ТЕОРИЯ ВЕРОЯТНОСТЕЙ
И
МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Лекция 14

Проверка статистических гипотез

Статистическая гипотеза — предположение о некоторой закономерности, относящейся к одной или нескольким случайным величинам.

Заключение о правильности или неправильности гипотезы делается по данным выборок на основе того, какова вероятность получить эти выборки (одну или несколько), если гипотеза справедлива.

Гипотеза может состоять из нескольких предположений, часть из которых принимается без доказательства, а другая часть проверяется по данным выборок.

Пример гипотезы

Выборка: x = 160, 160, 167, 170, 173, 176, 178, 178, 181, 181.

Объем выборки N = 10.

Гипотеза — распределение x таково: $x \sim \mathbf{N}\left(167, \sigma_x^2\right)$

Вид распределения принимается без доказательства,

σ_x^2 надо оценить из выборки x,

главная часть гипотезы: $\left\langle x \right\rangle = 167$.

Далее обозначим гипотетическое математическое ожидание

как $\left\langle x \right\rangle^{(h)}$. Здесь $\left\langle x \right\rangle^{(h)} = 167$.

Идея проверки гипотезы:

вероятность того, что предполагаемое распределение породит такую выборку, должна быть достаточно большой.

Если же такая выборка маловероятна, то гипотеза отвергается.

Выборочное среднее: $\overline{x} = 172.4$

Вопрос: отличие $\overline{x}$ от 167 может быть отнесено к чистой случайности, или нет.
Если да — гипотеза верна.

Будем исходить из предположения, что гипотеза верна.

Выберем критерий того, что такая выборка возможна.
Основой его должна быть случайная величина $\overline{x}$.

$\overline{x}$ — случайная величина, распределенная также по нормальному закону с $\langle \overline{x} \rangle = \langle x \rangle$ и $\sigma_{\overline{x}}^2 = \sigma_x^2 / N$:

$$\overline{x} \sim \mathbf{N}\left(\langle x \rangle^{(h)},\ \sigma_x^2 / N \right)$$

Надо сделать линейное преобразование $\overline{x}$,
такое, чтобы получилась случайная величина,

распределенная по закону $z \sim \mathbf{N}(0,1)$

$$z = \frac{\overline{x} - \langle x \rangle^{(h)}}{\sigma_{\overline{x}}} = \frac{\overline{x} - \langle x \rangle^{(h)}}{\sigma_x / \sqrt{N}}$$

Но дисперсия σ_x^2 неизвестна,
известна лишь выборочная дисперсия:

$$s_x^2 = \frac{1}{N-1} \sum_{n=1}^{N} \left(x_n - \overline{x} \right)^2$$

Тогда вместо z берем следующую величину:

$$t = \frac{\bar{x} - \langle x \rangle^{(h)}}{s_x / \sqrt{N}} = \frac{\bar{x} - \langle x \rangle^{(h)}}{\sigma_x / \sqrt{N}} \cdot \frac{\sigma_x}{s_x} = \underbrace{\frac{\bar{x} - \langle x \rangle^{(h)}}{\sigma_x / \sqrt{N}}}_{z} \cdot \sqrt{\frac{\sigma_x^2}{s_x^2}} = \frac{z}{\sqrt{U}}$$

где $U = \frac{s_x^2}{\sigma_x^2} = \frac{1}{N-1} \sum_{n=1}^{N} \frac{\left(x_n - \bar{x}\right)^2}{\sigma_x^2} = \frac{1}{N-1} \chi^2_{N-1}$

Найдем распределение $w_U\left(U\right)$

Известно распределение $w_X\left(X\right)$

Преобразование: $Y=\frac{1}{A}X$

$$w_Y\left(Y\right)dY=w_X\left(X\right)dX=$$

$$=w_X\left(AY\right)d\left(AY\right)=Aw_X\left(AY\right)dY$$

$$w_Y\left(Y\right)=Aw_X\left(AY\right)$$

$$w_Y\left(Y\right) = Aw_X\left(AY\right)$$

Здесь $X = \chi^2_{N-1}, \quad Y = U, \quad A = N-1, \quad \chi^2_{N-1} = \left(N-1\right)U$

$$w\left(\chi^2_{N-1}\right) = \frac{1}{2^{\frac{N-1}{2}}\Gamma\left(\frac{N-1}{2}\right)}\left(\chi^2_{N-1}\right)^{\frac{N-1}{2}-1} e^{-\chi^2_{N-1}/2}$$

$$w_U\left(U\right) = \left(N-1\right)w\left(\left(N-1\right)\chi^2_{N-1}\right) = \frac{\left(N-1\right)^{\frac{N-1}{2}}}{2^{\frac{N-1}{2}}\Gamma\left(\frac{N-1}{2}\right)}U^{\frac{N-1}{2}-1} e^{-\left(N-1\right)U/2}$$

$$t = \frac{z}{\sqrt{U}}$$

$$w(z,U) = w_z(z)\,w_U(U)$$

$$w_z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

$$w_U(U) = \frac{(N-1)^{\frac{N-1}{2}}}{2^{\frac{N-1}{2}}\,\Gamma\left(\frac{N-1}{2}\right)} U^{\frac{N-1}{2}-1} e^{-(N-1)U/2}$$

Преобразование переменных: $(z,U) \Rightarrow \left(\frac{z}{\sqrt{U}}, U\right)$

то есть $(z,U) \Rightarrow (t,U)$

Распределение Стьюдента (N — число данных):

$$w(t)=\int_0^\infty w(t,U)\,dU=\frac{1}{\sqrt{\pi(N-1)}}\frac{\Gamma\left(\frac{N}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}\frac{1}{\left(1+\frac{t^2}{N-1}\right)^{\frac{N}{2}}}$$

Другой вид распределения Стьюдента:

$$w(t)=\frac{1}{\sqrt{\pi\nu}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}},\quad \text{где}\quad \nu=N-1$$

— число степеней свободы

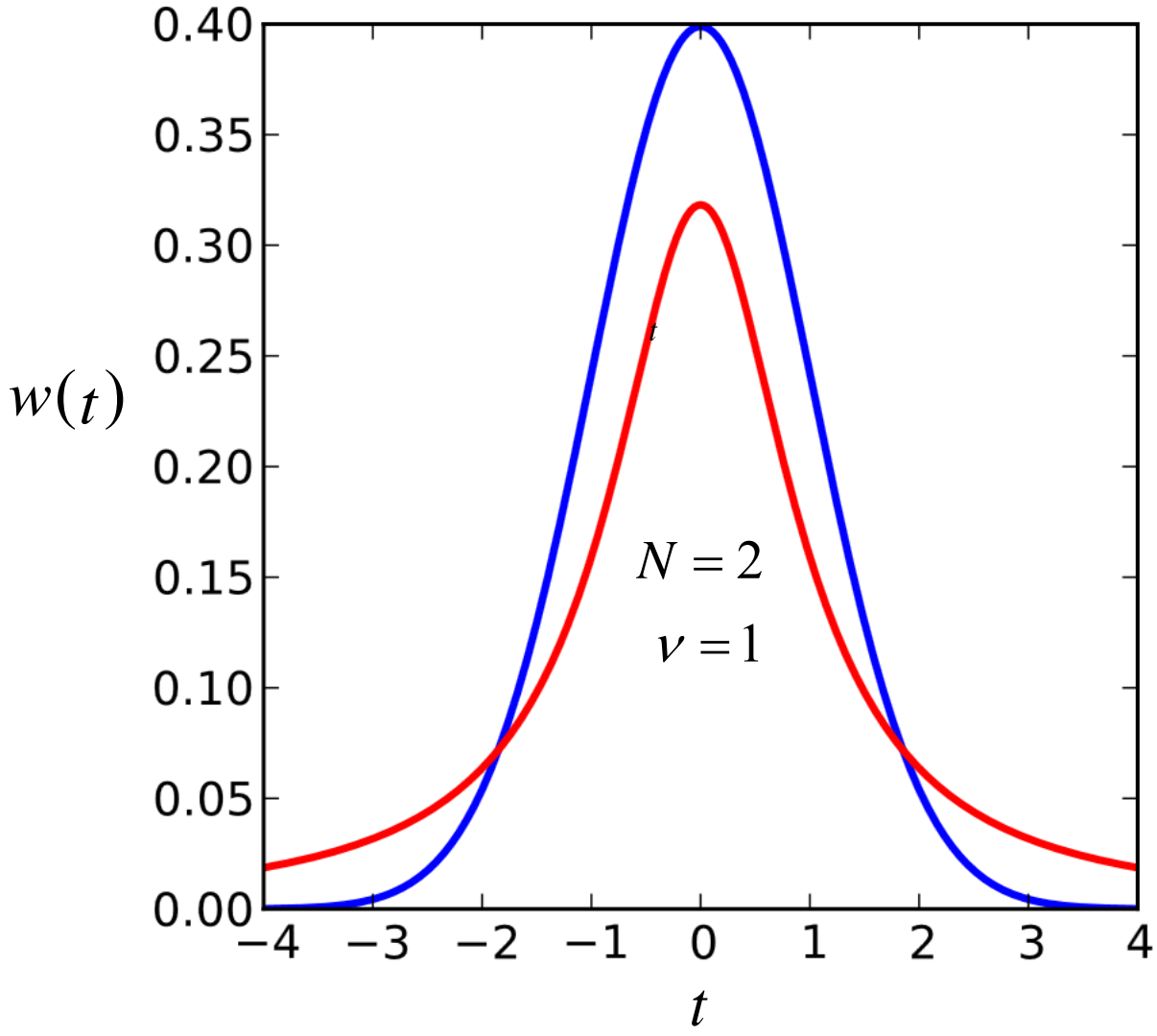
$$w(t) = \frac{1}{\sqrt{\pi(N-1)}} \frac{\Gamma\left(\frac{N}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)} \frac{1}{\left(1+\frac{t^2}{N-1}\right)^{\frac{N}{2}}}$$

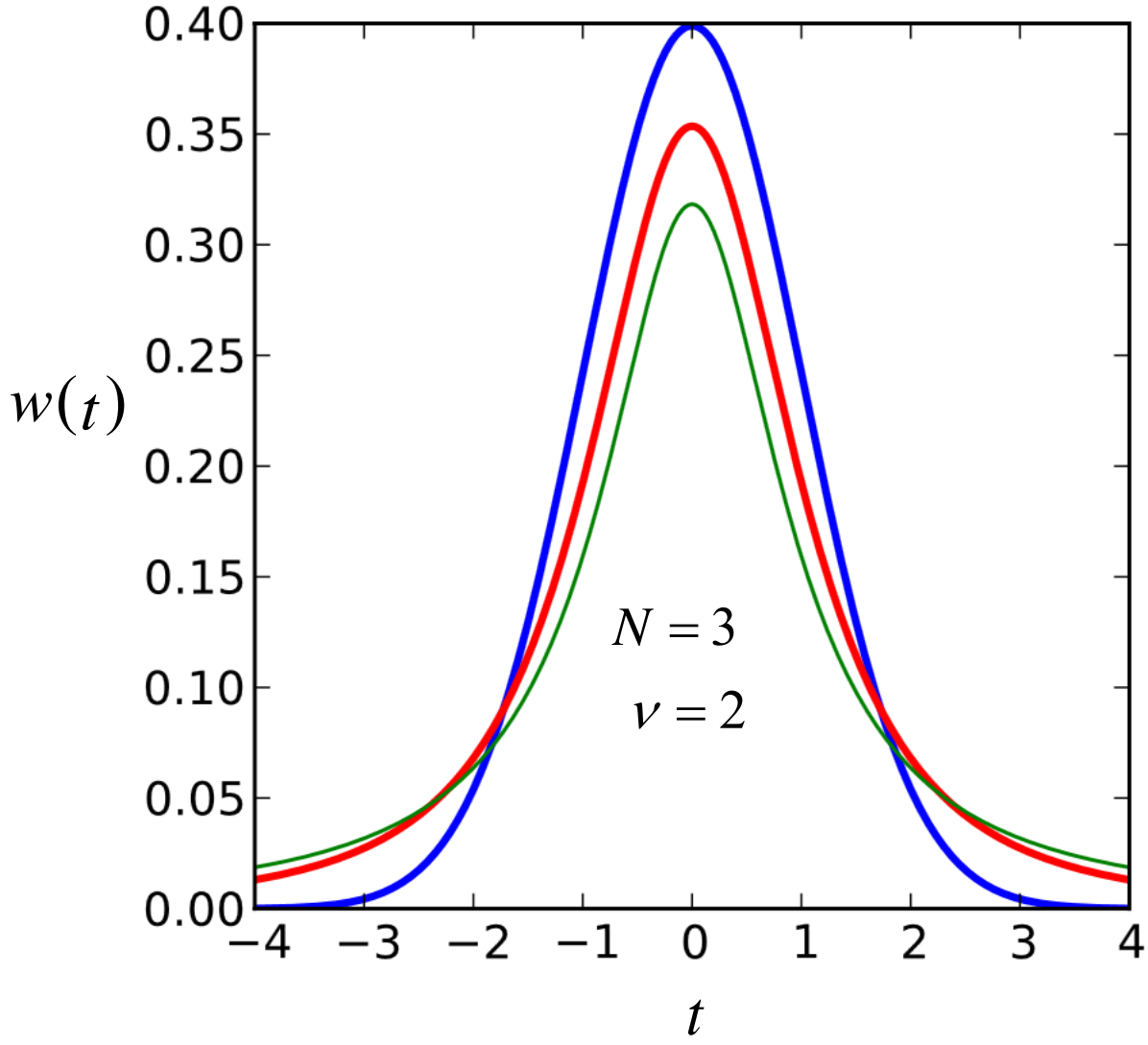
$$t = \frac{\overline{x} - \langle x \rangle^{(h)}}{s_x / \sqrt{N}} \quad , \qquad \text{где} \quad \overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad , \quad s_x^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(x_n - \overline{x}\right)^2$$

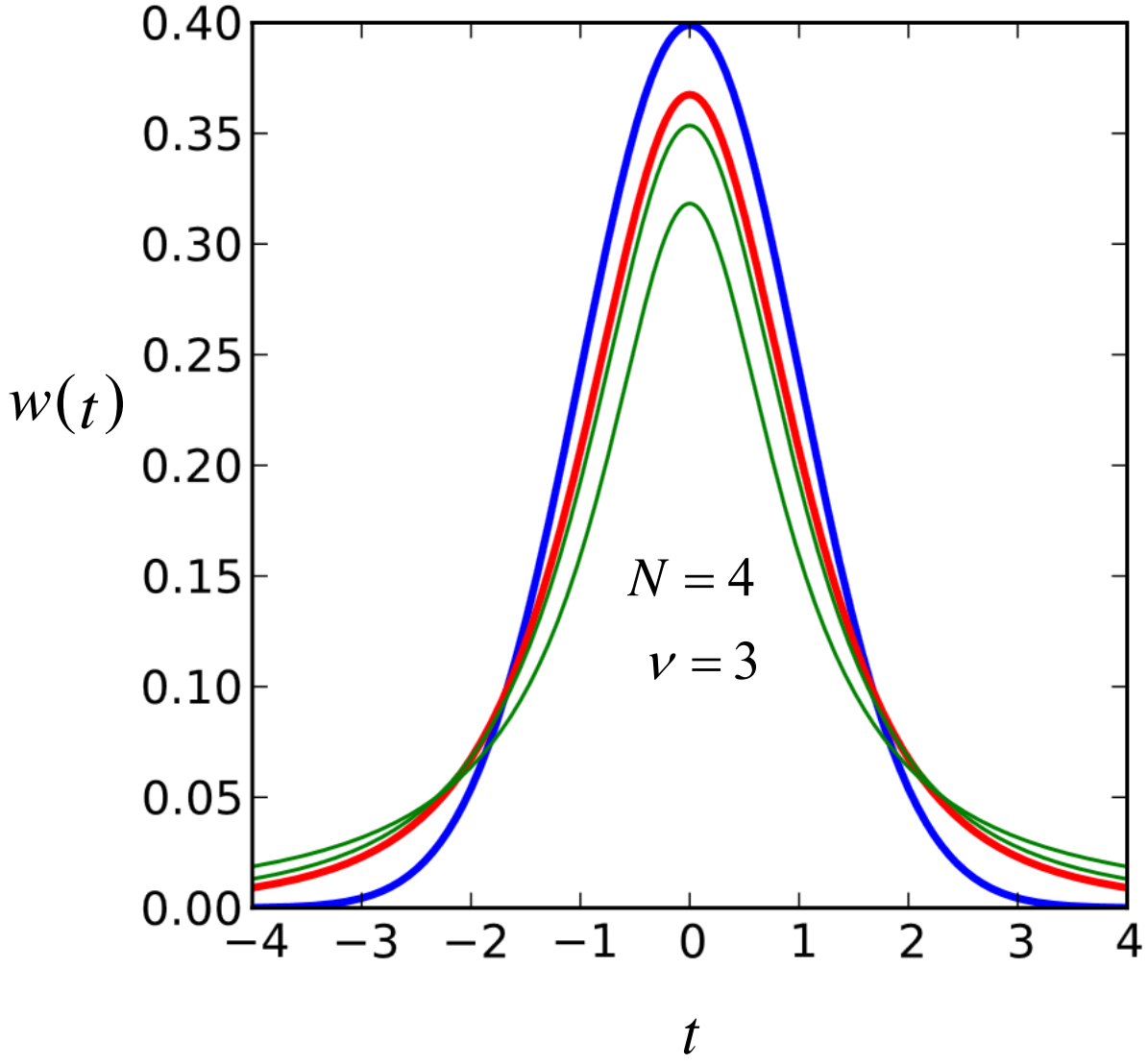
N — число данных,

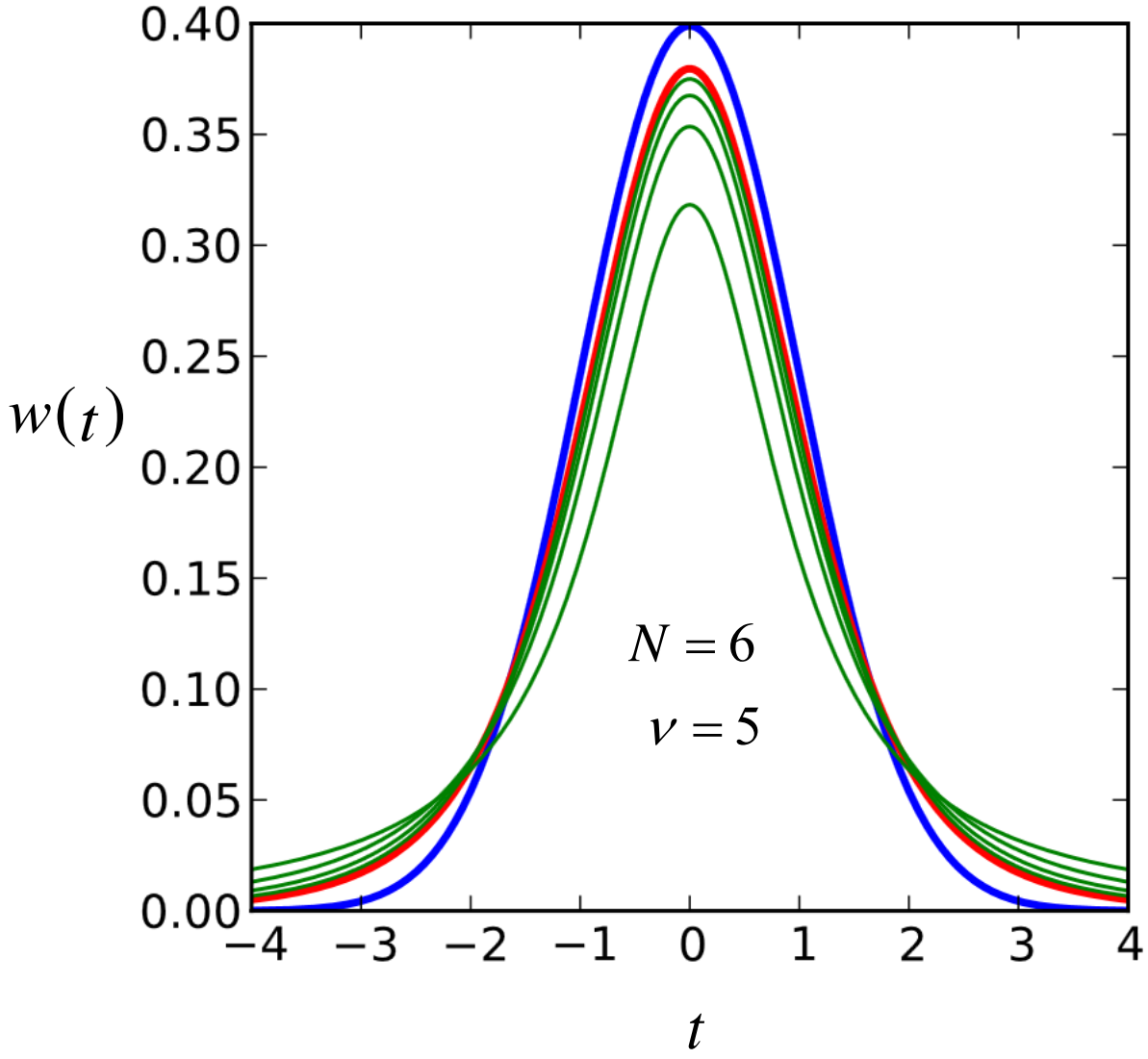
$\nu = N-1$ — число степеней свободы

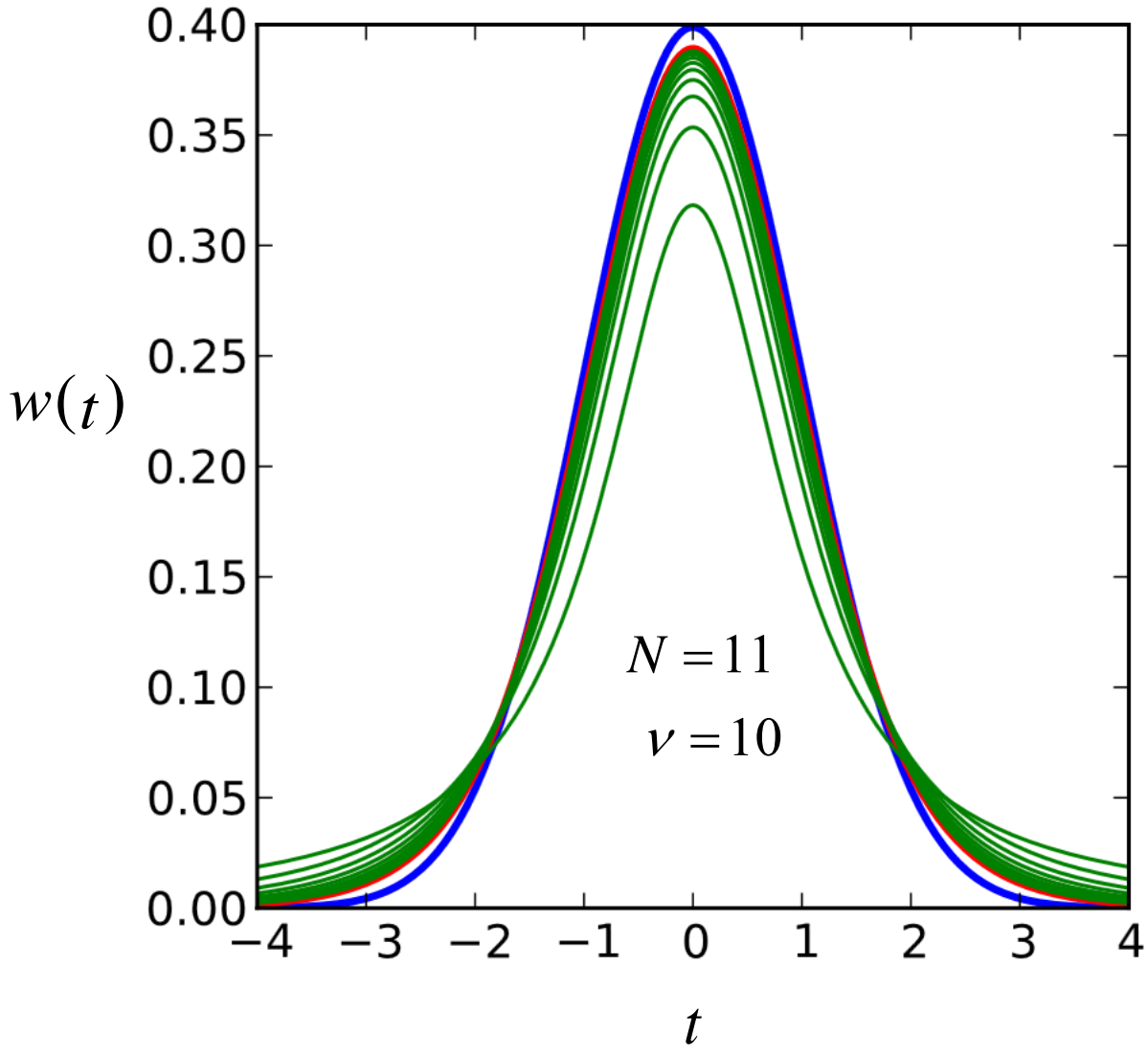
Стандартное нормальное распределение (синяя кривая)
и распределение Стьюдента с разным числом степеней свободы ν

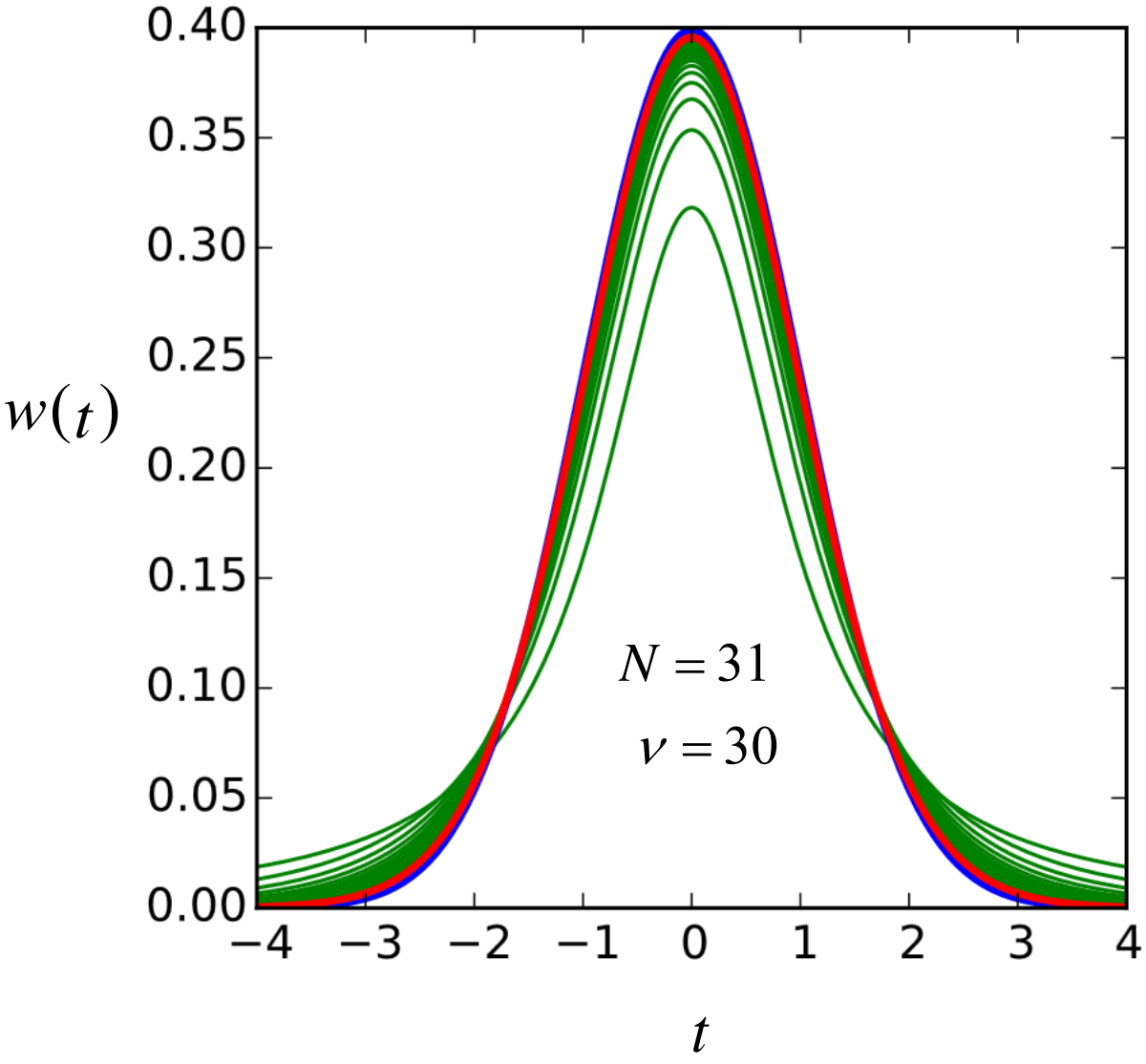



0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
w(t)
−4
−3
−2
−1
0
1
2
3
4
t
N = 3
ν = 2


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
w(t)
−4
−3
−2
−1
0
1
2
3
4
t
N = 4
ν = 3

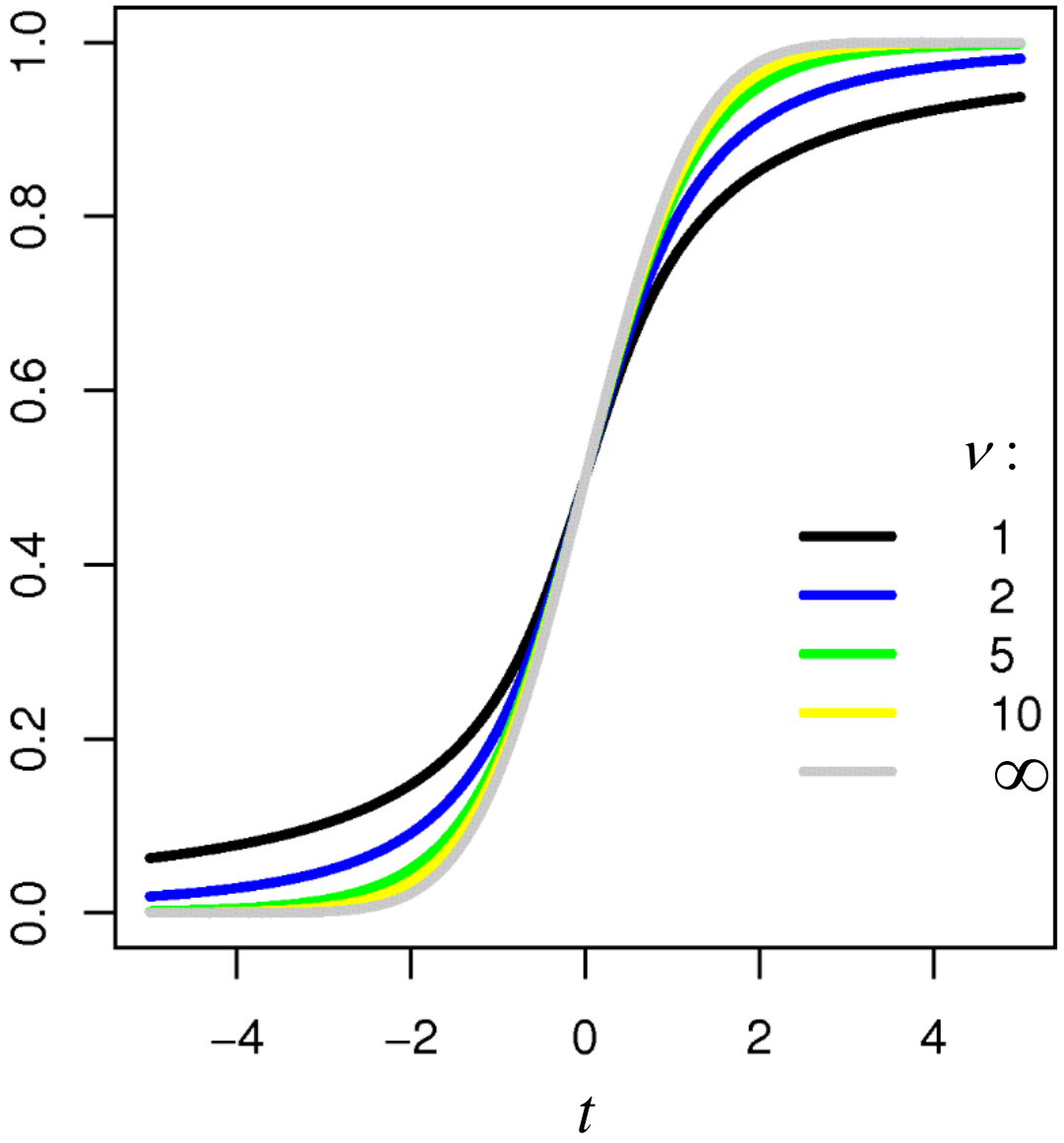

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
w(t)
−4
−3
−2
−1
0
1
2
3
4
t
N = 6
ν = 5


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
w(t)
−4
−3
−2
−1
0
1
2
3
4
t
N = 11
ν = 10


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
w(t)
−4
−3
−2
−1
0
1
2
3
4
t
N = 31
ν = 30

Функция распределения:

$$F_{-\infty}(t) = \int_{-\infty}^{t} w(\tau)\, d\tau$$



Взятый здесь пример

Выборка: x = 160, 160, 167, 170, 173, 176, 178, 178, 181, 181.
Объем выборки $N = 10,$ число степеней свободы $\nu = 9.$

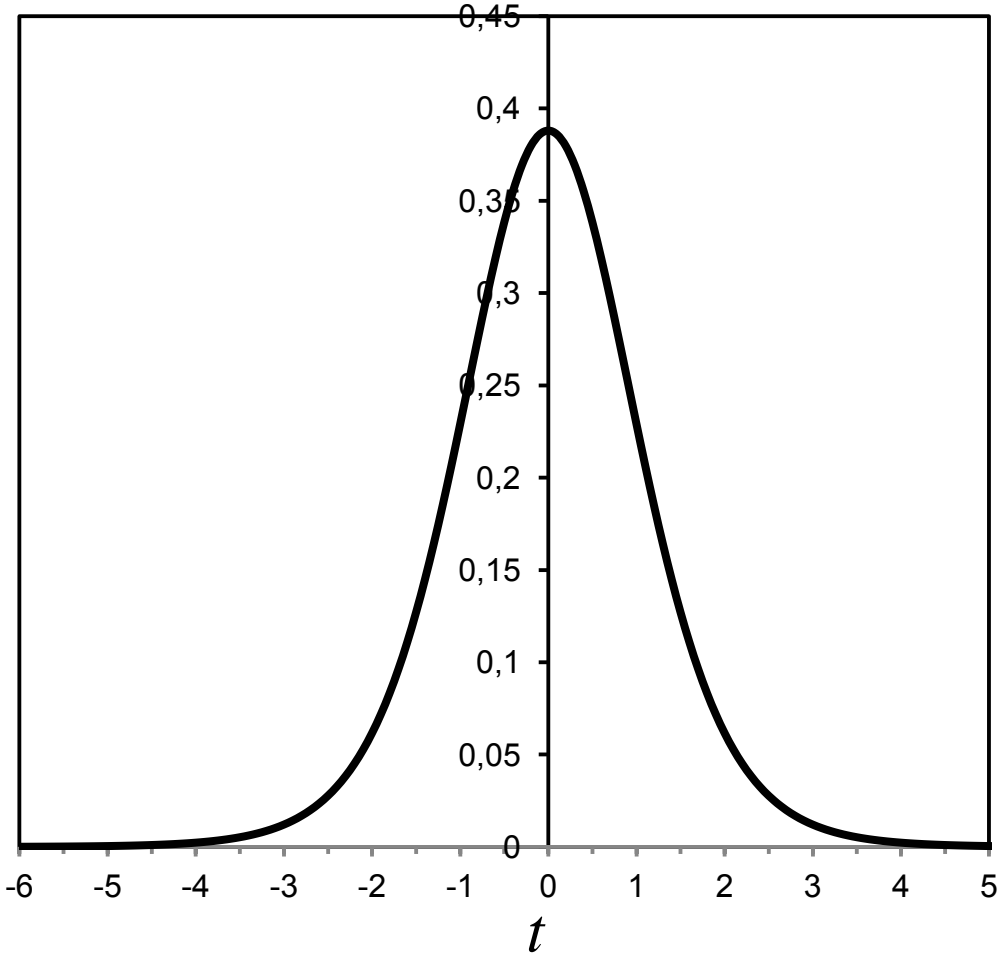
$$\bar{x} = 172.4, \qquad s_x^2 = 62.93, \qquad s_x = 7.93$$

$$\langle x \rangle^{(h)} = 167$$

$$t_9 = \frac{\bar{x} - \langle x \rangle^{(h)}}{s_x / \sqrt{N}} = \frac{172.4 - 167}{7.93 / \sqrt{10}} = 2.153$$

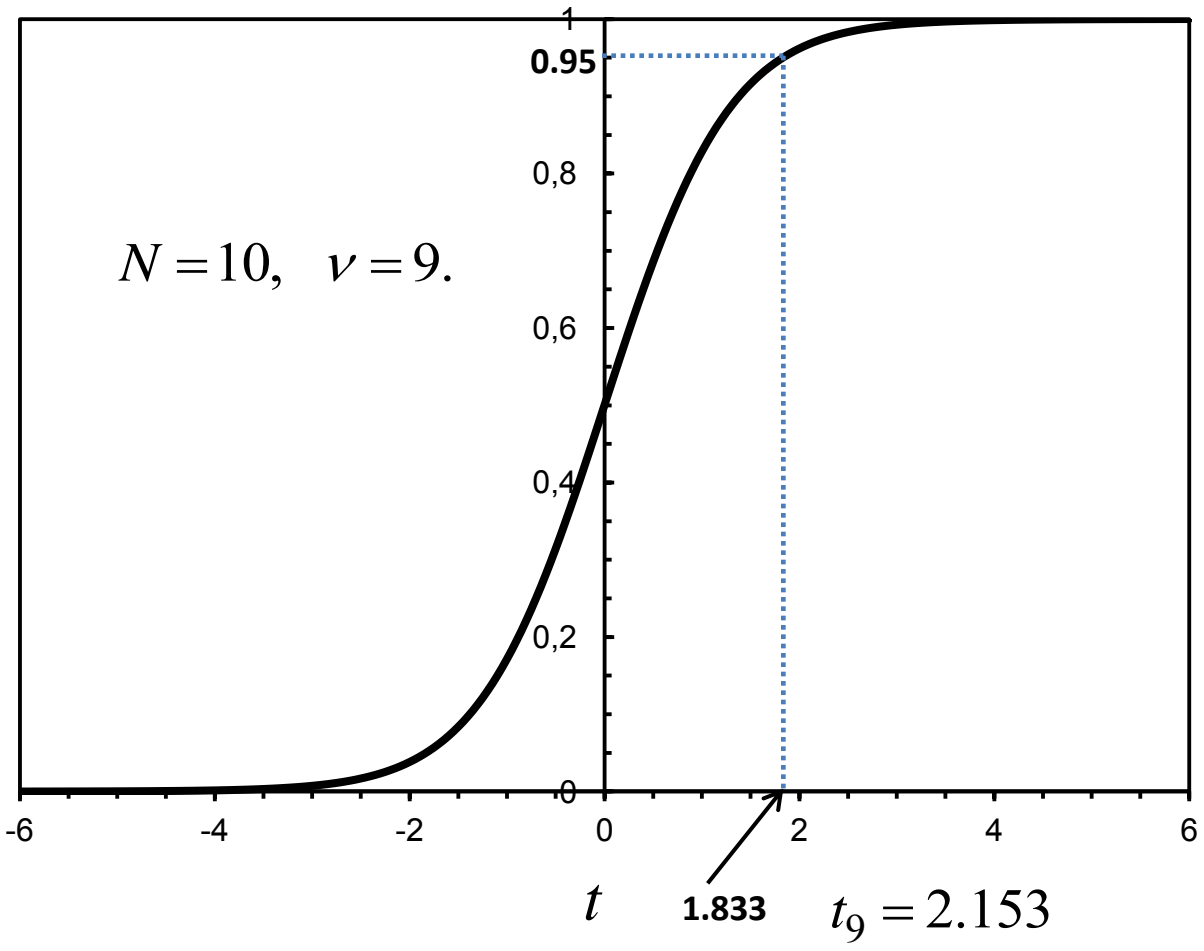
Для взятого здесь примера $N = 10, \quad \nu = 9.$

Плотность распределения



Функция распределения

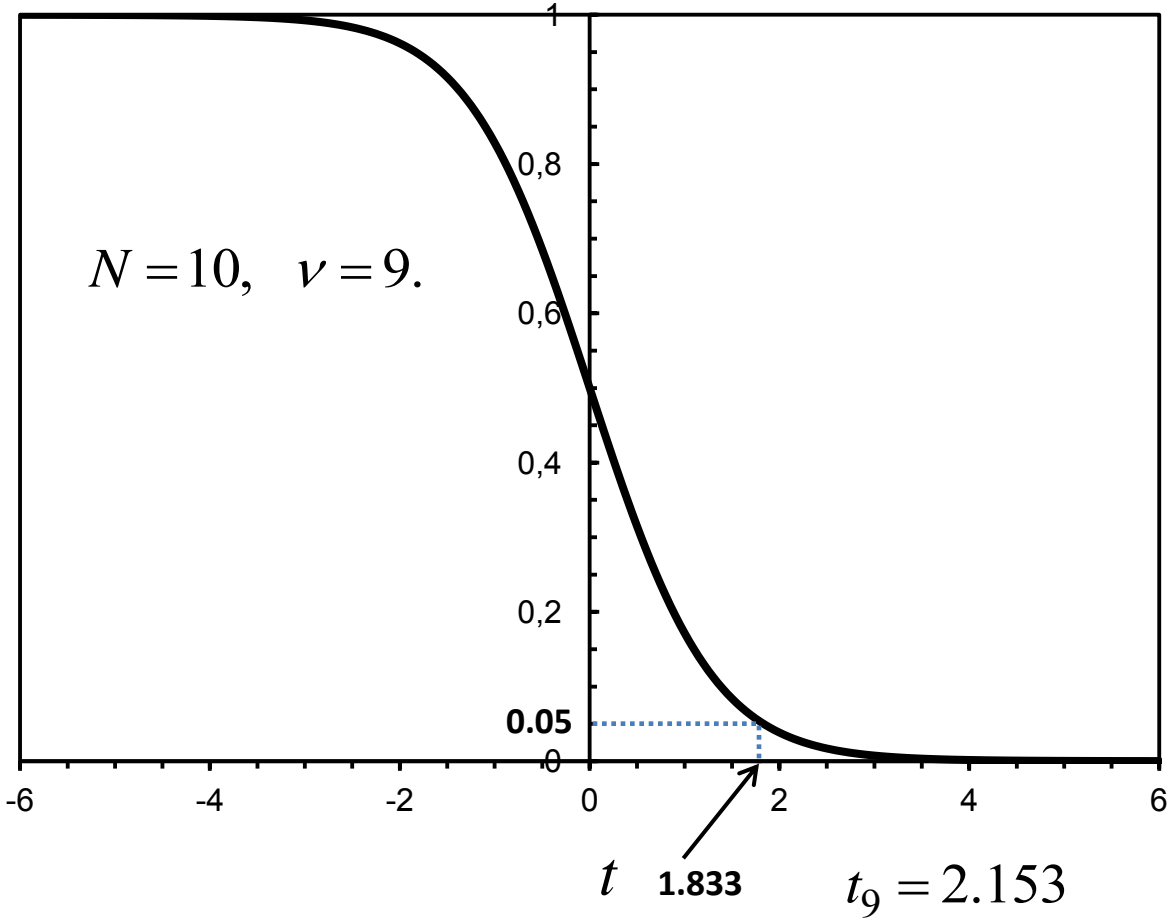
$$F_{-\infty} = \int_{-\infty}^{t} w(\tau)\, d\tau$$



Отклонение $\overline{x}$ от $\langle x \rangle^{(h)}$
статистически значимо на 5% уровне значимости

Функция распределения

$$F^{+\infty} = \int_{t}^{+\infty} w(\tau)\, d\tau$$



Отклонение $\overline{x}$ от $\langle x \rangle^{(h)}$
статистически значимо на 5% уровне значимости

Более распространенный случай – проверка гипотезы о том, что две совокупности данных являются выборками двух случайных величин, распределенных нормально с одинаковым математическим ожиданием и одинаковой дисперсией (или реализациями одной и той же случайной величины, распределенной нормально).

Пример:

x_1	15.7	10.3	12.6	14.5	12.6	13.8	11.9
x_2	12.3	13.7	10.4	11.4	14.9	12.6	

Два набора данных: x_{1k} и x_{2l}, где $k = 1,2,...,K,\ l = 1,2,...,L.$

Выборочные средние: $\overline{x}_1$ и $\overline{x}_2$.

Гипотеза: $\langle x_1 \rangle = \langle x_2 \rangle = m^{(1)}, \quad \sigma_{x_1}^2 = \sigma_{x_2}^2 = \sigma^2,$

$$x_1 \sim \mathbf{N}\left(m^{(1)}, \sigma^2\right), \quad x_2 \sim \mathbf{N}\left(m^{(1)}, \sigma^2\right)$$

Рассмотрим $z = \dfrac{\overline{x}_1 - \overline{x}_2}{\sigma_{\overline{x}_1 - \overline{x}_2}}$. Имеет место: $z \sim \mathbf{N}(0,1)$

Если $x \sim \mathbf{N}\left(m^{(1)}, \sigma^2\right)$, то $M_x^{(I)} = e^{\left(\frac{\sigma^2 u^2}{2} + u m^{(1)}\right)}$

Если $x = \sum_{n=1}^{N} x_n$, то $M_{\Sigma x}^{(I)} = e^{\left(\frac{\sigma^2 u^2}{2} + u m^{(1)}\right) N} = e^{\left(\frac{N\sigma^2 u^2}{2} + u N m^{(1)}\right)}$

Следовательно, $\Sigma x \sim \mathbf{N}\left(N m^{(1)}, N\sigma^2\right)$

$$\sigma^2_{\bar{x}_1-\bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \frac{\sigma^2_{x_1}}{K} + \frac{\sigma^2_{x_2}}{L} = \sigma^2\left(\frac{1}{K}+\frac{1}{L}\right)$$

Оценка σ^2 (дисперсия и x_1, и x_2): $$s^2 = \frac{\sum_{k=1}^{K}\left(x_{1k}-\bar{x}_1\right)^2 + \sum_{l=1}^{L}\left(x_{2l}-\bar{x}_2\right)^2}{K+L-2}$$

Оценка дисперсии $\bar{x}_1$:

$$s^2_{\bar{x}_1} = \frac{1}{K}s^2$$

Оценка дисперсии $\bar{x}_2$:

$$s^2_{\bar{x}_2} = \frac{1}{L}s^2$$

$$s^2_{\bar{x}_1-\bar{x}_2} = s^2_{\bar{x}_1} + s^2_{\bar{x}_2} = s^2\left(\frac{1}{K}+\frac{1}{L}\right)$$

Тогда $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}}$ заменяется на $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$,

где $s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2_{\bar{x}_1 - \bar{x}_2}} = \sqrt{s^2 \left(\dfrac{1}{K} + \dfrac{1}{L} \right)}$

$$s^2 = \frac{\sum_{k=1}^{K} \left(x_{1k} - \bar{x}_1 \right)^2 + \sum_{l=1}^{L} \left(x_{2l} - \bar{x}_2 \right)^2}{K + L - 2}$$

Пример:

x_1	15.7	10.3	12.6	14.5	12.6	13.8	11.9
x_2	12.3	13.7	10.4	11.4	14.9	12.6	

$\bar{x}_1 = 13.06,\ \ \bar{x}_2 = 12.55,\ \ \bar{x}_1 - \bar{x}_2 = 0.51.$ Случайно ли это различие?

$$s^2 = \frac{18.98 + 12.86}{6 + 7 - 2} = 2.89, \quad s^2_{\bar{x}_1 - \bar{x}_2} = 2.89\left(\frac{1}{6} + \frac{1}{7}\right) = 0.895, \quad s_{\bar{x}_1 - \bar{x}_2} = 0.95$$

Отсюда $t_{11} = \dfrac{0.51}{0.95} = 0.54$

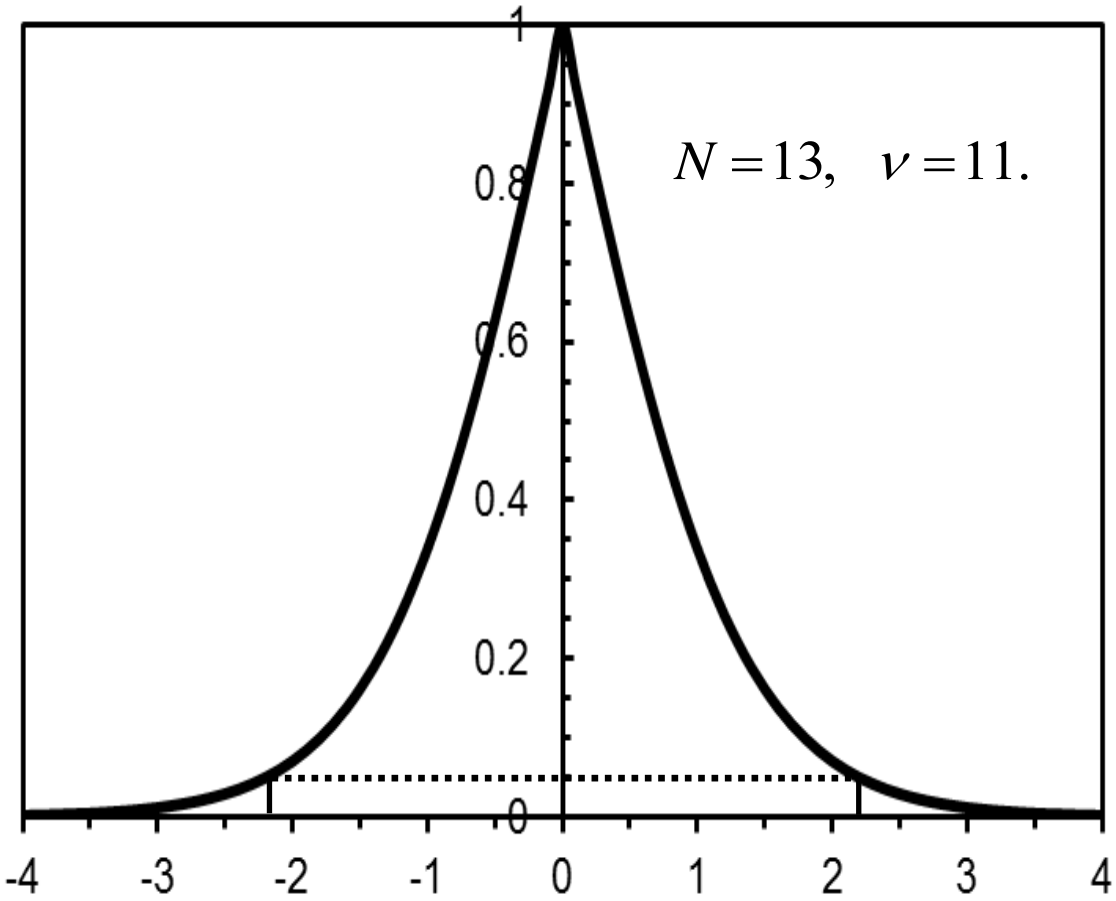
Двустороннее распределение Стьюдента

$$F_2 = \int_{-\infty}^{-t} w(\tau)\,d\tau + \int_{t}^{+\infty} w(\tau)\,d\tau$$

$-2.2 < t_{11} < 2.2$

здесь $t_{11} = 0.54$

Различие $\overline{x}_1$ и $\overline{x}_2$ незначимо.



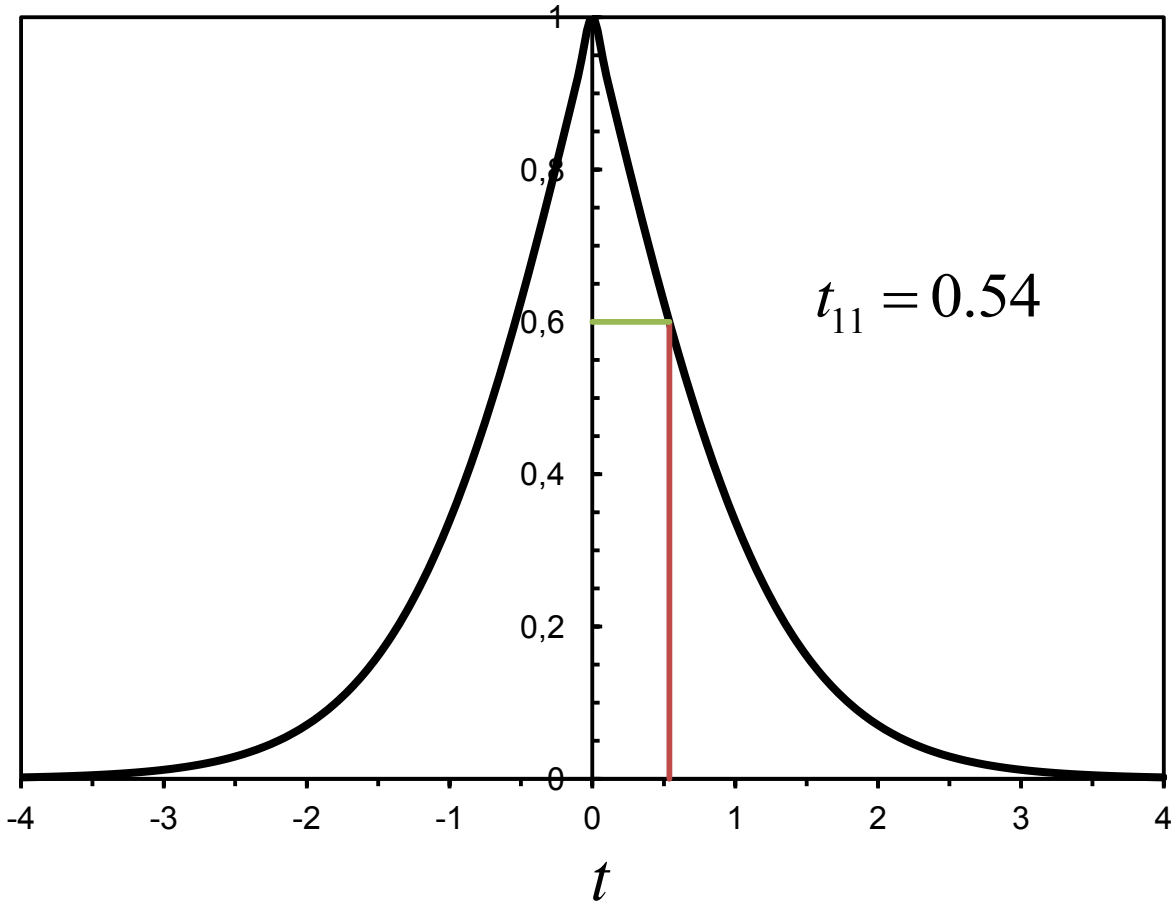
Поскольку $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$, то $\bar{x}_1 - \bar{x}_2 = t \cdot s_{\bar{x}_1 - \bar{x}_2}$

$$t^{(1)} s_{\bar{x}_1 - \bar{x}_2} < \bar{x}_1 - \bar{x}_2 < t^{(2)} s_{\bar{x}_1 - \bar{x}_2}$$

Здесь $-2.09 < \bar{x}_1 - \bar{x}_2 < 2.09,$

$$\bar{x}_1 - \bar{x}_2 = 0.51$$

Другой способ проверки значимости различия — t-тест, то есть, вероятность того, что t будет иметь рассчитанное по выборкам значение. Если эта вероятность меньше 0.05, то гипотеза отвергается.



Доверительный интервал

Новая гипотеза: разность между $\overline{x}_1$ и $\overline{x}_2$ равна δ .
Тогда в качестве критерия возьмем

$$t_{K+L-2} = \frac{\overline{x}_1 - \overline{x}_2 - \delta}{s_{\overline{x}_1 - \overline{x}_2}}$$

С вероятностью 0.95 $$t^{(1)} < \frac{\overline{x}_1 - \overline{x}_2 - \delta}{s_{\overline{x}_1 - \overline{x}_2}} < t^{(2)}$$

$$\left(\overline{x}_1 - \overline{x}_2\right) - t^{(2)} s_{\overline{x}_1 - \overline{x}_2} < \delta < \left(\overline{x}_1 - \overline{x}_2\right) - t^{(1)} s_{\overline{x}_1 - \overline{x}_2}$$

$$\left(\bar{x}_1 - \bar{x}_2\right) - t^{(2)} s_{\bar{x}_1 - \bar{x}_2} < \delta < \left(\bar{x}_1 - \bar{x}_2\right) - t^{(1)} s_{\bar{x}_1 - \bar{x}_2}$$

Здесь $t_{11}{}^{(1)} = -2.2, \quad t_{11}{}^{(2)} = 2.2, \quad \bar{x}_1 - \bar{x}_2 = 0.51$

$$-1.57 < \delta < 2.59$$

Это – 95% доверительный интервал.