抽样调查 第三周作业

20307100013 蒋翌坤

§2.13 Exercises：34

（a）The normal approximation CI is:

（i） $\hat{p} \pm 1.96\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}} = 0.5 \pm 1.96\sqrt{\frac{0.5\times(1-0.5)}{10-1}} = [0.173, 0.827]$

（ii） $0.9 \pm 1.96\sqrt{\frac{0.9\times(1-0.9)}{10-1}} = [0.704, 1.096]$

（iii） $0.5 \pm 1.96\sqrt{\frac{0.5\times(1-0.5)}{100-1}} = [0.402, 0.598]$

（iv） $0.9 \pm 1.96\sqrt{\frac{0.9\times(1-0.9)}{100-1}} = [0.841, 0.959]$

（v） $0.99 \pm 1.96\sqrt{\frac{0.99\times(1-0.99)}{100-1}} = [0.9704, 1.0096]$

（vi） $0.99 \pm 1.96\sqrt{\frac{0.99\times(1-0.99)}{1000-1}} = [0.9838, 0.9962]$

（vii） $0.999 \pm 1.96\sqrt{\frac{0.999\times(1-0.999)}{1000-1}} = [0.99704, 1.00096]$

（b）The above normal approximation CI is given when $\alpha = 2 \times \left(1 - \Phi(1.96)\right) = 0.05$
The Clopper-Pearson CI is:

（i） $\left[\frac{n\hat{p}F_{0.025}(2n\hat{p}, 2(n-n\hat{p}+1))}{n\hat{p}F_{0.025}(2n\hat{p}, 2(n-n\hat{p}+1))+n(1-\hat{p})+1}, \frac{(n\hat{p}+1)F_{0.975}(2(n\hat{p}+1), 2n(1-\hat{p}))}{(n\hat{p}+1)F_{0.975}(2(n\hat{p}+1), 2n(1-\hat{p})+n(1-\hat{p})}\right] = [0.187, 0.813]$

（ii） $\left[\frac{n\hat{p}F_1}{n\hat{p}F_1+n(1-\hat{p})+1}, \frac{(n\hat{p}+1)F_2}{(n\hat{p}+1)F_2+n(1-\hat{p})}\right] = [0.555, 0.997]$

（iii） $\left[\frac{n\hat{p}F_1}{n\hat{p}F_1+n(1-\hat{p})+1}, \frac{(n\hat{p}+1)F_2}{(n\hat{p}+1)F_2+n(1-\hat{p})}\right] = [0.398, 0.602]$

（iv） $\left[\frac{n\hat{p}F_1}{n\hat{p}F_1+n(1-\hat{p})+1}, \frac{(n\hat{p}+1)F_2}{(n\hat{p}+1)F_2+n(1-\hat{p})}\right] = [0.824, 0.951]$

（v） $\left[\frac{n\hat{p}F_1}{n\hat{p}F_1+n(1-\hat{p})+1}, \frac{(n\hat{p}+1)F_2}{(n\hat{p}+1)F_2+n(1-\hat{p})}\right] = [0.9455, 0.9997]$

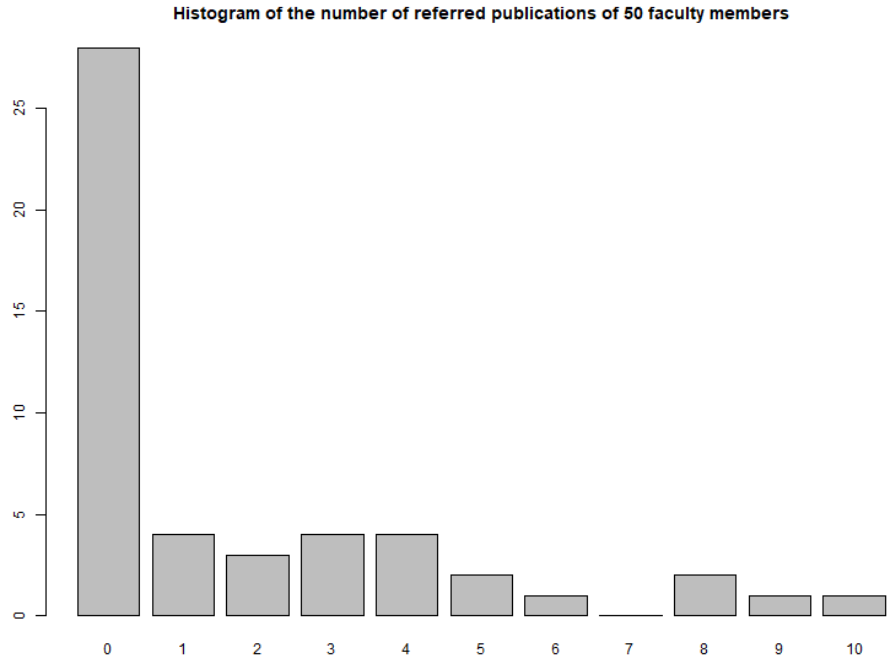
（vi） $\left[\frac{n\hat{p}F_1}{n\hat{p}F_1+n(1-\hat{p})+1}, \frac{(n\hat{p}+1)F_2}{(n\hat{p}+1)F_2+n(1-\hat{p})}\right] = [0.9817, 0.9952]$

（vii） $\left[\frac{n\hat{p}F_1}{n\hat{p}F_1+n(1-\hat{p})+1}, \frac{(n\hat{p}+1)F_2}{(n\hat{p}+1)F_2+n(1-\hat{p})}\right] = [0.99444, 0.99997]$

（c）In situation (i), (iii), (iv), (vi), the normal approximation CI is similar to Clopper-Pearson CI. The upper bond of the normal approximation CI for situations (ii), (v) and (vii) exceeds 1, which is impossible given that p ranges from 0 to 1.

§ 2.13 Exercises: 6

（a） Histogram of the data is in the follow graph. The data is skewed heavily to the left. Most faculty's referred publication is 0. Number of referred publication from 1 to 10 is distributed similarly.



（b）Using sample mean to estimate mean number of publications per faculty member.

$\bar{y} = \frac{1}{50}(0 \times 28 + 1 \times 4 + 2 \times 3 + 3 \times 4 + 4 \times 4 + 5 \times 2 + 6 \times 1 + 7 \times 0 + 8 \times 2 + 9 \times 1 + 10 \times 1) = 1.78$. $SE(\bar{y}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}} = \sqrt{\left(1 - \frac{50}{807}\right) \times \frac{7.34}{50}} = 0.367$

Therefore, mean number of publications per faculty member is 1.78 and SE is 0.367.

（c）$\bar{y}$ is not approximately normally distributed. Even though sample size $n = 50$ is sufficiently large enough for approximation, the data is highly skewed. Using Sugden et al. (2000) to get a minimum sample size, $n_{\min} = 28 + 25 \times G^2 > 50$, where $G = 1.593$ is the skewness of the sample (to estimate the skewness of the population).

（d）We can use normal approximation CI to estimate the proportion of faculty members with no publications. $\hat{p} \pm 1.96\sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}} = \frac{28}{50} \pm 1.96 \times \sqrt{(1 - \frac{50}{807})\frac{\frac{28}{50} \times (1 - \frac{28}{50})}{50-1}} = [0.425, 0.695]$ given 95% CI.

§ 2.13 Exercises: 7

（a）$\hat{p} \pm 1.96\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}} = 0.175 \pm 1.96 \times \sqrt{\frac{0.175\times(1-0.175)}{1000-1}} = [0.151, 0.199]$ Therefore, 95% CI for the percentage of entries that comes from South is $[0.151, 0.199]$.

（b）30% is larger than the upper limit of 95% CI for the South percentage. Therefore, the percentage of entries from the South is evidently smaller than the percentage of persons living in the South.

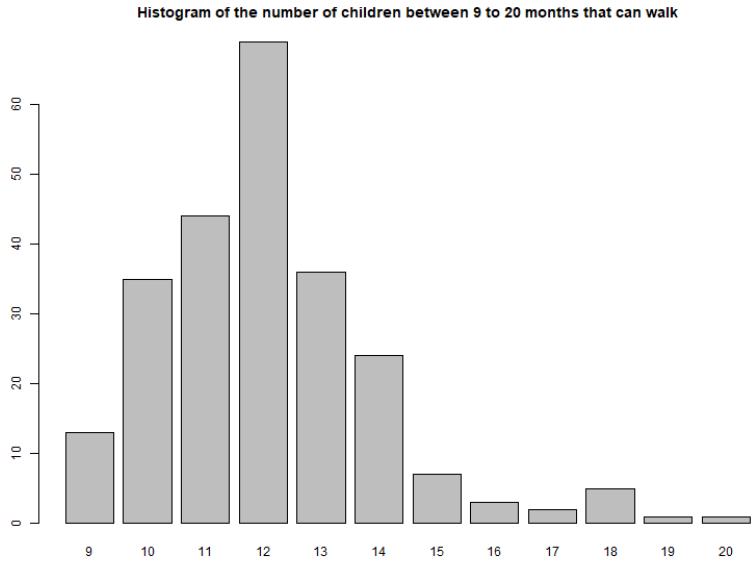
§ 2.13 Exercises: 14

（a）$\hat{p} \pm 1.96\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}} = \frac{23}{30} \pm 1.96 \times \sqrt{\frac{\frac{23}{30}\times(1-\frac{23}{30})}{30-1}} = [0.613, 0.921]$. Therefore, the estimation is $[0.613, 0.921]$ given 95% CI.

（b）The following assumption are needed. (i) The genetic database solely consists of people living in the U.S. (ii) The record of genetic database is collected in a way such that it can represent a typical person in U.S. because 1.28 million is less than the number of U.S. population.

§ 2.13 Exercises: 15

（a）Histogram of the distribution of age at walking is in the follow graph. The data is not normally distributed because it is skewed to the left. Sampling distribution of the sample average is normally distributed because Using Sugden et al. (2000) to get a minimum sample size, $n_{\min} = 28 + 25 \times G^2 < 240$, where $G = 1.13$ is the skewness of the sample (to estimate the skewness of the population).



（b）$\bar{y} = 12.08,\ SE(\bar{y}) = 0.124,\ CI_{95\%}: \bar{y} \pm 1.96SE(\bar{y}) = [11.84,12.32]$

§ 2.13 Exercises：23

（a）The sampling weight for each record is $w = \frac{N}{n} = \frac{7{,}048{,}107}{5000} = 1{,}409.621$ since the sampling process is SRS.

（c）Number of burglaries in the sample is 295. $\hat{t} = 7{,}048{,}107 \times \frac{295}{5000} \approx 415{,}838.$

$$SE(\hat{t}) = 7{,}048{,}107 \times \sqrt{(1 - \frac{5000}{7{,}048{,}107}) \times \frac{\frac{295}{5000} \times (1 - \frac{295}{5000})}{5000 - 1}} = 23{,}480$$

$CI_{95\%}: \hat{t} \pm 1.96SE(\hat{t}) = [369{,}817,\ 461{,}859]$

Therefore, the estimation for total number of burglaries known to police between 2001 and 2019 is 415,838, with a 95% CI $[369{,}817,\ 461{,}859]$.

（d）The percentage of crimes that were domestic-related is 13.1%.

$$\widehat{CV} = \sqrt{(1 - \frac{5000}{7{,}048{,}107}) \frac{0.131 \times (1 - 0.131)}{5000}} / 0.131 = 3.641 \times 10^{-2}$$

§ 2.13 Exercises：37

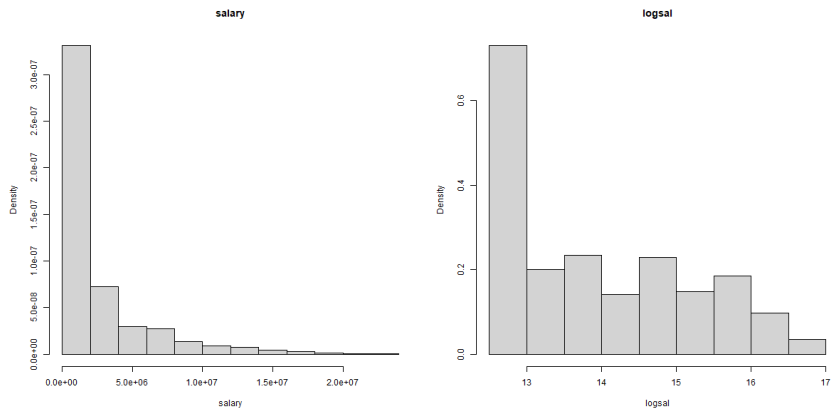
（c）The mean of *logsal* $\bar{y} = 13.77.\ CI_{95\%}: \bar{y} \pm 1.96SE(\bar{y}) = [13.590,13.950]$

（d）The estimation of proportion of players who are pitchers $\hat{p} = \frac{66}{150},\ CI_{95\%}: [0.368,0.512]$

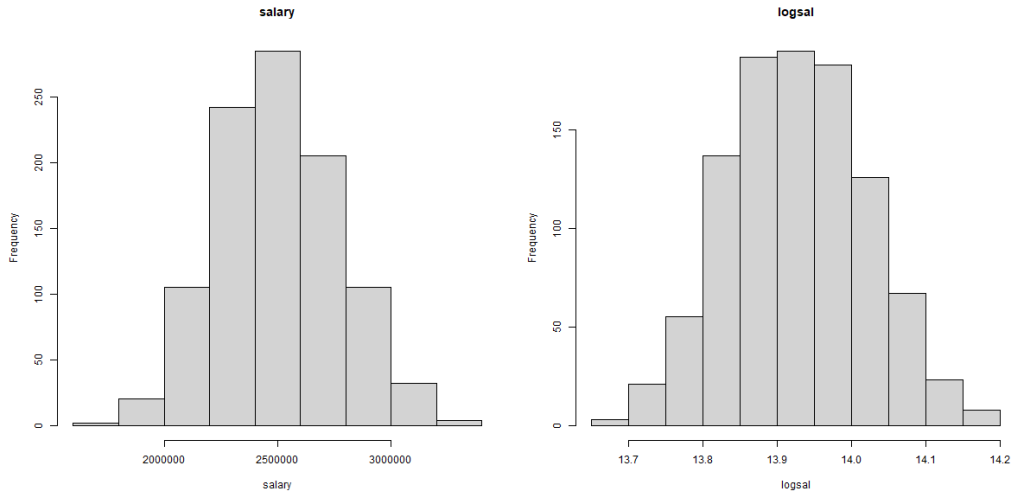
（e）True mean of *logsal* for the population $\overline{y_u} = 13.93$, True proportion of players who are pitchers for the population $p = 0.472$. Both CIs in (c) and (d) contain the true values respectively.

另外补充：

（1）*salary* 和 *logsal* 两个变量的总体分布直方图如下图所示。



（3）*salary* 和 *logsal* 两个变量的样本均值的频率分布直方图如下图所示。



salary 样本均值的均值：2,498,235，方差：7.101×10^{10}，样本方差的均值：1.250×10^{13}
logsal 样本均值的均值：13.93，方差：8.63×10^{-3}，样本方差的均值：1.535

可以发现 *salary* 和 *logsal* 的样本均值的均值、样本均值的方差、方差的均值都和真实值（总体的均值、$\frac{1-f}{n}S^2$、总体的方差）相近，相对误差比真实值小了几个数量级。这印证了课件上定理 3.1 的结论。

验证项目	变量	由 1000 个样本得到的估计值	真实值	相对误差
(1)$E(\bar{y}) = \overline{y_u}$	*salary*	2,498,235	2,497,669	0.0227%
	logsal	13.93	13.93	0.00142%
(2)$V(\bar{y}) = \frac{1-f}{n}S^2$	*salary*	7.101×10^{10}	6.77×10^{10}	4.805%
	logsal	8.63×10^{-3}	8.31×10^{-3}	3.750%
(3)$E(s^2) = S^2$	*salary*	1.250×10^{13}	1.250×10^{13}	0.0257%
	logsal	1.535	1.534	0.0184%

（4）*salary* 在 1000 个样本置信区间覆盖真实总体均值的比例为 94.3%，*logsal* 在 1000 个样本置信区间覆盖真实总体均值的比例为 95%。可以发现 *logsal* 的覆盖比例更接近 95%，这可能是因为 *logsal* 的分布虽然左偏，但比 *salary* 的左偏程度低，这导致 *logsal* 的样本均值的分布更加接近正态分布，使得 *logsal* 的覆盖比例更接近 95%。

验证项目	变量	由 1000 个样本得到的估计值	真实值	相对误差
(4) $\overline{y_u}$置信区间的覆盖范围	*salary*	94.2%	95%	0.842%
	logsal	95%	95%	0

附录：

解答题目所使用的 R 代码及输出请见 https://thisiskunmeng.github.io/sampling/hw3.html