Lecture 10, Part II: Special Distributions

Prof. Esther Duflo

14.310x

What is so special about us?

- Some distributions are special because they are connected to others in useful ways
- Some distributions are special because they can be used to model a wide variety of random phenomena.
- This may be the case because of a fundamental underlying principle, or because the family has a rich collection of pdfs with a small number of parameters which can be estimated from the data.
- Like network statistics, there are always new candidate special distributions! But to be really special a distribution must be mathematically elegant, and should arise in interesting and diverse applications
- Many special distributions have standard members, corresponding to specified values of the parameters.
- Today's class is going to end up being more of a reference class than a conceptual one...

We have seen some of them -we may not have named them!

- Bernouili
- Binomial
- Uniform
- Negative binomial
- Geometric
- Normal
- Log-normal
- Pareto

Bernouilli

Two possible outcomes ("success" or "failure"). The probability of success is p, failure is q (or: $1-p$)

$$f(x;p) = p^x q^{1-x} \quad \text{for } x \in \{0,1\}$$

$$0 \quad \text{otherwise}$$

$$\mathrm{E}(X) = p$$

(because: $\mathrm{E}[X] = \Pr(X=1)\cdot 1 + \Pr(X=0)\cdot 0 = p\cdot 1 + q\cdot 0 = p$)

$$\mathrm{E}[X^2] = \Pr(X=1)\cdot 1^2 + \Pr(X=0)\cdot 0^2 = p\cdot 1^2 + q\cdot 0^2 = p$$

and

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = p - p^2 = p(1-p) = pq$$

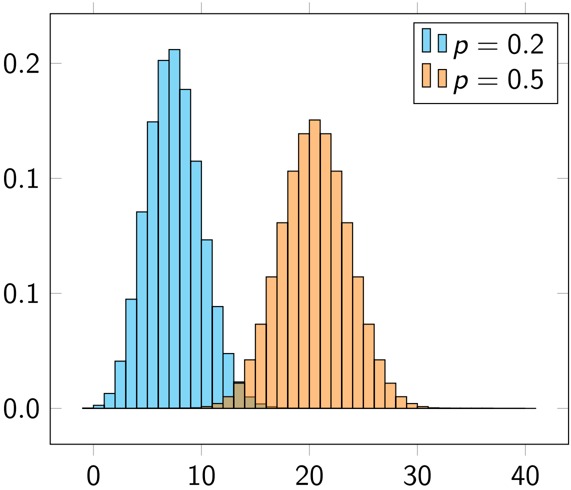
Binomial

Results: If $X_1, \ldots, X_n$ are independent, identically distributed (i.i.d.) random variables, all Bernoulli distributed with success probability p, then $X = \sum_{k=1}^{n} X_k \sim \mathrm{B}(n, p)$ (binomial distribution). The Bernoulli distribution is simply $\mathrm{B}(1, p)$
The binomial distribution is number of successes in a sequence of n independent (success/failure) trials, each of which yields success with probability p.
$f(x; n, p) = \Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, 3, \ldots, n$
$f(x; n, p) = 0$ otherwise.
where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$
Since the binomial is a sum of i.i.d Bernoulli, the mean and variance follows from what we know about these operators:

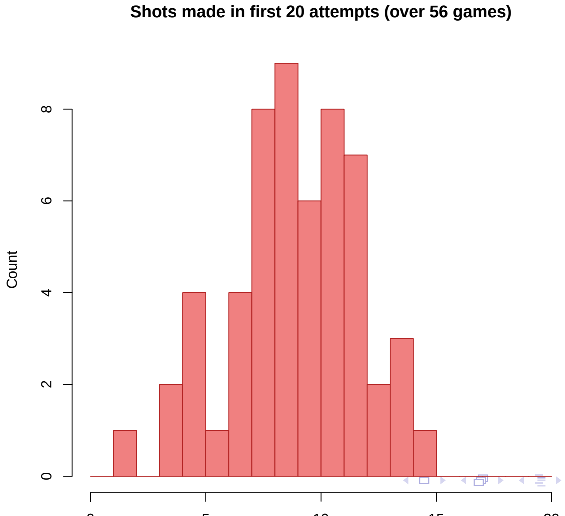
$$E(X) = np$$

$$Var(X) = npq$$

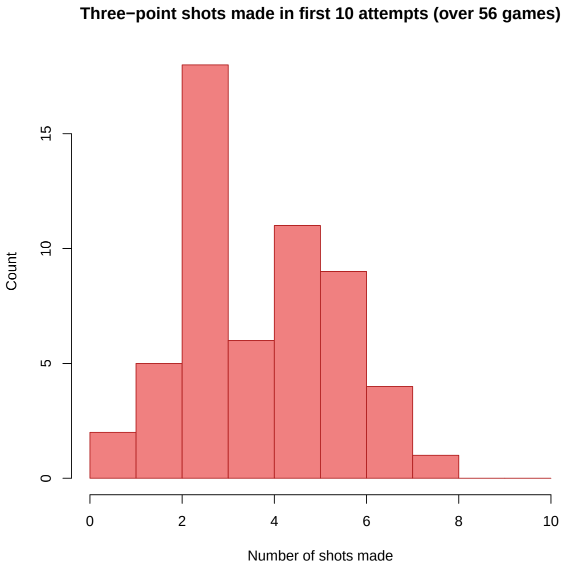
Binomial



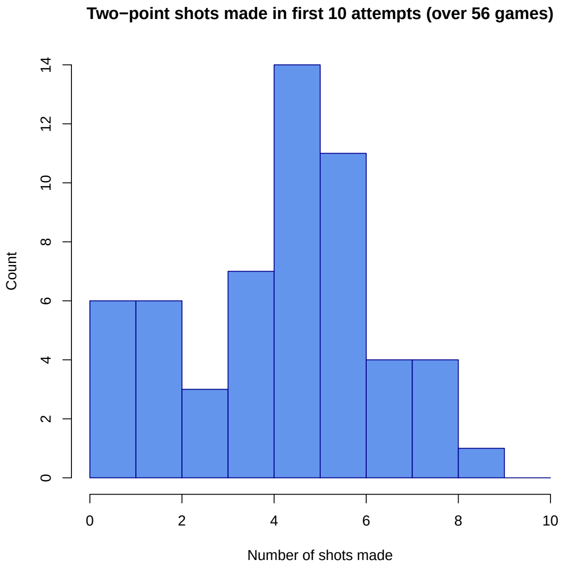
Does the number of Steph Curry's successful shot follows a binomial distribution?



But it is not likely–3pt success



But it is not likely–2pt success



Hypergeometric

- The binomial distribution is used to model the number of successes in a sample of size n *with replacement*
- If you sample *without* replacement, you get the hypergeometric distribution (e.g. number of red balls taken from an urn, number of vegetarian toppings on pizza)

let A be the number of successes and B the number of failure (you may want to define $N = A + B$), n the number of draws, then:

$$f(X|A, B, n) = \frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}},$$

$E(X) = \frac{nA}{A+B}$ and $V(X) = n(\frac{A}{A+B})(\frac{B}{A+B})(\frac{A+B-n}{A+B-1})$
Notice the relationship with the binomial, with $p = \frac{A}{A+B}$ and $q = \frac{B}{A+B}$.

- Note that if N is much larger than n, the binomial becomes a good approximation to the hypergeometric distribution

Negative Binomial

Consider a sequence of independent Bernouilli trials, and let X be the number of trials necessary to achieve r successes
$f_X(x) = \binom{x-1}{r-1} p^r q^{x-r}$ if $x = r, r+1 \dots$, and 0 otherwise.
$p^{r-1}q^{x-r}$ is the probability of any sequence with $r-1$ success and $x-r$ failures.
p is the probability of success after $r-1$ failures.
$\binom{x-1}{r-1}$ is the number of possibility of sequences where $r-1$ are success

$$E(X) = \frac{rq}{p}$$

$$V(X) = \frac{rq}{p^2}$$

(Alternatively, some textbooks/people can define is at the number of failures needed to achieve r successes.)

Geometric

- A negative binomial distribution with $r = 1$ is a geometric distribution [number of failures before the first success]
- $f(x;p) = pq^x$ if $x = 0, 1, 2, 3, ..$; 0 otherwise
  $E(X) = \frac{q}{p}$ $V(X) = \frac{q}{p^2}$
- The sum of r independent Geometric (p) random variables is a negative binomial (r, p) random variable
- By the way, if X_i are iid, and negative binomial (r_i, p), then $\sum X_i$ is distributed as a negative binomial $(\sum r_i, p)$
- Memorylessness: Suppose 20 failures occured on first 20 trials. Since all trials are independent, the distribution of the *additional* failures before the first success will be geometric.

Poisson

The poisson distribution expresses the probability of a given number of events occuring in a fixed interval of time if (1) the event can be counted in whole numbers (2) the occurences are independent and (3) the average frequency of occurrence for a time period is known.

Poisson

Formally, for $t \geq 0$, let N_t be an integer-valued random variables. If it satisfies the following properties

1. $N_0 = 0$
2. for $s < t$, N_s and $N_t - N_s$ are independent

Poisson

Formally, for $t \geq 0$, let N_t be an integer-valued random variables. If it satisfies the following properties

1. $N_0 = 0$
2. for $s < t$, N_s and $N_t - N_s$ are independent [arrival are independent in disjoint intervals]

Poisson

Formally, for $t \geq 0$, let N_t be an integer-valued random variables. If it satisfies the following properties

1. $N_0 = 0$
2. for $s < t$, N_s and $N_t - N_s$ are independent [arrival are independent in disjoint intervals]
3. N_s and $N_{t+s} - N_t$ have identical distribution

Poisson

Formally, for $t \geq 0$, let N_t be an integer-valued random variables. If it satisfies the following properties

1. $N_0 = 0$
2. for $s < t$, N_s and $N_t - N_s$ are independent [arrival are independent in disjoint intervals]
3. N_s and $N_{t+s} - N_t$ have identical distribution [the number of arrivals depend only on period length]

Poisson

Formally, for $t \geq 0$, let N_t be an integer-valued random variables. If it satisfies the following properties

1. $N_0 = 0$
2. for $s < t$, N_s and $N_t - N_s$ are independent [arrival are independent in disjoint intervals]
3. N_s and $N_{t+s} - N_t$ have identical distribution [the number of arrivals depend only on period length]
4. $\lim_{t \to 0} \frac{P(N_t = 1)}{t} = \gamma$

Poisson

Formally, for $t \geq 0$, let N_t be an integer-valued random variables. If it satisfies the following properties

1. $N_0 = 0$
2. for $s < t$, N_s and $N_t - N_s$ are independent [arrival are independent in disjoint intervals]
3. N_s and $N_{t+s} - N_t$ have identical distribution [the number of arrivals depend only on period length]
4. $\lim_{t \to 0} \frac{P(N_t=1)}{t} = \gamma$ [γ is the arrival rate, and it is constant for small interval]

Formally, for $t \geq 0$, let N_t be an integer-valued random variables. If it satisfies the following properties

1. $N_0 = 0$
2. for $s < t$, N_s and $N_t - N_s$ are independent [arrival are independent in disjoint intervals]
3. N_s and $N_{t+s} - N_t$ have identical distribution [the number of arrivals depend only on period length]
4. $\lim_{t \to 0} \frac{P(N_t = 1)}{t} = \gamma$ [γ is the arrival rate, and it is constant for small interval]
5. $\lim_{t \to 0} \frac{P(N_> 1)}{t} = 0$ No simultaneous arrival

Poisson

If N_t satisfies:

1. $N_0 = 0$
2. for$s < t$, N_s and $N_t - N_s$ are independent
3. N_s and $N_{t+s} - N_t$ have identical distribution
4. $\lim_{t\to 0} \frac{P(N_t=1)}{t} = \gamma$
5. $\lim_{t\to 0} \frac{P(N_>1)}{t} = 0$

then for any non-negative integer k

$$P(N_t = k) = \frac{(\gamma t)^k e^{-\gamma t}}{k!}$$

Note: γ and t always appear together so we combine them into one parameter, $\lambda = \gamma t$. γ is the propensity to arrive per unit of time. t is the number of units of time, and λ is the propensity to arrive in some amount of time.

Poisson

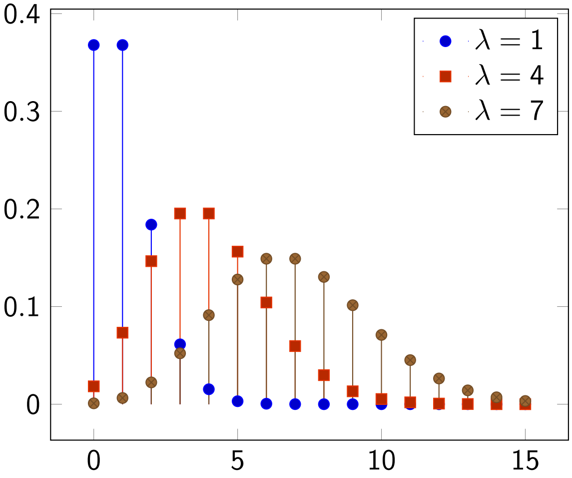
If N_t satisfies:

1. $N_0 = 0$
2. for$s < t$, N_s and $N_t - N_s$ are independent
3. N_s and $N_{t+s} - N_t$ have identical distribution
4. $\lim_{t \to 0} \frac{P(N_t = 1)}{t} = \gamma$
5. $\lim_{t \to 0} \frac{P(N_> 1)}{t} = 0$

then for any non-negative integer k

$$P(N_t = k) = \frac{(\gamma t)^k e^{-\gamma t}}{k!}$$

Note: γ and t always appear together so we combine them into one parameter, $\lambda = \gamma t$. γ is the propensity to arrive per unit of time. t is the number of units of time, and λ is the propensity to arrive in some amount of time.


0.4
0.3
0.2
0.1
0
0
5
10
15
$\lambda = 1$
$\lambda = 4$
$\lambda = 7$

Some properties

- $E[N_t] = \lambda$
- $V[N_t] = \lambda$
- It is asymetrical –skewed–(it cannot be negative!), but closer and closer to being symmetric as λ increases

Relationship between Poisson and Binomial

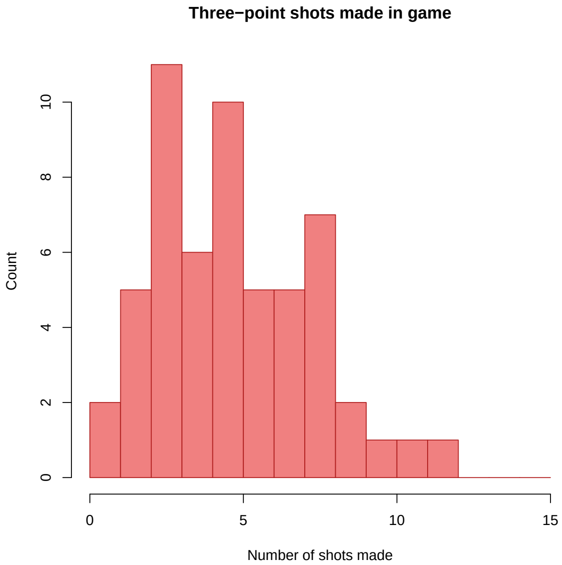
- Divide the interval $[0, t]$ into n subintervals so small that the probability of two occurences in each subinterval is approximately zero.
- The probability of success in each subinterval is now $\frac{\gamma t}{n} = \frac{\lambda}{n}$, and the probability of $n_t = k$ successes in $[0, t]$ is approximately binomial
- $P(N_t = k) \approx \binom{n}{k}(\frac{\lambda}{n})^k(1 - \frac{\lambda}{n})^{n-k}$
- we could prove that the limit of this as the number of subintervals goes to infinity is $\frac{\lambda^k e^{-\lambda}}{k!}$
- In other words, for each nonnegative integer k,

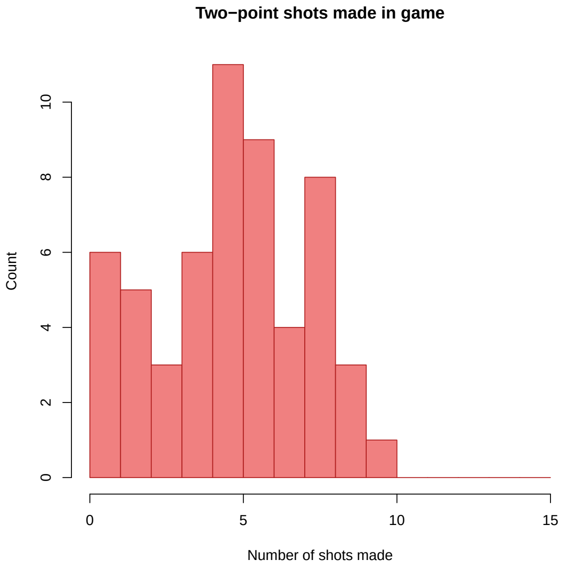
$$\lim_{n \to \infty} p^k (1-p)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

where $p = \frac{1}{\lambda}$, λ is fixed, n is positive.
- For small values of p, the Poisson distribution can simulate the Binomial distribution and it is easier to compute...

When do we use a Poisson distribution?

- Poisson distributions are useful with count data: Number of goals in a soccer match; Number of ideas that a researcher has in a month; number of accidents
- The parameter λ governs both the mean and the variance, so some times that it not what you want (you cannot increase the mean without increasing the variance)
- The negative binomial can be thought of as a generalization that does not have this property
- Some count data won't work well with Poisson: e.g. number of students who arrive at the coop (students arrive together; the events are not independent).


Three−point shots made in game
Count
0
2
4
6
8
10
0
5
10
15
Number of shots made


Two−point shots made in game
Count
0
2
4
6
8
10
0
5
10
15
Number of shots made

Waiting time between two events in a Poisson process:
$f_x = \lambda e^{-\lambda x}$ if $x > 0$ and 0 otherwise

$$E(X) = \frac{1}{\lambda}$$
$$V(X) = \frac{1}{\lambda^2}$$

The exponential distribution is Memoryless:$(P(X \geq t) = e^{-\lambda t}$
therefore $P(X \geq t + h | X \geq t) = P(X \geq h)$
It is a special case of an **Gamma** distribution the "waiting time" before a number (not necessary an integer number) of occurences.
We are skipping the mathematical description of the gamma distribution for now...

Continuous distributions

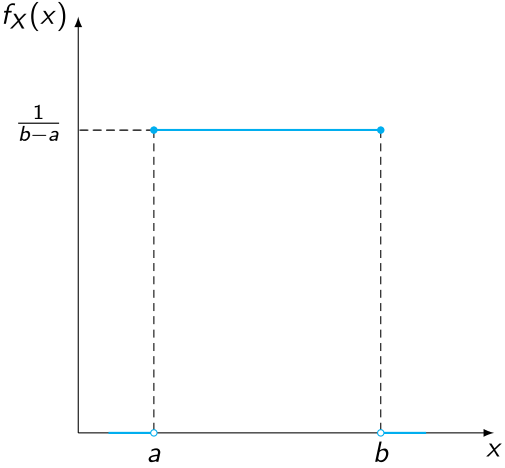
- Uniform
- Normal

The probability that X is in a certain sub-interval $[a; b]$ depends only on the length of that interval.

$$f_x(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b \end{cases}$$

Uniform distribution: density



Properties

- Mean
$$E(X) = \frac{1}{2}(a+b)$$
- $$E(X^2) = \frac{1}{3}\frac{b^3 - a^3}{b-a}$$
- Variance
$$V(X) = \frac{1}{12}(b-a)^2$$
- Set $a = 0$ and $b = 1$. The resulting distribution $U(0,1)$ is called standard uniform distribution. Note that if u_1 is standard uniform, so is $1 - u_1$.

Applications

- Many many: very useful in hypothesis testing for example.
- An important one: Quasi-random number generators. Computers don't really know random numbers... Many programming languages have the ability to generate pseudo-random numbers, which are really draw from a standard uniform distribution
- So the uniform distribution is very useful for example when you want to create a sample of treated and control observations (an example in R follows in one slide).
- As we have learnt, from a uniform distribution, you can use the inverse CDF method to get a sample for many (not all) distributions you are interested in

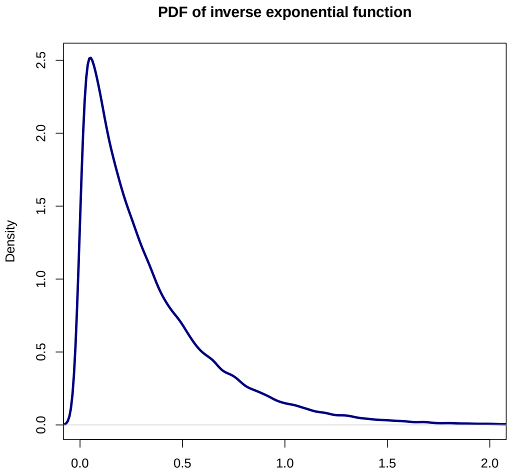
Applications

- Many many: very useful in hypothesis testing for example.
- An important one: Quasi-random number generators. Computers don't really know random numbers... Many programming languages have the ability to generate pseudo-random numbers, which are really draw from a standard uniform distribution
- So the uniform distribution is very useful for example when you want to create a sample of treated and control observations (an example in R follows in one slide).
- As we have learnt, from a uniform distribution, you can use the inverse CDF method to get a sample for many (not all) distributions you are interested in
- ... or you can just directly sample from the relevant distribution in R. [note that R does not always use the inverse transform method...]

sampling from an exponential using the inverse sampling method

```
## Random draws from uniform distribution
u <- runif(100000,0,1)

## Plot the inverse of CDF of the exponential
pdf("runiform_inverse_exponential.pdf")
inverse_exponential_cdf <- function(x,lambda) -log(x)/lambda
y <- inverse_exponential_cdf(u,3)
density_y <- density(y)
plot(density_y,type="l",xlim=c(0,2),
    main="PDF of inverse exponential function",
    lwd=3,col="navyblue",xlab="")
hide<-dev.off()
```

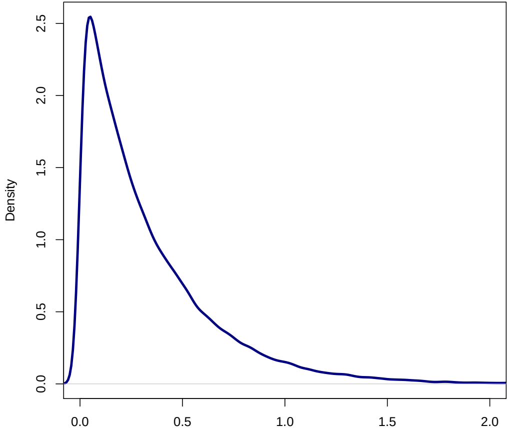
sampling from an exponential using the inverse sampling method



OR...

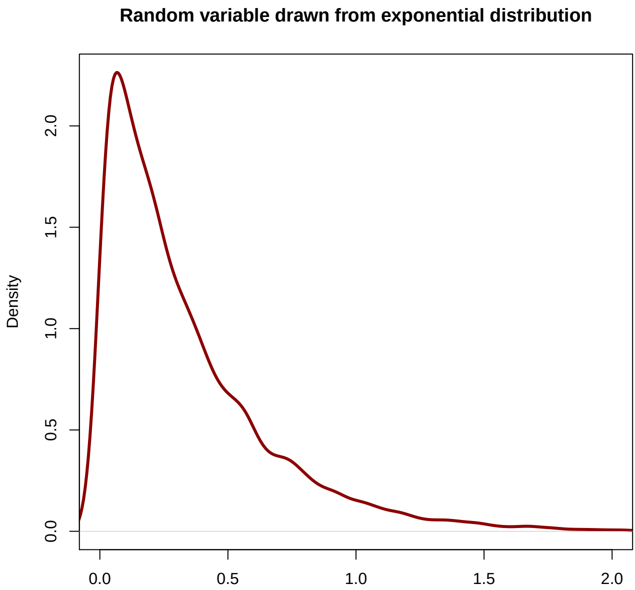
```
## Plot the inverse of CDF of the exponential using q_exp
pdf("runiform_inverse_exponential_qexp.pdf")
y_qexp <- qexp(u,rate=3)
density_y_qexp <- density(y_qexp)
plot(density_y_qexp,type="l",xlim=c(0,2),
    main="PDF of inverse exponential function",
    lwd=3,col="navyblue",xlab="")
hide<-dev.off()
```

PDF of inverse exponential function



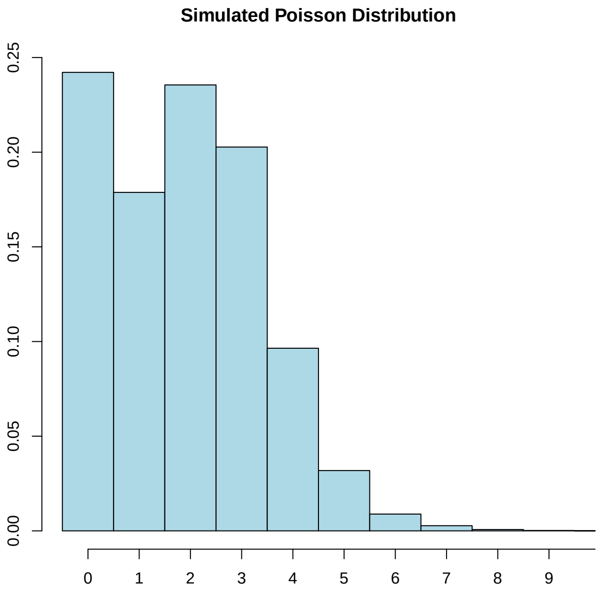
OR

```
## Compare to random draws straight from the exponential distribution
pdf("random_from_exponential.pdf")
y_rexp <- rexp(10000,rate=3)
density_y_rexp <- density(y_rexp)
plot(density_y_rexp,type="l",xlim=c(0,2),
    main="Random variable drawn from exponential distribution",
    lwd=3,col="darkred",xlab="")
hide<-dev.off()
```


Random variable drawn from exponential distribution
Density
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0

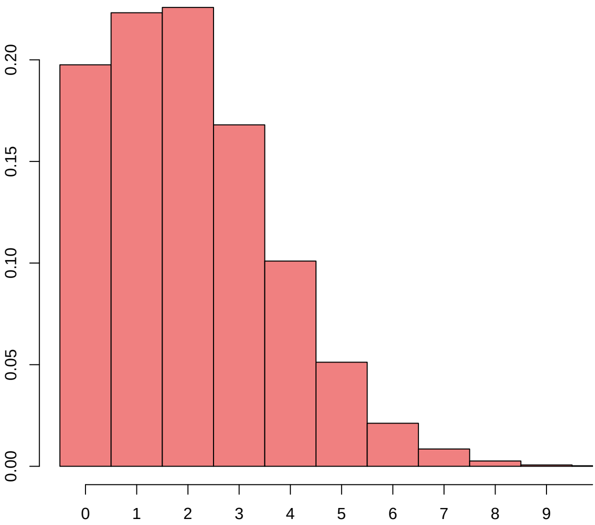
```r
## Poisson simulation
poisson<-numeric(1000000)

lambda<-2
c <- (0.767-0.336/lambda)
beta <- pi/sqrt(3.0*lambda)
alpha <- beta*lambda
k <- (log(c) - lambda - log(beta))

set.seed(20)
u <- runif(100000,0,1)
x <- (alpha-log((1.0-u)/u)/beta)
n <- floor(x+0.5)
set.seed(42)
v <- runif(100000,0,1)
y <- alpha-beta*x
lhs <- y + log(v/(1.0+exp(y)^2))
rhs <- k + n*log(lambda)-log(factorial(n))

j <- 1
for (i in 1:100000) {
    if (n[i]>=0) {
        if (lhs[i]<=rhs[i]) {
            poisson[j] <- n[i]
            j <- j+1
        }
    }
}
poisson <- poisson[1:j]
```

```r
## Plot the simulated Poisson random variable
pdf("runiform_poisson_simulation.pdf")
hist<-hist(poisson,
    main="Simulated Poisson Distribution",
    xlim=c(0,10),breaks=0:(max(poisson)+1),
    freq=FALSE,
    xlab="", ylab="",
    col="lightblue",
    xaxt="n")
axis(1,at=hist$mids,labels=0:max(poisson))
hide<-dev.off()
```


Simulated Poisson Distribution
0.00
0.05
0.10
0.15
0.20
0.25
0
1
2
3
4
5
6
7
8
9

```r
## Compare to random draws from the Poisson distribution
pdf("random_from_poisson.pdf")
y_rpois <- rpois(100000,3)
hist <- hist(y_rpois,
    main="Random variable drawn from Poisson distribution",
    xlim=c(0,10),breaks=0:(max(y_rpois)+1),
    freq=FALSE,
    xlab="",ylab="",
    col="lightcoral",
    xaxt="n")
axis(1,at=hist$mids,labels=0:max(y_rpois))
hide<-dev.off()
```

Random variable drawn from Poisson distribution



Choosing a random sample

```
## Sample 25 of 50 States Code, with and without replacement

## Read in list of state names
states <- read.csv("states.csv")

## Sample 25 without replacement, 25 with replacement
states_without_replacement <- list(sample(states$state_name,25,replace=FALSE))
states_with_replacement <- sample(states$state_name,25,replace=TRUE)

## Print output
print(states_without_replacement)
print(states_with_replacement)
```

Choosing a random sample

```
> ## Sample 25 of 50 States Code, with and without replacement
>
> ## Read in list of state names
> states <- read.csv("states.csv")
>
> ## Sample 25 without replacement, 25 with replacement
> states_without_replacement <- list(sample(states$state_name,25,replace=FALSE))
> states_with_replacement <- sample(states$state_name,25,replace=TRUE)
>
> ## Print output
> print(states_without_replacement)
[[1]]
 [1] Alaska         North Carolina New Jersey     Missouri       Louisiana      Virginia       Massachusetts
 [8] Mississippi    Idaho          Delaware       California     Iowa           South Dakota   South Carolina
[15] Illinois       Wyoming        New Mexico     Georgia        Michigan       Indiana        Ohio
[22] Utah           West Virginia  Minnesota      Arizona
50 Levels: Alabama Alaska Arizona Arkansas California Colorado Connecticut Delaware Florida Georgia ... Wyoming

> print(states_with_replacement)
 [1] Missouri       North Carolina Massachusetts  Texas          South Carolina Maryland       Wyoming
 [8] South Carolina Massachusetts  South Carolina Alabama        Vermont        California     Mississippi
[15] Nebraska       Tennessee      New Hampshire  South Dakota   North Carolina Colorado       South Carolina
[22] Maryland       Oklahoma       Oklahoma       Oklahoma
50 Levels: Alabama Alaska Arizona Arkansas California Colorado Connecticut Delaware Florida Georgia ... Wyoming
```

Continuous distributions

- Uniform
- Normal

Theorem

Lex $X \sim \mathrm{B}(n,p)$, *for any number* c *and* d*:*

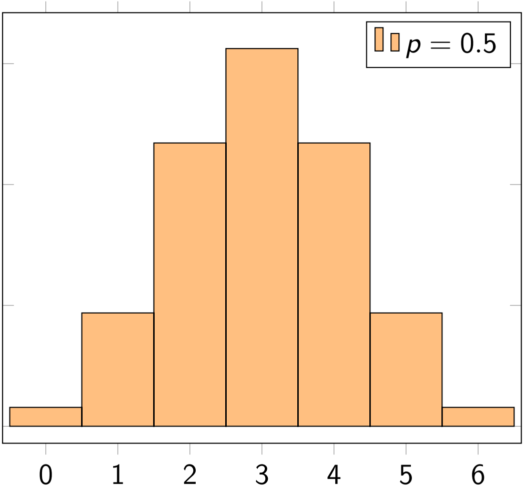
$$\lim_{n\to\infty} P(c \leq \frac{X-np}{\sqrt{np(1-p)}} < d) = \int_c^d \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}} dx$$

$\frac{X-np}{\sqrt{np(1-p)}}$ is the standardized version of the binomial. Keeps mean at zero and variance at 1.

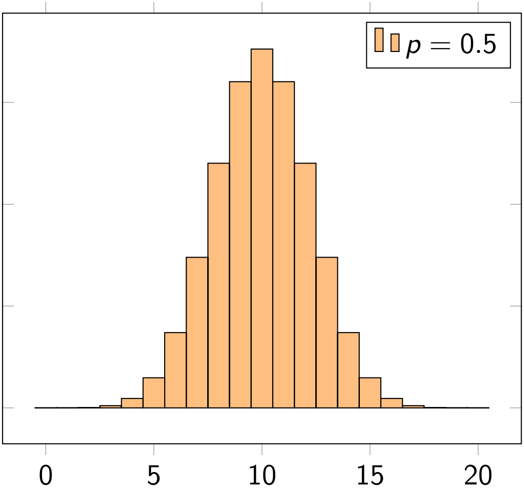
We note: $f_z(y) = \phi(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}$ and $F_Z(y) = \Phi(y) =$ for $-\infty < y < \infty$

$E(Z) = 0$ and $V(Z) = 1$

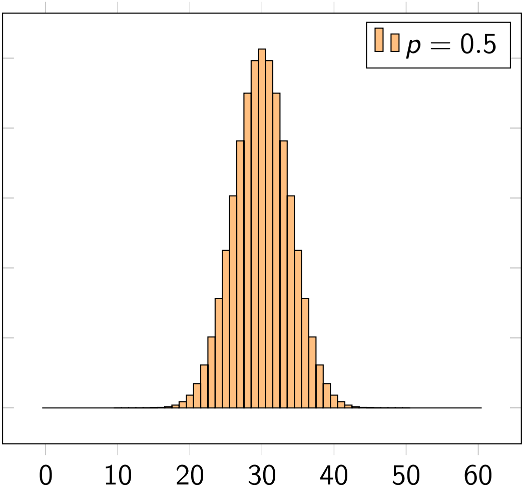
Binomial



Binomial

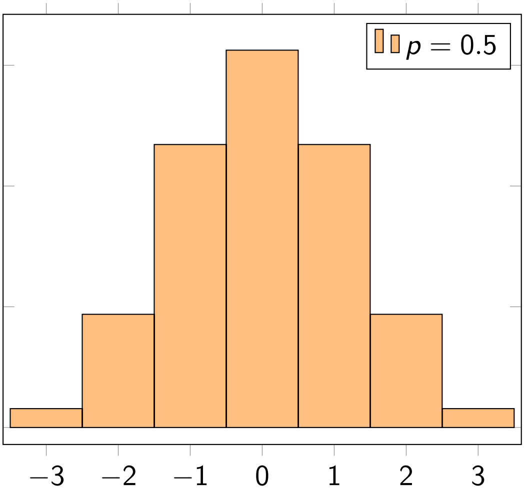


Binomial

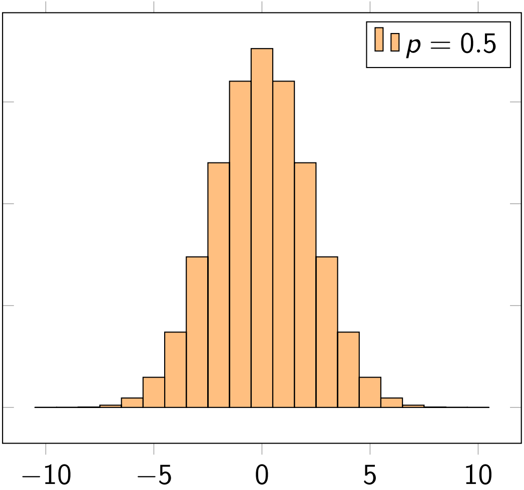


Binomial

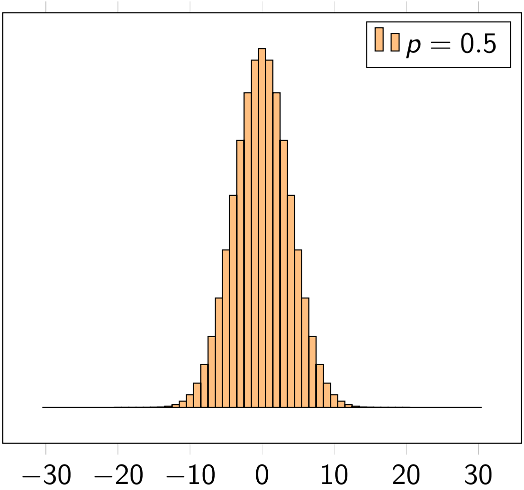
now standardize



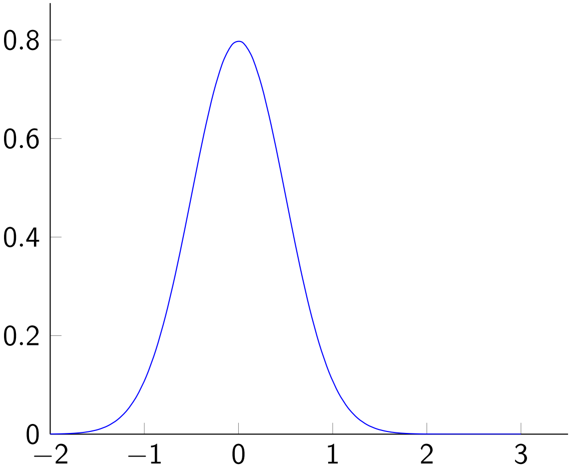
now standardize



now standardize



Standard Normal distribution



Normal distributions

We call any random variable $X = \mu + \sigma Z$ where Z is standard normal with $\sigma \neq 0$ normal as well.

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

for $-\infty < x < \infty$
Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$

Normal distributions

We call any random variable $X = \mu + \sigma Z$ where Z is standard normal with $\sigma \neq 0$ normal as well.

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

for $-\infty < x < \infty$

Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$ X distributed normal with parameters μ and σ^2

$$E(X) = E(Z) + \mu = \mu$$

$$Var(X) = \sigma^2 * Var(Z) = \sigma^2$$

Some properties

- If X_1 is normal, and $X_2 = a + bX_1$ is also normal, with mean $a + bE(X_1)$ and variance $b^2 Var(X_1)$

Theorem

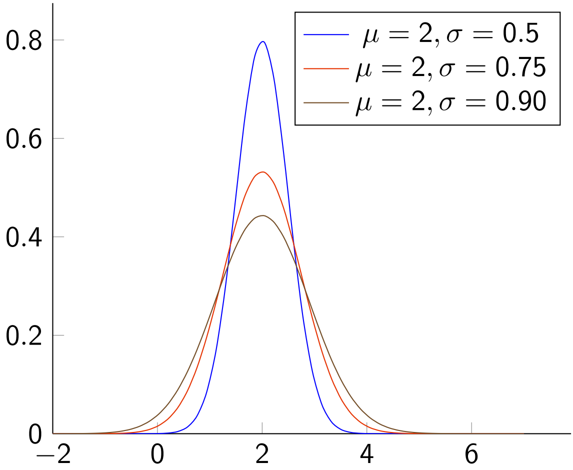
Let $X_1..X_n$ are iid and $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then

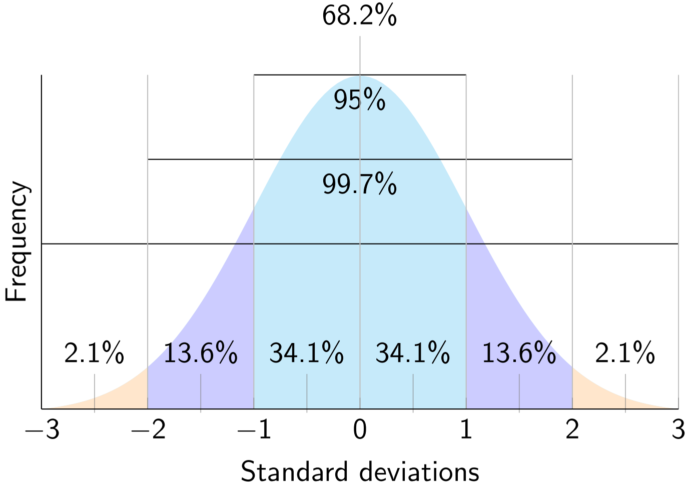
$$Y = \sum_i X_i \sim \mathcal{N}(\sum_i \mu_i, \sum_i \sigma_i^2)$$

We already knew the mean and the variance (by general properties of these operators) but we now also know that the pdf of a sum of normal remains normal.

- Normal distribution are symmetric, unimodal, "bell-shaped", have thin tails, and the support is $\mathbb{R}$

Same mean, different variances





(courtesy: John Canning for the tikzpicture code!)

Finding the area under the curve

- The integral of $\phi(x)$ over regions of $\mathbb{R}$ cannot be expressed in closed-form
- Therefore we use tables (or software...) to figure out the answer we are looking for.
- For example, from the standard normal table, suppose you want $P(Z < -1.23)$.
  - go down the left column to -1.2
  - and the top row to 0.03
  - the answer is

Finding the area under the curve

- The integral of $\phi(x)$ over regions of $\mathbb{R}$ cannot be expressed in closed-form
- Therefore we use tables (or software...) to figure out the answer we are looking for.
- For example, from the standard normal table, suppose you want $P(Z < -1.23)$.
  - go down the left column to -1.2
  - and the top row to 0.03
  - the answer is 1.093

Finding the area under the curve

- The integral of $\phi(x)$ over regions of $\mathbb{R}$ cannot be expressed in closed-form
- Therefore we use tables (or software...) to figure out the answer we are looking for.
- For example, from the standard normal table, suppose you want $P(Z < -1.23)$.
  - go down the left column to -1.2
  - and the top row to 0.03
  - the answer is 1.093

Finding the area under the curve

- what if you wanted $P(Z > -1.68)$

Finding the area under the curve

- what if you wanted $P(Z > -1.68)$
    - $P(Z > -1.68) = 1 - P(Z < -1.68)$

Finding the area under the curve

- what if you wanted $P(Z > -1.68)$
  - $P(Z > -1.68) = 1 - P(Z < -1.68)$
- What if you wanted positive numbers and I had not given you the positive numbers, e.g. $P(Z < 1.45)$

Finding the area under the curve

- what if you wanted $P(Z > -1.68)$
  - $P(Z > -1.68) = 1 - P(Z < -1.68)$
- What if you wanted positive numbers and I had not given you the positive numbers, e.g. $P(Z < 1.45)$
  - Exploit symmetry:
    $P(Z < 1.45) = P(Z > -1.45) = 1 - P(Z < -1.45)$
- What if you wanted $P(-1.23 < Z < 1.45)$

Finding the area under the curve

- what if you wanted $P(Z > -1.68)$
  - $P(Z > -1.68) = 1 - P(Z < -1.68)$
- What if you wanted positive numbers and I had not given you the positive numbers, e.g. $P(Z < 1.45)$
  - Exploit symmetry:
    $P(Z < 1.45) = P(Z > -1.45) = 1 - P(Z < -1.45)$
- What if you wanted $P(-1.23 < Z < 1.45)$
  - $P(-1.23 < Z < 1.45) = P(Z < 1.45) - P(Z < -1.23)$

Finding the area under the curve

- what if you wanted $P(Z > -1.68)$
  - $P(Z > -1.68) = 1 - P(Z < -1.68)$
- What if you wanted positive numbers and I had not given you the positive numbers, e.g. $P(Z < 1.45)$
  - Exploit symmetry: $P(Z < 1.45) = P(Z > -1.45) = 1 - P(Z < -1.45)$
- What if you wanted $P(-1.23 < Z < 1.45)$
  - $P(-1.23 < Z < 1.45) = P(Z < 1.45) - P(Z < -1.23)$
- what if you had a non standard normal?
  - First normalize it. Then use the table.

Useful R command about the Normal distribution

	PURPOSE	SYNTAX	EXAMPLE
RNORM	Generates random numbers from normal distribution	rnorm(n, mean, sd)	rnorm(1000, 3, .25) Generates 1000 numbers from a normal with mean 3 and sd=.25
DNORM	Probability Density Function (PDF)	dnorm(x, mean, sd)	dnorm(0, 0, .5) Gives the density (height of the PDF) of the normal with mean=0 and sd=.5.
PNORM	Cumulative Distribution Function (CDF)	pnorm(q, mean, sd)	pnorm(1.96, 0, 1) Gives the area under the standard normal curve to the left of 1.96, i.e. ~0.975
QNORM	Quantile Function – inverse of pnorm	qnorm(p, mean, sd)	qnorm(0.975, 0, 1) Gives the value at which the CDF of the standard normal is .975, i.e. ~1.96

```
> pnorm(1.96, lower.tail=TRUE)
[1] 0.9750021
> pnorm(1.96, lower.tail=FALSE)
[1] 0.0249979
```

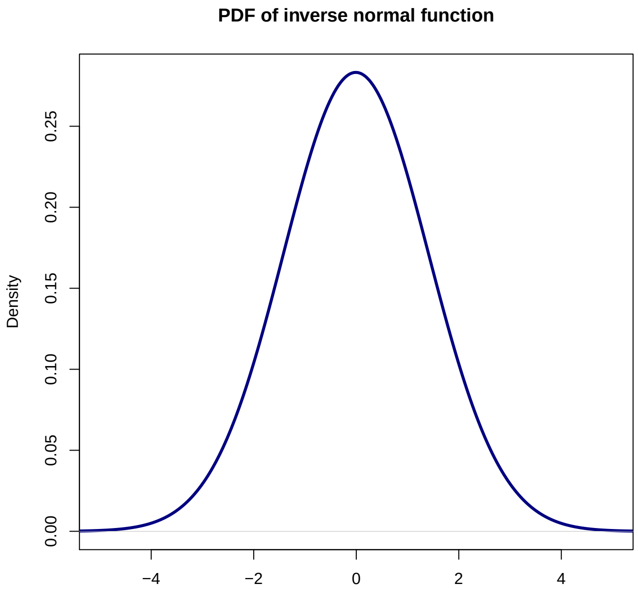
```r
## Compute probabilities from normal distribution

## Characterize distribution
x_mean <- 2
x_sd <- 0.5

## Set inputs
x1 <- 1.2
x2 <- 1.34
x3 <- 1.46
x4 <- 2.08

## Probability less than x1?
pnorm(x1,x_mean,x_sd)

## Probability between x2 and x3?
pnorm(x3,x_mean,x_sd)-pnorm(x2,x_mean,x_sd)

## Probability greater than x4?
pnorm(x4,x_mean,x_sd,lower.tail=FALSE)
```

```
> ## Characterize distribution
> x_mean <- 2
> x_sd <- 0.5
>
> ## Set inputs
> x1 <- 1.2
> x2 <- 1.34
> x3 <- 1.46
> x4 <- 2.08
>
> ## Probability less than x1?
> pnorm(x1,x_mean,x_sd)
[1] 0.05479929
>
> ## Probability between x2 and x3?
> pnorm(x3,x_mean,x_sd)-pnorm(x2,x_mean,x_sd)
[1] 0.04665358
>
> ## Probability greater than x4?
> pnorm(x4,x_mean,x_sd,lower.tail=FALSE)
[1] 0.4364405
```

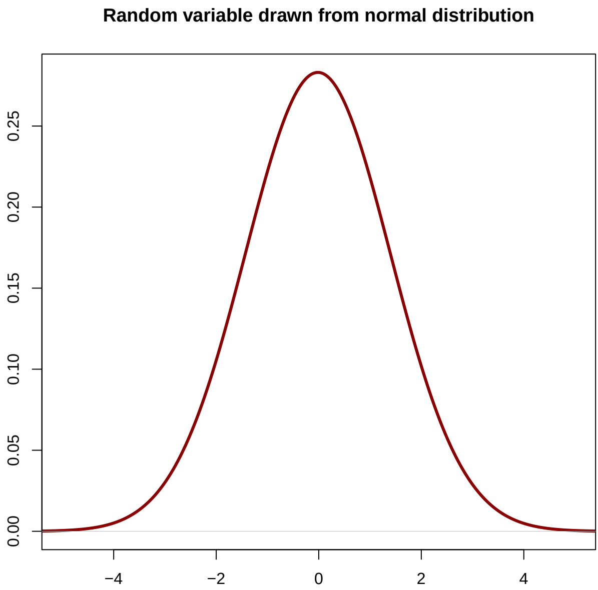
Sampling from a normal distribution in R

- In theory you can use the inverse sampling methods.
- In practice this would take much longer than using the built in command in R that uses a different algorithm.

```r
## Inverse of CDF of normal using qnorm
pdf("runiform_inverse_normal_qnorm.pdf")
y_qnorm <- qnorm(u)
density_y_qnorm <- density(y_qnorm,bw=1)
plot(density_y_qnorm,type="l",xlim=c(-5,5),
    main="PDF of inverse normal function",
    lwd=3,col="navyblue",xlab="")
hide<-dev.off()
```


PDF of inverse normal function
Density
0.00
0.05
0.10
0.15
0.20
0.25
−4
−2
0
2
4

```r
## Compare to random draws straight from the normal distribution
pdf("random_from_normal.pdf")
y_rnorm <- rnorm(10000,0,1)
density_y_rnorm <- density(y_rnorm,bw=1)
plot(density_y_rnorm,type="l",xlim=c(-5,5),
    main="Random variable drawn from normal distribution",
    lwd=3,col="darkred",xlab="",ylab="")
hide<-dev.off()
```


Random variable drawn from normal distribution
0.00
0.05
0.10
0.15
0.20
0.25
−4
−2
0
2
4