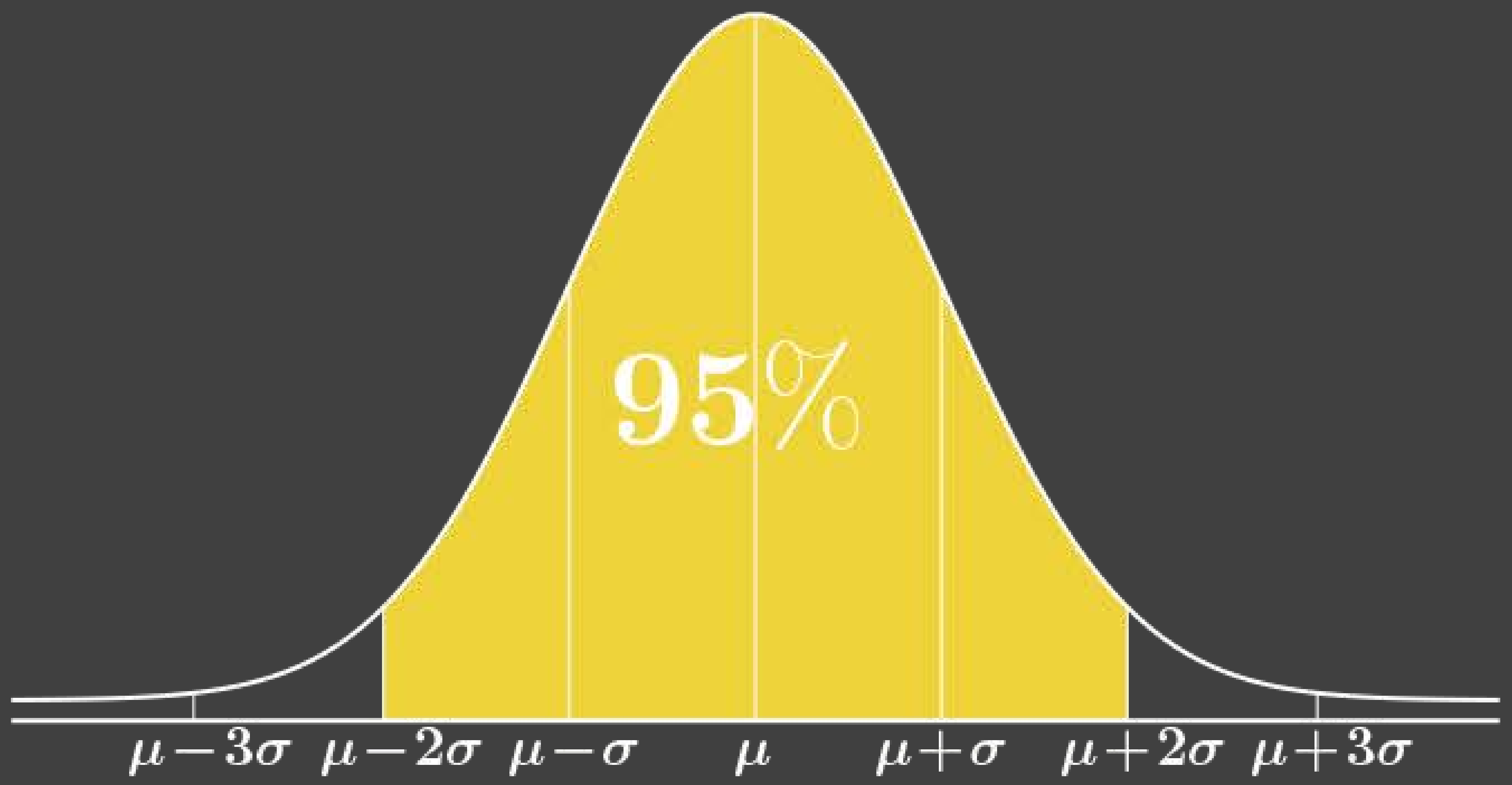
ИНТЕРВАЛЬНЫЕ ОЦЕНКИ

ПРАВИЛО ДВУХ СИГМ

$$X \sim N\left(\mu, \sigma^2\right) \Rightarrow \quad \mathrm{P}(\mu - 2\sigma \le X \le \mu + 2\sigma) \approx 0.95$$



Можно получить более точную оценку для любой вероятности

Квантиль

- Квантиль порядка $\alpha \in (0, 1)$ — такое число X_α, что:

$$\mathrm{P}(X \le X_\alpha) \ge \alpha$$

$$\mathrm{P}(X \ge X_\alpha) \ge 1 - \alpha$$

- Функция распределения X:

$$F(x) = \mathrm{P}(X \le x) \Rightarrow$$

- Эквивалентное определение квантиля:

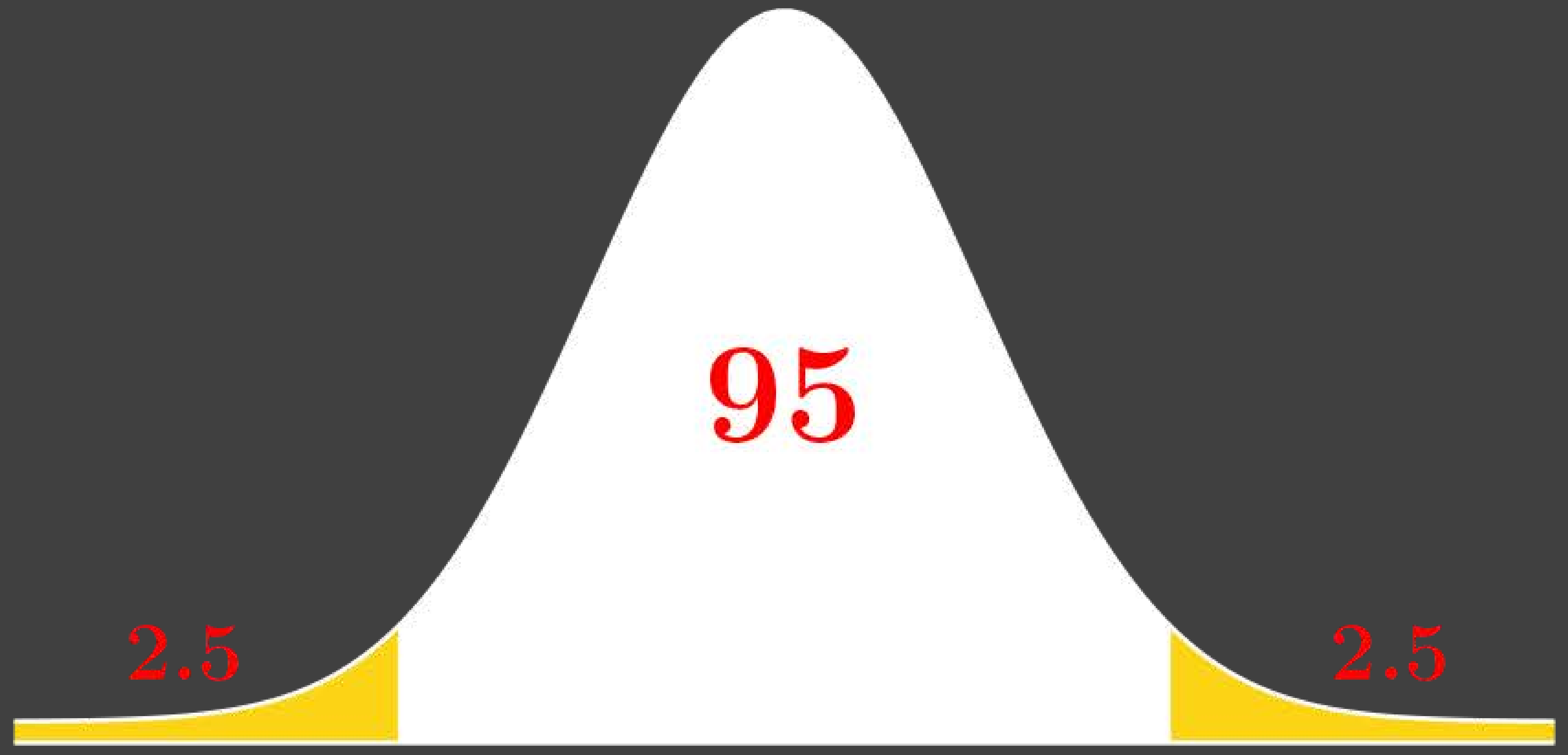
$$X_\alpha = F^{-1}(\alpha) = \inf\{x \colon F(x) \ge \alpha\}$$

Уточнение правила двух сигм

$$\mathrm{P}(? \leq X \leq ?) = 0.95$$

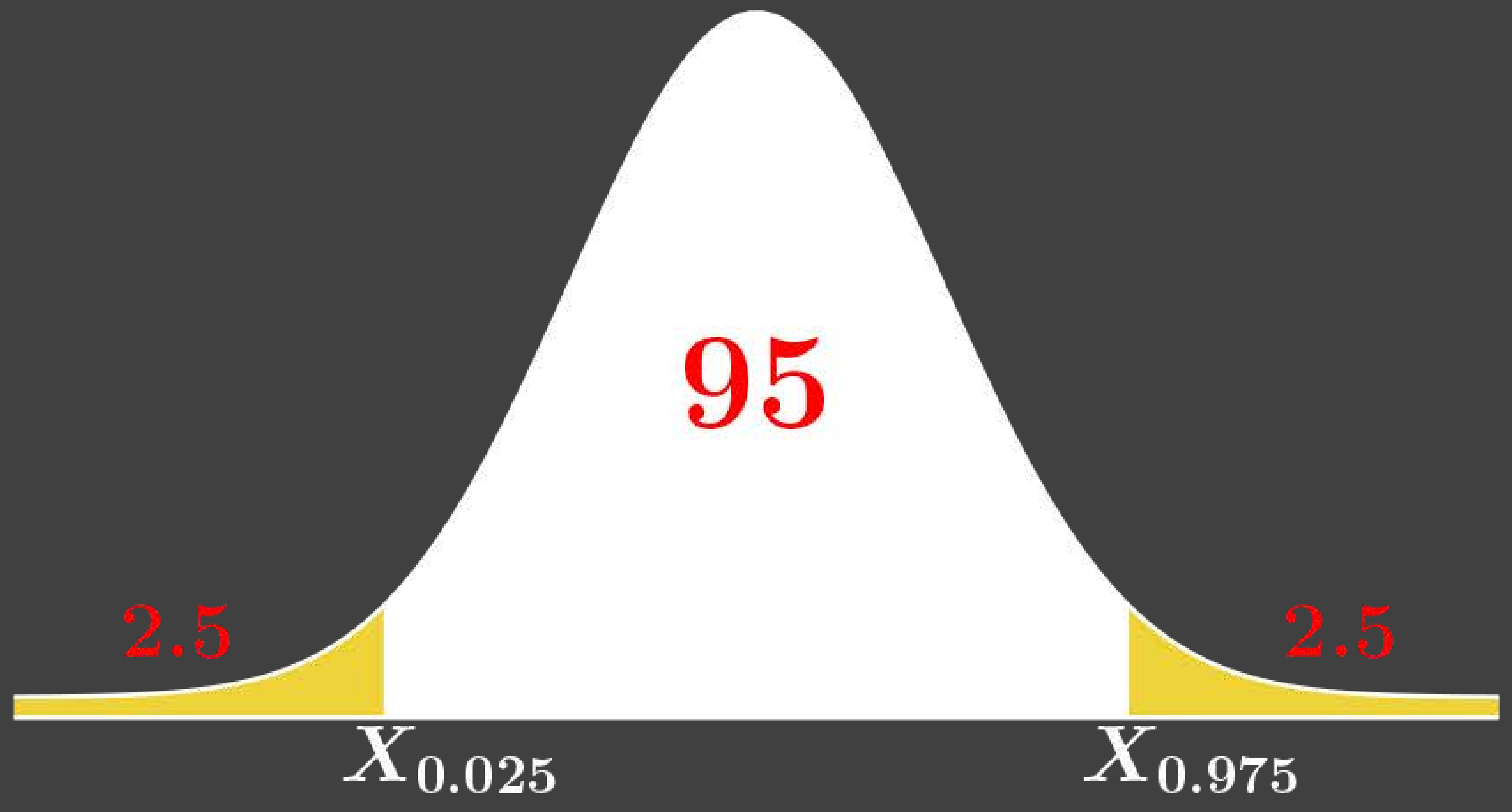
Уточнение правила двух сигм

$$\mathrm{P}(? \leq X \leq ?) = 0.95$$



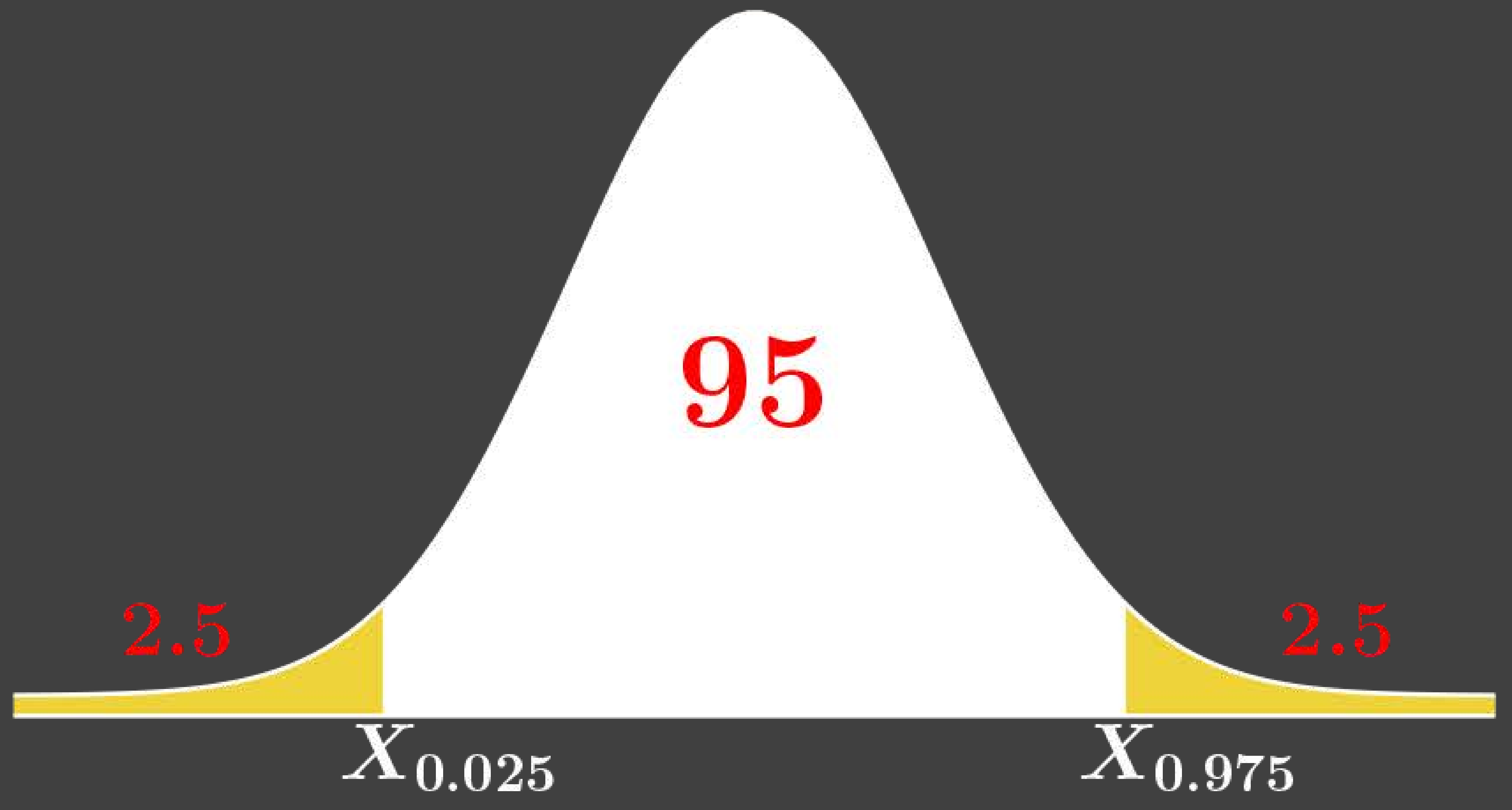
Уточнение правила двух сигм

$P(? \leq X \leq ?) = 0.95$



Уточнение правила двух сигм

$$\mathrm{P}(X_{0.025} \le X \le X_{0.975}) = 0.95$$



Предсказательный интервал

- $X \sim F(x) \Rightarrow \quad \mathrm{P}\left(X_{\frac{\alpha}{2}} \leq X \leq X_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$

  $\left[X_{\frac{\alpha}{2}}, X_{1-\frac{\alpha}{2}}\right]$ — предсказательный интервал порядка $1 - \alpha$

Предсказательный интервал

- $X \sim F(x) \Rightarrow \quad \mathrm{P}\left(X_{\frac{\alpha}{2}} \leq X \leq X_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$
  $\left[X_{\frac{\alpha}{2}}, X_{1-\frac{\alpha}{2}}\right]$ — предсказательный интервал порядка $1 - \alpha$
- $X \sim N\left(\mu, \sigma^2\right) \Rightarrow$
  $\mathrm{P}\left(\mu - z_{1-\frac{\alpha}{2}}\sigma \leq X \leq \mu + z_{1-\frac{\alpha}{2}}\sigma\right) = 1 - \alpha$
  z_α — квантиль стандартного нормального распределения $N\left(0, 1\right)$
  $z_{0.975} \approx 1.95996 \approx 2$

РЕЗЮМЕ

- Использование квантилей для построения интервальных оценок
- Предсказательный интервал
- Уточнение правила двух сигм
- Далее: доверительные интервалы

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

ТОЧЕЧНЫЕ ОЦЕНКИ

- $X \sim F(x, \theta)$, θ — неизвестный параметр
- $\theta = ?$

Точечные оценки

- $X \sim F(x, \theta)$, θ — неизвестный параметр
- $\theta = ?$
- $X^n = (X_1, \ldots, X_n)$
- $\hat{\theta}$ — оценка θ по выборке

Точечные оценки

- $X \sim F(x, \theta)$, θ — неизвестный параметр
- $\theta = ?$
- $X^n = (X_1, \ldots, X_n)$
- $\hat{\theta}$ — оценка θ по выборке

- Например, для $\theta = \mathbb{E}X$:

$$\hat{\theta} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$ — хорошая оценка

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

- Доверительный интервал для параметра θ — пара таких статистик C_L, C_U, что

$$\mathrm{P}(C_L \leq \theta \leq C_U) \geq 1 - \alpha$$

- Как оценить C_L и C_U по выборке?

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

- $\mathrm{P}(C_L \le \theta \le C_U) \ge 1 - \alpha$

- Как оценить C_L и C_U по выборке?

- Если $\hat{\theta}$ — оценка θ и мы знаем её распределение $F_{\hat{\theta}}(x)$, то:

  $$\mathrm{P}\left(F_{\hat{\theta}}^{-1}\left(\frac{\alpha}{2}\right) \le \theta \le F_{\hat{\theta}}^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

Для нормального распределения

- $X \sim N\left(\mu, \sigma^2\right)$, $X^n = (X_1, \ldots, X_n)$
- $\bar{X}_n$ — оценка $\mathbb{E}X = \mu$

Для нормального распределения

- $X \sim N\left(\mu, \sigma^2\right)$, $X^n = (X_1, \ldots, X_n)$
- $\bar{X}_n$ — оценка $\mathbb{E}X = \mu$
- $\bar{X}_n \sim N\left(\mu, \dfrac{\sigma^2}{n}\right) \Rightarrow$

ДЛЯ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

- $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow$
- Предсказательный интервал для $\bar{X}_n$:

$$\mathrm{P}\left(\mu - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Для нормального распределения

- Предсказательный интервал для $\bar{X}_n$:

$$\mathrm{P}\left(\mu - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \bar{X}_n \le \mu + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

- Доверительный интервал для μ:

$$\mathrm{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

ДОВЕРИТЕЛЬНЫЙ И ПРЕДСКАЗАТЕЛЬНЫЙ ИНТЕРВАЛЫ

- $X \sim N\left(\mu, \sigma^2\right),\ X^n = (X_1, \ldots, X_n)$
-

ДОВЕРИТЕЛЬНЫЙ И ПРЕДСКАЗАТЕЛЬНЫЙ ИНТЕРВАЛЫ

- $X \sim N\left(\mu, \sigma^2\right)$, $X^n = (X_1, \ldots, X_n)$

- Предсказательный интервал для X:

$$\mathrm{P}\left(\mu - z_{1-\frac{\alpha}{2}}\sigma \leq X \leq \mu + z_{1-\frac{\alpha}{2}}\sigma\right) = 1 - \alpha$$

Доверительный и предсказательный интервалы

- $X \sim N\left(\mu, \sigma^2\right)$, $X^n = (X_1, \ldots, X_n)$
- Предсказательный интервал для X:

$$\mathrm{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\sigma \leq X \leq \bar{X}_n + z_{1-\frac{\alpha}{2}}\sigma\right) = 1 - \alpha$$

ДОВЕРИТЕЛЬНЫЙ И ПРЕДСКАЗАТЕЛЬНЫЙ ИНТЕРВАЛЫ

- Предсказательный интервал для X:

$$\mathrm{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\sigma \leq X \leq \bar{X}_n + z_{1-\frac{\alpha}{2}}\sigma\right) = 1 - \alpha$$

- Доверительный интервал для μ:

$$\mathrm{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Для любого распределения

- $X \sim F(x)$, $X^n = (X_1, \ldots, X_n)$
- $\bar{X}_n$ — оценка $\mathbb{E}X$
- $\bar{X}_n \approx\sim N\left(\mathbb{E}X, \frac{\mathbb{D}X}{n}\right)$ (ЦПТ) $\Rightarrow$

ДЛЯ ЛЮБОГО РАСПРЕДЕЛЕНИЯ

- $\bar{X}_n \approx\sim N\left(\mathbb{E}X, \frac{\mathbb{D}X}{n}\right)$ (ЦПТ) $\Rightarrow$
- Доверительный интервал для $\mathbb{E}X$:

$$\mathrm{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\mathbb{D}X}{n}} \le \mathbb{E}X \le \bar{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\mathbb{D}X}{n}}\right) \approx 1-\alpha$$

РЕЗЮМЕ

- Построение доверительных интервалов
- Разница между доверительными и предсказательными интервалами
- Интервалы для нормального распределения

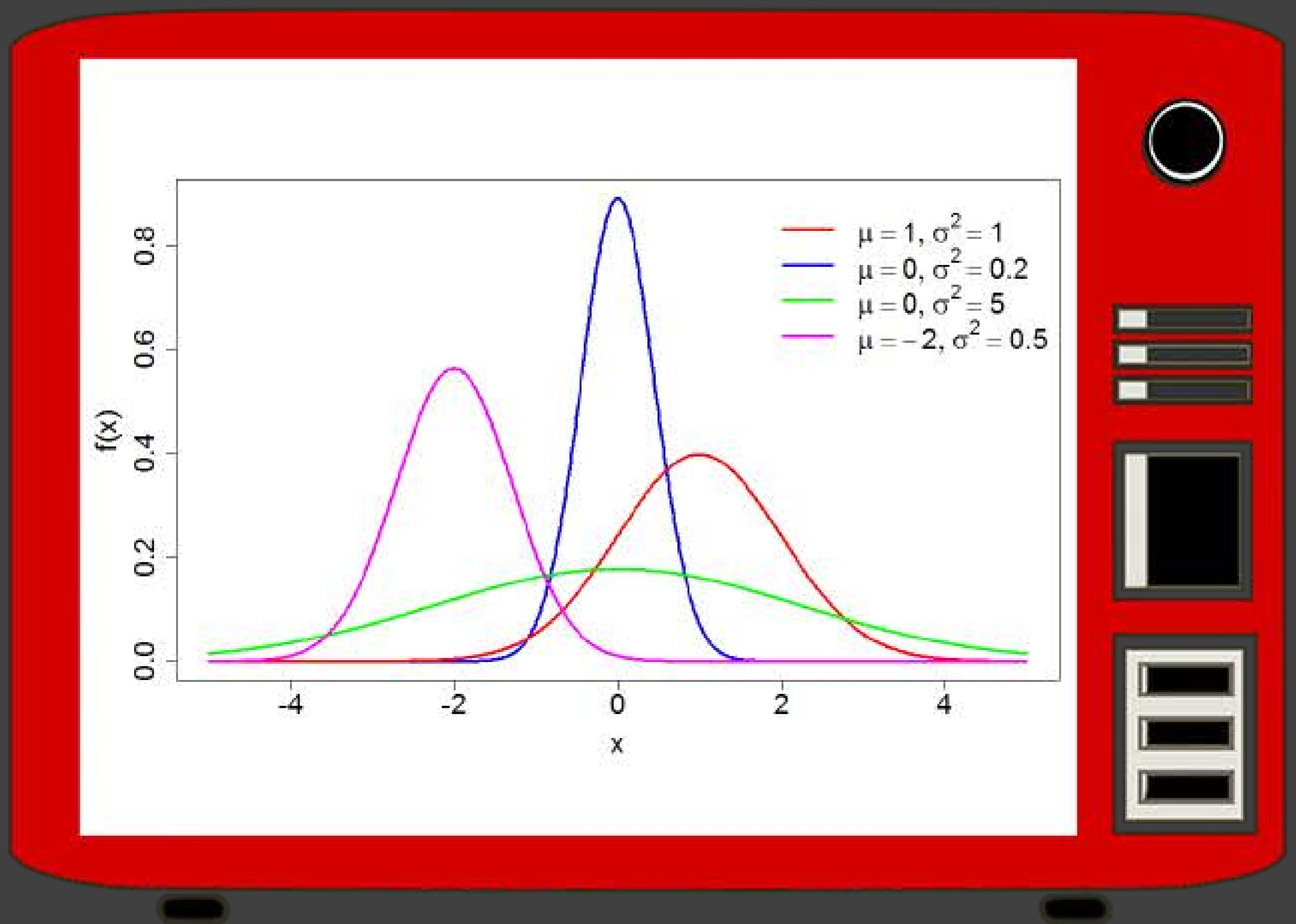
РЕЗЮМЕ

- Разница между доверительными и предсказательными интервалами
- Интервалы для нормального распределения
- Далее: распределения, производные от нормального

Распределения, производные от нормального

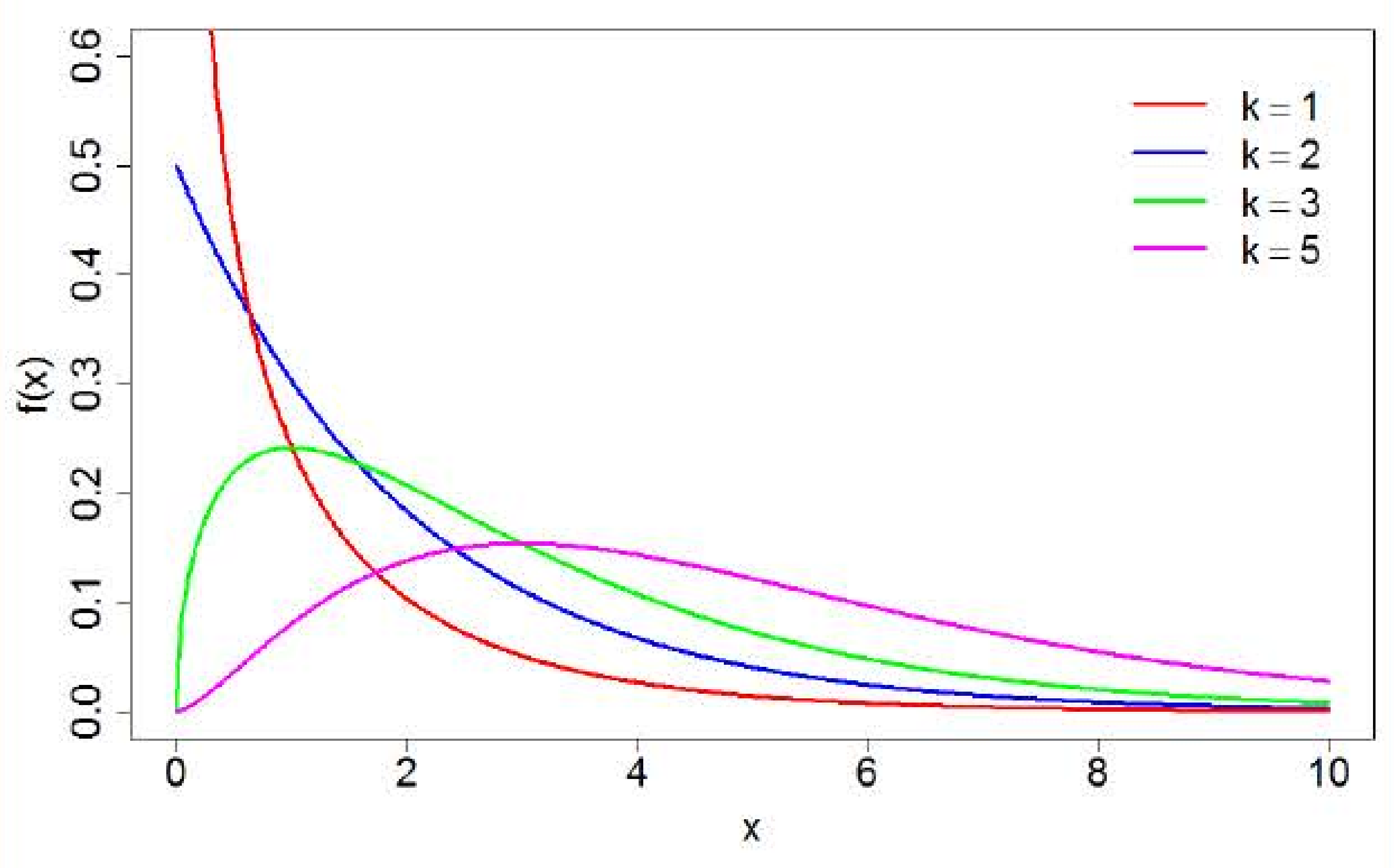
Нормальное распределение

- $X \sim N\left(\mu, \sigma^2\right)$
- $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $F(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$


μ = 1, σ² = 1
μ = 0, σ² = 0.2
μ = 0, σ² = 5
μ = −2, σ² = 0.5
f(x)
x
0.0
0.2
0.4
0.6
0.8
-4
-2
0
2
4

Распределение хи-квадрат

- $X_1, X_2, \ldots, X_k \sim N(0,1)$ независимы
- $X = \sum\limits_{i=1}^{k} X_i^2 \sim \chi_k^2$ — распределение хи-квадрат с k степенями свободы



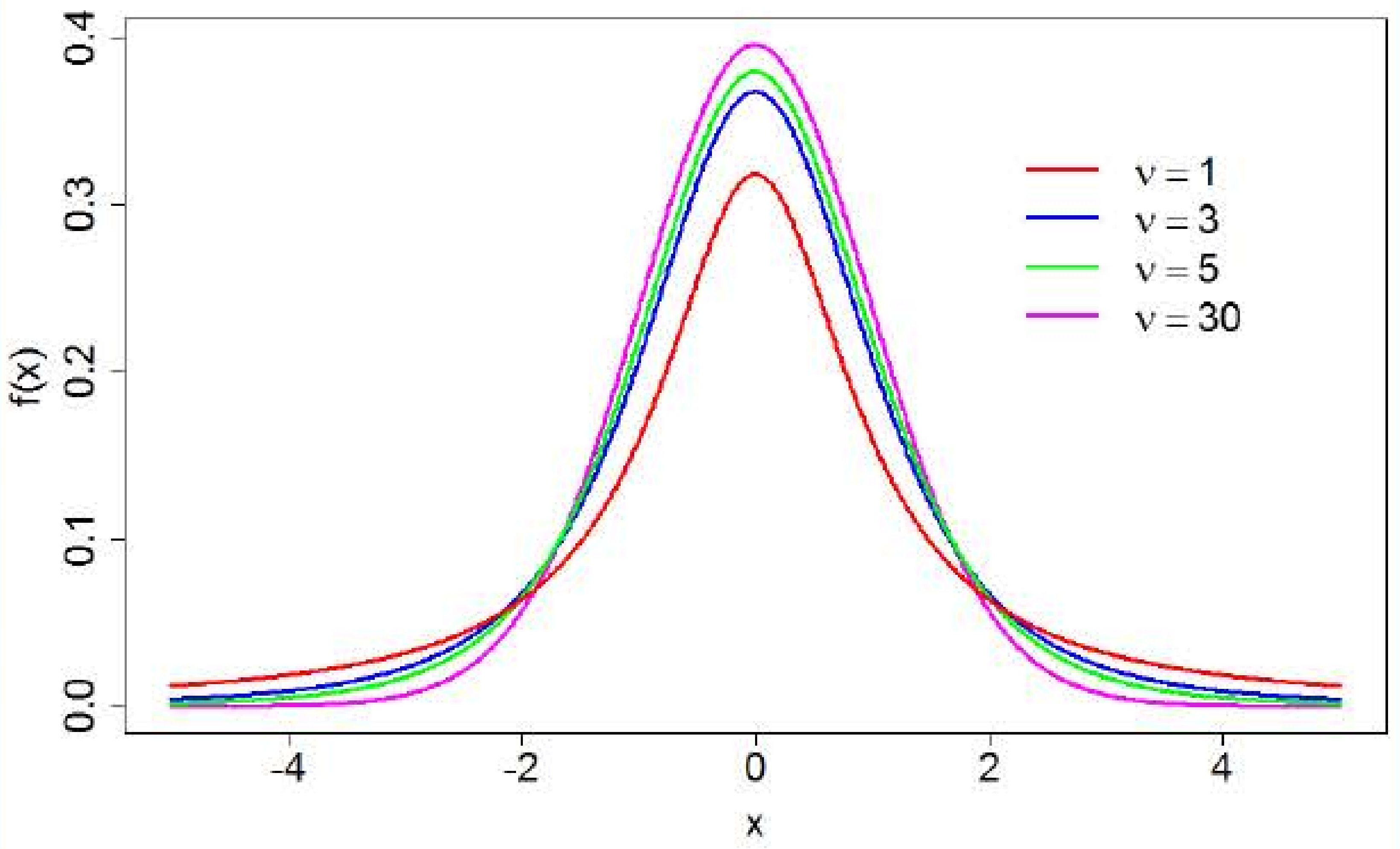
k = 1
k = 2
k = 3
k = 5
f(x)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
10
x

Распределение Стьюдента

- $X_1 \sim N(0,1)$, $X_2 \sim \chi^2_\nu$ независимы
- $X = \dfrac{X_1}{\sqrt{X_2/\nu}} \sim St(\nu)$ — распределение Стьюдента с ν степенями свободы

Распределение Стьюдента

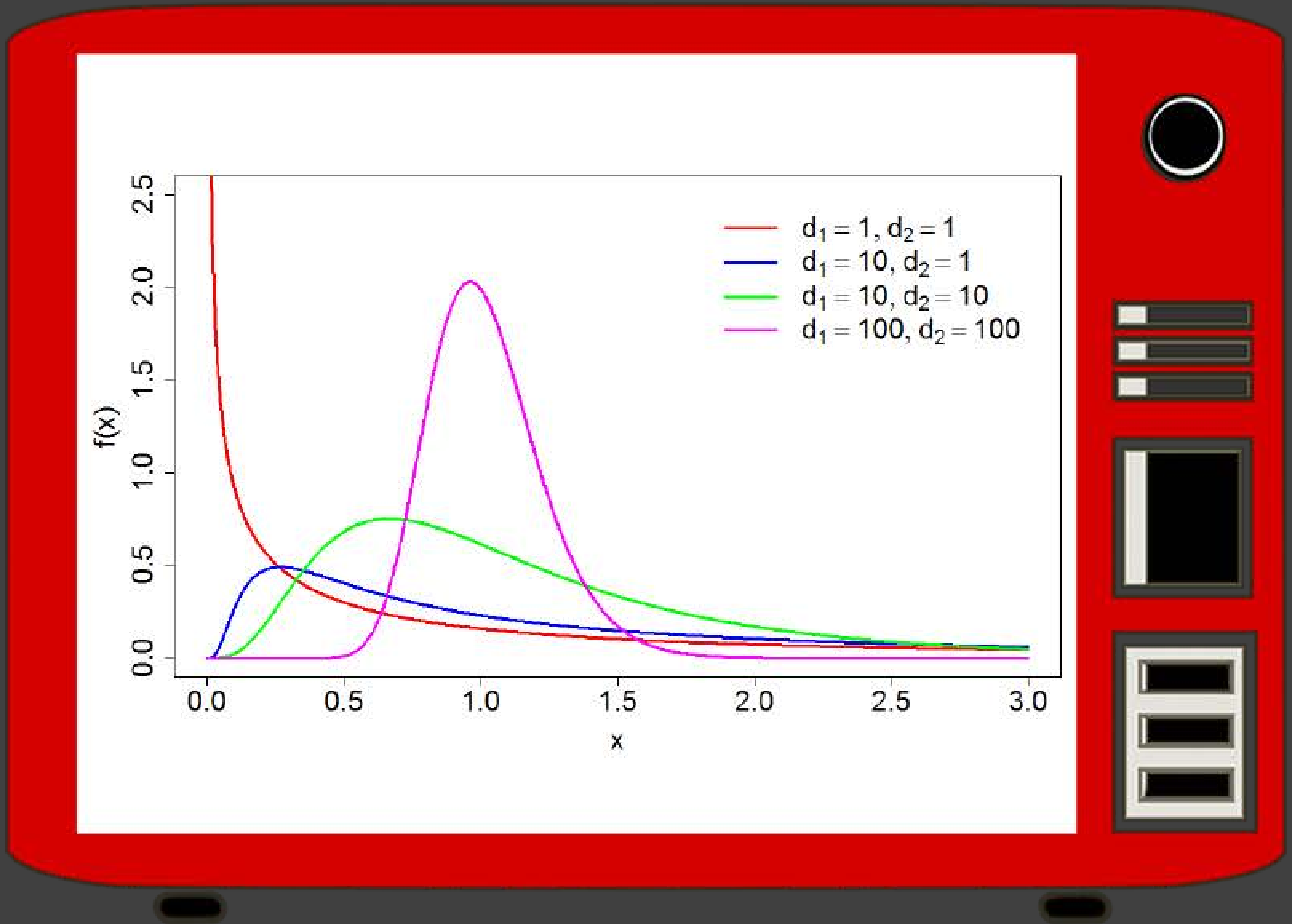
- При больших ν очень похоже на $N(0, 1)$



0.0
0.1
0.2
0.3
0.4
f(x)
-4
-2
0
2
4
x
ν = 1
ν = 3
ν = 5
ν = 30

Распределение Фишера

- $X_1 \sim \chi^2_{d_1}$, $X_2 \sim \chi^2_{d_2}$ независимы
- $X = \dfrac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2)$ — распределение Фишера с d_1, d_2 степенями свободы



f(x)
x
0.0
0.5
1.0
1.5
2.0
2.5
3.0
$d_1 = 1, d_2 = 1$
$d_1 = 10, d_2 = 1$
$d_1 = 10, d_2 = 10$
$d_1 = 100, d_2 = 100$

ЗАЧЕМ?

- $X \sim N\left(\mu, \sigma^2\right)$, $X^n = (X_1, \ldots, X_n)$

ЗАЧЕМ?

- $X \sim N\left(\mu, \sigma^2\right)$, $X^n = (X_1, \ldots, X_n)$
- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

ЗАЧЕМ?

- $X \sim N\left(\mu, \sigma^2\right)$, $X^n = (X_1, \ldots, X_n)$
- $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$
- $S_n^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$

ЗАЧЕМ?

- $X \sim N\left(\mu, \sigma^2\right), X^n = (X_1, \ldots, X_n)$
- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- $(n-1)\frac{S_n^2}{\sigma^2} = \sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 \sim \chi^2_{n-1}$

ЗАЧЕМ?

- $(n-1)\frac{S_n^2}{\sigma^2} = \sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 \sim \chi^2_{n-1}$
- $T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim St\,(n-1)$

ЗАЧЕМ?

- $X_1 \sim N\left(\mu_1, \sigma_2^2\right)$, $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$
- $X_2 \sim N\left(\mu_2, \sigma_2^2\right)$, $X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$
- $\dfrac{S_1^2}{S_2^2} \sim F\left(n_1 - 1, n_2 - 1\right)$

РЕЗЮМЕ

- Распределения хи-квадрат, Стьюдента и Фишера
- Их связь со статистиками нормальных выборок
- Далее: доверительные интервалы для среднего и дисперсии

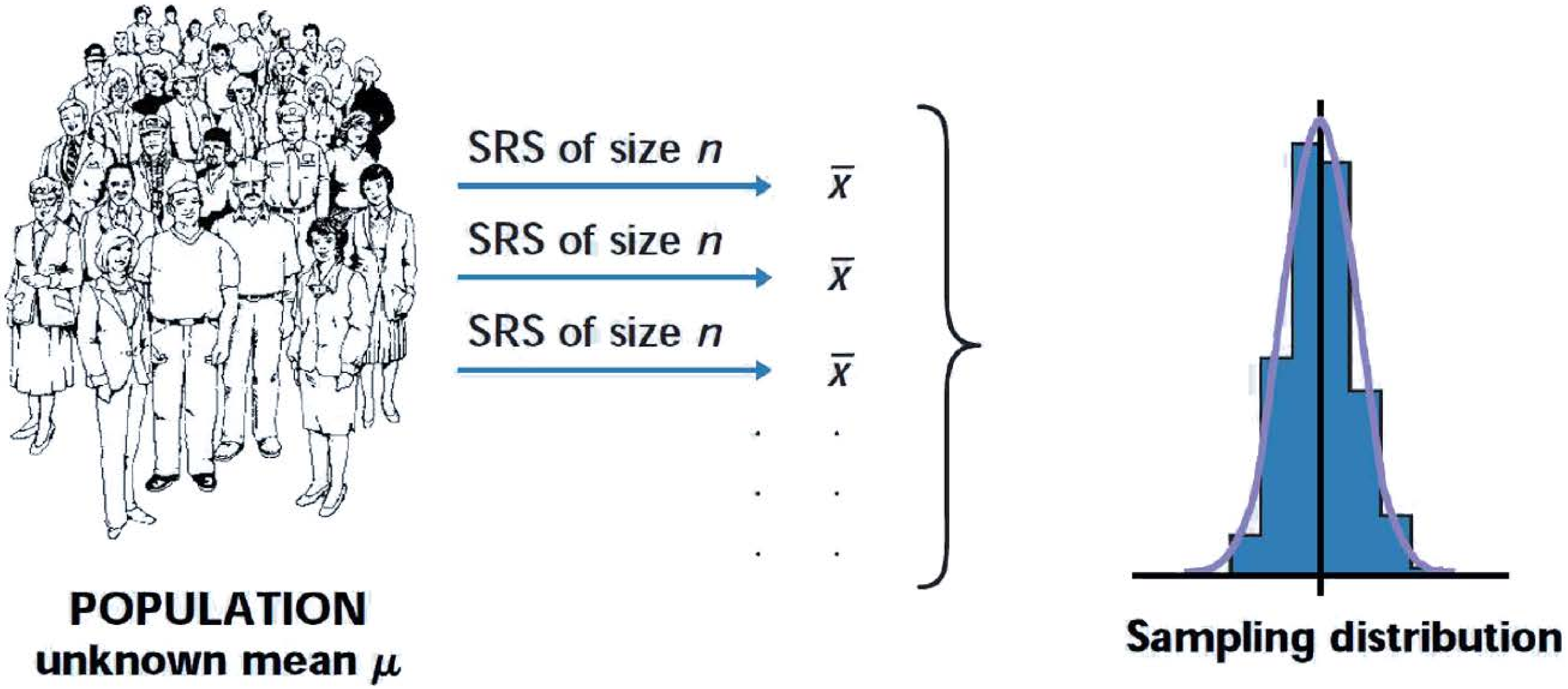
ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ НА ОСНОВЕ БУТСТРЕПА

ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

- Чтобы построить доверительный интервал для статистики $T_n = T(X^n)$, нужно знать её выборочное распределение $F_{T_n(x)}$
- Как его оценить?

ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

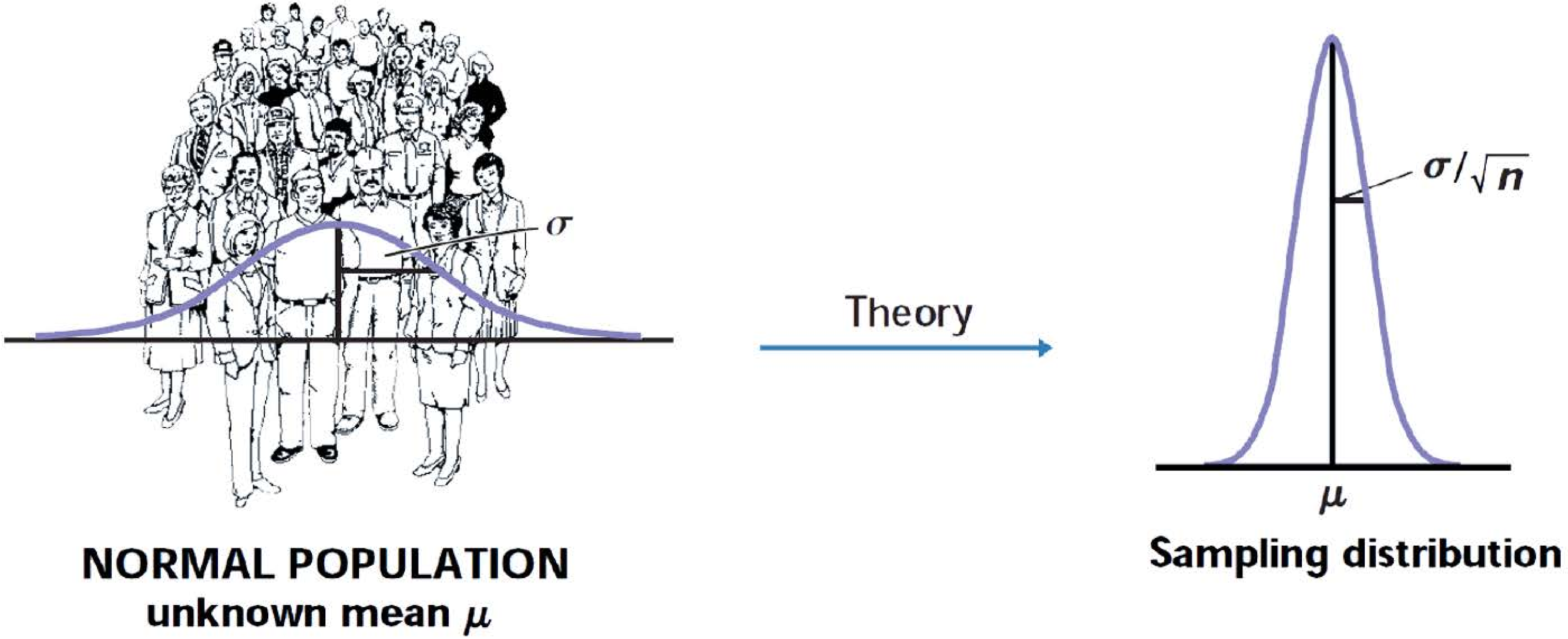
- Наивный метод:
  - Извлечь из генеральной совокупности N выборок объёма n и оценить выборочное распределение T_n эмпирическим



ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

- Параметрический метод:

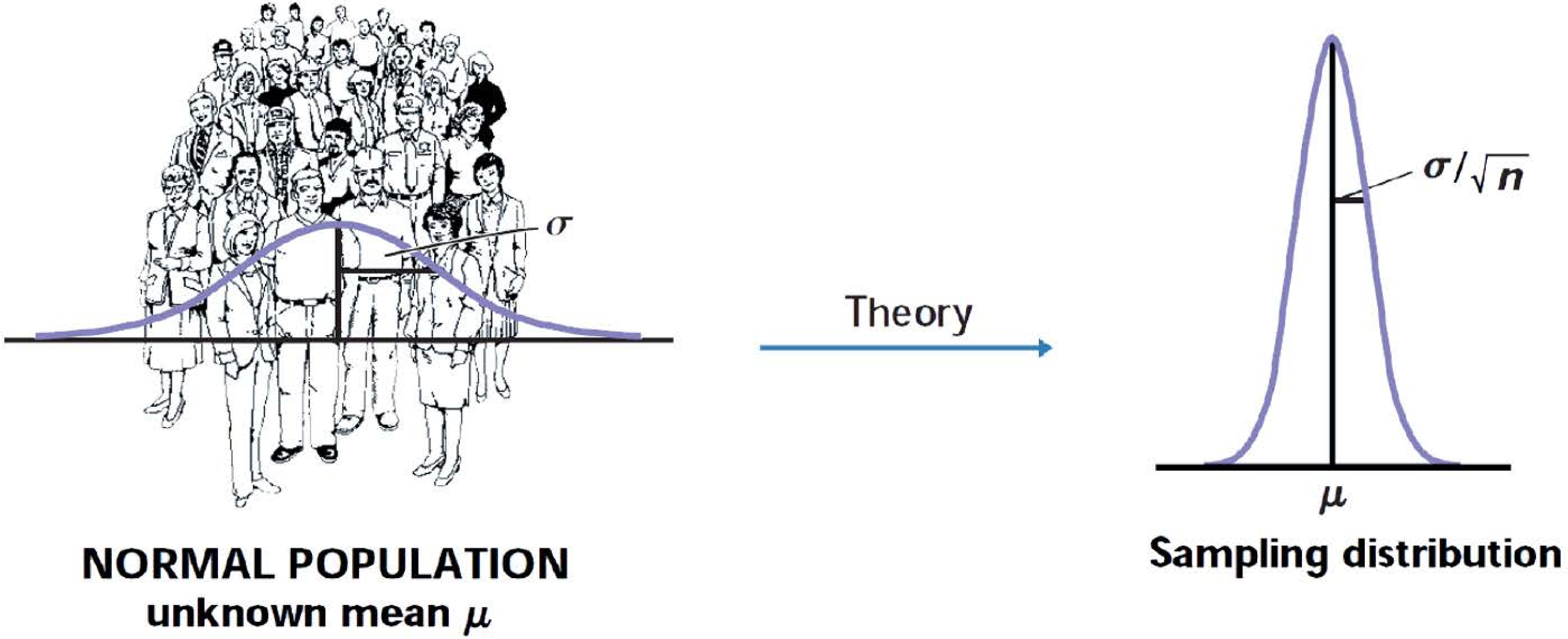
  - Сделать предположение, что X распределена по закону $F_X(x)$, при выполнении которого закон распределения известен



ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

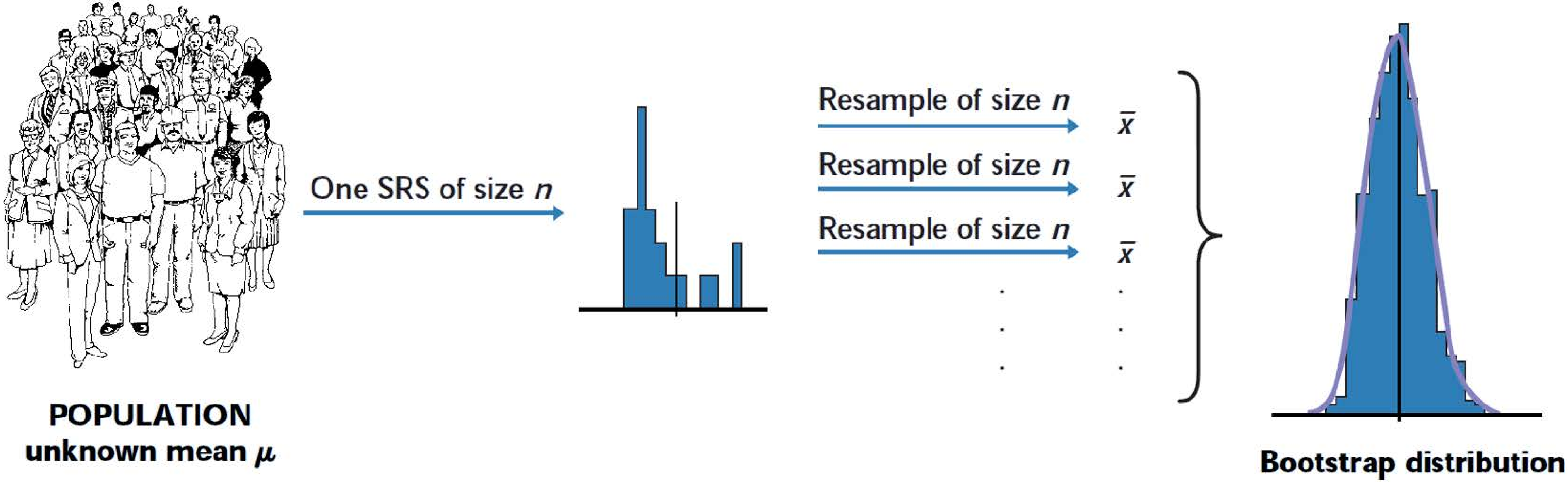
- Параметрический метод:

  - Сделать предположение, что X распределена по закону $F_X(x)$, при выполнении которого закон распределения T_n известен



ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

- Бутстреп:
  - Сгенерировать N “псевдовыборок” объёма n и оценить выборочное распределение T_n “псевдоэмпирическим”



Бутстреп

- Извлечение выборок из генеральной совокупности — сэмплирование из неизвестного распределения $F_X(x)$
- Лучшая оценка $F_X(x)$, которая у нас есть — $F_{X_n}(x)$
- Будем сэмплировать из неё. Это то же самое, что делать из X^n выборки с возвращением объёма n