Интервальные оценки в статистике

Независимые и зависимые (парные) выборки.

Независимые выборки (Independent Samples):

Две выборки считаются независимыми, если наблюдения в одной **не зависят** от наблюдений в другой.

Примеры:

- Сравнение среднего балла **школьников и студентов**.
- Измерение уровня гемоглобина у **двух разных групп** пациентов.
- Сравнение продаж в **двух разных магазинах**.

Математическая модель:

$$X_1, X_2, \dots, X_n \sim N(\mu_1, \sigma_1^2)$$

$$Y_1, Y_2, \dots, Y_m \sim N(\mu_2, \sigma_2^2)$$

Зависимые (парные) выборки (Dependent / Paired Samples)

Наблюдения в одной выборке **соотносятся напрямую** с наблюдениями в другой. Часто — это **измерения «до» и «после»** для одного и того же объекта.

Примеры:

- До и после приёма лекарства у **одного пациента**.
- Время реакции **одного человека** до и после кофе.
- Тестирование **двух методов обучения** на **одних и тех же студентах**.

Математическая модель:

Работаем с разностями:

$$D_i = X_i - Y_i, i = 1, \dots , n$$

Задача сводится к анализу одной выборки разностей!

Ключевые различия:

Характеристика	**Независимые выборки**	**Зависимые (парные) выборки**
Структура данных	Разные группы	Те же объекты в разных условиях
Размер выборки	Обычно разное: $n1 \neq n2$	Всегда одинаково: $n1 = n2 = n$
Анализируемое значение	Разность средних: $\bar{X} - \bar{Y}$	Среднее разностей: $\bar{D}$
Рассеяние/шум	Больше	Меньше (так как «шум» совпадает)
Метод оценки	Статистика двух выборок	Статистика одной (разностей)

Вывод:

- Используйте **независимые выборки**, если группы не связаны.
- Используйте **парный анализ**, если наблюдения «связаны по смыслу»: до/после, близнецы, левый/правый глаз и т.п.
- Парный подход **уменьшает дисперсию**, повышает **мощность теста** и **сужает доверительные интервалы**.

Доверительные интервалы на разность математический ожиданий.

Условие: независимые выборки

Пусть:

$$X_1, X_2, \dots, X_n \sim N(\mu_1, \sigma_1^2) - \text{первая выборка}$$

$$Y_1, Y_2, \dots, Y_m \sim N(\mu_2, \sigma_2^2) - \text{вторая выборка}$$

Обе выборки независимы.

1. Задача: построить интервал для $\boldsymbol{\mu 1 - \mu 2}$

Оценка разности математических ожиданий:

$$\mu 1 - \mu 2 = \bar{X} - \bar{Y}$$

2. Распределение оценки:

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n}\right), \quad \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

⇒ Разность $\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$

Случай 1: дисперсии известны (σ_1, σ_2 известны)

Тогда можно стандартизировать:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1)$$

Для доверительного интервала уровня $1 - \alpha$:

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha$$

⇒ Подставляем выражение Z, преобразуем:

$$P\left(\hat{\delta} - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2 < \hat{\delta} + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right) = 1 - \alpha$$

Финальная формула:

$$\boxed{\mu_1 - \mu_2 \in \left[\bar{X} - \bar{Y} \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right]}$$

Случай 2:

Доверительный интервал для разности математических ожиданий $\boldsymbol{\mu_1 - \mu_2}$ по **двум независимым выборкам с неизвестными и неравными дисперсиями**.

Условие задачи:

У нас есть **две независимые выборки**:

$$X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2) - \text{первая выборка}$$

$$Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma_2^2) - \text{вторая выборка}$$

где:

- μ_1, μ_2 — неизвестны,
- σ_1^2, σ_2^2 — тоже **неизвестны**, и **предполагаются неравными**,
- X_i и Y_j — **независимы** между собой.

Цель:

Построить доверительный интервал уровня $1 - \alpha$ для параметра:

$$\boldsymbol{\mu_1 - \mu_2}$$

Этап 1. Оценка разности средних:

$$\hat{\delta} = \bar{X} - \bar{Y}$$

где:

- $\bar{X} = \frac{1}{n} \sum X_i$,
- $\bar{Y} = \frac{1}{m} \sum Y_j$

Этап 2. Выборочные дисперсии:

Так как σ_1^2 и σ_2^2 неизвестны, мы используем несмещённые оценки:

$$S_1^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{m-1}\sum(Y_j - \bar{Y})^2$$

Этап 3. Распределение разности средних

Так как выборки независимы, а дисперсии неравны, обобщённая формула дисперсии разности:

$$\mathrm{Var}(\bar{X} - \bar{Y}) \approx \frac{S_1^2}{n} + \frac{S_2^2}{m}$$

Соответственно, используем следующую **t-статистику:**

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$$

Этап 4. Распределение t

Вот где важный момент: **распределение этой статистики — не стандартное t**, а приближённое **t-распределение с дробным числом степеней свободы**, по формуле **Уэлча (Welch–Satterthwaite):**

$$\nu = \frac{\left(\frac{S_1^2}{n} + \frac{S_2^2}{m}\right)^2}{\frac{\left(\frac{S_1^2}{n}\right)^2}{n-1} + \frac{\left(\frac{S_2^2}{m}\right)^2}{m-1}}$$

Это приближение позволяет использовать критические значения t (α/2,df).

Этап 5. Формула доверительного интервала

$$\boxed{\mu_1 - \mu_2 \in \left[\bar{X} - \bar{Y} \pm t_{\alpha/2,\,\nu} \cdot \sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}} \right]}$$

где:

- $\bar{X}, \bar{Y}$ — выборочные средние,
- S_1^2, S_2^2 — выборочные дисперсии,
- $t_{\alpha/2,\,\nu}$ — квантиль t-распределения с ν степенями свободы.

Случай 3:

Доверительный интервал для разности математических ожиданий $\boldsymbol{\mu_1 - \mu_2}$ по **двум независимым выборкам с неизвестными, но равными дисперсиями**.

Условие:

Пусть:

$$X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma^2) - \text{первая выборка}$$

$$Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma^2) - \text{вторая выборка}$$

где:

- Выборки независимы,
- μ_1, μ_2 — неизвестны,
- σ^2 одинакова, но неизвестна.

Цель:

Построить доверительный интервал для $\boldsymbol{\mu_1 - \mu_2}$ с уровнем доверия 1−α

Шаг 1: оценка разности средних:

$$\hat{\delta} = \bar{X} - \bar{Y}$$

где:

- $\bar{X} = \frac{1}{n} \sum X_i,$
- $\bar{Y} = \frac{1}{m} \sum Y_j$

Шаг 2: выборочные дисперсии

$$S_1^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{m-1}\sum(Y_j - \bar{Y})^2$$

Шаг 3: объединённая (пулевая) дисперсия

Так как σ^2 считается **одинаковой**, берём её **общую оценку** из обеих выборок:

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

— это несмещённая оценка общей дисперсии, учитывающая оба объёма.

Шаг 4: распределение t-статистики

Поскольку объединённая дисперсия используется, применима классическая t-статистика:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

Шаг 5: Доверительный интервал

$$\boxed{\mu_1 - \mu_2 \in \left[\bar{X} - \bar{Y} \pm t_{\alpha/2,\, n+m-2} \cdot S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}\right]}$$

Вывод использования. Если:

- выборки независимы,
- дисперсии одинаковы, но неизвестны,
- мы можем использовать объединённую дисперсию и t-распределение с $n + m - 2$ степенями свободы.

Это более узкий интервал, чем у Уэлча, но требует предположения о равенстве дисперсий.

Доверительный интервал для разности долей (пропорций) в двух независимых выборках.

Условие

У нас есть две независимые выборки:

- В первой выборке объём n_1, число «успехов» x_1, доля $\hat{p}_1 = \frac{x_1}{n_1}$
- Во второй выборке объём n_2, число «успехов» x_2, доля $\hat{p}_2 = \frac{x_2}{n_2}$

Цель

Построить доверительный интервал для разности долей $p_1 - p_2$:

Формула (асимптотический интервал)

Если выборки достаточно большие, и выполняется условие:

- $\boldsymbol{n_1 \widehat{p_1} \geq 10}$,
- $\boldsymbol{n_1(1 - \widehat{p_1}) \geq 10}$,
- $\boldsymbol{n_2 \widehat{p_2} \geq 10}$,
- $\boldsymbol{n_2(1 - \widehat{p_2}) \geq 10}$

то мы используем нормальное приближение:

$$\boxed{(p_1 - p_2) \in \left[(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}\right]}$$

где:

- $z_{\alpha/2}$ — квантиль стандартного нормального распределения (например, 1.96 при 95% уровне доверия)

Доверительный интервал для разности математических ожиданий μ1−μ2 по **парной (зависимой) выборке**

Когда это используется?

Когда мы сравниваем две выборки, в которых наблюдения идут парами:

- До и после (эксперимент/терапия),
- Левая и правая рука одного человека,
- Один и тот же человек в двух условиях (время, режим, продукт и т.п.).

Условие

Пусть даны пары наблюдений:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

где:

- X_i, Y_i — результаты для одного и того же объекта,
- $D_i = X_i - Y_i$ — **разности внутри каждой пары**.

Цель

Построить доверительный интервал для:

$$\boldsymbol{\mu_D = \mu_1 - \mu_2}$$

(то есть для математического ожидания разностей).

Методика

Поскольку всё сводится к одной выборке разностей Di, задача превращается в обычный одновыборочный случай:

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i, \quad S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$$

Формула доверительного интервала:

$$\boxed{\mu_D \in \left[\bar{D} \pm t_{\alpha/2,\, n-1} \cdot \frac{S_D}{\sqrt{n}}\right]}$$

где:

- $\bar{D}$ — среднее разностей,
- S_D — стандартное отклонение разностей,
- $t_{\alpha/2,n-1}$ — квантиль t-распределения.

Доверительный интервал для разности (отношений) дисперсий $\boldsymbol{\sigma_1^2 - \sigma_2^2}$:

📌 **Обычно доверительные интервалы строят не для разности дисперсий $\sigma_1^2 - \sigma_2^2$, а для их отношения:**

$$\frac{\sigma_1^2}{\sigma_2^2}$$

Это связано с тем, что разность дисперсий — не положительная величина, и её распределение сложно описать. А вот отношение дисперсий можно протестировать и оценить с помощью F-распределения.

Интервал для отношения дисперсий (F-интервал)

Пусть:

- $X_1, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$
- $Y_1, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$

и S_1^2, S_2^2 — выборочные дисперсии.

Тогда статистика:

$$F = \frac{S_1^2}{S_2^2} \sim F(n-1,\ m-1)$$

Доверительный интервал для отношения:

$$\boxed{\left[\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{1-\alpha/2;\ n-1,\ m-1}},\ \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{\alpha/2;\ n-1,\ m-1}}\right]}$$

где:

- $F_{\alpha/2}, F_{1-\alpha/2}$ — квантиль F-распределения.