Статистика

1. Нужно ограничить список тем
2. Попробовать найти источник по теории.

3. Изучить теорию:
    - Зная определения и законы делать задания можно будет сходу, просто подумав. Где в Statgraphics кнопки ты и так уже знаешь.
    - При этом теорию нужно охватить полностью и после нее пройтись по вопросам и практике, поэтому она не должна занять неограниченное количество времени

4. Собрать теор вопросы, на которые нужно знать ответ
5. Пройти лабы, чтобы научиться практической работе
6. Научиться объяснять результаты при помощи теории

ПЕРЕД ЗАНЕСЕНИЕМ В Anki ПРОДУМАТЬ ТЕМЫ И ТЭГИ!

1 часть
Узнать про доверительные интервалы, проверку гипотез.
Разобрать лабы 1-5

2 часть
Дисперсионный анализ (лаба 6), Корреляционный анализ (лаба 7), Регрессионный анализ (лаба 8-9)

3 часть
11 - КЛАСТЕРНЫЙ АНАЛИЗ
12 - МЕТОД ГЛАВНЫХ КОМПОНЕНТ
13 - ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Темы лабораторных работ

1-5 более-менее базовые вещи, про них можно почитать у Лагутина (Часть 3-4). У него же есть Классификация, Корреляция, Регрессия (Часть 5). Ещё есть дополнительный раздел “Некоторые сведения из теории вероятностей”.

У Черновой - Распределения, связ. с нормальными, Проверка гипотез, Критерий согласия, Линейная регрессия

У Горяинова - Доверительные инт. и интервальные оценки, Проверка гипотез (параметрические, непараметрические), Основы корреляционного анализа, Основы регрессионного анализа, Основы дисперсного анализа (однофакторный, двуфакторный)

1. ВВОД, ПЕРВИЧНАЯ ОБРАБОТКА И ГРАФИЧ. ПРЕДСТАВЛЕНИЕ СТАТИСТИЧ. ДАННЫХ
2. СВЯЗЬ СТАТИСТИКИ С ТЕОРИЕЙ ВЕРОЯТНОСТЕЙ
    - Определение достаточного объема выборки
    - *типичные распределения, генерация выборок, гистограмма и полигон*
    - Проверка нормальности через Колмогорова-Смирнова и хи2
3. ДОВЕРИТЕЛЬНЫЕ **ИНТЕРВАЛЫ**

4. ПРОВЕРКА СТАТИСТИЧЕСКИХ **ГИПОТЕЗ**. ОШИБКИ ПЕРВОГО И ВТОРОГО РОДА
    - Проверка параметрических гипотез (основная и альтернативная) о мат.ожидании нормального распределения, ошибка 1-го рода
    - Ошибки 2-го рода, связь вероятностей ошибок 1-го и 2-го рода.
5. ПРОВЕРКА ГИПОТЕЗ О ТИПЕ РАСПРЕДЕЛЕНИЯ (большая работа)
    - Глазомерный метод проверки нормальности
    - Критерии согласия хи2 и Колмогорова–Смирнова
    - Проверка нормальности выборки
    - Проверка однородности двух выборок
    - Сравнение нескольких выборок с помощью критерия хи2

6. ОДНОФАКТОРНЫЙ АНАЛИЗ
    - Однофакторный **дисперсионный** анализ
    - Непараметрические критерии проверки однородности.
        - Критерий Краскела–Уоллиса
        - Критерий Фридмана
7. ВЫЯВЛЕНИЕ СВЯЗИ МЕЖДУ ПРИЗНАКАМИ
    - **Корреляционный** анализ (*корреляционные матрицы, частные коэффициенты корреляции, коэффициенты корреляции Пирсона*)
    - Анализ связи по таблицам сопряженности (*коэффициент сопряженности и коэффициент Крамера*)
    - Коэффициенты ранговой корреляции (*матрица ранговых коэффициентов корреляции Спирмена, ранговый коэффициент Кендалла*)
8. **РЕГРЕССИЯ**
    - Простая регресия
    - Множественная регрессия
9. РЕГРЕССИЯ (ПРОДОЛЖЕНИЕ)
    - Нелинейная регрессия (*тренд временного ряда, метод наименьших квадратов*)
    - Полиномиальная регрессия
    - Пошаговая множественная регрессия
    - Пошаговый отбор переменных

9.1. ЛОГИТ И ПРОБИТ МОДЕЛИ

- Ридж-регрессия, мультиколлинеарность

10. КЛАСТЕРНЫЙ АНАЛИЗ
11. МЕТОД ГЛАВНЫХ КОМПОНЕНТ
    - ФАКТОРНЫЙ АНАЛИЗ
12. ДИСКРИМИНАНТНЫЙ АНАЛИЗ
    - дискриминантные функции
    - линейные дискриминантные функции Фишера (классифицирующие).

13. ВРЕМЕННЫЕ РЯДЫ
    - *тесты на нерегулярность, тренды, автокорреляция, остатки, линейный тренд*
    - *метод перехода от исходного ряда к ряду разностей соседних значений ряда*
    - Выделение сезонной компоненты, *мультипликативная модель*

Рекомендуемая литература

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика. Основы эконометрики. В 2 т. М. “Юнити -Дана”, 2001. http://ecsocman.hse.ru/text/33442857 1 том - **656 с.**
2. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. М.Физматлит, 2006, 816с. (Серия: Современные методы в математике)
3. Горяинов В.Б. и др. Математическая статистика. М.Изд-во МГТУ им.Баумана, 2001, 423с. (Серия:Математика в техническом университете, вып.17)
4. Ивченко Г.И., Медведев Ю. И. Введение в математическую статистику: Учебник. М.: Издательство ЛКИ, 2010. 600 с.
5. Лагутин М.Б. Наглядная математическая статистика М. БИНОМ, Лаборатория знаний, 2007, 472с.

   https://alleng.org/d/math-stud/math-st889.htm (**Глава 11. Доверительные интервалы. Часть III. Проверка гипотез, Глава 12. Критерии согласия, Глава 13. Альтернативы, Часть IV. Однородность выборок, Глава 14. Две независимые выборки, Глава 16. Несколько независимых выборок Часть V. Анализ многомерных данных, Глава 19. Классификация, Глава 20. Корреляция, Глава 21. Регрессия**)
6. Магнус Я.Р., Катышев П.К., Пересецкий А.А.. Эконометрика. Начальный курс. Москва, Изд-во “Дело” 2004, 575с.
7. Катышев П.К., Магнус Я.Р.,Пересецкий А.А.Сборник задач к начальному курсу эконометрики. Москва, Изд-во “Дело” 2003, 207с..
8. Доугерти К. Введение в эконометрику. М.: Инфра-М, Экономический факультет МГУ, 2001,402с.
9. Кремер Н.Ш., Путко Б.А. Эконометрика. М., ЮНИТИ, 2002, 311 с.
10. Тюрин Ю.Н , Макаров А.А.Анализ данных на компьютере. / Изд. 3-е перераб. и дополн. / Под редакцией В.Э.Фигурнова, – М.: «ИНФРА-М», 2003 – 540 с.
11. Бродская Л.И., Бродский Ю.И., Логинов М.И.,Шелементьев Г.С. Анализ данных в пакете Statgraphics . Екатеринбург, УрГУ, 2004, 132с.

Ссылочки

Статистика 7 семест Univer https://drive.google.com/drive/folders/0BxZGCrhoyRfPMDlqdFN4ajBjVmM
Краткие содержания лаботаторных https://www.luminpdf.com/viewer/5e06fcc3a3277d001970182b
https://drive.google.com/open?id=0BxZGCrhoyRfPdktRTjdsSkdMVW8
Есть несколько доков со всякими вопросами и ответами

К консультации

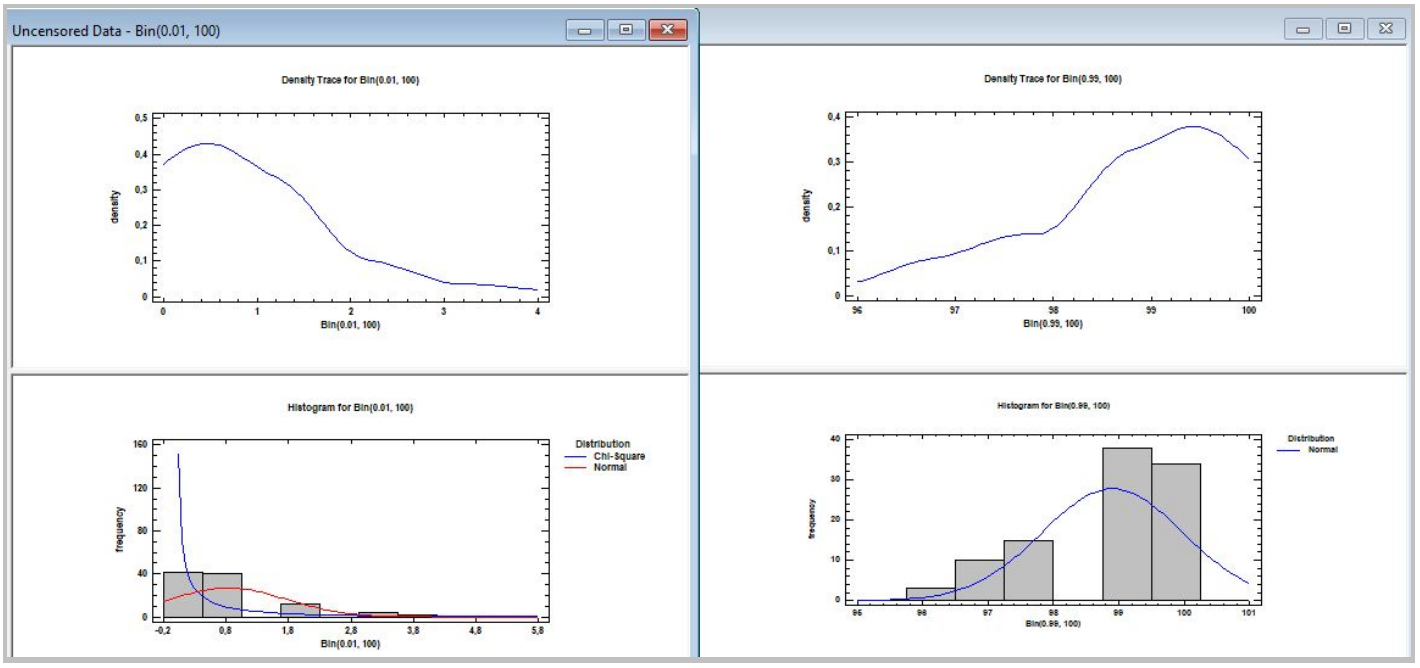
1. Из 2-ой лабы

Bin(0.5, 100), Bin(0.01, 100), Bin(0.99, 100)

А) Для какой из выборок гистограмма «похожа» на нормальную кривую? - для первой Почему это можно было ожидать (вспомните предельные теоремы из теории вероятностей (какую???).

Б) На какое распределение должна быть «похожа» гистограмма для второго распределения? Наложите это распределение на гистограмму.

В) Почему нормальная аппроксимация дает плохой результат для третьей выборки?



2.
3.

Вопросы из лабораторных

Лаб 1

1. Как записывается выборочное среднее для не сгруппированных данных?
2. Как записывается несмещенная выборочная дисперсия для не сгруппированных данных?
3. Что такое (выборочная) мода (можно на примере).
4. Что такое (выборочная) медиана (можно на примере).
5. Что характеризуют асимметрия и эксцесс?
6. Для чего используется коэффициент вариации?
7. Что вы понимаете под репрезентативностью выборки?
8. Что такое гистограмма частостей, статистическим аналогом чего она является?
9. Что такое кумулята частостей, статистическим аналогом чего она является?
10. Как записывается выборочное среднее для сгруппированных данных?

Лаб 2

1. Каков содержательный смысл распределения Бернулли?
2. Почему для распределения Бернулли выборочное среднее совпадает с частостью?
3. Почему в приложениях чаще других встречается нормальное распределение?
4. Что происходит с частостью появления единицы для распределения Бернулли при увеличении объема выборки? В какой теореме из теории вероятностей обосновывается полученное утверждение?
5. Каков содержательный смысл параметров биномиального распределения?
6. Какой смысл имеют параметры a, b в распределении N(a,b)?
7. Что происходит с графиком плотности нормального распределения, если увеличивать математическое ожидание? Дисперсию?
8. Что происходит с графиком плотности распределения Стьюдента при увеличении числа степеней свободы?
9. Что происходит с графиком плотности распределения χ^2 при увеличении числа степеней свободы?

Лаб 3

1. Для чего нужно вычислять доверительный интервал оценки?
2. Что такое доверительная вероятность?
3. Как записывается доверительный интервал для математического ожидания?
4. Какое распределение используется для построения доверительного интервала для математического ожидания?
5. В каком случае требование нормальности распределения изучаемой случайной величины существенно?
6. Какое распределение используется при построении доверительного интервала для дисперсии?
7. Во сколько раз следует увеличить объем выборки, чтобы на порядок уменьшить длину доверительного интервала? Что происходит с длиной доверительного интервала при увеличении доверительной вероятности?

Лаб 4

(задание 1)Обратите внимание на величину разности между X_(выборочным средним) и m0 и на то, как это повлияло на результаты проверки гипотез, а также на то, что в некоторых случаях результат проверки основной гипотезы зависит от вида альтернативной.

(задание 2) Как зависит частость ошибки второго рода от вероятности ошибки первого рода (α):
При увеличении вероятности ошибки 1-го рода вероятность ошибки 2-го рода уменьшается (если увеличивается вся крит. область) либо остается прежней (если только та часть, что относится к P(S|H).
При уменьшении увеличивается либо остается прежней.

(задание 3) В каком случае применение критерия Стьюдента строго обосновано? При каком условии его можно использовать приближенно?

1. Что такое статистическая гипотеза?
2. Что такое простая гипотеза?
3. Что такое параметрическая гипотеза? Приведите пример.
4. Что такое ошибка первого рода? второго рода при проверке статистических гипотез?
5. Что такое критическая область?
6. Что такое наилучшая критическая область (область принятия решения)?
7. Что происходит с вероятностью ошибки второго рода при уменьшении вероятности ошибки первого рода?
8. Что такое непараметрическая гипотеза? Приведите пример.
9. Можно ли по выборочному среднему спрогнозировать, в каких случаях гипотеза H0 будет отвергаться?

Лаб 5

1. Что такое критерии согласия?
2. Какая гипотеза проверяется с помощью критерия согласия 2? Как следует группировать данные для применения этого критерия?
3. Какие критерии проверки однородности вы знаете для независимых (непарных) наблюдений?
4. Какие критерии проверки однородности вы знаете для парных наблюдений?
5. В чем «идея» критерия знаков?
6. В чем «идея» критерия знаковых ранговых сумм?
7. В чем разница между парными и независимыми наблюдениями?
8. Параметрические или непараметрические гипотезы проверяются с помощью критерия Пирсона? Обоснуйте ответ.
9. Что вы будете делать, если при проверке гипотез о математическом ожидании у вас нет нормальности распределений?

Инструменты Statgraphics

0.
Edit > Preferences > отключить Use Six Sigma Menu.

1.

Описательный стат.анализ введенных данных, числовые характеристики выборки
Describe > Numeric Data > One-Variable Analysis
Tables and graphs > Frequency Tabulation, Frequency Histogram
  Pane Options > Polygon

Наглядное представление о распределении (диаграмма)
Describe > Categorical Data > Tabulation

Характеристики нескольких величин одновременно
Describe > Numeric Data > Multiple Variable Analyses > выбрать в левой части окна оба столбца данных
В открывшемся диалоговом окне выберите Summary Statistics.
С помощью Pane Options выведите на экран все нужные выборочные характеристики

2.

Определение объема выборки
Tools > Sample Size Determination > One Sample

Генератор случайных чисел
Plot > Probability Distributions
  Save results > включить Random Numbers for Dist.1.

Выборка другого объема
Table and graphs > Random Numbers; в раскрывшемся окне в контекстном меню Pane Options

Вычислить формулу в таблице
Правой кнопкой по названию свободного столбца, в контекстном меню Generate Data. Окно Generate Data, в правой части в поле Operators выберите AVG(?) (среднее), щелкните дважды, название перейдет в поле Expression. В этом поле сотрите в скобках вопросительный знак, щелкните дважды в поле Variables по названию нужного вам столбца.

Полигоны для всех выборок на одном экране
правой кнопкой по графику, в контекстном меню Copy Pane to Gallery, раскроется окно, щелкните в левом верхнем углу правой кнопкой, выберите Paste (+ для второго и последующих выберите Overlay).

Аппроксимирующая кривая
Describe > Distribution Fitting > Fitting Uncensored Data

Графики плотностей и функций распредления
Plot > Probability Distributions > Normal
Щелкните правой кнопкой, выберите Analysis Options.

Построение доверительных интервалов

Describe > Numeric Data > One-Variable Analysis
Tables and graphs > Confidence Interval
Для выборок равного объема можно получить все доверительные интервалы одновременно
Describe > Numeric Data > Multivariable Analysis

4.

Проверка параметрических гипотез
Describe > Numeric Data > One-Variable Analysys
Tables and graphs > Hypothesis Tests
Pane Options > Hypothesis Tests Options
можно выбирать нужные значения среднего (mean), уровня значимости (alpha), а также варианты альтернативных гипотез (Not Equal, Less Than, Greater Then)

5.

Глазомерный метод проверки нормальности
Describe > Numeric Data > One-Variable Analysys
Tables and graphs > Normal Probability Plot

Критерий Колмогорова-Смирнова проверки нормальности (проверки типа распределения)
Describe > Distribution Fitting > Fitting Uncensored Data > Normal
Tables and graphs > Goodness-of-Fit Tests > Pane Options > Kolmogorov-Smirnov
p-value < 0,05 => отвергаем гипотезу о нормальности распределения.
Критерий χ^2
в окне Goodness-of-Fit Tests > Pane Options > Chi-squared
p-value < 0,05 => критерий χ^2 также отвергает нормальность распределения

Подобрать распределение генеральной совокупности, из которой взята выборка
Tables and graphs > Comparison of Alternative Distributions
Здесь результаты метода макс. правд. (Log Likelihood) и макс. отклонение из метода Колмогорова-Смирнова (KS D).

Вывести столбец со значениями P-Value для критерия χ^2 **и упорядочить по этим результатам**
Pane Options > Tests > Include Chi-squared, Sort By Chi-squared

Проверка нормальности выборки (тест Шапиро-Уилка и χ^2 **)**
Describe > Distribution Fitting > Fitting Uncensored Data
Tables and graphs > Tests for Normality
p-value > 0,05 => нельзя отвергать гипотезу о нормальности распределения

Pane Options > Chi-squared
аналогично, если p-value > 0,05 => нельзя отвергать гипотезу о нормальности распределения.

ПРОВЕРКА ОДНОРОДНОСТИ ДВУХ ВЫБОРОК

Непарные наблюдения. Критерий Стьюдента для проверки равенства средних.

Compare > Two Samples > Independent Samples

1) **Проверка нормальности распределений**
2) **Проверка гипотезы о равенстве дисперсий (эту проверку можно опустить, если объемы выборок достаточно велики и не сильно отличаются друг от друга)**

Tables and Graphs > Comparison of Standard Deviations **(F-test)**

p-value > 0,05 => нет оснований отвергнуть гипотезу о том, что дисперсии выборок равны. Теперь можно воспользоваться критерием Стьюдента (t–тестом) для проверки равенства средних:

Tables and Graphs > Comparison of Means

В качестве альтернативной гипотезы разумно рассмотреть mean1> mean2 (это следует из того, что выборочная средняя 1 больше, чем выборочная средняя 2, вопрос в том, насколько эта разница статистически значима).

Comparison of Means > Pane Option > Alt. Hypothesis

Непарные наблюдения. Критерий Уилкоксона (Манна-Уитни) для проверки равенства медиан.

Compare > Two Samples > Independent Samples

Tables > Comparison of Medians

P-value > 0,05 => нет оснований отвергнуть нулевую гипотезу, следовательно, по данным нельзя сказать, что выборки отличаются.

Однородность с помощью критерия Колмогорова-Смирнова

Tables and graphs > Kolmogorov-Smirnov Test

P-value > 0,05 => нет оснований отвергать гипотезу о равенстве распределений, следовательно, выборки однородные.

Парные наблюдения: проверка изменений

Compare > Two Samples > Paired Samples

Tables and graphs > Hypothesis Tests

Если есть нормальность (и в т.ч. достаточно наблюдений): Критерий Стьюдента (t–тест)

Выборочная средняя разности выборок > нуля, поэтому в качестве альтернативной гипотезы естественно выбрать Greater Than. Значение p-value < 0,05, говорит о статистически значимой разнице средних выборок.

Sign test — критерий знаков

Signed rank test — критерий знаковых ранговых сумм

Наблюдения (непарные) из одной выборки по условию (столбец X1 - например, пол)

Compare > Two Samples > Independent Samples.

В раскрывшемся окне выберите Y в поле Data, в нижней части окна выберите Data and Code Columns, X1 в поле Sample Code.

Нет нормальности - медианный критерий

Tables > Comparison of Medians

Таблицы сопряженности (хи2 для нескольких выборок)

Describe > Categorical Data > Contingency Table

Tests of Independence

p-value в окне Tests of Independence больше или равно 0,1 => мы не можем отклонить гипотезу о независимости данных столбцов от использованной методики => методика не влияет.

6.

Однофакторный дисперсионный анализ. Проверьте законность применения однофакторного дисперсионного анализа с помощью критерия Кочрена.

ANOVA (Дисперсионный анализ)
Compare, в раскрывшемся меню выберите Multiple Samples, Multiple-Sample Comparison,

необходимо проверить равенство дисперсий
Tables > Variance Check.
Pane Options В этом окне можно видеть четыре варианта тестирования. Поскольку у нас объемы выборок равны, воспользуемся критерием Кочрена. Если гипотезу о равенстве дисперсий следует отклонить, применение однофакторного дисперсионного анализа нельзя считать законным.
В нашем случае P-value = 0,395821 говорит о том, что нет основания отвергнуть нулевую гипотезу о равенстве дисперсий (альтернативная – двусторонняя). Следовательно, мы можем воспользоваться критерием ANOVA.

Tables > ANOVA Table
В нашем случае как раз p-value = 0,0000<0,05, отсюда следует, что гипотезу о равенстве математических ожиданий следует отвергнуть. Следовательно, день недели влияет на выручку.

Compare > Analyses of Variance > One-Way ANOVA
Dependent Variable (зависимая переменная), затем Factor.
Tables > Variance Check > ANOVA Table

НЕПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ ПРОВЕРКИ ОДНОРОДНОСТИ.
Критерий Краскела–Уоллиса
Compare > MultipleSample, раскроется диалоговое окно, в котором в поле Sample
Tables > Kruskal-Wallis and Friedman Test
Так как p-value=0,0007363 меньше чем 0,05, то существует статистически значимое различие между медианами с 95 % доверительным интервалом. Следовательно, можно уверенно отвергнуть гипотезу об однородности.

Критерий Фридмана
то же самое, но потом Pane Options > Friedman Test
т.к. P-Value = 0,246597 < 0,05 нет оснований отвергнуть нулевую гипотезу (которая состояла в том, что медианы равны). Следовательно, по данным этой выборки нельзя сказать, что какие-то предметы студенты прогуливают чаще.