总体分布，样本分布，抽样分布

Polulation, sample, and sampling distributions

Qingyao Zhang

2026-03-27

目录

1 总体

我们捏造一个人数为 10000 人的总体，为 10000 名被试设定编号 ID，并捏造出其 IQ 数据，设定 IQ 的均值为 100，标准差为 15。

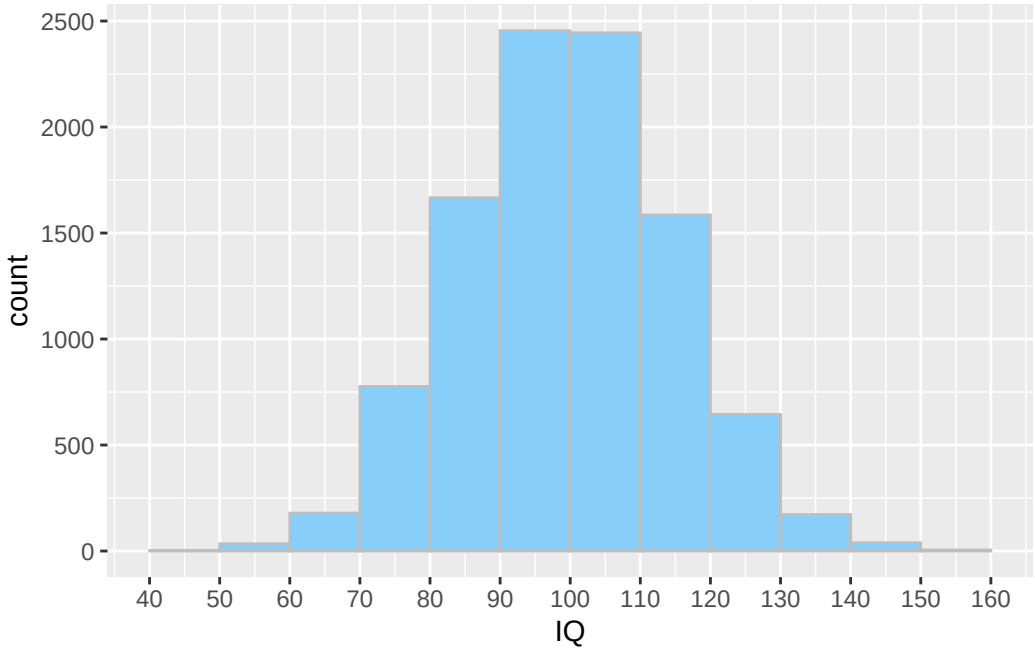
```
# Population, 总体
# 总体均值设置为 100
mu <- 100
# 总体标准差设置为 15
sigma <- 15
# 随机种子值设置为 20240906
set.seed(20251107)
# 将 ID 与 IQ 存入数据表中
population10000 <- data.frame(
  # 生成 10000 个 ID (被试编号)
  ID = seq(1:10000),
  # 生成 10000 个均值为 mu、标注差为 delta 的 IQ 分数
  IQ = round(rnorm(10000, mean = mu, sd = sigma), 0)
  )
# 查看数据表
# View(population10000)
# 查看数据表前 6 行
head(population10000)
##   ID  IQ
## 1  1  57
## 2  2 109
## 3  3 105
## 4  4  68
## 5  5 105
## 6  6 135
```

2 总体分布

总体分布 (population distribution) 是总体 (10000 名被试 IQ 得分) 的分布。注意该分布的横纵坐标的全距。

```
# 作总体的直方图
# 总体分布是 10000 个数据点的分布。
library(ggplot2)
ggplot(population10000, aes(IQ)) +
  geom_histogram(breaks = seq(40, 160, 10),
```

```
                  fill = "lightskyblue", color = "gray") +
  scale_x_continuous(breaks = seq(40, 160, 10)) +
  coord_cartesian(xlim = c(40, 160))
```



3 取样过程

接下来，我们模拟取样的过程。我们从 $N = 10000$ 的总体中随机抽取一个样本量 $n = 100$ 的样本。我们称这个样本为 sample1。

```
# 从总体中随机抽取一个样本量为 100 的样本，抽取其 ID
sample1ID <- sample(population10000$ID, 100)
# 根据 ID 选出 sample1 的数据
sample1 <- population10000[sample1ID, ]
```

4 样本

查看 sample1 中的数据。注意：被试 ID 的顺序是乱的，这是由随机取样导致的。

```r
# 查看样本
head(sample1)
##        ID  IQ
## 2974 2974 110
## 6384 6384 117
## 8008 8008  84
## 7773 7773  94
## 7247 7247 117
## 3739 3739 115
```

5 基于样本估计总体均值与标准差

样本的均值是总体均值的无偏估计。样本的标准差总是小于总体的标准差，因此使用样本的无偏标准差作为总体标准差的无偏估计。尽管名为“无偏”，实际上还是有偏差的。

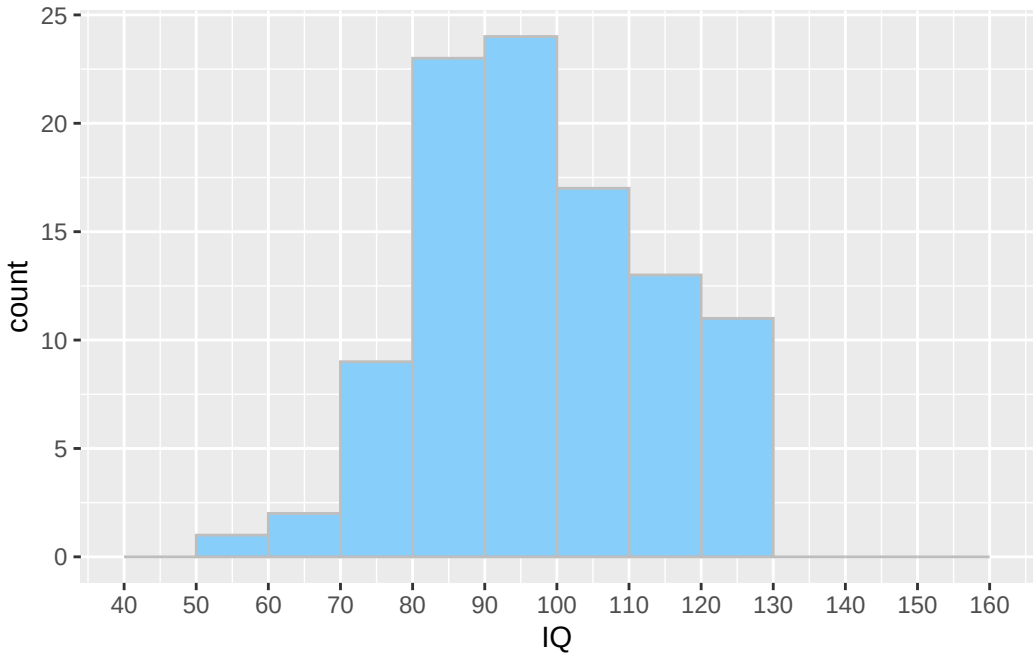
```r
# 计算 sample1 的 IQ 的均值
sample1_mean <- mean(sample1$IQ)
sample1_mean
## [1] 98.26
# 计算 sample1 的 IQ 的标准差
sample1_sd <- sqrt(sum((sample1$IQ - sample1_mean)^2)/100)
sample1_sd
## [1] 15.76745
# sample1 的无偏标准差（基于 sample1 估计总体的标准差）
sample1_sd_unbiased <- sqrt(sum((sample1$IQ - sample1_mean)^2)/(100 - 1))
sample1_sd_unbiased
## [1] 15.84688
# 使用`sd()`函数计算 sample1 的无偏标准差（总体标准差的估计值）
sd(sample1$IQ)
## [1] 15.84688
```

6 样本分布

样本分布（sample distribution）是样本中 100 个被试 IQ 得分的分布。

```r
ggplot(sample1, aes(IQ)) +
  geom_histogram(breaks = seq(40, 160, 10),
                 fill = "lightskyblue", color = "gray") +
```

```
  scale_x_continuous(breaks = seq(40, 160, 10)) +
  coord_cartesian(xlim = c(40, 160))
```



7 重复取样

使用一个样本计算得到的均值估计总体的均值产生的偏差可能较大，如果我们重复取样的过程，抽取多个样本，将多个样本的均值求均值，那么，我们将得到更为准确的估计值。下面，我们从人数 $N = 10000$ 的总体中取出 $m = 1000$ 个样本量 $n = 100$ 的样本。

```
# 从总数为 10000 的样本中累计抽取 1000 个样本量为 100 的样本，存入 samples
samples <- list()
for (index in 1:1000) samples[[index]] <- population10000[sample.int(10000, 100), ]
# 查看 samples
length(samples)
## [1] 1000
head(samples[[1]])
##        ID  IQ
## 2238 2238  74
## 5186 5186 105
## 4033 4033 138
## 2375 2375 118
```

```
## 7640 7640 131
## 556   556  93
```

8 1000 个样本的均值的均值

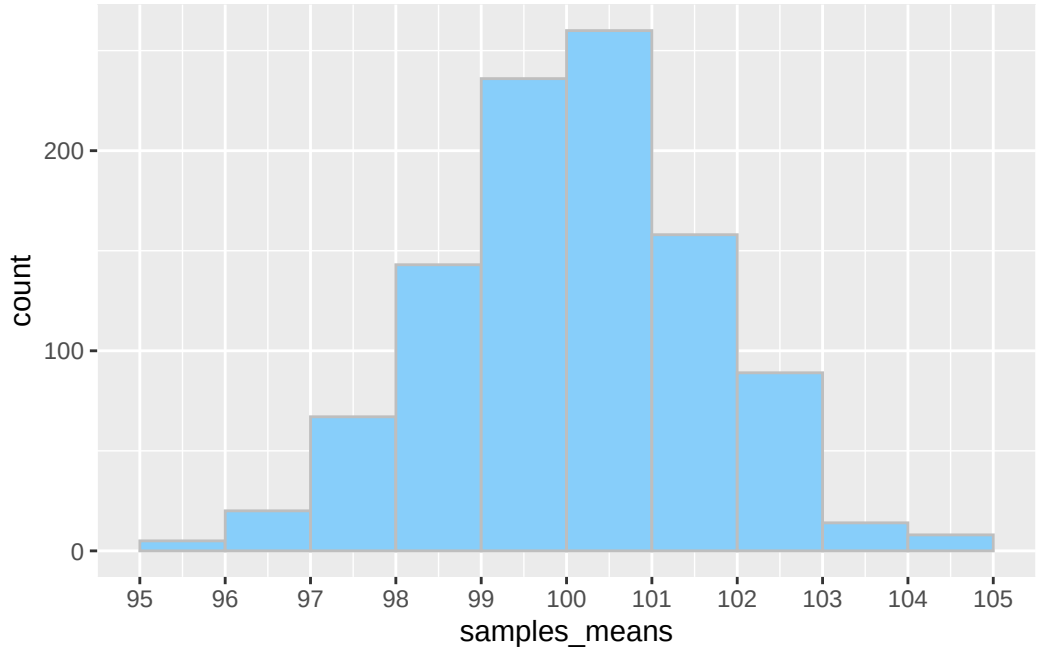
与单个样本（例如：sample1）相比，1000 个实际样本的均值的均值更加接近总体均值。

```
# 分别计算 1000 个样本的均值，一共得到 1000 个均值，存入 samples_means
samples_means <- sapply(samples, function(x) mean(x$IQ))
range(samples_means)
## [1]  95.49 104.44
head(samples_means)
## [1]  99.54 101.41 101.60 100.98  99.20  99.97

# 1000 个实际样本的均值的均值。
mean(samples_means)
## [1] 100.0849
```

9 抽样分布

抽样分布是 1000 个样本的均值的分布。这个分布中的每一个数据点是一个样本中 100 名被试的 IQ 的均值，而不是一个被试的 IQ 得分，这个分布的中心是是 1000 个样本的均值的均值，这个分布的标准差是 1000 个样本的均值的标准差。

```
# 抽样分布。即，1000 个样本的均值的分布。
ggplot(data.frame(samples_means), aes(samples_means)) +
  geom_histogram(breaks = seq(95, 105, 1),
                 fill = "lightskyblue", color = "gray") +
  scale_x_continuous(breaks = seq(95, 105, 1))
```



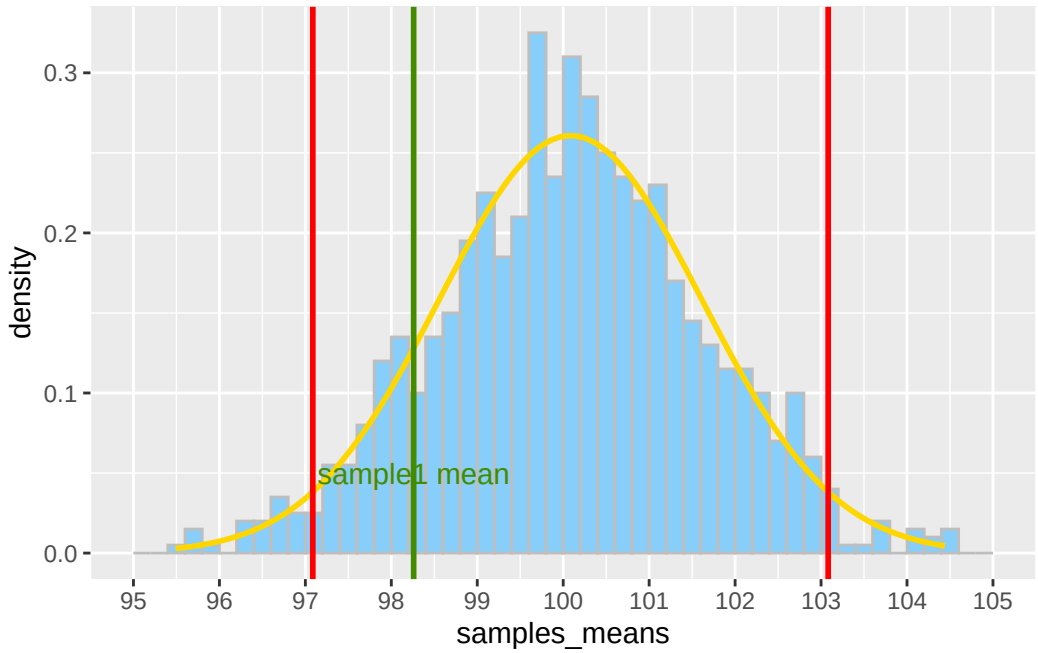

```r
#  + coord_cartesian(xlim = c(40, 160))

# 抽样分布的标准差。
# 即，1000 个均值的标准差。
sqrt(sum((samples_means - mean(samples_means)) ^ 2) / 1000)
## [1] 1.528941
```

10 差异显著的概率标准

理想条件下，抽样分布服从正态分布，呈现出“左侧低分占少数、右侧高分占少数、中间中等分占大多数”的分布特点。得分的极端程度与概率相关，极端低分与极端高分所占比例较小。统计上判断一个分数是否极端的标准为概率 $p < 0.05$，p 是 probability 的缩写。意即，分布两侧尾端的极端得分（含左侧极端低分与右侧极端高分）占总数的 5%。由于正态分布具有对称性，因此左侧尾端（左侧红线以左）占 2.5%，右侧尾端（右侧红线以右）占 2.5%。当得分落在分布两侧尾端 5% 的区域时，我们就说该得分与分布的中心（分布的均值）具有显著差异。换言之，如果我们知道了一个得分 x 在分布中的位置及其切割出来的尾部的概率，我们就可以判断这个得分 x 与分布的中心是否具有显著差异。

```r
# 抽样分布。即，1000 个样本的均值的分布。
ggplot(data.frame(samples_means), aes(samples_means)) +
  geom_histogram(aes(y = after_stat(density)),
                 breaks = seq(95, 105, 0.2),
```

```
                 fill = "lightskyblue", color = "gray") +
  scale_x_continuous(breaks = seq(95, 105, 1)) +
  stat_function(fun = dnorm,
                args = list(mean = mean(samples_means), sd = sd(samples_means)),
                linewidth = 1,
                color = "gold") +
  geom_vline(xintercept = mean(samples_means) - qnorm(0.975)*sd(samples_means),
             color = "red", linewidth = 1) +
  geom_vline(xintercept = mean(samples_means) + qnorm(0.975)*sd(samples_means),
             color = "red", linewidth = 1) +
  geom_vline(xintercept = sample1_mean,
             color = "chartreuse4", linewidth = 1) +
  annotate("text", x = sample1_mean, y = 0.05,
           label = "sample1 mean", color = "chartreuse4")
```



11 sample1 均值与总体均值的直观比较

sample1 均值为 98.26，比总体均值小。但是，这一差异是否显著？我们需要观察 sample1 的均值在抽样分布中的位置。上图蓝色线标出了 sample1 均值的位置。可见，sample1 均值并未落在双侧尾部，故 sample1 均值与总体均值没有显著差异。

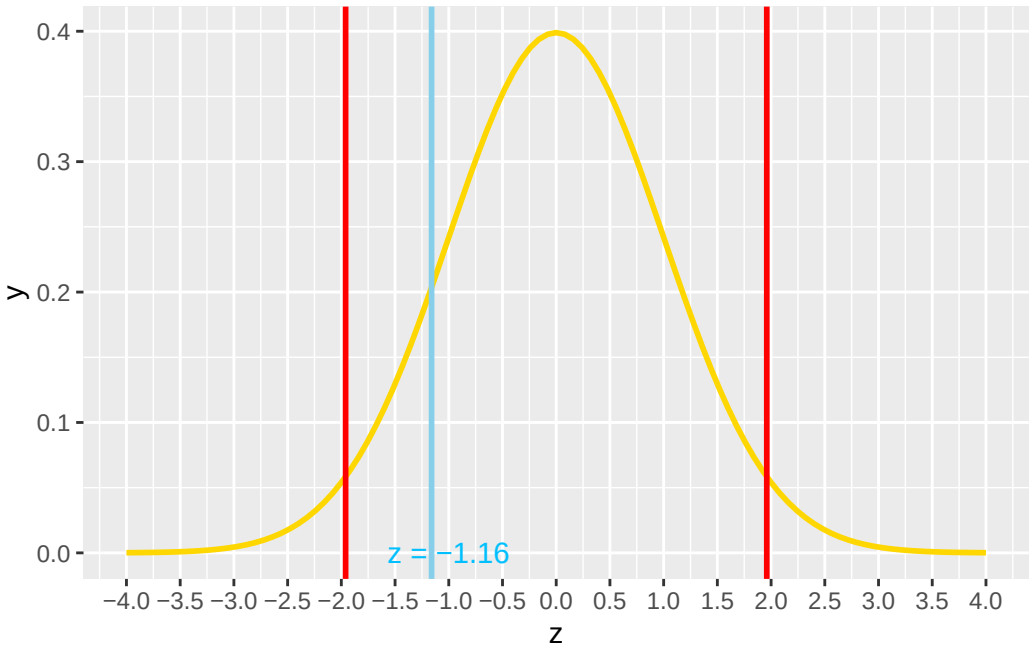
12 z 分数与单样本 z 检验

理想情况下，抽样分布服从正态分布，正态分布的形状是固定的，横坐标与其概率具有严格的对应关系，只要我们知道了横坐标，我们就可以计算出这一横坐标所切割出来的尾部的精确概率。由于众多变量（例如：智商、年龄、身高）的单位不同，为了便于计算，我们可以将变量得分转化为 z 分数，将数据的分布转化为标准正态分布，从而计算 z 分数所对应的尾部的概率。z 分数转换的公式如下，其中 M_x 为 x 的均值，SD_x 为 x 的标准差：

$$z = \frac{x - M_x}{SD_x}$$

转换为 z 分数后，z 分数的均值为 0，标准差为 1。z 分数的分布如下。

```
ggplot(data.frame(z = seq(-4, 4, 0.1)), aes(z)) +
  scale_x_continuous(breaks = seq(-4, 4, 0.5)) +
  stat_function(fun = dnorm,
                linewidth = 1,
                color = "gold") +
  geom_vline(xintercept = qnorm(0.025),
             color = "red", linewidth = 1) +
  geom_vline(xintercept = qnorm(0.975),
             color = "red", linewidth = 1) +
  geom_vline(xintercept = (sample1_mean - mu)/(sigma/sqrt(100)),
             color = "skyblue", linewidth = 1) +
  annotate("text", x = -1, y = 0,
           label = "z = -1.16",
           color = "deepskyblue")
```



z 分数的均值（分布的中心）为 0，z 的取值范围大致为 [-4, 4]。z 反映了一个得分在分布中的位置。当 $z < 0$ 时，z 位于分布中心的左侧，当 $z > 0$ 时，z 位于分布中心的右侧。已知 z，pnorm() 函数可以计算累积概率 p（即上图中金色正态曲线与垂直线 z 包围的面积）。pnorm() 默认计算从左到右的累积概率，即左侧尾部的概率。例如，当 z = -1 时，左尾的概率为 0.1586553。另外，已知 p，qnorm() 函数可以计算 z。

因此，我们可以将 sample1 均值转换为 z 分数，然后计算 sample1 均值在抽样分布中的尾部概率。**理想情况下**（例如：样本量 n 更大，抽取的样本的数量 m 更大），抽样分布的均值为 $\mu = 100$，抽样分布的标准差为 $\frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = 1.5$。抽样分布的标准差通常被称为标准误 (Standard Error)，统计符号为 SE。样本均值的标准误反映了样本均值估计总体均值的标准误差。

```
# 使用总体标准差估计均值的标准误
sigma / sqrt(100)
## [1] 1.5
# 基于总体的标准差，对 sample1 的均值进行单样本 z 检验
z <- (sample1_mean - mu)/(sigma/sqrt(100))
z
## [1] -1.16
# z 左尾概率
p_one_tailed <- pnorm(z)
p_one_tailed
## [1] 0.1230244
# z 双尾 p 值
p_two_tailed <- p_one_tailed*2
```

```r
p_two_tailed
## [1] 0.2460488
```

可得，z = -1.16，p = 0.25，sample1 均值与总体均值 100 没有显著差异。

这里，我们计算 z 分数，继而计算 p 值的统计检验方法就是单样本 z 检验。单样本 z 检验的作用是在总体均值、标准差已知的情况下，比较样本均值与总体均值的差异。

13 t 分数与单样本 t 检验

但是，总体均值与标准差通常是未知的，需要用样本的均值与无偏标准差进行估计。按照与单样本 z 检验相同的方法，我们可以计算出 “z” 及其 p 值。由于样本对总体的估计有偏差，由样本均值与样本无偏标准差估计得到的抽样分布与基于总体均值与总体标准差估计得到的抽样分布有偏差。因此，此时得到的 “z” 不是 z，而是 t。这一检验就是单样本 t 检验。

```r
# 使用样本 1 的标准差估计样本 1 均值的标准误
sample1_sd_unbiased / sqrt(100)
## [1] 1.584688
# 基于 sample1 无偏标准差，对 sample1 的均值进行单样本 t 检验
t <- (sample1_mean - mu)/(sample1_sd_unbiased/sqrt(100))
t
## [1] -1.098008
# t 的双尾 p 值
p_t <- pt(t, df = 100 - 1)*2
p_t
## [1] 0.2748644
```

可见，单样本 z 检验计算得到的 z 值、p 值与单样本 t 检验计算得到的 z 值、p 值是接近的，统计结论是相同的。

事实上，我们不必手写上述略微复杂的代码来进行单样本 t 检验，t.test() 函数可以帮我们做：

```r
t.test.out <- t.test(sample1$IQ, alternative = "two.sided", mu = 100)
print(t.test.out)
##
##  One Sample t-test
##
## data:  sample1$IQ
## t = -1.098, df = 99, p-value = 0.2749
## alternative hypothesis: true mean is not equal to 100
## 95 percent confidence interval:
```

```
##   95.11564 101.40436
## sample estimates:
## mean of x
##     98.26
```

可见，`t.test()` 检验的结果与我们编辑公式代码计算的结果是一样的。

14 自由度

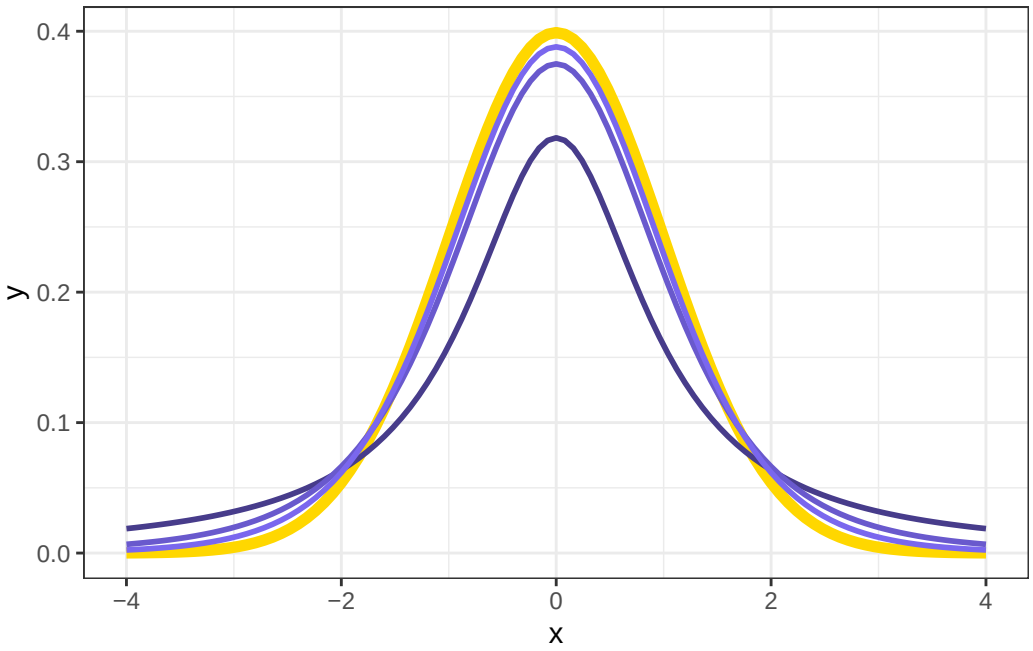
在上面的 t 检验中，p 值的计算会使用自由度 (degree of freedom, df) 这一统计量。什么是自由度？自由度是数据可以自由变化的程度。在前文中，我们对 `sample1` 均值进行单样本 t 检验，这一检验需要我们首先计算 `sample1` 的均值。当 `sample1` 的均值保持不变时，样本中只有 n - 1 = 100 - 1 = 99 个人的 IQ 值可以自由变化，因此自由度 df = 99。我们以三个人的数据为例：假设张三、李四、王二的 IQ 的均值为 100，我们在猜测张三、李四的 IQ 得分时可以任意猜测，张三、李四的 IQ 得分可以自由变化，假定张三、李四的 IQ 得分分别为 140、90。那么接下来，王二的 IQ 得分不能任意猜测，王二的 IQ 得分不再自由，因为张三、李四、王二的 IQ 的均值为 100，所以王二的 IQ 得分必须是：100*3 - 140 - 90 = 70。

15 t 分布的自由度

在单样本 t 检验中，df = n - 1，其中 n 为样本量。样本的样本量可能不同，由样本均值与样本无偏标准差估计得到的抽样分布会随着样本量发生变化，样本量越大，由样本均值与样本无偏标准差估计得到的抽样分布会越接近基于总体均值与总体标准差估计得到的抽样分布。

在下图中，金色曲线为 z 分布，蓝色线由深变浅为自由度为 1、4、9 时的 t 分布。可见，与 z 分布相比，样本量越小（df 较小），t 分布的尾部会越厚，这就造成了 z 检验与 t 检验的 p 值的差异。

```
ggplot(data.frame(x = seq(-4, 4, 0.1)), aes(x)) +
  stat_function(fun = dnorm, color = "gold", linewidth = 2) +
  stat_function(fun = dt, args = list(df = 1),
                color = "slateblue4", linewidth = 1) +
  stat_function(fun = dt, args = list(df = 4),
                color = "slateblue3", linewidth = 1) +
  stat_function(fun = dt, args = list(df = 9),
                color = "slateblue2", linewidth = 1) +
  theme_bw()
```



本文到此结束。