

总体分布，样本分布，抽样分布

Polulation, sample, and sampling distributions

Qingyao Zhang

2026-05-23

目录

1 总体	1
2 总体分布	2
3 取样过程	3
4 样本 1	4
5 基于样本 1 估计总体均值与标准差	4
6 样本分布：样本 1 的分布	5
7 样本分布与总体分布的关系	5
8 重复取样	6
9 抽样分布：1000 个样本的均值的均值与标准差	6
10 抽样分布	7
11 抽样分布与总体分布、样本分布的关系	8
12 三种分布的比较	8

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。

```

1 # Population, 总体
2 # 总体均值设置为 100
3 mu <- 100
4 # 总体标准差设置为 15
5 sigma <- 15
6 # 随机种子值设置为 2026
7 set.seed(202605)
8 # 将 ID 与 IQ 存入数据表中
9 population10000 <- data.frame(
10   # 生成 10000 个 ID (被试编号)
11   ID = seq(1:10000),
12   # 生成 10000 个均值为 mu、标注差为 delta 的 IQ 分数
13   IQ = round(rnorm(10000, mean = mu, sd = sigma), 0)
14 )
15 # 查看数据表
16 # View(population10000)
17 # 查看数据表前 6 行
18 range(population10000$IQ)
19 ## [1] 38 156
20 head(population10000)
21 ##   ID  IQ
22 ## 1  1 126
23 ## 2  2  94
24 ## 3  3  93
25 ## 4  4 119
26 ## 5  5 104
27 ## 6  6  81
28 # 总体的实际均值
29 mean(population10000$IQ)
30 ## [1] 99.8009
31 # 总体的实际标准差
32 sqrt(sum((population10000$IQ - mean(population10000$IQ))^2)/10000)
33 ## [1] 14.94354

```

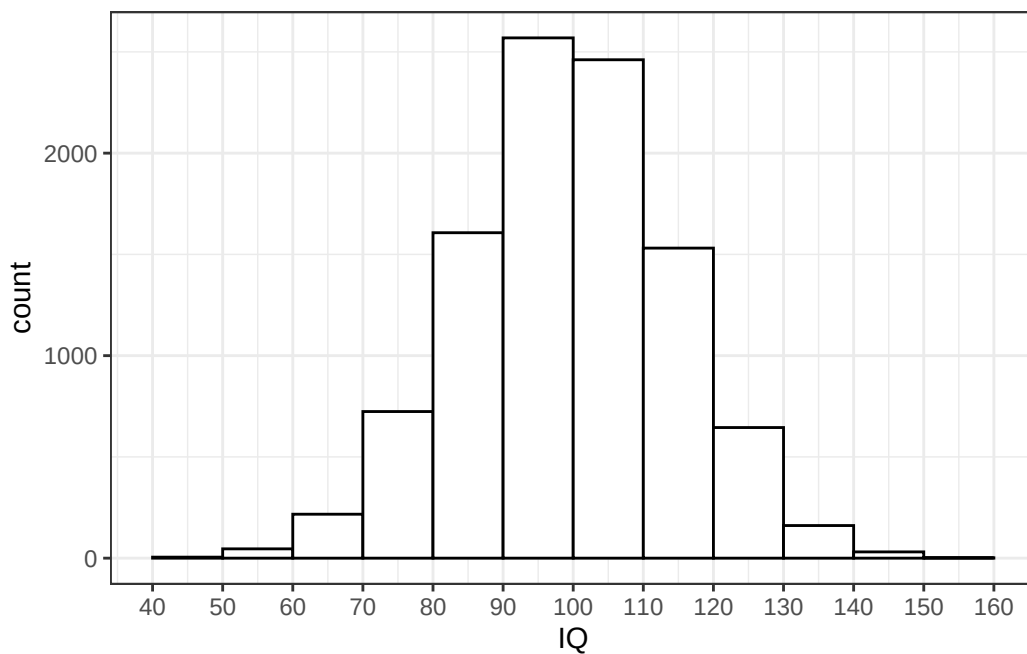
2 总体分布

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。

```

1 # 作总体的直方图
2 # 总体分布是 10000 个数据点的分布。
3 library(ggplot2)
4 ggplot(population10000, aes(x = IQ)) +
5   geom_histogram(breaks = seq(40, 160, 10),
6                 fill = "white",
7                 color = "black") +
8   scale_x_continuous(breaks = seq(40, 160, 10)) +
9   coord_cartesian(xlim = c(40, 160)) +
10  theme_bw()

```



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`。

```

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

```

4 样本 1

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。

```
1 # 查看样本
2 head(sample1)
3 ##           ID  IQ
4 ## 1553 1553 116
5 ## 7307 7307  98
6 ## 5822 5822  90
7 ## 6361 6361 104
8 ## 2752 2752 124
9 ## 242   242  99
```

5 基于样本 1 估计总体均值与标准差

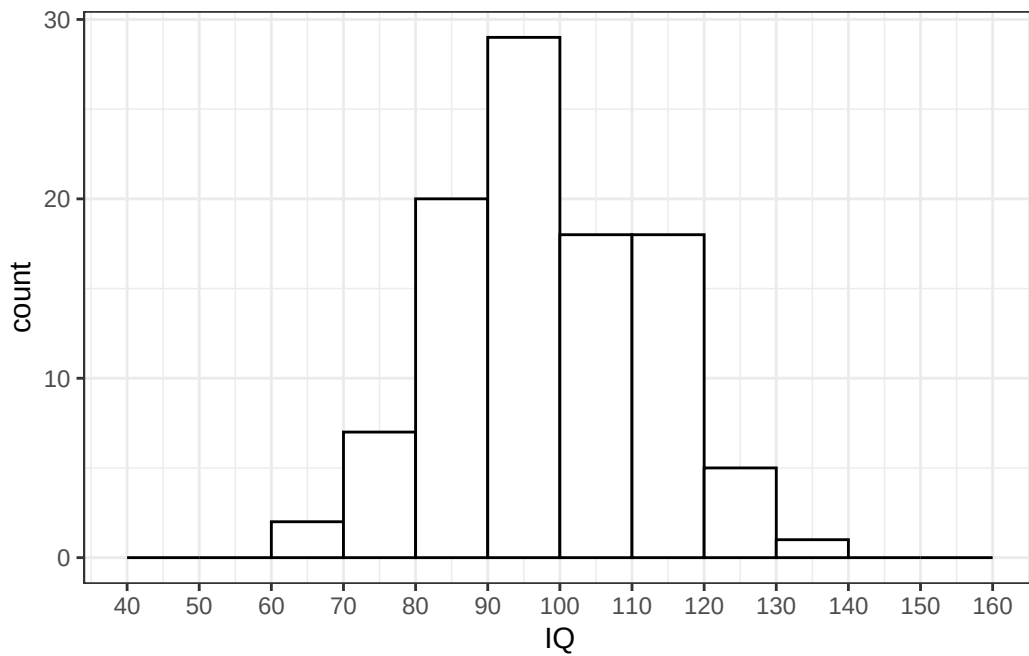
样本的均值是总体均值的无偏估计。样本的标准差总是小于总体的标准差，因此使用样本的无偏标准差作为总体标准差的无偏估计。尽管名为“无偏”，实际上还是有偏差的。

```
1 # 取值范围
2 range(sample1$IQ)
3 ## [1]  67 139
4 # 计算 sample1 的 IQ 的均值
5 sample1_mean <- mean(sample1$IQ)
6 sample1_mean
7 ## [1] 98.93
8 # 计算 sample1 的 IQ 的标准差
9 sample1_sd <- sqrt(sum((sample1$IQ - sample1_mean)^2)/100)
10 sample1_sd
11 ## [1] 14.04511
12 # sample1 的无偏标准差（基于 sample1 估计总体的标准差）
13 sample1_sd_unbiased <- sqrt(sum((sample1$IQ - sample1_mean)^2)/(100 - 1))
14 sample1_sd_unbiased
15 ## [1] 14.11587
16 # 使用`sd()`函数计算 sample1 的无偏标准差（总体标准差的估计值）
17 sd(sample1$IQ)
18 ## [1] 14.11587
```

6 样本分布：样本 1 的分布

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

```
1 ggplot(sample1, aes(IQ)) +  
2   geom_histogram(breaks = seq(40, 160, 10),  
3                 fill = "white",  
4                 color = "black") +  
5   scale_x_continuous(breaks = seq(40, 160, 10)) +  
6   coord_cartesian(xlim = c(40, 160)) +  
7   theme_bw()
```



7 样本分布与总体分布的关系

若总体的均值 μ 与标准差 σ 是未知的，我们可以用样本的均值与无偏标准差来估计总计的均值与标准差：

$$\mu = M_{sample} \quad (1)$$

$$\sigma = s \quad (2)$$

在上式中, M_{sample} 是样本的均值, s 是样本的无偏标准差。

8 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。

```
1 # 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
2 samples <- list()
3 for (index in 1:1000) samples[[index]] <- population10000[sample.int(10000, 100), ]
4 # 查看 samples
5 length(samples)
6 ## [1] 1000
7 head(samples[[1]])
8 ##      ID  IQ
9 ## 989   989 107
10 ## 194   194 115
11 ## 3489 3489 104
12 ## 4818 4818 144
13 ## 7487 7487 122
14 ## 3070 3070 117
```

9 抽样分布: 1000 个样本的均值的均值与标准差

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。

```
1 # 分别计算 1000 个样本的均值, 一共得到 1000 个均值, 存入 samples_means
2 samples_means <- sapply(samples, function(x) mean(x$IQ))
3 head(samples_means)
4 ## [1] 101.21 98.14 100.45 100.09 100.32 98.57
5 range(samples_means)
6 ## [1] 95.35 104.40
7 head(samples_means)
8 ## [1] 101.21 98.14 100.45 100.09 100.32 98.57
9
10 # 1000 个实际样本的均值的均值。
11 mean(samples_means)
12 ## [1] 99.8118
```

```

13 # 1000 个实际样本的均值的方差。
14 sum((samples_means-mean(samples_means))^2)/1000
15 ## [1] 2.023013
16 # 1000 个实际样本的均值的标准差。
17 sqrt(sum((samples_means-mean(samples_means))^2)/1000)
18 ## [1] 1.422327

```

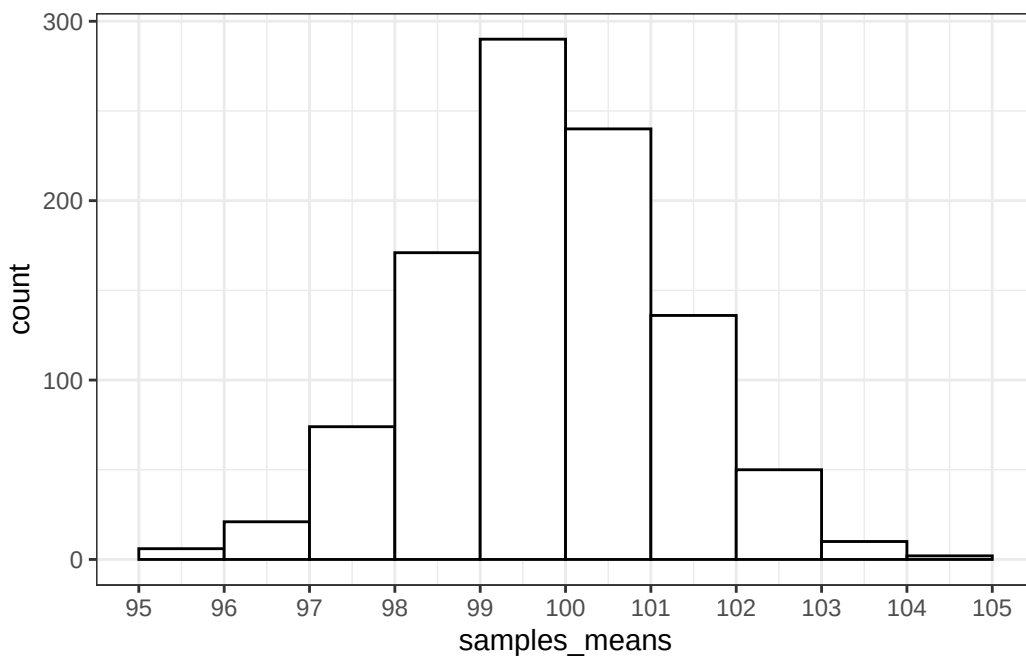
10 抽样分布

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 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。

```

1 # 抽样分布。即，1000 个样本的均值的分布。
2 ggplot(data.frame(samples_means), aes(samples_means)) +
3   geom_histogram(breaks = seq(95, 105, 1),
4                 fill = "white",
5                 color = "black") +
6   scale_x_continuous(breaks = seq(95, 105, 1)) +
7   theme_bw()

```



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

11 抽样分布与总体分布、样本分布的关系

若总体的均值 μ 与标准差 σ 是已知的:

$$M_{sampling} = \mu \quad (3)$$

$$SD_{sampling} = \frac{\sigma}{\sqrt{n}} \quad (4)$$

在上式中, n 是样本量。

若总体的均值 μ 与标准差 σ 是未知的, 我们可以用样本的均值与无偏标准差来估计总计的均值与标准差:

$$M_{sampling} = M_{sample} \quad (5)$$

$$s_{sampling} = \frac{s}{\sqrt{n}} \quad (6)$$

在上式中, M_{sample} 是样本的均值, s 是样本的无偏标准差。

12 三种分布的比较

比较三种分布的频数分布图:

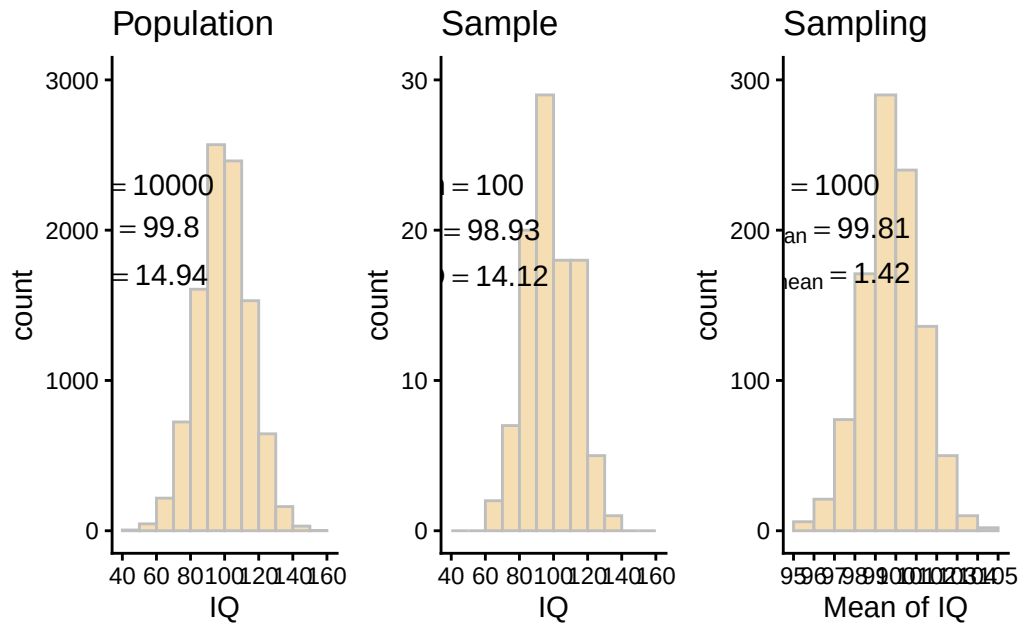
```
1 library(patchwork)
2 plot_pop <- ggplot(population10000, aes(x = IQ)) +
3   geom_histogram(breaks = seq(40, 160, 10),
4     fill = "wheat",
5     color = "grey") +
6   scale_x_continuous(breaks = seq(40, 160, 20)) +
7   coord_cartesian(xlim = c(40, 160), ylim = c(0, 3000)) +
8   labs(title = "Population") +
9   annotate(geom = "text",
10     label = c("N == 10000", "mu == 99.80", "sigma == 14.94"),
11     x = 55, y = c(2300, 2000, 1700), parse = TRUE) +
```



```

12   theme_classic()
13 plot_sample <- ggplot(sample1, aes(IQ)) +
14   geom_histogram(breaks = seq(40, 160, 10),
15                 fill = "wheat",
16                 color = "grey") +
17   scale_x_continuous(breaks = seq(40, 160, 20)) +
18   coord_cartesian(xlim = c(40, 160), ylim = c(0, 30)) +
19   labs(title = "Sample") +
20   annotate(geom = "text",
21           label = c("n == 100", "M == 98.93", "SD == 14.12"),
22           x = 55, y = c(23, 20, 17), parse = TRUE) +
23   theme_classic()
24 plot_sampling <- ggplot(data.frame(samples_means), aes(samples_means)) +
25   geom_histogram(breaks = seq(95, 105, 1),
26                 fill = "wheat",
27                 color = "grey") +
28   scale_x_continuous(breaks = seq(95, 105, 1)) +
29   coord_cartesian(ylim = c(0, 300)) +
30   labs(title = "Sampling", x = "Mean of IQ") +
31   annotate(geom = "text",
32           label = c("m == 1000", "M[mean] == 99.81", "SD[mean] == 1.42"),
33           x = 96.3, y = c(230, 200, 170), parse = TRUE) +
34   theme_classic()
35 plot_pop + plot_sample + plot_sampling

```



```
1 ggsave("compare3distributions.png", width = 12, height = 3)
```

本文到此结束。