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:

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

- 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:

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

1. 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.

   Решение:

   - Распределение при нулевой гипотезе, когда мы абсолютно уверены в том, что она верна:

   $$\bar{X} \sim \mathcal{N}\left(100, \tfrac{1}{16}\right)$$

   - Построение выражения для нахождения критической границы. Строим с помощью зафиксированной ошибки первого рода.

     Идея: Предположим, что базовая гипотеза верна. Но из-за того, что выборочное среднее - случайная величина, то чисто случайно нам может достаться партия банок с кофе, где у всех случайный перевес, и это нас конечно убедит в альтернативной гипотезе - но это ошибка первого рода. С другой стороны, у нас есть уровень значимости α, это шанс ошибки первого рода, который нас "психологически"устраивает. Мы уравниваем две вероятности: вероятность получить какие-то чрезмерно высокие значения среднего при верной базовой гипотезе (шанс случайного экстремального "выброса") и уровень ошибки первого рода.

   $$P_{H_0}\left(\bar{X} > K\right) = \alpha$$
   $$P_{H_0}(\bar{X} > K) = P_{H_0}\left(\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{K-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z > z_\alpha\right)$$
   $$K = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

   - Критическая точка стандартного нормального распределения: $z_{0.05} = 1.645$
   - Граница критической области: $K = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} = 100 + 1.645 \cdot \frac{1}{\sqrt{16}} = 100 + 0.411 = 100.411$
   - Правило принятия решения: Отклонить H_0, если $\bar{x} > 100.411$

2. 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?

   Решение:

   - Должен был использовать: $z_{0.05} = 1.645$ (стандартное нормальное распределение)
   - Использовал: $t_{(6,0.05)} \approx 1.943$ (распределение Стьюдента с 6 степенями свободы)
   - Истинный уровень значимости: $\alpha_{real} = P(Z > 1.943) \approx 0.026$ (около 2.6%)

3. Случайная выборка из десяти студентов показала следующие значения времени (в часах), потраченного на изучение в неделю перед финальными экзаменами:

$$28 \quad 57 \quad 42 \quad 35 \quad 61 \quad 39 \quad 55 \quad 46 \quad 49 \quad 38.$$

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*, *m.e.* путём преобразования $\bar{x}$ в шкалу выбранного распределения.

(f) Покажите, что результаты проверки в двух подходах идентичны.

Решение:

(a) Выборочное среднее: $\bar{x} = \frac{28+57+42+35+61+39+55+46+49+38}{10} = \frac{450}{10} = 45$ часов

Выборочная дисперсия: $s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - 45)^2 = \frac{1}{9}[(28-45)^2 + \ldots + (38-45)^2] \approx 111.1$

Выборочное стандартное отклонение: $s \approx \sqrt{111.1} \approx 10.54$ часов

(b)
- Исследуемая случайная величина X - время, затрачиваемое студентом на подготовку в экзаменационную неделю. Истинная дисперсия X неизвестна. Так как известно, что у X нормальное распределение, то по свойству устойчивости можем сказать, что для случайной выборки размера 10 выполняется:
$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{10}\right)$$
- Гипотезы: $H_0 : \mu = 40$ против $H_1 : \mu > 40$ (правосторонний тест).
- Распределения при нулевой гипотезе, когда мы абсолютно уверены в том, что она верна:
$$\bar{X} \sim \mathcal{N}\left(40, \tfrac{\sigma^2}{10}\right), \quad \text{а также} \quad \tfrac{\bar{X}-40}{S/\sqrt{10}} \sim t_9$$
- Построение границы критической области теста относительно установленного уровня значимости:
$$P_{H_0}\left(\bar{X} > K\right) = \alpha$$
$$P_{H_0}(\bar{X} > K) = P_{H_0}\left(\tfrac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}} > \tfrac{K-\mu_0}{\frac{S}{\sqrt{n}}}\right) = P\left(t > t_{(n-1,\alpha)}\right)$$
$$K = \mu_0 + t_{(n-1,\alpha)}\tfrac{S}{\sqrt{n}}$$
- Критические точки t-распределения с 9 степенями свободы: $t_{(9,0.01)} \approx 2.821$, $t_{(9,0.05)} \approx 1.833$, $t_{(9,0.10)} \approx 1.383$
Границы критической области:
$$K_{1\%} = 40 + 2.821 \cdot \frac{10.54}{\sqrt{10}} = 40 + 2.821 \cdot 3.33 \approx 40 + 9.40 = 49.40$$
$$K_{5\%} = 40 + 1.833 \cdot \frac{10.54}{\sqrt{10}} = 40 + 1.833 \cdot 3.33 \approx 40 + 6.10 = 46.10$$
$$K_{10\%} = 40 + 1.383 \cdot \frac{10.54}{\sqrt{10}} = 40 + 1.383 \cdot 3.33 \approx 40 + 4.61 = 44.61$$

- При $\alpha = 1\%$: $\bar{x} = 45 < K_{1\%} = 49.40$, не отклоняем H_0
- При $\alpha = 5\%$: $\bar{x} = 45 < K_{5\%} = 46.10$, не отклоняем H_0
- При $\alpha = 10\%$: $\bar{x} = 45 > K_{10\%} = 44.61$, отклоняем H_0

- Проверка с помощью t-статистики (score):
  t-статистика: $t_{\text{score}} = \frac{45-40}{10.54/\sqrt{10}} \approx 1.50$
  Сравнение с критическими точками:
  - При $\alpha = 1\%$: $t_{\text{score}} = 1.50 < t_{(9,0.01)} = 2.821$, не отклоняем H_0
  - При $\alpha = 5\%$: $t_{\text{score}} = 1.50 < t_{(9,0.05)} = 1.833$, не отклоняем H_0
  - При $\alpha = 10\%$: $t_{\text{score}} = 1.50 > t_{(9,0.10)} = 1.383$, отклоняем H_0

4. 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 $\alpha = 0.10$)

   Решение:

   (a) Гипотезы: $H_0 : p = 0.40$ против $H_1 : p < 0.40$ (левосторонний тест)

   (b) Данные: $n = 337$, $\tilde{p} = \frac{133}{337} \approx 0.395$, $p_0 = 0.40$, $\alpha = 0.10$

   Распределение при нулевой гипотезе: $\hat{p} \sim \mathcal{N}\left(0.40, \frac{0.40 \cdot 0.60}{337}\right)$

   Построение границы критической области теста относительно установленного уровня значимости:

$$P_{H_0}(\hat{p} < K) = \alpha$$

$$P_{H_0}(\hat{p} < K) = P_{H_0}\left(\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < \frac{K - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right) = P\left(Z < -z_\alpha\right)$$

$$K = p_0 - z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$$

   (c) Критическая точка: $z_{0.10} = 1.282$

   Граница критической области: $K = 0.40 - 1.282 \cdot \sqrt{\frac{0.40 \cdot 0.60}{337}} \approx 0.366$

   Проверка: $\tilde{p} = 0.395 > K = 0.366$, поэтому не отклоняем H_0

   (d) Проверка с помощью z-статистики: $z_{\text{score}} = \frac{0.395 - 0.40}{0.0267} \approx -0.187 > -z_{0.10} = -1.282$, не отклоняем H_0

   Вывод: Данные не предоставляют достаточных доказательств для утверждения, что менее 40% жителей защищены страховкой.

5. 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, предоставляют 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%.

   Решение:

   - Исследуемая случайная величина X - дневная стоимость для одного пациента в больницах Нью-Йорка. Истинная дисперсия X известна: $\sigma = 333$. Так как размер выборки $n = 30$, по ЦПТ можем сказать, что для случайной выборки размера 30 выполняется:

   $$\bar{X} \sim \mathcal{N}\left(\mu, \frac{333^2}{30}\right)$$

   - Гипотезы: $H_0 : \mu = 951$ против $H_1 : \mu > 951$ (правосторонний тест)
   - Распределение при нулевой гипотезе, когда мы абсолютно уверены в том, что она верна:

   $$\bar{X} \sim \mathcal{N}\left(951, \frac{333^2}{30}\right)$$

   - Построение границы критической области теста относительно установленного уровня значимости:

   $$P_{H_0}\left(\bar{X} > K\right) = \alpha$$
   $$P_{H_0}(\bar{X} > K) = P_{H_0}\left(\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{K-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z > z_\alpha\right)$$
   $$K = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

   - Критическая точка стандартного нормального распределения: $z_{0.05} = 1.645$
     Граница критической области: $K = 951 + 1.645 \cdot \frac{333}{\sqrt{30}} = 1051$
     Проверка: $\bar{x} = 1185 > K = 1051$, поэтому отклоняем H_0
   - Проверка с помощью z-статистики (score):
     z-статистика: $z_{\text{score}} = \frac{1185-951}{333/\sqrt{30}} \approx 3.85$
     Сравнение с критической точкой: $z_{\text{score}} = 3.85 > z_{0.05} = 1.645$, отклоняем H_0
   - Вывод: Данные предоставляют достаточные доказательства для заключения, что средняя стоимость в больницах Нью-Йорка превышала национальное среднее.

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.

Решение:

(a)
- Исследуемая случайная величина X - вес мятной конфеты. Истинная дисперсия X известна: $\sigma = 0.31$ грамма. Немного аппроксимируем, и говорим, что размер выборки достаточно большой, чтобы использовать ЦПТ. Тогда для случайной выборки размера 28 выполняется:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{0.31^2}{28}\right)$$

- Гипотезы: $H_0 : \mu = 20$ против $H_1 : \mu > 20$ (правосторонний тест)
- Распределение при нулевой гипотезе, когда мы абсолютно уверены в том, что она верна:

$$\bar{X} \sim \mathcal{N}\left(20, \frac{0.31^2}{28}\right)$$

- Построение границы критической области теста:

$$P_{H_0}\left(\bar{X} > K\right) = \alpha$$

$$P_{H_0}(\bar{X} > K) = P_{H_0}\left(\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{K-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z > z_\alpha\right)$$

$$K = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

- Критическая точка стандартного нормального распределения: $z_{0.05} = 1.645$
  Граница критической области: $K = 20 + 1.645 \cdot \frac{0.31}{\sqrt{28}} \approx 20.096$ граммов
  Проверка: $\bar{x} = 21.45 > K = 20.096$, поэтому отклоняем H_0
- Проверка с помощью z-статистики (score):
  z-статистика: $z_{\text{score}} = \frac{21.45-20}{0.31/\sqrt{28}} \approx 24.75$
  Сравнение с критической точкой: $z_{\text{score}} = 24.75 > z_{0.05} = 1.645$, отклоняем H_0

(b) P-значение: $p\text{-value} = P_{H_0}\left(\bar{X} > \bar{x}\right) = P\left(Z > \frac{21.45-20}{0.31/\sqrt{28}}\right) = P(Z > 24.75) \approx 0$ (практически равно нулю)

(c) При $\mu = 22$ (альтернативная гипотеза верна):
- Распределение при альтернативной гипотезе:

$$\bar{X} \sim \mathcal{N}\left(22, \frac{0.31^2}{28}\right)$$

- Вероятность ошибки II рода. Идея: Пусть верна альтернативная гипотеза. Из-за того, что выборочное среднее - случайная величина, то чисто случайно нам может достаться выборка конфет, где средний вес меньше границы критической области, найденной ранее. Это нас конечно послужит аргументом не отклонять нулевую гипотезу. Но так мы совершаем ошибку второго рода.

$$\beta = P_{H_1}(\bar{X} < K) = P\left(\frac{\bar{X}-22}{0.31/\sqrt{28}} < \frac{20.096 - 22}{0.31/\sqrt{28}}\right) = P(Z < -32.50) \approx 0$$

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 $H_1 : p_2 > 0.55$.

Решение:

(a) Данные: $n_1 = 100$ (Тверь), $\tilde{p}_1 = \frac{52}{100} = 0.52$; $n_2 = 200$ (Москва), $\tilde{p}_2 = \frac{125}{200} = 0.625$
Доверительный интервал для $p_1 - p_2$:

$$(p_1 - p_2) \in \left(\tilde{p}_1 - \tilde{p}_2 - z_{0.025}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}}, \tilde{p}_1 - \tilde{p}_2 + z_{0.025}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}} \right)$$

$z_{0.025} = 1.96$, подкоренное выражение: $\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}} \approx 0.06$
Интервал: $(0.52 - 0.625) \pm 1.96 \cdot 0.06 = (-0.2237, 0.0137)$

(b) Исследуемые случайные величины: X — индикатор отношения против курения в ресторанах для жителя Твери, Y — индикатор отношения против курения в ресторанах для жителя Москвы. Обе случайные величины имеют распределение Бернулли. При размерах выборок $n_1 = 100$ и $n_2 = 200$ (обе больше 30), по ИТМЛ можем получить распределения выборочных долей:

$$\hat{p}_1 \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \quad \hat{p}_2 \sim \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

- Гипотезы: $H_0 : p_1 = p_2$ против $H_1 : p_1 < p_2$ (эквивалентно $H_1 : p_1 - p_2 < 0$) (левосторонний тест)
- Распределение при нулевой гипотезе, когда мы абсолютно уверены в том, что она верна:

$$D = \hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(0, p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

  где p_c — объединённая доля при нулевой гипотезе $p_1 = p_2 = p_c$.
  Объединённая доля: $p_c = \frac{52+125}{100+200} = 0.59$
- Проверка с помощью z-статистики (score):
  z-статистика: $z_{\text{score}} = \frac{\tilde{p}_1 - \tilde{p}_2}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
  Подставляя значения: $z_{\text{score}} = \frac{0.52 - 0.625}{\sqrt{0.59 \cdot 0.41 \cdot \left(\frac{1}{100} + \frac{1}{200}\right)}} \approx -1.743$
  Критическая точка: $-z_{0.05} = -1.645$
  Решение: $z_{\text{score}} = -1.743 < -1.645$, поэтому отклоняем H_0 на уровне 5%.

(c) Гипотезы: $H_0 : p_2 = 0.55$ против $H_1 : p_2 > 0.55$ (правосторонний тест)

Распределение при нулевой гипотезе: $\hat{p}_2 \sim \mathcal{N}\left(0.55, \frac{0.55 \cdot 0.45}{200}\right)$

z-статистика: $z_{\text{score}} = \frac{0.625 - 0.55}{\sqrt{\frac{0.55 \cdot 0.45}{200}}} \approx 2.13$

Критическая точка: $z_{0.025} = 1.96$

Решение: $z_{\text{score}} = 2.13 > 1.96$, поэтому отклоняем H_0 на уровне 2.5%.

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 $\{X_1, \ldots, X_{11}\}$, а вторая — $\{Y_1, \ldots, Y_{10}\}$. 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, *т.е.* 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.

Решение:

(a) Распределения: $\bar{X} \sim \mathcal{N}\left(\mu_X, \frac{\sigma_X^2}{11}\right)$, $\bar{Y} \sim \mathcal{N}\left(\mu_Y, \frac{\sigma_Y^2}{10}\right)$, $D = \bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{11} + \frac{\sigma_Y^2}{10}\right)$

(b) Гипотезы: $H_0 : \mu_X = \mu_Y$ (или $\mu_X - \mu_Y = 0$) против $H_1 : \mu_X < \mu_Y$ (или $\mu_X - \mu_Y < 0$) (левосторонний тест)

(c) Распределения при нулевой гипотезе, когда мы абсолютно уверены в том, что она верна:

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(0, \frac{\sigma_X^2}{11} + \frac{\sigma_Y^2}{10}\right), \quad \text{а также} \quad \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{11} + \frac{S_Y^2}{10}}} \sim t_k$$

где число степеней свободы k задаётся формулой:

$$k \approx \frac{(V_X + V_Y)^2}{\frac{V_X^2}{n-1} + \frac{V_Y^2}{m-1}}, \text{ где } V_X = \frac{S_X^2}{n},\ V_Y = \frac{S_Y^2}{m}$$

(d) Проверка с помощью t-статистики (score):
t-статистика: $t_{\text{score}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$
Подставляя значения: $t_{\text{score}} = \frac{5.3 - 7.2}{\sqrt{\frac{2.4}{11} + \frac{2.8}{10}}} \approx -2.69$
Сравнение с критическими точками:

- При $\alpha = 1\%$: $t_{\text{score}} = -2.69 < -t_{(18,0.01)} = -2.552$, отклоняем H_0
- При $\alpha = 5\%$: $t_{\text{score}} = -2.69 < -t_{(18,0.05)} = -1.734$, отклоняем H_0
- При $\alpha = 10\%$: $t_{\text{score}} = -2.69 < -t_{(18,0.10)} = -1.330$, отклоняем H_0

(e) Построение границы критической области теста:

$$P_{H_0}\left(\bar{X} - \bar{Y} < K\right) = \alpha$$

$$P_{H_0}(\bar{X} - \bar{Y} < K) = P_{H_0}\left(\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{11} + \frac{S_Y^2}{10}}} < \frac{K}{\sqrt{\frac{S_X^2}{11} + \frac{S_Y^2}{10}}}\right) = P\left(t < -t_{(k,\alpha)}\right)$$

$$K = -t_{(k,\alpha)}\sqrt{\frac{S_X^2}{11} + \frac{S_Y^2}{10}}$$

(f) Вычисление степеней свободы:

$$V_X = \frac{2.4}{11} \approx 0.218, \quad V_Y = \frac{2.8}{10} = 0.28$$

$$k = \frac{(0.218 + 0.28)^2}{\frac{0.218^2}{10} + \frac{0.28^2}{9}} \approx 18.4 \approx 18$$

(g) Критические точки t-распределения с 18 степенями свободы: $t_{(18,0.01)} \approx 2.552$, $t_{(18,0.05)} \approx 1.734$, $t_{(18,0.10)} \approx 1.330$ (нет в той таблице, что отправляли, но можно найти в других)
Границы критической области:

$$K_{1\%} = -2.552 \cdot \sqrt{\frac{2.4}{11} + \frac{2.8}{10}} \approx -1.80$$

$$K_{5\%} = -1.734 \cdot \sqrt{\frac{2.4}{11} + \frac{2.8}{10}} \approx -1.22$$

$$K_{10\%} = -1.330 \cdot \sqrt{\frac{2.4}{11} + \frac{2.8}{10}} \approx -0.94$$

Проверка: $\bar{x} - \bar{y} = 5.3 - 7.2 = -1.9$

- При $\alpha = 1\%$: $-1.9 < K_{1\%} = -1.80$, отклоняем H_0
- При $\alpha = 5\%$: $-1.9 < K_{5\%} = -1.22$, отклоняем H_0
- При $\alpha = 10\%$: $-1.9 < K_{10\%} = -0.94$, отклоняем H_0

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 A	Район B
38	32
38	38
29	22
45	30
42	34
33	28
27	32
32	34
32	24
34	нет данных

Решение аналогично предыдущей задаче

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):

	Ремень безопасности использован		
Тяжёлая или смертельная травма	Да	Нет	Всего
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?

Решение:

(a) Гипотезы:

- Словами: H_0: Доля тяжёлых травм одинакова для пассажиров с ремнями и без ремней, H_1: Доля тяжёлых травм с ремнями меньше, чем без ремней
- Формализованно: $H_0 : p_1 = p_2$ против $H_1 : p_1 < p_2$, где p_1 — доля тяжёлых травм с ремнями, p_2 — без ремней

(b) Исследуемые случайные величины: X — индикатор тяжёлой травмы для пассажира с ремнями, Y — индикатор тяжёлой травмы для пассажира без ремней. Обе случайные величины имеют распределение Бернулли. При размерах выборок $n_1 = 832$ и $n_2 = 9168$ (обе больше 30), по ИТМЛ можем получить распределения выборочных долей:

$$\hat{p}_1 \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \quad \hat{p}_2 \sim \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

(c) Распределение при нулевой гипотезе, когда мы абсолютно уверены в том, что она верна:

$$D = \hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(0, p_c(1-p_c)\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)\right)$$

где p_c — объединённая доля при нулевой гипотезе $p_1 = p_2 = p_c$.

(d) Данные: $n_1 = 832$, $\tilde{p}_1 = \frac{3}{832} \approx 0.00361$; $n_2 = 9168$, $\tilde{p}_2 = \frac{119}{9168} \approx 0.01298$
Объединённая доля: $p_c = \frac{3+119}{832+9168} \approx 0.0122$

(e) Проверка с помощью z-статистики (score):
z-статистика: $z_{\text{score}} = \frac{\tilde{p}_1 - \tilde{p}_2}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
Подставляя значения: $z_{\text{score}} = \frac{0.00361 - 0.01298}{\sqrt{0.0122 \cdot 0.9878 \cdot \left(\frac{1}{832}+\frac{1}{9168}\right)}} = \frac{-0.00937}{0.00395} \approx -2.36$
Критические точки: $-z_{0.10} = -1.282$, $-z_{0.05} = -1.645$, $-z_{0.01} = -2.326$
Сравнение с критическими точками:

- При $\alpha = 1\%$: $z_{\text{score}} = -2.36 < -z_{0.01} = -2.326$, отклоняем H_0
- При $\alpha = 5\%$: $z_{\text{score}} = -2.36 < -z_{0.05} = -1.645$, отклоняем H_0
- При $\alpha = 10\%$: $z_{\text{score}} = -2.36 < -z_{0.10} = -1.282$, отклоняем H_0

P-значение: p-value $= P(Z < -2.36) \approx 0.0091$ (около 0.91%)
Вывод: Сильные статистические доказательства положительного эффекта автомобильных ремней безопасности.

(f) Доверительный интервал для $p_1 - p_2$:
$(1-\alpha)100\%$ доверительный интервал:

$$(p_1 - p_2) \in \left(\tilde{p}_1 - \tilde{p}_2 - z_{\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}}, \tilde{p}_1 - \tilde{p}_2 + z_{\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}}\right)$$

Примечание: В отличие от теста гипотез, здесь мы используем отдельные выборочные доли $\tilde{p}_1$ и $\tilde{p}_2$, а не объединённую долю p_c.

Подкоренное выражение: $\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}} \approx 0.00239$

90%: $(p_1 - p_2) \in -0.00937 \pm 1.645 \cdot 0.00239 = (-0.0133, -0.0054)$

95%: $(p_1 - p_2) \in -0.00937 \pm 1.96 \cdot 0.00239 = (-0.0141, -0.0047)$

99%: $(p_1 - p_2) \in -0.00937 \pm 2.576 \cdot 0.00239 = (-0.0155, -0.0032)$

Все интервалы не содержат нуль и находятся полностью в отрицательной области, что дополнительно подтверждает, что $p_1 < p_2$ (ремни безопасности снижают долю тяжёлых травм).