Основы статистического анализа

Доц. В.В. Викентьев

Описательная статистика

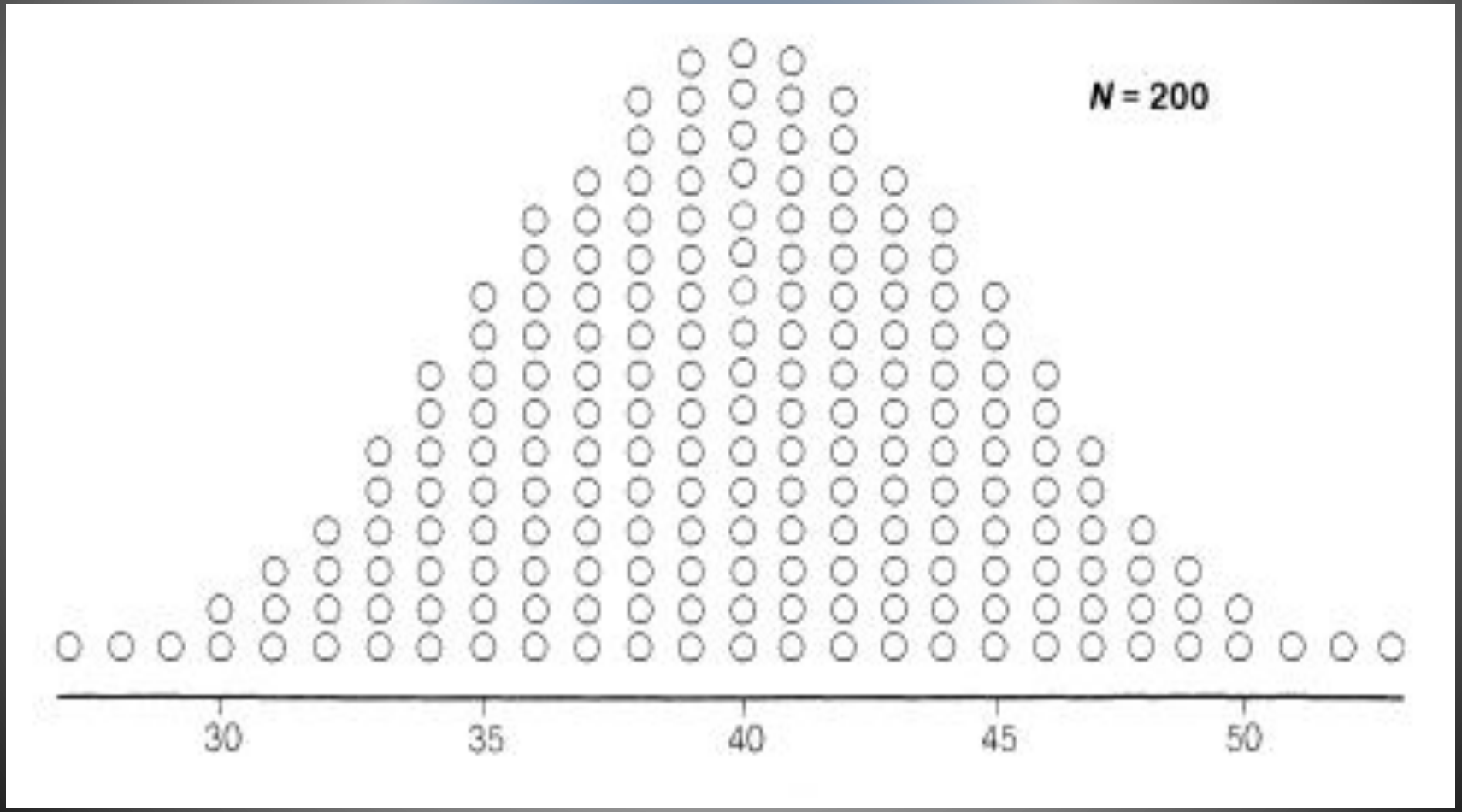
Статистика решает 2 типа задач:

1. Как сжато описать данные?
2. Как оценить значимость различий при сравнении данных?

Описание данных

- Наиболее часто мы анализируем количественные признаки
- Наиболее простое решение проблемы – построить график распределения признака в зависимости от его значения

Пример распределения признака



Описание данных

- Встает проблема как коротко описать данные?
- Среднее по совокупности - **μ**, это среднее арифметическое всех имеющихся значений
- Тем не менее, важно описать каков разброс значений в полученной совокупности относительно среднего
- Эту проблему решает среднее квадратичное (стандартное) отклонение - **σ**

Среднее по совокупности и среднее квадратичное отклонение

$$\mu = \frac{\Sigma X}{N},$$

X – значение признака
N – число членов совокупности

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

Дисперсия

- σ^2 - характеризует разброс значений относительно среднего, как в сторону увеличения значения, так и уменьшения
- Разброс именуют дисперсией
- Пользоваться квадратом числа не удобно
- Корень квадратный из дисперсии называется стандартным отклонением

Стандартное отклонение

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

Нормальное распределение

Нормальное, или Гауссово распределение описывается формулой:

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

Что же такое нормальное распределение?

- При нормальном распределении отклонения значений ряда от среднего подчиняются следующему правилу:

68% значений ряда отличаются от среднего не более чем на одно стандартное отклонение

95% значений ряда отличаются от среднего не более чем на 2 стандартных отклонения

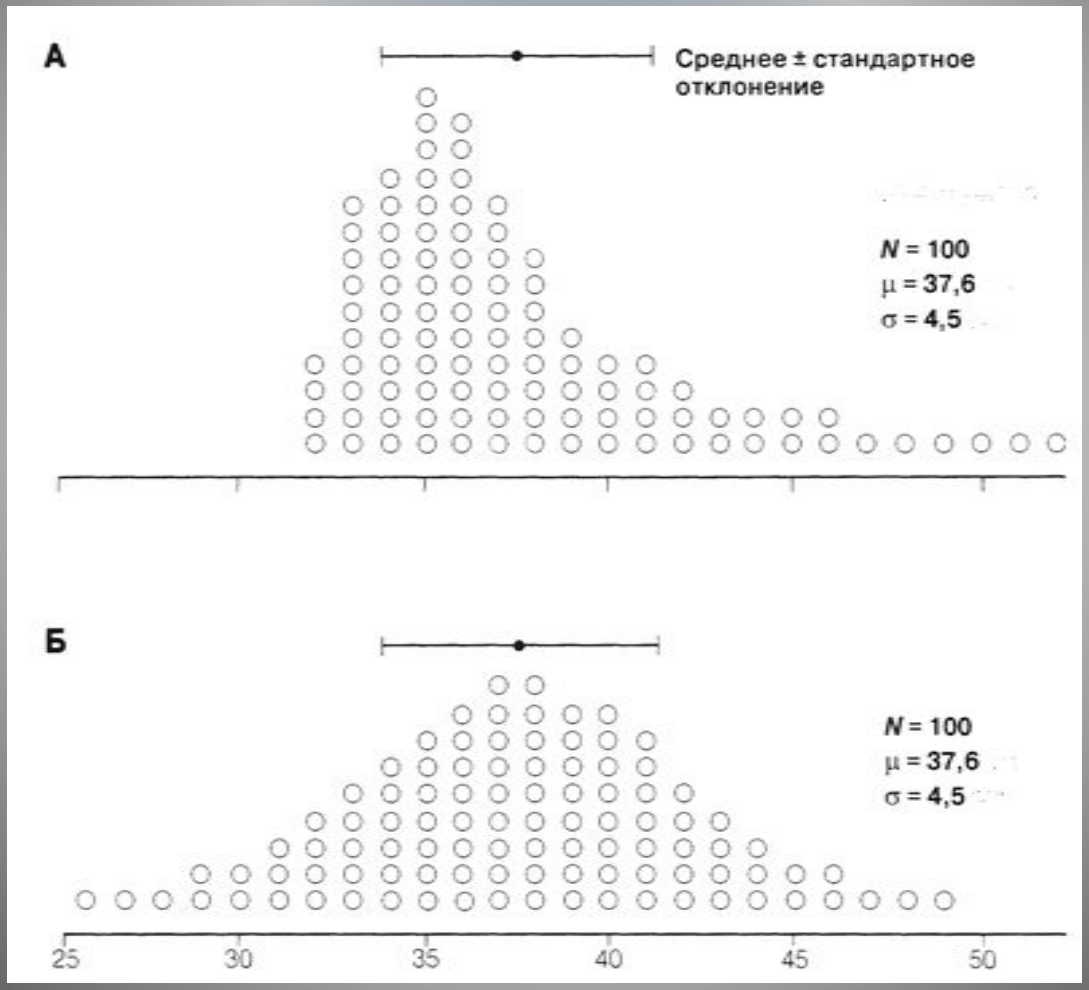
Описание данных

- Таким образом, при наличии нормального распределения данные могут быть описаны при помощи следующих основных параметров:
- 1. среднего значения
- 2. стандартного отклонения
- 3. количества значений ряда

Описание данных

μ	σ	N
15	3	25

Что делать если распределение отличается от нормального?



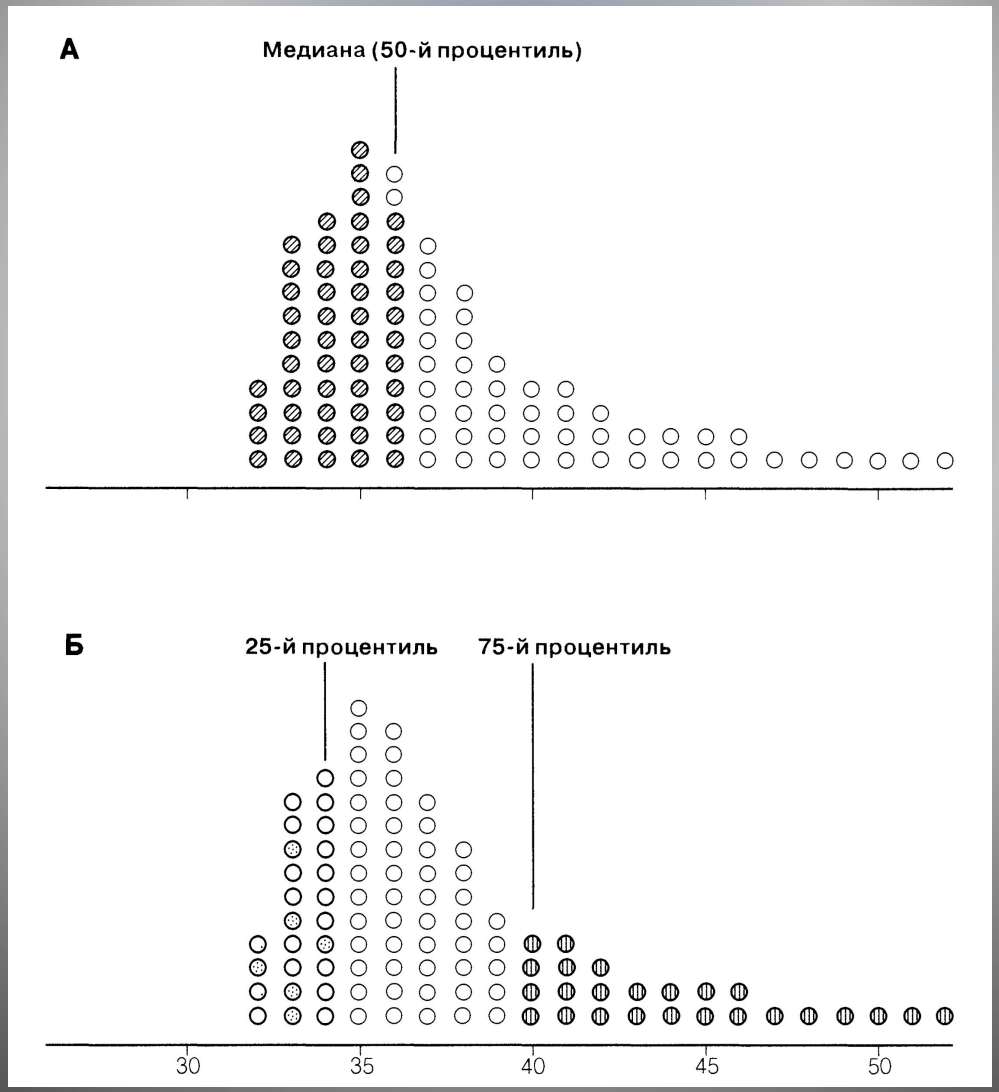
Ассиметричное распределение

- Полагаться на среднее значение и стандартное отклонение нельзя.
- Необходимо использовать другие показатели.

Медиана и процентили

- Медиана это значение признака которое делит ряд значений пополам
- Обычно используют также 25-й и 75-й процентили: это значения признака которое не превышают соответственно 25 и 75 % значений ряда
- Медиана, таким образом, сама является 50-м процентилем

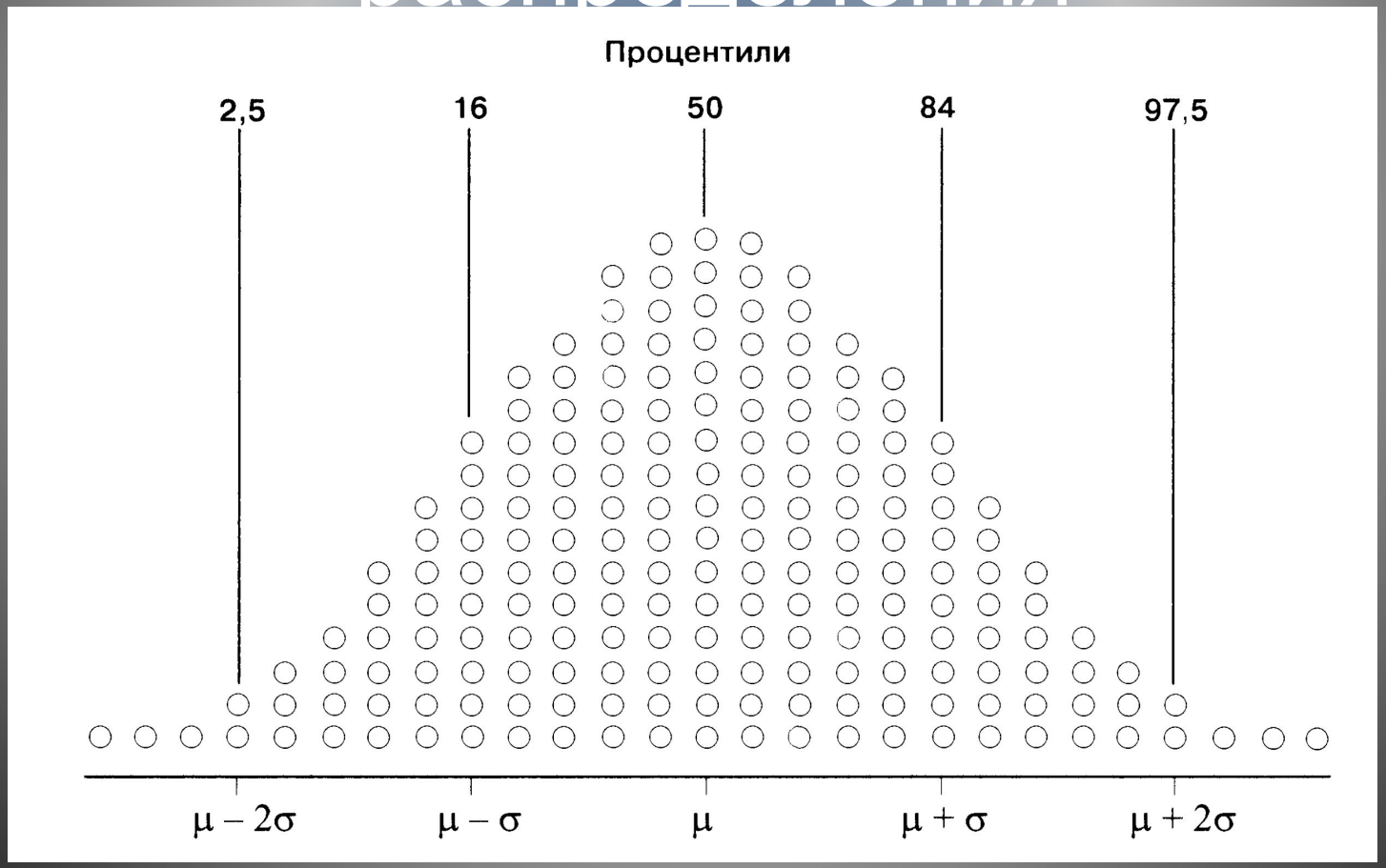
Способы описания ассиметричного распределения



Соответствие между процентилями, медианой и показателями нормального распределения

Процентили	Отклонения от среднего
2,5	$\mu - 2\sigma$
16	$\mu - \sigma$
50	μ
84	$\mu + \sigma$
97,5	$\mu + 2\sigma$

Соответствие между процентилями, медианой и показателями нормального распределения



Выборочные оценки

- Обследовать всю совокупность практически никогда не удается
- Приходится обследовать лишь часть совокупности
- Показатели выборки называют выборочными

Обозначение выборочных оценок

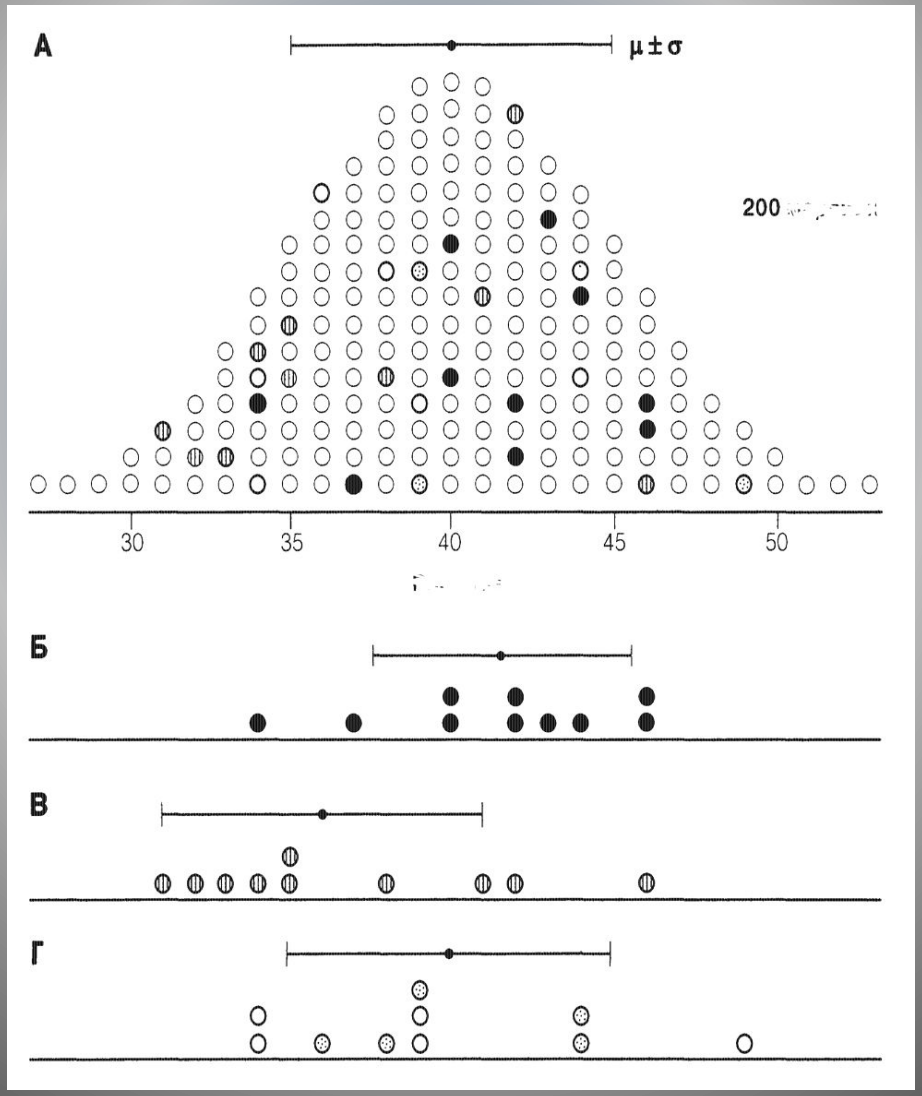
$$\bar{X} = \frac{\Sigma X}{n}$$

Обозначение выборочных оценок

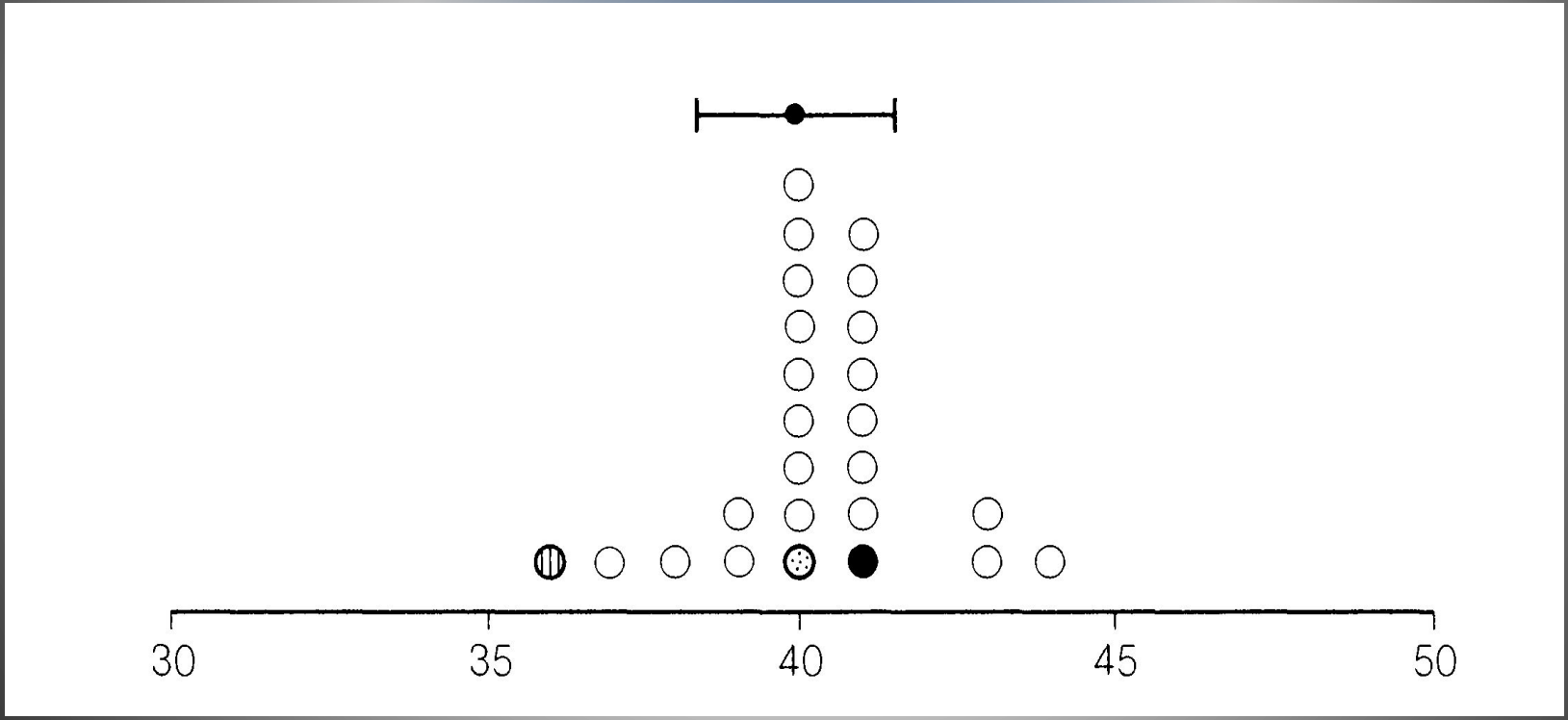
$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}$$

Разброс значений в выборке никогда не бывает столь значительным как во всей
совокупности и поэтому деление на n-1 компенсирует возникающее занижение
оценки стандартного отклонения

Точность выборочных оценок



Точность выборочных оценок



Стандартная ошибка среднего

- Позволяет оценить точность с которой выборочное среднее характеризует значение среднего по совокупности
- Это мера точности с которой выборочное среднее является оценкой среднего по совокупности
- Чем больше выборка, тем ближе выборочная средняя к средней по совокупности, тем меньше ошибка среднего

Стандартная ошибка среднего

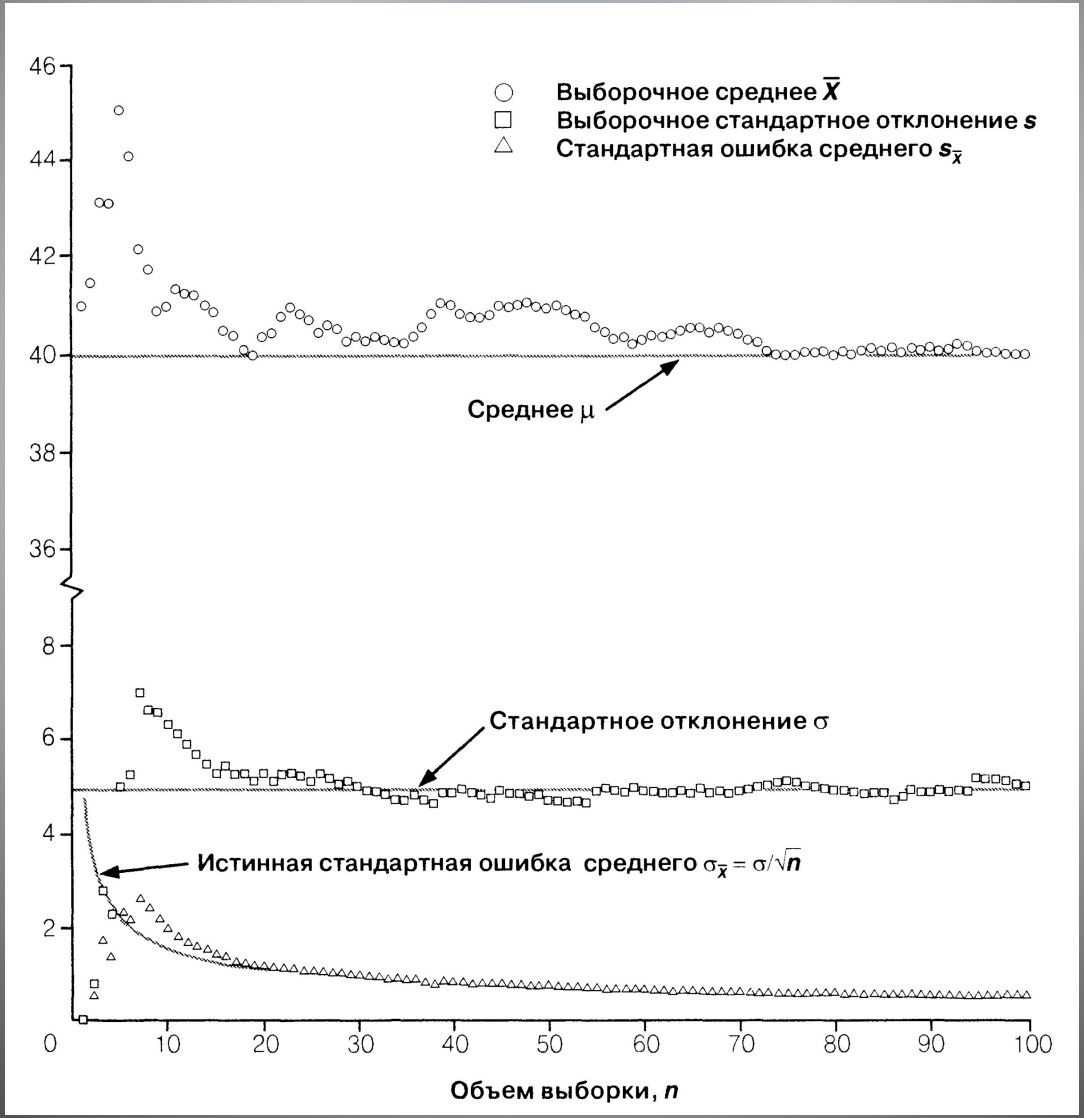
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Центральная предельная теорема

1. Выборочные средние имеют приближенно нормальное распределение независимо от распределения исходной совокупности
2. Среднее значение всех возможных выборочных средних равно среднему исходной совокупности
3. Стандартное отклонение всех возможных средних по выборкам (называемое стандартной ошибкой среднего) зависит от объема выборки и от стандартного отклонения совокупности

Стандартная ошибка среднего



Выводы

1. При нормальном распределении лучше использовать выборочное среднее и стандартное отклонение
2. Если распределение отличается от нормального следует использовать медиану 25 и 75 процентили
3. Стандартная ошибка среднего ничего не говорит о разбросе данных – она лишь характеризует точность выборочной оценки среднего