Введение в математическую статистику III

Computer Science Club, 27 ноября 2021

Частотный подход

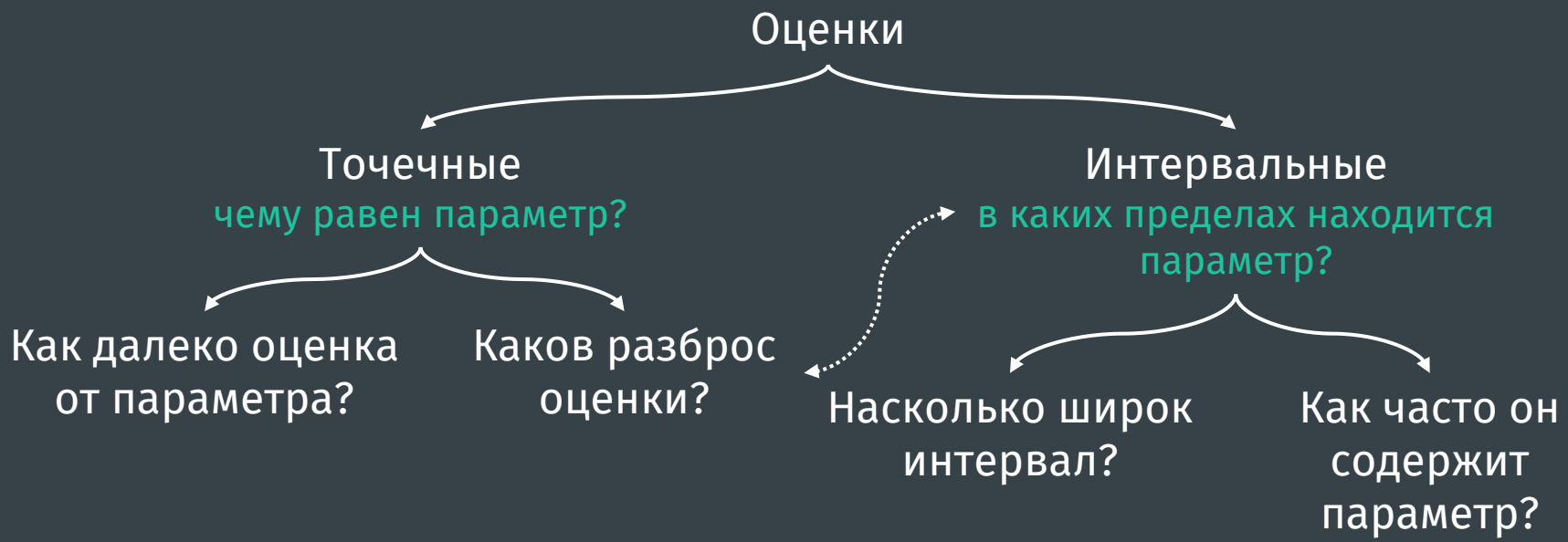
Основная идея — вероятность есть только у событий, которые можно (хотя бы виртуально) провести бесконечное число раз.

Примеры:

– монетку можно подбросить много раз
– провести один и тот же матч Челси—Зенит 8 декабря 2021 года много раз совсем нельзя

Одним из следствий такого подхода является отказ от априорных распределений. Аналогично, всякая гипотеза либо верна, либо не верна. Здесь нет никакой вероятности.

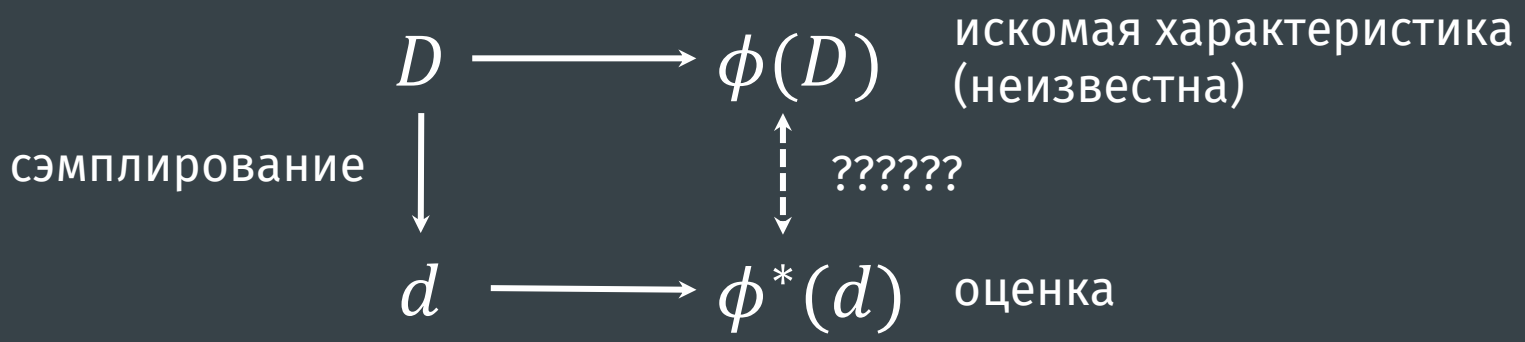
Оценка параметров



PDMI COMPUTER SCIENCE CLUB

Как анализировать оценки?

Параметр или характеристика распределения ϕ — это функционал от этого распределения. Функция ϕ^* от данных d называется статистикой.



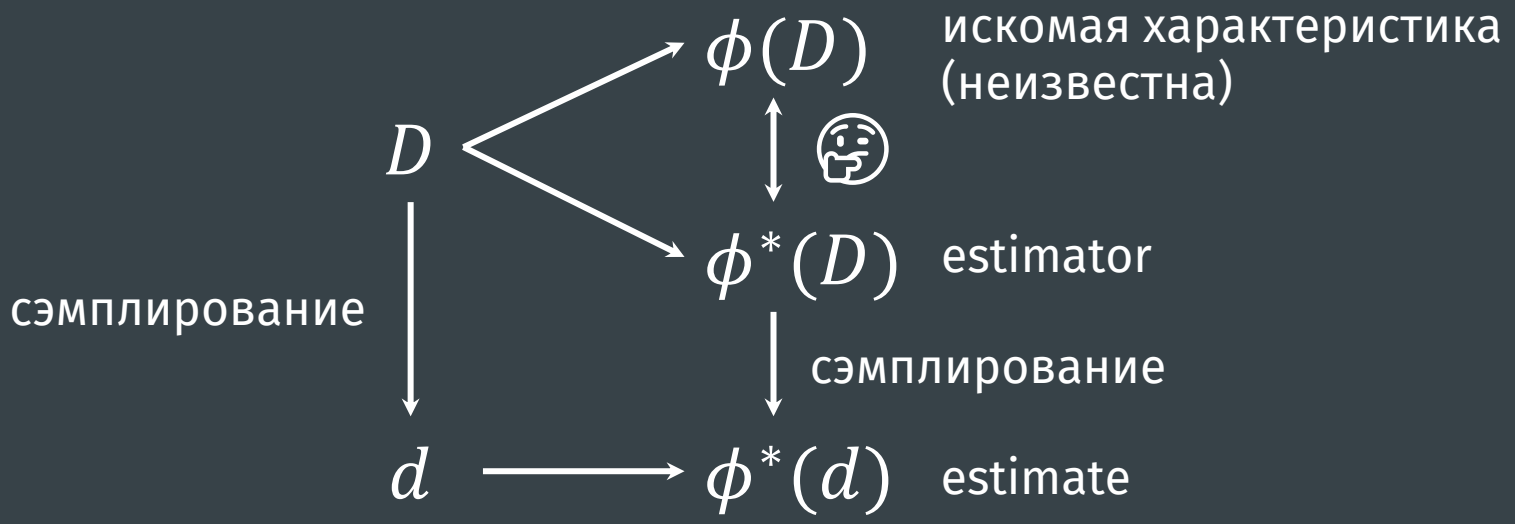
Примеры:

- $D = X_{[n]}$, $\phi^*(x_{[n]}) = \bar{x}$ — статистика
- $D = X_{[n]}$, $\phi^*(x_{[n]}) = \mathbb{E}X_1$ — не статистика

PDMI
COMPUTER
SCIENCE
CLUB

Как анализировать оценки?

Основная идея анализа оценок — это интерпретация $\phi^*(d)$ как реализации случайной величины $\phi^*(D)$.



Пример:

- $D = X_{[5]}$, $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $d = [4.21, -0.97, 2.7, 2.25, 0.43]$,
  $\phi(D) = \mathbb{E}X_i = \mu$, $\phi^*(d) = \bar{x} = 1.72$, $\theta^*(D) = \bar{X} \sim \mathcal{N}(\mu, \sigma^2/5)$

PDMI COMPUTER SCIENCE CLUB

Смещение и разброс

Пусть $X_{[n]}$ — выборка из $X \sim \mathcal{P}$, а $\phi(\mathcal{P})$ — искомая характеристика.

Величина $b(\phi^*) = \mathbb{E}\phi^*\left(X_{[n]}\right) - \phi(\mathcal{P})$ называется смещением.

Разброс оценки типично измеряется ее дисперсией $\mathbb{D}\phi^*\left(X_{[n]}\right)$.

Чем меньше смещение, тем лучше; чем меньше разброс, тем лучше.

Величина $MSE(\phi^*) = \mathbb{E}\left(\phi^*\left(X_{[n]}\right) - \phi(\mathcal{P})\right)^2$ называется среднеквадратичной ошибкой. Можно показать, что

$$MSE(\phi^*) = \mathbb{D}\phi^*\left(X_{[n]}\right) + b^2(\phi^*),$$

то есть MSE сочетает в себе и смещение и разброс оценки. Из двух оценок имеет смысл выбирать ту, у которой меньше MSE. Часто вместо MSE рассматривают $RMSE = \sqrt{MSE}$.

PDMI COMPUTER SCIENCE CLUB

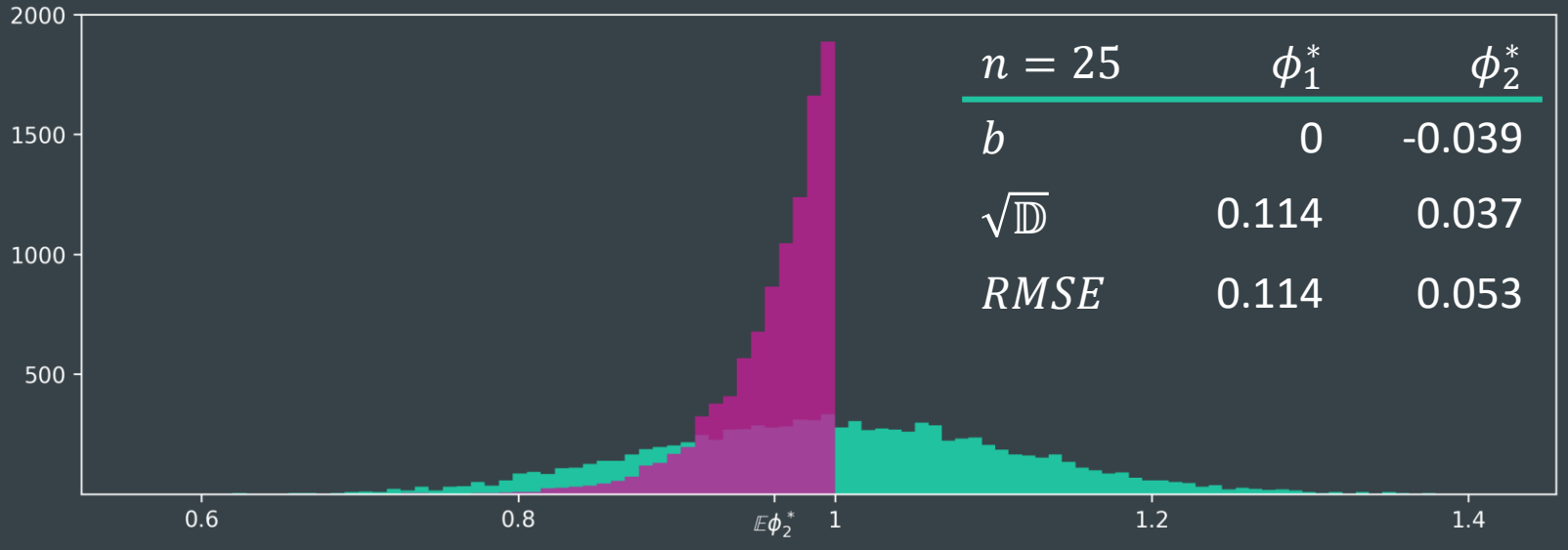
Пример

Пусть $X \sim U([0,\theta])$, $\theta \in (0,\infty)$. Рассмотрим две оценки $\phi = \theta$:

– $\phi_1^*(x_{[n]}) = 2\bar{x}$

– $\phi_2^*(x_{[n]}) = \max\{x_{[n]}\}$

Заметим, что $\phi_i^*(\lambda \cdot x_{[n]}) = \lambda\phi_i^*(x_{[n]})$, поэтому для их анализа достаточно рассмотреть какой-нибудь один θ (например, $\theta = 1$).



$n = 25$	ϕ_1^*	ϕ_2^*
b	0	-0.039
$\sqrt{\mathbb{D}}$	0.114	0.037
$RMSE$	0.114	0.053

PDMI COMPUTER SCIENCE CLUB

Из-за чего возникает смещение

Есть три основных причины смещенности оценок:

– Использование нелинейных преобразований

  Пример: $\mathbb{E}\left(\frac{1}{X}\right) \neq \frac{1}{\mathbb{E}X}$, поэтому $\frac{1}{\bar{x}}$ будет смещенной оценкой $\frac{1}{\mathbb{E}X}$

– Использование оптимизации — если оптимизировать значение некоторой величины, то получившееся ее значение будет смещено

  Пример: $\frac{1}{n}\sum_i (x_i - \mathbb{E}X)^2$ является несмещенной оценкой $\mathbb{D}X$, но $\frac{1}{n}\sum_i (x_i - \bar{x})^2$ уже смещена: дело в том, что $\bar{x} = \arg\min_c \frac{1}{n}\sum_i (x_i - c)^2$

– Использование плохой модели

PDMI
COMPUTER
SCIENCE
CLUB

Оценка смещения и разброса

В предыдущем примере мы попользовались особенностью модели: распределение оценок очень удобно зависит от θ. Это дало нам возможность оценить все подряд.

Такое бывает не всегда. В частности, возможны следующие трудности:

- Распределение ϕ^* однозначно определяется ϕ, но сложным образом от него зависит
- Распределение ϕ^* не однозначно определяется ϕ

  Пример: $X \sim \mathcal{N}(\mu \in \mathbb{R}, \sigma^2 > 0)$, $\phi = \mu$, $\phi^* = \bar{x}$
- Модель не параметрическая и распределение ϕ^* вообще не определяется конечномерным вектором параметров

Бутстрап

Пусть ϕ^* является plug-in оценкой, то есть имеет конкретный вид

$$\phi^*(x_{[n]}) = \phi\left(\tilde{\mathcal{P}}(x_{[n]})\right),$$

где $\tilde{\mathcal{P}}$ это некоторая оценка истинного распределения по выборке $x_{[n]}$. Во всех перечисленных случаях оценить ее необходимые характеристики можно с помощью техники, которая называется бутстрап.

Мы с ней уже сталкивались кучу раз, когда оценивали модель с помощью данных, сгенерированных этой моделью.

Бутстрап

Идея — Приближаем распределение $\phi^*(X_{[n]})$ с помощью бутстраповской выборки $\phi^*_{[N]} = \{\phi^*(x^*_{[n],i}), i = 1, \dots, N\}$, где $x^*_{[n],i}$ сэмплируются из $\tilde{\mathcal{P}}(x_{[n]})$.

Основной принцип — Величина $\phi^*(X^*_{[n]})$ относится к $\phi^*(x_{[n]})$ так же, как $\phi^*(X_{[n]})$ относится к $\phi(\mathcal{P})$.

Виды бутстрапа:

- Непараметрический — $\tilde{\mathcal{P}} = \mathcal{P}^*_n$

  Пример: Модель 1 + Модель 1
- Параметрический — $\tilde{\mathcal{P}} = \mathcal{P}(\theta^*)$, где θ^* это оценка параметра θ модели $\mathcal{P}(\theta)$

  Пример: Модель 3 + Модель 3

PDMI COMPUTER SCIENCE CLUB

Оценка смещения

По определению, $b(\phi^*) = \mathbb{E}\phi^*\left(X_{[n]}\right) - \phi(\mathcal{P})$.

Применяя основной принцип, оценим эту величину с помощью

$$\mathbb{E}\phi^*\left(X^*_{[n]}\right) - \phi^*\left(x_{[n]}\right) \approx \frac{1}{N}\sum_i \phi^*\left(x^*_{[n],i}\right) - \phi^*\left(x_{[n]}\right).$$

Пример

$X \sim Exp(\lambda), \lambda > 0; \phi = \frac{1}{\mathbb{E}X} = \lambda, \phi^*\left(x_{[n]}\right) = \frac{1}{\bar{x}}.$

Рассмотрим параметрический бутстрап $-\ \tilde{\mathcal{P}} = Exp(1/\bar{x})$:

```
n = 25

def phi(x):
    return 1 / np.mean(x, axis=1)

truth = ss.expon(scale=1.337)
sample = truth.rvs(size=n)
boot = ss.expon(scale=sample.mean())

print(phi(truth.rvs(size=(10000, n))).mean() - 1 / truth.mean())
print(phi(boot.rvs(size=(10000, n))).mean() - 1 / boot.mean())
```

```
0.028897219618772674
0.02661362039964732
```

```
n = 25

def phi(x):
    return 1 / np.mean(x, axis=1)

truth = ss.expon(scale=13.37)
sample = truth.rvs(size=n)
boot = ss.expon(scale=sample.mean())

print(phi(truth.rvs(size=(10000, n))).mean() - 1 / truth.mean())
print(phi(boot.rvs(size=(10000, n))).mean() - 1 / boot.mean())
```

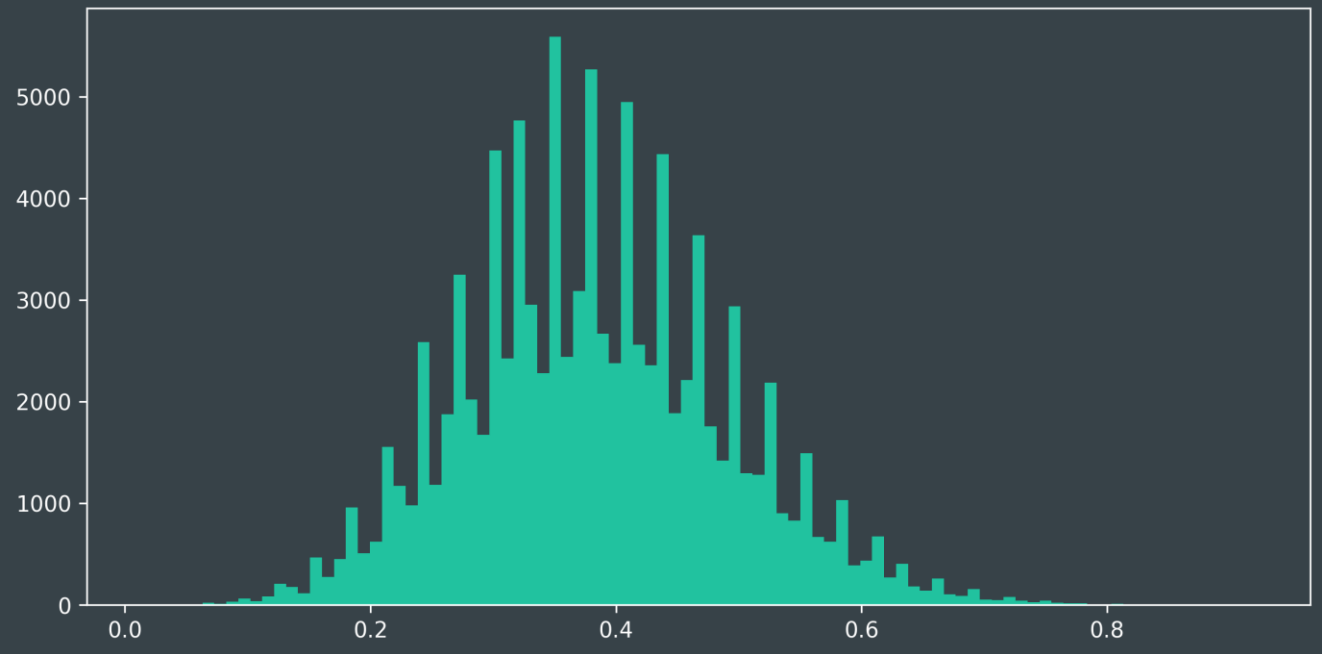
```
0.002918852929327323
0.0031658162228872383
```

PDMI COMPUTER SCIENCE CLUB

Оценка смещения

Бутстрап не может определить смещение, обусловленное плохой моделью.

Пример: Модель 1 + Модель 1



PDMI
COMPUTER
SCIENCE
CLUB

Оценка разброса

В качестве оценки меры разброса берем нужную меру разброса бутстраповской выборки.

Пример

```
n = 25

def phi(x):
    return np.median(x, axis=1)

truth = ss.norm()
sample = truth.rvs(size=n)
boot = ss.rv_discrete(values=(np.arange(n), np.ones(n)/n))

print(phi(truth.rvs(size=(10000, n))).std())
print(phi(sample[boot.rvs(size=(10000, n))]).std())
```

```
0.24966999967451106
0.24417308697867648
```

PDMI COMPUTER SCIENCE CLUB

Ошибка бутстрапа

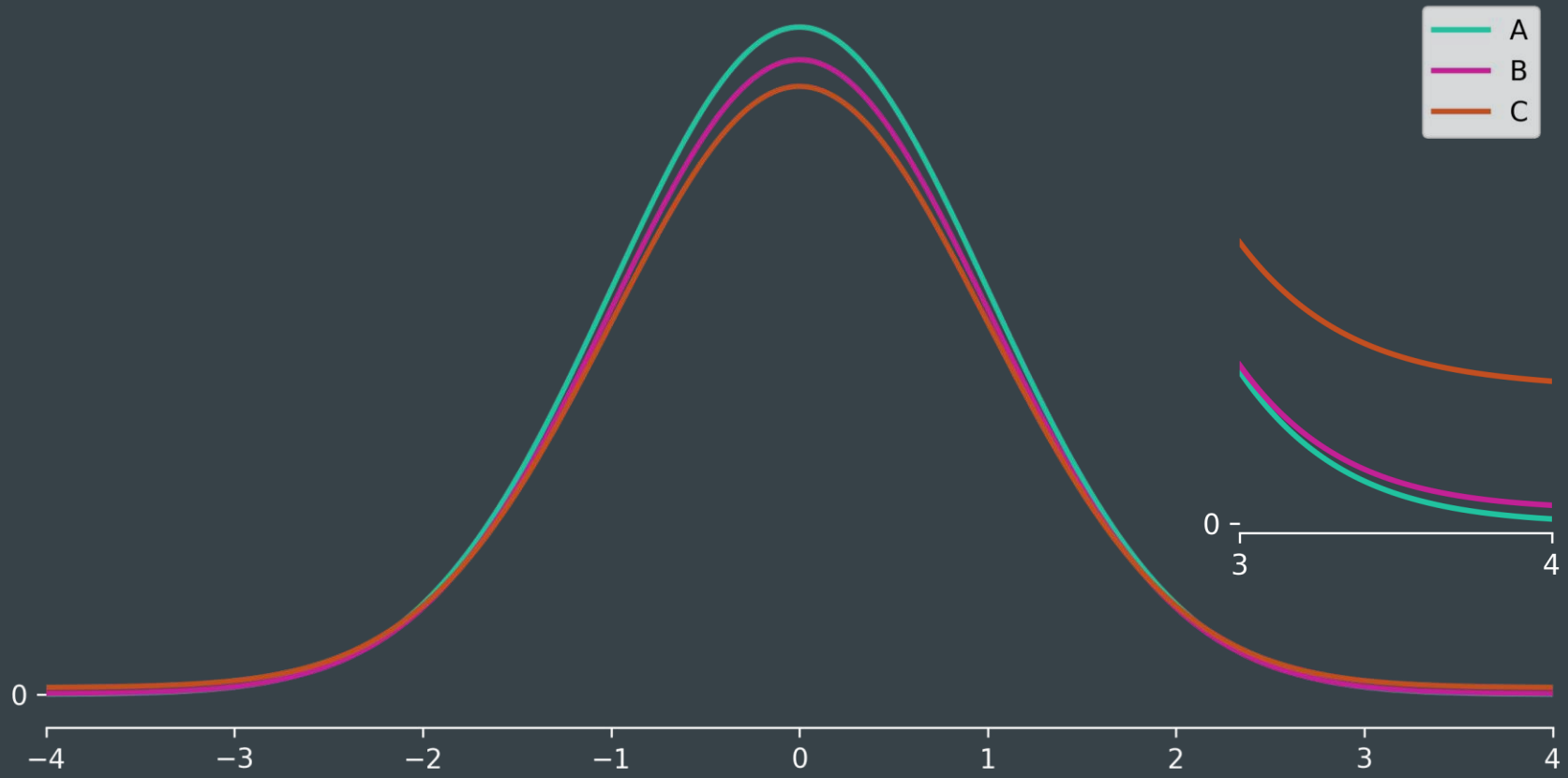
Ошибка бутстрапа складывается из неустранимой ошибки, связанной с тем, что мы взяли $\tilde{\mathcal{P}}$ вместо $\mathcal{P}$, и устранимой ошибки, связанной с тем, что мы взяли N выборок, а не ∞.

PDMI
COMPUTER
SCIENCE
CLUB

Робастность

Робастность

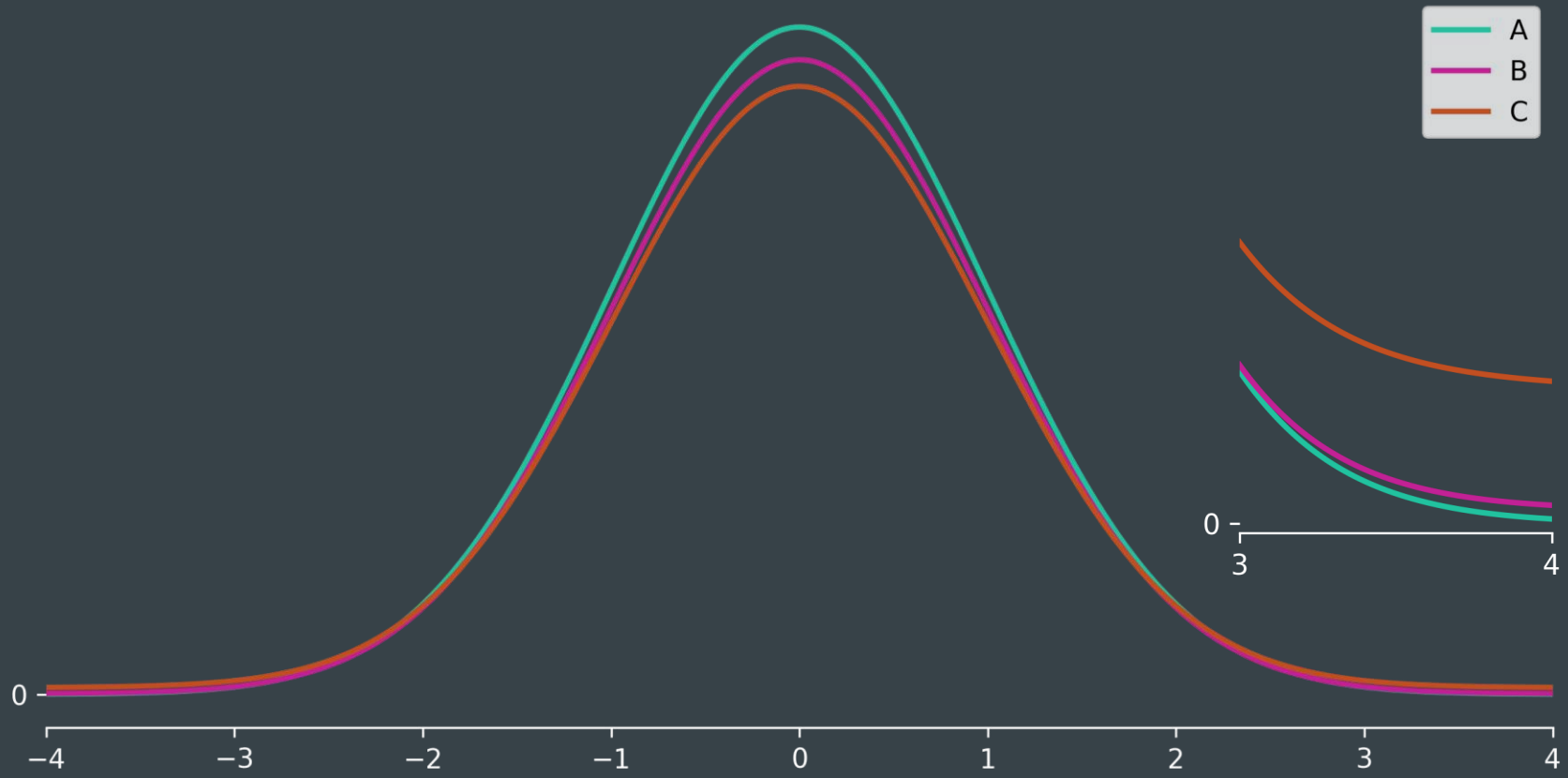
Чему равны стандартные отклонения?



PDMI
COMPUTER
SCIENCE
CLUB

Робастность

Стандартные отклонения равны 1, 3 и 11. У кого какое?



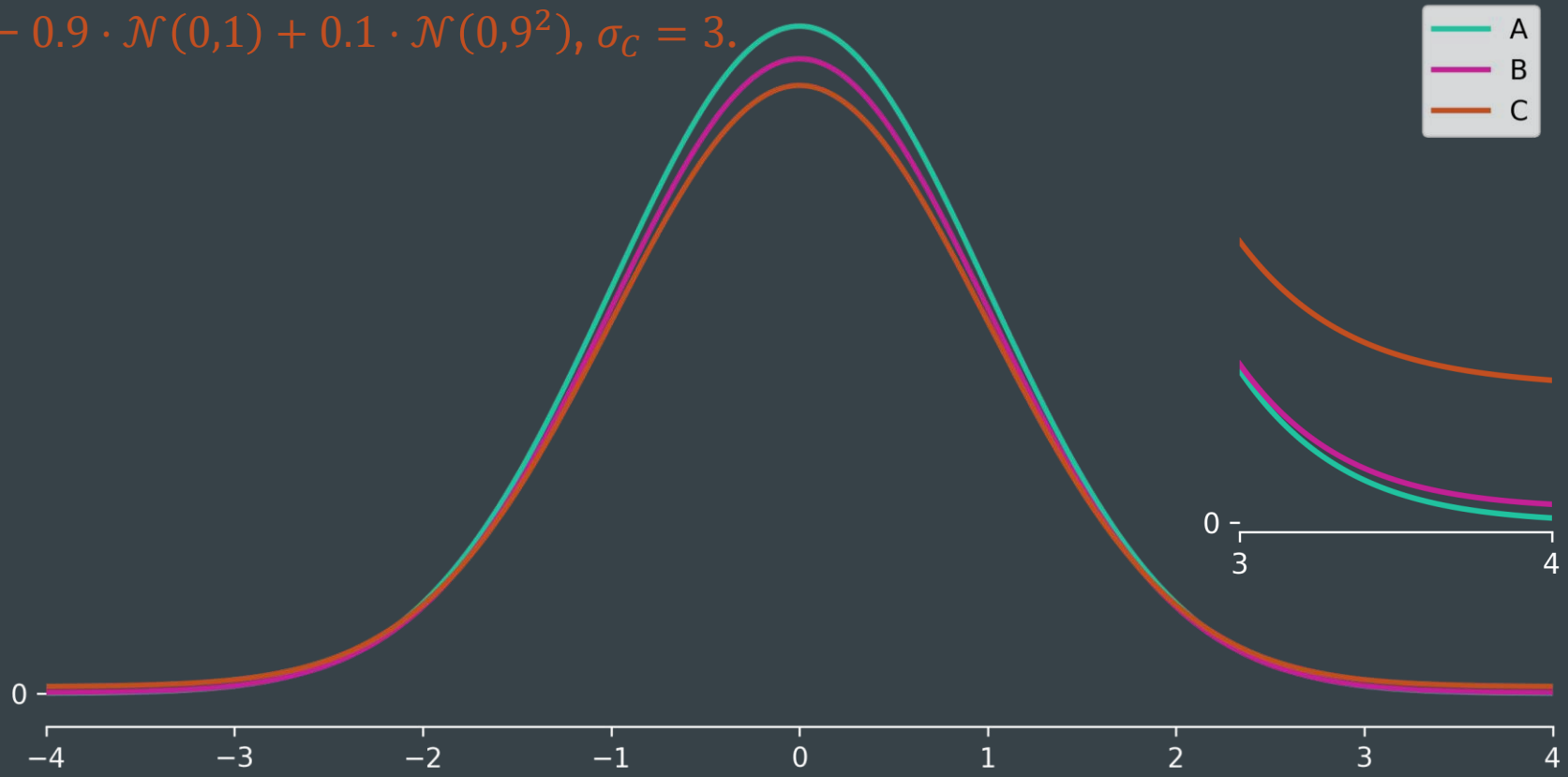
PDMI
COMPUTER
SCIENCE
CLUB

Робастность

A — $\mathcal{N}(0,1)$, $\sigma_A = 1$;

B — $0.95 \cdot \mathcal{N}(0,1) + 0.05 \cdot \mathcal{N}(0,49^2)$, $\sigma_B = 11$;

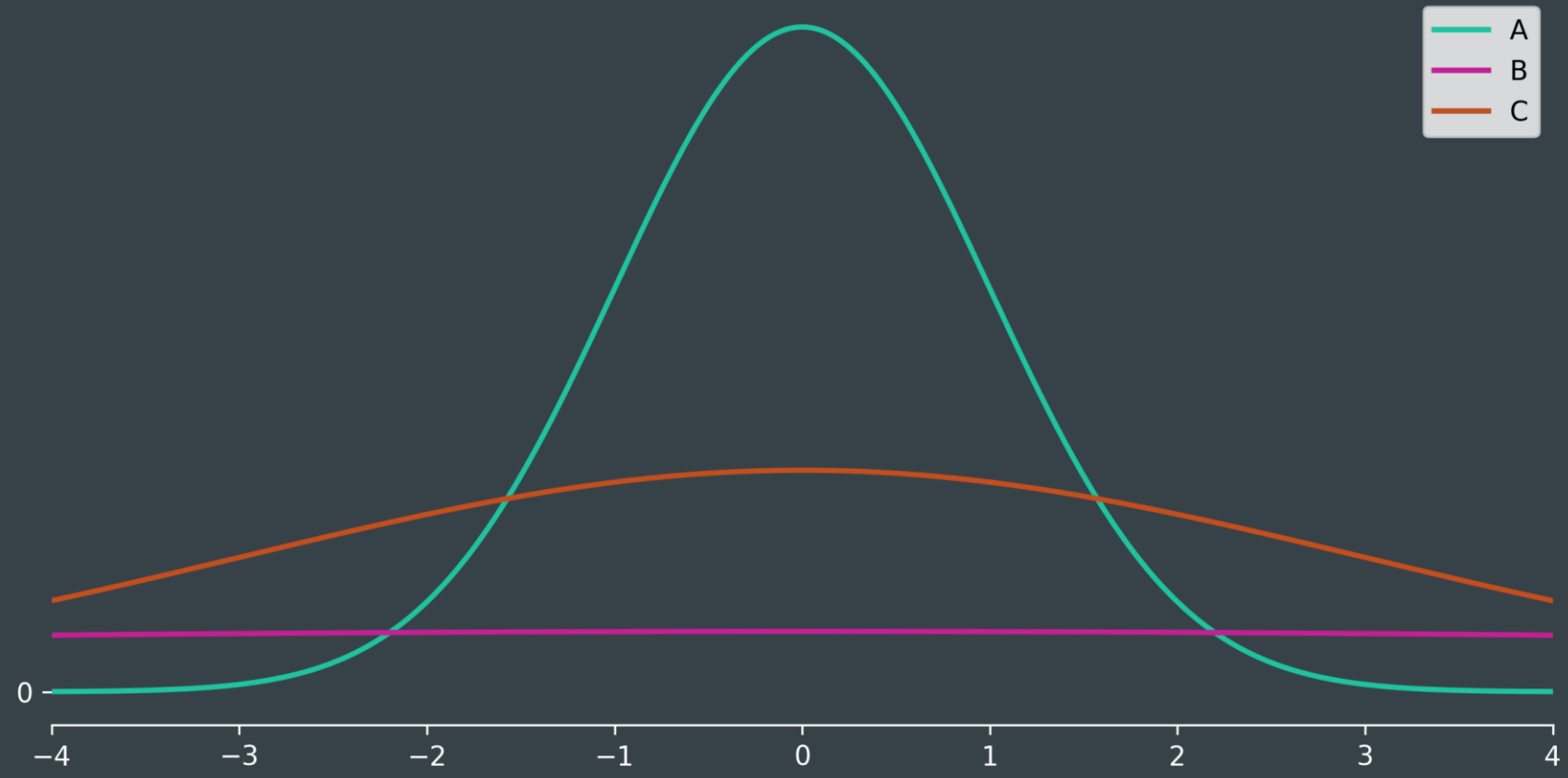
C — $0.9 \cdot \mathcal{N}(0,1) + 0.1 \cdot \mathcal{N}(0,9^2)$, $\sigma_C = 3$.



PDMI
COMPUTER
SCIENCE
CLUB

Робастность

Если бы все распределения были нормальными, то картинка выглядела бы вот так:



PDMI
COMPUTER
SCIENCE
CLUB

Робастность

Вывод: со стандартным отклонением что-то не то — оно очень чувствительно к хвостам распределения.

Свойство характеристики или оценки быть устойчивой к хвостам распределения называется робастностью.

Пороговая точка

Один из способов анализировать робастность оценки — двигать значения нескольких элементов выборки и смотреть, как ведет себя оценка.

Минимальная доля τ элементов выборки, контроль над которой позволяет устремить оценку к $\pm\infty$ называется пороговой точкой.

Примеры:

- Выборочное среднее — $\tau = \frac{1}{n}$.
- k-Усеченное среднее $\frac{1}{n-2k}\sum_{i=k+1}^{n-k-1} x_{(i)}$ — $\tau = \frac{k}{n}$, для медианы $\tau \approx 0.5$.
- Квантиль уровня α — $\tau \approx \min\{\alpha, 1-\alpha\}$.

PDMI COMPUTER SCIENCE CLUB

Робастные оценки

Базовые методы робастного оценивания:

- Отслеживание выбросов и исключение (усечение) или замена их на менее экстремальное значение (винзоризация)
- Квантили в качестве параметров положения
- Абсолютные отклонения вместо квадратичных
- Медиана вместо среднего: $\tilde{s}^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \to MAD = \text{med}\{|x - med^*|\}$

PDMI COMPUTER SCIENCE CLUB

Интервальные оценки

Доверительные интервалы

Пара статистик (ϕ_L^*, ϕ_R^*) называется доверительным интервалом для $\phi(\mathcal{P})$ с уровнем доверия γ если для любого $\mathcal{P} \in \mathfrak{P}$ выполняется

$$\mathbb{P}\left(\phi_L^*\left(X_{[n]}\right) \leq \phi(\mathcal{P}) \leq \phi_R^*\left(X_{[n]}\right) \middle| X \sim \mathcal{P}\right) = \gamma.$$

Если $\mathbb{P}\left(\phi_L^*\left(X_{[n]}\right) \leq \phi(\mathcal{P}) \leq \phi_R^*\left(X_{[n]}\right)\right) \xrightarrow[n\to\infty]{} \gamma$, то интервал называется асимптотическим.

Интервал называется точным, если хочется подчеркнуть, что он не асимптотический.

PDMI
COMPUTER
SCIENCE
CLUB

Доверительные интервалы

$\mathbb{P}\left(\phi_L^*(X_{[n]}) \le \phi(\mathcal{P}) \le \phi_R^*(X_{[n]}) \middle| X \sim \mathcal{P}\right) = \gamma$ означает, что если сгенерировать N выборок из распределения $\mathcal{P}$ и для каждой из них построить доверительный интервал, то примерно γN из них накроют истинное значение $\phi(\mathcal{P})$.

Конкретный доверительный интервал $\left(\phi_L^*(x_{[n]}), \phi_R^*(x_{[n]})\right)$ либо накрывает истинное значение, либо нет, никакой вероятности у этого события нет: нельзя говорить «Этот интервал с вероятностью 90% содержит истину».

PDMI
COMPUTER
SCIENCE
CLUB

Доверительные интервалы

Если $\mathbb{P}\left(\theta(\mathcal{P}) < \theta_L^*(X_{[n]})\right) = \mathbb{P}\left(\theta_R^*(X_{[n]}) < \theta(\mathcal{P})\right)$, то интервал называется центральным.

Если $\mathbb{P}\left(\theta(\mathcal{P}) < \theta_L^*(X_{[n]})\right) = 0$, то интервал называется левым (часто $\theta_L^*(X_{[n]}) = -\infty$).

Если $\mathbb{P}\left(\theta_R^*(X_{[n]}) < \theta(\mathcal{P})\right) = 0$, то интервал называется правым (часто $\theta_R^*(X_{[n]}) = \infty$).

Как строить доверительные интервалы?

Во-первых, бутстрап.

Эфронов доверительный интервал:

Пусть ϕ_L^* — α_1-квантиль бутстраповской выборки, а ϕ_R^* — $1 - \alpha_2$-квантиль.

Тогда (ϕ_L^*, ϕ_R^*) это асимптотический доверительный интервал с уровнем доверия $1 - \alpha_1 - \alpha_2$.

Как строить доверительные интервалы?

Во-вторых, метод центральной функции.

Пример:

Пусть $X \sim \mathcal{N}(\mu \in \mathbb{R}, \sigma^2 > 0)$, $\phi = \mu$.

Заметим, что распределение величины $t\left(X_{[n]}, \mu, \sigma\right) = \frac{\bar{X}-\mu}{S}$ не зависит от μ и σ.

Сгенерируем кучу выборок из $\mathcal{N}(0,1)$, посчитаем для каждой из них величину $t\left(x_{[n]}, 0,1\right)$ и оценим ее 0.025 и 0.975 квантили $t^*_{0.025}$ и $t^*_{0.975}$.

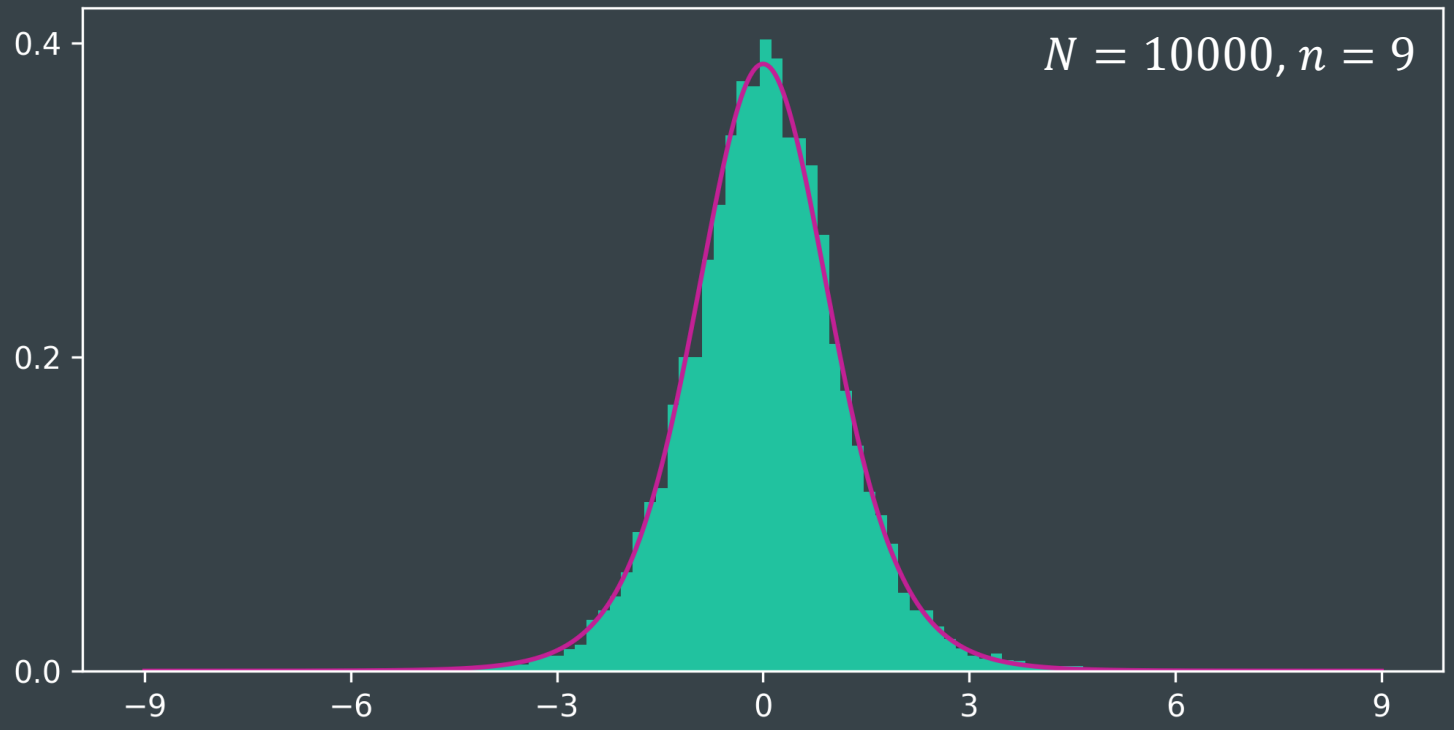
Тогда $\mathbb{P}\left(t^*_{0.025} < \frac{\bar{X}-\mu}{S} < t^*_{0.975}\right) = \mathbb{P}(\bar{X} - t^*_{0.975}S < \mu < \bar{X} - t^*_{0.025}S) \approx 0.95.$

Стало быть, $(\bar{x} - t^*_{0.975}s, \bar{x} - t^*_{0.025}s)$ точный* 95% центральный доверительный интервал.

PDMI
COMPUTER
SCIENCE
CLUB

Как строить доверительные интервалы?

Распределение величины $\sqrt{n}t$ можно посчитать аналитически. Оно называется распределением Стьюдента с $n - 1$ степенью свободы. Это распределение похоже на нормальное, но имеет тяжелые хвосты.



PDMI
COMPUTER
SCIENCE
CLUB

Проверка гипотез

Гипотеза

Пусть $\mathfrak{P}$ — модель. Гипотеза — это утверждение вида $H\colon \mathcal{P}_X \in \mathfrak{P}_0 \subset \mathfrak{P}$.

Примеры:

- $\mathfrak{P} = \{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 \in (0, \infty)\}$, $\mathfrak{P}_0 = \{\mathcal{N}(0, \sigma^2) \mid \sigma^2 \in (0, \infty)\}$,
- $\mathfrak{P} = \left\{\mathcal{P} = \mathcal{P}_\xi^{\otimes n_1} \otimes \mathcal{P}_\eta^{\otimes n_2}\right\}$, $\mathfrak{P}_0 = \left\{\mathcal{P} = \mathcal{P}_\xi^{\otimes n_1} \otimes \mathcal{P}_\eta^{\otimes n_2} \,\middle|\, \mathbb{E}\xi = \mathbb{E}\eta\right\}$,
- $\mathfrak{P} = \left\{\mathcal{P}^{\otimes n} \middle| \mathcal{P} \in \mathbb{R}^2\right\}$, $\mathfrak{P}_0 = \left\{\mathcal{P} = \mathcal{P}_\xi^{\otimes n} \otimes \mathcal{P}_\eta^{\otimes n}\right\}$,
- $\mathfrak{P} = \left\{\mathcal{P}_{\{Y,X\}}^{\otimes n} \middle| Y = aX + b + \varepsilon, \mathbb{E}\varepsilon = 0\right\}$, $\mathfrak{P}_0 = \left\{\mathcal{P}_{\{Y,X\}}^{\otimes n} \middle| Y = b + \varepsilon, \mathbb{E}\varepsilon = 0\right\}$.

Если $|\mathfrak{P}_0| = 1$, то гипотеза называется простой, иначе сложной.

PDMI COMPUTER SCIENCE CLUB

Гипотеза

Гипотеза H_0, которую мы хотим проверить, называется нулевой гипотезой или просто гипотезой.

Проверка гипотезы — процесс принятия решения о том, противоречит ли она наблюдаемой выборке данных. Обычно, мы проверяем гипотезы, которые хотим опровергнуть.

В качестве нулевой гипотезы часто берут

- положение по умолчанию: мы выкатили супер обновление и хотим убедиться, что стало лучше, то есть хотим опровергнуть утверждение, что ничего не произошло.
- неблагоприятный исход: мы производим клетки для тигров и хотим убедиться, что они прочные, то есть хотим опровергнуть утверждение, что они хлипкие.

Альтернатива

Гипотеза может быть неверна по-разному, при этом не все отклонения от нее могут быть одинаково интересны. Чтобы понять, что же именно мы утверждаем нашей гипотезой, полезно после нее добавлять «а не»:

Гипотеза H_1, которая отражает, какие отклонения от нулевой гипотезы нам интересны, называется альтернативой.

PDMI COMPUTER SCIENCE CLUB

Еще пример

PDMI
COMPUTER
SCIENCE
CLUB

Критерий

Отображение $\varphi: d \mapsto \{\text{не отклоняем}, \text{отклоняем}\} = \{0,1\}$ называется критерием.

Часто критерий устроен так: имеется статистика критерия T, которая характеризует отклонение от гипотезы в сторону альтернативы, и критическое множество C, которое характеризует область противоречащих H_0 отклонений. При этом

$$\varphi(d) = [T(d) \in C] = [d \in T^{-1}(C)].$$

PDMI
COMPUTER
SCIENCE
CLUB

Значимость

Критическая область определяет, какие отклонения от нулевой гипотезы следует считать значимыми. Мы рассмотрим два вида значимости:

– статистическая значимость или просто значимость — характеризуется уровнем значимости $\mathbb{P}(T(D) \in C | H_0)$

  «Значимые — те, которые редко случаются»

– практическая значимость или значительность — характеризуется минимальным практически интересным размером эффекта: $C = (c_{min}, \infty)$, $\varphi(d) = 1 \Leftrightarrow T(d) > c_{min}$.

  «Значительные — те, которые сильно отклоняются от нуля»

PDMI COMPUTER SCIENCE CLUB

Статистическая значимость

Уровень значимости $\alpha := \mathbb{P}(\varphi(D) = 1 \mid H_0)$ обычно является параметром критерия, т.е. задавая его, мы определяем критическое множество C_α такое, что $\mathbb{P}(T(D) \in C_\alpha \mid H_0) = \alpha$.

Таким образом, для критерия определено целое семейство критических областей $\{C_\alpha | \alpha \in [0,1]\}$, при этом $C_\alpha \subset C_{\alpha'}$, если $\alpha < \alpha'$.

PDMI
COMPUTER
SCIENCE
CLUB

P-value

Уровень значимости является характеристикой критерия, а нам иногда хочется узнать, насколько значимы наши данные: вместо вердикта «отклонили»/«не отклонили» хочется оценить степень, насколько гипотеза противоречит наблюдаемым данным.

В качестве характеристики этого противоречия используют p-значение или p-value:

$$\text{p-value} := \arg\min\{\alpha \in [0,1] | T(d) \in C_\alpha\}.$$

То есть p-value это минимальное значение уровня значимости для данного значения статистики критерия, при котором H_0 может быть отвергнута.
Чем меньше p-value, тем больше гипотеза противоречит данным.

PDMI COMPUTER SCIENCE CLUB

Проверка гипотезы через значимость

1. выбираем H_0 и H_1,
2. выбираем критерий φ, чувствительный к отклонениям от H_0 в сторону H_1,
3. задаем уровень значимости α,
4. вычисляем p-value статистики критерия $T(d)$
5. сравниваем p-value с α: если меньше, то отклоняем гипотезу, иначе нет

PDMI
COMPUTER
SCIENCE
CLUB

Размер эффекта

Во многих ситуациях важно не только (и не столько) информация о p-value, но и величина наблюдаемого эффекта.

Размеры эффекта бывают разные и использование того или иного размера эффекта зависит от контекста.

Пример:

Витя собирает кубик Рубика в среднем за 6.7 секунд с $s/\sqrt{n} = 0.3$ секунд.

Проверяя гипотезу о равенстве среднего 6 секундам, можно сообщить:

- $\bar{x} - \mu_0 = 0.7$ секунд,
- $\frac{(\bar{x}-\mu_0)}{s/\sqrt{n}} = 2.33$,
- $\frac{\bar{x}-\mu_0}{\mu_0} \times 100\% = 11.67\%$.

PDMI COMPUTER SCIENCE CLUB

Проверка гипотезы через размер эффекта

1. выбираем H_0 и H_1,
2. выбираем эффект, которым мы будем измерять отклонение от H_0 к H_1,
3. выбираем минимальный практически значимый размер эффекта c,
4. вычисляем размер эффекта на данных
5. сравниваем размер эффекта с c, если больше, то отклоняем H_0, иначе нет

Значимость ≠ значительность

Эффект может быть значимо ненулевым, но с практической точки зрения слишком маленьким!

И наоборот, эффект может быть с практической точки зрения не нулевым, но статистически не значимым.

Ошибка первого рода

Событие $\varphi(D) = 1 \mid H_0$ называется ошибкой первого рода. В народе ошибку первого рода величают «Ложная тревога». $\mathbb{P}(\varphi(D) = 1 | H_0)$ обычно обозначают через α.

Пример: тест на коронавирус выдал положительный результат на здоровом человеке.

В теории уровень значимости критерия должен совпадать с вероятностью ошибки первого рода (такой критерий называется точным). На практике такое бывает не всегда.

PDMI
COMPUTER
SCIENCE
CLUB

Ошибка второго рода

Событие $\varphi(D) = 0 \mid H_1$ называется ошибкой второго рода. В народе ошибку второго рода величают «Пропуск цели». $\mathbb{P}(\varphi(D) = 0 | H_1)$ обычно обозначают через β.

Вероятность $1 - \beta$ отклонить H_0 при условии, что верна H_1, называется мощностью критерия.

PDMI
COMPUTER
SCIENCE
CLUB

Выбор критерия

Проверка гипотез значимость:

1. выбираем уровень значимости α
2. ищем критерий с наибольшей мощностью

Проверка гипотез через значительность:

1. выбираем минимальный размер эффекта (мощность ≈ 0.5)
2. ищем критерий с наименьшей ошибкой первого рода

Таблички, которые везде рисуют

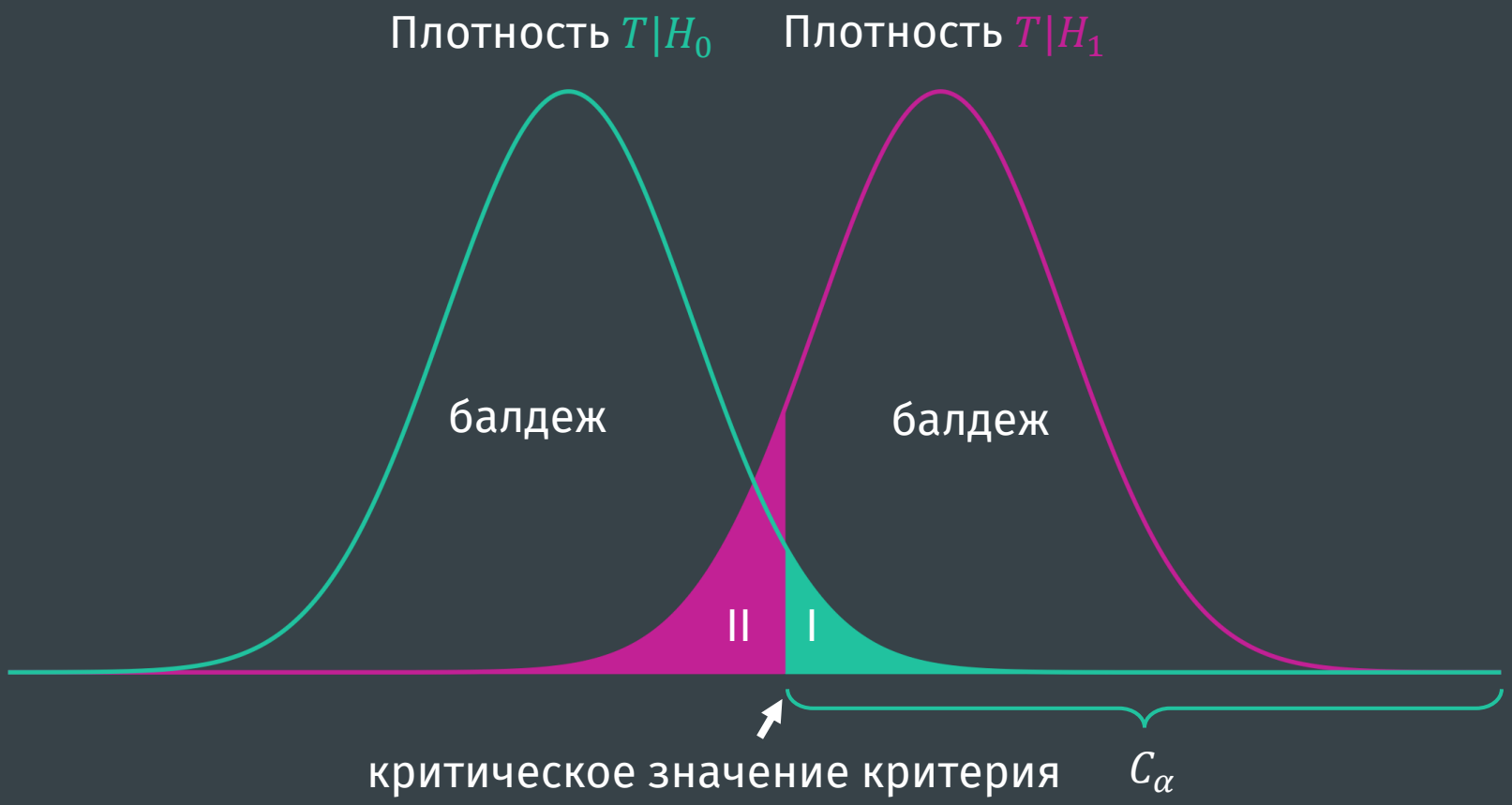
	Результат применения критерия	
Истина	H_0 не отвергнута	H_0 отвергнута
H_0	балдеж	ошибка I рода
H_1	ошибка II рода	балдеж

	Результат применения критерия	
Истина	H_0 не отвергнута	H_0 отвергнута
H_0	True Negative (TN)	False Positive (FP)
H_1	False Negative (FN)	True Positive (TP)

PDMI COMPUTER SCIENCE CLUB

Картинка, которую везде рисуют

$H_0: X \sim \mathcal{N}(\mu_0, \sigma^2)$, $H_1: X \sim \mathcal{N}(\mu_1, \sigma^2)$, $T(x_{[n]}) = \bar{x}$, $T|H_i \sim \mathcal{N}(\mu_i, \sigma^2)$.



PDMI COMPUTER SCIENCE CLUB

Точность и мощность

Точность и мощность это неотъемлемые составляющие качества критерия!

Важно не только выбрать критерий так, чтобы ошибка первого рода была равна заявленному уровню значимости α, но и чтобы мощность была как можно больше.

Критерий, который немного врет с уровнем значимости, но имеет хорошую мощность, гораздо лучше критерия, который выдает абсолютно честную ошибку первого рода, но при этом имеет околонулевую мощность.

Что делать со сложными гипотезами?

Если гипотеза сложная, то ошибка первого рода не определена, кроме случая, когда $\mathbb{P}(T(D) \in C | D \sim \mathcal{P}) = \alpha$ для всех $\mathcal{P} \in \mathfrak{P}_0$. С альтернативой аналогично.

Тем не менее, часто полагают $\alpha = \sup\{\mathbb{P}(T(D) \in C | D \sim \mathcal{P}) \mid \mathcal{P} \in \mathfrak{P}_0\}$,
$\beta = \sup\{\mathbb{P}(T(D) \notin C | D \sim \mathcal{P}) \mid \mathcal{P} \in \mathfrak{P}_1\}$.
Соответственно, p-value для сложной гипотезы это максимальное значение p-value по простым гипотезам из нее.

PDMI
COMPUTER
SCIENCE
CLUB

Какие бывают гипотезы?

Простая гипотеза и простая альтернатива

Самый простой вариант: гипотеза состоит из одного распределения $\mathcal{P}_0$ и альтернатива состоит из одного распределения $\mathcal{P}_1$.

В этом случае мы умеем строить наиболее мощный критерий для заданного уровня значимости!

Простая гипотеза и простая альтернатива

Пусть $d = x_{[1]}$, а $\alpha = 0.1$. Какое нужно выбирать C_α, чтобы мощность была максимальна?

Подсказка: Пусть $C_\alpha = \{7\}$. Чему равны ошибка первого рода и мощность?

распределения $x_{[1]}|H_0$ и $x_{[1]}|H_1$

PDMI COMPUTER SCIENCE CLUB

Простая гипотеза и простая альтернатива

Если $\mathcal{P}_0$ непрерывно, то наиболее мощный критерий является критерием отношением правдоподобия:

- $T(x_{[n]}) = \frac{p(x_{[n]}|H_1)}{p(x_{[n]}|H_0)}$,
- $C_\alpha = (c, \infty)$, где c такая, что $\mathbb{P}(T(X_{[n]}) > c|H_0) = \alpha$.

Как искать c? Можно аналитически, а можно и замонтекарлить. С вычислительной точки зрения удобнее работать с

$$\log T(x_{[n]}) = \log p(x_{[n]}|H_1) - \log p(x_{[n]}|H_0) = \sum \log p(x_i|H_1) - \sum \log p(x_i|H_0).$$

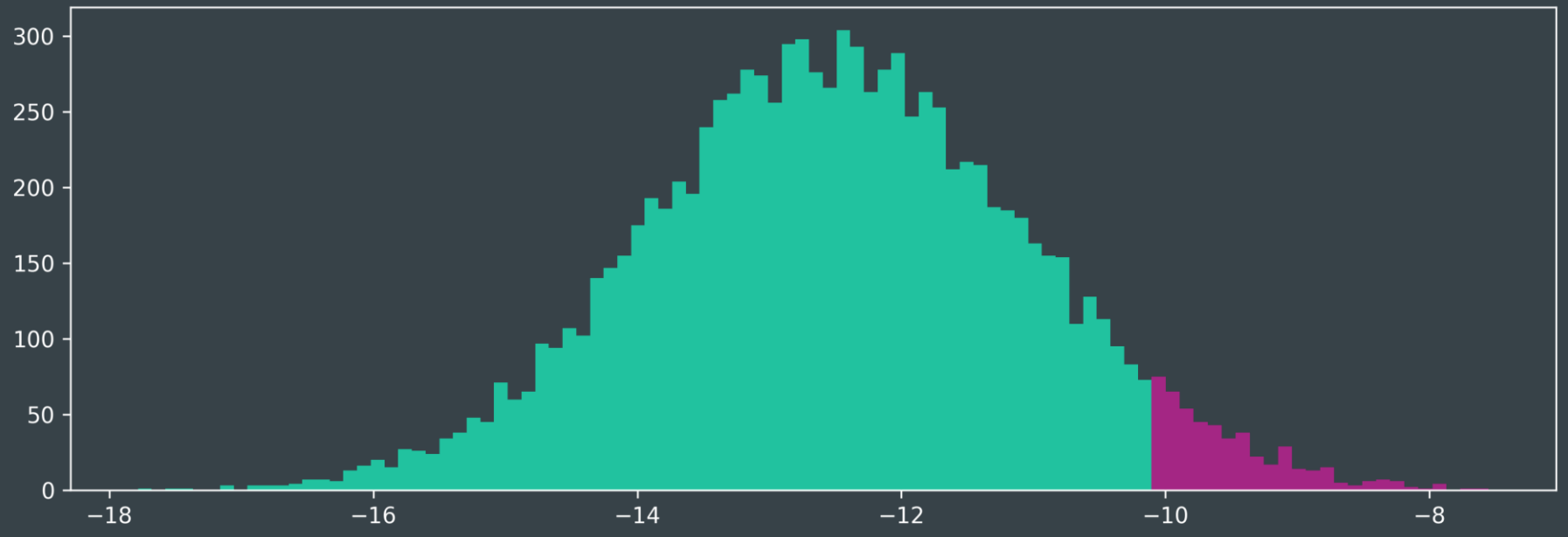
PDMI COMPUTER SCIENCE CLUB

Пример

Пусть $\mathcal{P}_0 = U([0,1])$, $\mathcal{P}_1 = Exp(1)$, $n = 25$, $\alpha = 0.05$.

Сгенерируем N выборок объема n из $\mathcal{P}_0$; для каждой из них посчитаем $\log T$; у полученной выборки найдем $1 - \alpha$ квантиль.

Получилось $\log c \approx -10.1$.



PDMI
COMPUTER
SCIENCE
CLUB

Гипотезы о характеристиках

Пусть ϕ это характеристика распределения F_X.

Нулевая гипотеза имеет вид: $H_0: \phi = \phi_0$.

Типичные альтернативы:

- $H_1: \phi = \phi_1 \neq \phi_0$,
- $H_>: \phi > \phi_0$,
- $H_<: \phi < \phi_0$,
- $H_{\neq}: \phi \neq \phi_0$.

Гипотезы о характеристиках

Рассмотрим два способа разбираться с такими гипотезами.

Первый:

1. Выбирается оценка ϕ^* параметра ϕ, распределение которой при условии H_0 (приближенно) известно.
2. В зависимости от альтернативы строится критическое множество:

   $H_1: \phi = \phi_1 > \phi_0$ или $H_>$: $C_\alpha = (\phi^*_{1-\alpha}, \infty)$ — правое критическое множество;

   $H_1: \phi = \phi_1 < \phi_0$ или $H_<$: $C_\alpha = (-\infty, \phi^*_\alpha)$ — левое критическое множество;

   $H_{\neq}$: $C_\alpha = \left(-\infty, \phi^*_{\alpha/2}\right) \cup \left(\phi^*_{1-\alpha/2}, \infty\right)$ — двустороннее критическое множество.
3. Если $\phi^*\left(x_{[n]}\right) \in C_\alpha$, то гипотезу можно отклонить, иначе нельзя.

ϕ^*_x — это квантиль уровня x распределения $\theta^*|H_0$.

PDMI COMPUTER SCIENCE CLUB

Бутстрап из нулевой гипотезы

Непараметрический:

1. Назначим каждому наблюдению x_i в выборке вероятность p_i.
2. Из пар (x_i, p_i) изготовим дискретное распределение F_p^* (F_n^* это частный случай при $p = \left(\frac{1}{n}, \dots, \frac{1}{n}\right)$).
3. Подберем p_i так, чтобы, с одной стороны, $\phi(F_p^*) = \phi_0$, а с другой, чтобы p_i максимизировали правдоподобие выборки $p(x_{[n]}|p) = p_1 p_2 \cdots p_n$.
4. Набутстрапим кучу выборок из получившегося F_p^* и посчитаем по ним $\phi_{[N]}^*$.
5. Построим критическое множество в зависимости от альтернативы и проверим, лежит ли в нем $\phi^*(x_{[n]})$.

Бутстрап из нулевой гипотезы

Параметрический:

1. Выберем параметр θ так, чтобы, с одной стороны, $\phi(F_\theta) = \phi_0$, а с другой, чтобы θ максимизировал правдоподобие выборки $p\left(x_{[n]}\middle|\theta\right)$.
2. Набутстрапим кучу выборок из получившегося F_θ и посчитаем по ним $\phi^*_{[N]}$.
3. Построим критическое множество в зависимости от альтернативы и проверим, лежит ли в нем $\phi^*\left(x_{[n]}\right)$.

Пример

Метод максимального правдоподобия

Чтобы найти лучшую модель $\mathcal{M} \in \mathfrak{M}_0$, придется максимизировать

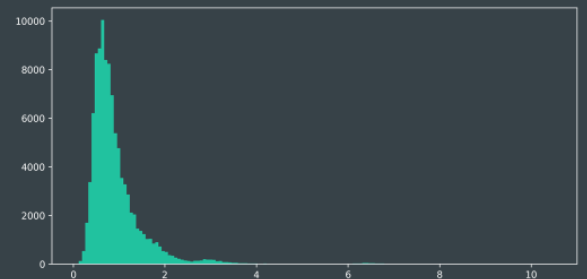
$$\prod_{x,w} (p_x^w)^{n_x^w} \to \max_p,$$

с ограничением $\sum_{x,y,z} reward(x,y,z) p_x^1 p_y^2 p_z^3 = 0.92$.

Не совсем понятно, как это делать аналитически, но можно написать программу, которая посчитает результат численно!

Проверка гипотезы

Теперь, когда мы нашли лучшую модель из $\mathfrak{M}_0$ мы можем оценить распределение среднего выборки объема 138.



Метод максимального правдоподобия

Вот что получилось!

Вероятности лучшей модели из $\mathfrak{M} \times 138$ (то есть исходные кратности)

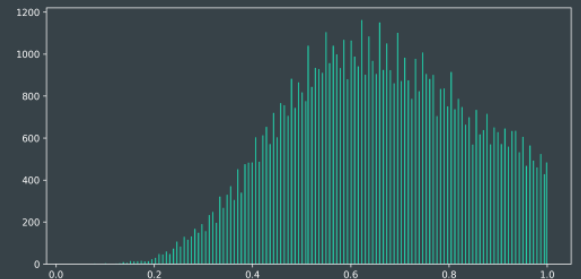
Окно	0	bar	bar x2	bar x3	7	🍒	💎💎
1	59	49	14	6	6	1	3
2	85	8	24	16	4	0	1
3	77	39	6	1	7	3	5

Вероятности лучшей модели из $\mathfrak{M}_0 \times 138$

Окно	0	bar	bar x2	bar x3	7	🍒	💎💎
1	56	49.7	14.6	6.2	5.9	1.1	4.4
2	80	8.7	25.5	16.9	4.1	0	1.9
3	72.9	40.1	6.4	1.1	6.8	3.3	7.4

Проверка гипотезы

Рассмотрим только часть, левее 1. Видно, что средние довольно регулярно бывают меньше 0.4. Доля средних меньших 0.384 составила ≈ 0.0469.



PDMI COMPUTER SCIENCE CLUB

Гипотезы о характеристиках

Второй способ:

1. В зависимости от альтернативы строится доверительный интервал с уровнем доверия $\gamma = 1 - \alpha$:

   $H_1: \phi = \phi_1 > \phi_0$ или $H_>$: (ϕ_L^*, ∞) — **правый** доверительный интервал;

   $H_1: \phi = \phi_1 < \phi_0$ или $H_<$: $C_\alpha = (-\infty, \phi_R^*)$ — **левый** доверительный интервал;

   $H_{\neq}$: $C_\alpha = (\phi_L^*, \phi_R^*)$ — центральный доверительный интервал.
2. Если θ_0 не лежит в интервале, то гипотезу можно отклонить, иначе нельзя

В качестве примера можно рассмотреть эфронов доверительный интервал.

Беды с проверкой гипотез через значимость

Множественные сравнения

Если вместо одной гипотезы вы проверяете k гипотез с уровнем значимости α, то в среднем αk гипотез будут отвергнуты, даже если они все верны.

Что делать?

– Избегать множественных сравнений
– Поправка Бонферрони: уровень значимости меняем на α/k. Более мощный вариант: поправка Холма—Бонферрони и иже с ними.

P-hacking

Если вы смотрите в данные, то даже если вам кажется, что вы не проводите множественные сравнения, скорее всего вы их все-таки проводите.
При большом упорстве вы обязательно найдете что-то значимое даже в шуме!

Что делать?

- ~~Не смотреть в данные~~
- Проверять гипотезы кучей, а не по очереди
- Проверять найденные отклонения на новых данных

PDMI
COMPUTER
SCIENCE
CLUB