1965-2000 年上海市降水的统计分析

摘　要：通过了解当地的降水特征，有助于合理运用降水资源和防范极端降水带来的灾害、因地制宜发展农业和旅游业、建造基础设施。本文利用上海市气象局历年来的逐年春秋季降水数据，选取其中 1965-2000 年的数据作为样本，并通过分析数据分布情况、求解样本期望方差、参数估计、对其进行估计量评价并进行置信区间的估计和假设检验来对样本进行分析，在过程中得出上海市春秋季节的降水特征并对其进行了比较。

关键词：降水；上海；季节

Statistical analysis of precipitation in Shanghai from 1965 to 2000

Abstract: Understanding the local characteristics of precipitation can contribute to the rational use of precipitation resources, prevention of disasters caused by extreme precipitation, tailored development of agriculture and tourism, and construction of infrastructure. This article utilizes the annual spring and autumn precipitation data from the Shanghai Meteorological Bureau, selecting the data from 1965 to 2000 as the sample. The article analyzes the sample through analyzing the data distribution, determining the sample mean and variance, parameter estimation, evaluating the estimators, and conducting confidence interval estimation and hypothesis testing. In the process, it derives the characteristics of spring and autumn precipitation in Shanghai and compares them.

Key words: precipitation; Shanghai; season

一、研究问题与数据抽样

天气是与我们的日常生活息息相关的气候现象，在一定程度会影响我们的日常生活，如晴朗的天气适合出游，阴雨天则适合室内活动。长期的天气状况反映到一个地区则构成了这个地区的气候特征，对这个地区的气候特征进行分析，可以因地制宜种植最适合当地的农作物、相应地发展特色旅游业、做好应对极端天气的准备。本文针对上海市的气候状况，以“历年来上海市春季的季节降水总量”和“历年来上海市秋季的季节降水总量”为两个研究总体，分别记为X和Y。从总体中选取“1965-2000 年的春季的季节降水总量”和“1965-2000 年的秋季的季节降水总量”两个容量为 36 的样本，分别记为$(X_1, X_2, \dots, X_{36})$和$(Y_1, Y_2, \dots, Y_{36})$，根据中国气象局提供的数据，可以得到$(x_1, x_2, \dots, x_{36})$和$(y_1, y_2, \dots, y_{36})$两个样本观测值。由于气候在不同的历史时期有不同的变化趋势和不同的平均季节降水，所选取的样本仅能反映当代的上海降水特征，不具有更早的时期降水情况的反映能力。

二、抽样数据和分布状况

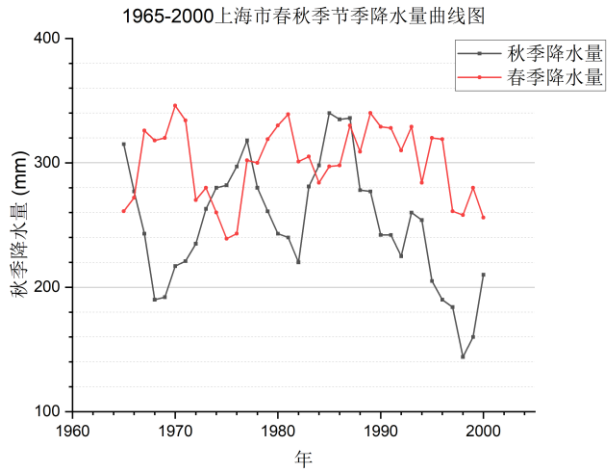
1. 抽样数据

年份	1965	1966	1967	1968	1969	1970	1971	1972	1973
春季	261	272	326	318	320	346	334	270	280
秋季	315	277	243	190	192	217	221	235	263
年份	1974	1975	1976	1977	1978	1979	1980	1981	1982
春季	260	239	243	302	300	319	330	339	301
秋季	280	282	297	318	280	261	243	240	220
年份	1983	1984	1985	1986	1987	1988	1989	1990	1991
春季	305	284	297	298	330	309	340	329	328
秋季	281	298	340	335	336	278	277	242	242
年份	1992	1993	1994	1995	1996	1997	1998	1999	2000
春季	310	329	284	320	319	261	258	280	256
秋季	225	260	254	205	190	184	144	160	210

降水量单位：mm

2. 分布状况

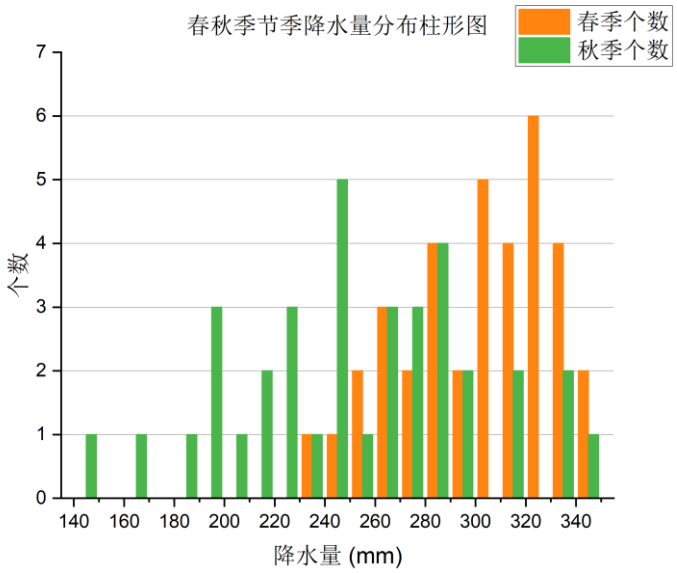
用折线图表示春秋季节季降水量随年份的变化，得到：



按每 10mm 的降水量为一个区间，分别统计春秋季节降水量到达该区间的年数：

降水区间	140-149	150-159	160-169	170-179	180-189	190-199	200-209
春季	0	0	0	0	0	0	0
秋季	1	0	1	0	1	3	1
降水区间	210-219	220-229	230-239	240-249	250-259	260-269	270-279
春季	0	0	1	1	2	3	2
秋季	2	3	1	5	1	3	3
降水区间	280-289	290-299	300-309	310-319	320-329	330-339	340-349
春季	4	2	5	4	6	4	2
秋季	4	2	0	2	0	2	1

为了便于形成降水量分布的柱状图，取各区间的中位数为该区间年份春秋季的季降水量，如 190-199 区间所有年份的该季节的季降水量均为 195mm。得到柱状图：



3. 初步分析

从“1965-2000 上海市春秋季节季降水量曲线图”分析，春季降水量随年份变化的趋势比秋季降水量的变化趋势稳定，在大多数年份里，春季降水量要高于秋季降水量。且不论是春季降水量还是秋季降水量，其都在波动变化。

从“春秋季节季降水量分布柱形图”分析，春季降水量全都分布在 230-350mm 之间，而秋季降水量在 140-350mm 之间均有分布。春季降水量在 325mm 处的年份数量达到最大，频率达到 16.7%，在 235、245mm 处的年份数量最小，频率均只有 2.8%，秋季降水量在 245mm 处的年份数量达到最大，频率为 13.9%，分别在 155、175mm 处的年份数量降为 0，频率也降至 0%。并且不管是春季降水量还是秋季降水量，都大致呈现正态分布的特征，其中春季的μ大于秋季，春季的σ^2小于秋季。

三、基本分析

1. 分析公式

由于数据为离散数据，所以引入离散数学期望：

$$\mathrm{E}(X)=\sum_{i=1}^{+\infty} x_i \,\mathrm{P}(x_i)$$

方差：

$$\mathrm{D}(X)=\frac{1}{n}\sum_{i=1}^{+\infty}\big(x_i-E(X)\big)^2$$

样本均值：

$$\overline{\mathrm{X}}=\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{X_i}$$

样本方差：

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

样本协方差：

$$S_{XY}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

样本相关系数：

$$\rho_{XY}=\frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}$$

2. 1965-2000 年的春季的季节降水总量

首先对春季降水量进行基本分析。以第一张表格中的数据进行分析，可以得到：

$$\bar{X}=299.92$$

$$S_X^2=906.08$$

3. 1965-2000 年的秋季的季节降水总量

然后对秋季降水量进行基本分析，得到：

$$\bar{Y}=244.19$$

$$S_Y^2=2410.54$$

4.样本协方差与样本相关系数

计算得到：

$$S_{XY}=-91.86$$

$$\rho_{XY}=-0.06$$

5.分析

X 的样本均值 299.92，对比与 Y 的样本均值 244.19，验证了先前分析春季降水量的均值大禹秋季降水量均值的结论，两者的差值为 55.73，意味这在 1965-2000 这段时间内，每年春季降水量平均要比秋季降水量大 55.73㎜。对比 X 的样本方差 906.08 和 Y 的样本方差 2410.54，可知春季降水量随年变化较稳定，没有太大的起伏，相比于秋季可以更好的对未来春季降水量进行预测。通过求出 X 和 Y 的相关系数可知，相关系数小于 0 且接近 0，可知春季降水量和秋季降水量只有很小的负相关关系，可以视作相互独立，互不影响。

四、参数估计

由柱状图的形状特征分析，可以近似认为 1965-2000 上海市春季降水量和秋季降水量的分布都呈正态分布。设$X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，即春季降水量服从均值为$\mu_1$，方差为$\sigma_1^2$的正态分布，秋季降水量服从均值为$\mu_2$，方差为$\sigma_2^2$的正态分布。

1. 春季降水量分布参数的估计

用矩估计法估计μ_1:

$$E(X)=\mu_1$$

$$\widehat{\mu_1}=M_1=\bar{X}=299.92$$

同理，σ_1^2:

$$D(X)=E(X^2)-[E(X)]^2$$

$$\widehat{\sigma_1^2}=M_2-M_1^2=880.91$$

用最大似然估计法估计μ_1:

由正态分布的概率密度函数$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

似然函数$L(\mu_1,\sigma_1^2)=\prod_{i=1}^{36}f(x_i)=\left(\frac{1}{\sqrt{2\pi}\sigma_1}\right)^{36}e^{-\frac{\sum_{i=1}^{36}(x_i-\mu_1)^2}{2\sigma_1^2}}$

$$\frac{\partial \ln L}{\partial \mu_1}=-\frac{\sum_{i=1}^{36}(\mu_1-x_i)}{\sigma_1^2}=0$$

$$\sum_{i=1}^{36}(x_i-\mu_1)=0$$

$$\widehat{\mu_1}=\frac{1}{36}\sum_{i=1}^{36}x_i=\bar{X}=299.92$$

用最大似然估计法估计σ_1^2:

$$\frac{\partial lnL}{\partial \sigma_1^2}=0$$

$$-\frac{n}{2\sigma_1^2}+\frac{\sum_{i=1}^{36}(x_i-\mu_1)^2}{2{\sigma_1}^4}=0$$

$$\sigma_1^2=\frac{1}{36}\sum_{i=1}^{36}(x_i-\mu_1)^2$$

$$\widehat{\sigma_1^2}=\frac{1}{36}\sum_{i=1}^{36}(x_i-\widehat{\mu_1})^2=\frac{1}{36}\sum_{i=1}^{36}(x_i-\bar{X})^2=(CM)_2=880.91$$

通过以上两种方式计算得到的两个估计量相等，其矩估计值和最大似然估计值均为$(299.92,880.91)$。

2. 秋季降水量分布参数的估计

类似的，用矩估计法估计μ_2:

$$\widehat{\mu_2}=M_2=\bar{Y}=244.19$$

同理，σ_2^2:

$$\widehat{\sigma_2^2} = M_2 - M_1^2 = 2389.59$$

用最大似然估计法估计μ_2:

$$似然函数L(\mu_2, \sigma_2^2) = \prod_{i=1}^{36} f(y_i) = \left(\frac{1}{\sqrt{2\pi}\sigma_2}\right)^{36} e^{-\frac{\sum_{i=1}^{36}(y_i-\mu_2)^2}{2\sigma_2^2}}$$

$$\frac{\partial lnL}{\partial \mu_2} = -\frac{\sum_{i=1}^{36}(\mu_2 - y_i)}{\sigma_2^2} = 0$$

$$\widehat{\mu_2} = \frac{1}{36}\sum_{i=1}^{36} y_i = \bar{Y} = 244.19$$

用最大似然估计法估计σ_2^2:

$$\frac{\partial lnL}{\partial \sigma_2^2} = 0$$

$$-\frac{n}{2\sigma_2^2} + \frac{\sum_{i=1}^{36}(y_i - \mu_2)^2}{2{\sigma_2}^4} = 0$$

$$\widehat{\sigma_2^2} = \frac{1}{36}\sum_{i=1}^{36}(y_i - \widehat{\mu_2})^2 = \frac{1}{36}\sum_{i=1}^{36}(y_i - \bar{Y})^2 = (CM)_2 = 2389.59$$

通过以上两种方式计算得到的两个估计量相等，其矩估计值和最大似然估计值均为$(244.19, 2389.59)$。

3.分析

通过两种估计方法，得出在正态分布下，春季平均降水量的估计值为 299.92㎜，秋季平均降水量的估计值为 244.19㎜，和在基本分析中计算得到的期望相同。但是在方差上面，方差的估计值在春季和秋季分别为 880.91 和 2389.59，对比于基本分析中得到的样本方差 906.08，2410.54。不难发现，方差的估计值$=\frac{35}{36}$样本方差，这也是由于计算的公式差异不同所导致。总体而言，估计值和基本分析中得到的结果能很好的相符。

五、估计量的评价标准

在通过矩估计法和最大似然估计法获得春秋两季的降水量的估计量后，需进行对估计量优劣性的评判，这里将从无偏性、有效性和一致性评价上述估计量。

1.无偏性

由于两种方法所得的$\widehat{\mu_1}$、$\widehat{\mu_2}$、$\widehat{\sigma_1^2}$、$\widehat{\sigma_2^2}$一致，所以仅需进行一次判定。

$$\mathrm{E}(\widehat{\mu_1}) = \mathrm{E}(\bar{X}) = \mathrm{E}(X) = \mu_1$$

$$E(\widehat{\sigma_1^2}) = E\left(\frac{1}{36}\sum_{i=1}^{36}(y_i - \bar{Y})^2\right) = \frac{35}{36}E(S_X^2) = \frac{35}{36}D(X) = \frac{35}{36}\sigma_1^2$$

所以$\widehat{\mu_1}$是μ_1的无偏估计量，而$\widehat{\sigma_1^2}$不是σ_1^2的无偏估计量，对其进行修正：

$$\sigma_1^2 = \frac{36}{35}E(\widehat{\sigma_1^2})$$

同理，$\widehat{\mu_2}$是μ_2的无偏估计量，而$\widehat{\sigma_2^2}$不是σ_2^2的无偏估计量，修正为：$\sigma_2^2 = \frac{36}{35}E(\widehat{\sigma_2^2})$。

2.有效性

由于只有μ_1、μ_2是无偏估计量，故仅对μ_1、μ_2进行有效性分析。由 Rao-Cramer 不等式，任意的无偏估计量的方差都有一个非零下界：$\frac{1}{nE[g(X)^2]}$。

$$所以任意\mu_1的估计值的方差D(\widehat{\mu_1}) \geq \frac{1}{36E\left[\left(\frac{\partial lnf(x;\mu_1)}{\partial \mu_1}\right)^2\right]} = \frac{\sigma_1^2}{36}$$

$$之前所求得到的估计值的方差D(\widehat{\mu_1}) = D(\bar{X}) = \frac{\sigma_1^2}{36}$$

所以$\widehat{\mu_1}$为μ_1的有效估计量，同理$\widehat{\mu_2}$为μ_2的有效估计量。

3. 一致性

将样本容量拓展到 n，则$D(\widehat{\mu_1}) = D(\bar{X}) = \frac{\sigma_1^2}{n}$，由于$\sigma_1^2$是常数，故$\lim_{n\to\infty} D(\widehat{\mu_1}) = \lim_{n\to\infty} \frac{\sigma_1^2}{n} = 0$。

由切比雪夫不等式得到：$P(|\widehat{\mu_1} - \mu_1| \geq \epsilon) = P(|\widehat{\mu_1} - E(\widehat{\mu_1})| \geq \epsilon) \leq \frac{D(\widehat{\mu_1})}{\epsilon^2} = 0$，所以$\widehat{\mu_1}$是$\mu_1$的一致估计量，同理$\widehat{\mu_2}$是$\mu_2$的一致估计量。

六、置信区间

1. 春季降水量的置信区间

仍然假设春季降水量分布服从正态分布，$X \sim N(\mu_1, \sigma_1^2)$。其中，$\mu_1$和$\sigma_1^2$均未知。构造$\mu_1$的置信度为 95%的置信区间$(\widehat{\mu_L}, \widehat{\mu_U})$。

$$枢轴量T = \frac{\bar{X}-\mu_1}{S_X/\sqrt{36}} \sim t(35)$$

$$T的可能取值范围\left(-t_{0.025}(35), t_{0.025}(35)\right)，可信度为\ 95\%$$

求解$T > -t_{0.025}(35)$和$T < t_{0.025}(35)$，解得置信区间：

$$\left(\bar{X} - \frac{S_1}{\sqrt{n}} t_{0.025}(35), \bar{X} + \frac{S_1}{\sqrt{n}} t_{0.025}(35)\right)$$

查表，带入具体数据，解得置信区间为:$(289.74, 310.10)$

构造σ_1^2的置信度为 95%的置信区间$\left(\widehat{\sigma_L^2}, \widehat{\sigma_U^2}\right)$。

$$枢轴量\chi^2 = \frac{1}{\sigma_1^2}\sum_{i=1}^{36}(X_i - \bar{X})^2 \sim \chi^2(35)$$

$$\chi^2的可能取值范围\left(\chi^2_{0.975}(35), \chi^2_{0.025}(35)\right)，可信度为\ 95\%$$

解得置信区间：

$$\left(\frac{\sum_{i=1}^{36}(X_i - \bar{X})^2}{\chi^2_{0.025}(35)}, \frac{\sum_{i=1}^{36}(X_i - \bar{X})^2}{\chi^2_{0.975}(35)}\right)$$

解得置信区间为$(596.07, 1541.77)$

2. 秋季降水量的置信区间

假设$Y \sim N(\mu_2, \sigma_2^2)$，$\mu_2$的 95%的置信区间$(\widehat{\mu_L}, \widehat{\mu_U}) = (227.58, 260.80)$

σ_2^2的 95%的置信区间$\left(\widehat{\sigma_L^2}, \widehat{\sigma_U^2}\right) = (1616.93, 4182.28)$

3. 分析

春季降水量均值的 95%的置信区间为$(289.74, 310.10)$，方差的 95%置信区间为$(596.07, 1541.77)$。秋季降水量均值的 95%置信区间为$(227.58, 260.80)$，方差的 95%置信区间为$(1616.93, 4182.28)$。通过对置信区间的计算，可以得到降水量大概率落在哪个范围，以及大概率会有多大的起伏，稳定性如何。计算发现，春季降水量和秋季降水量的均值的 95%置信区间没有重叠，这表明在很多时候，春季的降水量要大于秋季的降水量，但由于秋季降水量方差的置信区间落在数值较高的区间，所以秋季更可能会有降水量忽大忽小的情况。

七、假设检验

通过上海市气象局提供的资料，闵行区 1965-2000 年的春季平均降水量和冬季平均降水量分别为 304.3mm 和 247mm，是否有理由认为上海市的季节平均降水量和闵行区的季节平均降水量相等？

首先分析春季降水，取显著性水平$\alpha = 0.05$。

$$H_0: \mu_1 = 304.3,\ H_1: \mu_1 \neq 304.3$$

$$T = \frac{\bar{X} - 304.3}{S_X/\sqrt{36}} \sim t(35)。$$

$$C = \frac{S_X}{\sqrt{36}} \mathrm{t}_{\alpha/2}(35) = 10.18$$

代入$\bar{X} = 299.92$, 得到$|\bar{X} - 304.3| < 10.18$

所以不在拒接域内，所以接受假设。同理可得秋季降水同样不在拒接域内，接受假设。

其次分析方差，是否有理由认为春季平均降水量的方差不超过 950？

$$H_0: \sigma_1^2 \leq 950, H_1: \sigma_1^2 > 950$$

$$\chi^2 = \frac{35S_X^2}{950} \sim \chi^2(35)$$

拒接域:$\frac{35S_X^2}{950} > 49.82$

代入$S_X^2 = 906.08$, 得:$\frac{35S_X^2}{950} = 33.38 < 49.82$

不在拒接域，所以接受H_0，认为春季平均降水量的方差不超过 950。

同理，可以计算秋季降水方差，$H_0: \sigma_2^2 \leq 2750, H_1: \sigma_2^2 > 2750$

$$\frac{35S_Y^2}{2750} = 30.68 < 49.82$$

不在拒接域，接受H_0，秋季平均降水量的方差不超过 2750。

八、总结

本文基于上海市 1965-2000 年春秋季的平均降水量的数据，通过进行统计分析、参数估计、参数评价、置信区间计算和假设检验，从统计学角度分析了上海市春秋季降水的特征，包括均值、方差和协方差等。从均值和方差的对比可以看出：春季平均降水量高于秋季，方差显著小于秋季，这意味着春季平均降水量在年与年之间的差异相比于秋季不会过大。同样，也通过假设分布和参数估计得出了数据可能满足的一种正态分布，通过对参数的评价，得出了无偏、有效和一致性的结论。然后进行了置信区间的计算，得到了数值最可能分布的区间。最后提出了自己对均值和方差的假设并进行了检验，进而证明了假设的正确性。

随着现代社会的不断发展，降水的不确定性大大增加，也加大了天气预测的难度。为了减小降水的不均匀性以及极端天气对城市造成过大的影响，应相应发展天气预警技术，同时

也应考虑未来城市的建设和城市生态的维护，尤其要考虑降水带来的影响，实现充分利用降水，完善极端降水下的应对方式和保护设施。