
清華大學
Tsinghua University

# 数据抽样和假设检验
# Data Sampling and Hypothesis Testing

吴志勇
清华大学深圳国际研究生院

# Example 1: 微信大数据（2017年9月数据）

- 微信日登录用户超9亿，
- 日发送消息380亿条，
- 日发送语音61亿次，
- 日成功通话次数超2亿，
- 微信用户日发表朋友圈视频次数6800万次

# Example 1: 微信大数据（2017年9月数据）

# Example 2: 产品品控

- 某厂有一批产品，共10000件，需检验合格后方能出厂。按规定要求次品率不得超过2%。
- 这批产品的次品率是多少?

# Example 3: 缺斤少两?

- 某奶制品厂生产奶粉，按规定每罐奶粉的标准质量为500g。
- 如何检测某批次奶粉的质量是否的确为500g?

# Example 4: “奶茶”妹妹 (Lady Testing Tea)

- “奶茶”妹妹宣称她可以辨别
  - 是奶倒入茶中 (milk into the tea)

  抑或
  - 是茶倒入奶中 (tea into the milk)

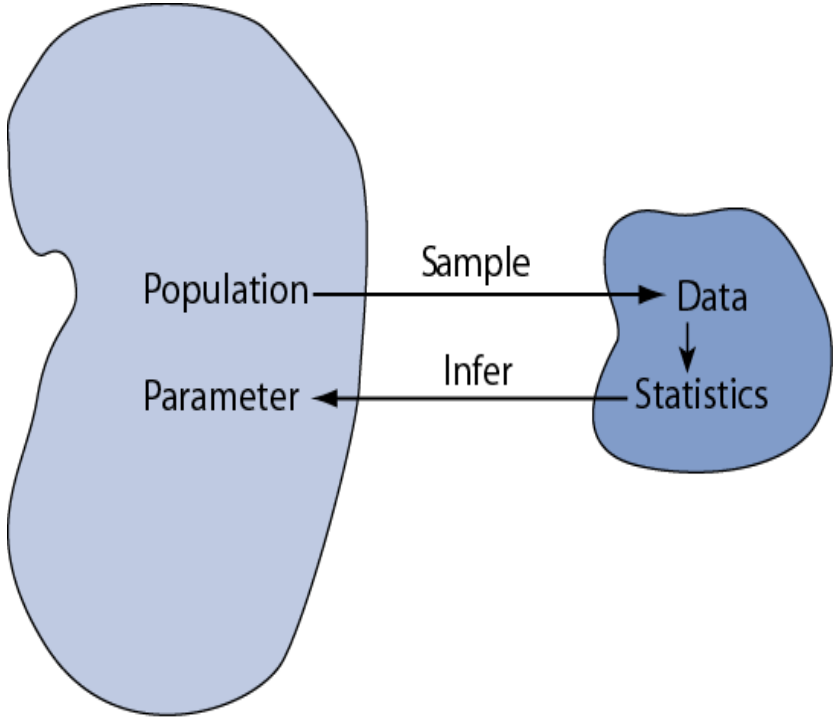
清華大學
Tsinghua University

# Statistical Inference
# 统计推断

# Statistical Inference (统计推断)

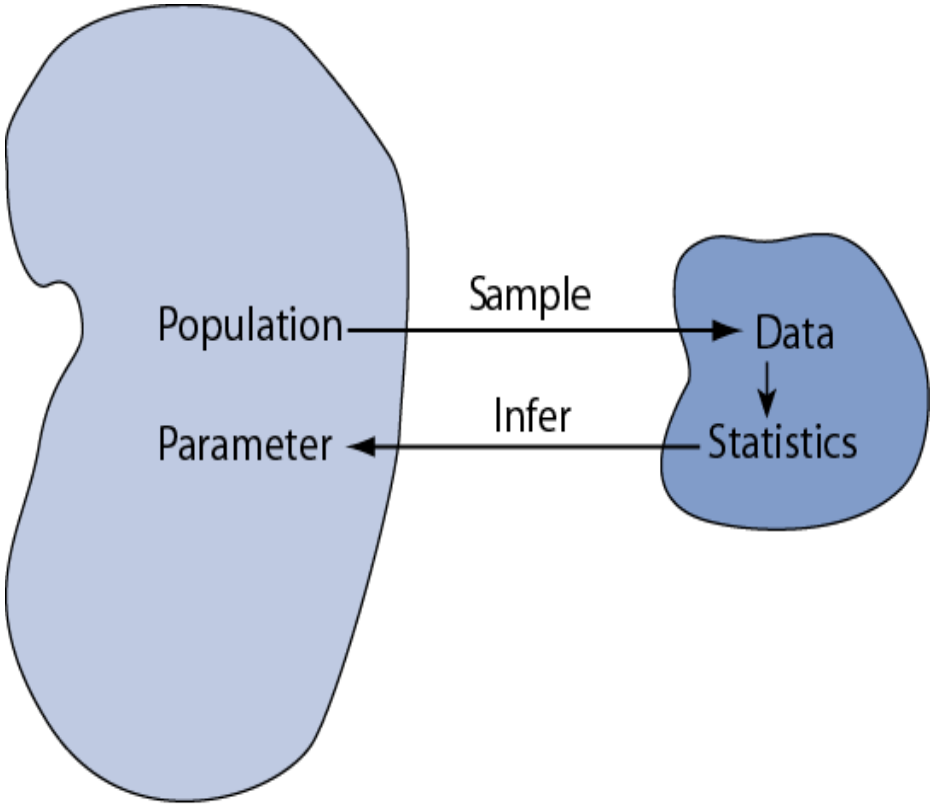
- **Statistical inference** is the act of generalizing from a **sample** (样本) to a **population** (总体) with calculated ***degree of certainty*** (置信度).

We want to learn about population *parameters…*



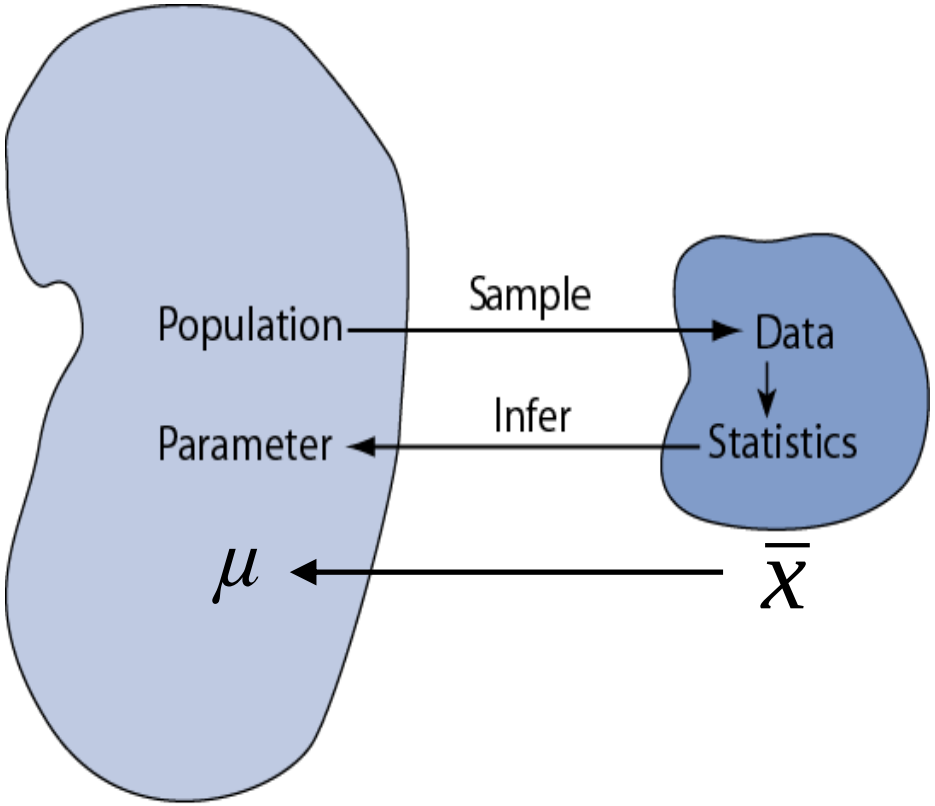

…but we can only calculate *sample statistics*

# Parameters (总体参数) vs. Statistics (样本统计量)




| | Parameters | Statistics |
|---|---|---|
| Source | Population | Sample |
| Calculated? | No | Yes |
| Constants? | Yes | No |
| Examples | $\mu, \sigma$ | $\bar{x}, s$ |

# Parameters (总体参数) vs. Statistics (样本统计量)




How well a given sample mean "x-bar" reflects an underlying population mean μ?

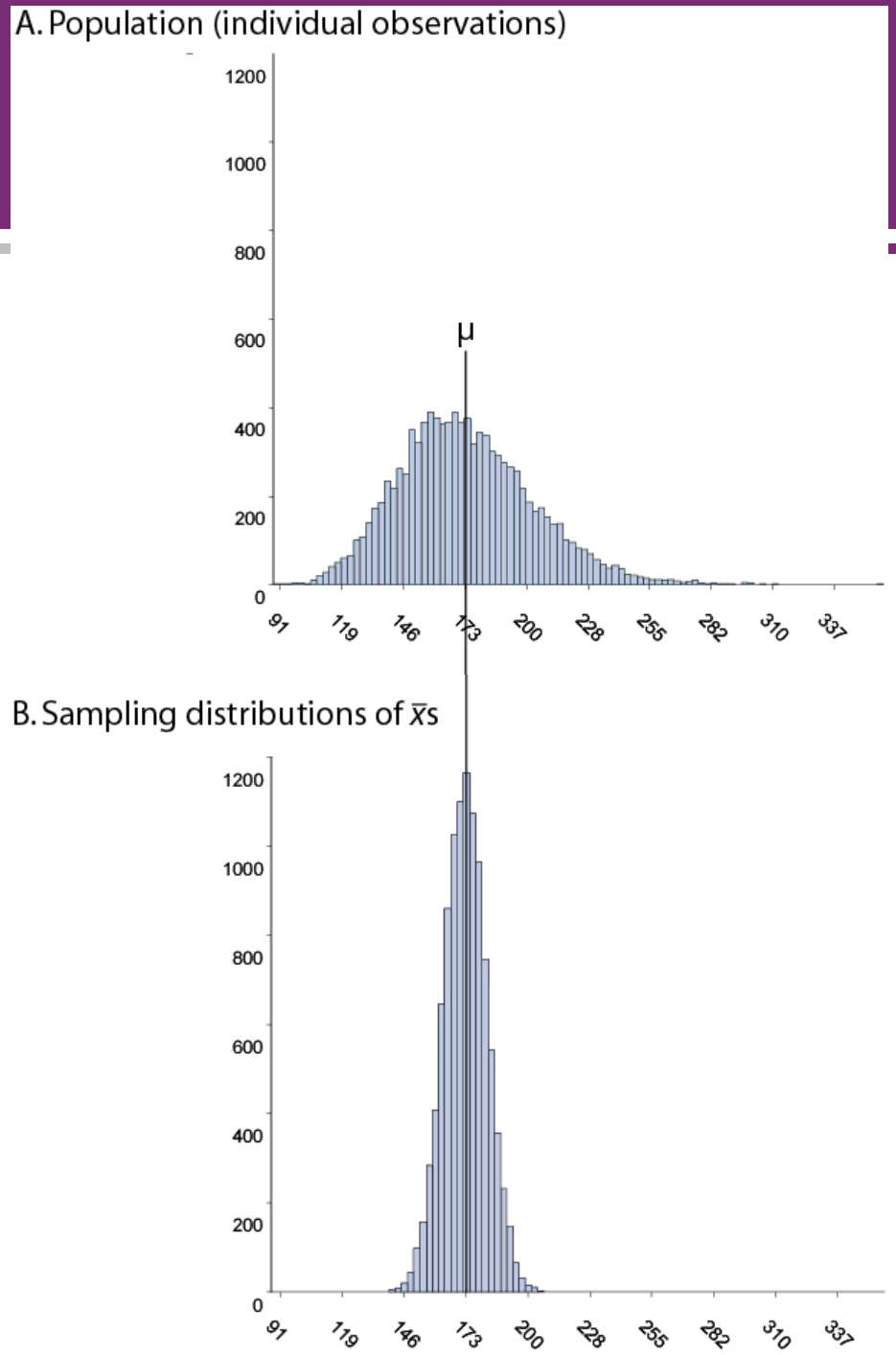
# Precision (精确度) and Reliability (可靠性)

- How precisely does a given sample mean (x-bar) reflect underlying population mean (μ)?
- How reliable are our inferences?

**Sampling Distribution of a Mean (SDM)**
**样本平均值的抽样分布**

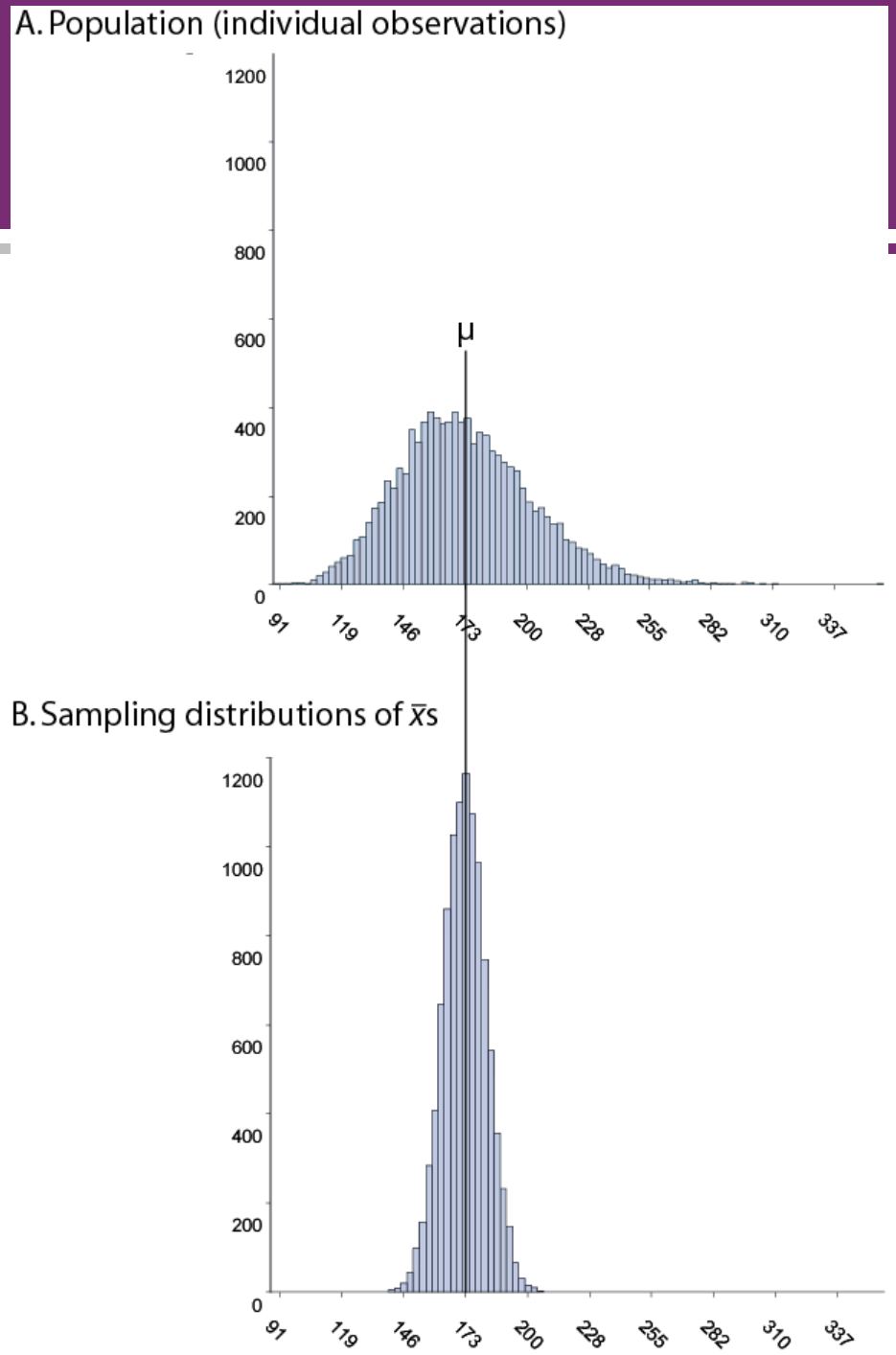
# Simulation Experiment of SDM

- Population (Fig. A)
  - N = 10,000
  - Lognormal shape (positive skew)
  - μ = 173
  - σ = 30
- Take repeated simple random samplings (SRSs), each with n = 10
- Calculate x-bar in each sample
- Plot x-bars (Fig. B)



# Simulation Experiment Results

- Distribution B is more Normal than distribution A
  ⇔ **Central Limit Theorem**
- Both distributions centered on μ
  ⇔ **x-bar is unbiased estimator of μ**
- Distribution B is skinnier than distribution A
  ⇔ related to "**square root law**"



# Reiteration of Key Findings

- **Finding 1:** central limit theorem, 中心极限定理
  - The sampling distribution of x-bar tends toward Normality even when the population is not Normal (esp. strong in large samples).
- **Finding 2:** unbiasedness, 无偏性
  - The expected value of x-bar is μ.
- **Finding 3:** square root law, 平方根法则

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

# Standard Deviation of the Mean

- Assume the sample data are Gaussian $N(\mu_0 = 173, \sigma^2 = 30^2)$
- Then, the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n=10} X_i \sim N(\mu_0 = 173, \frac{\sigma^2}{n} = \frac{30^2}{10} = 90)$

# Standard Deviation of the Mean

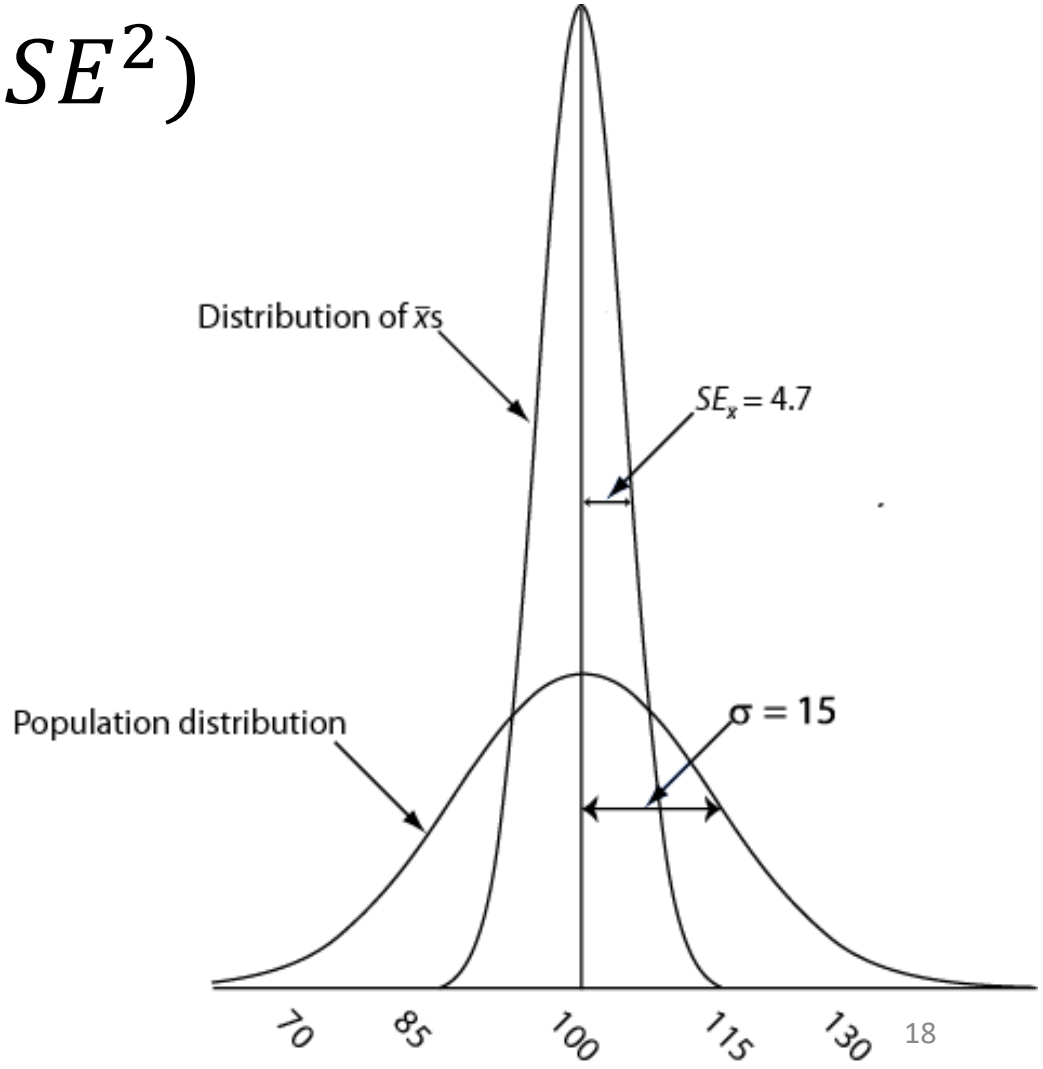
- The standard deviation of the sampling distribution of the mean (SDM) has a special name: **standard error of the mean** (denoted $\sigma_{xbar}$ or $SE_{xbar}$)
- The square root law (平方根法则) says:

$$\sigma_{\bar{x}} \equiv SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The precision of the sample mean is
inversely proportional to the square root of the sample size!

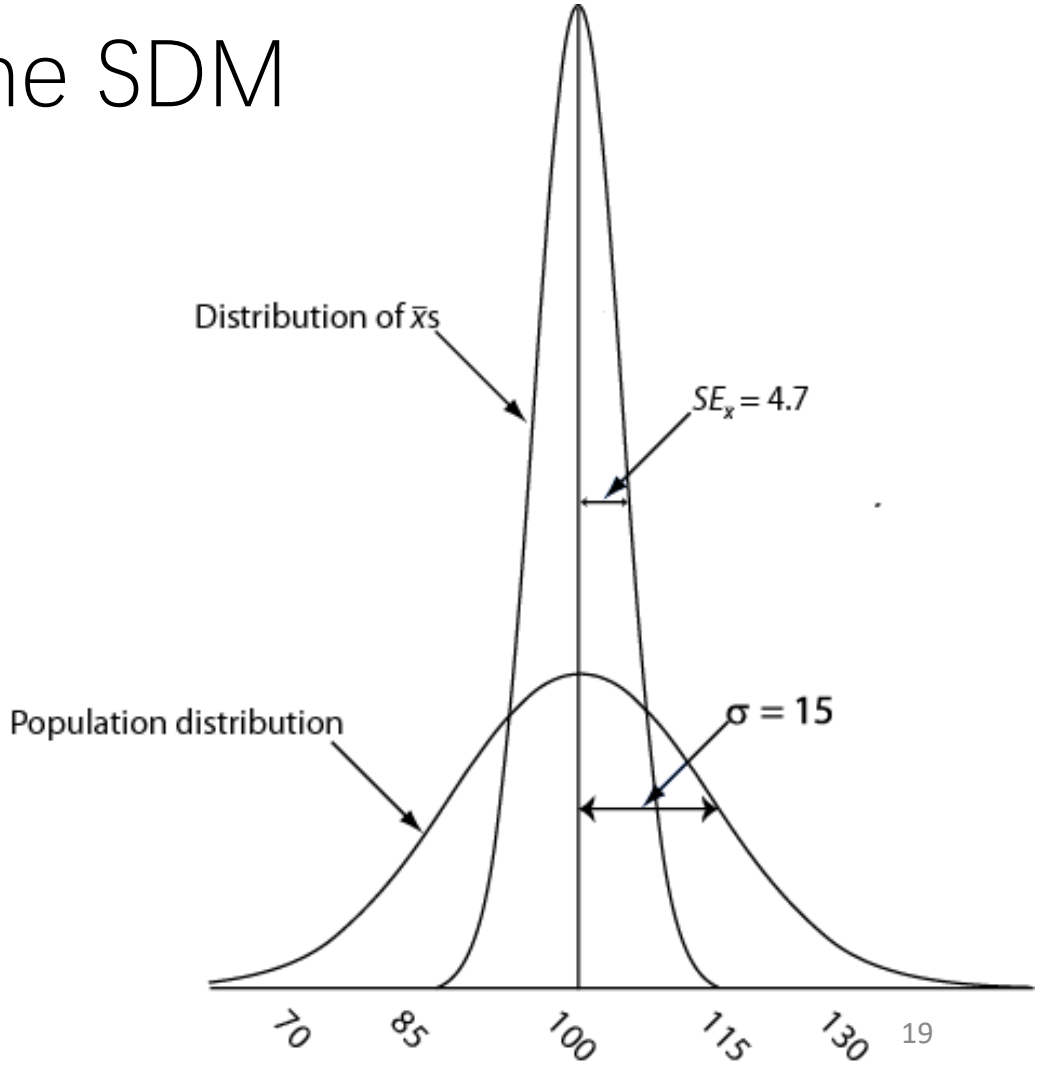
# Putting it together: $\bar{x} \sim N(\mu, SE^2)$

- The sampling distribution of $\bar{x}$ tends to be Normal with mean $\mu$ and stand deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
- Example:
- Let $X \sim N(100, 15^2)$
  - Take an SRS of $n = 10$
  - $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ = 15/ √ 10 = 4.7
  - Thus, $\bar{x} \sim N(100, 4.7^2)$



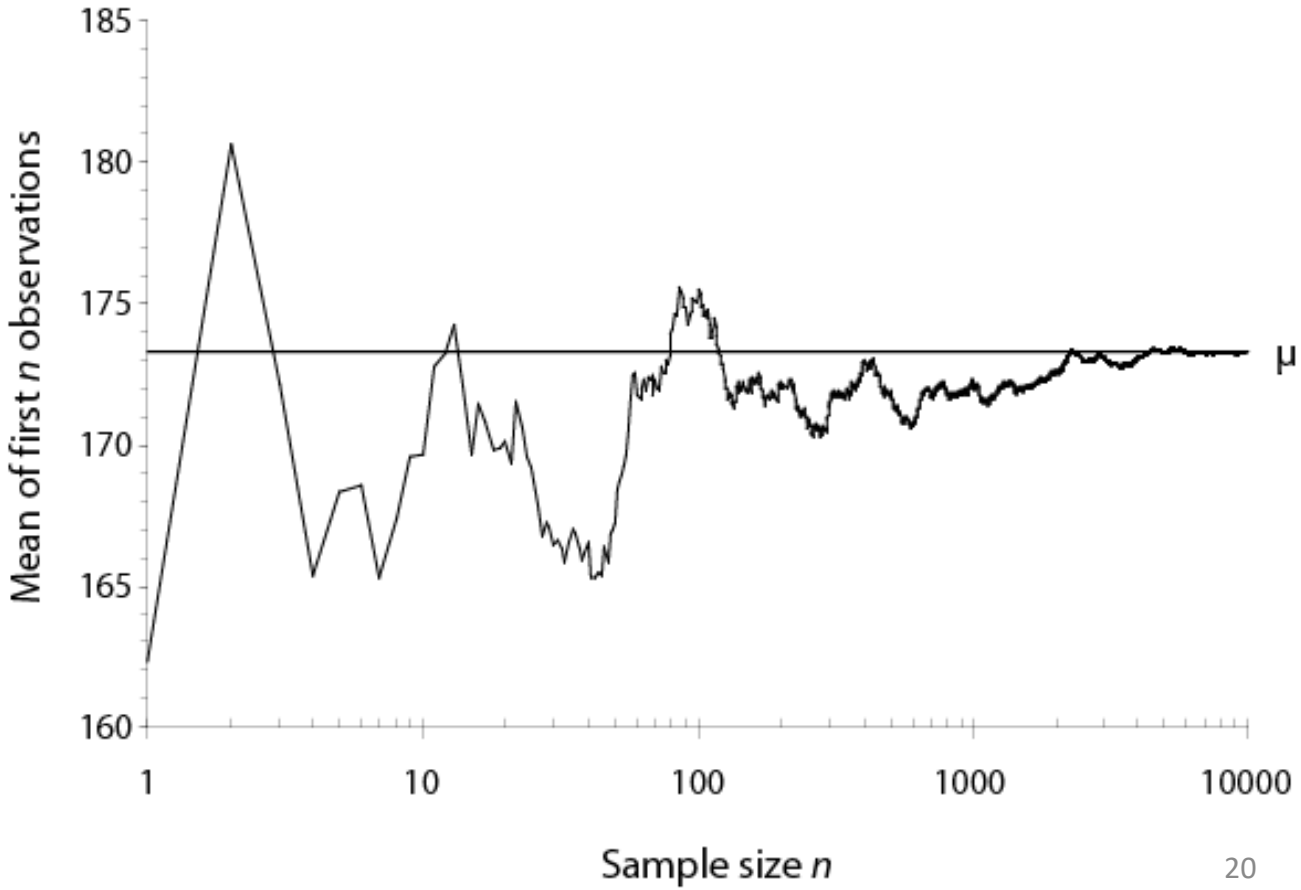
# 68-95-99.7 rule applied to the SDM

- We've established $\bar{x} \sim N(100, 4.7^2)$. Therefore,
- 68% of $\bar{x}$s within $\mu \pm \sigma_{\bar{x}}$
  = 100 $\pm$ 4.7
  = 95.3 to 104.7
- 95% of $\bar{x}$s within $\mu \pm 2 \cdot \sigma_{\bar{x}}$
  = 100 $\pm$ (2·4.7)
  = 90.6 to 109.4
- 99.7% of $\bar{x}$s within $\mu \pm 3 \cdot \sigma_{\bar{x}}$
  = 100 $\pm$ (3·4.7)
  = 85.9 to 114.1



# Law of Large Numbers (大数定律)

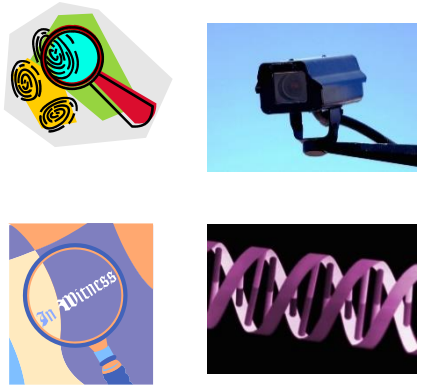
- As a sample gets larger and larger, the x-bar approaches μ.



# Hypothesis Testing
# 假设检验

# Example 1 (A Non-Statistical One)

- In a criminal trial, based on the evidence presented, a jury need to decide between two hypotheses:
  - the defendant is innocent
  - the defendant is guilty




# Example 2 (A Statistical One)

A quality engineer would like to determine

- whether the production process he is charged of monitoring is still producing products whose mean response value is supposed to be $\mu_0$ (process is in-control),
- or whether it is producing products whose mean response value is now different from the required value of $\mu_0$ (process is out-of-control).

- Statement 1 (Null): $\mu = \mu_0$ (process in-control)
- Statement 2 (Alternative): $\mu \neq \mu_0$ (process out-of-control)

# Hypothesis Testing (假设检验)

- Given sample data from a population (equivalently a distribution) with a parameter of interest (the mean, variance, standard deviation, etc.), we would like to decide/choose between two complementary statements concerning the parameter.
- These statements are called statistical hypotheses.

- 假设检验就是这样一种统计推断方法, 根据样本提供的信息对所提出的假设做出判断: 是接受, 还是拒绝.

## 1. 假设检验的基本原理

**小概率推断原理: $0 < \alpha \leq 0.05$**

小概率事件 (概率接近0的事件), 在一次实验中, 实际上可认为不会发生 (这是人们长期积累起的普遍经验!)

## 2. 假设检验的基本思想方法

**采用概率性质的反证法:**

先提出假设$H_0$, 再根据一次抽样得到的样本值进行计算. 若导致小概率事件发生, 则拒绝(否决)假设$H_0$; 否则, 接受假设$H_0$.

例3 某厂有一批产品, 共有10000件, 需检测合格方能出厂. 按规定次品率不得超过2%. 今从中任取100件, 发现有5件次品, 问这批产品能否出厂?

**分析**

从直观上分析, 这批产品不能出厂. 因为抽样得到的次品率: 5/100>2%

然而, 由于样本的随机性, 如何才能根据抽样结果判断总体 (所有产品) 的次品率是否 ≤2%?

**解** 用假设检验法, **步骤**:

1. 提出假设$H_0: p \leq 2\%$

   其中$p$为总体的次品率.

2. 设 $X_i = \begin{cases} 1, \text{第}i\text{次抽取的产品是次品} \\ 0, \text{否则} \end{cases}$

   则 $X_i \sim B(1, p) \quad (i = 1,2,3, \ldots, 100)$

   令 $Y = X_1 + X_2 + \cdots X_{100}$ = {抽取的100件产品中的次品数}

   则 $Y \sim B(100, p)$

3. 在假设$H_0$成立的条件$H_0$: $p\leq 2\%$, $Y\sim B(100,p)$下, 计算得到

$f(p)=P\{Y=5;\ p\}=C_{100}^{5}p^{5}(1-p)^{95}$,

因为$\frac{df(p)}{\mathrm{d}p}=C_{100}^{5}p^{4}(1-p)^{94}(5-100p)$

故当$p\leq 2\%$时, $f(p)$单调增加,

有 $f(p)=P\{Y=5;\ p\}=C_{100}^{5}p^{5}(1-p)^{95}\leq f(0.02)\approx 0.035<\alpha=0.05$ $(p\leq 2\%)$

从而 $P\{Y=5;p\}<\alpha=0.05$

故 $\{Y=5\}$ 是小概率事件.

4. 作判断

由于在假设$H_0$成立的条件下，$\{Y = 5\}$是小概率事件，

而实际情况是: 小概率事件竟然在一次试验中发生了，这违背了小概率原理，是不合理的.

故应该否定原假设$H_0$，认为产品的次品率 $p > 2\%$.

所以，这批产品不能出厂.

例4 某奶制品厂生产奶粉, 按规定每罐奶粉的标准质量为500g. 由以往经验可知, 该厂生产的罐装奶粉的质量服从正态分布 $N(500,4)$.
现随机抽取5罐, 其质量分别为 (单位: g): 501, 507, 498, 502, 504
能否认为该厂生产的奶粉每罐标准质量为500g?

**分析** 假设该厂生产的罐装奶粉平均质量 μ=500g,
则问题变为检验假设$H_0$: μ=500 是否成立?

由题设中以往经验可知, 标准差 $\sigma$=2, 则 $x$~$N(\mu, 2^2)$, 其中 $\mu$ 未知.

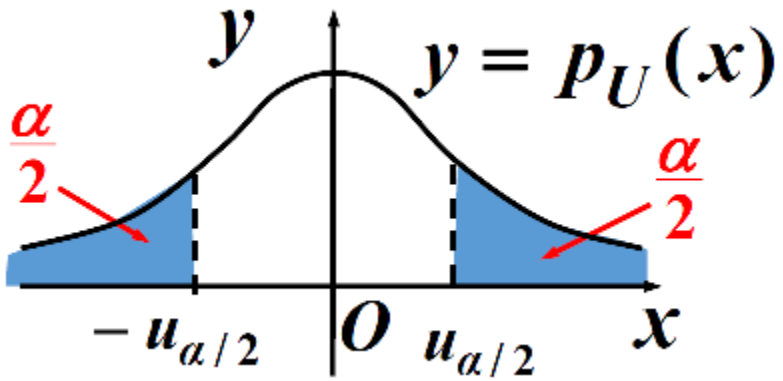
**问题**: 根据样本值判断 $\mu$=500 还是 $\mu \neq 500$ ?

**解**

1. 提出两个对立假设 $H_0$: $\mu=\mu_0$=500 , $H_1$: $\mu \neq \mu_0$;
2. 因为: $\bar{x}$ 是 $\mu$ 的无偏估计量,
   所以: 若 $H_0$ 为真, 则 $|\bar{x}-\mu_0|$ 不应太大,

   根据平方根法则, 衡量 $|\bar{x}-\mu_0|$ 大小可归结为衡量 $\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}}$ 的大小.

当 $H_0$ 为真时, $U=\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}}\sim N(0,1)$




因为$P\{|U|>u_{\alpha/2}\}=\alpha$,

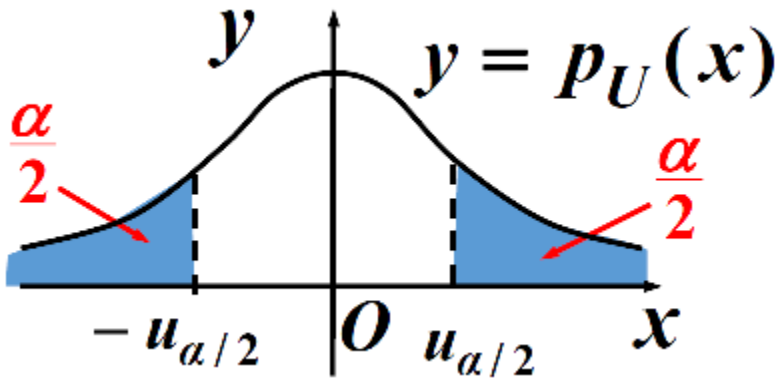
当$\alpha>0$ 很小时, $\{|U|>u_{\alpha/2}\}$是个小概率事件 (如上图).

根据小概率原理, 可以认为如果 $H_0$ 为真, 则由一次试验得到满足不等式 $|u|=\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}}>u_{\alpha/2}$ 的观察值几乎不会发生.

若在一次实验中得到了满足不等式

$|u| = \frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} > u_{\alpha/2}$ 的观察值 $\bar{x}$,

则我们有理由怀疑原来的假设 $H_0$ 的正确性, 因而拒绝 $H_0$.




若出现观察值满足不等式 $|u| = \frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} \leq u_{\alpha/2}$, 则没有理由拒绝假设 $H_0$, 因而只能接受 $H_0$.

当 $\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} > u_{\alpha/2}$ 时, 拒绝 $H_0$;

当 $\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} \leq u_{\alpha/2}$ 时, 接受 $H_0$.

若取定 $\alpha = 0.05$, 则 $u_{\alpha/2} = u_{0.025} = 1.96$ .

3. 在假设 $H_0$ 成立的条件下, 由样本计算

$$|u| = \frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} = 2.68 > 1.96 = u_{\alpha/2} = u_{0.025},$$

故拒绝假设 $H_0$ , 认为该厂罐装奶粉的标准质量不是500g.

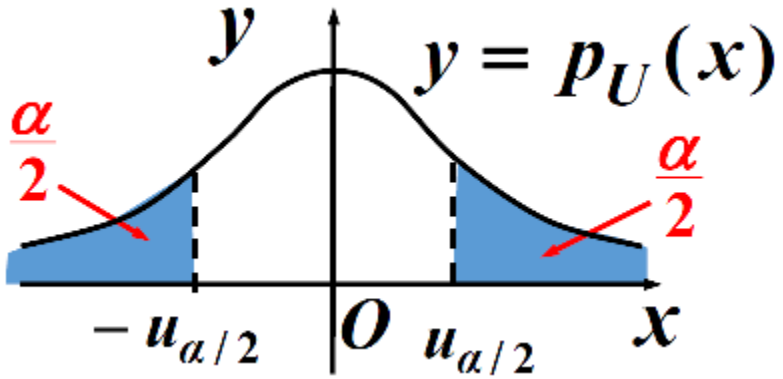
$n$=5, $\sigma$=2,
$\bar{x}$=502.4, $\mu$=500

当 $\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} > u_{\alpha/2}$ 时, 拒绝 $H_0$;
当 $\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} \leq u_{\alpha/2}$ 时, 接受 $H_0$.
若取定 $\alpha = 0.05$, 则 $u_{\alpha/2} = u_{0.025} = 1.96$ .
3. 在假设 $H_0$ 成立的条件下, 由样本计算
$|u| = \frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} = 2.68 > 1.96 = u_{\alpha/2} = u_{0.025}$,
故拒绝假设 $H_0$ , 认为该厂罐装奶粉的标准质量不是500g.

# 3. 假设检验的基本概念

- Significance Level (显著性水平)
- Test Statistic (检验统计量)
- Statistical Hypothesis (统计假设)
  - Null Hypothesis: $H_0$ (原假设/零假设)
  - Alternative Hypothesis: $H_1$ (备择假设)
- Critical Value (临界值/临界点)
- Type I and Type II Error (两类错误)
- P-value（P值）

## Significance Level (显著性水平)




当 $H_0$ 为真时, $U = \frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}} \sim N(0,1)$

因为$P\{|U| > u_{\alpha/2}\} = \alpha$, 当$\alpha > 0$ 很小时, 如果$\{|U| > u_{\alpha/2}\}$这个小概率事件都发生了, 说明$\bar{x}$与$\mu_0$的差异是显著的.

- The **significance level α** is the probability of rejecting $H_0$ given $H_0$ is correct

$$\alpha = P\{\text{拒绝} H_0 \mid H_0 \text{正确}\}$$

对于例4, 当 $H_0$: $\mu$=500 为真时, $U=\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}}\sim N(0,1)$

因为 $P\{|U|>u_{\alpha/2}|H_0\text{为真}\}=\alpha$,

如果 $|u|=\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}}>u_{\alpha/2}$ , 则称$\bar{x}$与$\mu_0$的差异是显著的, 则拒绝 $H_0$ .

反之, 如果 $|u|=\frac{|\bar{x}-\mu_0|}{\sigma/\sqrt{n}}\leq u_{\alpha/2}$ , 则称$\bar{x}$与$\mu_0$的差异是不显著的, 则接受 $H_0$ .

上述$\bar{x}$与$\mu_0$有无显著差异的判断是在显著性水平 $\alpha$ 之下做出的.

# Test Statistic (检验统计量)

- 用于假设检验的统计量, 称为检验统计量.
  - Quantity computed from sample data that measures the agreement of the sample data with the null hypothesis specification.
- Commonly used statistics:
  - $\bar{x}$ or equivalently the normalized version $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ (z-statistic) when $\sigma$ is known
  - t-statistic: $\mathrm{t} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ where $s$ is the sample standard deviation when $\sigma$ is unknown

# Statistical Hypotheses (统计假设)

- 假设检验问题通常叙述为: "在显著性水平 $\alpha$ 下, 检验假设 $H_0$: $\mu=\mu_0$, $H_1$: $\mu \neq \mu_0$"
- 或叙述为: "在显著性水平 $\alpha$ 下, 针对 $H_1$ 检验 $H_0$".
- $H_0$: Null hypothesis (原假设或零假设)
  - Is usually the hypothesis that corresponds to the status quo, the standard, the desired level/amount, or it represents the statement of "no difference."
- $H_1$: Alternative hypothesis (备择假设)
  - Is the complement of $H_0$.

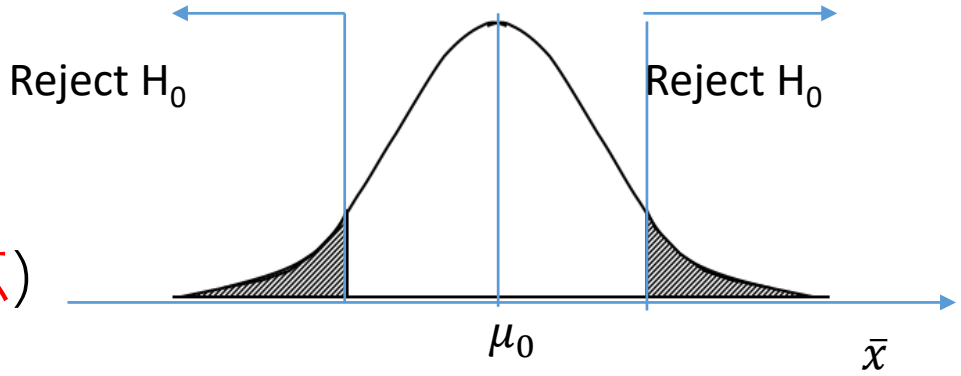
# Statistical Hypotheses

- In example 2 of quality control with $\mu_0 = 350$
  - Null $H_0$: $\mu = \mu_0$ (process in-control)
  - Alterative $H_1$: $\mu \neq \mu_0$ (process out-of-control)

- Two possible decisions based on sample data
  - There is enough evidence to reject the null (support the alternative hypothesis)
  - There is not enough evidence, and we fail to reject the null

# When to Reject $H_0$?

- We reject $H_0$ when the distance between $\bar{}$ and $\mu_0$ is large, or equivalently, |z| is large




- But how "large" is "large"?
  - Threshold $C$ (critical value, 临界值/临界点)
  - Reject $H_0$ when $|z| > C$

  - 如在前面例4中，临界值 $C = u_{\alpha/2} = u_{0.025} = 1.96$

# State of Reality and Decisions Made

| | | State of Reality | |
|---|---|---|---|
| | | $H_0$ is true | $H_1$ is true |
| Decisions Made | Do not reject $H_0$ | Correct decision | Type II error ($\beta$) |
| | Reject $H_0$ | Type I error ($\alpha$) | Correct decision |

- Type I error (false positive) $\alpha$:
  - reject a true null hypothesis (accuse an innocent)
- Type II error (false negative) $\beta$:
  - don't reject a false null hypothesis (set a guilty person free )

# Type I and Type II Error (两类错误)

- Type I error (false positive, 假阳性) $\alpha$:
  - Reject the null hypothesis when it is true
  - 把合法的判断为非法的: 误报
- Type II error (false negative, 假阴性) $\beta$:
  - Fail to reject the null hypothesis when it is false
  - 把非法的判断为合法的: 漏报

# Assessing the Errors

- These two types of errors are inversely related.
  - For a fixed sample size, decreasing one increases the other.
  - To decrease both, we have to increase the sample size (collect more evidence).
- In the court trial, a Type I error (false positive, convicting an innocent) is considered to be a ***more serious*** type of error.
- Therefore, we try to minimize the probability of committing the Type I error ($\alpha$).

## Steps for Hypothesis Testing (假设检验的一般步骤)

- Step 1: 根据实际问题要求, 提出待检验的原假设$H_0$和备择假设$H_1$;
- Step 2: 选择适当的统计量, 在$H_0$成立的条件下确定其概率分布;
- Step 3: 给定显著性水平$\alpha$, 确定拒绝原假设的临界值;
- Step 4: 根据样本观察值计算统计量的值;
- Step 5: 根据统计量值是否超过临界值, 做出拒绝或接受$H_0$的判断.

# Sampled Data of Example 2

- Assume that a random sample $(X_1, X_2, \ldots, X_n)$ of size $n$=25 are obtained with sample mean $\bar{x}$=370.16.
  - Assume data is Gaussian ($X_i \sim N(\mu, \sigma^2)$)
  - Assume the population standard deviation is $\sigma$=75.
  - If we don't know the true variance, we can use the sample variance.

# Statistical Hypotheses

- In example 2 of quality control with $\mu_0 = 350$
  - Null $H_0$: $\mu = \mu_0$ (process in-control)
  - Alterative $H_1$: $\mu \neq \mu_0$ (process out-of-control)

- Two possible decisions based on sample data
  - There is enough evidence to reject the null (support the alternative hypothesis)
  - There is not enough evidence, and we fail to reject the null

# The Decision Rule

- To decide whether the mean is not 350
  - a large sample mean $\bar{x}$ (say, 600) would provide enough evidence,
  - a $\bar{x}$ close to 350 (e.g., 355) does not provide enough evidence to infer that the mean is different from 350.

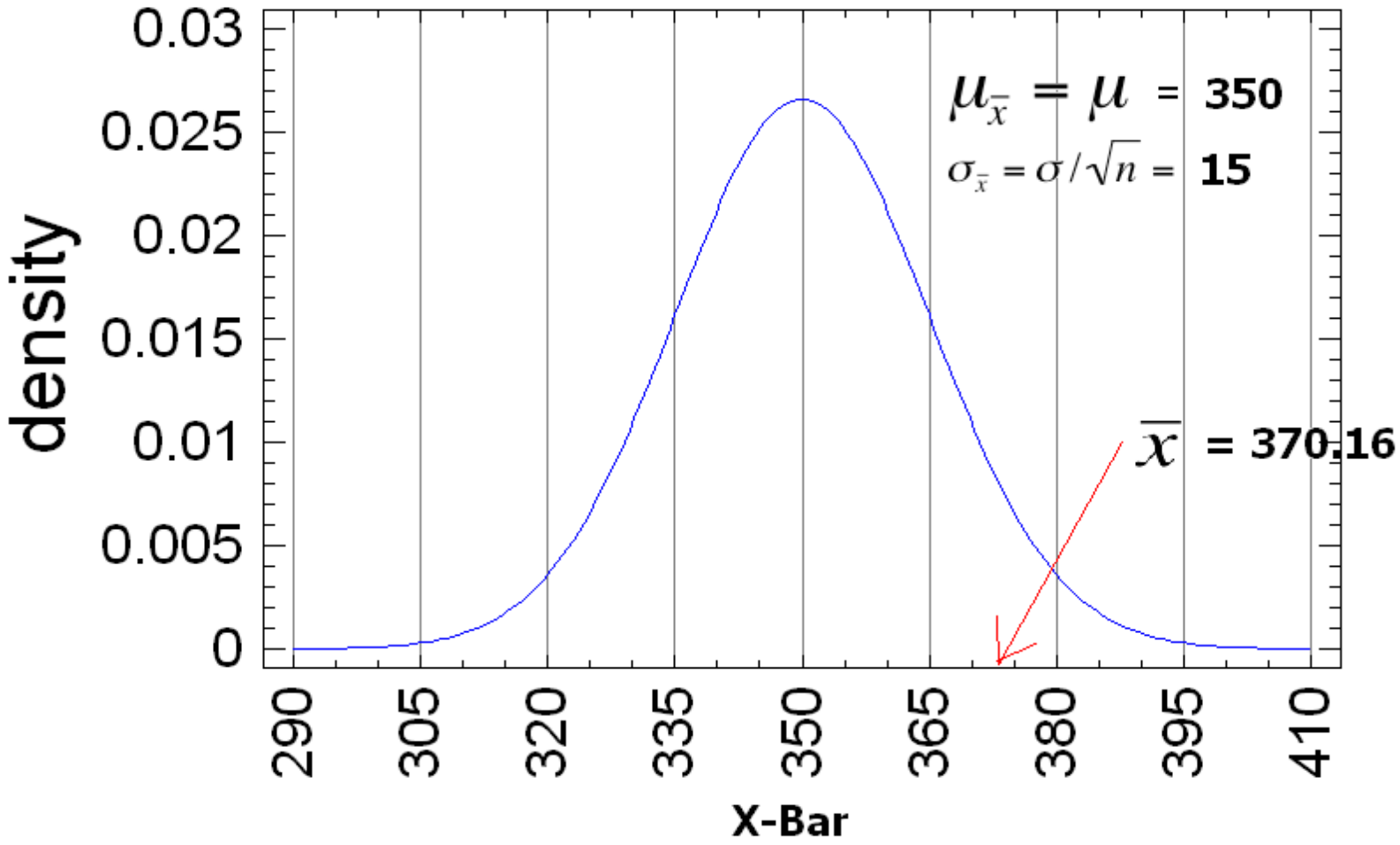
- The decision rule is the procedure that states when the null hypothesis $H_0$ will be rejected on the basis of the sample data.

# The Decision Rule

- The testing starts with the assumption that the null hypothesis is true (i.e., $\mu = \mu_0 = 350$)
  - Assume the sample (noisy) data are Gaussian $N(\mu_0 = 350, \sigma^2 = 75^2)$
  - Then, the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n=25} X_i \sim N(\mu_0 = 350, \frac{\sigma^2}{n} = \frac{75^2}{25} = 225)$

The defendant is presumed to be innocent until proven guilty

# Sampling Distribution of X-Bar




Do
the sample data agree with the null hypothesis spec?

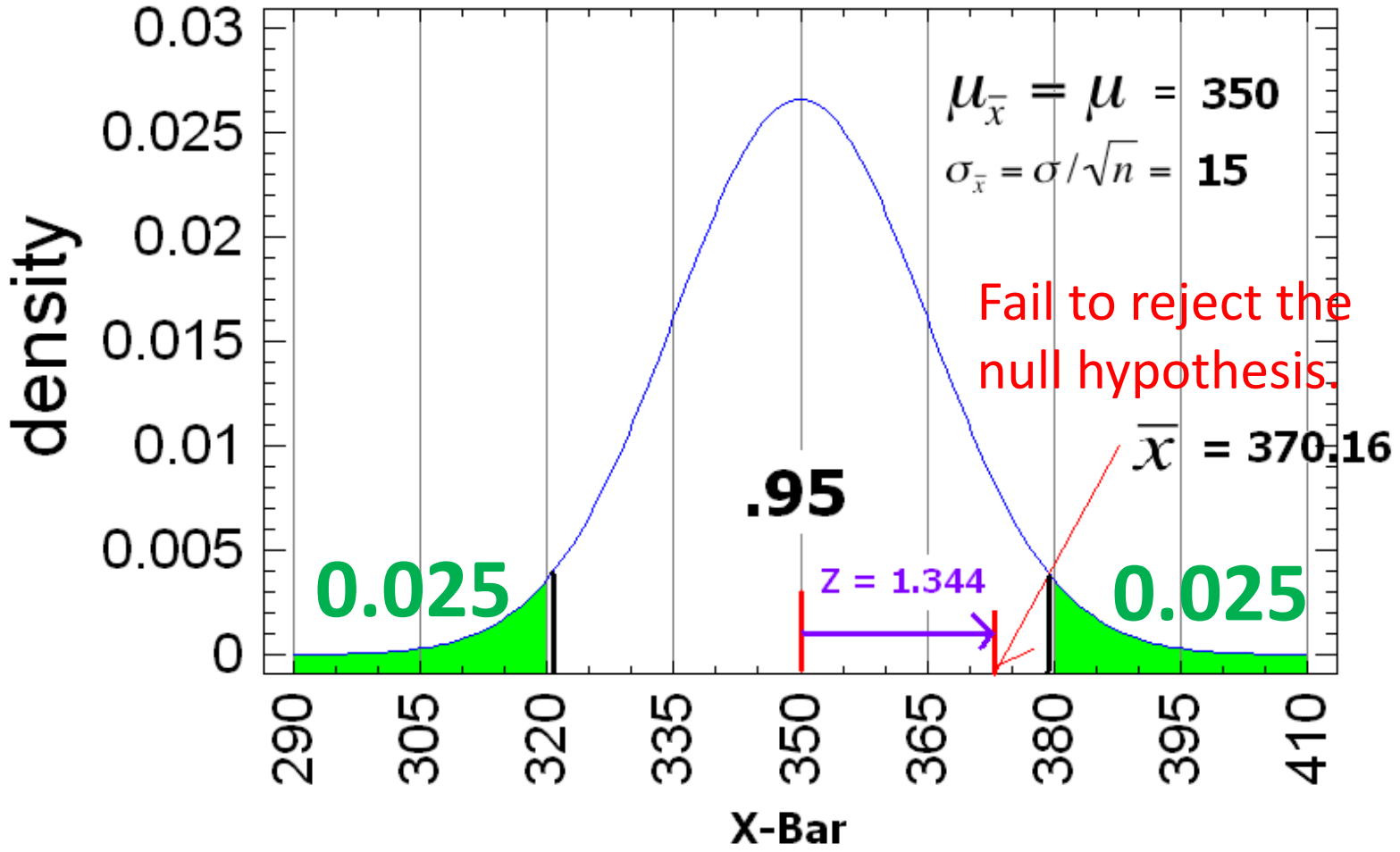
# Z statistic

- Assuming the null hypothesis is true → $z \sim N(0,1)$
- Choose the Critical Value (threshold) $C$ such that the Type I error probability is $\alpha = 0.05$.
  - Type I error probability

    $=P\{|z| > C|H_0\} = 2\int_C^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = 2Q(C) = 0.05$

    Q function
  - Choose $C = Z_{\alpha/2}$ where $Z_{\alpha/2} = Q^{-1}\left(\frac{\alpha}{2}\right) = 1.96$
  - Reject $H_0$ if |z|>C

# Sampling Distribution of X-Bar

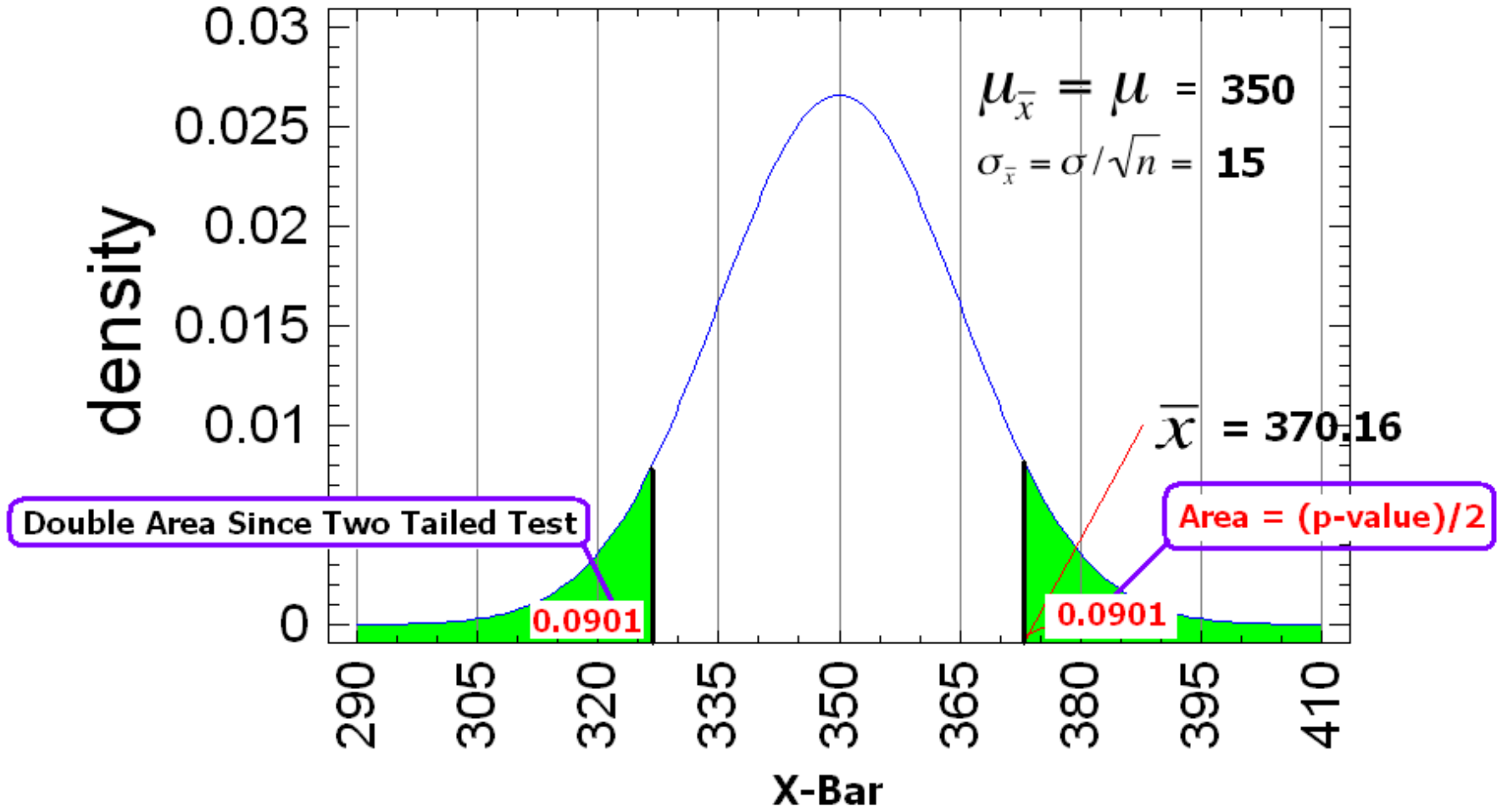


# P-value

- Change the "rejection region" (critical value $C$) such that it "captures/includes" the sample mean $\bar{x}$ ($z$=1.344 in our example).
- Then, calculate the Type I error rate
  - P-value = $P(|z| > 1.344) = 2*Q(1.344)$
    $= 2* 0.0901 = 0.1802 > 0.05$

- If P-value < $\alpha$: enough evidence to reject $H_0$
- Otherwise, fail to reject $H_0$

# Sampling Distribution of X-Bar

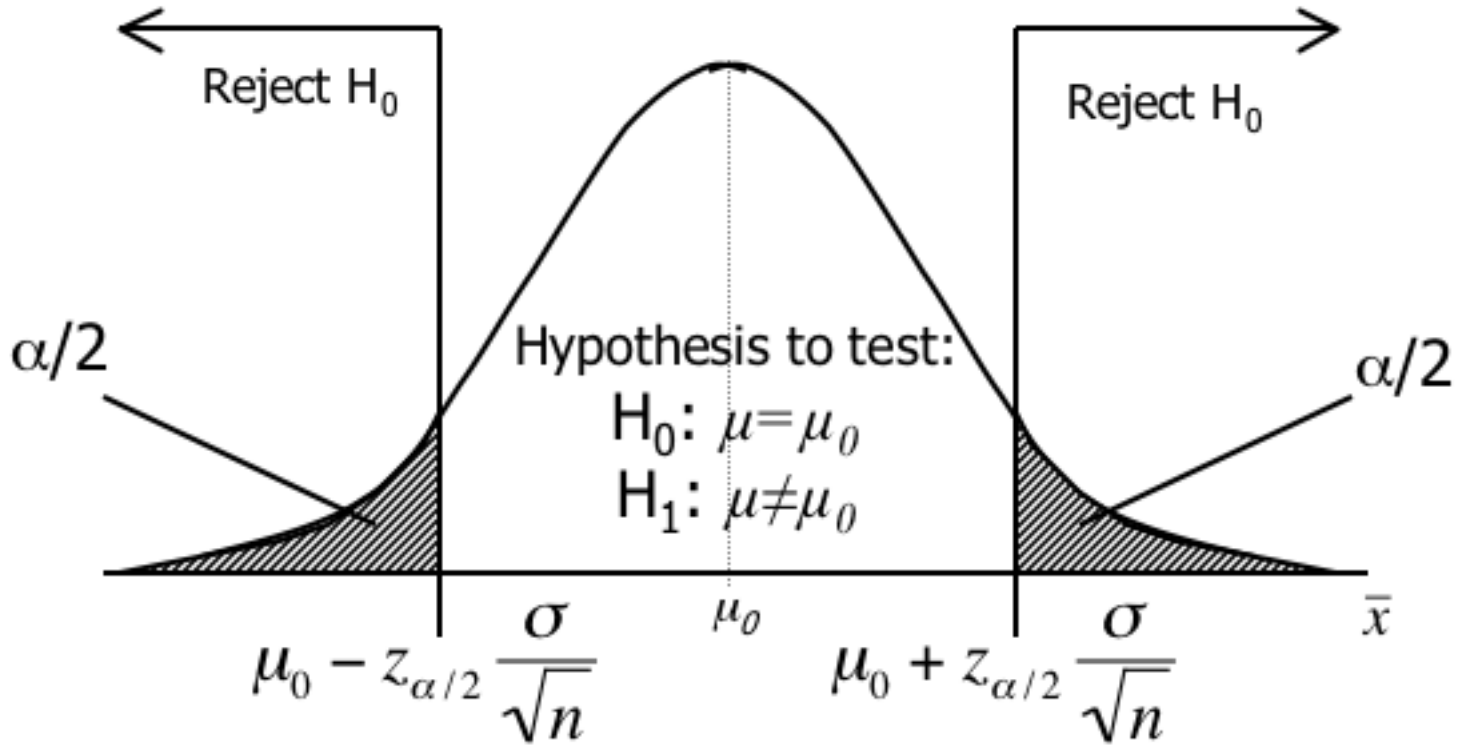


# Interpreting the P-Value

- The smaller the *P*-value, the more statistical evidence exists to against the null hypothesis $H_0$ (support the alternative hypothesis).
  - $P > 0.10 \Rightarrow$ ***no evidence*** against $H_0$
  - $0.05 < P \leq 0.10 \Rightarrow$ ***weak evidence*** against $H_0$
  - $0.01 < P \leq 0.05 \Rightarrow$ ***significant evidence*** against $H_0$
  - $P \leq 0.01 \Rightarrow$ ***highly significant evidence*** against $H_0$

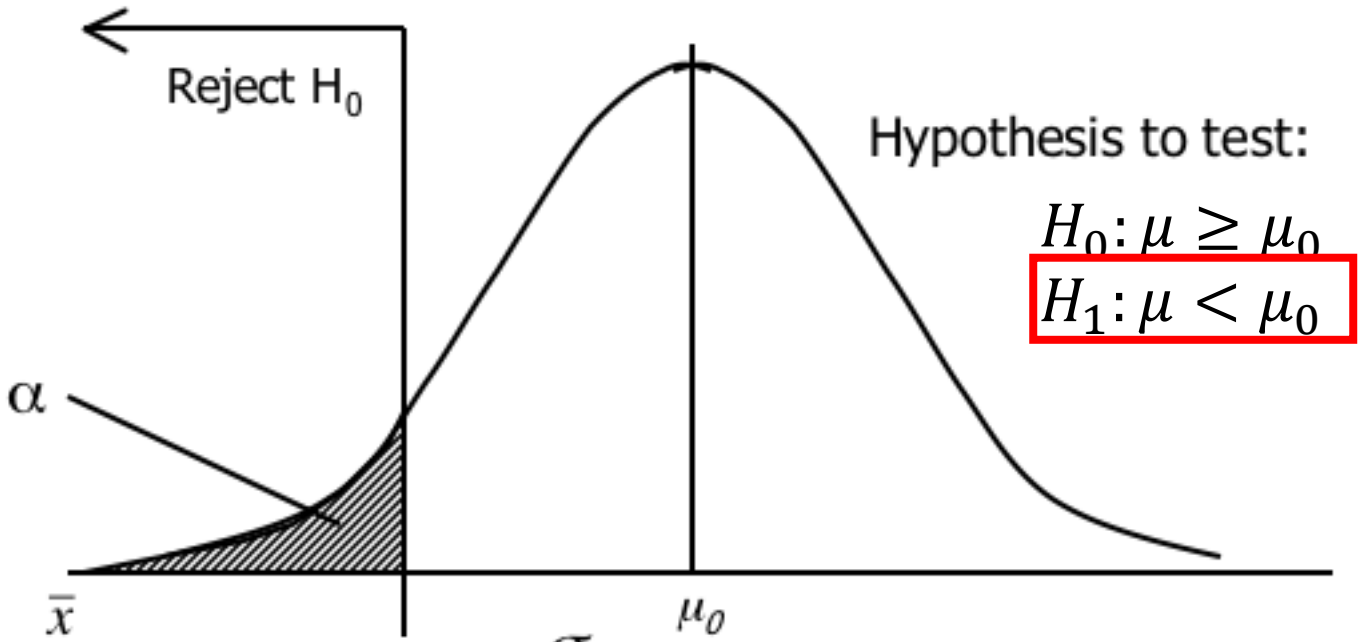
# Two-tail Test

- The rejection region is split equally between the two tails.



# Left-tail Test

- The rejection region is in the left tail



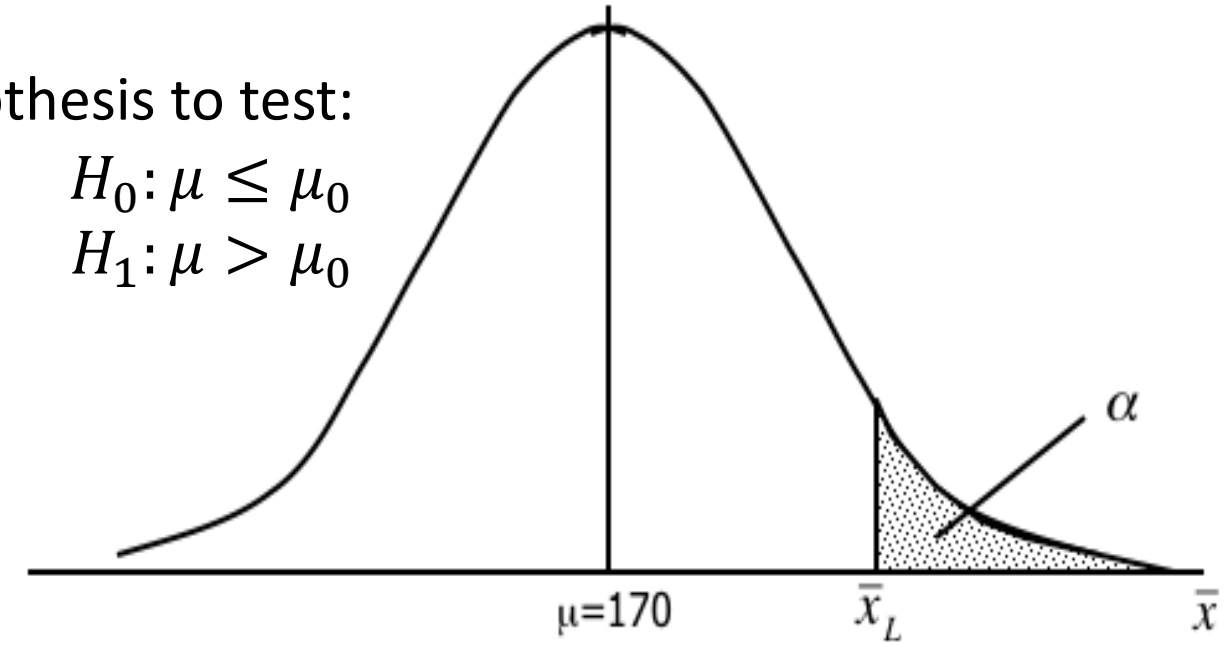
# Right-tail Test

- The rejection region is in the right tail

Hypothesis to test:

$H_0: \mu \leq \mu_0$

$H_1: \mu > \mu_0$



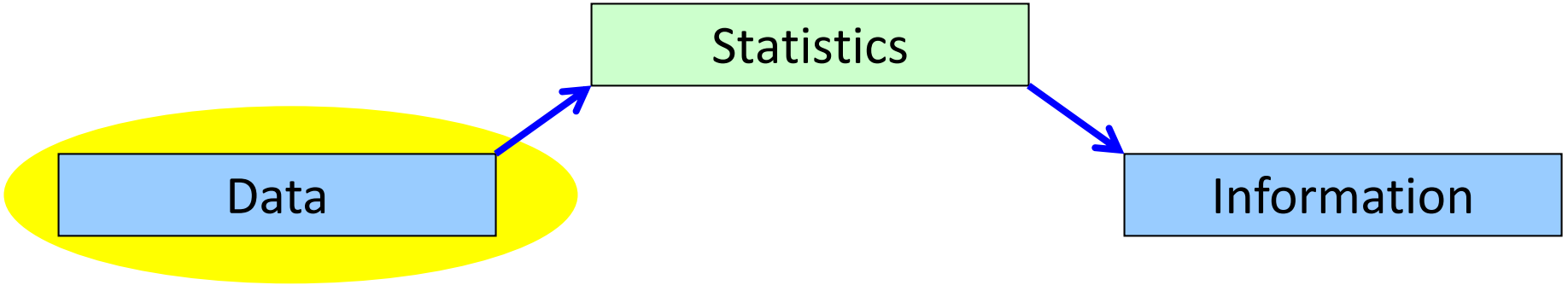
# References


- Introduction to Hypothesis Testing
  - David F. Groebner, Patrick W. Shannon, Philip C. Fry and Kent D. Smith, *Business Statistics*, 7e, Prentice Hall, Chap. 9, http://www.prenhall.com/divisions/bp/app/chapters/groebner7/Chapter09.pdf
  - Frederick J. Gravetter and Larry B. Wallnau, *Statistics for the Behavioral Sciences*, Sage Pub., Chapter 8, http://www.sagepub.com/upm-data/40007_Chapter8.pdf

- More mathematics and statistics:
  - Steven Kay, *Fundamentals of Statistical Signal Processing, Vol. II: Detection Theory*, Prentice Hall

# A Few Words about Data Collection and Sampling
# 数据获取与数据抽样

# Data Collection




- Statistics is a tool for converting ***data*** into ***information***:
  - But where does ***data*** come from?
  - How is it gathered?
  - How do we ensure it's accurate?
  - Is the data reliable?
  - Is it representative of the population from which it was drawn?

# Data Sampling

- Statistical inference permits us to draw conclusions about a population based on a ***sample***.
- Sampling
  - Selecting a sub-set of a whole population
  - Is often done for reasons of ***cost***
    - It's less expensive to sample 1,000 TV viewers than 100 million TV viewers.
  - and ***practicality***
    - Performing a crash test on every automobile produced is impractical.
- In any case, the ***sampled population*** and the ***target population*** should be ***similar*** to one another.

# Sampling Plans

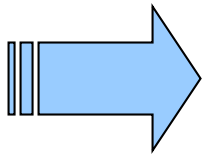
- A sampling plan is just a method or procedure for specifying how a sample will be taken from a population.
- Methods:
  - Simple random sampling (简单随机抽样): by far the most common one used
  - Systematic random sampling (周期系统随机抽样)
  - Stratified random sampling (分层随机抽样)
  - Cluster sampling (整群抽样)

# Simple Random Sampling (简单随机抽样)

- Every possible sample of the same size is equally likely to be chosen.
- Example:
  - Drawing 3 names from a hat containing all names of the students in the class

- A government income tax auditor must choose a sample of 5 of 11 returns to audit … [Can do in many different ways]

| Person | Generate Random # |
|---|---|
| baker | 0.87487 |
| george | 0.89068 |
| ralph | 0.11597 |
| mary | 0.58635 |
| sally | 0.34346 |
| joe | 0.24662 |
| andrea | 0.47609 |
| mark | 0.08350 |
| greg | 0.53542 |
| aaron | 0.37239 |
| kim | 0.73809 |



| | Person | Sorted Random # |
|---|---|---|
| 1 | mark | 0.08350 |
| 2 | ralph | 0.11597 |
| 3 | joe | 0.24662 |
| 4 | sally | 0.34346 |
| 5 | aaron | 0.37239 |
| | andrea | 0.47609 |
| | greg | 0.53542 |
| | mary | 0.58635 |
| | kim | 0.73809 |
| | baker | 0.87487 |
| | george | 0.89068 |

# Systematic Random Sampling (周期系统随机抽样)

- Randomly select a starting point, then select every *n*-th individual.
  - *n*: sampling interval (pop. size/sample size)
- Example
  - Number all students in a list,
  - Randomly select a stating point in the list, and
  - Select every 5th individual (selecting 20% of the student population)

# Stratified Random Sampling (分层随机抽样)

- Separate the population into mutually exclusive sets or strata, and then drawing simple random samples from each stratum.

| Strata 1 : Gender | Strata 2 : Age | Strata 3 : Occupation |
|---|---|---|
| Male | < 20 | professional |
| Female | 20-30 | clerical |
| | 31-40 | blue collar |
| | 41-50 | other |
| | 51-60 | |
| | > 60 | |

We can acquire about the total population, make inferences within a stratum or make comparisons across strata

# Stratified Random Sampling

- After the population has been stratified, we can use ***simple random sampling*** to generate the complete sample:

| Income Category | Population Proportion | Sample Size n = 400 | n = 1000 |
|---|---|---|---|
| under $25,000 | 25% | 100 | 250 |
| $25,000 - $39,999 | 40% | 160 | 400 |
| $40,000 – $60,000 | 30% | 120 | 300 |
| over $60,000 | 5% | 20 | 50 |

If we only have sufficient resources to sample 400 people total, we would draw 100 of them from the low income group…

…if we are sampling 1000 people, we'd draw 50 of them from the high income group.

# Cluster Sampling (整群抽样)

- Population is organized into groups, where groups are randomly selected and all members of the group are sampled.
  - Example:
    - The Government randomly selects 5 high schools in Jiangsu Province and surveys each teacher in those schools.
- Useful when it is difficult/costly to develop a complete list of the population, or when the population is widely dispersed geographically.
- May increase sampling error due to similarities among cluster members

## Sample Size

- It is suffice to say that the larger sample size is, the more accurate we can expect the sample estimates to be.

**Law of Large Numbers (大数定律)**

# Sampling and Non-sampling Errors

- Two major types of error: *Sampling error* and *Non-sampling error*
- Sampling Error
  - Differences between the sample and the population that exist only because of the observations selected for the sample.
  - Random and can be reduced by a larger sample size.
- Non-sampling Error
  - Due to **mistakes** made in the acquisition of data or due to the sample observations being selected **improperly**.
  - Mostly likely caused by poor planning, etc.

# Non-sampling Error

- Three types of non-sampling errors
  - Errors in data acquisition
  - Non-response errors
  - Selection bias

- Increasing the sample size **will not** reduce this type of error.

# Errors in Data Acquisition

- Arises from the recording of incorrect responses, due to:
  - incorrect measurements being taken because of faulty equipment,
  - mistakes made during transcription from primary sources,
  - inaccurate recording of data due to misinterpretation of terms, or
  - inaccurate responses to questions concerning sensitive issues.

# Non-response Error

- Refers to error (or ***bias***) introduced when responses are not obtained from some members of the sample
  - The collected sample observations may not be representative of the target population.

- The ***Response Rate*** is a key survey parameter and helps understand the validity of the survey and sources of nonresponse error.

# Selection Bias

- Occurs when the sampling plan is flawed and some members of the target population cannot possibly be selected for inclusion in the sample.

# Q&A?