

# Прикладная статистика. Занятие 1. Введение.

8 февраля 2011 г.

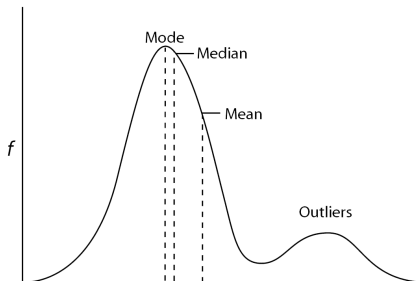
8.2	8.2	8.5	8.1	8.2	7.2	6.8	7.5	6.7	7.1	6.4	8.3	8.6	6.5	6.6
7.9	8.9	7.7	8.2	7.9	7.3	6.4	7.2	6.6	7.2	7.3	8	8.7	7.1	6.4
7.6	8.3	8.2	8.3	8.3	7.4	7.5	6.7	7.1	7	7.2	8.5	8.8	7.2	7.4
8	7.8	8.7	8.4	7.6	8.1	7.1	7.5	6.9	6.8	6.8	7.7	8.2	7.4	6.8
8	8	8.2	9.3	8	7.2	6.9	7.2	7	7.2	7	8.2	7.7	6.9	6.8
7.6	8.3	6.2	8	7.5	7.5	7.1	7.6	6.5	7.1	7	8.6	8.5	7.1	7.1
7.9	7.8	8.6	7.7	8.2	8.1	7.1	6.6	7.5	5.6	8.3	8.2	8.4	6.7	7.2
7.7	7.9	8.6	8.3	7.5	7	7.4	6.8	7.2	6.8	7.1	6.6	8.6	7.6	6.9
7.5	8.2	9.1	7.9	7.8	7.7	6.9	7	8.4	7	6.7	7	8.7	7.4	7.7
7.5	8.6	8.2	7.9	8.2	7.1	6.5	7.5	6.9	6.9	7.2	7.2	8.4	6.6	6.8
7.8	8.1	8.4	8.4	7.6	7.2	6.8	6.9	6.5	7.2	6.5	7.1	8.2	7.1	7
8.5	8.8	8.7	8.5	8.3	7.2	6.5	7.1	6.7	6.8	7.2	7	7.7	7.7	7.1
7.7	8	8.1	9.1	8	7.8	7.1	6.9	7.8	7.3	7	7.1	8	7	6.7
8.1	8.6	7.8	7.9	8	6.9	7.5	7.2	7.5	6.7	7.1	7.1	8.6	7	8
8.2	8.3	9	8.4	7.2	7	7.3	7	6.9	6.9	6.7	7.3	7.9	6.6	9
8.2	7.9	8.4	8	7.9	8.8	7	6.8	6.4	6.9	7.8	7.2	7.9	7.1	8.6
8.1	8.1	8.7	7.9	7.6	7.2	7.1	7.2	7	6.9	7.8	7.6	8.1	7.3	8.4
7.7	8.9	8.7	8.9	7.3	7.5	7.2	6.5	7.2	7.3	8.1	6.7	7.7	7.1	7.2
8	8.5	8.7	8.3	7.6	7.1	6.7	7.7	7.1	7.2	8	7.2	7.7	6.7	7.2
7.6	8.3	8.1	7.8	7.8	7.8	6.9	7	6.5	7	8.2	7.5	5.5	7	7
8.6	9.1	8.6	8.8	7.4	6.9	6.8	6.7	7.1	6.6	7.8	8	8	6.9	7.1
8	8	8.1	8	7.6	7.3	7.1	6.4	7	7.1	7.9	7.7	7.5	7.4	7.3
7.9	8.3	8.2	7.8	7.3	7.2	7.1	7.6	7.2	7	8.3	7.6	7.7	7.3	8
8.6	8.3	8.1	8	7.3	7.4	7	6.9	7.3	6.6	7.8	8.6	8.6	8.4	8.2
8.2	8.3	8.1	8.2	7.4	7	7.2	7.3	6.4	6.8	8.3	8.4	7.9	8.3	8.4
8	8.9	8.2	9.1	8	7	6.9	6.9	7.1	7.9	8.1	7.8	7.2	7	7.6
7.5	7.1	8.8	7.7	7.5	6.9	7.1	7.3	7	8	8.3	8.3	7.3	7.7	7.4
8.3	8.2	9.2	7.9	7.5	7.4	7.1	6.8	6.6	9	7.9	8.5	7.2	5.9	
7.6	8.3	9	8.9	7.4	7.1	7.1	7.7	6.1	8.3	9.1	8.1	6.9	7.7	
7.7	7.8	8.5	7.3	7.9	7.2	6.8	6.8	7.1	8.2	8.3	8.1	7.4	7.1	
8.1	8.2	8.4	8.2	7.8	6.5	6.8	6.9	6.8	8.6	7.9	8.4	6.8	7.2	
8.4	8.4	8.6	8.2	7.5	6.6	7.2	7.2	7	7.9	8.3	9	7.7	7.1	

## Оценки центральной тенденции

Выборочное среднее — среднее значение в выборке.

Медиана — средний элемент вариационного ряда выборки.

Мода — самое распространённое значение в выборке.



# How to lie with statistics (Huff, 1954)



\$45,000



\$15,000



\$10,000



← **ARITHMETICAL AVERAGE**

\$5,700



\$5,000



\$3,700



← **MEDIAN** (the one in the middle)  
(12 above him, 12 below)

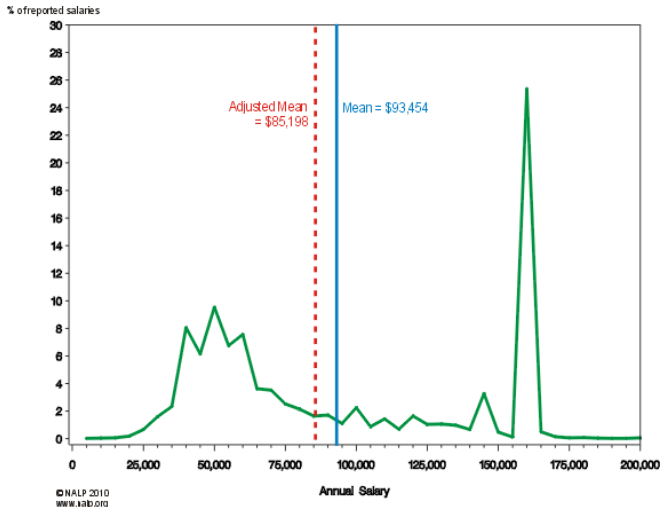
\$3,000



\$2,000

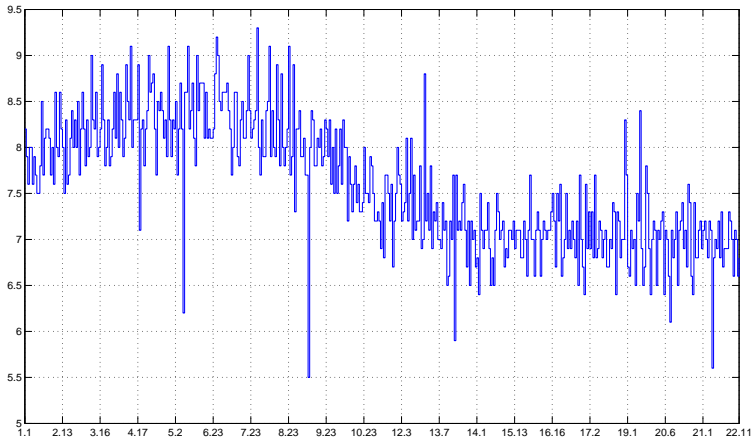
← **MODE**  
(occurs most frequently)

## Данные NALP (The Association for Legal Career Professionals)



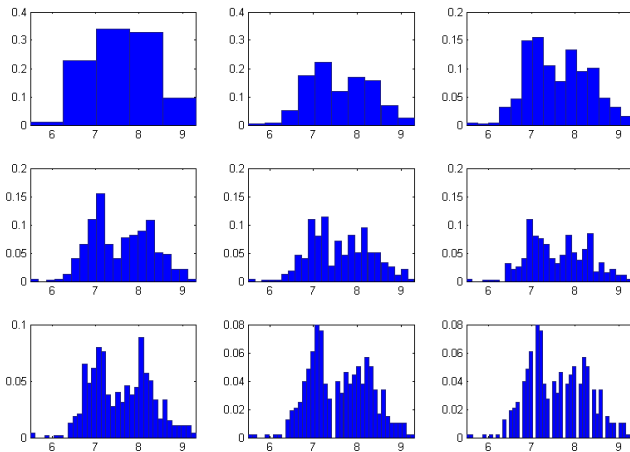
Salary Distribution Curve for the Class of 2009 Shows Relatively Few Salaries Were Close to the Mean

## Гистограмма: входные данные



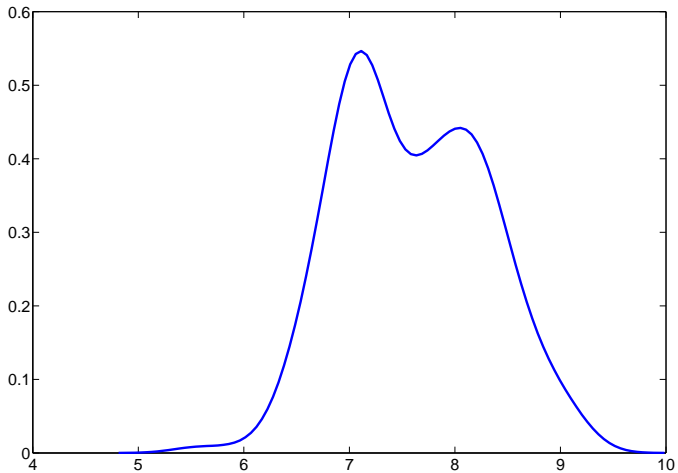
Рейтинги 475 эпизодов "Симпсонов" на [imdb.com](http://imdb.com)

# Гистограмма: число карманов



Правило Стурджеса:  $1 + \log_2 N$  (10)  
 Правило Райса:  $2\sqrt[3]{N}$  (15)

## Сглаженные оценки плотности распределения





# Проверка гипотез



## Shaken, not stirred

Джеймс Бонд говорил, что предпочитает мартини смешанным, но не взболтанным.

Слепой тест:  $n$  раз предложим Джеймсу Бонду пару напитков и выясним, какой из двух он предпочитает.

Выборка: бинарный вектор длины  $n$ , “1” — Джеймс Бонд предпочёл смешанный, “0” — взболтанный.

Нулевая гипотеза  $H_0$ : Джеймс Бонд не различает два вида мартини, т.е., выбирает наугад.

Статистика  $t$  — число единиц в выборке, т.е., число испытаний, в которых Джеймс Бонд предпочёл смешанный мартини.

## Нулевое распределение

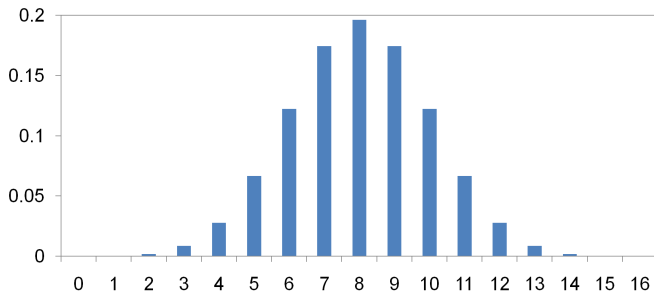
Если нулевая гипотеза справедлива и Джеймс Бонд не различает два вида мартини, то равновероятны все выборки длины  $n$  из нулей и единиц.

Пусть  $n = 16$ , тогда существует  $2^{16} = 65536$  равновероятных варианта.

0000000000000000	0100000000000000	1000000000000000	1100000000000000
0000000000000001	0100000000000001	1000000000000001	1100000000000001
0000000000000010	0100000000000010	1000000000000010	1100000000000010
0000000000000011	0100000000000011	1000000000000011	1100000000000011
0000000000000100	0100000000000100	1000000000000100	1100000000000100
0000000000000101	0100000000000101	1000000000000101	1100000000000101
0000000000000110	0100000000000110	1000000000000110	1100000000000110
0000000000000111	0100000000000111	1000000000000111	1100000000000111
0000000000001000	0100000000001000	1000000000001000	1100000000001000
0000000000001001	0100000000001001	1000000000001001	1100000000001001
0000000000001010	0100000000001010	1000000000001010	1100000000001010
0000000000001011	0100000000001011	1000000000001011	1100000000001011
...	...	...	...

## Нулевое распределение

Статистика  $t$  принимает значения от 0 до 16.

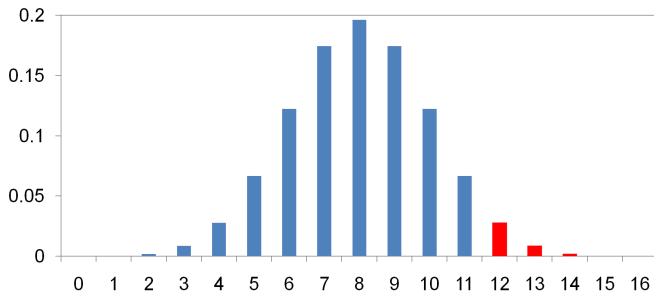


## Односторонняя альтернатива

Проверим нулевую гипотезу против альтернативы  $H_1$ : Джеймс Бонд предпочитает смешанный мартини.

При справедливости такой альтернативы более вероятны большие значения статистики  $t$  (большие значения  $t$  свидетельствуют против  $H_0$  в пользу  $H_1$ ).

Например, вероятность, что Джеймс Бонд предпочтёт смешанный мартини в 12 или более случаях при справедливости  $H_0$  равна  $\frac{2517}{65536} \approx 0.0384$ .



0.0384 — **достигаемый уровень значимости** при реализации статистики  $t = 12$ .

## Достигаемый уровень значимости

Чем ниже достигаемый уровень значимости, тем сильнее данные свидетельствуют против справедливости нулевой гипотезы в пользу альтернативы.

0.0384 — вероятность реализации достаточно большого значения статистики ( $t \geq 12$ ) **при условии, что нулевая гипотеза справедлива**, т.е., Джеймс Бонд выбирает мартини наугад.

Достигаемый уровень значимости нельзя интерпретировать как вероятность справедливости нулевой гипотезы!

## Достигаемый уровень значимости

Пример: утверждается, что осьминог предсказывает результаты матчей чемпионата мира по футболу с участием сборной Германии, выбирая кормушку с флагом страны-победителя. По результатам 13 испытаний ему удаётся верно угадать результаты 11 матчей. Соответствующий достигаемый уровень значимости —  $p \approx 0.0112$ .

0.0112 — не вероятность того, что осьминог выбирает кормушку наугад! Исходя из здравого смысла, стоит ожидать, что эта вероятность близка к единице.



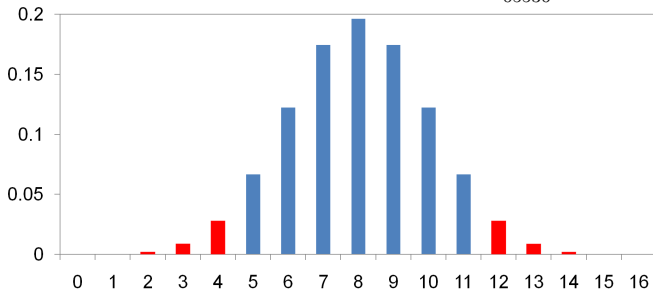
0.0112 — вероятность того, что осьминог угадает победителя в 11 и более матчах из 13 при выборе наугад.

## Двусторонняя альтернатива

Проверим нулевую гипотезу против альтернативы  $H_1$ : Джеймс Бонд предпочитает какой-то определённый вид мартини.

При справедливости такой альтернативы и большие, и маленькие значения статистики  $t$  свидетельствуют против  $H_0$  в пользу  $H_1$ .

Вероятность, что Джеймс Бонд предпочтёт какой-то один вид мартини в 12 или более случаях при справедливости  $H_0$  равна  $\frac{5034}{65536} \approx 0.0768$ .



0.0768 — **достижимый уровень значимости** при реализации статистики  $t = 12$ .



## Ошибки I и II рода

Задача проверки гипотез несимметрична относительно пары  $(H_0, H_1)$ .

	$H_0$ верна	$H_0$ неверна
$H_0$ принимается	$H_0$ верно принята	Ошибка второго рода
$H_0$ отвергается	Ошибка первого рода	$H_0$ верно отвергнута

Вероятность ошибки первого рода жёстко ограничивается малой величиной  $\alpha$  —  $H_0$  отвергается при  $p \leq \alpha$ .

Вероятность ошибки второго рода минимизируется путём выбора достаточно мощного критерия.

Если величина  $p$  достаточно мала, то данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Если величина  $p$  недостаточно мала, то данные не свидетельствуют против нулевой гипотезы в пользу альтернативы.

Интерпретация больших значений  $p$ -value

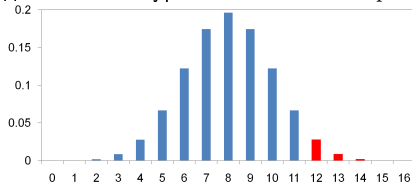
Пример: пусть Джеймс Бонд выбирает смешанный мартини в 51% случаев (ненаблюдаемая вероятность).

Пусть по итогам 100 испытаний Джеймс Бонд выбрал смешанный мартини 49 раз. Достижимый уровень значимости против односторонней альтернативы —  $p \approx 0.6178$ . Нулевая гипотеза не отвергается, при этом сказать, что она верна, было бы ошибкой — Джеймс Бонд выбирает смешанный или взболтанный мартини не с одинаковыми вероятностями!

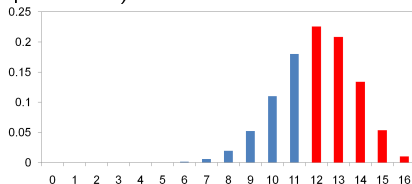
При помощи инструмента проверки гипотез нельзя доказать справедливость нулевой гипотезы.

## Мощность

Проверяя нулевую гипотезу о случайном выборе мартини против односторонней альтернативы, мы отвергаем  $H_0$  при значениях  $t \geq 12$ , что обеспечивает нам достигаемый уровень значимости  $p$  меньше  $\alpha = 0.05$ .



Пусть Джеймс Бонд выбирает смешанный мартини в 75% случаев (ненаблюдаемая вероятность).



Мощность равна  $\approx 0.6302$ , т.е., при многократном повторении эксперимента гипотеза будет отклонена только в 63% случаев.

# Мощность

Мощность зависит от следующих факторов:

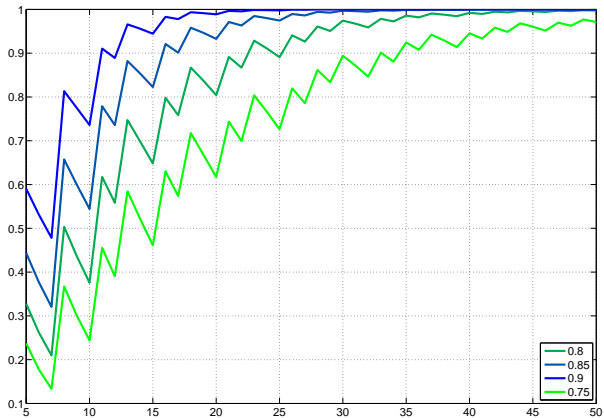
- альтернатива (односторонняя или двусторонняя)
- чувствительность статистики критерия
- размер отклонения от нулевой гипотезы
- размер выборки

## Размер выборки

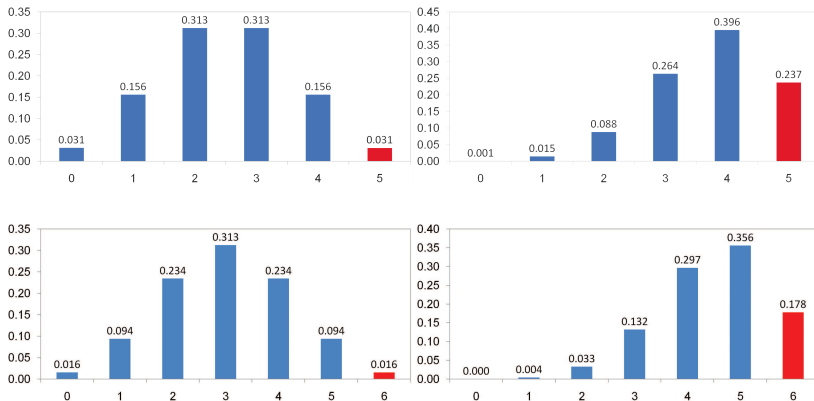
Нематематические соображения: 1 порция мартини содержит 55 мл джина и 15 мл вермута — суммарно около 25 мл спирта. Смертельная доза алкоголя при массе тела 80 кг составляет от 320 до 960 граммов спирта в зависимости от толерантности (от 13 до 38 мартини).

Обеспечение требуемой мощности: размер выборки подбирается так, чтобы при размере отклонения от нулевой гипотезы не меньше заданного (например, вероятность выбора смешанного мартини не меньше 0.75) мощность была не меньше заданной.

# Шокирующий график



# Падение мощности: объяснение



# Прикладная статистика

## Семинар 1. Введение.

Рябенко Евгений  
riabenko.e@gmail.com