**1 слайд**

Приветствие.

**2 слайд**

В данной работе рассматривается проблема статистического анализа малых выборок. Актуальность данной проблемы обоснована необходимостью уметь работать с данными и правильно их интерпретировать даже при ограниченных объемах информации.

Работать с малыми выборками приходится по разным причинам.

Во-первых, сбор достаточно большого количества данных может быть достаточно долгим и нецелесообразным.

Во-вторых, зачастую, получение больших выборок может быть требовать больших финансовых расходов, например проведение множества дорогостоящих экспериментов.

В-третьих, нередко ставится задача вроде «здесь и сейчас», в которой необходимо получить быстрый результат с удовлетворяющей погрешностью в точности.

В данной работе малыми выборками называем выборки, объем которых не более 200 значений, очень малыми – не более 30.

**3 слайд**

Целью работы является изучение способов обработки информации ограниченного объема, а также комбинация и использование различных статистических методов для решения различных практических задач.

В работе рассматривается три обширных и общих задачи, а именно точечное и интервальное оценивание, а также проверка гипотез и комплексные методы, позволяющие получить более сложную и практически ценную информацию.

**4 слайд**

Объектом исследования являются статистические методы анализа данных, а предметом – методы, которые можно использовать для работы с малыми выборками. Часть из приведенных в работе методов может быть использована самостоятельно, а часть, для лучшей точности, следует комбинировать.

**5 слайд**

Для точечного оценивания рассматривается два метода оценок: ММП, ММ и три типа распределения: нормальное (в большей части методов) как самое распространенное и полезное практически, экспоненциальное и равномерное. Полученные по каждому методу зависимости для оценок приведены на экране.

Работая с малыми выборками более предпочтительными являются интервальные оценки, которые обладают меньшей чувствительностью к размеру данных. Для интервального оценивания в работе используются два критерия: т критерий для построения ДИ для МО, и хи квадрат для дисперсии, распределения и квантили стьюдента и хи квадрат соответственно. Зависимости для границ на экране. В этот блок также вошел метод построения толерантных интервалов.

**6 слайд**

Наибольшее внимание в работе уделялось методам проверки статистических гипотез по малым выборкам. Это группа универсальных методов, которая позволяет регулировать точность полученных результатов через варьирование уровня значимости. С помощью него можно получить численное выражение вероятности верности полученных результатов.

Для проверки гипотез о параметрах, в частности о МО по одной и двум малым выборкам, использовались z и t статистики соответственно.

Более объемной получился подраздел проверки гипотез о распределении. Здесь приведены различные методы, а именно:  
проверка гипотезы об однородности распределения двух выборок – критерий Уилкоксона

**7 слайд**

проверка гипотезы о типе распределения – приведены два метода: основывающийся на критерии хи квадрат и критерий Крамера-Мизеса-Смирнова, эти методы подходят для работы с малыми выборками установленного в работе объема. Также в работе приведены модификации данных методов, которые обеспечивают более точные результаты при работе с очень малыми выборками.

Метод проверки малой выборки на нормальность – критерий Шапиро-Уилка

Для обобщения и дальнейшего использования информации, полученной по вышеозвученным методам, в работе используются дисперсионный анализ и линейная регрессиия.

Комбинация и правильное использование методов, приведенных в работе, позволяет решать различные практические задачи.

**8 слайд**

Для примера рассмотрим проблему разработки проекта новой социальной сети. Перед разработчиками рано или поздно встанет проблема выбора дизайна и внешнего вида. Пусть наша сеть только начинает свой путь на рынке, поэтому она находится в бета-тесте и пользователей немного, допустим 1000 человек, из них 150 человек пригласим на тест. А также пусть у нас есть три варианта дизайна. Мы разделим пользователей на группы по 50 человек и дадим каждой группе доступ к своему типу дизайна. Будем отслеживать показатели среднесуточного использования соцсети, а также их активные действия и переходы по рекламным ссылкам. После завершения тестирования необходимо будет понять, насколько сильно и статистически значимо отличаются показатели ежедневного онлайна для различных типов оформления.

Также мы исследуем зависимость количества рекламных переходов от онлайна и активности.

**9 слайд**

Для начала каждую группу проверим на нормальное распределение (пользуемся модификацией критерия Крамера-Мизеса-Смирнова, в работе он показал наилучшие результаты). У каждой группы мы найдем МО онлайна а также его ДИ. В условиях малых выборок информации о пользователей и для большей точности мы начнем проверку гипотез о конкретных значениях и подберем наиболее верное (в данном случае необходимо получить минимально возможный показатель z статистики в сравнении с критическим значением). Проверяем значения из полученного ДИ начиная С ТО МО. Это и есть то самое комбинирование методов, о котором упоминалось выше. Проверки и вычисления приведены на слайде.

Далее для выяснения, являются ли полученные Расхождения в МО статистически значимыми, воспользуемся однофакторным дисперсионным анализом. После вычислений получаем p-value < 0.05 (подтверждает статистические различия), а также графики бокс плот (В данном случае можно сказать, что показатели дизайна С статистически значимо отличаются от показателей дизайна В). Полученная информация позволит понять, какой из типов дизайна и оформления наиболее приятен пользователям и способствует их ежедневному использованию соцсети и удержанию в ней.

**10 слайд**

Другим тестом может стать проверка вида зависимости рекламных переходов от онлайна и активности пользователя. С точки зрения логики зависимость доказана, мы хотим узнать ее вид и построить модель.

Для этого воспользуемся множественной линейной регрессией по МНК. Для начала проверим наши данные на линейную зависимость, нормальность распределения остатков, нормальный тип распределения (уже проверяли ранее) и мультиколлинеарность. Найдем вектор коэффициентов нашей модели путем матричного умножения элементов выборки. После этого мы можем построить модель, найти исправленный КД и f-value для теста нашей модели. Оптимальная модель позволит владельцам соцсети прогнозировать количество рекламных кликов и понимать, как сильно и в каком виде на него влияют показатели онлайна и активности. Это очень полезно при предоставлении различных рекламных услуг.

**11 слайд**

Подводя итог важно отметить, что методы обработки малых выборок не ограничиваются приведенными в работе. Однако я считаю, что в работу включены полезные в практическом плане методы, которые по отдельности или в комбинациях позволяют решать широкий список проблем и задач из сферы бизнеса и исследований, что можно было увидеть на примере. Данные методы просты в понимании и программной реализации, поэтому они будут полезны для многих задач, в которых предстоит работать с малыми выборками.

Благодарю за внимание.