Глава 1

Теория вероятностей

1.1 Базовая вероятность

1. Вы изучаете не числа, а закономерности. Числа – лишь появление этих закономерностей. Что такое закономерность в нашем случае? Мы рассматриваем окружающий мир как вероятностный, когда, в основном, не происходит что-то, что детерминированное (однозначное), а явления случайные, т.е. возможны разные исходы.

2. Статистика: как по проявлениям вероятностной закономерности (вероятностной модели) узнать что-то про эту закономерность.

3. Соответственно, нужно как-то формализовать эти объекты – вероятностные модели и их проявления (выборку).

4. Итак, что такое вероятностная закономерность? Это вероятностное распределение (тут формулы). В дискретном случае – это состояния и вероятности. В непрерывном случае – плотность. Удобно говорить, что случайная величина имеет такое распределение, но это, скорее, формальность.

5. **Задание**. Приведите пример таких закономерностей, которые неизвестны, но хотелось бы узнать? Вернее, вы предполагаете, что закономерность такая-то (такое-то распределение, непрерывное или дискретное), но хотели бы в этом убедиться. Подпишите ось x какими-то реальными числами, как вы себе представляете эту закономерность (плотность или дискретное распределение).

6. Пример – баллы за тест на экзамене (возможно, опустить). Вопрос – зависят ли результаты теста, если проверяющий знает, на каком числе баллов граница между оценками (пусть по пятибальной системе)?

7. Итак, есть вероятностные распределения. Кстати, они необязательно одномерные. Нас интересуют какие-то их характеристики.

8. Вер.пространство, Независимость - $\mathsf{P}(AB) = \mathsf{P}(A)\mathsf{P}(B)$, несовместность - $\mathsf{P}(AB) = 0$. Очевидно, что если множества не нулевой вероятности, но несовместные события не могут быть независимыми. Пусть есть две монеты, результаты -

   $$\{(0,0),(0,1),(1,0),(1,1)\}.$$

   Если ввести случайную величину, то можно говорить, что $\mathsf{P}(\{(i,j)\}) = 1/4$. **Задание**: Приведите два независимых (докажите) и два несовместных события в случае двух бросаний монеты (элементарное событие - пара чисел).

9. Для примеров удобно пользоваться геометрическим определением равномерного распределения в области, когда вероятность пропорциональна размеру (длине, площади, объему) области. Я буду рисовать область, заштриховывая ее – имея в виду, что в этой области равномерное распределение. Можно говорить, что на квадрате задано равномерное распределение. А можно (удобно), что $(\xi, \eta) \in A$). Пусть распределение в прямоугольнике. На самом деле, штриховка означает двумерную плотность $p(x, y)$. Т.е. представьте себе площадку над заштрихованной областью. $p(x, y) = c$, если (x, y) лежит в заштихованной области, и 0 иначе. Так как площадь под графиком равна 1 ($\int\int p(x, y) = 0$), то константа равна 1, деленное на площадь заштрихованной области. Для равномерного распределения вероятность A равна площади A, деленной на площадь все области (говорят: вероятность пропорциональна площади). Например, вероятность половины прямоугольника - это его площадь, делённая на всю площадь, т.е. 0.5.

10. Очень важное понятие – это условные вероятности, они задают структуру того, что происходит. Определение: $\mathsf{P}(A \mid B) = \mathsf{P}(AB)/\mathsf{P}(B)$. Если A и B независимы, то $\mathsf{P}(A \mid B) = \mathsf{P}(A)$. Условие сужает область, делает как бы перенормировку. От этого вероятность может как увеличиться, так и уменьшится.

11. **Задание**. Что больше, вер-ть квадратика в круге или в полукруге? Нарисуйте такой квадратик, когда условная вероятность больше обычной и когда условная вероятность меньше оббычной.

12. Формула полной вероятности Пусть дано вероятностное пространство $(\Omega, \mathcal{F}, \mathbb{P})$, и полная группа попарно несовместных событий $\{B_i\}_{i=1}^k \subset \mathcal{F}$, таких что
$\forall i\ \mathbb{P}(B_i) > 0$;
$\forall j \neq i\ B_i \cap B_j = \varnothing$;
$\bigcup_{i=1}^k B_i = \Omega$.

    Пусть $A \in \mathcal{F}$ — интересующее нас событие. Тогда получим:

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

    Очень полезно для создания модели. Мы можем структурировать задачу, знаем, что на входе, и можем посчитать, что получается на выходе.

13. Теорема Байеса. Мы знаем, что на выходе, а можем посчитать, что было на входе.

$$P(B_i \mid A) = \frac{P(A \mid B_i)\, P(B_i)}{P(A)} = \frac{P(A \mid B_i)\, P(B_i)}{\sum\limits_{i=1}^{k} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}.$$

14. **Задание** – пример про прибор, вероятность быть здоровым.
    Пусть существует заболевание с частотой распространения среди населения 0,001 и метод диагностического обследования, который с вероятностью 0,9 выявляет больного, но при этом имеет вероятность 0,01 ложноположительного результата — ошибочного выявления заболевания у здорового человека. Найти вероятность того, что человек здоров, если он был признан больным при обследовании.

    Обозначим событие, что обследование показало, что человек болен, как «Б» с кавычками, Б — событие, что человек действительно больной, З — событие, что человек

действительно здоров. Тогда заданные условия переписываются следующим образом:

$$\begin{aligned} P(\text{«Б»} \mid \text{Б}) &= 0.9 \\ P(\text{«Б»} \mid \text{З}) &= 0.01 \\ P(\text{Б}) &= 0.001 \\ P(\text{З}) &= 1 - P(\text{Б}), \text{значит} : P(\text{З}) = 0.999 \end{aligned}$$

Вероятность того, что человек здоров, если он был признан больным равна условной вероятности:

$$P(\text{З} \mid \text{«Б»})$$

Чтобы её найти, вычислим сначала полную вероятность признания больным:

$$\begin{aligned} P(\text{«Б»}) &= P(\text{«Б»} \mid \text{З}) \cdot P(\text{З}) + P(\text{«Б»} \mid \text{Б}) \cdot P(\text{Б}) \\ &= 0.01 \times 0.999 + 0.9 \times 0.001 = 0.01089 \end{aligned}$$

Вероятность, что человек здоров при результате «болен»:

$$\begin{aligned} P(\text{З} \mid \text{«Б»}) &= \frac{P(\text{«Б»} \mid \text{З}) \cdot P(\text{З})}{P(\text{«Б»})} \\ &= \frac{0.01 \times 0.999}{0.01089} \approx 0.917 \end{aligned}$$

Таким образом, 91.7 % людей, у которых обследование показало результат «болен», на самом деле здоровые люди. Причина этого в том, что по условию задачи вероятность ложноположительного результата хоть и мала, но на порядок больше доли больных в обследуемой группе людей.

Если ошибочные результаты обследования можно считать случайными, то повторное обследование того же человека будет давать независимый от первого результат. В этом случае для уменьшения доли ложноположительных результатов имеет смысл провести повторное обследование людей, получивших результат «болен». Вероятность того, что человек здоров после получения повторного результата «болен», также можно вычислить по формуле Байеса:

$$\begin{gathered} P\big((\text{З} \mid \text{«Б»}) \mid \text{«Б»}\big) = \\ = \frac{P(\text{«Б»} \mid \text{З}) \cdot \big(P(\text{«Б»} \mid \text{З}) \cdot P(\text{З})\big)}{P(\text{«Б»} \mid \text{З}) \cdot \big(P(\text{«Б»} \mid \text{З}) \cdot P(\text{З})\big) + P(\text{«Б»} \mid \text{Б}) \cdot \big(P(\text{«Б»} \mid \text{Б}) \cdot P(\text{Б})\big)} = \\ = \frac{0.01 \times 0.01 \times 0.999}{0.01 \times 0.01 \times 0.999 + 0.9 \times 0.9 \times 0.001} \approx 0.1098 \end{gathered}$$

1.2 Случайные величины

1. Случайные величины. Их распределения.
   Случайная величина $\xi : (\Omega, \mathcal{F}, P) \mapsto (\mathbb{R}, \mathcal{B})$ — измеримое отображение. Задает распределение $(\mathbb{R}, \mathcal{B}, \mathcal{P}_\xi)$. Поэтому говорят о $\mathcal{P}_\xi(A) = P(\xi \in A)$ и можно рассматривать дискретные распределения (дискретные случайные величины), непрерывные распределения (непрерывные случайные величины), имеющие плотность $p(x) = p_\xi(x)$.

2. **Задание**. Случайная величина —- время из (вашего) дома до мат-меха. Нужно нарисовать плотность распределения случайной величины и отметить на оси x значение времени, за которое вы будете выходить из дома 1) на лекцию, 2) на контрольную работу.

3. Функция распределения. Определение $F_\xi(x) = P(\xi < x)$ (но более стандартно $F_\xi(x) = P(\xi \le x)$, разница только для дискретных распределений). Рисунки функций распределения. Для дискретного распределения, разница в том, $<$ или $<=$. **Задание**. Функция распределения случайной величины, которая всегда равна 1?

4. Характеристики положения (мат.ож., медиана), характеристики разброса (дисперсия, среднее абсолютное отклонение). Свойства мат.ож. и дисперсии.
   //Свойства дисперсии. Минимум достигается на мат.ож. **Задание**: нарисовать две плотности и соответствующие им функции распределения, с отличием только в среднем или с отличием только в разбросе.

5. Асимметрия и эксцесс. Соотношение мат.ож. и медианы.
   //Задание: пусть коэффициенты равны .... Чему они равны для $2\xi + 5$? **Задание**: Логнормальное распределение зарплаты (плотность нарисована мной). Вам предлагают в качестве начальной зарплаты среднюю зарплату или медианную. Какую вы выберете?

6. Примеры распределений: Нормальное, правило k сигм. Свойства хорошо известны. В частности, плотность имеет вид

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-a)^2}{2\sigma^2}},$$

   математическое ожидание равно a, дисперсия σ^2, асимметрия и эксцесс равны 0.

   Рассмотрим вопрос про измерение расстояния в сигмах. Будет говорить, что точка далеко от мат.ожидания, если это и более далекие значения маловероятны.

   Формально, пусть $\xi \sim N(a, \sigma^2)$. Рассмотрим $\mathsf{P}(|\xi - a| > k\sigma)$. Эта вероятность не зависит от σ и равна $2(1 - \Phi(k))$, где $\Phi(x)$ — функция стандартного нормального распределения $\mathrm{N}(0, 1)$.

   Значения $\mathsf{P}(|\xi - a| > k\sigma)$:

k	Вероятность
1	0.317
1.64	0.101
1.96	0.050
2	0.046
3	2.70E-03
6	1.97E-09

Отсюда правило двух сигм (вероятность быть на расстоянии от мат.ож. больше двух сигм примерно равна 0.05), правило трех сигм, правило шести сигм. **Задание**. Нарисуйте плотность распределения $N(1, 4)$.
Рост человека может иметь нормальное распределение?

7. Экспоненциальное (что происходит с мат.ож, когда lambda растет?), $p(x) = \lambda e^{-\lambda x}$, если $x \geq 0$; 0 иначе. $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$ (0 иначе)
пуассоновское $P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ $(k = 0, 1, 2, \ldots, \infty)$, Бернулли, биномиальное как сумма Бернулли, геометрическое и пр. **Задание**: какое мат.ож. и какая дисперсия у биномиального распределения (через мат.ож. и дисперсию суммы)? **Задание**: Одна из формул правильная для мат.ож. геометрического распределения (p/q или q/p). Какая (по смыслу) и почему? Считать ничего не нужно.

8. Сходимость в разном смысле. Очень важно, потому что часто мы что-то знаем про предел и можем этим пользоваться. Например, S_n (накопленные суммы) или среднее арифметическое. Или последовательность случайных величин, которая стремится к константе.

9. Пример: сходимость по вероятности к константе – начиная с некоторого момента вероятность попасть в eps-окрестность константы стремится к 1.

10. Сходимость по распределению: функция распределения стремится к предельной ф.р. кроме точек разрыва предельного распределения. **Задание** ξ_n, $\mathsf{E}\xi = 1$, $\mathsf{D}\xi = 1/n$. Докажите с помощью картинок, что есть сходимость и по вероятности, и по распределению к 0.

11. Закон больших чисел. На его основе считаем среднее арифметическое для оценки мат.ож. Неравенство Чебышёва. Из него следует ЗБЧ.

12. Центральная предельная теорема. На ней основано огромное количество теоретических результатов. Также, она объясняет, почему нормальная модель хорошо описывает большое количество явлений.

13. Сходимость для биномиального распределения к нормальному. **Задание**. Чему примерно равна вероятность того, что при 400 попытках число успехов больше 220.

14. Если останется время, то про мат.ож. и дисп., про медиану и абс.откл., про виды случ.величин (порядковые, качественные) и их характеристики.

1.3 Двумерные случайные величины. Зависимость

1. Переход к двумерным распределениям. Про плотности (найти картинку двумерной плотности). Про изображение плотности линиями уровня, или цветом, либо скаттерплотом из точек. Напомнить, про изображение плотности при равномерном распределении в области (или неравномерном).

2. Про одномерные (margin) распределения, формула. **Задание**. Я рисую плотность (нарисовать плотность (заштриховать ромб)) Нужно нарисовать одномерные распределения.

3. Про условные распределения, формула плотности. Суть – нормированный срез. Условное математическое ожидание – обычное мат.ож. условного распределения (например, с условной плотностью).

4. Независимость через произведение плотностей и через одинаковость условных вероятностных распределений. **Задание**: написать, где есть независимость, где ее нет. (по картинкам).

5. Зависимость как вид условного мат.ож. Регрессия: $y = \mathsf{E}(\eta|\xi = x)$. **Задание**. Пример с параллелепипедом. Нарисовать линию регрессии.

6. Виды регрессии – линейная, нелинейная, полиномиальная, ... Меры зависимости. Корреляция, корреляционное отношение. Линейная регрессия и просто регрессия. Свойства корреляции, некоррелированности и независимость, нарисовать картинки.

7. Коэффициент корреляции, через МНК. 2. Ковар. и корреляц.матрица. Изображение цветом. Задание. Изобразите корреляц.матрицу для признаков (вес, рост, возраст) для распределения характеристик взрослого человека 30-40 лет и для распределения всех живущих в большом доме в спальном районе. 3. Многомерное нормальное распределение. Расстояние Махаланобиса (=линии уровня плотности). Линейная регрессия. Формула плотности в невырожденном случае. Двумерный случай, как выглядит плотность распределения в зависимости от мат.ож, дисперсий и корреляции. Стандартное нормальное распределение. Показываю, как связаны параметры и линии уровня. (Рисую три графика с линиями уровня.) Варианты для корреляции 0.9, 0.4, 0, -0.4, -0.9, сопоставляю корреляц.матрицу, вектор дисперсий и вектор мат.ож.

8. Ковар. и корреляц.матрица. Изображение цветом. **Задание**. Изобразите корреляционные матрицы для признаков (вес, рост, возраст) для распределения характеристик взрослого человека и детей. Или: Изобразите корреляц.матрицу для признаков (вес, рост, возраст) для распределения характеристик взрослого человека 30-40 лет и для распределения всех живущих в большом доме в спальном районе.

9. Многомерное нормальное распределение. Расстояние Махаланобиса (=линии уровня плотности). Линейная регрессия. Формула плотности в невырожденном случае. Двумерный случай, как выглядит плотность распределения в зависимости от мат.ож, дисперсий и корреляции. Стандартное нормальное распределение. Показываю, как связаны параметры и линии уровня. (Рисую три графика с линиями уровня.) Варианты для корреляции 0.9, 0.4, 0, -0.4, -0.9, сопоставляю корреляц.матрицу, вектор дисперсий и вектор мат.ож.

   Задание: нарисовать линиями уровня плотность при мат.ож (1,2), стандартных отклонениях (2,1), корреляции - 0.8.

Глава 2

Статистика

2.1 Базовая статистика

1. Нас интересует вероятностное распределение (распределение случайной величины). Выборка, выборка одномерная, выборка двумерная, как реализации скаттерплот. **Задание**: нарисовать выборку согласно нормальному распределению с корр. -0.4, средними 1 и 2, стандартными отклонениями 2 и 1.

2. Теории на числах не построить. Нужно теперь связать выборку с закономерностью. Поэтому в мат.стат. выборка воспринимается в абстрактном смысле как n независимых случайных величин (и в двумерном случае). **Задание**: решите задание из теста: $\mathsf{E}(x_1 + x_3)$ и $\mathsf{E}(x_1 x_2) = ?$

3. Эмпирическое распределение. Связываем выборку и случайную величину через эмпирическое распределение. (одномерный и двумерный случай.) Функция распределения. Гистограмма как эмпирич. распределение.

4. Таким образом, схема такая: 1) нас интересует характеристика ξ
   2) подставляя вместо неизвестной ξ известную эмпирическую случайную величину, получаем оценку
   3) рассматриваем выборку как абстрактную (это случайный вектор), получается оценка – случайная величина. Выясняем, насколько оценка хорошая.

5. Пример 1) и 2) для мат.ож., выборочное среднее.

6. Примеры со значением параметра и плотностями оценок. (смещ., не смещ., разная дисперсия). **Задание**: упорядочить оценки по качеству, с объяснением.

7. Объяснением: сравниваем по MSE. Получаем сумма дисп. и смещения в квадрате.

8. Итак, свойства выб.среднего как оценки мат.ож. Нулевое смещение и поэтому дисперсия. Свойства оценок – несмещенность, состоятельность, ... **Задание**: Посчитайте дисперсию.

9. Выписываю то же самое для выборочной дисперсии.

10. Переходим к корреляции. Все то же самое. (на этом закончили)

11. Переходим к регрессии (МНК). Напоминаю для случ. величин. То же самое для выборки.

12. Проблемы с выбросами и нелинейными зависимостями.
13. Пример с ирисами
14. Надо бы пример с деньгами и логарифмированием.

Глава 3

Проверка гипотез

3.1 Общие сведения

3.1.1 Примеры гипотез

Пусть H_0 — это гипотеза (**h**ypothesis), т.е. некоторое предположение о случайной величине ξ, которое мы хотим проверить (модель — это предположение, которое считается верным без проверки). Она называется нулевой (null hypothesis), потому что позднее появится альтернативная к ней.

Важно, что гипотеза — предположение о неизвестном законе распределения ξ, а не о выборке.

Например, гипотеза о том, что мат.ож. давления до и после приема лекарств одинаково. Или гипотеза о том, что распределение ошибки прибора нормальное, или о том, что распределение генератора псевдослучайных чисел равномерное на [0,1], или что . Или о том, что зависимости между временем на смотрением анимэ и успехами в учебе нет. Тут прослеживается такая особенность: в гипотезе обычно предполагается, что эффекта нет, так как решение принимается, если гипотеза отвергается.

Задание 1: Напишите, какую гипотезу вам было бы интересно проверить и какие данные для этого нужно было бы собрать?

3.1.2 Критерий

Для проверки гипотезы применяется критерий. Критерий — это правило, по которому гипотеза либо отвергается, либо не отвергается. Про построение такого правила — позже.

Правило строится на основе выборки. Разумное решение: нам нужно построить правило, которое, как уже говорилось, показывает, насколько выборка отличается от тех предположений, которые постулируются в гипотезе. Если отличается сильно, то гипотеза отвергается.

Измерять отличие удобно в числах. Поэтому вводится статистика критерия (функция от выборки, на основе которой строится критерий) $t = t(x_1, \ldots, x_n)$, которая выборке сопоставляет число.

Например, гипотеза про то, что $H_0 : \mathsf{E}\xi = 0$ (например, условия тренировки не влияют на результат (средняя разница в результатах нулевая)). В этом случае, 0 — ожидаемое значение (expected), выборочное среднее $\bar{x}$ — наблюдаемое значение (observed). Их разница как раз измеряет отличие. Однако, просто по разнице не сказать, отличие большое или нет. И вообще, числа могут получиться случайно, на их основе теорию не построить. Поэтому нужен статистический подход, который мы опишем ниже.

На основе результатов критерия принимают решения. Например, если лекарство показало эффективность (гипотезу о том, что оно неэффективно, отвергли), то его запускают в производство.

3.1.3 Нет безошибочных решений

Проблема: случайно может произойти что угодно, т.е. безошибочных решений практически не бывает. Приходится задавать максимальный уровень вероятности ошибки, на который можно согласиться при принятии решения.

Задаем маленький уровень значимости (significance level) $0 < \alpha < 1$ и соглашаемся, что с вероятностью α будем принимать неправильное решение.

Что такое маленький? Это зависит от критичности ошибки при принятии решения. Например, принять решение о полете и полететь на неисправном самолете или принять решение взять зонт и зря носить зонт в сумке весь день.

Задание 2: Придумайте ситуацию (гипотезу), когда она верна, а вы ошибочно считаете, что на не верна и поэтому принимаете неверное решение. В одном случае на вероятность ошибочного решение больше 0.001 вы бы точно не согласились. Вторая ситуация — когда согласились бы и на 0.2, но, пожалуй, не больше.

3.1.4 Статистический критерий

Чтобы построить теорию и отвечать на вопрос, маленькое или большое значение статистики критерия, отвергать гипотезу или нет, нужно перейти на теоретический язык. Т.е., нужно рассматривать абстрактную выборку, 'до эксперимента', где x_i — одинаково распределенные независимые случайные величины с тем же распределением, что и у ξ.

Таким образом, и статистика критерия $t = t(x_1, \ldots, x_n)$ — тоже случайная величина. Если верна H_0, то t имеет некоторое распределение и принимает некоторый диапазон значений.

Например, для модели $\xi \sim N(a, \sigma^2)$ и гипотезы $H_0 : \mathsf{E}\xi = 0$, статистика критерия $t = \sqrt{n}\bar{x}/\sigma$ имеет распределение $N(0, 1)$. Видим, что возможны любые значения. Однако, если мы допускаем некоторую вероятность ошибочно отвергнуть верную нулевую гипотезу, то критерий можем построить. Обычно главное — контролировать эту вероятность.

Разбиваем значения статистики критерия на две части, доверительную и критическую область так, что вероятность для статистика критерия попасть в критическую область равна α.

Задание 3. Нарисуйте плотность распределения статистики критерия $t = \sqrt{n}\bar{x}/\sigma$ при условии, что верна нулевая гипотезы, и разбейте область значений на две части, доверительную и критическую. Сделайте это разумным образом, чтобы было разумно значения из критической области считать не соответствующими справедливости нулевой гипотезы.

Формальное определение Назовем критерием разбиение области значений статистики критерия на две части, $A_\alpha^{(\text{крит})}$ и $A_\alpha^{(\text{дов})}$, такие что вероятность ошибки первого рода $\alpha_I = \mathsf{P}_{H_0}(t \in A_\alpha^{(\text{крит})}) = \alpha$.

После того как критерий построен, пользуемся им уже в режиме 'после эксперимента', когда выборка и значение статистики критерия — числа. Если число t попадает в критическую область, то гипотеза отвергается. Иначе — не отвергается (но нельзя говорить, что принимается, это обсудим позднее).

Итак, важно (!): разбиение на доверительную и критическую область строится на теор. языке, для абстрактной выборки. А используется это разбиение уже для конкретной выборки, чисел.

Допустимо строить разбиение так, чтобы выполнялось $\alpha_I \leq \alpha$ (тогда критерий называется консервативным).

Часто удается построить только асимптотический критерий, когда $\alpha_I \to \alpha$ при $n \to \infty$. В этом случае критерий можно применять при достаточно (для критерия) большом объеме выборки, где допустимый объем выборки зависит от скорости сходимости.

Ниже более подробно.

3.2 Схема построение критерия на основе статистики критерия

1. Строим статистику критерия t так, что:

   - Статистика критерия t должна измерять то, насколько выборка соответствует гипотезе. В этом случае мы получаем значение статистики критерия для «идеального соответствия».

     Например, если гипотеза про математическое ожидание $H_0 : \mathsf{E}\xi = a_0$, то $t = \bar{x} - a_0$ подходит под это требование. Если гипотеза про дисперсию $H_0 : \mathsf{D}\xi = \sigma_0^2$, то соответствие правильнее измерять отношением и поэтому подошло бы $t = s^2/\sigma_0^2$.

     Пример. Пусть $H_0 : \mathsf{E}\xi = a_0$; тогда $t = \bar{x} - a_0$ и «идеальное значение» $t = 0$.

   - Распределение t при верной H_0 должно быть известно хотя бы асимптотически. Из-за этого часто преобразовывают меры несоответствия, приведенные выше. Для $H_0 : \mathsf{E}\xi = a_0$ в модели $\xi \sim N(a, \sigma^2)$ с известной дисперсией σ^2 удобно использовать статистику критерия $t = \sqrt{n}(\bar{\mathbf{x}} - a_0)/\sigma \sim N(0, 1)$.

     Для $H_0 : \mathsf{D}\xi = a_0$ в модели $\xi \sim N(a, \sigma^2)$ известно распределение статистики критерия $t = ns^2/\sigma_0^2 \sim \chi^2_{n-1}$.

2. Строим разбиение области значений статистики критерия t так, что:

   - $\mathsf{P}(t \in A_\alpha^{\text{крит}}) = \alpha$.
   - Если альтернативная гипотеза H_1 (см. про нее в след.разделе) не конкретизирована, то $A_\alpha^{(\text{крит})}$ следует выбрать так, чтобы она располагалась как можно дальше от идеального значения.

     Пример. Обозначения: pdf (probability distribution function) — это плотность, а cdf (cumulative distribution function) — это функция распределения.

     В случае $t \sim \mathrm{N}(0, 1)$ при идеальном значении 0, разумно определить $A_\alpha^{(\text{крит})}$ «на хвостах» графика плотности $\mathrm{pdf}_{\mathrm{N}(0,1)}$ симметрично по обе стороны от 0 так, что для $A_\alpha^{(\text{крит})} = (-\infty, -t_\alpha) \cup (t_\alpha, \infty)$

     $$\alpha/2 = \int_{-\infty}^{-t_\alpha} \mathrm{pdf}_{\mathrm{N}(0,1)}(y)\,\mathrm{d}y = \int_{t_\alpha}^{+\infty} \mathrm{pdf}_{\mathrm{N}(0,1)}(y)\,\mathrm{d}y.$$

     Иными словами,

     $$\alpha/2 = 1 - \mathrm{cdf}_{\mathrm{N}(0,1)}(t_1) \implies t_1 = \mathrm{cdf}^{-1}_{\mathrm{N}(0,1)}(1 - \alpha/2)$$

     и аналогично для t_0. Границы доверительной области часто называют критическими значениями.

- На будущее: если H_1 известна, то $A_\alpha^{(\text{крит})}$ выбирается так, чтобы максимизировать мощность критерия против альтернативы H_1, определения будут позже.

3.2.1 Пример с числами

Общая схема всех примеров будет как написано ниже.

- Модель/предположения: (необязательно, но если есть, то это нужно проверять/обсуждать до использования критерия)
- Гипотеза: H_0 : ...
- Статистика критерия t = ...:
- Ее распределение при условии, что верная H_0: ... — выписано распределение. Если распределение асимптотическое, то при применении критерия нужно обращать внимание на объем выборки.
- Разбиение значений статистики критерия на доверительную и критическую области.
- Дана выборка, дан уровень значимости.
- Задание: проверить гипотезу, сказать, отвергается она или нет.

Как решать:

1. Теор.часть, выборка абстрактная, уровень значимости α тоже произвольный. По виду статистики критерия вы понимаете, какое значение соответствует 'идеальному' соответствию данных гипотезе. Рисуете график плотности статистики критерия и разбиваете значения на доверит. и крит. части, чтобы вероятность попасть в крит. область была равна α. В крит.область включаете значения, наиболее далекие от 'идеального'.
2. Практическая часть, выборка состоит из чисел, уровень значимости — конкретное число. Подставляете в формулу статистики критерия числа, получаете число, обозначим t_0. Затем считаем, чему равны критические значения (граница(ы) между критической и доверительной областями). Эти числа выражаются через обратную функцию распределения (это квантили). Значения можно вычислить в R или Python, см. ниже приложение. Рисуем снова график плотности статистики критерия, отмечаем там найденные числа, показываем, где критическая область, где доверительная. На основе того, куда попало t_0, делаем вывод, отвергается или нет нулевая гипотеза.

Ниже я буду давать три варианта, в зависимости от остатка деления дня рождения на 3.

Задание 4: Провести эту схему (записать то, что выше, с самого начала, со слова Модель) для уже разобранного выше критерия со статистикой критерия $t = \sqrt{n}(\bar{\mathbf{x}} - a_0)/\sigma$. Там дисперсия σ^2 предполагается известной. Пусть она равна 1.44. Гипотеза: $H_0 : \mathsf{E}\xi = a_0$, где (0) $a_0 = -1$, (1) 0.5, (2) 1. Примените критерий для выборки $(0, 2, 1, -1, -2)$ и уровня значимости (0) $\alpha = 0.05$, (1) 0.1, (2) 0.2.

Задание 5: Провести эту схему для следующей постановки задачи.

Модель: ξ имеет распределение Бернулли с неизвестным параметром, вероятностью успеха p. Напомню, что это означает, что она принимает значения 0 и 1, 1 (успех) с вероятностью p.

Гипотеза: $H_0 : p = p_0$

Статистика критерия:

$$t = \sqrt{n}\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}},$$

где $\hat{p} = \bar{x}$, что логично, так как сумма значений выборки — это в точности число успехов.

Ее распределение при условии, что H_0 верна: $t \xrightarrow{\text{d}} \text{N}(0,1)$ (т.е. это асимптотический критерий).

Дана выборка в виде: число успешных собеседований 45, неуспешных — 55. Проверить гипотезу (0) $H_0 : p = 0.45$, (1) 0.5, (2) 0.4, уровень значимости (0) $\alpha = 0.2$, (1) 0.05, (2) 0.1.

Задание: проверить гипотезу, сказать, отвергается она или нет.

Задание 6: Провести эту схему для следующей постановки задачи.

Модель: нет (но предполагается, что ξ принимает конечное число значений).

Гипотеза:

$$H_0 : \mathcal{P}_\xi = \mathcal{P}_0, \text{ где } \mathcal{P}_0 : \begin{pmatrix} x_1^* & \dots & x_k^* \\ p_1 & \dots & p_k \end{pmatrix}.$$

Статистика критерия:

$$T = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}.$$

Здесь такие обозначения: выборка $\mathbf{x}$ сгруппирована, т.е. каждому x_i^* сопоставляем *наблюдаемую* абсолютную частоту n_i (сколько раз оно встретилось в выборке); np_i — *ожидаемая* абсолютная частота.

Ее распределение при условии, что H_0 верна: $T \xrightarrow{\text{d}} \chi^2(k-1)$. (т.е. это асимптотический критерий). По поводу свойств и вида плотности распределения хи-квадрат отсылаем к википедии https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D1%85%D0%B8-%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82.

Дана выборка в виде: число успешных собеседований 45, неуспешных — 55. Проверить гипотезу, что это распределение Бернулли с $p = 0.5$. Проверить для значений α от 0.01 до 0.99 с шагом 0.01.

Задание: проверить гипотезу, сказать, отвергается она или нет. Когда надоест перебирать уровни значимости, найдите такое пороговое значение, называемое p-значение, что при меньших уровнях значимости гипотеза не отвергается, а при бОльших — отвергается.

3.3 Понятие вероятностного уровня p-value.

Определение. *p-value* — это такой значение, что при значениях уровня значимости α, больших p-value, H_0 отвергается (по причине попадания t в $A_\alpha^{\text{крит}}$), а при меньших — не отвергается.

p-value — не вероятность, это пороговое значение. Неформально его можно интерпретировать как меру согласованности H_0 и выборки. Например, при больших значениях p-value практически при всех разумных уровнях значимости гипотеза не отвергается. При близких к нулю значениях p-value, наоборот, гипотеза будет отвергаться.

p-value — максимальное значение уровня значимости, при котором гипотеза не отвергается (значение статистики критерия попадает в доверит. область). Или, что эквивалентно, минимальное значение уровня значимости, при котором гипотеза отвергается.

Если критическая область определяется через превышение статистики критерия некоторого значения t_0, то есть еще определение p-value как вероятности того, что при повторных экспериментах статистика критерия будет больше, чем значение в текущем эксперименте. Это определение написано и в wikipedia, но оно не универсальное. Тем не менее, лучше его знать.

Задание 7 В заданиях 4, 5 и 6 найти p-value и сформулировать ответ в виде: при таких-то уровнях значимости гипотеза отвергается, при таких-то — не отвергается.

3.4 Приложение. Вычисление функции распределения и обратной к ней

https://rdrr.io/snippets/

По этому адресу можно делать вычисления он-лайн, вставив туда нужную часть кода

```
###normal distribution N(a, sd^2)
a <- 0
sd <- 1
x <- 2

#cumulative distribution function (cdf)
cdf <- pnorm(x, mean = a, sd = sd) print(cdf)

#inverse to this cdf
x <- qnorm(cdf, mean = a, sd = sd)
print(x)

###chi-square distribution chi2(m), where m is degree of freedom
x <- 240
m <- 200

#cumulative distribution function (cdf) of chi2(m)
cdf <- pchisq(x, df = m) print(cdf)

#inverse to this cdf
x <- qchisq(cdf, df = m)
print(x)
```

Просто онлайн калькуляторы:
https://planetcalc.ru/4986/, https://www.statdistributions.com/normal/ (для ф.р. нужен left tail) — нормальное распределение,
https://www.statdistributions.com/chisquare/ (для ф.р. нужен left tail) — распределение хи-квадрат.

3.5 Ошибки 1 и 2 рода. p-value

3.5.1 p-value

1. Повторяем про p-value для примера с хи-квадрат.

2. Возвращаемся к гипотезе про мат.ож. Вопрос – как там посчитать p-value, если $t_0 = 1$? Если $t_0 = -1.5$?

3. P-value – мера согласия данных с гипотезой. Минимальный уровень значимости, при котором гипотеза отвергается. Проверили гипотезу, что производительность труда не зависит от вознаграждения. Получили p-value 0.01. Что это означает? Значимость коэффициента корреляции. Гипотеза о том, что корреляция времени на дорогу в кафе и времени, проведенном в кафе, незначима. Получили p-value 0.8. Что это

означает? Задание: Проверяли много верных гипотез, каждый раз считали p-value. Какие p-value могли получиться? (приведите какие-нибудь 10)

4. Т.о., p-value строится так, что если alpha>p, то гипотеза отвергается. Но чему должна быть равна вероятность того, что alpha>p ? Она должна быть равна alpha, по определению (вероятность отвергнуть H0, если она верна). P – функция от выборки, поэтому случайная величина. Получаем равномерное распределение. Напомним, что ошибка 1 рода - .... Для точного критерия она равна альфе. Нам важно понять, точный ли критерий? Возможны разные ситуации. Можно провести моделирование (случайное разыгрывание ситуации). Как оценивать вероятность? Задание. 10 раз моделировали, получились такие p-value: 0.27, 0.34, 0.5, 0.7, 0.15, 0.65, 0.1, 0.55, 0.01, 0.45. Постройте график эмпирической функции распределения p-value и скажите, можно ли пользоваться таким критерием? А если 0.9, 0.25, 0.33, 0.5, 0.67, 0.11, 0.78, 0.44, 0.82, 0.99? Пусть строят для своих 10 чисел.

   Научить читать график распределения p-value. Консервативный и радикальный критерии. Задание. Попросить нарисовать распределение p-value для этих случаев.

3.5.2 Ошибки 1 и 2 рода

1. Про ошибки 1 и 2 рода. Мощность, состоятельность. Показать картинку, объяснить, что на ней нарисовано. **Задание**: найти ошибку 2 рода против альтернативы (указать ее), по вариантам. Пусть дисперсия у всех 4, объем выборки 100.

   (0) a0 = 1, a1 = 0.5
   (1) a0 = -1, a1 = -1.2
   (2) a0 = 0, a1 = 0.1

2. Зависимость мощности от ....

3. 1 рода – контролируем, 2 рода – какая получится.

4. Пример с самолетом.

5. Если известно, с какой стороны альтернатива, то ...

6. Распределение p-value для нахождения мощности критерия. **Задание**. Предлагаю три графика. Характеризуйте критерий, точный, консервативны, радикальный. Маленькая мощность, большая мощность (рисую три пары картинок, причем радикальным нельзя пользоваться).

7. FPTN и пр.

3.6 Доверительные инntервалы

1. Про доверительные интервалы, в целом. Пример, как построить доверит. интервал для матем.ожидания, без модели и с моделью (что все равно, если нет нормальности). Про распределение Стьюдента, коротко. Исправленная дисперсия.

2. Про асимптотические доверит.интервалы для мат.ож.

3. Про использование доверит.интервалов для проверки гипотез. Пример с числами

4. **Задание**. Сделать выводы (по файлу, приготовить и выложить перед занятием).

5. Про p в Бернулли. **Задание**: написать, как будет выглядеть дов.интервал. Рассказать про более точный интервал Wilson'a.

6. **Задание**: Построить доверит.интервал для p, если $\bar{x} = 0.5$, $n = 100$ (0) $\gamma = 0.9$, (1) $\gamma = 0.95$ (2) $\gamma = 0.99$. Проверить гипотезу с соотв. альфой, что (0) $p_0 = 0.6$ (1) $p_0 = 0.4$ (2) $p_0 = 0.7$.

7. Интересные примеры.

8. Проверка гипотезы про нулевую вероятность, крит.область справа.

9. Про одинаковые числа подряд.

10. $\gamma = 0.2$, $c_\gamma = 0.25$. $0.2^5 = 0.0003$, $0.2^{10} = 10^{-7}$

11. $\gamma = 0.7$, $c_\gamma = 1$. $0.7^5 = 0.17$, $0.7^{10} = 0.03$.

3.7 Построение оценок

1. Методы построения оценок. Метод подстановки. Метод моментов. **Задание**: равномерное распределение, Пуассоновское распределение, экспоненциальное распределение.

2. Определение функции правдоподобия. Максимальное правдоподобие – что делать, если нет учителя. Ответ: максимизировать правдоподобие.

3. Пример: распределение Бернулли, нормальное распределение при известной дисперсии. Пример: пуассоновское распределение, экспоненциальное распределение. **Задание**. Найти ОМП. Совпадают ли с оценкой по ММ? Пример р.р. на $[0, \theta]$, где разные.

4. **Задание**. Есть оценка θ, несмещенная. Утв. существует константа c, такая что дисперсия больше c. (или меньше c). Выберите то, что считаете верным. Неравенство Рао-Крамера.

5. **Задача**: есть модель: $a + (laplace(\lambda)), a + N(0, \sigma^2)$. Какие наилучшие оценки для a?

6. Выборка не повторная, разные дисперсии. **Задание**. Найти оценку ОМП.

7. Использование ОМП: хи-квадрат, ... (вряд ли что-то еще), model-based кластеризация, тематическое моделирование, классификация с помощью логистической регрессии, информационные критерии для выбора модели.

8. Информационные критерии. $AIC = 2k - 2\ln L$, $BIC = k\ln n - 2\ln L$. **Задание**. Вам нужно сравнить две модели данных, экспоненциальную и логнормальную. Объем выборки 55 (ln 55 примерно равен 4). Exp: подставили ОМП в L и получили exp(-13). Lognorm: exp(-10), т.е. правдоподобие в случае логнормального больше.

9. Нарисовать картинки и посчитать AIC через SSE.

3.8 Робастность

1. Что такое выброс. Выброс по отношению к закономерности. Выброс и неоднородность. Задание. Нарисовать картинки и спросить, где выбросы и по отношению к чему?

2. Робастность. Выб.ср. и выб.медиана. T-test и MW.

3. Коэф.корр.Пирсона и Спирмена. **Задание**: какой коэффициент корреляции больше по модулю?

4. M-оценки.

3.9 Множественное тестирование

1. Множественное тестирование. Проблема: ошибка FWER. Задача: посчитать FWER для независимых тестов. Комикс. **Задание**: объяснить комикс.

2. Решение: изменить уровень значимости для одного теста. **Задание**: изменить на поправку p-value. Этот тест точный, т.е. FWER = alpha

3. А если тесты не независимые? Тогда поправка Бонферрони. **Задание**. В каком случае поправка Бонферрони приведет к максимально консервативному тесту? Если все тесты полностью зависимы (выдают одинаковые p-values).

4. Иногда получается построить точный тест для зависимых сравнений. Post-hoc comparisons. Таблица с **заданием**.

5. (распечатать 6 и 7 из файла)