数学实验 统计推断 实验报告

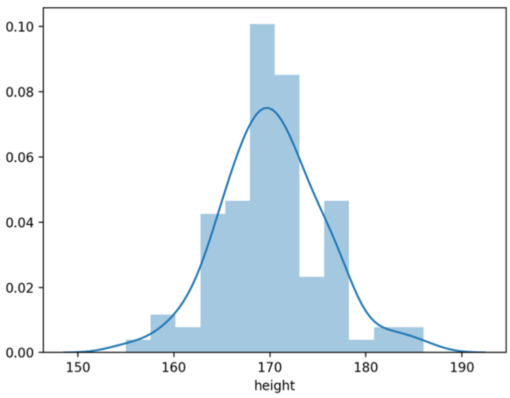
计 65 赖金霖 2016011377

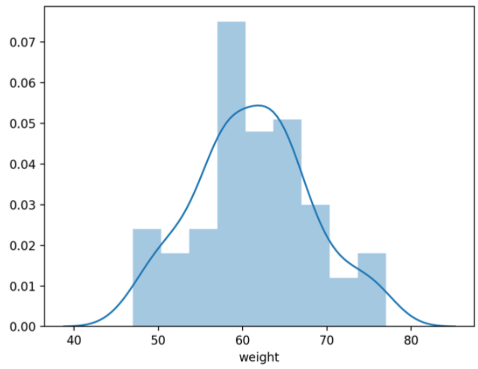
实验 12

6.

（1）

身高和体重分别的频数直方图和拟合分布曲线如下图所示





从形状上看，这两个分布都接近正态分布。

根据 python 的 scipy.stats.kstest(Kolmogorov-Smirnov 测试)计算，这两个分布都符合正态分布（p 值在计算精度下均为 0）。

（2）

以身高为例，设其真实均值为μ，它的观测均值 x^*满足

$$\frac{x^*-\mu}{s/\sqrt{n}} \sim t(n-1)$$

其中 n 为样本数，s 为样本标准差。取α=0.05，可以通过 t(n-1)计算μ的置信区间。经过计算，身高和体重估计如下

	平均值	置信区间
身高	170.25	[169.18,171.32]
体重	61.27	[59.91,62.63]

(3)

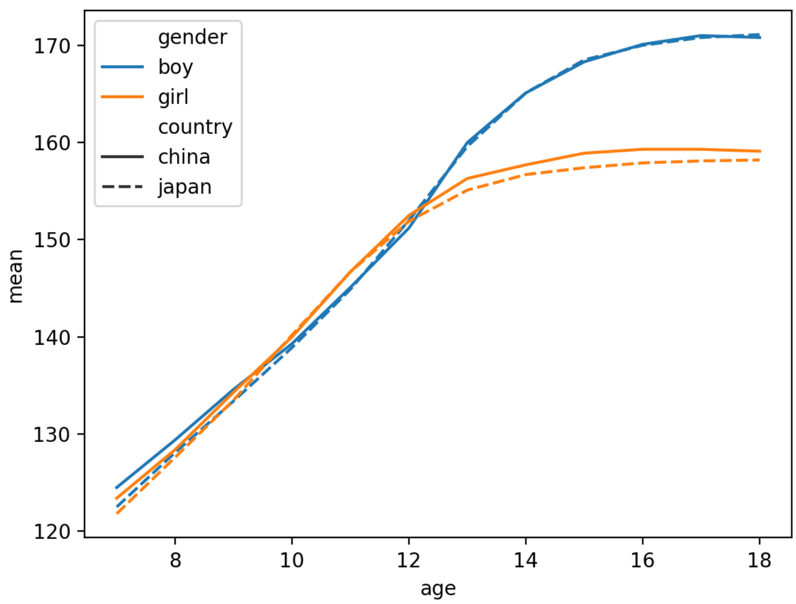
以身高为例，可进行假设检验：H_0：平均身高=167.5=μ_0；H1：平均身高≠167.5。而

$$\frac{x^*-\mu_0}{s/\sqrt{n}} \sim t(n-1)$$

因为10年前的平均身高不在(2)中的置信区间内，所以学生平均身高明显上升了；而体重同理，10年前的平均体重60.2在(2)中的置信区间内，所以学生的平均体重没有明显变化。

10.

各类学生平均身高随年龄的变化曲线如下所示：



从肉眼上看，中日两国男生的身高在年龄小时有差别，女生的身高也有细微差别，中国女生在 12 岁以后要略高于日本女生。

设中国样本数量 n_1=8333，均值为 μ_1，标准差为 d_1；日本样本数量 n_2=28983,均值为 μ_2，标准差为 d_2，则

$$z = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$where \quad s^2 = \frac{(n_1-1)d_1^2 + (n_2-1)d_2^2}{n_1 + n_2 - 2}$$

可以对 μ_1 是否等于 μ_2 进行假设检验。在 α=0.05 时，上述设置下各年龄是否有显著差别如下：

	7	8	9	10	11	12	13	14	15	16	17	18
男	有	有	有	有	有	有	有	无	有	无	有	有
女	有	有	有	有	无	有	有	有	有	有	有	有

可以看出，统计方法比肉眼更加敏感。而男性小年龄和女性大年龄时的差别都体现出来了。

补充习题

3.

（1）

理想情况下取值小于 T 的样本数占总样本数的比例为

$$1 - e^{-\lambda T}$$

设观测值为 K，则 $1/\lambda$ 的估计为

$$\frac{1}{\lambda} \approx -\frac{T}{ln(1-K)}$$

(2)

理想情况下，取值小于 T 的样本的均值为

$$\frac{-Te^{-\lambda T}-\frac{1}{\lambda}e^{-\lambda T}+\frac{1}{\lambda}}{1-e^{-\lambda T}}$$

若观测值为 K，则解上式=K，可以得到λ的估计。

（3）

理想情况下，样本的平均值为（包括没有失效的元件）

$$-\frac{1}{\lambda}e^{-\lambda T}+\frac{1}{\lambda}$$

若观测值为 K，则解上式=K，也可以得到λ的估计。

分别取 T=500，800，1000，1200，1500，对上述三种策略下的 1/λ 进行一次估计（置信区间α取 0.05）：

设置	T	平均值	置信区间
（1）	500	1020.789	[1009.942 , 1031.637]
（1）	800	1013.771	[1004.696 , 1022.847]
（1）	1000	1009.272	[1000.949 , 1017.594]
（1）	1200	1000.51	[992.579 , 1008.441]
（1）	1500	1012.351	[1004.508 , 1020.195]
（2）	500	695.791	[267.928 , 1123.654]
（2）	800	3342.624	[-80.89 , 6766.137]
（2）	1000	11974.961	[-10451.74 , 34401.662]

（2）	1200	734.085	[-93.193 , 1561.364]
（2）	1500	1123.774	[1087.99 , 1159.557]
（3）	500	1039.926	[1027.918 , 1051.933]
（3）	800	1013.093	[1003.552 , 1022.633]
（3）	1000	1005.515	[996.965 , 1014.066]
（3）	1200	1013.945	[1005.684 , 1022.206]
（3）	1500	1009.915	[1002.445 , 1017.385]

从数据中可以看出：

- （2）中的估计特别不稳定，只有当 T 较大时才能获得可观的置信区间。
- 从整体上看，随着 T 增大，置信区间缩短，估计变得更准确。
- （1）的方法和（3）的方法的估计准确度差不多，而（1）的计算成本更小，所以（1）会是一个更好的估计。

代码可在

https://github.com/lll6924/math_exp/blob/master/exp8

下找到