Review – distribution of discrete random variable

Binomial distribution:

- n trials (n is fixed in advance)
- each trial: either success or failure
- Probability of success (p) remains the same throughout the trials
- All trials are independent

Binomial Random Variable:

X = the number of successes after n trials

X~Binom(n,p)   P(X=k) = ?

E(X) = np

Var(X) = np(1-p)

Review – distribution of discrete random variable

Geometric distribution:

- each trial: either success or failure
- Probability of success (p) remains the same throughout the trials
- All trials are independent

Geometric Random Variable:

X = the number of trials until the first success occurs

X~Geom(p)   P(X=k) = ?

E(X) = 1/p

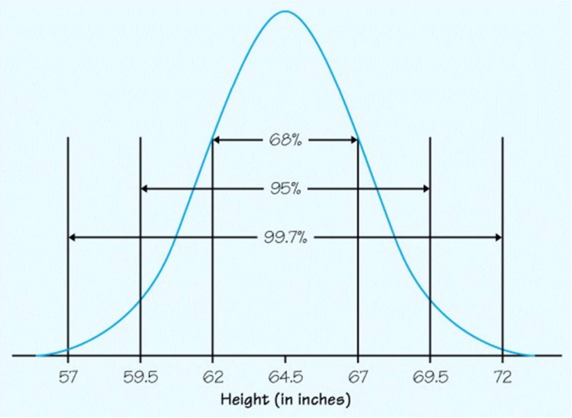
Var(X) = $\frac{1-p}{p^2}$

Distribution of

continuous

random variable

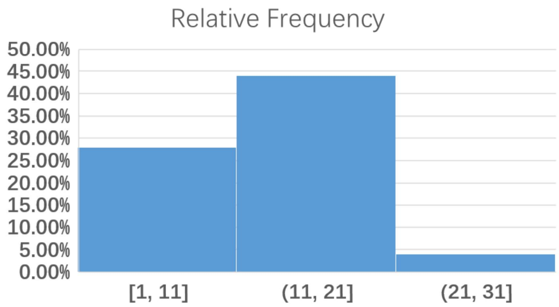
Density Curves

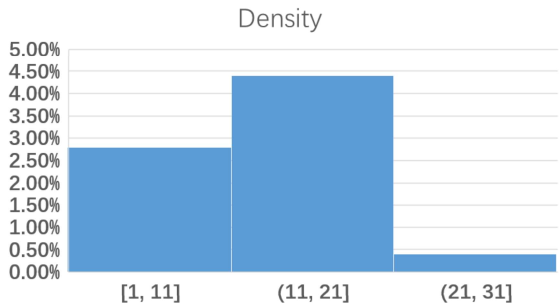


Density

$$\text{density} = \text{rectangle height} = \frac{\text{relative frequency of class interval}}{\text{class interval width}}$$

Relative Frequency = density * interval width





先回顾一下Density是什么？
在学习histogram的时候，我们把我们的数据分类成不同的区间
横坐标就是区间
纵坐标我们可以设置成Frequency，也就是去计数，不同区间的数据有多少
也可以设置成relative Frequency，也就是不同区间的数据所占比例是多少，比如左图：数据落在11,21这个区间大概有44%
那我们的vertical axis也可以设置成density，不知道大家还有没有印象… density是 relative Frequency 除以 区间的长度，也就是relative Frequency 等于我的纵坐标density乘以区间长度。这个是不是就是对应的矩形面积？

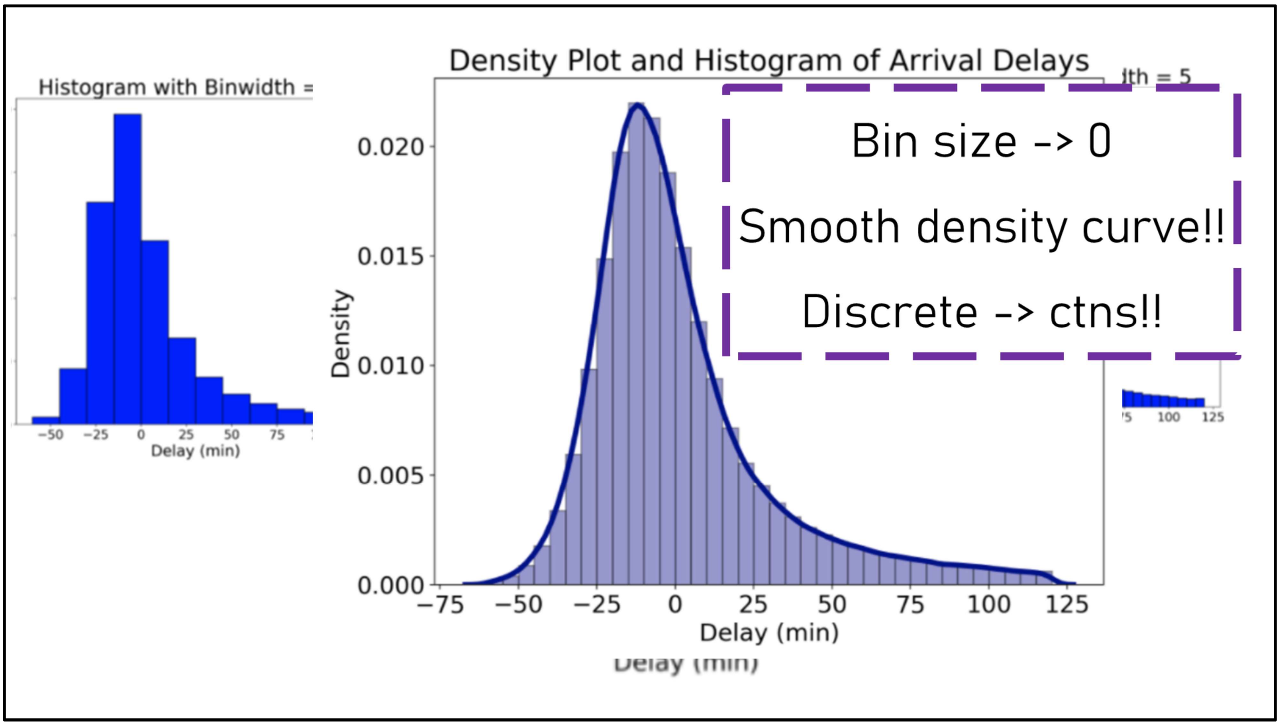
Density Curves

Density Histogram

working with smooth curves is much easier than jagged histograms

Density Curve

那我们看一下density histogram是怎么转化到density curve的
这个有点像微积分的思路



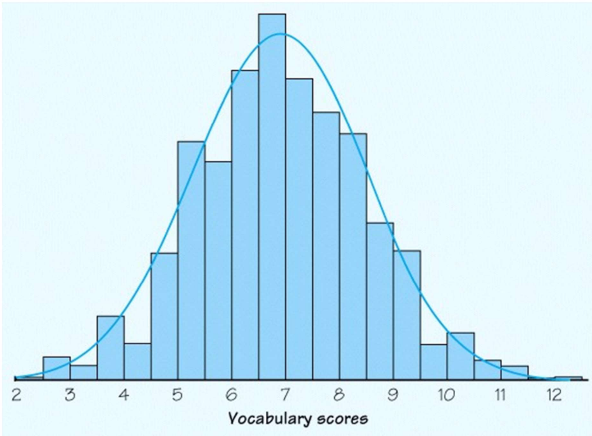
当我们的bin size是15的时候，density histogram是这样的
随着bin size 的减小，也就是分类越来越细，我们的density histogram就会变得越来越平滑，断层越来越不明显

当bin size 趋近于0的时候，我们就得到了一条平滑的曲线，a smooth density curve!
也就从离散变成了连续。
对于continuous random variable，我们就用density curve to describe their Distributions
这条曲线的方程也就叫做概率密度函数 pdf： Probability density function

Density Curves

A density curve is a **smooth curve** that describes the **probability distribution for a continuous random variable** X.

The function that defines this curve is denoted by f(x) and it is called the **density function.**



看一下正式的定义：
A density curve is a **smooth curve** that describes the **probability distribution for a continuous random variable** X.

The function that defines this curve is denoted by f(x)，我们会用fx来表示density function，这个是固定的，不要乱用其他的，and it is called the **density function**.

Density Curves -- Properties

1. $f(x) \geq 0$

看一下它的性质

1. 大于等于0

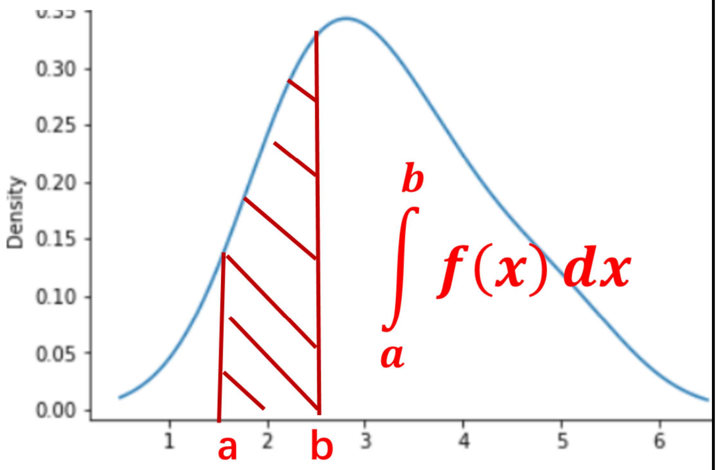
Density Curves -- Properties

2. The probability that X falls in an interval is the area under the density curve and above the horizontal axis.

$$density = \frac{relative\ frequency}{interval\ width}$$

$$= \frac{probability}{dx}$$

$$\mathrm{f(x)dx} = \mathrm{probability}$$

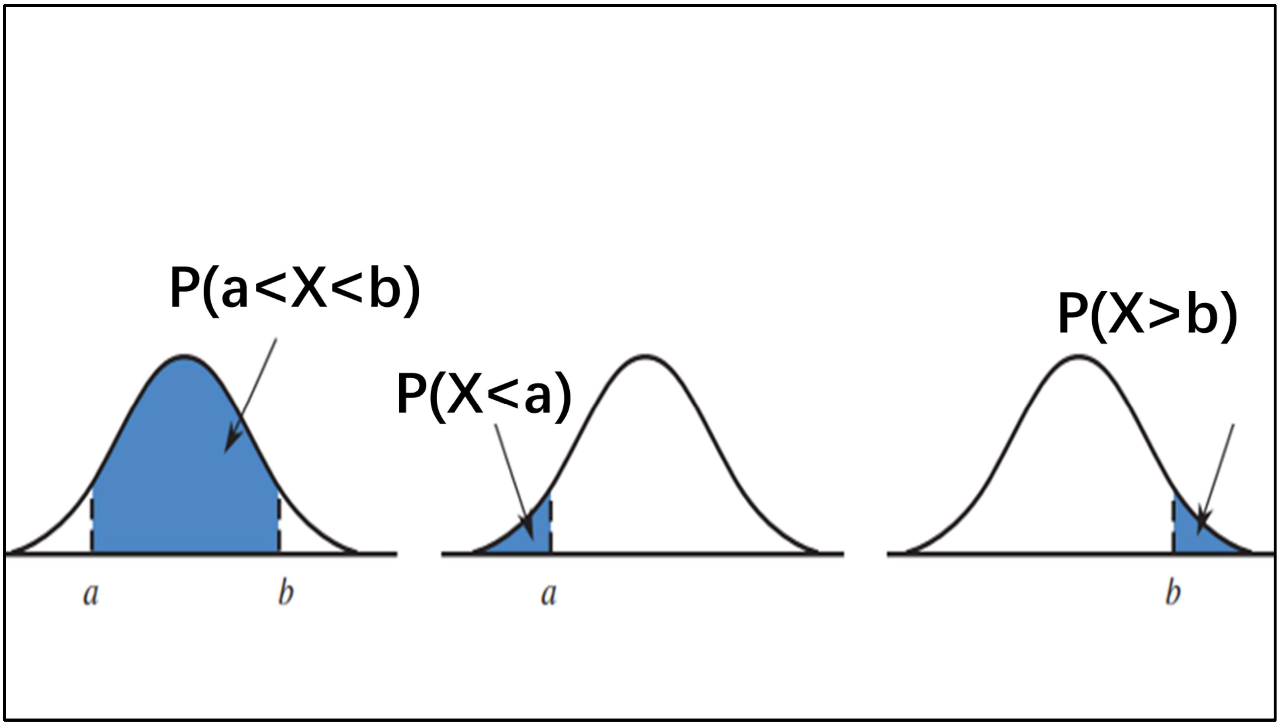


The probability that X falls in an interval is the area under the density curve and above the horizontal axis.

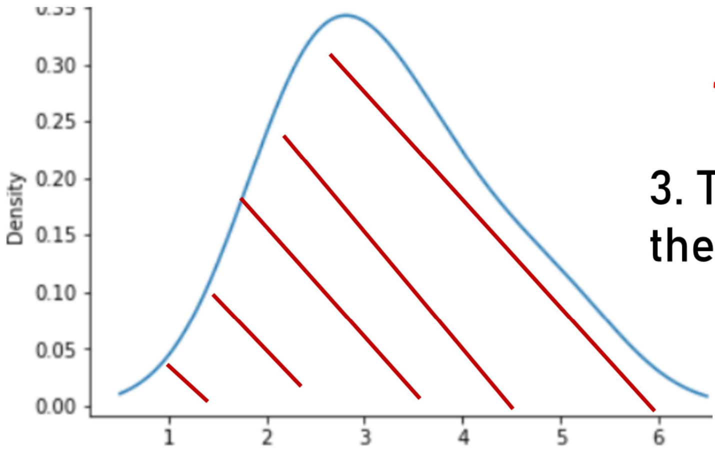
因为如果纵坐标是density的话，我们的面积对应的就是概率

density = relative Frequency / interval width 在连续变量的情况下，我们让区间长度趋近于0了，所以就是 dx,
也就是说 density function f(x) * dx = 这一段dx对应的概率

那区间a,b对应的概率应该就是 $\int_a^b f(x)\,dx$,也就是阴影部分的面积


P(a<X<b)
P(X<a)
P(X>b)
a
b
a
b

Density Curves -- Properties



$\int_{-\infty}^{+\infty} f(x)\, dx = ?$

3. The total area under the density curve is 1

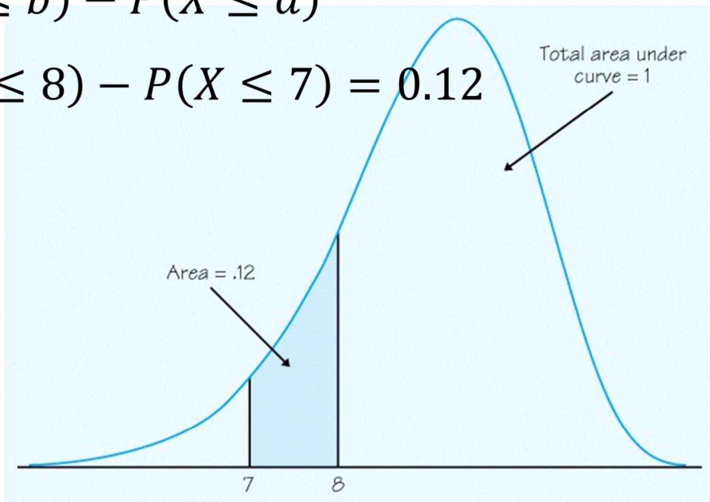
问：f(x) 下的面积总和是多少？

$\int f\ dx = \sum p$，全集的概率应该就是1

Density Curves -- Properties

4. $P(a < X \leq b) = P(X \leq b) - P(X \leq a)$

$P(7 < X \leq 8) = P(X \leq 8) - P(X \leq 7) = 0.12$



Density Curves -- Properties

$$P(x=a) = 0$$

If x is a continuous random variable, then for any two numbers a and b with $a < b$,

$$P(a \le x \le b) = P(a < x \le b) = P(a \le x < b) = P(a < x < b)$$

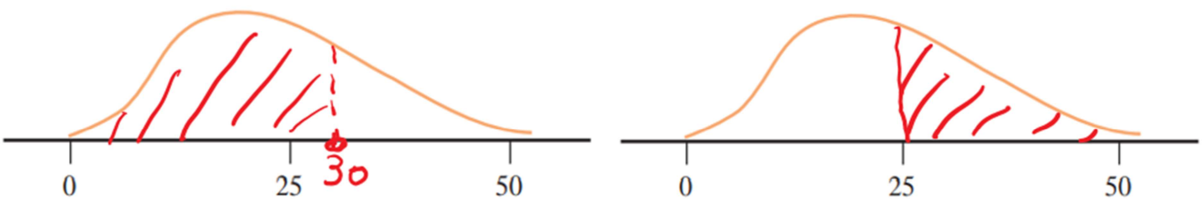
对于连续型随机变量来说，某一个点的概率是0
所以:

Practice:

Let X denote the lifetime (in thousands of hours) of a certain type of fan used in diesel engines. The density curve of X is as pictured.

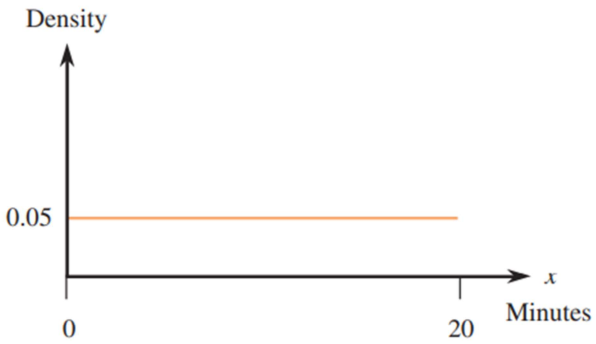
Shade the area under the curve corresponding to each of the following

1. $P(X<30)$
2. The probability that the lifetime is at least 25,000 hours


0
25
30
50
0
25
50

Practice:

Let X be the amount of time (in minutes) that a particular San Francisco commuter must wait for a BART train. Suppose that the density curve is as pictured (a uniform distribution):



a. What is the probability that X is less than 10 minutes? more than 15 minutes?

b. What is the probability that X is between 7 and 12 minutes?

c. Find the value c for which $P(X<c) = 0.9$

Practice:

What is the probability that X is less than 10 minutes?

$P(X<10) = 0.5$

more than 15 minutes?

$P(X>15) = 0.25$

What is the probability that X is between 7 and 12 minutes?

$P(7<X<12) = 0.25$

Find the value c for which $P(X<c) = 0.9$

$c = 18$

7.24 Let x denote the amount of gravel sold (in tons) during a randomly selected week at a particular sales facility. Suppose that the density curve has height $f(x)$ above the value x, where

$$f(x) = \begin{cases} 2(1 - x) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

a. $P\left(x < \frac{1}{2}\right)$

b. $P\left(x \leq \frac{1}{2}\right)$

c. $P\left(x < \frac{1}{4}\right)$

d. $P\left(\frac{1}{4} < x < \frac{1}{2}\right)$

Hint: Use the results of Parts (a)–(c)

Mean

Discrete R.V.: $\mu = \mathrm{E(X)} = \sum_i x_i p_i$

Continuous R.V.: probability $p = \mathrm{f(x)dx}$

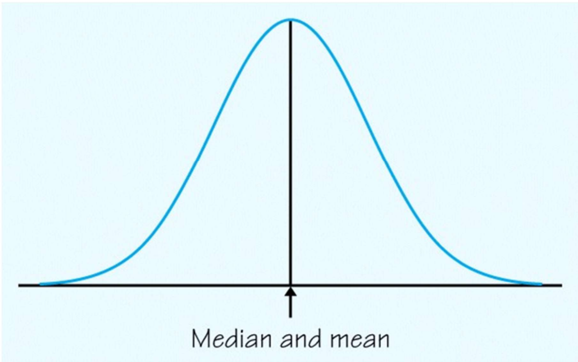
$$\mu = \mathrm{E(X)} = \int_{-\infty}^{+\infty} x f(x)\, dx$$

Median

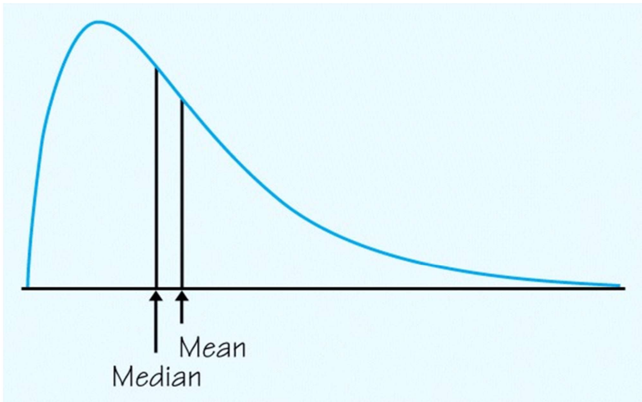
The median of a continuous Random Variable is the value of **a** such that:

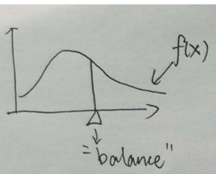
$$\text{probability} = \int_{-\infty}^{a} f(x)\,dx = 0.5$$

For a symmetric Density curve…



For a skewed density curve…





Assume the balance point is $x = \mu$.

G(solid on the left side) * moment of G_{left} should be equal to G_{right} * $moment_{right}$

$$\int_{-\infty}^{\mu} |x-\mu| f(x)\, dx = \int_{\mu}^{+\infty} |x-\mu| f(x)\, dx$$

moment

$$\therefore \int_{-\infty}^{\mu} (\mu - x) f(x)\, dx = \int_{\mu}^{+\infty} (x-\mu) f(x)\, dx$$

$$\therefore \int_{-\infty}^{\mu} x f(x)\, dx + \int_{\mu}^{+\infty} x f(x)\, dx = \int_{-\infty}^{\mu} f(x) \cdot \mu\, dx + \int_{\mu}^{+\infty} \mu f(x)\, dx$$

Left side $= \int_{-\infty}^{+\infty} x f(x)\, dx = E(x)$

Right side $= \mu \int_{-\infty}^{+\infty} f(x)\, dx = \mu * 1 = \mu$

Thus, $E(x)$ is the balance point of the solid material.