1 常用代码及公式

$\frac{1}{\sigma^2}\sum(X_i-\mu)\sim\chi^2_n$
$\frac{1}{\sigma^2}\sum(X_i-\overline{X})\sim\chi^2_{n-1}$
$\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t_{n-1}$ ($n\geq 30$ 当 $\mathcal{N}$)
`qnorm`: 生成 normal 数据

2 Measurement

Scales of Measurement

1. nominal: 没有 order 的 categories
2. ordinal: 有 order
3. interval: 数值按照等长区间分类
4. ratio: 单点的数值

数据的分类

1. Categorical / Qualitative: Nominal / Ordinal
2. Numerical / Quantitative: Discrete / Continuous

Basic Quantities

`quantile(arr, 0.25)`: Q_1
$Q_{1,2,3}$: 25%, 50%, 75% percentile
IQR $= Q_3 - Q_1$
Skewness: 看尾巴在哪边

1. Left-Skewed: Negative Skewness
2. Right-Skewed: Positive Skewness

Why trimmed mean?

1. May have a lower SE when data is not normal
2. Balance between median and mean, protect against outliers

画图

1. Stem and leaf plot: 左边是数字第一位，右边是后面的，中间用 | 隔开（`stem(x)`）
2. Histogram: `hist(x)`

transformation

log 把中心往右，exp 把中心往左
Log-normal distribution: $\log X\sim\mathcal{N}(\mu,\sigma^2)$
$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\log x-\mu)}{2\sigma^2}}$
$\mu=e^{\mu+\frac{\sigma^2}{2}}$, $\sigma^2=[\exp(\sigma^2)-1]\exp(2\mu+\sigma^2)$
Coefficient of Variation (CV): $\frac{\sigma}{\mu}$
Geomean $=\sqrt[n]{\prod X_i}$
Imposing a Normal PDF on the Histogram

```
hist(x)
xpt <- seq(from, to, by=by)
n_den <- dnorm(xpt, mean(return), sd(return))
ypt <- n_den * length(x) * 10
# We notice that each data point in the return
  dataset represents an area of 1 * 10, so the
  total area of the histogram would be * 10.
lines(xpt, ypt, col="blue")
```

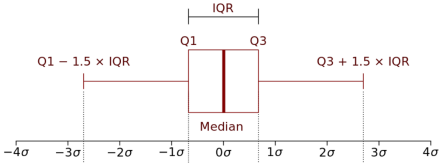
QQplot: $\left(\Phi^{-1}(q_i),\hat{F}_x^{-1}(q_i)\right)$

1. 左侧越低表示 longer left tail
2. 右侧越高表示 longer right tail

left skew 是两侧都高，right skew 是两侧都低
t 两个尾巴都长，是左低右高
Shapiro-Wilk Test: `shapiro.test(x)`
$W=\frac{\left(\sum_{i=1}^n a_i x_i\right)^2}{\sum_{i=1}^n(x_i-\overline{x})^2}$，$a_i$ 是这个系统自带的常数
p 越小越 normal
Limitations:

1. Adversely affected when there are tied data
2. Has a bias by sample size. Statistically significant result, large sample.

box-plot: `boxplot(x~group, data=x)`



Outliers

classic tech: $|x_i-\overline{x}|>2\cdot\text{sd}$
boxplot rule: $x_i<Q_1-1.5\cdot\text{IQR}$ or $x_i>Q_3+1.5\cdot\text{IQR}$