Замечание. *Полную версию факультатива можно найти на* `github.com/bolychevanton/LMSH54`

Говорят, Джордж Р.Р. Мартин, автор цикла "Песнь Льда и Огня", истребляет Старков: чаще "убивает" персонажей, относящихся к этому дому, чем персонажей других домов. В таблице 1 приведено количество персонажей, относящихся к тому или иному дому, упомянутых за первые 4 книги, а так же количество погибших персонажей. Предлагается протестировать отличие уровня смертности дома Старков от уровня смертности каждого из других домов на 5% уровне значимости. Необходимо привести значения оценок вероятностей смертельных исходов для всех домов, найти p-value для каждого их трех тестов, а также проделать данные эксперименты с использованием метода Бонферрони.

Дом	Упомянутые персонажи	Погибшие персонажи
House Stark	72	18
House Lannister	49	11
House Greyjoy	41	12
Night's Watch	105	41

Таблица 1: Данные взяты из датасета `https://www.kaggle.com/mylesoneill/game-of-thrones`

Замечание 1. *В таком прекрасном разделе математики как теория вероятностей случайные величины, которые принимают всего два значения* 0 *и* 1*, называют* ***бернуллиевскими*** *случайными величинами. При этом считается, что единица принимается с конкретной вероятностью* p*, а ноль, соответственно, с вероятностью* $1-p$*. Число* p *называют параметром бернуллиевской случайной величины.*

Задача 1. *Нарисуйте график функции* $\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$*.*

Задача 2. *Пусть* $F(x)$ *— это площадь под графиком функции* $\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$ *от* $-\infty$ *до* x*.*

1. *Как будет выглядеть график функции* $F(x)$*?*

   P.S. Без доказательства считаем, что $F(0) = 0.5$ *и* $F(x) \to 1$ *при* $x \to +\infty$*, или, по-другому, что прямая* $y = 1$ *— это горизонтальная асимптота* $F(x)$*.*

2. *А как будет выглядеть график функции* $F^{-1}(\alpha)$*, обратной к* $F(x)$*?*

Задача 3. *Пусть* $X_1, X_2, X_3, \ldots, X_N$ *— бернуллиевские случайные величины с параметром* p*.*

1. *Какие значения может принимать* $X_1 + X_2 + \cdots + X_N$*? С какими вероятностями?*
2. *А какие значения может принимать* $\frac{X_1+X_2+\cdots+X_N}{N}$*? И с какими вероятностями?*
3. *Нарисуйте график функции* $f(t)$*, где* $f(t) = \mathsf{P}(\frac{X_1+X_2+\cdots+X_N}{N} = t)$ *— вероятность того, что* $\frac{X_1+X_2+\cdots+X_N}{N} = t$*. Если* $\frac{X_1+X_2+\cdots+X_N}{N}$ *не может принимать конкретное значение* t*, то тогда считаем, что* $f(t) = 0$*. Для простоты положим* $N = 5$ *и* $p = \frac{1}{2}$*.*

Теорема 1 (Центральная предельная теорема для бернуллиевских случайных величин)**.** *Пусть* $X_1, X_2, \ldots, X_N$ *— бернуллиевские случайные величины с параметром* p*. Обозначим* $\hat{p} = \frac{X_1+X_2+\cdots+X_N}{N}$*, тогда*

$$\mathsf{P}\left(\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}} \leq F^{-1}(\alpha)\right) \approx \alpha$$

при достаточно больших N*. Как правило, данная теорема хорошо работает при* $N \geqslant 30$*.*

Теорема 2 (Другая центральная предельная теорема для бернуллиевских случайных величин)**.** *Пусть* $X_1, \ldots, X_N$ *— бернуллиевские случайные величины с параметром* p_X*, а* $Y_1, \ldots Y_M$ *— бернуллиевские случайные величины с параметром* p_Y*. Обозначим* $\hat{p}_X = \frac{X_1+\cdots+X_N}{N}$ *и* $\hat{p}_Y = \frac{Y_1+\cdots+Y_M}{M}$*, тогда*

$$\mathsf{P}\left(\frac{(\hat{p}_X-\hat{p}_Y)-(p_X-p_Y)}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{N}+\frac{\hat{p}_Y(1-\hat{p}_Y)}{M}}} \leq F^{-1}(\alpha)\right) \approx \alpha$$

Пусть $X_1, X_2, \ldots, X_N$ — бернуллиевские случайные величины aka персонажи Старков. Будем считать, что каждый Старк умирает с вероятностью p_S, то есть $\mathsf{P}(X_i = 1) = p_S$ и $\mathsf{P}(X_i = 0) = 1 - p_S$ для всех i. Аналогично, пусть $Y_1, Y_2, \ldots, Y_M$ — бернуллиевские случайные величины aka персонажи Ланнистеров с $\mathsf{P}(Y_i = 1) = p_L$, $\mathsf{P}(Y_i = 0) = 1 - p_L$ для всех i.

Задача 4. *Пусть* $\hat{p}_S = \frac{X_1 + \cdots + X_N}{N}$ *и* $\hat{p}_L = \frac{Y_1 + \ldots Y_M}{M}$. *По таблице 1 определите, чему равны* $\hat{p}_S$ *и* $\hat{p}_L$.

Итак, можно ли утверждать, что Старки умирают так же, как Ланнистеры? Сформулируем данный вопрос на математическом языке:

$$H_0 : p_S = p_L, \qquad H_1 : p_S \neq p_L \tag{1}$$

В данном аккуратном вопросе, начинающемся с «можно ли», скрыто ещё одно важное обстоятельство: нам нужен конкретный критерий, согласно которому мы считаем, когда *можно*, а когда *нельзя*. Данный критерий носит название *уровня значимости*. Его, как правило берут равным 5%, как мы в общем и сделаем. А сейчас попробую расшифровать, что это такое.

Обозначим $\alpha = 0.05$ и сделаем следующее: построим множество, в которое при верности H_0 некий крокодил попадёт с вероятностью не менее $1 - \alpha$. Данное множество назовём *критическим*. Далее посмотрим на крокодила, попутно проверяя, принадлежит ли он критическому множеству или нет. Если принадлежит, то принимают H_0, иначе — отвергают H_0 в пользу H_1. В качестве так называемого крокодила предлагается взять разность $p_S - p_L$, которая должна быть равна 0 в случае верности H_0. Итак, если H_0 верна, то, согласно теореме 2, получаем

$$\mathsf{P}\left(0 \in \left[(\hat{p}_S - \hat{p}_L) - F^{-1}(1-\alpha)\sqrt{\frac{\hat{p}_S(1-\hat{p}_S)}{N} + \frac{\hat{p}_L(1-\hat{p}_L)}{M}}, +\infty\right)\right) \approx 1 - \alpha$$

Задача 5. *Проверьте* (1) *на* 5% *уровне значимости, если* $F^{-1}(0.95) \approx 1.645$.

Задача 6. *Проверьте на* 5% *уровне значимости, действительно ли Старки умирают чаще*

(a) Грейджоев?

(b) Ночного Дозора?

Мы уже находимся на финишной прямой. Мне хотелось бы рассказать про ещё одну очень популярную и нужную штуковину. Она называется p-value. Но прежде давайте вспомним про уровень значимости. Отметим, что чем он больше, тем меньше критическое множество, и, соответсвенно, тем вероятнее мы отвергнем H_0. Так вот, p-value — это минимальный уровень значимости, при котором мы отвергаем H_0. Давайте поясним на примере нашей задачи. Итак, мы отвергаем H_0, когда

$$0 < (\hat{p}_S - \hat{p}_L) - F^{-1}(1-\alpha)\sqrt{\frac{\hat{p}_S(1-\hat{p}_S)}{N} + \frac{\hat{p}_L(1-\hat{p}_L)}{M}}$$

И нам нужно найти минимальное α, при котором данное неравенство будет выполнятся. Это и будет искомое p-value. Строго говоря, минимум не может быть достигнут в принципе в силу строгого неравенства. По факту это то же самое, что и потребовать найти минимальное действительное x, при котором $x > 2$. На самом деле, в правильном определении используется не минимум, а так называемый инфимум, но об этом вам расскажут позже на мехмате[1], поэтому мы (пока) закроем глаза на математическую строгость высказывания и положим, что p-value — это решение уравнения

$$0 = (\hat{p}_S - \hat{p}_L) - F^{-1}(1-\text{p-value})\sqrt{\frac{\hat{p}_S(1-\hat{p}_S)}{N} + \frac{\hat{p}_L(1-\hat{p}_L)}{M}}$$

По факту это будет «пограничным значением» уровня значимости: если p-value меньше заданного в условии уровня значимости α, то мы отвергаем H_0, иначе — принимаем. Посчитаем p-value для каждого из случаев

	Старки vs Ланнистеры	Старки vs Грейджои	Старки vs Ночной Дозор
p-value	0.37255026511742795	0.6871878448525566	0.9779318062843195

Мы по отдельности сравнили Старков с Ланнистерами, Грейджоями и Ночным Дозором. По факту у нас есть 3 отдельные гипотезы с тремя отдельными p-value, однако мы хотим оценить ситуацию в общем! Воспользуемся методом Бонферонни: если все найденные p-value меньше чем $\alpha/3$ (тройка в знаменателе — количество гипотез), то тогда признаем, что Старки не умирают чаще остальных персонажей, иначе — сделаем вывод, что Мартин чрезмерно жесток к Старкам.

Задача 7. *Действительно ли Джордж Мартин жесток к Старкам?*

[1] Если вы, конечно, поступите, что вряд ли