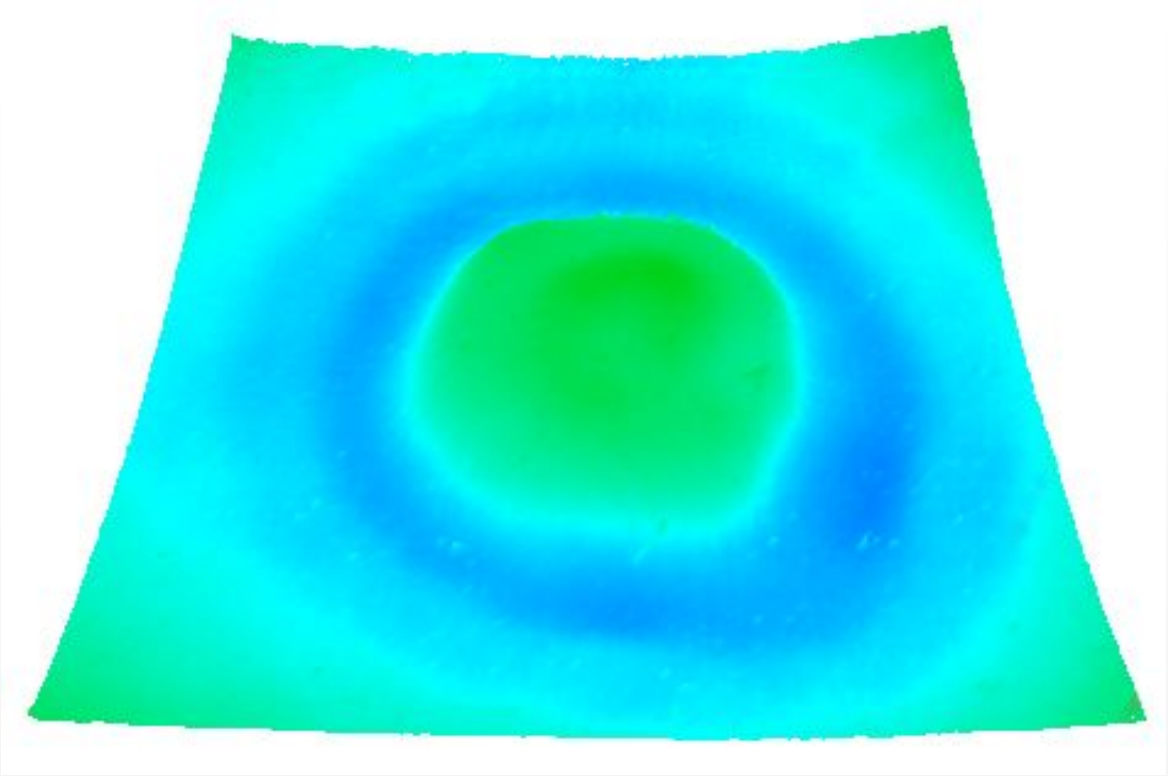
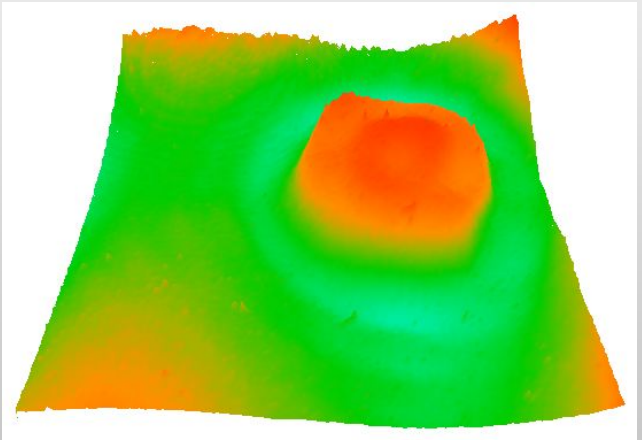


Обработка оптических изображений

Лекция III

Несколько слов о статистике



Несколько слов о статистике

Кратко о том как охарактеризовать и сравнить Ваши данные
Почти инструкция

В последние годы требования к статистике при публикации результатов ужесточились, не все российские ученые адаптировались к этим требованиям
К счастью появилось большое количество программ, в которых все считают за вас, даже указывая на применимость или неприменимость метода.
Для исследователя сейчас важно знать терминологию, чтобы нажать правильную кнопку (границы применимости методов тоже – увы, автоматический режим не всегда работает)

Программы, где можно достаточно просто обработать и представить свои данные
(хотя статистические модули есть во всех уважающих себя программах)

GraphPad Prism 6.07, GraphPad Software – проста, есть необходимый минимум и не только, есть подробный хелп по программе и статистике на сайте

Microcal Origin Pro 2016 – мощная программа для представления и обработки данных, есть подробный и понятный хелп

StatSoft, Inc. STATISTICA 10 – программа для статистических расчетов, неплохая подборка материалов о статистике на сайте

MedCalc Statistical Software version 15.8 – неплохая небольшая программа для статистических расчетов

Статистический анализ данных

Включает несколько этапов. Один из наиболее важных для вас разделов это

Описательная (дескриптивная) статистика

Основная задача данного раздела– предоставление сжатой, концентрированной и наглядной характеристики экспериментальных и контрольных выборок в числовом и графическом виде

Индуктивная статистика

Основная задача данного раздела– проверка статистических гипотез о законе распределения, а основной областью применения – использование в медико-биологических исследованиях для сравнения двух разных выборок на предмет принадлежности к общей генеральной совокупности (достоверны ли отличия между группами).

Исследование зависимостей между переменными *(корреляционный, регрессионный и в какой-то степени факторный анализ)*

Снижение размерности *(задача сократить количество оцениваемых переменных, это делает факторный анализ)*

Классификация и прогноз *(группировка – когортные исследования, дискриминация – дискриминантный анализ, кластеризация – кластерный анализ)*

Анализ выживаемости *(анализ времени до наступления вероятного события)*

Типы данных

количественные — *Имеют некоторое числовое значение*

- **дискретные** — *Принимают строго определенные, как правило, целочисленные значения*
- **непрерывные** — *Данные могут быть представлены любыми численными значениями*

качественные (категориальные) — *применяются для описания состояния объекта путем отнесения его к определенной категории. Объект относится только к одной категории исследования*

- **номинальные** — *Категории не упорядочены, обозначают состояние объекта и не упорядочивают это состояние, например, по полу: 1 – мужской, 2 – женский.*
- **порядковые (ранговые)** — *Категории могут быть упорядочены, обозначают состояние объекта (например самочувствие - 1 – хорошее, 2 – удовлетворительное, 3 – плохое). На практике часто используются для перевода количественных данных в качественные категориальные, например, при расчётах пороговых значений.*

! Для каждого типа данных необходимо выбирать соответствующую процедуру обработки

Описательная (дескриптивная) статистика

Важно учитывать тип данных и параметры распределения, характеризующиеся показателями асимметрии и гистограммой распределения

Распределение данных можно (условно) разделить на:

Нормальное (логнормальное) распределение

Для обработки используются параметрические методы

У качественных переменных есть стандартное отклонение и станд. ошибка среднего тоже есть но считаются по-другому

Все остальные

Для обработки используются непараметрические методы

Поэтому вначале проверяем является ли распределение данных в нашей выборке нормальным!

Существуют специальные тесты для проверки на нормальность

д'Агустино-Пирсона (там целое семейство) – наиболее популярный в настоящее время
Шапиро-Уилка
Комогорова-Смирнова – сейчас не рекомендуется, но иногда используется

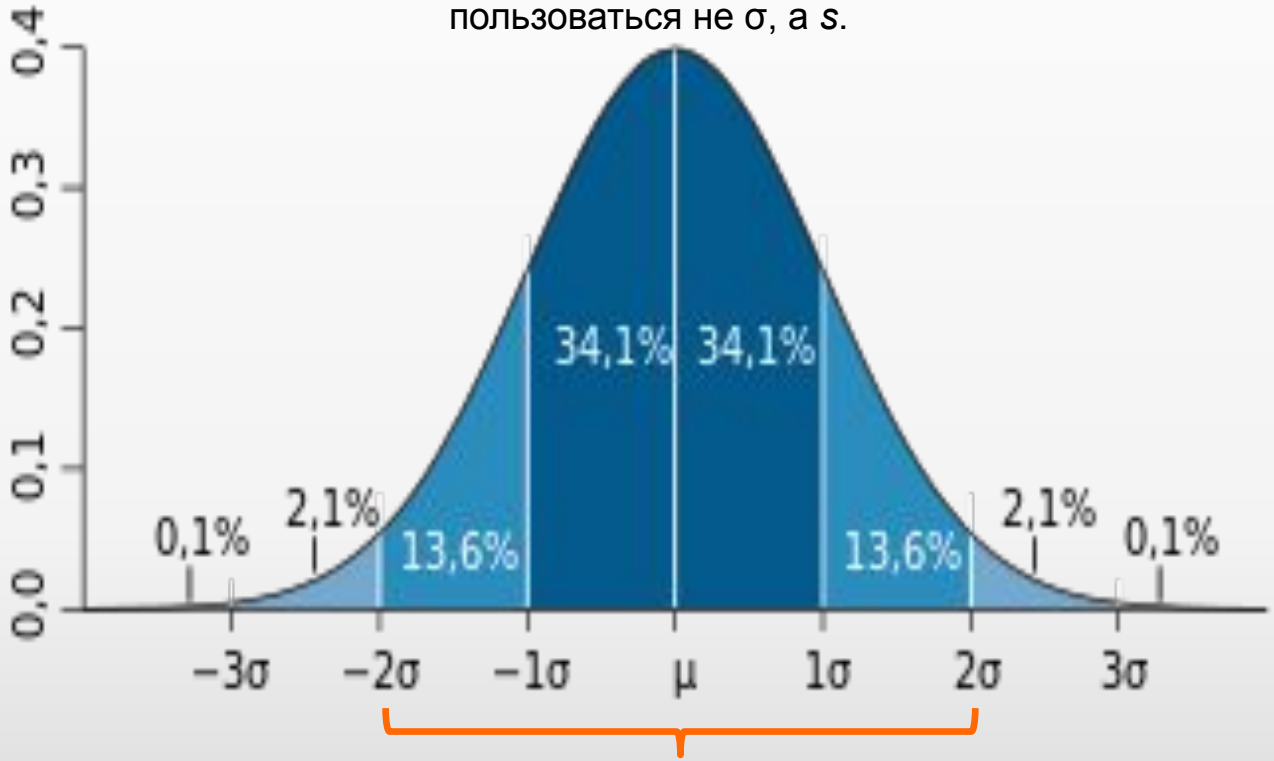
! Программа автоматически проверит все за вас, и с учетом количества выборки тоже, но стоит осторожно относиться к выводу, что распределение нашей выборки не противоречит нормальному распределению, если у Вас менее 12 (а еще лучше 20) объектов

Нормальное распределение

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

Правило трёх сигм (трех *s*)

практически все значения нормально распределённой случайной величины лежат в интервале 3σ (0,9973). Если же истинная величина неизвестна, то следует пользоваться не σ, а *s*.



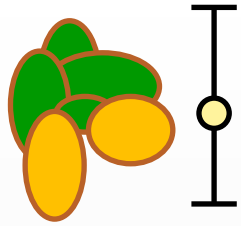
доверительный интервал

Процентили, Медиана, среднее, мода

В идеальном н.р. Медиана=Среднее=Мода

Если нет правильнее использовать медиану. Указание медианы означает сомнение в нормальности распределения для признака

Нормальное распределение

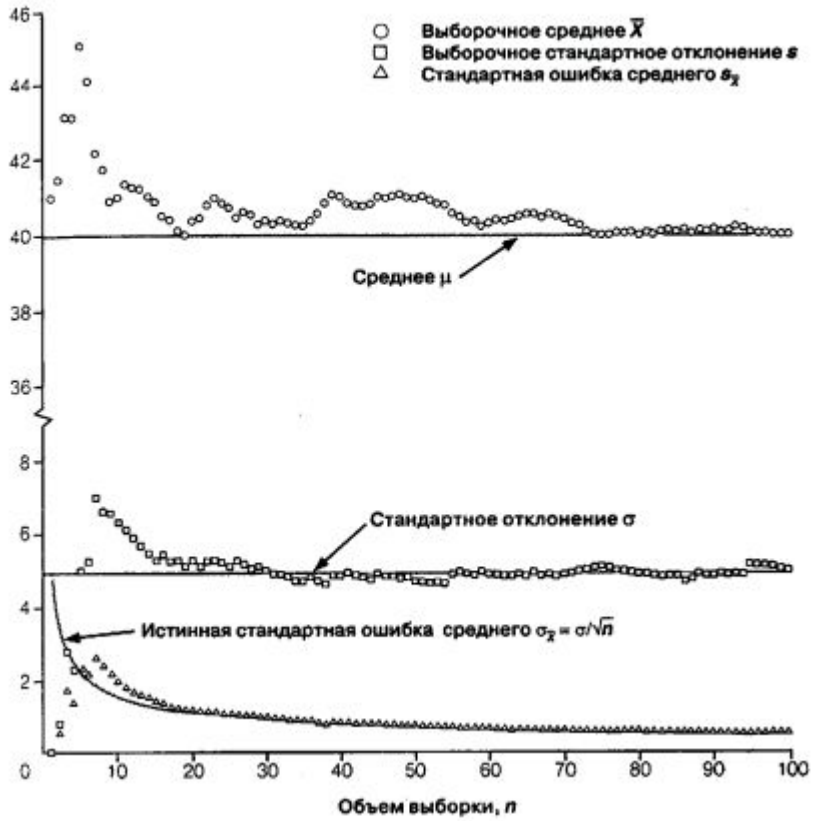


Среднее $\mu = \frac{\sum X_i}{N}$ $\overline{X} = \frac{\sum X_i}{n}$

(выборочное) Стандартное отклонение $\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$ $s = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$

Стандартная ошибка среднего

$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$ $s_{\overline{X}} = \frac{s}{\sqrt{n}}$



Доверительный интервал

диапазон значений, который с определённой исследователем вероятностью включает в себя все значения параметра в популяции

при небольшом объёме выборки предпочтителен. Обычно в настоящее время принимается представление 95% доверительного интервала с указанием нижней (5%) и верхней (95%) границы.

! В случае больших выборок (больше 200) даже очень малые различия будут достоверны

Меньше 17 объектов – не рекомендуют пользоваться нормальным распределением (ваш выбор), меньше 7 – не пользуются

Описательная (дескриптивная) статистика

Параметрические и непараметрические методы

Параметрические методы анализируют *нормально распределенные количественные* признаки

Непараметрические методы используются во всех остальных случаях (для анализа количественных и качественных признаков независимо от вида их распределения)

Непараметрические методы считаются менее мощными по сравнению с параметрическими, т.е. иногда они не позволяют выявить статистические закономерности, которые могут быть выявлены с помощью параметрических методов.

Непараметрические методы более надежны в случаях, когда есть сомнения в том, что анализируемый признак имеет нормальное распределение.

Для нормально распределенных признаков параметрические и непараметрические методы дают близкие результаты

Описательная (дескриптивная) статистика

Показатели описательной статистики

- показатели положения экспериментальных данных на числовой оси

максимальный и минимальный элементы
среднее значение
Медиана
Мода
геометрическое среднее
и др.;

- показатели разброса, описывающие степень разброса данных

выборочная дисперсия
разность между минимальным и максимальным элементами (размах, интервал выборки)
доверительный интервал
интерквартильный размах
и др.

- показатели асимметрии

положение медианы относительно среднего
и др.

- графическое представление результатов

гистограмма
частотная диаграмма
и др.

Описательная (дескриптивная) статистика

Показатели положения экспериментальных данных на числовой оси

Среднее арифметическое — ***показатель центральной тенденции**, полученный делением суммы всех значений данных на число этих данных. Адекватно если у нас **нормальное (!)** распределение*

Медиана — *центральное значение признака в последовательном ряду всех полученных значений (половина объектов больше, а половина меньше). Как вариант: медиана - 50-м перцентиль (0,5-квантиль) или второй квартиль выборки или распределения. Медиана вместе с квартилями используется для представления дискретных или количественных переменных при ненормальном распределении.*

Мода — *наиболее часто встречаемое значение в выборке. В некоторых случаях может быть две или более мод, что может свидетельствовать о наличии двух (нескольких) самостоятельных групп.*

Максимальное и минимальное значение

Среднее геометрическое *(как правило применяется для описания логнормального распределения)* — *Потенцированная величина среднего арифметического рассчитанного из логарифмов значений переменной в выборке*

Описательная (дескриптивная) статистика

показатели разброса, описывающие степень разброса данных

Доверительный интервал

В биологических исследованиях значения параметра достаточно сильно варьирует, поэтому наиболее оптимальным описанием величины является диапазон, в который укладывается большинство значений исследуемого признака, т.е. ширина распределения. 95% доверительный интервал.

Стандартное отклонение

$$s = \sqrt{\frac{\sum\left(X_i - \overline{X}\right)^2}{n-1}}$$

Только для нормального распределения! Оценивает широту распределения, характеризует разброс данных

Стандартная ошибка среднего

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

Только для нормального распределения! Характеризует точность нахождения среднего (если ошибка обусловлена случайными причинами)

Квантили, квартили (интерквартильный размах)

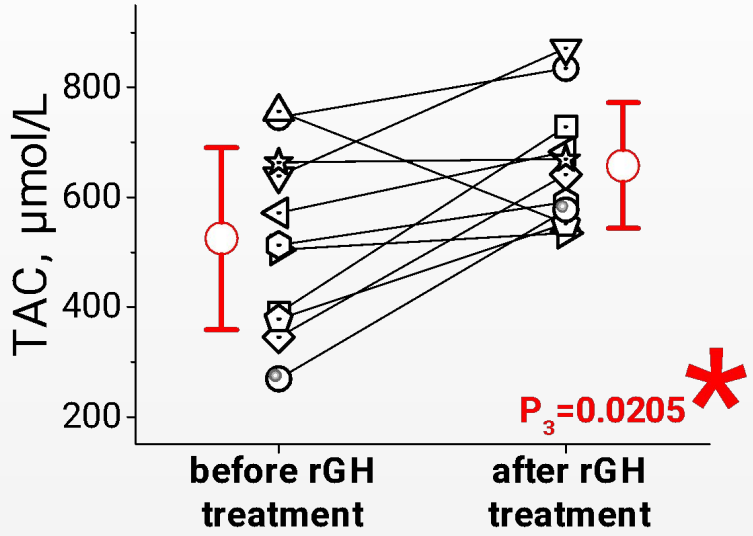
Квантили характеризует собой частоту попадания значений переменной в определённые интервалы. Чаще всего используется разделение на 4 интервала (25%, 50%, 75%).

При разделении на четыре квантиля (именуемых квартилями) для предоставления оценки центральной тенденции, ширины и асимметрии распределения результатов достаточно трёх чисел: нижний квартиль (попало 25% самых маленьких значений), 50% квартиль, который соответствует медиане (попало 50% значений), и верхний квартиль (попало 75% самых маленьких значений). Интерквантильный размах - разность между верхней и нижней квартилью.

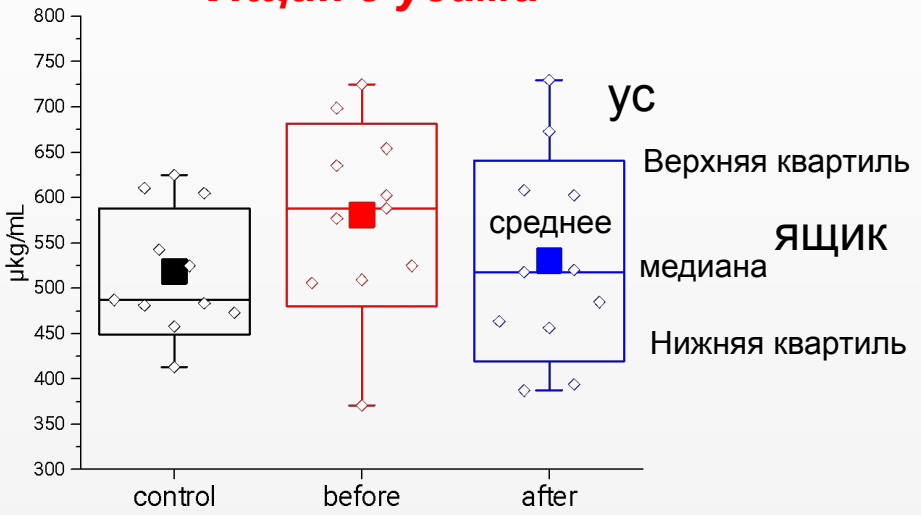
Описательная (дескриптивная) статистика

Графическое представление результатов

Количественные данные



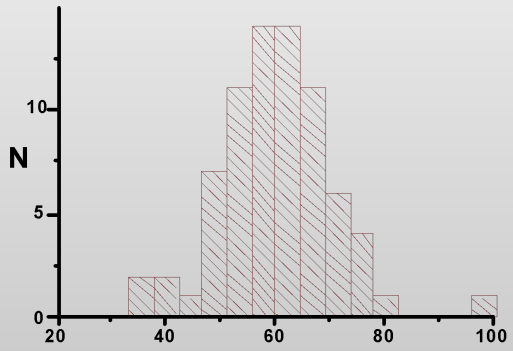
Ящик с усами



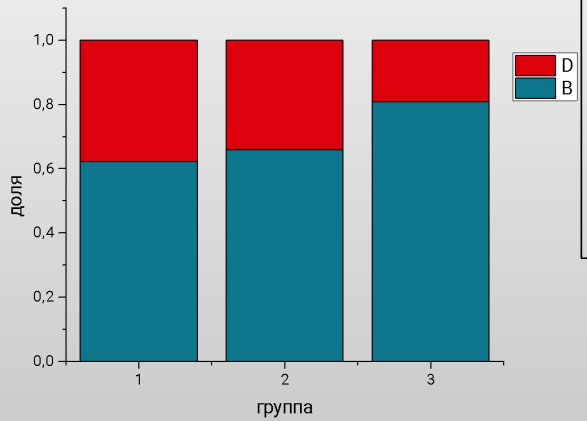
Качественные данные

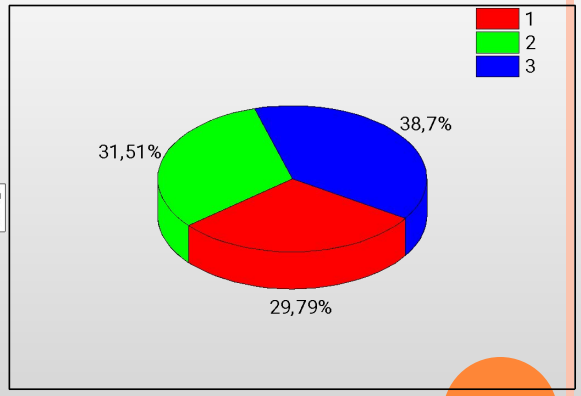
Количественные данные

Гистограмма



Диаграммы





Индуктивная статистика

Основная область применения – использование для сравнения двух (или более) выборок для определения их принадлежности к общей генеральной совокупности

Принадлежность выборок к одной генеральной совокупности свидетельствует об отсутствии различия между ними

Для проверки принадлежности формулируют статистические гипотезы:

гипотеза об отсутствии (случайности) различий между выборками- H_0 (*нулевая гипотеза*)

гипотеза о значимости различий - H_1 (*альтернативная гипотеза*)

Количественную характеристику случайности различий показывает статистическая значимость (***p***). Чем больше p, тем больше вероятность отсутствия различий (истинности нулевой гипотезы), чем меньше ***p***, тем больше вероятность наличия различий (истинности альтернативной гипотезы

Индуктивная статистика

Типы ошибок

Ошибка – обязательный компонент статистического анализа
Допустимый уровень ошибок выбирается исследователем.
Обычно принято использовать два вида ошибок:

ошибка первого рода

которой соответствует понятие
уровня статистической значимости *α*

Вероятность ошибочного признания альтернативной гипотезы (различий нет, но мы думаем что есть)

При p≤ α различия принимаются статистически значимыми

Традиционно в качестве порога (уровня) значимости традиционно выбирается уровень 0,05 (допускает наличие ошибки в 5 случаях из 100)

В предварительных исследованиях допускается уровень значимости *α*=0,1 для выявления намечающихся различий с целью дальнейшего планирования на их основе новых исследований с достаточной значимостью.

ошибка второго рода *β*

которой соответствует понятие
статистической мощности 1-*β*

Вероятность ошибочного признания нулевой гипотезы (различия есть но мы думаем что нет)
обусловлено недостаточным количеством данных

Необходима для определения адекватного объёма выборки. При достаточной статистической мощности отсутствие статистически значимых различий действительно признаётся таковым

Обычно в качестве критического порога принимается значение β
равное 0,1 или 0,2 (допускает наличие ошибки в 10 или 20 случаях из 100, соответственно)

Индуктивная статистика (сравнение групп)

смещение признака

односторонние тесты

Априорно предполагается, что в одной из групп распределение признака смещено в определенную сторону (большую или меньшую) по отношению к другой

двусторонние тесты

Отсутствует априорная информация о смещении групп относительно друг друга

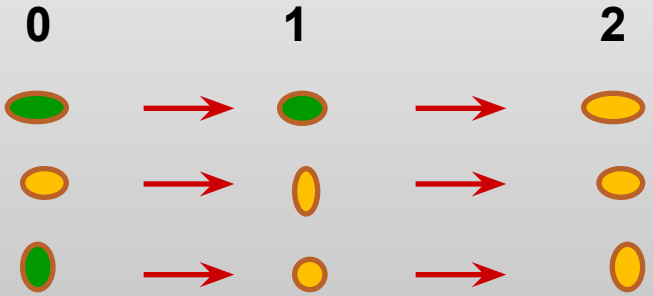
Вычисляемое для односторонних тестов значение статистической значимости примерно в 2 раза меньше, чем для двусторонних тестов, что позволяет при обосновании использования одностороннего теста чаще выявлять достоверные различия. Двусторонние тесты более универсальны. Рекомендуется использовать двусторонние тесты (выбор за вами).

Тип выборки

Выборки могут быть независимыми (несвязанными) или зависимыми (связанными)

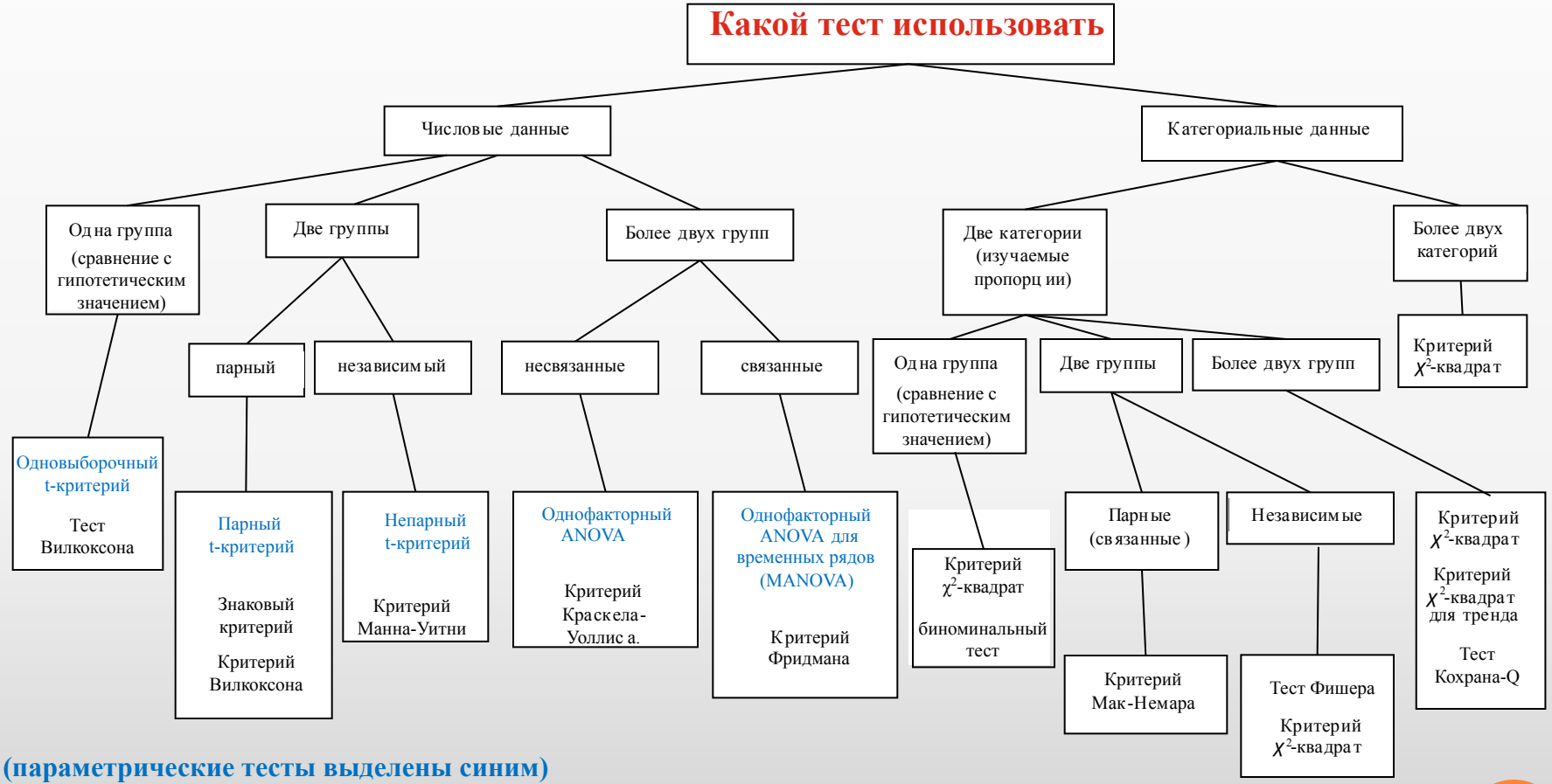
Сравниваем между собой (или с референсными значениями две или несколько проб

Изменение пробы во времени



Индуктивная статистика (сравнение групп)

Одной из главных задач исследователя заключается в формулировке статистических гипотез и выборе правильного статистического критерия для проверки этих гипотез



(параметрические тесты выделены синим)

Статистическая обработка

Как правильно обработать статистические данные?

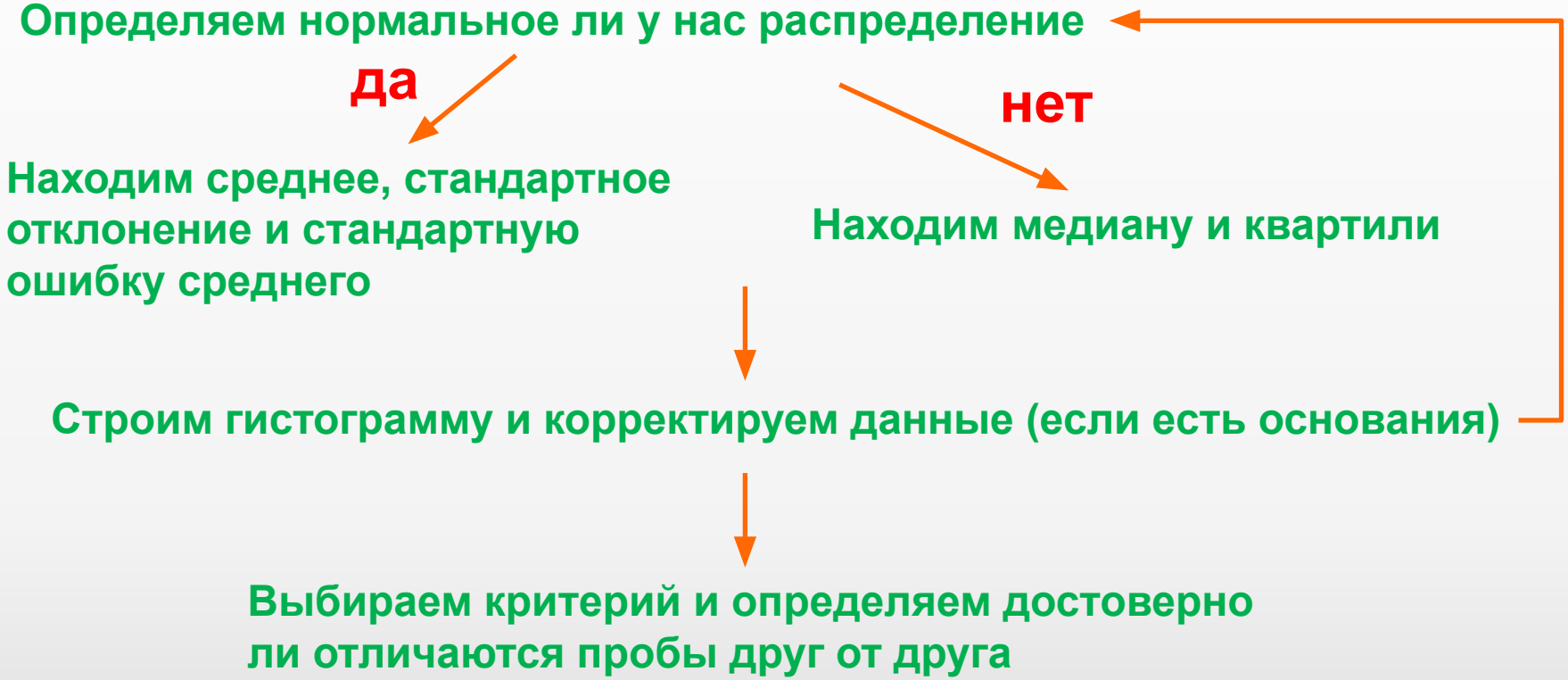
однозначного ответа нет, зависит от формы проведения эксперимента, количества экспериментальных данных, приборной погрешности и т.д.

Ниже приведены некоторые соображения по обработке статистических результатов применительно к конкретной задаче практикума по микроскопии

Тем не менее подобный подход применим к обработке любых микроскопических данных

Статистическая обработка

Статистическая обработка измеренного параметра



Здесь не упоминается метрология и основы обработки сигнала

Где и что про это можно прочитать

Ищем методички для медиков, там мало объясняют, зато пишут что чем обработать и сколько человек должно быть минимум
Если не хватает то, например:

Гланц. Медико-биологическая статистика – есть в интернете бесплатно
Учебник по статистике на www.statsoft.ru – на русском иногда сложноват
Intuitive Biostatistics. Harvey J. Motulsky – надо искать бесплатную версию, на английском, неплоха, http://www.intuitivebiostatistics.com
Русская выжимка из нее: http://pubhealth.spb.ru/SAS/InBio.htm

Origin и мануалы к нему www.originlab.com
Мануалы к Prism http://graphpad.com/data-analysis-resource-center/ - на английском, но написаны довольно понятно, насколько это возможно; много справочной информации по Prism и статистике в целом
Мануалы к Medcalc https://www.medcalc.org/manual/index.php - информации по статистике меньше, но интересующие разделы стоит посмотреть
Последние две программы это члены всяких статобществ, поэтому что там прописано и как считается это некий стандарт

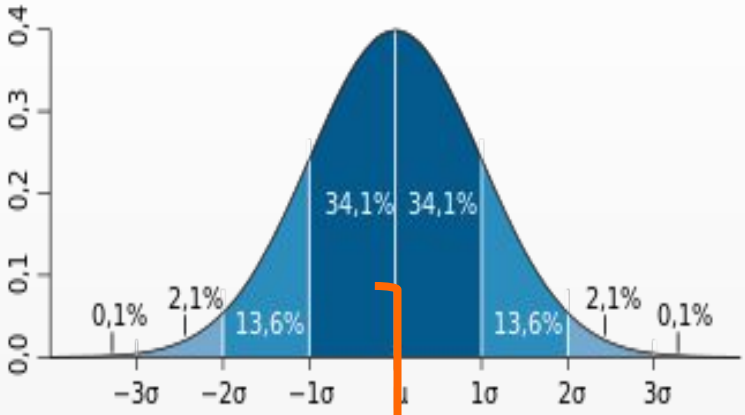
Статистическая обработка

Референсные значения

К этому тесно примыкает понятие

Доверительный интервал

Если распределение данных соответствует нормальному распределению, то это интервал в который укладывается 95% экспериментальных значений (±2s)



Позволяет количественно оценить различия

доверительный интервал

Если величина не входит в референсный (доверительный) интервал, значит различия (с указанной вероятностью) достоверны

Как посчитать доверительный интервал

Критерий Стьюдента

(частный случай дисперсионного анализа)

$$\Delta = \overline{X} \pm t_a \cdot s_{\overline{X}}$$

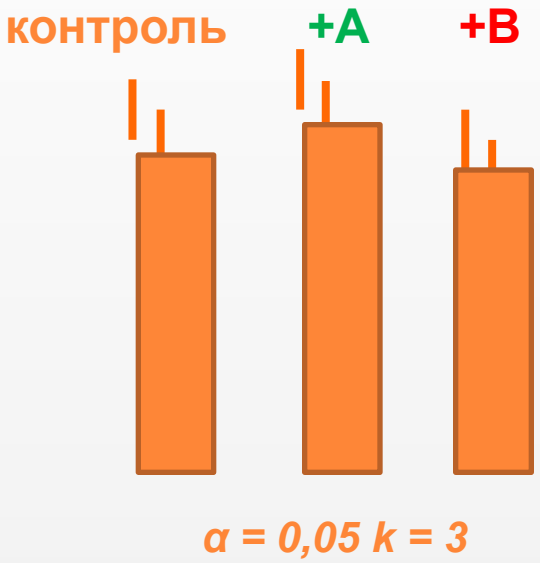
t_a - зависит от количества степеней свободы системы (зависит от количества объектов в пробе и количества экспериментов) – определяется из специальных таблиц, если объектов больше 200 практически не меняется

Говорит в какой интервал входит и с какой вероятностью

Статистическая обработка

Эффект множественных сравнений

Опасность попарного сравнения



α = 0,05 k = 3

Вероятность ошибиться в одном из трех случаев

$$P=1-(1-\alpha)^k \text{ или } P=\alpha k$$

α – вероятность ошибки в одном случае
k – количество сравнений

Неравенство Бонферрони

Вероятность хотя бы один раз ошибочно выявить различия

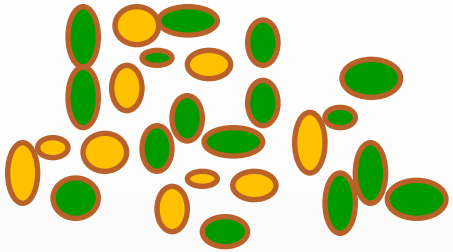
$$\alpha' < \alpha k$$

Что делать?

Пользоваться другими методами (см. дисперсионный анализ)
Ужесточать требования к α

Статистическая обработка

оценка качественных переменных



Оценивают не количество, а доли!

Своя специфика в математике

Стандартное отклонение и станд. Ошибка среднего тоже есть но считается по-другому

Для ситуации есть признак (1) – нет признака (0)

P- доля членов совокупности, обладающее признаком

Стандартное отклонение

$$s = \sqrt{p(1-p)}$$

Стандартная ошибка доли

$$s_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Когда корректно использовать?

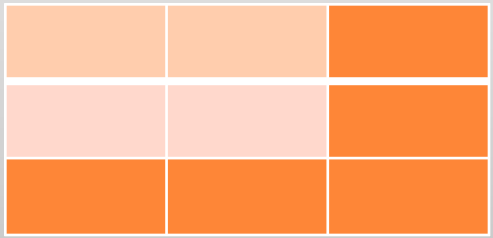
$n\bar{p}>5$ и $n(\bar{p}-1)>5$

Для оценки достоверности тоже есть Стьюдент, z, но свой с поправкой Йейтса

А если больше 2 признаков?

Читаем книжки и изучаем все про χ2

Составляем специальные таблицы



Программы для обработки изображений

FIJI (ImageJ)

Gwiddion

Femtoskan

SPIP

Продукция компании Мекос

Семейство программ Image Pro Plus

Metamorph

И др.

Обработка изображений

вычитание фоновой плоскости

Данная процедура позволяет устранить дефекты, обусловленные следующими причинами

Постоянная составляющая

Обусловлена наличием:
Жидкости ячейки, обладающей конечной толщиной, заполненной жидкостью с определенным показателем преломления

Удаляется из кадра путем вычитания

$$Z'_{ij} = Z_{ij} - \bar{Z} \qquad \bar{Z} = \frac{1}{N^2}\sum_{ij} Z_{ij}$$

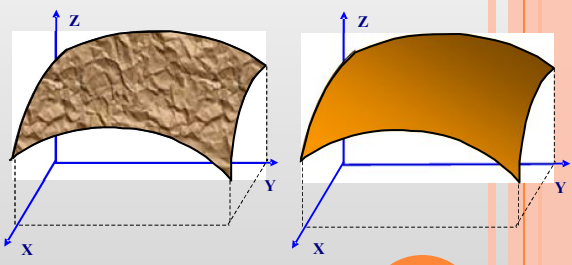
Постоянный наклон

Обусловлен наличием:
- неровностью подложки
- неточной установки образца относительно луча света

Искажения, связанные с неравномерностью освещения

Обусловлен наличием:
- Неравномерности освещения

Удаляется из изображения путем вычитания постоянного наклона.
Для этого находится аппроксимирующая плоскость, которая вычитается из плоскости фазового изображения



Процедура позволяет увеличить точность и улучшить детализацию изображений

Обработка изображений

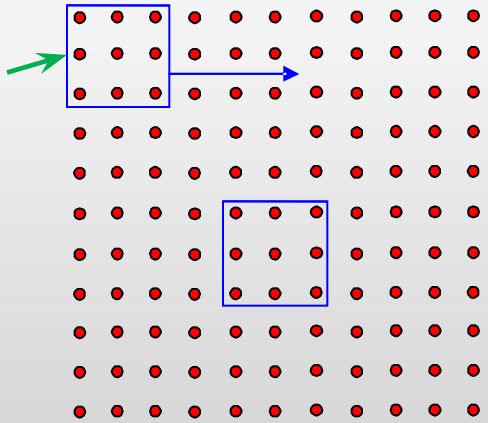
Фильтрация случайных помех при помощи различных фильтров

Случайные помехи обусловлены следующими причинами

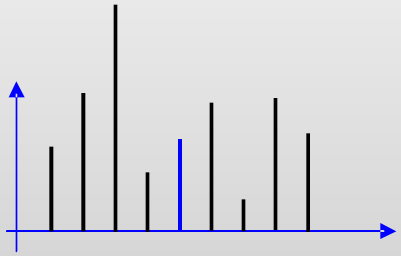
- Шумы аппаратуры
- Дефекты на матрице
- Внешние акустические шумы и вибрации

Устраняется из изображения в результате применения различных фильтров

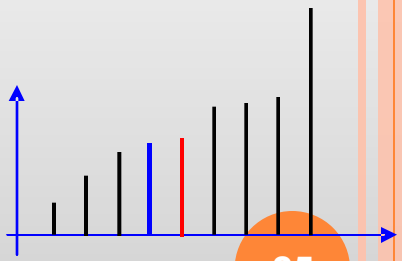
Участок применения фильтра



расположение элементов в неотсортированном массиве (синим цветом помечен центральный элемент)



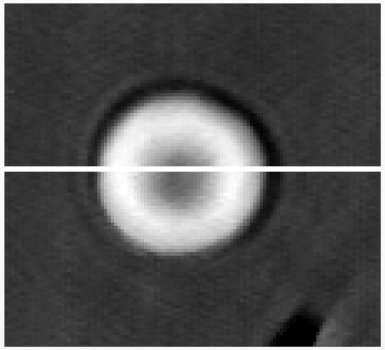
расположение элементов в отсортированном массиве (новый центральный элемент помечен красным цветом)



Процедура позволяет увеличить точность изображений

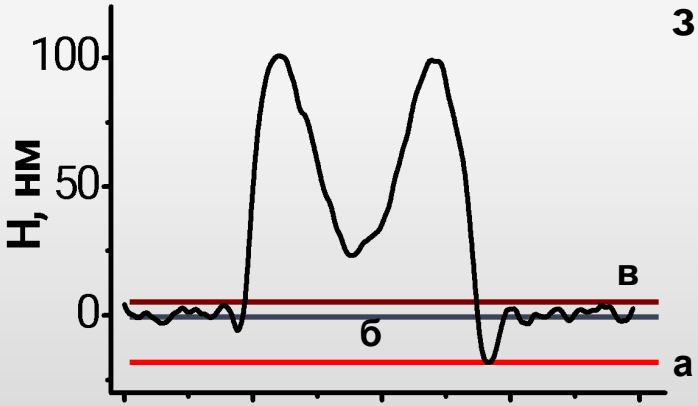
Определение размеров клеток

Определение границ объектов



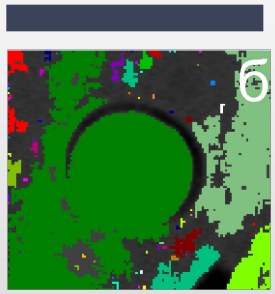
Пороговый алгоритм и его варианты

Самый простой и используемый способ оценить площадь объектов это вычислять их, используя фоновое значение. Каждый объект, представляется как “остров”, окруженный “морем”, т.е. участками изображения, имеющего фоновое значение (или как “озеро”, окруженное “сушей”, в случае пор).

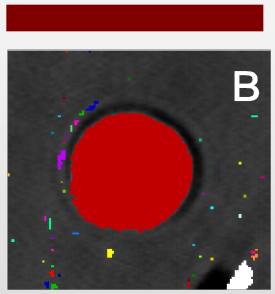


Изменение фонового значения приводит к значительным изменениям площади объектов.

H=-20 нм



H= 0 нм

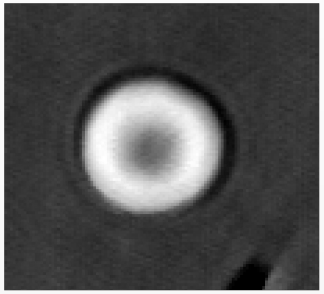


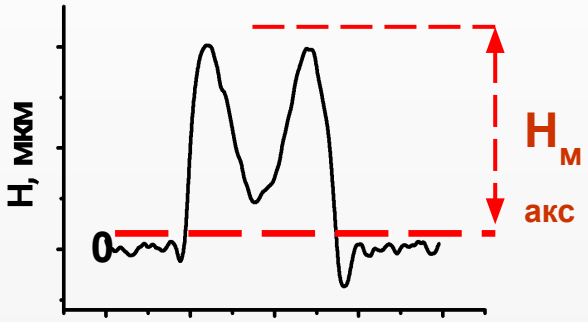
H= 5 нм

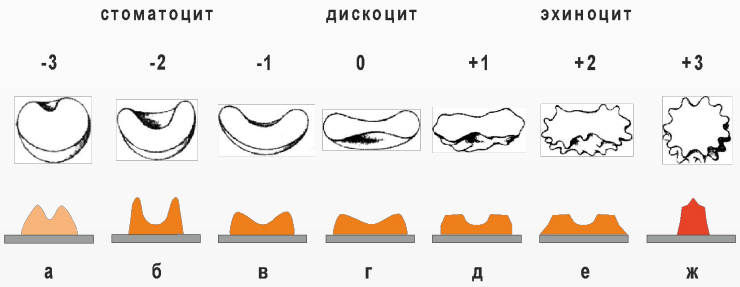
Определение размеров клеток

Что можно посчитать?

Да что угодно







Площадь фазового изображения эритроцитов;
среднее ОРХ клетки;
содержание гемоглобина: $m = \frac{\rho_{Hb}}{(k_s - k_0)} OPD_{mean} S$

количественные

качественные

порядковые

(оценка их доли)

Заключение

Все что Вы услышали не очень подробный обзор как более или менее корректно оценивать полученные микроскопические данные.

Что-то можно применять и не для микроскопа.

Необходимо читать самому!