



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY



生物医学统计概论(BI148-2)

Fundamentals of Biomedical Statistics

2022 春季

授课老师：林关宁



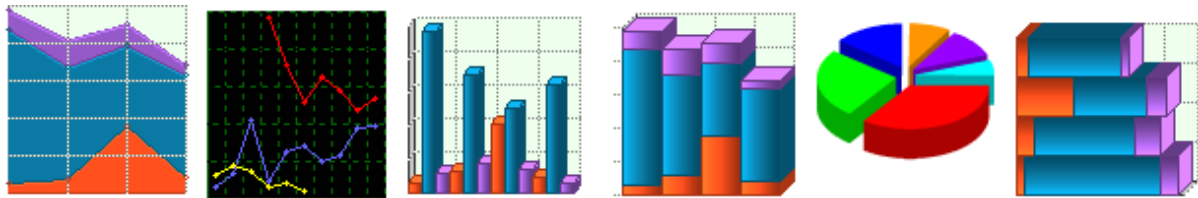
课程内容安排



| 上课日期 | 章节 | 教学内容 | 教学要点 | 作业 | 随堂测 | 学时 |
|---------|----|-----------------|---|-----------|-----------|----|
| 2.16 | 1 | 数据可视化, 描述性统计 | 1. 课程介绍 & 数据类型 | 作业1 (8%) | 测试1 (8%) | 2 |
| 2.23 | | | 2. 描述性统计 Descriptive Statistics & 数据常用可视化 | | | 2 |
| 3.2 | | | 3. 常用概率分布 | | | 2 |
| 3.9 | | | 4. 大数定理 & 中心极限定理 | | | 2 |
| 3.16 | 2 | 推断性统计, 均值差异检验 | 5. 统计推断基础-1: 置信区间 Confidence Interval * | 作业2 (12%) | 测试2 (12%) | 2 |
| 3.23 | | | 6. 统计推断基础-2: 假设检验 Hypothesis Test | | | 2 |
| 3.30 | | | 7. 数值数据的均值比较-1: 单样本t-检验 | | | 2 |
| 4.6 | | | 8. 数值数据的均值比较-2: 独立双样本t-检验, 配对样本t-检验 | | | 2 |
| 4.13 | | | 9. 数值数据的均值比较-3: One-Way ANOVA | | | 2 |
| 4.20 | | | 10. 数值数据的均值比较-4: Two-way ANOVA | | | 2 |
| 4.27 | 3 | 比例差异检验 | 11. 类别数据的比例比较-1: 单样本比例推断 * | 作业3 (4%) | 测试3 (4%) | 2 |
| 5.7 (调) | | | 12. 类别数据的比例比较-2: 联立表的卡方检验 | | | 2 |
| 5.11 | 4 | 协方差, 相关分析, 回归分析 | 13. 相关分析 (Pearson r, Spearman rho, Kendal' s tau) * | 作业4 (6%) | 测试4 (6%) | 2 |
| 5.18 | | | 14. 简单回归分析 | | | 2 |
| 5.25 | | | 15. 多元回归 Multiple Regression | | | 2 |
| 6.1 | 5 | Course Summary | 16. 课程总结 * | | | 2 |
| | | | Total | 30% | 30% | 32 |

* 随堂测试

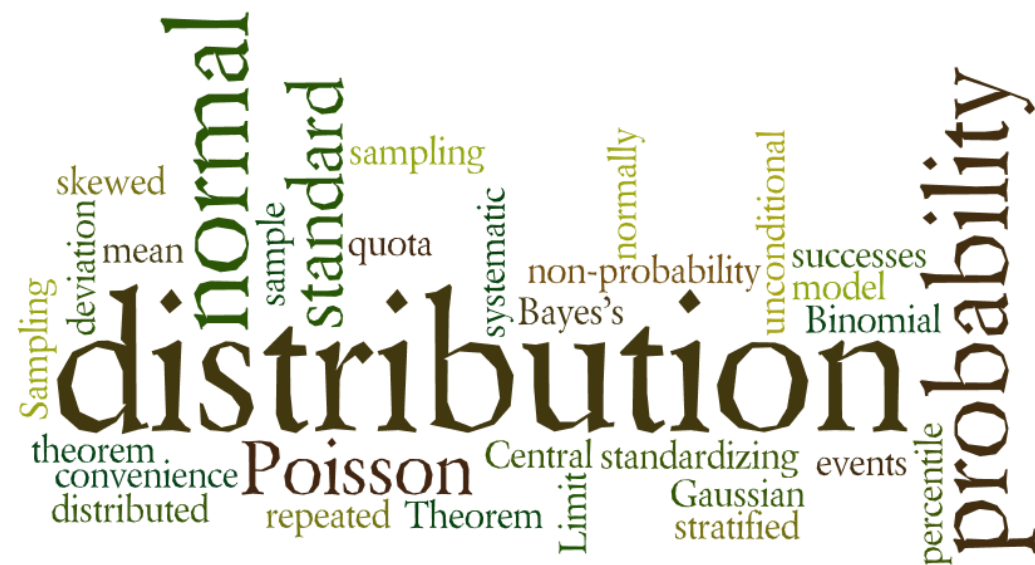




Week 3

常用概率分布

(参考书第3章)



今日摘要

1. Random Variables (随机变量)
2. The Bernoulli trials
3. The Binomial distribution
4. The Poisson distribution
5. The Normal distribution

- 伯努利分布
- 二项分布
- 泊松分布



- 正态分布

Random variables 随机变量

1. **Definition of random variables**
2. Distributions of random variables
3. Mean, variance, and standard deviation for random variables



DEFINITION OF A RANDOM VARIABLE

A *random variable* is a function that maps each event in a sample space to a number.


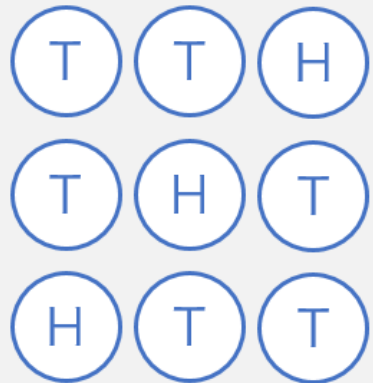
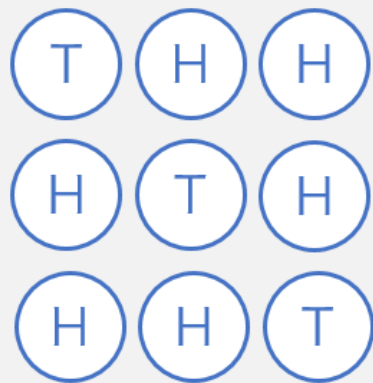

离散随机变量

- A *discrete random variable* takes on a finite number of values.

Suppose X is the number of heads in 3 tosses of a fair coin.

- X can take on the values 0, 1, 2, 3.

- 表示随机试验各种结果的函数
- 随机事件不论与数量是否直接有关，都可以数量化
- 随机变量就是量化随机事件的函数

| | | | |
|---|---|--|---|
|  |  |  |  |
| $X = 0$ | $X = 1$ | $X = 2$ | $X = 3$ |

Random variables 随机变量

1. Definition of random variables
- 2. Distributions of random variables**
3. Mean, variance, and standard deviation for random variables



DISTRIBUTION OF A DISCRETE RANDOM VARIABLE

离散随机变量的分布

The distribution of a *discrete random variable* is the collection of its **values and the probabilities** associated with those values. 离散随机变量的分布是**其值和与出现这些值相关的概率的集合**。

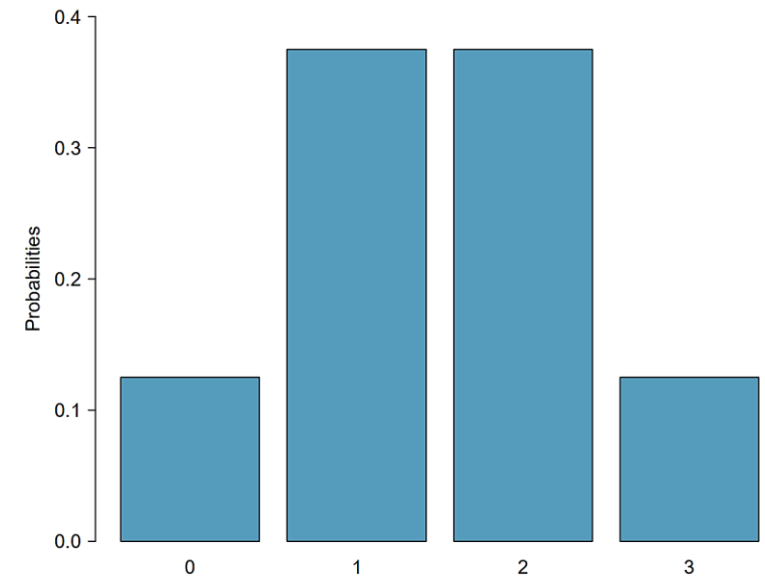
The probability distribution for X is as follows:

离散概率分布

| x_i | 0 | 1 | 2 | 3 |
|--------------|-------|-------|-------|-------|
| $P(X = x_i)$ | $1/8$ | $3/8$ | $3/8$ | $1/8$ |

$$\sum_{x=0}^3 P(X = x_i) = 1$$

BAR GRAPH SHOWING A DISTRIBUTION



Random variables 随机变量

1. Definition of random variables
2. Distributions of random variables
3. **Mean, variance, and standard deviation for random variables**



EXPECTATION OF A RANDOM VARIABLE 随机变量的期望值

If X has outcomes x_1, \dots, x_k with probabilities $P(X = x_1), \dots, P(X = x_k)$, the **expected value of X** is the sum of each outcome multiplied by its corresponding probability:

x 的期望值是每个结果乘以相应的概率的总和

$$E(X) = x_1 P(X = x_1) + \dots + x_k P(X = x_k) = \sum_{i=1}^k x_i P(X = x_i)$$

The Greek letter μ may be used in place of the notation $E(X)$ and is sometimes written μ_X .

In the coin tossing example,

$$\begin{aligned} E(X) &= 0P(X = 0) + 1P(X = 1) + 2P(X = 2) + 3P(X = 3) \\ &= (0)(1/8) + (1)(3/8) + (2)(3/8) + (3)(1/8) \\ &= 12/8 \\ &= 1.5 \end{aligned}$$



VARIANCE AND SD OF A RANDOM VARIABLE

随机变量的方差,
标准差 (σ)

If X takes on outcomes x_1, \dots, x_k with probabilities $P(X = x_1), \dots, P(X = x_k)$ and expected value $\mu = E(X)$, then the variance of X , denoted by $\text{Var}(X)$ or σ^2 , is

$$\begin{aligned}\text{Var}(X) &= (x_1 - \mu)^2 P(X = x_1) + \dots + (x_k - \mu)^2 P(X = x_k) \\ &= \sum_{j=1}^k (x_j - \mu)^2 P(X = x_j)\end{aligned}$$

The standard deviation of X , written as $\text{SD}(X)$ or σ , is the square root of the variance. It is sometimes written σ_X .

In the coin tossing example,

$$\begin{aligned}\sigma_X^2 &= (x_1 - \mu_X)^2 P(X = x_1) + \dots + (x_4 - \mu)^2 P(X = x_4) \\ &= (0 - 1.5)^2 (1/8) + (1 - 1.5)^2 (3/8) + (2 - 1.5)^2 (3/8) + (3 - 1.5)^2 (1/8) \\ &= 3/4\end{aligned}$$

The standard deviation is $\sqrt{3/4} = \sqrt{3}/2 = 0.866$.



今日摘要

1. Random Variables (随机变量)
2. The Bernoulli trials
3. The Binomial distribution
4. The Poisson distribution
5. The Normal distribution

- 伯努利分布
- 二项分布
- 泊松分布



- 正态分布



(1) 伯努利分布 (Bernoulli distribution)、两点分布

耶鲁大学心理学家斯坦利·米尔格拉姆 (Stanley Milgram) 在1963年开始了一系列实验，以探讨个人对权威人物的服从情况。在实验中，参与者分“教师”和“学生”，权威人物会命令“教师”对学习出错的“学生”进行一系列越来越严重的电击。米尔格拉姆发现，只有大约35%的参与者会抵抗权威，在达到最大电压之前停止对受罚者进行电击。后来在其他数个研究也显示，这一数字 (35%) 在不同社区和不同时间大致一致。

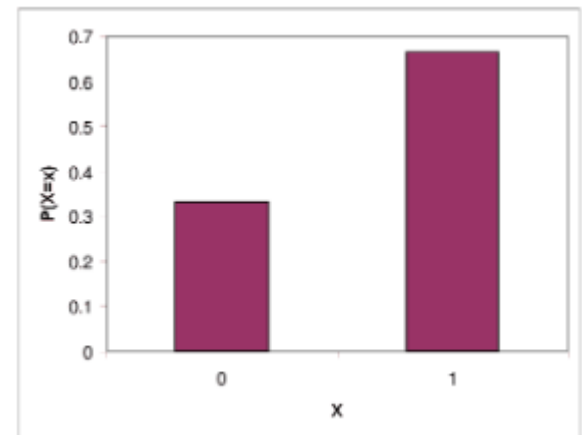
米尔格拉姆实验中的每个人都可以被认为是一次试验 (trail)。假设如果一个参与者拒绝实施最严重的电击，试验被标记为成功 (success, 1)；如果该参与者照命令进行了最严重的电击，试验则是失败 (failure, 0)。成功的概率可以写为 $p = 0.35$ 。失败的概率有时用 $q = 1 - p$ 表示。

当单个试验只有两种可能的结果时，称为伯努利随机变量 (Bernoulli random variable)。

The sample proportion, \hat{p} , is the sample mean of these observations:

$$\hat{p} = \frac{\# \text{ of successes}}{\# \text{ of trials}} = \frac{0 + 1 + 1 + 1 + 1 + 0 + 1 + 1 + 0 + 0}{10} = 0.6.$$

- 伯努利试验 (Bernoulli trail)
 - 伯努利试验是在同样的条件下重复地、各次之间相互独立地进行的试验
 - 随机试验的结果只有两种结果 (成功，或者失败)。
 - 最常见的例子为抛硬币 (正面或者反面)
- 期望: $E(X) = p$
 - 方差: $\text{Var}(X) = \sigma^2 = p(1-p) = pq$
 - 其中, p 为每一次事件成功的概率, q 为每一次事件失败的概率。



(1) 伯努利分布 (Bernoulli distribution)、两点分布

EXAMPLE 3.9

Suppose that four individuals are randomly selected to participate in Milgram's experiment. What is the chance that there will be exactly one successful trial, assuming independence between trials? Suppose that the probability of success remains 0.35.



今日摘要

1. Random Variables (随机变量)
2. The Bernoulli trials
3. The Binomial distribution
4. The Poisson distribution
5. The Normal distribution

- 伯努利分布
- 二项分布
- 泊松分布



- 正态分布

(2) 二项分布 (Binomial distribution)

The Bernoulli distribution is unrealistic in all but the simplest of settings. However, it is a useful building block for other distributions. The binomial distribution describes the probability of having exactly k successes in n independent Bernoulli trials with probability of a success p . In Example 3.9, the goal was to calculate the probability of 1 success out of 4 trials, with probability of success 0.35 ($n = 4, k = 1, p = 0.35$).

Like the Bernoulli distribution, the binomial is a discrete distribution, and can take on only a finite number of values. A binomial variable has values $0, 1, 2, \dots, n$.

二项分布描述了在 n 个独立的伯努利试验中获得 k 次成功的概率，成功概率为 p 。
前面举的米尔格拉姆实验例子里，就是计算每4次试验有一次反对的概率是多少。
(Example 3.9)

The binomial coefficient $\binom{n}{x}$ is the number of ways to choose x items from a set of size n , where the order of the choice is ignored.

Mathematically,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- $n = 1, 2, \dots$
- $x = 0, 1, 2, \dots, n$
- For any integer m , $m! = (m)(m-1)(m-2) \cdots (1)$



(2) 二项分布 (Binomial distribution)

Let x = number of successes in n trials

$$P(x \text{ successes}) = \binom{\# \text{ of trials}}{\# \text{ of successes}} p^{\# \text{ of successes}} (1 - p)^{\# \text{ of trials} - \# \text{ of successes}}$$

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

Parameters of the distribution:

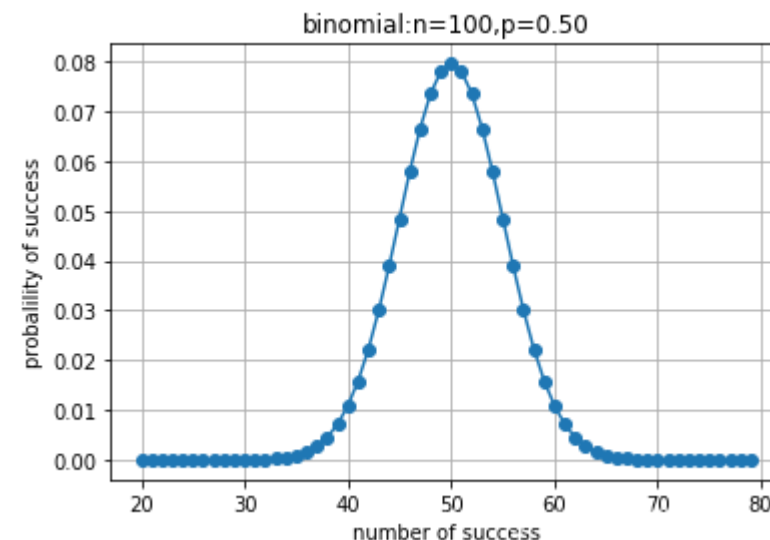
- n = number of trials
- p = probability of success

Shorthand notation: $X \sim \text{Bin}(n, p)$



(2) 二项分布 (Binomial distribution)

- 如何检验是二项分布？
 - 做某件事的次数是固定的，且 n 次事件是相互独立的（次数用 n 表示）。
 - 每一次事件都有两个可能的结果（成功，或者失败）
 - 每一次事件成功的概率都相等，成功的概率用 p 表示
 - 求**发生 k 次的概率**，服从二项分布
(PS: 二项分布就是 n 重伯努利分布。伯努利分布是 **$n=1$** 时的二项分布的特殊情况)
- 期望: $E(X) = np$
 - 方差: $\text{Var}(X) = np(1-p) = npq$
 p 为每一次事件成功的概率, q 为每一次事件失败的概率
- 比如: 抛硬币的问题, 做100次实验, 正反面概率都为0.5
可以看到, 对于 $n = 100$ 次实验中,
有50次成功的概率 (正面向上) 的概率最大



(2) 二项分布 (Binomial distribution)

What is the probability that 3 of 8 randomly selected participants will refuse to administer the worst shock?

First, check the conditions for applying the binomial model. The number of trials is fixed ($n = 8$) and each trial outcome can be classified as either success or failure. The sample is random, so the trials are independent, and the probability of success is the same for each trial.

For the outcome of interest, $k = 3$ successes occur in $n = 8$ trials, and the probability of a success is $p = 0.35$. Thus, the probability that 3 of 8 will refuse is given by

$$\begin{aligned} P(X = 3) &= \binom{8}{3} (0.35)^3 (1 - 0.35)^{8-3} = \frac{8!}{3!(8-3)!} (0.35)^3 (1 - 0.35)^{8-3} \\ &= (56)(0.35)^3 (0.65)^5 \\ &= 0.28. \end{aligned}$$



```
▶ import numpy as np
import pandas as pd
from scipy.stats import binom, norm, poisson
```

```
▶ # What is the probability that 3 of 8 randomly selected participants
# will refuse to administer the worst shock?

#calculate binomial probability of  $P(x=4)$ 
result = binom.pmf(k=3, n=8, p=0.35)

#Print the result
print("Binomial Probability: ", result)
```

Binomial Probability: 0.27858577906250004

```
▶ # What is the probability that less than 3 of 8 randomly selected participants
# will refuse to administer the worst shock?
result = binom.cdf(k=2, n=8, p=0.35)

#Print the result
print("Binomial Probability: ", result)
```

Binomial Probability: 0.42781365707031244



(2) 二项分布 (Binomial distribution)

BINOMIAL DISTRIBUTION

Suppose the probability of a single trial being a success is p . The probability of observing exactly k successes in n independent trials is given by

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k}. \quad (3.11)$$

Additionally, the mean, variance, and standard deviation of the number of observed successes are, respectively

$$\mu = np \qquad \sigma^2 = np(1 - p) \qquad \sigma = \sqrt{np(1 - p)}. \quad (3.12)$$

A binomial random variable X can be expressed as $X \sim \text{Bin}(n, p)$.



今日摘要

- Random Variables (随机变量)
- The Bernoulli trials
- The Binomial distribution
- **The Poisson distribution**
- **The Normal distribution**

- **伯努利分布**
- **二项分布**
- **泊松分布**



- **正态分布**

(3) 泊松分布 (Poisson distribution)

The Poisson distribution is used to calculate probabilities for rare events that accumulate over time.

泊松分布用于计算随时间累积的罕见事件的概率

It is used most often in settings where events happen at a rate λ per unit of population and per unit time, such as the annual incidence of a disease in a population. 它最常用于以每单位人口和每单位时间 λ 的速率发生的事件，例如人口中疾病的年发病率。

- Typical example: for children ages 0 - 14, the incidence rate of acute lymphocytic leukemia (ALL) was approximately 30 diagnosed cases per million children per year in 2010. 2010年，0-14岁儿童急性淋巴细胞白血病 (ALL) 的发病率约为30例/百万儿童/年



(3) 泊松分布 (Poisson distribution)

EXAMPLE: OUTBREAKS OF CHILDHOOD LEUKEMIA

Fortunately, childhood cancers are rare.

For children ages 0 - 14, the incidence rate of acute lymphocytic leukemia (ALL) was approximately 30 diagnosed cases per million children per year in the decade from 2000 - 2010. Approximately 20% of the US population are in this age range. 从2000年到2010年的十年间, 0-14岁儿童的急性淋巴细胞白血病 (ALL) 发病率约为每年每百万儿童中有30例确诊病例。大约20%的美国人口处于这个年龄段。

- What is the incidence rate over a 5 year period? 五年内的发病率是多少?
- In a small city of 75,000 people, what is the probability of observing exactly 8 cases of ALL over a 5 year period? 在一个有75000人的小城市里, 在5年时间里观察到8个病例的概率是多少?
- In the small city, what is the probability of observing 8 or more cases over a 5 year period? 在小城市里, 5年内观察到8个或更多病例的概率是多少?



(3) 泊松分布 (Poisson distribution)

当事件在一段时间内累积, 使得事件在一个时间间隔内发生的概率与时间间隔的长度成正比, 且非重叠时间间隔内的事件数是独立的, 则参数 λ (希腊字母 lambda) 表示单位时间内的平均事件数 (每单位时间的速率)

Suppose events occur over time in such a way that

1. The probability an event occurs in an interval is proportional to the length of the interval.
2. Events occur independently at a rate λ per unit of time.

Then the probability of exactly x events in one unit of time is

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

The probability of exactly x events t units of time is

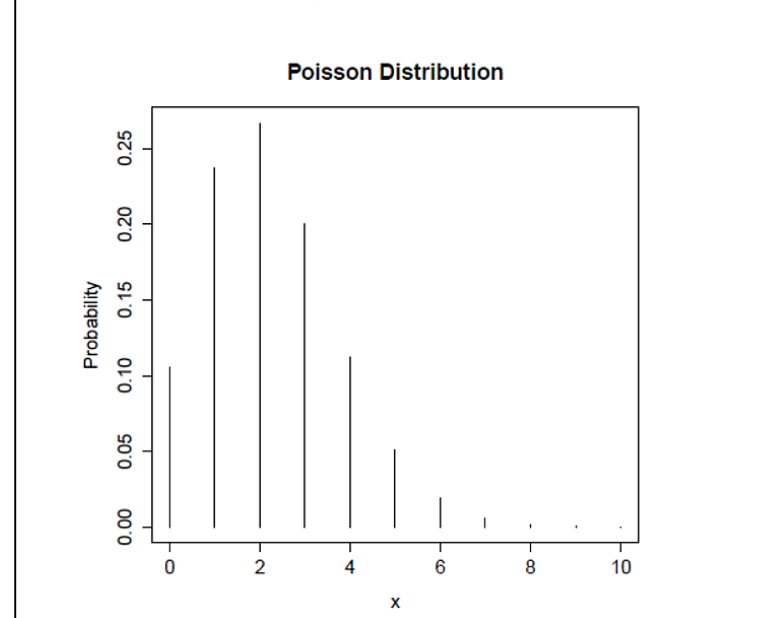
$$P(X = x) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}, \quad x = 0, 1, 2, \dots$$

For the Poisson distribution modeling the number of events in one unit of time:

- The mean is λ .
- The standard deviation is $\sqrt{\lambda}$.

In t units of time, the mean and standard deviation are, respectively, λt and $\sqrt{\lambda t}$.

POISSON DISTRIBUTION WITH $\lambda = 2.25$



(3) 泊松分布 (Poisson distribution)

EXAMPLE: OUTBREAKS OF CHILDHOOD LEUKEMIA

Fortunately, childhood cancers are rare.

For children ages 0 - 14, the incidence rate of acute lymphocytic leukemia (ALL) was approximately 30 diagnosed cases per million children per year in the decade from 2000 - 2010. Approximately 20% of the US population are in this age range. 从2000年到2010年的十年间, 0-14岁儿童的急性淋巴细胞白血病 (ALL) 发病率约为每年每百万儿童中有30例确诊病例。大约20%的美国人口处于这个年龄段。

- What is the incidence rate over a 5 year period? 五年内的发病率是多少?



(3) 泊松分布 (Poisson distribution)

EXAMPLE: OUTBREAKS OF CHILDHOOD LEUKEMIA

Fortunately, childhood cancers are rare.

For children ages 0 - 14, the incidence rate of acute lymphocytic leukemia (ALL) was approximately 30 diagnosed cases per million children per year in the decade from 2000 - 2010. Approximately 20% of the US population are in this age range. 从2000年到2010年的十年间, 0-14岁儿童的急性淋巴细胞白血病 (ALL) 发病率约为每年每百万儿童中有30例确诊病例。大约20%的美国人口处于这个年龄段。

- What is the incidence rate over a 5 year period? 五年内的发病率是多少?
- In a small city of 75,000 people, what is the probability of observing exactly 8 cases of ALL over a 5 year period? 在一个有75000人的小城市里, 在5年时间里观察到8个病例的概率是多少?



(3) 泊松分布 (Poisson distribution)

EXAMPLE: OUTBREAKS OF CHILDHOOD LEUKEMIA

Fortunately, childhood cancers are rare.

For children ages 0 - 14, the incidence rate of acute lymphocytic leukemia (ALL) was approximately 30 diagnosed cases per million children per year in the decade from 2000 - 2010. Approximately 20% of the US population are in this age range. 从2000年到2010年的十年间, 0-14岁儿童的急性淋巴细胞白血病 (ALL) 发病率约为每年每百万儿童中有30例确诊病例。大约20%的美国人口处于这个年龄段。

- What is the incidence rate over a 5 year period? 五年内的发病率是多少?
- In a small city of 75,000 people, what is the probability of observing exactly 8 cases of ALL over a 5 year period? 在一个有75000人的小城市里, 在5年时间里观察到8个病例的概率是多少?
- In the small city, what is the probability of observing 8 or more cases over a 5 year period? 在小城市里, 5年内观察到8个或更多病例的概率是多少?



(3) 泊松分布 (Poisson distribution)

- 如何检验是泊松分布?
 - 发生的事件是独立事件
 - 在任何相同的时间范围内, 某事件发生的概率相同
 - 求**某个时间范围内, 发生某事件k次的概率**, 服从泊松分布
(PS: 当二项分布的n很大而p很小时, 泊松分布可作为二项分布的近似, 其中 λ 为np。通常当 $n \geq 20$, $p \leq 0.05$ 时, 就可以用泊松公式近似得计算。事实上, 泊松分布正是由二项分布推导而来的)
- 期望: λ
 - λ , 其中, 参数 λ 是单位时间(或单位面积)内随机事件的平均发生率。
泊松分布适合于描述单位时间内随机事件发生的次数。

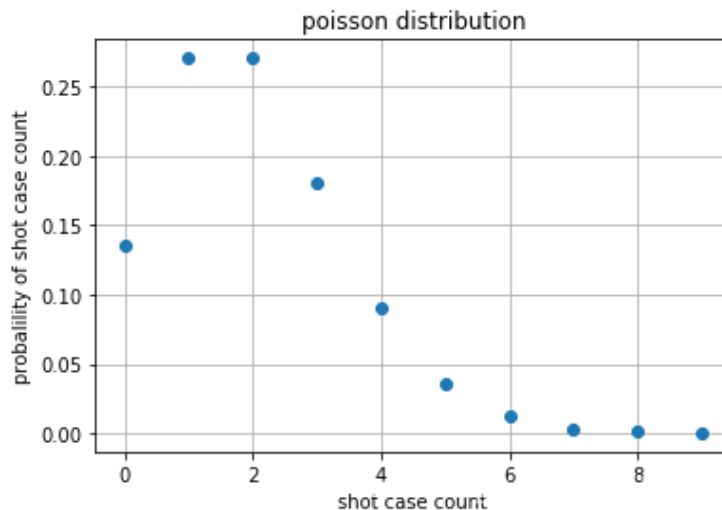


(3) 泊松分布 (Poisson distribution)

- 如何检验是泊松分布？
 - 发生的事件是独立事件
 - 在任何相同的时间范围内，某事件发生的概率相同
 - 求**某个时间范围内，发生某事件k次的概率**，服从泊松分布
(PS: 当二项分布的n很大而p很小时，泊松分布可作为二项分布的近似，其中 λ 为np。通常当 $n \geq 20$, $p \leq 0.05$ 时，就可以用泊松公式近似得计算。事实上，泊松分布正是由二项分布推导而来的)
- 期望： λ
 - λ ，其中，参数 λ 是单位时间(或单位面积)内随机事件的平均发生率。
泊松分布适合于描述单位时间内随机事件发生的次数。

例：假设某地区，一年中发生枪击案的平均次数为2

一年内的枪击案发生次数的分布如上所示。可以看到1次和2次的枪击案发生概率最高

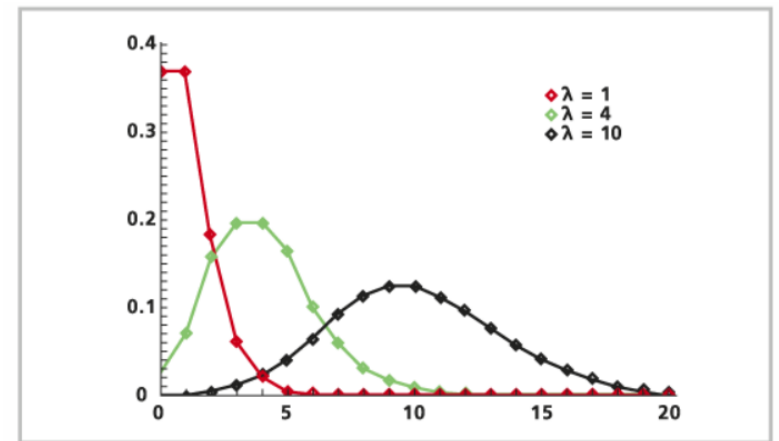
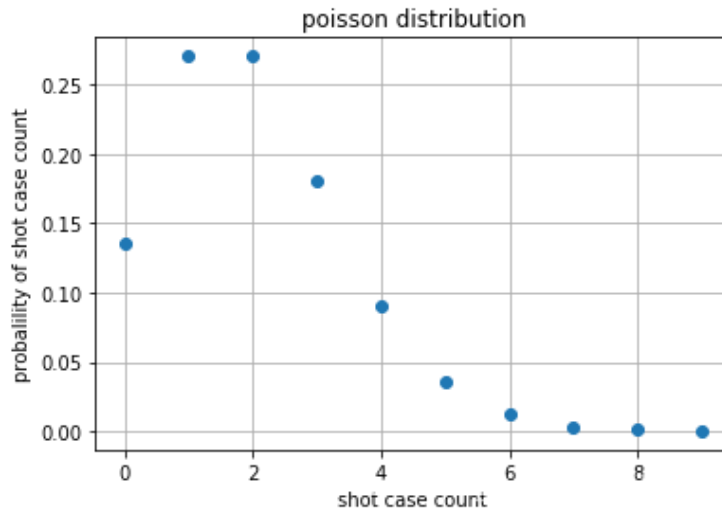


(3) 泊松分布 (Poisson distribution)

- 如何检验是泊松分布？
 - 发生的事件是独立事件
 - 在任何相同的时间范围内，某事件发生的概率相同
 - 求**某个时间范围内，发生某事件k次的概率**，服从泊松分布
(PS: 当二项分布的n很大而p很小时，泊松分布可作为二项分布的近似，其中 λ 为np。通常当 $n \geq 20$, $p \leq 0.05$ 时，就可以用泊松公式近似得计算。事实上，泊松分布正是由二项分布推导而来的)
- 期望： λ
 - λ ，其中，参数 λ 是单位时间(或单位面积)内随机事件的平均发生率。
泊松分布适合于描述单位时间内随机事件发生的次数。

例：假设某地区，一年中发生枪击案的平均次数为2

一年内的枪击案发生次数的分布如上所示。可以看到1次和2次的枪击案发生概率最高



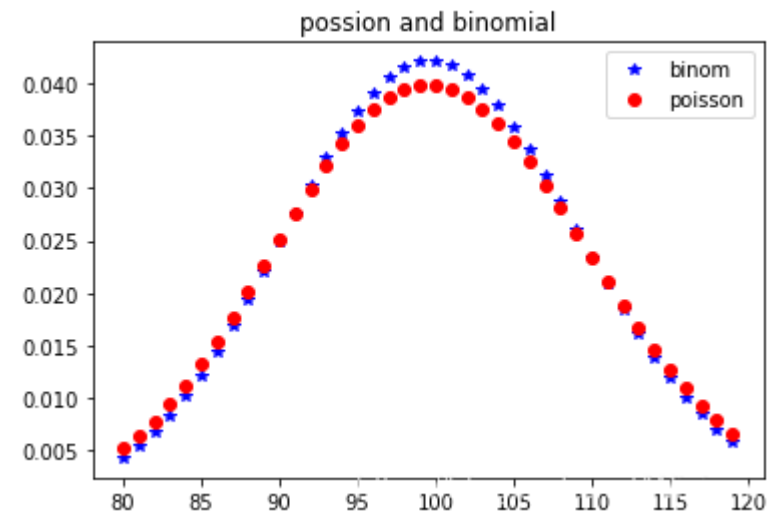
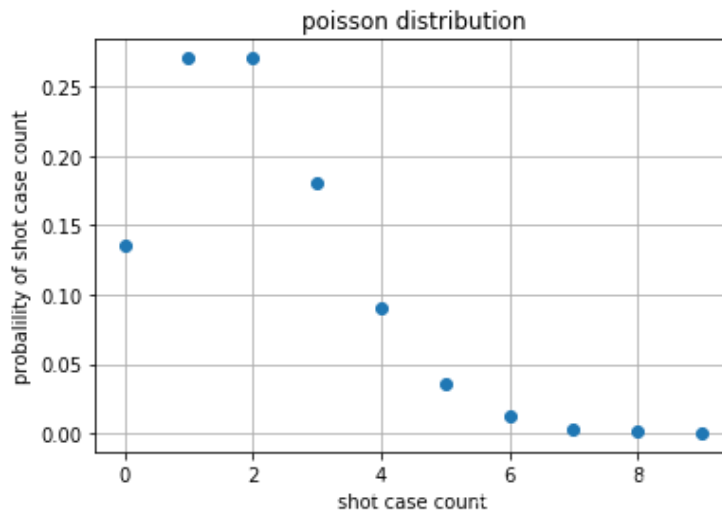
泊松分布

(3) 泊松分布 (Poisson distribution)

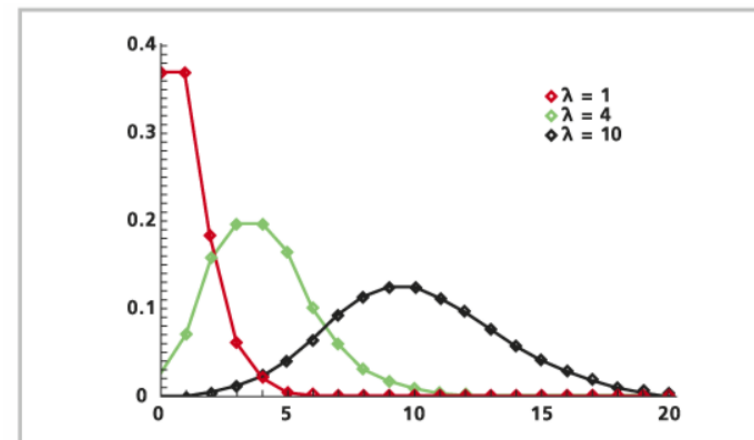
- 如何检验是泊松分布？
 - 发生的事件是独立事件
 - 在任何相同的时间范围内，某事件发生的概率相同
 - 求**某个时间范围内，发生某事件k次的概率**，服从泊松分布
(PS: 当二项分布的 n 很大而 p 很小时，泊松分布可作为二项分布的近似，其中 λ 为 np 。通常当 $n \geq 20$, $p \leq 0.05$ 时，就可以用泊松公式近似得计算。事实上，泊松分布正是由二项分布推导而来的)
- 期望： λ
 - λ ，其中，参数 λ 是单位时间(或单位面积)内随机事件的平均发生率。
泊松分布适合于描述单位时间内随机事件发生的次数。

例：假设某地区，一年中发生枪击案的平均次数为2

一年内的枪击案发生次数的分布如上所示。可以看到1次和2次的枪击案发生概率最高



可以看到这里当 $n=1000$, $p=0.1$ 时, $\lambda=100$, 泊松分布和二项分布已经很接近了



泊松分布



今日摘要

- Random Variables (随机变量)
- The Bernoulli trials
- The Binomial distribution
- The Poisson distribution
- **The Normal distribution**

- **伯努利分布**
- **二项分布**
- **泊松分布**



- **正态分布**



Continuous random variable 连续随机变量

A discrete random variable takes on a finite number of values.

离散随机变量

- Number of heads in a set of coin tosses
- Number of people who have had chicken pox in a random sample

A continuous random variable can take on any real value in an interval.

连续随机变量

- Height in a population
- Blood pressure in a population

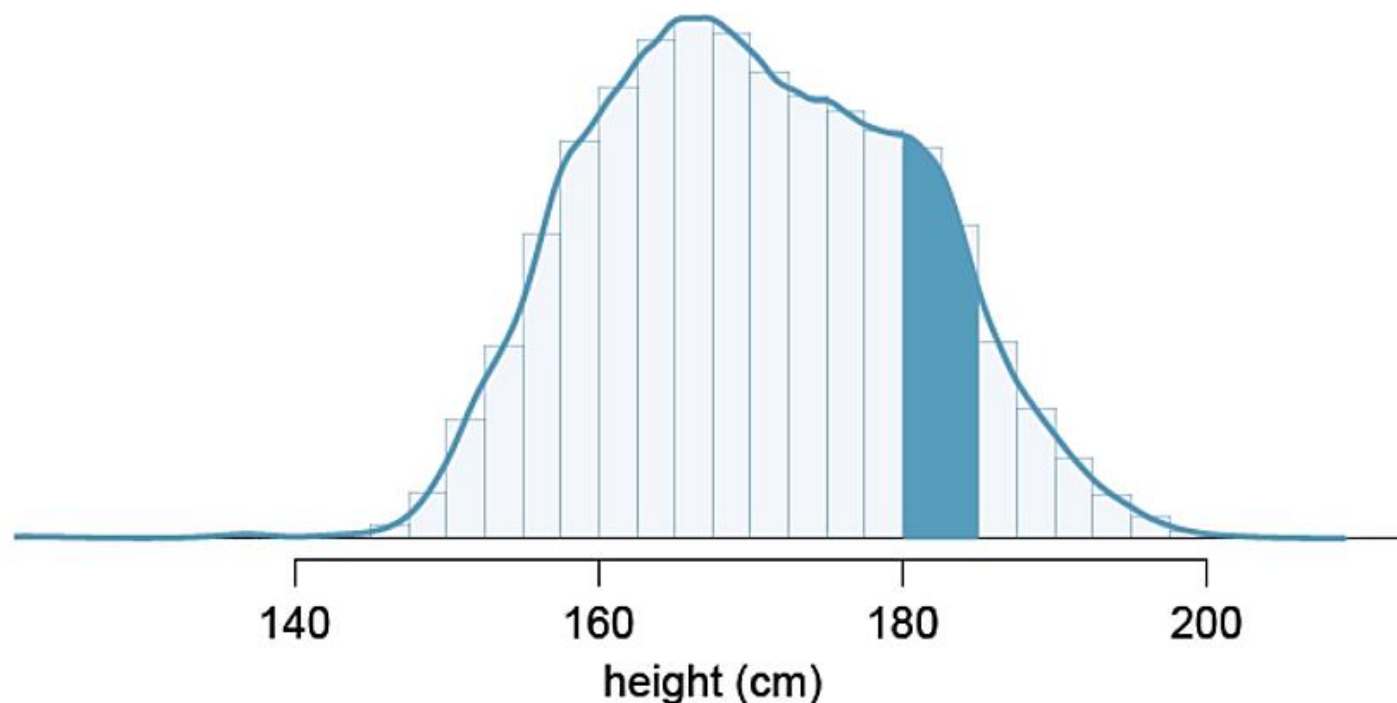
A general distinction to keep in mind: discrete random variables are *counted*, but continuous random variables are *measured*.



Probabilities for continuous distributions 连续分布的概率

Two important features of continuous distributions:

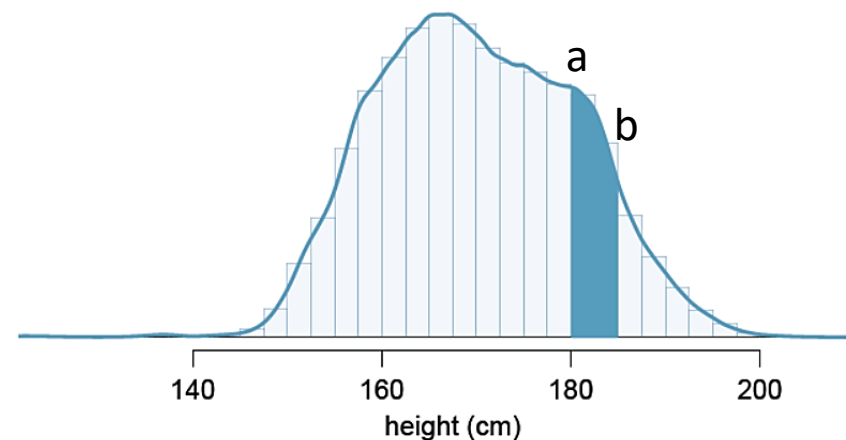
- The total area under the density curve is 1.
- The probability that a variable has a value within a specified interval is the area under the curve over that interval.



Probabilities for continuous distributions 连续分布的概率

When working with continuous random variables, the probability is found for intervals of values rather than individual values. 当分析连续随机变量时, 概率是针对值的**区间而不是单个值**

- The probability that a continuous r.v. X takes on any single individual value is 0. That is, $P(X = x) = 0$.
- Thus, $P(a < X < b)$ is equivalent to $P(a \leq X \leq b)$.



正态分布 (Normal distribution)

- 正态分布，也称“常态分布”，又称高斯分布
- 正态分布是生活中最常见的一种数据分布形态，随机变量可以是产品的质量数值、智商、人的身高和体重等
- 正态分布的概率计算
 - 确定分布和概率范围
 - 使其标准化，求标准分 Z
 - 查 Z 表格，获取概率
- 期望： μ ，方差： σ^2
记作， $N(\mu, \sigma^2)$ ，**标准正态分布**记作， $N(0, 1)$

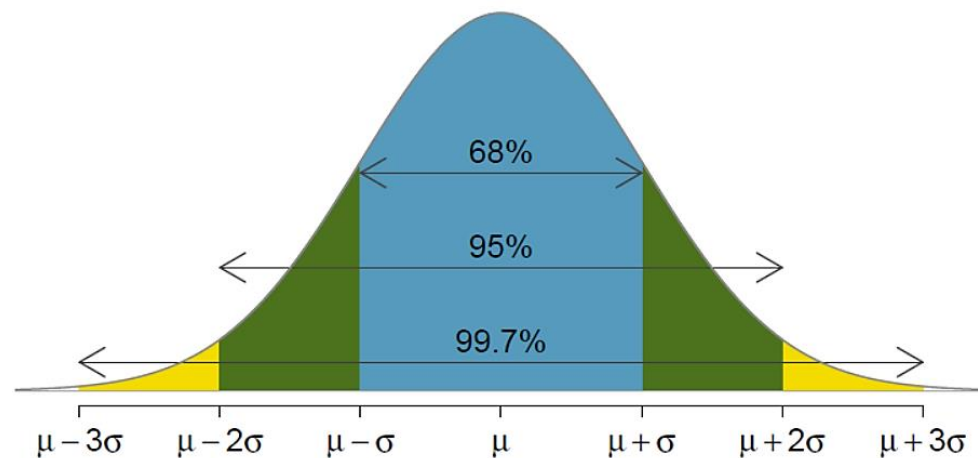
A *standard normal* distribution is defined as a normal distribution with mean 0 and variance 1. It is often denoted as $Z \sim N(0, 1)$.

Any normal random variable X can be transformed into a standard normal random variable Z .

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

According to the Empirical Rule, for any normal distribution,

- approximately 68% of the data are within 1 SD of the mean
- approximately 95% of the data are within 2 SDs of the mean
- approximately 99.7% of the data are within 3 SDs of the mean



Calculating normal probabilities

The normal distribution is a continuous probability distribution.

the total area under the density curve is always equal to 1, and the probability that a variable has a value within a specified interval is the area under the curve over that interval. By using either statistical software or normal probability tables, the normal model can be used to identify a probability or percentile based on the corresponding Z-score (and vice versa).



Figure 3.9: The area to the left of Z represents the percentile of the observation.

| Z | Second decimal place of Z | | | | | | | | | |
|-----|---------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| | 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 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

Figure 3.10: A section of the normal probability table. The percentile for a normal random variable with $Z = 0.43$ has been *highlighted*, and the percentile closest to 0.8000 has also been *highlighted*.



Normal probability calculation examples

- **正态分布主要有两类相关问题：**
 1. Calculating probabilities from a given value (whether X or Z)
 2. Identifying the observation that corresponds to a particular probability



Normal probability calculation examples

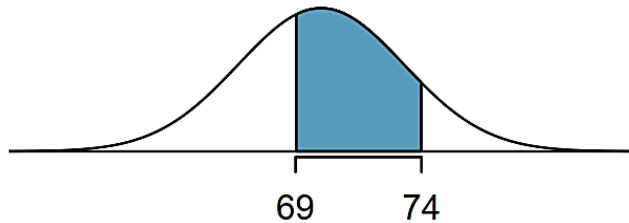
- 正态分布主要有两类相关问题：

1. Calculating probabilities from a given value (whether X or Z)

EXAMPLE 3.25

The height of adult males in the United States between the ages of 20 and 62 is nearly normal, with mean 70 inches and standard deviation 3.3 inches.¹⁶ What is the probability that a random adult male is between 5'9" and 6'2"?

These heights correspond to 69 inches and 74 inches. First, draw the figure. The area of interest is an interval, rather than a tail area.

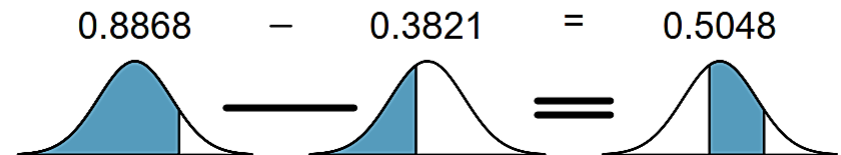


To find the middle area, find the area to the left of 74; from that area, subtract the area to the left of 69.

First, convert to Z-scores:

$$Z_{74} = \frac{x - \mu}{\sigma} = \frac{74 - 70}{3.3} = 1.21, \quad Z_{62} = \frac{x - \mu}{\sigma} = \frac{69 - 70}{3.3} = -0.30.$$

From the normal probability table, the areas are respectively, 0.8868 and 0.3821. The middle area is $0.8868 - 0.3821 = 0.5048$. The probability of being between heights 5'9" and 6'2" is 0.5048.



Normal probability calculation examples

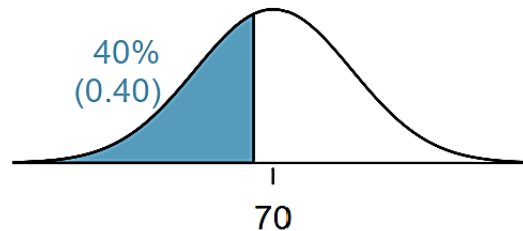
- 正态分布主要有两类相关问题：

1. Calculating probabilities from a given value (whether X or Z)
2. Identifying the observation that corresponds to a particular probability

EXAMPLE 3.27

How tall is a man with height in the 40th percentile?

First, draw a picture. The lower tail probability is 0.40, so the shaded area must start before the mean.



Determine the Z-score associated with the 40th percentile. Because the percentile is below 50%, Z will be negative. Look for the probability inside the negative part of table that is closest to 0.40: 0.40 falls in row -0.2 and between columns 0.05 and 0.06. Since it falls closer to 0.05, choose $Z = -0.25$.

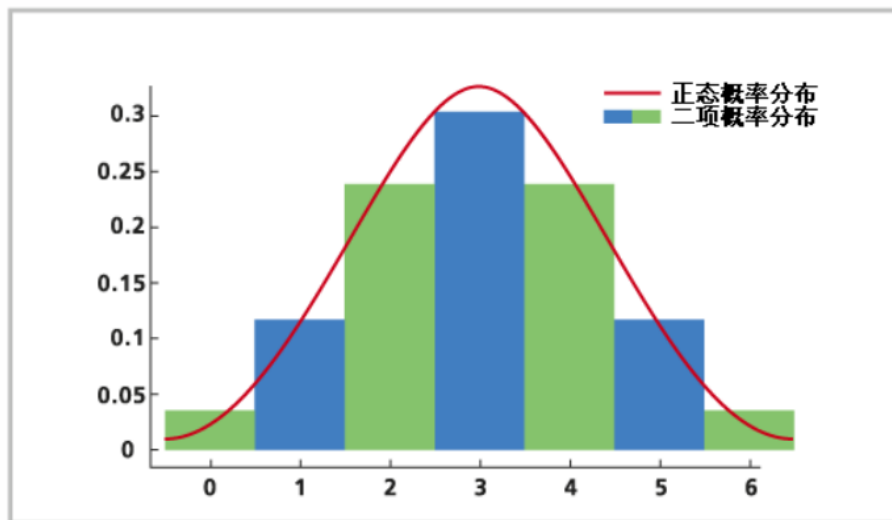
Convert the Z-score to X, where $X \sim N(70, 3.3)$.

$$X = \mu + \sigma Z = 70 + (-0.25)(3.3) = 69.18.$$

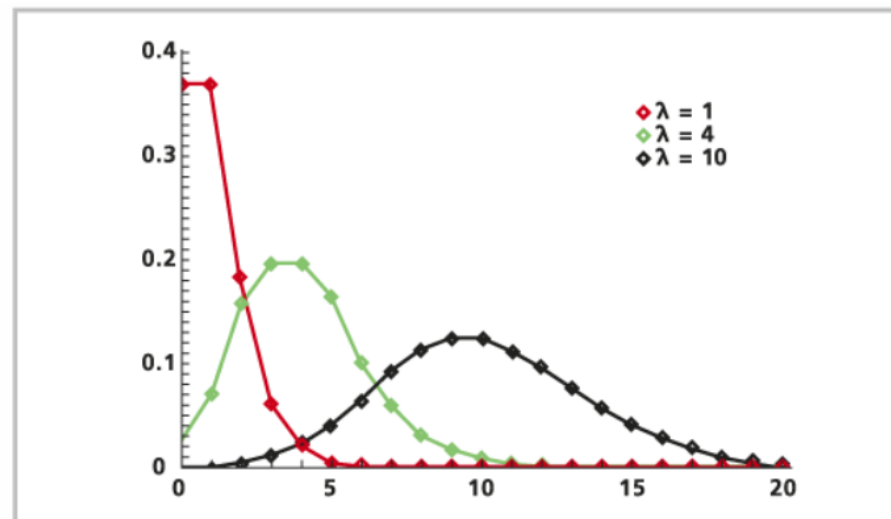
A man with height in the 40th percentile is 69.18 inches tall, or about 5' 9".

| Second decimal place of Z | | | | | | | | | | Z |
|---------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|------|
| 0.09 | 0.08 | 0.07 | 0.06 | 0.05 | 0.04 | 0.03 | 0.02 | 0.01 | 0.00 | |
| 0.0002 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | -3.4 |
| 0.0003 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0005 | 0.0005 | 0.0005 | -3.3 |
| 0.0005 | 0.0005 | 0.0005 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0007 | 0.0007 | -3.2 |
| 0.0007 | 0.0007 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0009 | 0.0009 | 0.0009 | 0.0010 | -3.1 |
| 0.0010 | 0.0010 | 0.0011 | 0.0011 | 0.0011 | 0.0012 | 0.0012 | 0.0013 | 0.0013 | 0.0013 | -3.0 |
| 0.0014 | 0.0014 | 0.0015 | 0.0015 | 0.0016 | 0.0016 | 0.0017 | 0.0018 | 0.0018 | 0.0019 | -2.9 |
| 0.0019 | 0.0020 | 0.0021 | 0.0021 | 0.0022 | 0.0023 | 0.0023 | 0.0024 | 0.0025 | 0.0026 | -2.8 |
| 0.0026 | 0.0027 | 0.0028 | 0.0029 | 0.0030 | 0.0031 | 0.0032 | 0.0033 | 0.0034 | 0.0035 | -2.7 |
| 0.0036 | 0.0037 | 0.0038 | 0.0039 | 0.0040 | 0.0041 | 0.0043 | 0.0044 | 0.0045 | 0.0047 | -2.6 |
| 0.0048 | 0.0049 | 0.0051 | 0.0052 | 0.0054 | 0.0055 | 0.0057 | 0.0059 | 0.0060 | 0.0062 | -2.5 |
| 0.0064 | 0.0066 | 0.0068 | 0.0069 | 0.0071 | 0.0073 | 0.0075 | 0.0078 | 0.0080 | 0.0082 | -2.4 |
| 0.0084 | 0.0087 | 0.0089 | 0.0091 | 0.0094 | 0.0096 | 0.0099 | 0.0102 | 0.0104 | 0.0107 | -2.3 |
| 0.0110 | 0.0113 | 0.0116 | 0.0119 | 0.0122 | 0.0125 | 0.0129 | 0.0132 | 0.0136 | 0.0139 | -2.2 |
| 0.0143 | 0.0146 | 0.0150 | 0.0154 | 0.0158 | 0.0162 | 0.0166 | 0.0170 | 0.0174 | 0.0179 | -2.1 |
| 0.0183 | 0.0188 | 0.0192 | 0.0197 | 0.0202 | 0.0207 | 0.0212 | 0.0217 | 0.0222 | 0.0228 | -2.0 |
| 0.0233 | 0.0239 | 0.0244 | 0.0250 | 0.0256 | 0.0262 | 0.0268 | 0.0274 | 0.0281 | 0.0287 | -1.9 |
| 0.0294 | 0.0301 | 0.0307 | 0.0314 | 0.0322 | 0.0329 | 0.0336 | 0.0344 | 0.0351 | 0.0359 | -1.8 |
| 0.0367 | 0.0375 | 0.0384 | 0.0392 | 0.0401 | 0.0409 | 0.0418 | 0.0427 | 0.0436 | 0.0446 | -1.7 |
| 0.0455 | 0.0465 | 0.0475 | 0.0485 | 0.0495 | 0.0505 | 0.0516 | 0.0526 | 0.0537 | 0.0548 | -1.6 |
| 0.0559 | 0.0571 | 0.0582 | 0.0594 | 0.0606 | 0.0618 | 0.0630 | 0.0643 | 0.0655 | 0.0668 | -1.5 |
| 0.0681 | 0.0694 | 0.0708 | 0.0721 | 0.0735 | 0.0749 | 0.0764 | 0.0778 | 0.0793 | 0.0808 | -1.4 |
| 0.0823 | 0.0838 | 0.0853 | 0.0869 | 0.0885 | 0.0901 | 0.0918 | 0.0934 | 0.0951 | 0.0968 | -1.3 |
| 0.0985 | 0.1003 | 0.1020 | 0.1038 | 0.1056 | 0.1075 | 0.1093 | 0.1112 | 0.1131 | 0.1151 | -1.2 |
| 0.1170 | 0.1190 | 0.1210 | 0.1230 | 0.1251 | 0.1271 | 0.1292 | 0.1314 | 0.1335 | 0.1357 | -1.1 |
| 0.1379 | 0.1401 | 0.1423 | 0.1446 | 0.1469 | 0.1492 | 0.1515 | 0.1539 | 0.1562 | 0.1587 | -1.0 |
| 0.1611 | 0.1635 | 0.1660 | 0.1685 | 0.1711 | 0.1736 | 0.1762 | 0.1788 | 0.1814 | 0.1841 | -0.9 |
| 0.1867 | 0.1894 | 0.1922 | 0.1949 | 0.1977 | 0.2005 | 0.2033 | 0.2061 | 0.2090 | 0.2119 | -0.8 |
| 0.2148 | 0.2177 | 0.2206 | 0.2236 | 0.2266 | 0.2296 | 0.2327 | 0.2358 | 0.2389 | 0.2420 | -0.7 |
| 0.2451 | 0.2483 | 0.2514 | 0.2546 | 0.2578 | 0.2611 | 0.2643 | 0.2676 | 0.2709 | 0.2743 | -0.6 |
| 0.2776 | 0.2810 | 0.2843 | 0.2877 | 0.2912 | 0.2946 | 0.2981 | 0.3015 | 0.3050 | 0.3085 | -0.5 |
| 0.3121 | 0.3156 | 0.3192 | 0.3228 | 0.3264 | 0.3300 | 0.3336 | 0.3372 | 0.3409 | 0.3446 | -0.4 |
| 0.3483 | 0.3520 | 0.3557 | 0.3594 | 0.3632 | 0.3669 | 0.3707 | 0.3745 | 0.3783 | 0.3821 | -0.3 |
| 0.3859 | 0.3897 | 0.3936 | 0.3974 | 0.4013 | 0.4052 | 0.4090 | 0.4129 | 0.4168 | 0.4207 | -0.2 |
| 0.4247 | 0.4286 | 0.4325 | 0.4364 | 0.4404 | 0.4443 | 0.4483 | 0.4522 | 0.4562 | 0.4602 | -0.1 |

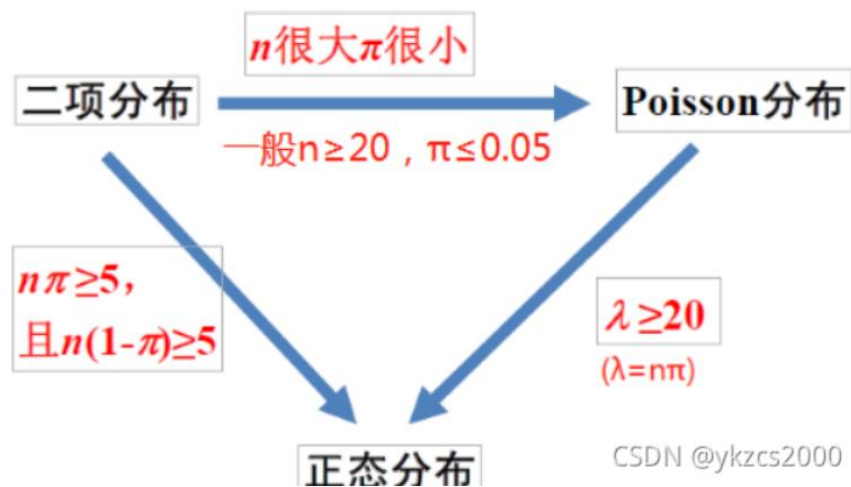
常见概率分布的相关性



二项分布



泊松分布



CSDN @ykcscs2000

Summary:

1. 伯努利是扔一次硬币
2. 二项分布是多次伯努利，即扔多次硬币
3. 泊松分布是 p 很小的二项，即扔好多好多次硬币，且扔出正面概率极小
4. 正态分布是 n 很大的二项，即扔好多好多次硬币，且硬币是完全相同的

DISTRIBUTIONS SUMMARY TABLE

| | Binomial | Normal | Poisson |
|--------------------|------------------|---------------------|-----------------------|
| Parameters | n, p | μ, σ | λ |
| Possible values | $0, 1, \dots, n$ | $(-\infty, \infty)$ | $0, 1, \dots, \infty$ |
| Mean | np | μ | λ |
| Standard Deviation | $\sqrt{np(1-p)}$ | σ | $\sqrt{\lambda}$ |



numpy 小贴士

| 包 | 方法 | 说明 |
|--------------|----------|----------------|
| numpy | array | 创造一组数 |
| numpy.random | normal | 创造一组服从正态分布的定量数 |
| numpy.random | randint | 创造一组服从均匀分布的定性数 |
| numpy | mean | 计算均值 |
| numpy | median | 计算中位数 |
| scipy.stats | mode | 计算众数 |
| numpy | ptp | 计算极差 |
| numpy | var | 计算方差 |
| numpy | std | 计算标准差 |
| numpy | cov | 计算协方差 |
| numpy | corrcoef | 计算相关系数 |

谢谢，下周见！

期待的搓搓手

