Лекция №4 Непрерывная случайная величина

Цель лекции:

- ✓ познакомиться с основными понятиями функции распределения вероятностей и плотности распределения вероятностей
- ✓ изучить нормальное распределение
- ✓ научиться работать с таблицей z значений
- ✓ изучить правило трех сигм
- ✓ изучить центральную предельную теорему
- ✓ познакомиться с равномерным распределением

Материал прошлого урока:

На прошлом занятии мы познакомились с разведочным анализом, обсудили важность его роли в статистическом анализе. Закончили прошлый урок построением самых простых и в то же время максимально информативных графиков, боксплот и гистограмма для случайной величины «рост». Эта случайная величина следует нормальному распределению, о котором сегодня пойдет речь.

План урока:

1. **Определения**
2. **Функция плотности распределения вероятностей**
3. **Свойства нормального распределения**
4. **Правило трех сигм**
5. **Стандартное нормальное распределение**
6. **Нормирование**
7. **Центральная предельная теорема**
8. **Равномерное распределение**

Как всегда начнем с некоторых определений.

Определения

Непрерывная случайная величина – это величина, которая может принимать все возможные значения, содержащиеся на промежутке, причем этот промежуток может быть, как конечным, так и бесконечным.

Мы помним, что случайные величины (СВ) принято делить на 2 большие категории. Дискретные и непрерывные. Напомню примеры дискретных СВ : количество учащихся по школам, количество бракованных изделий в партии, количество интернет-заказов по дням.

Непрерывные случайные величины – это СВ, которые мы можем измерить с той точностью, что позволяет нам прибор, или с той точностью, что нам нужна.

Обе эти категории СВ следуют своим законам распределения. Для дискретной СВ мы изучили основные распределения, такие как биномиальное распределение и распределение

Пуассона, которое является частным случаем биномиального распределения. А сегодня речь пойдет о распределениях, свойственных непрерывной случайной величине. И одним из самых важных распределений статистики является нормальное распределение или распределение Гаусса.

Но прежде познакомлю вас еще с 2мя понятиями, с которыми мы будем работать.

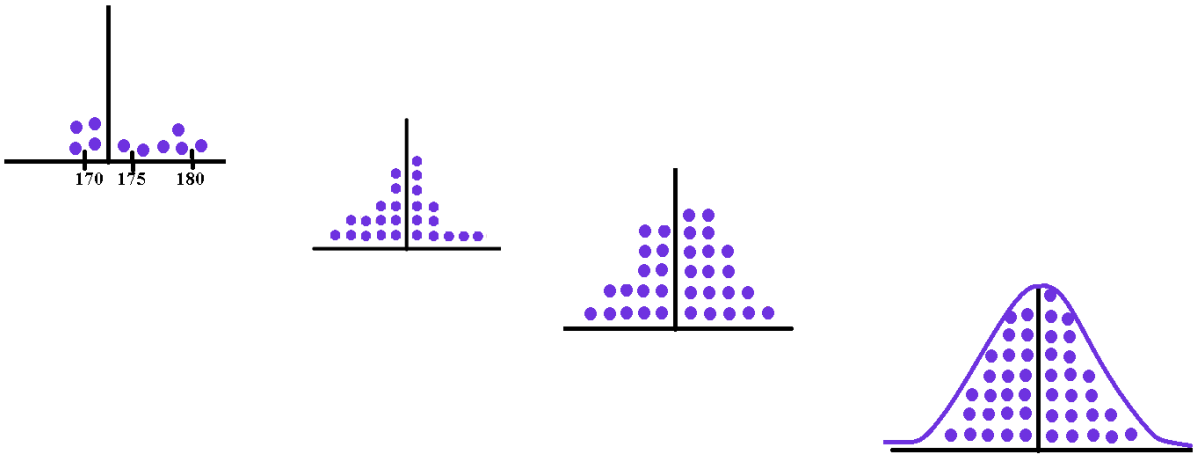
Функция распределения вероятностей - это такая функция $F(x)$, которая для каждого значения x показывает, какова вероятность того, что случайная величина меньше или равна x.

Плотность распределения вероятностей - это функция $f(x)$, которая равна производной функции распределения вероятностей.

По сути 2 этих понятия синонимизируют в статистике.

Функция плотности распределения вероятностей

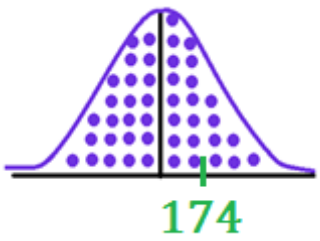
Давайте разбираться с этими понятиями. Взгляните на картинку ниже:



Предположим перед нами стояла задача измерить рост людей из некоторого сообщества. В первый день мы взяли выборку из 10 человек и измерили их рост. Данные стали размещать на гистограмме. И так мы делали некоторый период времени. Мы копили данные, размещали их на гистограмме, и они, в конце концов, образовали колокол. С помощью колоколообразной кривой (Г) мы можем описать распределение наших данных. Эта кривая задается функцией плотности распределения $f(x)$.

Вся площадь под кривой равна 1, т.е. под этой кривой лежат 100% значений случайной величины рост. Если мы возьмем интеграл от $f(x)$ на участке ab, мы найдем площадь под кривой, которая будет показывать, какая доля значений случайной величины лежит в этом промежутке или, иными словами, какая вероятность, что СВ «рост» попала в отрезок ab.

Например, ниже 174 лежит 38 значений (кружки на рисунке ниже) из 45 или это, приблизительно, 84,4%. С помощью функции плотности распределения вероятности мы можем находить эту вероятность (долю).

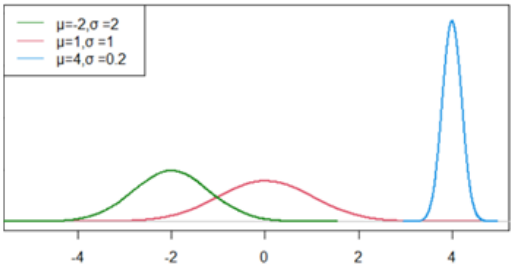


Функция плотности распределения вероятностей для нормального распределения выглядит так

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}},$$

Нормальное распределение или его еще называют распределение Гаусса - это распределение вероятностей непрерывной случайной величины X, плотность вероятности которой подчиняется этой формуле.

Здесь а – это математическое ожидание, σ^2 — дисперсия. Т.о. зная, математическое ожидание и дисперсию (а из нее можем получить среднее квадратичное отклонение), можем задать нормальное распределение, как показано на рисунке ниже.



Примерами нормального распределения случайной величины служат такие случайные величины, которые обычно описывают биологические физические, химические процессы:

- ✓ рост
- ✓ вес людей
- ✓ показатели IQ
- ✓ скорость движения молекул в газах и жидкостях и т.д.

Свойства нормального распределения

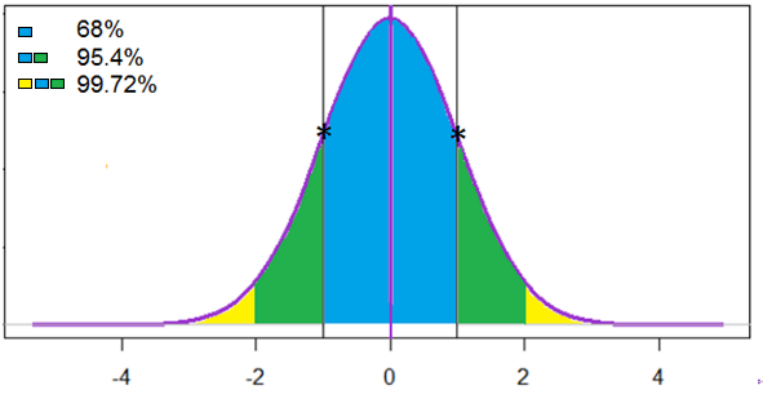
Свойствами нормального распределения являются:

- ✓ Колоколообразная форма графика
- ✓ График симметричен оси y
- ✓ Площадь под дугой равна единице
- ✓ График показывает долю (вероятность) СВ меньше x
- ✓ Значения среднего, медианы и моды совпадают

Правило трех сигм

Считается, что нормальное распределение подчиняется правилу трех сигм, которое в свою очередь гласит следующее: на отрезке

- ✓ от −σ до +σ расположено около 68% наблюдений
- ✓ от −2σ до +2σ - 95.4%
- ✓ от −3σ до +3σ - 99.72 %



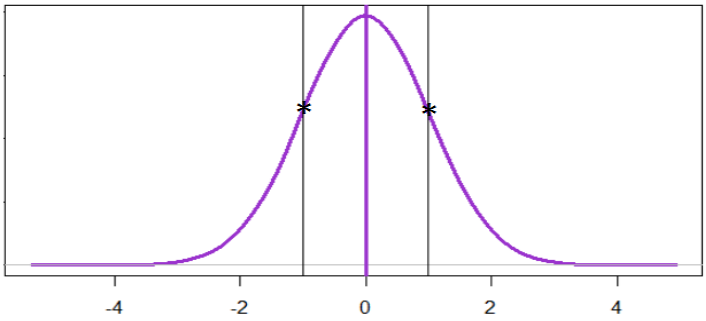
Например, используя правило трех сигм, мы можем найти долю значений, которые лежат выше 2 сигм (площадь правого желтого треугольника).

$(1 - 0.954)/2 = 0.023$ или ,приблизительно, 2.3%

Но в реальной жизни не всегда нам нужно найти долю значений выше целого числа сигм. Например, а какая доля лежит ниже -1.7. Брать интеграл от такой громоздкой функции, как плотность распределения вероятностей очень неудобно. И здесь на помощь приходит стандартное нормальное распределение.

Стандартное нормальное распределение

Стандартное нормальное распределение - это нормальное распределение со средним арифметическим = 0, стандартным отклонением =1



Например, чтобы найти долю ниже любого значения по оси x, мы можем воспользоваться таблицей z-значений или ее еще называют таблицей накопленного нормального распределения.

Фрагмент из этой таблице приведен ниже.

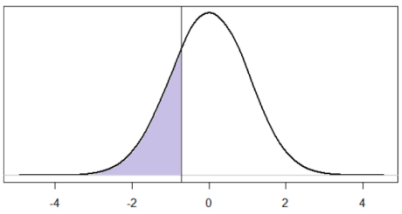
Таблица накопленного нормального распределения
N (x) при x < 0

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823

Именно в этой таблице z значение обозначено x. Пусть это вас не смущает. Эта таблица для нормального стандартного распределения. Т.е. мы можем искать по ней долю значений ниже любого заданного значения СВ, которая следует стандартному нормальному распределению. В этой таблице x (значение, ниже которого будем искать долю значений СВ) как бы состоит из 2х частей. Первая часть по вертикали (до первого знака после запятой), а вторая часть (сотые) по горизонтали. В некоторых таблицах до тысячных.

А значения до десятитысячных, которые лежат внутри таблицы показывают, какая доля значений лежит ниже заданного x. Т.е. на пересечении составных частей x находится значения, которое показывают ту долю, что ниже x.

Например, найдем по этой таблице вероятность, что случайная величина, которая следует нормальному стандартному распределению, меньше или равна -0.72. Или иными словами, долю значений, которая лежит не выше - 0.72 с помощью этой таблицы.



x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,468
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,428
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,385
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,352
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,315
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,281
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,248
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,217
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,189
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,163
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,140
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,119
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,100
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,083

Ниже -0.72 лежит 0.2358 или 23.58% всех значений нашей СВ. Или мы можем сказать, что вероятность того, что наша СВ меньше или равна -0.72 будет, приблизительно, 23.58%.

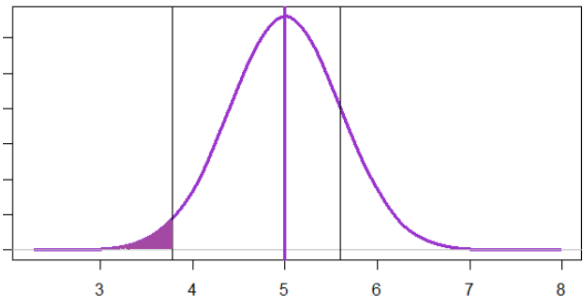
Ну а чтобы найти, а какая доля лежит выше, чем -0.72 просто нужно 1 – 0.2358 = 0.7642

Но в реальной жизни со стандартным нормальным распределением нам приходится сталкиваться нечасто. Но пропорции (доли значений ниже некоторого количества сигм) соблюдены во всех нормальных распределениях. Так же как и правило 3 сигм работает для любого нормального распределения с любым средним и любым стандартным отклонением. Поэтому таблица накопленных значений нормального распределения будет работать для любого нормального распределения, но для этого мы должны его привести к «стандартному виду»

Нормирование

Давайте сразу это сделаем на примере задачи.

Диаметр выпускаемых гаек следует нормальному распределению с mu = 5 мм, дисперсией 0.36 кв.мм. Найти пропорцию гаек с размером менее 3,78 мм.



Для этого распределения среднее равно 5, а стандартное отклонение = 0.6 (квадратный корень из дисперсии 0.36 кв.мм). Нам нужно поместить это распределение в масштабы

стандартного нормального распределения, где мю равно нулю, а среднее квадратичное отклонение (далее ско) равно 1.

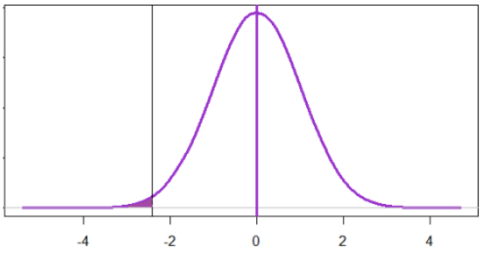
Сначала сместим среднее на ноль:

Если мы из каждого диаметра гаек вычтем 5, а потом посчитаем среднее арифметическое, то оно будет равно нулю.

А чтобы эта СВ также лежала в масштабах распределения, где ско равно единице, нам надо посмотреть, а в скольких сигмах лежит каждое значение от среднего арифметического. Например, у нас в задаче надо найти долю значений ниже 3,78. Взглянем, а как далеко 3.78 мм лежит от среднего 5 мм. Расстояние будем измерять в количестве сигм, т.к. все нормальные распределения лежат практически полностью в пределах $\pm 3\boldsymbol{\sigma}$, только сигма для каждого своя.

3.78 - 5 = -1.22. А теперь -1.22 /0.6 ≈ -2.03. Т.е. между 3.78 и 5 мм помещается около 2.03 ско. Знак минус показывает, что 3.78 лежит ниже среднего арифметического. Важно вычитать в правильном порядке. Сейчас поймете почему.

Итак, мы можем воспользоваться этой нормированной величиной -2.03 и по таблице ниже определим, что 0.0212 – такая доля гаек имеет диаметр меньше 3.78 мм.



z	0,00	0,01	0,02	0,03	0,04	0,05	0,
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,

Если бы мы потеряли знак минус, то мы бы искали долю значений, которая лежит ниже 2.03, а это уже будет очень большая величина. А предположим, что все, что меньше 3.78 мм –

это брак. И потеряв знак минус при нормировании, мы бы получили, что почти вся совокупность у нас – брак.

Подведем итог того, что мы сделали, как получили значение z (в этой таблице оно обозначено x)

$$Z = \frac{X-\mu}{\sigma}$$

Т.е. z показывает, а в скольких сигмах лежит то или иное значение СВ от среднего арифметического.

Это нас приводит к теореме

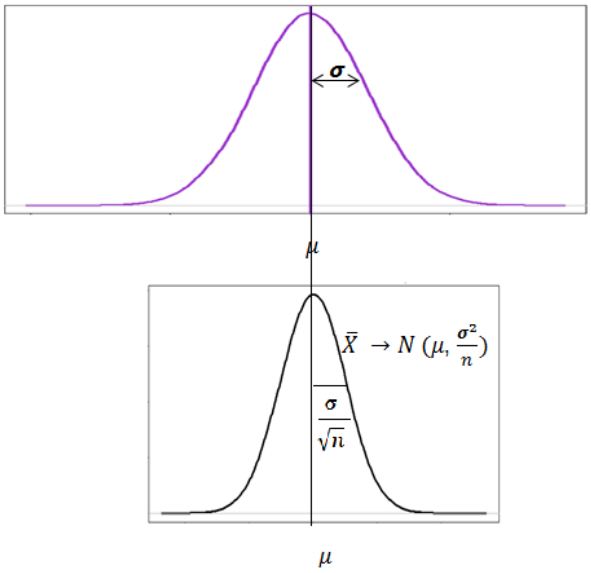
Если $X \sim N\ (\mu\ ,\ \sigma^2)$, ***тогда*** $Z = \frac{X-\mu}{\sigma}$ ***~ N (0,1)***

Центральная предельная теорема

Пусть генеральная совокупность имеет любое распределение с средним арифметическим μ и дисперсией σ^2, тогда среднее выборочное стремится к нормальному распределению с мю таким же, как у генеральной совокупности и дисперсией, равной дисперсии генеральной совокупности, разделенной на объем выборки.

$\underline{X}\ \rightarrow N\ (\mu, \frac{\sigma^2}{n})$

Давайте объясним на примере с гайками. Диаметр гаек – это некая случайная величина. А вот если мы будем брать выборки некоторого объема, предположим, объемом 30 и измерять среднее по ним, то у нас получится новая случайная величина- среднее выборочное. Т.е. мы взяли одну выборку объемом 30, померили по ней среднее, другую выборку объемом 30 и померили по ней среднее, третью выборку и померили по ней среднее и т.д. Эти средние арифметические и будут образовывать новую СВ, которая будет стремиться к нормальному распределению с средним , как у генеральной совокупности, из которой мы брали эти выборки и с дисперсией , равной дисперсию генеральной совокупности разделить на объем выборки. Кстати, корень из этой «новой» дисперсии $\frac{\sigma}{\sqrt{n}}$ называется стандартной ошибкой среднего.

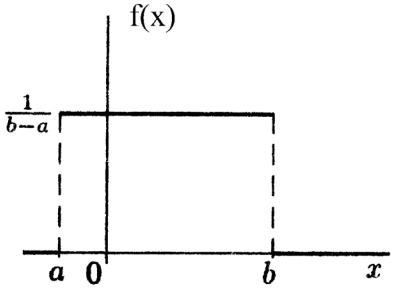


Равномерное распределение

И еще одно распределение, которое мы с вами рассмотрим для непрерывной СВ, это равномерное распределение. Одно из самых простых распределений статистики.

Непрерывная случайная величина равномерно распределена на отрезке ab, если плотность распределения вероятностей ее равна нулю за пределами отрезка и равна постоянной величине внутри него. Постоянная величина равна 1/(b-a).

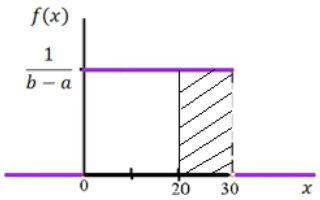
$$f(x) = \begin{cases} 0, \text{если } x \le a; \\ \frac{1}{b-a}, \text{если } a < x \le b; \\ 0, \text{ если } x > b. \end{cases}$$



Вероятность попадания равномерно распределенной СВ в интервал между ab равна площади под графиком функции. Т.е. равна 1.

Рассмотрим сразу пример решения задачи на равномерное распределение.

Посадка на самолет задерживается на 30 минут. Найти вероятность, что посадка начнется между 20 и 30 минутами, предполагая, что время посадки распределено равномерно.



Т.е. для ответа на вопрос задачи, нам нужно найти площадь заштрихованной фигуры.

Площадь прямоугольника равна произведению длины и ширины.

Одна сторона прямоугольника 10, а другая (по оси y) $\frac{1}{30-0} = \frac{1}{30}$

Вероятность, что посадка начнется между 20 и 30 минутами равна $10 * \frac{1}{30} = \frac{1}{3}$

Ну и как всегда, чтобы получить как можно больше информации о той величине, что мы изучаем, будем использовать описательную статистику.

$M(x) = \frac{a+b}{2}$

$$D = \frac{(b-a)^2}{12}$$

Решим задачу: Найти математическое ожидание и дисперсию равномерно распределенной величины на отрезке от 2 до 8. (Обратите внимание, что не сказано, входят эти значения или нет. Т.к. мы работаем с непрерывной СВ, то используем эти края отрезка в формулах для нахождения необходимых параметров).

$$a = 2$$

$$b = 8$$

$$M(x) = \frac{2+8}{2} = 5$$

$$D = \frac{(8-2)^2}{12} = 3$$

На этом уроке мы познакомились с распределениями, свойственными непрерывной случайной величине. С распределением Гаусса (нормальное распределение) и центральной предельной теоремой мы продолжим работать на следующем уроке, когда будем учиться проводить тестирование гипотез и строить доверительные интервалы.