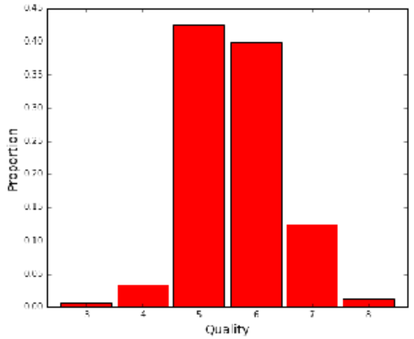
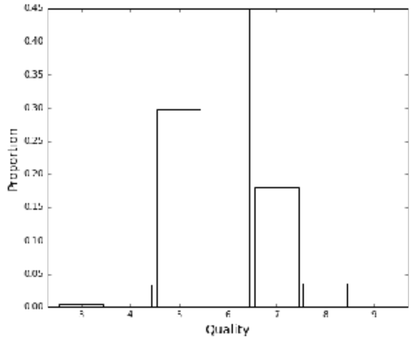
R для терьера и матстата

Посиделка четвёртая: Метод максимального правдоподобия

Задача матстата

$$\longrightarrow X$$





- Сундук плюётся! Надо по плевкам восстановить его внутренности!

	Type	Fixed acidity (g/l)	Volatile acidity (g/l)	Citric acid (g/l)	Residual sugar (g/l)	Chlorides (g/l)	Free sulfur dioxide (mg/l)	Total sulfur dioxide (mg/l)	Density (g/cm3)	pH	Sulphates (g/l)	Alcohol (%)	quality
1038	red	8.7	0.41	0.41	6.2	0.078	25.0	42.0	0.99530	3.24	0.77	12.6	7
4292	white	6.7	0.23	0.25	1.6	0.036	28.0	143.0	0.99256	3.30	0.54	10.3	6
5960	white	6.5	0.20	0.33	1.5	0.039	36.0	110.0	0.99008	3.22	0.65	12.0	6
2218	white	7.4	0.19	0.30	1.4	0.057	33.0	135.0	0.99300	3.12	0.50	9.6	6
743	red	11.6	0.41	0.58	2.8	0.096	25.0	101.0	1.00024	3.13	0.53	10.0	5

Как их восстановить?

1. Собрать репрезентативную выборку (что такое репрезентативность?).

2. Предположить из какого распределения эта выборка, проверить гипотезу об этом.

3. Оценить параметры распределения (или модели) методом максимального правдоподобия или другим методом.

4. Проверить выполнены ли все предпосылки, проверить все гипотезы о свойствах сундука, которые нас интересовали.

Что хочет статистик?

1. Несмещённости
2. Состоятельности
3. Эффективности

Кексы

Кекс 1

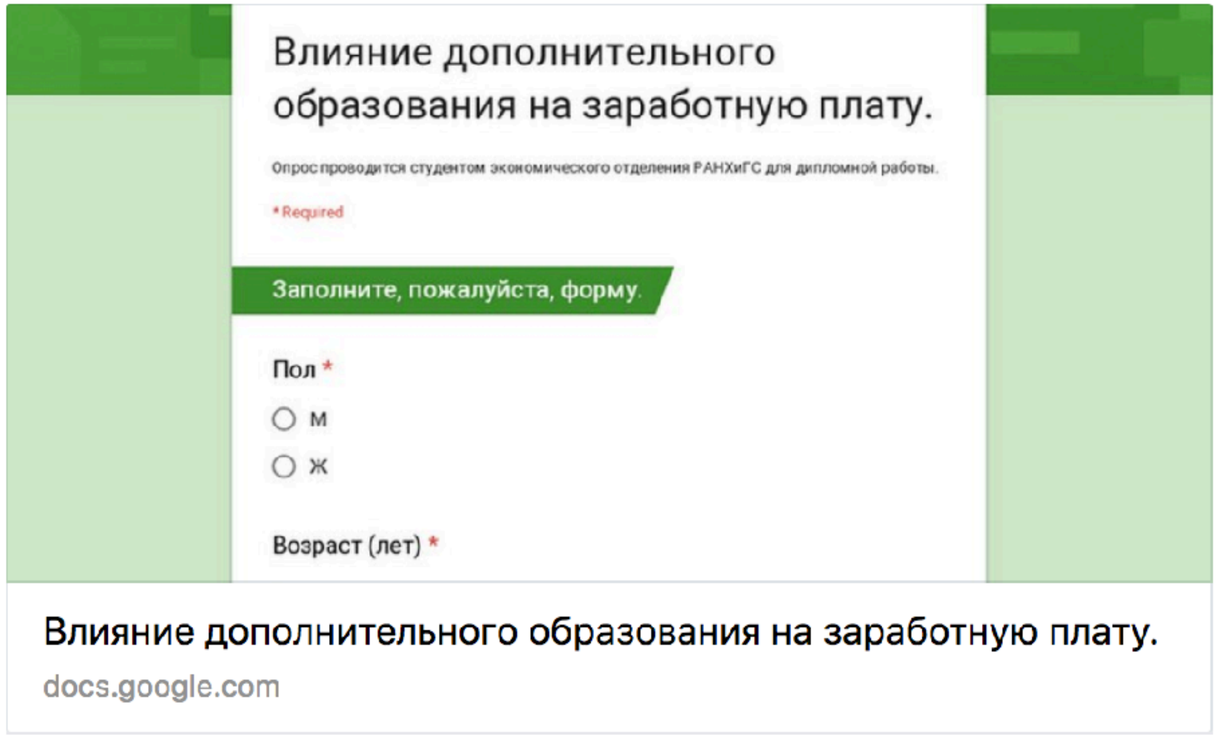
- Журнал «Литерари Дайджест» опросил 10 млн. человек насчёт выборов президента.
- Предсказал победу республиканцу Альфу Лэндону со значительным перевесом (60 на 40),
- Выборы выиграл демократ Франклин Рузвельт - как раз с таким же перевесом, но в обратную сторону.
- Как думаете, почему?

Кекс 2

- Часто вконтакте можно увидеть такие посты.

- Как считаете, какие проблемы возникнут у исследователя с выборкой? Удастся ли ему получить хорошие оценки?

Помогите нашему студенту в рисерче. Пожалуйста пройдите небольшой опрос, он займет менее 5 минут.
https://docs.google.com/forms/d/e/1FAIpQLSfnp-8-jzV_Y..



Влияние дополнительного образования на заработную плату.
docs.google.com

Кекс 3

- Фермер хочет оценить урожайность пшеницы от количества внесённых удобрений.

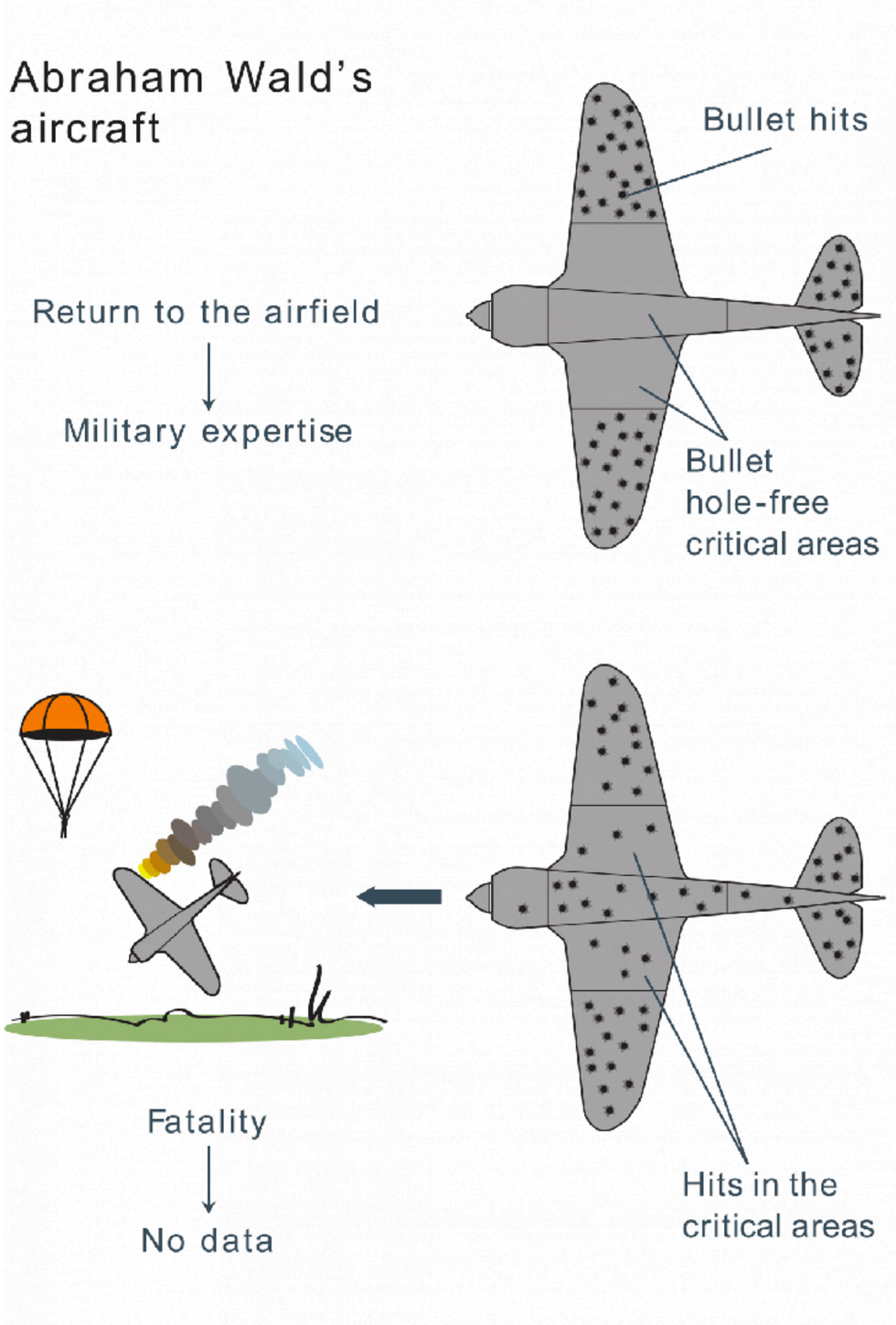
- Для этого он проводит эксперимент по выращиванию пшеницы. Он делит поле на две части. На правую он вносит удобрения, на левую нет.

- Как думаете, у исследователя получится адекватно оценить влияние удобрений на урожайность?

Кекс 4

- Во время Второй Мировой войны американские военные собрали статистику попаданий пуль в фюзеляж самолёта.
- По самолётам, вернувшимся из полёта на базу, была составлена карта повреждений среднестатистического самолёта.
- С этими данными военные обратились к статистику Абрахаму Вальду с вопросом, в каких местах следует увеличить броню самолёта.
- Что посоветовал Абрахам Вальд и почему?

- Сбитые самолёты не прилетают. Надо укреплять те части, в которых нет дырок.



Кекс 4

- Хотим рекомендательную систему! Что делать? Какие проблемы?

j

	Вечернее платье	Поднос для писем	iPhone 6s	Шуба D&G
Маша	1		1	
Юля	1	1		1
Вова		1	1	
Коля	1	?	1	
Петя		1	1	
Ваня			1	1

Правдоподобие

Задача про фонтан

- Саша приехал в южный город и увидел, что там есть фонтан и он работает!

- А как часто он работает?

Задача про фонтан

- Саша приехал в южный город и увидел, что там есть фонтан и он работает!

- А как часто он работает?
- Гипотезы:

1. Фонтан работает раз в году и нам повезло приехать в правильный день
2. Фонтан работает каждые выходные, а мы приехали в выходные
3. Фонтан работает всегда

Задача про фонтан

- Саша приехал в южный город и увидел, что там есть фонтан и он работает!

- А как часто он работает?

- Гипотезы:

	Вероятности:
1. Фонтан работает раз в году и нам повезло приехать в правильный день	$\frac{1}{365}$
2. Фонтан работает каждые выходные, а мы приехали в выходные	$\sim \frac{1}{3}$
3. Фонтан работает всегда	1

Задача про фонтан

- Саша приехал в южный город и увидел, что там есть фонтан и он работает!

- А как часто он работает?
- Гипотезы:

Вероятности:

1. Фонтан работает раз в году и нам повезло приехать в правильный день

$$\frac{1}{365}$$

2. Фонтан работает каждые выходные, а мы приехали в выходные

$$\sim \frac{1}{3}$$

3. Фонтан работает всегда

$$1$$

Максимальная вероятность увидеть работающий фонтан при верности гипотезы

Чуть более формально

Выборка: $x_1, \ldots, x_n$

Предположили, что она взята из распределения с плотностью $f(x \mid \theta)$

Параметр θ — это константа, про которую ничего не знаем и хотим оценить

Правдоподобие данных, то есть вероятность пронаблюдать именно эту выборку:

$$L(\theta \mid x_1, \ldots, x_n) = f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Чуть более формально

Выборка: $x_1, \ldots, x_n$

Предположили, что она взята из распределения с плотностью $f(x \mid \theta)$

Параметр θ — это константа, про которую ничего не знаем и хотим оценить

Правдоподобие данных, то есть вероятность пронаблюдать именно эту выборку:

$$L(\theta \mid x_1, \ldots, x_n) = f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Чуть более формально

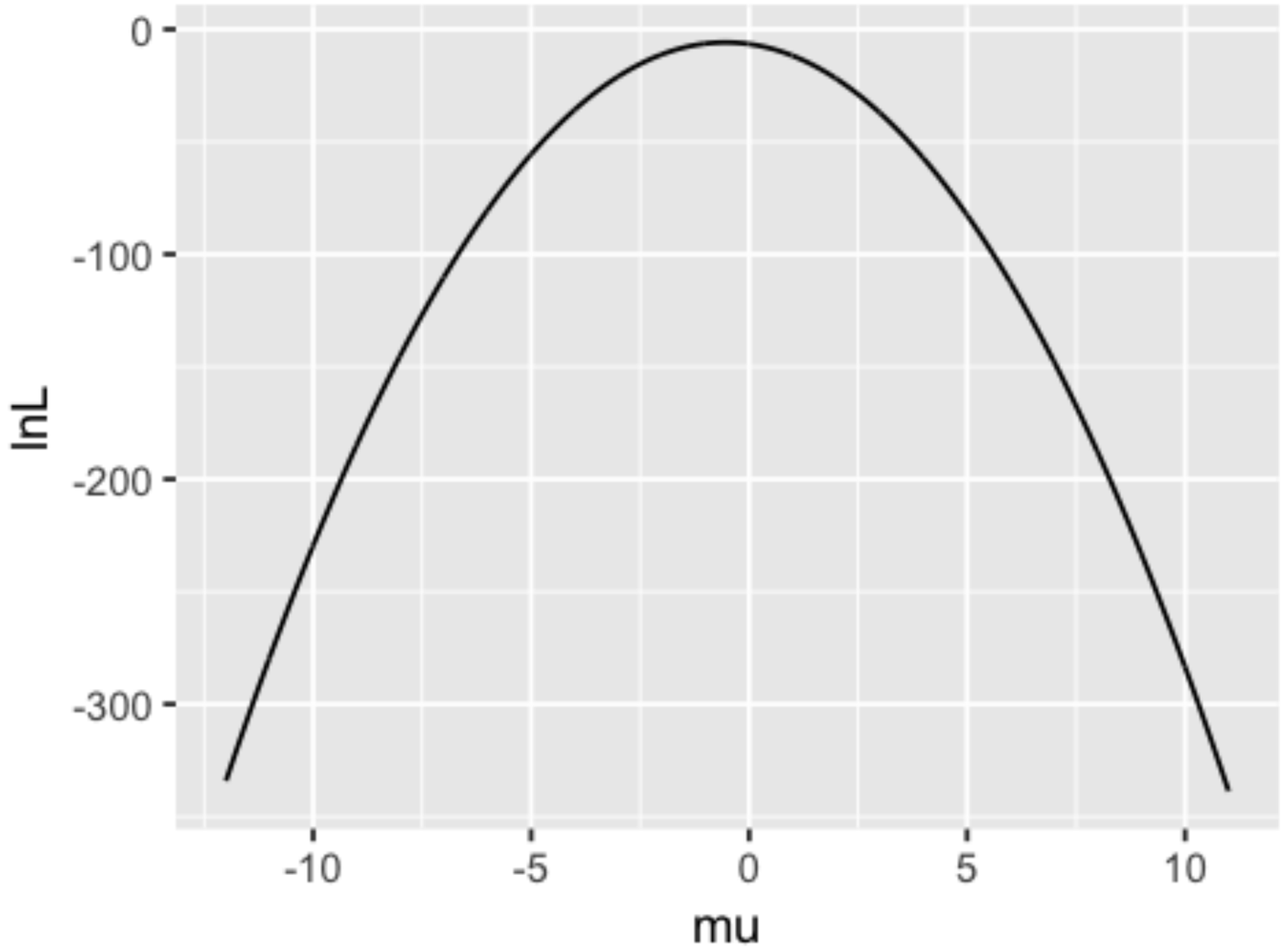
Для максимизации надо брать производные. Их приятнее брать от логарифма.

$$\ln L(\theta \mid x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i \mid \theta).$$

Решив уравнение правдоподобия, получим $\hat{\theta}^{ML}$

$$\frac{\partial \ln L}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i \mid \theta)}{\partial \theta} = 0$$

Как выгляди правда?



Одно слагаемое - тоже правдоподобие!

Сумма логарифмов!

$$\ln L(\theta \mid x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i \mid \theta).$$

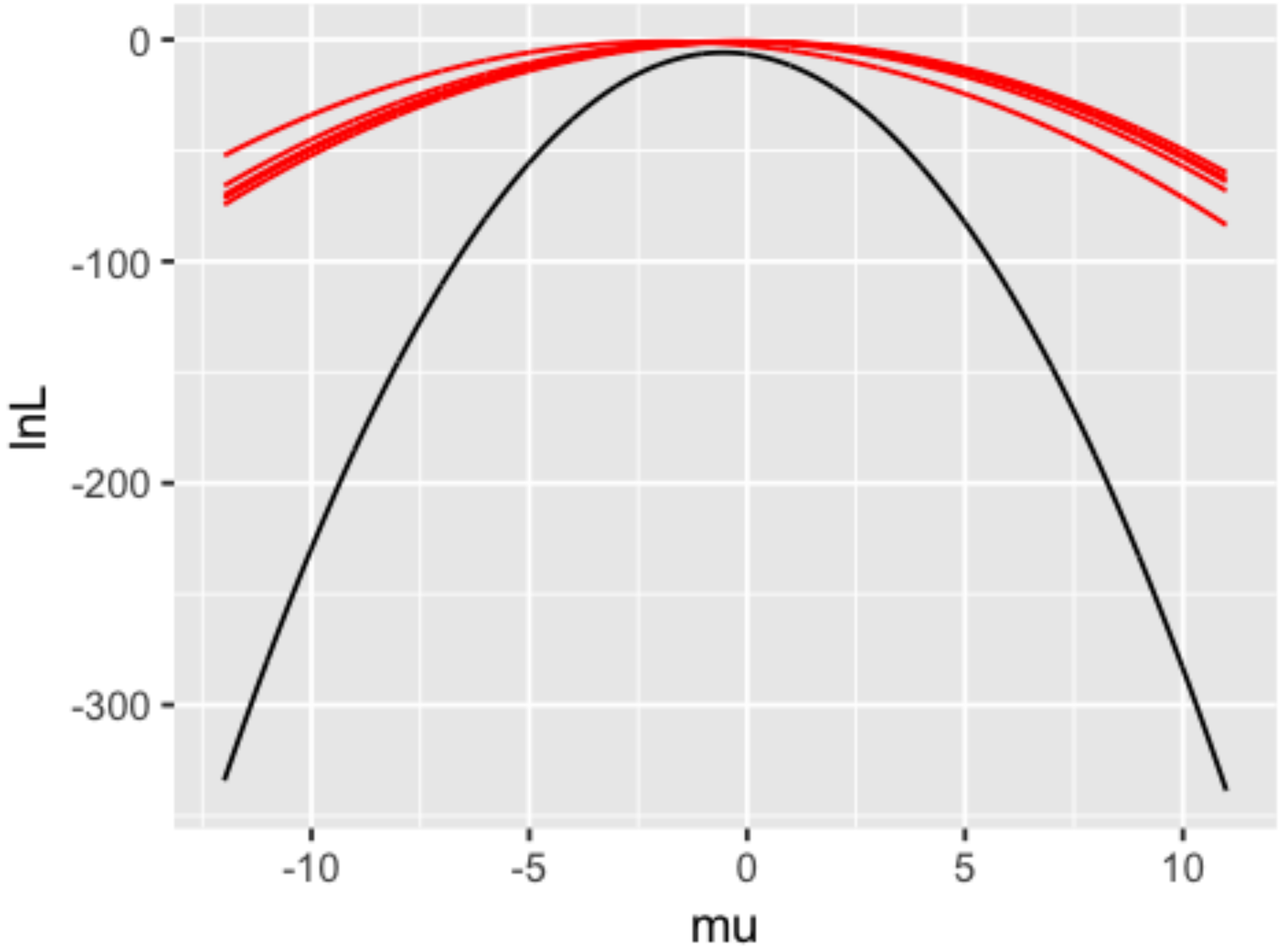
Одно слагаемое $\ln f(x_i \mid \theta)$ можно интерпретировать как логарифм правдоподобия для одного наблюдения x_i.

Как выгляди правда?

- Логарифмическая функция правдоподобия (черная линия) для всей выборки равна сумме логарифмических функция правдоподобия для отдельных наблюдений (красные линии).

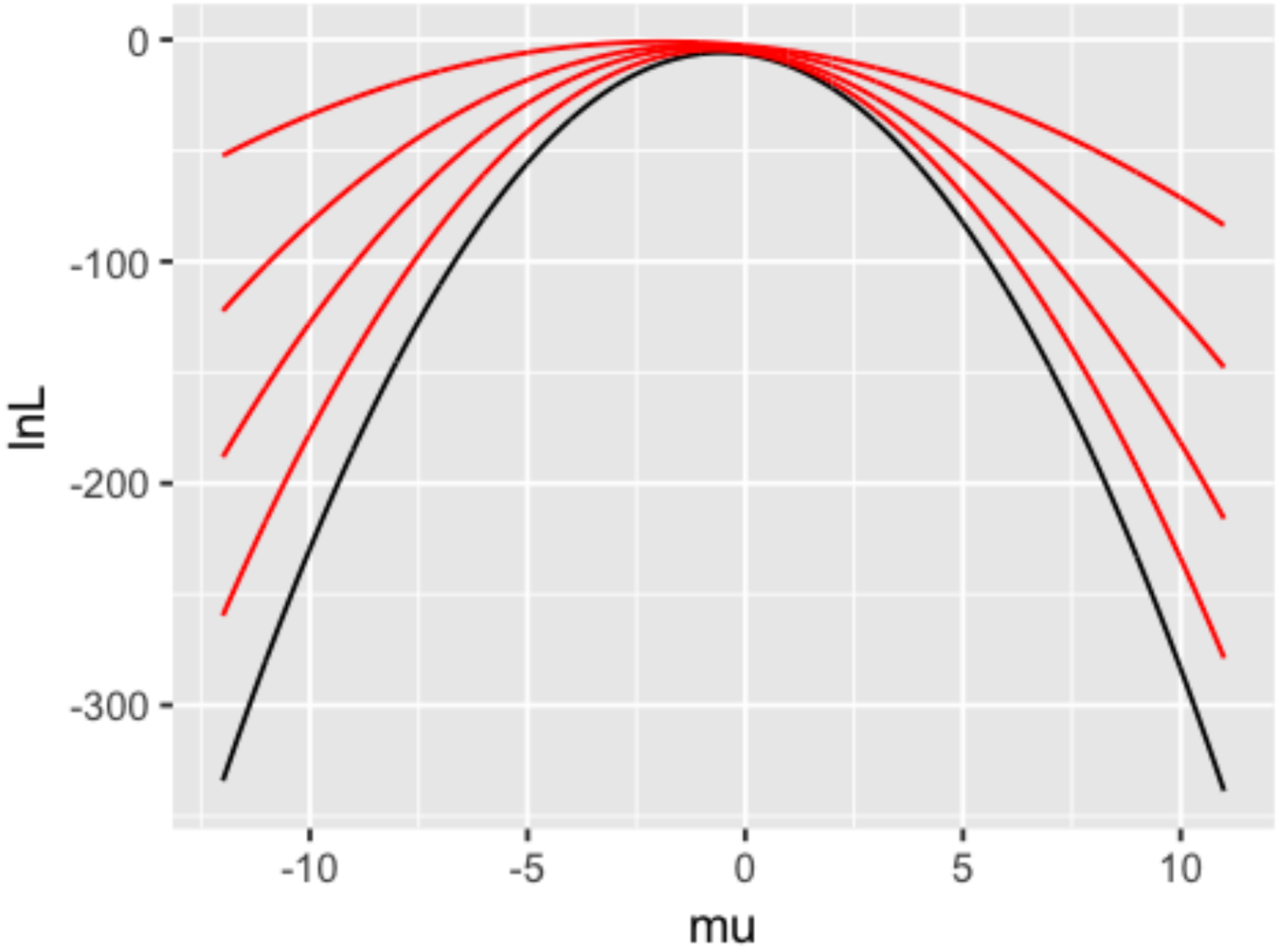
- У каждого слагаемого своя выпуклость. Чем более выпукло слагаемое, тем более острым оно делает пик при суммировании

- Общая функция правдоподобия имеет более выраженный максимум по сравнению с функциями правдоподобия для отдельных наблюдений



Как выгляди правда?

- Каждая красная линия - добавление к сумме нового слагаемоего.
- С каждым слагаемым максимум становится всё более выраженым.
- Каждое слагаемое добавляет нам информации. Объём этой информации зависит от выпуклости функции правдоподобия для отдельного наблюдения.



Выпуклость функции

Что отвечает за выпулость функции? Правильно! Вторая производная. Именно её, взятую со знаком минус, интерпретируют как наблюдённую информацию.

$$I_o(\theta) = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)$$

Со знаком минус, так как у нас максимум и вторая производная отрицательна.

Если параметр векторный, то в этом случае наблюдэнная информация представляется наблюдённой информационной матрицей:

$$I_o(\theta) = -\left(\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right) = -H$$

Выпуклость функции

Математическое ожидание этой матрицы по распределению x называется **информационной матрицей Фишера.**

$$I(\theta) = -E\left(\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right) = -E(H)$$

Ожидаемая информация зависит только от закона распределения наблюдений. Она отражает то, какую информацию в среднем вносит в наше правдоподобие, каждое наблюдение.

Выпуклость функции

Если функция плотности $f(x \mid \theta)$ удовлетворяет условиям регулярности, то тогда для любой несмещённой оценки $\hat{\theta}$ выполняется неравенство Рао-Фреше-Крамера:

$$Var(\hat{\theta}) \geq [I(\theta)]^{-1}$$

Более того, в этом случае

$$I(\theta) = E\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right) = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right],$$

но доказывать это здесь, мы конечно же не будем.

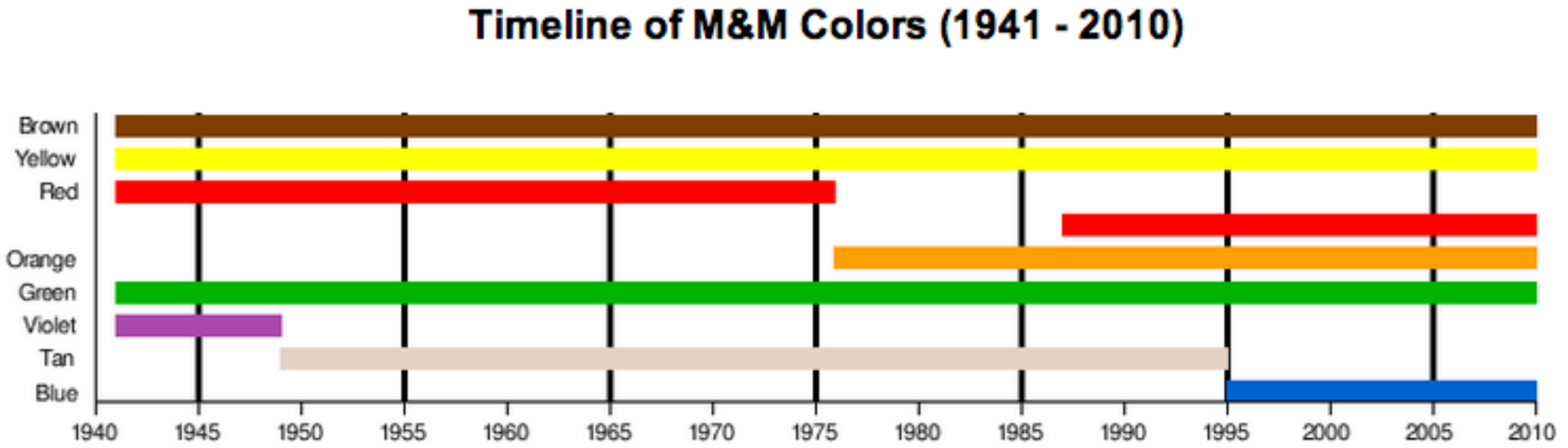
Приятные свойства оценки

- Состоятельность
- Асимптотическая несмещённость
- Асимптотическая эффективность
- Асимптотическая нормальность

$$\hat{\theta} \to N(\theta, I(\theta)^{-1})$$

Задача про M&Ms

- А как часто в упаковке попадаются жёлтые, красные и другие цвета?
- Я принёс выборку, её можно есть :)



Напоследок! Задача про наркотики.

- Все доставайте монетки и слушайте, что я буду говорить!!!