STA 371G: Statistics and Modeling

Review of Basic Probability and Statistics: Normal and Binomial Distributions

Mingyuan Zhou
McCombs School of Business
The University of Texas at Austin

http://mingyuanzhou.github.io/teaching

Continuous Random Variables

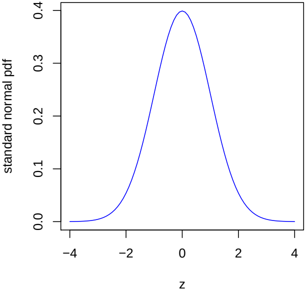
- Suppose we are trying to predict tomorrow's return on the S&P500...
- Question: What is the random variable of interest?
- Question: How can we describe our uncertainty about tomorrow's outcome?
- Listing all possible values seems like a crazy task... we'll work with intervals instead.
- These are called continuous random variables.
- The probability of an interval is defined by the area under the probability density function.

Discrete and Continuous Random Variables (optional)

- Discrete random variables
  - A finite (or countably infinite) number of possible outcomes
  - Each possible value is associated with a probability
  - Bernoulli/binomial/multinomial ...
  - Poisson/negative binomial ...
- Continuous random variables
  - An uncountably infinite number of outcomes
  - Probability density function (PDF) $f_X(x)$
  - The probability that X falls in some interval: $P(a \leq X \leq b) = \int_a^b f_X(x)dx$
  - $P(X = x) = 0$
  - Uniform distribution
  - Normal distribution $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $\int_{-\infty}^{\infty} f_X(x) = 1$
  - Gamma/beta/exponential ....

Normal Distribution

- A random variable is a number we are NOT sure about but we might have some idea of how to describe its potential outcomes. The Normal distribution is the most used probability distribution to describe a random variable.
- The probability the number ends up in an interval is given by the area under the curve (pdf), which is always symmetric, unimodal and bell shaped.

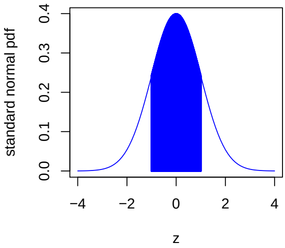


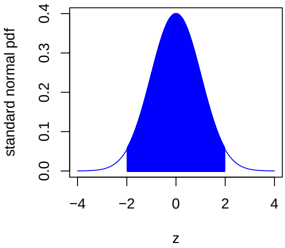
Normal Distribution

- The standard Normal distribution has mean 0 and variance 1. Its probability density function is $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$
- Notation: If $Z \sim \mathcal{N}(0,1)$ (Z is the random variable)

$$Pr(-1 < Z < 1) = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0.68$$

$$Pr(-1.96 < Z < 1.96) = \int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0.95$$





Normal Distribution

Note:

For simplicity we will often use $P(-2 < Z < 2) \approx 0.95$

Questions:

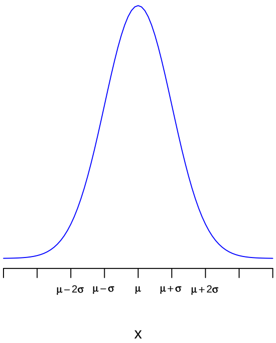
- What is $Pr(Z < 2)$?
- What is $Pr(Z = 2)$?
- How about $Pr(Z \leq 2)$?
- What is $Pr(Z < 0)$?

Normal Distribution

- The standard normal is not that useful by itself. When we say "the normal distribution", we really mean a family of distributions.
- We obtain pdfs in the normal family by shifting the bell curve around and spreading it out (or tightening it up).

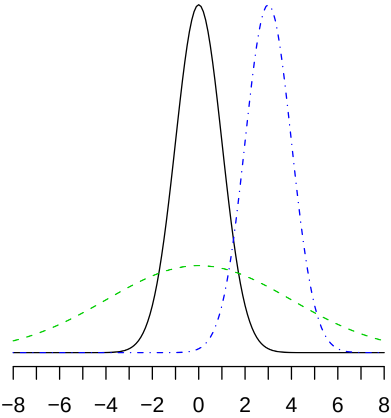
Normal Distribution

- $X \sim \mathcal{N}(\mu, \sigma^2)$: normal distribution with mean μ and variance σ^2. Its PDF is $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- The parameter μ determines the center of the curve.
- The parameter σ determines how spread out the curve is. The area under the curve in the interval $(\mu - 2\sigma, \mu + 2\sigma)$ is 95%. $Pr(\mu - 2\,\sigma < X < \mu + 2\,\sigma) \approx 0.95$



Normal Distribution

- Example: Below are the pdfs of $X_1 \sim \mathcal{N}(0, 1)$, $X_2 \sim \mathcal{N}(3, 1)$, and $X_3 \sim \mathcal{N}(0, 16)$.
- Which pdf goes with which X?

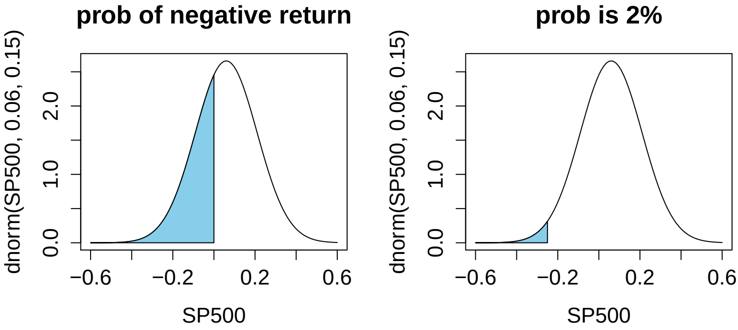


Normal Distribution – Example

- Assume the annual returns on the SP500 are normally distributed with mean 6% and standard deviation 15%. SP500 $\sim \mathcal{N}(0.06, (0.15)^2)$.
- Two questions: (i) What is the chance of losing money on a given year? (ii) What is the value that there's only a 2% chance of losing that or more?

- (i) $Pr(SP500 < 0) = ?$
- (ii) $Pr(SP500 < ?) = 0.02$

Normal Distribution – Example



- (i) $Pr(SP500 < 0) =$ 0.34 and (ii) $Pr(SP500 < -0.25) = 0.02$
- In Excel2011: NORMDIST(0,0.06,0.15,TRUE) and NORMINV(0.02,0.06,0.15)
- In R: pnorm(0,0.06,0.15) and qnorm(0.02,0.06,0.15)

Normal Distribution

1. Note: In

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

   μ is the mean and σ^2 is the variance.

2. Standardization: if $X \sim \mathcal{N}(\mu, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

3. Summary:

   $X \sim \mathcal{N}(\mu, \sigma^2)$:

   μ: where the center of the curve is

   σ: how spread out the curve is

   95% chance $X \in \mu \pm 2\sigma$.

Normal Distribution – Another Example

Prior to the 1987 crash, monthly S&P500 returns (r) followed (approximately) a normal with mean 0.012 and standard deviation equal to 0.043. How extreme was the crash of -0.2176 under the normal assumption? The standardization helps us interpret these numbers...

$$r \sim \mathcal{N}(0.012, (0.043)^2), \quad Z = \frac{r - 0.012}{0.043} \sim \mathcal{N}(0, 1)$$

For the crash,

$$z = \frac{-0.2176 - 0.012}{0.043} = -5.27$$

How extreme is this z value? 5 standard deviations away!!

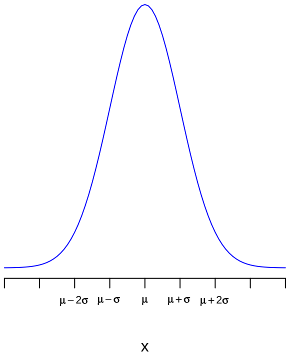
In R: $pnorm(-0.2176, 0.012, 0.043) = 4.66 \times 10^{-8}$, i.e.,

$P(r \leq -0.2176) = P(Z \leq -5.27) = 4.66 \times 10^{-8}$.

In Excel2011: NORMDIST(-0.2176, 0.012, 0.043,TRUE) or NORMSDIST((-0.2176- 0.012)/0.043,TRUE)

Mean and Variance of Random Variables

- Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$.
- Suppose someone asks you for a prediction of X. What would you say?



- Suppose someone asks you how sure you are. What would you say?

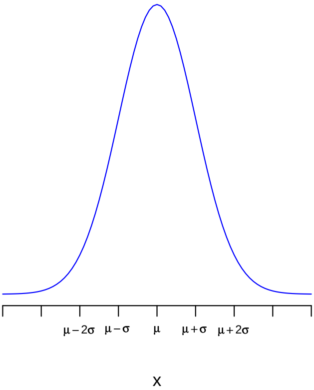
Mean and Variance of Random Variables

- For the normal family of distributions we can see that the parameter μ talks about "where" the distribution is *located* or *centered*.
- We often use μ as our best guess for a *prediction*.
- The parameter σ talks about how *spread out* the distribution is. This gives us an indication about how *uncertain* or how *risky* our prediction is.
- If X is any random variable, the mean will be a measure of the location of the distribution and the variance will be a measure of how spread out it is.

Mean and Variance of Normal Random Variables

The Mean and Variance of a Normal

If $X \sim \mathcal{N}(\mu, \sigma^2)$ then $E(X) = \mu, \quad Var(X) = \sigma^2, \quad sd(X) = \sigma$



Two More Formulas

Let X and Y be two random variables:

- $E(aX + bY) = aE(X) + bE(Y)$
- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab \times Cov(X, Y)$
- If X and Y are independent, then $Cov(X, Y) = 0$.

We will get back to this later...

Binomial Distribution

- A binomial random variable $X \sim \text{Binomial}(n, p)$ describes the random number of success in n independent trials, each of which succeeds with probability p.
- It is a discrete random variable that takes value k with probability $P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$.
- Mean and variance:

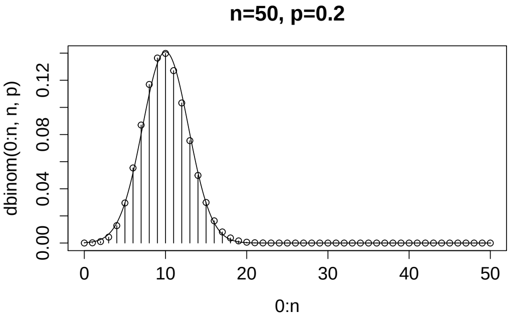
$$E(X) = np, \quad Var(X) = np(1-p)$$

- Examples:
  - number of heads in $n = 100$ coin tosses.
  - number of votes for Democrats in a survey of $n = 1000$ voters.

Binomial Distribution

- When n is sufficiently large and p is not too close to 0 or 1, $X \sim \text{Binomial}(n, p)$ can be approximated with

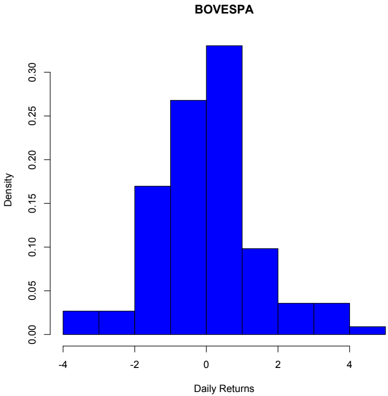
$$X \sim \mathcal{N}(np, np(1-p))$$

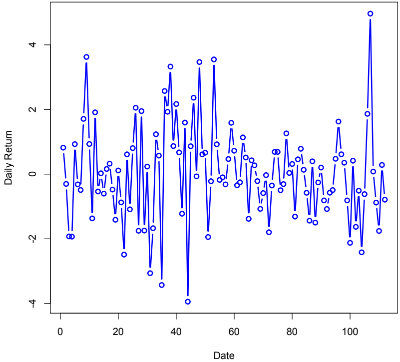


A First Modeling Exercise

- I have US$ 1,000 invested in the Brazilian stock index, the IBOVESPA. I need to predict tomorrow's value of my portfolio.
- I also want to know how risky my portfolio is, in particular, I want to know how likely am I to lose more than 3% of my money by the end of tomorrow's trading session.
- What should I do?

IBOVESPA - Data





- As a first modeling decision, let's call the random variable associated with daily returns on the IBOVESPA X and assume that returns are independent and identically distributed as

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

- Question: What are the values of μ and σ^2 ?
- We need to estimate these values from the sample in hands (n=113 observations)...

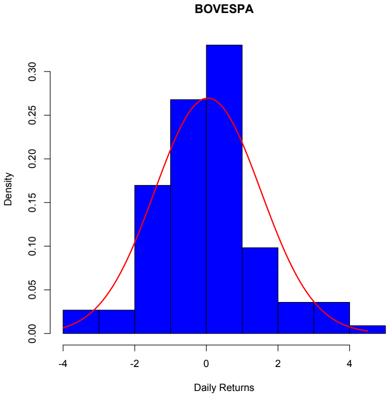
- Let's assume that the random samples $\{x_1, x_2, x_3, \ldots, x_n\}$ are *independently and identically distributed* according to the model above, i.e., $x_i \sim \mathcal{N}(\mu, \sigma^2)$
- An usual strategy is to estimate μ and σ^2, the mean and the variance of the distribution, via the sample mean ($\bar{X}$) and the sample variance (s^2)... (their sample counterparts)

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

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{X}\right)^2$$

- We will discuss more about the sample mean $\bar{X}$ and sample variance s^2 later.

For the IBOVESPA data in hands,



$\bar{X} = 0.04$ and $s^2 = 2.19$

- The red line represents our "model", i.e., the normal distribution with mean and variance given by the estimated quantities $\bar{X}$ and s^2.
- What is $Pr(X < -3)$?

Models, Parameters, Estimates...

In general we talk about unknown quantities using the language of probability... and the following steps:

- Define the random variables of interest
- Define a model (or probability distribution) that describes the behavior of the RV of interest
- Based on the data available, we estimate the parameters defining the model (Statistical Inference)
- We are now ready to describe possible scenarios, generate predictions, make decisions, evaluate risk, etc...