# Lecture 07:

# Estimation

- Estimation
  - Point estimation
  - Confidence Interval (interval estimation)

# Interpretation of Statistical Inference: The sample is one of the possibilities

It is critical to note that the sample is thought of as one of many possible samples we may have gotten by drawing a random sample from the population under the study. The total number of possible samples is $\binom{N}{n}$

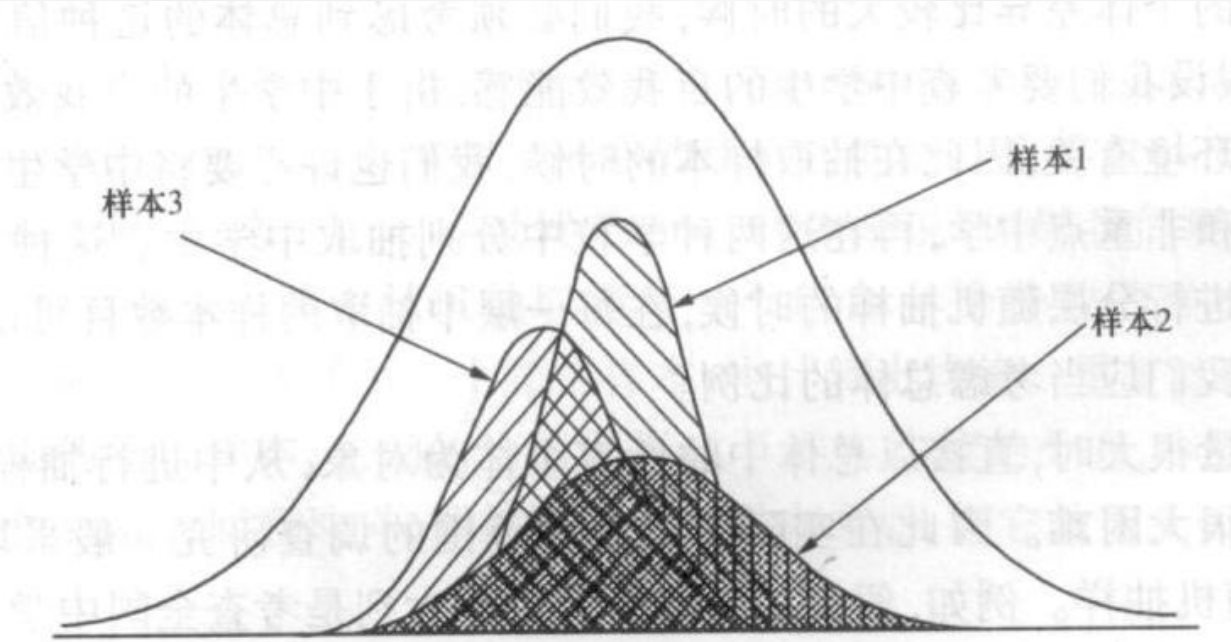



图 5.1　从同一总体中抽取的三个不同的样本

Population Distribution: 总体分布

Sample Distribution: 样本分布

Sampling Distribution: 取样分布， 多个样本的估值的分布

- 首先我觉得假设检验的流程本身比较死板(？)，因为我完全可以在不做任何关于显著性水平的预计的前提下，根据我得到的p值来断言我的实验结果关于多大的alpha显著。比如我在单侧检验的前提下发现p＝0.04，那我当然可以去说这个实验结果是关于alpha＝0.05显著的，而不必说我一定要在实验之前提前说明我想要多大的alpha(如果不考虑估计样本量这个操作的话)。
  所以这就涉及到了和p-hacking擦边的一个操作：我本来想做双侧检验，但是我得到了p＝0.08，对于0.05不显著；此时我直接在paper上改口，说这个实验结果对于0.05的单侧检验显著(p＝0.04)。从数学上我真没找到有什么问题……所以想请教一下，这样的操作到底算不算“hacking”？真正的科研可不可以这么搞？

- 老师但是我其实还是没有打心底理解这件事为什么不对。我觉得我之前描述的这个被称为p-hacking的过程在数学上完全正确呀，统计上也都说得通，只是不符合大家规定的假设检验的步骤而已。因为无论我选择单侧检验还是双侧检验，都不会对实验数据产生任何的影响，所以在看到这一组实验数据的时候，我完全可以从双侧检验的标准放宽到单侧检验的标准，我只需要看这组数据满足什么样的显著性条件就可以了呀。这样做得到的结论也没有任何问题：这组数据不满足双侧检验的0.05，但是满足单侧检验的0.05。而我只在论文里说出上述两个结论之一：“它符合单侧检验的标准”，也是客观事实呀，即使我本来做的是双侧检验。这就是我觉得我可能存在误解的地方？我觉得可以不严格按照假设检验的流程来(而是根据实验数据反推)，仍然能得到准确且合理的实验结果。
- 我更多地觉得，假设检验只是一个科学的框架，但是如果一定要按照它的顺序来操作的话，是不是有点过分较真？(我只是单纯觉得自己没真正理解，不是在抬杠……)🌹☕
-

# Estimation

- In general terms, **estimation** uses a sample statistic as the basis for estimating the value of the corresponding population parameter.
- Although estimation and hypothesis testing are similar in many respects, they are complementary inferential processes.
- A hypothesis test is used to determine whether or not a treatment has an effect, while estimation is used to determine how much effect.

# Estimation (cont.)

- Case 1: estimation is used **after** a hypothesis test that resulted in rejecting the null hypothesis. Thus, the hypothesis test has established that a treatment effect exists and the next logical step is to determine how much effect.
- Case 2: it is also common to use estimation in situations where a researcher **simply wants to learn about an unknown population**. Thus, a sample is selected from the population and the sample data are then used to estimate the population parameters.

# Estimation (cont.)

- You should keep in mind that even though estimation and hypothesis testing are inferential procedures, these two techniques differ in terms of the type of question they address.
- A hypothesis test, for example, addresses the somewhat academic question concerning the *existence* of a treatment effect.
- Estimation, on the other hand, is directed toward the more practical question of *how much* effect.

**TABLE 12.1**

The logic behind hypothesis tests and estimation.

| Hypothesis Test | Estimation |
|---|---|
| *Goal:* To test a hypothesis about an unknown parameter, usually the null hypothesis, which states that the treatment has no effect.<br>A. Begin by hypothesizing a value for the unknown parameter.<br>B. The hypothesized value is substituted into the formula and the value for $t$ is calculated.<br>C. If the hypothesized value produces a reasonable outcome for $t$ (near zero), we conclude that the hypothesis was reasonable and we fail to reject $H_0$. If the outcome is an extreme, low-probability value for t, we reject $H_0$. | *Goal:* To estimate the value of an unknown parameter.<br>A. Do not attempt to calculate a $t$ statistic. Instead, begin by estimating what the $t$ value ought to be. The strategy for making this estimate is to pick a reasonable, high-probability value for $t$ (near zero).<br>B. The reasonable value for $t$ is substituted into the formula and the value for the unknown parameter is calculated.<br>C. Because we used a reasonable value for $t$, it is assumed that the calculations will produce a reasonable estimate of the population parameter. |

# Estimation of the mean of a distribution

Have sample $x_1, \ldots x_n$ from population of size N

Sample mean, $\bar{X}$ is ONE estimate of the population mean $\mu$

$$X = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{u}$$

**Point Estimation!** （点估计）

$$\exp(\bar{X}) = \mu$$

Recall that [$x_1, \ldots, x_n$ ] is ONE sample from the target population, one of many possible samples. Thus, think of a population of sample means.

The average of these means is $\mu$.

$$E(\bar{X}) = \mu$$

$$\bar{X} \ \ is \ \ an \ \ unbiased \ \ estimator \ \ of \ \ \mu.$$

*1). Unbiased means that if we use estimates of many sample distributions to estimate the parameters of a population, the error on average is zero.*

*2). Assuming the underlying distribution of the target population is Normal, the median is also an unbiased estimator of μ, and there are many more. Considering all unbiased estimators, the one with the smallest variance is the mean. The mean is referred to as the MVUE (Minimum variance unbiased estimator) of μ.*

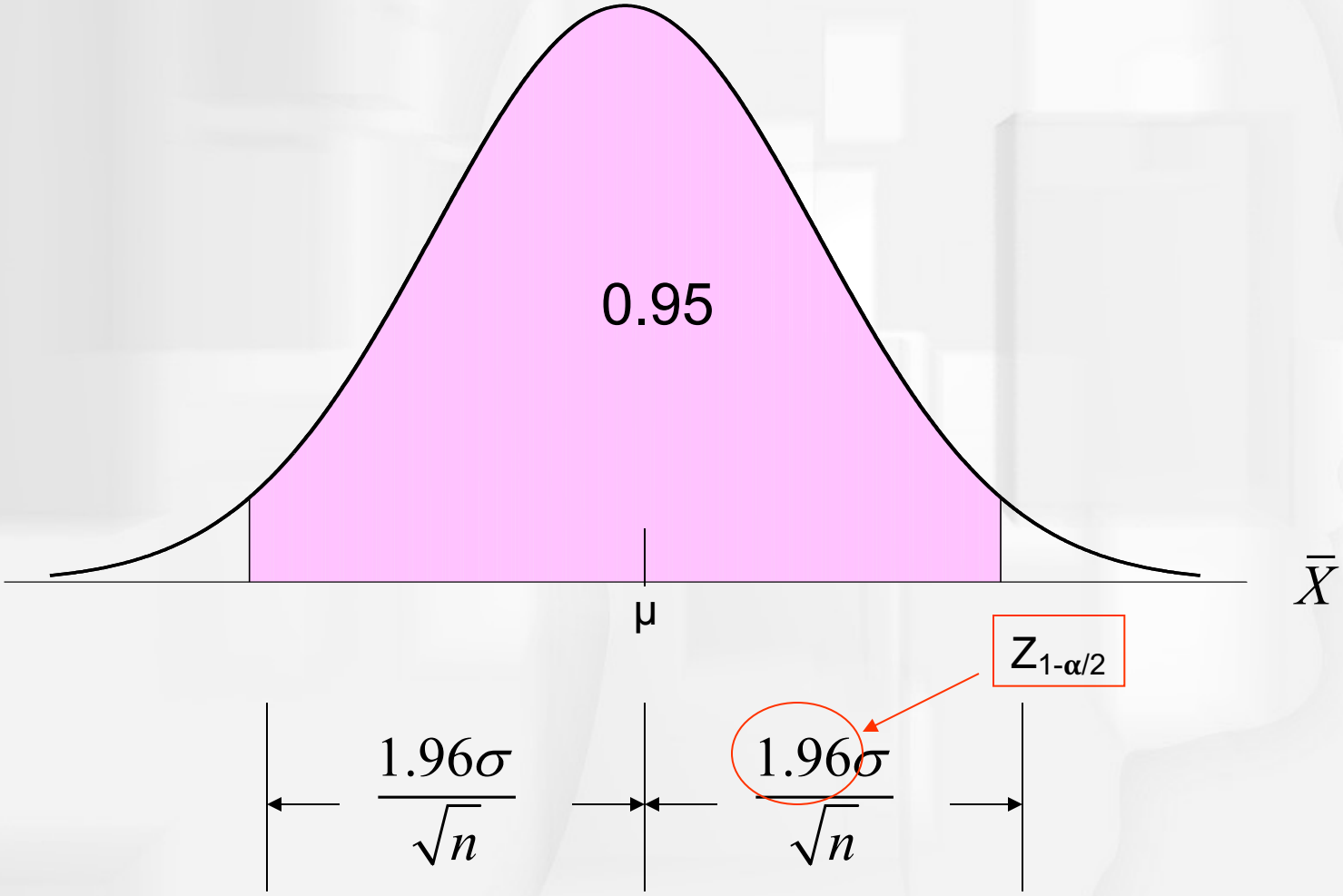
MVUE is about maximum precision and maximum accuracy. Please visualize the bullet scatter around a target

# Interval Estimation (区间估计，with known variance and mean)

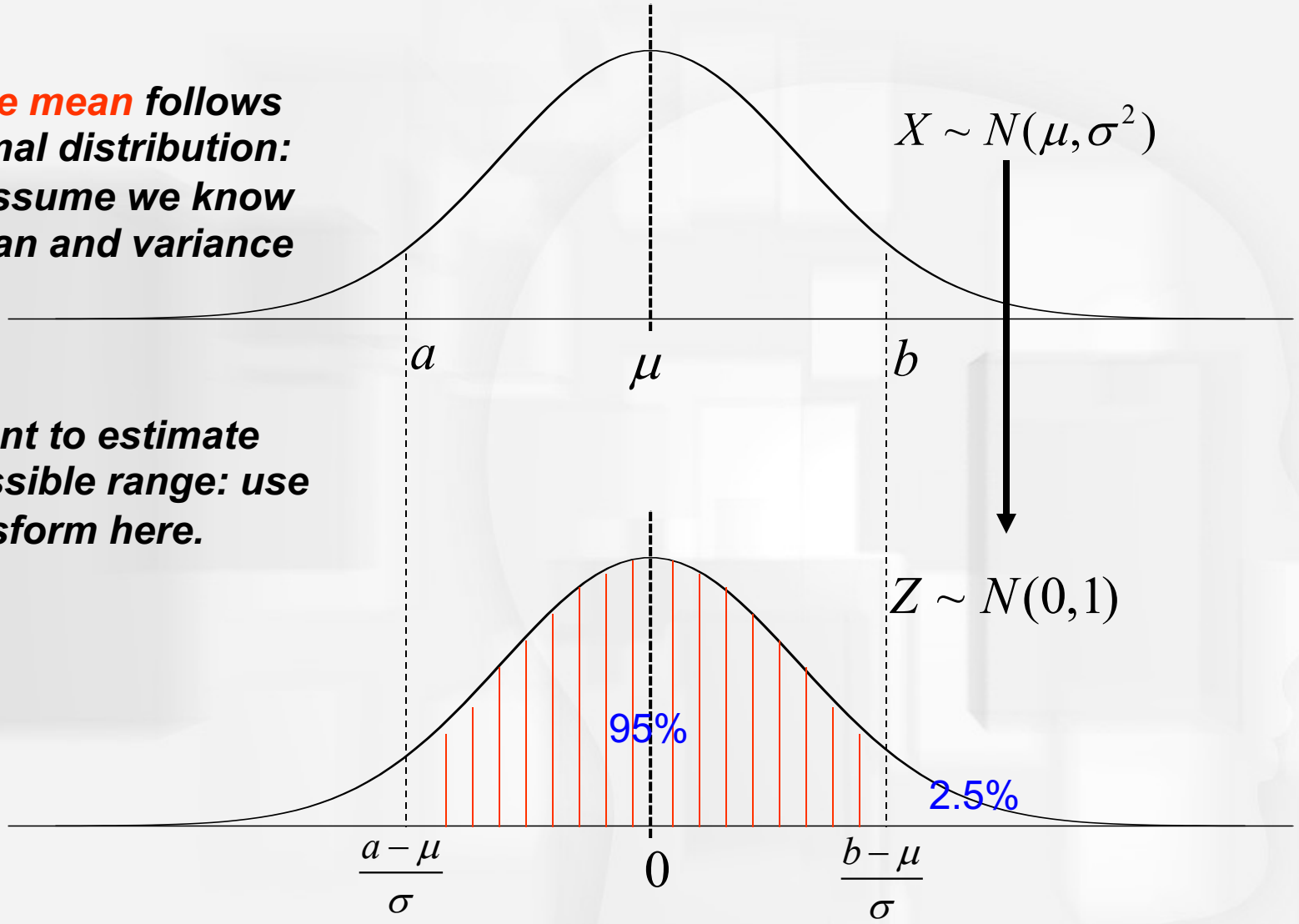
A *population* of children <1 yr. old in a province has a mean height of 20 inches and a SD of 4 inches. Now, we have a plan to choose a random sample of 64 children for a study.

Question: What is the range of sample mean for 95% of all possible samples?

Sample mean follows a Gaussian distribution with a NEW σ



***Sample mean* follows a Normal distribution: now assume we know its mean and variance**




***We want to estimate its possible range: use Z transform here.***

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

唯一不同

In the question, $\sigma = 4$, $n = 64$, $u = 20$, $z_{1-\alpha/2} = 1.96$.

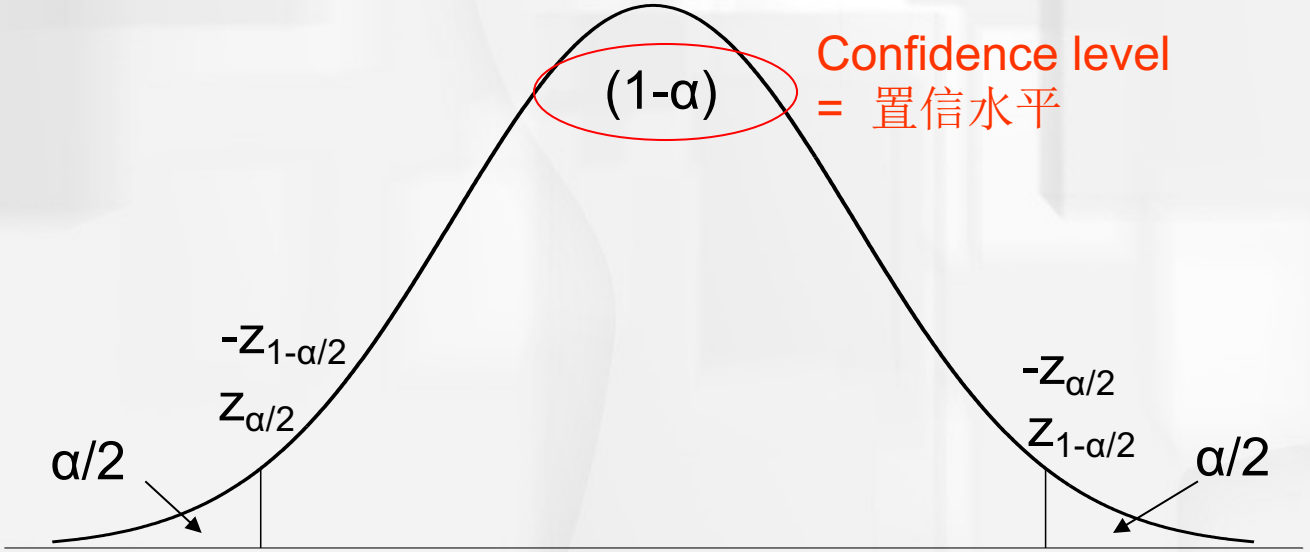
$$\left( \bar{X} - (z_{1-\alpha/2}) \frac{\sigma}{\sqrt{n}}, \bar{X} + (z_{1-\alpha/2}) \frac{\sigma}{\sqrt{n}} \right)$$

# Summary：Sample Distribution of $\bar{X}$

A $(1-\alpha)100\%$ large-sample Confidence Interval for $\mu$ is

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ or } \bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = Z$$




Confidence Limits (置信界限) for μ

# Definition of Confidence Interval

A $\left(1-\alpha\right)100\%$ Confidence Interval for $\mu$ is defined by the interval

$$\left(\bar{X}-z_{1-\alpha/2}\cdot\sigma/\sqrt{n},\bar{X}+z_{1-\alpha/2}\cdot\sigma/\sqrt{n},\right)$$

where $z_{1-\alpha/2}$ is $1-\alpha/2$ percentile of $N\left(0,1\right)$ distribution.

The length of a $\left(1-\alpha\right)100\%$ CI is $2\cdot z_{1-\alpha/2}\cdot\sigma/\sqrt{n}$,

and is a function of $n,\ \sigma,\ and\ \alpha.$

***In essence, CI is all about probability!***

L = Length of CI = $2 \cdot z_{1-\alpha/2} \cdot \sigma / \sqrt{n}$

L↓ as n↑

L↓ as σ↓

L↓ as α↑

L↓ as (1 - α )↓

*Previously constructed CI for mean of a Normal Distribution when variance is known. But population variance is rarely known.*

***What to do?***

We used the fact that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ to construct CI's. Now $\sigma^2$ is unknown and estimated by $s^2$.

How to compute?

But,

$\frac{\bar{X}-\mu}{s/\sqrt{n}}$ is NOT distributed as Normal.

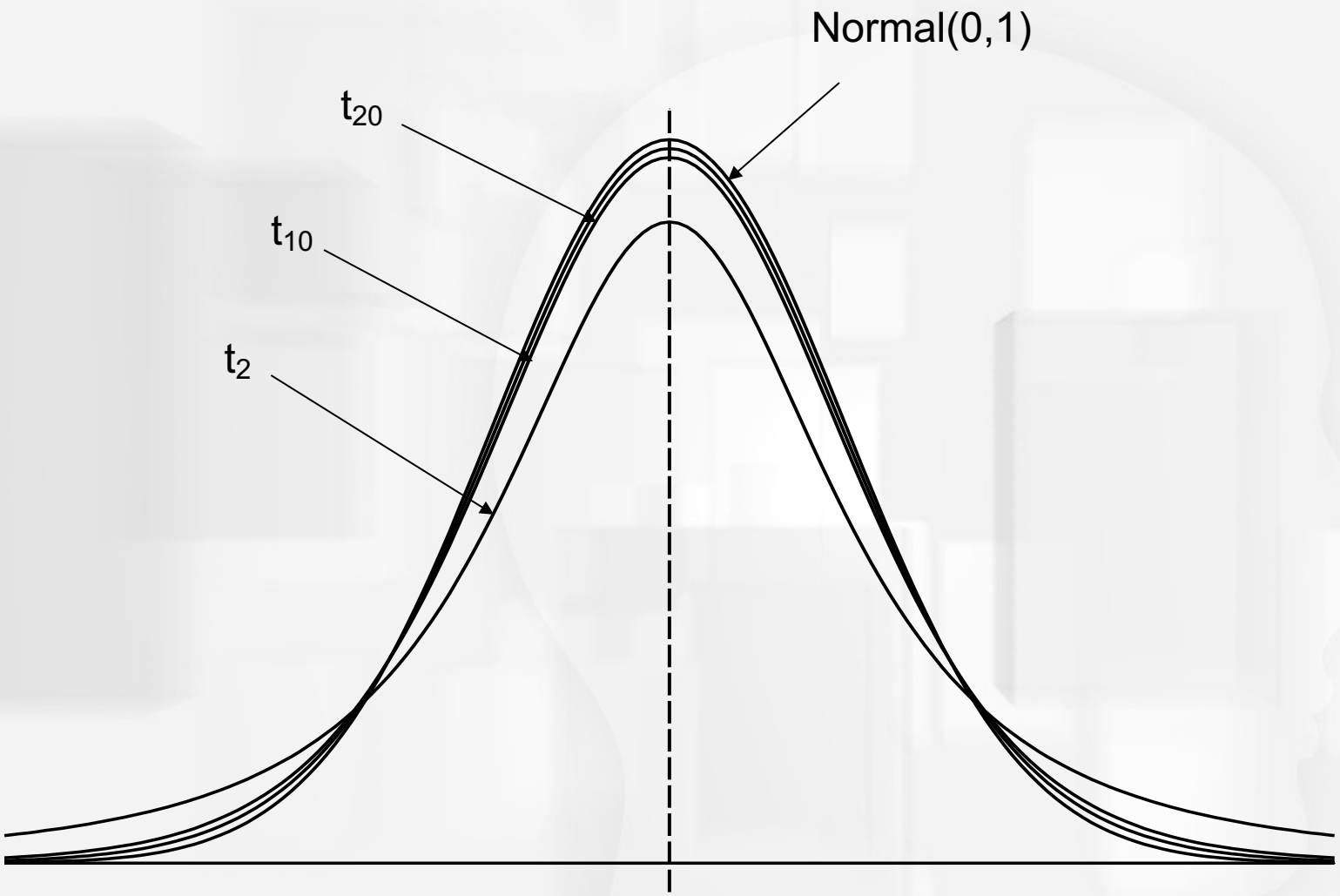
William Gossett under Pseudonym of "Student" showed that this statistic follows a student's t distribution or t distribution.

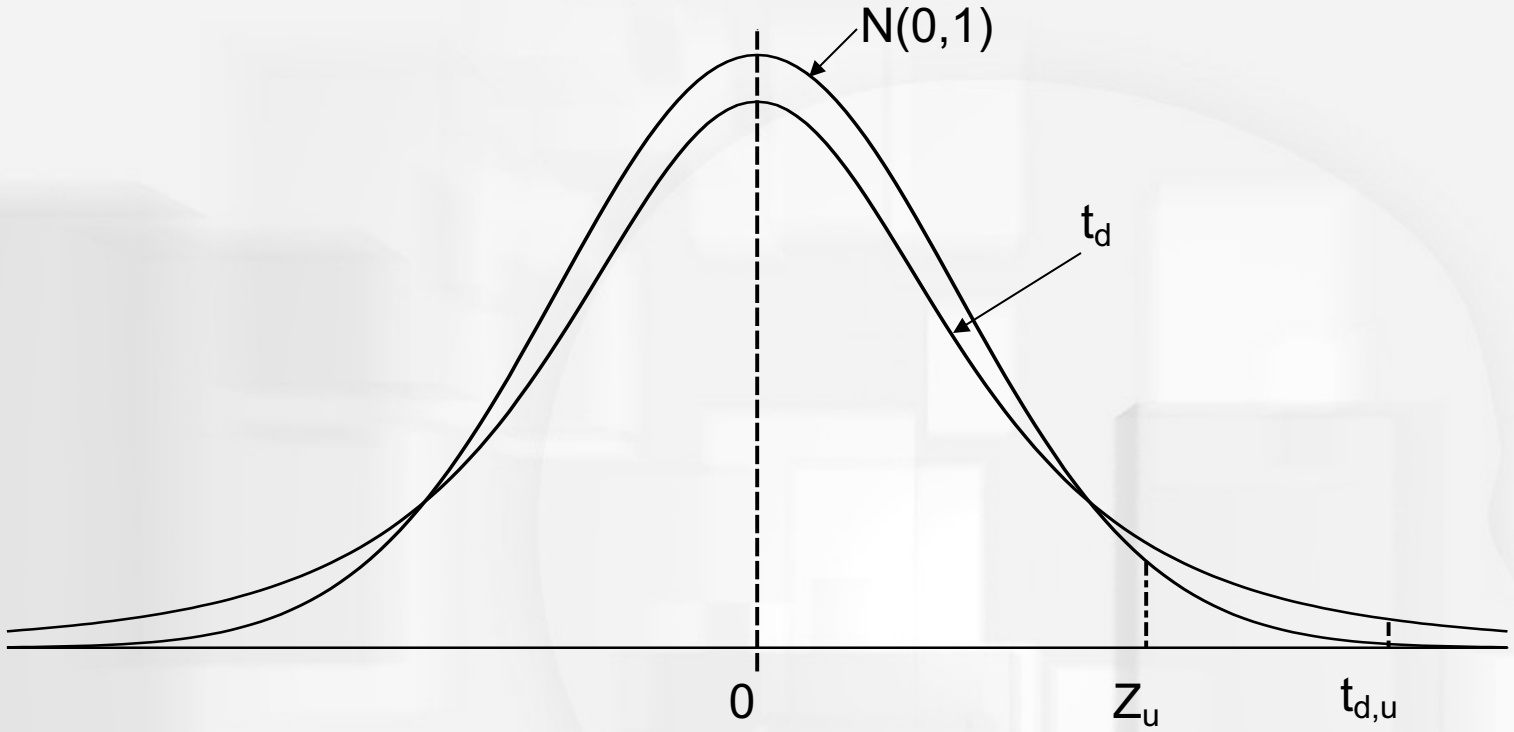
# t Distribution

Note that the shape of the t distribution depends on n; and thus t is a family of distributions with 1 parameter called the 'Degree of Freedom' (df) of the distribution.

If $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$ are independent, then $\frac{(\bar{X} - \mu)}{s/\sqrt{n}} \sim t(n-1)$ or $t_{n-1}$.

As the df increases, the t distribution tends towards the normal distribution. When df = $\infty$, t ($\infty$) is normal.


Normal(0,1)
$t_{20}$
$t_{10}$
$t_2$




$U^{th}$ percentile of a t-distribution with d degrees of freedom is denoted by $t_{d,u}$, where $Pr(t_d \leq t_{d,u}) \equiv u$.

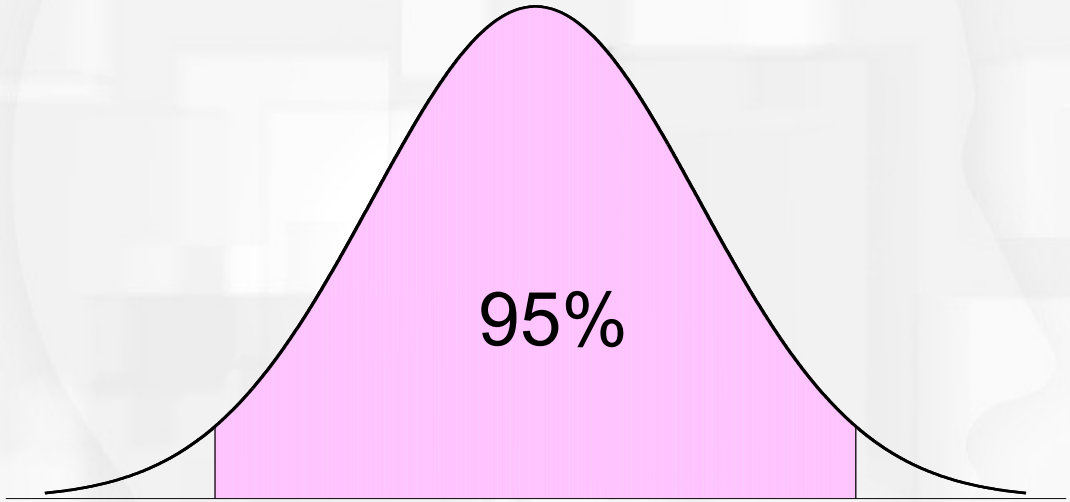
$t_{20,\ .95}$ is the $95^{th}$ percentile of a t-distribution with 20 df.

$t_{20,\ .95} = 1.725$ $\qquad$ $z_{.95}=1.645$

$t_{20,\ .975}= 2.086$ $\qquad$ $z_{.975}=1.96$

**EX**

| n | |
|---|---|
| 2 | $t_{1,0.975} = 12.706$ |
| 6 | $t_{5,0.975} = 2.571$ |
| 11 | $t_{10,0.975} = 2.228$ |
| 21 | $t_{20,0.975} = 2.086$ |
| 31 | $t_{30,0.975} = 2.042$ |
| 121 | $t_{120,0.975} = 1.98$ |
| $\infty$ | $t_{\infty,0.975} = 1.96$ |



# Interval Estimation (with unknown variance)

Similar to the case of know variance, a 100(1-α)% CI for the mean μ of a Normal distribution with unknown variance is

$$\left( \overline{X} - (t_{n-1,1-\alpha/2}) \frac{s}{\sqrt{n}}, \overline{X} + (t_{n-1,1-\alpha/2}) \frac{s}{\sqrt{n}} \right)$$

df determines what t distribution to use

# Descriptive statistics

Definition: Graphical, tabular and numerical approaches to summarizing data (to describe the main features of a collection of data)

# What is bootstrap?

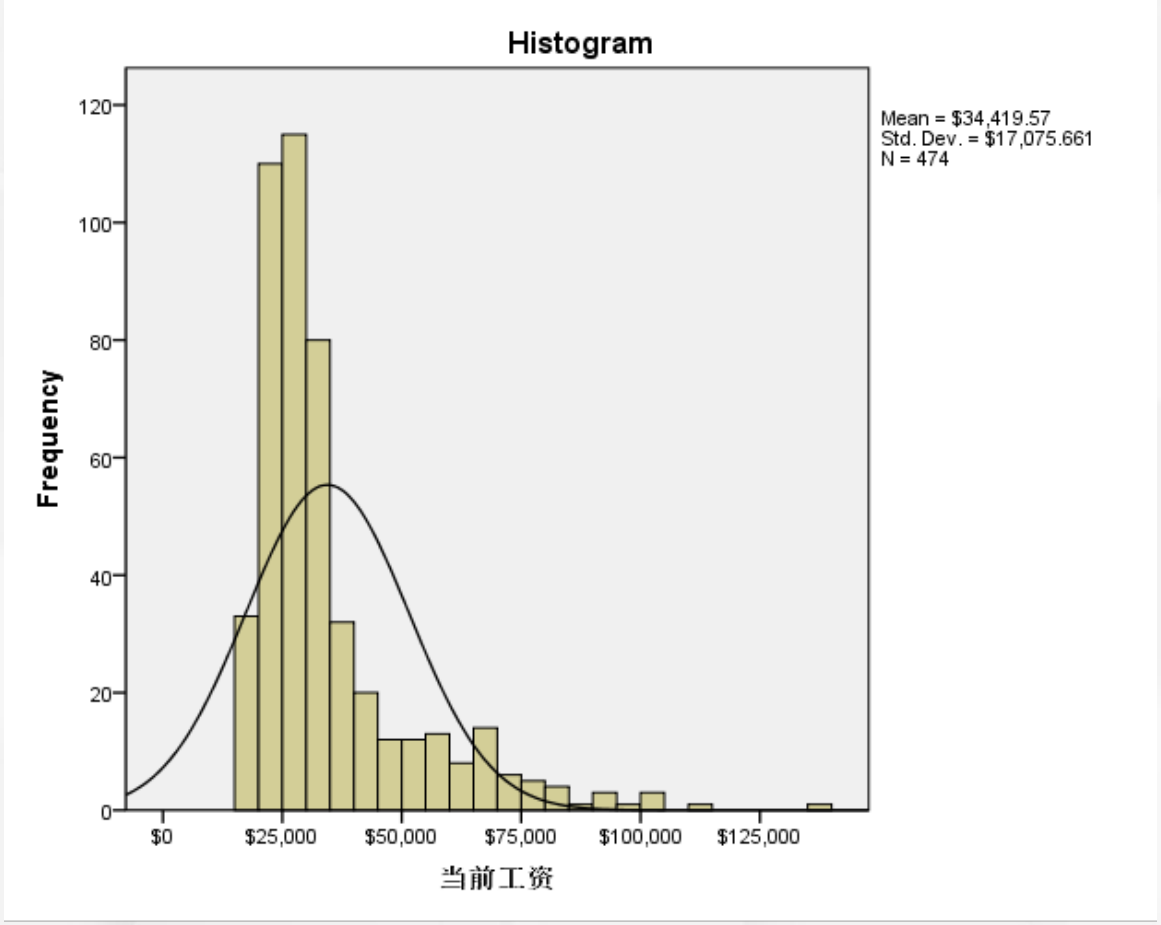
- A computation method to **resample** the data (e.g., your original sample) and to derive the confidence interval of parameters (such as mean, median, mode, variance, correlation, etc.).
- The CIs of these parameters are very hard to get by analytical methods; thus we compute them by using numeric methods.
- **When to use**: any time when you need to estimate CI of something **other than** a mean under a normal distribution.
- When not to use: the original sample is too biased (when garbage-in-and-garbage-out will happen) or too small.

# Typical outputs without bootstrap

**Statistics**

当前工资

| | | |
|---|---|---|
| N | Valid | 474 |
| | Missing | 0 |
| Mean | | $34,419.57 |
| Std. Error of Mean | | $784.311 |
| Median | | $28,875.00 |
| Mode | | $30,750 |
| Std. Deviation | | $17,075.661 |
| Variance | | 291578214.5 |
| Skewness | | 2.125 |
| Std. Error of Skewness | | .112 |
| Kurtosis | | 5.378 |
| Std. Error of Kurtosis | | .224 |
| Range | | $119,250 |
| Minimum | | $15,750 |
| Maximum | | $135,000 |
| Sum | | $16,314,875 |
| Percentiles | 25 | $24,000.00 |
| | 50 | $28,875.00 |
| | 75 | $37,162.50 |



# Outputs with bootstrap

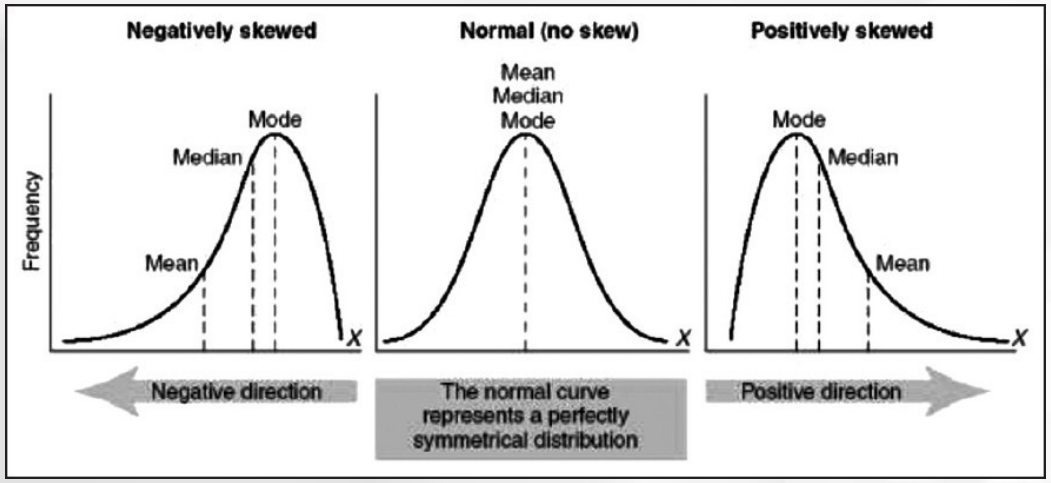
**Statistics**

当前工资

| | | Statistic | Bootstrap[b] Bias | Bootstrap[b] Std. Error | Bootstrap[b] 95% Confidence Interval Lower | Bootstrap[b] 95% Confidence Interval Upper |
|---|---|---|---|---|---|---|
| N | Valid | 474 | 0 | 0 | 474 | 474 |
| | Missing | 0 | 0 | 0 | 0 | 0 |
| Mean | | $34,419.57 | $10.03 | $808.40 | $32,784.98 | $36,103.36 |
| Std. Error of Mean | | $784.311 | | | | |
| Median | | $28,875.00 | $8.48 | $560.43 | $27,750.00 | $30,000.00 |
| Mode | | $30,750 | | | | |
| Std. Deviation | | $17,075.661 | -$89.733 | $1,071.268 | $15,013.581 | $19,017.424 |
| Variance | | 291578214.5 | -1909967.738 | 36457199.13 | 225407631.7 | 361662412.0 |
| Skewness | | 2.125 | -.035 | .223 | 1.710 | 2.542 |
| Std. Error of Skewness | | .112 | | | | |
| Kurtosis | | 5.378 | -.249 | 1.577 | 2.604 | 8.479 |
| Std. Error of Kurtosis | | .224 | | | | |
| Range | | $119,250 | | | | |
| Minimum | | $15,750 | | | | |
| Maximum | | $135,000 | | | | |
| Sum | | $16,314,875 | | | | |
| Percentiles | 25 | $24,000.00 | $54.07 | $339.77 | $23,250.00 | $24,673.09 |
| | 50 | $28,875.00 | $8.48 | $560.43 | $27,750.00 | $30,000.00 |
| | 75 | $37,162.50 | $108.31 | $1,651.06 | $34,800.00 | $40,237.50 |

b. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

# Skewness （偏度）




population

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{(\mathrm{E}[(X-\mu)^2])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}},$$

sample

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{3/2}}$$

How to interpret:

■Measures the symmetry of a unimodal distribution. 0 is perfectly symmetrical, negative is for left long tail, positive is for right long tail.

■Regard it an asymmetrical distribution if skewness is larger than 2*SE(skewness).

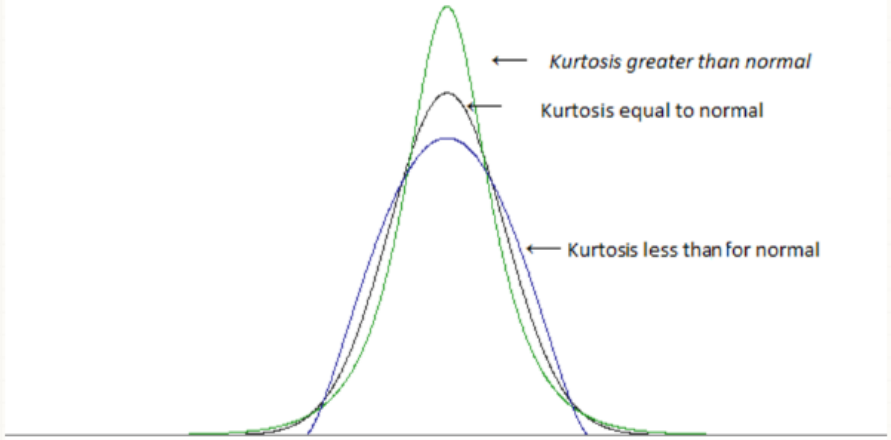
■多讲一个经验法则: |skewness|<0.5，基本对称; [0.5 1],中等偏斜; >1, 高度偏斜.

■However, the larger the sample size, the easier it is to have a skewed distribution since a large n will make SE small.

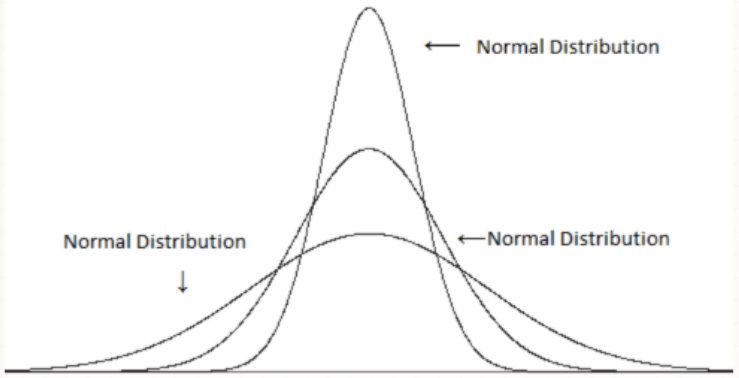
# Kurtosis (峰度)

$$\mathrm{Kurt}[X] = \frac{\mu_4}{\sigma^4} = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2}$$

**Kurtosis=3, <3 and >3**




**Kurtosis=3**




**How to interpret:**

■Measures the thickness of tails (peaked or flat around a center). 不是衡量分布的尖峭程度！是尾部的厚度，也就是极端值出现的概率！

■Normal distribution: 3 (or commonly kurtosis-3, which is **0**);

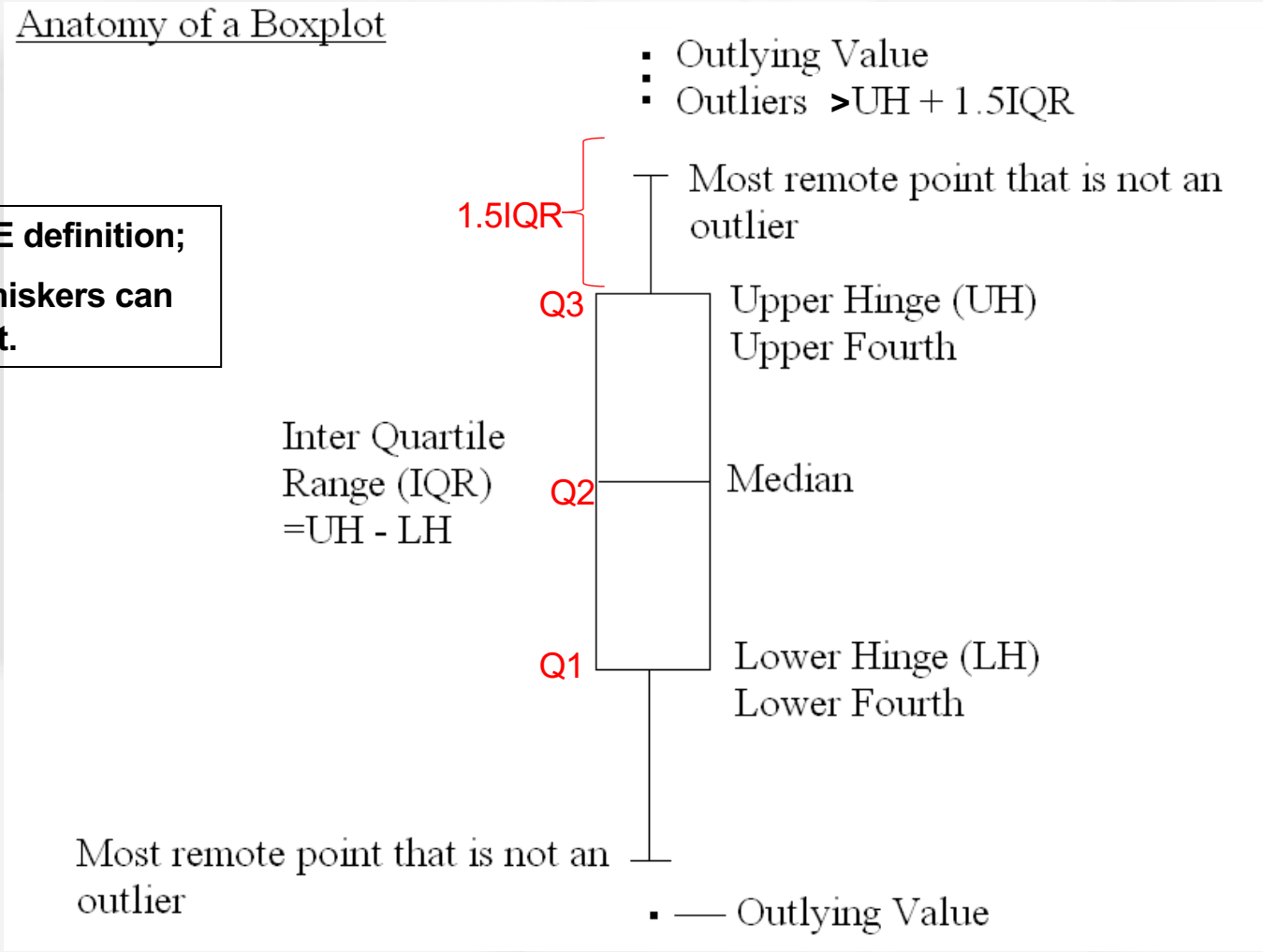
■More clustered and longer tails: >3 (+). E.g., Laplace distribution, approaching zero more slowly with more outliers;

■Less clustered and shorter tails: <3 (-). E.g., uniform distribution, with fewer or zero outliers.

■Note some people use Kurt(x) – 3, instead of Kurt(x)

**This is ONE definition;**

**The two whiskers can be different.**

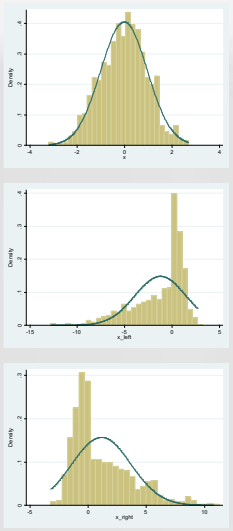


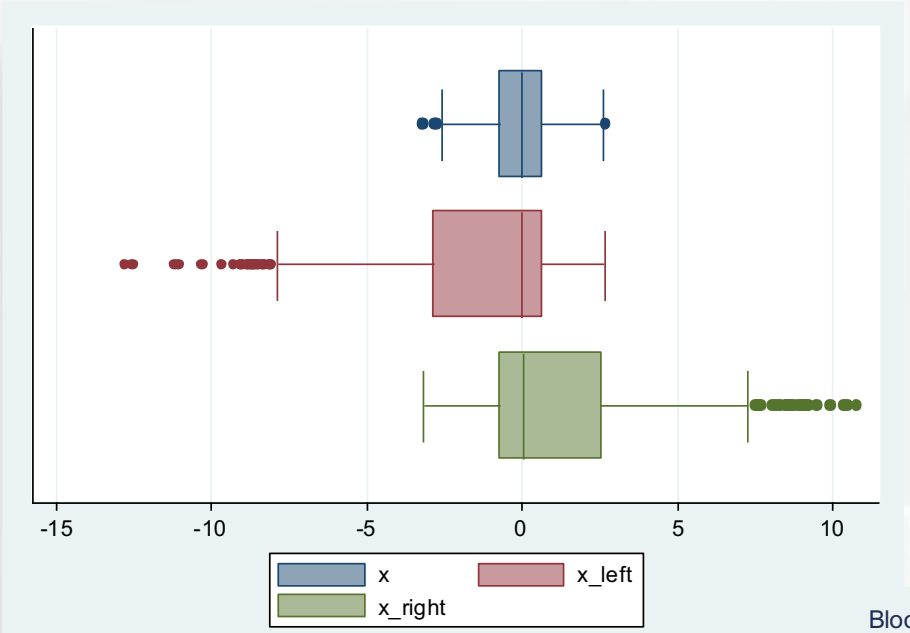
# Symmetry of a Distribution from Boxplot

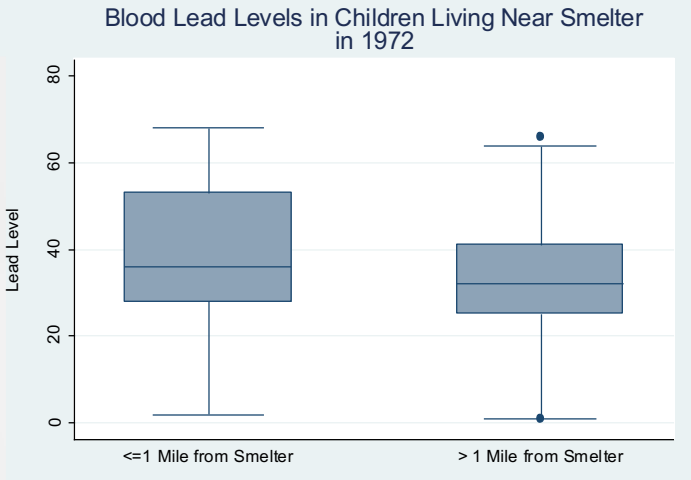
- Symmetric – Upper and lower hinges equally spaced from median
- Positively Skewed/right skewed – Upper hinge farther from Median than lower hinge.

- Negatively skewed/left skewed - Lower hinge farther from Median than upper hinge.

# Example: Showing Boxplots with Skewed Data

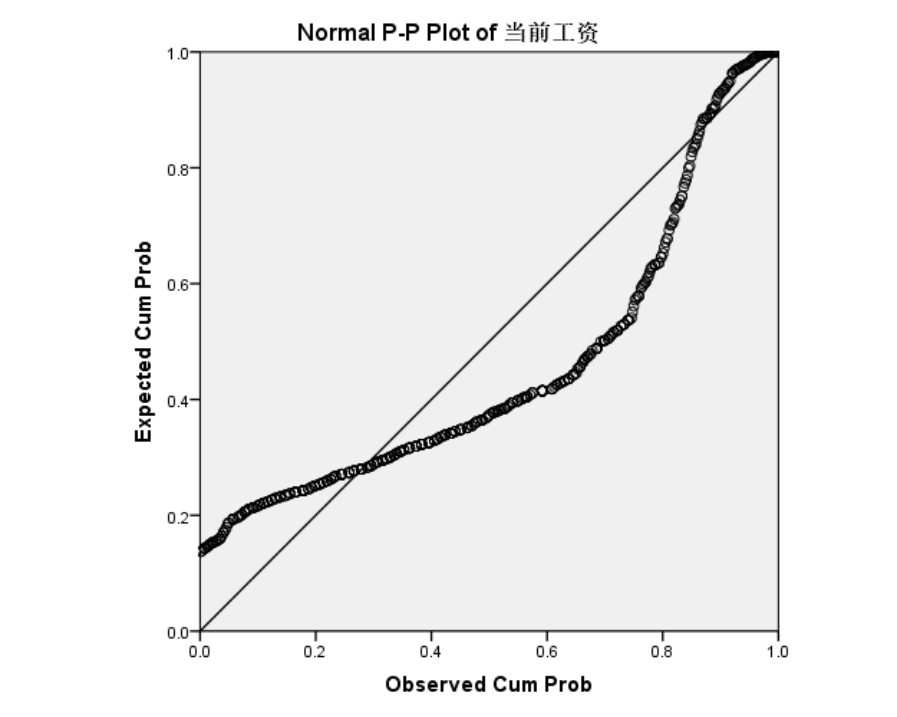


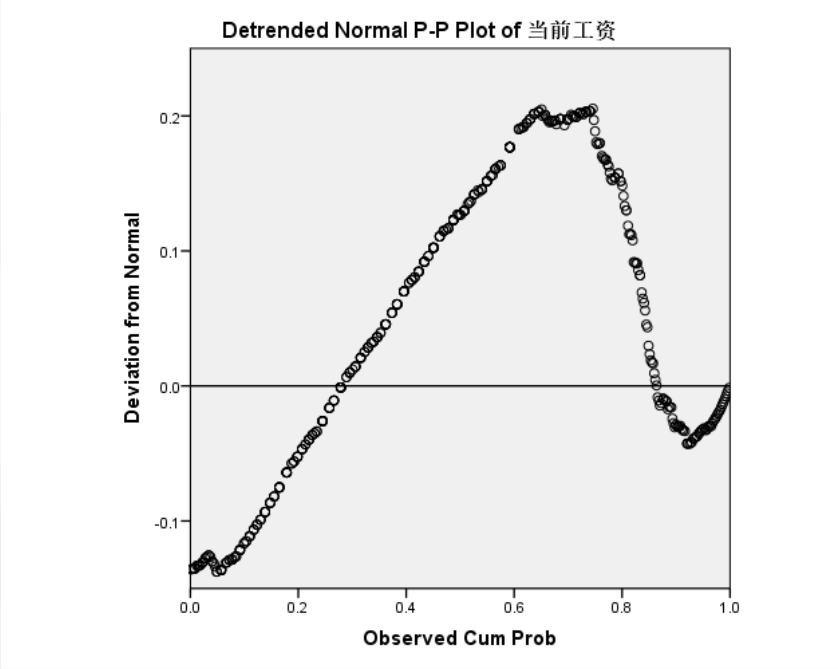


不常用，对分布中间更敏感

# P-P图

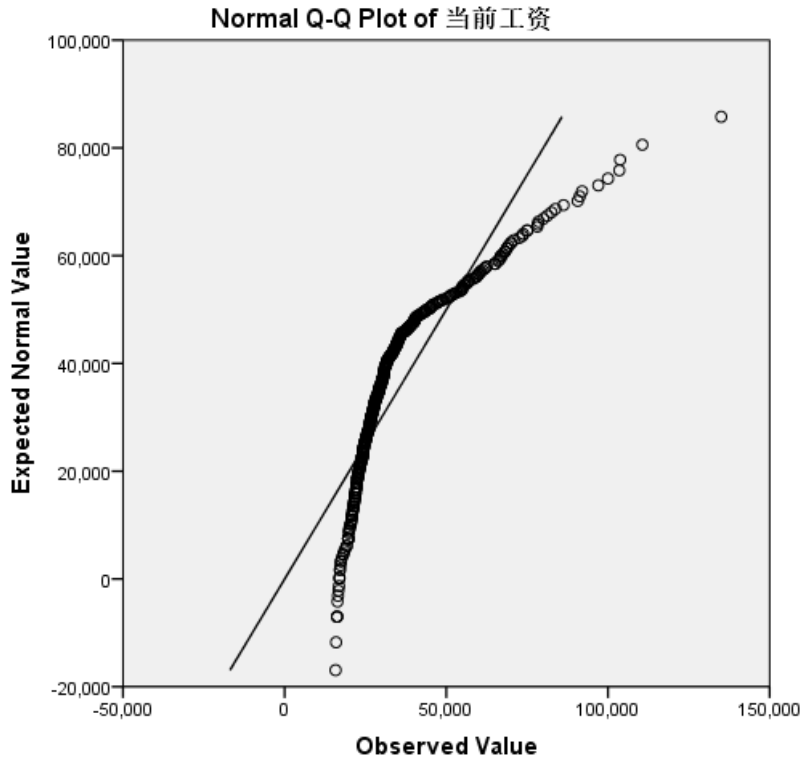


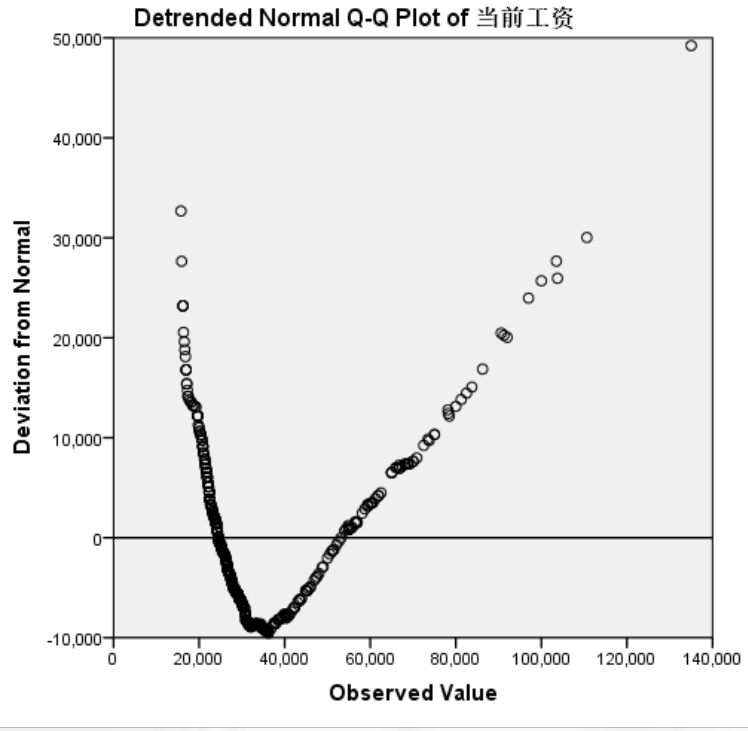




- 左图：画一个散点图，理论上的累计概率函数（某理论分布的CDF）与观察到的累计概率函数；如果实际数据和某分布一致，那么散点图应该在y=x的直线上。
- 右图：和P-P图不同的是，Y轴是两个CDF之间的差别。这样可以看出来，在什么地方两个分布不一致。如果完全一致，就是y=0的横线。
- 可检验的分布包括：正态、beta、卡方、指数、gamma、半正态、Laplace、Logistic、Lognormal等。

# Q-Q图

更常用，尾部更敏感




- 左图：画一个散点图，理论上的**分位数(quantile，这是数值)**与观察到的**分位数**；如果实际数据和某分布一致，那么散点图应该在y=x的直线上。
- 右图：detrended图还是和以前一样，Y是两条百分位数线的差异。如果一致，散点图应该是横线。
- 可检验的分布包括：正态、beta、卡方、指数、gamma、半正态、Laplace、Logistic、Lognormal等。

# Summary of Descriptive

- Screen data for recording errors, outliers, and distributional anomalies by using estimated numbers, tests, and graphs
- Descriptive: look at the basic info of variables
- Assumption checking: normal? symmetrical?