Статистические гипотезы

Лекция 2

1. Общее представление о статистических гипотезах. Статистика и параметры.
2. Гипотезы о среднем. Распределение Стьюдента.
3. Сравнение двух выборок. Структурная модель Стьюдента.
4. Сравнение дисперсий. F-распределение.

Вопросы для обсуждения

ВОПРОС №1

Общее представление о статистических гипотезах.

- *Статистическая гипотеза* – это предположение по поводу *параметров* распределения случайной величины.
- Проверка статистических гипотез осуществляется путем сбора *статистики*.

Статистическая гипотеза

Параметры и статистика

Параметры

- Теоретическая величина характеризующая распределение случайной величины
- Имеет отношение к генеральной совокупности
- Практически никогда не известна

Статистика

- Эмпирическая характеристика, оценка параметра распределения случайной величины
- Имеет отношение к выборке
- Измеряется в ходе эксперимента

- $\mu = A$
- $\mu_1 = \mu_2$
- $\sigma^2 = 0$
- $\sigma^2 \neq 0$
- $\mu_1 \neq \mu_2$

Примеры гипотез

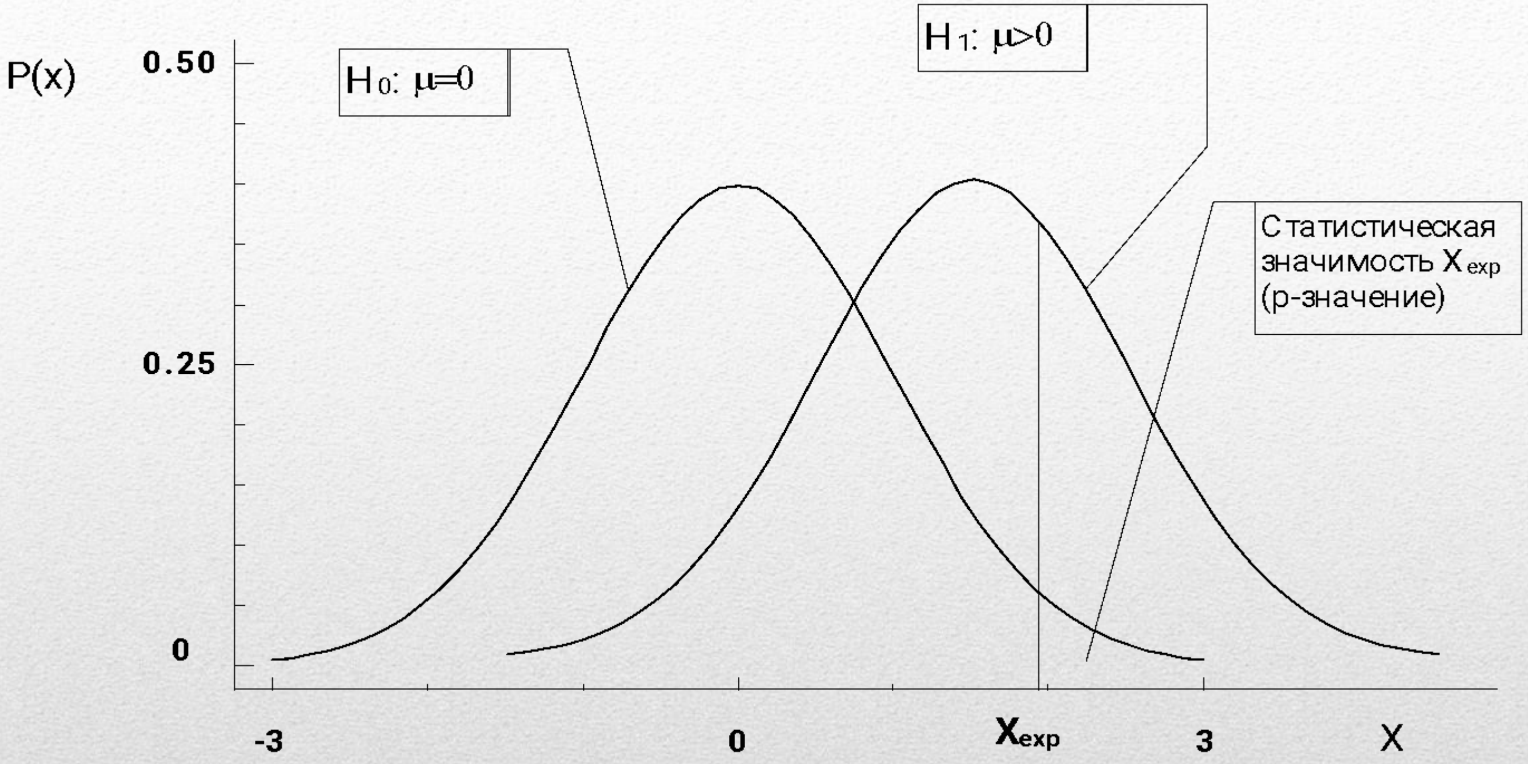
Нулевая (H_0)

- Утверждает что-то конкретное о параметрах распределения
- Истинность определяется на основе оценки статистики

Альтернативная (H_1)

- Утверждает что-то противоречащее нулевой гипотезе, менее конкретна
- Истинность определяется на основе рассмотрения нулевой гипотезы

Виды гипотез



Проверка гипотез

Гипотезы	H_0 принимается (H_1 отвергается)	H_0 отвергается (H_1 принимается)
H_0 верна (H_1 неверна)	Правильное принятие H_0 (правильное отвержение H_1)	Ошибка первого рода (α-ошибка)
H_0 неверна (H_1 верна)	Ошибка второго рода (β-ошибка)	Правильное отвержение H_0 (правильное принятие H_1).

Матрица исходов

- Теоретически не существует возможности со 100% вероятностью выбрать истинную гипотезу. Вне зависимости от установленного критерия всегда остается вероятность ошибки первого или второго рода.
- Уменьшая вероятность ошибки первого рода, мы увеличиваем вероятность ошибки второго рода и наоборот.

Статистическая надежность

P>0,10	H_0 принимается
P<0,05	H_1, как правило, принимается. Статистический вывод при этом признается *надежным*
P<0,01	H_1 принимается. Статистический вывод считается *высоко надежным*
0,05<p<0,10	Не представляется возможным принять ни H_0, ни H_1. Результат находится на границе уровней значимости (*маргинально* значим)

Уровни статистической надежности

ВОПРОС №2

Гипотезы о среднем. Распределение Стьюдента.

- Пусть есть вектор данных *X*
- Допустим, что *X* извлечены из нормальной совокупности с параметрами μ и σ^2
- Предположим: H_0: μ=A
- Тогда: H_1: $\neq$ A

Гипотезы о среднем

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

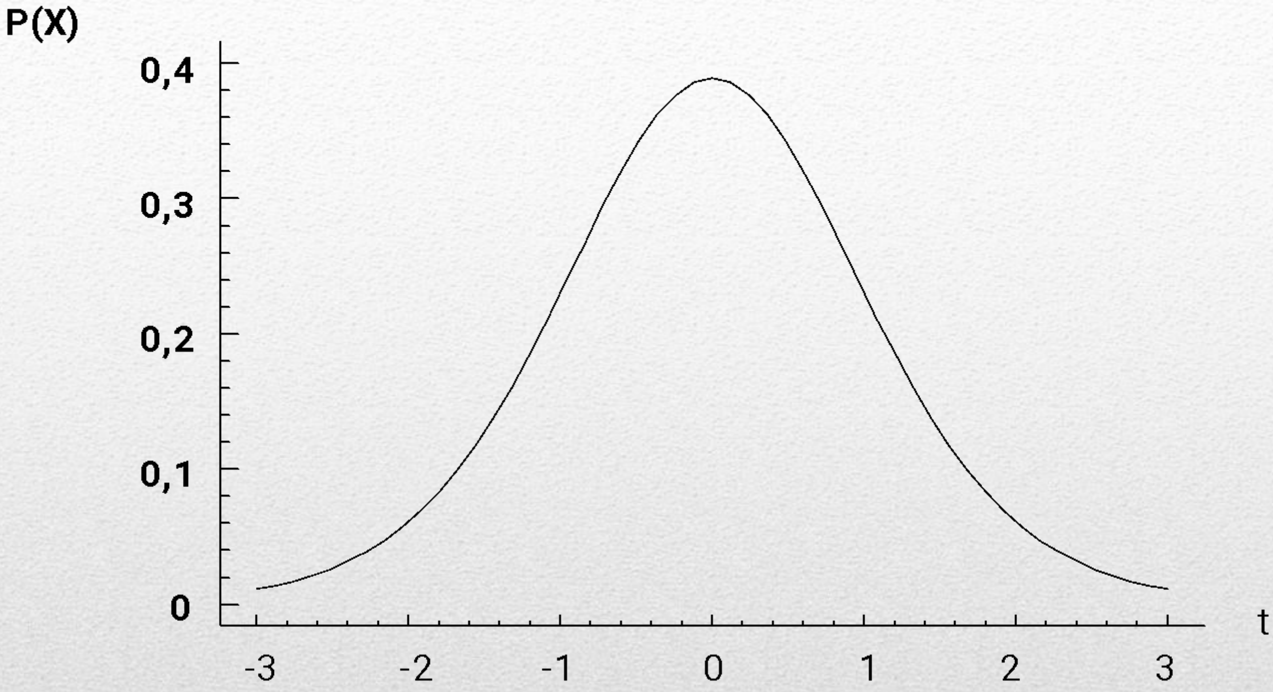
Случай №1: σ известна

$$t(n-1) = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Случай №2: σ неизвестна

- Распределение *t*-статистики отличается от нормального.
- Это распределение принято называть распределением Стьюдента, или просто *t*-распределением.
- Распределение Стьюдента симметрично относительно среднего и имеет небольшой положительный эксцесс.
- Оно характеризуется степенями свободы (обозначается *df*, от англ. *degrees of freedom*).
- Для данного случая число степеней свободы *t*-статистики на одну меньше объема выборки, т.е. равно *n-1*.

Статистика Стьюдента



t-распределение

ВОПРОС №3

Сравнение двух выборок. Структурная модель Стьюдента.

- Пусть есть два вектора данных – *X* и *Y*
- Допустим, что *X* и *Y* извлечены из нормальной совокупности с параметрами соответственно μ_X и σ_X и μ_Y и σ_Y
- Предположим: H_0: $\mu_X = \mu_Y$
- Тогда: H_1: $\mu_X \neq \mu_Y$

Сравнение двух выборок

- Пусть: $X_i = \mu_X + \varepsilon_{Xi}$ и $Y_j = \mu_Y + \varepsilon_{Yj}$
- Где: ε – величина статистической ошибки, распределенной в соответствии с нормальным законом с параметрами 0 и σ_ε

Структурная модель

$$\bar{x} = \mu_x + \bar{\varepsilon}_x$$

$$\bar{y} = \mu_y + \bar{\varepsilon}_y$$

$$\bar{x} - \bar{y} = \mu_x - \mu_y + \bar{\varepsilon}_x - \bar{\varepsilon}_y$$

$$(\bar{x} - \bar{y}) - (\mu_x - \mu_y) = \bar{\varepsilon}_x - \bar{\varepsilon}_y$$

Тогда…

- Сделаем неочевидное, но правдоподобное допущение, что дисперсии X и Y одинаковы.
- Поскольку дисперсии X и Y определяются дисперсией статистической ошибки ε, то

$$\sigma_{\varepsilon} = \sqrt{\sigma^2_{pooled}\left(\frac{1}{n}+\frac{1}{m}\right)}$$

Допустим…

$$s^2_{pooled} = \frac{\sum\left(x-\bar{x}\right)^2 + \sum\left(y-\bar{y}\right)^2}{(n-1)+(m-1)}$$

$$t(n+m-2) = \frac{\left(\bar{x}-\bar{y}\right)-\left(\mu_x-\mu_y\right)}{\sqrt{s^2_{pooled}\left(\frac{1}{n}+\frac{1}{m}\right)}}$$

Отсюда…

ВОПРОС №4

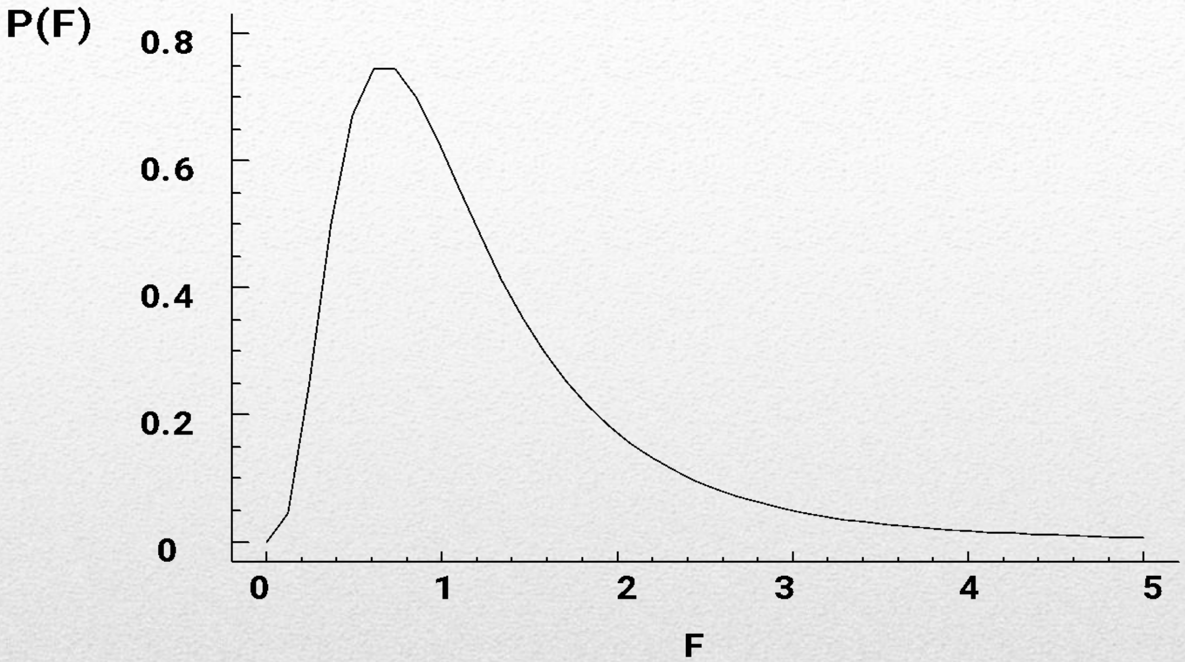
Сравнение дисперсий. F-распределение

- Пусть есть два вектора данных – *X* и *Y*
- Допустим, что *X* и *Y* извлечены из нормальной совокупности с параметрами соответственно μ_X и σ_X и μ_Y и σ_Y
- Предположим: H_0: $\sigma_X = \sigma_Y$
- Тогда: H_1: $\sigma_X \neq \sigma_Y$

Сравнение дисперсий

$$F\left(n-1, m-1\right) = \frac{s_x^2}{s_y^2}$$

F-статистика



F-распределение

www.ebbinghaus.ru