Business Statistics 41000

Lecture 5

Mladen Kolar

Continuous Random Variables

Previously, we considered discrete random variables.

We used them to model uncertainty in situations where there were a countable number of outcomes of the random event.

In many situations, a random event can take on a continuous range of values.

We will use a continuous random variable to model such events.

What is the probability of *Yellow*?

What is the probability that the spinner lands in any *particular* place?

We will use **continuous random variables** to capture uncertainty in outcomes that can fall into a continuous range of values.

In the discrete case, we listed the possible outcomes and assigned a probability to each outcome. Now, the number of possible outcomes is so large that we cannot make a list and give each outcome a probability!

- Instead we talk about the probability of intervals.
- Individual points have zero probability.

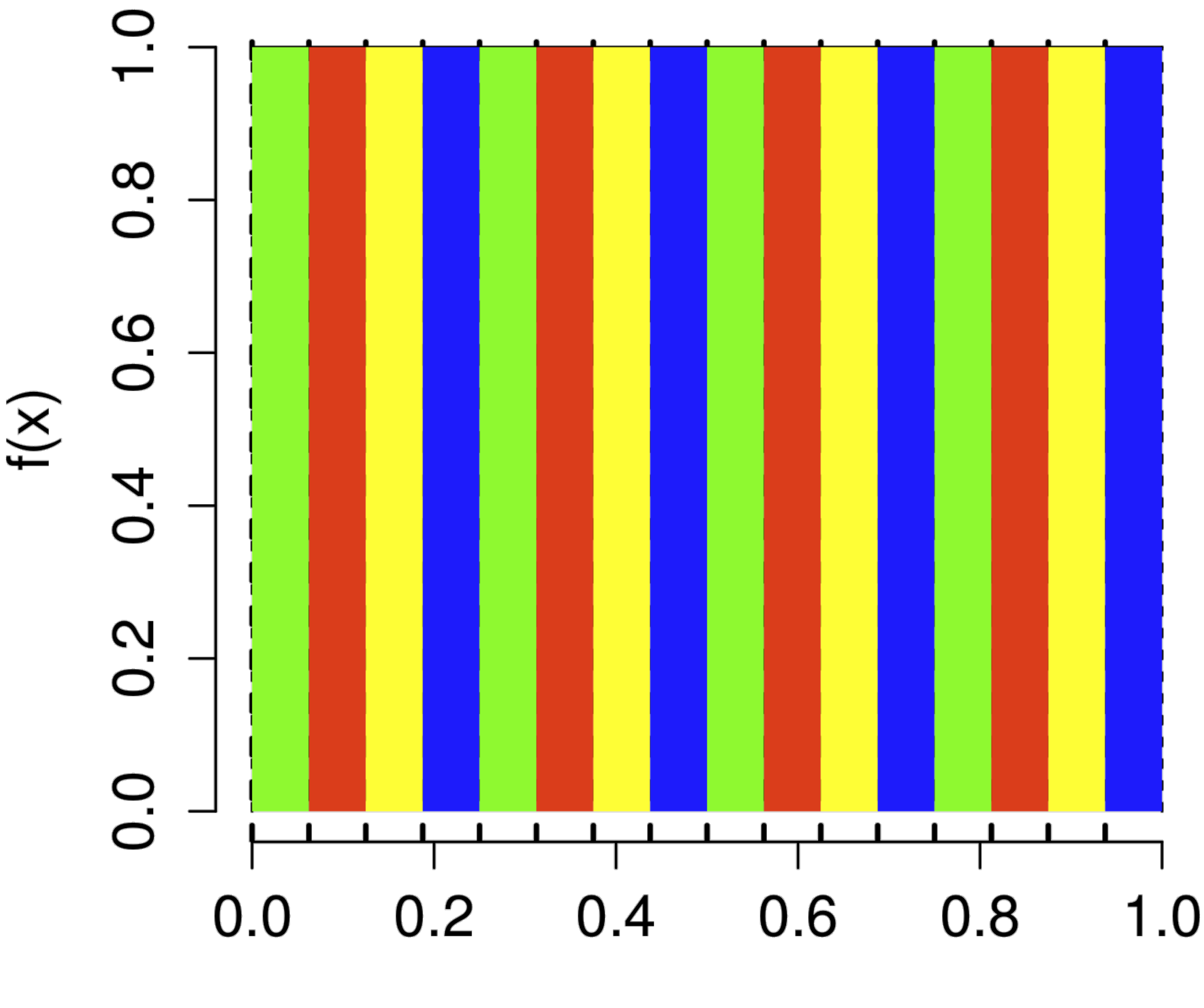
As before, we will use the model X (continuous random variable) to stand for a numerical quantity that we are modelling, but we need a new way to describe what the probabilities are for the outcomes.

In the discrete case, we had $P(X = x) = 0.1$ whereas in the continuous case we have

$$P(0.25 < X < 0.5) = 0.25$$

How do we specify the probability that X falls in an interval?

From Twister to Density — Probability as Area



Probability density function

One convenient way to specify the probability of an interval is with a **probability density function** (pdf).

- We will use $f(x)$ to denote the pdf of X.

The probability of an interval is the area under the pdf over that interval.

$$P(a < X < b) = \int_a^b f(x)\mathrm{d}x$$

Properties of a probability density function

- The area under the pdf curve must be equal to 1.
- $f(x) \geq 0$ for any x.
- **Extremely Important:** For any realization x, $P(X = x) = 0 \neq f(x)$!

The uniform distribution

A random variable X is said to have a uniform distribution if it has probability density function

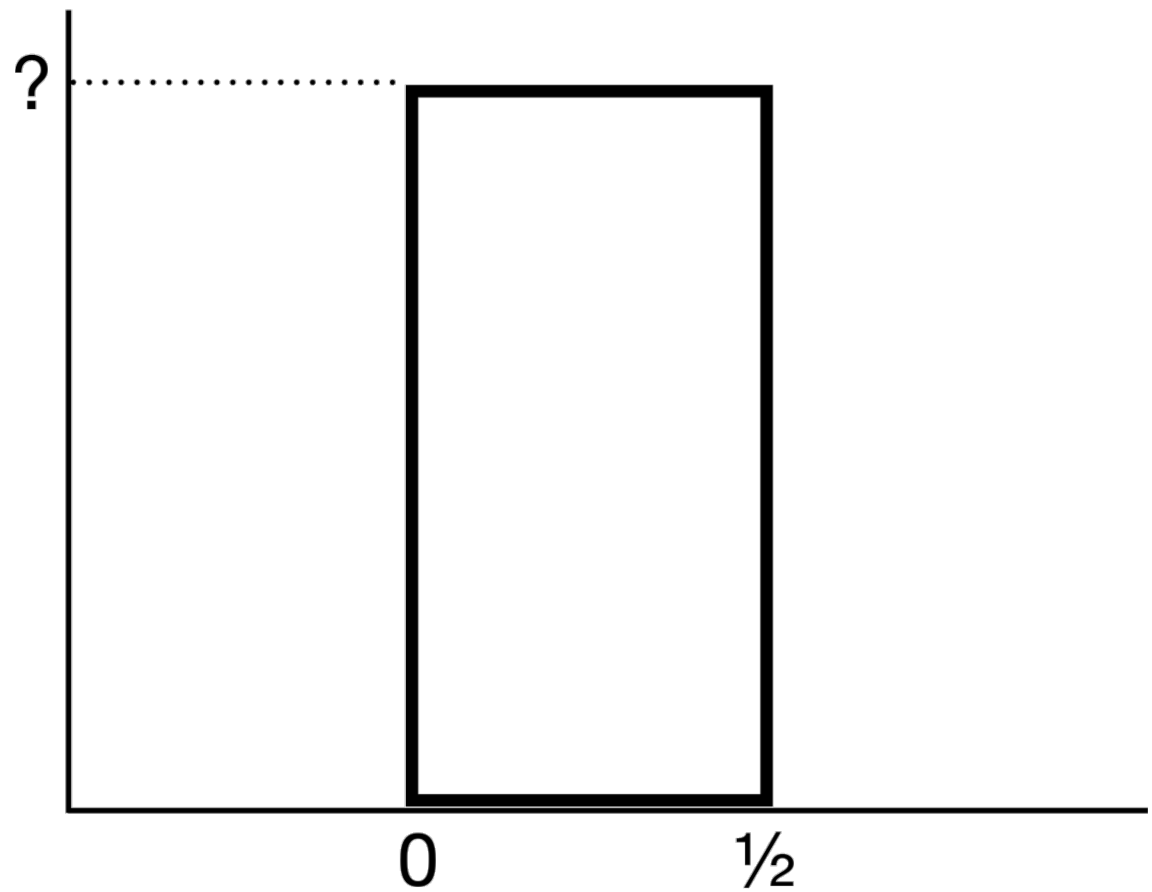
$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

We write $X \sim U(a, b)$.

The parameters of the uniform distribution are the two endpoints a and b.

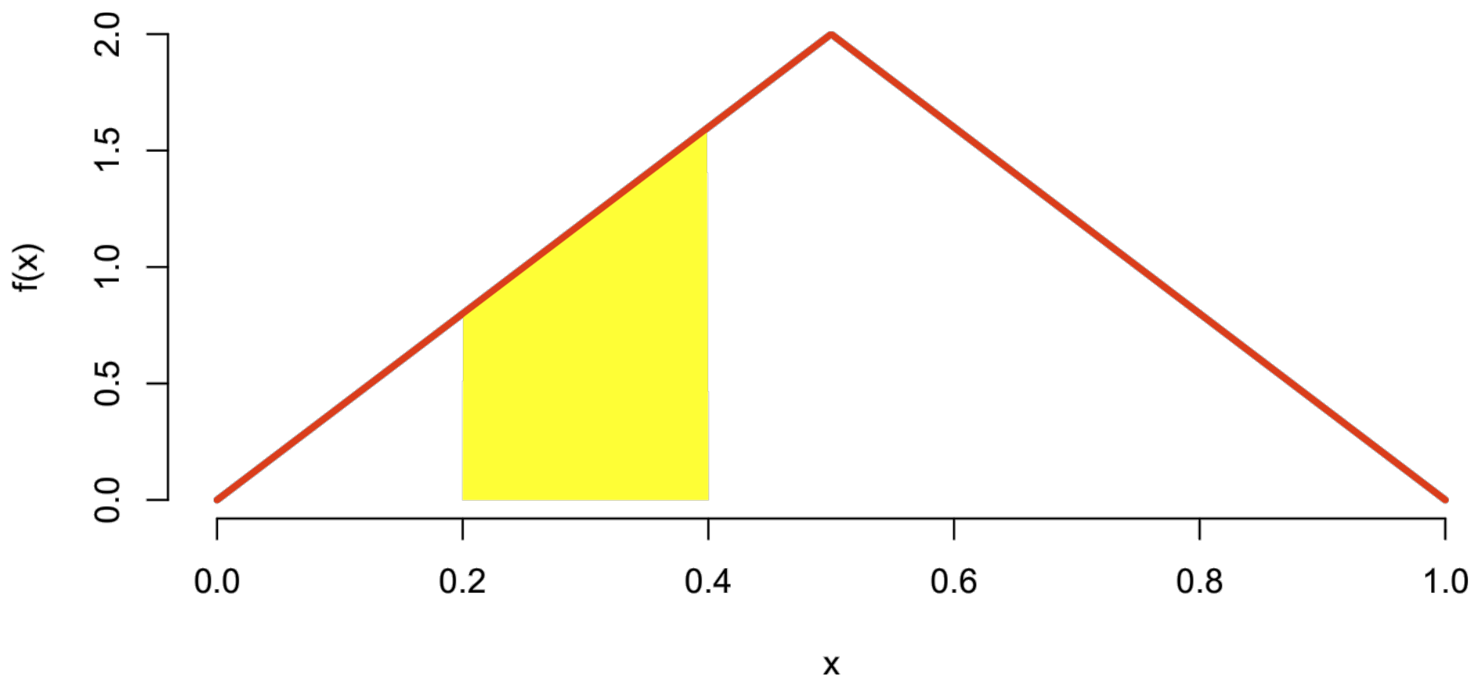
Example

Consider randomly selected points between 0 and 1/2 where any point is as likely as any other. What must be the height of the density?



Example: triangular distribution

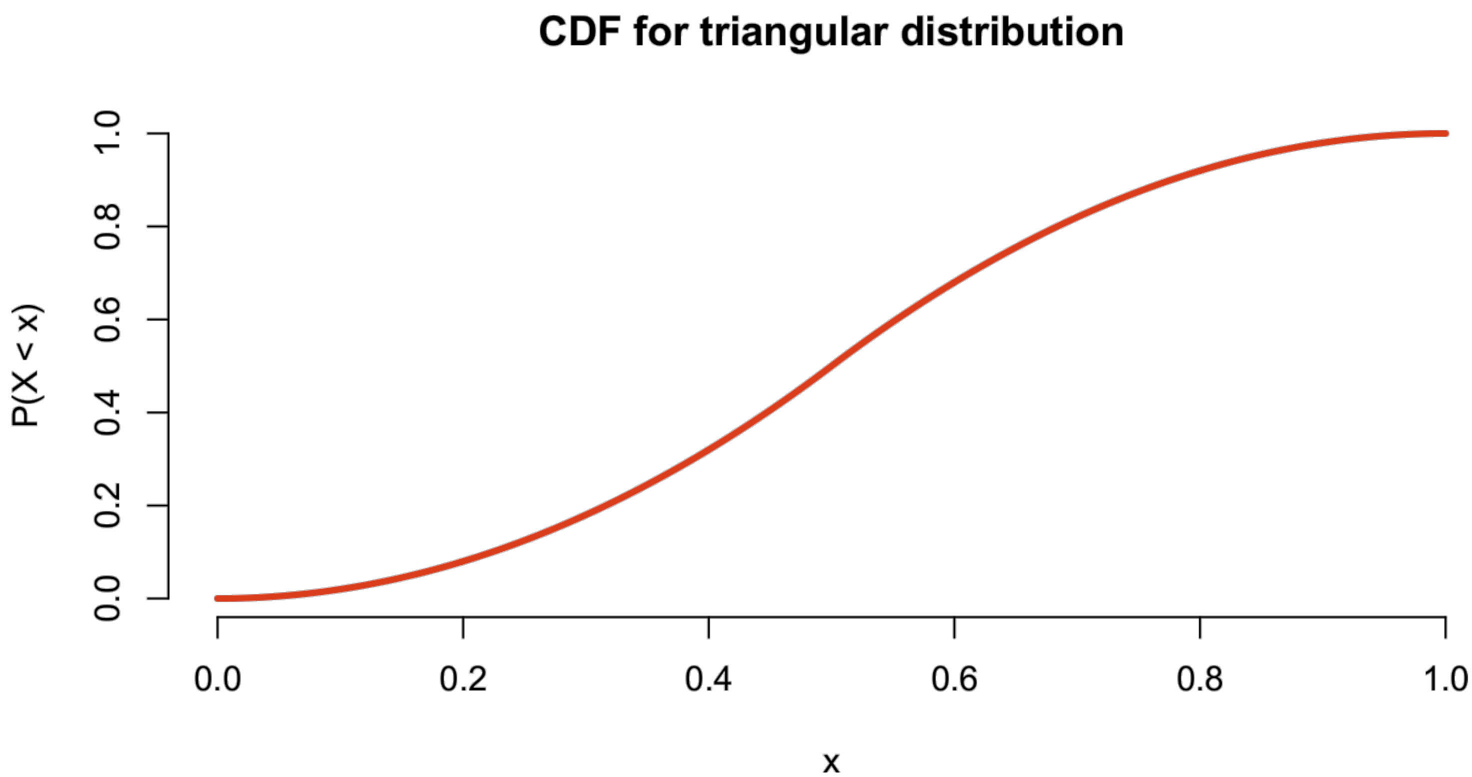
The shaded region here is the set $B = \{x \mid 0.2 \leq x \leq 0.4\}$.



What is $P(B)$?

Cumulative distribution function

Another way to visualize the distribution of a continuous random variable is to plot $P(X \leq x)$ across all possible x values.



The **CDF** of the triangular density from the previous slide is shown above.

Cumulative distribution function

$F(x) = P(X \leq x)$ describes the distribution of a random variable X if

- $0 \leq F(x) \leq 1$ for all x,
- $F(x)$ is increasing in x.

We can compute the probability of any interval via the expression:

$$P(a \leq X \leq b) = F(b) - F(a)$$

PDFs and CDFs

Probability density functions and cumulative distribution functions are two different ways to represent the distribution of a continuous random variable.

$$\text{PDF of triangular distribution} \qquad \text{CDF of triangular distribution}$$

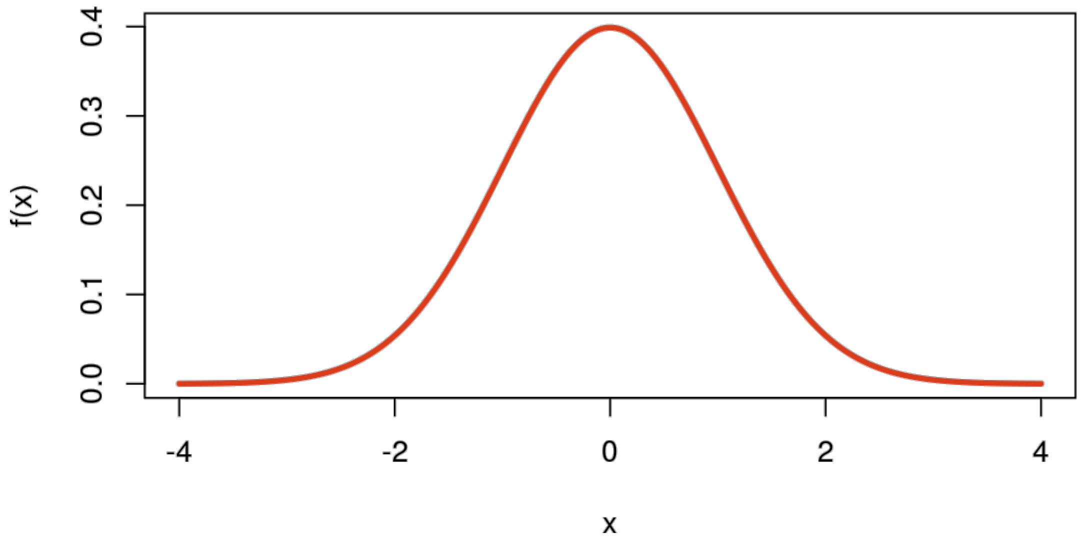
$$f(x) = \begin{cases} 4x & \text{if } x \le \frac{1}{2} \\ 4 - 4x & \text{if } x > \frac{1}{2} \end{cases} \quad F(x) = \begin{cases} 2 \cdot x^2 & \text{if } x \le \frac{1}{2} \\ 1 - 2(1-x)^2 & \text{if } x > \frac{1}{2} \end{cases}$$

PDFs are better for visualization while CDFs are better for determining specific probabilities.

More examples: https://mlakolar.shinyapps.io/distributionCalculator/

The Normal distribution

This the pdf of the **standard normal distribution**.



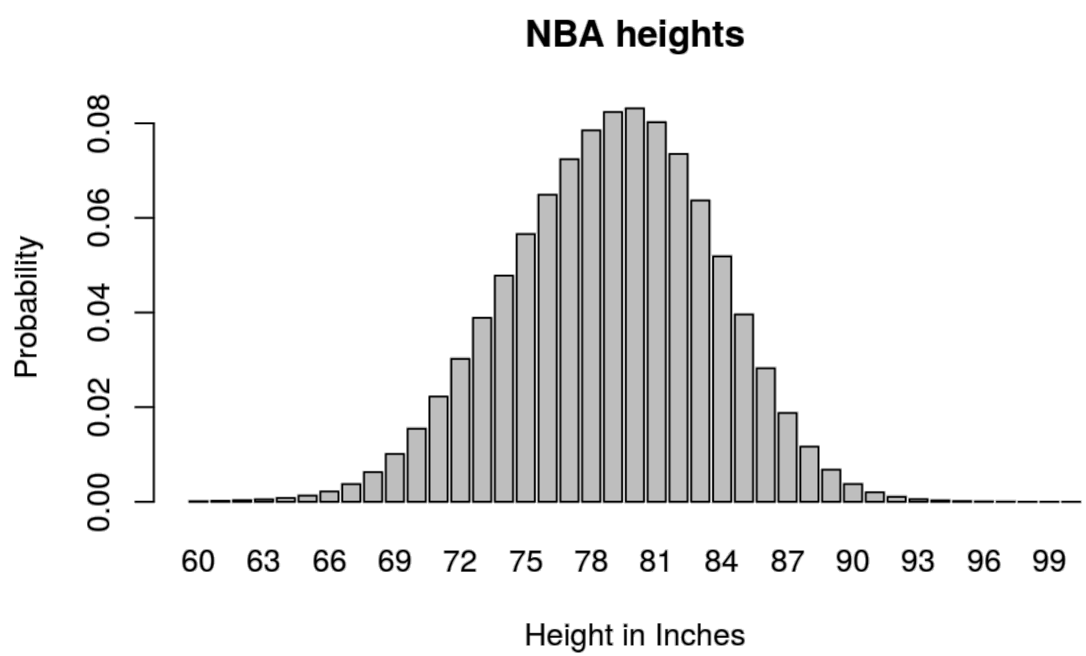
We often use the letter Z to denote the random variable which has this pdf.

The normal distribution has a "bell-shaped curve."

Any value in the interval $(-\infty, \infty)$ is "possible."

It is the "standard" normal because it has mean 0 and variance 1.

Example: NBA player heights



The Normal distribution

If X is a Normal random variable, we write:

$$X \sim N(\mu, \sigma^2)$$

We can obtain X as a linear function of $N(0,1)$ r.v.

$$X \sim N(\mu, \sigma^2) \equiv \mu + \sigma \cdot Z$$

where $Z \sim N(0,1)$ and μ, σ are constants.

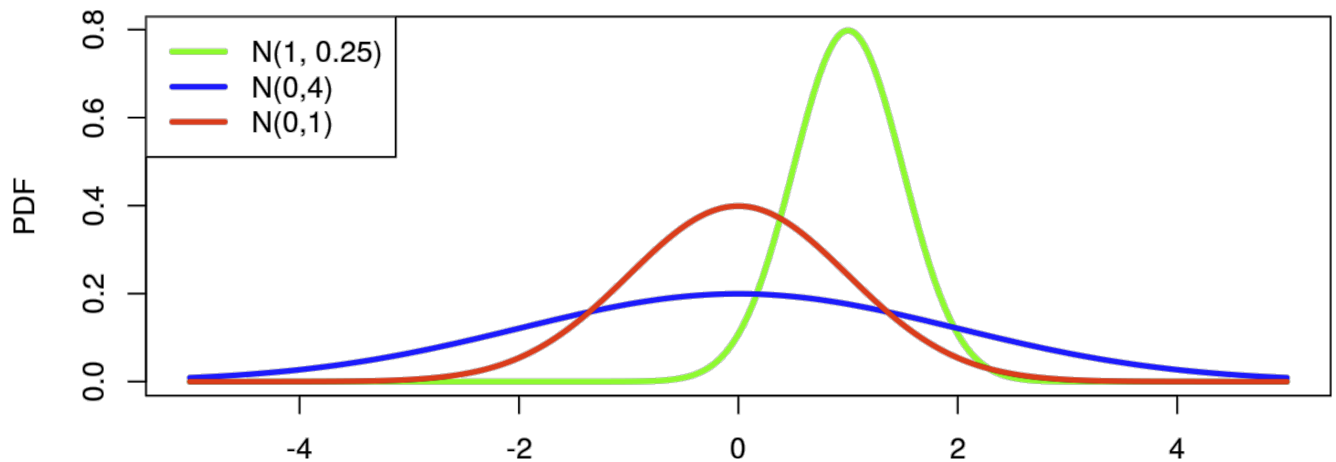
Properties of $N(\mu, \sigma^2)$

- Parameters: Expected values = μ, Variance = σ^2
- Symmetric and bell-shaped
- Any value in the interval $(-\infty, \infty)$ is "possible."
- $N(0,1)$ is the special case where $\mu = 0$, and $\sigma^2 = 1$

The Normal distribution is also known as the Gaussian distribution.

The normal family has two parameters.

- μ determines where the curve is centered.
- σ^2 determines how spread out the curve is



Interpreting the Normal Distribution

$$X \sim N(\mu, \sigma^2)$$

There is a 95% chance X is in the interval $\mu \pm 1.96\sigma$:

- μ describes what you think will happen on average;
- $\pm 1.96\sigma$ describes how wrong you could be.

μ determines the "location of the distribution.
σ determines the "spread."

$$Z \sim N(0, 1)$$

On average, outcomes will be close to 0.

$$P(-1 < Z < 1) = 0.68$$

$$P(-1.96 < Z < 1.96) = 0.95$$

$$P(-3 < Z < 3) = 0.9974$$

Where does the Empirical Rule come from?

Empirical Rule

- Approximately 68% of observations within $\mu \pm \sigma$
- Approximately 95% of observations within $\mu \pm 2\sigma$
- Nearly all observations within $\mu \pm 3\sigma$

We are just using a normal distribution as an approximation.

R Commands for the Normal RV

Let $X \sim N(\mu, \sigma^2)$

PDF $f(x)$ of X: `dnorm(x, mean =`μ`, sd =`σ`)`

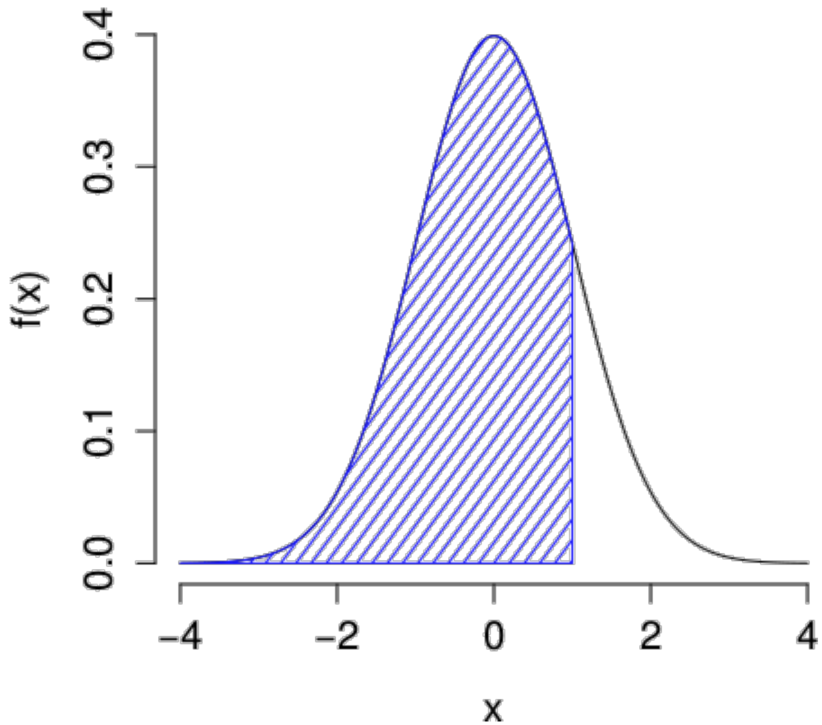
- Mnemonic: `d` $=$ density, `norm` $=$ normal

CDF $F(x)$ of X: `pnorm(x, mean =`μ`, sd =`σ`)`

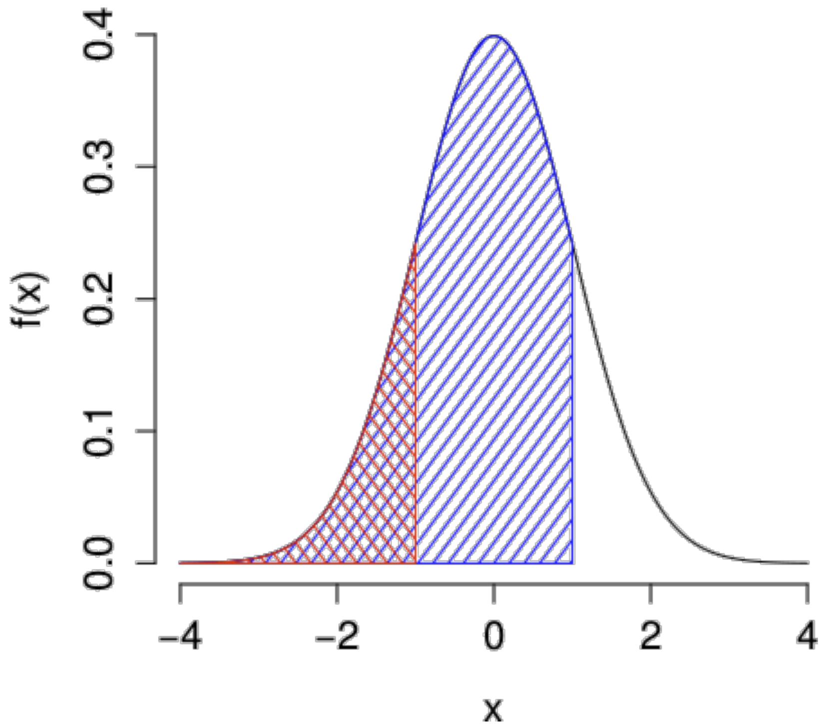
- Mnemonic: `p` $=$ probability, `norm` $=$ normal

Generate a random number from $N(\mu, \sigma^2)$: `rnorm(x, mean =`μ`, sd =`σ`)`

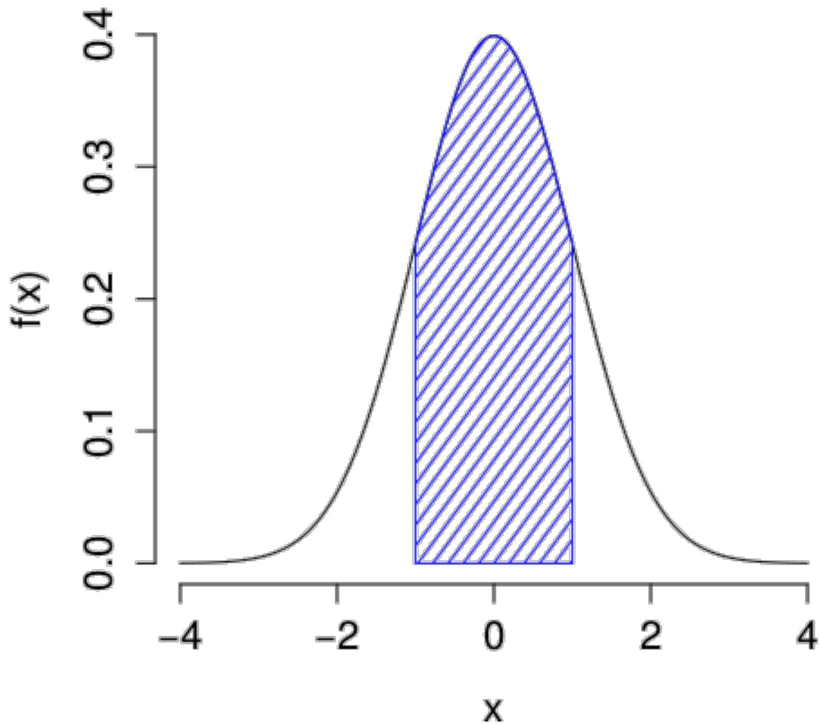
- Mnemonic: `r` $=$ random, `norm` $=$ normal `pnorm(1)` ≈ 0.84



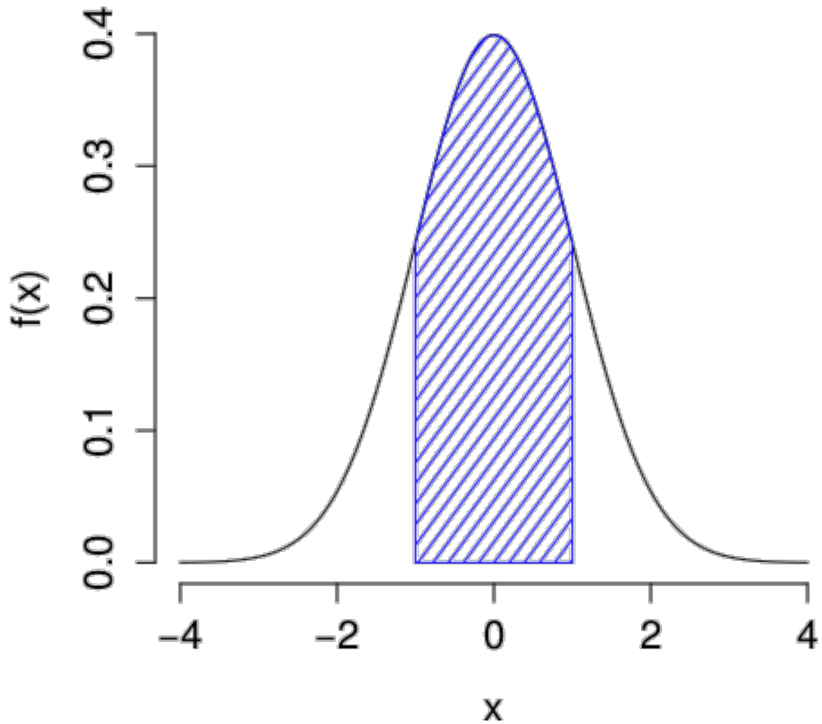
`pnorm(1) - pnorm(-1)` $\approx 0.84 - 0.16$



`pnorm(1) - pnorm(-1)` ≈ 0.68



Middle 68% of $N(0,1) \rightarrow$ approx. $(-1,1)$



Suppose $X \sim N(0, 1)$

$$\begin{aligned} P(-1 \leq X \leq 1) &= \texttt{pnorm(1) - pnorm(-1)} \\ &\approx 0.683 \end{aligned}$$

$$\begin{aligned} P(-2 \leq X \leq 2) &= \texttt{pnorm(2) - pnorm(-2)} \\ &\approx 0.954 \end{aligned}$$

$$\begin{aligned} P(-3 \leq X \leq 3) &= \texttt{pnorm(3) - pnorm(-3)} \\ &\approx 0.997 \end{aligned}$$

Linear Function of Normal RV is a Normal RV

Suppose that $X \sim N(\mu, \sigma^2)$.
Then if a and b constants,

$$a + bX \sim N(a + b\mu, b^2\sigma^2)$$

For **any** r.v. X,

$$E[a + bX] = a + bE[X]$$

and

$$\text{Var}[a + bX] = b^2\text{Var}[X].$$

Key point: linear transformation of normal is still normal!

For example, linear transformation of Binomial is **not** Binomial!

Example

Suppose $X \sim N(\mu, \sigma^2)$ and let $Z = (X - \mu)/\sigma$.
What is the distribution of Z?

1. $N(\mu, \sigma^2)$
2. $N(\mu, \sigma)$
3. $N(0, \sigma^2)$
4. $N(0, \sigma)$
5. $N(0, 1)$

Probability of an Interval: $X \sim N(\mu, \sigma^2)$

$$
\begin{aligned}
P(a \leq X \leq b) &= F_X(b) - F_X(a) \\
&= P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\
&= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\
&= F_Z\left(\frac{b-\mu}{\sigma}\right) - F_Z\left(\frac{a-\mu}{\sigma}\right) \\
&= \mathtt{pnorm}((b-\mu)/\sigma) - \mathtt{pnorm}((a-\mu)/\sigma)
\end{aligned}
$$

Where Z is a standard normal random variable.

$$
P(a \leq X \leq b) = \mathtt{pnorm}(b, \mu, \sigma) - \mathtt{pnorm}(a, \mu, \sigma)
$$

Suppose $X \sim N(\mu, \sigma^2)$

What is $P(\mu - \sigma \leq X \leq \mu + \sigma)$?

Suppose $X \sim N(\mu, \sigma^2)$

What is $P(\mu - \sigma \leq X \leq \mu + \sigma)$?

$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &= P\left(-1 \leq \frac{X - \mu}{\sigma} \leq 1\right) \\ &= P\left(-1 \leq Z \leq 1\right) \\ &= \texttt{pnorm(1)} - \texttt{pnorm(-1)} \\ &\approx 0.68 \end{aligned}$$

Linear Combinations of Multiple Independent Normals

Let $X \sim N(\mu_x, \sigma_x^2)$ independent of $Y \sim N(\mu_y, \sigma_y^2)$. Then if a, b, c are constants:

$$aX + bY + c \sim N(a\mu_x + b\mu_y + c, a^2\sigma_x^2 + b^2\sigma_y^2)$$

- Result assumes independence
- Particular to Normal RV
- Extends to more than two Normal RVs

Example

Suppose $X_1, X_2, \sim \text{iid } N(\mu, \sigma^2)$

Let $\bar{X} = (X_1 + X_2)/2$. What is the distribution of $\bar{X}$?

1. $N(\mu, \sigma^2/2)$
2. $N(0, 1)$
3. $N(\mu, \sigma^2)$
4. $N(\mu, 2\sigma^2)$
5. $N(2\mu, 2\sigma^2)$

IID Normal Model

We want to use a Normal distribution to model data in the real world.

Surprisingly often, data looks like i.i.d. draws from a Normal distribution.

We can have i.i.d. draws from any distribution. When we write

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2) \quad \text{i.i.d.}$$

we mean that each random variable X_i for $i = 1, \ldots, n$ will be an independents draw from the same Normal distribution.

We have not formally defined independence for continuous distributions, but our intuition is the same as for discrete random variables.

Each draw has no effect on the others, and the same normal distribution describes what we think each variable will turn out to be.

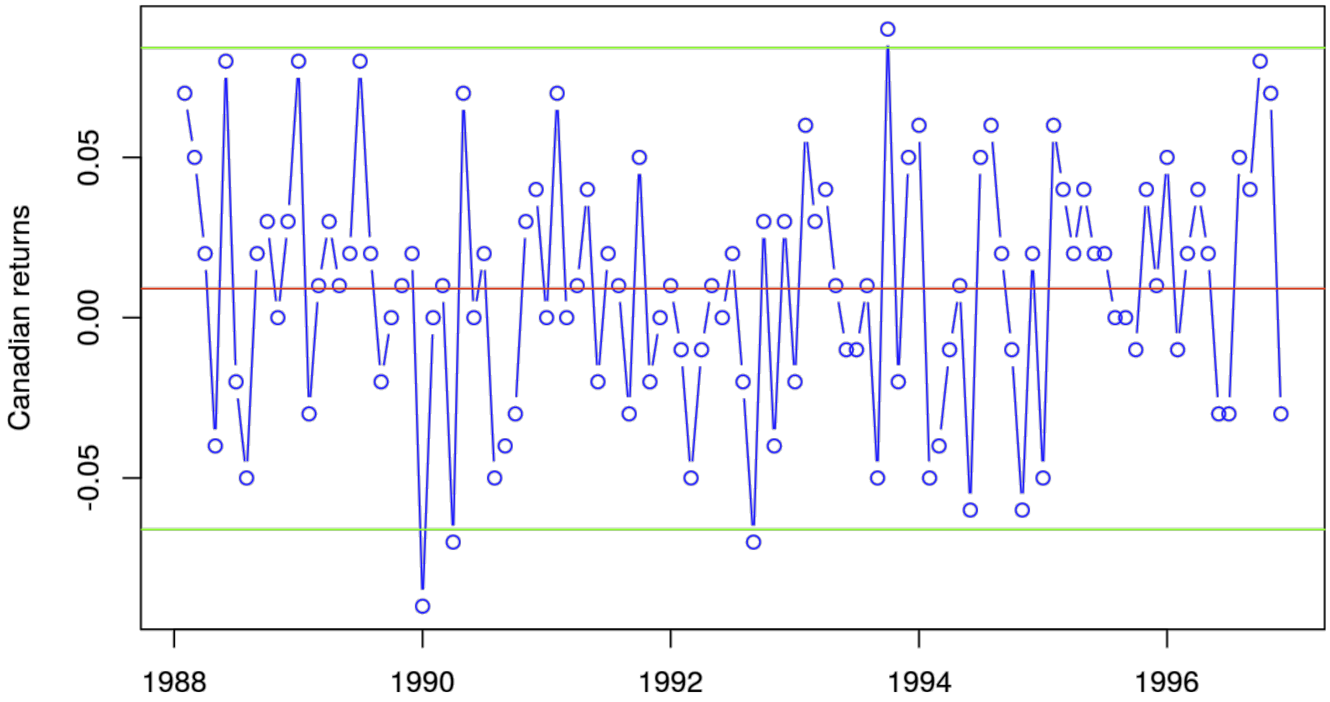
What do i.i.d. Normal draws look like?

See: https://mlakolar.shinyapps.io/RV_distributions/

We check if i.i.d. Normal model is appropriate we:

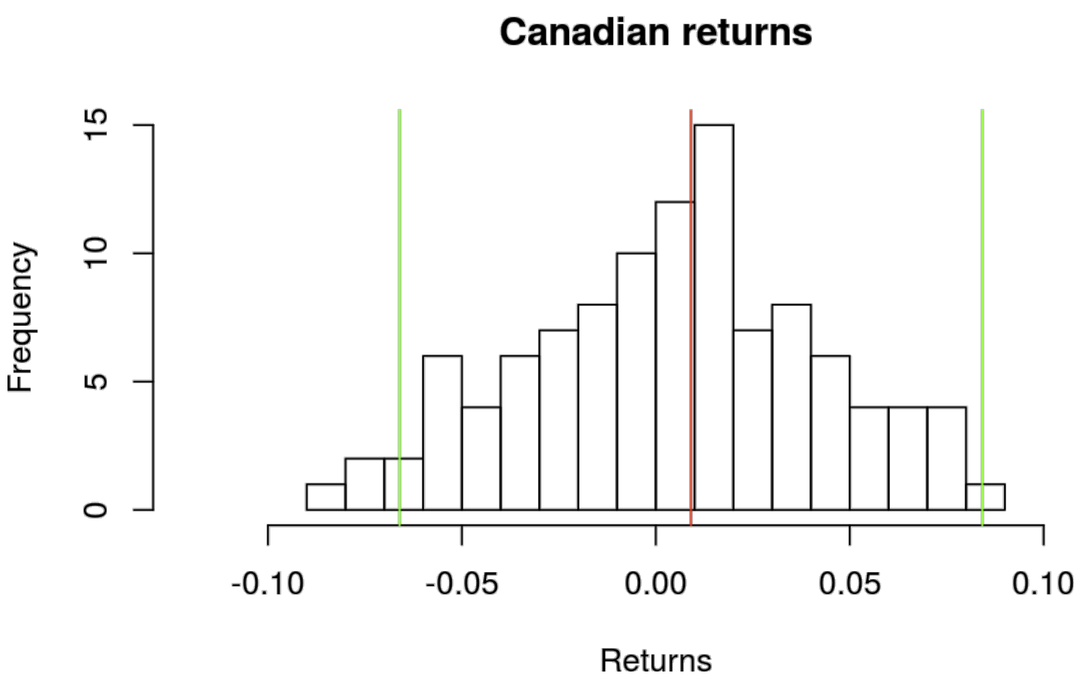
1. Do the data look i.i.d.? — use time series plot and check that there are no obvious patterns
2. Do the data look Normally distribution? — plot histogram and see if it looks like a Normal pdf

Example: Country returns



Do the data look i.i.d.?

By looking at a time series plot, are there obvious patterns.

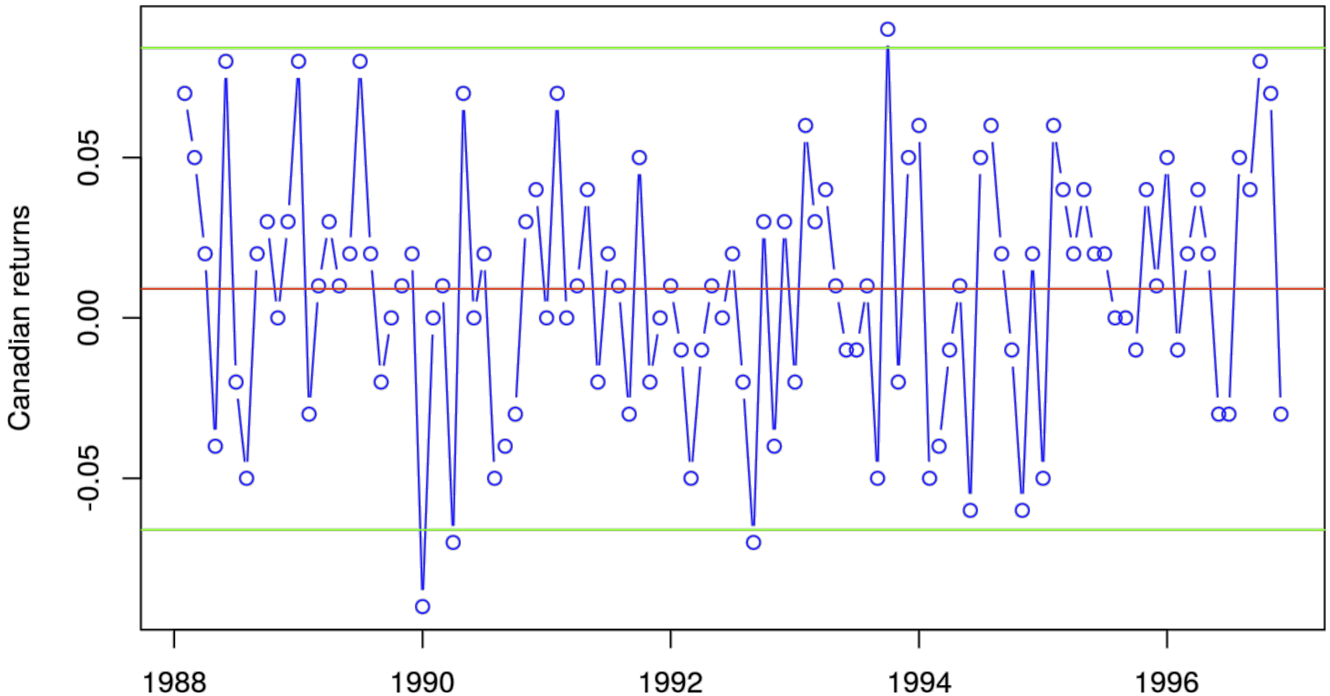


By looking at a histogram, do data appear to be approximately normally distributed?

We squint our eyes and say the returns look i.i.d. Normal (at least as a first approximation).

The sample mean and standard deviation are 0.00907 and 0.03833.

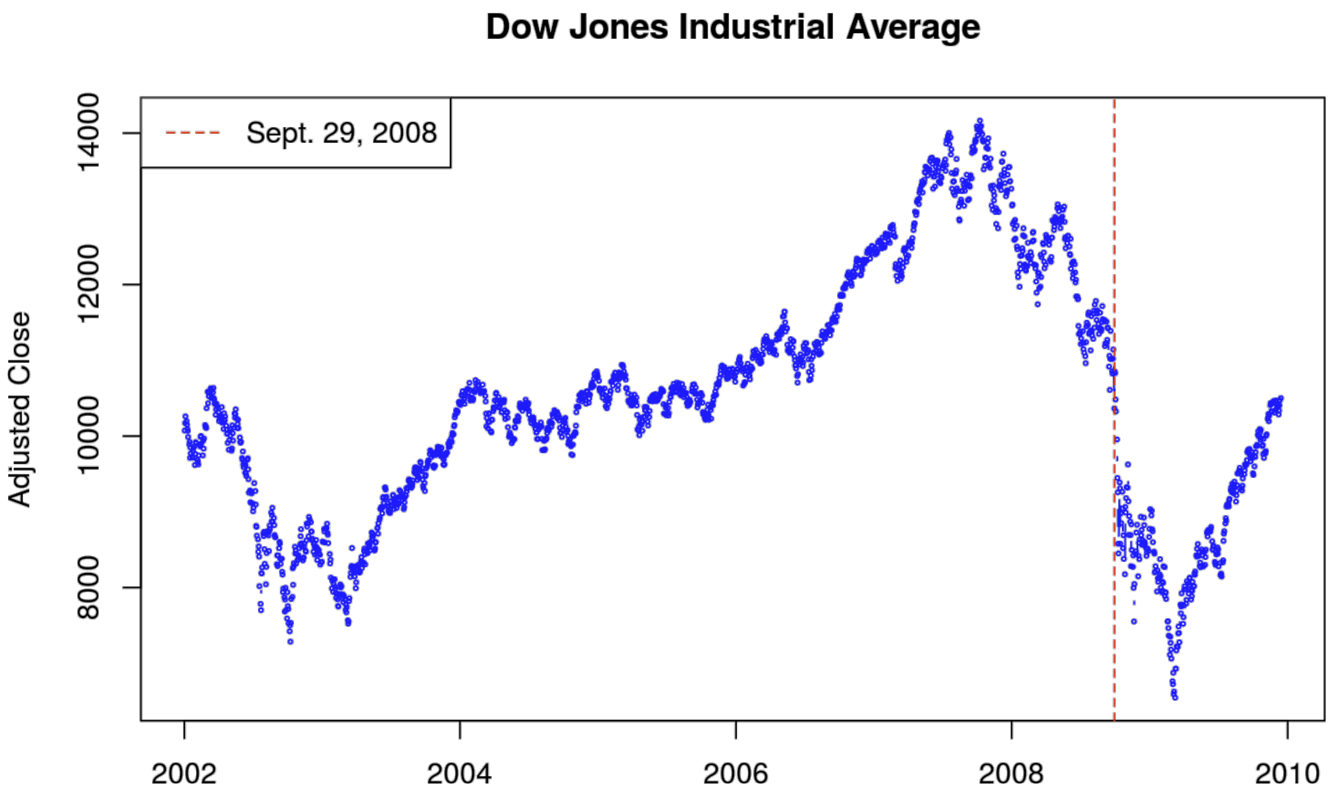
Based on this, our model is $R_t \sim N(0.01, 0.04^2)$.



Our prediction for the next return is 0.01.

Our 95% interval is for the next return is $(-0.07, 0.09)$

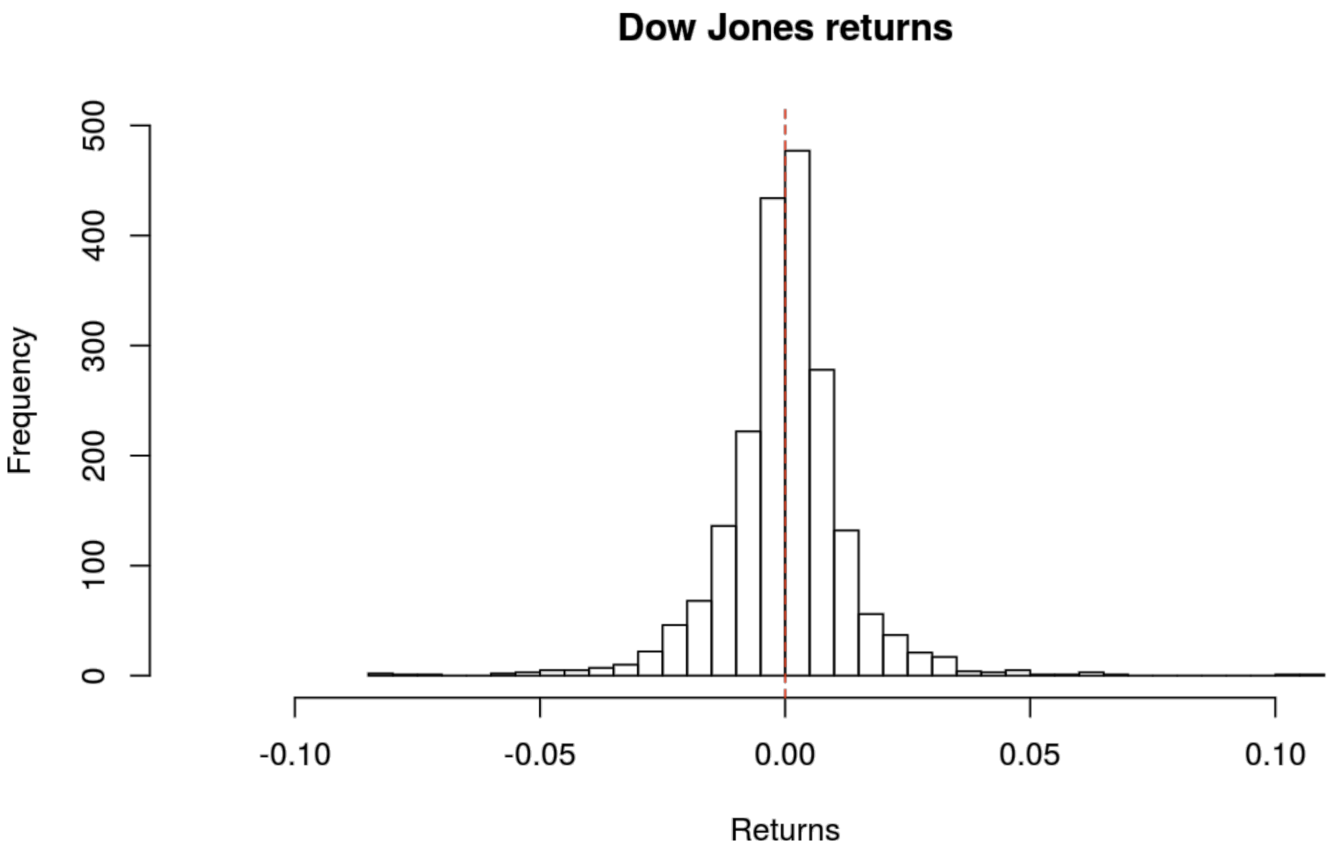
Example: Dow Jones returns



We can often "transform" data such that the i.i.d. Normal model is a better approximation.

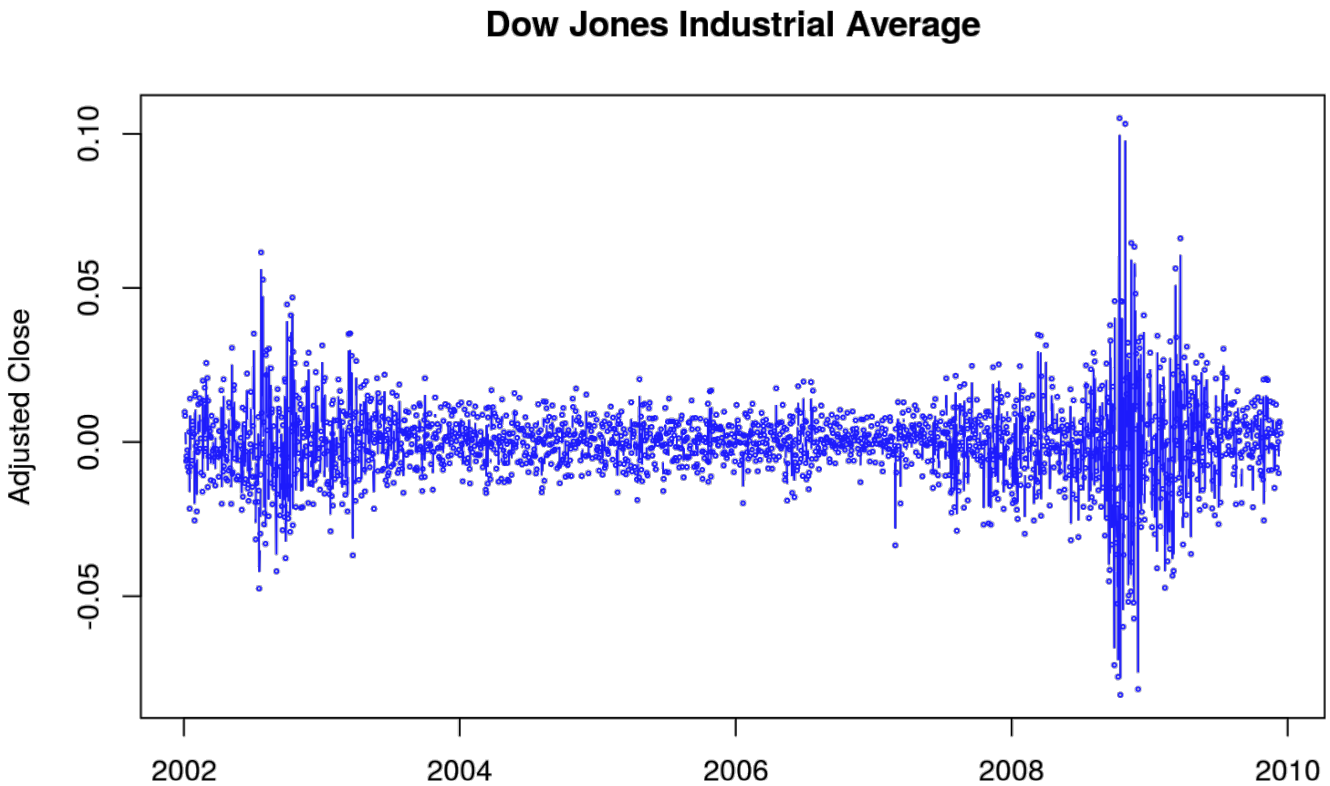
We can transform Dow Jones Industrial Average levels X_t to returns

$$R_t = \log(X_t) - \log(X_{t-1})$$



The normal approximation appears to be better.

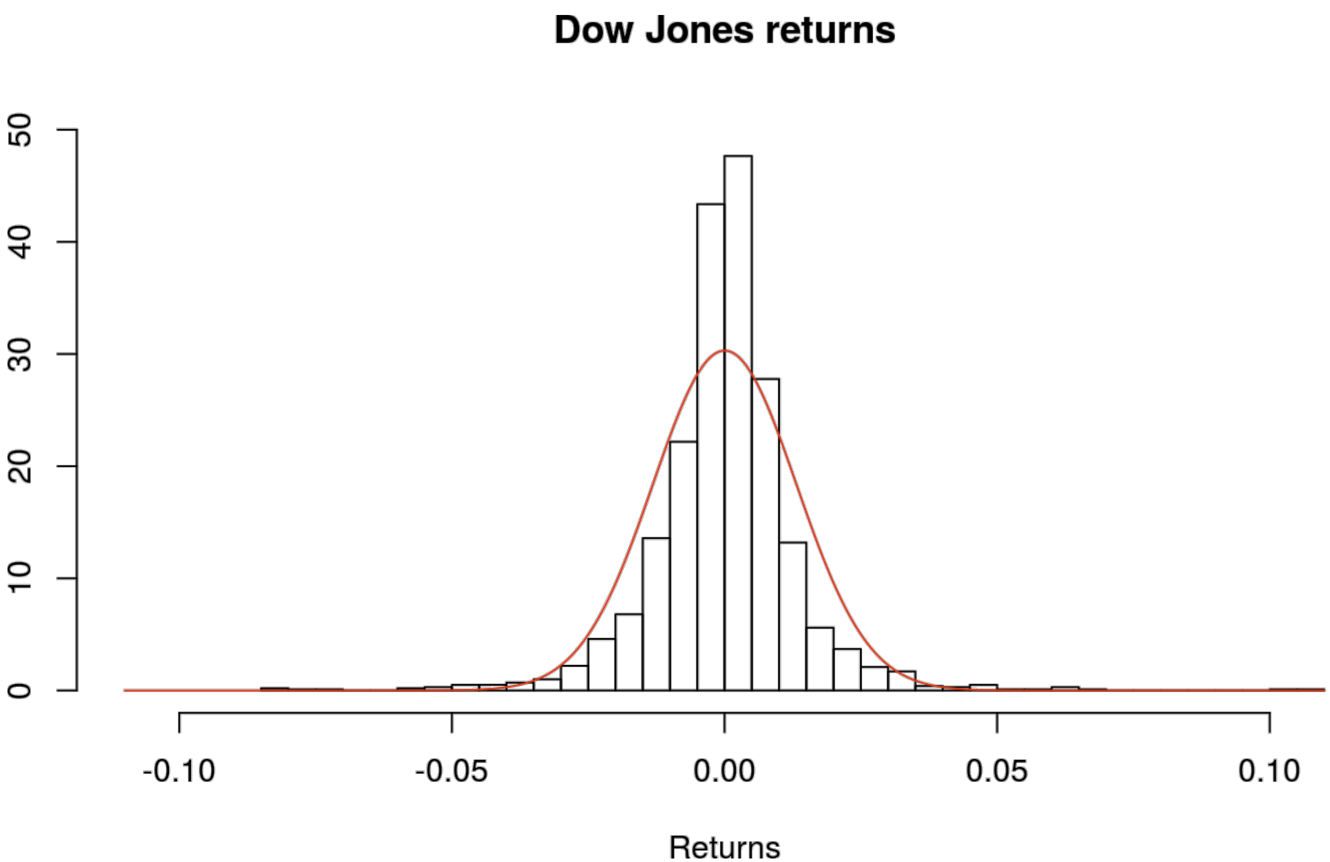
But, appearances can still be deceiving.



Are the returns really i.i.d.?

Does it look like they have the same variance σ^2?

Here is the histogram with the best fitting Normal distribution.



$$R_t \sim N\left(2.076 \cdot 10^{-5}, 1.733 \cdot 10^{-4}\right)$$

Return on September 29, 2008 was -8.2%.

Using the normal model, $R_t \sim N\left(2.076 \cdot 10^{-5}, 1.733 \cdot 10^{-4}\right)$, as an approximation, how unusual was the crash?

We will standardize R_t to answer that question.

$$Z_t = \frac{R_t - E(R_t)}{\sqrt{\mathrm{Var}(R_t)}} = \frac{R_t - 2.076 \cdot 10^{-5}}{0.0132}$$

We calculate the observed z-value

$$Z_t = \frac{-0.082 - 2.076 \cdot 10^{-5}}{0.0132} = -6.23$$

This is a 6 sigma event!

Standardization

For $X \sim N(\mu, \sigma^2)$, the z-value corresponding to a value x is

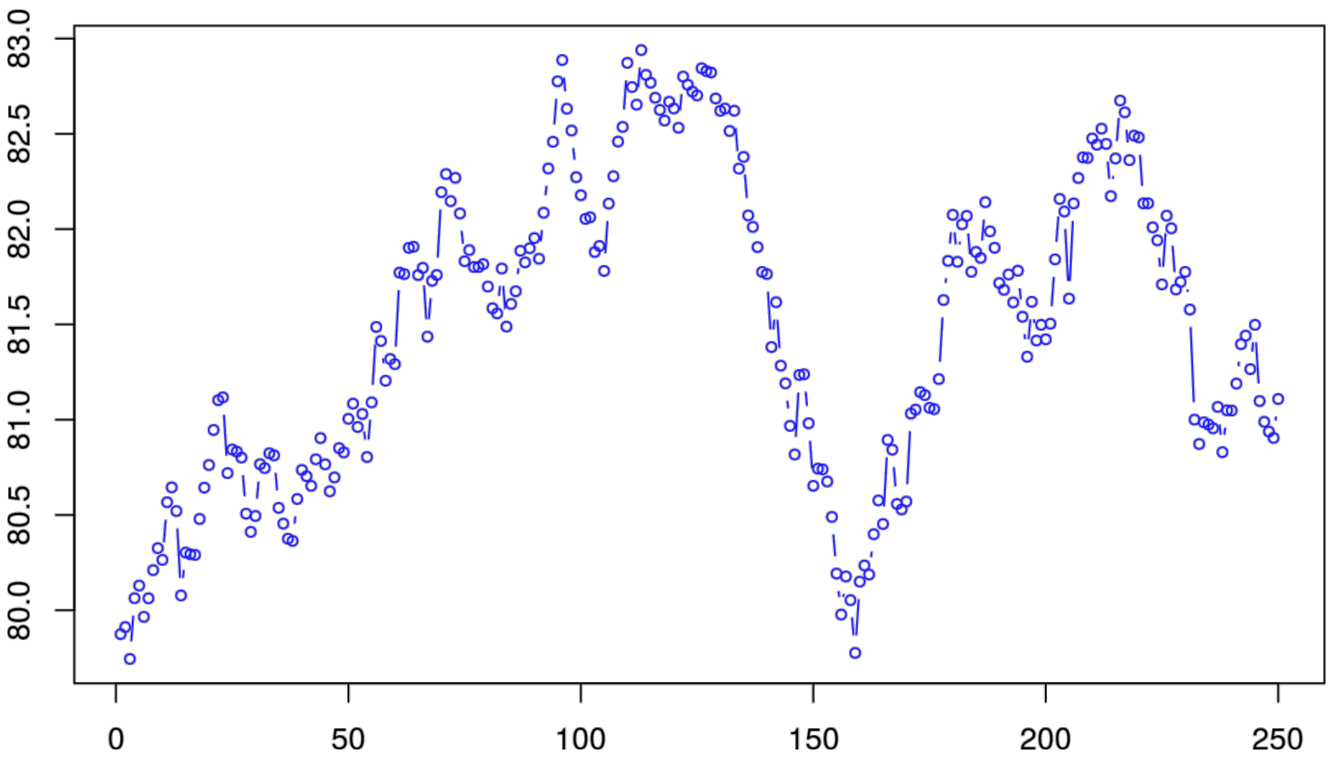
$$z = \frac{x - \mu}{\sigma}.$$

Any time someone says "z-value" or "z-score," they are just talking about how many standard deviations we are away from the mean under a bell curve.

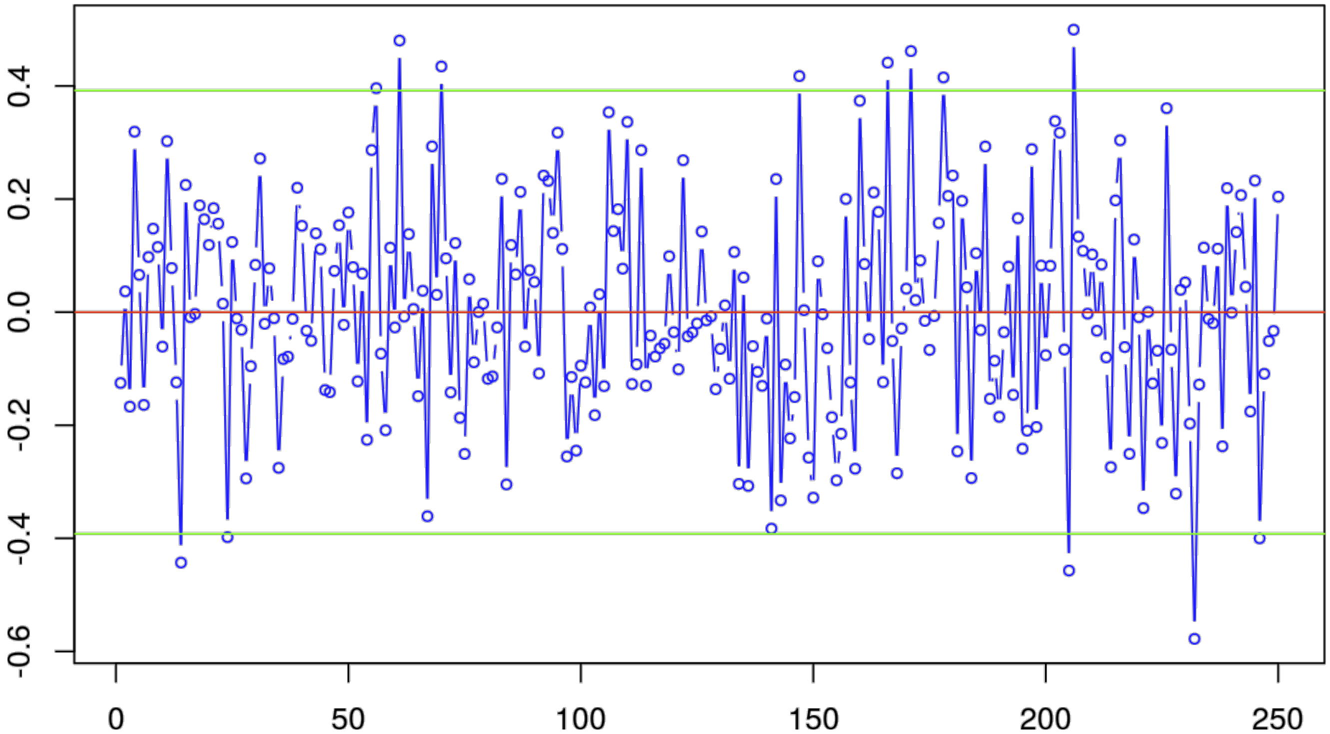
The Random Walk Model

$$P_t = P_{t-1} + D_t$$

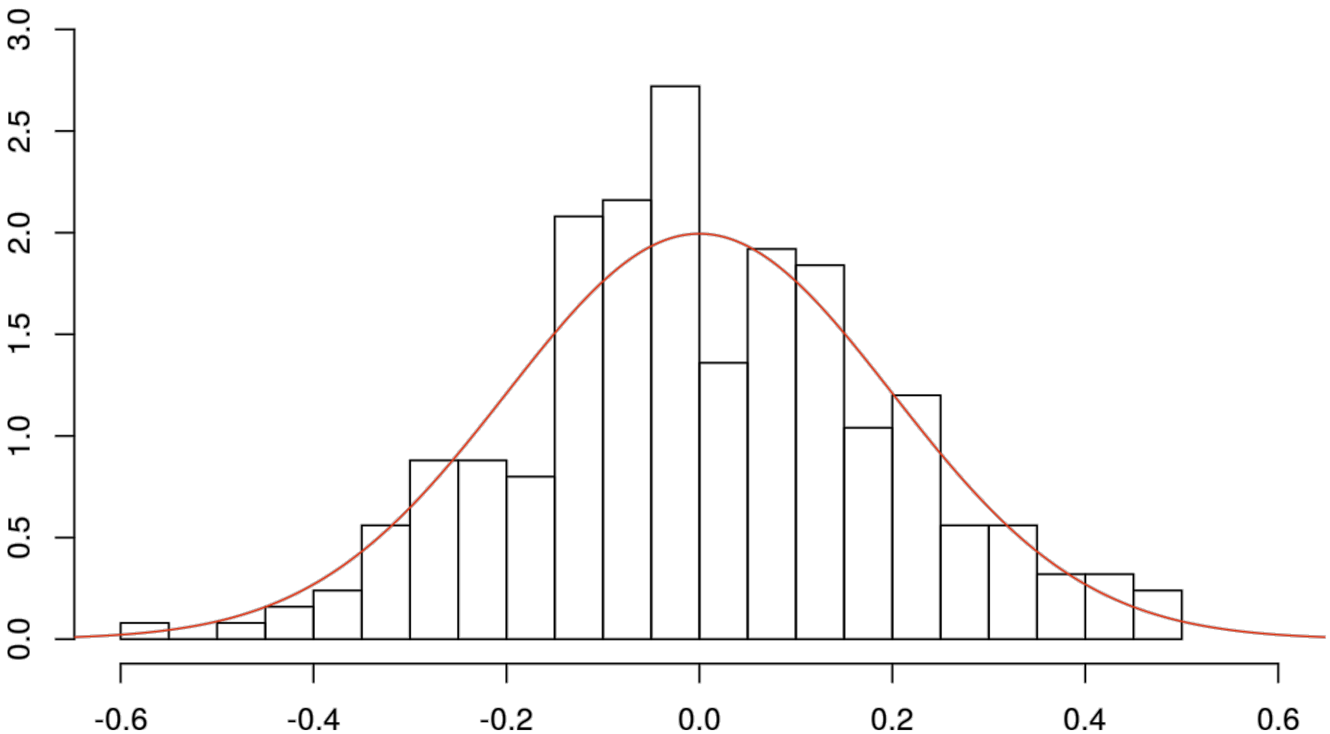
D_t can have a continuous distribution. Consider $D_t \sim N(0, 0.04)$ to be a sequence of i.i.d. Normal random variables for $t = 1, 2, \ldots$



Differences



Differences



$$P_t = P_{t-1} + D_t, \qquad D_t \sim N(0, 0.04), \quad \text{i.i.d.}$$

Given we know the price today $P_t = p_t$, what is the expected value of the price tomorrow based on this model?

What is the conditional distribution of $P_{t+1} \mid P_t = p_t$?

What is the conditional distribution of $P_{t+2} \mid P_t = p_t$?

The Central Limit Theorem

The **central limit theorem** (CLT) says that the average of a large number of independent random variables is (approximately) Normally distributed.

Another way of saying this is:

Suppose that $X_1, X_2, \cdots, X_n$ are independent random variables and let

$$Y = \frac{1}{n} \sum_i X_i$$

For large n, we have that

$$Y \approx N(\mu_Y, \sigma_Y^2)$$

where

$$\mu_Y = \frac{1}{n} \sum_{i=1}^{n} E(X_i)$$
$$\sigma_Y^2 = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i)$$

The Central Limit Theorem

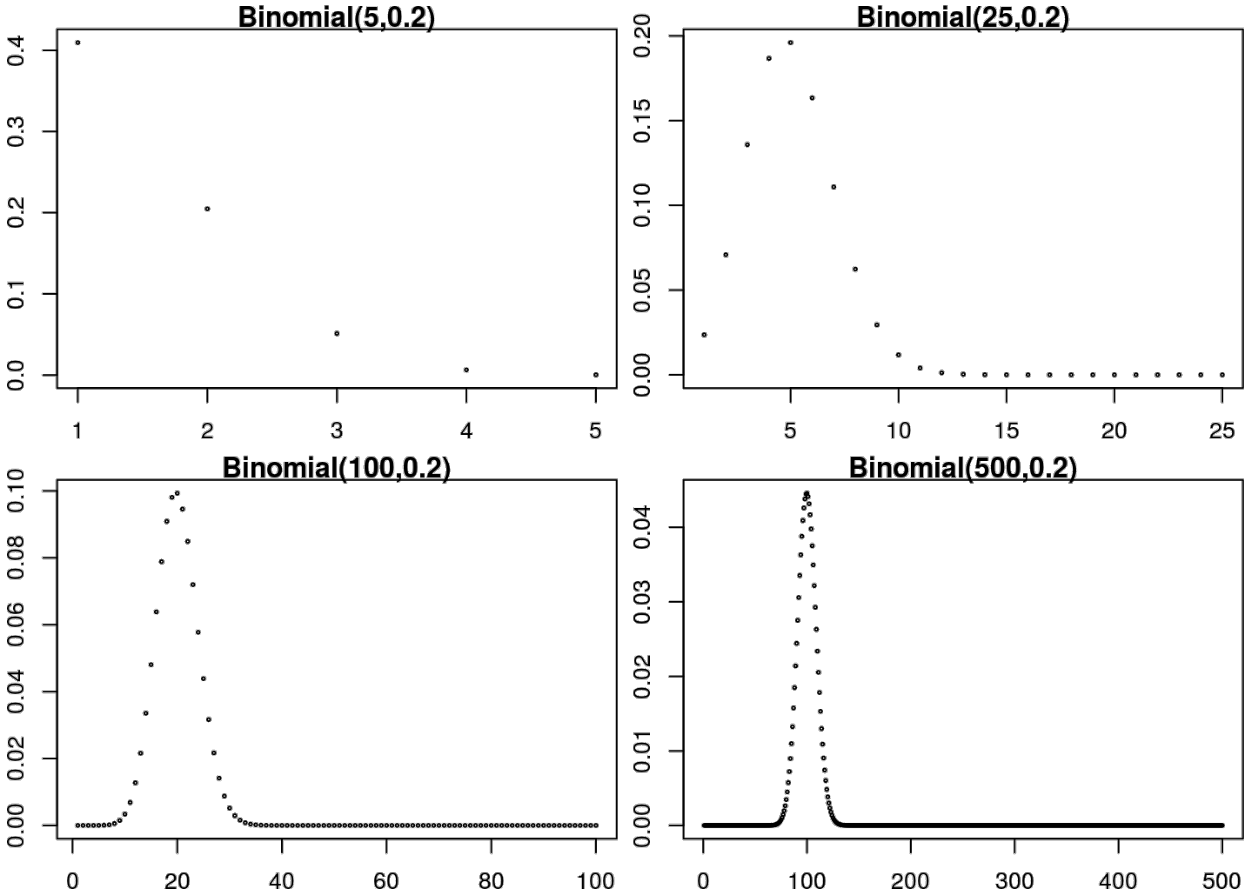
This is really amazing. Tell me why? I want to learn more.

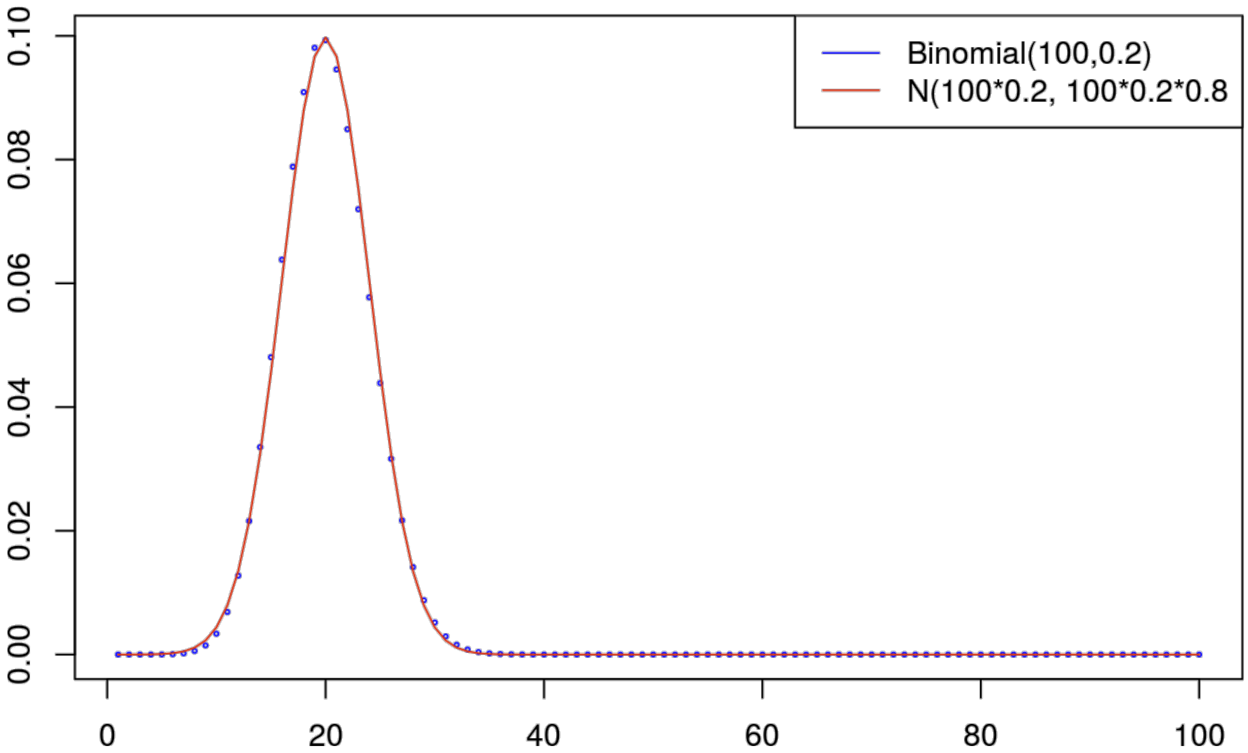
Notice that we only assume that r.v.s X_i are independent, we **DID NOT** say what distribution they have. They may have **different** distributions.

The average of a large number of independent random variables is (approximately) normally distributed, **no matter what distribution the individual random variables have**!

Example: The binomial distribution

$$Y = \sum_{i=1}^{n} X_i, \qquad X_i \sim \text{Bernoulli}(p), \quad \text{i.i.d.}$$





As we increase n, the distribution of Y gets closer and closer to a normal distribution with the same mean and variance as the binomial.

How good is the approximation?

Your company is about to manufacture 100 parts.

Suppose defects are i.i.d. $X_i \sim \mathrm{Bernoulli}(0.1)$.

Let $Y = \sum_{i=1}^{n} X_i$ be the number of defects. $Y \sim \mathrm{Binomial}(100, 0.1)$

$$E[Y] = np = 10 \qquad V[Y] = np(1-p) = 9$$

Even though $Y \sim \mathrm{Binomial}(100, 0.1)$, let us use the normal approximation, first.

Let the Normal distribution have the same mean and variance. $Y \approx N(10, 9)$

Based on the normal approximation, there is a 95% chance that the number of defects is in the interval:

$$\mu \pm 2\sigma_Y = 10 \pm 6 = (4, 16).$$

We can compare this to the exact answer based on binomial probabilities.

$$P(4 < Y < 16) = F(16) - F(4) = 0.9794 - 0.0237 = 0.9557$$

Approximation is pretty good!