生物医学信号处理与分析

Biomedical Signal Processing and Analysis

2025.09.02 – 2025.10.23

第三讲 随机信号基础1

概率与统计与本课程相关的主要概念回顾

（自学环节）

学习要点

- 生物医学信号随机性的典型来源
- 概率论与统计学的应用要点
  - 随机变量
  - 随机样本与抽样分布
  - 统计推断1——参数估计
  - 统计推断2——假设检验
- 模型与仿真实践
- 主要参考资料
  - 盛骤编著《概率论与数理统计》教材，同学们根据自己的情况复习相关内容

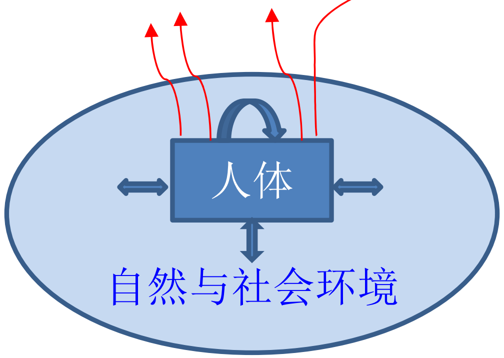
医学信号随机特性的来源

- 个体差异
  - 年龄、性别、人种、生活习惯、职业等不同，例如运动员心率相对较低，肺活量较高
- 状态变化
  - 同一个人不同时间的心率、血压、体温和脑电等信号随着清醒-睡眠、紧张-放松、高兴-悲伤、健康-疾病等状态的变化而变化
- 环境影响
  - 温度、气压、噪声水平等
- 时间变化
  - 同一人的体温、身高等测量数据在一天内呈现准周期变化
- 测量设备
  - 精度、稳定性、噪声水平、环境干扰等

医学信号产生与观测模型

医学数据产生

是人体自身各部分，以及人体与环境之间相互作用的结果，无时无刻都在产生数据。



医学数据采集

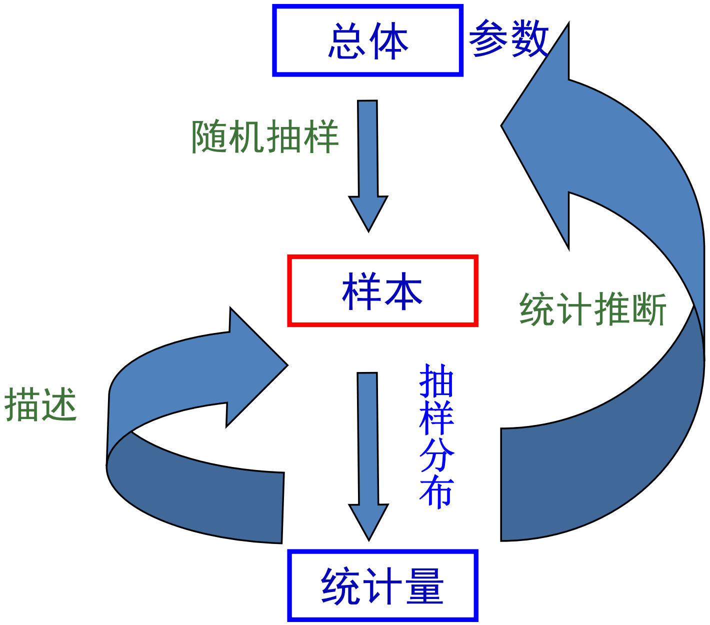
仅仅是对无时无刻产生的数据的一个极其有限的采样：

（1）有限种类：例如脑电、心电

（2）有限通道：例如32通道

（3）有限时间：例如1分钟至几天

基于数据的医学诊断与研究



目标： 临床诊断、基础研究

方法： 从有限的观测数据（随机样本）得到关于样本的基本知识——统计量，进而通过统计推断获得关于总体的参数，推测数据产生的机制（疾病诊断，人体的内在机制及其与环境的相互作用等）

数学基础：

（1）概率论与数理统计

（2）随机信号（引入时间变量，下次课）

随机变量

- 随机事件
  - 随机试验中，可能出现也可能不出现，但是大量重复试验中具备某些规律性的事件
- 随机事件的样本空间
  - 所有可能发生事件的集合
- 随机变量
  - 用于描述随机事件的数学工具
  - 设E为随机试验，样本空间$S = \{e_i\}$，如果对于每一个$e_i \in S$，有唯一一个实数$X(e_i)$与之对应，则$X(e_i)$称为随机变量，简写为X

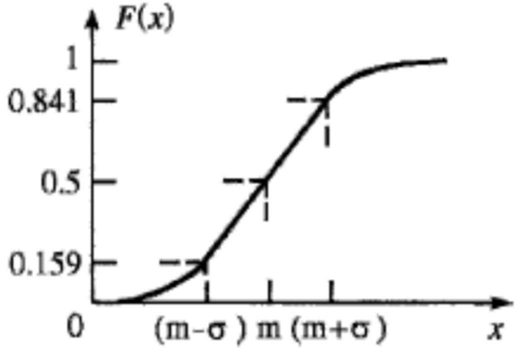
离散型随机变量的分布律

- 离散型随机变量
  - 有限个数的取值，或者可列无限多个取值

- 离散型随机变量X的分布律
  - $P(X=x_k) = p_k$, $k = 1, 2, 3, \ldots$

概率分布函数性质

- 定义：$F(\alpha) = P(x \leq \alpha)$

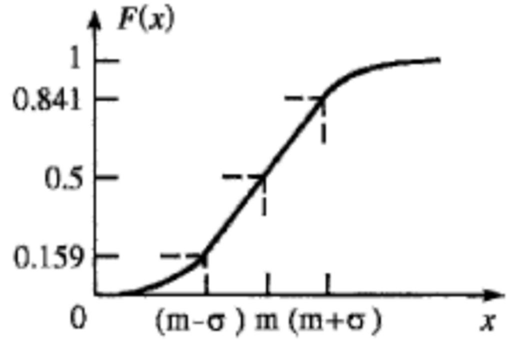


- 性质：
  - 完备刻画样本空间中的任何数值的发生概率

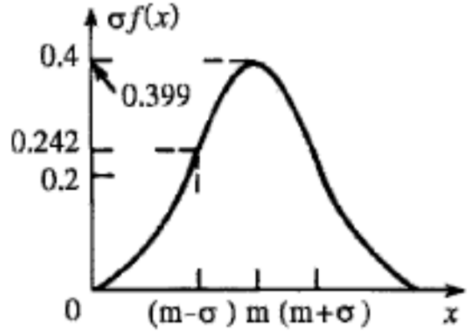
1. $0 \leq F_x(\alpha) \leq 1, \ -\infty \leq \alpha \leq -\infty;$
2. $F_x(-\infty) = 0, \quad F_x(\infty) = 1;$
3. $F_x(\alpha)$ is nondecreasing with α;
4. $P[\alpha_1 \leq x \leq \alpha_2] = F_x(\alpha_2) - F_x(\alpha_1)$

连续随机变量的概率密度函数

- **概率密度函数** (Prob. density function, PDF)
  - 定义： $f(\alpha)=\frac{dF(\alpha)}{d\alpha}$
  - 性质：



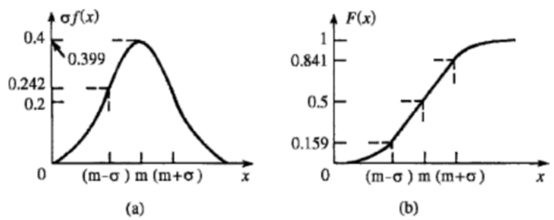
1. $f_x(\alpha)\geq 0 \quad -\infty\leq\alpha\leq-\infty;$
2. $\int\limits_{-\infty}^{\infty} f_x(u)du=1;$
3. $F_x(\alpha)=\int\limits_{-\infty}^{\alpha} f_x(u)du;$
4. $P[\alpha_1\leq x\leq\alpha_2]=\int\limits_{\alpha_1}^{\alpha_2} f_x(u)du;$



高斯随机变量

- 概率密度函数可表达为如下公式（正态分布函数）的随机变量称作高斯随机变量

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-m)^2}{2\sigma^2}), -\infty \le x \le \infty$$



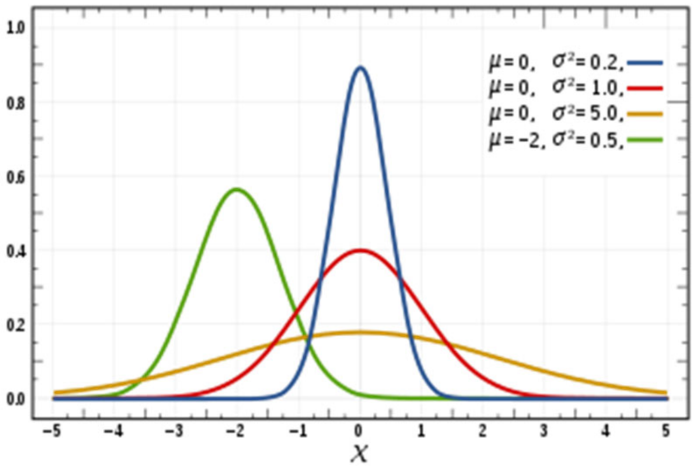
- 优点：均值和方差完全确定概率密度函数

- 生物医学信号是大量独立随机事件相互作用的结果，如果每个因素对信号的贡献都很小，总的贡献构成测量到的信号，这个信号的幅值特性可以用正态分布描述

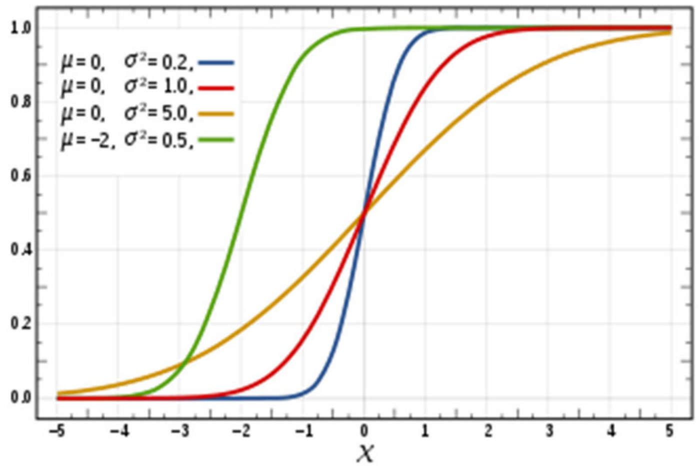
- 在描述随机噪声和生物变异等许多现象中很有用

不同均值m和方差σ^2的高斯随机变量

概率密度函数



累计概率分布函数



$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-m)^2}{2\sigma^2}), -\infty \le x \le \infty$$

随机变量的数字特征

- 概率分布能完全确定随机变量的特性
- 可以根据概率分布计算获得随机变量的某种特征
- 数字特征——刻画随机变量某一方面特征的常数
  - 数学期望（均值）、方差
  - 矩、原点矩、中心矩
  - 随机变量间的协方差、相关系数、协方差矩阵

随机变量的矩

- 矩（moment），可用来刻画随机变量的性质，随机变量的一种数字特征

- 定义
  - 随机变量x的函数g(x)的广义矩定义为其数学期望算子

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

    - $g(x)$ 常常是多项式形式，例如x, x^2, $(x-m)^2$
    - f(x) 是概率密度函数

常见矩

- 均值（一阶矩）
  - 在广义矩的定义中

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

  - 定义 $g(x) = x$，是一次方，因此是一阶矩

$$E[x] = m_1 = m = \int_{-\infty}^{\infty} xf(x)dx$$

- 均方值（二阶矩）
  - 定义 $g(x) = x^2$，则为二阶矩

$$E[x^2] = m_2 = \int_{-\infty}^{\infty} x^2 f(x)dx$$

其它常见矩

- 中心矩（central moments） – 相对于均值的矩
- 方差（二阶中心矩）

$$E[(x-m)^2] = \sigma^2 = \int_{-\infty}^{\infty} (x-m)^2 f(x)dx$$

- 偏度（三阶中心矩）

$$\mu_3 = \int_{-\infty}^{\infty} (x-m)^3 f(x)dx$$

  - 衡量概率密度分布函数（PDF）的对称性
  - 左边拖尾的PDF存在负偏度，右边拖尾的PDF存在正偏度（positive skewness）

概率论 → 统计学

概率论	统计学
随机变量	随机样本（→推断总体，对应随机变量）
分布函数（分布律）、概率密度函数	抽样分布—统计量的概率分布
数字特征	基于统计量的推断（参数估计和假设检验）
前提：**已知**随机变量的分布，在此前提下研究**随机变量**的性质、特点和规律	条件：假设随机变量分布，但是**具体参数未知**，需要根据**随机样本**对其（即总体性质）进行**推测**

随机抽样

- 随机样本
  - 对总体（随机变量）X进行n次放回抽样$X_1, X_2, \ldots, X_n$，构成样本容量为n的一个简单随机样本
  - 这n次抽样可以认为相互独立，而且与X同分布
- 统计量 （样本矩） – 样本的函数（函数中不含有未知数）
  - 均值、方差、标准差、原点矩、中心矩（理论上/定义的公式）
  - 统计量的观察值（将观察值$x_1, x_2, \ldots, x_n$ 代入对应的公式）
- 抽样分布 – 统计量的概率分布
  - 统计量是随机样本的函数，因此仍然是一个随机变量
  - 当总体（随机变量）的分布形式（假设）已知，则抽样分布可以确定。因为统计量是随机样本的函数，因此抽样分布是随机变量的函数的分布 —— 概率论研究过的内容（本课程仅要求会应用）

统计推断1——参数估计

- 假设已知总体X的分布函数的形式，例如高斯分布$N(\mu, \sigma^2)$，但是函数中的参数μ和 σ未知。参数估计的任务即是找到根据观察到的随机样本计算这些未知参数

- 点估计
  - 计算这些未知参数的值

- 区间估计
  - 计算这些未知参数所处的区间

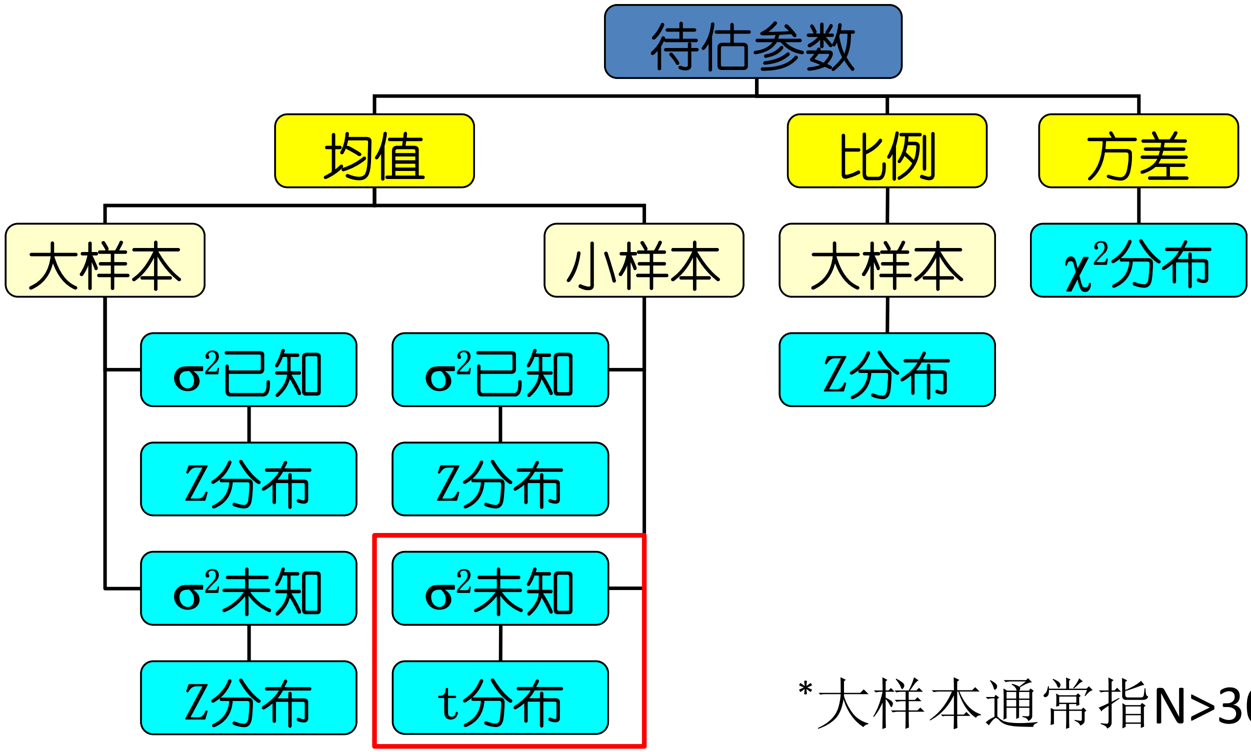
参数点估计

- 对于随机样本$\{x_j, j=1, 2, \ldots, N\}$，常见参数点估计的任务，实质是获得参数的**近似值**
  - 均值 $$E[x] \approx \hat{m} = \frac{1}{N}\sum_{j=1}^{N} x_j$$
  - 方差 $$\hat{\sigma}^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - m)^2 \qquad \hat{\sigma}^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \hat{m})^2$$
  - 偏度 $$\hat{\mu}_3 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \hat{m})^3$$

区间估计

- 区间估计是点估计（近似值）的补充
- 目标：根据抽样分布确定统计量取值的精确范围
- 我们仅考虑高斯信号处理，掌握正态分布总体的均值和方差的区间估计
- 常用抽样分布
  – 均值的抽样分布Z, $t(n-1)$
  – 方差的抽样分布$\chi^2(n)$

常见区间估计的条件和抽样分布

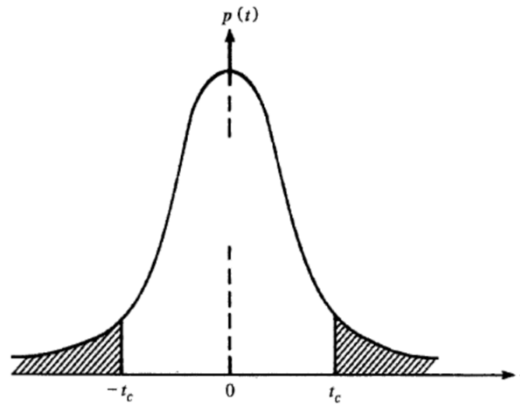


*大样本通常指N>30

均值估计的置信区间

- 小样本方差未知情况下，均值估计量符合$t(n\text{-}1)$分布
- 均值估计的置信区间 $[m_l, m_u]$
  - 满足 的t_c是临界值
    - 1-α 为置信 $P[|t| \ge t_c] = 0.05 = \alpha$
  - 非阴影区域为置信区
  - 对于 α 的置信下界和置信上界

$$\begin{cases} m_l = \hat{m} - t_c \dfrac{\hat{\sigma}}{\sqrt{N}} \\ m_u = \hat{m} + t_c \dfrac{\hat{\sigma}}{\sqrt{N}} \end{cases}$$



例：流量均值估计

样本数N = 30，临界水平 α = 0.05，估计的样本均值和方差计算结果为

$$\hat{m} = \frac{1}{30}\sum_{i=1}^{30} x_i = 1913$$

$$\hat{\sigma} = \sqrt{\frac{1}{29}\sum_{i=1}^{30}(x_i - \hat{m})^2} = 700.8$$

解：查阅t-分布表，自由度 $v = N\text{-}1 = 29$, 对应α = 0.05 的$t_c = 2.045$

$$m_l = \hat{m} - t_c\frac{\hat{\sigma}}{\sqrt{N}} = 1913 - 2.045\cdot\frac{700.8}{\sqrt{30}} = 1651.3$$

$$m_u = \hat{m} + t_c\frac{\hat{\sigma}}{\sqrt{N}} = 1913 + 2.045\cdot\frac{700.8}{\sqrt{30}} = 2174.7$$

$$P[1651.3 \le m \le 2174.7] = 0.95$$

根据30个测量值，推断在95%概率下，真实均值在 [1651.3 2174.7] 范围内，该范围即置信水平为0.95的置信区间。

方差的区间估计

样本方差S^2标准化误差的抽样分布为卡方分布

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

给定置信水平1-α

$$P\left\{\chi^2_{1-\frac{\alpha}{2}}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}(n-1)\right\} = 1-\alpha$$

$$即P\left\{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right\} = 1-\alpha$$

因此,方差σ^2的置信度为$1-\alpha$的置信区间为

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right)$$

标准差的区间估计

已知某产品重量的分布服从正态分布，随机其中25个样品，样品的方差s^2 =93.21，计算95%的置信水平下该产品重量标准差的置信区间

解：临界水平α=1-95%=0.05，样本数n=25，方差的置信区间公式

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$$

$$\chi^2_{\frac{\alpha}{2}}(n-1) = \chi^2_{0.025}(25-1) = 39.3641$$

$$\chi^2_{1-\frac{\alpha}{2}}(n-1) = \chi^2_{0.975}(25-1) = 12.4011$$

上述结果代入方差的置信区间公式

$$\frac{(25-1)\times 93.21}{39.3641} \le \sigma^2 \le \frac{(25-1)\times 93.21}{12.4011} \quad \rightarrow \quad 56.83 \le \sigma^2 \le 180.39$$

标准差的置信区间是方差置信区间的平方根

在95%的置信水平下，产品标准差的置信区间为[7.54 13.43]

区间估计的要点

- 统计量的分布的对应关系
  - 与参数（均值、比例、方差）有关
  - 与样本大小（自由度）有关
  - 与方差已知与否有关
- 查分布表对应的临界值
  - 与临界水平/置信水平有关
  - 与样本大小有关
- 代入区间估计公式计算置信区间的边界

模型与仿真

成年健康人的收缩压范围为90-140 mmHg，舒张压为60-90 mmHg。假设大二学生人群（总体）的收缩压服从正态分布，均值为116.3 mmHg，标准差为8.2 mmHg。请完成下列四项任务。

（1）编写MATLAB程序，仿真从大二学生人群中随机抽样52人的一次收缩压测试数据；

（2）用这组数据估计总体的收缩压均值和方差；

（3）用这组数据估计总体的收缩压均值和方差在95%置信水平下的置信区间；

（4）分析仿真数据与题干数值的关系。

参考解答要点

- Knowns
  - Matlab randn() ~ N(0, 1)
  - meanSysPres = 116.3;
  - stdSysPres = 8.2;
  - sampSize = 52;
- Simulation
  - sNormSamp = randn(1, sampSize); % standard normal distribution sample
  - % transform to the non-standard normal distribution
  - % since sNormSamp = (x-meanSysPres)/stdSysPres
  - % x = sNormSamp*stdSysPres + meanSysPres;
  - sampData = sNormSamp*stdSysPres + meanSysPres;
- Visualization
  - figure, hist(sampData, 10);
  - ylabel('counts'); xlabel('systolic pressure');
  - title('simulation of the systolic pressure sample data');
- Estimation
  - % Point estimation: mean and variance or standard deviation
  - % confidence interval estimation: mean and variance or standard deviation

参考答案要点

- %% point estimation
  - meanSysPres_Est = sum(sampData)/sampSize;
  - stdSysPres_Est = sqrt(sum((sampData-meanSysPres_Est).^2)/(sampSize-1));

- %% confidence interval estimation
  - % As sampSize = 52 > 30, use Z-distribution for the CI of mean value, and kai-square for variance
  - % Given the confidence level P = 95%, Zc at alpha/2 = (1-95%)/2 = 0.025 is
  - P = 0.95; halfAlpha = (1-P)/2;
  - Zc_Upper = norminv((1-halfAlpha), 0, 1); Zc_Lower = norminv(halfAlpha, 0, 1);
  - meanSysPres_Upper = meanSysPres_Est + Zc_Upper*stdSysPres_Est/sqrt(sampSize);
  - meanSysPres_Lower = meanSysPres_Est - Zc_Upper*stdSysPres_Est/sqrt(sampSize);
  - % CI of standard deviation (sqrt of variance)
  - chi_Lower = chi2inv(1-halfAlpha, sampSize-1); chi_Upper = chi2inv(halfAlpha, sampSize-1);
  - var_Est = stdSysPres_Est^2;
  - var_Upper = (sampSize-1)*var_Est/chi_Upper;
  - var_Lower = (sampSize-1)*var_Est/chi_Lower;
  - stdSysPres_Upper = sqrt(var_Upper);
  - stdSysPres_Lower = sqrt(var_Lower);

参考答案要点

- 比较仿真结果与假设的关系
  - 假设条件：收缩压服从正态分布，均值为116.3 mmHg，标准差为8.2 mmHg
  - 基于仿真样本的点估计结果与假设的值一致吗？为什么？
  - 基于仿真样本的区间估计包含假设值吗？为什么？
    - 提示：仅95%的置信区间
    - 如果将置信水平调整到99%，如何？
    - 如果增加仿真的样本量：52 → 520 → 5200
  - 提高置信水平，同样的仿真样本估计结果更容易包含假设给定的参数值
  - 提高仿真的样本量，更容易逼近假设给定的参数，估计的区间更容易包含假设给定的参数值

小结

- 我们测量的生物医学信号是对无时不刻发生的生命过程/现象的有限采样
  - 时间、空间、种类（传感器决定）有限
  - 随机性
- 数学基础
  - 概率论（关于随机变量的基础理论）
  - 数理统计（关于随机样本的统计量的研究）
    - 参数估计
  - 随机过程（下一讲）