社会统计学及SPSS软件应用
STATISTICS WITH SPSS

Instructor:王荣欣

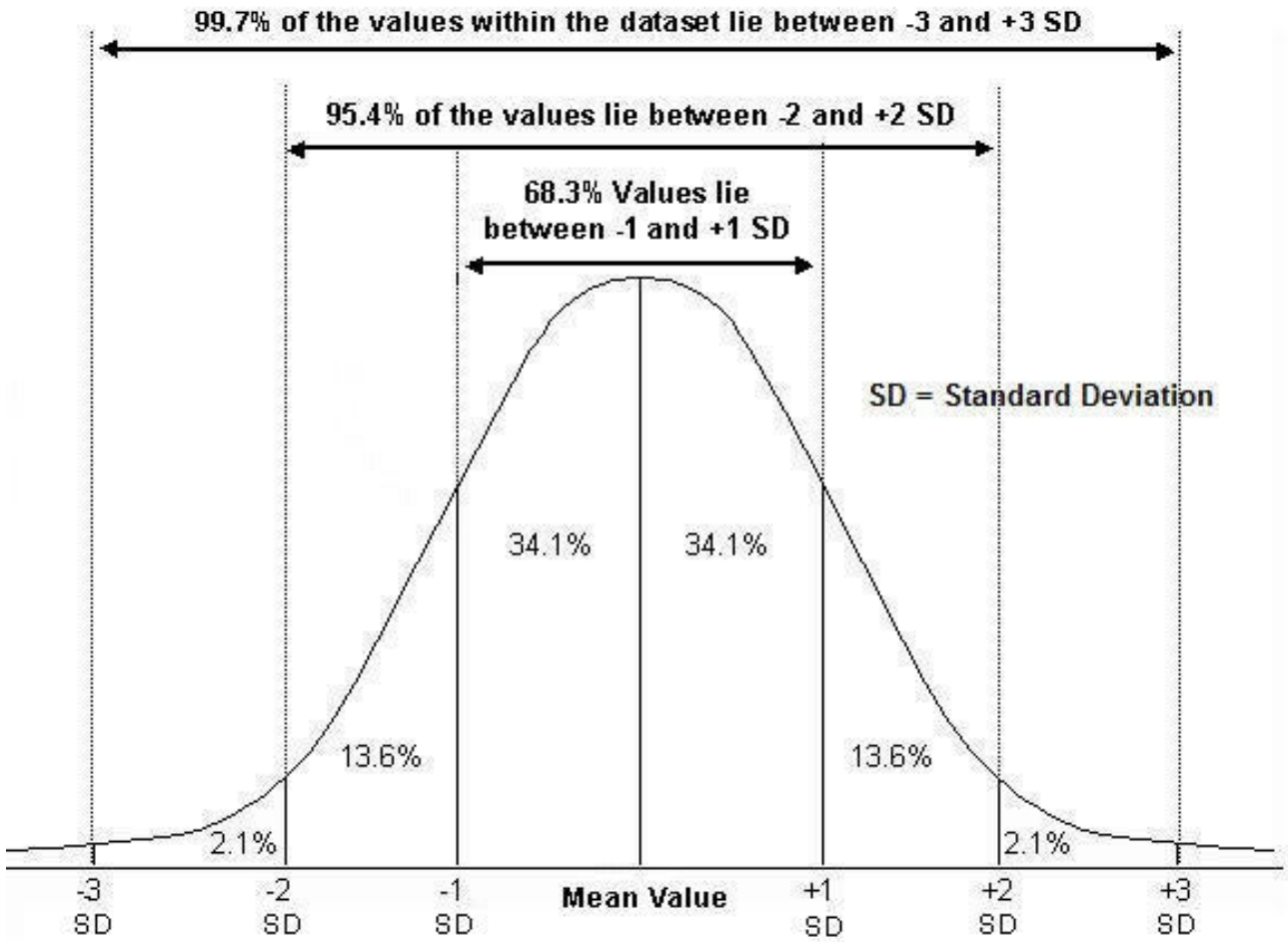
Email: rxwang@qq.com

周二3-4节、单周周四3-4节，3A106-2

2020年9月28日

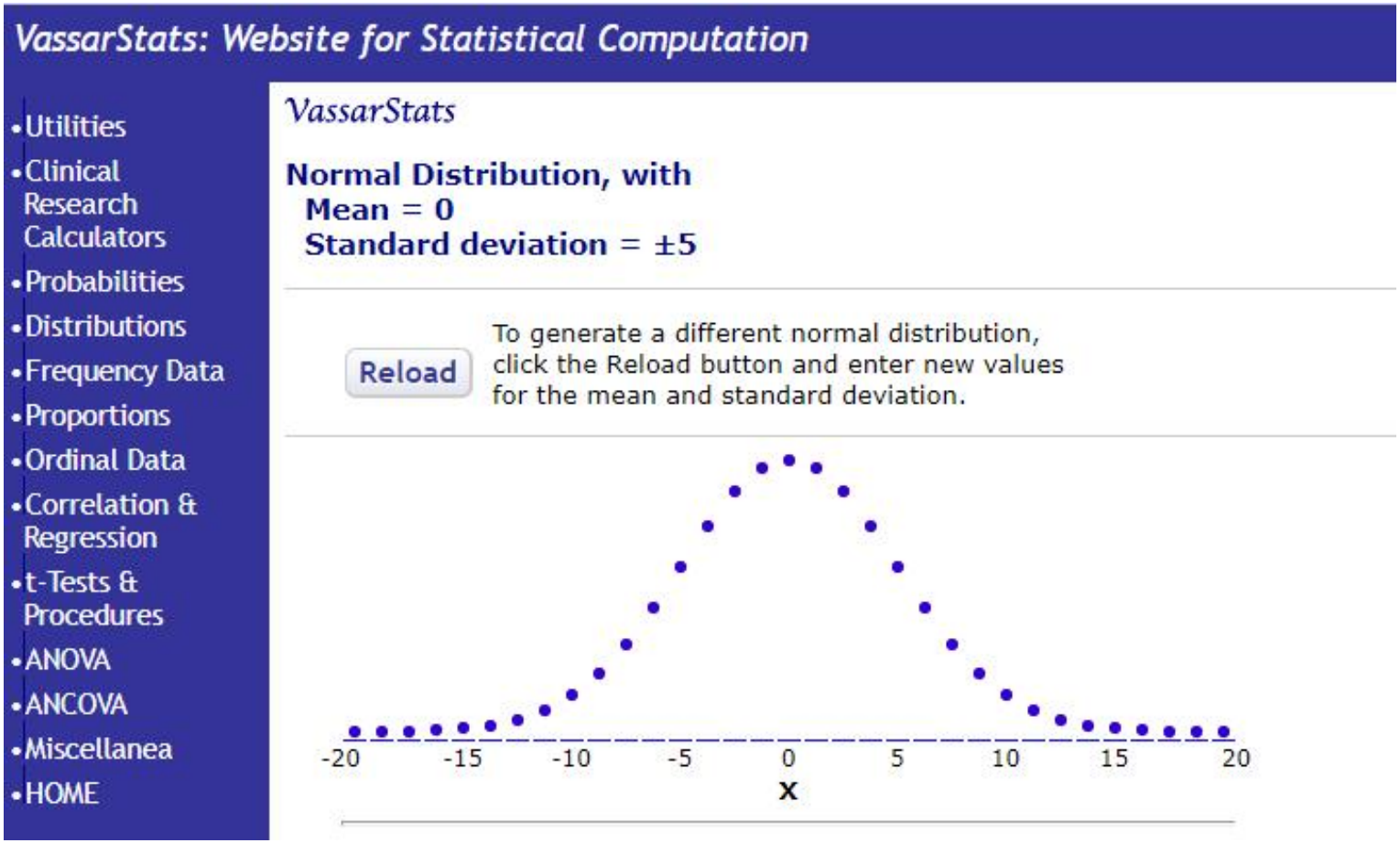
CONTENTS

- Hypothesis Testing(1)


99.7% of the values within the dataset lie between -3 and +3 SD
95.4% of the values lie between -2 and +2 SD
68.3% Values lie between -1 and +1 SD
SD = Standard Deviation
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
-3 SD
-2 SD
-1 SD
Mean Value
+1 SD
+2 SD
+3 SD

NORMAL DISTRIBUTION

1. The area to the right of the mean is 50 percent.
2. The area within 1 standard deviation of the mean (that is, $\mu - \sigma < x < \mu + \sigma$) is approximately 68 percent.
3. The area within 2 standard deviation of the mean (that is, $\mu - 2\sigma < x < \mu + 2\sigma$) is approximately 95 percent.
4. The area within 3 standard deviation of the mean (that is, $\mu - 3\sigma < x < \mu + 3\sigma$) is approximately 99 percent.



http://vassarstats.net/

例子：标准化

Example 2.1

你参加了两次考试，第一次考了60分，第二次考了85分；第一次考试全班的平均分是30分，标准差15分，第二次考试全班的平均分是70分，标准差15分。

请问，哪一次成绩较好？

（提示：计算标准分，也就是对数据进行“标准化”）

例子：标准化

Example 2.2

期末考试中，小明语文成绩*76*分，数学成绩*84*分；全班语文平均分*70*分，标准差*8*分，数学平均分*77*分，标准差*10*分。

请问，小明哪一科成绩较好？

（提示：计算标准分，也就是对数据进行“标准化”，然后比较相对位置）

Standardized Random Variable

若一个随机变量X具有平均值μ和标准差$\sigma(X)$，新的变量$z = \frac{X-\mu}{\sigma(X)}$被看作随机变量的标准化形式。

标准化后的新变量变成了一个均值为0、方差为1的变量。

Example 2.3

数据：*chip.dta*

```
sum earn
gen earn_st=(earn-3687.005)/3489.139
sum earn_st, detail
```

```
. sum earn
```

Variable	Obs	Mean	Std. Dev.	Min	Max
earn	20,421	3687.005	3489.139	154.8	76300

```
. gen earn_st=(earn-3687.005)/3489.139

. sum earn
```

Variable	Obs	Mean	Std. Dev.	Min	Max
earn	20,421	3687.005	3489.139	154.8	76300

```
. sum earn_st
```

Variable	Obs	Mean	Std. Dev.	Min	Max
earn_st	20,421	-1.19e-07	1	-1.012343	20.81115

标准正态分布

- 所有可能样本平均数$\bar{y}$服从（$\mu, \frac{\sigma_Y^2}{n}$）的正态分布。
  $\bar{y} \sim N(\mu, \frac{\sigma_Y^2}{n})$
- 变换为标准正态分布，即平均数为0，方差为1的z分布。
  $z = \frac{\bar{y}-\mu}{\sigma(\bar{y})} = \frac{\bar{y}-\mu}{\frac{\sigma_Y}{\sqrt{n}}} \sim N(0,1)$

$\sigma(\bar{y})$ read "standard deviation of $\bar{y}$".

Example 2.4

估计一个培训班学生的平均年龄。总体=250人。*

比如，现在在班级随机抽取25人**，从中得到他们的平均年龄为27岁。

由此回答，班级学生的平均年龄是多少？

* 假定总体是得不到的。对总体的估计，需要通过能够代表总体的样本来得到。

**在250名学生中抽取25人，共有250^{25}种可能的样本。

SAMPLING DISTRIBUTION（定义）

1. 假设我们对总体进行重复抽样，每次用同样的公式计算样本统计量，那么从所有这些样本中得到的统计量就构成了一个分布，该分布被称为抽样分布。
2. 对于一个总体而言，**所有可能样本**的统计量数值的概率分布被称作抽样分布。
3. A sampling distribution is a mathematical description of **all possible sampling event** outcomes and the probability of each one.

统计推断的理论依据

依靠抽样分布，我们就能够将实际观测到的样本结果与其他**所有可能的样本**结果进行比较，从而建立起单一样本与总体之间的联系。

Example 2.5

样本平均数的抽样分布是：

$$\bar{y} \sim N(\mu, \frac{\sigma_Y^2}{n})$$

[等式左边是样本统计量，右边是总体参数，这就搭建了样本与总体之间的桥梁]

样本统计量（平均值）服从以总体平均数μ为中心、方差为$\frac{\sigma_Y^2}{n}$的正态分布。σ_Y^2为总体的方差。

中心极限定理

1. 所有可能样本平均数$\bar{y}$的分布是正态分布。
2. 所有可能样本平均数的平均数等于总体平均数。

$$E(\bar{y}) = \bar{\bar{y}} = \mu$$

3. 所有可能样本平均数的方差等于总体的方差除以样本规模。

标准误 STANDARD ERROR

所有可能样本的（某个变量的）平均数的标准差为 $\sigma(\bar{y}) = \frac{\sigma_Y}{\sqrt{n}}$，也称为标准误（standard error, 简称SE）。

Example 2.6

举例：计算*SE* 数据：*chip.dta*

sum age

mean age

display 9.542207/sqrt(20421)

```
. sum age
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	20,421	38.05729	9.542207	20	59

```
. mean age

Mean estimation                 Number of obs    =     20,421
```

	Mean	Std. Err.	[95% Conf.	Interval]
age	38.05729	.0667745	37.92641	38.18818

```
. display 9.542207/sqrt(20421)
.06677445
```

INFERENTIAL STATISTICS

- Hypothesis testing is a scientific procedure for making rational decisions about two different claims. （先假设总体的情况，然后进行抽样，检验原有的假设是否成立）
- Estimation theory is a branch of statistics that deals with estimating the values of parameters. （先看样本情况，再问总体的情况）

参数估计是利用样本信息来推断未知的总体参数。
假设检验则是先对总体参数提出一个假设，然后利用样本信息来判断这一假设是否成立。[1]

INFERENTIAL STATISTICS

- Hypothesis testing is a scientific procedure for making rational decisions about two different claims. （先假设总体的情况，然后进行抽样，检验原有的假设是否成立）
- Estimation theory is a branch of statistics that deals with estimating the values of parameters. （先看样本情况，再问总体的情况）

参数估计是利用样本信息来推断未知的总体参数。
假设检验则是先对总体参数提出一个假设，然后利用样本信息来判断这一假设是否成立。[1]

INFERENTIAL STATISTICS

1. 科学发现的基本逻辑：证伪。
2. 科学家先针对特定问题提出假设或猜想，再依据事实对假设进行检验，并在检验过程中不断淘汰或修改原有的假设。
3. 区分科学和非科学的标准，在于其理论是否具有“可证伪性”。

(1) empirical science: falsify
   - reject null hypotheses

(2) formal science(mathematics, logic): prove

例子：女士品茶

Example 3.1

*20*世纪*20*年代，在英国剑桥，一群人在品茶。在品茶过程中，一位女士坚称，把茶加进奶，或把奶加进茶，会使茶的味道不同。

*Ronald Fisher*设计了一个实验：煮了*8*杯茶，有*4*杯是先加茶，有*4*杯是先加奶。让这位女士进行评断。

1935，*Fisher*在他的*The Design of Experiments*一书中，描述了“女士品茶”的实验。

HYPOTHESIS TESTING (JERZY NEYMAN AND EGON PEARSON)

Neyman提出至少要有两个可能的假设。

1. 被检验的假设称为null hypothesis。
2. 其他可能的假设为alternative hypothesis。

推荐阅读：统计之都（CapStat）

1. 系列文章《漫谈现代统计“四大天王”》

HYPOTHESIS TESTING

Example 3.2

H_0：该女士完全分不出哪些先加茶，哪些先加奶？（该女士只是猜测。）

H_1：该女士可以分出哪些先加茶，哪些先加奶。

女士品茶的排列组合

X	组合数公式	组合数结果	概率
0	$C_4^0 \times C_4^4$	1	0.0143
1	$C_4^1 \times C_4^3$	16	0.2286
2	$C_4^2 \times C_4^2$	36	0.5143
3	$C_4^3 \times C_4^1$	16	0.2286
4	$C_4^4 \times C_4^0$	1	0.0143
总数	C_8^4	70	1

检验女士是否真能品尝出茶的区别（1）

- 假设检验是在假设“待检验的假设为真”的前提下，计算观测结果发生的概率。
- 当观测结果发生的概率很低时，可以得出原假设不成立的结论。
- 对某个待检验假设，统计分析用significant（显著的）这个词来表示结果发生的概率很低。

检验女士是否真能品尝出茶的区别（2）

- 如果我们之前的假设H_0是正确的，那么出现这种结果（X=4）的概率只有0.0143。
- 如果要断定H_0是错误的，那么，这个判断犯错误的概率是0.0143。
- 这个概率很小，所以，我们有充分的理由认为“该女士可以分出哪些先加茶，哪些先加奶”。

THE BASIC OF HYPOTHESIS TESTING

- The goal is to find a decision rule about whether the null hypothesis should be ruled out based on the available data.
- When the null hypothesis is rejected, this means you decide that the corresponding alternative hypothesis is accepted.

Example 3.3

零假设（*null hypothesis*）：总体的平均年龄μ=24岁
备择假设（*alternative hypothesis*）：$\mu \neq$24岁
已知总体方差σ^2=100, 样本规模=25

THE BASIC OF HYPOTHESIS TESTING

- The goal is to find a decision rule about whether the null hypothesis should be ruled out based on the available data.
- When the null hypothesis is rejected, this means you decide that the corresponding alternative hypothesis is accepted.

Example 3.3

零假设（*null hypothesis*）：总体的平均年龄μ=24岁
备择假设（*alternative hypothesis*）：$\mu \neq$24岁
已知总体方差σ^2=100, 样本规模=25

THE BASIC OF HYPOTHESIS TESTING

- The goal is to find a decision rule about whether the null hypothesis should be ruled out based on the available data.
- When the null hypothesis is rejected, this means you decide that the corresponding alternative hypothesis is accepted.

Example 3.3

零假设（*null hypothesis*）：总体的平均年龄μ=24岁
备择假设（*alternative hypothesis*）：$\mu \neq$24岁
已知总体方差σ^2=100, 样本规模=25

DECISION RULE

- A natural strategy is to try to determine how small the sample mean must be in order to reject the hypothesis that μ is greater than or equal to 24. （前述培训班学生平均年龄的例子）
- But rather than work with $\bar{y}$, it is more convenient to work with

$$z = \frac{\bar{y} - 24}{\frac{\sigma_Y}{\sqrt{n}}}$$

- The issue of whether the sample mean is sufficiently small to reject the null hypothesis is the same as whether Z is sufficiently small.

假设检验 HYPOTHESIS TESTING

1. 在假定H_0成立的条件下，计算某个统计量的值，并确定它的概率分布。
2. 计算由样本得到的统计量的值所发生的概率，又称之为显著性水平（significance level），一般用α表示。
3. 若统计量的值所发生的概率低于我们事先设定的概率标准（如0.10、0.05和0.01），就说明统计显著，于是倾向于拒绝或否定原假设。

否定域与显著性水平（1）

- 否定域在整个抽样分布中所占的比例，叫做显著性水平α（或显著度α），代表样本的统计值落在否定域的可能性。
- The p-value is just the smallest α value for which the null hypothesis is rejected.
- Alternatively, the p-value is the probability of a Type I error if the observed value of Z is used as a critical value.

否定域与显著性水平（1）

- 否定域在整个抽样分布中所占的比例，叫做显著性水平α（或显著度α），代表样本的统计值落在否定域的可能性。
- The p-value is just the smallest α value for which the null hypothesis is rejected.
- Alternatively, the p-value is the probability of a Type I error if the observed value of Z is used as a critical value.

否定域与显著性水平（2）

- A critical value is the value used to determine whether the null hypothesis should be rejected.
- If you want the probability of a Type I error to be α=0.025, the critical value is -1.96.
- If α=0.005, then the critical value is -2.58. [2]

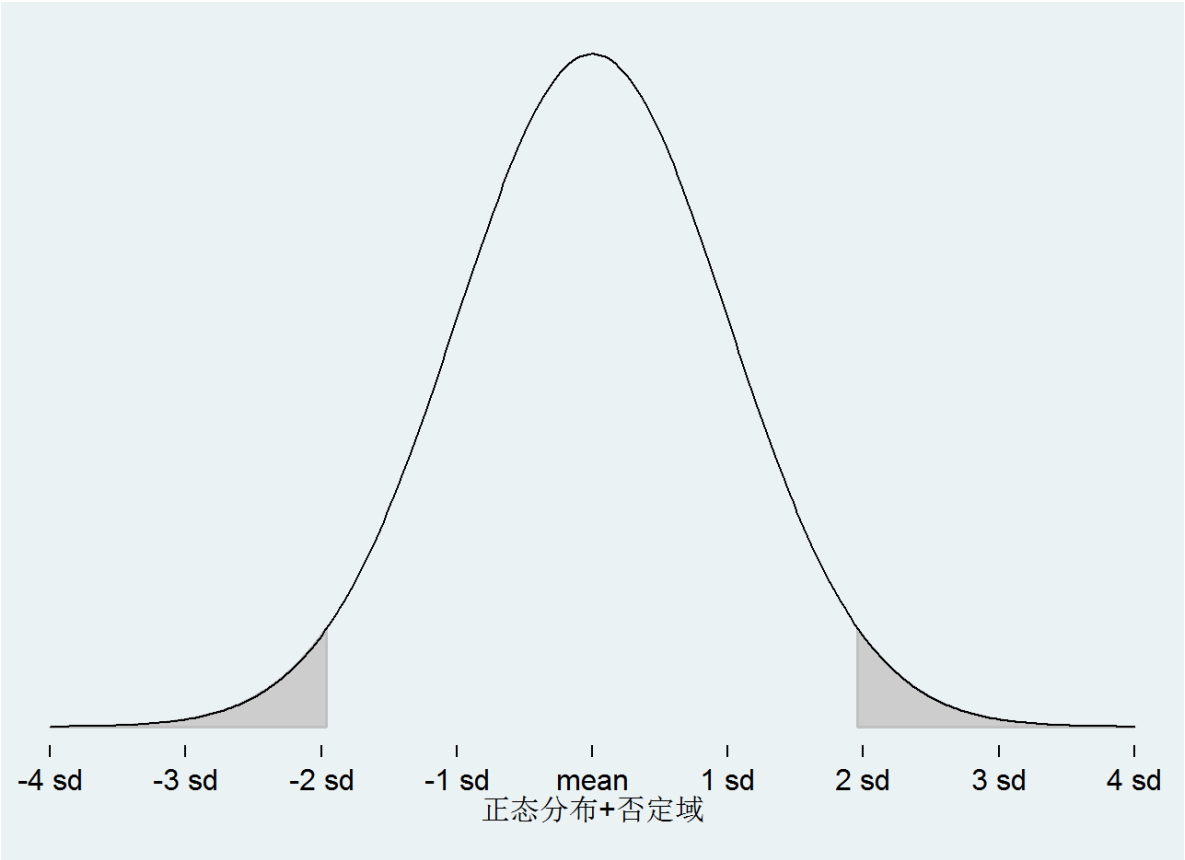


Figure 3.1: 否定域

否定域与显著性水平（3）

1. Decision rule 1: Reject the null hypothesis H_o: $\mu \geq \mu_0$ if Z $\leq c$, where c is the α quantile of a standard normal distribution.
2. Decision rule 2: Reject the null hypothesis H_o: $\mu \leq \mu_0$ if Z $\geq c$
3. Decision rule 3: Reject the null hypothesis H_o: $\mu = \mu_0$ if Z $\leq -c$ or $Z \geq c$
   - eg: Compare
     $z = \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}}$, if Z is outside the range of -1.96 and 1.96, the hypothesis is rejected.

Example 3.4

零假设（*null hypothesis*）：总体的平均年龄μ =*24*岁
已知总体方差σ^2=*100*, 样本规模=*25*, 样本均值$\bar{y}$=*26*岁

$$Z = \frac{26-24}{\frac{10}{\sqrt{25}}} = 1$$

$$P(Z \leq -1) + P(Z \geq 1) = 0.1587 + 0.1587 = 0.3174$$
$$P = 2(1 - P(Z \leq 1)) = 2 \times (1 - 0.8413) = 0.3174$$

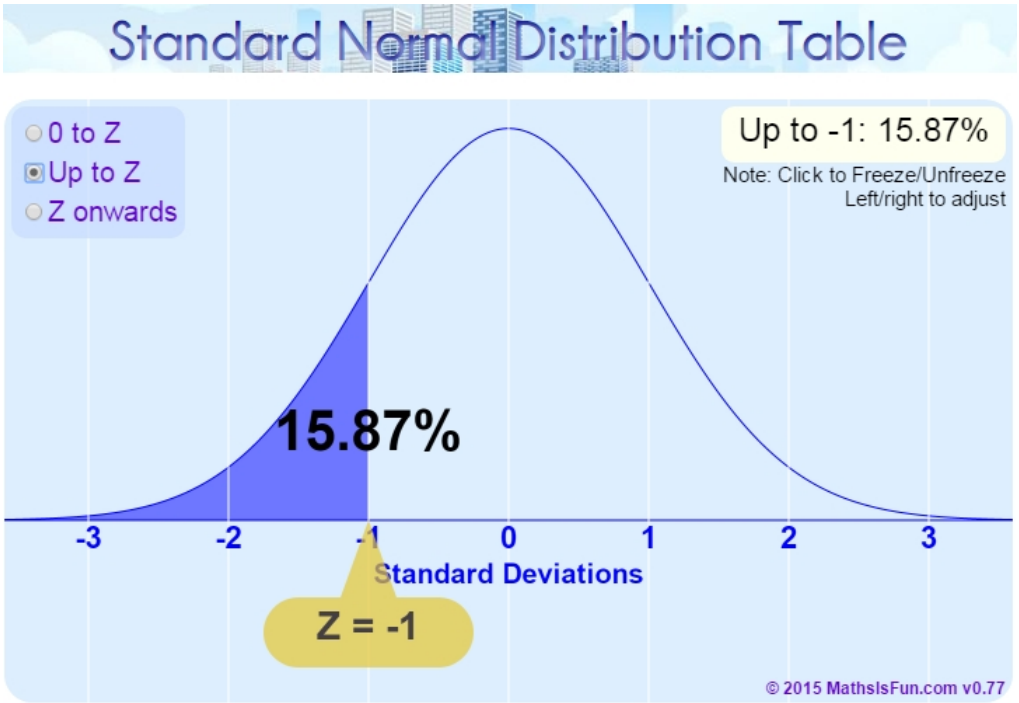


Figure 3.2: 标准正态分布

https://www.mathsisfun.com/data/standard-normal-distribution-table.html

Example 3.5

零假设：总体的平均年龄μ =24岁
已知总体方差σ^2=100, 样本规模=25, 样本均值$\bar{y}$=28岁

$$Z = \frac{28-24}{\frac{10}{\sqrt{25}}} = 2$$

$$P(Z \leq -2) + P(Z \geq 2) = 0.0228 + 0.0228 = 0.0456$$
$$P = 2(1 - P(Z \leq 2)) = 2 \times (1 - 0.9772) = 0.0456$$

So we can reject hypothesis and say μ (of age) is significant different from 24.

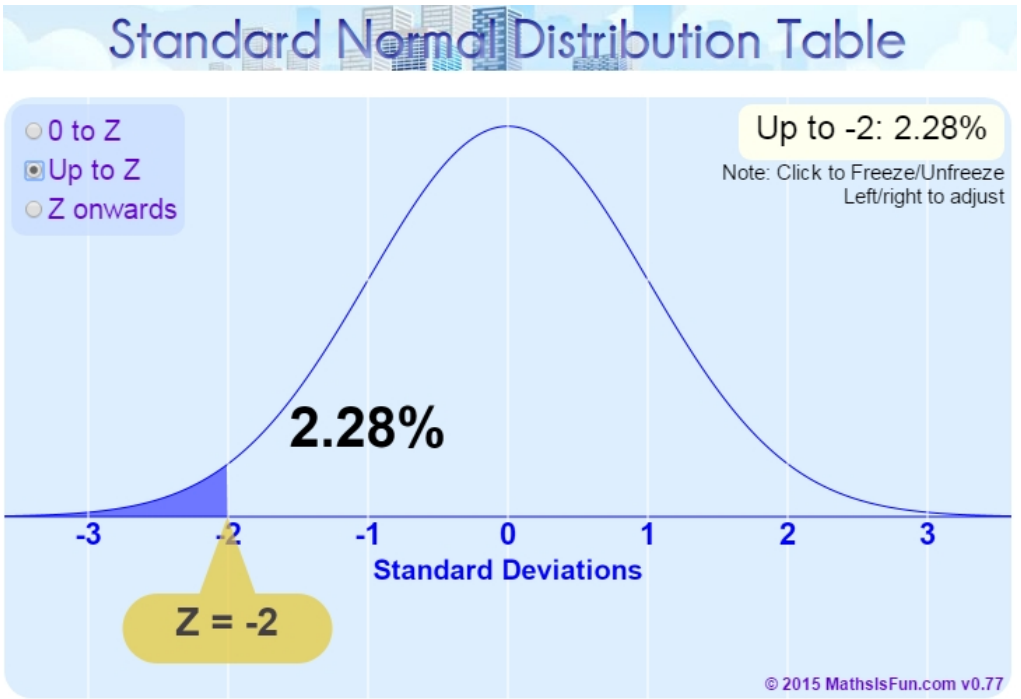


Figure 3.3: 标准正态分布

假设检验

- $z = \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

Example 3.6

H_0：总体的平均年龄μ=24岁
H_1：$\mu \neq$24岁
已知总体方差σ^2=100, 样本规模=25

如果我们的样本来自这个总体，那么它的平均数就应该在24岁附近，即落在一个标准差内的概率为68%，落在1.96个标准差内的概率为95%。
如果样本平均数小于26岁，z=1，发生的概率为68%。如果样本平均数大于28岁，z=2，即可能性不到5%。

1. z test：Z检验法，又称正态分布检验
   - Gosset introduced his z-ratio (or z-test) to determine if there was a significant difference between the sample mean and the population mean.
2. t test: 若不知道总体标准差，用样本标准差s代替总体方差，此时所有可能样本平均数的分布满足t分布，即t检验。

Testing hypothesis about the mean:

- When σ is known, we can test hypotheses about the population mean if we can determine the distribution of

$$z = \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

- When σ is not known, we estimate σ with s, and we can test hypotheses if the distribution of

$$T = \frac{\bar{y}-\mu}{\frac{s}{\sqrt{n}}}$$

can be determined.

Testing hypothesis about the mean:

- When σ is known, we can test hypotheses about the population mean if we can determine the distribution of

$$z = \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

- When σ is not known, we estimate σ with s, and we can test hypotheses if the distribution of

$$T = \frac{\bar{y}-\mu}{\frac{s}{\sqrt{n}}}$$

can be determined.