What to do today (Oct 26, 2020)?

1. *Introduction*

2. *Probability and Distribution (Chp 1-3)*

3. **Essential Topics in Mathematical Statistics**

3.1 Elementary Statistical Inferences (Chp 4)

- *3.1.1 Sampling and Statistics*
- *3.1.2 Confidence Interval*
- *3.1.3 Order Statistics*
- **3.1.4 Hypothesis Testing**
- **3.1.5 Statistical Simulation and Bootstrap**

3.2 Consistency and Limiting Distributions (Chp 5)

3.3 Maximum Likelihood Methods (Chp 6)

3.1.4 Hypothesis Testing: Basic Setup

- **Population.** Suppose r.v. $X \sim F(\cdot;\theta)$, where θ is an unknown *parameter*.

- **Data (observations).** Suppose $X_1, \cdots, X_n$ are iid and arise from $F(\cdot;\theta)$.

- **Goal.** To test on $H_0 : \theta \in \Omega_0$ vs $H_1 : \theta \in \Omega_1$ using the random sample. $(\Omega_0 \bigcap \Omega_1 = \emptyset)$

How to achieve the goal?

$\Longrightarrow$ testing procedures for making an inference on **the null hypothesis** H_0 vs **the alternative hypothesis** H_1?

3.1.4 Hypothesis Testing: Basic Concepts

Let $\mathcal{D} = \{$*all possible realizations of* $(X_1, \ldots, X_n)\}$.

Definition. Set $\mathcal{C} \subseteq \mathcal{D}$ is called the **rejection region** for a hypothesis test if the test's decision rule is as follows:

Reject H_0 (Accept H_1) if $(x_1, \ldots, x_n) \in \mathcal{C}$;

Acdept H_0 (Reject H_1) if $(x_1, \ldots, x_n) \notin \mathcal{C}$.

- The **Type I error** of a test occurs if H_0 is rejected when H_0 is true; the **Type II error** of a test occurs if H_0 is accepted when H_1 is true.
- We say the rejection region $\mathcal{C}$ is of **size** (or significance level) α if $\alpha = \max_{\theta \in \Omega_0} P_\theta\{(X_1, \ldots, X_n) \in \mathcal{C}\}$.
  The **power function** of the test is
  $power(\theta) = P_\theta\{(X_1, \ldots, X_n) \in \mathcal{C}\}$ for $\theta \in \Omega_1$.

3.1.4 Hypothesis Testing: Examples

Example 3.3 To test for a binomial proportion of success at size α: $X \sim B(1, \theta)$ with $H_0 : \theta = \theta_0$ vs $H_1 : \theta < \theta_0$, provided a random sample $X_1, \ldots, X_n$

Approach 1. The decision rule should be "Reject H_0 in favor of H_1 if $\sum_{i=1}^n X_i \leq k$" with k determined by $\alpha = P_{H_0}(\sum_{i=1}^n X_i \leq k)$.

Approach 2. Since $\theta = E(X)$, $\bar{X}$ is a "good estimator" for it and with approximate distn $N(\theta, \frac{\theta(1-\theta)}{n})$.

The decision rule should be "Reject H_0 in favor of H_1 if $T < c$" with c determined by $\alpha = P_{H_0}(T < c)$:

$$T = (\bar{X} - \theta_0) \Big/ \sqrt{\theta_0(1 - \theta_0)/n}$$

Example 3.4 To test at $\alpha = .05$ on whether a six-face die is even by rolling it 60 times indptly with the outcomes

face	1	2	3	4	5	6
count	13	19	11	8	5	4

Formulation:

- **Population.** r.v. X = the number from a cast of the die: X is discrete with pmf $p(x)$ for $x = 1, \ldots, 6$.
- **Data (observations).** Suppose $X_1, \cdots, X_{60}$ are iid, from $p(\cdot)$, and with realizations given in the table above.
- **Goal.** To test on $H_0 : p(x) = 1/6$ for $x = 1, \ldots, 6$ vs $H_1 : \textit{otherwise}$.

Example 3.5 A measure of suspended particles in $\mu g/m^3$ is used by the World Health Organization air quality monitoring project. Let X and Y be the measure in the city center of Melbourne and Houston, respectively. Suppose X and Y are indpt. Test $H_0 : \mu_X = \mu_Y$ vs $\mu_x < \mu_Y$ at $\alpha = .05$ with $n = 13$ observations from Melbourne and $m = 16$ observations from Houston: $\bar{x} = 72.9$ and $\bar{y} = 81.7$, assuming $\sigma_X = 25.6$ and $\sigma_Y = 28.3$.

3.1.4 Hypothesis Testing: Comments

- In practice, $\alpha = 0.05$ is often used to "protect" H_0, and 80% is a commonly used standard for a satisfactory power.
- Consider a hypothesis test with the test statistic T.

  Instead of to construct a rejection region to "make a decision", a **significance test** includes the following:

  - calculate the p-value as

    $p = P_{H_0}(T \text{ the same as } T_{obs} \text{ or leaning toward } H_1 morethanit);$

  - conclude based on the p-value: if p is smaller than a predetermined significance level α, there's strong evidence against H_0; otherwise, there's no strong evidence against H_0 from the data.

3.1.4 Hypothesis Testing: Comments

- There is a *duality* between CI of a population parameter θ and the hypothesis testing on $H_0 : \theta = \theta_0$.
  - Given that $\widehat{\theta}_L(X_1, \ldots, X_n)$ and $\widehat{\theta}_U(X_1, \ldots, X_n)$ are the lower and upper limits of a 95% CI of θ, consider the rejection region

    $$\mathcal{C} = \left\{(x_1, \ldots, x_n) : \theta_0 \notin (\widehat{\theta}_L, \widehat{\theta}_U)\right\},$$

    which gives a test of size .05.
  - If the rejection region $\mathcal{C}$ with size of .05 can be presented as $\mathcal{C} = \left\{(x_1, \ldots, x_n) : \theta_0 \notin (\widehat{\theta}_L, \widehat{\theta}_U)\right\}$, the following interval is then a 95% CI for θ:

    $$(\widehat{\theta}_L(X_1, \ldots, X_n), \widehat{\theta}_U(X_1, \ldots, X_n))$$

3.1 Elementary Statistical Inferences (Chp 4)

What do we care about a random variable X?

Its **distribution**: its pattern of taking different values, that is, what values X takes and how often it takes a particular value.

How do we find out X's distribution from its observations (**data**: $x_1, \ldots, x_n$)?

- (i) by descriptive analysis: plotting/tabulating the data; summarizing the data with statistics
- (ii) by **making inference** on X's disnt $F(\cdot)$

  (iia) to approximate (estimate) $F(\cdot)$ by point/interval estimation;
  (iib) to choose between (test on) two contradictory claims about $F(\cdot)$ by hypothesis testing

How to verify a conclusion? How to assess performance of an inference procedure?

3.1.5 Statistical Simulation and Bootstrap: Monte Carlo Methods

- **Monte Carlo** refers to an area of Monaco, where the *Monte Carlo Casino* is located.
- **Monte Carlo methods** (or Monte Carlo experiments) are a class of computational algorithms that obtain numerical results by repeated random sampling.
- Monte Carlo methods are especially useful for simulating phenomena with significant uncertainty in inputs and random systems.

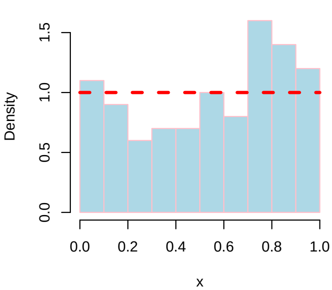
How does an Monte Carlo method work?

- How to simulate a particular system?
- After quantifying the system by a rv, how to simulate the rv?

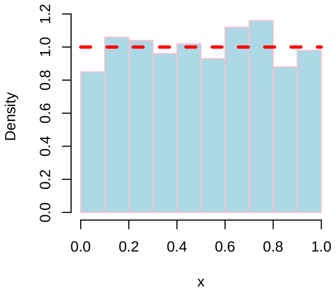
3.1.5 Statistical Simulation and Bootstrap: Monte Carlo Generation

Uniform generator. eg the one in the software package *R*: "runif(n,min,max)"

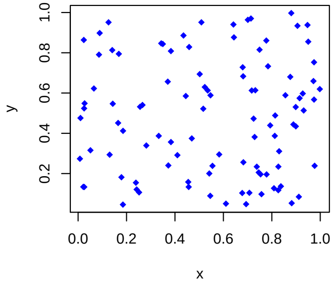
```
x=runif(100);
hist(x, freq=FALSE, ... ...)
curve(dunif(x), col = 2, lty = 2, lwd = 3, add = TRUE)
y=runif(100);
plot(x,y, xlab=x, ylab=y, pch=18, col=4, sub=(a2). n=100)
x=runif(1000);
hist(x, freq=FALSE, ... ...)
curve(dunif(x), col = 2, lty = 2, lwd = 3, add = TRUE)
y=runif(1000);
plot(x,y, xlab=x, ylab=y, pch=18, col=4, sub=(b2). n=1000)
```



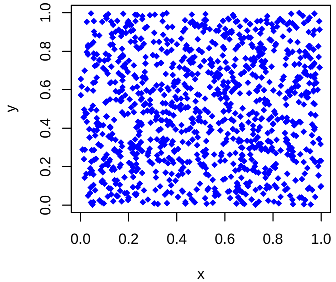
(a1). n=100



(b1). n=1000



(a2). n=100



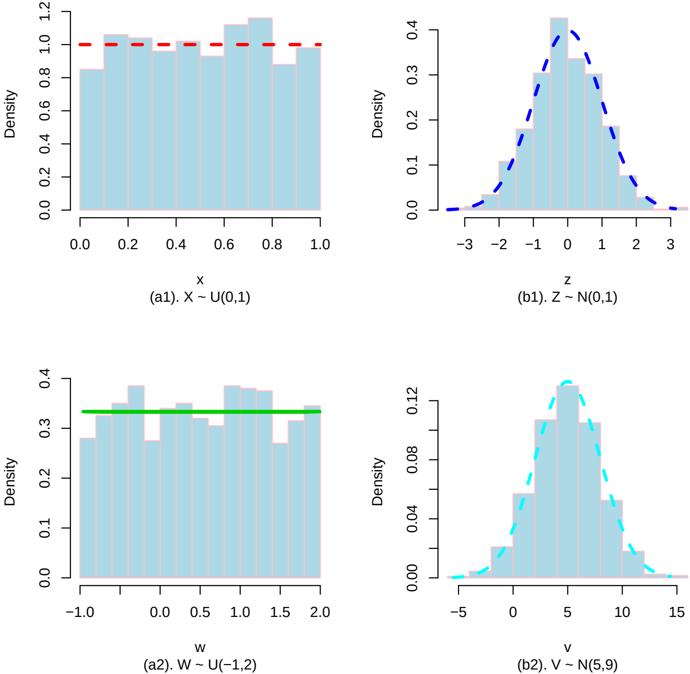
(b2). n=1000

3.1.5 Statistical Simulation and Bootstrap: Monte Carlo Generation

How to generate random variables?

- *R* has generators of most commonly used rvs: eg. "rnorm(n,mean,sd)"
- Use transformations of commonly used rvs: for example,

```
x=runif(1000);
w=3*x-1;
z=rnorm(1000);
v=3*z+5;
hist(x, freq=FALSE, ...)
curve(dunif(x), col = 2, lty = 2, lwd = 3, add = TRUE)
hist(w, freq=FALSE, ...)
lines(w,rep(1/3,1000), col = 3, lty = 2, lwd = 3)
hist(z, freq=FALSE, ...)
curve(exp(-x^2/2)/(2*pi)^.5, min(z),max(z),
                    col = 4, lty = 2, lwd = 3, add = TRUE)
hist(v, freq=FALSE, ...)
curve(exp(-(x-5)^2/(2*9))/(2*pi*9)^.5, min(v),max(v),
                     col = 5, lty = 2, lwd = 3, add = TRUE)
```

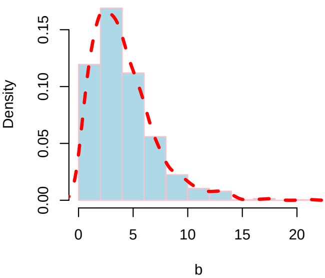

Density
x
(a1). X ~ U(0,1)
Density
z
(b1). Z ~ N(0,1)
Density
w
(a2). W ~ U(−1,2)
Density
v
(b2). V ~ N(5,9)

3.1.5 Statistical Simulation and Bootstrap: Monte Carlo Generation

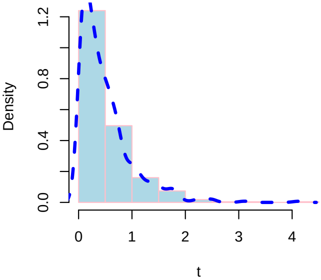
How to generate random variables?

- Use transformations of commonly used rvs: for example,
  - If given a cdf $F(\cdot)$, $Y = F^{-1}(U)$ with $U \sim U(0,1)$ has $Y \sim F(\cdot)$

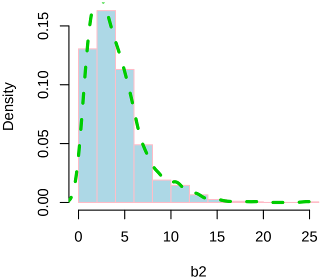
```
ztmp=matrix(rnorm(4000),ncol=4)
b=apply(ztmp^2,1,sum)
x=runif(1000);
t=-log(1-x)/2
hist(b, freq=FALSE, ...)
lines(density(b), col = 2, lty = 2, lwd = 3)
b2=rchisq(1000,df=4)
hist(b2, freq=FALSE, ...)
lines(density(b2), col = 3, lty = 2, lwd = 3)
hist(t, freq=FALSE, ...)
lines(density(t), col = 4, lty = 2, lwd = 3)
t2=rexp(1000,2)
hist(t2, freq=FALSE, ...)
lines(density(t2), col = 5, lty = 2, lwd = 3)
```



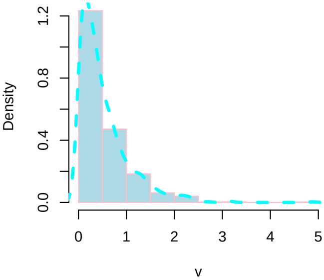
(a1). B ~ chisq(4)



(b1). T ~ NE(2)



(a2). B2 ~ chisq(4)



(b2). T2 ~ NE(2)

3.1.5 Statistical Simulation and Bootstrap: Monte Carlo Generation

How to generate random variables?

- **Accept-Reject Algorithm.** If $f(\cdot)$ is a pdf and $f(x) \leq Mg(x)$ with M a constant and $g(\cdot)$ the *instrumental* pdf.

  Step 1. Generate $Y \sim g(\cdot)$ and $U \sim U(0,1)$ indptly.

  Step 2. If $U \leq \frac{f(Y)}{[Mg(Y)]}$, take $X = Y$ and go to Step 3; otherwise, return to Step 1.

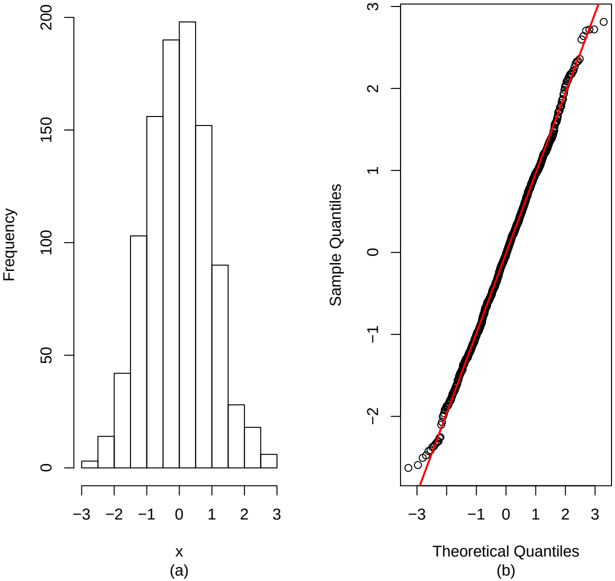
  Step 3. Obtain X, which follows $f(\cdot)$.

To prove it? (p298, the textbook by Hogg et al)

Example. Suppose $X \sim N(0,1)$ with pdf
$f(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, and $Y \sim \mathit{Cauchy}(0,1)$ with pdf
$g(y) = \pi^{-1}(1+x^2)^{-1}$.
Note that $f(x) \leq Mg(x)$ with $M = \frac{\pi}{\sqrt{2\pi}}(1.213) = 1.520$.

Use the Accept-Reject Algorithm to generate 1000 observations from N(0,1):

```
x<-rep(0,1000)
for(i in 1:1000){
        y<-rcauchy(1, location = 0, scale = 1);
        u<-runif(1, min=0, max=1)
        while(u>(exp(-y^2/2)/(2*pi)^.5/1.520*pi*(1+y^2))){
        y<-rcauchy(1, location = 0, scale = 1);
        u<-runif(1, min=0, max=1)
}
        x[i]<-y
}
```



(a) Histogram and (b) QQNorm Plot of the generated 1000 observations from $N(0, 1)$.

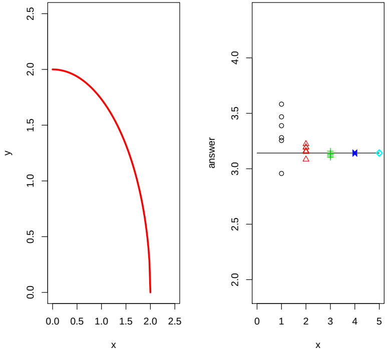
Monte Carlo Integration

How to calculate $\int_a^b g(x)dx$?

Example.

$$\int_0^2 \sqrt{4-x^2}dx = 2\int_0^2 \sqrt{4-x^2}\left(\frac{1}{2}\right)dx = 2E\{\sqrt{4-X^2}\} \quad (\pi)$$

provided that $X \sim U(0,2)$.



(a) y=(4−x^2)^.5

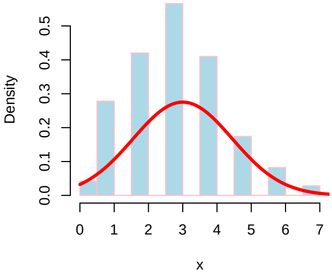
(b) true answer=pi

(b) Approximates to the integral by generating n observations from $U(0,2)$, with $n = 10^k$ for $k = 1, \ldots, 5$.

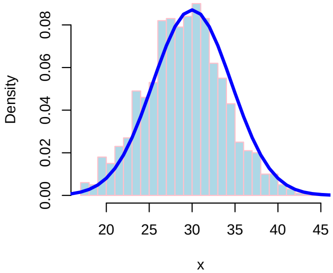
3.1.5 Statistical Simulation and Bootstrap: Simulation Example 1

To verify the normal approximation to binomial distn:

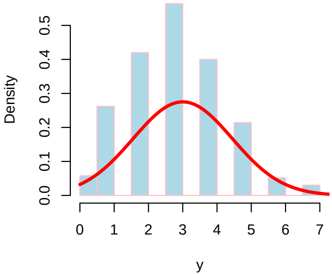
```
xtmp=matrix(ifelse(runif(1000*10)<.3,1,0),ncol=10)
x=apply(xtmp,1,sum)
hist(x, freq=FALSE, ...)
y=rbinom(n=1000,size=10,prob=0.3)
hist(y, freq=FALSE, breaks=11, ...)

xtmp=matrix(ifelse(runif(1000*100)<.3,1,0),ncol=100)
x=apply(xtmp,1,sum)
hist(x, freq=FALSE,breaks=20, ...)
curve(exp(-(x-30)^2/2/(30*.7))/(2*pi*21)^.5, 0, 100, lty=1, col=4,
        lwd=3, add=TRUE)
y=rbinom(n=1000,size=100,prob=0.3)
hist(y, freq=FALSE,breaks=20, ...)
curve(exp(-(x-30)^2/2/(30*.7))/(2*pi*21)^.5, 0, 100, lty=1, col=4,
        lwd=3, add=TRUE)
```



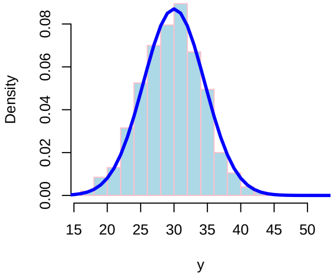
(a1). X ~ B(10,0.3)



(a2). X ~ B(100,0.3)



(b1). Y ~ B(10,0.3)

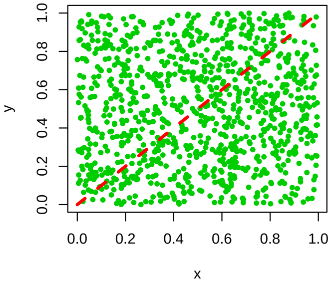


(b2). Y ~ B(100,0.3)

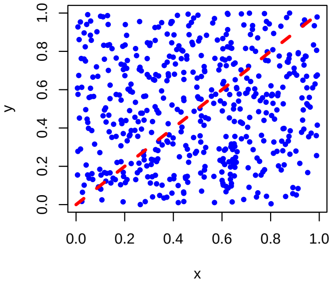
3.1.5 Statistical Simulation and Bootstrap: Simulation Example 2

What can data mssing result in?

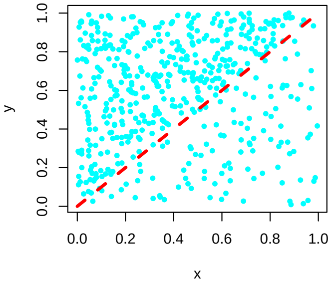
```
x=runif(1000);
y=runif(1000);
w=ifelse(x<y,1,0); sum(w)/1000

r0=ifelse(runif(1000)<.5,1,0)
x0=x[r0==1];y0=y[r0==1];
w0=ifelse(x0<y0,1,0);sum(w0)/sum(r0);

r1=rep(0,1000)
r1[x<y]=rbinom(length(x[x<y]),size=1,prob=.8)
r1[x>=y]=rbinom(length(x[x>=y]),size=1,prob=.2)
x1=x[r1==1];y1=y[r1==1];
w1=ifelse(x1<y1,1,0);sum(w1)/sum(r1);

r2=rep(0,1000)
r2[x<y]=rbinom(length(x[x<y]),size=1,prob=.3)
r2[x>=y]=rbinom(length(x[x>=y]),size=1,prob=.7)
x2=x[r2==1];y2=y[r2==1];
w2=ifelse(x2<y2,1,0);sum(w2)/sum(r2);
```



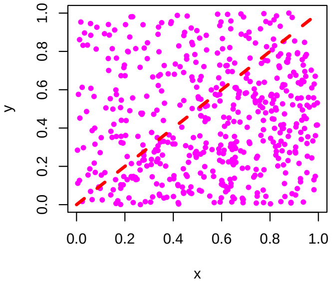
(a).all obs: rate of 'x<y'=.495



(b).random missing 50%: rate of 'x<y'=.486



(c).nonrandom missing 50%: rate of 'x<y'=.789



(d).nonrandom missing 50%: rate of 'x<y'=.337

What will we study next class?

1. Introduction

2. Probability and Distribution (Chp 1-3)

3. Important Topics in Mathematical Statistics (Chp 4-6)

- **3.1 Elementary Statistical Inferences**
    - *3.1.1 Sampling and Statistics*
    - *3.1.2 Confidence Interval*
    - *3.1.3 Order Statistics*
    - *3.1.4 Hypothesis Testing*
    - **3.1.5 Statistical Simulation and Bootstrap**
- **3.2 Consistency and Limiting Distributions**
- *3.3 Maximum Likelihood Methods*

4. Further Topics, Selected from Chp 7-11