Район В
38	32
38	38
29	22
45	30
42	34
33	28
27	32
32	34
32	24
34	нет данных

10. *Насколько помогают ремни безопасности?* Чтобы ответить на этот вопрос, было проведено исследование автомобилей, которые были оборудованы ремнями безопасности (поясные и плечевые ремни) и впоследствии попали в аварии. Случайная выборка из 10,000 пассажиров показала следующие показатели травматизма (восстановлено из U.S. Department of Transportation, 1981):

Тяжёлая или смертельная травма	Ремень безопасности надет		Всего
	Да	Нет	
Да	3	119	122
Нет	829	9049	9878
Всего	832	9168	10000

- (a) Как бы вы проверили положительный эффект использования ремней безопасности с помощью теста на разность долей? Сформулируйте H_0 словами и формализованно.
- (b) Выполните проверку гипотез при различных уровнях значимости $\alpha = 0.1, 0.05, 0.01$. Какое получилось P -значение в вашем выбранном тесте? Прокомментируйте результаты.
- (c) Постройте соответствующий доверительный интервал. Исследуйте поведение границ при различных уровнях доверия: 90%, 95%, 99%. Какие выводы можно сделать?