计量经济学

第三讲：统计学复习

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

办公室　粤海校区汇文楼2613
E-mail　huangjp@szu.edu.cn
Website　https://huangjp.com

主要内容

- 估计
  - 估计量与估计值
  - 估计量的性质
  - 最小二乘估计量
- 推断：假设检验
  - 原假设和备择假设
  - p 值
- 推断：置信区间
- 结果的展示：论文中的图与表

统计学的作用

- 统计学是数据科学。统计学包含对数据的收集、描述、分析、以及由此得出的结论。
- 典型问题：硕士毕业生的平均年薪是多少?
  - 无法针对总体（population）进行全数调查
  - 利用合适的样本（sample），可以在一定程度上了解总体的分布特征
  - 简单随机抽样（simple random sampling）：总体的每一个成员都有同样的机会被选中
- 统计学可以帮助我们回答关于总体分布的未知特征值的问题
- 统计学的几个重要主题:
  - **估计（estimation）**：通过样本猜测总体的特征值
  - **推断（inference）**：假设检验和置信区间
  - 仿真（simulation）

估计

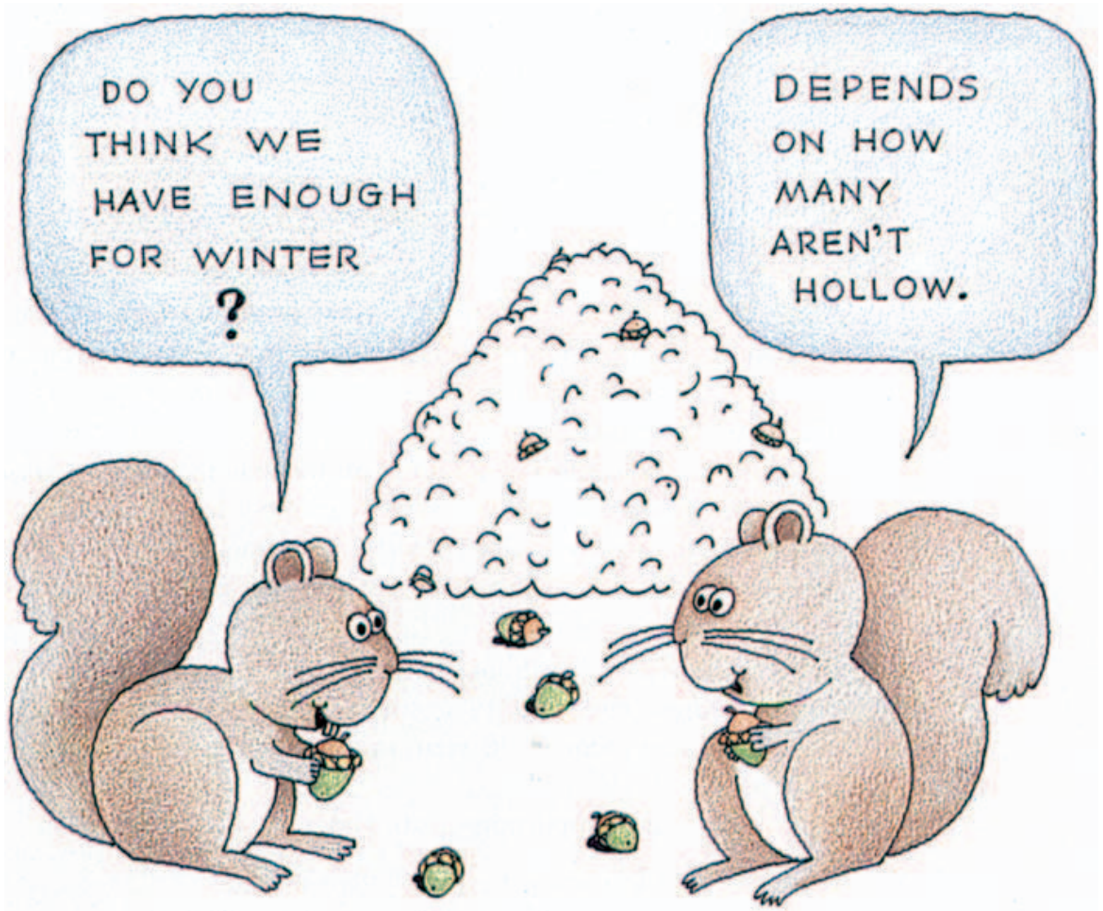
估计：最佳猜测值

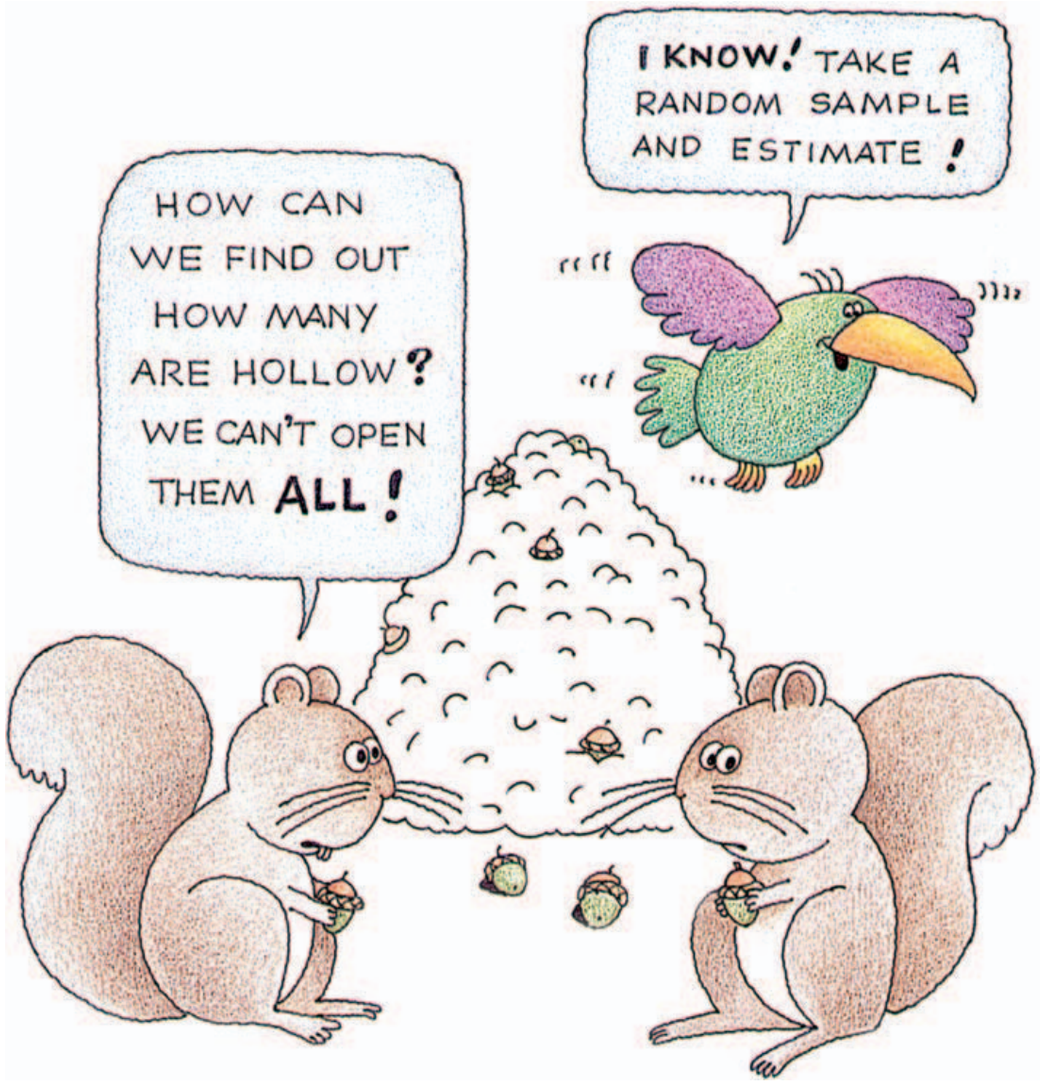
Estimation

- 假设你想了解总体中某个变量 Y 的均值 μ_Y。你可以获得 n 个随机样本 $Y_1, Y_2, \ldots, Y_n$，但是无法获取总体中全部成员所对应的 Y 值。

- 估计的目的就是利用样本，对总体的特征值做出的“最佳”猜测。针对总体均值 μ_Y，一个自然的猜测是样本均值 $\overline{Y}$。然而同时也存在其他可能的猜测，如样本中第一个观测值 Y_1。

估计量（estimator）是从总体中随机抽取的样本数据的函数。**估计值（estimate）**是基于某一特定的样本数据实际计算得出的估计量的数值。由于抽样的随机性，估计量是一个随机变量，而估计值是一个非随机的数。





Pictures: Ross, S., Introductory Statistics, 3rd Edition, 2010, Academic Press.

什么是好的估计量?

- 一个好的估计量应当满足以下三个性质：
  - **无偏性（unbiasedness）**：在均值意义上估计正确
  - **一致性（consistency）**：样本增大时估计得更准确
  - **有效性（efficiency）**：分布更集中的更有效

设 $\hat{\mu}_Y$ 为 μ_Y 的一个估计量。

- 若 $E(\hat{\mu}_Y) = \mu_Y$，则 $\hat{\mu}_Y$ 是 μ_Y 的一个**无偏估计量**（unbiased estimator）。
- 若 $\hat{\mu}_Y \xrightarrow{p} \mu_Y$，则 $\hat{\mu}_Y$ 是 μ_Y 的一个**一致估计量**（consistent estimator）。
- 令 $\tilde{\mu}_Y$ 为 μ_Y 的另一个估计量，且两者都无偏。若 $\mathrm{var}(\hat{\mu}_Y) < \mathrm{var}(\tilde{\mu}_Y)$，则称 $\hat{\mu}_Y$ 比 $\tilde{\mu}_Y$ **有效**（efficient）。

$\overline{Y}$ 的性质

- $\overline{Y}$ 是无偏的：$E(\overline{Y}) = \mu_Y$
- $\overline{Y}$ 是一致的：由大数定律可得，$\overline{Y} \xrightarrow{p} \mu_Y$ 。
- $\overline{Y}$ 是 BLUE 的：**BLUE** (**B**est **L**inear **U**nbiased **E**stimator)，即**最佳线性无偏估计量**。在所有 $Y_1, Y_2, \ldots, Y_n$ 加权平均类无偏估计量中，$\overline{Y}$ 是最有效的。（Gauss-Markov定理，后述）
- $\overline{Y}$ 是 μ_Y 的**最小二乘估计量（least squares estimator）**：

  若一个估计量 m 使 $\sum_{i=1}^{n} (Y_i - m)^2$ 的值最小，则称 m 为最小二乘估计量。

证明 $\overline{Y}$ 是 μ_Y 的最小二乘估计量

对 $\sum_{i=1}^{n}(Y_i - m)^2$ 进行求导，可得

$$\begin{aligned}\frac{d}{dm}\sum_{i=1}^{n}(Y_i - m)^2 &= \frac{d}{dm}\sum_{i=1}^{n}(Y_i^2 - 2Y_i m + m^2)\\ &= \sum_{i=1}^{n}(-2Y_i + 2m)\\ &= -2\sum_{i=1}^{n}Y_i + 2nm\end{aligned}$$

最小化的一阶条件为 $-2\sum_{i=1}^{n}Y_i + 2nm = 0$，由此可得出当

$$m = \frac{1}{n}\sum_{i=1}^{n}Y_i = \overline{Y}$$

时，$\sum_{i=1}^{n}(Y_i - m)^2$达到最小。

随机抽样的重要性

- 我们假定 $Y_1, Y_2, \ldots, Y_n$ 为简单随机抽样的结果，因此是 i.i.d.。如果样本不是随机的，则有可能造成估计偏差，以至于估计结果失去应有的效果。这被称为选择偏差（selection bias）。

- 例1：同学 A 在做毕业论文研究时需要对深圳市高层次人才进行问卷调查，他设计好网络问卷并委托认识的老师（被选为深圳市高层次人才）在朋友圈中扩散。这样得到的调查结果是有偏的，因为回答问卷的人可能多数是老师们的朋友，因此在工作性质上（偏向学术界）和工作领域上（偏向老师所在专业）无法满足随机抽样。

- 例2：记者在春运的火车上询问乘客：“你是否顺利地买到了回家的火车票？”

推断：假设检验

假设

Hypothesis

- 假设是对总体分布的某些特征值提出的可以用“是”或“否”回答的问题。

- **原假设（null hypothesis）**：要检验的假设，例如

$$H_0$$：总体均值等于 0，即 $E(Y) = 0$

- **备择假设（alternative hypothesis）**：当原假设不成立时成立的假设，例如

$$H_1$$：总体均值不等于 0，即 $E(Y) \neq 0$

假设检验

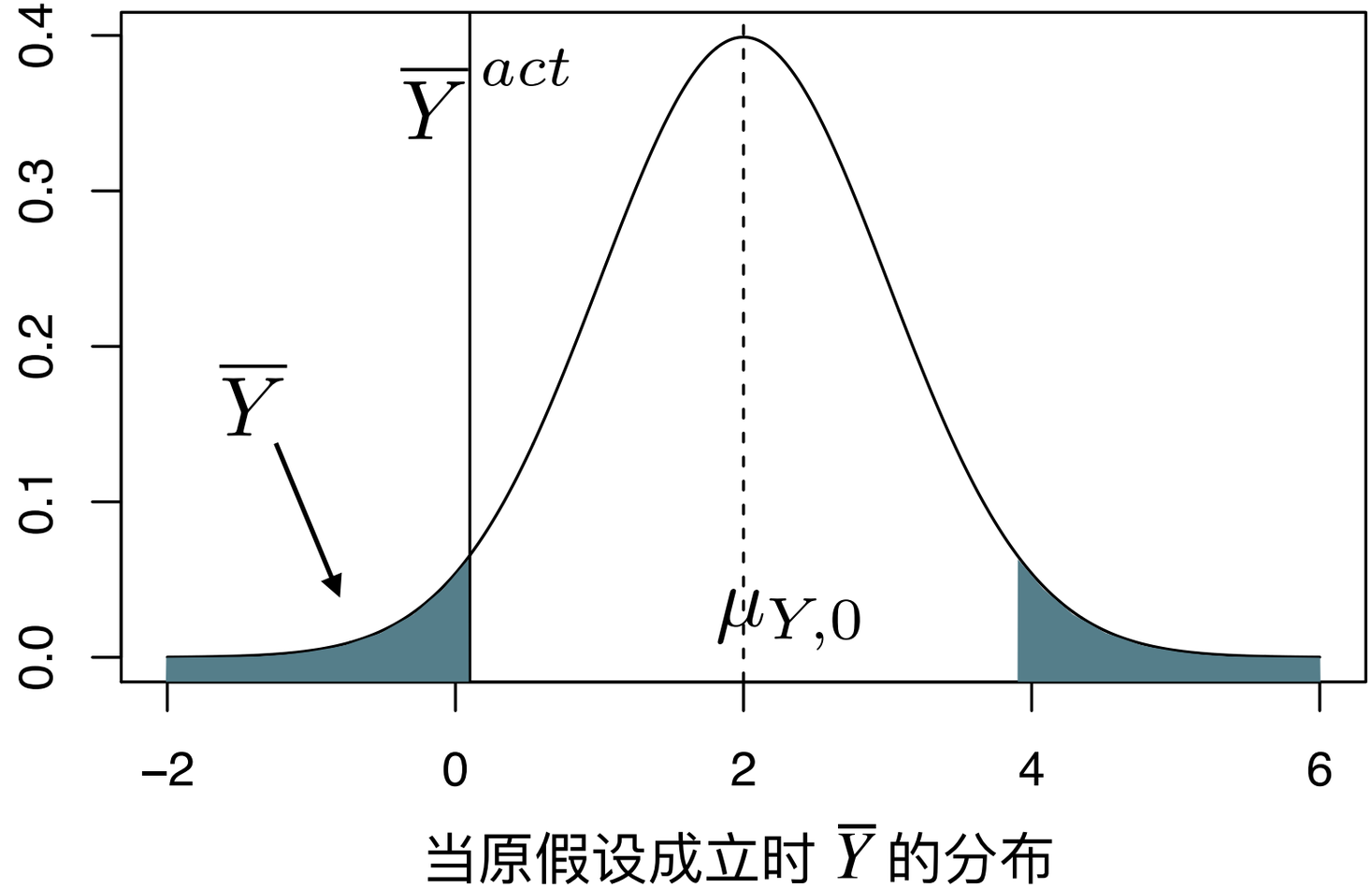
- 原假设可以通过随机抽样的样本数据进行检验，即拒绝或无法拒绝。
- 拒绝原假代表有证据支持原假设是错误的，此时接受备择假设。
- “接受”原假设仅代表暂时还没找到否定它的证据，并不证明它是正确的。

因此我们只有两种结论：
拒绝原假设，或没有足够拒绝原假设的证据

p 值

- p 值，也称为**显著性概率（significance probability）**，是在原假设成立的情况下，抽样得到的统计量与原假设之间的距离至少等于其样本计算值与原假设之间的距离的概率。

$$H_0 : \mathrm{E}(Y) = \mu_{Y,0}$$



当原假设成立时 $\overline{Y}$ 的分布

p 值

- 令 $\overline{Y}^{act}$ 表示利用手头的数据实际计算出的样本均值，$\Pr_{H_0}$ 表示原假设成立时计算出来的概率（即在计算中假定 $E(Y_i) = \mu_{Y,0}$），则

$$p \text{ 值} = \Pr_{H_0}\left[|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{act} - \mu_{Y,0}|\right]$$

即 p 值是在原假设下 $\overline{Y}$ 的分布位于 $|\overline{Y}^{act} - \mu_{Y,0}|$ 之外的尾部面积。

- p 值较大则观测到的 $\overline{Y}^{act}$ 与原假设相符，p 值较小则不相符。当 p 值非常小时，若原假设成立，则很难从总体中抽样观测到 $\overline{Y}^{act}$，但我们观测到了，因此原假设不成立的可能性很大。

p 值的计算：总体方差已知时

- 计算 p 值需要知道 $\overline{Y}$ 的分布。当样本容量较大时，根据中心极限定理，在原假设下 $\overline{Y} \sim N\left(\mu_{Y,0}, \sigma_Y^2/n\right)$。
- 标准化后的统计量 $\dfrac{\overline{Y}-\mu_{Y,0}}{\sigma_Y/\sqrt{n}}$ 服从标准正态分布，因此

$$
\begin{aligned}
p \text{ 值} &= \Pr_{H_0}\left[\left|\frac{\overline{Y}-\mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right| > \left|\frac{\overline{Y}^{act}-\mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right] \\
&= 2\Phi\left(-\left|\frac{\overline{Y}^{act}-\mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right)
\end{aligned}
$$

p 值的计算：总体方差未知时

- 在实践中，往往无法知道总体方差 σ_Y^2 。此时我们需要先估计总体方差。
- **样本方差（sample variance）** s_Y^2 是总体方差的一致估计量：

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 , \qquad s_Y^2 \xrightarrow{p} \sigma_Y^2$$

- $\overline{Y}$ **的标准误（standard error of** $\overline{Y}$ **）** $s_Y/\sqrt{n}$ 可以用作 $\overline{Y}$ 的标准差的估计量：

$$\mathrm{SE}(\overline{Y}) = \hat{\sigma}_{\overline{Y}} = s_Y/\sqrt{n}$$

p 值的计算：总体方差未知时

- 在总体方差 σ_Y^2 未知时，我们可以在 p 值的计算中用 s_Y^2 替代它。此时，

$$p \text{ 值} = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right|\right) = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{\mathrm{SE}(\overline{Y})}\right|\right)$$

- t 统计量：

$$t = \frac{\overline{Y} - \mu_{Y,0}}{\mathrm{SE}(\overline{Y})}$$

t 统计量的大样本分布

- 由中心极限定理可知，在原假设下，

$$n \text{ 较大时 } t \text{ 近似服从标准正态分布 } N(0,1)$$

- p 值可以用 t 统计量来表示。令 t^{act} 表示实际计算的得到的 t 统计量值，即

$$t^{act} = \frac{\overline{Y}^{act} - \mu_{Y,0}}{\mathrm{SE}(\overline{Y})}$$

则

$$p \text{ 值} = 2\Phi(-|t^{act}|)$$

总体均值的假设检验的一般步骤

1. 确定原假设和备择假设

$$H_0 : E(Y) = \mu_{Y,0} \quad H_1 : E(Y) \neq \mu_{Y,0}$$

2. 根据样本计算 t^{act} 和 p 值
3. 选择**显著水平（significance level）**，例如 5%

   显著水平为可容忍的第一类错误（原假设为真时拒绝了原假设）的概率
4. 如果 p 值 $<$ 显著水平（或 $|t^{act}| > 1.96$），则拒绝原假设。

需要注意的是，t^{act} 随样本容量增加而变大。因此样本越大，在同一显著水平下原假设越容易被拒绝。在大样本时，我们不仅要考虑5%显著水平是否合适，也要考虑 $\bar{Y}^{act}$ 本身是否具有现实意义。

不同总体的均值比较

- 大学刚毕业的男生和女生的平均收入相同吗？这一问题涉及了两个不同的总体均值的比较。

- 均值之差的假设检验

$$H_0 : \mu_m - \mu_w = d_0 \quad H_1 : \mu_m - \mu_w \neq d_0$$

- $\overline{Y}_m - \overline{Y}_w$ 是 $\mu_m - \mu_w$ 的估计量

- $\overline{Y}_m - \overline{Y}_w$ 的标准误

$$\mathrm{SE}(\overline{Y}_m - \overline{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$$

因果关系的测量：条件期望之差

- 在随机对照试验中，处理组（$X = 1$）的期望值 $E(Y \mid X = 1)$ 与对照组（$X = 0$）的期望值 $E(Y \mid X = 0)$ 之差被称作平均处理效应（average treatment effect），可以用来衡量处理 X 对随机变量 Y 的因果效应（causal effect）。
- 可以用处理组和对照组的样本平均之差来估计因果效应。

$$H_0 : E(Y \mid X = 1) - E(Y \mid X = 0) = 0$$
$$H_1 : E(Y \mid X = 1) - E(Y \mid X = 0) \neq 0$$

推断：置信区间

置信区间

Confidence interval

- 描述总体均值一定需要先设置假设然后再检验吗？不需要。

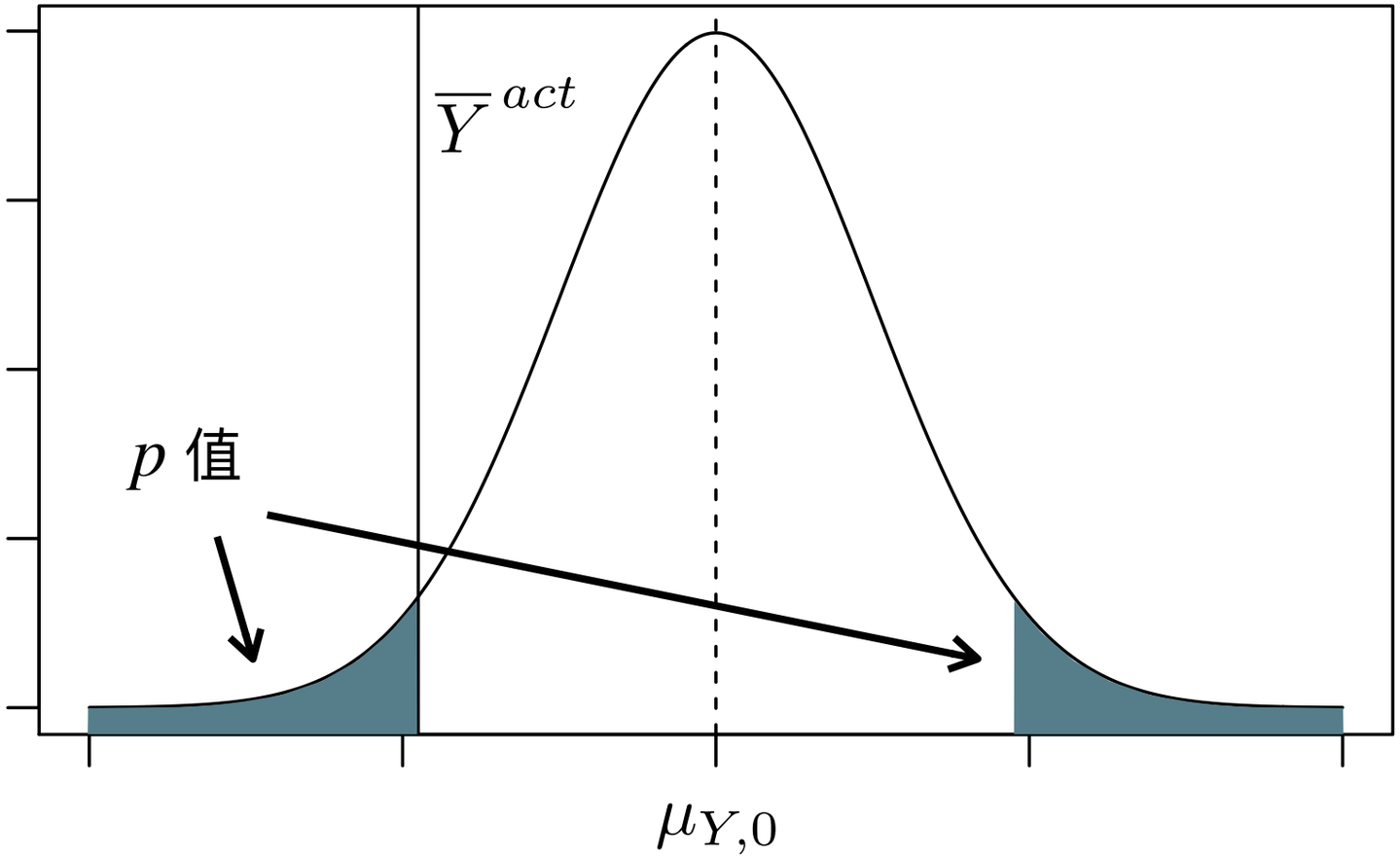
- 我们可以利用随机样本数据构造一个数值集合，使其以指定概率包含真实的总体均值 μ_Y 。这样的集合被称为**置信集（confidence set）**，指定概率被称作**置信水平（confidence level）**。

- μ_Y 的置信集是均值在上限和下限之间所有可能的取值，因此置信集是实数轴上的一个区间，也被称为**置信区间（confidence interval）**。

置信区间与 *p* 值

p 值

$$p\text{-value} = \Pr_{H_0}\left[|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{act} - \mu_{Y,0}|\right]$$

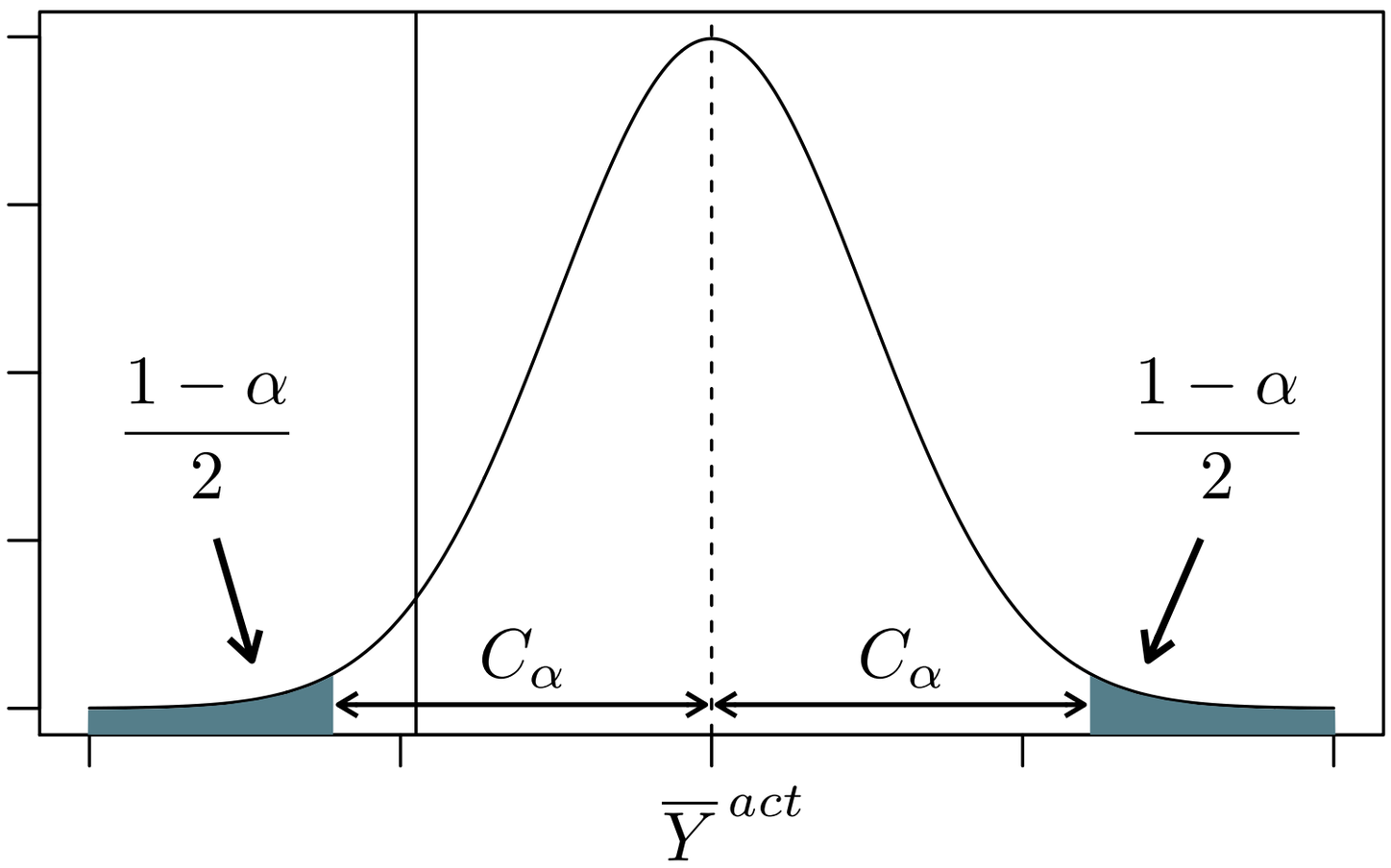


在原假设的前提下（即确定 $\mu_{Y,0}$），计算样本均值 $\overline{Y}^{act}$，得出 p 值

置信区间与 *p* 值

μ_Y 的 α 置信区间的一般表现形式是

$$[\overline{Y} - C_\alpha \times SE(\overline{Y}), \overline{Y} + C_\alpha \times SE(\overline{Y})]$$

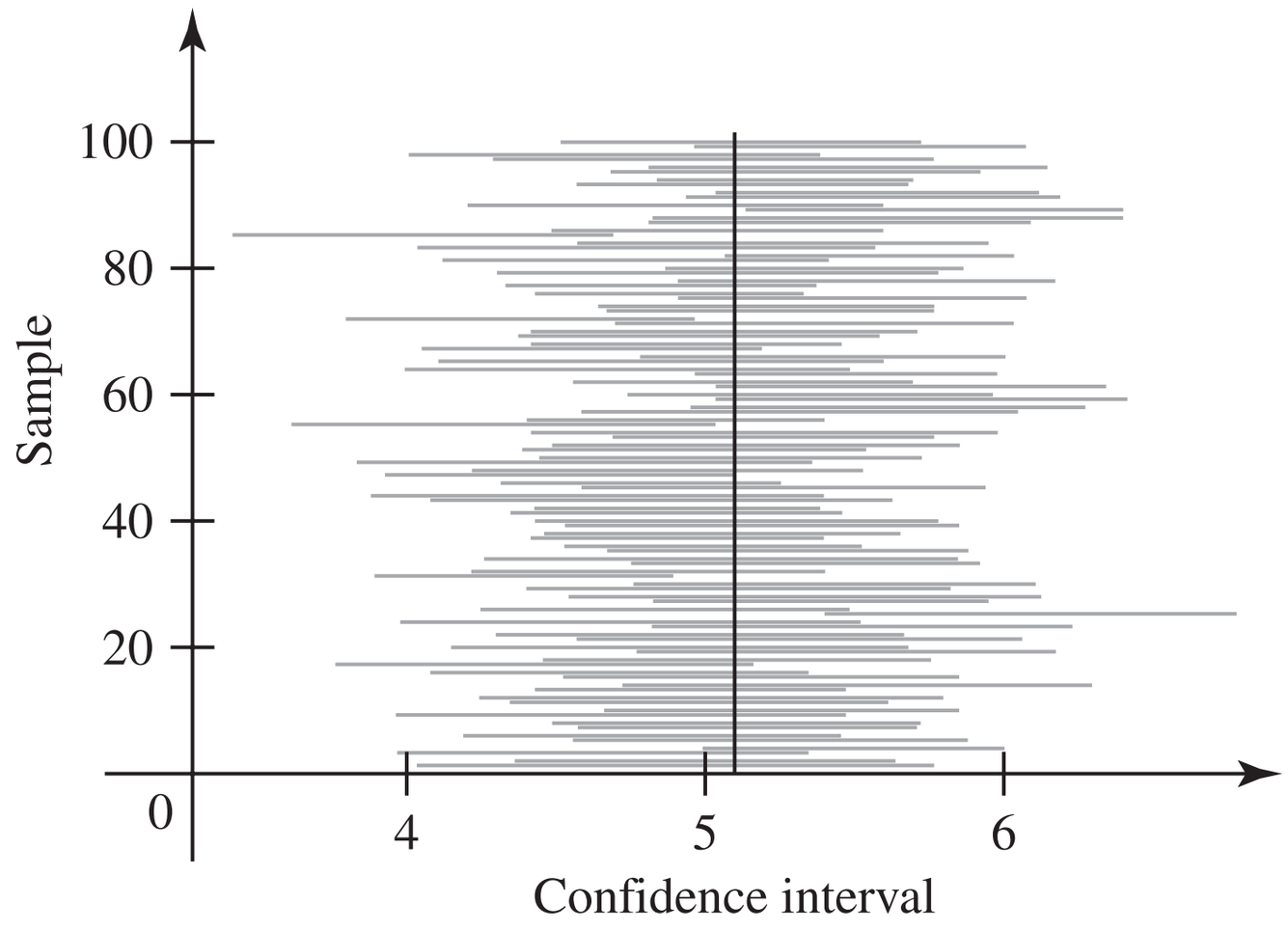


在指定置信水平的前提下（即确定 α），计算样本均值 $\overline{Y}^{act}$，得出置信区间

置信区间的含义

Figure 8.5 in DeGroot & Schervish (2012).

Figure 8.5 A sample of one hundred observed 95% confidence intervals based on samples of size 26 from the normal distribution with mean $\mu = 5.1$ and standard deviation $\sigma = 1.6$. In this figure, 94% of the intervals contain the value of μ.



根据正态分布 $N(\mu = 5.1, \sigma^2 = 1.6^2)$ 随机生成 26 个观测值的样本，然后构造置信区间。
图中包含了100个这样的置信区间，其中94个包含真实的分布均值 $\mu = 1.5$。

大样本下的置信区间

- μ_Y 的 α 置信区间的一般表现形式是

$$[\overline{Y} - C_\alpha \times SE(\overline{Y}), \overline{Y} + C_\alpha \times SE(\overline{Y})]$$

我们需要知道如何计算 C_α 。

- 在样本容量较大时，根据中心极限定理，

$$C_\alpha = \Phi^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) = -\Phi^{-1}\left(\frac{1-\alpha}{2}\right)$$

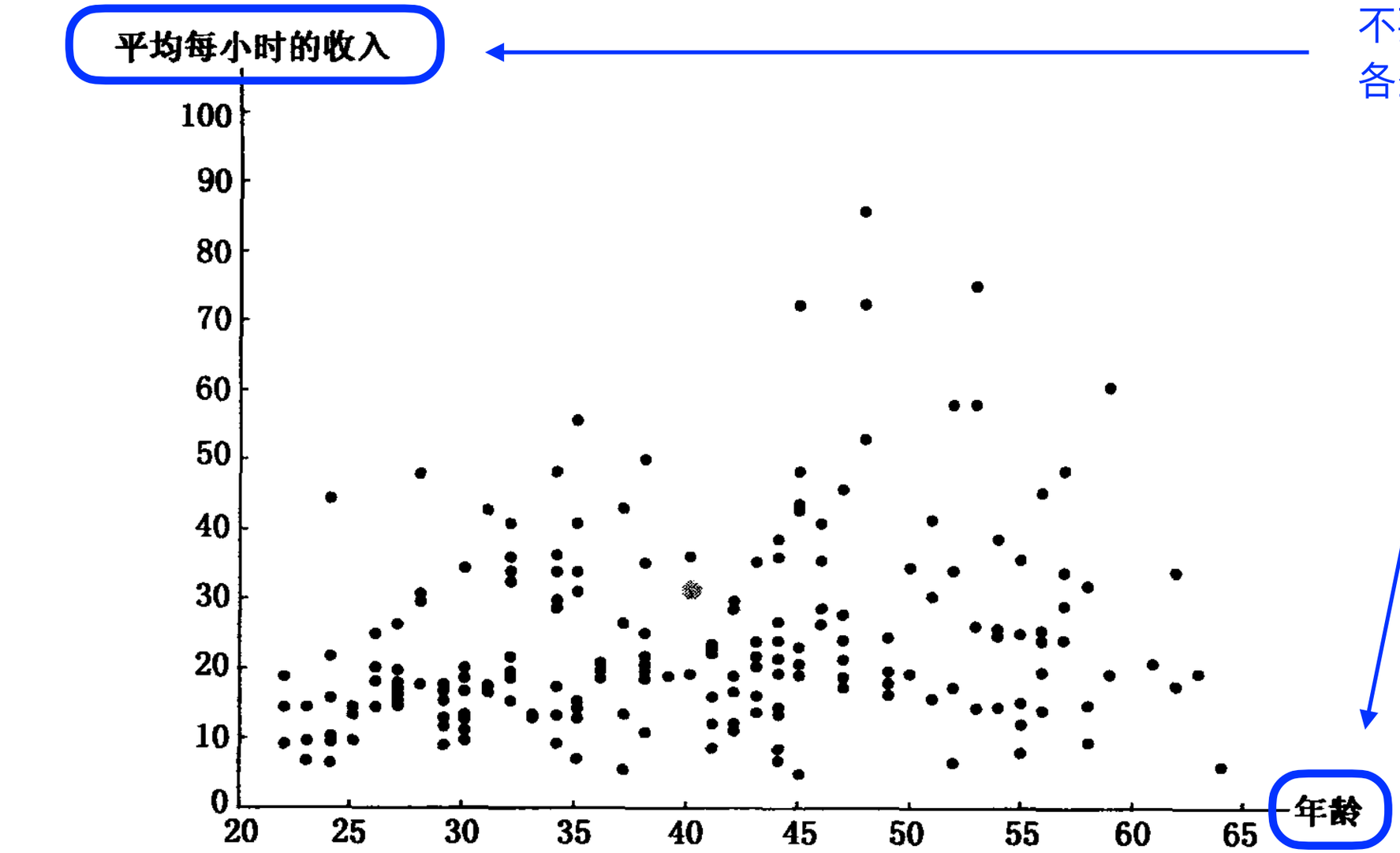
- $C_{95\%} = 1.96, C_{90\%} = 1.64, C_{99\%} = 2.58$

结果的展示：

论文中的图与表

散点图

Scatter plot



不要忘记标注
各坐标轴的含义

注:图中的每点表示 200 个工人样本中每个工人的年龄和平均收人。中间的灰色点对应收人为 31.25 美元/时的 40 岁的工人。数据来源于 2005 年 3 月的 CPS。

图 3.2　平均每小时收入对年龄的散点图

图的标题放在图下面

表

Table

表的标题放在表上面

表 9.1　加利福尼亚和马萨诸塞州测试成绩数据集的概括统计量

	加利福尼亚州		马萨诸塞州	
	均　值	标准差	均　值	标准差
测试成绩	654.1	19.1	709.8	15.1
学生/教师比	19.6	1.9	17.3	2.3
英语学习者(%)	15.8	18.3	1.1	2.9
接受午餐资助(%)	44.7	27.1	15.3	15.1
地区平均收入(美元)	15 317	7 226	18 747	5 808
观测个数	420		220	
年份	1999		1998	

三条主横线必须有，其他辅助横线可适当添加。
不应该出现竖线（例外如长表折返为两列时）

Tax Burden in Selected Countries*

Country	Income tax (%)	Social Security (%)	Total payment† (%)	Country	Income tax (%)	Social Security (%)	Total payment† (%)
Denmark	33	11	43	Czech Republic	11	13	24
Belgium	28	14	41	United States	17	8	24
Germany	21	21	41	United Kingdom	16	8	23
Finland	26	6	32	Iceland	22	0	22
Poland	6	25	31	Luxembourg	8	14	22
Sweden	23	7	30	Switzerland	10	12	22
Turkey	15	15	30	New Zealand	20	0	20
Netherlands	7	22	29	Slovak Republic	7	13	19
Norway	21	8	29	Spain	13	6	19
Austria	11	18	29	Greece	1	16	17
Hungary	17	13	29	Portugal	6	11	17
Italy	19	9	28	Ireland	11	5	16
France	13	13	27	Japan	6	10	16
Canada	19	7	26	Korea	2	7	9
Australia	24	0	24	Mexico	2	2	4

* Does not include taxes not listed, such as sales tax or VAT. Rates shown apply to a single person with average earnings.
† Totals may not add due to rounding.
Source: Organization for Economic Cooperation and Development, 2002.

Table: Ross, S., Introductory Statistics, 3rd Edition, 2010, Academic Press.