

Mean

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

Continuous R.V.: probability $p = f(x)dx$

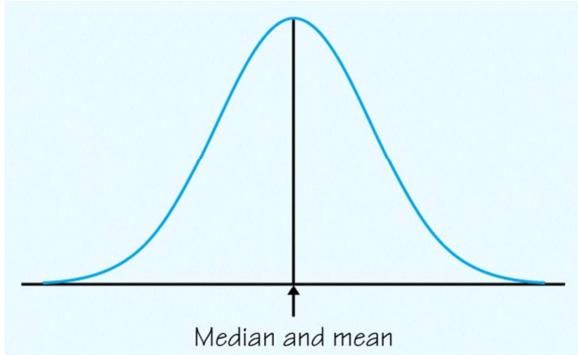
$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x) dx$$

Median

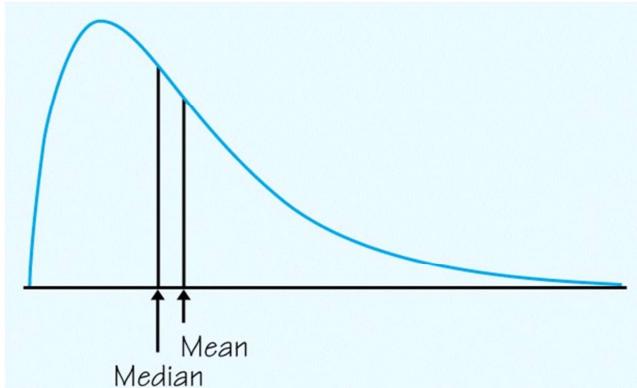
The median of a continuous Random Variable is the value of **a** such that:

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

For a symmetric Density curve...



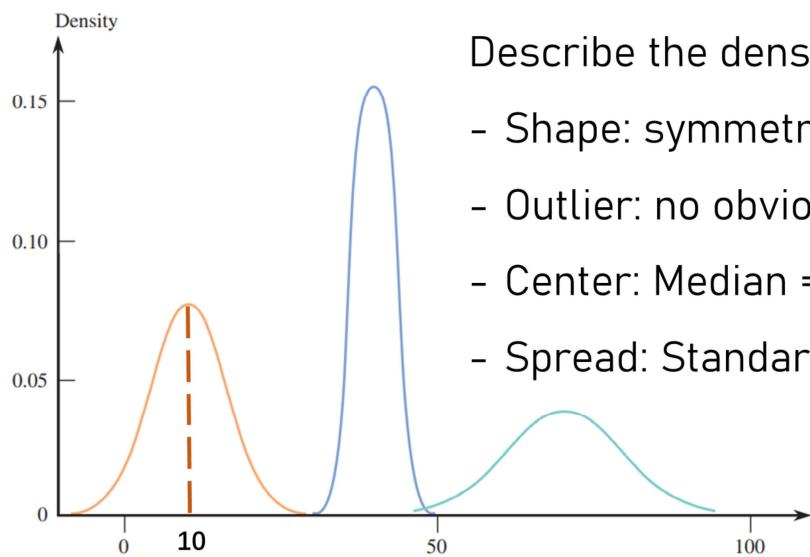
For a skewed density curve...



Normal Distribution

那我们来看一下density curve里面最典型的一个分布：正态分布，normal Distribution

Normal Curves



Describe the density curve:

- Shape: symmetric, single peak
- Outlier: no obvious outliers
- Center: Median = Mean
- Spread: Standard deviation, Variance...

这三个都是normal Distribution, 大家来想一下我们怎么去描述他们的分布
Hint: 想一下我们用于描述Distribution的四点：SOCS

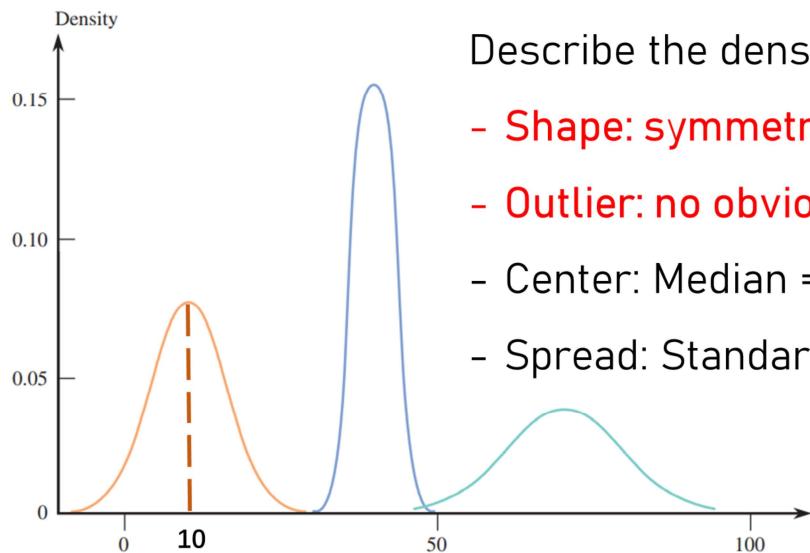
- **Shape: symmetric, single peak**
- **Outlier: no obvious outliers**
- **Center: Median = Mean**
- **Spread: Standard deviation, Variance...**

这四点里有哪些是这三个density curve的共同特点？

- Shape 和 Outlier

那这两个合起来就是 normal curve的定义。

Normal Curves



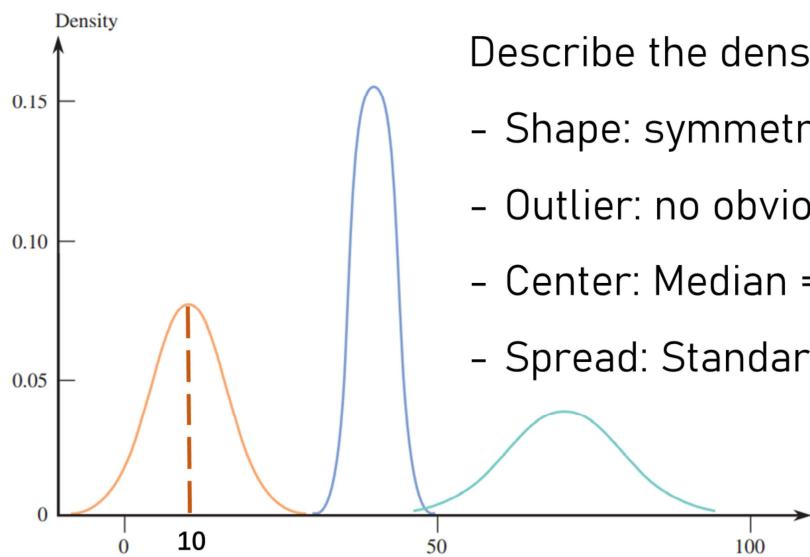
Describe the density curve:

- Shape: symmetric, single peak
- Outlier: no obvious outliers
- Center: Median = Mean
- Spread: Standard deviation, Variance...

Normal Curves

- A Normal distribution is described by a **symmetric, single-peaked, bell-shaped** density curve called a Normal curve.
- All Normal distributions have **the same overall shape**.

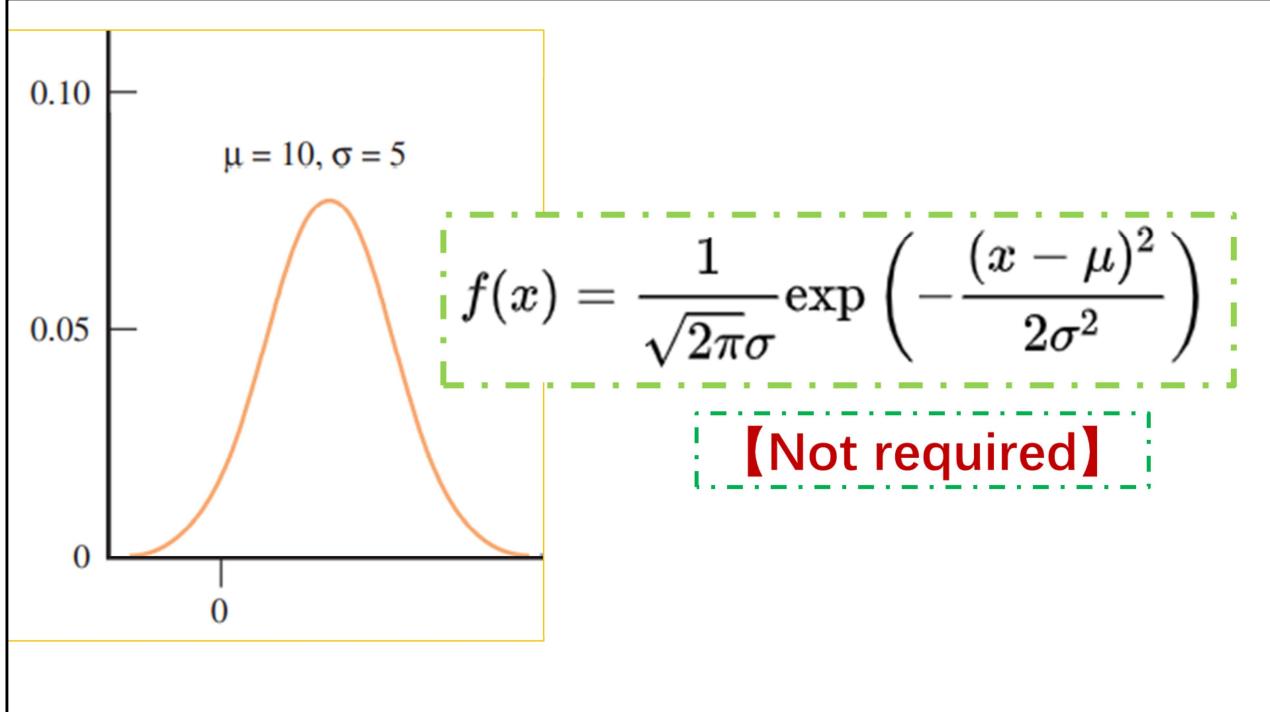
Normal Curves



Describe the density curve:

- Shape: symmetric, single peak
- Outlier: no obvious outliers
- Center: Median = Mean
- Spread: Standard deviation, Variance...

那这里面剩下的两点也就决定了 normal curve的具体形状，也就是说，如果我已知 center 和 spread的值，知道 μ 和 σ ，我就知道这个normal curve的方程式了

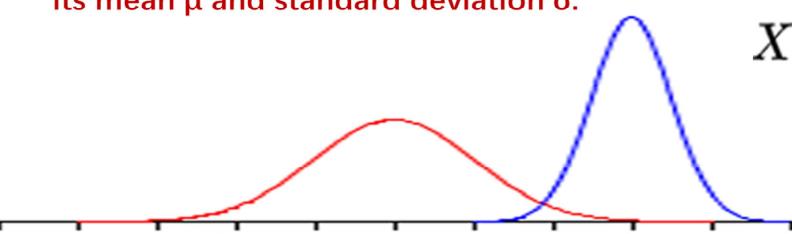


它的density function是这样的，我们可以看一下，这个不会考，我们就看一下，这里面只有两个系数mu和sigma，所以知道这俩，我们的function就定了

Normal Curves

- A Normal distribution is described by a **symmetric, single-peaked, bell-shaped** density curve called a Normal curve.
- All Normal distributions have **the same overall shape**.
- Any Normal distribution is completely specified by two numbers:
its mean μ and standard deviation σ .

$$X \sim N(\mu, \sigma^2)$$



Any Normal distribution is completely specified by two numbers: its mean μ and standard deviation σ .

所以我们也把随机变量 X 写成 $X \sim N(\mu, \sigma^2)$

这里需要注意一下，如果大家在写这个normal Distribution的时候，在 $N(,)$

第二个空这里写了一个数，那就默认为 σ^2

这个是有标准的，也是很容易在考试中丢一些比较冤的分

如果你非得写标准差，那你就明确的写出来 $\sigma = \dots$

那对于这两个系数，

均值 决定了这个density curve的中心在x轴上的位置

σ 决定了这个curve是高高瘦瘦还是矮矮胖胖的

这两个curve，哪个的 σ 要大一些？

左边的！因为数据分布的比较分散

右边的，数据更集中在中心

Properties

All Normal Distributions have the same overall shape.

If X follows the normal distribution,
 $aX+b$ still follows the normal distribution.

$$X \sim N(\mu, \sigma^2) \quad aX+b \sim N(?, ?)$$

$$\begin{aligned} E(aX+b) &= ? & a\mu + b \\ \text{Var}(aX+b) &= ? & a^2\sigma^2 \end{aligned} \quad aX+b \sim N(a\mu + b, a^2\sigma^2)$$

因为我们刚刚提到的， All Normal Distributions have the same overall shape.
看一下它的性质：

所以我们可以对这个normal curve进行平移和拉伸或压缩并不会改变它的分布类别，
它还是normal

也就是说， If X follows the normal distribution, $aX+b$ still follows the normal distribution.

这个证明我们不要求掌握，感兴趣的的同学自行知乎，有多种证明方式，最基础的是用cdf的导数来证

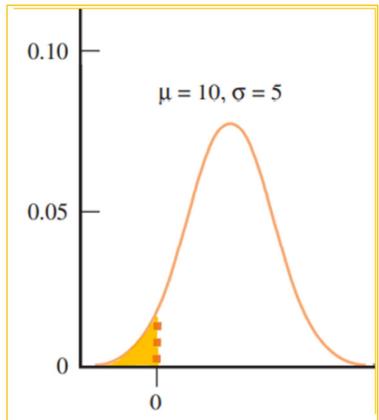
所以我们根据这个性质，能得到结论：

如果 $X \sim N(\mu, \sigma^2)$, 那 $aX+b$ 一定也是 $\sim N(a\mu + b, a^2\sigma^2)$

那具体的分布我们就要计算一下 $aX+b$ 的期望和方差了

大家试一下计算一下

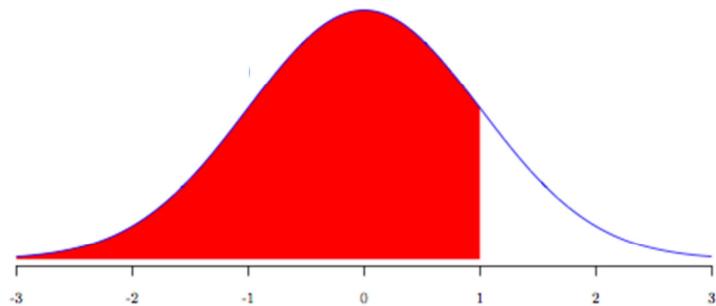
What is the value of $P(X < 0)$?



$$\int_{-\infty}^0 f(x) dx$$

DEFINITION Standard Normal distribution

The **standard Normal distribution** is the Normal distribution with mean 0 and standard deviation 1.



$$Z \sim N(0,1)$$

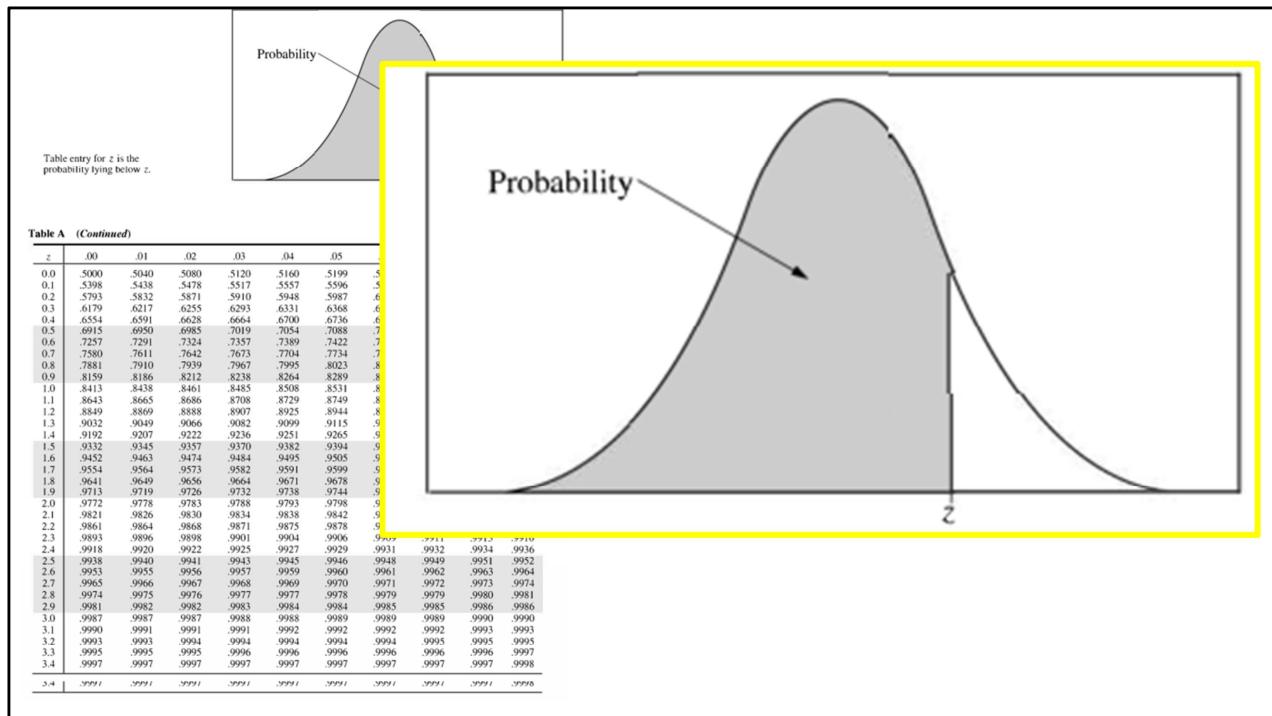
再来看一个定义：

这是一个特殊的normal Distribution，算是normal curve里面的一个基准
当均值为0，方差为1的时候，就叫做standard normal Distribution，标准正态分布

我们一般来说会把服从标准正态分布的随机变量定义为Z，这个和我们之前接触的Z-score也是有一定联系的

它的图像是这样的，红色这部分的面积就是 $X \leq x$ 的概率，这个应该是个特别复杂的积分，我们不用自己算，可以查表

我们考试的时候会有一个正态分布的表，我们来看一下咋看



这个就是这个表，大家在我们之前发的公示表里也能看到
上面是一个图，示意的是我们这个表里的概率指的都是 \leq 某个值的概率

	z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
T	0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
P	0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
Table	0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
—	0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
—	0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
—	0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
—	0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
—	0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
—	0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
—	0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
—	1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
—	1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
—	1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
—	1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
—	1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
—	1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
—	1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
—	1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
—	1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
—	1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
—	2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

下面的这部分

首先我们先看一下最左边的一列和最上面的一行，这个是Z的取值
中间则是对应的概率

比如蓝色框框里的，0.6179就是Z=0.30的概率，
它右边的0.6217对应的就是Z=0.31的概率

$$P(X \leq 1.33) = 0.9082$$



<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5635	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9082	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

举个例子

如果想求 $X \leq 1.33$ 的概率，我们怎么找

首先我们要找到1.33，我们就需要找到1.3 和 0.03

那对应的0.9082就是它的概率了

Use the table of Standard Normal Curve Areas to find

$$P(-1.76 < Z < 0.58)$$

$$P(-1.76 < Z < 0.58)$$

$$= P(Z < 0.58) - P(Z < -1.76)$$

$$= P(Z < 0.58) - [1 - P(Z < 1.76)]$$

$$= 0.7190 - 0.0392$$

$$= 0.6798$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

练习一下

**We only have the distribution table of the
STANDARD NORMAL DISTRIBUTION!!!**

但是我们不可能有一个分布，我就去算一个分布的表格，所以我们只有标准正态分布的表格

那如果我们遇到了均值不是0的随机变量了，想求概率的话咋办？

$X \sim N(\mu, \sigma^2)$ $X \stackrel{?}{\leftrightarrow}$ standard normal random variable

$X - \mu \sim ? \quad N(0, \sigma^2)$

Z-value (Z-score)

$\frac{X - \mu}{\sigma} \sim ? \quad N(0, 1)$

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

$P(X < x) = P\left(\frac{X - \mu}{\sigma} < \frac{x - \mu}{\sigma}\right) = P(Z < \frac{x - \mu}{\sigma})$ Then use the table!

怎么能把X和标准正态分布联系到一起呢?

Hint: $x - \mu$ 的分布是什么样的?

我们也把这个叫做Z-score或者Z-value

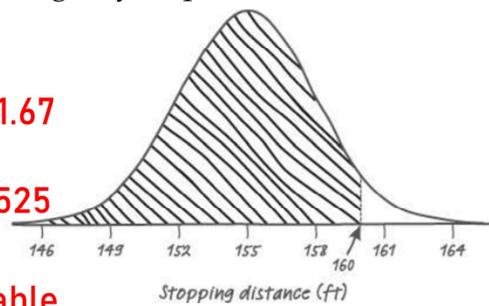
如果我想求 $X < x$ 的概率我可以两边同时 $-\mu / \sigma$, 那我就会得到一个 $N(0, 1)$ 的变量 $< x - \mu / \sigma$ 的概率
然后就可以查表了

Studies on automobile safety suggest that stopping distances follow an approximately Normal distribution. For one model of car traveling at 62 mph, the mean stopping distance is $\mu=155$ ft with a standard deviation of $\sigma=3$ ft . Danielle is driving one of these cars at 62 mph when she spots a wreck 160 feet in front of her and needs to make an emergency stop. About what percent of cars of this model when going 62 mph would be able to make an emergency stop in less than 160 feet? Is Danielle likely to stop safely?

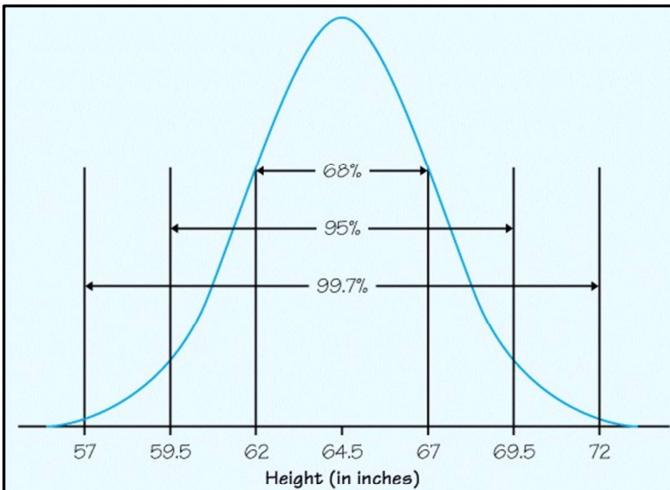
$$Z = (160 - 155)/3 = 1.67$$

Using the table, we have the area for $Z < 1.67$ is 0.9525.

$$P(\text{stopping distance} < 160) = P(Z < 1.67) = 0.9525$$



About 95% of cars of this model would be able to make an emergency stop within 160 feet. So Danielle is likely to be able to stop safely.



The 68–95–99.7 RULE

In a Normal distribution with mean μ and standard deviation σ :

- Approximately **68%** of the observations fall within σ of the mean μ .
- Approximately **95%** of the observations fall within 2σ of the mean μ .
- Approximately **99.7%** of the observations fall within 3σ of the mean μ .

This result is known as the **68–95–99.7 rule**.

大家对68 95 99.7还有没有印象

这个其实是源于normal Distribution

对于正态分布来说，

均值+-一倍的标准差，会包含68%的数据

均值+-2倍的标准差，会包含95%的数据

均值+-3倍的标准差，会包含99.7%的数据

Percentiles

The *percentile* of a score, x , is the percentage of scores which fall at or below the score.

The k th percentile P_k :

$$\int_{-\infty}^{P_k} f(x) dx = k\%$$

Percentile 百分位数， P_k 对应的应该就是排在 $k\%$ 这个位置的值，所以：

Determine the value of each of the following percentiles for the standard normal distribution (Hint: If the cumulative area that you must look for does not appear in the z table, use the closest entry):

- a. The 91st percentile
- b. The 77th percentile
- c. The 50th percentile
- d. The 9th percentile
- e. What is the relationship between the 70th z percentile and the 30th z percentile?

Determine the value of each of the following percentiles for the standard normal distribution (Hint: If the cumulative area that you must look for does not appear in the z table, use the closest entry):

- a. The 91st percentile **1.341**
- b. The 77th percentile **0.737**
- c. The 50th percentile **0**
- d. The 9th percentile **-1.34**

e. What is the relationship between the 70th z percentile and the 30th z percentile?

70th z percentile = - 30th z percentile

Data from the paper “Fetal … Composition” suggest that a normal distribution with mean 3500 grams and standard deviation 600 grams is a reasonable model for the probability distribution of the continuous numerical variable X = birth weight of a randomly selected full-term baby. What proportion of birth weights are between 2900 and 4700 grams?

$$\begin{aligned}
 a^* &= \frac{a - \mu}{\sigma} = P(2900 < x < 4700) = P(-1.00 < z < 2.00) \\
 &\quad = (\text{z curve area to the left of } 2.00) \\
 &\quad \quad - (\text{z curve area to the left of } -1.00) \\
 b^* &= \frac{b - \mu}{\sigma} = .9772 - .1587 \\
 &\quad = .8185
 \end{aligned}$$

Garbage trucks entering a particular waste management facility are weighed and then they offload garbage into a landfill. Data from the paper “Estimating … GPS” suggest that a normal distribution with mean 13 minutes and standard deviation 3.9 minutes is a reasonable model for the probability distribution of the random variable X = total processing time for a garbage truck at this waste management facility /total

process weight descr the lc corre: For the standard normal distribution, the largest 10% are those with z values greater than $z^* = 1.28$ (from Appendix Table 2, based on a cumulative area of .90). Then $x^* = \mu + z^*\sigma$ $= 13 + 1.28(3.9)$ $= 13 + 4.992$ $= 17.992$: to 10% with 1 times ation.

About 10% of the garbage trucks using this facility would have a total processing time of more than 17.992 minutes.

Determine the value of z^* such that

- a.- z^* and z^* separate the middle 95% of all z values from the most extreme 5%
- b.- z^* and z^* separate the middle 90% of all z values from the most extreme 10%
- c.- z^* and z^* separate the middle 98% of all z values from the most extreme 2%
- d.- z^* and z^* separate the middle 92% of all z values from the most extreme 8%

Determine the value of z^* such that

- a.- z^* and z^* separate the middle 95% of all z values from the most extreme 5% $z^*=1.96$
- b.- z^* and z^* separate the middle 90% of all z values from the most extreme 10% $z^*=1.645$
- c.- z^* and z^* separate the middle 98% of all z values from the most extreme 2% $z^*=2.33$
- d.- z^* and z^* separate the middle 92% of all z values from the most extreme 8% $z^*=1.75$