Lecture 9 – 正态分布及t-检验

- 内容大纲
  - 正态分布 Normal Distribution
  - 单样本t-检验 One sample t-test
  - 总结 Summary
  - R Lab & Discussion

生物统计学

李 勤

生态与环境科学学院

1. 回顾 – 基于列联表的独立性检验

- 类型变量的相关假设检验

一维频数数据
- 两个类别的频数
  - 二项检验
  - 卡方检验
  - → 频数或比例是否与H_0中的值一致?
- 两个或两个以上类别的频数或单位时间/空间内的频数

二维频数数据
- 两个分类变量间的独立性检验
  - 关联强度
    - 相对风险*RR*/比值比*OR*
  - 关联显著性
    - $R \times C$列联表的卡方检验 (the χ^2 contingent test)
    - 2× 2列联表的Fisher's exact test

表 1. 一个变量的检验方法 (一组数据)

Data type 数据类型	Goal	目标	Test	检验方法
Categorical 类型变量 （分类变量）	Use frequency data to test whether a population proportion equals a null hypothesized value	使用频数数据检测总体中的比例是否等于零假设中的值	Binomial test (7) χ^2 Goodness-of-fit test with two categories (use if sample size is too large for the binomial test) (8)	二项检验 (7) χ^2 拟合度检验，应用于变量分为两类（如果样本量过大，无法进行二项检验，则使用该检验） (8)
	Use frequency data to test the fit of a specific population model	使用频数数据检测特定总体模型的拟合程度	χ^2 Goodness-of-fit test (8)	χ^2 拟合度检验 (8)
Numerical 数值变量	Test whether the mean equals a null hypothesized value when data are approximately normal (possibly only after a transformation) (13)	当数据近似正态分布（或经过转换后符合）时，检测平均值是否等于零假设中的值 (13)	One-sample t-test (11)	单样本t检验 (11)
	Test whether the median equals a null hypothesized value when data are not normal (even after transformation)	当数据不符合正态分布（即使经过转换），检测中位数是否等于零假设中的值	Sign test (13)	符号检验 (13)
	Use frequency data to test the fit of a discrete probability distribution	使用频率数据测试离散概率分布的拟合程度	χ^2 Goodness-of-fit test (8)	χ^2 拟合度检验 (8)
	Use data to test the fit of the normal distribution	检测数据是否符合正态分布	Shapiro-Wilk test (13)	Shapiro-Wilk 检验 (13)

表 2. 两个变量相关性的检验方法

		Type of explanatory variable 解释变量			
		Categorical	类型变量	Numerical	数值变量
Type of response variable 响应变量	Categorical 类型变量	Contingency analysis (9)	独立性检验 (9)	Logistic regression (17)	逻辑斯蒂回归 (17)
	Numerical 数值变量	t-tests, ANOVA, Mann-Whitney U-test, etc. [See Table 3 for more details.]	t检验 方差分析 U检验等 [更多细节见表 3]	Linear and nonlinear regression (17) Linear correlation (16) Spearman's rank correlation (when data are not bivariate normal) (16)	线性和非线性回归 (17) 线性相关 (16) Spearman 秩相关（当数据不是二元正态分布时） (16)

表 3. 两个变量相关性的检验方法及其前提假设（或检验多组数据的差异）

Number of treatments 处理组数量	Tests assuming normal distribution	假设数据符合正态分布的检验	Tests not assuming normal distributions	假设数据不符合正态分布的检验
Two treatments (independent samples) 两独立样本	Welch's t-test (12) Two-sample t-test (use when variance is equal in the two groups) (12)	Welch's t检验 (12) 双样本t检验（两组方差相等时使用） (12)	Mann-Whitney U-test (Wilcoxon rank-sum test) (13)	U检验（秩和检验）(13)
Two treatments (paired data) 两配对样本	Paired t-test (12)	配对t检验 (12)	Sign test (13)	符号检验 (13)
More than two treatments 超过两组设置	ANOVA (15)	方差分析 (15)	Kruskal-Wallis test (15)	Kruskal-Wallis 检验 (15)

学习目标

- 生物学中正态分布的应用场景
- 数值变量的假设检验 t-test
  - R coding

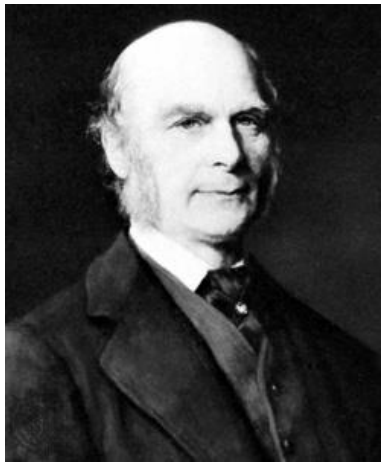
```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE,
       var.equal = FALSE, conf.level = 0.95, ...)
```

2. 正态分布 Normal distribution

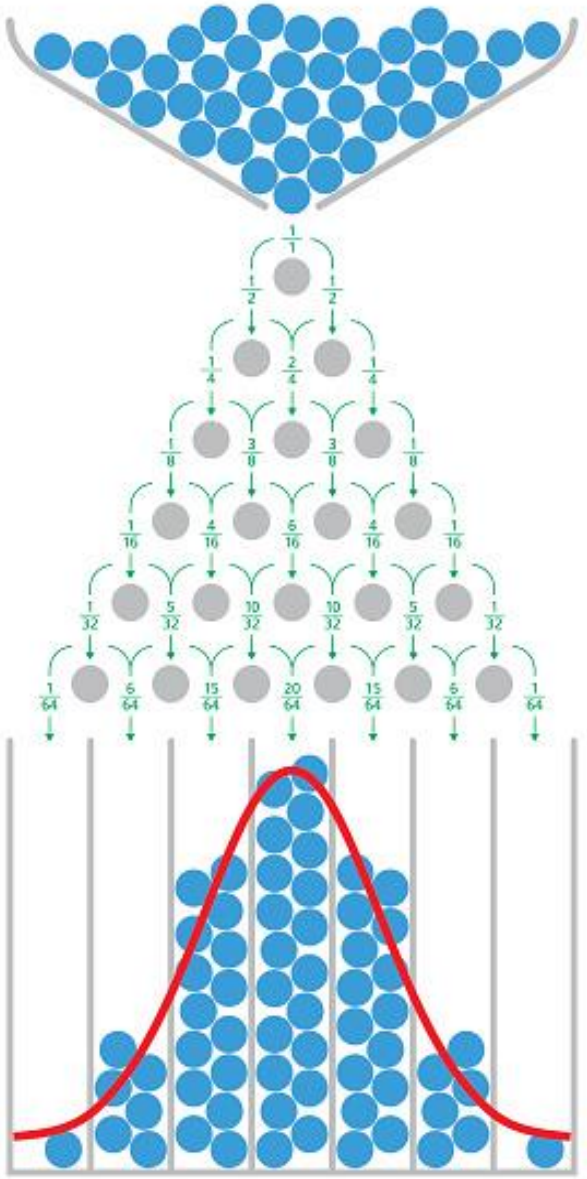
- 生物学中的数值数据，例如：
  - 婴儿的身长和体重
  - 燕子的速度
  - 从感染到出现症状的时间
  - 松树上的球果数量

2.1 正态分布 Normal distribution

- 中心极限定理 the central limit theorem
  - 是概率论 (probability theory) 一个非常重要的结论
  - 它指出在一定条件下，独立 (independent) 随机变量随样本量 (sample size) 变大会趋向正态分布 (normal distribution) [e.g., 抛硬币]
- 高尔顿版 Galton board



Sir Francis Galton



2.1 正态分布 Normal distribution

- 正态分布是一种描述钟形曲线的连续概率分布。它是许多生物变量频率分布的良好近似值。
  - 可以说，正态分布比任何其他数学函数都更能描述自然，因此在生物统计学中占据着重要地位。

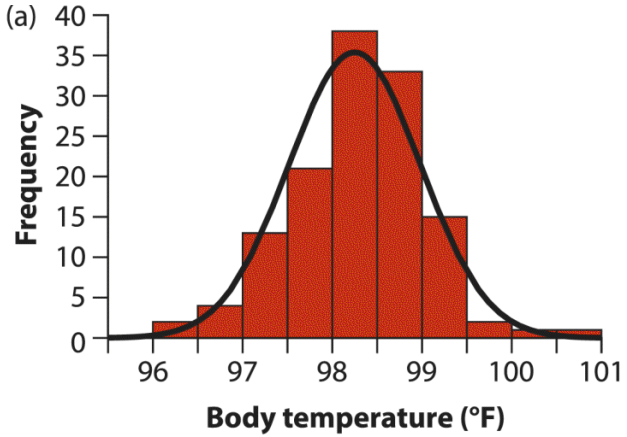
- 更重要的是，正态分布可以用来近似估计值的抽样分布，尤其是样本平均值。

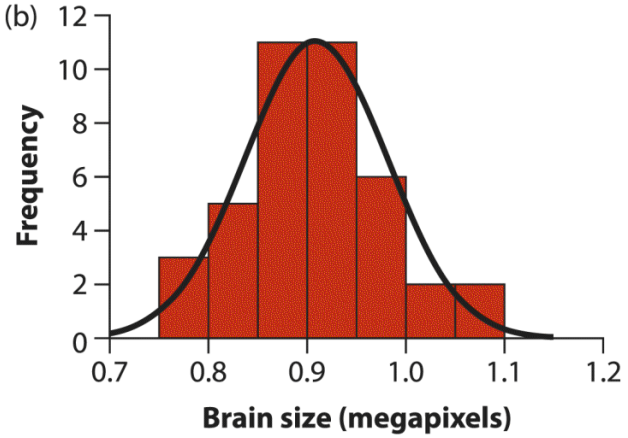
- 目前已有许多统计技术来处理具有正态抽样分布的变量。

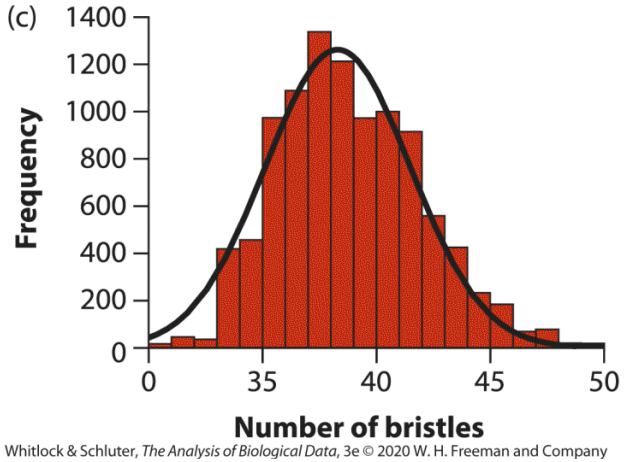
2.1 正态分布 Normal distribution

- 许多自然界中的频率分布都可以用正态分布来近似。
  - 人的体温
  - 人脑大小
  - 果蝇身体第四节和第五节上的刚毛数量

- 其它例子?

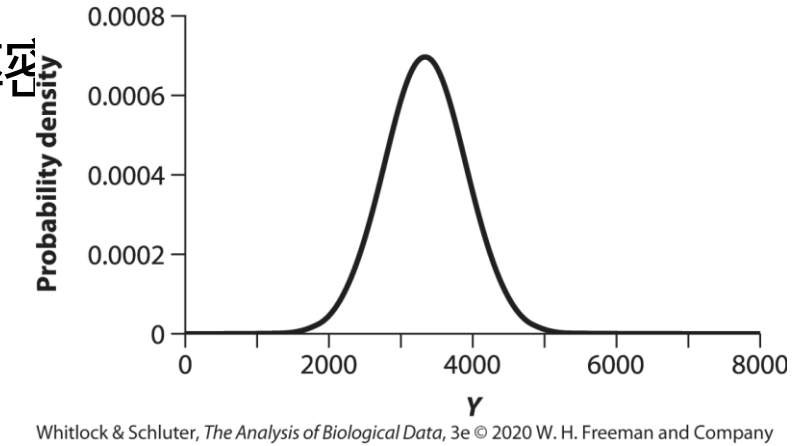






2.1 正态分布 Normal distribution

- 描述许多钟形曲线的理论概率分布即正态分布。
  - 正态分布是一种连续概率分布，这意味着它描述了连续数值变量的概率分布。
  - 它围绕均值呈对称：一个值离均值越远，其概率密
- 2个参数
  - 集中趋势（location）：均值 μ
  - 变异程度（spread）： 标准差 σ

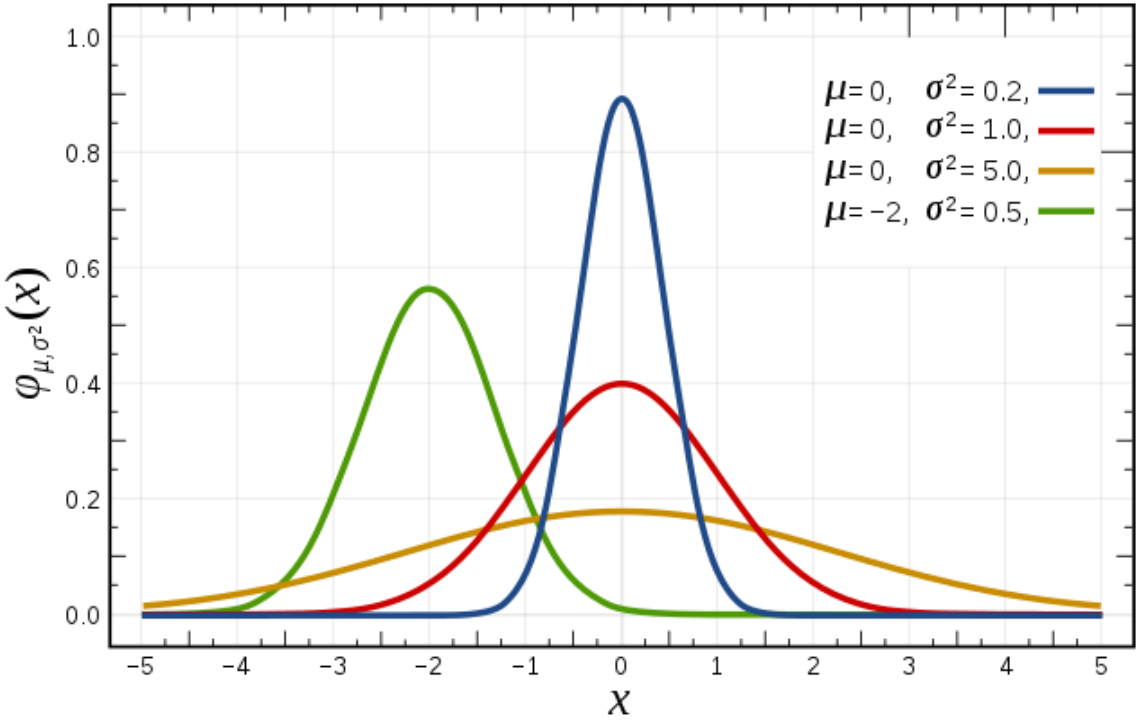


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

- 概率：曲线下某一段的面积
  - 正态分布显示的是概率密度， 而数据表示的是计数。
  - 正态分布在其整个范围内（从负无穷大到无穷大）的积分等于 1。

2.2 正态分布的一些特征

- 公式和符号
  - 正态分布：$Y \sim N(\mu, \sigma^2)$
  - 其中 μ 为分布的均值; σ 为标准差
  - Y 值对应的概率密度
    - the probability density $f(Y)$
    - $f(Y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{Y-\mu}{\sigma})^2}$

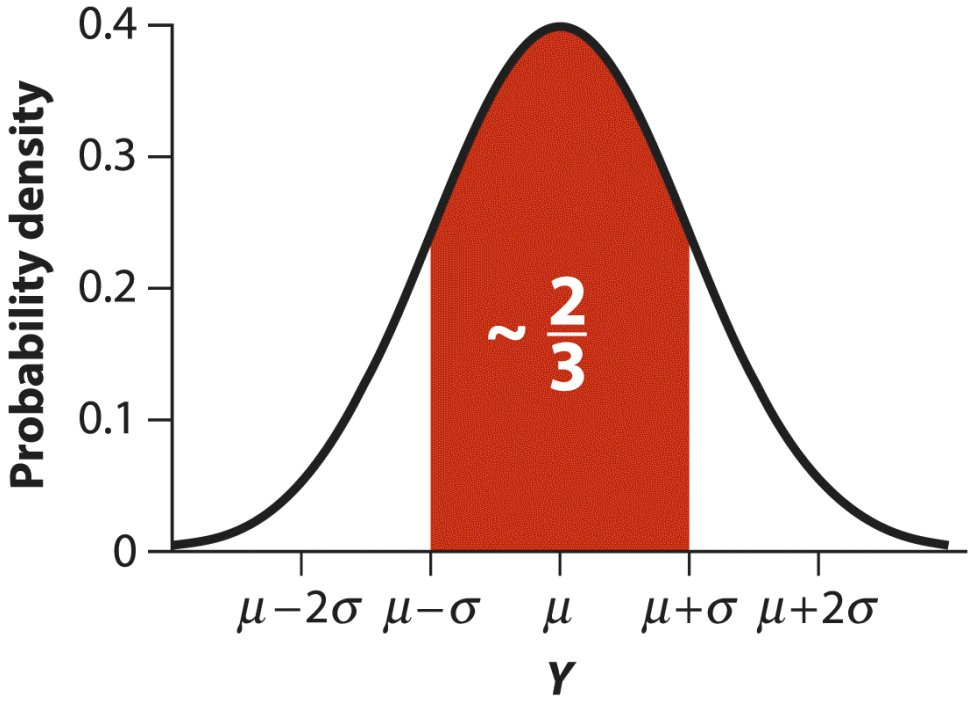


2.2 正态分布的一些特征

- 正态分布的一些特征
  - 它是一种连续分布；
  - 概率等于曲线下的面积（而不是曲线的高度来衡量）；
  - 它围绕均值对称；
  - 概率密度在均值处是最高的（单峰）；
  - 正态分布的平均数 (mean)、中位数 (median) 和众数 (mode) 是相等的。

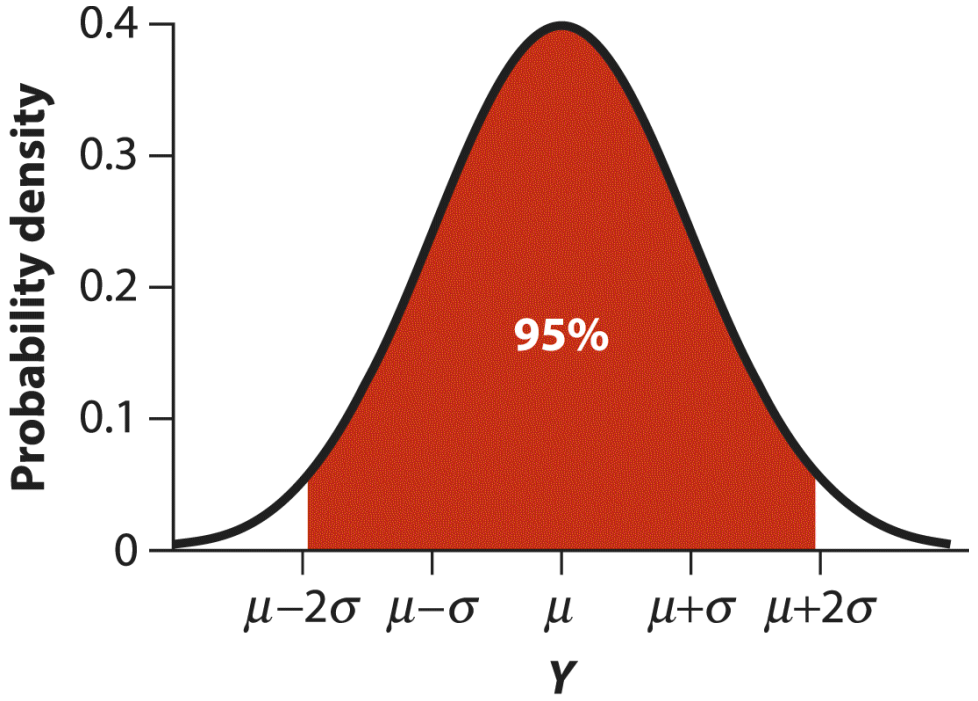
2.2 正态分布的一些特征

- 正态分布中 2/3的概率（更准确的说是68.3%）位于均值的一个标准差以内。
  - 从正态分布中随机抽取的观测值的测量值在$\mu - \sigma$ 和 $\mu + \sigma$ 之间的概率为 0.683。



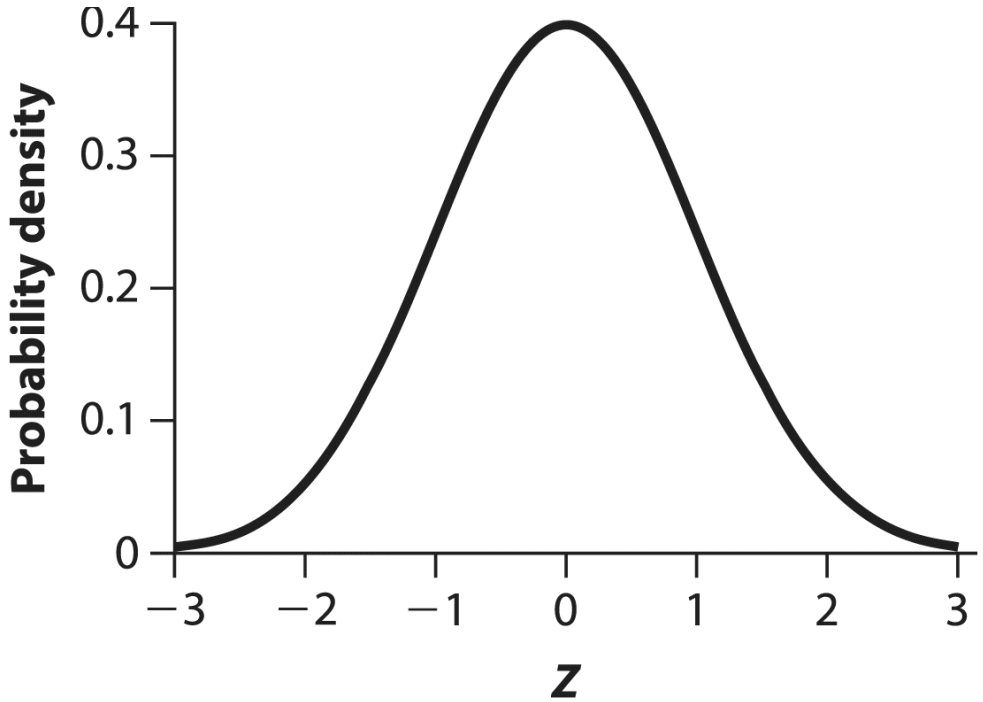
2.2 正态分布的一些特征

- 正态分布中 95% 的概率位于均值的两个标准差以内（更准确地说，是 1.96 个标准差以内）。
  - 从正态分布中随机抽取的观测值的测量值在 $\mu - 1.96\,\sigma$ 和 $\mu + 1.96\,\sigma$ 之间的概率为 0.95。



2.3 标准正态分布及其统计表

- 标准正态分布 the standard normal distribution
  - 均值为0、标准差为1的正态分布（$\mu = 0;\ \sigma = 1$）
  - 通常用符号 Z 来表示具有标准正态分布的变量

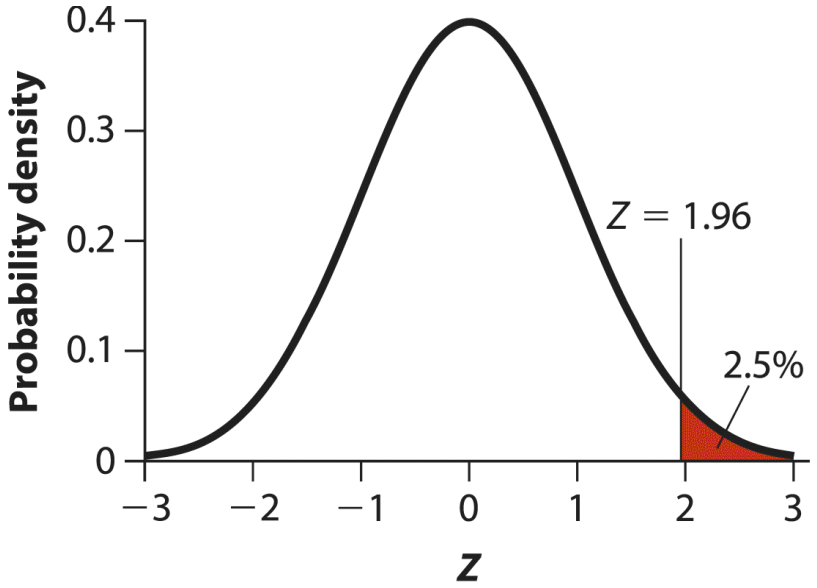


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 标准正态分布及其统计表

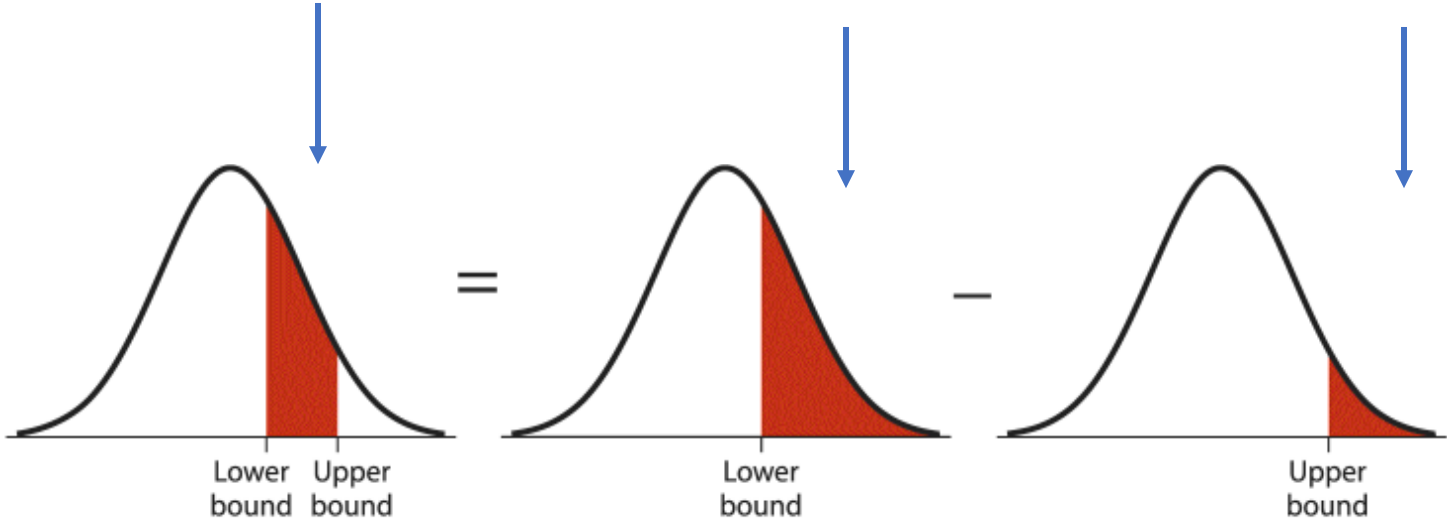
- 统计表 statistical tables
  - 与泊松分布和二项分布不同，从正态分布中取样时特定事件发生的概率很难手工计算，因为它需要对复杂的函数进行积分。
  - 相反，我们使用统计表或计算机来获取正态曲线下的概率。
  - 统计表 B (Statistical Table B) 给出了大于某个正值Z的概率： 1 - Pr[Z]

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
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
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857



2.3 标准正态分布及其统计表

- 对称性
  - 标准正态分布中随机观测值小于-1.96 的概率与观测值大于 1.96 的概率相同： $\Pr[Z < -1.96] = \Pr[Z > 1.96] = 0.025$
- 计算某一区间的概率（两步法）
  - $\Pr[\text{lower bound} < Z < \text{upper bound}] = \Pr[Z > \text{lower bound}] - \Pr[Z > \text{upper bound}]$



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 标准正态分布及其统计表

- 转换任何正态分布为标准正态分布
  - 转换公式：$Z = \frac{Y-\mu}{\sigma}$
  - 该公式将具有均值 μ 和标准差 σ 的正态分布的 Y 转换为具有标准正态分布的Z；
  - 这个标准化的 Z 值称为标准正态偏差
    - 以标准差的数量来衡量 Y 与平均值的距离（i.e., Z 个 σ）
  - 对于所有正态分布来说（包括标准正态分布），获得与均值相差 Z 个标准差的观测值的概率是相同的。

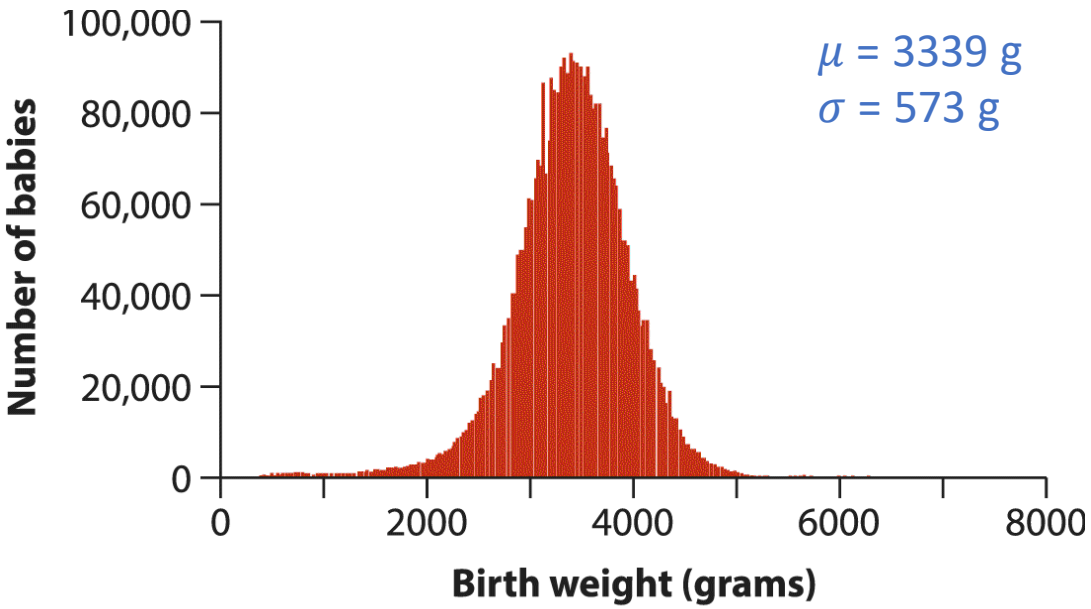
3. 样本均值的抽样分布与正态分布

- 正态分布最重要的应用之一
  - 它可以用来描述许多估计值的抽样分布，其中最主要的就是样本均值。
  - 估计值的抽样分布列出了我们在对总体进行抽样时的所有可能的值，并描述了它们出现的概率

3. 正态分布和抽样分布

- 例子：新生儿体重
  - n = 4,017,264（1991年）
    - 数据很多，直方图的区间宽度也很窄 → 分布看起来接近一条平滑的曲线。

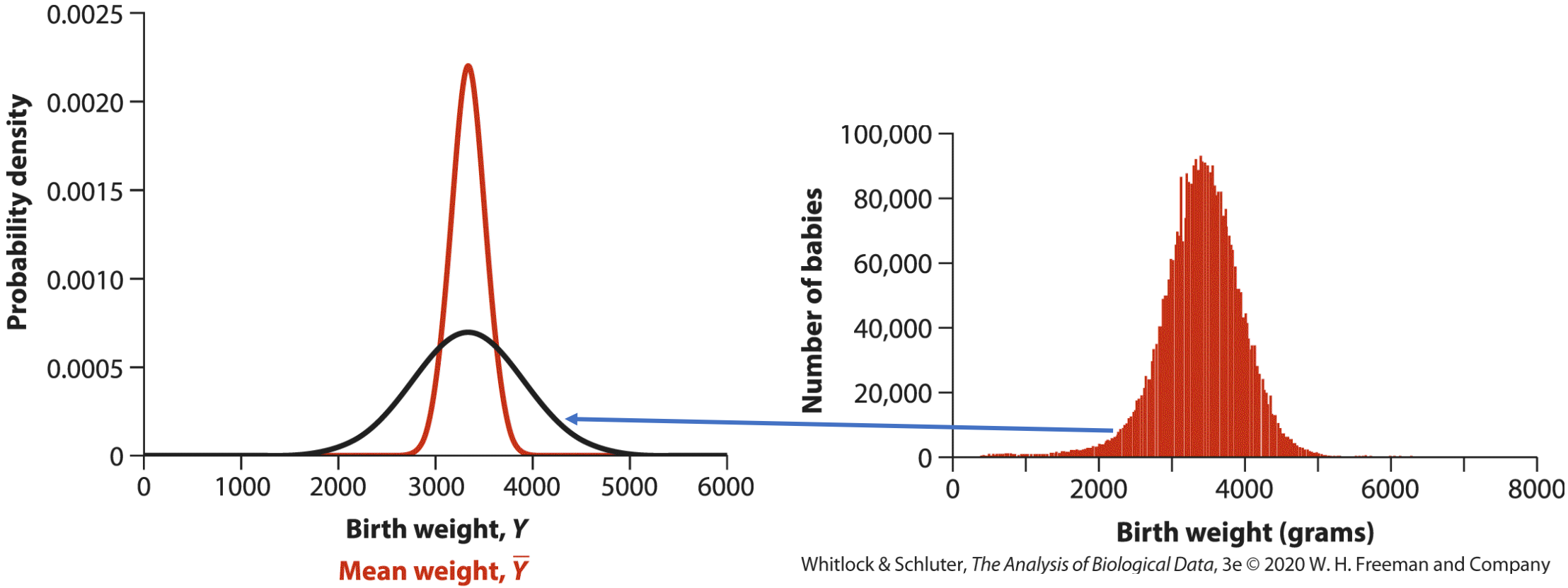
- 均值 mean
- 中位数 median
- 众数 mode



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

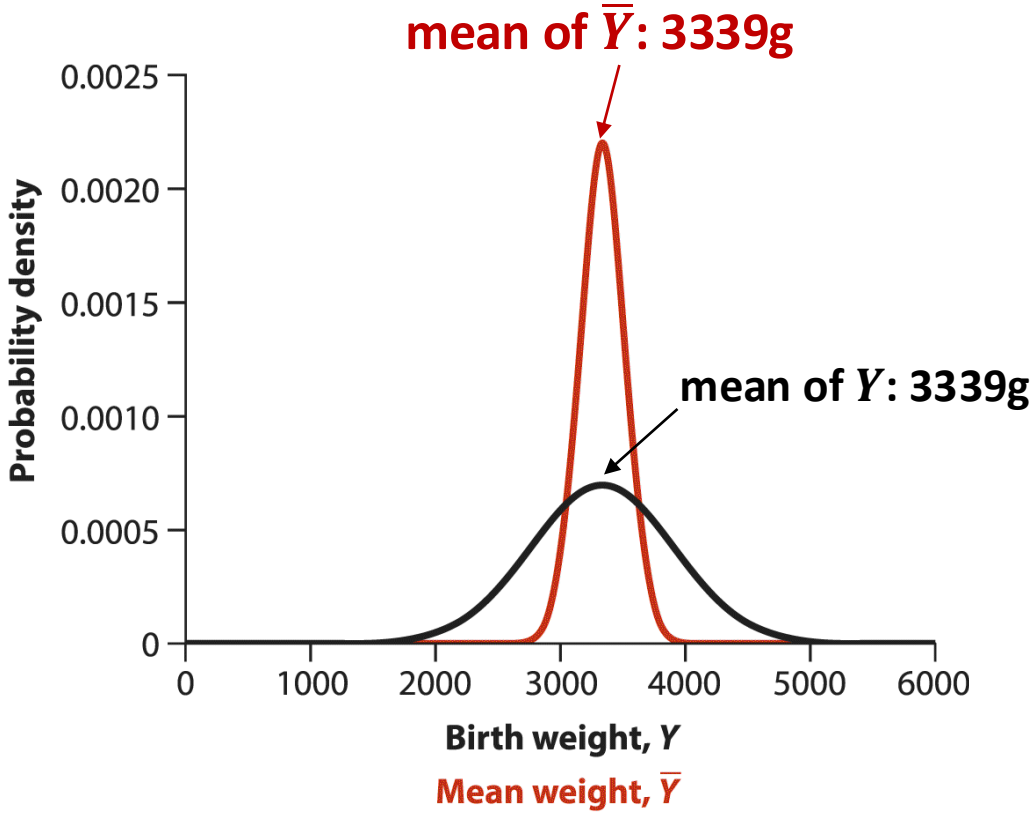
3. 正态分布和抽样分布

- 每个样本的平均值都会偶然地与真实平均值不同；如果我们抽取不同的样本，它也会以不同的方式与真实值不同。
- 不同样本可能得到的 $\bar{Y}$ 的不同值及其相关概率，构成了 $\bar{Y}$ 的抽样分布。



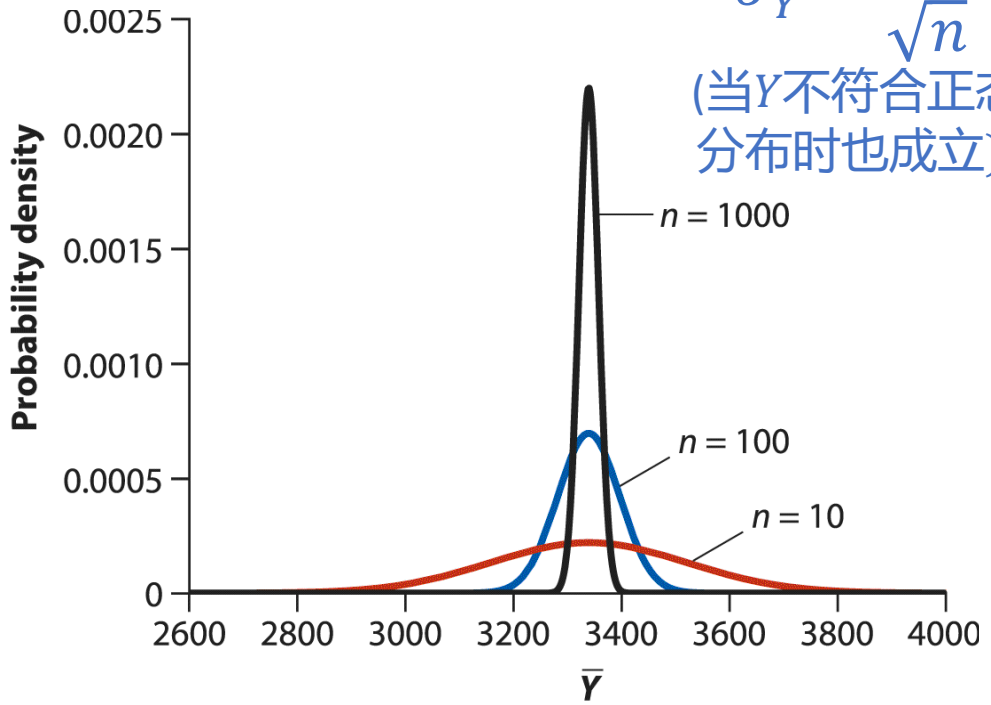
3. 正态分布和抽样分布

- 抽样分布： $\bar{Y}$的抽样分布的均值等于 Y 的原始分布的均值；表明，从正态分布的总体中抽取的随机样本的均值即为总体均值 μ 的无偏估计；
- 样本大小：影响抽样分布的标准差（即均值估计的标准误）；



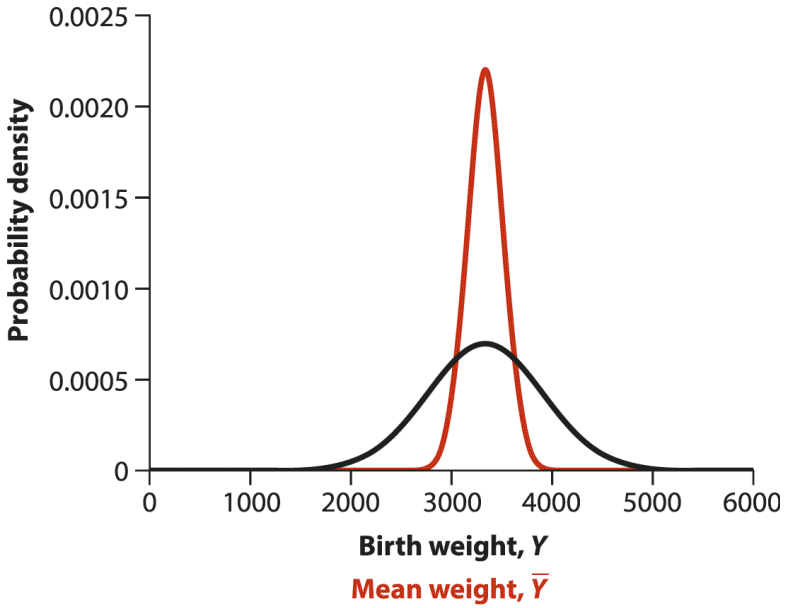
$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

(当Y不符合正态分布时也成立)

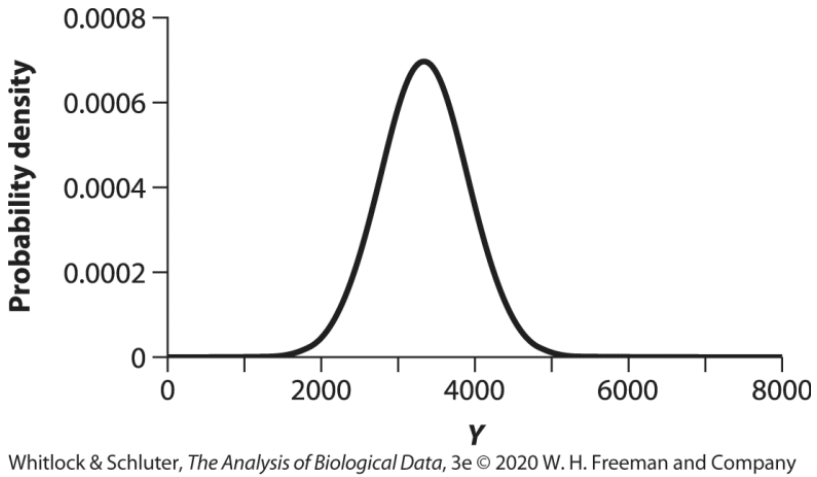


3.1 样本均值的抽样分布

- 当总体本身是正态分布时，样本均值的抽样分布就是正态分布。
  - 那么，可以利用“当 Y 为正态时，$\bar{Y}$ 的抽样分布为正态分布”这一事实，来计算总体均值的置信区间，并计算针对均值的检验假设的P 值。
- 样本均值 $\bar{Y}$ 的标准正态转换
  - $Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$



$$Z = \frac{Y - \mu}{\sigma}$$ 样本数据的Z转换

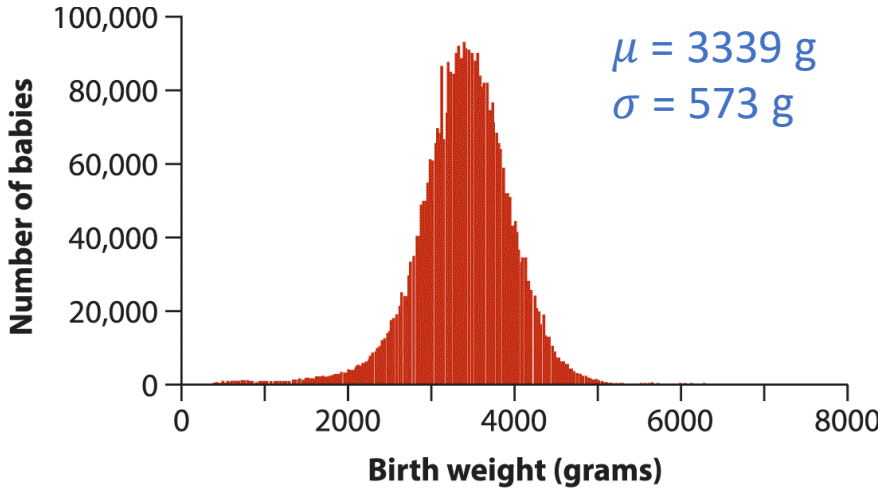


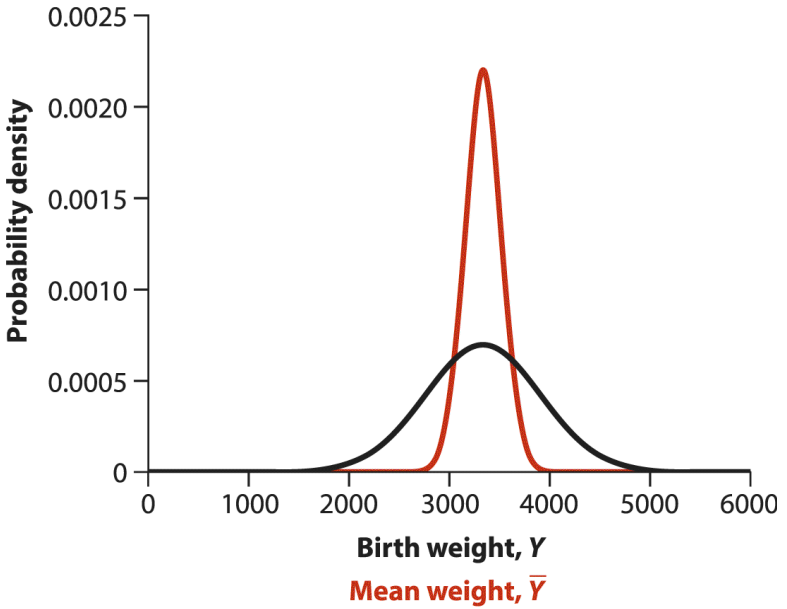
3.1 样本均值的抽样分布

- 样本均值 $\bar{Y}$ 的标准正态转换
  - 样本： n = 80个新生儿
  - 问题：样本均值大于 3370 g 的概率是多少?

    Pr[$\bar{Y}$ > 3370] = ?
  - $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{573}{80} = 64.1\text{g}$
  - $Z = \frac{\bar{Y}-\mu}{\sigma_{\bar{Y}}} = \frac{3370-3339}{64.1} = 0.48$

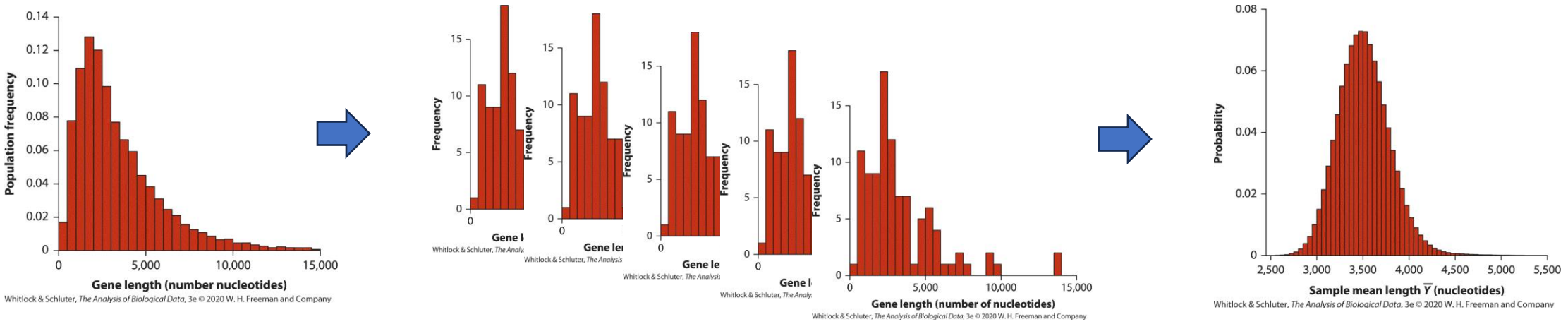
- **Pr[$\bar{Y}$ > 3370] = Pr[$Z = \frac{\bar{Y}-\mu}{\sigma_{\bar{Y}}}$ > 0.48]**
  - 借助统计表， Pr[Z > 0.48] = 0.316
  - 从这个总体中抽取 n = 80的 （多个） 样本，其中样本均值会超过3370 g 的比例为31.6%。





3.2 中心极限定理 the central limit theorem

- 从正态分布中抽取的样本平均值本身也是正态分布。
- 中心极限定理：只要样本量足够大，即使单个数据点的 分布不是正态分布，样本均值 $\bar{Y}$ 的抽样分布也近似正态分布。
- 样本量多大才算足够大，取决于总体中观测值分布的形状：
  - 原始分布与正态分布越相似，要得到近似于正态分布的样本均值的抽样分布所需的样本量就越小。



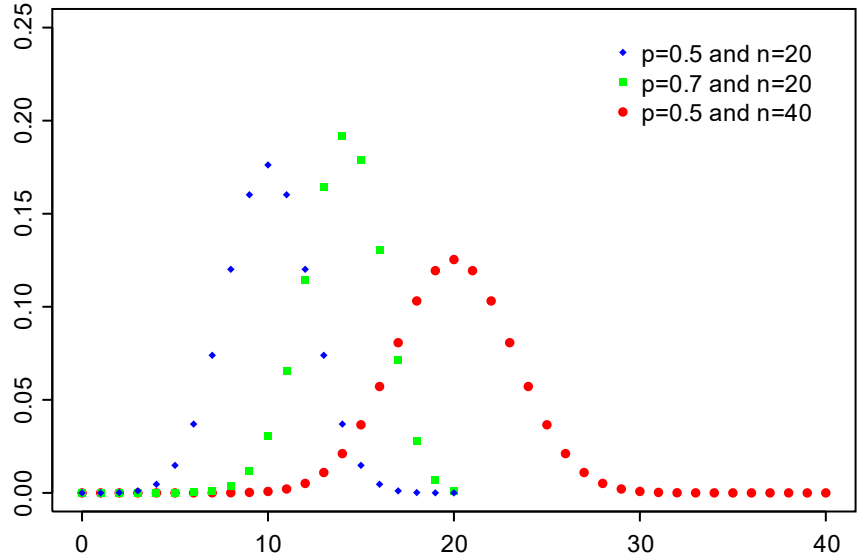
3.2 中心极限定理 the central limit theorem

- 多数强大的用于分析生物数据的统计方法都假定样本均值的分布（以及其它一些估计值的分布）遵循正态分布。

- 所以，中心极限定理的妙处在于，如果样本量足够大，那么即使我们的数据是从非正态分布的总体中抽取的，也可以使用这些强大的方法。

3.3 二项分布的正态近似值

- 二项分布是一种离散概率分布
  - 它描述了 n 次独立试验中“成功事件”的次数；
  - 其中 p 是任何一次试验成功的概率。
- 正态近似
  - 根据中心极限定理，n值较大的二项分布近似于正态分布。
  - 因为成功的次数是一种和：如果每次成功都标记为 "1"，每次失败都标记为 "0"，那么成功次数的计数就是所有个体的 "1 "和 "0 "之和。



记号	$B(n,p)$
参数	$n \geq 0$ 试验次数 (整数) $0 \leq p \leq 1$ 成功概率 (实数)
值域	$k \in \{0,\ldots,n\}$
概率质量函数	$\binom{n}{k} p^k (1-p)^{n-k}$

3.3 二项分布的正态近似值

- 正态逼近的条件:
  - 经验法则 (Rule of thumb)
    - 若要判断P是否小于0.05（即当 $\alpha = 0.05$时），np 和 $n(1-p)$ 均需 > 5;
    - 若要计算精确P值或 $\alpha < 0.05$，则需要更大的 np;
- 正态分布参数:
  - 均值 mean (μ): np
  - 标准差 standard deviation (σ): $np(1-p)$

4. 单样本t-检验

- 对正态分布变量进行的最简单的假设检验
  - 可以询问样本数据的测量值是否与假设的总体均值一致
  - 如何对标准差进行统计说明

4.1 t 分布 (Student's t-distribution)

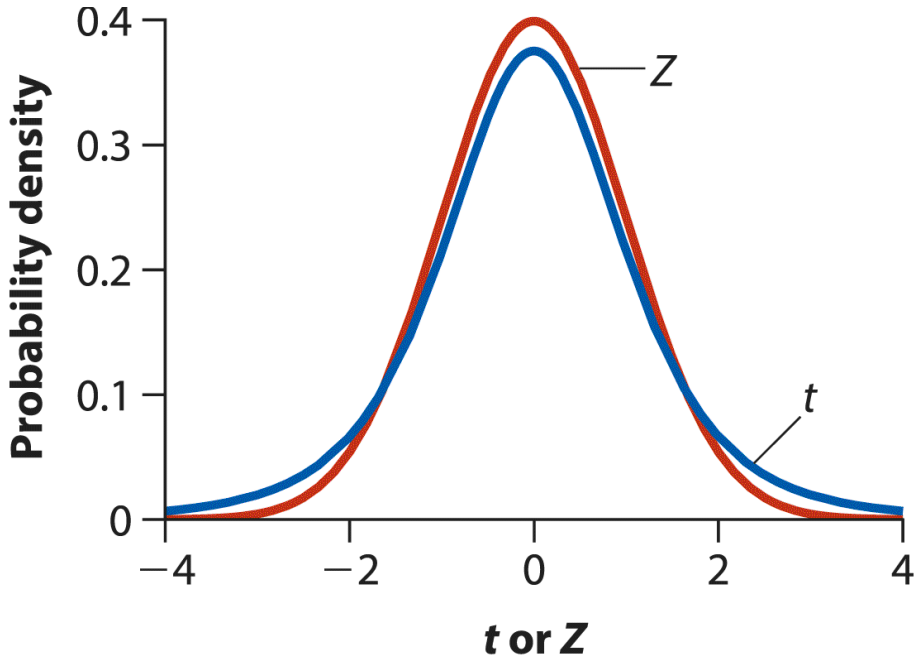
- 当 Y 为正态时，样本均值 $\bar{Y}$ 的抽样分布为正态分布
  - 对 $\bar{Y}$ 进行如下标准正态转换：$Z = \frac{\bar{Y}-\mu}{\sigma_{\bar{Y}}}$，但其中 $\sigma_{\bar{Y}}$ 通常是未知的；
  - 所以，我们参考t分布；
- 统计量 Student's t
  - 统计量 t 被称为 "Student's t"，源于最先发现其特性的人的笔名 "Student"。(实际上，"Student "是都柏林健力士啤酒公司的员工威廉-戈塞特（William Gosset）。戈塞特之所以使用笔名，是因为吉尼斯禁止员工发表文章，因为几年前另一名员工在未经授权的情况下泄露了一些酿酒秘密。

4.1 t 分布 (Student's t-distribution)

- 统计量 Student's t
  - $\boldsymbol{t = \frac{\bar{Y} - \mu}{\mathrm{SE}_{\bar{Y}}}}$
  - 其中 $\mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$，$s$为样本的标准差；
  - 自由度为 $df = n - 1$，因为计算了样本的s；
  - 虽然 t 的计算公式与 Z 相似，但重要的区别在于该统计量的抽样分布不是正态分布。
    - 考虑了小样本情况下的不确定性和总体方差未知的情况，提供了一种在不确定条件下进行有效推断的方法
    - 使用$\mathrm{SE}_{\bar{Y}}$来近似 $\sigma_{\bar{Y}}$会增加 t 的抽样误差；随着样本量的增加，t会更趋近于Z；

4.1 t 分布 (Student's t-distribution)

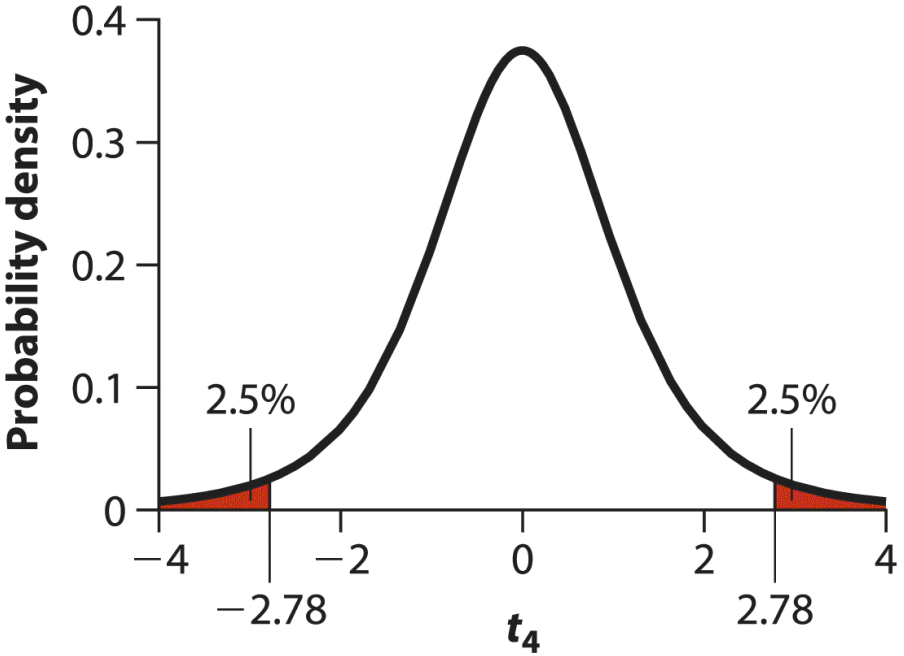
- 统计量 Student's t
  - $\boldsymbol{t = \frac{\bar{Y} - \mu}{\mathrm{SE}_{\bar{Y}}}}$
  - 其中 $\mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$，　s为样本的标准差；
  - 自由度为 $df = n - 1$；
  - 虽然 t 的计算公式与 Z 相似，但重要的区别在于该统计量的抽样分布不是正态分布。
    - t 分布的尾部更宽些



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

4.1 t 分布 (Student's t-distribution)

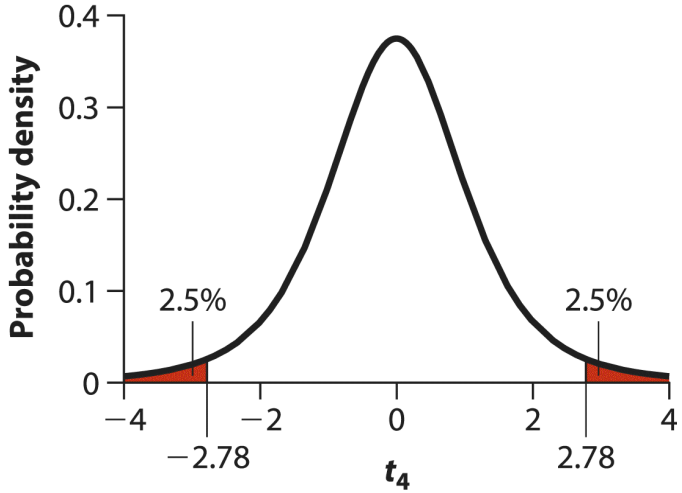
- t分布的一些特征
  - t 分布的尾部更宽些
  - $df = 4$ 的 t 分布见右图（n=5）
    - 95%的面积在区间 [-2.78, 2.78] 之间;
    - 即，从符合正态分布的总体中获得重复随机样本（n=5），那么 95% 的样本得到的均值 $\bar{Y}$ 将在真实的总体均值 (μ) 两侧的2.78 个估计标准误范围内。
    - Z 分布的95%的面积在[-1.96, 1.96]区间内。
    - 与Z分布相比，t 分布的取值范围更大，这是因为标准误估计的不确定性。



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

4.1 t 分布 (Student's t-distribution)

- t分布的关键值/临界值
  - $t_{0.05(2),\,df}$
    - 表示自由度为df的t分布的5%的关键值；
    - (2) 表示5%的面积分布在曲线两侧；
    - $t_{0.05(2),\,df=4}$ = 2.78
  - 统计表C (Statistical Table C)



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

one-tailed α	0.10	0.05	0.025	0.01	0.005	0.0005
two-tailed α	0.20	0.10	0.05	0.02	0.01	0.001
df						
1	3.078	6.314	12.71	31.82	63.66	636.62
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959

4.1 t 分布 (Student's t-distribution)

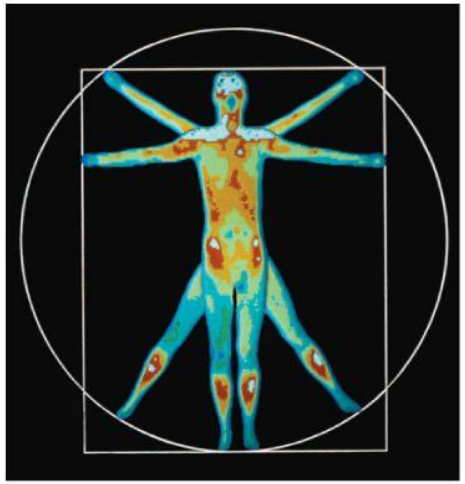
- 均值估计的95%置信区间
  - 2SE法则（近似估计）：
    - $\bar{Y} - 2\mathrm{SE}_{\bar{Y}} < \mu < \bar{Y} + 2\mathrm{SE}_{\bar{Y}}$
  - 而标准化差异 $\frac{\bar{Y}-\mu}{\mathrm{SE}_{\bar{Y}}}$ 符合 t 分布，因此可更加精确给出95%CI：
    - $-t_{0.05(2),\,df} < \frac{\bar{Y}-\mu}{\mathrm{SE}_{\bar{Y}}} < t_{0.05(2),\,df}$
    - $\bar{Y} - t_{0.05(2),\,df}\ \mathrm{SE}_{\bar{Y}} < \mu < \bar{Y} + t_{0.05(2),\,df}\ \mathrm{SE}_{\bar{Y}}$
    - （用关键值$t_{0.05(2),\,df}$ 来替代了2）

4.2 单样本 t-test

- 单样本 t-test (the one sample t-test)
  - 基于t-分布来比较一个随机样本的均值和零假设中的总体均值。
- 零假设和备择假设
  - H_0: The true mean equals μ_0 （总体均值为μ_0）；
  - H_A: The true mean does not equal μ_0 （总体均值不为μ_0）；
- 检验统计量为 t
  - $t = \frac{\bar{Y} - \mu_0}{\mathrm{SE}_{\bar{Y}}}$
- P值
  - 对比观测 t 值与 Student’s t-distribution

4.2 单样本 t-test

- 例子：人体温度
  - 人的正常体温为 37°C (98.6°F)，但实际数据是否很好地支持这个陈述呢?
  - 体温并非都等同于98.6 华氏度，但测量结果与上述中总体均值98.6°F是否一致?

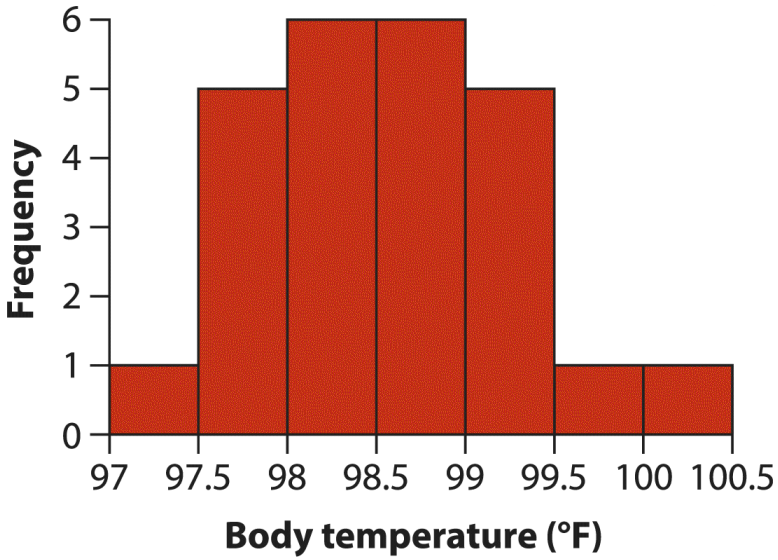


Dr. Ray Clark FRPS & Mervyn de Calcina-Goff FRPS/Science Source

individual	temperature		
1	98.4	13	99.5
2	98.6	14	99.4
3	97.8	15	98.4
4	98.8	16	99.1
5	97.9	17	98.4
6	99	18	97.6
7	98.2	19	97.4
8	98.8	20	97.5
9	98.8	21	97.5
10	99	22	98.8
11	98	23	98.6
12	99.2	24	100
		25	98.4

4.2 单样本 t-test

- 例子：人体温度 $(n = 25)$
- 1. 零假设和备择假设
  - H_0: 总体均值$\mu_0 = 98.6°F$。
  - H_A: 总体均值$\mu_0 \neq 98.6°F$。

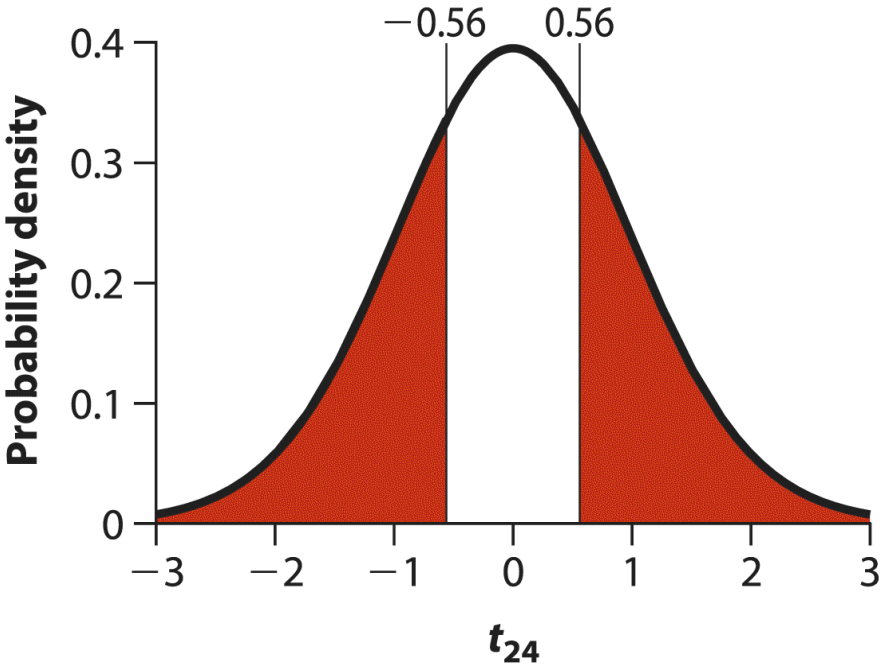


  - 设置假设相关说明
    - 零假设并不是随意设置的，因为我们检验的是正常体温均值为 98.6°F这样一个常识。
    - 该检验是双侧的，即如果样本均值远高于或远低于98.6 °F，结论就会是拒绝零假设。

- 2. 计算检验统计量 t
  - $\bar{Y}$=98.524， $s = 0.678$
  - $\mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{0.678}{\sqrt{25}} = 0.136$
  - $t = \frac{\bar{Y} - \mu_0}{\mathrm{SE}_{\bar{Y}}} = \frac{98.524 - 98.6}{.0136} = -0.56$

4.2 单样本 t-test

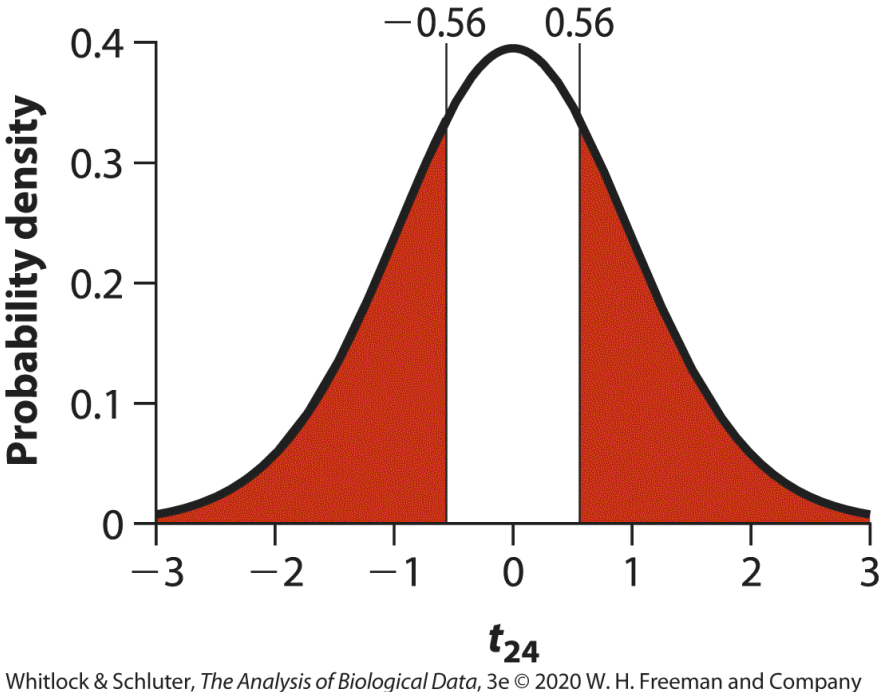
- 例子：人体温度 $(n = 25)$
  - 1. H_0: $\mu_0 = 98.6°F$; H_A: $\mu_0 \neq 98.6°F$;
  - 2. 检验统计量 $t = \frac{\bar{Y} - \mu_0}{SE_{\bar{Y}}} = -0.56$;
  - 3. P值?
    - 如果零假设成立，则 t应具有 $df = n - 1$ 的分布;
    - 如果样本均值$\bar{Y}$与零假设的值 μ_0完全吻合，则 t 等于零;
    - 如果 $\bar{Y}$ 和 μ_0之间的差异太大，那么观测值将落在t分布的某一个尾部;
    - P 值是假定零假设成立，得到与t = -0.56 一样极端或更极端结果的概率。



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

4.2 单样本 t-test

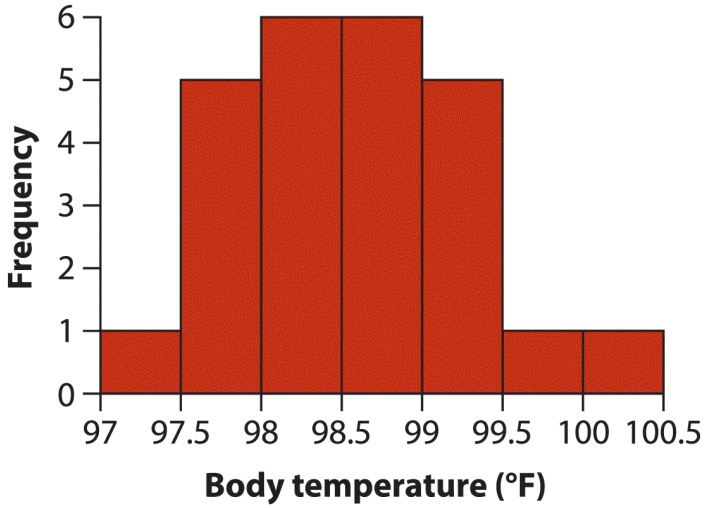
- 例子：人体温度 ($n = 25$)
  - 1. H_0: $\mu_0 = 98.6°F$; H_A: $\mu_0 \neq 98.6°F$;
  - 2. 检验统计量 $t = \frac{\bar{Y} - \mu_0}{SE_{\bar{Y}}} = -0.56$;
  - 3. P值?
    - P = Pr[t < -0.56] + Pr[t > 0.56] = 2 × Pr[t > 0.56]
      - t分布是以0为中心对称的分布
    - 统计软件：P = 0.58
    - 统计表C中关键值：$t_{0.05(2),\, df=24}$ = 2.06
    - → P > 0.05
      - 如果 t 值比关键值/临界值离零更远，那么我们就可以拒绝零假设，反之则不拒绝。



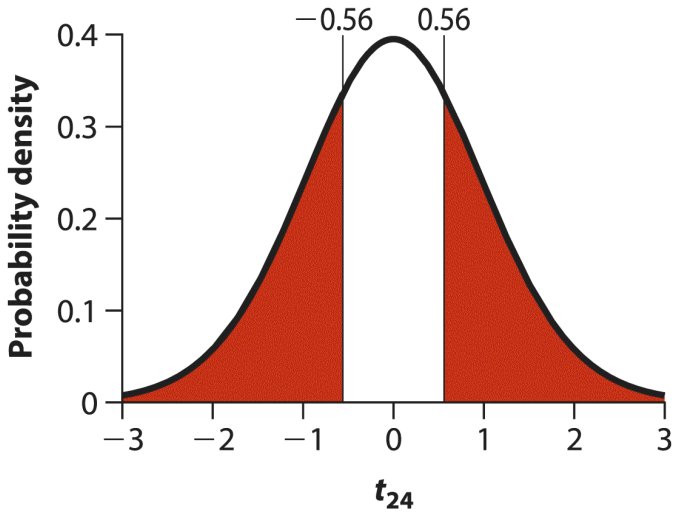
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

4.2 单样本 t-test

- 例子：人体温度 $(n = 25)$
  - 1. H_0: $\mu_0 = 98.6°F$; H_A: $\mu_0 \neq 98.6°F$;
  - 2. 检验统计量 $t = \frac{\bar{Y} - \mu_0}{SE_{\bar{Y}}} = -0.56$;
  - 3. P值
    - P > 0.05
  - 4. 结论
    - 不拒绝零假设
    - 即人体正常体温为 98.6°F / 37°C。



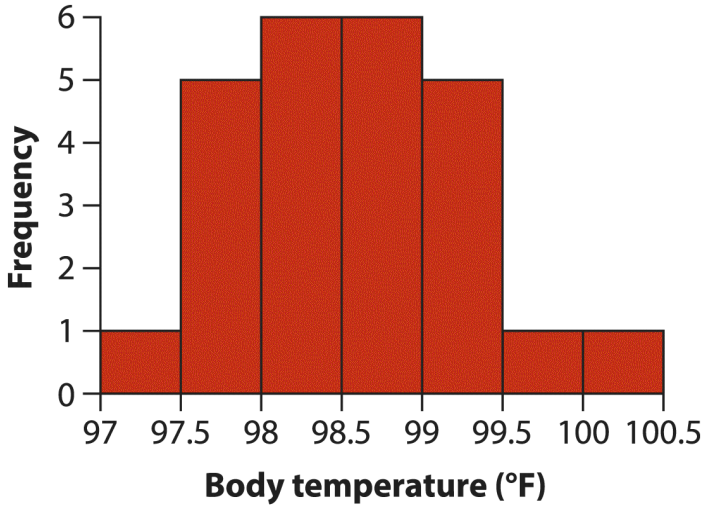
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company



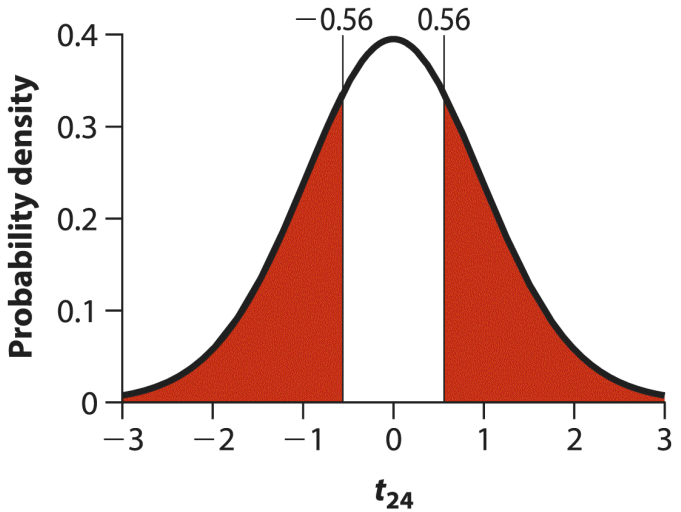
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

4.2 单样本 t-test

- 例子：人体温度 ($n = 25$)
  - 1. H_0: $\mu_0 = 98.6°F$; H_A: $\mu_0 \neq 98.6°F$;
  - 4. 结论: 不拒绝零假设。
  - 这一结果是否意味着关于人体体温的常识是正确的?
    - 不一定；只是我们现有的数据无法拒绝零假设;
    - 零假设仍可能是不成立，而我们可能缺乏足够的效力 (power) 来发现差异。
  - 从数据来看，平均体温合理的取值范围?
    - 均值的 95% 置信区间
    - $\bar{Y} - t_{0.05(2),df}\ \mathrm{SE}_{\bar{Y}} < \mu < \bar{Y} + t_{0.05(2),df}\ \mathrm{SE}_{\bar{Y}}$
    - $98.24°F < \mu < 98.80°F$



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

4.2 单样本 t-test

- 例子：人体温度；更大的样本 $n = 130$
  - 1. H_0: $\mu_0 = 98.6°F$; H_A: $\mu_0 \neq 98.6°F$;
  - 2. $\bar{Y}$ =98.25°F； $s = 0.733°F$

    $t = \frac{\bar{Y} - \mu_0}{SE_{\bar{Y}}} = -5.44$;
  - 3. P值 = 0.000016 < 0.05
  - 4. 结论: 拒绝零假设。
  - 均值的 95% 置信区间
    - $\bar{Y} - t_{0.05(2),df}\ SE_{\bar{Y}} < \mu < \bar{Y} + t_{0.05(2),df}\ SE_{\bar{Y}}$
    - 98.12°F < μ < 98.38°F （范围：0.26 °F）
    - 对比：n=25时的95% CI：98.24°F < μ < 98.80°F （范围：0.56 °F）

4.2 单样本 t-test

- 例子：人体温度；更大的样本 $n = 130$
  - 1. H_0: $\mu_0 = 98.6°F$; H_A: $\mu_0 \neq 98.6°F$;
  - 2. $\bar{Y}$ =98.25°F； $s = 0.733°F$

    $t = \frac{\bar{Y} - \mu_0}{SE_{\bar{Y}}} = -5.44$;
  - 3. P值 = 0.000016 < 0.05
  - 4. 结论: 拒绝零假设。
  - 均值的 95% 置信区间
    - $\bar{Y} - t_{0.05(2),df}\ SE_{\bar{Y}} < \mu < \bar{Y} + t_{0.05(2),df}\ SE_{\bar{Y}}$
    - 98.12°F < μ < 98.38°F （范围：0.26 °F）
    - 对比：n=25时的95% CI：98.24°F < μ < 98.80°F （范围：0.56 °F）

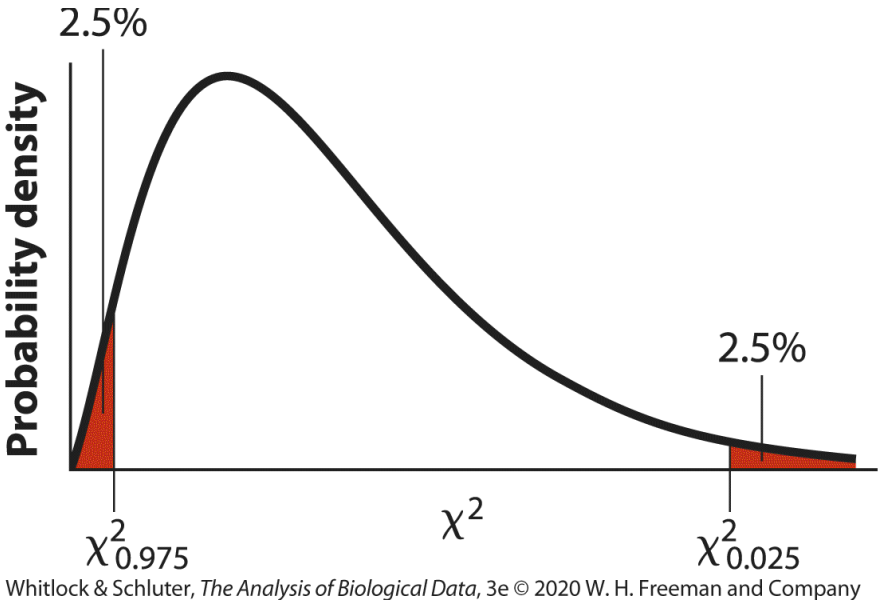
- 随着样本量的增加，均值估计的抽样误差趋于减小。
- 即使小样本的均值估计与大样本的均值估计相差不大，但大样本的均值估计的范围（即置信区间）要小得多。
- 因此，大样本更有可能拒绝错误的零假设。
- 假定对大样本的分析结果是真实的，那么小样本就更容易出现II类错误（即未能拒绝错误的零假设）。

4.3 单样本 t 检验的前提假设 (Assumptions)

- 单样本 t 检验及计算总体均值的前提假设：
  - 数据是从总体中随机抽取的样本；
    - （本门课中的每种统计推断方法都有这一假设）
  - 变量在总体中呈正态分布；
    - 评估方法：直方图或其它图形方法来检查数据的频率分布；
    - 判断是否有偏斜、双峰或其它偏离正态分布的情况；
- 现实：
  - 生物学中很少有变量与正态分布完全吻合；
  - 但完全吻合并不是必须的。

4.4 正态分布总体的标准差和方差的估计

- 总体的标准差(standard deviation): σ
  - 用样本的标准差来估计： s
- 总体的方差 (variance): σ^2
  - 用样本的方差来估计： s^2
  - 置信区间 $(1 - \alpha)$
    - 如果Y来自正态分布，那么以下估计值符合卡方分布：
      - $\chi^2 = (n-1)s^2/\sigma^2$
    - $\frac{df \times s^2}{\chi^2_{\alpha/2,df}} < \sigma^2 < \frac{df \times s^2}{\chi^2_{1-\alpha/2,df}}$

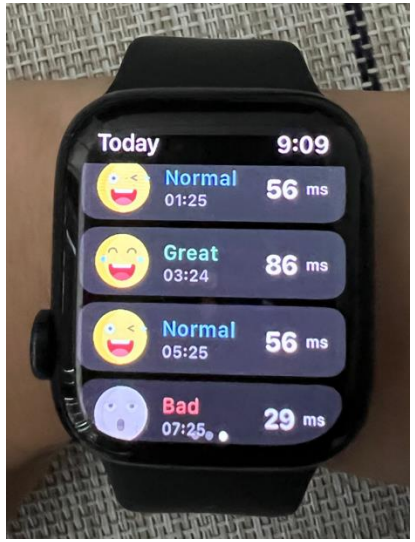


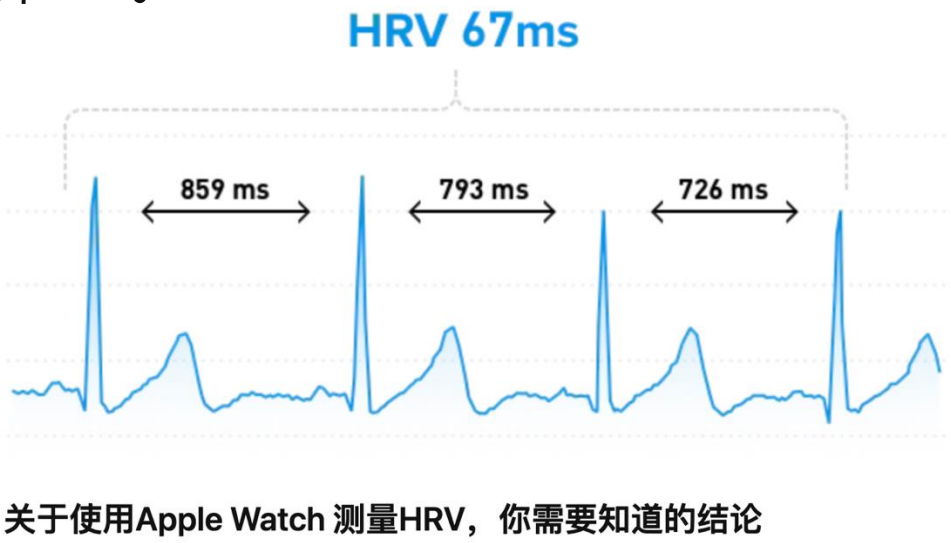
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

卡方分布的范围为[0, + ∞)，
并非像正态分布那样是围绕0对称的；

4.4 正态分布总体的标准差和方差的估计

- 总体的标准差σ 和样本的标准差s
- HRV = Heart Rate Variability = 心率变异性
  - 是判断自主神经活动的常用的定量指标;
  - 研究证实，较低的HRV与焦虑、压力，甚至心血管疾病有关;
  - 使用心率传感器测量到的心跳间隔的标准差来计算HRV。





关于使用Apple Watch 测量HRV，你需要知道的结论

https://zhuanlan.zhihu.com/p/145708916

5. 小结 – 正态分布与 t 检验

- 正态分布 $Y \sim N(\mu, \sigma^2)$
  - 一条钟形曲线，一种连续的概率分布，近似于自然界中许多变量的分布；
  - 如果一个变量是正态分布，那么它的平均值、中位数和众数都是相同的；
  - 正态分布围绕其均值μ对称，约 95% 的个体在均值的两个标准差σ以内；
- 标准正态分布
  - 均值为 0，标准差等于 1 的正态分布 $Y \sim N(0, 1)$
  - 所有正态分布都可以转换成标准正态分布；
  - 方法是计算每个测量值偏离平均值的标准差数：
    - $Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$，即标准正态偏差；
- 当n较大时的二项分布可近似为正态分布（np和n(1-p)均>5）

5. 小结 – 正态分布与 t 检验

- 如果连续数值变量 Y 在总体中呈正态分布，均值为μ，那么多次随机抽样（样本大小为 n）的样本均值 $\bar{Y}$ 也呈正态分布（抽样分布），其均值等于μ（与Y的真实均值相同），标准误为 $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$，其中 σ 是总体的真实标准差。
- 样本均值标准误的估计值为 $\mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$。
- 如果总体呈正态分布，则标准化样本均值 $t = \frac{\bar{Y} - \mu_0}{\mathrm{SE}_{\bar{Y}}}$ 符合自由度 $df = \mathrm{n} - 1$的Student-t 分布。

5. 小结 – 正态分布与 t 检验

- 单样本 t 检验
  - 将样本均值与零假设中提出的总体均值的特定值μ_0进行比较；
  - 检验统计量为$t = \frac{\bar{Y} - \mu_0}{\mathrm{SE}_{\bar{Y}}}$，零假设下其抽样分布为$t$ 分布（*df = n-1*）；
  - t 分布还用来计算均值的置信区间；
- 总体分布的方差估计的置信区间基于卡方分布来计算；
  - 标准差即为方差估计的平方根；
- 均值的 t 检验、置信区间及方差/标准差估计的前提假设：
  - 随机样本；
  - 变量在总体中呈正态分布；

R coding

```
> t.test(x, y = NULL,
        alternative = c("two.sided", "less", "greater"),
        mu = 0, paired = FALSE,
        var.equal = FALSE, conf.level = 0.95, ...)
```

2.3 标准正态分布及其统计表

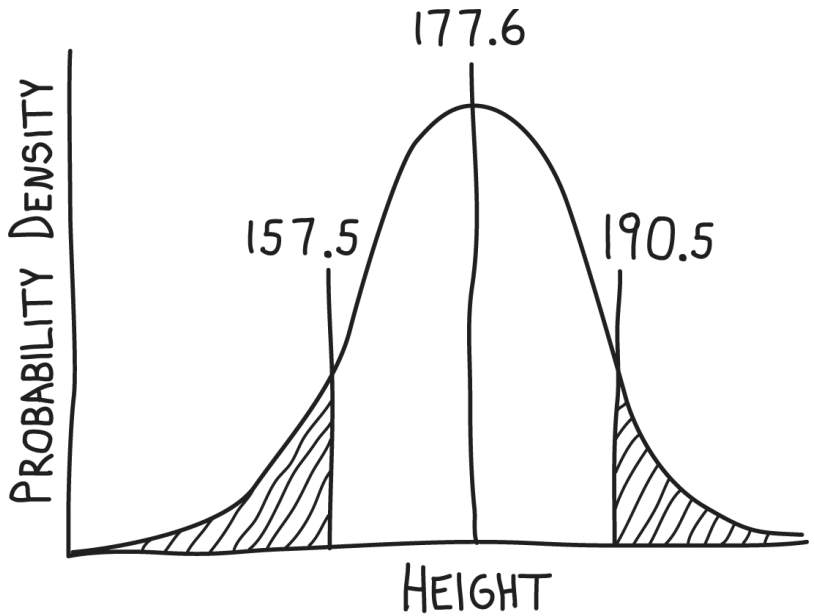
- 例子：宇航员的身高
  - 在一个性别和年龄组中，成年人身高的分布合理地近似于正态分布
    - 20 至 29 岁美国男性的平均值为 177.6 厘米，标准差为 9.7 厘米;
    - 20 至 29 岁美国女性的平均身高为 163.2 厘米，标准差为 10.1 厘米;
    - （麦克道尔等人，2008 年）
  - 美国国家航空航天局规定，身高低于 62 英寸（ 157.5 厘米）和超过 75 英寸（和 190.5 厘米）则不能成为宇航员飞行员（NASA，2004 年）。
  - 157.5 厘米 ≤ 宇航员 ≤ 190.5 厘米
  - 提问： 20至29岁的美国人中，因NASA这样的身高限制而无法成为宇航员的人群占多大比例?

补充资料——2.3 标准正态分布及其统计表

- 转换问题
  - 157.5 厘米 ≤ 宇航员身高 ≤ 190.5 厘米
  - Pr[不能成为宇航员] = Pr[身高 < 157.5] + Pr[身高 > 190.5]
- 转换为<u>标准正态分布</u>
  - Too tall?
  - Pr[身高 > 190.5] = Pr[$Z > \frac{190.5-177.6}{9.7}$]

    = Pr[$Z > 1.33$] = 0.09176
    - 190.5 厘米的数值出现在男性平均身高的 1.33 个标准差之上
    - 超过这个身高（1.33个标准差）的概率是0.09176
    - R code： pTooTall <- 1 - pnorm(190.5, mean = 177.6, sd = 9.7)

20 至 29 岁美国男性身高
平均值为 177.6 厘米
标准差为 9.7 厘米



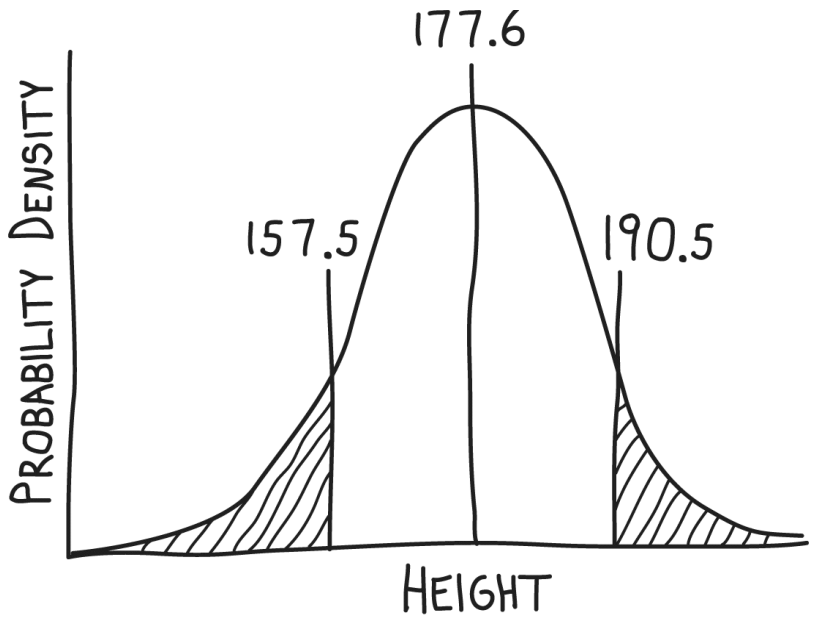
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 标准正态分布及其统计表

- 转换问题
  - 157.5 厘米 ≤ 宇航员身高 ≤ 190.5 厘米
  - Pr[不能成为宇航员] = Pr[身高 < 157.5] + Pr[身高 > 190.5]
- 转换为<u>标准正态分布</u>
  - Too short?
  - Pr[身高 < 157.5] = Pr[$Z < \frac{157.5-177.6}{9.7}$]

    = Pr[$Z < 2.07$] = 0.01923
    - 157.5 厘米的数值出现在男性平均身高的 2.07 个标准差之下
    - 低于这个身高（1.33个标准差）的概率是0.01923
    - R code： pTooShort <- pnorm(157.5, mean = 177.6, sd = 9.7)

20 至 29 岁美国男性身高
平均值为 177.6 厘米
标准差为 9.7 厘米



2.3 标准正态分布及其统计表

- 转换问题
  - 157.5 厘米 ≤ 宇航员身高 ≤ 190.5 厘米
  - Pr[不能成为宇航员] = Pr[身高 < 157.5] + Pr[身高 > 190.5]

20 至 29 岁美国男性身高
平均值为 177.6 厘米
标准差为 9.7 厘米

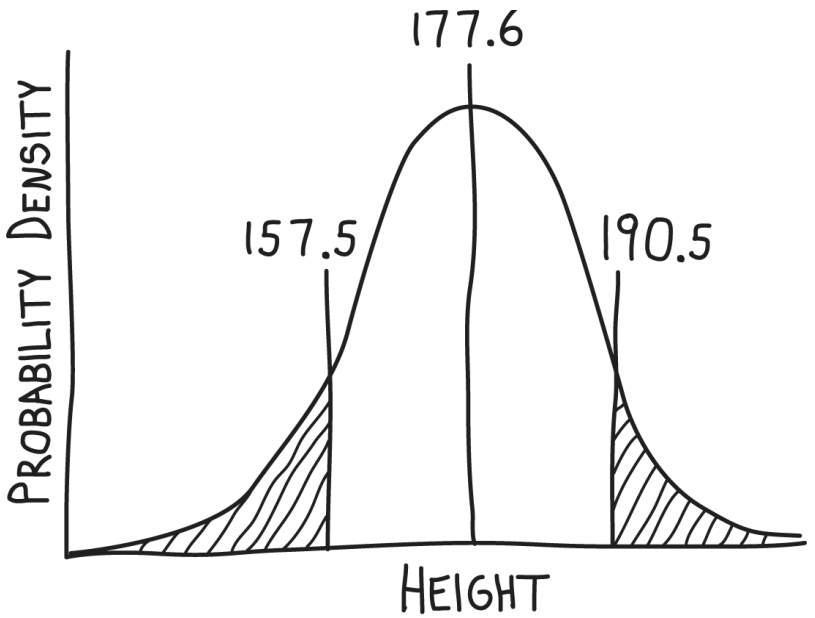
- 转换为<u>标准正态分布</u>
  - 男性： Pr[身高 > 190.5或身高 < 157.5]

    = Pr[Z > 1.33] + Pr[Z < 2.07]

    = 0.09176 + 0.01923 = 0.11099
  - 女性： Pr[身高 > 190.5或身高 < 157.5]

    = Pr[Z > 2.7] + Pr[Z < 0.56]

    = 0.00347 + 0.28774 = 0.291
  - 总计：男性11.1%，女性 29.1%不能成为宇航员。



Next Lecture

- Comparing two means 两个样本的t检验
- Handling violations of assumptions 不满足前提假设的相关处理
- Designing experiment 实验设计