Empirical Distribution function

Let $X_1, X_2, X_3 \ldots, X_n$ be random variables.

Random variable Y, has the "empirical distribution"

- Descriptive statistics
- infer things about the distribution from the Empirical distribution

$$\text{Range}(Y) = \{X_1, X_2, \ldots, X_n\},$$

$$P(Y = t) = \frac{1}{n} \mid \{i : X_i = t\} \mid,$$

and

$$\hat{F}_n(x) = P(Y \leq x) = \frac{|\{i : X_i \leq x\}|}{n}$$

Sample Mean

Let $X_1, X_2, X_3 \ldots, X_n$ be i.i.d. random variables. The "sample mean"

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

- Regular average of $X_1, \ldots, X_n$
- E[Y] - (Y - previous slide)

Suppose $E[X] = \mu$ and $Var[X] = \sigma^2$.

Then

$$E[\bar{X}] = \mu \quad \text{and} \quad SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}.$$

unbiased estimate

– Variance reduction

χ^2-distribution

Let $n \geq 1$, $X_1, X_2, \ldots, X_n$ be i.i.d. Normal(μ, σ^2). Consider the sample mean and variance

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Then

- $\overline{X}_n$ is a Normal random variable with mean μ and variance $\frac{\sigma^2}{n}$. - sketch given last class
- $\frac{(n-1)}{\sigma^2} s_n^2$ has the χ^2_{n-1} distribution. ~ Ex: n=2
- $\overline{X}_n$ and s_n^2 are independent.

Proof: - Omit □

Observation

Let $n \geq 1$, $X_1, X_2, \ldots, X_n$ be i.i.d. Normal$(0,\sigma^2)$. $\equiv$ { Reason for choosing data like this

Then $Z_i = \frac{X_i}{\sigma}, i = 1, 2, \ldots, n$ are i.i.d. Normal$(0,1)$.

$$U_X = \frac{\bar{X}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}} \quad \text{and} \quad U_Z = \frac{\bar{Z}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2}}$$

What can you say about U_X and U_Y ?

Observation

Let $n \geq 1$, $X_1, X_2, \ldots, X_n$ be i.i.d. Normal$(0,\sigma^2)$.

Then $Z_i = \frac{X_i}{\sigma}, i = 1, 2, \ldots, n$ are i.i.d. Normal$(0,1)$.

$$U_X = \frac{\bar{X}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}} \quad \text{and} \quad U_Z = \frac{\bar{Z}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2}}$$

$U_X = U_Z$ and U_Z has distribution independent of σ

t-distribution

Let X_1 be a Normal random variable with mean 0 and variance 1. Let X_2 be an independent χ^2_n random variable. Let

$$Z = \frac{X_1}{\sqrt{\frac{X_2}{n}}}.$$

Constructing the t_n distribution

Z is said to have t-distribution with n-degrees of freedom if its p.d.f. is given by

$$f_Z(z) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{z^2}{n}\right)^{-\frac{n+1}{2}} \xrightarrow[n\to\infty]{} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

(Ex.)

for $z \in \mathbb{R}$.

t-distribution

Corollary :-

Let $n \geq 1$, $X_1, X_2, \ldots, X_n$ be i.i.d. Normal(μ, σ^2). Consider the sample mean and variance

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Then

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S}$$

has the t_{n-1} distribution.

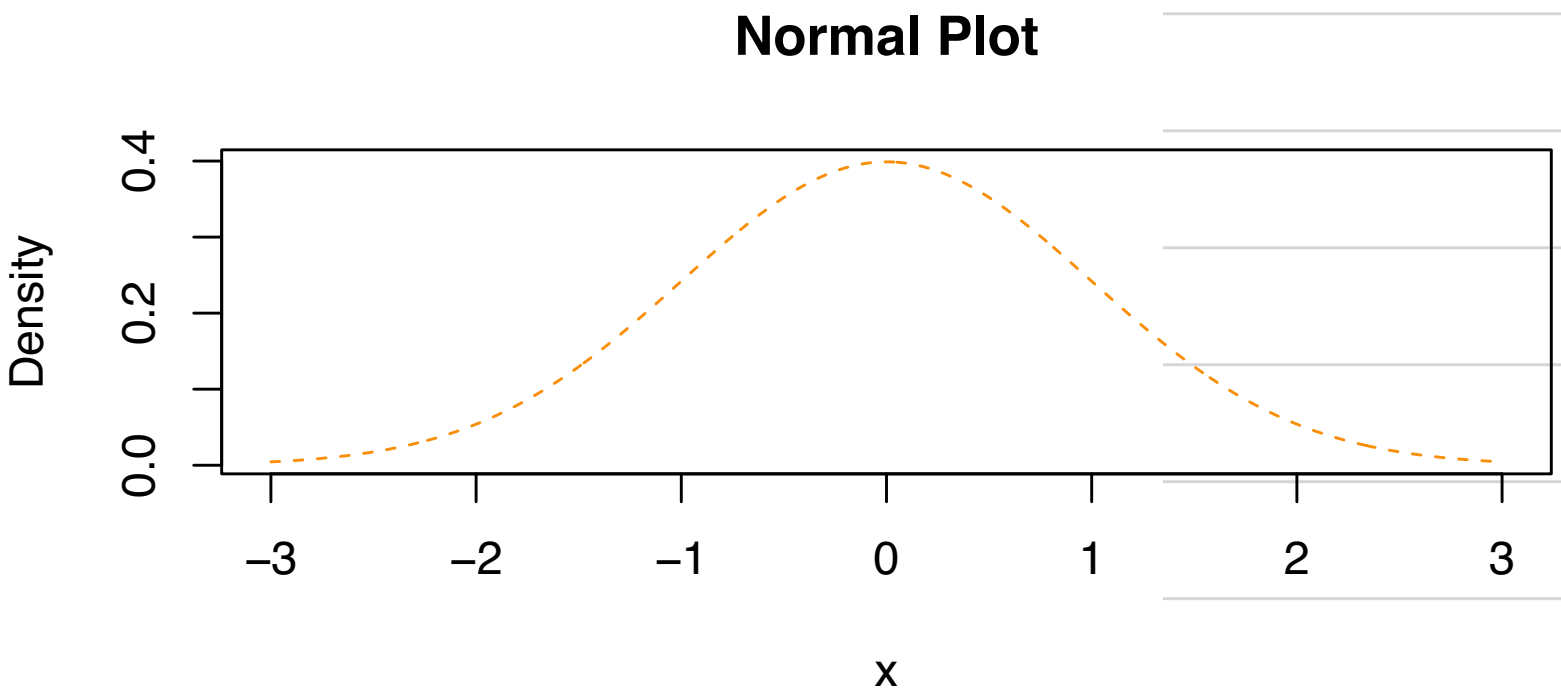
- $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$
- $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{d}{\equiv} N(0,1)$
- $\overline{x}$ and s^2 are independent

$$= \frac{\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{n-1}{\sigma^2}S^2}{n-1}}} \sim t_{n-1}$$

(Previous slide)

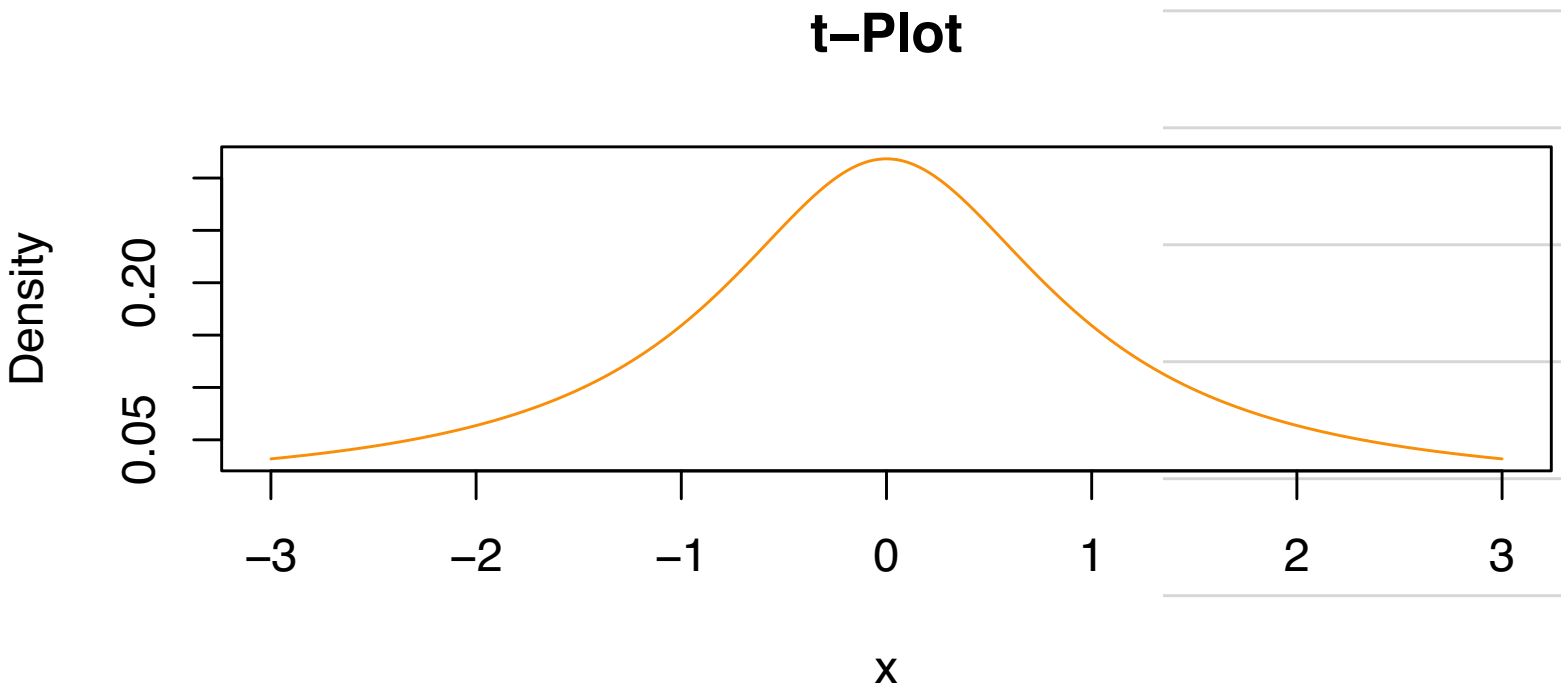
Normal

```
> x= seq(-3, 3, length=1000);densityx = dnorm(x)
> plot(x, densityx, type="l", lty=2, col="#f98e09",
+ ylab="Density",
+ main="Normal Plot")
```



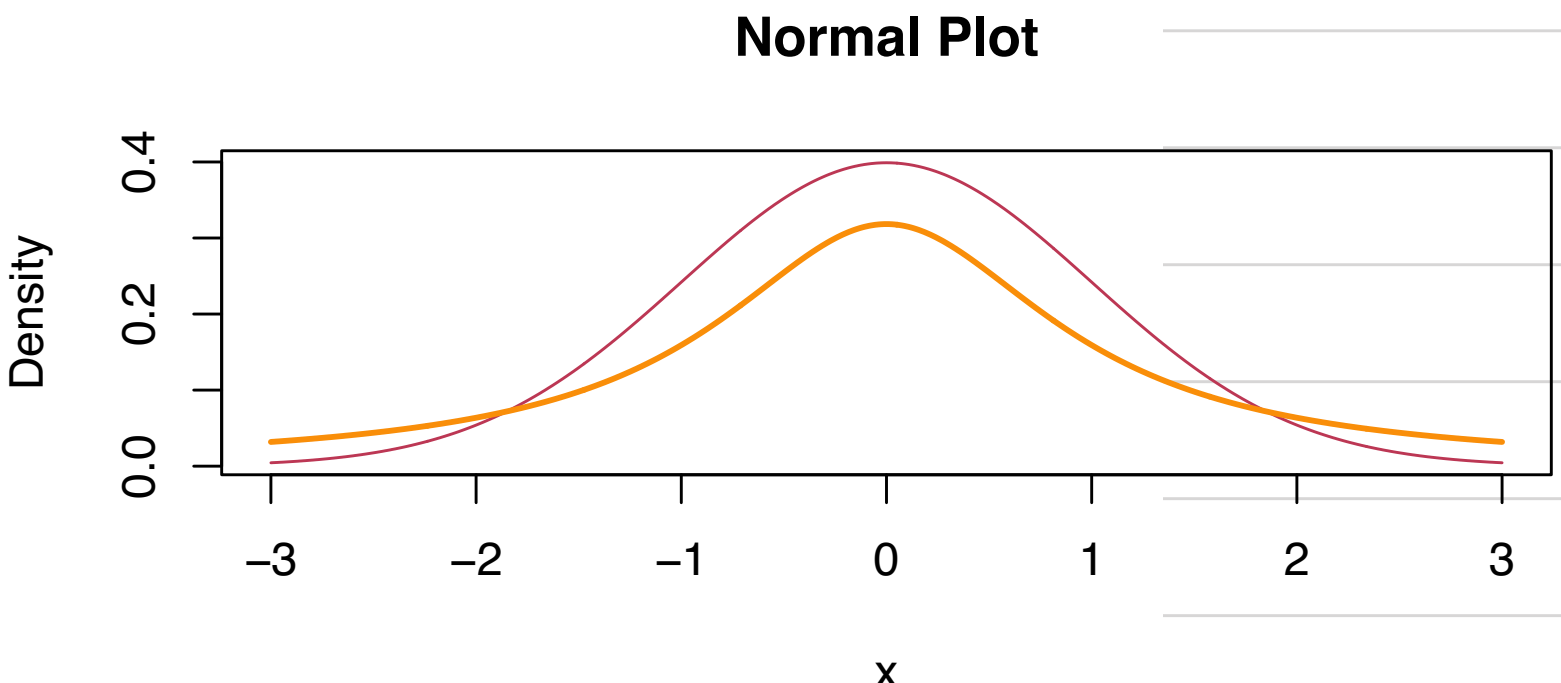
t_1

```
> x= seq(-3, 3, length=1000);dtx = dt(x,1)
> plot(x, dtx, type="l", col="#f98e09",
+ ylab="Density",
+ main="t-Plot")
```



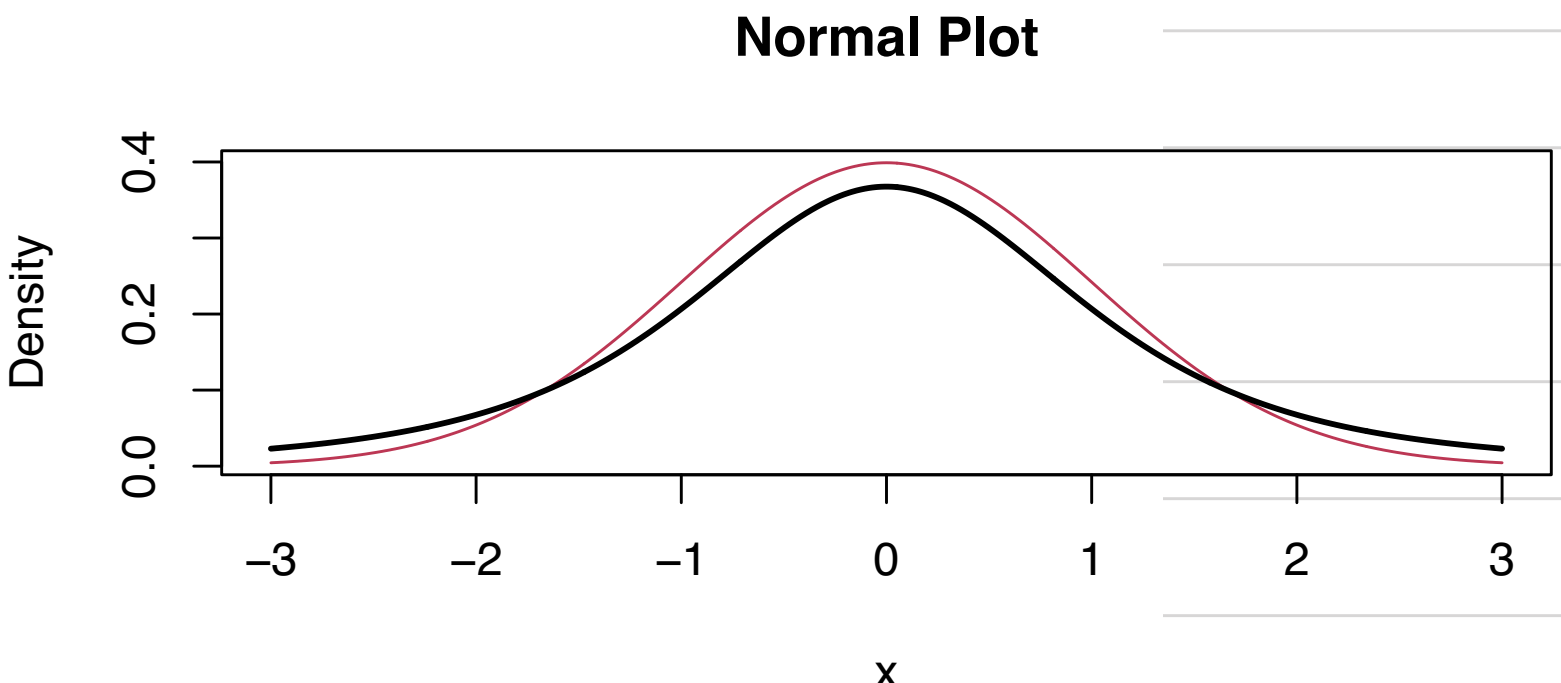
t_1-versus Normal

```
> x= seq(-3, 3, length=1000);densityx = dnorm(x)
> plot(x, densityx, type="l",  col="#bc3754",
+ ylab="Density",
+ main="Normal Plot")
> lines(x, dt(x,1), lwd=2, col="#f98e09")
```



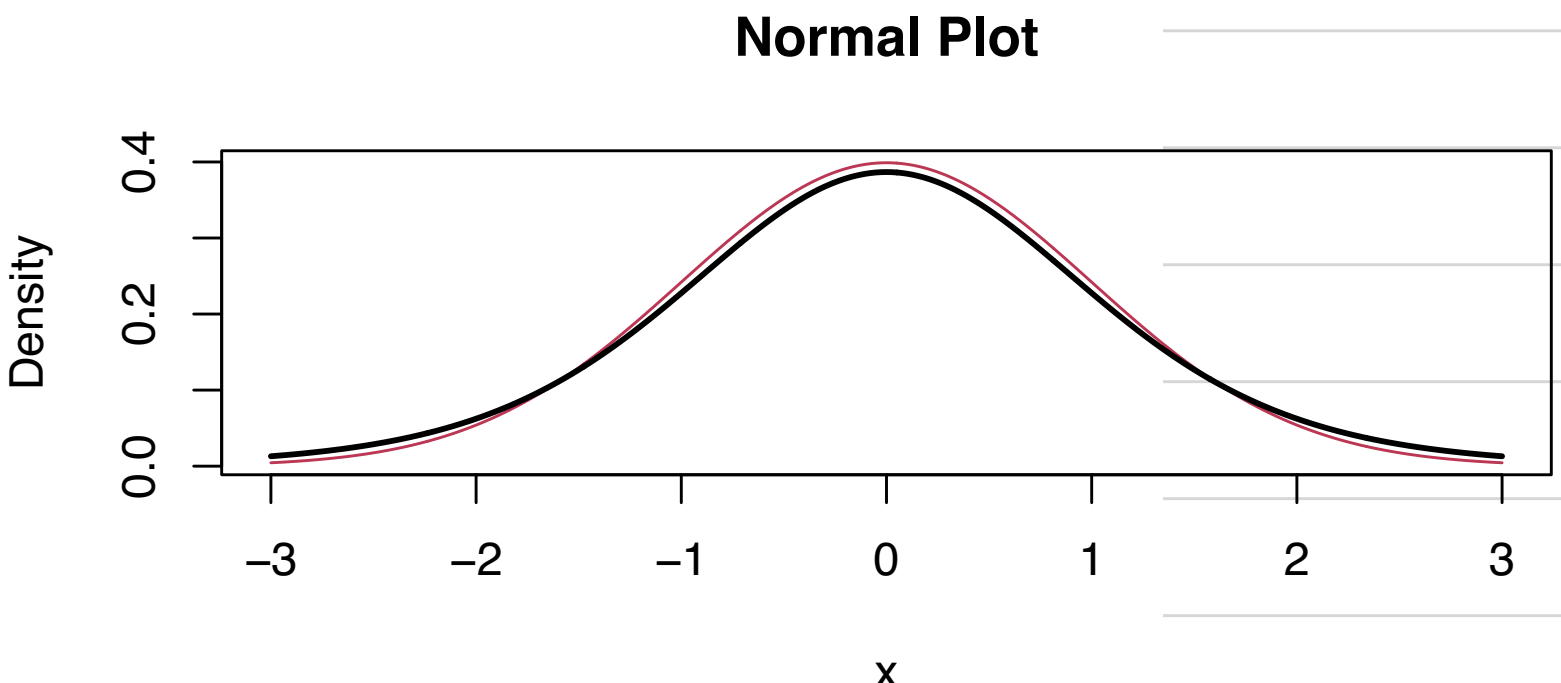
t_3-versus Normal

```
> x= seq(-3, 3, length=1000);densityx = dnorm(x)
> plot(x, densityx, type="l",  col="#bc3754",
+ ylab="Density",
+ main="Normal Plot")
> lines(x, dt(x,3), lwd=2)
```



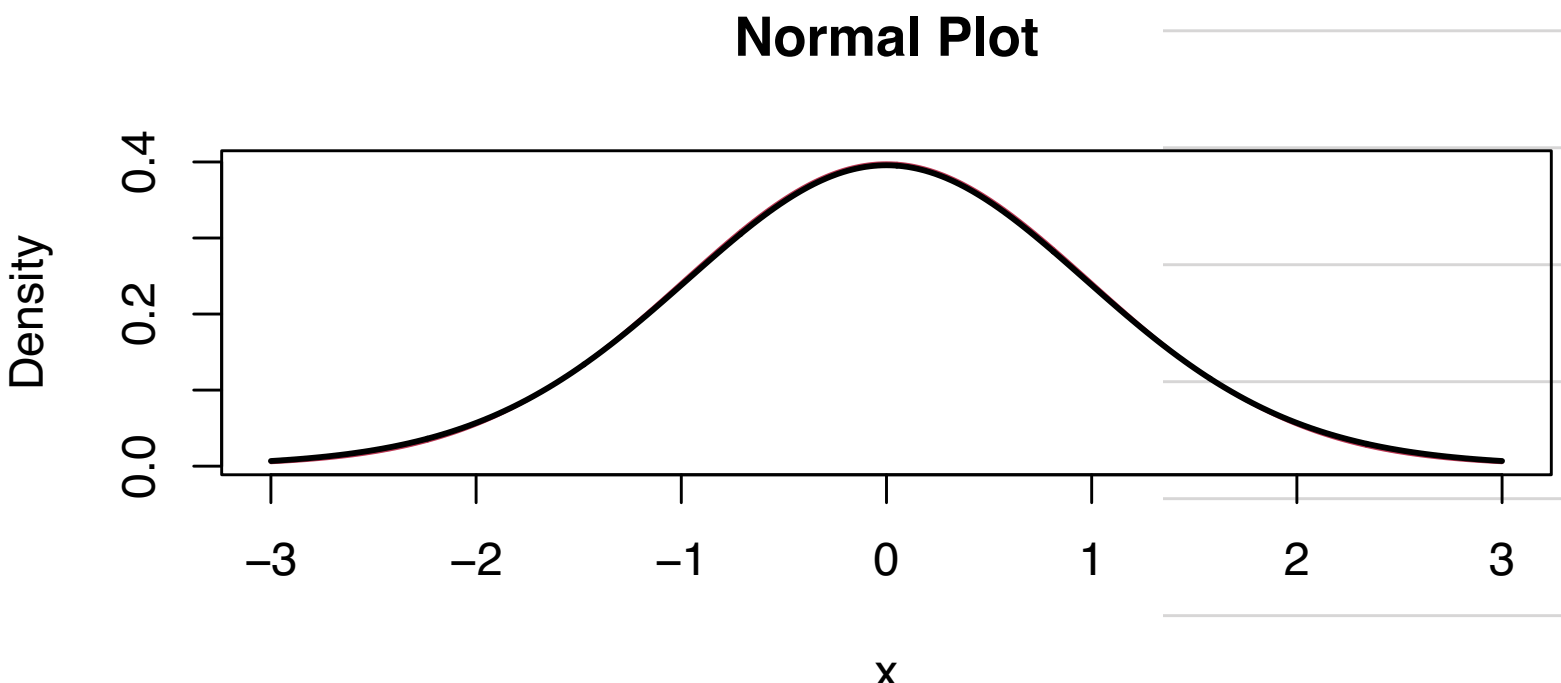
t_8-versus Normal

```
> x= seq(-3, 3, length=1000);densityx = dnorm(x)
> plot(x, densityx, type="l",  col="#bc3754",
+ ylab="Density",
+ main="Normal Plot")
> lines(x, dt(x,8), lwd=2)
```

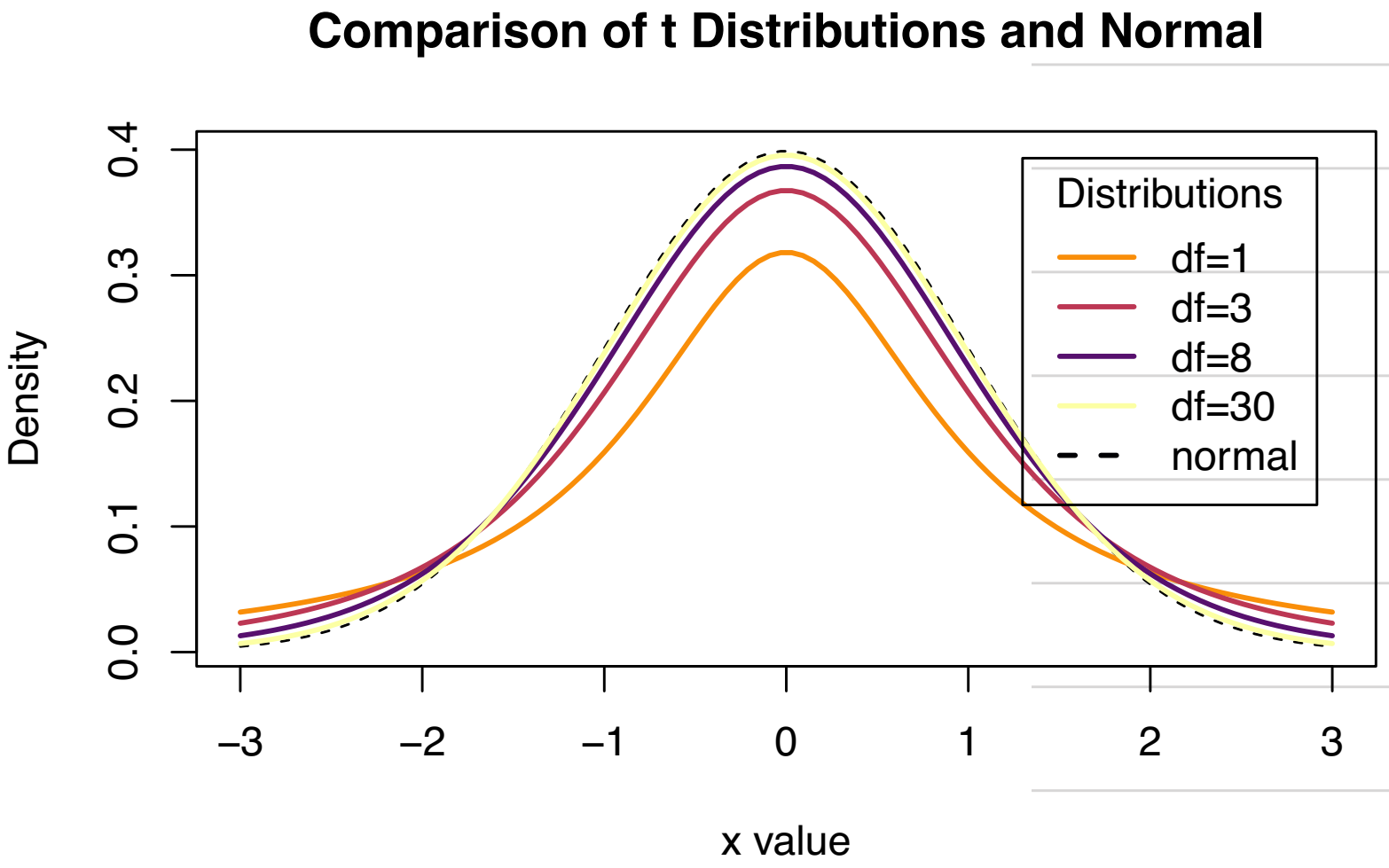


t_{30}-versus Normal

```
> x= seq(-3, 3, length=1000);densityx = dnorm(x)
> plot(x, densityx, type="l",  col="#bc3754",
+ ylab="Density",
+ main="Normal Plot")
> lines(x, dt(x,30), lwd=2)
```



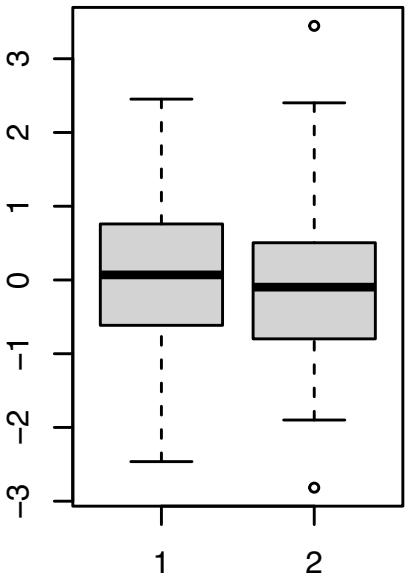
t-versus Normal

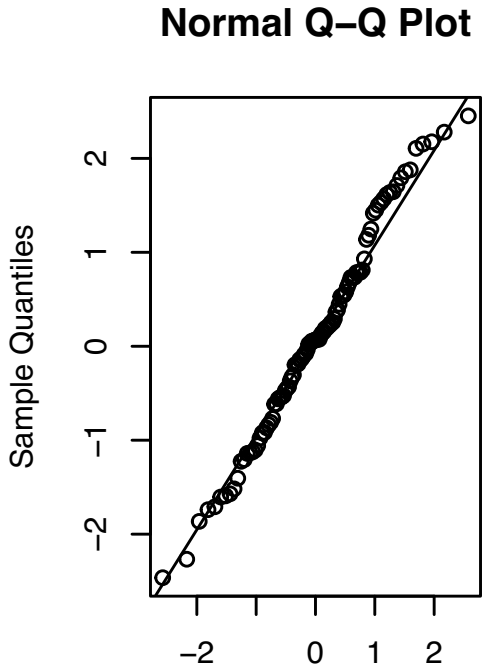


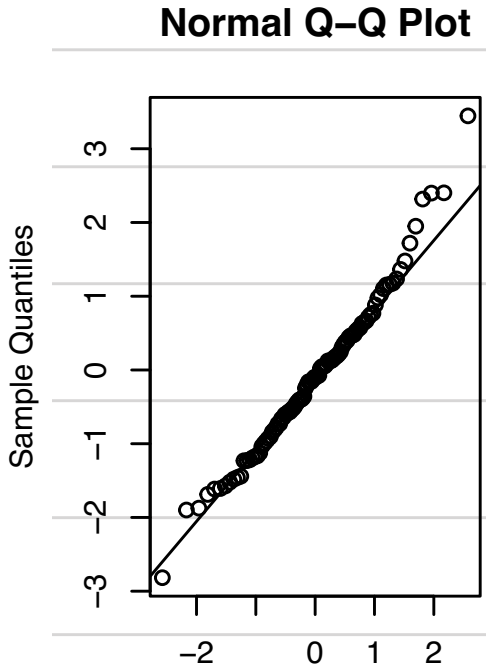
- $t_n \to N(0,1)$ as $n \to \infty$

t-versus Normal

```
> x = rnorm(100); y = rt(100,30)
> par(mfrow=c(1,3))
> boxplot(x,y)
> qqnorm(x);qqline(x)
> qqnorm(y);qqline(y)
```







Markov Inequality

- X - random variable
  - (Discrete) $X : \Omega \to T$, T - countable, $\mathbb{P}(X = t)$, $t \in T$
  - (Continuous) $X : \Omega \to \mathbb{R}$, $\mathbb{P}(X \in (a,b)) = \int_a^b f(x)\,dx$

- $\mu = \mathbb{E}[X]$ is finite.

- $X \geq 0$ $\equiv$ $\mathbb{P}(X = t) > 0 \iff t \geq 0$ $\equiv$ $f(x) > 0 \iff x \geq 0$

let $c > 0$ be given.

$$\mathbb{P}(X \geq c) \leq \frac{\mu}{c}$$

Upper Bound meaningless if $\mu > c$

Proof of Markov Inequality.

Assume X – continuous. (Ex: Prove when X – discrete)

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{0}^{\infty} x f(x)\,dx \quad (\because X \geq 0)$$

(1)

$$\int_{0}^{\infty} x f(x)\,dx \overset{(\mu < \infty)}{=} \underbrace{\int_{0}^{c} x f(x)\,dx}_{\geq 0} + \int_{c}^{\infty} x f(x)\,dx \quad (c > 0)$$

$x \geq c$

$$(h \geq l;\ \textstyle\int h(x)dx \geq \int l(x)dx)$$

$$\geq 0 + c \underbrace{\int_{c}^{\infty} f(x)\,dx}_{P(X \geq c)}$$

$$= c\,P(X \geq c)$$

(2)

Proof of Markov Inequality.

From ① and ②

$$\mu \overset{①}{=} \int_0^\infty x f(x)\,dx \overset{②}{\geq} c\,\mathbb{P}(X \geq c)$$

$$\Rightarrow \quad \frac{\mu}{c} \geq \mathbb{P}(X \geq c) \qquad c > 0$$

$\equiv$

- Markov Inequality is a powerful tool
- has many applications

Tschebychev Inequality

X r.v. $\quad E[X] = \mu < \infty \quad$ and $\quad Var[X] = \sigma^2 < \infty$

$k > 0$; $\quad P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$

Proof. let $k \geq 0$ Set $Y = \frac{|X-\mu|^2}{\sigma^2}$

. Note $Y \geq 0$ & $\{|X-\mu| \geq k\sigma\} = \{Y \geq k^2\}$

$$P(|X-\mu| \geq k\sigma) = P(Y \geq k^2)$$

Apply markov Inequality to Y

$$\leq \frac{E[Y]}{k^2}$$

$$= \frac{E|X-\mu|^2}{\sigma^2 k^2}$$

$$= \frac{\sigma^2}{\sigma^2 \cdot k^2}$$

$$= \frac{1}{k^2}$$

$X_1, \dots, X_n$ i.i.d $X \equiv \mu = E[X]$, $\sigma^2 = \mathrm{var}[X]$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$E[\bar{X}] = \mu$ (Consistency)

$\mathrm{var}[\bar{X}] = \frac{\sigma^2}{n}$

$X_i = \begin{cases} 1 & \text{if } A \text{ occurred} \\ 0 & \text{otherwise} \end{cases}$

$\cdot\ p = \mathbb{P}(A)$

$E[\bar{X}] = p$

$\mathrm{var}[\bar{X}] = \frac{p(1-p)}{n}$

Relative frequency — True Probability

let $\varepsilon > 0$ be given

$$\mathbb{P}(|\bar{X} - \mu| \geq \varepsilon)$$

$$= \mathbb{P}\left(\frac{|\bar{X}-\mu|}{\sigma} \geq \frac{\varepsilon}{\sigma}\right)$$

$$= \mathbb{P}\left(|\bar{X}-\mu| \geq \left(\frac{\varepsilon}{\sigma}\sqrt{n}\right)\frac{\sigma}{\sqrt{n}}\right)$$

$\left(\mu = E[\bar{X}],\ \sigma/\sqrt{n} = \mathrm{var}(\bar{X})\right.$ Tschebyscheff to $\bar{X}$ with $K = \frac{\varepsilon}{\sigma}\sqrt{n}$$\left.\right)$ $\leq \frac{1}{K^2} = \frac{\sigma^2}{\varepsilon^2}\cdot\frac{1}{n}$

$$0 \leq \mathbb{P}(|\bar{X}-\mu| > \varepsilon) \leq \left(\frac{\sigma^2}{\varepsilon^2}\right)\frac{1}{n}$$

as $n \to \infty$: $0 \to 0$; $\mathbb{P}(|\bar{X}-\mu| > \varepsilon) \to 0$ (Sandwich / Squeeze) ; $\frac{1}{n} \to 0$ as $n \to \infty$

Weak Law of Large Numbers

So we have shown:-

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Assume that X_1 has finite mean μ and finite variance σ^2. Then for any $\epsilon > 0$

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0, \tag{1}$$

i.e. Let $A_n = \{|\bar{X} - \mu| > \epsilon\}$

then $\lim_{n \to \infty} \mathbb{P}(A_n) = 0.$

Law of Large Numbers

```
> runningmean = function (x,N){
+ y = rpois(N,x)
+ # y =runif(N,2,3)
+
+ c = cumsum(y)
+ n = 1:N
+ c/n
+ }
> u = runningmean(1, 1000)
```

$X_1, X_2, \ldots X_n$ are i.i.d Poisson(1)

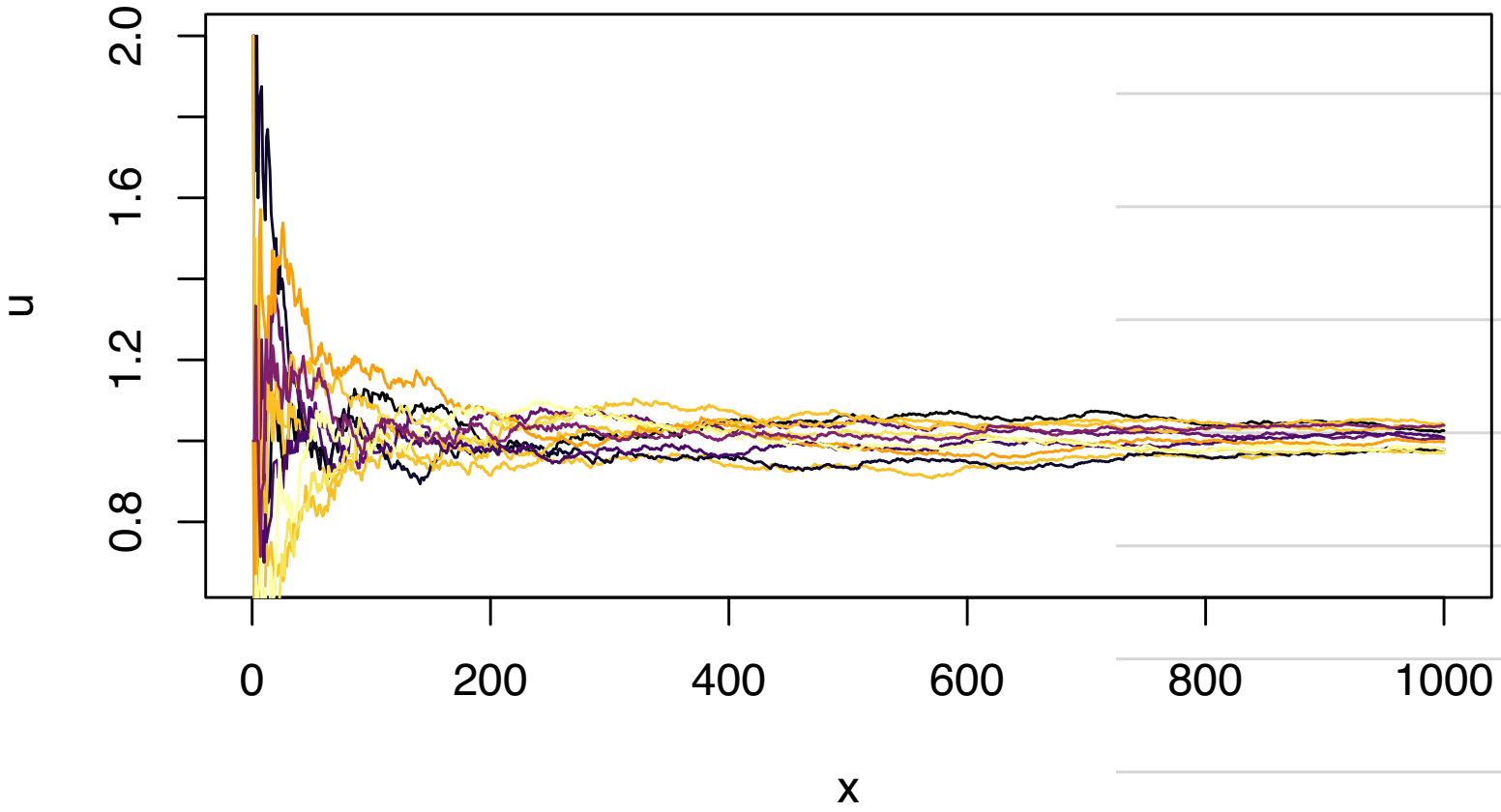
$E[X_1] = 1$

$Var[X_1] = 1$

$P(|\bar{X}_n - 1| > \varepsilon) \longrightarrow 0$ as $n \to \infty$

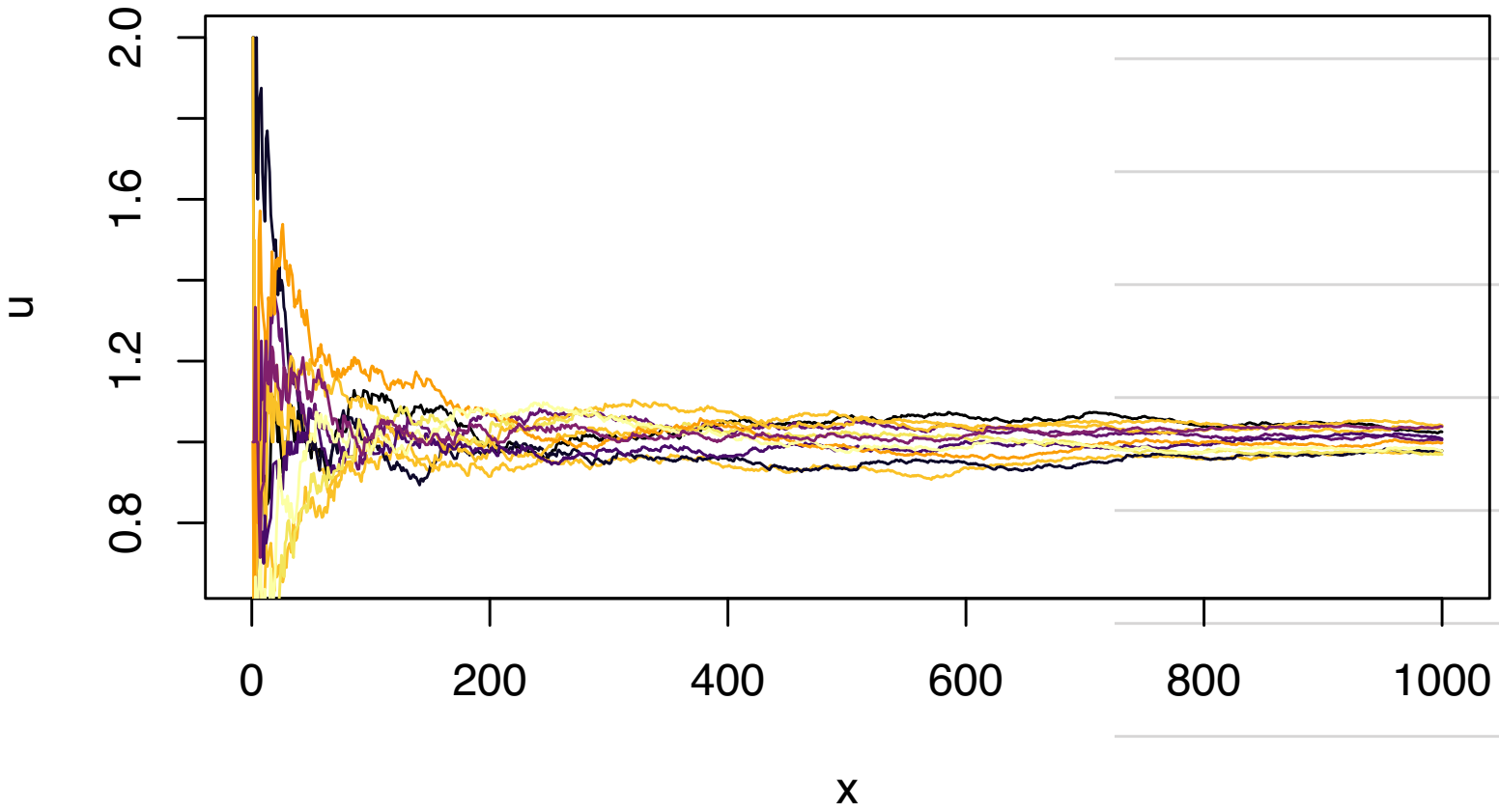
Law of Large Numbers

```
> x=1:1000; plot(u~x, type="l");
>
```

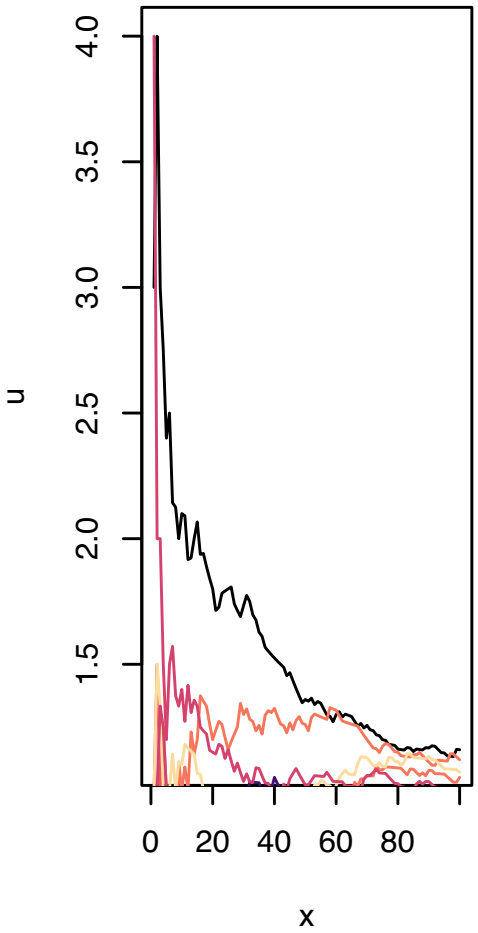


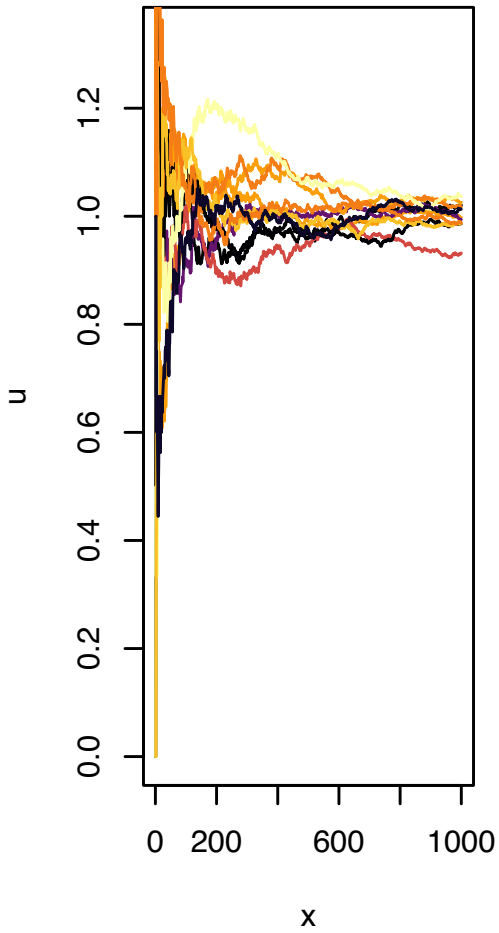
Law of Large Numbers

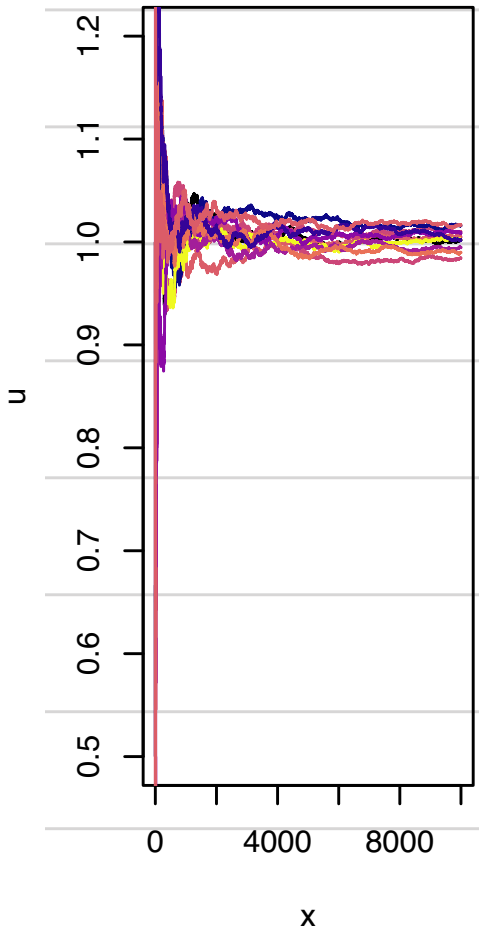
```
> library(viridis)
> x=1:1000; plot(u~x, type="l");
> replicate(10, lines(runningmean(1, 1000)~x, type="l",    col = sample(viridis(15, option="B",1))))
```



Law of Large Numbers







Strong Law of Large Numbers

Recall: Fix $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| \le \varepsilon) = 1$

(weak law of large numbers)

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Assume that X_1 has finite mean μ and $E \mid X_1 \mid < \infty$

$$A = \left\{ \lim_{n\to\infty} \frac{X_1 + X_2 + \ldots + X_n}{n} = \mu \right\},$$

then

$$P(A) = 1.$$

Stronger statement

$\mathbb{P}(\forall \varepsilon > 0 \ \exists N \ge 1 \text{ s.t. } \forall n \ge N \ |\bar{X}_n - \mu| < \varepsilon)$

$= \mathbb{P}\left(\bigcap_{\varepsilon>0} \bigcup_{N=1}^{\infty} \bigcap_{n=N} \{|\bar{X}_n - \mu| < \varepsilon\} \right) = 1$

Recall

$X_1, \dots, X_n$ i.i.d X $\quad E[X] = \mu, \quad Var[X] = \sigma^2$

- ✓ $E[\bar{X}_n] = \mu \quad , \quad Var[\bar{X}_n] = \frac{\sigma^2}{n}$

$$\left(\mu - \frac{\sigma}{\sqrt{n}} \; , \; \mu + \frac{\sigma}{\sqrt{n}}\right) \equiv \text{Effective range of } \bar{X}$$

- ✓ $\varepsilon > 0$: $\quad \mathbb{P}(|\bar{X} - \mu| > \varepsilon) \longrightarrow 0$

  ↳ $\bar{X}$ Concentrates around mean

- Fact: If: $X \overset{d}{=} \text{Normal}(\mu, \sigma^2)$ then $\bar{X} \overset{d}{=} \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$

(and)

$$\sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) \equiv \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{d}{=} \text{Normal}(0,1)$$

Is this always true?

i.e.

$X_1, \dots, X_n$ i.i.d X $\quad E[X] = \mu, \quad Var[X] = \sigma^2$

$$\sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) \equiv \text{looks like} \equiv \text{Normal}(0,1) \quad \text{for large } n$$

Central Limit Theorem

Suppose each X_i was distributed as Bernoulli (p) random variable. Then S_n is a Binomail(n,p) random variable. Let us check for what p does

$$\frac{S_n - np}{\sqrt{np(1-p)}}$$

is close to a Normal distribution.

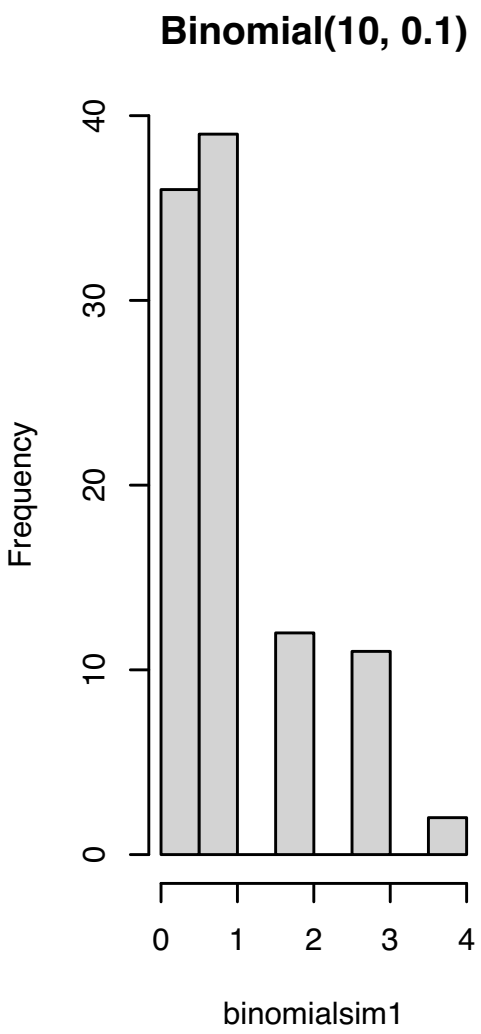
Next:- Verification via simulation

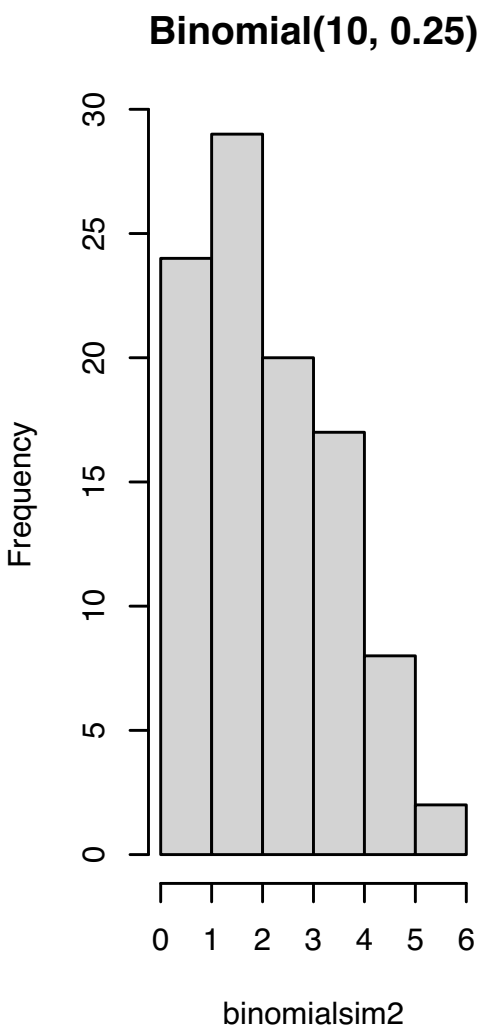
Central Limit Theorem

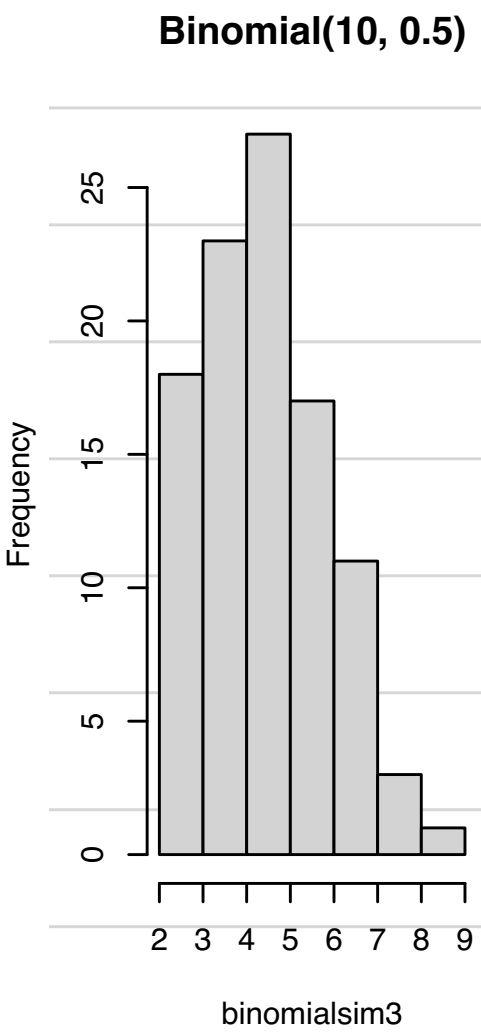
We may simulate Binomial samples either direclty by `rbinom` command or usi ng the `replicate` and `rbinom` command.

```
> binomialsim1 = rbinom(100,10,0.1)
> # generates 100 Binomial (10,0.1) samples
>
> binomialsim2 = rbinom(100,10,0.25)
> # generates 100 Binomial (10,0.25) samples
>
> binomialsim3 = rbinom(100,10,0.5)
> # generates 100 Binomial (10,0.5) samples
>
```

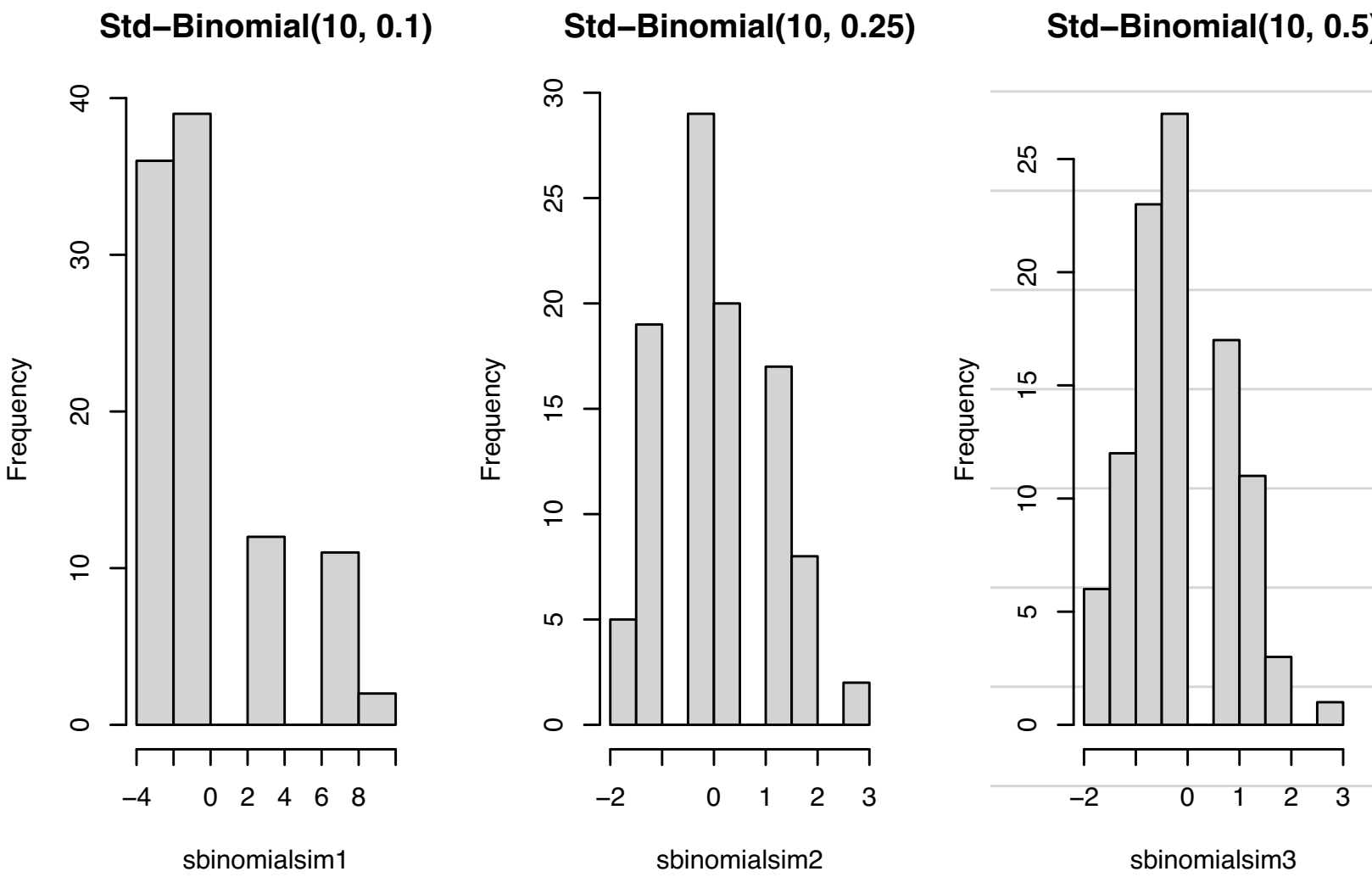
Histogram of all three simulations



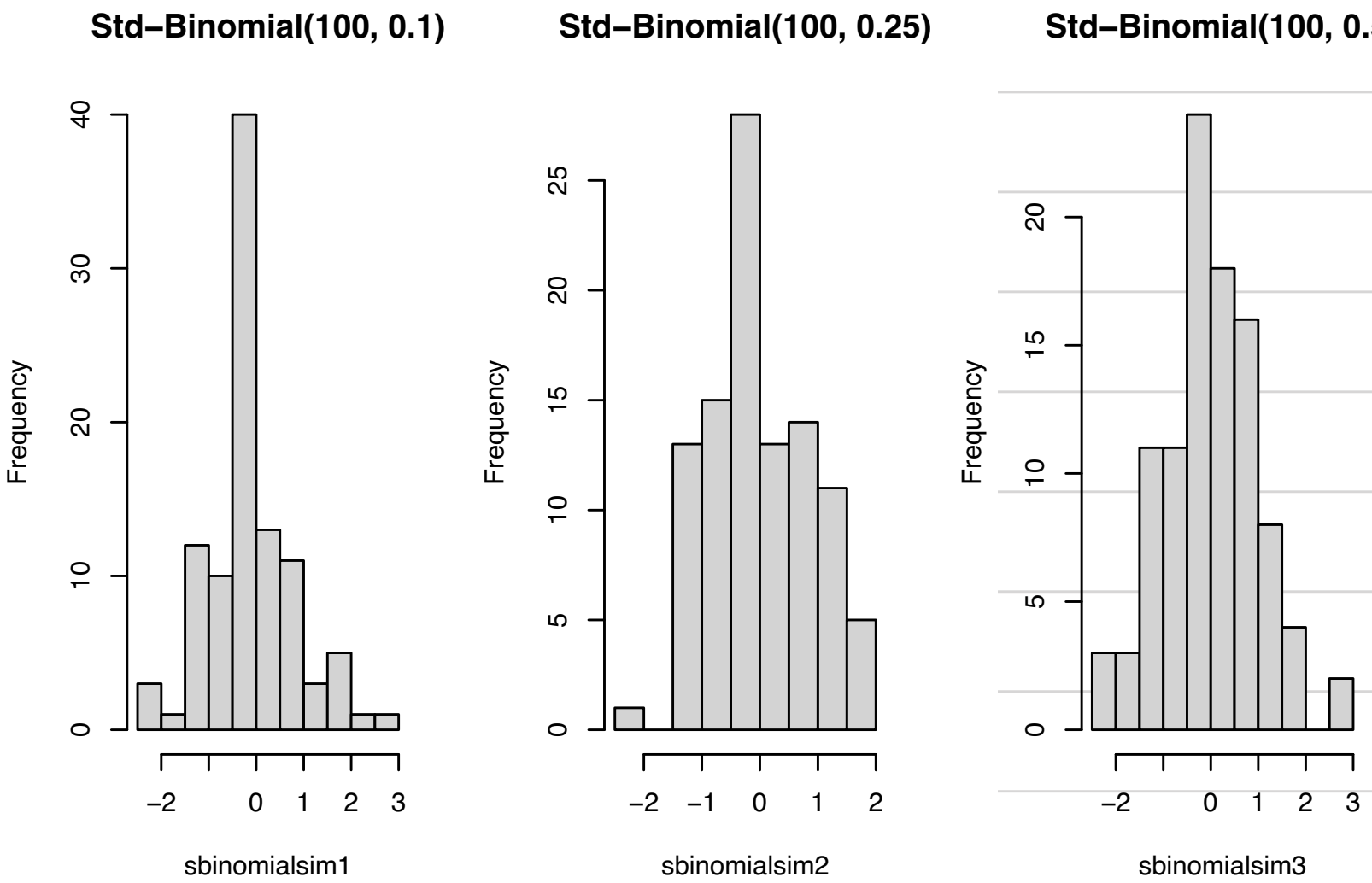




Standardised Histograms: Binomial n=10 and p=0.1, 0.25, 0.5

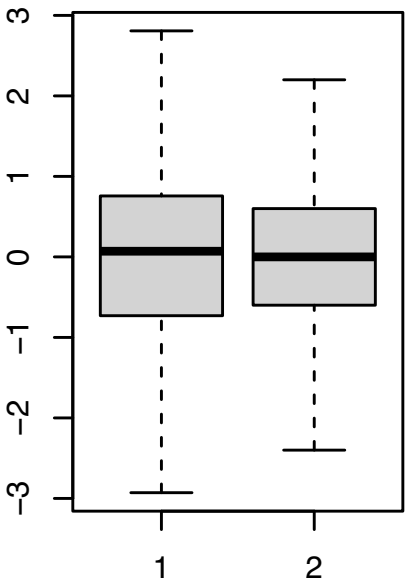


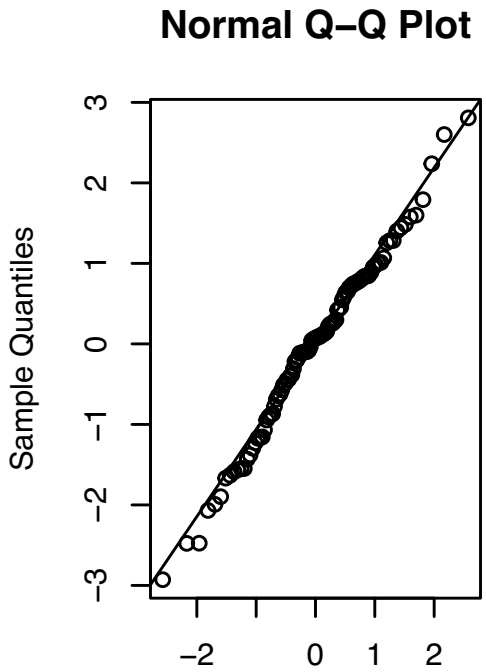
Standardised Histograms: Binomial n=100 and p=0.1, 0.25, 0.5

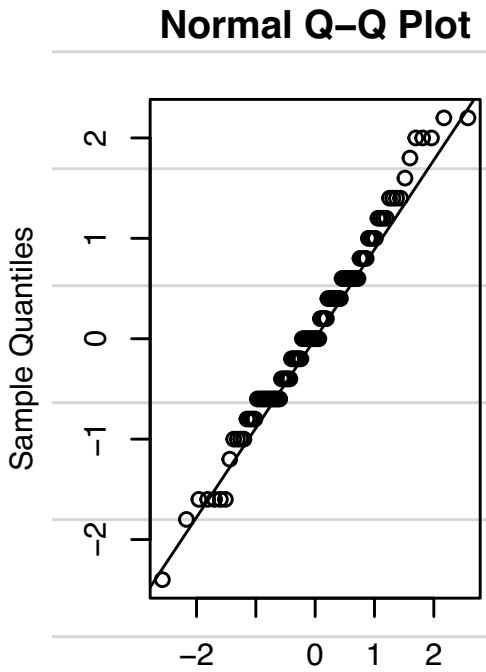


QQ-plot, Binomial(100,0.5)-versus Normal

```
> x = rnorm(100); y = (rbinom(100,100,0.5)-100*0.5)/(sqrt(100*0.5*0
> par(mfrow=c(1,3))
> boxplot(x,y)
> qqnorm(x);qqline(x)
> qqnorm(y);qqline(y)
```







Skewness-Kurtosis, Binomial(100,0.5)-versus Normal

```
> x = rnorm(100)
> y = (rbinom(100,100,0.5)-100*0.5)/(sqrt(100*0.5*0.5))
> library(moments)
> c(skewness(x),skewness(y))
[1] 0.04235202 0.08212817
> c(kurtosis(x), kurtosis(y))
[1] 2.725994 2.626574
```

Role of n versus p

Binomial Random variable is close to Normal when the distribution is symmetric. That is when p is close to 0.5. Otherwise the general rule that we can apply is that when

$$np \geq 5 \text{ and } n(1-p) \geq 5.$$

then Binomial(n,p) is close to Normal distribution.

Reconcile Simulation

$p = 0.1$

$n = 10$, 100, 1000

- $S_n \overset{d}{=}$ Binomial (n, p)

- $S_n = \sum_{i=1}^{n} x_i$ $\quad x_i \overset{d}{=}$ Bernoulli $(p) \equiv X$ independent

$E[X] = p$

$Var(X) = p(1-p)$

$= n \overline{X}$

- $\sqrt{n}\left(\dfrac{\overline{X} - p}{\sqrt{p(1-p)}}\right) \overset{"d"}{=}$ Normal $(0,1)$

$\Big\|$

$\sqrt{n}\left(\dfrac{\frac{S_n}{n} - p}{\sqrt{p(1-p)}}\right) = \dfrac{S_n - np}{\sqrt{n\,p(1-p)}}$

$= \dfrac{S_n - n\mu}{\sqrt{n}\,\sigma}$

- $S_n \simeq$ Normal $(np, np(1-p))$

Central limit theorem

Let $X_1, \{X_n\}_{n \geq 1}$ be a sequence of iid random variables.

$$E[X] = \mu \quad \text{and} \quad Var[X] = \sigma^2$$

$$\frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma} = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

For all $x \in \mathbb{R}$

$$\mathbb{P}\left(\frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma} \leq x\right) \xrightarrow[n \to \infty]{} \int_{-\infty}^{x} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}\, dy$$

Proof: [OMIT] - chapter 8 of book □

Simulation:

- Fix $n \geq 1$
- Generate "lots" sample of

$$\frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma} = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

1, 2, 3 - SD rule

Kurtosis Skewness

Plot Histogram | QQ-QR plot | Box plot

Check: if they agree with

- Generate "lots" sample of

Normal $(0,1)$

Plot Histogram | QQ-QR plot | Box plot

Ex: x – Uniform $(0,1)$ $\quad \mu = E[x] = \frac{1}{2}$, $\quad \text{Var}[x] = \frac{1}{12}$

- Generate 100 samples of
  - $n = 100$

$$\boxed{\frac{S_{100} - 50}{10 \cdot \sqrt{1/12}}} = \boxed{10\left(\frac{\bar{X}_{100} - 0.5}{\sqrt{1/12}}\right)}$$

Find
- plot histogram
  Q-Q plot
  skewness / kurtosis

Central Limit Theorem

Suppose we want to verify the below result via simulations:

Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite mean μ, finite variance σ^2.

Then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} Z, \tag{2}$$

where $\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$ and $Z \sim \text{Normal}\,(0, 1)$.

Central Limit Theorem

(Asymptotic normal)

Say: $S_n = A.N.(n\mu, n\sigma^2)$

We could rephrase the result as:

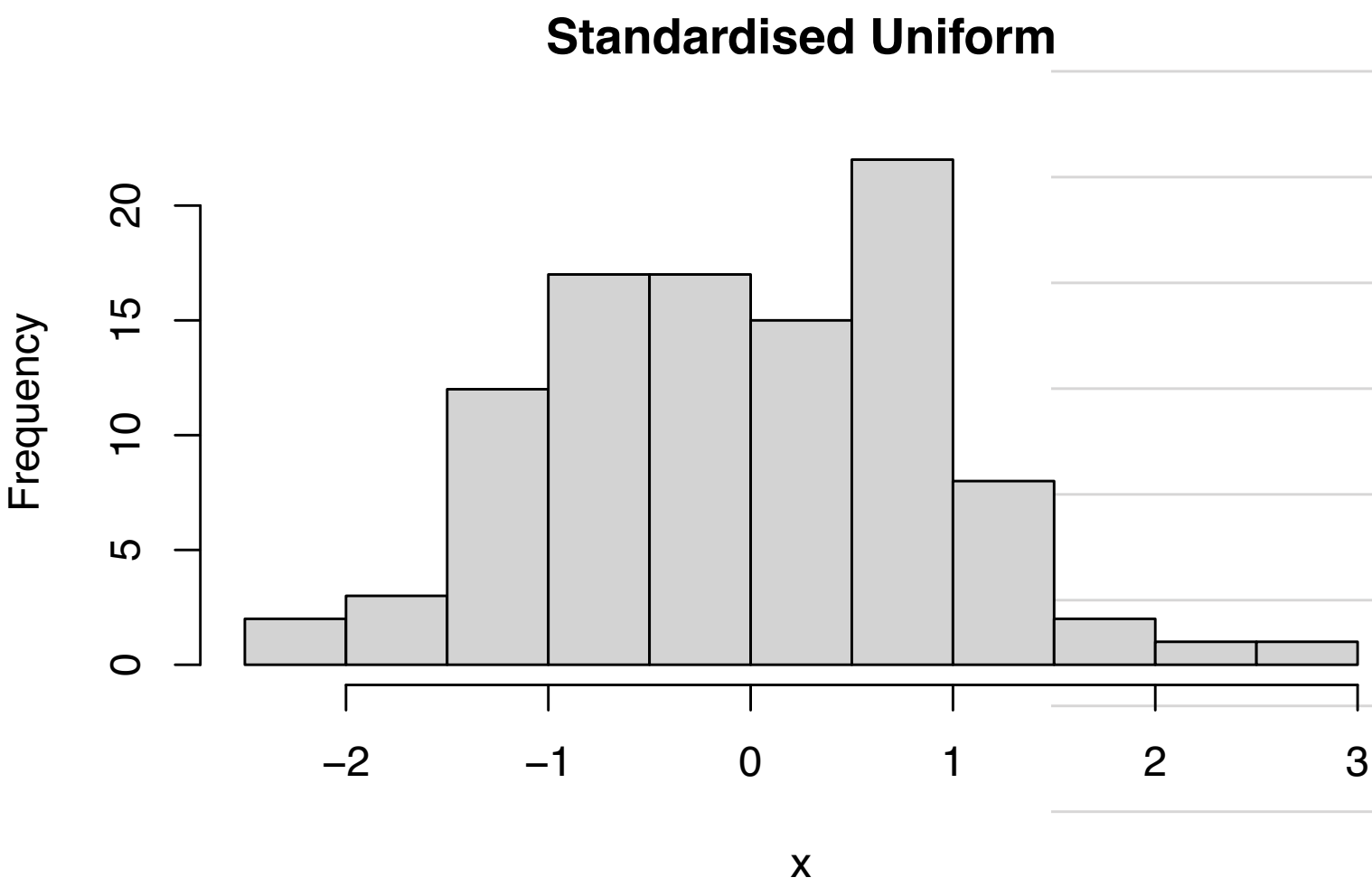
Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite mean μ, finite variance σ^2.

Then

$$\frac{(S_n - n\mu)}{\sqrt{n}\sigma} \xrightarrow{d} Z, \tag{3}$$

where $S_n = X_1 + X_2 + \ldots + X_n$ and $Z \sim \text{Normal}\,(0, 1)$.

Standardised Histograms: Uniform$(0, 1)$



Skewness-Kurtosis, Uniform$(0, 1)$-versus Normal

```
> x = rnorm(100)
> muniform = replicate(100, mean(runif(100)))
> suniform = sqrt(100)*(muniform -0.5)/(sqrt(1/12))
> library(moments)
> c(skewness(x),skewness(suniform))
[1]  0.2805206 -0.1589919
> c(kurtosis(x), kurtosis(suniform))
[1] 3.131999 2.511119
```