Jackknife & Bootstrap

本文用于课后复习，简单明了的记录自己学习的东西，数学不太严谨，请多多指教。

这篇文章仅介绍Jackknife，Bootstrap请看这里。

Jackknife以及Bootstrap均是重抽样（resample）的一种方法。
那么为什么要使用重抽样呢？

假如我们想要统计初中生身高的平均值，由于能采取的样本有限（比如在一所中学中抽取了200名学生）。自然而然地你将两百个身高平均之后得出答案（比如说170cm），那么我就会质疑这个结果的准确性。

你想要向我证明结果的准确性，要怎么办呢？我们只有样本，也只能在样本上下文章。于是你灵光一闪，那我们不如用样本的样本来进行统计。具体的，你在两百个数据中随机抽取150个作为一组，这样有放回的抽取10次，那么可以得到10组1500个数据。这样我们就可以计算出每一组身高的均值以及他们的方差。然后你用这些数据去汇报：“看，1500个数据，我算出了均值和方差，同时可以算出置信区间，你没理由不相信我了！”

这里用样本的样本进行统计，其实就是重抽样。那么下面就是对重抽样的两种方法Jackknife和Bootstrap进行介绍。

Jackknife

方差

Jackknife又叫做刀切法，那么怎么切呢？

我们将数据随机切出去一个，用剩余的作为一组新数据。200个学生用Jackknife随机剔除一个，剩余199个学生作为新的一组。

$\hat{\theta} = s(x)$ 表示对样本x要估计的参数$\hat{\theta}$，这里我们可以看作要估计**均值**，利用刀切法估计的参数表示为$\hat{\theta_{(i)}} = s(x_{(i)})$，其中$i$表示为此次估计剔除了第$i$个样本后进行估计$(x_1, x_2 \dots x_{i-1}, x_{i+1}, \dots, x_n)$。即：

$$\begin{aligned}\hat{\theta_{(i)}} = s(x_{(i)}) &= \frac{1}{n-1}\sum_{j\neq i} x_j \\ &= \frac{n\bar{x} - x_i}{n-1}\end{aligned}$$

这里$\bar{x}$是样本的均值，并且记$\hat{\theta_{(\cdot)}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta_{(i)}}$，也就是重抽样后所有组进行估计后得到的估计值的均值。

自然而然地可以得到他们的方差，我们记作$\hat{se}_{jack}$：
$\hat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta_{(i)}} - \hat{\theta_{(\cdot)}})}$
在这里我们要计算样本的均值，所以先进行推导：

$$
\begin{aligned}
\hat{\theta_{(\cdot)}} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\theta_{(i)}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{n\bar{x}-x_i}{n-1} \\
&= \frac{1}{n}(\frac{n^2\bar{x}}{n-1}-\frac{\sum_i x_i}{n-1}) \\
&= \frac{n\bar{x}-\bar{x}}{n-1} = \bar{x}
\end{aligned}
$$

$$
\begin{aligned}
\therefore \hat{se}_{jack} &= \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta_{(i)}}-\hat{\theta_{(\cdot)}})^2} \\
&= (\frac{n-1}{n}\times\sum\frac{(x_i-\bar{x})^2}{(n-1)^2})^{\frac{1}{2}} \\
&= (\frac{1}{n(n-1)}\sum(x_i-\bar{x})^2)^{\frac{1}{2}} \\
&= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}
\end{aligned}
$$

也就是说“样本的样本”的方差为$\frac{\sigma}{\sqrt{n}}$，这与我们之前所学的$se(\bar{x}) = \frac{\sigma}{\sqrt{n}}$相同（$\sigma$是样本的标准差，就是那200个学生的标准差）。

由于我们重抽样出多组新的数据，对于每一组数据都有一个均值，那么这些均值的方差也就有了意义。

偏差

接下来对$\hat{\theta_{(\cdot)}}$是否无偏（bias）进行讨论。

如果$\hat{\theta} = s(x)$是θ的无偏估计，即$E(\hat{\theta}) = \theta$，那么：

$$
\begin{aligned}
E(\hat{\theta_{(\cdot)}}) &= E(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta_{(i)}}) \\
&= \frac{1}{n}E(\sum_{i=1}^{n}\hat{\theta_{(i)}}) \\
&= \frac{1}{n}\sum_{i=1}^{n}E_\theta(\hat{\theta_{(i)}}) \\
&= E(\hat{\theta_{(i)}}) \\
&= \theta \leftarrow E(\hat{\theta})
\end{aligned}
$$

那么可以看出$E(\hat{\theta_{(\cdot)}})$也是无偏的。

如果$\hat{\theta} = s(x)$是有偏的，记：

$$
\hat{bias}_{jack}(\hat{\theta}) = (n-1)(\hat{\theta_{(\cdot)}}-\hat{\theta})
$$

可以理解为每组数据的偏差之和。

那么经过偏差修正，我们可以得到（bias corrected jackknife estimator）：

$$
\hat{\theta}_{jack} = \hat{\theta} - \hat{bias}_{jack}(\hat{\theta})
$$

有偏差的估计减去他的偏差就可以得到无偏差的估计$\hat{\theta}_{jack}$。

Pseudo-values

此外，还有一个叫pseudo-values的东西来实现jackknife。他被看作是一个无偏的估计（感觉就和$\hat{\theta}_{jack}$是一种东西）。

我们定义：$ps_i = n\varphi_n(X) - (n-1)\varphi_{n-1}(X_{(i)})$

或者：$ps_i = \varphi_n(X) - (n-1)(\varphi_n(X) - \varphi_{n-1}(X_{(i)})$

这个公式与$\hat{\theta}_{jack} = \hat{\theta} - bi\hat{a}s_{jack}(\hat{\theta})$简直一摸一样。。。。

其中$\varphi_n(X)$是对n个样本进行的估计，$\varphi_{n-1}(X_{(i)})$是对缺少第i个数据的样本进行估计，ps_i指第i个pseudo-value的估计。

那么n个pseudo-value估计的均值为：

$$\begin{aligned}\overline{ps} &= \frac{1}{n}\sum_{i=1}^{n} ps_i \\ &= \frac{1}{n}(\sum(n\hat{\theta} - (n-1)\hat{\theta_{(i)}})) \\ &= \frac{1}{n}(n^2\hat{\theta} - n(n-1)\hat{\theta_{(\cdot)}}) \\ &= n\hat{\theta} - (n-1)\hat{\theta_{(\cdot)}} \\ &= \hat{\theta}_{jack}\end{aligned}$$

由此可以证明此估计确实是无偏的。

同样的，我们来计算他的方差：

$$\begin{aligned}s^2_{ps-jack}(ps_i) &= \frac{1}{n-1}\sum(ps_i - \overline{ps})^2 \\ &= \frac{1}{n-1}\sum(n\hat{\theta} - (n-1)\hat{\theta_{(i)}} - \hat{\theta}_{jack})^2 \\ &= \frac{1}{n-1}\sum(n\hat{\theta} - (n-1)\hat{\theta_{(i)}} - n\hat{\theta} + (n-1)\hat{\theta_{(\cdot)}})^2 \\ &= \frac{1}{n-1}\sum((n-1)\hat{\theta_{(\cdot)}} - (n-1)\hat{\theta_{(i)}})^2 \\ &= (n-1)\sum(\hat{\theta_{(\cdot)}} - \hat{\theta_{(i)}})^2\end{aligned}$$

而对于$\overline{ps}$来说，其标准差为：

$$\begin{aligned}\hat{se}_{ps-jack}(\overline{ps}) &= \frac{\hat{se}_{ps-jack}(ps_i)}{\sqrt{n}} \\ &= \sqrt{\frac{(n-1)\sum(\hat{\theta_{(\cdot)}} - \hat{\theta_{(i)}})^2}{n}}\end{aligned}$$

可以看到此公式与上面非pseudo-value方法里的$\hat{se}_{jack}$相同，即：

$$\hat{se}_{ps-jack}(\hat{\theta}) = \hat{se}_{jack}(\hat{\theta})$$

置信区间

可以很简单的得到：

$$CI = [\hat{\theta} \pm t_{\frac{\alpha}{2}}(n-1)\hat{se}_{jack}(\hat{\theta})]$$

对于Jackknife就介绍到这里，过几天写一篇文章总结Bootstrap，现在就很迷茫，推导公式之后也记不住，不知道有什么好的方法可以扎扎实实的学习统计这门课。共勉吧。